



Universidade Federal de Pelotas

Programa de Pós-Graduação em Epidemiologia

Doutorado em Epidemiologia



Aspectos Genéticos e Epigenéticos da Amamentação

TESE DE DOUTORADO

FERNANDO PIRES HARTWIG

Pelotas, RS

Março de 2018

UNIVERSIDADE FEDERAL DE PELOTAS
FACULDADE DE MEDICINA
PROGRAMA DE PÓS-GRADUAÇÃO EM EPIDEMIOLOGIA

ASPECTOS GENÉTICOS E EPIGENÉTICOS DA AMAMENTAÇÃO

Doutorando: Fernando Pires Hartwig

Orientador: Cesar Gomes Victora

A apresentação desta tese é exigência do
Programa de Pós-Graduação em
Epidemiologia da Universidade Federal de
Pelotas para obtenção do título de Doutor.

Pelotas, RS

Março de 2018

Universidade Federal de Pelotas / Sistema de Bibliotecas
Catalogação na Publicação

H337a Hartwig, Fernando Pires

Aspectos genéticos e epigenéticos da amamentação /
Fernando Pires Hartwig ; Cesar Gomes Victora, orientador.
— Pelotas, 2018.

301 f. : il.

Tese (Doutorado) — Programa de Pós-Graduação em
Epidemiologia, Faculdade de Medicina, Universidade
Federal de Pelotas, 2018.

1. Epidemiologia. 2. Amamentação. 3. Metilação do
DNA. 4. Polimorfismos genéticos. 5. Interação gene-
ambiente. I. Victora, Cesar Gomes, orient. II. Título.

CDD : 614.4

Tese apresentada ao Programa de Pós-Graduação em Epidemiologia da Universidade Federal de Pelotas para obtenção do título de Doutor.

Banca examinadora:

Prof. Dr. Alexandre da Costa Pereira

Universidade de São Paulo

Prof. Dr. Bernado Lessa Horta

Universidade Federal de Pelotas

Prof^a. Dr^a. Luciana Tovo Rodrigues

Universidade Federal de Pelotas

Prof. Dr. Cesar Gomes Victora (orientador)

Universidade Federal de Pelotas

Pelotas, 9 de março de 2018.

“Para ser sábio, é preciso primeiro temer a Deus, o SENHOR.” (Provérbios 1:7a)

Agradecimentos

Agradeço a todos os membros da equipe que compõe o Pós-Graduação em Epidemiologia da UFPel (PPGE). O bom-humor, a atenção e o trabalho de vocês sempre favoreceu um ambiente de trabalho mais leve, produtivo e prazeroso. Em especial, agradeço a todos os professores do PPGE por terem passado seus conhecimentos a mim, durante o Mestrado e o Doutorado, e aos meus colegas.

Agradeço a todos os colegas de Mestrado e Doutorado pelos momentos de parceria e coleguismo, bem como os de discussões e debates científicos. Foi um privilégio e uma oportunidade de aprendizado muito grande ter convivido com vocês durante estes anos.

Agradeço aos professores e funcionários do curso de graduação em Biotecnologia da UFPel pela formação que me foi concedida e que foi fundamental para chegar até este momento.

Agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelas bolsas de estudos recebidas ao longo destes 3 anos. Também agradeço a *Integrative Epidemiology Unit Medical Research Council* (Reino Unido) pelo financiamento integral referente ao Doutorado Sanduíche. Sem estes auxílios, não teria sido possível dedicar-me integralmente ao Doutorado.

Agradeço aos professores George Davey-Smith e Caroline Relton e aos pesquisadores Neil Davies e Jack Bowden da Universidade de Bristol (Reino Unido) pela orientação durante o período de Doutorado Sanduíche em Bristol. Agradeço, também, pelo constante apoio e oportunidades de colaboração. Foi trabalhando com vocês que percebi meu entusiasmo por inferência causal, campo no qual pretendo atuar e tentar, minimamente, contribuir a desenvolver no Brasil.

Agradeço a Ingrid, Simon, Tom e Laila por me terem recebido em sua casa durante o Doutorado Sanduíche. Sempre me trataram como parte da família, e de fato era assim que eu me sentia. Agradeço à Maria Carolina, Ana Luiza, Maria Clara, Esther,

Alexandra, James, Sri, Marcus, Sam, Kirsty e tantas outras amigas que tive o privilégio de construir, e que foram essenciais para que meus dias em Bristol fossem mais produtivos, leves e felizes.

Agradeço ao professor Bernardo Lessa Horta por oportunizar minha participação em diversas pesquisas do grupo EPIGEN, as quais têm acrescentado muito à minha formação. Agradeço também pela confiança em mim depositada, oportunizando-me ministrar algumas aulas em disciplinas do PPGE.

Ao professor Cesar Gomes Victora, faltam-me palavras para agradecer. Lembro-me, logo após a defesa do Mestrado (quando também tive o privilégio de ser orientado pelo professor Cesar), que o Cesar disse “Não sei se é bom para ti que eu continue como teu orientador, porque tu trabalhas em uma área que eu não domino, e fico pensando se eu conseguiria te orientar adequadamente.” Essa frase (e tantas outras frases e atitudes) revelam a humildade do Cesar, o que é muito admirável dado seu merecido reconhecimento mundial como pesquisador. O professor Cesar aceitou orientar um egresso da Biotecnologia, uma área um tanto quanto “diferente” do perfil geral dos alunos do PPGE, durante 5 anos, no Mestrado e no Doutorado. Da minha parte, só posso agradecer pela paciência, pela confiança no meu trabalho, me concedendo liberdade para organizar meu tempo e participar de outras pesquisas, e por teus conselhos sobre a carreira de pesquisador. Como orientado de um dos maiores (se não o maior) pesquisadores brasileiros, aprendi a estabelecer as prioridades certas, a ser crítico, a questionar meus próprios métodos e a manter o bom-humor, a humildade e a humanidade.

Agradeço de forma muito especial aos meus familiares, cujo apoio e amor incondicional estiveram presentes não só durante o Doutorado, mas durante toda a minha vida. Agradeço ao aporte material, psicológico, familiar e espiritual recebido dos meus pais Dari e Cynthia, e à amizade e companheirismo dos meus irmãos Marcelo e Felipe.

Agradeço, de forma póstuma, ao meu avô Udo. Da sua maneira, ele acompanhou o início da jornada que resultou nesta tese. Meu avô foi uma pessoa humilde, mas

sempre disposta a ajudar os outros, e ciente do que temos de mais valioso, que é a comunhão com Deus, algo que ele sempre ativamente apoiou.

Agradeço à Comunidade Cristo Redentor da Igreja Evangélica Luterana do Brasil por todo o apoio espiritual concedido ao longo de toda a minha vida. Agradeço à Juventude Cristo Redentor, pelos momentos de comunhão e amizade. Também agradeço à comunidade St. Matthews (Bristol), onde encontrei apoio e grandes amizades durante o Doutorado Sanduíche.

Agradeço aos demais que, direta ou indiretamente, contribuíram neste trabalho. Certamente muitos outros agradecimentos seriam cabíveis.

Agradeço principalmente a Jesus Cristo, meu Senhor, por me ter permitido chegar até este momento, e a ser uma pessoa realizada profissionalmente e feliz.

Resumo

HARTWIG, Fernando Pires. **Aspectos Genéticos e Epigenéticos da Amamentação**. 2018. Tese (Doutorado). Programa de Pós-Graduação em Epidemiologia. Universidade Federal de Pelotas (UFPel).

A amamentação traz claros benefícios à saúde e capital humano, tanto em curto quanto em longo prazo. Entre os possíveis mecanismos biológicos responsáveis por estes efeitos, estão as modificações epigenéticas – incluindo a metilação do DNA – e a presença de ácidos graxos poli-insaturados de cadeia longa (cuja sigla em inglês é LC-PUFAs) no leite materno. Nesta tese, foi investigada a relação entre amamentação e metilação do DNA da criança através de uma revisão sistemática da literatura (artigo 1) e de um estudo original, avaliando níveis de metilação do DNA em centenas de milhares de regiões ao longo do genoma (artigo 2). O papel dos LC-PUFAs na associação entre amamentação e quociente de inteligência (QI) também foi avaliada através de um estudo de interação entre polimorfismos no gene *FADS2* (que codifica uma enzima chave para a síntese endógena desses ácidos graxos) e amamentação, tendo QI como desfecho (artigo 3). Adotou-se a hipótese de adequação nutricional, de acordo com a qual o benefício da amamentação no QI seria maior nos indivíduos portadores de genótipos associados a uma menor síntese endógena de LC-PUFAs (e, portanto, mais dependentes de LC-PUFAs pré-formados). No artigo 1, verificou-se que a literatura sobre a relação entre amamentação e metilação do DNA é escassa, e os poucos estudos apresentam limitações importantes. No artigo 2, que utilizou dados de uma coorte inglesa de nascimentos, foram encontradas associações entre amamentação e níveis de metilação do DNA no sangue periférico aos 7 anos, e algumas dessas associações persistiram até a adolescência. No artigo 3, baseado em uma meta-análise *de novo* de dados publicados e não-publicados, observou-se uma maior média de QI entre os indivíduos que foram amamentados em ambos os grupos genéticos, sem evidência de interação amamentação-*FADS2*, contrariando a hipótese de adequação nutricional. Porém, análises complementares sugeriram que é possível que esta hipótese seja correta, mas que detectar a interação investigada requereria que a média de duração da amamentação nos estudos incluídos fosse maior. Os resultados dos três artigos indicam que amamentação está associada com modificações epigenéticas persistentes, e que a amamentação está positivamente associada com QI em todos os genótipos quanto aos polimorfismos estudados.

Palavras-chave: Amamentação; Metilação do DNA; *FADS2*; Polimorfismos genéticos; Interação gene-ambiente.

Abstract

HARTWIG, Fernando Pires. **Genetic and Epigenetic Aspects of Breastfeeding**. 2018. Thesis (Doctoral Thesis). Postgraduate Programme in Epidemiology. Federal University of Pelotas (UFPel).

Breastfeeding has clear short and long-term benefits to health and human capital. Possible biological mechanisms underlying these effects are epigenetic modifications – including DNA methylation – and the presence of long-chain polyunsaturated fatty acids (LC-PUFAs) in breast milk. In this, the relationship between breastfeeding and DNA methylation in the offspring was investigated through a systematic literature review (paper 1) and an original study using data on DNA methylation levels in hundreds of thousands of regions in the genome (paper 2). The role of LC-PUFAs in the association between breastfeeding and intelligence quotient (IQ) was also evaluated through the interaction between polymorphisms in the *FADS2* gene (which encodes a key enzyme for the endogenous synthesis of LC-PUFAs) and breastfeeding, with IQ being the outcome (paper 3). This analysis tested a nutritional adequacy hypothesis, which postulates that the benefits of breastfeeding on IQ are larger among carriers of genotypes associated with lower endogenous synthesis of LC-PUFAs (and therefore more dependent on pre-formed LC-PUFAs). In paper 1, it was verified that the literature on the relationship between breastfeeding and DNA methylation is scarce, and the few available studies have important limitations. In paper 2, which used data from a British birth cohort, breastfeeding was found to be associated with blood DNA methylation levels at the age of 7 years, and some of these associations persisted until adolescence. In paper 3, which was based on a *de novo* meta-analysis including both published and unpublished data, IQ was on average higher among individuals that were breastfed in both genetic groups, with no indication of the *FADS2*-breastfeeding interaction, thus arguing against the nutritional adequacy hypothesis. However, complementary analyses suggested that this hypothesis might be true, but detecting the interaction would require that the average duration of breastfeeding in the included studies were higher. Collectively, the results from these three papers indicate that breastfeeding is associated with persistent epigenetic modifications, and that breastfeeding is positively associated with IQ benefits in all genotypes with respect to the studied polymorphisms.

Keywords: Breastfeeding; DNA methylation; *FADS2*; Genetic polymorphisms; Gene-environment interaction.

Apresentação

A presente tese de Doutorado, exigência para obtenção do título de Doutor pelo Programa de Pós-Graduação em Epidemiologia, é composta pelos seguintes itens:

- 1) Projeto de Pesquisa, apresentado e defendido no dia 8 de agosto de 2016, com incorporação das sugestões dos revisores, professor Bernardo Lessa Horta e professora Luciana Tovo Rodrigues.
- 2) Relatório de atividades como analista de dados genéticos pelo projeto EPIGEN-Brasil e como pesquisador associado à Universidade de Bristol (Reino Unido).
- 3) Artigo de revisão: *Breastfeeding effects on DNA methylation in the offspring: A systematic literature review*, publicado no periódico *PLOS ONE*.
- 4) Artigo original 1: *Association between breastfeeding and DNA methylation over the life course: findings from the Avon Longitudinal Study of Parents and Children (ALSPAC)*, a ser submetido ao periódico *Scientific Reports*.
- 5) Artigo original 2: *Effect modification of FADS2 polymorphisms on the association between breastfeeding and intelligence: results from a collaborative meta-analysis*, submetido para o periódico *International Journal of Epidemiology*.
- 6) Comunicado para a imprensa.

Sumário

1 – Projeto de Pesquisa	1
2 – Relatório de atividades	113
3 – Artigo de revisão.....	130
4 – Artigo original 1	177
5 – Artigo original 2	205
6 – Comunicado para a imprensa.....	288

1 – Projeto de Pesquisa

SUMÁRIO

RESUMO	3
ARTIGOS	4
TERMOS E ABREVIATURAS	5
INTRODUÇÃO	9
EPIDEMIOLOGIA GENÉTICA E EPIDEMIOLOGIA EPIGENÉTICA	10
AMAMENTAÇÃO E EPIGENÉTICA	14
AMAMENTAÇÃO, INTELIGÊNCIA E <i>FADS2</i>	17
MODELO CONCEITUAL	21
JUSTIFICATIVA	24
OBJETIVOS	26
HIPÓTESES	26
METODOLOGIA	27
ASPECTOS ÉTICOS	34
PARTICIPAÇÃO DO DOUTORANDO NO PROJETO EPIGEN-Brasil	34
CRONOGRAMA	37
DIVULGAÇÃO DOS RESULTADOS	37
FINANCIAMENTO	38
REFERÊNCIAS	39
ANEXOS	52

RESUMO

A amamentação traz claros benefícios à saúde, tanto em curto quanto em longo prazo. Um dos possíveis mecanismos biológicos para estes efeitos seriam modificações epigenéticas, incluindo a metilação do DNA. Porém, a literatura disponível sobre o tema é escassa e nunca foi revisada sistematicamente. Ainda, estudos amplos de associação do epigenoma tendo amamentação como variável de exposição nunca foram realizados. Além disso, atualmente existe grande interesse na associação positiva entre amamentação e inteligência, suportado por uma meta-análise de estudos observacionais, uma comparação entre coortes com diferentes estruturas de confundimento e um estudo de intervenção. O leite materno é fonte de ácidos graxos poli-insaturados de cadeia longa (LC-PUFAs), relacionados ao desenvolvimento cerebral. É possível que esta associação difira conforme a capacidade da criança de sintetizar LC-PUFAs, influenciada por variantes genéticas, incluindo polimorfismos no gene *FADS2* (envolvido na síntese endógena destes ácidos graxos a partir de precursores nutricionais). Porém, os estudos que investigaram esta interação são inconsistentes. O objetivo deste projeto é investigar possíveis mecanismos biológicos dos efeitos duradouros da amamentação, incluindo: a) revisão sistemática da literatura sobre os efeitos epigenéticos da amamentação, focando na metilação do DNA; b) avaliação da associação entre amamentação e o epigenoma através de um estudo de varredura epigenômica em uma coorte inglesa, e se estas associações se mantêm ao longo do tempo; c) avaliação da interação entre variantes genéticas no gene *FADS2* e amamentação em diversas coortes, tendo inteligência cognitiva como desfecho. Os resultados auxiliarão na compreensão dos mecanismos biológicos ligando a amamentação a desfechos futuros.

ARTIGOS

Artigo 1

Breastfeeding and epigenetics: a systematic literature review.

Amamentação e epigenética: uma revisão sistemática da literatura.

Artigo 2

An epigenome-wide association study of breastfeeding in the Avon Longitudinal Study of Parents and Children.

Um estudo amplo de associação do epigenoma para amamentação no Estudo Longitudinal de Pais e Crianças de Avon.

Artigo 3

Breastfeeding x FADS2 gene interaction regarding intelligence: results from a collaborative meta-analysis.

Interação amamentação x *FADS2* quanto à inteligência: resultados de uma meta-análise colaborativa.

TERMOS E ABREVIATURAS

- Alelo: diferentes formas de uma mesma região do genoma presentes em uma população.
- ALSPAC: estudo longitudinal de pais e crianças de Avon (*Avon Longitudinal Study of Parents and Children*).
- ARIES: banco acessível para estudos integrados de epigenômica (*Accessible Resource for Integrated Epigenomic Studies*).
- DAG: gráfico acíclico direcionado (*directed acyclic graph*).
- *de novo*: no contexto do presente projeto (meta-análise *de novo*), significa que não utilizará resultados de análises prévias, mas sim resultados novos, gerados especificamente para serem incluídos no estudo.
- DHA: ácido docosa-hexaenóico (*docosahexaenoic acid*).
- DNA: ácido desoxirribonucleico (*deoxyribonucleic acid*).
- DOHaD: origens desenvolvimentistas da saúde e doença (*developmental origins of health and disease*).
- Epigenética: engloba mecanismos de regulação da expressão gênica que são transmissíveis durante o processo de divisão celular (ou seja, passam da célula mãe para as células filhas), mas que não envolvem modificação na sequência de DNA (ou seja, não são mutações).
- Epigenoma: conjunto de marcas epigenéticas presentes em um determinado tipo celular. A ciência que estuda o epigenoma se chama epigenômica. O processo de mensuração de uma região de metilação, por exemplo, se chama epigenotipagem.
- EWAS: Estudo de associação amplo do epigenoma (*Epigenome-wide association study*). Consiste em avaliar a associação entre cada marca epigenética identificada (epigenotipada) em uma varredura epigenômica e um desfecho de interesse.

- *FADS1, FADS2, FADS3*: genes que codificam as enzimas chamadas desaturase 1, desaturase 2 e desaturase 3 de ácidos graxos (*fatty acid desaturase 1, fatty acid desaturase 2, fatty acid desaturase 3*).
- Genética: estudo dos genes, variações genéticas e mecanismos de herança em organismos vivos.
- Genótipo: conjunto de alelos em uma região certa do genoma para um dado indivíduo. O processo de mensuração de um genótipo chama-se genotipagem.
- Genoma: conjunto completo do DNA de um organismo, incluindo todos os genes e outros elementos funcionais. Contém toda a informação genética necessária para o desenvolvimento e manutenção do organismo. A ciência que estuda o genoma se chama genômica.
- GWAS: Estudo de associação amplo do genoma (*Genome-wide association study*). Consiste em avaliar a associação entre cada variante genética identificada (genotipada) em uma varredura genômica e um desfecho de interesse.
- Heterogeneidade celular: diferenças entre material biológico coletado quanto às proporções de diferentes tipos celular que compõem um determinado tecido (por exemplo, diferentes amostras de sangue periférico podem apresentar diferentes proporções dos tipos celulares que compõem o sangue). Estas diferenças podem ser sistemáticas (por exemplo, caso uma variável de exposição influencie essas proporções, ou caso as amostras tenham sido coletadas em tecidos diferentes) ou aleatórias. Em estudos de epidemiologia epigenética, é importante tentar reduzir e/ou ajustar para essas diferenças (principalmente as sistemáticas), pois diferentes tipos celulares (mesmo dentro de um mesmo tecido) podem apresentar diferentes perfis epigenéticos.
- Ilhas CpG: regiões do DNA ricas em dinucleotídeos citosina (C) e guanina (G) (isto é, nucleotídeo C seguido imediatamente por um nucleotídeo G).
- Interação gene-ambiente: estudos de epidemiologia genética em que avalia-se a interação entre uma ou mais variantes genéticas e um ou mais fatores ambientais

(neste contexto, geralmente definidos como qualquer variável não-genética), tendo algum fenótipo de interesse como desfecho.

- LC-PUFAS: ácidos graxos poli-insaturados de cadeia longa (*long-chain polyunsaturated fatty acids*).
- methQTL ou mQTL: variantes genéticas associadas com níveis de metilação em uma ou mais regiões do genoma (*methylation quantitative trait locus*).
- Metiloma: conjunto de marcas epigenéticas do tipo metilação do DNA presentes em um determinado tipo celular.
- Microbioma: conjunto de microorganismos que naturalmente habitam determinado órgão de seres humanos e outros animais, muitas vezes tendo importantes papéis fisiológicos. A ciência que estuda o microbioma se chama microbiômica.
- POMC: gene que codifica o precursor proteico chamado pró-opiomelanocortina (*proopiomelanocortin*).
- PROBIT: promoção do aleitamento materno: um ensaio randomizado (*promotion of breastfeeding intervention trial*).
- QI: quociente de inteligência.
- RNA: Ácido ribonucleico (*ribonucleic acid*).
- SNP: polimorfismo de nucleotídeo único (*single nucleotide polymorphism*).
- *Splicing*: processamento do RNA mensageiro bruto, que envolve remoção de íntrons (regiões não-codificantes) e ligação dos éxons (regiões codificantes). *Splicing* alternativo refere-se a fenômeno em que, a partir de um mesmo RNA mensageiro bruto, diferentes RNA mensageiros podem ser gerados devido a diferenças no processamento (por exemplo, inclusão ou remoção de um éxon).
- Transcriptoma: conjunto de transcritos (ou seja, de moléculas de RNA), tanto codificantes (ie, que serão usados como moldes para síntese proteica) como não-codificantes (com funções estruturais, de transporte de moléculas ou de regulação

da expressão gênica), que estão presentes (em níveis variáveis) em um dado tipo celular. A ciência que estuda o transcriptoma se chama transcriptômica.

- Variante genética: região no genoma que pode apresentar diferentes alelos em uma determinada população.

INTRODUÇÃO

A amamentação beneficia a saúde da criança e da mãe [1]. Os benefícios mais imediatos da amamentação incluem a redução da morbimortalidade infantil. Estes efeitos são bem estabelecidos e ocorrem principalmente através da redução da incidência de diarreia, infecções respiratórias, otite média e má-oclusões [1]. Quanto à saúde materna, estudos epidemiológicos apontam um benefício no intervalo interpartal e no risco de câncer de mama, além de possíveis efeitos no câncer de ovário e diabetes tipo 2 [1].

A amamentação, por trazer benefícios à saúde e ser mais comum em países de renda mais baixa, pode contribuir para diminuir desigualdades em saúde [1]. Tendo em vista seus benefícios, promover a amamentação poderia diminuir custos associados a diversos problemas de saúde, particularmente porque é possível obter melhorias nas práticas de amamentação através de intervenções já disponíveis [2]. A partir de estudos avaliando qual seria a duração ideal da amamentação, as atuais recomendações orientam que o leite materno seja o único alimento oferecido até a idade de seis meses. Após este período, recomenda-se a continuação da amamentação juntamente com outros alimentos até os dois anos ou mais de idade [3, 4].

Estudos recentes demonstram associações entre amamentação e desfechos tardios, incluindo sobrepeso/obesidade, diabetes tipo 2 e quociente de inteligência (QI) [1]. Isto amplia o escopo de benefícios conferidos pela amamentação e confere maior prioridade à sua promoção, principalmente considerando que já existem exemplos bem sucedidos neste campo [5]. Além dos benefícios à saúde, os benefícios com relação ao QI, de modo particular, sugerem que intervenções visando favorecer a amamentação também podem ser vistas como investimentos em capital intelectual que terão um retorno econômico positivo [2, 6]. Também devem ser considerados os benefícios ambientais da amamentação, tendo em vista a necessidade do uso de recursos naturais e/ou energia nas etapas de produção, empacotamento, distribuição e preparação dos substitutos do leite materno, bem como os resíduos industriais produzidos ao longo destes processos [2].

Considerando os diversos benefícios da amamentação, já existe um sólido argumento para que a mesma seja facilitada e estimulada sempre que possível. Porém, há bastante relutância por parte de algumas mães e cientistas de países de alta renda em aceitar a evidência de que estas associações sejam de causa e efeito [7]. Novos estudos acerca dos mecanismos biológicos da amamentação (principalmente com relação a seus efeitos em longo prazo) podem contribuir ainda mais na compreensão dos seus efeitos, tanto com relação a associações já detectadas como também para identificar outros desfechos potencialmente influenciados pela amamentação. O presente projeto propõe avaliar alguns destes mecanismos biológicos nos contextos de epidemiologia genética e epidemiologia epigenética.

EPIDEMIOLOGIA GENÉTICA E EPIDEMIOLOGIA EPIGENÉTICA

Genética e epidemiologia genética

A genética é o campo da biologia voltado ao estudo dos genes, variações genéticas e mecanismos de herança em organismos vivos (apesar de que também é possível estudar a genética dos vírus, cuja classificação como ser vivo ou não é discutível) [8]. Gregor Mendel é, em geral, considerado o “pai” da genética moderna devido à importância de seus famosos estudos com ervilhas com base nos quais identificou que características fenotípicas eram herdadas em unidades discretas de herança e postulou as famosas “Leis de Mendel” [9]. Seus conceitos foram confirmados em estudos posteriores, onde os mecanismos biológicos responsáveis pelos fenômenos descritos por Mendel foram identificados [8].

Epidemiologia genética pode ser definida como o estudo da distribuição de variantes genéticas e dos determinantes genéticos do processo saúde-doença em populações humanas ou de animais. Seus objetivos incluem avaliar se uma ou mais variantes genéticas estão associadas com determinado fenótipo de interesse, assim como estudar a distribuição de determinantes genéticos já conhecidos, entre distintas populações ou subgrupos de uma mesma população [10].

A epidemiologia genética tem diversas aplicações. Uma delas é aumentar a compreensão dos mecanismos biológicos da determinação de doenças ou outros

fenótipos. Por exemplo, associações entre um determinado desfecho e variantes genéticas que influenciam uma determinada rota (“*pathway*”) bioquímica sugerem que esta rota tem relação com o desfecho em questão [11]. Neste sentido, destacam-se os estudos amplos de associação do genoma (GWAS, sigla para *Genome-wide association study*), nos quais são avaliadas associações entre um determinado desfecho e milhões de variantes genéticas – na sua maioria, do tipo polimorfismo de base única (SNP, sigla para *Single nucleotide polymorphism*) – distribuídas ao longo do genoma [12]. Apesar de enfrentar algumas dificuldades, particularmente a necessidade de tamanhos de amostra cada vez maiores, estes estudos têm contribuído para o aumento da compreensão das bases genéticas de fenótipos multifatoriais [13, 14].

A epidemiologia genética pode contribuir para prever a susceptibilidade genética de um indivíduo à determinada(s) doença(s), resultando eventualmente em estratégias de prevenção primária voltadas a indivíduos de alto risco, ou como ferramentas complementares de diagnóstico. Porém, tais aplicações ainda estão em fase inicial de pesquisa e requerem mais avanços para que possam ser aplicadas na prática clínica [15-17]. A epidemiologia genética pode também contribuir para a adequação de medidas terapêuticas de forma individualizada, visando maximizar a razão benefício/risco do paciente [18]. Um dos exemplos mais conhecidos é a farmacogenômica, que já apresenta aplicações, por exemplo, no tratamento em oncologia [19].

Estudos de epidemiologia genética apresentam algumas vantagens com relação a estudos observacionais sobre outros tipos de exposição. Estas vantagens incluem: a) temporalidade bem definida mesmo em estudos transversais, pois os genótipos de variantes genéticas germinativas são determinados na fecundação; b) a alta precisão das técnicas modernas de genotipagem, diminuindo problemas associados com erro de medida, que são comuns em vários campos da epidemiologia (como a epidemiologia nutricional); c) robustez contra fatores de confusão “convencionais” (tais como variáveis demográficas, socioeconômicas e comportamentais), tendo em vista a aleatoriedade do processo de alocação dos alelos nos gametas durante a meiose [20-22].

Por outro lado, a epidemiologia genética tem suas limitações, incluindo confundimento introduzido por estratificação da população (ou seja, associação da etnia tanto com as frequências genotípicas como também com a doença, independentemente da variante genética em questão) e/ou por parentesco entre os indivíduos da amostra no caso de estudos de base populacional ou casos e controles (ou seja, estudos que não são em famílias). No entanto, estas limitações são bem conhecidas, e existem métodos não somente para corrigir estes vieses, mas também para explorar a existência de estratificação populacional e/ou a estrutura familiar da amostra de forma a beneficiar a análise [23-25]. Outra limitação comum é o baixo poder estatístico, principalmente em estudos de associação entre variantes genéticas comuns e fenótipos multifatoriais [26].

Outra aplicação da genética na epidemiologia é o uso de variantes genéticas como variáveis instrumentais para aumentar a robustez da inferência causal em estudos observacionais. Esta estratégia, conhecida como randomização mendeliana, visa obter uma estimativa do efeito causal de uma exposição modificável em um desfecho de interesse. Diversas estratégias para fortalecer a validade e robustez das estimativas obtidas através de randomização mendeliana vêm sendo propostas [27, 28].

Assim como a randomização mendeliana, estudos de interação gene-ambiente representam uma importante contribuição da epidemiologia genética para identificar mecanismos causais de doenças, sob a perspectiva de intervenção [22]. Por exemplo, se a associação entre o consumo de determinado alimento e o desfecho estudado é dependente da presença de um nutriente específico, seria esperado que a magnitude da associação fosse maior em pessoas com menor capacidade de síntese endógena deste nutriente, quando comparadas a pessoas com maior capacidade. Esta interação pode ser avaliada utilizando-se marcadores genéticos da capacidade endógena de sintetizar o nutriente em questão [22].

Epigenética e epidemiologia epigenética

A epigenética se refere a uma série de mecanismos de regulação da expressão gênica que se caracterizam por serem mitoticamente herdáveis, ou seja, passam da célula-mãe para a célula-filha durante o processo de mitose. Estes mecanismos não envolvem

mudanças na sequência de DNA. Atualmente, os mecanismos epigenéticos mais estudados são: metilação do DNA, acetilação de um conjunto de proteínas chamado histonas (conjunto de proteínas envolvidas na organização estrutural do DNA) e ação dos RNAs não-codificantes. O foco do presente projeto será na metilação do DNA, definida como a adição de um radical metil ($-\text{CH}_3$) na posição 5' de uma base nitrogenada do tipo citosina, tipicamente dentro das chamadas ilhas CpG, ou regiões do DNA ricas neste par de nucleotídeos. Este processo epigenético se dá através de uma ligação covalente, que, uma vez que tenha ocorrido, é relativamente estável ao longo do tempo [29-31]. Isto permite com que padrões de metilação do DNA estabelecidos durante o desenvolvimento embrionário (período em que o metiloma apresenta elevada plasticidade) perdurem. Porém, o metiloma ainda apresenta plasticidade após o nascimento, de modo que pode ser influenciado por fatores ambientais diversos (tais como o tabagismo[32]) [33], além de grande variabilidade entre diferentes tecidos em um mesmo indivíduo [34].

Epidemiologia epigenética pode ser definida como o estudo da distribuição de perfis epigenéticos e dos determinantes epigenéticos do processo saúde-doença em populações humanas ou animais. Atualmente, tais estudos buscam avaliar se variantes epigenéticas estão associadas com determinada variável de interesse, seja como exposição – buscando compreender o papel da epigenética em desfechos em saúde – ou como desfecho – avaliando o papel de determinadas exposições (podendo incluir variantes genéticas) no epigenoma [35, 36].

De forma similar à epidemiologia genética, a epidemiologia epigenética também visa aumentar a compreensão acerca dos mecanismos biológicos relacionados com a variabilidade fenotípica. De fato, as marcas epigenéticas, por não serem completamente fixadas ao longo do tempo, devem ser tratadas como fenótipos em um estudo epidemiológico. Assim, pode-se estudar tanto as causas como as consequências das modificações epigenéticas em populações [35-37]. Isto levanta a possibilidade de investigar os mecanismos epigenéticos como potenciais mediadores, por exemplo, dos efeitos de exposições precoces em desfechos mais tardios, conforme discutido na seção “AMAMENTAÇÃO E EPIGENÉTICA”.

Devido à natureza fenotípica do epigenoma, a epidemiologia epigenética constitui um campo da epidemiologia observacional convencional, e não uma subdivisão dentro da epidemiologia genética. Assim, as vantagens da epidemiologia genética em termos de evitar alguns tipos de viés (por exemplo, causalidade reversa e confundimento residual) não se aplicam à epidemiologia epigenética [35, 38]. Estas limitações não afetam as aplicações do “arquivo biossocial”, as quais são derivadas de simples associações, não necessariamente oriundas de relações de causa-efeito. No entanto, estes vieses afetam a inferência causal em estudos de epidemiologia epigenética.

AMAMENTAÇÃO E EPIGENÉTICA

Como mencionado na Introdução, existem evidências de que a amamentação tenha efeitos em longo prazo, apresentando associações com desfechos que se estabelecem após o desmame. Porém, muitas vezes não é possível separar os efeitos biológicos (devidos a componentes do leite materno) daqueles devidos à interação mãe-criança.

Em um estudo de intervenção britânico iniciado em 1982 por Lucas e colaboradores [39], crianças nascidas pré-termo foram aleatoriamente alocadas para receberem leite humano (de doadoras não-aparentadas) ou fórmula para pré-termos. Isto permitiu isolar o efeito biológico do leite humano dos demais efeitos da amamentação, tais como o vínculo emocional entre mãe e criança. Neste estudo, crianças que receberam leite humano apresentaram melhorias no perfil de risco cardiovascular aos 13-16 anos de idade quando comparadas com o grupo que recebeu fórmula, incluindo melhor perfil lipoprotéico [40] e menor pressão arterial [41]. Recentemente, este mesmo estudo observou melhorias na morfologia e função cardíaca no início da idade adulta no grupo que recebeu leite humano quando criança [42].

Estes resultados suportam a existência de efeitos biológicos causais e duradouros da amamentação, colocando-a como um fator importante no contexto das origens desenvolvimentistas da saúde e da doença (DOHaD, sigla para *Developmental Origins of Health and Disease*). Devido à lacuna de tempo entre a amamentação e o desfecho, postula-se que o efeito causal da amamentação requeira um mecanismo de “memória biológica”. Ou seja, é necessário que a amamentação cause modificações no organismo que se mantêm ao longo do tempo e que influenciam a ocorrência de um

desfecho anos após o desmame. Modificações epigenéticas constituem um possível mecanismo para explicar estes efeitos, e têm recebido grande atenção no contexto de DOHaD [43-45]. Por exemplo, tabagismo materno na gestação foi associado a modificações epigenéticas persistentes até, pelo menos, a adolescência [46].

Os potenciais efeitos epigenéticos da amamentação foram pouco estudados. A única revisão da literatura sobre este tema foi realizada em 2014 [47]. Esta revisão narrativa, não sistemática, não relatou nenhum estudo comparando crianças (ou animais) que receberam ou não leite materno em termos de marcadores epigenéticos. Os autores levantam a possibilidade da presença de efeitos epigenéticos para explicar associações entre amamentação e desfechos tardios ou expressão gênica, sem que mecanismos epigenéticos tenham sido avaliados diretamente. Os autores também descrevem a presença no leite materno de substâncias com possíveis efeitos epigenéticos identificados, por exemplo, em estudos *in vitro*.

Outros autores discutem o potencial papel da microbiota como mediadora da associação entre amamentação e desfechos tardios [48, 49]. De fato, o leite materno é fonte de uma microbiota particular [50], e a flora gastrointestinal difere entre crianças que foram amamentadas quando comparadas a crianças que não foram amamentadas [51, 52]. Embora alguns autores se refiram à relação entre amamentação e o microbioma como um efeito epigenético da amamentação [48], isso é conceitualmente inadequado tendo em vista a definição de epigenética. Por outro lado, outros argumentam que os efeitos da amamentação sobre a microbiota podem também promover modificações epigenéticas, conectando microbiômica e epigenômica [49].

O presente projeto inclui a proposta de uma revisão sistemática da literatura sobre os efeitos epigenéticos da amamentação. Esta revisão atualmente está em fase inicial e será completada dentro do próximo ano. Alguns artigos relevantes, já identificados, são brevemente descritos a seguir.

Um estudo holandês envolveu 120 pares mães-crianças (50 meninas, idade média de 17 meses) detectou que a duração da amamentação reduziu a metilação no gene que codifica a leptina – um hormônio relacionado à saciedade – em DNA extraído do

sangue periférico [53]. Isso poderia implicar em níveis circulantes mais altos de leptina em indivíduos amamentados, explicando a associação inversa entre amamentação e obesidade evidenciada por estudos epidemiológicos [54].

Efeitos epigenéticos da amamentação também foram descritos em outros contextos. Por exemplo, em 639 mulheres norte-americanas com câncer de mama, a probabilidade de o promotor do gene que codifica a proteína p16 (um importante supressor tumoral) estar metilado no tecido tumoral foi maior em mulheres que não haviam sido amamentadas. No entanto, esta associação só foi observada em mulheres pré-menopausa [55]. Em um estudo sobre asma em 245 adolescentes do sexo feminino, observou-se uma interação entre variantes genéticas e amamentação na metilação (em DNA extraído do sangue periférico) de ilhas CpG localizadas na região 17q12, sugerindo que os potenciais efeitos epigenéticos da amamentação incluem modulação dos efeitos epigenéticos de variantes genéticas [56]. Por fim, observou-se que amamentação também pode estar relacionada com perfis “globais” de metilação (também no sangue), estimados a partir de uma análise de redução de dimensionalidade chamada quadrados mínimos parciais, em crianças tchecas (n=200) de 7 a 15 anos de idade [57]. Dos quatro estudos, este foi o único a realizar epigenotipagem ao longo do metiloma, enquanto que os outros utilizaram a abordagem de uma determinada região candidata.

Estes estudos, embora esparsos e abordando temas distintos, sugerem que a amamentação pode apresentar efeitos epigenéticos sobre distintos sistemas do organismo humano. A revisão sistemática proposta como parte do presente projeto irá trazer maiores aportes sobre este tópico.

AMAMENTAÇÃO, INTELIGÊNCIA E *FADS2*

Como mencionado anteriormente, existem evidências de que a amamentação estaria positivamente associada com inteligência. Uma meta-análise recente demonstrou que esta associação persiste em indivíduos de 10-19 anos de idade [58]. Outro estudo recente detectou uma associação entre amamentação e QI aos 30 anos de idade [59]. Na meta-análise em questão, não foi detectado viés de publicação. Porém, estudos que não ajustaram para QI materno apresentaram uma estimativa meta-analítica de 4,1 pontos de QI, comparado com 2,6 para estudos ajustados. Isso reforça a importância de ajustar para QI materno ao estudar a associação entre amamentação e QI da criança, além da possibilidade de confusão residual caso o QI materno ou outros fatores de confusão importantes (como posição socioeconômica) não sejam bem medidos ou modelados nas análises.

Por outro lado, no único estudo de intervenção randomizado neste assunto, o estudo PROBIT (sigla para *Promotion of Breastfeeding Intervention Trial*), o QI aos 6,5 anos de idade foi, em média, 5,9 pontos maior nas crianças alocadas ao grupo que recebeu promoção da amamentação comparando com o grupo controle [60]. Além deste estudo, a associação entre amamentação e QI foi consistente entre as coortes de nascidos vivos em Pelotas no ano de 1993 [61, 62] e no Estudo Longitudinal de Pais e Crianças de Avon (ALSPAC, sigla para *Avon Longitudinal Study of Parents and Children*) [63, 64]; em Pelotas, a amamentação não esteve associada com posição socioeconômica, enquanto que no estudo ALSPAC esta associação estava presente [65]. Outro argumento a favor de uma associação causal é a ausência de diferenças no QI de crianças cujas mães tentaram, mas não conseguiram amamentar, e aquelas cujas mães optaram inicialmente pelo uso de fórmula [66].

Um possível mecanismo biológico dos efeitos da amamentação na inteligência é a presença no leite materno de LC-PUFAs, tais como o ácido docosa-hexaenóico (DHA, sigla para *docosahexaenoic acid*), o qual faz parte da família de ácidos graxos ômega-3 [67]. O DHA é um importante componente da membrana de células do sistema nervoso central e da retina [68, 69]. Estudos em animais e seres humanos sugerem que níveis adequados de DHA influenciam o desenvolvimento cognitivo de várias formas,

incluindo biogênese de membranas celulares, manutenção da fluidez celular, neurogênese, neurotransmissão e proteção contra estresse oxidativo [69, 70].

Estudos de intervenção randomizados já foram realizados para avaliar o efeito causal do DHA e/ou outros LC-PUFAs em indicadores de função cognitiva. Uma revisão sistemática e meta-análise de estudos de intervenção, realizada por Jiao e colaboradores [71], as estimativas agrupadas sugerem que a suplementação de PUFAs do tipo ômega-3 melhora o desenvolvimento cognitivo, com relação a todas as medidas avaliadas, em crianças de até quatro anos de idade (sete estudos, totalizando 567 indivíduos no grupo tratamento e 464 no grupo controle). Porém, foram detectadas associações com nenhum dos quatro domínios cognitivos estudados (memória, funções executivas, atenção e velocidade de processamento), além de desfechos secundários, em crianças com mais de quatro anos de idade ou adultos (15 estudos, com um total de 1517 crianças e 3657 adultos).

Qawasmi e colaboradores, através de uma revisão sistemática e meta-análise de estudos de intervenção (totalizando 19 estudos e 1949 crianças de até 1 ano de idade), detectaram que a suplementação de fórmulas substitutivas do leite materno com LC-PUFAs melhorou a acuidade visual [72]. Em conjunto, os estudos de Jiao [71] e Qawasmi [72] apontam que a infância é o período mais importante para os efeitos da suplementação com LC-PUFAs.

Além das fontes nutricionais, os níveis dos LC-PUFAs também dependem de fatores que influenciam sua síntese endógena, tais como variantes genéticas. Estudos de gene candidato e de GWAS identificaram SNPs que influenciam este processo metabólico na região 11q12-11q13.1, mais especificamente onde se localiza um *cluster* de genes da família *FADS* (*FADS1*, *FADS2* e *FADS3*) [73, 74]. Nestes estudos, foram observadas associações negativas entre os alelos menos frequentes (comparando com alelos mais frequentes) em diferentes SNPs do gene *FADS2* e menores níveis de PUFAs em fosfolipídeos no plasma e na membrana celular de eritrócitos [73, 74]. Estes genes codificam enzimas que atuam na dessaturação de ácidos graxos, um importante passo na síntese de LC-PUFAs. A enzima delta-6 dessaturase, codificada pelo gene *FADS2*, catalisa uma reação de conversão do ácido tetracosa-pentaenóico (24:5(n-3)) para o

ácido ácido tetracosahexaenóico (24:6(n-6)), que por sua vez é convertido em DHA através de uma reação de beta-oxidação (Figura 1) [75, 76]. A delta-6 dessaturase atua de forma similar na via metabólica dos ácidos graxos do tipo ômega 6, culminando na síntese do ácido docosa-pentaenóico (22:5(n-6)) (Figura 1) [75, 76]. A atividade das enzimas dessaturases delta-5 (codificada pelo gene *FADS1*) e delta-6 é considerada um fator chave na síntese endógena de LC-PUFAs [75, 76].

Assumindo que a associação entre amamentação e inteligência se deva, ao menos parcialmente, pelo fato do leite materno ser fonte de LC-PUFAs pré-formados tais como o DHA, seria plausível existir uma interação entre variantes genéticas nos genes da família *FADS* e amamentação. Mais especificamente, se esperaria um efeito mais evidente da amamentação sobre a inteligência em indivíduos portadores de genótipos associados a uma menor síntese endógena de LC-PUFAs. Estes indivíduos seriam mais dependentes do DHA e outros LC-PUFAs presentes no leite materno para atingir os níveis necessários destes ácidos graxos para o desenvolvimento cognitivo adequado [77]. Esta hipótese pressupõe que, uma vez que os requerimentos nutricionais de LC-PUFAs são atingidos, aumentos adicionais não conferem benefícios [77]. Assim, seria possível investigar a importância dos LC-PUFAs na associação entre amamentação e inteligência a partir de uma análise de interação *FADS*-amamentação.

No primeiro estudo a avaliar esta interação gene-ambiente, Caspi e colaboradores [78] avaliaram dois SNPs no gene *FADS2*: rs174575 (alelo mais frequente C e menos frequente G) e rs1535 (alelo mais frequente A e menos frequente G). Assim, de acordo com a hipótese dos LC-PUFAs, seria esperado que a associação entre amamentação e inteligência fosse mais evidente em indivíduos portadores do alelo G. Ao contrário do que seria esperado, não foi evidenciada associação entre amamentação e QI aos 5-6 anos de idade em indivíduos homozigotos para o alelo G do SNP rs174575, mas a associação esteve presente em indivíduos CC ou CG. O resultado foi similar em duas amostras independentes (n=858 e n=1848) para o SNP rs174575, mas não para o SNP rs1535 [78].

Os resultados de Caspi e colaboradores não foram replicados em estudos subsequentes. O estudo ALSPAC utilizou QI aos 8 anos como desfecho e apresentou as

The diagram illustrates the metabolic pathways of eicosanoids, starting from the n-6 and n-3 pathways. The n-6 pathway begins with C18:2 n-6 (LA), which is converted to C18:3 n-6 (GLA) by D6D (FADS2). GLA is then converted to C20:3 n-6 (DGLA) by Elongase-5 (ELOVL5). DGLA can be converted to C20:4 n-6 (AA) by D5D (FADS1), which is then converted to C22:4 n-6 (DTA) by Elongase-5 (ELOVL5). DTA is converted to C22:5 n-6 by β-Oxidation, which is then converted to C24:5 n-6 by D6D (FADS2). C24:5 n-6 is converted to C24:6 n-3 by β-Oxidation. The n-3 pathway begins with C18:3 n-3 (ALA), which is converted to C18:4 n-3 (SA) by D6D (FADS2). SA is then converted to C20:4 n-3 (ETA) by Elongase-5 (ELOVL5). ETA is converted to C20:5 n-3 (EPA) by D5D (FADS1), which is then converted to C22:5 n-3 (DPA) by Elongase-5 (ELOVL5). DPA is converted to C22:6 n-3 (DHA) by β-Oxidation, which is then converted to C24:6 n-3 by D6D (FADS2). C24:6 n-3 is converted to C24:5 n-3 by β-Oxidation. The diagram also shows the conversion of C20:2 n-6 to C20:3 n-6 (DGLA) by D8D (FADS2) and C20:2 n-3 to C20:4 n-3 (ETA) by D8D (FADS2). The final products are categorized into Anti-inflammatory Eicosanoids (DGLA, ETA, EPA, DHA), Pro or anti-inflammatory Eicosanoids (AA, DTA), and Less or anti-inflammatory Eicosanoids (C24:5 n-3, C24:6 n-3).

```

graph TD
    subgraph n6_Pathway [n-6 Pathway]
        LA[C18:2 n-6 LA] -->|D6D (FADS2)| GLA[C18:3 n-6 GLA]
        GLA -->|Elongase-5 (ELOVL5)| DGLA[C20:3 n-6 DGLA]
        DGLA -->|D5D (FADS1)| AA[C20:4 n-6 AA]
        AA -->|Elongase-5 (ELOVL5)| DTA[C22:4 n-6 DTA]
        DTA -.->|β-Oxidation| C22_5_n6[C22:5 n-6]
        C22_5_n6 -->|D6D (FADS2)| C24_5_n6[C24:5 n-6]
    end
    subgraph n3_Pathway [n-3 Pathway]
        ALA[C18:3 n-3 ALA] -->|D6D (FADS2)| SA[C18:4 n-3 SA]
        SA -->|Elongase-5 (ELOVL5)| ETA[C20:4 n-3 ETA]
        ETA -->|D5D (FADS1)| EPA[C20:5 n-3 EPA]
        EPA -->|Elongase-5 (ELOVL5)| DPA[C22:5 n-3 DPA]
        DPA -.->|β-Oxidation| DHA[C22:6 n-3 DHA]
        DHA -->|D6D (FADS2)| C24_6_n3[C24:6 n-3]
        C24_6_n3 -->|β-Oxidation| C24_5_n3[C24:5 n-3]
    end
    C24_5_n3 --> Anti_inflammatory_Docosanoids[Anti-inflammatory Docosanoids]

    LA -->|Elongation| C20_2_n6[C20:2 n-6]
    C20_2_n6 -->|D8D (FADS2)| DGLA
    ALA -->|Elongation| C20_2_n3[C20:2 n-3]
    C20_2_n3 -->|D8D (FADS2)| ETA

    DGLA --> Anti_inflammatory_Eicosanoids[Anti-inflammatory Eicosanoids]
    AA --> Pro_or_Anti_inflammatory_Eicosanoids[Pro or anti-inflammatory Eicosanoids]
    ETA --> Less_or_Anti_inflammatory_Eicosanoids[Less or anti-inflammatory Eicosanoids]
    EPA --> Less_or_Anti_inflammatory_Eicosanoids
    DHA --> Anti_inflammatory_Docosanoids
  
```


MODELO CONCEITUAL

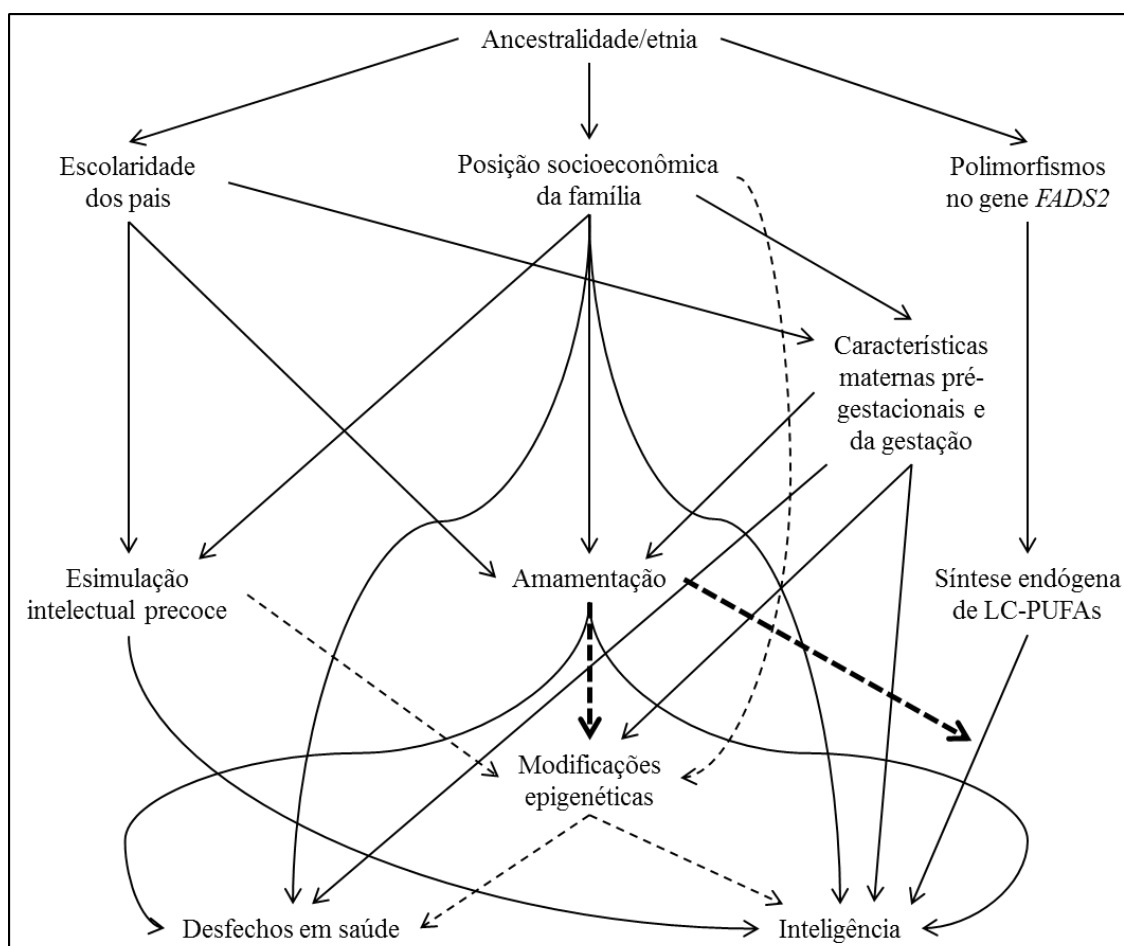
Com base no exposto acima, elaborou-se um modelo conceitual (Figura 2) que orientará as análises propostas neste projeto e a interpretação dos resultados. O modelo foi construído na forma de um DAG (*directed acyclic graph*, ou gráfico acíclico direcionado) para traduzir as relações sendo postuladas em estratégias de análise [85].

A variável mais distal (e exógena no DAG) é ancestralidade/etnia. Postula-se que esta variável teria efeitos causais diretos nas variáveis socioeconômicas e genéticas. Amplas evidências históricas e sociológicas confirmam que a ancestralidade/etnia é um importante determinante da posição socioeconômica na sociedade brasileira [86, 87] e em outros países [88, 89]. A não ser através do confundimento pela ancestralidade/etnia, é conceitualmente improvável que o genótipo de um indivíduo esteja causalmente associado com variáveis socioeconômicas na época do seu nascimento.

Em um segundo nível hierárquico, foram posicionadas variáveis socioeconômicas precoces, bem como polimorfismos genéticos no gene *FADS2*. A escolaridade dos pais foi separada da posição socioeconômica da família por poder ter efeitos independentes tanto na estimulação da criança quanto na amamentação, relacionados, por exemplo, com conhecimento de formas mais adequadas de estimulação intelectual e da importância da amamentação. A posição socioeconômica da família também foi apontada como tendo efeitos causais diretos sobre estimulação intelectual precoce [90] e amamentação. O sentido da associação entre posição socioeconômica e duração da amamentação varia conforme o nível de riqueza do país, sendo direta em países ricos e inversa em países pobres [1]. A posição socioeconômica também está associada com inteligência [91-94] e em desfechos em saúde de forma geral [95, 96]. Tanto a posição socioeconômica quanto a escolaridade dos pais também estão associados com características maternas pré-gestacionais e características da gestação [97-100]. Os polimorfismos genéticos apresentam, no DAG, efeitos causais diretos na síntese endógena de LC-PUFAs, demonstrados por estudos de epidemiologia genética [73, 74].

No terceiro nível, foram posicionadas as variáveis referentes a características maternas pré-gestacionais (tais como o índice de massa corporal e paridade) e da gestação (tais como fumo materno na gestação, tipo de parto e peso ao nascer). Existem evidências de que estas variáveis podem influenciar níveis de amamentação [101-107] e marcas epigenéticas [46, 108-113]. Logo, são potenciais variáveis de confusão da associação entre amamentação e traços epigenéticos. Também foram postulados efeitos das características maternas e da gestação em desfechos em saúde e inteligência [99, 114-117].

Figura 2. Modelo conceitual para a relação entre amamentação e modificações epigenéticas e para a interação entre amamentação e polimorfismos no gene *FADS2*.



No quarto nível, foram posicionadas as variáveis de estimulação intelectual precoce, amamentação e síntese endógena de LC-PUFAs. Postula-se que a estimulação tem efeitos causais diretos na inteligência [90] e possivelmente sobre alterações epigenéticas, de forma que a estimulação seria um mediador dos efeitos das variáveis socioeconômicas, assim como da ancestralidade/etnia. Também foram postulados efeitos diretos da amamentação em desfechos em saúde e na inteligência [1]. Cabe ressaltar que os efeitos epigenéticos da amamentação foram pouco estudados até o momento, tendo sido hipotetizados no modelo conceitual por serem um dos objetos de estudo do presente projeto. O mesmo se aplica à estimulação intelectual precoce, tendo seus efeitos epigenéticos sido postulados para evitar a exclusão de um potencial caminho enviesante da associação entre amamentação e epigenética.

A síntese endógena de LC-PUFAs foi considerada como tendo efeitos causais diretos na inteligência devido aos potenciais efeitos do DHA e outros ácidos graxos no desenvolvimento cognitivo [69]. Assumiu-se, ainda, que a amamentação modificaria o efeito da síntese endógena dos LC-PUFAs na inteligência [77, 78]. De acordo com resultados do maior estudo já publicado sobre o assunto [77], o efeito dos polimorfismos genéticos no gene *FADS2* é aparente apenas em indivíduos que não foram amamentados, concordando com a hipótese que a amamentação supre os requerimentos de LC-PUFAs para um desenvolvimento cognitivo adequado mesmo em crianças com menor síntese endógena destes ácidos graxos. Porém, outros estudos encontraram resultados diferentes [78, 80-82], de forma que esta interação é um dos objetos de estudo deste projeto.

Por fim, assumiu-se que modificações epigenéticas têm efeitos causais tanto em desfechos em saúde quanto na inteligência. Assim, estas modificações seriam potenciais mediadores dos efeitos em longo prazo da amamentação, bem como da estimulação e da posição socioeconômica familiar, e dos determinantes distais destas variáveis. Apesar do potencial da epigenética na elucidação dos mecanismos biológicos de fenótipos multifatoriais [35], foram realizados poucos estudos longitudinais robustos avaliando mecanismos epigenéticos, principalmente em nível de epigenoma completo.

JUSTIFICATIVA

Apesar dos aparentes efeitos em longo prazo da amamentação, seus potenciais mecanismos biológicos têm sido, até o momento, pouco estudados. A epidemiologia demonstra que não é obrigatório conhecer tais mecanismos para que uma associação possa ser considerada como de causa e efeito. Porém, como a maioria das evidências sobre efeitos da amamentação advém de estudos observacionais, elucidar os mecanismos biológicos aumenta a plausibilidade biológica dos achados epidemiológicos.

Conforme mencionado na seção “AMAMENTAÇÃO E EPIGENÉTICA”, mecanismos epigenéticos vêm sendo propostos como potenciais mediadores de associações entre exposições na infância e desfechos mais tardios, inclusive com relação aos efeitos da amamentação [1, 48, 118]. Algumas revisões narrativas sobre o tema sugerem efeitos epigenéticos da amamentação com base em evidências indiretas (discutido na seção “AMAMENTAÇÃO E EPIGENÉTICA”). No entanto, nenhuma revisão sistemática da literatura sobre o assunto foi realizada até o momento, de forma que conclusões mais sólidas sobre os potenciais efeitos epigenéticos da amamentação não são possíveis. Esta constatação cria a oportunidade de realizar uma revisão sistemática sobre o tema (Artigo 1).

Ainda dentro da epidemiologia epigenética, estudos amplos de associação do epigenoma (EWAS: *Epigenome-wide association study*) [119] constituiriam uma alternativa para avaliar a associação entre amamentação e vários fatores epigenéticos simultaneamente. Tais estudos podem fornecer evidências sobre uma ampla variedade de possíveis efeitos epigenéticos, os quais poderiam ser confirmados através de estudos de replicação e em estudos que utilizem estratégias mais robustas de inferência causal. Tais achados podem, ainda, gerar hipóteses sobre efeitos da amamentação em desfechos até agora não estudados. Estas análises formarão o segundo artigo a ser proposto.

A identificação de componentes do leite materno responsáveis por determinados benefícios da amamentação favoreceria a elaboração de fórmulas que o substituam de forma mais adequada. Tal substituição é necessária, por exemplo, quando a

amamentação não é possível devido a questões biológicas e inexitem na comunidade bancos de leite humano. No entanto, a simples adição de componentes nutricionais presentes no leite materno não necessariamente mimetiza seus efeitos biológicos [120], pois os benefícios do leite humano resultam de um complexo equilíbrio entre seus vários componentes [1].

Um exemplo é a adição de LC-PUFAs a fórmulas industrializadas, devido ao papel creditado a estes nutrientes na associação entre amamentação e desenvolvimento do sistema nervoso [120, 121]. No entanto, tem sido postulado que a adição de LC-PUFAs não beneficia a todas as crianças. Estudos de interação entre variantes genéticas no gene *FADS2* e amamentação, tendo QI como desfecho, têm apresentado resultados inconsistentes, conforme descrito na seção “AMAMENTAÇÃO, INTELIGÊNCIA E *FADS2*”. Tal inconsistência reforça a importância da replicação de estudos já publicados para se obter evidências mais robustas [122].

Conclusões mais robustas acerca desta interação podem ser obtidas pela combinação de resultados de diferentes estudos, o que ampliaria o tamanho de amostra total e reduziria a probabilidade de achados devidos ao acaso. A questão do poder estatístico é especialmente importante em análises de interação, as quais requerem amostras muito grandes. Análises *de novo* seguindo um plano de análise determinado *a priori* seriam de grande valia tanto para harmonizar os estudos (minimizando a diluição das estimativas de efeito) quanto para diminuir a possibilidade de que associações detectadas em análises exploratórias não-planejadas sejam publicadas como se tivessem sido definidas *a priori*. Estas análises consistirão no terceiro artigo da tese.

OBJETIVOS

Objetivo geral

Investigar possíveis mecanismos biológicos dos efeitos em longo prazo da amamentação.

Objetivos específicos

- 1) Fazer uma revisão sistemática da literatura sobre os efeitos epigenéticos da amamentação.
- 2) Avaliar a associação entre amamentação e mais de 450 mil sítios de metilação ao longo do genoma, e determinar se estas associações mantêm ao longo do tempo.
- 3) Avaliar se existe uma interação entre variantes genéticas no gene *FADS2* e amamentação, tendo inteligência como desfecho.

HIPÓTESES

- 1) A literatura é relativamente escassa, mas indica que a amamentação pode ter efeitos epigenéticos.
- 2) Amamentação é associada com os níveis de metilação em alguns sítios do genoma, mesmo após ajuste para confundidores; esta associação se mantém ao longo do tempo.
- 3) Existe uma interação entre variantes genéticas no gene *FADS2* e amamentação, sendo que o benefício da amamentação na inteligência é maior em grupos com genótipos associados à menor síntese endógena de LC-PUFAs.

METODOLOGIA

A seguir, descreveremos aspectos metodológicos dos três artigos a serem publicados para constituir a tese proposta.

Revisão sistemática (Artigo 1)

Uma revisão sistemática da literatura será utilizada para identificar os estudos existentes sobre os potenciais efeitos epigenéticos da amamentação. Será utilizada a ferramenta Ovid (<https://ovidsp.tx.ovid.com/>), que realiza buscas nos seguintes bancos de dados: MEDLINE, Embase, Allied and Complementary Medicine Database, CAB ABSTRACTS, PsycINFO® e Philosopher's Index. Os campos nos quais a busca será realizada serão os utilizados de forma padrão pelo Ovid: título; título original; comentário sobre o título; resumo, palavra contida no *Subject Heading*, *MeSH Subject Headings*, palavras-chave, conceitos-chave, texto completo e outros. Serão levadas em contas as especificidades dos bancos de dados.

Os descritores são mostrados na Tabela 1. A busca será realizada na forma “AMAMENTAÇÃO AND EPIGENÉTICA”, ou seja, limitada a artigos que contêm pelo menos um descritor presente em cada tópico. Após a obtenção da lista de artigos e remoção de duplicatas, a revisão será realizada independentemente por dois avaliadores. Inicialmente será feita uma triagem pela leitura dos títulos e resumos, seguida da leitura na íntegra dos artigos pré-selecionados. Discordâncias serão resolvidas por consenso entre os avaliadores. Espera-se que a literatura no tema seja demasiadamente escassa e heterogênea para permitir uma meta-análise. Neste caso, os artigos identificados serão apresentados e discutidos narrativamente.

Serão incluídos tanto estudos em seres humanos como em animais de experimentação, sem restrição quanto a seus delineamentos e quanto ao tecido ou tipo celular utilizado como fonte de DNA. Não será estabelecido um critério de exclusão *a priori* quanto ao idioma, porém é possível que este seja um aspecto restritivo devido às próprias características dos bancos de dados. Serão excluídos estudos que: avaliaram apenas componentes do leite materno de forma isolada; não avaliaram modificações epigenéticas, mesmo que tenham estudado expressão gênica

ou perfis microbiômicos; não trazem dados novos, tais como artigos de revisão ou editoriais. Não obstante, artigos de revisão relevantes poderão ser utilizados para busca de artigos originais nas suas listas de referências.

Tabela 1. Descritores para busca por artigos sobre amamentação e epigenética utilizando a ferramenta Ovid.

Tópico	Descritores^a
AMAMENTAÇÃO	("breastfe\$" OR "breast fe\$" OR "bottle fe\$" OR "formula fe\$" OR "infant feeding" OR "human milk" OR "breast milk" OR "formula milk" OR "weaning")
EPIGENÉTICA	("epigenetic\$" OR "epigenom\$" OR "methyлат\$" OR "methQTL" OR "mQTL")

^aO símbolo "\$" ao final dos descritores informa ao OVID que todas as palavras formadas pela adição de qualquer combinação de qualquer número de caracteres (incluindo nenhum), estão sendo buscados. Por exemplo, "breastfe\$" engloba termos como "breastfe", "breastfeeding" e "breastfed".

Estudo amplo de associação do epigenoma (Artigo 2)

A associação entre amamentação e marcadores epigenéticos será avaliada através de um EWAS. Resumidamente, a técnica consiste em avaliar a associação entre uma variável de exposição de interesse e diversos sítios de metilação, ou seja, regiões do genoma que apresentam metilação variável entre indivíduos e/ou entre tecidos. Como cada sítio de metilação é analisado individualmente como variável dependente, o número total de testes estatísticos realizados é igual ao número de sítios de metilação multiplicado pelo número de modelos brutos e ajustados que se deseja testar.

Serão utilizados dados do estudo ARIES (*Accessible Resource for Integrated Epigenomic Studies*) [123], que contém informações epigenéticas da coorte ALSPAC. No estudo ARIES, 1018 pares mãe-criança foram epigenotipadas utilizando a plataforma Illumina Infinium HumanMethylation450K BeadChip, que avalia mais de 485 mil sítios de metilação localizadas em regiões regulatórias ao longo do epigenoma humano. O resultado fornecido por esta metodologia é a proporção de células cujo DNA estava metilado na região analisada.

Outras características sobre a epigenotipagem no estudo ARIES são mostradas na Tabela 2. O acesso a estes dados se dará pela forma de submissão de uma proposta de pesquisa ao comitê executivo do estudo ALSPAC através do endereço eletrônico <https://proposals.epi.bristol.ac.uk/>.

Tendo em vista a disponibilidade de dados epigenéticos em diferentes idades para os participantes do estudo ARIES, será possível avaliar se associações entre amamentação e sítios de metilação se mantêm ao longo do tempo. Isso é útil não apenas para avaliar possíveis efeitos duradouros da amamentação, mas também para detectar associações que provavelmente foram resultado do acaso. Isto porque, mesmo aplicando uma correção para inflação do erro tipo-I, associações a partir de múltiplos testes são sujeitas ao viés *winner's curse*, ou seja, as estimativas de efeito mais fortes tendem a serem valores extremos das respectivas distribuições de estimativas que seriam obtidas ao obter várias amostras da população. Por outro lado, uma associação inicial que não persiste não implica, necessariamente, que o primeiro achado tenha sido obra do acaso, pois é possível que a modificação epigenética seja transiente.

Uma associação que se mantém ao longo do tempo não é necessariamente causal, pois pode ser consequência, por exemplo, de outros determinantes precoces que também estão associados com a exposição (Figura 2). Isto será abordado neste projeto de duas formas. Inicialmente, será feita a comparação entre análises brutas e ajustadas para fatores sociodemográficos precoces, variáveis relacionadas à gestação, características da mãe e de estimulação infantil (Tabela 3). Em segundo lugar, como o estudo ARIES inclui amostra do sangue umbilical, coletado antes do início da amamentação, será possível comparar estes padrões de metilação com os observados aos 7,5 e 15,5 anos de idade; padrões que se mantiverem inalterados não poderiam ser atribuídos à amamentação. Apesar de que o perfil epigenético do sangue do cordão umbilical é diferente do sangue periférico, é improvável que isto introduza vieses importantes nas análises por duas razões principais: i) os dados do estudo ARIES foram ajustados para heterogeneidade celular utilizando um algoritmo desenvolvido em painéis de referência externos [124]; ii) as análises principais utilizarão os dados referentes aos 7,5 e 15,5 anos de idade, e ambos coletaram sangue periférico para extrair DNA.

Tabela 2. Características do estudo ARIES.

Subgrupo	Ponto no tempo (idade da mãe)	Tecido/tipo celular
Mães	Pré-natal (~29,2 anos)	Sangue periférico
	Idade do filho de ~15,5 anos (~47,5 anos)	Sangue periférico
Crianças	Ao nascer, ~40 semanas de gestação	Sangue do cordão umbilical
	~7,5 anos de idade	Sangue periférico
	~15,5 anos de idade	Sangue periférico

A amamentação será a variável de exposição, e será avaliada conforme mostrado na Tabela 4. Esta variável será categorizada em cinco grupos e utilizada tanto na forma categórica como na forma numérica para avaliação de tendência linear.

Primeiramente, será realizado um EWAS utilizando dados epigenéticos coletados aos ~7,5 anos de idade, tendo amamentação como exposição, sem ajustar para potenciais confundidores. Estas análises serão comparadas com as respectivas análises ajustadas. As associações encontradas, após ajuste para múltiplos testes, serão avaliadas também aos ~15,5 anos de idade, visando avaliar se as mesmas se mantêm. Tanto as associações transientes como as duradouras serão confrontadas com os resultados referentes aos dados epigenéticos coletados ao nascer, sendo que associações encontradas em ambos os casos serão classificadas como falso-positivas, ou seja, devidas a fatores outros que a amamentação.

Apesar de o principal objetivo deste artigo ser identificar sítios de metilação associados com amamentação, análises adicionais serão realizadas visando compreender o contexto biológico dos achados do EWAS. Tais análises envolvem anotação gênica, enriquecimento de função molecular, avaliação de rotas bioquímicas e do perfil de expressão gênica. Estas análises serão definidas posteriormente, durante o período de Doutorado sanduíche na Universidade de Bristol.

A interação entre SNPs no gene *FADS2* e amamentação será avaliada em uma meta-análise colaborativa *de novo*. Os resultados serão obtidos com base em um plano de análise e códigos padronizados, diminuindo heterogeneidade em função de metodologias distintas. Com base em buscas na literatura, em ferramentas de busca gerais (como o Google [www.google.com]) e de contatos dos pesquisadores que

Tabela 3. Lista de covariáveis a serem utilizadas no EWAS de amamentação.

Grupo	Variável^a
Socioeconômicas	Maior qualificação profissional (mãe e parceiro) Classe social (mãe e parceiro) Grupo socioeconômico (mãe e parceiro) Classe social com base na ocupação (mãe e parceiro)
Demográficas	Idade (mãe e criança) Sexo (criança) Etnia (criança)
Características maternas	Paridade Peso pré-gestacional Altura Índice de massa corporal pré-gestacional
Características da gestação	Tabagismo materno Tipo de parto Gemelaridade Idade gestacional Peso ao nascer
Relação mãe-criança	Escore de suporte social Escore de conexão mãe-criança Escore de parentalidade

^aDefinições operacionais não foram apresentadas porque isto dependerá tanto do banco de dados em si quanto da associação entre amamentação e diferentes categorizações de cada covariável.

Tabela 4. Categorização da variável amamentação para as análises de EWAS.

Categorização	Categorias
Amamentado	0=Nenhuma amamentação 1=Qualquer amamentação
Amamentado por 6 meses	0=Amamentado por menos de 6 meses 1=Amamentado por pelo menos 6 meses
Amamentação em categorias	0=Nenhuma amamentação 1=Amamentação >0 e <3 meses 2=Amamentação ≥3 e <6 meses 3=Amamentação ≥6 e <12 meses 4=Amamentação ≥12
Amamentação em meses	Variável numérica

participarão desta pesquisa, serão identificados estudos potencialmente elegíveis. Isto tende a minimizar viés de publicação e a obter poder estatístico suficiente. O protocolo detalhado deste estudo foi recentemente publicado [125] (anexo I). Estima-se que cerca de dez estudos diferentes podem contribuir com dados.

Meta-análise *de novo* (Artigo 3)

O projeto prevê cinco etapas gerais:

a) Contatar os coordenadores dos estudos identificados como potencialmente elegíveis para saber seu interesse em participar da pesquisa. b) Enviar aos interessados e elegíveis um plano de análise detalhado, bem como planilhas sobre características do estudo a serem preenchidas pelos analistas de cada estudo colaborador. c) Receber os resultados de cada estudo na forma de arquivos gerados automaticamente pelas rotinas de análises disponibilizadas aos analistas (mais detalhes abaixo). d) Avaliar os arquivos recebidos e contatar novamente estudos com resultados discrepantes dos demais (se houver) para discutir possibilidades de erros na análise. e) Realizar as análises finais, que serão divulgadas na forma de artigo científico.

Na etapa b), serão enviados cinco arquivos aos estudos colaboradores:

Planilha descritiva: o analista deverá preencher e enviá-la juntamente com os resultados gerados automaticamente pelos códigos disponibilizados (ver abaixo). As informações contidas na planilha serão usadas para descrever cada estudo no artigo científico e possivelmente utilizadas em análises de meta-regressão como potenciais fontes de heterogeneidade.

Plano de análise (anexo II): arquivo que detalha todos os procedimentos relacionados à análise dos dados. Abrange aspectos teóricos (por exemplo, a estratégia de modelagem estatística proposta) e práticos (principalmente instruções sobre como utilizar os demais arquivos disponibilizados – ver abaixo).

Instruções para formatação dos dados: a maior tarefa do analista de cada estudo colaborador será formatar os dados do seu estudo de forma correta, de modo que as rotinas de análise disponibilizadas funcionem adequadamente. Este arquivo fornece informações detalhadas de como formatar os dados.

Código do analista (anexo III): este arquivo contém a rotina da análise que deve ser realizada pelo analista de cada estudo colaborador. Ela foi escrita na linguagem do programa R (www.r-project.org), pois é um programa gratuito e amplamente utilizado. O código é muito simples, tendo apenas 166 linhas, sendo que a maioria delas são instruções de como utilizar o código. A tarefa do usuário é, principalmente, indicar a localização do seu banco de dados e a pasta onde deseja salvar os arquivos com os resultados da análise gerados automaticamente pelo código fornecido, além de fornecer algumas informações simples (por exemplo, a data em que está realizando as análises).

O código do analista realiza três principais etapas de forma automatizada: i) extensa verificação dos dados quanto a possíveis erros de formatação; ii) computar estatísticas descritivas, gerando arquivos contendo as mesmas; iii) realizar as análises de associação, gerando arquivos contendo os resultados.

Código com funções (anexo IV): para diminuir a carga de trabalho dos analistas dos estudos colaboradores, os mesmos deverão utilizar apenas o código do analista (descrito no item anterior). Para que isto fosse possível, foram geradas diversas funções (também na linguagem R) através de mais de 1100 linhas de código. A rotina contida no código do analista faz uso destas funções sem que o mesmo necessite manipulá-las (como se as mesmas fossem funções disponíveis na instalação de base do programa), simplificando a tarefa do analista colaborador. Isso também limita o código que será manipulado pelos analistas, diminuindo a possibilidade de heterogeneidade em função de adaptações locais da rotina de análise que não forem comunicadas à coordenação do estudo.

Dois aspectos da análise de cada estudo merecem destaque. Em estudos de interação gene-ambiente, é necessário incluir termos de interação gene-covariável e ambiente-covariável. Estas covariáveis estão listadas no anexo I. Embora isso seja recomendável para controlar potenciais efeitos de confundimento exercidos pela covariável, a literatura apresenta diversos casos onde isto não foi considerado [126]. No presente estudo, as covariáveis serão apropriadamente modeladas para reduzir confundimento residual. O segundo aspecto se refere à interpretabilidade dos coeficientes. As

covariáveis quantitativas serão recodificadas de modo que suas médias passarão a ser zero. Como termos de interação com cada covariável serão incluídos, os coeficientes de regressão da variante genética, da amamentação e da interação entre ambas (β_1 , β_2 e β_3 na equação mostrada no anexo I) referem-se portanto ao valor médio das covariáveis quantitativas, facilitando portanto a interpretação de covariáveis para as quais o valor de zero não faz sentido.

ASPECTOS ÉTICOS

Com relação ao EWAS, aprovações éticas do estudo ARIES foram obtidas de diversos comitês, incluindo o *Human Development Biology Resource*, o *Newcastle Brain Tissue Resource* e o *Leiden University Medical Center*. Aprovação ética da coorte de ALSPAC foi obtida do *ALSPAC study Ethics and Law committee* e comitês locais de ética em pesquisa.

Com relação à meta-análise colaborativa, estudos sem aprovação ética adequada serão excluídos. Como somente dados sumarizados serão compartilhados, os aspectos éticos se limitam aos estudos individuais.

PARTICIPAÇÃO DO DOUTORANDO NO PROJETO EPIGEN-Brasil

O projeto EPIGEN-Brasil é, atualmente, a maior iniciativa Latino Americana nos campos de genômica populacional e epidemiologia genética. O estudo envolve 6487 indivíduos com dados de genotipagem em larga escala (~2,5 milhões de variantes genéticas), além de 30 indivíduos com sequenciamento do genoma completo. Três estudos fazem parte do projeto EPIGEN-Brasil: coorte dos nascidos em Pelotas em 1982 (n=3736) [127, 128]; coorte de Bambuí, voltada ao estudo do envelhecimento (n=1442) [129]; e o estudo longitudinal Salvador-SCAALA (n=1309) [130].

Desde 2013, o doutorando têm participado no projeto EPIGEN-Brasil gerenciando e analisando o banco de dados de Pelotas, bem como participando de eventos. Com relação ao gerenciamento do banco de dados, as principais atividades até o momento foram:

- Limpeza do banco de dados através da aplicação de filtros de controle de qualidade, removendo tanto variantes genéticas quanto amostras que foram genotipadas com baixa qualidade.
- Imputação de variantes genéticas não-genotipadas utilizando dados do projeto 1000 genomas como painel de referência [131, 132]. Este processo não só aumentou o número de marcadores disponíveis de ~2,5 milhões para ~40 milhões como também aumentou a sobreposição de variantes disponíveis em estudos que utilizaram diferentes plataformas de genotipagem. A última condição é essencial para participar em estudos colaborativos, muito comuns em epidemiologia genética [132]. Apesar de que a imputação já ter sido realizada em 2014, novos painéis de referência são disponibilizados ao longo do tempo, de forma que a imputação também precisa ser atualizada. Inclusive, painéis de referência mais detalhados já estão disponíveis, porém ainda não estão sendo adotados pelos principais consórcios internacionais de epidemiologia genética.
- Fornecimento de dados genéticos para estudos realizados na coorte de 1982. Para tanto, foram desenvolvidos *scripts* na linguagem R que permitem reunir, em um mesmo banco de dados, diferentes SNPs de forma eficiente.

Com relação a análises do banco de dados, as principais atividades até o momento foram:

- Associação entre ancestralidade genômica e função pulmonar na coorte de 1982 [133].
- Associação entre variantes genéticas que influenciam níveis de homocisteína no sangue e pressão arterial na coorte dos nascidos em Pelotas em 1982 e utilizando dados já publicados de consórcios de epidemiologia genética [134].
- Associação entre uma variante genética relacionada com persistência da lactase e obesidade e pressão arterial na coorte dos nascidos em Pelotas em 1982 seguida por uma revisão sistemática e meta-análise [135].
- Participação, como analista de dados da coorte dos nascidos em Pelotas em 1982, em consórcios internacionais de epidemiologia genética baseados em GWAS sobre: densidade mineral óssea; índice de massa corporal; altura; hemoglobina glicada; função pulmonar; atividade física; quatro estudos de interação gene-ambiente; DNA

mitocondrial (envolvendo imputação de variantes no DNA mitocondrial e associação com desfechos metabólicos). Em cada um destes projetos, foram realizadas análises de associação entre cada variante genética (~40 milhões) e o desfecho estudado, sendo que, em geral, cada projeto envolvia vários modelos diferentes (por exemplo, incluindo diferentes covariáveis). Para realizar estas análises, foram desenvolvidos *scripts* na linguagem R para facilitar a elaboração dos *scripts* de análise propriamente ditos (principalmente visando realizar as análises em paralelo, ou seja, utilizando diferentes unidades de processamento simultaneamente), que são executados por programas como o SNPTTEST (https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html) ou ProbABEL (<http://www.genabel.org/packages/ProbABEL>).

- Participação, como analista de dados da coorte dos nascidos em Pelotas em 1982, em consórcios internacionais de epidemiologia genética de outras naturezas, como estudos de replicação ou de gene-candidato.

Tendo em vista o grande envolvimento do doutorando no projeto EPIGEN-Brasil, bem como a necessidade de alocar recursos humanos para trabalhar com os dados genéticos da coorte dos nascidos em Pelotas em 1982, pensou-se na possibilidade de que esta atividade substituísse o trabalho de campo, que comumente se dá pela participação do aluno em um acompanhamento de alguma das coortes de nascimentos em Pelotas. Esta sugestão foi aprovada pela coordenação do programa através da professora Iná dos Santos da Silva (então coordenadora) e do professor Pedro Curi Hallal (atual coordenador).

CRONOGRAMA

Quadro 1. Cronograma de atividades e períodos de execução

Atividade	2015					2016					2017					2018				
	Mar a Abr	Mai a Jun	Jul a Ago	Set a Out	Nov a Dez	Jan a Fev	Mar a Abr	Mai a Jun	Jul a Ago	Set a Out	Nov a Dez	Jan a Fev	Mar a Abr	Mai a Jun	Jul a Ago	Set a Out	Nov a Dez	Jan a Fev	Mar a Abr	Mai a Jun
Projeto EPIGEN																				
Revisão de literatura																				
Elaboração do projeto																				
Elaboração do plano e scripts de análise																				
Obtenção dos dados de ALSPAC																				
Rebecimento dos dados da meta-análise																				
Doutorado sanduíche																				
Análise de dados																				
Redação de artigos																				
Defesa de tese																				

DIVULGAÇÃO DOS RESULTADOS

Os resultados oriundos deste projeto serão submetidos a periódicos pertinentes para publicação como artigos científicos. Além disso, os resultados serão apresentados como Tese de conclusão do curso de Doutorado em Epidemiologia da Universidade Federal de Pelotas.

FINANCIAMENTO

Em 2015, o doutorando recebeu bolsa da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e, em 2016, do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). O período de doutorado sanduíche será financiado pelo *MRC Integrative Epidemiology Unit* da Universidade de Bristol.

A coorte de ALSPAC é majoritariamente financiada pelo *UK Medical Research Council*, o *Wellcome Trust* e a Universidade de Bristol, além de agências de fomento adicionais através de projetos específicos. O estudo ARIES é financiado pelo *Biotechnology and Biological Sciences Research Council (BBSRC)* no Reino Unido.

A coorte dos nascidos em Pelotas em 1982 é conduzida pelo Programa de Pós-Graduação em Epidemiologia da Universidade Federal de Pelotas, em colaboração com a Associação Brasileira de Saúde Coletiva (ABRASCO).

De 2004 a 2013, a fundação *Wellcome Trust* financiou o estudo. Financiamentos adicionais foram recebidos do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e da Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS). Fases anteriores do estudo foram financiadas pelo Programa de Apoio a Núcleos de Excelência (PRONEX), Ministério da Saúde, Organização Mundial da Saúde, União Europeia, *International Development Research Center* e *Overseas Development Administration*. A genotipagem foi financiada pelo Departamento de Ciência e Tecnologia (DECIT, Ministério da Saúde), Fundo Nacional de Desenvolvimento Científico e Tecnológico (FNDCT, Ministério da Ciência e Tecnologia), Financiadora de Estudos e Projetos (FINEP, Ministério da Ciência e Tecnologia) e Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, Ministério da Educação).

REFERÊNCIAS

1. Victora CG, Bahl R, Barros AJ, et al. Breastfeeding in the 21st century: epidemiology, mechanisms, and lifelong effect. *Lancet* 2016;387(10017):475-90.
2. Rollins NC, Bhandari N, Hajeebhoy N, et al. Why invest, and what it will take to improve breastfeeding practices? *Lancet* 2016;387(10017):491-504.
3. World Health Organization and UNICEF. Protecting, Promoting and Supporting Breastfeeding: The Special Role of Maternity Services. Geneva, Switzerland: 1989.
4. World Health Organization. The Optimal Duration of Exclusive Breastfeeding. Geneva, Switzerland: World Health Organization: 2001.
5. Rollins NC, Bhandari N, Hajeebhoy N, et al. Why invest, and what it will take to improve breastfeeding practices? *Lancet* 2016;387(10017):491-504.
6. Hansen K. Breastfeeding: a smart investment in people and in economies. *Lancet* 2016;387(10017):416.
7. Mullan Z. The debate that shouldn't be. *Lancet Glob Health* 2015;3(9):e501.
8. Pierce BA. Genetics: A Conceptual Approach. 4 ed. New York, NY, USA: W. H. Freeman and Company, 2012.
9. Weiling F. Historical study: Johann Gregor Mendel 1822-1884. *Am J Med Genet* 1991;40(1):1-25; discussion 26.
10. Burton PR, Tobin MD, Hopper JL. Key concepts in genetic epidemiology. *Lancet* 2005;366(9489):941-51.
11. Cordell HJ, Clayton DG. Genetic association studies. *Lancet* 2005;366(9491):1121-31.
12. Bush WS, Moore JH. Chapter 11: Genome-wide association studies. *PLoS Comput Biol* 2012;8(12):e1002822.
13. Nature Genetics. On beyond GWAS. *Nat Genet* 2010;42(7):551.

14. Welter D, MacArthur J, Morales J, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 2014;42(Database issue):D1001-6.
15. Da Y, Wang C, Wang S, et al. Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using SNP markers. *PLoS One* 2014;9(1):e87666.
16. Abraham G, Inouye M. Genomic risk prediction of complex human disease and its clinical application. *Curr Opin Genet Dev* 2015;33:10-6.
17. Chatterjee N, Shi J, Garcia-Closas M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat Rev Genet* 2016;17(7):392-406.
18. Sheng J, Li F, Wong ST. Optimal drug prediction from personal genomics profiles. *IEEE J Biomed Health Inform* 2015;19(4):1264-70.
19. Panczyk M. Pharmacogenetics research on chemotherapy resistance in colorectal cancer over the last 20 years. *World J Gastroenterol* 2014;20(29):9775-827.
20. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum Mol Genet* 2014;23(R1):R89-98.
21. Smith GD, Lawlor DA, Harbord R, et al. Clustered environments and randomized genes: a fundamental distinction between conventional and genetic epidemiology. *PLoS Med* 2007;4(12):e352.
22. Davey Smith G. Use of genetic markers and gene-diet interactions for interrogating population-level causal influences of diet on health. *Genes Nutr* 2011;6(1):27-43.
23. Price AL, Zaitlen NA, Reich D, et al. New approaches to population stratification in genome-wide association studies. *Nat Rev Genet* 2010;11(7):459-63.
24. Ott J, Kamatani Y, Lathrop M. Family-based designs for genome-wide association studies. *Nat Rev Genet* 2011;12(7):465-74.

25. Shriner D. Overview of admixture mapping. *Curr Protoc Hum Genet* 2013;Chapter 1:Unit 1 23.
26. Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. *Nat Rev Genet* 2010;11(12):843-54.
27. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol* 2015;44(2):512-25.
28. Bowden J, Davey Smith G, Haycock PC, et al. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet Epidemiol* 2016;40(4):304-14.
29. Relton CL, Hartwig FP, Davey Smith G. From stem cells to the law courts: DNA methylation, the forensic epigenome and the possibility of a biosocial archive. *Int J Epidemiol* 2015;44(4):1083-93.
30. Han L, Su B, Li WH, et al. CpG island density and its correlations with genomic features in mammalian genomes. *Genome Biol* 2008;9(5):R79.
31. Rakyan VK, Down TA, Balding DJ, et al. Epigenome-wide association studies for common human diseases. *Nat Rev Genet* 2011;12(8):529-41.
32. Breitling LP, Yang R, Korn B, et al. Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. *Am J Hum Genet* 2011;88(4):450-7.
33. Bjornsson HT, Sigurdsson MI, Fallin MD, et al. Intra-individual change over time in DNA methylation with familial clustering. *JAMA* 2008;299(24):2877-83.
34. Zhang B, Zhou Y, Lin N, et al. Functional DNA methylation differences between tissues, cell types, and across individuals discovered using the M&M algorithm. *Genome Res* 2013;23(9):1522-40.
35. Relton CL, Davey Smith G. Epigenetic epidemiology of common complex disease: prospects for prediction, prevention, and treatment. *PLoS Med* 2010;7(10):e1000356.

36. Relton CL, Davey Smith G. Is epidemiology ready for epigenetics? *Int J Epidemiol* 2012;41(1):5-9.
37. Ng JW, Barrett LM, Wong A, et al. The role of longitudinal cohort studies in epigenetic epidemiology: challenges and opportunities. *Genome Biol* 2012;13(6):246.
38. Heijmans BT, Mill J. Commentary: The seven plagues of epigenetic epidemiology. *Int J Epidemiol* 2012;41(1):74-8.
39. Lucas A, Gore SM, Cole TJ, et al. Multicentre trial on feeding low birthweight infants: effects of diet on early growth. *Arch Dis Child* 1984;59(8):722-30.
40. Singhal A, Cole TJ, Fewtrell M, et al. Breastmilk feeding and lipoprotein profile in adolescents born preterm: follow-up of a prospective randomised study. *Lancet* 2004;363(9421):1571-8.
41. Singhal A, Cole TJ, Lucas A. Early nutrition in preterm infants and later blood pressure: two cohorts after randomised trials. *Lancet* 2001;357(9254):413-9.
42. Lewandowski AJ, Lamata P, Francis JM, et al. Breast Milk Consumption in Preterm Neonates and Cardiac Shape in Adulthood. *Pediatrics* 2016;138(1).
43. Godfrey KM, Lillycrop KA, Burdge GC, et al. Epigenetic mechanisms and the mismatch concept of the developmental origins of health and disease. *Pediatr Res* 2007;61(5 Pt 2):5R-10R.
44. Gluckman PD, Hanson MA, Mitchell MD. Developmental origins of health and disease: reducing the burden of chronic disease in the next generation. *Genome Med* 2010;2(2):14.
45. Waterland RA, Michels KB. Epigenetic epidemiology of the developmental origins hypothesis. *Annu Rev Nutr* 2007;27:363-88.
46. Richmond RC, Simpkin AJ, Woodward G, et al. Prenatal exposure to maternal smoking and offspring DNA methylation across the lifecourse: findings from the

- Avon Longitudinal Study of Parents and Children (ALSPAC). *Hum Mol Genet* 2015;24(8):2201-17.
47. Verduci E, Banderali G, Barberi S, et al. Epigenetic effects of human breast milk. *Nutrients* 2014;6(4):1711-24.
 48. Tow J. Heal the mother, heal the baby: epigenetics, breastfeeding and the human microbiome. *Breastfeed Rev* 2014;22(1):7-9.
 49. Mischke M, Plosch T. More than just a gut instinct-the potential interplay between a baby's nutrition, its gut microbiome, and the epigenome. *Am J Physiol Regul Integr Comp Physiol* 2013;304(12):R1065-9.
 50. Martin R, Heilig HG, Zoetendal EG, et al. Cultivation-independent assessment of the bacterial diversity of breast milk among healthy women. *Res Microbiol* 2007;158(1):31-7.
 51. Favier CF, de Vos WM, Akkermans AD. Development of bacterial and bifidobacterial communities in feces of newborn babies. *Anaerobe* 2003;9(5):219-29.
 52. Hopkins MJ, Macfarlane GT, Furrie E, et al. Characterisation of intestinal bacteria in infant stools using real-time PCR and northern hybridisation analyses. *FEMS Microbiol Ecol* 2005;54(1):77-85.
 53. Obermann-Borst SA, Eilers PH, Tobi EW, et al. Duration of breastfeeding and gender are associated with methylation of the LEPTIN gene in very young children. *Pediatr Res* 2013;74(3):344-9.
 54. Horta BL, Loret de Mola C, Victora CG. Long-term consequences of breastfeeding on cholesterol, obesity, systolic blood pressure and type 2 diabetes: a systematic review and meta-analysis. *Acta Paediatr* 2015;104(467):30-7.
 55. Tao MH, Marian C, Shields PG, et al. Exposures in early life: associations with DNA promoter methylation in breast tumors. *J Dev Orig Health Dis* 2013;4(2):182-90.

56. Soto-Ramirez N, Arshad SH, Holloway JW, et al. The interaction of genetic variants and DNA methylation of the interleukin-4 receptor gene increase the risk of asthma at age 18 years. *Clin Epigenetics* 2013;5(1):1.
57. Rossnerova A, Tulupova E, Tabashidze N, et al. Factors affecting the 27K DNA methylation pattern in asthmatic and healthy children from locations with various environments. *Mutat Res* 2013;741-742:18-26.
58. Horta BL, Loret de Mola C, Victora CG. Breastfeeding and intelligence: a systematic review and meta-analysis. *Acta Paediatr Suppl* 2015;104(467):14-9.
59. Victora CG, Horta BL, Loret de Mola C, et al. Association between breastfeeding and intelligence, educational attainment, and income at 30 years of age: a prospective birth cohort study from Brazil. *The Lancet. Global health* 2015;3(4):e199-205.
60. Kramer MS, Aboud F, Mironova E, et al. Breastfeeding and child cognitive development: new evidence from a large randomized trial. *Arch Gen Psychiatry* 2008;65(5):578-84.
61. Victora CG, Hallal PC, Araujo CL, et al. Cohort profile: the 1993 Pelotas (Brazil) birth cohort study. *Int J Epidemiol* 2008;37(4):704-9.
62. Goncalves H, Assuncao MC, Wehrmeister FC, et al. Cohort profile update: The 1993 Pelotas (Brazil) birth cohort follow-up visits in adolescence. *Int J Epidemiol* 2014;43(4):1082-8.
63. Fraser A, Macdonald-Wallis C, Tilling K, et al. Cohort Profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *Int J Epidemiol* 2013;42(1):97-110.
64. Boyd A, Golding J, Macleod J, et al. Cohort Profile: the 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children. *Int J Epidemiol* 2013;42(1):111-27.

65. Brion MJ, Lawlor DA, Matijasevich A, et al. What are the causal effects of breastfeeding on IQ, obesity and blood pressure? Evidence from comparing high-income with middle-income cohorts. *Int J Epidemiol* 2011;40(3):670-80.
66. Lucas A, Morley R, Cole TJ, et al. Breast milk and subsequent intelligence quotient in children born preterm. *Lancet* 1992;339(8788):261-4.
67. Innis SM. Human milk: maternal dietary lipids and infant development. *Proc Nutr Soc* 2007;66(3):397-404.
68. Cetin I, Koletzko B. Long-chain omega-3 fatty acid supply in pregnancy and lactation. *Curr Opin Clin Nutr Metab Care* 2008;11(3):297-302.
69. Innis SM. Dietary (n-3) fatty acids and brain development. *J Nutr* 2007;137(4):855-9.
70. Innis SM. Dietary omega 3 fatty acids and the developing brain. *Brain Res* 2008;1237:35-43.
71. Jiao J, Li Q, Chu J, et al. Effect of n-3 PUFA supplementation on cognitive function throughout the life span from infancy to old age: a systematic review and meta-analysis of randomized controlled trials. *Am J Clin Nutr* 2014;100(6):1422-36.
72. Qawasmi A, Landeros-Weisenberger A, Bloch MH. Meta-analysis of LCPUFA supplementation of infant formula and visual acuity. *Pediatrics* 2013;131(1):e262-72.
73. Schaeffer L, Gohlke H, Muller M, et al. Common genetic variants of the FADS1 FADS2 gene cluster and their reconstructed haplotypes are associated with the fatty acid composition in phospholipids. *Hum Mol Genet* 2006;15(11):1745-56.
74. Tanaka T, Shen J, Abecasis GR, et al. Genome-wide association study of plasma polyunsaturated fatty acids in the InCHIANTI Study. *PLoS Genet* 2009;5(1):e1000338.
75. Sprecher H. Metabolism of highly unsaturated n-3 and n-6 fatty acids. *Biochim Biophys Acta* 2000;1486(2-3):219-31.

76. Nakamura MT, Nara TY. Structure, function, and dietary regulation of delta6, delta5, and delta9 desaturases. *Annu Rev Nutr* 2004;24:345-76.
77. Steer CD, Davey Smith G, Emmett PM, et al. FADS2 polymorphisms modify the effect of breastfeeding on child IQ. *PLoS One* 2010;5(7):e11570.
78. Caspi A, Williams B, Kim-Cohen J, et al. Moderation of breastfeeding effects on the IQ by genetic variation in fatty acid metabolism. *Proc Natl Acad Sci U S A* 2007;104(47):18860-5.
79. Chisaguano AM, Montes R, Perez-Berezo T, et al. Gene expression of desaturase (FADS1 and FADS2) and Elongase (ELOVL5) enzymes in peripheral blood: association with polyunsaturated fatty acid levels and atopic eczema in 4-year-old children. *PLoS One* 2013;8(10):e78245.
80. Martin NW, Benyamin B, Hansell NK, et al. Cognitive function in adolescence: testing for interactions between breast-feeding and FADS2 polymorphisms. *J Am Acad Child Adolesc Psychiatry* 2011;50(1):55-62 e4.
81. Groen-Blokhuis MM, Franic S, van Beijsterveldt CE, et al. A prospective study of the effects of breastfeeding and FADS2 polymorphisms on cognition and hyperactivity/attention problems. *Am J Med Genet B Neuropsychiatr Genet* 2013;162B(5):457-65.
82. Rizzi TS, van der Sluis S, Derom C, et al. Genetic Variance in Combination with Fatty Acid Intake Might Alter Composition of the Fatty Acids in Brain. *PLoS One* 2013;8(6):e68000.
83. Yokoyama Y, Wada S, Sugimoto M, et al. Breastfeeding rates among singletons, twins and triplets in Japan: A population-based study. *Twin Res Hum Genet* 2006;9(2):298-302.
84. Flidel-Rimon O, Shinwell ES. Breast feeding twins and high multiples. *Arch Dis Child Fetal Neonatal Ed* 2006;91(5):F377-80.
85. Shrier I, Platt RW. Reducing bias through directed acyclic graphs. *BMC Med Res Methodol* 2008;8:70.

86. Cardoso FH. Capitalismo e escravidão no Brasil meridional: o negro na sociedade escravocrata do Rio Grande do Sul. 6 ed. Rio de Janeiro, RJ: Civilização Brasileira, 2011.
87. Chor D, Lima CR. Aspectos epidemiológicos das desigualdades raciais em saúde no Brasil. *Cad Saude Publica* 2005;21(5):1586-94.
88. Williams DR, Mohammed SA, Leavell J, et al. Race, socioeconomic status, and health: complexities, ongoing challenges, and research opportunities. *Ann N Y Acad Sci* 2010;1186:69-101.
89. Quillian L. Segregation and Poverty Concentration: The Role of Three Segregations. *Am Sociol Rev* 2012;77(3):354-79.
90. Walker SP, Wachs TD, Gardner JM, et al. Child development: risk factors for adverse outcomes in developing countries. *Lancet* 2007;369(9556):145-57.
91. Furfey PH. The Pedagogical Seminary and Journal of Genetic Psychology. *Journal of Genetic Psychology* 1928;35(3):478-80.
92. Sewell WH, Shah VP. Socioeconomic Status, Intelligence, and the Attainment of Higher Education. *Sociology of Education* 1967;40(1):1-23.
93. Duyme M, Dumaret AC, Tomkiewicz S. How can we boost IQs of "dull children"? A late adoption study. *Proc Natl Acad Sci U S A* 1999;96(15):8790-4.
94. Heckman JJ. Skill formation and the economics of investing in disadvantaged children. *Science* 2006;312(5782):1900-2.
95. Marmot M, Friel S, Bell R, et al. Closing the gap in a generation: health equity through action on the social determinants of health. *Lancet* 2008;372(9650):1661-9.
96. Braveman P, Gottlieb L. The social determinants of health: it's time to consider the causes of the causes. *Public Health Rep* 2014;129 Suppl 2:19-31.

97. Raisanen S, Gissler M, Kramer MR, et al. Influence of delivery characteristics and socioeconomic status on giving birth by caesarean section - a cross sectional study during 2000-2010 in Finland. *BMC Pregnancy Childbirth* 2014;14:120.
98. Elshibly EM, Schmalisch G. The effect of maternal anthropometric characteristics and social factors on gestational age and birth weight in Sudanese newborn infants. *BMC Public Health* 2008;8:244.
99. Black RE, Allen LH, Bhutta ZA, et al. Maternal and child undernutrition: global and regional exposures and health consequences. *Lancet* 2008;371(9608):243-60.
100. Ng SK, Cameron CM, Hills AP, et al. Socioeconomic disparities in prepregnancy BMI and impact on maternal and neonatal outcomes and postpartum weight retention: the EFHL longitudinal birth cohort study. *BMC Pregnancy Childbirth* 2014;14:314.
101. Jones JR, Kogan MD, Singh GK, et al. Factors associated with exclusive breastfeeding in the United States. *Pediatrics* 2011;128(6):1117-25.
102. Michels KA, Mumford SL, Sundaram R, et al. Differences in infant feeding practices by mode of conception in a United States cohort. *Fertil Steril* 2016;105(4):1014-22 e1.
103. Kitano N, Nomura K, Kido M, et al. Combined effects of maternal age and parity on successful initiation of exclusive breastfeeding. *Prev Med Rep* 2016;3:121-6.
104. Oakley LL, Renfrew MJ, Kurinczuk JJ, et al. Factors associated with breastfeeding in England: an analysis by primary care trust. *BMJ Open* 2013;3(6).
105. Wojcicki JM. Maternal prepregnancy body mass index and initiation and duration of breastfeeding: a review of the literature. *J Womens Health (Larchmt)* 2011;20(3):341-7.
106. Castillo H, Santos IS, Matijasevich A. Maternal pre-pregnancy BMI, gestational weight gain and breastfeeding. *Eur J Clin Nutr* 2016;70(4):431-6.

- 107.Horta BL, Kramer MS, Platt RW. Maternal smoking and the risk of early weaning: a meta-analysis. *Am J Public Health* 2001;91(2):304-7.
- 108.Engel SM, Joubert BR, Wu MC, et al. Neonatal genome-wide methylation patterns in relation to birth weight in the Norwegian Mother and Child Cohort. *Am J Epidemiol* 2014;179(7):834-42.
- 109.Adkins RM, Thomas F, Tylavsky FA, et al. Parental ages and levels of DNA methylation in the newborn are correlated. *BMC Med Genet* 2011;12:47.
- 110.Markunas CA, Wilcox AJ, Xu Z, et al. Maternal Age at Delivery Is Associated with an Epigenetic Signature in Both Newborns and Adults. *PLoS One* 2016;11(7):e0156361.
- 111.Herbstman JB, Wang S, Perera FP, et al. Predictors and consequences of global DNA methylation in cord blood and at three years. *PLoS One* 2013;8(9):e72824.
- 112.Sharp GC, Lawlor DA, Richmond RC, et al. Maternal pre-pregnancy BMI and gestational weight gain, offspring DNA methylation and later offspring adiposity: findings from the Avon Longitudinal Study of Parents and Children. *Int J Epidemiol* 2015;44(4):1288-304.
- 113.Simpkin AJ, Suderman M, Gaunt TR, et al. Longitudinal analysis of DNA methylation associated with birth weight and gestational age. *Hum Mol Genet* 2015;24(13):3752-63.
- 114.Katz J, Lee AC, Kozuki N, et al. Mortality risk in preterm and small-for-gestational-age infants in low-income and middle-income countries: a pooled country analysis. *Lancet* 2013;382(9890):417-25.
- 115.Fall CH, Sachdev HS, Osmond C, et al. Association between maternal age at childbirth and child and adult outcomes in the offspring: a prospective study in five low-income and middle-income countries (COHORTS collaboration). *Lancet Glob Health* 2015;3(7):e366-77.
- 116.Adair LS, Fall CH, Osmond C, et al. Associations of linear growth and relative weight gain during early life with adult health and human capital in countries of

- low and middle income: findings from five birth cohort studies. *Lancet* 2013;382(9891):525-34.
117. Tyrrell J, Richmond RC, Palmer TM, et al. Genetic Evidence for Causal Relationships Between Maternal Obesity-Related Traits and Birth Weight. *JAMA* 2016;315(11):1129-40.
 118. Horta BL, Victora CG. Breastfeeding and adult intelligence - Authors' reply. *Lancet Glob Health* 2015;3(9):e522.
 119. Flanagan JM. Epigenome-wide association studies (EWAS): past, present, and future. *Methods Mol Biol* 2015;1238:51-63.
 120. Kent G. Regulating fatty acids in infant formula: critical assessment of U.S. policies and practices. *Int Breastfeed J* 2014;9(1):2.
 121. Morgan C, Davies L, Corcoran F, et al. Fatty acid balance studies in term infants fed formula milk containing long-chain polyunsaturated fatty acids. *Acta Paediatr* 1998;87(2):136-42.
 122. Ioannidis JP. How to make more published research true. *PLoS Med* 2014;11(10):e1001747.
 123. Relton CL, Gaunt T, McArdle W, et al. Data Resource Profile: Accessible Resource for Integrated Epigenomic Studies (ARIES). *Int J Epidemiol* 2015;44(4):1181-90.
 124. Houseman EA, Accomando WP, Koestler DC, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* 2012;13:86.
 125. Hartwig FP, Davies NM, Horta BL, et al. Effect modification of FADS2 polymorphisms on the association between breastfeeding and intelligence: protocol for a collaborative meta-analysis. *BMJ Open* 2016;6(6):e010067.
 126. Keller MC. Gene x environment interaction studies have not properly controlled for potential confounders: the problem and the (simple) solution. *Biol Psychiatry* 2014;75(1):18-24.

127. Victora CG, Barros FC. Cohort profile: the 1982 Pelotas (Brazil) birth cohort study. *Int J Epidemiol* 2006;35(2):237-42.
128. Horta BL, Gigante DP, Goncalves H, et al. Cohort Profile Update: The 1982 Pelotas (Brazil) Birth Cohort Study. *Int J Epidemiol* 2015;44(2):441, 41a-41e.
129. Lima-Costa MF, Firmo JO, Uchoa E. Cohort profile: the Bambui (Brazil) Cohort Study of Ageing. *Int J Epidemiol* 2011;40(4):862-7.
130. Barreto ML, Cunha SS, Alcantara-Neves N, et al. Risk factors and immunological pathways for asthma and other allergic diseases in children: background and methodology of a longitudinal study in a large urban center in Northeastern Brazil (Salvador-SCAALA study). *BMC Pulm Med* 2006;6:15.
131. Abecasis GR, Auton A, Brooks LD, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012;491(7422):56-65.
132. Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* 2010;11(7):499-511.
133. Menezes AM, Wehrmeister FC, Hartwig FP, et al. African ancestry, lung function and the effect of genetics. *Eur Respir J* 2015;45(6):1582-9.
134. Borges MC, Hartwig FP, Oliveira IO, et al. Is there a causal role for homocysteine concentration in blood pressure? A Mendelian randomization study. *Am J Clin Nutr* 2016;103(1):39-49.
135. Hartwig FP, Horta BL, Smith GD, et al. Association of lactase persistence genotype with milk consumption, obesity and blood pressure: a Mendelian randomization study in the 1982 Pelotas (Brazil) Birth Cohort, with a systematic review and meta-analysis. *Int J Epidemiol* 2016;45(5):1573-87.

ANEXOS

Anexo I – Protocolo do estudo de interação entre *FADS2* e amamentação (versão que foi aceita para publicação no periódico *BMJ Open*)

Effect modification of *FADS2* polymorphisms on the association between breastfeeding and intelligence: protocol for a collaborative meta-analysis

Fernando Pires Hartwig^{1*}, Neil Davies², Bernardo Lessa Horta¹, Cesar Gomes Victora¹ and George Davey Smith²

¹Postgraduate Program in Epidemiology, Federal University of Pelotas, Pelotas, Brazil.

²MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK.

*Corresponding author. Postgraduate Program in Epidemiology, Federal University of Pelotas, Pelotas (Brazil) 96020-220. Phone: 55 53 81347172. E-mail: fernandophartwig@gmail.com.

Keywords: Breast Feeding; Intelligence; *FADS2*; Docosahexaenoic Acids; Meta-analysis.

Word count: 4808.

Abstract

Introduction: Evidence from observational studies and randomized controlled trials suggests that breastfeeding is positively associated with IQ, possibly because breast milk is a source of long-chain polyunsaturated fatty acids. Different studies have detected gene-breastfeeding interactions involving *FADS2* variants and intelligence. However, findings are inconsistent regarding the direction of such effect modification.

Methods/Design: To clarify how *FADS2* and breastfeeding interact in their association with IQ, we are conducting a consortium-based meta-analysis of independent studies. Results produced by each individual study using standardized analysis scripts and harmonized data will be used.

Inclusion criteria: breastfeeding, IQ and either rs174575 or rs1535 polymorphisms available; and being of European ancestry. *Exclusion criteria:* twin studies; only poorly-imputed genetic data available; or unavailability of proper ethics approval.

Studies will be invited based on being known to have at least some of the required data or suggested by participating studies as potentially eligible. This inclusive approach will favour to achieve a larger sample size and be less prone to publication bias.

Discussion: Improving current understanding of *FADS2*-breastfeeding interaction may provide important biological insights regarding the importance of long-chain polyunsaturated fatty acids for the breastfeeding-IQ association. This meta-analysis will help to improve such knowledge by replicating earlier studies, conducting additional analysis and evaluating different sources of heterogeneity. Publishing this protocol will minimize the possibility of bias due to *post hoc* changes to the analysis protocol.

Strengths and limitations of this study

- Standardized statistical analysis of harmonized data will improve comparability between studies.
- Attempts to include both published and unpublished studies will minimize the possibility of publications bias.
- It will not be possible to fully harmonize exposure and outcomes measures.
- Additional sources of heterogeneity will likely remain.
- Elaborating and reporting the analytical plan before data analysis will protect against biased reporting.

Introduction

Consortium-based efforts have been proposed as a practice that may contribute to generate more reliable scientific findings.[1] Such approach has many desirable characteristics, including improve power by increasing sample size, harmonization of variables and analyses, and avoiding winner's curse bias. Adapting a similar approach used in a previous work on 5-HTTLPR, stress and depression,[2] this manuscript describes the protocol for a collaborative meta-analysis on the interaction between breastfeeding and *FADS2* polymorphisms when intelligence quotient (IQ) is the outcome. As described previously,[2] publishing the protocol is important for several reasons. These include: avoiding biased reporting by documenting study protocol and design, as well as primary analysis, prior to conducting and publishing the study; facilitate the understanding of the results of the study by its readership when it is completed; and help similar initiatives in the future to elaborate a protocol and encourage this practice as a means to improve transparency and commitment to the analysis plan defined *a priori*.

Background

There is substantial evidence of short-term health benefits of breastfeeding by reducing children morbidity and mortality from infectious diseases.[3, 4] Based on these evidence the World Health Organization[5] and United Nations Children's Fund[6] recommend that every child should be exclusively breastfed for 6 months, with partial breastfeeding continued until two years of age. More recently, associations between breastfeeding and positive health outcomes in adulthood suggest that breastfeeding might also have long-term effects.[4, 7-9]

Different epidemiological studies have detected positive associations between breastfeeding and intelligence-related outcomes.[7, 9] Residual confounding has been suggested to influence much of the findings involving breastfeeding and child cognitive development.[10] However, randomized trials provided evidence that breastfeeding causes increased motor development during the first year of life[11], as well as intelligence measured in healthy infants participating in the PROBIT trial.[12]

Additional evidence of health benefits of breastfeeding from randomized studies includes better cardiovascular risk profile (lipoprotein profile[13] and blood pressure[14]) in preterm-born children at 13-16 years. Long-term observational associations with intelligence quotient (IQ) have also been detected. For example, a recent population-based study in South Brazil (where breastfeeding is not associated with socioeconomic position at birth) identified a positive association with IQ in individuals aged 30-31 years; this association captured 72% of the association of breastfeeding with income.[15] This raises the possibility that breastfeeding not only influences health, but also intellectual human capital and economic productivity.[15]

Given the nature of the interventions in some of the aforementioned trials,[13, 14] at least some of the effects of breastfeeding are hypothesized to be biological. Regarding intelligence, a potential mechanism is that breast milk is a source of long-chain polyunsaturated fatty acids (LC-PUFAs) including docosahexaenoic acid (DHA), which have been implicated in brain development.[16, 17] It has been hypothesized that the association between breastfeeding and IQ could differ according to the capacity to synthesize DHA from metabolic precursors.[18] Special attention has been given to genetic variation in the *FADS2* gene, which encodes a protein involved in desaturation processes required for endogenous synthesis of LC-PUFAS from shorter chain fatty acids.[19, 20]

Caspi and colleagues provided evidence for *FADS2*-breastfeeding interaction involving two *FADS2* variants: rs174575 (major/minor allele: C/G) and 1535 (major/minor allele: A/G). In two independent samples, breastfeeding was positively associated with IQ in non-G carriers, but not in GG individuals.[18] However, these results were not in accordance with the DHA hypothesis, since rs174575-G allele has been associated with lower LC-PUFAs levels in serum[21] and plasma[22] in large studies, although smaller (and possibly underpowered) studies failed to detect such associations.[23, 24] Therefore, GG individuals would be expected to benefit more from breastfeeding than their counterparts. Indeed, a subsequent *FADS2*-breastfeeding interaction study using a larger sample obtained results consistent with this hypothesis, with the strongest association occurring in GG individuals.[20] On the other hand, three twin studies failed to detect any interaction.[25-27] One of them failed to demonstrate a dose-

response trend.[25] Another study observed a negative trend between breastfeeding and IQ at age 18, but confidence intervals were large and the same trend was not observed for educational attainment at age 12.[26]

There may be several heterogeneity sources that will be discussed below:

1. **Study design.** Several design aspects can influence results. One of such aspects is sample size, and publication bias is due to the selective publication of small studies with positive results. Sample size is particularly important for this meta-analysis because the ongoing debate relates to the association of breastfeeding and IQ among GG individuals (minor allele homozygotes), which prevalence is expected to be approximately 12.9% (rs1535) and 7.2% (rs174575) in European ancestry samples based on estimates from the 1000 Genomes Project (phase 3).

Another issue is that several of the published studies collected breastfeeding information retrospectively at different offspring ages (2 years,[26] 2-3 years,[18] 10 years,[27] 12 or 16 years,[25] and 5-33 years[26]), while one study used prospective data.[20] Retrospective measurements might be subjected to recall bias. We will evaluate the role of study design characteristics as sources of heterogeneity.

2. **Sample characteristics.** General sample characteristics may influence the results due to non-modelled interactions or different confounding structures. For example, a cross-cohort comparison evidenced that the association between breastfeeding and socioeconomic position is different between the British Avon Longitudinal Study of Parents and Children (from a high-income population) and the Brazilian 1993 Pelotas Birth Cohort (from a middle-income population).[28] Another important aspect is ethnicity because genetic epidemiology studies in multi-ethnic samples are subjected to bias from population stratification.[29] Moreover, samples from different ethnicities may differ regarding underlying linkage disequilibrium structure. In case of indirect association, this could introduce heterogeneity due to differential associations between the genotyped variant(s) with the causal variant(s) between ethnicities.[30]

Another point related to both sample characteristics and study design is twin studies. Systematic differences in breastfeeding have been observed comparing singletons and twins,[31, 32] which could limit the comparability of results. We therefore opted by limiting the meta-analysis to singletons of European ancestry. We will also investigate the contribution of other sample characteristics to between-study heterogeneity.

3. **Limited breastfeeding information.** In addition to breastfeeding prevalence, other factors such as duration and quality (eg, exclusive vs. non-exclusive) are important when studying the association of breastfeeding with any outcome of interest. Because all *FADS2*-breastfeeding interaction studies published so far used breastfeeding as a binary (never vs. ever breastfed), important information is likely being lost. For example, it is not possible to do a fair comparison using a binary breastfeeding variable when the samples greatly differ regarding average breastfeeding duration.

On the other hand, using three or more categories of breastfeeding may incur in power issues when evaluating interactions. Therefore, we will use (whenever available) more detailed breastfeeding data to gain insights such as whether there is a dose-response pattern given that power issues are likely to be reduced. We will also evaluate whether breastfeeding characteristics (eg, prevalence and duration) contribute to heterogeneity.

4. **Timing and nature of IQ measurements.** The aforementioned studies measured IQ using different tests or comprising different subtests and at different ages. These are potential sources of heterogeneity, which will be explored in our analysis. To improve numerical comparability across studies, IQ measurements will be converted to sample-specific Z-scores prior to analysis.

Study objectives

The general aim of our study is to contribute to clarify how *FADS2* variants and breastfeeding interact regarding their association with IQ. We will address this research question by conducting a collaborative meta-analysis using results from *de*

novo standardized analyses performed by collaborators using variables determined before data analysis.

Our study will test the following main hypotheses:

- The association between breastfeeding and IQ is different among GG individuals compared to non-G carriers;
- Using more detailed breastfeeding data rather than a dichotomous variable will provide additional insights (eg, whether or not a dose-response relationship exists);
- Factors associated with study design or sample characteristics are sources of between-study heterogeneity.

It is possible that *a posteriori* hypotheses based on exploratory analysis emerge. In case they occur, they will be clearly indicated as such when reporting results.

Methods/Design

Overview

The coordinating team defined the analytical plan, inclusion criteria and variables to be analysed *a priori*. The overall guideline for such definition was to properly replicate previous investigations based on a binary variable for breastfeeding (eg, [18] and [20]), as well as including additional analyses (eg, evaluation of dose-response), while adjusting for important potential confounders.

As previously described,[2] using *de novo* results in a collaborative meta-analysis has several desirable aspects. These include analysis of harmonized data using consistent analytical approaches (such as statistical tests and covariate adjustment), inclusion of unpublished data and possibility of performing secondary analysis. Statistical analysis of each individual study will be performed by its own investigators using standardized scripts developed by the coordinating team. A detailed analysis plan describing how to use the scripts provided and how they work will be distributed to the analysts.

Eligibility criteria

Studies will be considered eligible for this study if they meet all following criteria:

1. **Data availability.** The minimal data required for eligibility is:
 - Binary (never vs. ever) breastfeeding variable (either any or exclusive breastfeeding);
 - IQ measured using standard tests;
 - At least one of the two *FADS2* polymorphisms considered: rs174575 and rs1535 – both genotyped and imputed will be included.
2. **Ancestry.** To avoid population stratification and ancestry effects, only samples of European ancestry are eligible. Multi-ethnic studies will be eligible if they can identify a subsample of European ancestry. Whenever possible, such classification will be based on ancestry-informative principal components (see “Study variables” for details), although other indicators (eg, self-reported skin colour) will also be considered.
3. **Study design.** Prospective and retrospective cohort studies will be included.

Exclusion criteria for this study are:

1. **Genetic data.** The only genetic data available is imputed and its imputation quality (eg, r^2 and INFO metrics of MACH and IMPUTE, respectively [33]) is below 0.3.
2. **Study design.** Twin studies will not be included.
3. **Ethical issues.** Studies that do not have appropriate ethical approval to use their data as this study requires will be excluded.

Identifying studies

Our aim is to invite all eligible studies to participate, regardless of having published or not on this topic. Doing so will favour to achieve a larger sample size and minimize publication bias. Invitations will be sent to groups that are known by the coordinating team to have at least some of the data required available, and suggested by participating groups as possibly eligible. Although this approach is likely unspecific (ie,

we expect that some of the contacted studies are not eligible), it is useful for improving sensitivity.

Following an initial contact, the analysis plan will be distributed to studies interested in participating. This has two main goals: identify eligible studies and obtain feedback regarding the analysis plan. One or more individual studies will be invited to run preliminary analysis using the code developed by the coordinating team in order to identify and correct potential issues before distributing the code to all contributing studies.

Study variables

- 1. *Breastfeeding.*** The simplest form will be as a binary variable (never vs. ever breastfed). Whenever breastfeeding duration is available, four additional breastfeeding variables will be considered: binary (<6 months and ≥6 months) categorical (none, >0 & ≤1, >1 & ≤3, >3 & ≤6 months and >6 months), numerically-coded categorical (for linear trend tests) or numeric (in months) variable. For studies with information regarding breastfeeding quality (ie, any vs. exclusive), all breastfeeding variables will be generated twice, corresponding to each quality category.
- 2. *IQ.*** Different IQ measures that yield an approximately normally-distributed numerical variable will be included. To improve numerical comparability, such measures will be converted to sample Z-scores (ie, for each observation, subtract the mean and divide by the standard deviation). However, this does not imply in comparability regarding other aspects, such as type of test or subtests included. Since limiting based on such aspects would be too restrictive, we opted by being less stringent in this regard. The influence of such differences will be evaluated at the meta-analysis stage.
- 3. *FADS2 polymorphisms.*** We will use two variants in the *FADS2* gene: rs174575 and rs1535. Each SNP is a three-level variable, depending on how many copies an individual carries of the rarest (G) allele. The levels are: no copies of the G allele (ie,

two copies of the major allele); one copy of the G allele and one copy of the major allele (heterozygous) and two copies of the G allele (ie, homozygous G). The genotypes corresponding to each of these levels are CC, CG and GG for rs174575; and AA, AG and GG for rs1535.

G is expected to be the rarest allele in Europeans samples, with a frequency of about 25.5% and 35.0% for rs174575 and rs1535, respectively. Importantly, since C pairs with G, strand-orientation issues related to the rs174575 variant can only be detected by comparing observed with expected allele frequencies. As a quality control check, the analysis script will stop if the G-allele frequency is outside the range of 10% to 40%. Both genotyped and imputed SNPs will be considered. If imputed, dosages corresponding to the G allele rather than “best-guess” genotypes will be used.

Each polymorphism will be coded in four different forms, reflecting distinct genetic effects: additive or per-allele, corresponding to the number of copies of the G allele (AA or CC=0, AG or CG=1, GG=2); dominant, where G-allele carriers are compared to non-G-carriers (AA or CC=0, AG/GG or CG/GG=1); recessive, where GG individuals are compared to A- or C-allele carriers (AA/AG or CC/CG=0, GG=1); and overdominant, where heterozygous are compared to homozygous individuals (AA/GG or CC/GG=0, AG or CG=1).

4. Covariates. This study will include the following covariates:

- Sex (male/female);
- Age at IQ measurement (in years) and age² (to account for potential non-linear age effects);
- Ancestry-informative principal components[34] for studies with genome-wide genotyping data available. Such components (calculated within the European subsample using a subset of independent SNPs of minor allele frequency > 1%) will be used to account for residual population stratification.
- Measures of maternal education or maternal cognition. To achieve international comparability, maternal education will be coded according to the 1997 International Standard Classification of Education (ISCED) of the United Nations

Educational, Scientific and Cultural Organization.[35] To improve numerical comparability across studies (relevant to sensitivity analysis), maternal cognition will be converted to sample Z-scores. In studies that measured these variables more than once (ie, at different time points), the closest time point to offspring birth will be used. Adjusting for these variables will be performed similarly to age at IQ measurement to account for potential non-linear effects.

- A categorical indicator of field centre for multi-centric studies. This will be used to account for eventual batch effects.
- Any other recommended study-specific indicators, if considered necessary by the coordinating team.

Statistical analysis

1. **Overview and pre-analysis steps.** The scripts were written in R (www.r-project.org) due to its free availability and widespread use. Two scripts were produced. One is called “user’s script” and is aimed at being used by the analysts. It contains less than 200 lines of code, and the vast majority are comment lines explaining how to conduct each step with examples. The other is called “developer’s script”, which contains the actual functions that will perform quality control checks, calculate summary statistics and perform association analysis in more than 1000 lines of code. By providing a simplified script that uses more complicated functions from an accompanying script, we hope to reduce the work burden of contributing studies.

To ensure consistency across studies, only the coordinating team will make any eventual modifications in the developer’s script. So, in case an analyst identifies an issue, it will be reported to the coordinating team who will make any revisions if necessary and re-distribute the code.

The main task of the analysts will be to format the data for the analysis. The analysis plan will contain detailed instructions on how the data should be formatted. To minimize harmonization issues, the first step of the analysis will be a series of quality control checks regarding general data formatting, eligibility criteria, categorical variable levels, outliers (defined as being outside the range of ± 4

standard deviation from the mean) and impossible numbers (eg, negative IQ points) in continuous variables. After the quality control step, summary statistics for the sample and for the SNPs will be generated. These will be used at the meta-analysis stage to identify potential heterogeneity sources.

2. **Association analysis.** Association analysis will be performed by linear regression with heteroskedasticity robust standard errors. The main statistical model underlying all analysis is:

$$IQ = \beta_0 + \beta_1 BF + \beta_2 FADS2 + \beta_3 (BF \times FADS2) + \sum_{i=4}^{n+3} \beta_i cov_{i-3} + \sum_{i=n+4}^{2n+3} \beta_i (BF \times cov_{i-3-n}) + \sum_{i=2n+4}^{3n+3} \beta_i (FADS2 \times cov_{i-3-2n})$$

, where:

BF: breastfeeding (any or exclusive) as a binary, categorical, numerically-coded categorical variable or numeric (in months) variable.

FADS2: *FADS2* polymorphism (rs174575 or rs1535) coded in additive or recessive model.

cov: generic representation of a covariate.

n: Number of covariates included in the analysis.

Given that all analysis will be performed three times (unadjusted and two adjusted models), up to 240 regression analysis will be performed. For studies that meet the minimal eligibility criteria, this number will be 12. The potential confounding effect of covariates on the interaction between breastfeeding and *FADS2* will be properly modelled by including interaction terms of breastfeeding and *FADS2* polymorphism with each covariate.[36]

3. The primary analysis will use any breastfeeding in binary form and a recessive genetic model in unadjusted and adjusted models. This corresponds to a replication of the main analysis performed by Caspi and colleagues[18] and Steer and colleagues.[20] The remaining analyses are aimed at further exploring the *FADS2*-

breastfeeding interaction by evaluating different genetic models and whether or not there are dose-response breastfeeding effects. **Covariate adjustment.**

Regarding covariate adjustment, three analysis will be performed:

- Unadjusted (model 1);
- Adjusted for sex, age and age². Multi-centric studies or studies with genome-wide genotyping data available will also control for field centre or ancestry-informative principal components, respectively (model 2);
- Adjust for the same covariates listed above, and also for maternal education and (maternal education)², and/or maternal cognition and (maternal cognition)² (model 3).

4. **Meta-analysis.** Descriptive statistics will be checked for potential errors, which will be corrected before conducting the meta-analysis. We will then conduct a preliminary analysis to evaluate if there is heterogeneity due to a few studies; if so, the coordinating team will contact these studies individually for identification of potential errors or problems. In case no issues are identified, the study(ies) will be included in the meta-analysis.

After checking for these potential sources of artificial heterogeneity, we will then conduct the final meta-analysis. We will report both fixed- and random-effects, and use meta-regression to evaluate the following sources of heterogeneity: age, prevalence and duration of breastfeeding, retrospective vs. prospective breastfeeding information, measures of IQ, adjustment for principal components and continental region. The main statistics that we will report are the pooled linear regression coefficients for breastfeeding (corresponding to the effect among individuals in the baseline *FADS2* genotype), *FADS2* (corresponding to the effect among never breastfed individuals) and *FADS2*-breastfeeding interaction. We will also report heterogeneity statistics and subgroup-specific estimates, as well as descriptive statistics from each contributing study.

5. **Sensitivity analysis.** We will compare overall meta-analytical estimates with results obtained using subsets of all studies. In case heterogeneity is detected, we will also report estimates for homogeneous subgroups in order to understand if some

sources of heterogeneity could be attributed to bias. For example, subsetting based on sample size or length of recall of information on breastfeeding duration may yield insights on the influence of publication or recall bias (respectively) in the estimates.

To explore the possibility of bias due to gene-environment correlation, we will repeat *FADS2*-breastfeeding interaction analysis having maternal education (converted to US years of education based on ISCED standards, as reported previously[37]) and maternal cognition as the outcome variable. Since only models 1 and 2 will be performed for these outcomes, there will be 160 regression analyses for each. Added to the 240 analyses for IQ, *de novo* results from 560 regression analyses (performed automatically by the scripts provided) will be obtained from studies that contribute to all analyses.

Sample size calculation

Sample size requirements to detect a *FADS2*-breastfeeding interaction were evaluated through simulations (5,000 simulations per combination of parameters) using R version 3.2.4.

The following parameters were evaluated:

- a) Prevalence of ever being breastfed: 85% and 95%. These values are based on the estimates recently provided by Victora and colleagues[4] for high-income countries and for countries in all other income groups, respectively.
- b) Prevalence of the GG genotype: 7.2% and 12.9%. These values were obtained from the 1000 Genomes (phase 3) Project Browser for the rs174575 and rs1535 SNPs (respectively) in European populations.
- c) Mean difference in IQ according to *FADS2* polymorphism among never breastfed individuals, comparing GG individuals with non-G carriers: -2.15, -4.3 and -8.6. The intermediate value (-4.3) correspond to the results from Steer and colleagues,[20] which the largest study that evaluated the *FADS2*-

breastfeeding interaction on IQ to date. The remaining values correspond to half and twice of the effect reported by Steer et al. and were used to evaluate sample size requirements in case of weaker and stronger *FADS2* effects.

- d) Mean difference in IQ according to *FADS2* polymorphism among ever breastfed individuals: zero and half of the effect in the never breastfed group. Lack of *FADS2* effect among ever breastfed individuals correspond to the DHA hypothesis described above.
- e) Sample size (10,000, 12,500, 15,000, 17,500 and 20,000 individuals).

All possible combinations of the above parameters correspond to 120 simulation scenarios. In all of them, the outcome variable was normally distributed (mean=100 and standard deviation=10) and *FADS2* and breastfeeding were independent. P-values for the interaction coefficient were obtained from linear regression models (two-sided T-tests). Power was defined as the proportion of tests with P-values<0.05.

Among the 120 simulation scenarios, power was <80% in only 7 of them. The most critical scenario was when breastfeeding prevalence was 95%, GG prevalence was 7.2%, *FADS2* effect among never breastfed individuals was -2.15 and the effect among ever breastfed individuals was half of the latter (power=77.3% for a sample size of 20,000 individuals). When sample size was up to 12,500 individuals, power was also <80% when GG prevalence was 12.9%.

It is important to consider that none of the scenarios were underpowered when breastfeeding prevalence was 85%. This estimate is likely to apply to this study better than the value of 95% given that our focus is on individuals of European ancestry (therefore, samples from high-income countries are more likely to be eligible). Moreover, none of the scenarios was underpowered when *FADS2* effect was at least equal to the effect reported by Steer and colleagues (which is the best estimate currently available), as well as when there was no *FADS2* effect among ever breastfed individuals.

Therefore, in the majority of realistic scenarios, a sample size of 10,000 individuals would allow properly-powered primary analysis. Based on a preliminary identification of eligible studies, achieving such sample size is feasible.

Ethics statement

Only studies with appropriate ethical approval will be considered to participate. Only summary-level statistics (rather than individual-level data) will be shared between the individual study and the coordinating team. Therefore, the present study does not require additional ethical approval other than what has already been provided to participating studies individually. We will obtain all necessary institutional approvals to conduct the analysis.

Discussion

This collaborative meta-analysis has the potential to improve the understanding of the effect modification of *FADS2* variants on the association between breastfeeding and IQ. However, the study has some limitations.

To achieve a larger sample size and allow participation of different studies, some compromises are necessary. Particularly, we will include breastfeeding measures with different recall times, as well as IQ measures that differ regarding test, subtests included and/or age at measurement. Although a large sample size will contribute to minimize limitations due to heterogeneity (which will also be evaluated in detail), such inconsistencies might still influence the results.

Second, the analysis will be limited to singletons of European ancestry. This will likely reduce heterogeneity (eg, due to systematic differences in breastfeeding patterns comparing twins to singletons) and bias (eg, due to population stratification). Moreover, most genetic epidemiology studies to date have been conducted in Europeans, so it is unlikely that restricting to Europeans will incur in substantial sample size losses. However, it may limit the external validity of our findings.

Third, several heterogeneity tests will be performed. However, it is difficult to identify all potential sources of heterogeneity. Moreover, it may occur that, in some cases,

subsetting studies based on heterogeneity-associated factors result in small subgroups, thus yielding imprecise subgroup-specific estimates.

Fourth, availability of maternal education or maternal cognition measures was not included as one eligibility criterion. Although we recognize the importance of accounting for these variables in studies involving breastfeeding and IQ, we opted by allowing studies without these data to participate for two main reasons. First, it is likely that requiring these data would substantially reduce the sample size. Second, previous publications observed no major implication of such measures on *FADS2*-breastfeeding interaction.[18, 20] Therefore, we opted by an inclusive approach coupled with sensitivity analyses using the subset of studies with these data.

Finally, based on sample size calculations under a variety of realistic situations, we expect to have enough power to detect interaction effects. However, a lack of strong statistical association could be a result of small effects and/or heterogeneity that we fail to account for. Moreover, given the inconsistencies among published studies and the fact that we will properly control for confounding in the interaction setting, it is also possible that our meta-analysis suggests that there is no *FADS2*-breastfeeding interaction (although such strong conclusion might not be feasible due to sample size limitations).

Understanding the health effects – and associated mechanisms – of breastfeeding is important to obtain a more accurate view of the impact of breastfeeding promotion. This, in turn, may have implications regarding the extent to which investments on such promotion should be prioritized over other public health initiatives. Identifying the mechanisms could also be important to incorporate key nutritional components of breast milk into formula milk.

Regarding effect modification (if any) of *FADS2* variants on the association between breastfeeding and IQ, individual studies published to date are inconsistent. Improving current understanding of this interaction might yield biological insights regarding the importance of LC-PUFAs for breastfeeding effects. This research question will be addressed using a collaborative meta-analysis based on consistent *a priori* defined

analysis of harmonized data. Therefore, publishing this protocol will reduce potential biases associated with data mining, thus contributing to generate reliable evidence.

Footnotes

Contributors. GDS and CGV conceived the study. The manuscript was drafted by FPH and ND, and was revised by BLH, GDS and CGV. GDS will contact individual studies to participate. FPH will perform the statistical analysis of *de novo* results obtained from each individual study. All authors will critically revise and interpret the results. All authors approved the publication of the protocol.

Funding statement. This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors. NMD is supported by the Economics and Social Research Council (ESRC) via a Future Research Leaders Fellowship [ES/N000757/1]. The Integrative Epidemiology Unit is supported by the MRC and the University of Bristol (MC_UU_12013/1,9).

Competing interests. None.

References

1. Ioannidis JP. How to make more published research true. *PLoS Med* 2014;11(10):e1001747.
2. Culverhouse RC, Bowes L, Breslau N, et al. Protocol for a collaborative meta-analysis of 5-HTTLPR, stress, and depression. *BMC Psychiatry* 2013;13:304.
3. WHO Collaborative Study Team on the Role of Breastfeeding on the Prevention of Infant Mortality. Effect of breastfeeding on infant and child mortality due to infectious diseases in less developed countries: a pooled analysis. *Lancet* 2000;355(9202):451-5.
4. Victora CG, Bahl R, Barros AJ, et al. Breastfeeding in the 21st century: epidemiology, mechanisms, and lifelong effect. *Lancet* 2016;387(10017):475-90.
5. World Health Organization. The Optimal Duration of Exclusive Breastfeeding. Geneva, Switzerland: World Health Organization: 2001.

6. World Health Organization and UNICEF. Protecting, Promoting and Supporting Breastfeeding: The Special Role of Maternity Services. Geneva, Switzerland: 1989.
7. Horta BL, Bahl R, Martines JC, et al. Long-term effects of breastfeeding: a systematic review. Geneva, Switzerland: World Health Organization 2013.
8. Horta BL, de Mola CL, Victora CG. Long-term consequences of breastfeeding on cholesterol, obesity, systolic blood pressure, and type-2 diabetes: systematic review and meta-analysis. *Acta Paediatr* 2015.
9. Horta BL, de Mola CL, Victora CG. Breastfeeding and intelligence: systematic review and meta-analysis. *Acta Paediatr* 2015.
10. Walfisch A, Sermer C, Cressman A, et al. Breast milk and cognitive development--the role of confounders: a systematic review. *BMJ Open* 2013;3(8):e003259.
11. Dewey KG, Cohen RJ, Brown KH, et al. Effects of exclusive breastfeeding for four versus six months on maternal nutritional status and infant motor development: results of two randomized trials in Honduras. *J Nutr* 2001;131(2):262-7.
12. Kramer MS, Aboud F, Mironova E, et al. Breastfeeding and child cognitive development: new evidence from a large randomized trial. *Arch Gen Psychiatry* 2008;65(5):578-84.
13. Singhal A, Cole TJ, Fewtrell M, et al. Breastmilk feeding and lipoprotein profile in adolescents born preterm: follow-up of a prospective randomised study. *Lancet* 2004;363(9421):1571-8.
14. Singhal A, Cole TJ, Lucas A. Early nutrition in preterm infants and later blood pressure: two cohorts after randomised trials. *Lancet* 2001;357(9254):413-9.
15. Victora CG, Horta BL, Loret de Mola C, et al. Association between breastfeeding and intelligence, educational attainment, and income at 30 years of age: a prospective birth cohort study from Brazil. *Lancet Glob Health* 2015;3(4):e199-205.

16. Koletzko B, Agostoni C, Carlson SE, et al. Long chain polyunsaturated fatty acids (LC-PUFA) and perinatal development. *Acta Paediatr* 2001;90(4):460-4.
17. Isaacs EB, Fischl BR, Quinn BT, et al. Impact of breast milk on intelligence quotient, brain size, and white matter development. *Pediatr Res* 2010;67(4):357-62.
18. Caspi A, Williams B, Kim-Cohen J, et al. Moderation of breastfeeding effects on the IQ by genetic variation in fatty acid metabolism. *Proc Natl Acad Sci U S A* 2007;104(47):18860-5.
19. Sprecher H. Metabolism of highly unsaturated n-3 and n-6 fatty acids. *Biochim Biophys Acta* 2000;1486(2-3):219-31.
20. Steer CD, Davey Smith G, Emmett PM, et al. FADS2 polymorphisms modify the effect of breastfeeding on child IQ. *PLoS One* 2010;5(7):e11570.
21. Schaeffer L, Gohlke H, Muller M, et al. Common genetic variants of the FADS1 FADS2 gene cluster and their reconstructed haplotypes are associated with the fatty acid composition in phospholipids. *Hum Mol Genet* 2006;15(11):1745-56.
22. Tanaka T, Shen J, Abecasis GR, et al. Genome-wide association study of plasma polyunsaturated fatty acids in the InCHIANTI Study. *PLoS Genet* 2009;5(1):e1000338.
23. Rzehak P, Heinrich J, Klopp N, et al. Evidence for an association between genetic variants of the fatty acid desaturase 1 fatty acid desaturase 2 (FADS1 FADS2) gene cluster and the fatty acid composition of erythrocyte membranes. *Br J Nutr* 2009;101(1):20-6.
24. Gieger C, Geistlinger L, Altmaier E, et al. Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genet* 2008;4(11):e1000282.
25. Martin NW, Benyamin B, Hansell NK, et al. Cognitive function in adolescence: testing for interactions between breast-feeding and FADS2 polymorphisms. *J Am Acad Child Adolesc Psychiatry* 2011;50(1):55-62 e4.

26. Groen-Blokhuis MM, Franic S, van Beijsterveldt CE, et al. A prospective study of the effects of breastfeeding and FADS2 polymorphisms on cognition and hyperactivity/attention problems. *Am J Med Genet B Neuropsychiatr Genet* 2013;162B(5):457-65.
27. Rizzi TS, van der Sluis S, Derom C, et al. FADS2 Genetic Variance in Combination with Fatty Acid Intake Might Alter Composition of the Fatty Acids in Brain. *PLoS One* 2013;8(6):e68000.
28. Brion MJ, Lawlor DA, Matijasevich A, et al. What are the causal effects of breastfeeding on IQ, obesity and blood pressure? Evidence from comparing high-income with middle-income cohorts. *Int J Epidemiol* 2011;40(3):670-80.
29. Smith GD, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* 2003;32(1):1-22.
30. Costas J, Torres M, Cristobo I, et al. Relative efficiency of the linkage disequilibrium mapping approach in detecting candidate genes for schizophrenia in different European populations. *Genomics* 2005;86(3):280-6.
31. Yokoyama Y, Wada S, Sugimoto M, et al. Breastfeeding rates among singletons, twins and triplets in Japan: A population-based study. *Twin Res Hum Genet* 2006;9(2):298-302.
32. Flidel-Rimon O, Shinwell ES. Breast feeding twins and high multiples. *Arch Dis Child Fetal Neonatal Ed* 2006;91(5):F377-80.
33. Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* 2010;11(7):499-511.
34. Price AL, Patterson NJ, Plenge RM, et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006;38(8):904-9.
35. UNESCO Institute for Statistics. International Standard Classification of Education. Paris, France: United Nations Educational, Scientific and Cultural Organization: 2006

36. Keller MC. Gene x environment interaction studies have not properly controlled for potential confounders: the problem and the (simple) solution. *Biol Psychiatry* 2014;75(1):18-24.
37. Rietveld CA, Medland SE, Derringer J, et al. GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* 2013;340(6139):1467-71.

Anexo II – Plano de análise do estudo de interação entre *FADS2* e amamentação

ANALYSIS PLAN – 09/04/2016

Meta-analysis of effect modification of *FADS2* polymorphisms on the association between breastfeeding and intelligence

Deadline for sending results: April 09, 2016

1) Research question

Research question: do polymorphisms in the *FADS2* gene modify the effects of breastfeeding on child intelligence?

2) Eligibility criteria

Studies will be considered eligible for this study if they meet all following criteria:

- a) Data on breastfeeding, *FADS2* polymorphisms and intelligence quotient (IQ) – other intelligence/cognitive measures may also be considered.
- b) To avoid population stratification and ancestry effects, only samples of European ancestry are eligible. Multi-ethnic studies can still contribute as long as they can identify a subsample of European ancestry (see section 3.4 for details).
- c) Prospective and retrospective cohort studies.

Exclusion criteria are:

- a) Only poorly-imputed genetic data is available. Imputation quality (eg, r^2 and INFO metrics of MACH and IMPUTE, respectively) should be > 0.3 .
- b) Twin studies.
- c) Unavailability of proper ethics approval.

3) Data

Description of the data that will be used in this project is provided below. Instructions to format your data for analysis are provided in section 4.1).

3.1) Breastfeeding

Two definitions of exposure: binary, categorical and continuous.

- a) Binary: ever breastfed or never breastfed (required for eligibility).
- b) Continuous variable: any breastfeeding in months.

NOTE: If you have information for both any (ie, exclusive or non-exclusive) and exclusive breastfeeding, please provide both separately. See section 4.1) for details.

3.2) FADS2 polymorphisms

We will use two variants in the FADS2 gene:

- a) rs174575 (major/minor allele: C/G)
- b) rs1535 (major/minor allele: A/G)

- To be eligible, studies must have data on at least one SNP.

Each SNP is a three-level variable, depending on how many copies an individual carries of the rarest allele. For rs174575, the three levels are CC (two copies of C: known as 'homozygous C'), CG (1 copy of C, known as 'heterozygous') and GG (no copies of C, known as 'homozygous G'). For rs1535, the three levels are AA, AG and GG. When formatting your data (see section 4.1), please use "AG" and "CG" for heterozygotes rather than "GA" or "GC".

Since C pairs with G, strand-orientation issues associated with the rs174575 variant can only be detected by comparing observed with expected allele frequencies. In samples drawn from European populations, G is expected to be the rarest allele, with a frequency of about 23.3% and 34.9% for rs174575 and

rs1535 respectively. If, in your sample, the rarest allele is A (rs174575) or C (1535) and its frequency is similar to the expected frequency of the G allele, it is possible that genotypes were called in the negative strand in your study. If that's the case, just "flip" the alleles.

NOTE: Imputed SNPs will be considered if imputation quality (eg, r^2 and INFO metrics of MACH and IMPUTE, respectively) > 0.3 . In this case, provide dosages corresponding to the 'G' allele (ie, CC=0, CG=1 and GG=2 and AA=0, AG=1 and GG=2 for rs174575 and rs1535, respectively). If they correspond to the 'non-G' allele, subtract the dosages from 2 (ie, 2 - dosages). If imputation was not performed using MACH, convert imputed data into MACH-like dosages. If you need assistance with this, please contact fernandophartwig@gmail.com.

3.3) Outcome variable

IQ in points. Different measurements will be accepted in principle (required for eligibility).

In longitudinal studies, outcome variables might have been measured more than once. Choose a single visit that maximizes the sample size. Be careful to not include any individuals more than once.

3.4) Covariates (see section 4.1)

- a) Sex (male or female);
- b) Age in years (continuous);
- c) Ancestry-informative principal components (PCs) for samples with genome-wide genotyping data available. Calculate the top 10 PCs and include as many PCs as necessary to account for residual substructure within Europeans in your sample. To calculate PCs, use a subset of independent SNPs ($LD < 0.3$) of minor allele frequency > 0.01 .

- d) An indicator of ancestry. It can be based on PCs or, if not available, in other indicators such as self-reported skin colour. Please use the following levels: 'european', 'african', 'asian', 'hispanic', 'other'.
- e) An indicator of field centre (for multi-centre studies). Please use the following levels: 'fc1', 'fc2', ... 'fcN' (N=number of field centres).
- f) Maternal education. To allow international comparability, recode maternal education according to the 1997 International Standard Classification of Education (ISCED) of UNESCO.
To do so, search and download at <http://www.uis.unesco.org/Education/ISCEDMappings/Pages/default.aspx> a ".xls" file containing detailed instructions for the country corresponding to your sample. If you have more than one measure, use the closest one to offspring birth.
- g) Maternal cognition (continuous). Different measures that yield a continuous variable will be accepted in principle. If you have more than one measure, use the closest one to offspring birth.

NOTE: Please provide PCs (if genome-wide genotyping data is available) and an indicator of ancestry even if your sample is entirely of European ancestry.

4) Data analysis

Individual studies are asked to perform all analyses using code written using R (www.r-project.org/) we provided. There are three main reasons for this: reduce the chance of errors; increase comparability; reduce burden of work by the analyst.

Although some groups may be more familiar with other statistical software, we opted by using R because it is available for free. For those who do not have R installed, it is straightforward to do so by following the instructions in the website.

The analysis can be divided into five steps: Data preparation, User input, Data formatting check, Summary statistics and Association analysis. Details of each are provided below.

4.1) Data preparation

This is the only step where using R is optional. The aim is to generate a file that will be loaded into R for the later steps. Detailed instructions are provided in the “FADS2 x BF interaction on intelligence - Data formatting guidelines 22102015.xlsx” spreadsheet.

NOTE: Please follow formatting instructions strictly!

4.2) User input

From this section one, you should work on the “FADS2 x BF interaction on intelligence - Code for users 20160409.R” file after opening it in R. All steps pertaining to “User input” start with the letter A, going from A1 to A11.

Please follow the instructions (with examples) provided along this section to properly provide the information the codes need to run the analysis.

NOTES: This is the only section where users should modify the code. And please do so only when it is indicated, following the instructions.

4.3) Data formatting check

From this section on, the code provided in “Code for users” will use functions contained in the “FADS2 x BF interaction on intelligence - Functions 20160409.R” file, which is not intended to be modified by the users. If you identify errors or have questions, please inform us so we can verify, make corrections if necessary and re-distribute the code to all studies.

Since data formatting is critical, this step performs some checks attempting to identify formatting errors, so these can be fixed before the data analysis stage. The code will also check if you have provided input.

To run the data checks, just run the single-line code provided in section B of “Code for users”. Six major checks will be performed:

1. General formatting checks

Checks number and names of columns

2. Eligibility check

Evaluates eligibility criteria describe in the analysis plan

3. Categorical variables check

Checks if categorical variables present the correct levels

4. Continuous variables check

Checks if continuous variables present impossible values (eg, negative breastfeeding duration) and the presence of outliers (be outside the range of ± 4 standard deviations from the mean).

5. Consistency of breastfeeding variables check

Consistency of the breastfeeding data provided. For example, are all individuals positive for exclusive breastfeeding also positive for any breastfeeding?

6. G-allele frequency check

Checks if G-allele frequencies of the provided SNPs are within 10%-40%.

This code will produce several messages to the user along it executes its tasks. Please pay attention to these messages. There are two main types of messages that will appear:

1. Error messages (in red)

These messages appear when there are errors that the code cannot handle. If they appear, this means that the function stopped at the point where the error occurred and will not continue from there. There are two types of error messages in this code:

a. ## error messages

Messages in the form “Error: ## <number>.<number>: ## <message> ##” (eg, “Error: ## 6.4 rs1535_imp G-allele frequency lies outside the 10%-40% range! ##”) indicate formatting errors identified by the code. The function was programmed to issue these errors with intelligible information about the problem, so it will be easier to fix.

b. other error messages

Messages in red that do not present the format described above indicate errors that R encountered when running the code, rather than problems the function was programmed to look for.

2. Non-error messages (in black)

These messages appear to inform the user about something. They are not error messages, but they are also important. There are two types of non-error messages:

a. Progress messages

These messages appear after the function completes a data formatting check stage. When coupled with error messages, this is very useful to help the user to identify where a problem occurred. For example, if an error message appears right after the progress message “#The data passed general formatting checks! #”, this means that the error occurred in the next formatting check (in this case, it would be Step 2. Eligibility check).

b. Non-fatal warnings

The code might eventually issue messages in the form “## NON-FATAL WARNING: <number>.<number>: <message> ##”. These indicate situations that are not necessarily errors, but worth-informing to the user. For example: if the program identifies outliers, it will issue a non-fatal warnings message indicating the outlying values. It is a task of the user to decide whether these are correct or not, and, if not, correct them or set them to missing.

The code ends (in case of no errors) with a progress message.

NOTE: Please carefully check all the output provided during data formatting check – and make any corrections if necessary – before proceeding to the next step. If you make any corrections in the data, please submit the entire dataset to data formatting check again.

4.4) Summary statistics

The next step of the code generates summary statistics. To generate summary statistics, just run the single-line coded provided in section C of “Code for users”.

Two files will be generated: one with summary statistics of the sample and another with information about the SNPs.

Sample summary statistics

These are provided separately for individuals with non-missing values for each outcome (IQ and educational attainment) separately, and limited to individuals of European ancestry. For quantitative variables, the following statistics are calculated: minimum, maximum, mean, standard deviation, median and interquartile range. For categorical variables, number of individuals in each category is obtained.

SNP information

For each SNP, the information (also for each outcome) will be: exact P-value for the Hardy-Weinberg equilibrium, if the SNP was genotyped or imputed (if so, the imputation quality) and minor allele (ie, G) frequency.

These files will be saved in the directory indicated by the user in “User input” section. The name of the files will also use information from this section. Please do not change the names of the files nor their contents.

NOTE: Please check the contents of these files to see if they seem OK based on your knowledge of your study before going to “Association analysis”.

The files (including the file that will be generated after running association analysis) will be automatically named as follows:

FADS2_OUTPUT_TYPE_type_STUDY_study_id_DATE_date.txt

type: indicates the contents of the file. This will be “sample_descriptives”, “snp_descriptives” or “association_results”.

study_id: an identifier of your study, provided in the “User input” section.

date: the date when the files were generated, as provided in the “User input” section.

4.5) Association analysis

All association analysis will be performed by linear regression with heteroskedasticity robust standard errors. To perform association analyses and generate a file with the results, just run the single-line coded provided in section D of “Code for users”.

The code will perform models of the form:

Crude:

outcome = breastfeeding + FADS2 + FADS2*breastfeeding

Adjusted 1:

outcome = breastfeeding + FADS2 + FADS2*breastfeeding + sex + age + age² + PCs + study centre + (sex + age + age² + PCs + study centre)*breastfeeding + (sex + age + age² + PCs + study centre)*FADS2

Adjusted 2:

outcome = breastfeeding + FADS2 + FADS2*breastfeeding + sex + age + age² + PCs + study centre + maternal education + maternal education² + maternal cognition + maternal cognition² + (sex + age + age² + PCs + study centre + maternal education + maternal education² + maternal cognition + maternal cognition²)*breastfeeding + (sex + age + age² + PCs + study centre + maternal

education + maternal education² + maternal cognition + maternal cognition²)*FADS2

NOTE: If PCs, study centre, maternal education and/or maternal cognition were not provided, (ie, genome-wide genotyping data is not available and/or the study was not multi-centric) the analysis will still run properly.

outcome: IQ (main analysis), maternal education converted to US years of education or maternal cognition (sensitivity analyses).

breastfeeding: four different breastfeeding variables for each quality (any and exclusive) of breastfeeding: binary (no=0, yes=1); binary (<6 months=0; ≥6 months=1); categorical (0=none, 1=any up to one month, 2=more than one month and less than three months, 3=more than 3 months up to six months, 4=more than six months) numerically coded (for linear trend); continuous (months of duration). In total, there will be eight different breastfeeding variables.

FADS2: rs174575 and rs1535, each coded in four different models: additive (number of copies of the 'G' allele), dominant (non-G homozygotes=0, G carriers=1), recessive (non-G carriers=0, GG=1) and overdominant (homozygous genotypes=0; heterozygotes=1). In total, there will be eight FADS2 variables.

All combinations of (eight) breastfeeding variables and (eight) FADS2 variables result in 64 regression analysis. When IQ is the outcome, each of these will be tested in unadjusted and two adjusted models, in a total of 192 analyses. For maternal education and cognition, there will be only one adjusted model, resulting in 128 analyses for each. Therefore, 192 + 128 + 128 = 448 regression analyses will be performed in total.

NOTE: As long as the study meets the eligibility criteria, the code will run properly by automatically skipping analysis that would not be possible to be performed.

5) Meta-analysis

Descriptive statistics will be checked for potential errors before conducting meta-analysis; if there are, we will contact these studies individually for discussion. We will

then conduct a pre-meta-analysis. If substantial heterogeneity is identified, we will check if this is due to a few studies; if so, we will contact these studies individually for discussion.

After checking for these potential sources of artificial heterogeneity, we will then conduct the final meta-analysis. We will report both fixed- and random-effects.

6) Authorship conditions

Up to three co-authors from each individual study.

Anexo III – Código do analista para as análises do estudo de interação entre *FADS2* e amamentação

```
#####
#-----#
#Meta-analysis of effect modification of FADS2 polymorphisms on the association between
breastfeeding and intelligence-#
#-----#
#####

#####
#####
##INTRODUCTION##
#####
#####

#Thanks for contributing to this initiative! Go through this file to run the analyses required.
#The analysis are described in detail in the analysis plan (FADS2 x BF interaction on intelligence -
Analysis plan 20150522.docx)
#Please follow all instructions provided in the analysis plan and in this file. This will minimise the chance
of errors and quality control issues.
#Although the code does some checks, it is essential that the data analyst to be careful about the quality
of the data and to follow formatting instructions provided (FADS2 x BF interaction on intelligence - Data
formatting guidelines 20150513)
#Only modify the code in section A where indicated. In the remaining sections, please DO NOT modify
any of the code.
#Please DO NOT modify the code provided in "FADS2 x BF interaction on intelligence - Functions
20150522.R" file.
#In case of questions or errors, contact Fernando Pires Hartwig (fernandophartwig@gmail.com).

#####
#####
##A) USER INPUT##
#####
#####

#A1) Clean the working environment
rm(list=ls())

#A2) Install - if necessary - and load required packages
if(require('sandwich')==F) {install.packages('sandwich'); library('sandwich')} #necessary to obtain
heteroscedastic-robust standard errors
if(require('lmtest')==F) {install.packages('lmtest'); library('lmtest')} #necessary to obtain heteroscedastic-
robust standard errors
if(require('genetics')==F) {install.packages('genetics'); library('genetics')} #easy calculation of HWE P-
value

#A3) Load functions provided in the "FADS2 x BF interaction on intelligence - Functions 20160409.R" file.
#To do this, provide the full filename of this file by replacing 'functions_filename' below.
#For example: the full filename of this file is "C:/Users/User1/Desktop/FADS2 x BF interaction on
intelligence - Functions 20160409.R".
```

#In this case, the code below would be `load('C:/Users/User1/Desktop/FADS2 x BF interaction on intelligence - Functions 20160409.R')`
`load('functions_filename')`

#A4) Provide an identifier of your study by replacing 'your_study_id' below.

#Please use letters or numbers only.

#For example: the 1982 Pelotas Birth Cohort Study could be identified using "1982PELOTAS".

#In this case, the code below would be `study_id <- '1982PELOTAS'`

`study_id <- 'your_study_id'`

#A5) Provide the full path to the directory where you want to save the files resulting from this script.

#All files generated by this code start with 'FADS2_OUTPUT_'.

#For example: the full path to the data is "C:/Users/User1/Desktop/data.txt".

#In this case, the code below would be `path <- 'C:/Users/User1/Desktop/'`

`path <- 'path_to_data'`

#A6) Provide the date when you performed the analysis in DDMMYYYY format by replacing 'DDMMYYYY' below.

#For example: the analyses were completed on November 10, 2015.

#In this case, the code below would be `date <- '10112015'`

`date <- 'DDMMYYYY'`

#A7) IF LEAST ONE OF THE SNPS WAS IMPUTED, inform the software by replacing 'your_imp_software' below.

#Possible values: GENOTYPED, BEAGLE, IMPUTE2, MACH, etc

#For example: you used IMPUTE2 for imputation.

#In this case, the code below would be `imp_software <- 'IMPUTE2'`

#For example: you haven't done imputation

#In this case, the code below would be `imp_software <- 'GENOTYPED'`

`imp_software <- 'your_imp_software'`

#A8) IF rs174575 WAS IMPUTED, provide imputation quality of this SNP by replacing NULL below.

#IF rs174575 WAS NOT IMPUTED, do not replace the NULL below (but still run the code).

#For example: imputation quality of rs174575 was 0.8.

#In this case, the code below would be `imp_quality_rs174575 <- 0.8`

`imp_quality_rs174575 <- NULL`

#A9) IF rs1535 WAS IMPUTED, provide imputation quality of this SNP by replacing NULL below.

#IF rs1535 WAS NOT IMPUTED, do not replace the NULL below (but still run the code).

#For example: imputation quality of rs1535 was 0.8.

#In this case, the code below would be `imp_quality_rs1535 <- 0.8`

`imp_quality_rs1535 <- NULL`

#A10) Indicate whether or not your study is multi-centric (ie, data generated in different centres) by replacing 'multi_centric_info' below.

#Use 'yes' or 'no' to indicate if your study is or isn't multi-centric (respectively).

#For example: if your study is multi-centric, the code below would be `multi_centric <- 'yes'`

`multi_centric <- 'multi_centric_info'`

#A11) Load your data.

#To do this, provide the full filename of your data by replacing 'your_data_filename' below

#For example: the full filename of your data is "C:/Users/User1/Desktop/data.txt".

#In this case, the code below would be `data <- NULL; data_filename <- 'C:/Users/User1/Desktop/'; data <- read.table(data_filename, header=T, sep='\t')`

`data <- NULL; data_filename <- 'your_data_filename'; data <- read.table(data_filename, header=T, sep='\t')`

```
#####
#####
#####
###-----###
###-----###
###FROM THIS STAGE ON, YOU ARE NOT REQUIRED TO PROVIDE ANY ADDITIONAL INFORMATION. SO,
PLEASE, DO NOT CHANGE THE CODE.###
###-----###
###-----###
#####
#####
#####
#####
#####
#####
##B) DATA FORMATTING CHECK##
#####
#####
```

#This section uses the data_check() function to do some checks regarding data formatting.
 #Although the code does some checks, it is essential that the data analyst to be careful about the quality of the data and to follow formatting instructions.
 #For details about how this function works, please read the analysis plan.
 #PLEASE PAY ATTENTION TO THE OUTPUT OF THIS FUNCTION!
 #ERROR MESSAGES (IN RED)
 #Error messages (in red) in the format '## MESSAGE ##' may be useful to correct formatting errors. When they appear, the function will stop and not return any results.
 #Other error messages (in red) are automatically produced by R, but may also be useful to correct formatting errors. When they appear, the function will stop and not return any results.
 #MESSAGES IN BLACK
 #Messages in black will not stop the function.
 #Messages such as "# MESSAGE :-)" indicate that the data passed a set of formatting checks.
 #Messages such as "# NON-FATAL WARNING: MESSAGE #" indicate that there might be an error, so the user should check the data as indicated in the message to ensure there are no errors.
 #To run the data_check() function, just run the single line below
 data_check(data, study_id, path, date, multi_centric, imp_quality_rs174575, imp_quality_rs1535, imp_software)

```
#CAREFULLY CHECK THE OUTPUT PRODUCED BY data_check(data, study_id, path, date, multi_centric,
imp_quality_rs174575, imp_quality_rs1535, imp_software)
#CORRECT ANY ERRORS IF NECESSARY
#PROCEED TO STEP C)
```

```
#####
#####
##C) SUMMARY STATISTICS##
#####
#####
```

```
#-----> ONLY RUN THIS AFTER GOING THOROUGH STEP B) <-----
```

#This section uses the summary_stats() function to generate summary statistics of your data.
 #Two tab-delimited text files will be generated with proper names in the directory indicated by path.
 #For details about how this function works, please read the analysis plan.
 #THIS CODE IS NOT EXPECTED TO PRODUCE ERROR (IN RED) MESSAGES. SO PLEASE PAY ATTENTION TO THEM, BECAUSE THEY COULD ONLY BE GENERATED AUTOMATICALLY FROM R AND WILL STOP THE FUNCTION.
 summary_stats(data)

```
#CHECK THE FILES PRODUCED BY summary_stats(data)
#IF ANYTHING SEEMS WRONG TO YOU BASED ON YOUR KNOWLEDGE OF YOUR DATA, RE-CHECK THE
DATA PROVIDED. IF YOU DO CHANGE INPUT DATA, PLEASE RETURN TO STEP B)
#PROCEED TO STEP D)
```

```
#####
#####
##D) ASSOCIATION ANALYSES##
#####
#####
```

```
#-----> ONLY RUN THIS AFTER GOING THOROUGH STEP C) <-----
#This section uses the fads2_association_analysis() function to perform all association analyses and
generate results.
#One tab-delimited text file will be generated with a proper name in the directory indicated by path.
#For details about how this function works, please read the analysis plan.
#THIS CODE IS NOT EXPECTED TO PRODUCE ERROR (IN RED) MESSAGES. SO PLEASE PAY ATTENTION TO
THEM, BECAUSE THEY COULD ONLY BE GENERATED AUTOMATICALLY FROM R AND WILL STOP THE
FUNCTION.
fads2_association_analysis(data)
```

Anexo IV – Código para gerar funções específicas para as análises do estudo de interação entre *FADS2* e amamentação

```
#####
#####
#####
#####
#-----AUXILIARY FUNCTIONS-----#
#####
#####
#####
#####

rm(list=ls())

#Detect outliers in continuous variables based on distance in SD units from the mean
detect_outliers <- function(x, sd_limit=4) {
  x <- x[!is.na(x)]
  outliers_index <- x<(mean(x)-sd_limit*sd(x)) | x>(mean(x)+sd_limit*sd(x))
  if(sum(outliers_index)==0) {
    outliers <- NULL
  } else {
    outliers <- x[outliers_index]
  }
  return(outliers)
}

#Calculate summary statistics for continuous variables
summary_stats_cont <- function(x) {
  x <- x[!is.na(x)]
  res <- c(min=min(x), max=max(x), mean=mean(x), sd=sd(x), median=median(x), iqr=IQR(x))
  return(res)
}

#Conver ISCED categories to US years of schooling
convertISCED <- function(x) {
  x <- factor(x)
  US_years <- NA
  US_years[x=='0' & !is.na(x)] <- 1
  US_years[x=='1' & !is.na(x)] <- 7
  US_years[x=='2' & !is.na(x)] <- 10
  US_years[x=='3' & !is.na(x)] <- 13
  US_years[x=='4' & !is.na(x)] <- 15
  US_years[x=='5' & !is.na(x)] <- 19
  US_years[x=='6' & !is.na(x)] <- 22
  return(US_years)
}

#Convert numeric variables to Z-scores
convertZ <- function(x) {
  z <- (x-mean(x, na.rm=T))/sd(x, na.rm=T)
  return(z)
}
```

```

#Re-level a factor variable so its reference category is the one with the largest sample size
relevel_to_MaxN <- function(x) {
  x_table <- table(x)
  x_largest_level <- names(x_table)[x_table==max(x_table)]
  x <- relevel(x, ref=x_largest_level)
  return(x)
}

#####
#####
#####
#####
#-----MAIN FUNCTIONS-----#
#####
#####
#####
#####

data_check <- function(data, study_id, path, date, multi_centric,
  imp_quality_rs174575=NULL, imp_quality_rs1535=NULL, imp_software) {

  #####
  #1) General formatting checks#
  #####

  expected_names <- c('bf_any_bin', 'bf_any_con', 'bf_exc_bin', 'bf_exc_con', 'iq',
    'rs174575_gen', 'rs1535_gen', 'rs174575_imp', 'rs1535_imp',
    'sex', 'age', 'mat_edu', 'mat_cog',
    'pc1', 'pc2', 'pc3', 'pc4', 'pc5', 'pc6', 'pc7', 'pc8', 'pc9', 'pc10',
    'ancestry', 'field_centre')

  #Check if 'data', 'study_id', 'path', 'date' and 'multi_centric' were provided
  if(is.null(data)) {
    stop('## 1.1) No data provided! Are you sure that that path is correct? ##')
  }

  if(study_id=='your_study_id') {
    stop('## 1.2) study_id not provided! ##')
  }

  if(path=='path_to_data') {
    stop('## 1.3) path not provided! ##')
  }

  if(date=='DDMMYYYY') {
    stop('## 1.4) date not provided! ##')
  }

  if(multi_centric=='multi_centric_info') {
    stop('## 1.5) multi_centric not provided! ##')
  }

  if(imp_software=='your_imp_software') {
    stop('## 1.6) imp_software not provided! ##')
  }
}

```



```

#Check if 'data' has the expected number of variables
if(ncol(data)!=length(expected_names)) {
  stop(paste('## 1.7) data has ', ncol(data), ' columns instead of ', length(expected_names), '! ##',
sep=""))
}

#Check if 'data' has the expected variable names in the expected order
num_match_vars <- sum(colnames(data)==expected_names)
if(num_match_vars!=length(expected_names)) {
  stop(paste('## 1.8) column(s) ', (1:length(expected_names))[,!colnames(data)%in%expected_names], '
don\'t match to its(their) expected name(s)! ##', sep=""))
}

cat('#####', sep='\n')
cat('#The data passed general formatting checks! :-)', sep='\n')
cat('#####', sep='\n')

#####
#2) Eligibility check#
#####

#Check if there is information on breastfeeding, intelligence and FADS2
if(sum(!is.na(data$bf_any_bin))==0 & sum(!is.na(data$bf_any_bin))==0) {
  stop('## 2.1) No breastfeeding data! The study must have data for at least one of bf_any_bin or
bf_exc_bin to be eligible! ##')
}

if(sum(!is.na(data$iq))==0) {
  stop('## 2.2) No IQ data! The study must have data for iq to be eligible! ##')
}

if(sum(!is.na(data$rs174575_gen))==0 & sum(!is.na(data$rs1535_gen))==0 &
sum(!is.na(data$rs1535_imp))==0 & sum(!is.na(data$rs1535_imp))==0) {
  stop('## 2.3) No FADS2 data! The study must have data for at least one of rs174575_gen, rs1535_gen,
rs174575_imp or rs1535_imp to be eligible! ##')
}

#If using genotyped SNPs, check if imp_quality was provided
if(sum(!is.na(data$rs174575_gen))>0 & !is.null(imp_quality_rs174575)) {
  stop('## 2.4) rs174575 was genotyped, but imputation quality was informed! Are you sure you are
using genotyped data? ##')
}

if(sum(!is.na(data$rs1535_gen))>0 & !is.null(imp_quality_rs1535)) {
  stop('## 2.5) rs1535 was genotyped, but imputation quality was informed! Are you sure you are using
genotyped data? ##')
}

#If using imputed SNPs, check if imp_quality was provided
if(sum(!is.na(data$rs174575_imp))>0 & is.null(imp_quality_rs174575)) {
  stop('## 2.6) rs174575 was imputed, but imputation quality was not informed! ##')
}

if(sum(!is.na(data$rs1535_imp))>0 & is.null(imp_quality_rs1535)) {
  stop('## 2.7) rs1535 was imputed, but imputation quality was not informed! ##')
}

```

```

#Check if SNPs were provided twice, as genotyped and imputed
if(sum(!is.na(data$rs174575_gen))>0 & sum(!is.na(data$rs174575_imp))>0) {
  stop('## 2.8) Both rs174575_gen and rs174575_imp were provided! ##')
}

if(sum(!is.na(data$rs1535_gen))>0 & sum(!is.na(data$rs1535_imp))>0) {
  stop('## 2.9) Both rs1535_gen and rs1535_imp were provided! ##')
}

#Check if SNPs (in case they were imputed) pass the quality threshold
if(!is.null(imp_quality_rs174575)) {
  if (imp_quality_rs174575<0.3) {
    stop('## 2.10) Imputation quality of rs174575 was below 0.3! ##')
  }
}

if(!is.null(imp_quality_rs1535)) {
  if (imp_quality_rs1535<0.3) {
    stop('## 2.11) Imputation quality of rs1535 was below 0.3! ##')
  }
}

#age checks
if(sum(!is.na(data$age))==0) {
  stop('## 2.12) No age data! ##')
}

#IQ checks
if(sum(!is.na(data$iq))==0) {
  stop('## 2.13) No IQ data! ##')
}

cat("", sep='\n')
cat('#####', sep='\n')
cat('#The data passed Eligibility checks! :-)#', sep='\n')
cat('#####', sep='\n')

#####
#3) Categorical variables check#
#####

#Check genotypes of rs174575, in case it was genotyped
if(sum(!is.na(data$rs174575_gen))>0) {
  rs174575_levels <- as.character(sort(unique(data$rs174575_gen[!is.na(data$rs174575_gen)])))

  if(!identical(rs174575_levels, c('CC', 'CG', 'GG'))){
    stop(paste('## 3.1) Genotypes of rs174575_gen were "', paste(rs174575_levels, collapse=' '), '"! ##',
sep=""))
  }
}

#Check genotypes of rs1535, in case it was genotyped
if(sum(!is.na(data$rs1535_gen))>0) {
  rs1535_levels <- as.character(sort(unique(data$rs1535_gen[!is.na(data$rs1535_gen)])))

  if(!identical(rs1535_levels, c('AA', 'AG', 'GG'))){

```

```

    stop(paste('## 3.2) Genotypes of rs1535_gen were "', paste(rs1535_levels, collapse=' '), "'! ##',
sep=''))
  }
}

#Check bf_any_bin, if available
if(sum(!is.na(data$bf_any_bin))>0) {
  bf_any_bin_levels <- as.character(sort(unique(data$bf_any_bin[!is.na(data$bf_any_bin)])))
  if(!identical(bf_any_bin_levels, c('no', 'yes'))) {
    stop(paste('## 3.3) bf_any_bin levels were "', paste(bf_any_bin_levels, collapse=' '), "'! ##', sep=''))
  }
}

#Check bf_exc_bin, if available
if(sum(!is.na(data$bf_exc_bin))>0) {
  bf_exc_bin_levels <- as.character(sort(unique(data$bf_exc_bin[!is.na(data$bf_exc_bin)])))
  if(!identical(bf_exc_bin_levels, c('no', 'yes'))) {
    stop(paste('## 3.4) bf_exc_bin levels were "', paste(bf_exc_bin_levels, collapse=' '), "'! ##', sep=''))
  }
}

#Check sex
if(sum(!is.na(data$sex))==0) {
  stop(paste('## 3.5) No sex information! ##'))
}

sex_levels <- as.character(sort(unique(data$sex[!is.na(data$sex)])))
if(!identical(sex_levels, c('female', 'male'))) {
  stop(paste('## 3.6) sex levels were "', paste(sex_levels, collapse=' '), "'! ##', sep=''))
}

#Check ancestry
if(sum(!is.na(data$ancestry))==0) {
  stop('## 3.7) No ancestry information! ##')
}

expected_ancestry_levels <- c('european', 'african', 'asian', 'hispanic', 'other')
ancestry_levels <- as.character((unique(data$ancestry[!is.na(data$ancestry)])))
if(sum(ancestry_levels%in%expected_ancestry_levels)!=length(ancestry_levels)) {
  stop(paste('## 3.8) ancestry levels were "', paste(ancestry_levels, collapse=' '), "'! ##', sep=''))
}

if(sum(data$ancestry=='european', na.rm=T)==0) {
  stop('## 3.9) No individuals of European ancestry! ##')
}

#Check mat_edu
expected_mat_edu_levels <- as.character(0:6)
mat_edu_levels <- as.character((unique(data$mat_edu[!is.na(data$mat_edu)])))
if(sum(mat_edu_levels%in%expected_mat_edu_levels)!=length(mat_edu_levels) &
sum(!is.na(data$mat_edu))!=0) {
  stop(paste('## 3.10) mat_edu levels were "', paste(mat_edu_levels, collapse=' '), "'! ##', sep=''))
}

#Check field_centre
if(sum(!is.na(data$field_centre))>0 & multi_centric=='no') {
  stop('## 3.11) Study is not multi-centric, but field_centre information was provided! ##')
}

```

```

}

if(multi_centric=='yes') {
  if(sum(!is.na(data$field_centre)==0)) {
    stop('## 3.12) Study is multi-centric, but field_centre information was not provided! ##')
  }

  field_centre_levels <- as.character(sort(unique(data$field_centre[!is.na(data$field_centre)])))
  expected_field_centre_levels <- sort(paste('fc', 1:length(field_centre_levels), sep=''))

  if(!identical(field_centre_levels, expected_field_centre_levels)) {
    stop(paste('## 3.13) field_centre levels were "', paste(field_centre_levels, collapse=' '), '"! ##',
sep=""))
  }
}

cat("", sep='\n')
cat('#####', sep='\n')
cat('#The data passed Categorical variables checks! :-)#', sep='\n')
cat('#####', sep='\n')

#####
#4) Continuous variables check#
#####

#bf_any_con, if available
if(sum(!is.na(data$bf_any_con))>0) {

  if(sum(data$bf_any_con<0, na.rm=T)>0) {
    stop('## 4.1 bf_any_con has one or more negative values! ##')
  }

  bf_any_con_outliers <- detect_outliers(data$bf_any_con)
  if(!is.null(bf_any_con_outliers)) {
    cat("", sep='\n')
    cat(paste('## 4.2 NON-FATAL WARNING: bf_any_con has the following outlier(s): "',
paste(unique(bf_any_con_outliers), collapse=' '), '"! ##', sep=""), sep='\n')
  }
}

#bf_exc_con, if available
if(sum(!is.na(data$bf_exc_con))>0) {

  if(sum(data$bf_exc_con<0, na.rm=T)>0) {
    stop('## 4.3 bf_exc_con has one or more negative values! ##')
  }

  bf_exc_con_outliers <- detect_outliers(data$bf_exc_con)
  if(!is.null(bf_exc_con_outliers)) {
    cat("", sep='\n')
    cat(paste('## 4.4 NON-FATAL WARNING: bf_exc_con has the following outlier(s): "',
paste(unique(bf_exc_con_outliers), collapse=' '), '"! ##', sep=""), sep='\n')
  }
}

#iq
if(sum(data$iq<0, na.rm=T)>0) {

```

```

    stop('## 4.5 iq has one or more negative values! ##')
  }

  iq_outliers <- detect_outliers(data$iq)
  if(!is.null(iq_outliers)) {
    cat("", sep='\n')
    cat(paste('## 4.6 NON-FATAL WARNING: iq has the following outlier(s): ', paste(unique(iq_outliers),
collapse=' '), '! ##', sep=''), sep='\n')
  }

  #mat_cog, if available
  if(sum(!is.na(data$mat_cog))>0) {

    if(sum(data$mat_cog<0, na.rm=T)>0) {
      stop('## 4.7 mat_cog has one or more negative values! ##')
    }

    mat_cog_outliers <- detect_outliers(data$mat_cog)
    if(!is.null(mat_cog_outliers)) {
      cat("", sep='\n')
      cat(paste('## 4.8 NON-FATAL WARNING: mat_cog has the following outlier(s): ',
paste(unique(mat_cog_outliers), collapse=' '), '! ##', sep=''), sep='\n')
    }
  }

  #age
  if(sum(data$age<0, na.rm=T)>0) {
    stop('## 4.9 age has one or more negative values! ##')
  }

  age_outliers <- detect_outliers(data$age)
  if(!is.null(age_outliers)) {
    cat("", sep='\n')
    cat(paste('## 4.10 NON-FATAL WARNING: age has the following outlier(s): ',
paste(unique(age_outliers), collapse=' '), '! ##', sep=''), sep='\n')
  }

  #rs174575_imp, if available
  if(sum(!is.na(data$rs174575_imp))>0) {

    rs174575_imp_outliers_index <- data$rs174575_imp<0 | data$rs174575_imp>2

    if(sum(rs174575_imp_outliers_index, na.rm=T)>0) {
      stop(paste('## 4.11: rs174575_imp has the following outlier(s): ',
paste(unique(data$rs174575_imp[rs174575_imp_outliers_index & !is.na(data$rs174575_imp)]),
collapse=' '), '! ##', sep=''))
    }
  }

  #rs1535_imp, if available
  if(sum(!is.na(data$rs1535_imp))>0) {

    rs1535_imp_outliers_index <- data$rs1535_imp<0 | data$rs1535_imp>2

    if(sum(rs1535_imp_outliers_index, na.rm=T)>0) {

```

```

    stop(paste('## 4.12: rs1535_imp has the following outlier(s): ',
paste(unique(data$rs1535_imp[rs1535_imp_outliers_index & !is.na(data$rs1535_imp)]), collapse=' '),
"'! ##', sep=''))
  }

}

#pcs
pcs_names <- paste('pc', 1:10, sep='')
pcs_index <- apply(!is.na(data[,pcs_names]), 2, sum)>0
pcs_included <- pcs_names[pcs_index]
cat("", sep='\n')
if(length(pcs_included)==0) {
  cat('## 4.11 NON-FATAL WARNING: all pcs are entirely missing!', sep='\n')
} else {
  cat(paste('## 4.13 NON-FATAL WARNING: there is information for the following pc(s): ',
paste(pcs_included, collapse=' '), "'! ##', sep=''), sep='\n')
}

cat("", sep='\n')
cat('#####', sep='\n')
cat('#The data passed Continuous variables checks! :-)', sep='\n')
cat('#####', sep='\n')

#####
#5) Consistency checks of breastfeeding variables#
#####

#Check if all individuals non-missing for bf_any_con are also non-missing for bf_any_bin
if (sum(is.na(data$bf_any_bin) & !is.na(data$bf_any_con))>0) {
  stop('## 5.1 not all individuals non-missing for bf_any_con are also non-missing for bf_any_bin! ##')
}

#Check if all individuals non-missing for bf_exc_con are also non-missing for bf_exc_bin
if (sum(is.na(data$bf_exc_bin) & !is.na(data$bf_exc_con))>0) {
  stop('## 5.2 not all individuals non-missing for bf_exc_con are also non-missing for bf_exc_bin! ##')
}

#Check if there are individuals with 'yes' for bf_exc_bin but 'no' for bf_any_bin, if both are available
if(sum(!is.na(data$bf_any_bin)>0 & sum(!is.na(data$bf_exc_bin)>0) {

  if(sum(data$bf_any_bin=='no' & data$bf_exc_bin=='yes', na.rm=T)>0) {
    stop('## 5.3 one or more individuals are "yes" for bf_exc_bin but "no" for bf_any_bin! ##')
  }
}

#Check if there are individuals with larger values for bf_exc_con values than for bf_any_con, if both are
available
if(sum(!is.na(data$bf_any_con)>0 & sum(!is.na(data$bf_exc_con)>0) {

  if(sum(data$bf_any_con<data$bf_exc_con, na.rm=T)>0) {
    stop('## 5.4 one or more individuals presented larger values for bf_exc_cont than for bf_any_con!
##')
  }
}

cat("", sep='\n')

```

```

cat('#####', sep='\n')
cat('#The data passed Consistency checks of breastfeeding variables! :-)#', sep='\n')
cat('#####', sep='\n')

#####
#6) G-allele frequency check#
#####

#Check if G-allele frequency of all SNPs within Europeans are within 0.1 and 0.4
#Since this is the last check, the data can be limited to Europeans
data <- data[data$ancestry=='european',]

#rs174575_gen, if available
if(sum(!is.na(data$rs174575_gen))>0) {
  rs174575_gen_G_freq <- (sum(data$rs174575_gen=='GG', na.rm=T)*2 +
sum(data$rs174575_gen=='CG', na.rm=T))/(sum(!is.na(data$rs174575_gen))*2)

  if(rs174575_gen_G_freq>0.4 | rs174575_gen_G_freq<0.1) {
    stop('## 6.1 rs174575_gen G-allele frequency lies outside the 10%-40% range! ##')
  }
}

#rs1535_gen, if available
if(sum(!is.na(data$rs1535_gen))>0) {
  rs1535_gen_G_freq <- (sum(data$rs1535_gen=='GG', na.rm=T)*2 + sum(data$rs1535_gen=='AG',
na.rm=T))/(sum(!is.na(data$rs1535_gen))*2)

  if(rs1535_gen_G_freq>0.4 | rs1535_gen_G_freq<0.1) {
    stop('## 6.2 rs1535_gen G-allele frequency lies outside the 10%-40% range! ##')
  }
}

#rs174575_imp, if available
if(sum(!is.na(data$rs174575_imp))>0) {
  rs174575_imp_G_freq <- mean(data$rs174575_imp, na.rm=T)/2

  if(rs174575_imp_G_freq>0.4 | rs174575_imp_G_freq<0.1) {
    stop('## 6.3 rs174575_imp G-allele frequency lies outside the 10%-40% range! ##')
  }
}

#rs1535_imp, if available
if(sum(!is.na(data$rs1535_imp))>0) {
  rs1535_imp_G_freq <- mean(data$rs1535_imp, na.rm=T)/2

  if(rs1535_imp_G_freq>0.4 | rs1535_imp_G_freq<0.1) {
    stop('## 6.4 rs1535_imp G-allele frequency lies outside the 10%-40% range! ##')
  }
}

cat("", sep='\n')
cat('#####', sep='\n')
cat('#The data passed G-allele frequency check! :-)#', sep='\n')
cat('#####', sep='\n')

cat("", sep='\n')
cat("", sep='\n')

```

```

cat('#####',
sep='\n')
cat('#####',
sep='\n')
cat('#-----The data passed ALL checks! :-)------#', sep='\n')
cat('#####',
sep='\n')
cat('#-----PLEASE CHECK IF THERE ARE ANY NON-FATAL ERROR MESSAGES!-----#', sep='\n')
cat('#-----IF SO, PLEASE ADDRESS ANY POTENTIAL ISSUES APPROPRIATELY!-----#', sep='\n')
cat('#####',
sep='\n')
cat('#####',
sep='\n')
}

summary_stats <- function(data, size_snp=6) {

#####
#####
#1) Limit to European-ancestry individuals with breastfeeding, genetic and covariate information#

#####
#####

#1.1) Find the breastfeeding variable with the largest number of non-missing values
bf_notNA <- !is.na(data[,c('bf_any_bin', 'bf_any_con', 'bf_exc_bin', 'bf_exc_con')])
bf_notNA_sum <- apply(bf_notNA, 2, sum)
bf_notNA_max_names <- names(bf_notNA_sum[bf_notNA_sum==max(bf_notNA_sum)])
bf_notNA_index <- !is.na(data[, (sample(bf_notNA_max_names, 1))])

#1.2) Find the SNP variable with the largest number of non-missing values
gen_notNA <- !is.na(data[,c('rs174575_gen', 'rs1535_gen', 'rs174575_imp', 'rs1535_imp')])
gen_notNA_sum <- apply(gen_notNA, 2, sum)
gen_notNA_max_names <- names(gen_notNA_sum[gen_notNA_sum==max(gen_notNA_sum)])
gen_notNA_index <- !is.na(data[, (sample(gen_notNA_max_names, 1))])

#1.3) Find the PC with the largest number of non-missing values
pc_notNA <- !is.na(data[, paste('pc', 1:10, sep='')])
pc_notNA_sum <- apply(pc_notNA, 2, sum)
#If there is no PC data, don't use it as a criterion to exclude individuals
if(sum(pc_notNA_sum==0)==length(pc_notNA_sum)) {
  pc_notNA_index <- rep(T, nrow(data))
} else {
  pc_notNA_max_names <- names(pc_notNA_sum[pc_notNA_sum==max(pc_notNA_sum)])
  pc_notNA_index <- !is.na(data[, (sample(pc_notNA_max_names, 1))])
}

#Select the individuals that meet all criteria
data <- data[data$ancestry=='european' & !is.na(data$ancestry) &
  bf_notNA_index &
  gen_notNA_index &
  pc_notNA_index,]

#Generate bf_any_cat, bf_any_bin_6, bf_exc_cat and bf_exc_bin_6 if continuous counterparts are
available

```



```

if(sum(!is.na(data$bf_any_con))>0) {

  data$bf_any_cat <- data$bf_any_con
  data$bf_any_cat[data$bf_any_con>0 & data$bf_any_con<=1] <- 1
  data$bf_any_cat[data$bf_any_con>1 & data$bf_any_con<=3] <- 2
  data$bf_any_cat[data$bf_any_con>3 & data$bf_any_con<=6] <- 3
  data$bf_any_cat[data$bf_any_con>6] <- 4
  data$bf_any_cat <- factor(data$bf_any_cat, levels=0:4)

  data$bf_any_bin_6 <- NA
  data$bf_any_bin_6[data$bf_any_con<6] <- 0
  data$bf_any_bin_6[data$bf_any_con>=6] <- 1
  data$bf_any_bin_6 <- factor(data$bf_any_bin_6)

} else {
  data$bf_any_cat <- NA
  data$bf_any_bin_6 <- NA
}

if(sum(!is.na(data$bf_exc_con))>0) {

  data$bf_exc_cat <- data$bf_exc_con
  data$bf_exc_cat[data$bf_exc_con>0 & data$bf_exc_con<=1] <- 1
  data$bf_exc_cat[data$bf_exc_con>1 & data$bf_exc_con<=3] <- 2
  data$bf_exc_cat[data$bf_exc_con>3 & data$bf_exc_con<=6] <- 3
  data$bf_exc_cat[data$bf_exc_con>6] <- 4
  data$bf_exc_cat <- factor(data$bf_exc_cat, levels=0:4)

  data$bf_exc_bin_6 <- NA
  data$bf_exc_bin_6[data$bf_exc_con<6] <- 0
  data$bf_exc_bin_6[data$bf_exc_con>=6] <- 1
  data$bf_exc_bin_6 <- factor(data$bf_exc_bin_6)

} else {
  data$bf_exc_cat <- NA
  data$bf_exc_bin_6 <- NA
}

#Convert ISCED categories into US years of schooling
data$mat_edu <- convertISCED(data$mat_edu)

#####
#2) Recode imputed SNPs into genotypes#
#####

#2.1) rs174575, if available
if(sum(!is.na(data$rs174575_gen))==0 & sum(!is.na(data$rs174575_imp))==0) {
  data$rs174575 <- NA

} else if (sum(!is.na(data$rs174575_gen))>0) {
  data$rs174575 <- data$rs174575_gen

} else if (sum(!is.na(data$rs174575_imp))>0) {
  data$rs174575 <- round(data$rs174575_imp)
  data$rs174575 <- factor(data$rs174575, levels=0:2, labels=c('CC', 'CG', 'GG'))
}

```

```

#2.2) rs1535, if available
if(sum(!is.na(data$rs1535_gen))==0 & sum(!is.na(data$rs1535_imp))==0) {
  data$rs1535 <- NA

} else if (sum(!is.na(data$rs1535_gen))>0) {
  data$rs1535 <- data$rs1535_gen

} else if (sum(!is.na(data$rs1535_imp))>0) {
  data$rs1535 <- round(data$rs1535_imp)
  data$rs1535 <- factor(data$rs1535, levels=0:2, labels=c('AA', 'AG', 'GG'))
}

#####
#3) Create a data frame to store the results#
#####

variable <- rep(c('sex', 'age', 'outcome',
  'bf_any_bin', 'bf_any_cat', 'bf_any_con',
  'bf_exc_bin', 'bf_exc_cat', 'bf_exc_con',
  'rs174575', 'rs1535'),
  c(2, 6, 6,
    2, 5, 6,
    2, 5, 6,
    3, 3)
)

trait <- c(rep('iq', length(variable)+3),
  rep('mat_edu', length(variable)),
  rep('mat_cog', length(variable))
)

variable <- c(variable, c('pcs', 'mat_edu', 'mat_cog'), variable, variable)

study <- rep(study_id, length(trait))

stat_con_names <- c('min', 'max', 'mean', 'sd', 'median', 'iqr')
stat_names <- c('females', 'males', stat_con_names, stat_con_names,
  'no', 'yes', 0:4, stat_con_names,
  'no', 'yes', 0:4, stat_con_names,
  c('CC', 'CG', 'GG'),
  c('AA', 'AG', 'GG'))
stat_names <- c(c(stat_names, rep('availability', 3)), stat_names, stat_names)
value <- rep(NA, length(trait))

res <- data.frame(study, trait, variable, stat_names, value)

#####
#4) Also generate some summary statistics for the SNPs#
#####
res_snp <- data.frame(study=rep(study_id, size_snp), trait=rep(c('iq', 'mat_edu', 'mat_cog'),
each=size_snp/3),
  snp=rep(c('rs174575', 'rs1535'), 3),
  hwe=NA, imputed=NA, imputation_quality=NA, maf=NA)

#####
#5) Calculate summary statistics for non-missing individuals for the each trait#
#####

```

#5.1) Define traits

```
traits <- c('iq', 'mat_edu', 'mat_cog')
```

#5.2) Obtain summary statistics for each trait

```
for(cur.trait in traits) {
```

```
  #Get the location of non-missing observations for the current trait
```

```
  cur.index <- !is.na(data[,cur.trait])
```

```
  #Only calculate summary statistics if there is data for the current trait
```

```
  if(sum(cur.index)>0) {
```

```
    #Limit the data to non-missing observations for the current trait
```

```
    cur.data <- data[cur.index,]
```

```
    #Calculate stats for sex
```

```
    res$value[res$trait==cur.trait & res$variable=='sex'] <- table(cur.data$sex)
```

```
    #Calculate stats for age
```

```
    res$value[res$trait==cur.trait & res$variable=='age'] <- summary_stats_cont(cur.data$age)
```

```
    #Calculate stats for outcome
```

```
    res$value[res$trait==cur.trait & res$variable=='outcome'] <-
```

```
    summary_stats_cont(cur.data[,cur.trait])
```

```
    #Calculate stats for bf_any_bin
```

```
    res$value[res$trait==cur.trait & res$variable=='bf_any_bin'] <- table(cur.data$bf_any_bin)
```

```
    #Calculate stats for bf_any_cat and bf_any_con, if available
```

```
    if(sum(!is.na(data$bf_any_con))>0) {
```

```
      res$value[res$trait==cur.trait & res$variable=='bf_any_cat'] <- table(cur.data$bf_any_cat)
```

```
      res$value[res$trait==cur.trait & res$variable=='bf_any_con'] <-
```

```
      summary_stats_cont(cur.data$bf_any_con)
```

```
    } else {
```

```
      res$value[res$trait==cur.trait & res$variable=='bf_any_cat'] <- NA
```

```
      res$value[res$trait==cur.trait & res$variable=='bf_any_con'] <- NA
```

```
    }
```

```
    #Calculate stats for bf_exc_bin, if available
```

```
    if(sum(!is.na(data$bf_exc_bin))>0) {
```

```
      res$value[res$trait==cur.trait & res$variable=='bf_exc_bin'] <- table(cur.data$bf_exc_bin)
```

```
    } else {
```

```
      res$value[res$trait==cur.trait & res$variable=='bf_exc_bin'] <- NA
```

```
    }
```

```
    #Calculate stats for bf_exc_cat and bf_exc_con, if available
```

```
    if(sum(!is.na(data$bf_exc_con))>0) {
```

```
      res$value[res$trait==cur.trait & res$variable=='bf_exc_cat'] <- table(cur.data$bf_exc_cat)
```

```
      res$value[res$trait==cur.trait & res$variable=='bf_exc_con'] <-
```

```
      summary_stats_cont(cur.data$bf_exc_con)
```

```
    } else {
```

```
      res$value[res$trait==cur.trait & res$variable=='bf_exc_cat'] <- NA
```

```

res$value[res$trait==cur.trait & res$variable=='bf_exc_con'] <- NA

}

#Calculate stats for rs174575, if available
if(sum(!is.na(data$rs174575))>0) {
  res$value[res$trait==cur.trait & res$variable=='rs174575'] <- table(cur.data$rs174575)

  #res_snp: get HWE for the SNP
  res_snp$hwe[res_snp$trait==cur.trait & res_snp$snp=='rs174575'] <-
HWE.exact(as.genotype(cur.data$rs174575, alleles=c('C', 'G'), sep=''))$p.value

  #res_snp: get imputation information for the SNP
  if(!is.null(imp_quality_rs174575)) {
    res_snp$imputed[res_snp$trait==cur.trait & res_snp$snp=='rs174575'] <- imp_software
    res_snp$imputation_quality[res_snp$trait==cur.trait & res_snp$snp=='rs174575'] <-
imp_quality_rs174575
  } else {
    res_snp$imputed[res_snp$trait==cur.trait & res_snp$snp=='rs174575'] <- 'GENOTYPED'
  }

  #res_snp: get MAF information for the SNP
  if(sum(!is.na(data$rs174575_imp))>0) {
    res_snp$maf[res_snp$trait==cur.trait & res_snp$snp=='rs174575'] <-
mean(cur.data$rs174575_imp, na.rm=T)/2
  } else {
    res_snp$maf[res_snp$trait==cur.trait & res_snp$snp=='rs174575'] <-
(sum(cur.data$rs174575_gen=='GG', na.rm=T)*2 + sum(data$rs174575_gen=='CG',
na.rm=T))/(sum(!is.na(cur.data$rs174575_gen))*2)
  }

} else {
  res$value[res$trait==cur.trait & res$variable=='rs174575'] <- NA
}

#Calculate stats for rs1535, if available
if(sum(!is.na(data$rs1535))>0) {
  res$value[res$trait==cur.trait & res$variable=='rs1535'] <- table(cur.data$rs1535)

  #res_snp: get HWE for the SNP
  res_snp$hwe[res_snp$trait==cur.trait & res_snp$snp=='rs1535'] <-
HWE.exact(as.genotype(cur.data$rs1535, alleles=c('A', 'G'), sep=''))$p.value

  #res_snp: get imputation information for the SNP
  if(!is.null(imp_quality_rs1535)) {
    res_snp$imputed[res_snp$trait==cur.trait & res_snp$snp=='rs1535'] <- imp_software
    res_snp$imputation_quality[res_snp$trait==cur.trait & res_snp$snp=='rs1535'] <-
imp_quality_rs1535
  } else {
    res_snp$imputed[res_snp$trait==cur.trait & res_snp$snp=='rs1535'] <- 'GENOTYPED'
  }

  #res_snp: get MAF information for the SNP
  if(sum(!is.na(data$rs1535_imp))>0) {
    res_snp$maf[res_snp$trait==cur.trait & res_snp$snp=='rs1535'] <- mean(cur.data$rs1535_imp,
na.rm=T)/2
  } else {

```

```

    res_snp$maf[res_snp$trait==cur.trait & res_snp$snp=='rs1535'] <-
(sum(cur.data$rs1535_gen=='GG', na.rm=T)*2 + sum(cur.data$rs1535_gen=='AG',
na.rm=T))/(sum(!is.na(cur.data$rs1535_gen))*2)
  }

  } else {
    res$value[res$trait==cur.trait & res$variable=='rs1535'] <- NA
  }

  #Check availability of pcs, mat_edu and mat_cog if cur.trait==iq
  if(cur.trait=='iq') {
    res$value[res$trait==cur.trait & res$variable=='pcs'] <- sum(apply(!is.na(cur.data[,paste('pc', 1:10,
sep=")]), 2, sum))!=0
    res$value[res$trait==cur.trait & res$variable=='mat_edu'] <- sum(!is.na(cur.data[, 'mat_edu']))!=0
    res$value[res$trait==cur.trait & res$variable=='mat_cog'] <- sum(!is.na(cur.data[, 'mat_cog']))!=0
  }
}
}

#####
#6) Save summary statistics in a tab-delimited text file#
#####

#6.1) Define the filenames
sample_filename <- paste(path, 'FADS2_OUTPUT_TYPE_sample_descriptives_STUDY_', study_id,
'_DATE_', date, '.txt', sep='')
snp_filename <- paste(path, 'FADS2_OUTPUT_TYPE_snp_descriptives_STUDY_', study_id, '_DATE_',
date, '.txt', sep='')

#6.2) Write the file
write.table(res, sample_filename, row.names=F, sep='\t')
write.table(res_snp, snp_filename, row.names=F, sep='\t')

cat("", sep='\n')

cat('#####
###', sep='\n')

cat('#####
###', sep='\n')
cat('#-----SUMMARY STATISTICS GENERATED! :)-----#', sep='\n')
cat("", sep='\n')
cat('#-----Stored in the following files: -----#', sep='\n')
cat("", sep='\n')
cat(sample_filename, sep='\n')
cat("", sep='\n')
cat(snp_filename, sep='\n')
cat("", sep='\n')

cat('#####
###', sep='\n')
cat('#PLEASE VERIFY IF THE CONTENTS OF THESE FILES SEEM CORRECT TO YOU BEFORE CONTINUING#',
sep='\n')

cat('#####
###', sep='\n')

```

```

cat('#####
###', sep='\n')

}

fads2_association_analysis <- function(data) {

#####
#1) Limit the data to individuals of European ancestry and aged at least 7 years#
#####

data <- data[data$ancestry=='european' & !is.na(data$ancestry),]

#####
#2) For each adjusted analysis, store all available covariates together#
#####
covs_1 <- data[,c('age', paste('pc', 1:10, sep=''))]
#Add sex as a numeric variable to allow estimating population-average effects
covs_1$sex <- as.character(data$sex)
covs_1$sex[data$sex=='male'] <- 0
covs_1$sex[data$sex=='female'] <- 1
covs_1$sex <- as.numeric(covs_1$sex)
#Create age squared
covs_1$age2 <- covs_1$age^2

#Mean-center continuous covariates in covs_1
for (cur_cov_index in 1:ncol(covs_1)) {
  cur_cov <- covs_1[,cur_cov_index]
  cur_centred_cov <- cur_cov-mean(cur_cov, na.rm=T)
  covs_1[,cur_cov_index] <- cur_centred_cov
}

covs_1 <- data.frame(covs_1, field_centre=data[,c('field_centre')]) #Add categorical variables
covs_2 <- data.frame(covs_1, data[,c('mat_edu', 'mat_cog')]) #Include maternal covariates in a distinct
data frame

#Make sure to only include variables that are not entirely missing
covs_1 <- covs_1[,apply(!is.na(covs_1), 2, sum)>0]
covs_2 <- covs_2[,apply(!is.na(covs_2), 2, sum)>0]

#Standardize maternal variables, if available
if('mat_edu'%in%colnames(covs_2)) {
  covs_2$mat_edu_2 <- covs_2$mat_edu^2
  covs_2$mat_edu <- covs_2$mat_edu-mean(covs_2$mat_edu, na.rm=T)
  covs_2$mat_edu_2 <- covs_2$mat_edu_2-mean(covs_2$mat_edu_2, na.rm=T)
}

if('mat_cog'%in%colnames(covs_2)) {
  covs_2$mat_cog_2 <- covs_2$mat_cog^2
  covs_2$mat_cog <- covs_2$mat_cog-mean(covs_2$mat_cog, na.rm=T)
  covs_2$mat_cog_2 <- covs_2$mat_cog_2-mean(covs_2$mat_cog_2, na.rm=T)
}

#Make sure to use the field_centre level with the largest sample size as the reference category
if('field_centre'%in%colnames(covs_1)) {
  covs_1$field_centre <- relevel_to_MaxN(covs_1$field_centre)
}

```

```

covs_2$field_centre <- relevel_to_MaxN(covs_2$field_centre)
}

#####
#3) Recode/generate variables#
#####

# 3.1) bf_any_cat, bf_any_cat_trend and by_any_bin_6, if bf_any_con is available
if(sum(!is.na(data$bf_any_con))>0) {

  data$bf_any_cat_trend <- data$bf_any_con
  data$bf_any_cat_trend[data$bf_any_cat_trend>0 & data$bf_any_cat_trend<=1] <- 1
  data$bf_any_cat_trend[data$bf_any_cat_trend>1 & data$bf_any_cat_trend<=3] <- 2
  data$bf_any_cat_trend[data$bf_any_cat_trend>3 & data$bf_any_cat_trend<=6] <- 3
  data$bf_any_cat_trend[data$bf_any_cat_trend>6] <- 4

  data$bf_any_bin_6 <- NA
  data$bf_any_bin_6[data$bf_any_con<6] <- 0
  data$bf_any_bin_6[data$bf_any_con>=6] <- 1

} else {
  data$bf_any_cat_trend <- NA
  data$bf_any_bin_6 <- NA
}

# 3.2) bf_exc_cat and bf_any_cat_trend, if bf_exc_con is available
if(sum(!is.na(data$bf_exc_con))>0) {

  data$bf_exc_cat_trend <- data$bf_exc_con
  data$bf_exc_cat_trend[data$bf_exc_cat_trend>0 & data$bf_exc_cat_trend<=1] <- 1
  data$bf_exc_cat_trend[data$bf_exc_cat_trend>1 & data$bf_exc_cat_trend<=3] <- 2
  data$bf_exc_cat_trend[data$bf_exc_cat_trend>3 & data$bf_exc_cat_trend<=6] <- 3
  data$bf_exc_cat_trend[data$bf_exc_cat_trend>6] <- 4

  data$bf_exc_bin_6 <- NA
  data$bf_exc_bin_6[data$bf_exc_con<6] <- 0
  data$bf_exc_bin_6[data$bf_exc_con>=6] <- 1

} else {
  data$bf_exc_cat_trend <- NA
  data$bf_exc_bin_6 <- NA
}

# 3.3) rs174575_add, rs174575_dom, rs174575_rec and rs174575_over, if rs174575 is available
data$rs174575_add <- NA
data$rs174575_dom <- NA
data$rs174575_rec <- NA
data$rs174575_over <- NA

#if rs174575 was genotyped
if(sum(!is.na(data$rs174575_gen))>0) {
  #Convert to G-allele dosages: this is the additive effect
  data$rs174575_add <- as.numeric(as.character(factor(data$rs174575_gen, levels=c('CC', 'CG', 'GG'),
labels=0:2)))

  #Convert to CC=0 vs. CG/GG=1: this is the dominant effect
  data$rs174575_dom <- data$rs174575_add

```

```

data$rs174575_dom[data$rs174575_dom>=1] <- 1

#Convert to CC/CG=0 vs. GG=1: this is the recessive effect
data$rs174575_rec <- data$rs174575_add-1
data$rs174575_rec[data$rs174575_rec<0] <- 0

#Convert to CC/CG=0 vs. CG=1: this is the overdominant effect
data$rs174575_over <- abs(abs(data$rs174575_add-1)-1)
}

#if rs174575 was imputed
if(sum(!is.na(data$rs174575_imp))>0) {
  #The SNP is already coded as G-allele dosages: this is the additive effect
  data$rs174575_add <- data$rs174575_imp

  #Convert to CC=0 vs. CG/GG=1: this is the dominant effect
  data$rs174575_dom <- data$rs174575_add
  data$rs174575_dom[data$rs174575_dom>=1] <- 1

  #Convert to CC/CG=0 vs. GG=1: this is the recessive effect
  data$rs174575_rec <- data$rs174575_add-1
  data$rs174575_rec[data$rs174575_rec<0] <- 0

  #Convert to CC/CG=0 vs. CG=1: this is the overdominant effect
  data$rs174575_over <- abs(abs(data$rs174575_add-1)-1)
}

# 3.4) rs1535_add and rs1535_rec, if rs1535 available

data$rs1535_add <- NA
data$rs1535_dom <- NA
data$rs1535_rec <- NA
data$rs1535_over <- NA

#if rs1535 was genotyped
if(sum(!is.na(data$rs1535_gen))>0) {
  #Convert to G-allele dosages: this is the additive effect
  data$rs1535_add <- as.numeric(as.character(factor(data$rs1535_gen, levels=c('AA', 'AG', 'GG'),
labels=0:2))))

  #Convert to AA=0 vs. AG/GG=1: this is the dominant effect
  data$rs1535_dom <- data$rs1535_add
  data$rs1535_dom[data$rs1535_dom>=1] <- 1

  #Conver to AA/AG=0 vs. GG=1: this is the recessive effect
  data$rs1535_rec <- data$rs1535_add-1
  data$rs1535_rec[data$rs1535_rec<0] <- 0

  #Convert to AA/GG=0 vs. AG=1: this is the overdominant effect
  data$rs1535_over <- abs(abs(data$rs1535_add-1)-1)
}

#if rs1535 was imputed
if(sum(!is.na(data$rs1535_imp))>0) {
  #The SNP is already coded as G-allele dosages: this is the additive effect
  data$rs1535_add <- data$rs1535_imp

```



```

#Convert to AA=0 vs. AG/GG=1: this is the dominant effect
data$rs1535_dom <- data$rs1535_add
data$rs1535_dom[data$rs1535_dom>=1] <- 1

#Conver to AA/AG=0 vs. GG=1: this is the recessive effect
data$rs1535_rec <- data$rs1535_add-1
data$rs1535_rec[data$rs1535_rec<0] <- 0

#Convert to AA/GG=0 vs. AG=1: this is the overdominant effect
data$rs1535_over <- abs(abs(data$rs1535_add-1)-1)
}

# 3.5) Convert iq and mat_cog (if available) to Z_scores, and convert mat_edu to US years
data$iq <- convertZ(data$iq)
if(sum(!is.na(data$mat_cog))>0) {data$mat_cog <- convertZ(data$mat_cog)}
if(sum(!is.na(data$mat_edu))>0) {data$mat_edu <- convertISCED(data$mat_edu)}

#####
#4) Create a data frame to store the results#
#####
trait <- c('iq', 'mat_edu', 'mat_cog')

breastfeeding <- c('bf_any_bin', 'bf_any_bin_6', 'bf_any_cat_trend', 'bf_any_con',
                  'bf_exc_bin', 'bf_exc_bin_6', 'bf_exc_cat_trend', 'bf_exc_con')

model <- c('crude', 'adjusted_1', 'adjusted_2')
#In case pcs were provided, modify 'ajusted' level of model to reflect this
if(sum(substr(colnames(covs_1), 1, 2)=='pc')>0) { model[2:3] <- paste(model[2:3], '_with_pcs', sep='')}

snp <- c('rs174575', 'rs1535')

effect <- c('add', 'dom', 'rec', 'over')

col_names <- c('trait', 'breastfeeding', 'model', 'snp', 'effect',
              'n', 'eaf',
              'beta0', 'beta0_se', 'beta0_p',
              'bf_beta', 'bf_se', 'bf_p',
              'snp_beta', 'snp_se', 'snp_p',
              'int_beta', 'int_se', 'int_p',
              'cov_beta0_snp', 'cov_beta0_bf', 'cov_beta0_int',
              'cov_snp_bf', 'cov_snp_int',
              'cov_bf_int')

res <- data.frame(matrix(nrow=1000, ncol=length(col_names)))
colnames(res) <- col_names

#####
#5) Run the analysis#
#####

count<-1

for(cur.trait in trait) {#cur.trait <- trait[1]
  cur.trait.var <- data[,cur.trait]

  for(cur.bf in breastfeeding) {#cur.bf <- breastfeeding[1]
    cur.bf.var <- data[,cur.bf]

```

```

if(class(cur.bf.var)=='factor') {
  cur.bf.var <- as.numeric(cur.bf.var)-1
}

for(cur.model in model) {#cur.model <- model[1]

  if(cur.model%in%c('adjusted_2', 'adjusted_2_with_pcs')) {
    if(cur.trait!='iq' | sum(c('mat_edu', 'mat_cog')%in%colnames(covs_2))==0) {
      next
    }
  }

  for(cur.snp in snp) {#cur.snp<-snp[1]

    for(cur.effect in effect) {#cur.effect<-effect[1]
      cur.snp.var.name <- paste(cur.snp, cur.effect, sep='_')
      cur.snp.var <- data[,cur.snp.var.name]

      #Fill in columns with the current combination of data
      res$trait[count] <- cur.trait
      res$breastfeeding[count] <- cur.bf
      res$model[count] <- cur.model
      res$snp[count] <- cur.snp
      res$effect[count] <- cur.effect

      #Check if there is at least one non-missing observation for the current combination of trait,
      breastfeeding and FADs2
      #Only run analysis if there is
      if(sum(!is.na(cur.trait.var))>0 & sum(!is.na(cur.bf.var))>0 & sum(!is.na(cur.snp.var))>0) {

        #Fit the model, adjusting for covariates
        if(cur.model=='crude') {
          fit <- lm(cur.trait.var~cur.snp.var*cur.bf.var)
        } else if (cur.model%in%c('adjusted_1', 'adjusted_1_with_pcs')) {
          fit <- lm(cur.trait.var~cur.snp.var*cur.bf.var+cur.snp.var*cur.bf.var, data=covs_1)
        } else {
          fit <- lm(cur.trait.var~cur.snp.var*cur.bf.var+cur.snp.var*cur.bf.var, data=covs_2)
        }

        #Add effective sample size and EAF
        res$n[count] <- length(residuals(fit))
        if(cur.effect=='add') {
          res$seaf[count] <- mean(data[rownames(data)%in%(names(residuals(fit))), cur.snp.var.name],
na.rm=T)/2
        } else {
          res$seaf[count] <- res$seaf[count-1]
        }

        #Re-estimate variance-covariance matrix
        fit.vcovHC1 <- vcovHC(fit, type='HC1')

        #Obtain robust standard errors with a one degree of freedom correction
        fit.rob <- data.frame(rbind(NULL, coeftest(fit, fit.vcovHC1))) #Small trick to convert into a data
frame

        #Save regression output into res

```

```

#Intercept
res$beta0[count] <- fit.rob[1,1]
res$beta0_se[count] <- fit.rob[1,2]
res$beta0_p[count] <- fit.rob[1,4]

#SNP
res$snp_beta[count] <- fit.rob[2,1]
res$snp_se[count] <- fit.rob[2,2]
res$snp_p[count] <- fit.rob[2,4]

#Breastfeeding and BreastfeedingxSNP
#Force the inclusion of a BFXSNP interaction row in fit.rob
req_rows <- c(row.names(fit.rob)[3],
              paste('cur.snp.var:', row.names(fit.rob)[3], sep=''))
)
req_rows <- data.frame(req_rows)
#Add a column named 'req_rows' to fit.rob
fit.rob <- data.frame(req_rows=row.names(fit.rob), fit.rob)
#Merge
fit.rob <- merge(req_rows, fit.rob, all.x=T)
rownames(fit.rob) <- fit.rob$req_rows #re-assign rownames
#Remove req_rows column
fit.rob <- fit.rob[,-1]

#Now, results can be obtained painlessly
res$bf_beta[count] <- fit.rob[1,1]
res$bf_se[count] <- fit.rob[1,2]
res$bf_p[count] <- fit.rob[1,4]

res$int_beta[count] <- fit.rob[2,1]
res$int_se[count] <- fit.rob[2,2]
res$int_p[count] <- fit.rob[2,4]

#Extract covariance between selected coefficients
if(!is.na(res$int_beta[count])) {
  cur_target_rows <- row.names(fit.vcovHC1)[2:3]
  cur_target_rows[3] <- paste(cur_target_rows[1], cur_target_rows[2], sep=':')

  covs_int <- fit.vcovHC1[rownames(fit.vcovHC1)%in%cur_target_rows,
colnames(fit.vcovHC1)=='(Intercept)']
  res$cov_beta0_snp[count] <- covs_int[names(covs_int)==cur_target_rows[1]]
  res$cov_beta0_bf[count] <- covs_int[names(covs_int)==cur_target_rows[2]]
  if(cur_target_rows[3]%in%names(covs_int)) {
    res$cov_beta0_int[count] <- covs_int[names(covs_int)==cur_target_rows[3]]
  }

  covs_snp <- fit.vcovHC1[rownames(fit.vcovHC1)%in%cur_target_rows[2:3],
colnames(fit.vcovHC1)==cur_target_rows[1]]
  res$cov_snp_bf <- covs_snp[names(covs_snp)==cur_target_rows[2]]
  if(cur_target_rows[3]%in%names(covs_snp)) {
    res$cov_snp_int[count] <- covs_snp[names(covs_snp)==cur_target_rows[3]]

    res$cov_bf_int[count] <- fit.vcovHC1[rownames(fit.vcovHC1)%in%cur_target_rows[3],
colnames(fit.vcovHC1)==cur_target_rows[2]]
  }
}

```

```

    }
    #Contabilize this
    count<-count+1
  }
}
}
}

#Remove extra-rows and add study_id
res <- res[1:count,]
res <- data.frame(study=study_id, res)

#####
#6) Save summary statistics in a tab-delimited text file#
#####

#6.1) Define the filenames
res_filename <- paste(path, 'FADS2_OUTPUT_TYPE_association_results_', 'STUDY_', study_id, '_DATE_',
date, '.txt', sep='')

#6.2) Write the file
write.table(res, res_filename, row.names=F, sep='\t')

cat("", sep='\n')

cat('#####
###', sep='\n')

cat('#####
###', sep='\n')
cat('#-----ASSOCIATION RESULTS GENERATED! :)-----#', sep='\n')
cat("", sep='\n')
cat('#-----Stored in the following file: -----#', sep='\n')
cat("", sep='\n')
cat(res_filename, sep='\n')
}

save(list=c('detect_outliers', 'summary_stats_cont', 'convertISCED', 'convertZ', 'relevel_to_MaxN',
'data_check', 'summary_stats', 'fads2_association_analysis'),
file='/Users/Fernando/Dropbox/Fernando Hartwig/Doutorado/FADS2 meta-analysis/FADS2 x BF
interaction on intelligence - Functions 20160409.R')

```

2 – Relatório de atividades

PARTICIPAÇÃO DO DOUTORANDO NO PROJETO EPIGEN-Brasil

Conforme mencionado no Projeto de Pesquisa, foi acordado junto à coordenação do Programa de Pós-Graduação em Epidemiologia que o doutorando trabalharia com dados do projeto EPIGEN-Brasil, como gerente e analista do banco de dados genéticos da coorte de nascimentos em Pelotas em 1982, em substituição ao trabalho de campo. Aqui, são relatadas as atividades realizadas (além das mencionadas no Projeto).

Gerência do banco de dados

A imputação dos dados genéticos foi atualizada, passando a ter como painel de referência dados da fase 3 do projeto 1000 Genomas. Comparado com a fase 1 (referência utilizada no processo de imputação anterior), a fase 3 inclui mais indivíduos e uma maior cobertura do genoma. Isto permite imputar mais variantes genéticas de forma mais precisa, principalmente no caso de populações miscigenadas. Tanto cromossomos autossômicos como o cromossomo X foram imputados.

O processo inclui as seguintes etapas principais: i) limpeza do banco de dados de variantes genéticas genotipadas conforme filtros de qualidade mais recentes; ii) formatação dos dados de acordo com o programa que realiza a imputação; iii) imputação propriamente dita; iv) processamento pós-imputação, de modo a re-harmonizar os bancos de dados imputados referente a cromossomos autossômicos e o cromossomo X (necessário pois o cromossomo X é imputado de forma diferente).

Os programas Plink 1.9, BCFtools e VCFtools foram utilizados para limpeza e processamento dos dados. A imputação foi realizada utilizando os programas *Eagle2* e *SHAPEIT* (identificação de haplótipos) e Minimac3 (imputação propriamente dita), conforme implementados pelo *Michigan Imputation Server* [1].

Estudos empíricos

- *Genomic ancestry and the social pathways leading to major depression in adulthood: the mediating effect of socioeconomic position and discrimination* (publicado [2]).

Dados de ancestralidade genômica foram utilizados para investigar os determinantes sociais da associação entre etnia e depressão maior em adultos participantes da coorte de nascimentos em Pelotas em 1982. Observou-se que a posição socioeconômica modificou a associação entre ancestralidade africana e depressão, havendo um efeito de aumento de risco apenas entre o tercil mais rico, enquanto que não houve evidência de associações nos outros estratos socioeconômicos. Além disso, detectou-se que a percepção de sofrer discriminação racial explicou aproximadamente 84% desta associação, indicando um importante componente social na relação entre etnia e depressão. Este trabalho foi liderado pelo prof. Christian Loret de Mola (Universidade Federal de Pelotas). O doutorando contribuiu na análise de dados, interpretação dos resultados e revisão crítica do artigo.

- *PCSK9 genetic variants and risk of type 2 diabetes: a mendelian randomisation study* (publicado [3]).

Este consórcio incluindo mais de 550,000 indivíduos identificou que variantes genéticas no gene que codifica a enzima PCSK9 estão associados com menores níveis de colesterol LDL e risco aumentado de diabetes tipo 2, bem como maiores níveis de glicose em jejum, peso corporal e razão cintura-quadril. Estes resultados indicam que novos medicamentos para reduzir o colesterol LDL que atuam na enzima PCSK9 (atualmente sendo testados em ensaios clínicos) podem aumentar o risco de desenvolver diabetes. Este estudo foi liderado pelo Dr. Amand Floriaan Schmidt (*University College London*), e o doutorando participou realizando as análises dos dados da coorte de nascimentos em Pelotas em 1982 e revisão crítica do artigo.

- *Suggestive association between variants in IL1RAPL and asthma symptoms in Latin American children* (publicado [4]).

Este estudo complementa GWAS prévios sobre asma ao analisar a associação entre este desfecho e variantes genéticas no cromossomo X. Houve evidência de associação para o SNP rs12007907 (localizada no gene *IL1RAPL*) em homens, e esta variante também apresentou associação com níveis de interleucina 13. Estes resultados indicam que modificação do efeito de variantes genéticas no cromossomo X de acordo com sexo podem ter um papel na diferença de frequência de asma severa entre os

sexos. Este estudo foi liderado pela Dr^a. Cintia Rodrigues Marques (Universidade Federal da Bahia), e o doutorando contribuiu na análise de dados da coorte de nascimentos em Pelotas em 1982 e revisão crítica do artigo.

- *Breastfeeding moderates FTO-related adiposity: a birth cohort study with 30 years of follow-up* (publicado [5]).

Utilizando dados do último acompanhamento da coorte de nascimentos em Pelotas em 1982 (idade média: 30.2 anos), neste estudo investigou-se o potencial efeito modulador da amamentação na associação entre a variante genética rs9939609 (localizada no gene *FTO*) e medidas de obesidade. Foi observada uma atenuação do efeito obesogênico do alelo A (quando comparado com o alelo T) em indivíduos amamentados, com evidência estatística de interação para índice de massa corporal (IMC), índice de massa magra e circunferência da cintura. Os resultados para outras medidas de obesidade apresentaram a mesma tendência, mas não atingiram os limiares convencionais de significância estatística. O estudo indica que amamentação pode atenuar os efeitos de predisposição genética a ser obeso. Este estudo foi liderado pelo prof. Bernardo Horta (Universidade Federal de Pelotas), com participação do doutorando na análise de dados, interpretação dos resultados, redação e revisão crítica do artigo.

- *Genome-Wide Association Study of Blood Pressure Traits by Hispanic/Latino Background: the Hispanic Community Health Study/Study of Latinos* (publicado [6]).

Este foi o primeiro GWAS de pressão arterial em populações hispânicas e latino-americanas. A varredura genômica foi realizada utilizando dados do estudo norte americano *Hispanics Community Health Study/Study of Latinos* (HCHS/SOL; N=12,278), com replicação em três outros estudos (incluindo a coorte de nascimentos em Pelotas em 1982). Não foram detectadas variantes genéticas associadas com pressão arterial que atingiram critérios pré-definidos de significância estatística, mas algumas variantes apresentaram evidência sugestiva, indicando que possivelmente amostras maiores são necessárias. Além disso, várias associações detectadas em estudos prévios em populações europeias e chinesas foram replicadas, indicando que ao menos uma parte do componente genético da pressão arterial é comum a diversos grupos étnicos. Este

estudo foi liderado pela Dr^a. Tamar Sofer (*Washington University*) e o doutorando contribuiu na análise de dados da coorte de nascimentos em Pelotas em 1982 e revisão crítica do artigo.

- *A Genome-Wide Association Study in Hispanics/Latinos Identifies Novel Signals for Lung Function* (publicado [7]).

Neste GWAS de função pulmonar, realizou-se uma varredura genômica também utilizando dados do estudo HCHS/SOL (N=11,822), com replicação em três outros estudos (incluindo a coorte de nascimentos em Pelotas em 1982). Foram detectadas oito novos sítios do genoma associados com medidas de função pulmonar, dos quais três atingiram critérios pré-definidos de replicação. Além disso, várias associações detectadas em estudos prévios em populações europeias foram replicadas no estudo HCHS/SOL, indicando que ao menos uma parte do componente genético da função pulmonar é comum a diversos grupos étnicos. Este estudo foi liderado pela Dr^a. Kristin Burkart (*Columbia University*) e o doutorando contribuiu na análise de dados da coorte de nascimentos em Pelotas em 1982 e revisão crítica do artigo.

- *Life-course genome-wide association study meta-analysis of total body bone mineral density yields thirty-six novel loci and identifies age-specific effects* (publicado [8]).

Esta pesquisa, parte do consórcio *Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE)*, foi o primeiro estudo de GWAS a avaliar se associações entre variantes genéticas e densidade mineral óssea variam conforme a idade. Para isso, as análises foram realizadas estratificando em diferentes grupos etários em um total de 66,628 indivíduos. Ao combinar todos os grupos etários, foram identificados 36 novos sítios do genoma independentemente associados com densidade mineral óssea. Comparações entre os grupos etários indicaram que a idade pode modular o efeito de algumas variantes genéticas, mas na maioria dos casos não houve uma clara evidência de modificação de efeito. Este estudo foi liderado pela Dr^a. Carolina Medina Gomez (*Erasmus Medical Center*), com participação do doutorando através da análise de dados da coorte de nascimentos em Pelotas em 1982 e revisão crítica do artigo.

- *A large-scale multi-ancestry genome-wide study incorporating gene-smoking interactions identifies 139 genome-wide significant loci for systolic and diastolic blood pressure* (aceito para publicação [9]).

O grupo *Gene-Lifestyle Working Group*, também parte do consórcio *CHARGE*, realizou um estudo amplo de interação gene-tabagismo. Este é o primeiro consórcio de GWAS a estudar uma interação gene-ambiente. Em um total de 610,091 indivíduos, foram identificados 132 novos sítios do genoma independentemente associados com pressão arterial, com alguns desses sítios apresentando modificação de efeito por tabagismo. Estes resultados fornecem uma corroboração empírica à importância de avaliar interações gene-ambiente ao estudar os determinantes genéticos de desfechos multifatoriais. Este estudo foi liderado pela prof^a. Yun Ju Sung (*Washington University*) e o doutorando contribuiu na análise de dados da coorte de nascimentos em Pelotas em 1982 e revisão crítica do artigo.

- *Multiethnic Meta-analysis Identifies New Loci for Pulmonary Function* (submetido ao periódico *Nature Genetics*).

Também parte do consórcio *CHARGE*, este estudo de GWAS incluiu mais de 90 mil indivíduos de vários grupos étnicos. Mais de 50 novas regiões do DNA associadas com medidas de função pulmonar foram identificadas, e análises de bioinformática indicaram que o possível envolvimento de 16 genes cujos produtos são proteínas as quais são alvo de fármacos, indicando uma potencial significância clínica dos resultados. Este estudo foi liderado pela Dr^a. Annah Wyss (*National Institute of Environmental Health Sciences*), com participação do doutorando através da análise de dados da coorte de nascimentos em Pelotas em 1982 e revisão crítica do artigo.

- *Novel genetic associations for blood pressure identified via gene-alcohol interaction in up to 570K individuals across multiple ancestries* (artigo em redação).

Este estudo também foi conduzido pelo consórcio *Gene-Lifestyle Working Group*, envolvendo aproximadamente 130 mil indivíduos na fase de descobrimento e aproximadamente 440 mil indivíduos na fase de replicação. O desenho do estudo foi parecido com o descrito imediatamente acima, mas utilizando consumo de álcool ao

invés de tabagismo. Foram identificados 5 novos sítios do genoma associados com pressão arterial, bem como 18 potenciais novos sítios em populações africanas, porém não haviam estudos dessa etnia suficientes para replicação. Este estudo foi liderado pela prof^a. Mary Feitosa (*Washington University*) e o doutorando contribuiu na análise de dados da coorte de nascimentos em Pelotas em 1982 e revisão crítica do artigo.

- *Multi-ancestry genome-wide association study incorporating gene-alcohol interactions identifies 18 new lipid loci* (artigo em redação).

Neste estudo, também conduzido pelo consórcio *Gene-Lifestyle Working Group*, utilizou-se uma metodologia similar ao estudo supracitado, com a diferença de que os desfechos eram colesterol LDL, colesterol HDL e triglicerídeos. Através de uma análise de descobrimento incluindo 45 estudos juntamente seguida de uma análise de replicação incluindo 66 estudos, totalizando um tamanho de amostra de 394.914 indivíduos, foram identificadas 18 novas regiões independentes do genoma associadas com pelo menos uma das frações lipídicas estudadas. Este estudo foi liderado pelo Dr. Paul de Vries (*University of Texas*) e o doutorando contribuiu na análise de dados da coorte de nascimentos em Pelotas em 1982 e revisão crítica do artigo.

- *Multi-ancestry Genome-wide Association Study incorporating Gene $\times$ Smoking Interactions Identifies Novel Lipid Loci* (artigo em redação).

Este quarto estudo conduzido pelo consórcio *Gene-Lifestyle Working Group* utilizou a mesma metodologia que o estudo supracitado, porém avaliando modificação de associações genéticas pelo tabagismo ao invés de consumo de álcool. Foram identificados 13 novas regiões independentes do genoma associadas com pelo menos uma das frações lipídicas consideradas através de análises de descobrimento em 133.816 indivíduos replicadas em uma amostra independente de 253.467 indivíduos. Muitas dessas novas regiões apresentaram associações com frações lipídicas substancialmente diferentes conforme status de tabagismo, corroborando a importância de se considerar interações gene-ambiente em estudos de epidemiologia genética. Este estudo foi liderado pela Dr^a. Amy R. Bentley (*National Human Genome Research Institute*) e o doutorando contribuiu na análise de dados da coorte de nascimentos em Pelotas em 1982 e revisão crítica do artigo.

ATIVIDADES COMO PESQUISADOR ASSOCIADO À UNIVERISDADE DE BRISTOL

Desde fevereiro de 2013, o doutorando vem realizando pesquisa em colaboração com a *University of Bristol* (Reino Unido) junto ao *Medical Research Council Epidemiology Unit (MRC IEU)*. Em julho de 2016, o doutorando passou a ter um vínculo honorário com a universidade, passando a ser pesquisador associado ao *MRC IEU*. Inclusive, os três artigos que fazem parte desta tese têm participação de pesquisadores desta instituição.

As pesquisas do doutorando junto ao *MRC IEU* são voltadas a um delineamento de estudo conhecido como randomização (ou aleatorização) mendeliana.

Resumidamente, randomização mendeliana é um tipo de análise de variáveis instrumentais, em que variantes genéticas associadas com uma determinada exposição são utilizadas como instrumentos desta exposição para investigar seu efeito causal em um ou mais desfechos de interesse. Ou seja, o objetivo da randomização mendeliana não é estudar aspectos genéticos de um determinado desfecho, mas sim utilizar a genética para fazer inferências mais robustas sobre os efeitos causais de um fator de risco modificável.

A seguir, são listadas as pesquisas (em andamento ou já concluídas, além das mencionadas no Projeto) na área de randomização mendeliana que o doutorando participou junto ao *MRC IEU*:

Estudos empíricos

- *Body mass index and psychiatric disorders: a Mendelian randomization study* (publicado [10]).

Foram utilizadas 97 variantes genéticas identificadas no maior e mais recente GWAS de IMC realizado até o momento como variáveis instrumentais. Dados sumarizados (ou seja, coeficientes de regressão linear e os correspondentes erros padrão) foram obtidos deste GWAS, disponibilizado gratuitamente pelo consórcio *Genetic Investigation of ANthropometric Traits (GIANT)*. Dados sumarizados da associação entre cada uma destas variantes genéticas e transtornos mentais (transtorno bipolar, transtorno de depressão maior e esquizofrenia) foram obtidos no banco de dados de

resultados de GWAS disponibilizados gratuitamente pelo *Psychiatric Genetics Consortium* (PGC). Combinar dados sumarizados gerados em diferentes estudos é possível através de metodologias específicas de randomização mendeliana, que são análogas a uma meta-análise. Aplicou-se uma série de análises de sensibilidade, com destaque para diferentes estimadores que refletem diferentes pressupostos sobre fontes de vieses importantes em randomização mendeliana. Os resultados indicaram que o IMC não tem efeito causal (ou este efeito é muito pequeno) em transtorno bipolar e esquizofrenia, porém parece influenciar o risco de desenvolver depressão. Porém, a evidência estatística foi relativamente fraca, de forma que esta pergunta deve ser abordada novamente utilizando novos dados sumarizados (indisponíveis no momento da realização do estudo), estimados em amostras maiores. Este estudo foi liderado pelo doutorando.

- *Inflammatory Biomarkers and Risk of Schizophrenia: A 2-Sample Mendelian Randomization Study* (publicado [11]).

Este estudo avaliou o efeito causal da proteína C reativa (PCR) e outros marcadores inflamatórios no risco de desenvolver esquizofrenia. Foram utilizados dados sumarizados, como no estudo anterior. No caso da PCR, foram utilizados dois conjuntos de instrumentos: a) quatro variantes genéticas localizadas no gene que codifica a PCR, correspondendo a um conjunto mais conservador de instrumentos; b) 17 variantes genéticas localizadas em diferentes partes do genoma, identificados no maior GWAS de PCR até o momento; este é um conjunto mais liberal de instrumentos, que fornece mais poder para detectar uma associação. Dados sumarizados da associação entre cada instrumento genética e risco de esquizofrenia foram obtidos do PGC. Após aplicar uma série de análises de sensibilidade, os resultados indicaram um efeito causal protetor da PCR no risco de desenvolver esquizofrenia, contrário aos resultados comumente obtidos em estudos observacionais convencionais. A possibilidade de que PCR precoce influencia o risco de infecções é apontada como um potencial mecanismo deste efeito. Este estudo foi liderado pelo doutorando.

- *Education and coronary heart disease: a Mendelian randomization study* (publicado [12]).

Neste estudo, avaliou-se o efeito causal de escolaridade no risco de desenvolver doença arterial coronariana através de análises observacionais convencionais e randomização mendeliana. Diferentes estimadores de randomização mendeliana e análises observacionais convencionais indicaram que um aumento de 3.6 anos reduz o risco de doença arterial coronariana em 20-30%. Este estudo foi liderado pelo Dr. Taavi Tillmann (*University College London*). O doutorando participou das etapas de análise de dados, interpretação dos resultados, redação e revisão crítica do artigo.

- *The genetic architecture of osteoarthritis: insights from UK Biobank* (aceito para publicação [13]).

Este estudo foi um GWAS de osteoartrite envolvendo ~16,5 milhões de variantes genéticas, usando a primeira onda de dados do *UK Biobank* (N~150 mil) e uma amostra de replicação independente de ~400 mil indivíduos. Foram detectadas nove novas regiões associadas com risco de osteoartrite. As análises foram complementadas usando definições do desfecho mais específicas, bem como análises de transcriptoma (sequenciamento do transcriptoma completo) e proteoma (espectrometria de massa) de cultivo primário de tecido cartilaginoso extraído de lesões de 38 pacientes. Análises de randomização mendeliana corroboraram que obesidade aumenta o risco de osteoartrite. Este estudo foi liderado pela Dr^a. Eleni Zegnini (*Wellcome Sanger Institute*). O doutorando realizou todas as análises de randomização mendeliana, e também contribuiu na interpretação dos resultados, redação e revisão crítica do artigo.

Estudos metodológicos, teóricos e/ou sem dados empíricos

- *Why internal weights should be avoided (not only) in MR-Egger regression* (publicado [14]).

Utilizando dados já publicados e análises originais de dados sumarizados de GWAS, este estudo demonstrou que uma técnica recente de randomização mendeliana chamada regressão MR-Egger pode ser altamente influenciada por viés de instrumentos fracos. Mais especificamente, demonstrou-se que do efeito causal da altura na função pulmonar publicadas em outros trabalhos estavam provavelmente superestimadas devido a esse viés. Como a regressão MR-Egger apresenta

propriedades teóricas muito vantajosas, pesquisadores a vinham utilizando ignorando a influência do viés de instrumentos fracos (que é relevante em amostras finitas, ou seja, em dados reais) nos resultados obtidos por este método. Este trabalho foi liderado pelo doutorando.

- *Two-sample Mendelian randomization: avoiding the downsides of a powerful, widely applicable but potentially fallible technique* (publicado [15]).

Neste artigo, um exemplo real foi utilizado para ilustrar a importância da harmonização correta dos bancos de dados ao realizar análises de randomização mendeliana com dados sumarizados. Demonstrou-se que harmonização imperfeita tende a enviesar as estimativas de efeito na sua direção oposta (por exemplo, tornar uma estimativa positiva em negativa), e orientações detalhadas de como realizar o processo de harmonização de forma adequada foram fornecidas, bem como *scripts* que realizam este processo automaticamente. Também foi evidenciado o recente crescimento do uso de dados sumarizados em randomização mendeliana, corroborando a importância de divulgar orientações sobre como realizar o processo de harmonização de dados. Este trabalho foi liderado pelo doutorando.

- *Robust inference in summary data Mendelian randomisation via the zero modal pleiotropy assumption* (publicado [16]).

Esta publicação descreve um novo método para análise de randomização mendeliana com dados sumarizados, chamado MBE (*mode-based estimate*). O pressuposto do qual o método depende (chamado ZEMPA – *zero modal pleiotropy assumption*) é apresentado e comparado com os pressupostos de métodos já descritos. Foi realizado um estudo de simulação que indicou que o MBE é mais robusto contra diversos casos em que os pressupostos da análise de randomização mendeliana são violados, e sua aplicação foi ilustrada utilizando dados reais.

O MBE foi apresentado na edição de 2017 do *UK Causal Inference Meeting* (Colchester, UK) e do *Mendelian Randomization Conference* (Bristol, UK). Este trabalho foi liderado pelo doutorando.

- *Lactase Persistence and Body Mass Index: The Contribution of Mendelian Randomization* (publicado [17]).

Neste editorial referente a um estudo original [18], comenta-se sobre a contribuição da randomização mendeliana no estudo do efeito causal do consumo de leite e obesidade, utilizando uma variante genética associada com persistência da lactase como variável instrumental. Os pontos positivos e negativos do estudo supracitado são analisados no contexto dos principais estudos já publicados neste tópico, e perguntas de pesquisa ainda não respondidas foram apresentadas. Apesar de o doutorando ter primeira autoria, o artigo surgiu a partir de um convite ao segundo autor pelo corpo editorial do periódico para escrever este editorial.

- *Bias in Mendelian randomisation due to assortative mating* (submetido ao periódico *Genetic Epidemiology*).

Neste estudo, foi utilizado um modelo de simulação de dados para quantificar como processos de cruzamento preferencial (por exemplo, mulheres altas tendem a escolher homens altos) podem levar a vieses em análises de randomização mendeliana. Vários cenários diferentes foram considerados, sendo constatado que o viés pode acontecer mesmo que a exposição e o desfecho não sejam as variáveis que sofram seleção diretamente. Além disso, observou-se que a ocorrência de cruzamento preferencial ao longo de várias gerações pode amplificar o viés, que não é detectado por métodos tipicamente utilizados em análises de sensibilidade em investigações que utilizam a metodologia de randomização mendeliana. Porém, foi mostrado que é possível utilizar informação genética parental para corrigir este viés. Este estudo foi liderado pelo doutorando.

- *Instrumental variables estimation of causal effects in the presence of invalid instruments* (em andamento).

Este estudo visa comparar propriedades de diferentes estimadores de randomização mendeliana que exploram o pressuposto ZEMPA. Apesar de ainda estar em andamento, já foi possível identificar métodos que superam o MBE em algumas situações. Atualmente, simulações adicionais estão sendo realizadas para comparar os

métodos mais promissores em uma ampla variedade de cenários. Este estudo é liderado pelo professor Frank Windmeijer (*University of Bristol*).

- *Covariate-adjusted summary association results in two-sample Mendelian randomisation: a simulation study* (em andamento).

Muitos dos grandes consórcios de GWAS têm realizado análises ajustando para covariáveis herdáveis (ou seja, que têm um componente genético) na tentativa de obter efeitos das variantes genéticas nos desfechos de interesse que são independentes da covariável em questão. Porém, muitas vezes aspectos conceituais não são levados em conta, podendo levar a estimativas que não correspondem ao que os pesquisadores desejam estimar. Estudos de GWAS são a principal fonte de dados sumarizados para randomização mendeliana, e esta abordagem vem sendo aplicada a dados de GWAS que ajustaram para uma ou mais covariáveis herdáveis. Através de um estudo de simulação e análises empíricas, demonstra-se que o uso de dados sumarizados oriundos de GWAS ajustados podem enviesar os resultados de uma análise de randomização mendeliana de diversas formas, muitas vezes difíceis de prever e altamente dependentes da estrutura causal que gerou os dados.

Outros estudos

As atividades do doutorando junto ao *MRC IEU* também envolvem pesquisas em outras áreas, como pode ser verificado nos artigos que compõem esta tese. Além destes, outros trabalhos incluem:

- *From stem cells to the law courts: DNA methylation, the forensic epigenome and the possibility of a biosocial archive* (publicado [19]).

Aqui, foram descritos os conceitos “epigenoma forense” e “arquivo bio-social”, ambos baseados em efeitos epigenéticos duradouros que algumas exposições (tais como tabagismo) apresentam. É apresentada a teoria de que células tronco adultas podem ter um importante papel na manutenção de modificações epigenéticas ao longo do tempo. Também foram discutidas as implicações destes conceitos e teorias em epidemiologia. Este trabalho foi liderado pela prof^a. Caroline Relton (*University of Bristol*).

- *On applying Egger regression to evaluate pleiotropic effects of drugs* (submetido ao periódico *Arteriosclerosis, Thrombosis, and Vascular Biology*).

Esta carta discute limitações importantes de um artigo original [20], que propõe uma modificação da regressão de Egger para estudar efeitos pleiotrópicos de medicamentos. Como exemplo, os autores re-analisaram dados de 25 estudos de intervenção randomizados, e concluíram que o efeito protetor das estatinas sobre o risco de doenças cardiovascular é virtualmente totalmente mediado pela redução nos níveis de colesterol LDL. Na carta, é discutido que a aderência imperfeita à intervenção, bem como efeitos pleiotrópicos em função do próprio alvo biológico primário do medicamento. Este trabalho foi liderado pelo doutorando.

- *Meta-analysis in the presence of small study bias: the utility of reporting the mean, the median and the mode* (em andamento).

Os primeiros métodos de randomização mendeliana com dados sumarizados foram adaptados da literatura sobre meta-análise, porém métodos mais recentes foram desenvolvidos diretamente para randomização mendeliana. Através de simulações e análises de dados reais, este estudo apresenta como dois métodos de randomização mendeliana com dados sumarizados podem ser utilizados em meta-análises para obter estimativas mais robustas menos influenciadas por viés de publicação.

CONCLUSÕES E PERSPECTIVAS

As atividades junto ao projeto EPIGEN-Brasil permitiram ao doutorando ter um maior entendimento sobre os desafios de lidar com dados pesados e multidimensionais, bem como a aprender a desenvolver maneiras de lidar com dados dessa natureza. Isto envolveu o desenvolvimento de scripts eficientes de manejo e extração de dados, incluindo paralelização explícita de tarefas, além do uso de diferentes softwares livres voltados ao manejo e análises de dados genéticos amplos. Esta experiência contribuiu para aprender como solucionar, de forma independente, problemas no campo de manejo e análise de dados, principalmente na automação e otimização de tarefas. Tal conhecimento tem sido muito útil em algumas das pesquisas mencionadas nesta seção, que envolvem simulação de dados, bem como no Artigo original 1 que compõe

esta tese. O último foi baseado em dados amplos de metilação do DNA ao longo do genoma, e muitos dos métodos utilizados para otimizar o manejo e a análise desse tipo de dado são similares aos utilizados para dados genéticos.

O trabalho como gerente e analista de dados genéticos da coorte de nascimentos em Pelotas em 1982 também possibilitou ao doutorando participar dos estudos empíricos mencionados acima. Tal trabalho conferiu experiência em análises de varredura genômica, análises de replicação em epidemiologia genética, bem como expansão da rede de colaboradores e ganho de experiência com consórcios internacionais. Isto foi de grande valia no planejamento e realização do Artigo original 2 que compõe esta tese, que consiste de uma meta-análise *de novo* incluindo 11 estudos epidemiológicos.

As pesquisas junto ao *MRC IEU* foram, principalmente, na área de randomização mendeliana. O doutorando adquiriu conhecimento teórico e experiência prática neste campo, que ainda é incipiente no Brasil. O doutorando já ministrou aulas no PPGE sobre o assunto, com o intuito de contribuir na disseminação do conhecimento e na capacitação de mais pessoas para o uso desse delineamento. Além disso, o conhecimento e experiência adquiridos conferiram ao doutorando um entendimento mais profundo sobre randomização mendeliana, principalmente sobre suas limitações. Isto permitiu ao doutorando atuar também no campo metodológico, contribuindo no desenvolvimento de métodos que ajudam a reduzir as limitações desta abordagem. Esta atuação têm resultado em um grande aprendizado na área de inferência causal de forma geral, área de maior interesse do doutorando.

De forma geral, as atividades aqui relatadas conferiram ao doutorando: i) maior independência no manejo e análise de amplos bancos de dados genéticos; ii) experiência no planejamento e condução de consórcios; iii) expansão da rede de colaboradores; iv) capacitação na área de randomização mendeliana.

No total, o doutorando participou, durante o doutorado, de 33 estudos: 22 aceitos ou publicados, 5 submetidos e 6 em andamento ou redação.

REFERÊNCIAS

1. Das S, Forer L, Schonherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. *Nat Genet.* 2016;48(10):1284-1287.
2. Loret de Mola C, Hartwig FP, Goncalves H, Quevedo Lde A, Pinheiro R, Gigante DP, et al. Genomic ancestry and the social pathways leading to major depression in adulthood: the mediating effect of socioeconomic position and discrimination. *BMC Psychiatry.* 2016;16(1):308.
3. Schmidt AF, Swerdlow DI, Holmes MV, Patel RS, Fairhurst-Hunter Z, Lyall DM, et al. PCSK9 genetic variants and risk of type 2 diabetes: a mendelian randomisation study. *Lancet Diabetes Endocrinol.* 2017;5(2):97-105.
4. Marques CR, Costa GN, da Silva TM, Oliveira P, Cruz AA, Alcantara-Neves NM, et al. Suggestive association between variants in IL1RAPL and asthma symptoms in Latin American children. *Eur J Hum Genet.* 2017;25(4):439-445.
5. Horta BL, Victora CG, França GVA, Hartwig FP, Ong K, Rolfe EL, et al. Breastfeeding moderates *FTO* related adiposity: a birth cohort study with 30 years of follow-up. *Sci Rep.* 2018;8(1):2530.
6. Sofer T, Wong Q, Hartwig FP, Taylor K, Warren HR, Evangelou E, et al. Genome-Wide Association Study of Blood Pressure Traits by Hispanic/Latino Background: the Hispanic Community Health Study/Study of Latinos. *Sci Rep.* 2017;7(1):10348.
7. Burkart KM, Sofer T, London SJ, Manichaikul A, Hartwig FP, Yan Q, et al. A Genome-wide Association Study in Hispanics/Latinos Identifies Novel Signals for Lung Function. *The Hispanic Community Health Study/Study of Latinos. Am J Respir Crit Care Med.* 2018.
8. Medina-Gomez C, Kemp JP, Trajanoska K, Luan J, Chesi A, Ahluwalia TS, et al. Life-Course Genome-wide Association Study Meta-analysis of Total Body BMD and Assessment of Age-Specific Effects. *Am J Hum Genet.* 2018;102(1):88-102.
9. Sung YJ, Winkler TW, de las Fuentes L, Bentley AR, Brown MR, Kraja AT, et al. A Large-Scale Multi-ancestry Genome-wide Study Accounting for Smoking Behavior Identifies Multiple Significant Loci for Blood Pressure. *Am J Hum Genet.* 2018;In press.

10. Hartwig FP, Bowden J, Loret de Mola C, Tovo-Rodrigues L, Davey Smith G, Horta BL. Body mass index and psychiatric disorders: a Mendelian randomization study. *Sci Rep*. 2016;632730.
11. Hartwig FP, Borges MC, Horta BL, Bowden J, Davey Smith G. Inflammatory Biomarkers and Risk of Schizophrenia: A 2-Sample Mendelian Randomization Study. *JAMA Psychiatry*. 2017;74(12):1226-1233.
12. Tillmann T, Vaucher J, Okbay A, Pikhart H, Peasey A, Kubinova R, et al. Education and coronary heart disease: mendelian randomisation study. *BMJ*. 2017;358j3542.
13. Zengini E, Hatzikotoulas K, Tachmazidou I, Steinberg J, Hartwig FP, Southam L, et al. The genetic architecture of osteoarthritis: insights from UK Biobank. *bioRxiv*. 2017;10.1101/174755.
14. Hartwig FP, Davies NM. Why internal weights should be avoided (not only) in MR-Egger regression. *Int J Epidemiol*. 2016;45(5):1676-1678.
15. Hartwig FP, Davies NM, Hemani G, Davey Smith G. Two-sample Mendelian randomisation: avoiding the downsides of a powerful, widely applicable but potentially fallible technique. *Int J Epidemiol*. 2016;45(6):1717-1726.
16. Hartwig FP, Davey Smith G, Bowden J. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *Int J Epidemiol*. 2017;46(6):1985-1998.
17. Hartwig FP, Davey Smith G. Lactase Persistence and Body Mass Index: The Contribution of Mendelian Randomization. *Clin Chem*. 2018;64(1):4-6.
18. Dairy Consumption and Body Mass Index Among Adults: Mendelian Randomization Analysis of 184802 Individuals from 25 Studies. *Clin Chem*. 2018;64(1):183-191.
19. Relton CL, Hartwig FP, Davey Smith G. From stem cells to the law courts: DNA methylation, the forensic epigenome and the possibility of a biosocial archive. *Int J Epidemiol*. 2015;44(4):1083-1093.
20. Labos C, Brophy JM, Smith GD, Sniderman AD, Thanassoulis G. Evaluation of the Pleiotropic Effects of Statins: A Reanalysis of the Randomized Trial Evidence Using Egger Regression- Brief Report. *Arterioscler Thromb Vasc Biol*. 2018;38(1):262-265.

3 – Artigo de revisão

Breastfeeding effects on DNA methylation in the offspring: a systematic literature review

Short title: Breastfeeding effects on DNA methylation

Fernando Pires Hartwig^{1,2*}, Christian Loret de Mola¹, Neil Martin Davies^{2,3}, Cesar Gomes Victora¹, Caroline L. Relton^{2,3}

¹Postgraduate Programme in Epidemiology, Federal University of Pelotas, Pelotas, Brazil.

²MRC Integrative Epidemiology Unit, School of Social & Community Medicine, University of Bristol, Bristol, UK.

³School of Social and Community Medicine, University of Bristol, United Kingdom.

*Corresponding author. Postgraduate Program in Epidemiology, Federal University of Pelotas, Pelotas (Brazil). Zip code: 96020-220. Phone: 55 53 981068670. E-mail: fernandophartwig@gmail.com.

Abstract

Background: Breastfeeding benefits both infants and mothers. Recent research shows long-term health and human capital benefits among individuals who were breastfed. Epigenetic mechanisms have been suggested as potential mediators of the effects of early-life exposures on later health outcomes. We reviewed the literature on the potential effects of breastfeeding on DNA methylation.

Methods: Studies reporting original results and evaluating DNA methylation differences according to breastfeeding/breast milk groups (e.g., ever vs. never comparisons, different categories of breastfeeding duration, etc) were eligible. Six databases were searched simultaneously using Ovid, and the resulting studies were evaluated independently by two reviewers.

Results: Seven eligible studies were identified. Five were conducted in humans. Studies were heterogeneous regarding sample selection, age, target methylation regions, methylation measurement and breastfeeding categorisation. Collectively, the studies suggest that breastfeeding might be negatively associated with promoter methylation of *LEP* (which encodes an anorexigenic hormone), *CDKN2A* (involved in tumour suppression) and *Slc2a4* genes (which encodes an insulin-related glucose transporter) and positively with promoter methylation of the *Nyp* (which encodes an orexigenic neuropeptide) gene, as well as influence global methylation patterns and modulate epigenetic effects of some genetic variants.

Conclusions: The findings from our systematic review are far from conclusive due to the small number of studies and their inherent limitations. Further studies are required to understand the actual potential role of epigenetics in the associations of breastfeeding with later health outcomes. Suggestions for future investigations, focusing on epigenome-wide association studies, are provided.

Introduction

Breastfeeding has well-established short-term health benefits, and there is increasing evidence that it also has long-term effects on health and human capital [1]. For the effects of an early exposure to persist over time, the exposure must leave some kind of “mark” in the organism [2]. Epigenetics processes – i.e., mitotically heritable events other than changes in DNA sequence that regulate gene expression – have been proposed as important mediators in the developmental origins of health and disease (DOHaD) context [3-5]. Currently, the most frequently studied epigenetic process is DNA methylation, which (in mammals) is the addition of a methyl ($-\text{CH}_3$) group to DNA at the 5' position of a cytosine base. In mammals, DNA methylation most commonly occurs in cytosine-guanine (CpG) dinucleotides located in genomic regions called CpG islands – i.e., DNA sequences rich in CpG dinucleotides [6,7].

The notion of epigenetic effects of breastfeeding seems to be widely held, and a Google search (January 23, 2017) using the search terms “epigenetics breastfeeding” resulted in approximately 111,000 hits. There is indeed some evidence supporting the notion that breast milk influences DNA methylation. For example, early-life supplementation of omega-3 fatty acids (an important nutritional compound of breast milk) was associated with methylation profiles in pigs [8]. It has also been hypothesised that the microbiome mediates the effects of breast milk on DNA methylation, since there is evidence that breastfeeding influences the composition of the gut microbiota and that the latter influences DNA methylation [9]. Breast milk also contains long non-coding RNAs [10] and small non-coding RNAs called microRNAs [11], which are involved in gene expression regulation at the post-transcriptional level, suggesting that epigenetic effects of breast milk may not be restricted to DNA methylation.

Three separate literature reviews available to date have suggested the existence of epigenetic effects of breast milk [9,11,12]. However, these reviews were non-systematic and mostly based on evaluations of breast milk properties in isolation rather than comparisons of groups of humans or animals with different feeding modes. We therefore aimed at systematically reviewing the literature on the association between breastfeeding and DNA methylation in humans and animal models.

Methods

Search strategy

A systematic review of the literature was performed in August 22, 2016 through Ovid (<https://ovidsp.tx.ovid.com/>), which allows simultaneously searching of the following databases: MEDLINE, Embase, Allied and Complementary Medicine Database, CAB ABSTRACTS, PsycINFO®, and The Philosopher's Index. By default, Ovid searches the following fields (some of which are database-specific) when all of its databases are searched: Title, Original Title, Title Comment, Abstract, Subject Heading Word, MeSH Subject Headings, Keyword Heading, Keyword Heading Word, Key Concepts, Full Text, Cited Reference Author Word and others.

The following search terms were used for breastfeeding: “breastfe\$” OR “breast fe\$” OR “bottle fe\$” OR “formula fe\$” OR “infant feeding” OR “human milk” OR “breast milk” OR “formula milk” OR “weaning”. For epigenetics, the search terms were: “epigenetic\$” OR “epigenom\$” OR “methy\$” OR “methQTL” OR “mQTL”. Using the wildcard character “\$” retrieves any number (including zero) of characters after the stem word (e.g., “breastfe\$” retrieves “breastfeeding”, “breastfed”, etc). The two group of search terms were combined using the AND operator: “Breastfeeding” AND “Epigenetics”.

Study selection and data collection

The aim of our review was to identify studies on DNA methylation differences associated with breastfeeding. Studies were excluded if they met at least one of the following criteria: i) not reporting effects of breastfeeding on DNA methylation (e.g., studies of epigenetic determinants of breastfeeding, such as the association between methylation in promoters of genes involved in breast milk production); ii) being limited to specific breast milk components rather than breastfeeding or breast milk as a whole; iii) not reporting original data.

Eligibility was assessed independently by two reviewers (F.P.H. and C.L.M.), and disagreements were resolved by consensus. Initially, duplicate records were excluded,

titles screened and abstracts reviewed. For the remaining studies, full-texts were examined.

The following data were extracted from the included studies:

- i) First author's name and publication year.
- ii) Country where the study was conducted.
- iii) Study aim and design.
- iv) Species, number of individuals, % of females and age.
- v) Methylation region, DNA source, measurement method and outcome (e.g., proportion of methylated cells).
- vi) Breastfeeding categorisation (e.g., never vs. ever, duration in months, etc) and age at ascertainment.
- vii) Covariates.
- viii) Breastfeeding-methylation association results.

Data analysis

Given the lack of consistency between the designs and methods among the studies (as described below), we opted for a narrative review rather than attempting to perform a meta-analysis.

Results

We evaluated the sensitivity and specificity of our search strategy in a pilot search (S1 Appendix). Briefly, we noted that the Ovid filter to remove non-original publications would likely remove some studies with original data, while the English language filter would likely not substantially influence our study. This pilot search allowed us to reduce the number of publications retrieved in the main search without reducing its sensitivity.

Fig 1 displays a flow diagram of the study selection process. The initial search yielded 5348 records. Of these, 1076 were duplicates. Of the 4272 unique records, 884 were excluded because they were publication types unlikely to include original results according to our pilot search. The remaining 3388 records were screened based on their titles and abstracts, yielding 19 original publications. Another 29 non-original publications were selected only for reference list searching for additional eligible studies, thus totalizing 48 records (S1 Table). After evaluating the full-texts and reference lists, 7 records (6 journal articles and 1 conference abstract) were included (Table 1).

Fig 1. Flow diagram of study selection.

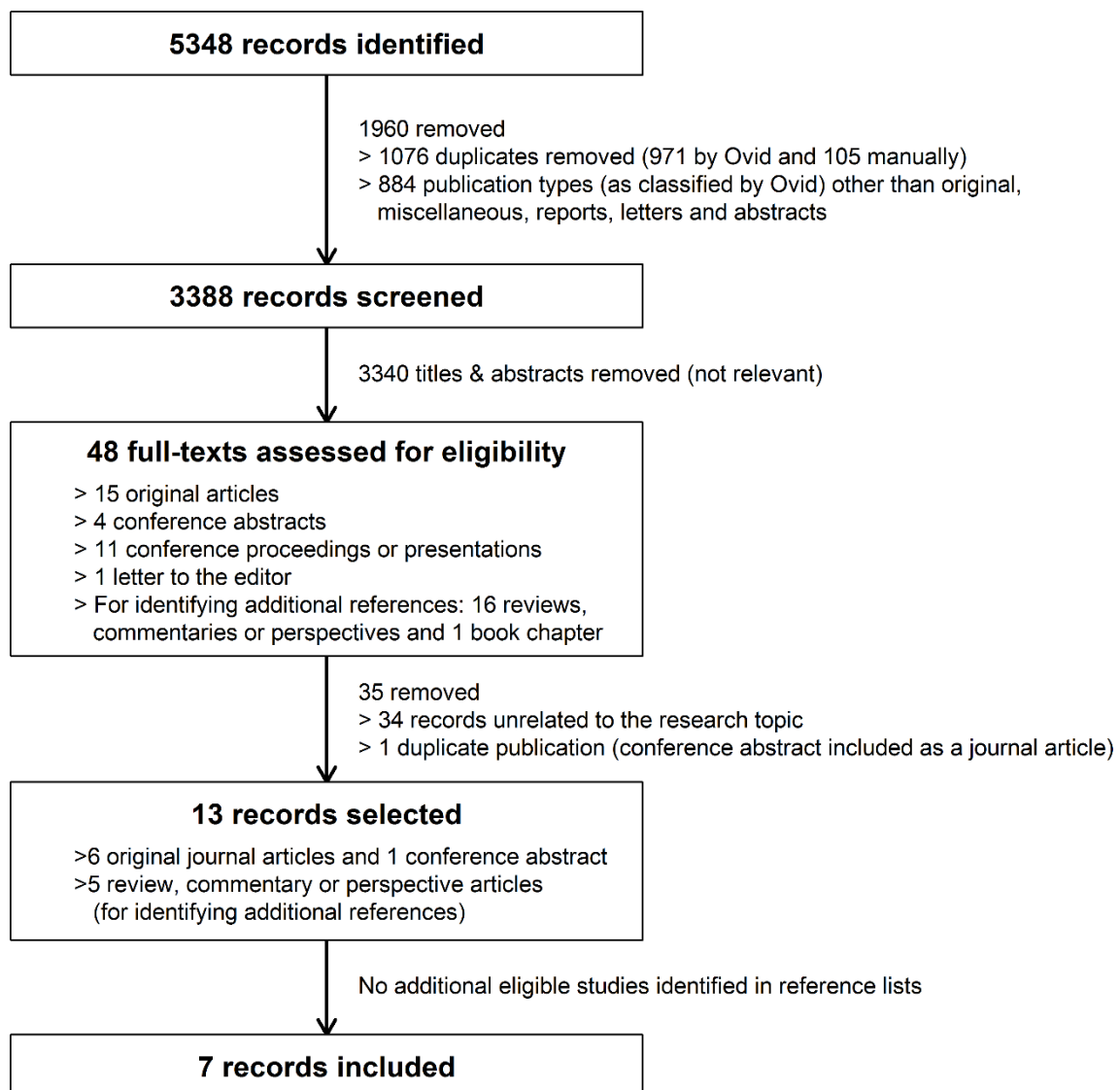


Table 1. Characteristics of studies included in the review.

Characteristic	First author, year						
	Obermann-Borst, 2013	Rossnerova, 2013	Soto-Ramirez, 2013	Tao, 2013	Simpkin, 2016	Mahmood, 2013	Raychaudhuri, 2014
Country	Netherlands	Czech Republic	England	USA	England ^e	USA	USA
Study aim	Evaluate the association of early-life factors with <i>LEP</i> promoter methylation in young children	Evaluate if there were methylation differences comparing regions with different levels of air pollution and asthma case/control groups. Other variables (including breastfeeding) were evaluated in secondary analyses	Evaluate potential interactions among genetic variants, CpG sites and breastfeeding, as well as their relationship with asthma	Evaluate the association of early-life factors with methylation in the promoter regions of three genes in breast tumour tissues	Evaluate the association of early-life factors with epigenetic age in children and adolescents.	Compare breast milk with a high-carbohydrate formula regarding epigenetic regulation of <i>Npy</i> and <i>Pomc</i> genes in the hypothalamus in rats	Compare breast milk with a high-carbohydrate formula regarding epigenetic regulation of <i>Slc2a4</i> gene in the skeletal tissue in rats
Sample characteristics							
Species	Humans	Humans	Humans	Humans	Humans	Rats	Rats
N	120	200 (100 asthmatics and 100 controls)	245	639 (all breast cancer cases)	Up to 974	32 (16 per group)	12 (6 per group)
% females	42	45	100	100	52	100	0
Mean age (SD)	1.4 years (0.2)	11.6 years (2.2)	18.0 years (NA)	57.5 years (11.3)	At birth (NA), 7.5 (0.15) and 17.14 years (1.01)	16 (0) and 100 (0) days	100 days (0)
Design	Cross-sectional	Cross-sectional ^b	Longitudinal	Case-case ^d	Longitudinal	Experimental	Experimental
Methylation							
Region	<i>LEP</i> promoter ^a	Global methylation ^c	CpG regions associated with 17q12 genetic variation	<i>CDH1</i> , <i>CDKN2A</i> and <i>RARB</i> promoters	353 CpG sites used to estimate epigenetic age	<i>Pomc</i> and <i>Npy</i> promoters	<i>Slc2a4</i> promoter
DNA source	Peripheral blood	Peripheral blood	Peripheral blood	Paraffin-embedded	Cord and peripheral blood	Hypothalamus	Skeletal muscle

Outcome	Proportion of methylated DNA copies	Principal component scores of multiple methylated regions	Proportion of methylated DNA copies	tumour tissue Methylation status (yes/no)	Epigenetic age acceleration (regression of epigenetically-predicted age on chronological age), in years	Proportion of methylated DNA copies	Difference of normalised methylation measures ^f
Measurement	Mass spectrometry-based quantification of PCR amplicons from bisulfite-converted DNA	Infinium HumanMethylation27 BeadChip	Infinium HumanMethylation450 BeadChip	Methylation-specific qPCR using bisulfite-converted DNA	Infinium HumanMethylation450 BeadChip	Mass spectrometry-based quantification of <i>in vitro</i> transcripts generated using PCR amplicons from bisulfite-converted DNA	Southern blot after methylation-sensitive enzymatic cleavage
Breastfeeding							
Categorisation	Score ranging from 0 to 4, corresponding to 0, >1 – <1, >1 – 3, >3 – 6 and >6 months of duration of any breastfeeding, respectively	Duration of full breastfeeding in months	Duration in weeks	0: Ever. 1: Never.	0: Never. 1: Ever.	Breast milk vs. high-carbohydrate milk formula (both from postnatal days 4 to 16 or 24)	Breast milk vs. high-carbohydrate milk formula (both from postnatal days 4 to 24)
Mean age at ascertainment	1.4 years	11.6 years	1-2 years	57.5 years	1.0 month	Not applicable	Not applicable
Covariates	Bisulfite batch,	None	None	Menopause	Epigenetic age	None	None

	CpG site, maternal education and smoking at birth, sex, birth weight, current BMI and serum leptin			status (stratification), age, education, race and estrogen receptor status	acceleration was adjusted for cellular heterogeneity		
Result	-0.6 (95% CI: -1.19; -0.01) percentage points in methylation per increment in breastfeeding duration category	Pooling asthmatic subjects and controls, breastfeeding was apparently associated with patterns of overall DNA methylation, although no statistical test was performed	There was an interaction between breastfeeding and mQTLs regarding the methylation levels of 10 CpG sites	Odds ratio of <i>CDKN2A</i> promoter methylation was 2.75 (95% CI: 1.14; 6.62) times higher in never breastfed women, but only in the premenopausal group (stratification on menopausal status was defined <i>a priori</i>)	Pearson's correlation coefficients (r) and associated P-values (P) were for the association between breastfeeding and epigenetic age acceleration were r=0.035 and P=0.301 (at birth), r=-0.010 and P=0.756 (in childhood), and r=0.026 and P=0.434 (adolescence)	<i>Nyp</i> promoter methylation was generally higher in the breast milk compared to the high-carbohydrate formula group. However, there was no strong evidence for methylation differences in the <i>Pomc</i> promoter	<i>Slc2a4</i> promoter methylation was lower in the breast milk compared to the high-carbohydrate formula group

PCR: polymerase chain reaction. qPCR: quantitative PCR. NA: not available. CpG site: genomic region rich in cytosine-guanine dinucleotides. mQTLs: methylation quantitative trait loci (i.e., genetic variants associated with methylation levels).

^aBased on seven CpG sites. The outcome for the primary analysis was average methylation across these sites in linear mixed models, although individual-site analyses were also performed.

^bEven though study participation also depended on asthma case/control status and region, the variables under consideration are methylation and breastfeeding.

^cPrincipal component analysis was performed to generate variables that represent global methylation patterns.

^dThe original study was a population-based case-control study, but the analyses involving breastfeeding and methylation were restricted to cases.

^eEven though Danish and German individuals were also studied in the replication stage, the analyses involving breastfeeding were performed in British individuals only.

^fMethylation differences were measured using the difference in Southern Blot signal detection between *HpaII*- (blocked by CpG methylation) and *MspI*- (methylation-insensitive) digested DNA, after normalisation to *Actb* gene.

There were five studies in humans and two in rats, all in high-income countries. Human studies included two cross-sectional studies, two longitudinal studies and one case-only study, with a mean age range of 0 (at birth) to 57.5 years. All studies evaluated distinct and limited genomic regions using six different measurement techniques, although five used methods that involved bisulfite DNA conversion. Four studies analysed blood samples, one analysed paraffin-embedded tumour tissues and the animal studies analysed skeletal muscle and the hypothalamus. Studies also differed regarding breastfeeding categorisation, mean age at ascertainment, selection of covariates and presentation of results.

Human studies

Obermann-Borst et al. (2013).

This was a cross-sectional study in 120 Dutch children (50 girls) at an average age of 1.4 years [13]. The outcome was methylation at the *LEP* gene (which encodes the anorexigenic hormone leptin) promoter in peripheral blood. Methylation was measured using a mass spectrometry-based method involving bisulfite conversion of DNA, yielding the proportion (from 0 to 1) of methylated DNA copies at the sites investigated. In the main analyses, seven different CpG sites in the *LEP* promoter were analysed simultaneously as the outcome variable, using linear mixed models to account for repeated measures. Therefore, the outcome variable can be interpreted as the average methylation in the *LEP* gene promoter as measured by those seven CpG sites. Batch and CpG site were adjusted for in all analyses as fixed effects. Each CpG site was individually evaluated in secondary analysis. Importantly, it is uncertain whether those seven CpG sites, which are within a <170 bp-long region [14], are representative of overall methylation status in this CpG island, which is 625 bp-long and contains 58 CpG sites. Features for this CpG island can be found at the USCS Genome Browser (GRCh38/ hg38 assembly) by searching using the following coordinates: chr7:128,240,698-128,241,322.

Breastfeeding was analysed as a score ranging from 0 to 4, corresponding to 0, >1 – <1, >1 – 3, >3 – 6 and >6 months of duration of any breastfeeding, respectively. Information was recorded when the child was 1.4 years old through self-administered questionnaires completed by the mothers. The following characteristics were also evaluated as exposure variables: education, folic acid supplementation and smoking at birth (maternal); sex, birth weight, age, serum leptin levels, growth rate and body mass index (BMI) (children).

In unadjusted analyses, each 1-unit increment in the breastfeeding score was associated with a reduction of 0.6 (95% confidence interval [CI]: 0.01; 1.19) percent points in the proportion of methylated copies of DNA. This corresponded to a relative reduction of 2.9% in DNA methylation. The results were virtually unchanged in analyses adjusting for maternal education and smoking, as well as sex, birth weight, BMI and serum leptin levels of the children. Because child BMI and leptin levels were measured at the average age of 1.4 years, they are not potential confounders of the breastfeeding-methylation association. Indeed, they are potential consequences of *LEP* gene methylation, so adjusting for them might have introduced bias. Nevertheless, it is reassuring that doing so had little effect on the results.

Rossnerova et al. (2013).

This Czech study [15] evaluated 200 individuals (mean age of 11.6 years; 89 girls), of whom 100 presented asthma and 100 did not. Half of cases and controls lived in a highly polluted region; the remaining individuals lived in a control region. Case/control status regarding asthma and region were the main exposure variables. Secondary analyses evaluated sex, length of gestation (weeks), birth weight (g), cotinine levels (ng/mg) and length of fully breastfeeding (months).

Methylation was measured in peripheral blood using the Infinium HumanMethylation27 BeadChip, which uses bisulfite DNA conversion and provides the proportion of methylated copies of DNA for approximately 27,800 methylation sites spanning approximately 14,500 genes. This technology has been superseded by a more comprehensive method (described below). For the analysis involving breastfeeding, methylation was evaluated as overall methylation patterns (rather than CpG-site

specific analysis) through partial least squares (PLS) with 3 latent factors (although results shown were limited to the 1st and 2nd factors only) and length of gestation, birth weight, cotinine levels and breastfeeding as outcome or response variables. Individuals who were breastfed for longer time had higher values of both factors. Even though this was graphically clear, none of the analyses involving breastfeeding and methylation used statistical tests, which would be essential to evaluate the possible role of chance in the findings.

Furthermore, evaluating the association of breastfeeding with DNA methylation using PLS has some limitations. PLS is not optimal for understanding the relationships between variables. Indeed, the apparently positive relationship of breastfeeding with the PLS factors is difficult to interpret beyond the simple observation that breastfeeding is related to overall patterns of methylation. Second, it is not mentioned in the publication how much of the variation in methylation the 3 PLS factors account for. If this value is low, it is possible that other PLS factors that would account for non-negligible amounts of variation in breastfeeding (which would be indicative of an association between breastfeeding and methylation) might be missed. Analysing the association of breastfeeding with each methylation site individually – a strategy known as epigenome-wide association study (EWAS) [7] – would have provided important and more interpretable biological insights into the potential epigenetic effects of breastfeeding and would have complemented the PLS findings. However, a EWAS in such sample size would likely be underpowered.

Soto-Ramirez et al. (2013).

This study (published as a conference abstract) [16] was performed in 245 females participating in the 1989 Isle of Wight Birth Cohort. Peripheral blood methylation data obtained at 18 years of age using the Infinium HumanMethylation450 BeadChip, which provides the proportion of methylated DNA copies for over 485,000 sites, covering 99% of RefSeq (<http://www.ncbi.nlm.nih.gov/refseq/>) genes. Based on a related publication using this cohort [17] it was possible to identify that breastfeeding was analysed as duration in weeks (probably any breastfeeding, although not specified), ascertained when

participants were 1-2 years of age. The overall aim of the study was to evaluate whether there are interactions among breastfeeding, genetic and epigenetic variants with respect to asthma risk. Some important aspects were unclear (possibly due to the brevity of the conference abstract). Following our contact, the authors of the study kindly provided clarifications and additional results, which are described below.

Firstly, eight genetic variants (selected using a linkage disequilibrium filter out of 20 genotyped variants) at the 17q21 locus were tested for association (one at a time) with methylation levels at 26 CpG sites (one at a time) in the same region. The model included the main effects of breastfeeding and genetic variants and an interaction term between these variables. 10 out of the 26 CpGs were influenced by interactions between breastfeeding and genetic variants. This suggests that breastfeeding may modulate the epigenetic effects of some methylation quantitative trait loci (ie, the epigenetic effects of those mQTLs vary according to breastfeeding status). However, it is also possible that some genetic profiles reduce the plasticity of the epigenome, thus mitigating the epigenetic effects of environmental factors. For example, a single nucleotide polymorphism may abrogate a CpG site, thus preventing it from being methylated regardless of the states of other determinants of methylation levels at this specific site. It was not possible to investigate the interaction mechanisms of these associations because neither regression coefficients nor stratified results were available.

Similarly to the study by Rossnerova and colleagues [15], performing an EWAS would have provided important additional biological insights, especially given that the Infinium HumanMethylation450 BeadChip was used, which is the current gold-standard for EWAS in epidemiology studies. Moreover, this study has not been yet published as a full, per-reviewed article, so it must be interpreted in its current form with caution. Study strengths included control of type-I error inflation using the false discovery rate and a relatively short recall period of breastfeeding measurement.

Tao et al. (2013).

Tao and colleagues [18] evaluated whether early-life factors are associated with promoter methylation of the *CDH1* (which encodes the cell-adhesion protein cadherin-

1), *CDKN2A* (which encodes important tumour suppression proteins such as p14 and p16) and *RARB* (which encodes a receptor for retinoic acid) genes. The analyses involving breastfeeding included 639 women (mean age of 57.5 years) with breast cancer participating in the Western New York Exposures and Breast Cancer Study. Methylation was measured in paraffin-embedded breast tumour tissues using bisulfite-converted DNA followed by methylation-specific quantitative polymerase chain reaction (qPCR). This yielded a binary variable (methylated/unmethylated) for each promoter region. Importantly, since breastfeeding occurred before disease onset, any potential epigenetic effects of breastfeeding would primarily affect healthy cells. Therefore, for associations between breastfeeding and methylation to be detectable in this study, they must still be discernible in tumour tissues. Given that epigenetic dysregulation occurs in many cancers [19,20], it is possible that methylation changes caused by the disease distorted breastfeeding-methylation associations. This issue would have been addressed by analysing paired non-cancerous tissues.

The associations of breastfeeding with promoter methylation were adjusted for age, education, race and estrogen receptor status, and were reported comparing never with ever (reference group) breastfed women. The analyses were also stratified according to menopausal status. In premenopausal women (n=205), odds ratio estimates were 1.21 (95% CI: 0.50; 2.93), 2.75 (95% CI: 1.14; 6.62) and 1.18 (95% CI: 0.53; 2.62) for *CDH1*, *CDKN2A* and *RARB* promoters, respectively. In postmenopausal women (n=434), the corresponding estimates were 1.06 (95% CI: 0.64; 1.77), 0.79 (95% CI: 0.49; 1.26) and 1.30 (95% CI: 0.83; 2.04). Analyses were also performed using a composite outcome variable: 1: ≥ 1 of the three promoters was methylated; 0: none of the promoters was methylated. In these analyses, the odds ratio estimates were 1.87 (95% CI: 0.91; 3.83) and 1.02 (95% CI: 0.67; 1.57) in premenopausal and postmenopausal women, respectively.

Although the above findings suggest that breastfeeding might be related to *CDKN2A* promoter methylation, there were some important limitations. The analyses involved three promoter regions, eight exposure variables, and stratification according to menopausal status. This adds up to 48 comparisons, thus inflating the type-I error rate, which was not corrected. Moreover, although there are conceptual reasons for

stratifying according to menopausal status, interaction tests would have been informative regarding whether or not the associations differ between the strata. It is also important to consider that case-control studies involve conditioning on a descendent of the outcome variable. In this study, this is even more pronounced, since it was conditioned on the outcome variable itself. In this situation, associations between breastfeeding and methylation profiles may be biased in different ways, depending on the underlying causal relationships [21]. Therefore, investigating the association between breastfeeding and methylation profiles using other study designs, such as cross-sectional or, ideally, longitudinal studies would be preferred [22].

Simpkin et al. (2016).

This study analysed the association between early-life factors with epigenetic age acceleration [23]. The analyses involving breastfeeding (0: never; 1: ever) were performed in up to 974 participants in the Accessible Resource for Integrated Epigenomic Studies (ARIES) project, a sub-study of the Avon Longitudinal Study of Parents and Children [24]. Individuals were epigenotyped using the Infinium HumanMethylation450 BeadChip at birth (cord blood), in childhood and adolescence (peripheral blood). Epigenetic age was estimated using 353 CpG sites applied using the Horvath method [25], and epigenetic age acceleration was computed as the residuals of regressing epigenetic on chronological age. Epigenetic age is an attempt to quantify biological age, and epigenetic age acceleration indicates how much an individual's epigenetic age is ahead (positive values) or behind (negative values) of his or her chronological age [23].

Breastfeeding was not associated with epigenetic age acceleration at any of the time points investigated in this study, with Pearson's correlation coefficients (P-values) ranging in magnitude from -0.010 (P=0.756) to 0.026 (P=0.434).

The heterogeneity in cell-type composition between cord and peripheral blood (as well as between-individual differences in cell-type composition in the same tissue) could distort associations between breastfeeding and epigenetic clock. In this study [23], epigenetic age was adjusted for cell-type composition estimated using DNA methylation data, as described elsewhere [24,26]. Although measured cell-type

composition would be ideal, the estimates used likely at least attenuate any potential confounding. Moreover, Horvath method to estimate epigenetic age is less affected by cell-type composition than Hannum method [27], thus attenuating the possibility of residual confounding even more. Furthermore, it is possible that this study was underpowered to detect modest effects of breastfeeding on epigenetic age acceleration. This problem could have been attenuated by statistical adjustment for covariates that temporally precede breastfeeding and were associated with epigenetic age acceleration in one or more time points. If those variables are also associated with breastfeeding, this would have also contributed to reducing negative confounding that might exist in the estimates.

Animal studies

We identified many studies evaluating epigenetic effects of different forms of early-life feeding in animal models, but only two [28,29] comparing breastfeeding with a breast milk substitute.

Mahmood et al. (2013).

This study [28] included two groups with sixteen female rats each: one received breast milk and the other received a high-carbohydrate formula. Half the animals in each group were weaned at postnatal day 16 and the other half at day 24, when animals started to receive standard laboratory rodent diet and water *ab libitum*. Epigenetic measures of the promoter regions of the *Pomc* (which encodes a precursor of many peptide hormones) and *Npy* (which encodes the neuropeptide Y) genes promoter were obtained 16 and 100 days after birth in the hypothalamus. Both genes are involved in many physiological processes, including energy homeostasis. Methylation was measured using Sequenom MassARRAY quantitative methylation analysis [30], which yields the proportion of methylated copies of DNA at a specific genomic site.

Rats that received breast milk were shown to display higher methylation in the *Nyp* promoter compared to the high-carbohydrate formula group. They also showed lower levels of *Nyp* mRNA and of histone acetylation (which is another epigenetic

marker). Regarding *Pomc* promoter methylation, there was no strong evidence of a difference. However, the breast milk group presented higher *Pomc* mRNA levels, possibly linked to the higher levels of histone acetylation in this group.

Raychaudhuri et al. (2014).

This study [29] design was similar to the aforementioned study,[28] with the following differences: i) all rats were males; ii) there were six rats in each feeding group; iii) weaning occurred at postnatal day 24 only; iv) epigenetic measures were taken 100 days after birth in skeletal muscle tissues. v) the *Slc2a4* gene (which encodes the Glut-4 protein, an insulin-regulated glucose transporter) promoter was evaluated.

Methylation was measured using methylation-sensitive enzymatic cleavage followed by Southern blot. The general idea is to use two enzymes that can cleave the DNA given the presence of specific DNA sequences (called restriction sites). However, the activity of one of such enzymes is blocked if the DNA is methylated, while the other is not. Therefore, DNA fragmentation patterns after enzymatic cleavage depend on methylation. By using a probe that binds to a specific region of the target gene promoter that contains the restriction site, it is possible to measure methylation differences in such promoter. Since the signal was normalised by dividing to a loading control (in this case, the *Actb* gene), the results were in arbitrary units. This form of measurement is semi-quantitative.

Using this strategy, Raychaudhuri and colleagues reported that *Slc2a4* promoter methylation was lower in rats that received breast milk compared to the high-carbohydrate formula group. They also showed higher levels of *Slc2a4* gene expression at both transcriptional and protein levels. Additional evaluations (such as differences in histone acetylation) complemented the results.

Given the experimental nature and the fact that they were performed in an animal model, the two animal studies could evaluate the epigenetic event in the target rather than in a surrogate tissue. They also showed that the observed epigenetic differences were associated with changes in gene expression, suggesting a functional implication of such intervention-mediated epigenetic events.

However, several factors in the two aforementioned animal studies must be considered before extrapolating their findings to humans. First, the purpose of feeding some animals with a high-carbohydrate formula was to evaluate the epigenetic effects of a high-carbohydrate diet in early life, rather than being an attempt to mimic rat milk effects as closely as possible (as in the case of human milk substitutes). This hampers the interpretation of the results, because the epigenetic differences between the two feeding groups could be due to either particular properties of rat milk (e.g., specific nutritional components that have epigenetic effects) or simply the high carbohydrate content in the formula. This issue would have been minimised if it had been an artificial rearing control group fed – i.e., pups artificially fed with rat milk or formula milk that is as similar as possible to rat milk (see below). There was no such group due to the absence of substantial differences between artificial rearing groups fed with a high-carbohydrate formula and with a formula that had a similar caloric distribution to that of rat milk in previous studies [31-33]. However, it may well be the case that the rearing mode is distorting the results because it is well-known that maternal care has epigenetic effects on the offspring [34-37]. Therefore, it is not possible to know if the epigenetic differences between the experimental groups were due to feeding (i.e., high-carbohydrate formula vs. rat milk) or to rearing (i.e., artificial vs. maternal nursing).

Discussion

Our study summarizes the current evidence regarding the association of breastfeeding with DNA methylation. Collectively, the studies we identified suggest that breastfeeding might be associated with promoter methylation of the *LEP* [13] (negatively) and *CDKN2A* [18] (negatively) genes in humans, and *Npy* [28] (positively) and *Slc2a4* [29] (negatively) genes in rats, as well as implicated in global methylation patterns [15] and in modulation of epigenetic effects of some genetic variants [16]. Moreover, in the *LEP*, *Npy* and *Slc2a4* studies, gene promoter methylation was also associated with higher gene expression levels. This is in agreement with the notion that gene promoter methylation is commonly, although not universally, associated with lower gene expression [38]. Higher gene expression levels of *LEP*, *Pomc* and

Slc2a4 genes and lower levels of the *Npy* gene in breastfed individuals is in agreement with other epidemiological evidence that breastfeeding might protect against obesity and diabetes [1]. *CDKN2A* products have important tumour suppression roles [39] so if breastfeeding really does increase *CDKN2A* expression via epigenetic changes, then it has the potential to protect against cancer. Nevertheless, given the small number of studies and their limitations, it would be premature to make any firm conclusions regarding epigenetic effects of breastfeeding.

In spite of the small number of studies directly addressing the association of breastfeeding with DNA methylation, some authors expressed high expectations regarding these associations (e.g., this commentary [40] and the Google search mentioned above). Although the studies we identified collectively indicate that breastfeeding might be associated with DNA methylation, our systematic review indicates that the evidence is far from compelling and much more research is needed on this topic. Importantly, the present review was focused on DNA methylation changes related to breastfeeding. Future reviews may also address DNA methylation differences due to other foodstuffs or to maternal diets, and to epigenetic changes other than DNA methylation.

In our search we prioritised sensitivity over specificity at the search stage, in order to minimise the possibility of failing to identify eligible studies, which would be particularly relevant in light of the small number of studies on the topic. For this purpose, we searched for studies in many literature databases and piloted our search criteria and filters to avoid excluding eligible studies. The fact that we identified (and included) an eligible abstract and a study that evaluated breastfeeding only in secondary analysis also argues in favour of the sensitivity of our search.

Although our systematic review suggests that breastfeeding might influence DNA methylation, its main conclusion is that more (and better) studies are needed. Particularly, given the focus to date on candidate gene studies or global (non-site specific) measures of methylation, EWAS studies would be very useful to identify regions of the methylome associated with (and possibly influenced by) breastfeeding. Furthermore, these studies must be adequately powered to identify subtle differences

in DNA methylation. We used the findings from Obermann-Borst et al. [13] to estimate the sample sizes required to detect DNA methylation differences according to breastfeeding in an EWAS in a total of 18 situations (S2 Appendix and S2 Table). In six of them, up to 1000 individuals were required, suggesting that existing resources (such as the ARIES project) may be properly powered. However, in other scenarios larger sample sizes would be required, and achieving them may be possible through collaborative effort and consortia-based science, examples of which are emerging in the epigenetic literature [41]. Importantly, our calculations are limited because the parameters were obtained from a single study evaluating a single methylation locus with a different method than that used in EWAS.

It is also important that EWAS studies of breastfeeding control for important potential confounding variables. S2 Fig displays postulated causal relationships among breastfeeding, DNA methylation and potential important confounders in the form of a directed acyclic graph [42]. It is well-known that ancestry/ethnicity is an important determinant of indicators of socioeconomic position (e.g., as income, educational attainment, etc) [43-45], and the allele frequencies of many genetic variants are associated with ancestry/ethnicity [46]. Moreover, socioeconomic position is associated with breastfeeding, with the direction of the association differing between income settings [1]. Therefore, if ancestry/ethnicity is associated with genetic variants with direct (i.e., not mediated by breastfeeding) effects on DNA methylation, it may act as a confounder.

Horizontally pleiotropic genetic variants [47] may also confound the association between breastfeeding and DNA methylation. Such horizontal pleiotropy could be mediated, for example, by maternal pre-pregnancy (such as body mass index and parity) and gestational factors (such as maternal smoking during pregnancy, type of delivery and birth weight). This is because epidemiological studies suggest that these factors may influence both breastfeeding [48-54] and epigenetic events [55-61]. Therefore, maternal pre-pregnancy and gestational factors may confound the association between breastfeeding and DNA methylation. Moreover, since family socioeconomic position is associated with those factors [62-65], the latter represent another pathway through which socioeconomic position and ancestry/ethnicity may

induce confounding. Another potential pathway is care/stimulation, given that it is associated with family socioeconomic position [66] and, according to studies in animal models, may lead to epigenetic modifications in the offspring. In this context, however, it is important to avoid adjusting for measures of mother-offspring bonding, which may be influenced by breastfeeding [67,68], and therefore mediate (at least partially) its epigenetic effects. Importantly, S2 Fig likely does not exhaust the list of all confounders. We opted by presenting a more parsimonious model focusing one potentially important confounders given the evidence that is currently available. Such model may serve as a basis for more comprehensive models as knowledge on the relationship between breastfeeding and DNA methylation improves.

Another important consideration for future EWAS of breastfeeding is the tissue used to extract DNA. Intra-individual variation (i.e., between tissues of the same individual) in epigenetic patterns is generally higher than variation between individuals [69,70] (although with some exceptions, such as the brain [71]), which limits investigations using easily accessible DNA sources (such as peripheral blood or saliva) when they are not the target tissue [72,73]. This may be an important limitation for epigenetic epidemiology studies of breastfeeding. For example, one of the most strongly supported long-term effects of breastfeeding is its positive association with IQ [74-76]. The optimal DNA source for studying the potential mediating role of DNA methylation in this association would clearly be the brain, but due to practical reasons large-scale epidemiological studies need to rely on easily accessible surrogate tissues. However, some studies suggest that the correlation between epigenetic signatures in the brain and in peripheral blood is generally low, with strong correlations occurring in only a few loci [77-79]. This suggests that, in the case of IQ, the epigenetic studies using DNA extracted from peripheral blood mononuclear cells may provide limited information about DNA methylation in the target tissue. However, this does not mean that such studies are of no utility, since results from some loci would still provide information relevant to the target tissue. Moreover, epidemiological studies suggest that breastfeeding may have long-term effects on other disease outcome, such as obesity and diabetes [1]. More generally, findings from surrogate tissues may provide important insights into the potential range of epigenetic effects of breastfeeding,

which may thus inform subsequent studies in tissues of difficult access such as the brain, as well as *in vitro* and *in vivo* studies in animal models. Combining evidence from studies in humans and animals, exploring the strengths of each, is likely to be a fruitful strategy to improve knowledge on the potential epigenetic effects of breastfeeding.

A well-designed and appropriately powered EWAS with good measures of important potential confounders of the association between breastfeeding and DNA methylation would provide important biological insights regarding the well-established associations of breastfeeding with a range of health outcomes [1], as well as to identify potential new biological pathways related to breastfeeding. Moreover, longitudinal DNA methylation data will allow not only identification regions in the methylome associated with breastfeeding, but whether or not such associations persist over time [22,55,61].

Our conclusion is that, in spite of epigenetic mechanisms being postulated by many to explain the links between breastfeeding and long-term outcomes, the literature supporting such claims is remarkably limited. With tempered expectations, adequate definitions and proper research, our understanding of the relationship between breastfeeding and the epigenome will likely improve.

References

1. Victora CG, Bahl R, Barros AJ, Franca GV, Horton S, Krasevec J, et al. Breastfeeding in the 21st century: epidemiology, mechanisms, and lifelong effect. *Lancet*. 2016;387:475-490.
2. Relton CL, Hartwig FP, Davey Smith G. From stem cells to the law courts: DNA methylation, the forensic epigenome and the possibility of a biosocial archive. *Int J Epidemiol*. 2015;44:1083-1093.
3. Godfrey KM, Lillycrop KA, Burdge GC, Gluckman PD, Hanson MA. Epigenetic mechanisms and the mismatch concept of the developmental origins of health and disease. *Pediatr Res*. 2007;61:5R-10R.

4. Gluckman PD, Hanson MA, Mitchell MD. Developmental origins of health and disease: reducing the burden of chronic disease in the next generation. *Genome Med.* 2010;2:14.
5. Waterland RA, Michels KB. Epigenetic epidemiology of the developmental origins hypothesis. *Annu Rev Nutr.* 2007;27:363-388.
6. Han L, Su B, Li WH, Zhao Z. CpG island density and its correlations with genomic features in mammalian genomes. *Genome Biol.* 2008;9:R79.
7. Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. *Nat Rev Genet.* 2011;12:529-541.
8. Boddicker RL, Koltes JE, Fritz-Waters ER, Koesterke L, Weeks N, Yin T, et al. Genome-wide methylation profile following prenatal and postnatal dietary omega-3 fatty acid supplementation in pigs. *Anim Genet.* 2016.
9. Mischke M, Plosch T. More than just a gut instinct-the potential interplay between a baby's nutrition, its gut microbiome, and the epigenome. *Am J Physiol Regul Integr Comp Physiol.* 2013;304:R1065-1069.
10. Karlsson O, Rodosthenous RS, Jara C, Brennan KJ, Wright RO, Baccarelli AA, et al. Detection of long non-coding RNAs in human breastmilk extracellular vesicles: Implications for early child development. *Epigenetics.* 2016:0.
11. Alsaweed M, Hartmann PE, Geddes DT, Kakulas F. MicroRNAs in Breastmilk and the Lactating Breast: Potential Immunoprotectors and Developmental Regulators for the Infant and the Mother. *Int J Environ Res Public Health.* 2015;12:13981-14020.
12. Verduci E, Banderali G, Barberi S, Radaelli G, Lops A, Betti F, et al. Epigenetic effects of human breast milk. *Nutrients.* 2014;6:1711-1724.
13. Obermann-Borst SA, Eilers PH, Tobi EW, de Jong FH, Slagboom PE, Heijmans BT, et al. Duration of breastfeeding and gender are associated with methylation of the LEPTIN gene in very young children. *Pediatr Res.* 2013;74:344-349.

14. Stoger R. In vivo methylation patterns of the leptin promoter in human and mouse. *Epigenetics*. 2006;1:155-162.
15. Rossnerova A, Tulupova E, Tabashidze N, Schmuczerova J, Dostal M, Rossner P, Jr., et al. Factors affecting the 27K DNA methylation pattern in asthmatic and healthy children from locations with various environments. *Mutat Res*. 2013;741-742:18-26.
16. Soto-Ramirez N, Karmaus W, Ziyab A, Lockett G, Arshad S, Holloway J, et al (2013) The interaction of breastfeeding, DNA methylation, and genetic variants in chromosome 17q12 and the risk of asthma in girls at age 18 years. American Thoracic Society 2013 International Conference. Philadelphia, USA: American Journal of Respiratory and Critical Care Medicine. pp. A:3517.
17. Soto-Ramirez N, Arshad SH, Holloway JW, Zhang H, Schaubberger E, Ewart S, et al. The interaction of genetic variants and DNA methylation of the interleukin-4 receptor gene increase the risk of asthma at age 18 years. *Clin Epigenetics*. 2013;5:1.
18. Tao MH, Marian C, Shields PG, Potischman N, Nie J, Krishnan SS, et al. Exposures in early life: associations with DNA promoter methylation in breast tumors. *J Dev Orig Health Dis*. 2013;4:182-190.
19. Sharma S, Kelly TK, Jones PA. Epigenetics in cancer. *Carcinogenesis*. 2010;31:27-36.
20. Berdasco M, Esteller M. Aberrant epigenetic landscape in cancer: how cellular identity goes awry. *Dev Cell*. 2010;19:698-711.
21. Cole SR, Platt RW, Schisterman EF, Chu H, Westreich D, Richardson D, et al. Illustrating bias due to conditioning on a collider. *Int J Epidemiol*. 2010;39:417-420.
22. Ng JW, Barrett LM, Wong A, Kuh D, Davey Smith G, Relton CL. The role of longitudinal cohort studies in epigenetic epidemiology: challenges and opportunities. *Genome Biol*. 2012;13:246.

23. Simpkin AJ, Hemani G, Suderman M, Gaunt TR, Lyttleton O, McArdle WL, et al. Prenatal and early life influences on epigenetic age in children: a study of mother-offspring pairs from two cohort studies. *Hum Mol Genet.* 2016;25:191-201.
24. Relton CL, Gaunt T, McArdle W, Ho K, Duggirala A, Shihab H, et al. Data Resource Profile: Accessible Resource for Integrated Epigenomic Studies (ARIES). *Int J Epidemiol.* 2015;44:1181-1190.
25. Horvath S. DNA methylation age of human tissues and cell types. *Genome Biol.* 2013;14:R115.
26. Jaffe AE, Irizarry RA. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol.* 2014;15:R31.
27. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sada S, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol Cell.* 2013;49:359-367.
28. Mahmood S, Smiraglia DJ, Srinivasan M, Patel MS. Epigenetic changes in hypothalamic appetite regulatory genes may underlie the developmental programming for obesity in rat neonates subjected to a high-carbohydrate dietary modification. *J Dev Orig Health Dis.* 2013;4:479-490.
29. Raychaudhuri N, Thamotharan S, Srinivasan M, Mahmood S, Patel MS, Devaskar SU. Postnatal exposure to a high-carbohydrate diet interferes epigenetically with thyroid hormone receptor induction of the adult male rat skeletal muscle glucose transporter isoform 4 expression. *J Nutr Biochem.* 2014;25:1066-1076.
30. Ehrlich M, Nelson MR, Stanssens P, Zabeau M, Liloglou T, Xinarianos G, et al. Quantitative high-throughput analysis of DNA methylation patterns by base-specific cleavage and mass spectrometry. *Proc Natl Acad Sci U S A.* 2005;102:15785-15790.
31. Vadlamudi S, Hiremagalur BK, Tao L, Kalhan SC, Kalaria RN, Kaung HL, et al. Long-term effects on pancreatic function of feeding a HC formula to rats during the preweaning period. *Am J Physiol.* 1993;265:E565-571.

32. Mitrani P, Srinivasan M, Dodds C, Patel MS. Role of the autonomic nervous system in the development of hyperinsulinemia by high-carbohydrate formula feeding to neonatal rats. *Am J Physiol Endocrinol Metab.* 2007;292:E1069-1078.
33. Srinivasan M, Mitrani P, Sadhanandan G, Dodds C, Shbeir-ElDika S, Thamotharan S, et al. A high-carbohydrate diet in the immediate postnatal life of rats induces adaptations predisposing to adult-onset obesity. *J Endocrinol.* 2008;197:565-574.
34. Champagne FA. Epigenetic mechanisms and the transgenerational effects of maternal care. *Front Neuroendocrinol.* 2008;29:386-397.
35. Champagne FA, Curley JP. Epigenetic mechanisms mediating the long-term effects of maternal care on development. *Neurosci Biobehav Rev.* 2009;33:593-600.
36. McGowan PO, Suderman M, Sasaki A, Huang TC, Hallett M, Meaney MJ, et al. Broad epigenetic signature of maternal care in the brain of adult rats. *PLoS One.* 2011;6:e14739.
37. Gudsnek K, Champagne FA. Epigenetic influence of stress and the social environment. *ILAR J.* 2012;53:279-288.
38. Wagner JR, Busche S, Ge B, Kwan T, Pastinen T, Blanchette M. The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome Biol.* 2014;15:R37.
39. Deng Y, Chan SS, Chang S. Telomere dysfunction and tumour suppression: the senescence connection. *Nat Rev Cancer.* 2008;8:450-458.
40. Tow J. Heal the mother, heal the baby: epigenetics, breastfeeding and the human microbiome. *Breastfeed Rev.* 2014;22:7-9.
41. Joubert BR, Felix JF, Yousefi P, Bakulski KM, Just AC, Breton C, et al. DNA Methylation in Newborns and Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-analysis. *Am J Hum Genet.* 2016;98:680-696.
42. Shrier I, Platt RW. Reducing bias through directed acyclic graphs. *BMC Med Res Methodol.* 2008;8:70.

43. Chor D, Lima CR. [Epidemiologic aspects of racial inequalities in health in Brazil]. *Cad Saude Publica*. 2005;21:1586-1594.
44. Williams DR, Mohammed SA, Leavell J, Collins C. Race, socioeconomic status, and health: complexities, ongoing challenges, and research opportunities. *Ann N Y Acad Sci*. 2010;1186:69-101.
45. Quillian L. Segregation and Poverty Concentration: The Role of Three Segregations. *Am Sociol Rev*. 2012;77:354-379.
46. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. A global reference for human genetic variation. *Nature*. 2015;526:68-74.
47. Paaby AB, Rockman MV. The many faces of pleiotropy. *Trends Genet*. 2013;29:66-73.
48. Jones JR, Kogan MD, Singh GK, Dee DL, Grummer-Strawn LM. Factors associated with exclusive breastfeeding in the United States. *Pediatrics*. 2011;128:1117-1125.
49. Michels KA, Mumford SL, Sundaram R, Bell EM, Bello SC, Yeung EH. Differences in infant feeding practices by mode of conception in a United States cohort. *Fertil Steril*. 2016;105:1014-1022 e1011.
50. Kitano N, Nomura K, Kido M, Murakami K, Ohkubo T, Ueno M, et al. Combined effects of maternal age and parity on successful initiation of exclusive breastfeeding. *Prev Med Rep*. 2016;3:121-126.
51. Oakley LL, Renfrew MJ, Kurinczuk JJ, Quigley MA. Factors associated with breastfeeding in England: an analysis by primary care trust. *BMJ Open*. 2013;3.
52. Wojcicki JM. Maternal prepregnancy body mass index and initiation and duration of breastfeeding: a review of the literature. *J Womens Health (Larchmt)*. 2011;20:341-347.
53. Castillo H, Santos IS, Matijasevich A. Maternal pre-pregnancy BMI, gestational weight gain and breastfeeding. *Eur J Clin Nutr*. 2016;70:431-436.

54. Horta BL, Kramer MS, Platt RW. Maternal smoking and the risk of early weaning: a meta-analysis. *Am J Public Health*. 2001;91:304-307.
55. Richmond RC, Simpkin AJ, Woodward G, Gaunt TR, Lyttleton O, McArdle WL, et al. Prenatal exposure to maternal smoking and offspring DNA methylation across the lifecourse: findings from the Avon Longitudinal Study of Parents and Children (ALSPAC). *Hum Mol Genet*. 2015;24:2201-2217.
56. Engel SM, Joubert BR, Wu MC, Olshan AF, Haberg SE, Ueland PM, et al. Neonatal genome-wide methylation patterns in relation to birth weight in the Norwegian Mother and Child Cohort. *Am J Epidemiol*. 2014;179:834-842.
57. Adkins RM, Thomas F, Tylavsky FA, Krushkal J. Parental ages and levels of DNA methylation in the newborn are correlated. *BMC Med Genet*. 2011;12:47.
58. Markunas CA, Wilcox AJ, Xu Z, Joubert BR, Harlid S, Panduri V, et al. Maternal Age at Delivery Is Associated with an Epigenetic Signature in Both Newborns and Adults. *PLoS One*. 2016;11:e0156361.
59. Herbstman JB, Wang S, Perera FP, Lederman SA, Vishnevetsky J, Rundle AG, et al. Predictors and consequences of global DNA methylation in cord blood and at three years. *PLoS One*. 2013;8:e72824.
60. Sharp GC, Lawlor DA, Richmond RC, Fraser A, Simpkin A, Suderman M, et al. Maternal pre-pregnancy BMI and gestational weight gain, offspring DNA methylation and later offspring adiposity: findings from the Avon Longitudinal Study of Parents and Children. *Int J Epidemiol*. 2015;44:1288-1304.
61. Simpkin AJ, Suderman M, Gaunt TR, Lyttleton O, McArdle WL, Ring SM, et al. Longitudinal analysis of DNA methylation associated with birth weight and gestational age. *Hum Mol Genet*. 2015;24:3752-3763.
62. Raisanen S, Gissler M, Kramer MR, Heinonen S. Influence of delivery characteristics and socioeconomic status on giving birth by caesarean section - a cross sectional study during 2000-2010 in Finland. *BMC Pregnancy Childbirth*. 2014;14:120.

63. Elshibly EM, Schmalisch G. The effect of maternal anthropometric characteristics and social factors on gestational age and birth weight in Sudanese newborn infants. *BMC Public Health*. 2008;8:244.
64. Black RE, Allen LH, Bhutta ZA, Caulfield LE, de Onis M, Ezzati M, et al. Maternal and child undernutrition: global and regional exposures and health consequences. *Lancet*. 2008;371:243-260.
65. Ng SK, Cameron CM, Hills AP, McClure RJ, Scuffham PA. Socioeconomic disparities in prepregnancy BMI and impact on maternal and neonatal outcomes and postpartum weight retention: the EFHL longitudinal birth cohort study. *BMC Pregnancy Childbirth*. 2014;14:314.
66. Walker SP, Wachs TD, Gardner JM, Lozoff B, Wasserman GA, Pollitt E, et al. Child development: risk factors for adverse outcomes in developing countries. *Lancet*. 2007;369:145-157.
67. Zetterstrom R. Breastfeeding and infant-mother interaction. *Acta Paediatr Suppl*. 1999;88:1-6.
68. Fergusson DM, Woodward LJ. Breast feeding and later psychosocial adjustment. *Paediatr Perinat Epidemiol*. 1999;13:144-157.
69. Byun HM, Siegmund KD, Pan F, Weisenberger DJ, Kanel G, Laird PW, et al. Epigenetic profiling of somatic tissues from human autopsy specimens identifies tissue- and individual-specific DNA methylation patterns. *Hum Mol Genet*. 2009;18:4808-4817.
70. Ziller MJ, Gu H, Muller F, Donaghey J, Tsai LT, Kohlbacher O, et al. Charting a dynamic DNA methylation landscape of the human genome. *Nature*. 2013;500:477-481.
71. Illingworth RS, Gruenewald-Schneider U, De Sousa D, Webb S, Merusi C, Kerr AR, et al. Inter-individual variability contrasts with regional homogeneity in the human brain DNA methylome. *Nucleic Acids Res*. 2015;43:732-744.

72. Relton CL, Davey Smith G. Epigenetic epidemiology of common complex disease: prospects for prediction, prevention, and treatment. *PLoS Med.* 2010;7:e1000356.
73. Heijmans BT, Mill J. Commentary: The seven plagues of epigenetic epidemiology. *Int J Epidemiol.* 2012;41:74-78.
74. Kramer MS, Aboud F, Mironova E, Vanilovich I, Platt RW, Matush L, et al. Breastfeeding and child cognitive development: new evidence from a large randomized trial. *Arch Gen Psychiatry.* 2008;65:578-584.
75. Brion MJ, Lawlor DA, Matijasevich A, Horta B, Anselmi L, Araujo CL, et al. What are the causal effects of breastfeeding on IQ, obesity and blood pressure? Evidence from comparing high-income with middle-income cohorts. *Int J Epidemiol.* 2011;40:670-680.
76. Horta BL, Loret de Mola C, Victora CG. Breastfeeding and intelligence: a systematic review and meta-analysis. *Acta Paediatr.* 2015;104:14-19.
77. Davies MN, Volta M, Pidsley R, Lunnon K, Dixit A, Lovestone S, et al. Functional annotation of the human brain methylome identifies tissue-specific epigenetic variation across brain and blood. *Genome Biol.* 2012;13:R43.
78. Walton E, Hass J, Liu J, Roffman JL, Bernardoni F, Roessner V, et al. Correspondence of DNA Methylation Between Blood and Brain Tissue and Its Application to Schizophrenia Research. *Schizophr Bull.* 2016;42:406-414.
79. Hannon E, Lunnon K, Schalkwyk L, Mill J. Interindividual methylomic variation across blood, cortex, and cerebellum: implications for epigenetic studies of neurological and neuropsychiatric phenotypes. *Epigenetics.* 2015;10:1024-1032.

Supporting information

S1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2009 checklist.

Section/topic	#	Checklist item	Section reported
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	Title
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	Abstract
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	Introduction
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	Introduction (3 rd parag.)
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	Study selection and data collection
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	Search strategy
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	Search strategy
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	Study selection and data collection
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	Study selection and data collection
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	Study selection and data collection
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	

Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	Results (2 nd parag.); Figure 1
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	Human studies; Animal studies; Table 1
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	Human studies; Animal studies; Discussion
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	Human studies; Animal studies; Table 1
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	Discussion (1 st parag.)
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	Discussion (1 st -3 rd parag.)
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	Discussion (1 st , 4 th -9 th parag.)
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit: www.prisma-statement.org.

S2 Appendix. Pilot literature search.

Methods

Using an Ovid filter to remove non-original publications, a pilot search was performed using the same search strategy described in the main text in August 27, 2015. We also wanted to evaluate if limiting the search to publications in English would be too restrictive.

4724 records were initially obtained. After removing duplicates in Ovid, 3876 remained (set 1). 2563 remained after removing non-original publications using the Ovid filter (set 2), and 2543 remained after further limiting to publications in English using another Ovid filter (set 3). Not all Ovid databases allow duplicate removal, so potential residual duplicates (classified as such if title and authors' names were the same) were manually removed. This reduced the number of records to 3806 (set 1), 2543 (set 2) and 2523 (set 3).

The 1263 publications present in set 2, but not in set 1, were classified as "supposedly non-original". They were distributed (according to the database from which they were retrieved) as follows: 1168 in Journals@OVID, 92 in OVID fulltext Journals@Bristol, and 3 in PsycARTICLES Full Text. 113 supposedly non-original publications (100 randomly sampled from Journals@OVID; 10 randomly sampled from OVID fulltext Journals@Bristol; and all PsycARTICLES Full Text) were analyzed in detail. All 20 publications contained in set 3, but not in set 2, were selected to evaluate the consequences of limiting the search to publications in English only.

Results

The distribution of the 113 supposedly (ie, according to Ovid filter) non-original publications (sampled from a total of 1263 studies) according to Ovid classification was: 70 reviews, 21 miscellaneous, 9 editorials, 8 reports, 4 letters and 1 abstract. Of the 21 Ovid-classified miscellaneous publications, 10 were reviews, 5 were original journal article, 1 was an abstract with no original data, 1 was a commentary and 1 was a case report. The remaining 3 were impossible to classify. The only information

available for them was their titles: (i) What's in breast milk? A new screening method helps find out (likely a review or a commentary); (ii) IN THIS ISSUE (likely an editorial); (iii) American Journal of Clinical Nutrition: VOL. 70, NO. 4, OCTOBER 1999 (not even a title; possibly an editorial). Although none of the 5 original journal articles were related to the topic of the present review, the fact that there were original publications excluded because they were classified as “miscellaneous” allows the possibility that at least a few relevant studies (not included in this sample) would be excluded by the Ovid filter. Of the 177 miscellaneous publications in the entire list of supposedly non-original publications, 42 (assuming a proportion of 5/21) would be expected to be original publications. Therefore, the main search included miscellaneous publications.

Of the 8 publications classified as reports by Ovid, with 3 being review-like articles, 3 were case reports, 1 was a collection of abstracts (none of them relevant to the topic of the present review) and 1 was an original journal article. Of the 176 publications classified as reports in the entire list of supposedly non-original publications, 22 (assuming a proportion of 1/8) would be expected to be original publications. Therefore, the main search included this publication type.

Of the 4 publications classified as letters by Ovid, 3 presented new data (although none of them was related to the topic of the present review) and 1 of them was a letter to the editor. Of the 29 letters in the entire list of non-original papers, 22 of these (assuming a proportion of 3/4) would be expected to present new data. Therefore, letters were included in the main search.

Regarding language, of the 20 publications in non-English languages according to Ovid, 4 were in Polish, 4 in Hungarian, 4 in French, 2 in Japanese, 1 in Chinese, 1 in German, 1 in Swedish, 1 in Italian, 1 in Spanish and 1 in English (evidencing some lack of specificity in this filter). 9 of them provided new data, but none were relevant to the present review. Therefore, limiting the search to English is not expected to substantially influence the findings from the present systematic review, although it might be important to look for English papers within papers classified as non-English by Ovid. Nevertheless, since the number of publications in languages other than English was small, we opted by not applying a language filter in principle.

S3 Appendix. Sample size calculations.

Based on the findings by Obermann-Borst and colleagues [1], we performed calculations to estimate the sample size requirements to detect DNA methylation patterns associated with breastfeeding in epigenome-wide association studies.

For simplicity, breastfeeding was treated as a binary variable (ever=1; never=0), with prevalence of ever breastfeeding $p \in \{0.8, 0.9\}$. The standard deviation of the DNA methylation outcome variable s in the promoter region of the *LEP* gene was $0.3/120 \cong 3.3$ [1]. Therefore, we evaluated the following values of s : 1.65, 3.3 and 4.95. The absolute mean change in *LEP* promoter methylation comparing a category with the immediately smaller category was 0.7 percentage points. Given that breastfeeding was treated as a binary variable in our calculations, using 0.7 as the mean difference in DNA methylation comparing ever with never breastfed individuals (denoted by β) would likely be an underestimation (given that an ever vs. never comparison is much more drastic than a comparison between categories of duration), we used it as the smallest value to be evaluated in the calculations, so that $\beta \in \{0.7, 1.4 \text{ and } 2.1\}$.

Using the Bonferroni correction would yield a statistical significance threshold (alpha level) of $0.05/480,000 \cong 1.4 \times 10^{-4}$. However, such alpha level is known to be over conservative because it does not account for the correlation between CpG sites. In the study by Richmond et al. [2], the false discovery rate cut-off of 0.05 corresponded to a P-value of approximately 2.0×10^{-6} , which was then used as the multiple testing-corrected alpha level in our calculations. Power was set to 90%.

References

1. Obermann-Borst SA, Eilers PH, Tobi EW, de Jong FH, Slagboom PE, Heijmans BT, et al. Duration of breastfeeding and gender are associated with methylation of the LEPTIN gene in very young children. *Pediatr Res*. 2013;74:344-349.
2. Richmond RC, Simpkin AJ, Woodward G, Gaunt TR, Lyttleton O, McArdle WL, et al. Prenatal exposure to maternal smoking and offspring DNA methylation across the lifecourse:

findings from the Avon Longitudinal Study of Parents and Children (ALSPAC). *Hum Mol Genet.* 2015;24:2201-2217.

S1 Table. List of records identified after screening based on titles and abstracts in ascending publication order.

Authors	Publication year	Title	Publication type	Included (reason for exclusion, if applicable)
Editors of Journal of Human Lactation.	2007	Abstracts of presentations at the 13th International Conference of the International Society for Research in Human Milk and Lactation (ISRHML).	Conference proceedings	No (does not address epigenetic effects of breastfeeding)
Gillman MW, Barker D, Bier D, Cagampang F, Challis J, Fall C, Godfrey K, Gluckman P, Hanson M, Kuh D, Nathanielsz P, Nestel P, Thornburg KL.	2007	Meeting Report on the 3rd International Congress on Developmental Origins of Health and Disease (DOHaD).	Conference summary	No (does not address epigenetic effects of breastfeeding)
Ledo A, Arduini A, Asensi MA, Sastre J, Escrig R, Brugada M, Aguar M, Saenz P, Vento M.	2009	Human milk enhances antioxidant defenses against hydroxyl radical aggression in preterm infants1-3.	Original	No (does not address epigenetic effects of breastfeeding)
Palou A, Sanchez J, Pico C.	2009	Nutrient-gene interactions in early life programming: Leptin in breast milk prevents obesity later on in life.	Book chapter	No (does not address epigenetic effects of breastfeeding)
Waterland RA, Kellermayer R, Rached MT, Tatevian N, Gomes MV, Zhang J, Zhang L, Chakravarty A, Zhu W, Laritsky E, Zhang W, Wang X, Shen L.	2009	Epigenomic profiling indicates a role for DNA methylation in early postnatal liver development.	Original	No (does not address epigenetic effects of breastfeeding)
Burdge GC, Lillycrop KA.	2010	Nutrition, epigenetics, and developmental plasticity: implications for understanding human disease.	Review	No (does not address epigenetic effects of breastfeeding)
Chmurzynska A.	2010	Fetal programming: link between early nutrition, DNA methylation, and complex diseases.	Review	No (does not address epigenetic effects of breastfeeding)

Ho SM.	2010	Environmental epigenetics of asthma: An update.	Review	No (does not address epigenetic effects of breastfeeding)
Kappeler L, Meaney MJ.	2010	Epigenetics and parental effects.	Perspective	No (does not address epigenetic effects of breastfeeding)
Mehler MF.	2010	Epigenetics and neuropsychiatric diseases: introduction and meeting summary.	Conference summary	No (does not address epigenetic effects of breastfeeding)
Kuzawa CW, Thayer ZM.	2011	Timescales of human adaptation: the role of epigenetic processes.	Review	No (does not address epigenetic effects of breastfeeding)
Lester BM, Tronick E, Nestler E, Abel T, Kosofsky B, Kuzawa CW, Marsit CJ, Maze I, Meaney MJ, Monteggia LM, Reul JM, Skuse DH, Sweatt DJ, Wood MA.	2011	Behavioral epigenetics.	Review	No (does not address epigenetic effects of breastfeeding)
Palou M, Pico C, McKay JA, Sanchez J, Priego T, Mathers JC, Palou A.	2011	Protective effects of leptin during the suckling period against later obesity may be associated with changes in promoter methylation of the hypothalamic pro-opiomelanocortin gene.	Original	No (does not address epigenetic effects of breastfeeding)
Anto JM, Pinart M, Akdis M, Auffray C, Bachert C, Basagana X, Carlsen KH, Guerra S, von Hertzen L, Illi S, Kauffmann F, Keil T, Kiley J, Koppelman G, Lupinek C, Martinez F, Nawijn M, Postma D, Siroux V, Smit H, Sterk P, Sunyer J, Valenta R, Valverde S, Akdis CA, Annesi-Maesano I, Ballester F, Benet M, Cambon-Thomsen A, Chatzi, L, Coquet J, Demoly P, Gan W, Garcia-Aymerich J, Gimeno-Santos EPT, Guihenneuc-Jouyaux C, Haahtela T, Heinrich J, Herr MP, Hohmann CDP, Jacquemin B, Just J, Kerkhof M, Kogevinas M, Kowalski ML, Lambrecht BN, Lau S,	2012	Understanding the complexity of IgE-related phenotypes from childhood to young adulthood: A Mechanisms of the Development of Allergy (MeDALL) Seminar.	Conference summary	No (does not address epigenetic effects of breastfeeding)

Lodrup Carlsen KC, Maier D, Momas I, Noel P, Oddie S, Palkonen S, Pin I, Porta D, Punturieri A, Ranciere FP, Smith RA, Stanic B, Stein RT, van de Veen W, van Oosterhout AJM, Varraso R, Wickman M, Wijmenga C, Wright J, Yaman G, Zuberbier T, Bousquet J, WHO Collaborating Centre on Asthma and Rhinitis (Montpellier).

Godfrey K.	2012	Perinatal nutrition, epigenetics & later metabolic risk.	Conference abstract	No (does not address epigenetic effects of breastfeeding)
Hartman C, Shamir R.	2012	Nutrition and growth: highlights from the first international meeting.	Conference summary	No (does not address epigenetic effects of breastfeeding)
Kasten CH.	2012	The National Children's Study. Abstracts of the National Children's Study Research Day 2011.	Conference proceedings	No (does not address epigenetic effects of breastfeeding)
Qin W, Zhang K, Kliethermes B, Ruhlen RL, Browne EP, Arcaro KF, Sauter ER.	2012	Differential expression of cancer associated proteins in breast milk based on age at first full term pregnancy.	Original	No (does not address epigenetic effects of breastfeeding)
Tao M, Marian C, Shields P, Postischman N, Nie J, Ambrosone C, Edge S, Krishnan S, Vito D, Trevisan M, Freudenheim J.	2012	Early life exposures and promoter methylation in breast cancer: the Western New York Exposures and Breast Cancer (WEB) Study.	Conference abstract	No (published as a full-text original article also identified in our literature search - Tao et al., 2013)
Baumgartel KL, Conley YP.	2013	The utility of breastmilk for genetic or genomic studies: a systematic review.	Systematic review	No (does not address epigenetic effects of breastfeeding)
Grove-White D, Curtis G, Argo C.	2013	Feeding the dairy calf up till weaning - is it time to re-think?	Conference presentation	No (does not address epigenetic effects of breastfeeding)
Jaquiere AL, Phua HH, Park SS, Berry MJ, Bloomfield FH.	2013	Brief nutritional supplementation of term lambs results in epigenetic modification of pancreatic genes regulating insulin secretion.	Conference abstract	No (does not address epigenetic effects of breastfeeding)
Mahmood S, Smiraglia DJ, Srinivasan M,	2013	Epigenetic changes in hypothalamic	Original	Yes ^a

Patel MS.		appetite regulatory genes may underlie the developmental programming for obesity in rat neonates subjected to a high-carbohydrate dietary modification.		
Mischke M, Plosch T.	2013	More than just a gut instinct-the potential interplay between a babys nutrition, its gut microbiome, and the epigenome.	Perspective	Yes ^b
Nauta AJ, Ben Amor K, Knol J, Garssen J, van der Beek EM.	2013	Relevance of pre- and postnatal nutrition to development and interplay between the microbiota and metabolic and immune systems.	Review	Yes ^b
Obermann-Borst SA, Eilers PH, Tobi EW, de Jong FH, Slagboom PE, Heijmans BT, Steegers-Theunissen RP.	2013	Duration of breastfeeding and gender are associated with methylation of the LEPTIN gene in very young children.	Original	Yes ^a
Rossnerova A, Tulupova E, Tabashidze N, Schmuczerova J, Dostal M, Rossner P Jr, Gmuender H, Sram RJ.	2013	Factors affecting the 27K DNA methylation pattern in asthmatic and healthy children from locations with various environments.	Original	Yes ^a
Soto-Ramirez N, Karmaus W, Ziyab A, Lockett GA, Arshad S, Holloway JW, Zhang H, Ewart S.	2013	The interaction of breastfeeding, DNA methylation, and genetic variants in chromosome 17q12 and the risk of asthma in girls at age 18 years.	Conference abstract	Yes ^a
Tao MH, Marian C, Shields PG, Potischman N, Nie J, Krishnan SS, Berry DL, Kallakury BV, Ambrosone C, Edge SB, Trevisan M, Winston J, Freudenheim JL.	2013	Exposures in early life: associations with DNA promoter methylation in breast tumors.	Original	Yes ^a
Yong SB, Wu CC, Wang L, Yang KD.	2013	Influence and mechanisms of maternal and infant diets on the development of childhood asthma.	Review	No (does not address epigenetic effects of breastfeeding)
Daniel ZC, Akyol A, McMullen S, Langley-Evans SC.	2014	Exposure of neonatal rats to maternal cafeteria feeding during suckling alters hepatic gene expression and DNA	Original	No (does not address epigenetic effects of breastfeeding)

		methylation in the insulin signalling pathway.		
Gao F, Zhang J, Jiang P, Gong D, Wang JW, Xia Y, Ostergaard MV, Wang J, Sangild PT.	2014	Marked methylation changes in intestinal genes during the perinatal period of preterm neonates.	Original	No (does not address epigenetic effects of breastfeeding)
McInerny TK.	2014	Breastfeeding, early brain development, and epigenetics--getting children off to their best start.	Commentary	Yes ^b
Raychaudhuri N, Thamotharan S, Srinivasan M, Mahmood S, Patel MS, Devaskar SU.	2014	Postnatal exposure to a high-carbohydrate diet interferes epigenetically with thyroid hormone receptor induction of the adult male rat skeletal muscle glucose transporter isoform 4 expression.	Original	Yes ^a
Shafai T, Mustafa M, Hild T, Mulari J, Curtis A.	2014	The association of early weaning and formula feeding with autism spectrum disorders.	Letter to the editor	No (does not address epigenetic effects of breastfeeding)
Singhal A.	2014	The global epidemic of noncommunicable disease: the role of early-life factors.	Workshop proceedings	No (does not address epigenetic effects of breastfeeding)
Tow J.	2014	Heal the mother, heal the baby: epigenetics, breastfeeding and the human microbiome.	Commentary	Yes ^b
UK Molecular Epidemiology Group.	2014	Abstracts of the UK Molecular Epidemiology Group (MEG) Winter Meeting on The Future of Epidemiology: Biomarkers meet Populations. Newcastle University, United Kingdom. December 6, 2013.	Conference proceedings	No (does not address epigenetic effects of breastfeeding)
Verduci E, Banderali G, Barberi S, Radaelli G, Lops A, Betti F, Riva E, Giovannini M.	2014	Epigenetic effects of human breast milk.	Review	Yes ^b
Wu AM, Yang M, Dalvi P, Turinsky AL, Wang W, Butcher D, Egan SE, Weksberg R,	2014	Role of STAT5 and epigenetics in lactation-associated upregulation of multidrug	Original	No (does not address epigenetic effects of breastfeeding)

Harper PA, Ito S.		transporter ABCG2 in the mammary gland.		
Alsaweed M, Hartmann PE, Geddes DT, Kakulas F.	2015	MicroRNAs in Breastmilk and the Lactating Breast: Potential Immunoprotectors and Developmental Regulators for the Infant and the Mother.	Original	No (does not address effects of breastfeeding on DNA methylation)
Langley-Evans SC.	2015	Nutrition in early life and the programming of adult disease: a review.	Review	No (does not address epigenetic effects of breastfeeding)
Lukyanova OL, Borovik TE.	2015	Nutritional epigenetics and epigenetic effects of human breast milk.	Review	Yes ^b
Montirosso R.	2015	XI. Relationship Between Feeding and Early Stress in Premature Infant: The Role of Epigenetic Factors.	Conference paper	No (does not address epigenetic effects of breastfeeding)
Remely M, Stefanska B, Lovrecic L, Magnet U, Haslberger AG.	2015	Nutrieigenomics: the role of nutrition in epigenetic control of human diseases.	Review	No (does not address epigenetic effects of breastfeeding)
Godfrey KM, Costello PM, Lillycrop KA.	2016	Development, Epigenetics and Metabolic Programming.	Workshop proceedings	No (does not address epigenetic effects of breastfeeding)
Moisá SJ, Shike DW, Shoup L, Loo JJ.	2016	Maternal Plane of Nutrition During Late-Gestation and Weaning Age Alter Steer Calf Longissimus Muscle Adipogenic MicroRNA and Target Gene Expression.	Original	No (does not address effects of breastfeeding on DNA methylation)
Simpkin AJ, Hemani G, Suderman M, Gaunt TR, Lyttleton O, Mcardle WL, Ring SM, Sharp GC, Tilling K, Horvath S, Kunze S, Peters A, Waldenberger M, Ward-Caviness C, Nohr EA, Sørensen TI, Relton CL, Davey Smith G.	2016	Prenatal and early life influences on epigenetic age in children: a study of mother-offspring pairs from two cohort studies.	Original	Yes ^a

^aEligible for inclusion in the systematic review.

^bSelected for searching for additional references.

S2 Table. Sample size requirements to detect DNA methylation differences according to breastfeeding (ever vs. never) in an epigenome-wide association study (power=90%; $\alpha=2\times 10^{-6}$).

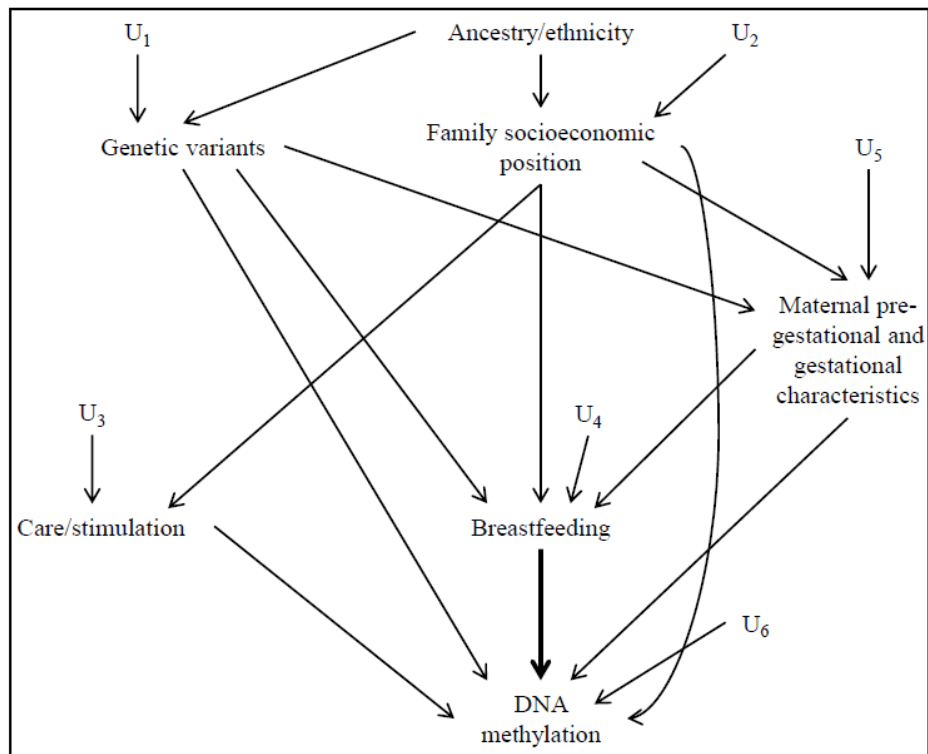
<i>p</i>	<i>s</i>	<i>β</i>		
		0.7	1.4	2.1
0.8	1.65	1265	317	142
0.8	3.3	5060	1265	563
0.8	4.95	11384	2847	1265
0.9	1.65	2249	563	250
0.9	3.3	8995	2249	1000
0.9	4.95	20237	5060	2249

p: Prevalence of ever breastfeeding.

s: Standard deviation of the outcome variable.

β: Mean absolute difference (in percentage points) in DNA methylation between the two breastfeeding groups.

S1 Fig. Directed acyclic graph depicting postulated causal relationships among breastfeeding, DNA methylation and potential important confounders.



U_N represents an unknown variable.

The thicker line indicates the target causal relationship.

4 – Artigo original 1

Association between breastfeeding and DNA methylation over the life course: findings from the Avon Longitudinal Study of Parents and Children (ALSPAC)

Fernando Pires Hartwig^{1,2*}, George Davey Smith², Andrew Simpkin^{2,3}, Cesar Gomes Victora¹, Caroline L. Relton², Doretta Caramaschi²

¹Postgraduate Programme in Epidemiology, Federal University of Pelotas, Pelotas, Brazil. ²MRC Integrative Epidemiology Unit, University of Bristol, Population Health Science, Bristol Medical School, Bristol, United Kingdom. ³Insight Centre for Data Analytics, National University of Ireland, Galway, Ireland.

*Corresponding author. Postgraduate Program in Epidemiology, Federal University of Pelotas, Pelotas (Brazil). Zip code: 96020-220. Phone: 55 53 981068670. E-mail: fernandophartwig@gmail.com; fh15144@bristol.ac.uk.

Abstract

Breastfeeding is associated with short and long-term health benefits. Long-term effects might be mediated by epigenetic mechanisms, yet a recent systematic review indicated that the literature on this topic is scarce. We performed the first epigenome-wide association study of breastfeeding, using peripheral blood DNA methylation data in childhood (age 7) and adolescence (age 15-17) from the Accessible Resource for Integrated Epigenomic Studies (ARIES) project within the Avon Longitudinal Study of Parents and Children (ALSPAC) cohort. We also analysed cord blood DNA methylation as a negative control. We found stronger associations when treating breastfeeding as a binary (ever vs. never) variable compared to other categorisations. Two methylation sites presented directionally-consistent associations with breastfeeding at ages 7 and 15-17, but not at birth. 12 differentially-methylated regions in relation to breastfeeding were identified, and for three of them there was evidence of directional concordance between ages 7 and 15-17, but not between birth and age 7. Our findings indicate that DNA methylation may play a role in mediating long-term associations between breastfeeding and health outcomes, but further studies with large enough samples for replication are required to identify robust associations.

Keywords: Breastfeeding; Life-course; DNA methylation; Epigenome-wide association study.

Introduction

Breastfeeding has clear short-term health benefits, particularly in reducing the risk of infections in childhood. Accumulating evidence indicates that breastfeeding may also have long-term effects on health outcomes and human capital, as well as benefit maternal health¹. For example, being breastfed has been associated with better performance in intelligence quotient (IQ) tests in a meta-analysis based on a systematic literature review², in population-based birth cohorts with different confounding structures³, and in the single randomized controlled trial on this subject⁴.

The mechanisms underlying the long-term effects of breastfeeding are not fully understood. Such mechanisms clearly must persist over time after weaning – in other words, become “imprinted” in the organism.⁵ In the case of other early-life exposures such as maternal smoking during pregnancy, there is evidence of long-term associations with offspring DNA methylation⁶ – i.e., addition of a methyl (–CH₃) group to DNA at the 5' position of a cytosine base, typically in cytosine-guanine (CpG) dinucleotides located in DNA sequences called CpG islands, which are rich in CpG dinucleotides^{7,8}. DNA methylation is one type of a broader class of biological processes known as epigenetics, which encompasses mitotically heritable events – other than changes in the DNA sequence itself – involved in gene expression regulation. Epigenetic processes play a key role in developmental processes^{9,10}, and have more recently been linked to disease processes¹¹⁻¹⁴.

Some evidence suggests that breastfeeding might influence DNA methylation through epigenetic effects of some of its nutritional components¹⁵ or through the microbiome, which is shaped by early feeding habits¹⁶. However, according to a recent systematic literature review¹⁷, the overall evidence on the epigenetic effects of breastfeeding is scarce. Our aim was to perform a genome-wide assessment of the association between breastfeeding and DNA methylation in childhood, characterise – if present – the pattern of this association and investigate whether it persists until adolescence in a population-based study in England.

Results

Description of study participants

Supplementary Table 1 displays the characteristics of the study participants. There were 702 (birth), 640 (age 7) and 709 (age 15-17) individuals with non-missing information for all study variables (corresponding to approximately 70% of all ARIES participants). In general, the subset included in our analysis was similar to the entire ARIES dataset. The largest differences were observed for maternal education at birth (with the mothers of included individuals having slightly higher educational attainment) and ethnicity (with the proportion of individuals of European ethnicity being slightly higher in the included individuals). Previous analysis indicated that ARIES is reasonably representative of the entire ALSPAC cohort.¹⁸

Association of breastfeeding with single CpG sites

Figures 1 and 2 provide an overall view of the EWAS results. There was no strong indication of genome-wide inflation for breastfeeding analysed in duration categories, assuming a linear trend (genomic inflation factor of 0.97), but there was some indication for the “ever breastfeeding” variable (genomic inflation factor of 1.10). Importantly, the bulk of the distribution closely resembled the expected under the null, with the deviation occurring in the right tail of the distribution of P-values. This may be due to breastfeeding having small effects on DNA methylation (in which case detection would require larger samples) in many regions of the genome, rather than due to the presence of systematic bias in the results.

Regarding ever breastfeeding, no CpGs achieved the conventional significance threshold of $FDR < 0.05$ (which approximately corresponds to a P-value of 1.0×10^{-7}) in the minimally-adjusted model, although a few ones achieved a $FDR < 0.20$ (which approximately corresponds to a P-value of 1.0×10^{-6}). In the fully-adjusted model (Table 1), one CpG (cg11414913) achieved a $FDR < 0.05$, and there was suggestive evidence of association for six additional ones (cg00234095, cg04722177, cg03945777, cg17052885, cg05800082 and cg24134845; see Supplementary Table 2 for a description of those CpGs). The results for breastfeeding coded as a categorical variable in duration categories (assuming a linear trend) were remarkably null, with no CpGs achieving even suggestive levels of association. This suggests that, if

breastfeeding is associated with peripheral blood DNA methylation, the association depends more on whether or not the individual was ever breastfed than breastfeeding duration.

Table 1 shows that methylation in the cg11414913 CpG was 3.19 percent points lower ($P=5.2\times10^{-8}$) in ever breastfed children. There was also suggestive evidence for association lower methylation in the cg00234095 ($\beta=-1.74$; $P=4.9\times10^{-7}$), cg04722177 ($\beta=-2.90$; 2.7×10^{-6}), and cg03945777 ($\beta=-0.84$; $P=3.2\times10^{-6}$) sites, and for higher methylation in the cg17052885 ($\beta=1.79$; $P=4.9\times10^{-6}$), cg05800082 ($\beta=1.05$; $P=5.8\times10^{-6}$), and cg24134845 ($\beta=0.23$; $P=3.3\times10^{-5}$) site. The evidence of an association virtually disappeared when breastfeeding was analysed continuously, and the regression coefficients were generally similar among different categories of breastfeeding duration. Those results indicate that the association between breastfeeding and peripheral blood DNA methylation does not follow a dose-response relationship, but presents a threshold (ever vs. never) pattern.

Table 2 displays the association between ever breastfeeding and peripheral blood methylation at different ages in the CpGs identified in the EWAS. The cg11414913 CpG presented a persistent, directionally-consistent association with breastfeeding at the age of 15-17 years ($\beta=-2.77$; $P=0.004$), and no evidence of association at birth ($\beta=-0.44$; $P=0.631$). The cg05800082 CpG presented a similar pattern, although the point estimate was attenuated compared to age 7 years, and presented rather weak statistical evidence of association at the age of 15-17 years ($\beta=0.56$; $P=0.083$). However, it was reassuring that its point estimate at birth ($\beta=-0.53$; $P=0.144$) was directionally inconsistent with the results at later ages. The CpGs cg00234095, cg03945777 and cg24134845 presented evidence of association only at age 7, suggesting that their association with breastfeeding does not persist until the ages of 15-17. DNA methylation at birth in the two remaining CpGs was associated with breastfeeding in the same direction as the association at the age of 7, suggesting that those associations are substantially influenced by some unaccounted bias source (e.g., unmeasured confounders).

Association between breastfeeding and methylation regions

Given that quantile-quantile plots were suggestive of small effects of breastfeeding on DNA methylation in many regions of the genome, we complemented the ever breastfeeding EWAS with a search for differentially methylated regions (DMRs) – i.e., two or more CpGs enriched for low P-values of the association with breastfeeding (see the Methods for details). 12 DMRs were identified (Table 3 and Supplementary Table 3). There was no strong indication that the association of breastfeeding with different CpGs in the same DMR was generally directionally consistent (Table 3). However, regarding directional concordance for each CpG across time points, four DMRs presented evidence of concordance between 7 and 15-17 years, but not between methylation and birth and at age 7: 18:106178-106850, 9:91296-92146, 22:255590-256045, and 8:409905-410098 (Table 4). For two DMRs (5:97867-98797 and 1:425524-426297), there was evidence for directional concordance between birth and 7 years of age, suggesting that the associations between breastfeeding and methylation at age 7 in the CpGs in those DMRs may be distorted by pre-natal confounders. For the remaining CpGs, there was no evidence for directional concordance between the any of the two comparisons, suggesting that the association between breastfeeding and methylation at age 7 in the CpGs in those DMRs may be transient or false-positives. A sensitivity analysis considering only the CpGs that achieved $P < 0.05$ in at least one time point corroborated the strongest directional consistency between 7 and 15-17 years observed for the four aforementioned DMRs, except the 8:409905-410098; importantly, this analysis involved only 3 CpGs for this DMR (Supplementary Table 4). Moreover, a fifth DMR – 19:365914-366989 – was identified in this analysis, suggesting that CpGs with weak associations could have diluted the association in the analysis considering all CpGs in the DMR.

Discussion

In this breastfeeding EWAS, ever breastfeeding was associated with peripheral blood methylation in the cg11414913 CpG at ages 7 and 15-17 years, but not at birth. There was suggestive evidence of association between ever breastfeeding and age 7 methylation in six additional CpGs, with one – the cg05800082 CpG – also presenting a directionally consistent (although attenuated) point estimate at age 15, but not at

birth. Moreover, 12 DMRs were identified, and three of them presented evidence of directional concordance between ages 7 and 15-17, but not between birth and age 7, in all sensitivity analyses. Our quantile-quantile plots indicated that the associational effect estimates between ever breastfeeding and peripheral blood DNA methylation are generally small. None of our analyses supported a dose-response relationship between breastfeeding and peripheral blood DNA methylation, but were consistent with an effect that depends on whether or not the child was ever breastfed.

The CpG cg11414913, which presented the most robust statistical evidence of association with breastfeeding, is located in an intergenic region, with the nearest gene being the *TTC34* gene. This gene is overexpressed in the testis, but largely unknown regarding its biological roles, although there is some indication of a relation with multiple sclerosis and lung cancer. The region around this CpG is highly conserved among vertebrates, and contains a 249 bp region (which includes the CpG) that presents DNase I hypersensitivity (which is related to more transcriptional activity) in six cell/tissue types, including lung carcinoma, prostate adenocarcinoma and pancreatic islets. The CpG cg05800082, which presented some evidence of persistent association with breastfeeding, is located within the *DST* gene, which is expressed in many tissues, including skin and brain. This gene encodes isoforms of cytoskeletal linker proteins that present tissue-specificity regarding expression and function: while some isoforms expressed in epithelial tissues anchor keratin-containing intermediate filaments to hemidesmosomes, other isoforms – mainly expressed in neural and muscle tissue – anchor neural intermediate filaments to the actin cytoskeleton. Mutations in the *DST* gene have also been implicated in neuronal and skin disorders. Moreover, the region spanning this CpG presents DNase I hypersensitivity in 5 cell/tissue types and enrichment of the H3K27Ac histone mark, which is also related to enhanced transcription.

Regarding DMRs, the 18:106,178-106,850 region is located within the *DUX4* gene, which encodes a transcriptional activator of *PITX1*, and is linked to autosomal dominant facioscapulohumeral muscular dystrophy (FSHD). It is expressed in the testis, and in muscle tissues of FSHD patients. The 9:91,296-92,146 region is located 1,719 bp away from the *PGM5P3-AS1* gene, which encodes a non-coding RNA of unknown

function. The 22:255,590-25,6045 region did not present any obvious important biological feature in a 100,000 bp window centred at the DMR. Two additional DMRs presented weaker evidence of a persistent association with breastfeeding. One was the 8:409,905-410,098 region located in the *FBXO25* gene, which encodes a protein that is overexpressed in the testis and belong to the family of F-box proteins, which are components of a ubiquitin protein ligase complex. The second was the 19:365,914-366,989 region located in the *THEG* gene, which encodes a nuclear protein specifically in the nucleus of haploid male germ cells, with a possible role in spermatogenesis.

The epidemiological literature on breastfeeding and health focuses on well-established effects against infectious diseases, as well as on potential impact on intelligence, obesity and diabetes, among other outcomes¹. In the present analyses, none of the regions where methylation was detected seem to be involved in the above conditions. This may be due to analysing a surrogate tissue, limited statistical power to detect more CpGs, and limited knowledge about the health effects of the methylation sites that were detected. Moreover, the effects of breastfeeding on health and development may be mediated through other epigenetic processes, such as non-coding RNAs^{19,20}, as well as a host of mechanisms other than epigenetics, including provision of nutrients (e.g., pre-formed long-chain polyunsaturated fatty acids, which are plausible mediators of the benefits on IQ²¹), antibodies and other immunoactive compounds, antimicrobials, and important effects on the gut microbiome¹.

One of the strengths of this study is that longitudinal measures of DNA methylation allowed not only identifying regions of the methylome associated with breastfeeding, but also assessing if those associations persist until adolescence. Dense phenotyping and genotyping of study participants allowed controlling for several covariates, which were selected using a conceptual model defined *a priori*. Moreover, DNA methylation data at birth was used to rule out associations likely driven by residual confounding due to pre-natal factors. However, residual confounding cannot be fully discarded, so triangulating our findings with those from future studies using designs prone to different potential sources of bias will be important to disentangle causality²².

In addition to the possibility of residual confounding, another important limitation of this study is that it was restricted to peripheral blood. As we discussed elsewhere¹⁷, DNA methylation in blood may not be a good proxy of DNA methylation in other tissues, such as the brain,²³⁻²⁵ thus limiting the capacity of any breastfeeding EWAS using peripheral blood to inform DNA methylation patterns in the target tissue^{11,26} – in this example, when assessing if the association between breastfeeding and IQ has an epigenetic component. This may also limit the capacity to identify true signals. However, epigenetic studies in surrogate tissues are important. These are frequently the only viable alternative in large epidemiological studies, also being able to provide useful information on the range of potential epigenetic effects of the exposure of interest, which may then guide future, specific studies such as *in vitro* studies in cells and *in vivo* studies in animal models¹⁷.

Another important limitation is that we did not perform a formal replication of our results. However, the fact that some hits (both in the CpG and DMR analysis) at age 7 years did not present evidence of association at age 15-17 years indicates that inflation of type-I error due to multiple-testing alone was not sufficient for a hit in one age to also present evidence of association in other ages. Therefore, CpGs and DMRs that presented evidence of persistent associations are less likely to be a sole product of multiple testing. However, this reasoning is less clear for transient associations, which could be truly transient effects or merely false-positives that do not carry over to adolescence. Although persistent associations are likely to be more robust from a methodological perspective in our study, this does not mean that transient effects are irrelevant. For example, the latter could trigger the actual processes that will lead to long-term effects (e.g., influences on brain development and IQ in adulthood). Moreover, in our context transient effects are defined as associations observed at the age of 7 years which did not persist until adolescence, but associations at age 7 may already be regarded as persistent effects of breastfeeding.

This study provided important insights into the shape and persistence of the association between breastfeeding and peripheral blood DNA methylation. Rather than providing definitive answers on their own, our results will serve to motivate future studies using different designs to improve causal inference, as well as

consortium-based efforts – examples of which are already available in the epigenetic epidemiology literature^{27,28} – to achieve sample sizes large enough to both improve power and allow replication. Such future efforts will complement and expand our findings by providing robust evidence on the potential effects of breastfeeding on DNA methylation, which may contribute to understand the biological basis of long-term associations between breastfeeding and health and human capital outcomes, and potentially also reveal new biological aspects of breastfeeding.

Methods

Study setting and participants

Study subjects were part of the Accessible Resource for Integrated Epigenomic Studies (ARIES)¹⁸, a representative sub-sample of the Avon Longitudinal Study of Parents and Children (ALSPAC) for which methylation data were collected. ALSPAC is a population-based, prospective birth cohort of women and their children²⁹⁻³¹. All pregnant women living in the geographical area of Avon (UK) with expected to delivery date between 1 April 1991 and 31 December 1992 were invited to participate. Approximately 85% of the eligible population was enrolled, totalling 14,541 pregnant women who gave informed and written consent. Information on the data collection and availability can be found at <http://www.bris.ac.uk/alspac/researchers/data-access/data-dictionary/>. Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees.

Our analysis was focused on the offspring born in 1991-1992. The analyses were restricted to singletons or only to one participant out of a twin pair, selected at random. Individuals with missing information for the exposure, outcome or covariates (described below) were excluded.

Study variables

DNA methylation

DNA methylation in white blood cells was measured in ARIES offspring at three time points: at birth (cord blood), and at 7 and 15-17 years of age (peripheral blood). DNA

samples underwent bisulphite conversion using the Zymo EZ DNA methylationTM kit (Zymo, Irvine, CA). The Illumina HumanMethylation450 BeadChip was used for genome-wide epigenotyping. The arrays were scanned using an Illumina iScan, and initial quality checks performed using GenomeStudio version 2011.1. We excluded single nucleotide polymorphisms, probes with a high detection P-value (ie, P-value > 0.05 in more than 5% samples) and sex chromosomes. Methylation data normalisation was carried out using the “Tost” algorithm to minimise non-biological between-probe differences³², as implemented in the “watermelon” R package³³. All processing steps used the “meffil” R package³⁴.

The outcome variables of this study were cord and peripheral blood (ages 7 and 15) DNA methylation levels in ~470,000 CpG sites. Methylation was analysed as beta values, which vary from 0 to 1 and indicate the proportion of cells methylated at a particular CpG³⁵. Regression coefficients and standard errors were multiplied by 100, so that they can be interpreted as percent point differences in DNA methylation.

Breastfeeding

Breastfeeding data was collected through questionnaires answered by the mothers when their offspring were (on average) four weeks, six months and 15 months old. These data were used to define four different breastfeeding categorisations:

- i) A binary indicator of whether the individual was ever breastfed (regardless of duration).
- ii) Breastfeeding duration groups, defined as follows: 0=never breastfed; 1=1 day to 3 months of duration; 2=3.01 to 6 months; 3=6.01 to 12 months; and 4=more than 12 months.
- iii) Same as ii), but coding each category as a number, thus assuming a linear trend.
- iv) Breastfeeding duration in months, as a continuous variable.

Covariates

Covariates were selected mostly based on a conceptual model that we defined previously¹⁷. The following covariates were used:

- i) Sociodemographic: an indicator of whether the participant had European ethnic background (informed by mothers at 32 weeks of gestation), and the top two ancestry-informative principal components estimated using genome-wide genotyping data³⁶.
- ii) Family socioeconomic position: to avoid collinearity issues, we used only the mother's highest educational qualification (informed by the mothers themselves at 32 weeks of gestation).
- iii) Maternal characteristics: parity (informed by the mothers at 18 weeks of gestation), height, pre-pregnancy weight (informed by the mothers themselves at 12 weeks of gestation), age at birth (calculated from mother's date of birth and date of delivery) and folic acid supplementation (informed by the mothers at 18 and 32 weeks of gestation).
- iv) Gestational characteristics: maternal smoking during pregnancy (informed by the mothers at 18 weeks of gestation), type of delivery (informed by the mothers when their offspring were eight weeks old), gemelarity, gestational age (calculated from the date of the mother's last menstrual period reported at enrolment; when the mother was uncertain of this or when it conflicted with clinical assessment, the ultrasound assessment was used; where maternal report and ultrasound assessment conflicted, an experienced obstetrician reviewed clinical records and provided an estimate) and birthweight (from obstetric data, measures from the ALSPAC team and notifications or clinical records).

Although not mentioned in the DAG, participant's sex and age at blood collection were also selected as covariates. Given that they are associated with DNA methylation but are not influenced by breastfeeding, adjusting for those two covariates may improve power by reducing variance in DNA methylation. We also adjusted for estimated cell counts using Bakulski's³⁷ (for cord blood) or Houseman's (for peripheral blood)³⁸ methods to account for methylation differences due to cell composition. Finally, a surrogate variable analysis was performed on the methylation data using the "sva" R package, and the surrogate variables not associated with breastfeeding were additionally included as covariates to adjust for batch effects³⁹.

Statistical analyses

We conducted an epigenome-wide association study (EWAS) of breastfeeding. The main EWAS analyses considered breastfeeding as the exposure in two categorisations: i) none vs. any; ii) duration categories, assuming a linear trend. The outcome was DNA methylation measured at ~470,000 CpG sites in peripheral blood at the age of 7 years. CpGs with suggestive evidence, here defined as achieving a $P\text{-value} < 5.0 \times 10^{-6}$, were then re-analysed to explore additional breastfeeding categorisations and to investigate whether the signal persisted until 15 years of age. Cord blood methylation was analysed as a negative control, under the assumption that at least some of possible pre-natal residual confounding would result in associations between breastfeeding and cord blood methylation. Two analysis models were performed: i) adjusting only for estimated cell composition and batch effects, and ii) adjusting for all covariates. These models are hereafter referred to as minimally-adjusted and fully-adjusted, respectively. All analyses were performed using heteroskedasticity-consistent standard errors, implemented using the “lmtest”, “MASS” and “sandwich” R packages.

The EWAS results were further used to identify DMRs in relation to breastfeeding. DMRs were identified using the Comb-P method, which tags regions enriched for low P-values while accounting for auto-correlation and multiple testing^{40,41}. Following the criteria used by Sharp et al.⁴², a region was classified as a DMR if: i) it contained at least two CpGs; ii) all CpGs in the region are within 1000 bp of at least another CpG in the same region; and iii) the auto-correlation and multiple-testing corrected (upon applying Stouffer-Liptak-Kechris and Sidak methods, respectively) P-value for the region was < 0.05 . The CpGs belonging to the identified DMRs analysed further to assess if breastfeeding had a consistent effect across the DMR (ie, if CpGs in the DMR generally presented greater or lower levels of methylation according to breastfeeding) using linear mixed models to account for the correlation between CpGs assuming that they are nested within individuals, implemented using the “nlme” R package. This was complemented by evaluating, for each DMR, the directional consistency of each CpG across time points using a sign test. Analyses were performed using R (<http://www.r-project.org/>).

Biological characterisation

The nearest gene to each CpG site was extracted from the annotation file provided by Illumina. We used the The UCSC Genome Browser (<https://genome.ucsc.edu/cgi-bin/hgGateway>; GRCh37/hg19 Assembly) when no genes were available, and to identify other biological features – focusing on DNase I hypersensitivity, presence of binding sites of transcription factors, and conservation among vertebrates – of the regions containing the identified CpGs and DMRs. Features of identified genes (and encoded proteins) were extracted from GeneCards®: The Human Gene Database (<http://www.genecards.org/>) and from GeneEntrez (<https://www.ncbi.nlm.nih.gov/gene>). Linked diseases were identified using the Online Mendelian Inheritance in Man (OMIM) database (<https://www.ncbi.nlm.nih.gov/omim/>). This characterisation is presented in the Discussion.

References

1. Victora, C. G. *et al.* Breastfeeding in the 21st century: epidemiology, mechanisms, and lifelong effect. *Lancet* **387**, 475-490 (2016).
2. Horta, B. L., Loret de Mola, C. & Victora, C. G. Breastfeeding and intelligence: a systematic review and meta-analysis. *Acta Paediatr* **104**, 14-19 (2015).
3. Brion, M. J. *et al.* What are the causal effects of breastfeeding on IQ, obesity and blood pressure? Evidence from comparing high-income with middle-income cohorts. *Int J Epidemiol* **40**, 670-680 (2011).
4. Kramer, M. S. *et al.* Breastfeeding and child cognitive development: new evidence from a large randomized trial. *Arch Gen Psychiatry* **65**, 578-584 (2008).
5. Relton, C. L., Hartwig, F. P. & Davey Smith, G. From stem cells to the law courts: DNA methylation, the forensic epigenome and the possibility of a biosocial archive. *Int J Epidemiol* **44**, 1083-1093 (2015).
6. Richmond, R. C. *et al.* Prenatal exposure to maternal smoking and offspring DNA methylation across the lifecourse: findings from the Avon Longitudinal Study of Parents and Children (ALSPAC). *Hum Mol Genet* **24**, 2201-2217 (2015).

7. Han, L., Su, B., Li, W. H. & Zhao, Z. CpG island density and its correlations with genomic features in mammalian genomes. *Genome Biol* **9**, R79 (2008).
8. Rakyan, V. K., Down, T. A., Balding, D. J. & Beck, S. Epigenome-wide association studies for common human diseases. *Nat Rev Genet* **12**, 529-541 (2011).
9. Kiefer, J. C. Epigenetics in development. *Dev Dyn* **236**, 1144-1156 (2007).
10. Huang, K. & Fan, G. DNA methylation in cell differentiation and reprogramming: an emerging systematic view. *Regen Med* **5**, 531-544 (2010).
11. Relton, C. L. & Davey Smith, G. Epigenetic epidemiology of common complex disease: prospects for prediction, prevention, and treatment. *PLoS Med* **7**, e1000356 (2010).
12. Tollefsbol, T. *Epigenetics in Human Disease*. (Academic Press, 2012).
13. Kaelin, W. G., Jr. & McKnight, S. L. Influence of metabolism on epigenetics and disease. *Cell* **153**, 56-69 (2013).
14. Tobi, E. W. *et al.* DNA methylation as a mediator of the association between prenatal adversity and risk factors for metabolic disease in adulthood. *Sci Adv* **4**, eaao4364 (2018).
15. Verduci, E. *et al.* Epigenetic effects of human breast milk. *Nutrients* **6**, 1711-1724 (2014).
16. Mischke, M. & Plosch, T. More than just a gut instinct-the potential interplay between a baby's nutrition, its gut microbiome, and the epigenome. *Am J Physiol Regul Integr Comp Physiol* **304**, R1065-1069 (2013).
17. Hartwig, F. P., Loret de Mola, C., Davies, N. M., Victora, C. G. & Relton, C. L. Breastfeeding effects on DNA methylation in the offspring: A systematic literature review. *PLoS One* **12**, e0173070 (2017).
18. Relton, C. L. *et al.* Data Resource Profile: Accessible Resource for Integrated Epigenomic Studies (ARIES). *Int J Epidemiol* **44**, 1181-1190 (2015).
19. Karlsson, O. *et al.* Detection of long non-coding RNAs in human breastmilk extracellular vesicles: Implications for early child development. *Epigenetics*, 0 (2016).
20. Alsaweed, M., Hartmann, P. E., Geddes, D. T. & Kakulas, F. MicroRNAs in Breastmilk and the Lactating Breast: Potential Immunoprotectors and Developmental Regulators for the Infant and the Mother. *Int J Environ Res Public Health* **12**, 13981-14020 (2015).

21. Innis, S. M. Dietary (n-3) fatty acids and brain development. *J Nutr* **137**, 855-859 (2007).
22. Lawlor, D. A., Tilling, K. & Davey Smith, G. Triangulation in aetiological epidemiology. *Int J Epidemiol* **45**, 1866-1886 (2016).
23. Davies, M. N. *et al.* Functional annotation of the human brain methylome identifies tissue-specific epigenetic variation across brain and blood. *Genome Biol* **13**, R43 (2012).
24. Walton, E. *et al.* Correspondence of DNA Methylation Between Blood and Brain Tissue and Its Application to Schizophrenia Research. *Schizophr Bull* **42**, 406-414 (2016).
25. Hannon, E., Lunnon, K., Schalkwyk, L. & Mill, J. Interindividual methylomic variation across blood, cortex, and cerebellum: implications for epigenetic studies of neurological and neuropsychiatric phenotypes. *Epigenetics* **10**, 1024-1032 (2015).
26. Heijmans, B. T. & Mill, J. Commentary: The seven plagues of epigenetic epidemiology. *Int J Epidemiol* **41**, 74-78 (2012).
27. Joubert, B. R. *et al.* DNA Methylation in Newborns and Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-analysis. *Am J Hum Genet* **98**, 680-696 (2016).
28. Gruzieva, O. *et al.* Epigenome-Wide Meta-Analysis of Methylation in Children Related to Prenatal NO₂ Air Pollution Exposure. *Environ Health Perspect* **125**, 104-110 (2017).
29. Golding, J., Pembrey, M. & Jones, R. ALSPAC--the Avon Longitudinal Study of Parents and Children. I. Study methodology. *Paediatr Perinat Epidemiol* **15**, 74-87 (2001).
30. Boyd, A. *et al.* Cohort Profile: the 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children. *Int J Epidemiol* **42**, 111-127 (2013).
31. Fraser, A. *et al.* Cohort Profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *Int J Epidemiol* **42**, 97-110 (2013).
32. Touleimat, N. & Tost, J. Complete pipeline for Infinium((R)) Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics* **4**, 325-341 (2012).
33. Pidsley, R. *et al.* A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics* **14**, 293 (2013).

34. Min, J., Hemani, G., Davey Smith, G., Relton, C. L. & Suderman, M. Meffil: efficient normalisation and analysis of very large DNA methylation samples. *bioRxiv*: 10.1101/125963 (2017).
35. Du, P. *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11**, 587 (2010).
36. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-909 (2006).
37. Bakulski, K. M. *et al.* DNA methylation of cord blood cell types: Applications for mixed cell birth studies. *Epigenetics* **11**, 354-362 (2016).
38. Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, 86 (2012).
39. Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet* **3**, 1724-1735 (2007).
40. Pedersen, B. S., Schwartz, D. A., Yang, I. V. & Kechris, K. J. Comb-p: software for combining, analyzing, grouping and correcting spatially correlated P-values. *Bioinformatics* **28**, 2986-2988 (2012).
41. Jaffe, A. E. *et al.* Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int J Epidemiol* **41**, 200-209 (2012).
42. Sharp, G. C. *et al.* Distinct DNA methylation profiles in subtypes of orofacial cleft. *Clin Epigenetics* **9**, 63 (2017).

Tables

Table 1. Association of peripheral blood DNA methylation at age 7 and breastfeeding. Regression coefficients (β) are average percent point differences in DNA methylation.

Breastfeeding	Statistic	CpG						
		cg11414913	cg00234095	cg04722177	cg03945777	cg17052885	cg05800082	cg24134845
Binary (ever vs. never)	P-value	5.2×10^{-8}	4.9×10^{-7}	2.7×10^{-6}	3.2×10^{-6}	4.9×10^{-6}	5.8×10^{-6}	3.3×10^{-5}
	β (SE)	-3.19 (0.59)	-1.74 (0.35)	-2.90 (0.62)	-0.84 (0.18)	1.79 (0.39)	1.05 (0.23)	0.23 (0.06)
Categories	P-value	-	-	-	-	-	-	-
	β (SE)	0 (Ref.)	0 (Ref.)	0 (Ref.)	0 (Ref.)	0 (Ref.)	0 (Ref.)	0 (Ref.)
0	P-value	1.5×10^{-6}	1.2×10^{-7}	5.3×10^{-4}	2.9×10^{-5}	8.2×10^{-6}	1.7×10^{-6}	6.8×10^{-5}
0.01-3 months	β (SE)	-3.19 (0.66)	-2.02 (0.38)	-2.45 (0.71)	-0.85 (0.20)	1.85 (0.41)	1.19 (0.25)	0.25 (0.06)
	P-value	5.4×10^{-7}	3.3×10^{-5}	5.8×10^{-5}	0.005	6.8×10^{-5}	6.4×10^{-4}	0.011
3.01-6 months	β (SE)	-3.50 (0.70)	-1.88 (0.45)	-3.22 (0.80)	-0.66 (0.23)	1.85 (0.47)	0.94 (0.28)	0.17 (0.07)
	P-value	2.5×10^{-5}	3.2×10^{-4}	5.9×10^{-5}	7.4×10^{-5}	6.1×10^{-6}	0.001	2.2×10^{-4}
6.01-12 months	β (SE)	-3.00 (0.71)	-1.59 (0.44)	-3.05 (0.76)	-0.90 (0.23)	2.02 (0.45)	0.87 (0.27)	0.24 (0.06)
	P-value	5.8×10^{-4}	0.037	1.1×10^{-6}	1.2×10^{-4}	0.008	0.001	4.4×10^{-4}
>12 months	β (SE)	-2.96 (0.86)	-0.93 (0.44)	-3.79 (0.78)	-0.99 (0.26)	1.29 (0.49)	1.04 (0.31)	0.25 (0.07)
Linear trend of categories	P-value	0.036	0.832	1.7×10^{-4}	0.007	0.067	0.230	0.020
	β (SE)	-0.42 (0.20)	-0.02 (0.11)	-0.70 (0.19)	-0.16 (0.06)	0.19 (0.10)	0.08 (0.07)	0.04 (0.02)
Continuous (in monhts)	P-value	0.080	0.766	2.5×10^{-4}	0.035	0.966	0.399	0.289
	β (SE)	-0.09 (0.05)	0.01 (0.03)	-0.18 (0.05)	-0.03 (0.02)	0.00 (0.03)	0.01 (0.02)	0.00 (0.00)

SE: standard error.

Table 2. Association between DNA methylation at different ages and ever breastfeeding. Regression coefficients (β) are average percent point differences in DNA methylation.

CpG	Time point	β	SE	P-value
cg11414913	At birth	-0.44	0.91	0.631
	7 years	-3.19	0.59	5.2×10^{-8}
	15-17 years	-2.47	0.85	0.004
cg00234095	At birth	0.59	0.57	0.296
	7 years	-1.74	0.35	4.9×10^{-7}
	15-17 years	0.29	0.43	0.505
cg04722177	At birth	-1.50	0.70	0.032
	7 years	-2.90	0.62	2.7×10^{-6}
	15-17 years	-1.05	0.78	0.180
cg03945777	At birth	0.42	0.3	0.158
	7 years	-0.84	0.18	3.2×10^{-6}
	15-17 years	0.10	0.29	0.742
cg17052885	At birth	1.32	0.57	0.022
	7 years	1.79	0.39	4.9×10^{-6}
	15-17 years	-0.29	0.47	0.547
cg05800082	At birth	-0.53	0.36	0.144
	7 years	1.05	0.23	5.8×10^{-6}
	15-17 years	0.56	0.32	0.083
cg24134845	At birth	0.04	0.07	0.535
	7 years	0.23	0.06	3.3×10^{-5}
	15-17 years	0.00	0.08	0.991

SE: standard error.

Table 3. Association between peripheral blood DNA methylation at different ages at each differentially methylation region (DMR) and ever breastfeeding. Regression coefficients (β) are average percent point differences in DNA methylation averaged across CpGs that belong to the DMR.

DMR (Chr:Start-End ^a)	At birth			7 years			15-17 years		
	β	SE	P-value	β	SE	P-value	β	SE	P-value
5:97,867-98,797	0.30	0.21	0.146	0.43	0.21	0.043	0.30	0.21	0.158
19:365,914-366,989	-0.01	0.34	0.975	0.05	0.34	0.881	-0.04	0.35	0.897
18:106,178-106,850	-0.08	0.77	0.913	0.14	0.75	0.855	0.23	0.77	0.767
1:425,524-426,297	0.26	0.62	0.673	0.33	0.61	0.590	0.16	0.62	0.800
9:91,296-92,146	-0.10	0.33	0.759	-0.18	0.33	0.578	-0.10	0.34	0.755
17:222,498-222,991	-0.01	0.37	0.983	0.00	0.36	0.994	-0.04	0.36	0.913
4:136,643-137,027	-0.03	0.41	0.951	-0.37	0.38	0.324	-0.31	0.41	0.448
22:255,590-256,045	0.40	0.71	0.577	1.18	0.70	0.095	1.06	0.71	0.136
4:33,482-33,808	0.13	2.05	0.950	0.06	2.00	0.978	0.08	2.04	0.967
8:409,905-410,098	0.82	1.31	0.530	1.05	1.32	0.425	1.04	1.32	0.433
1:224,191-225,190	0.03	0.45	0.940	-0.03	0.44	0.951	-0.03	0.45	0.948
9:61,093-61,964	-0.39	0.50	0.432	-0.44	0.49	0.369	-0.39	0.50	0.435

^aHuman Genome Assembly GRCh37.

Chr: Chromosome. SE: standard error.

Table 4. Directional concordance (in %) between time points for each individual CpG belonging to the same differentially methylated region (DMR).

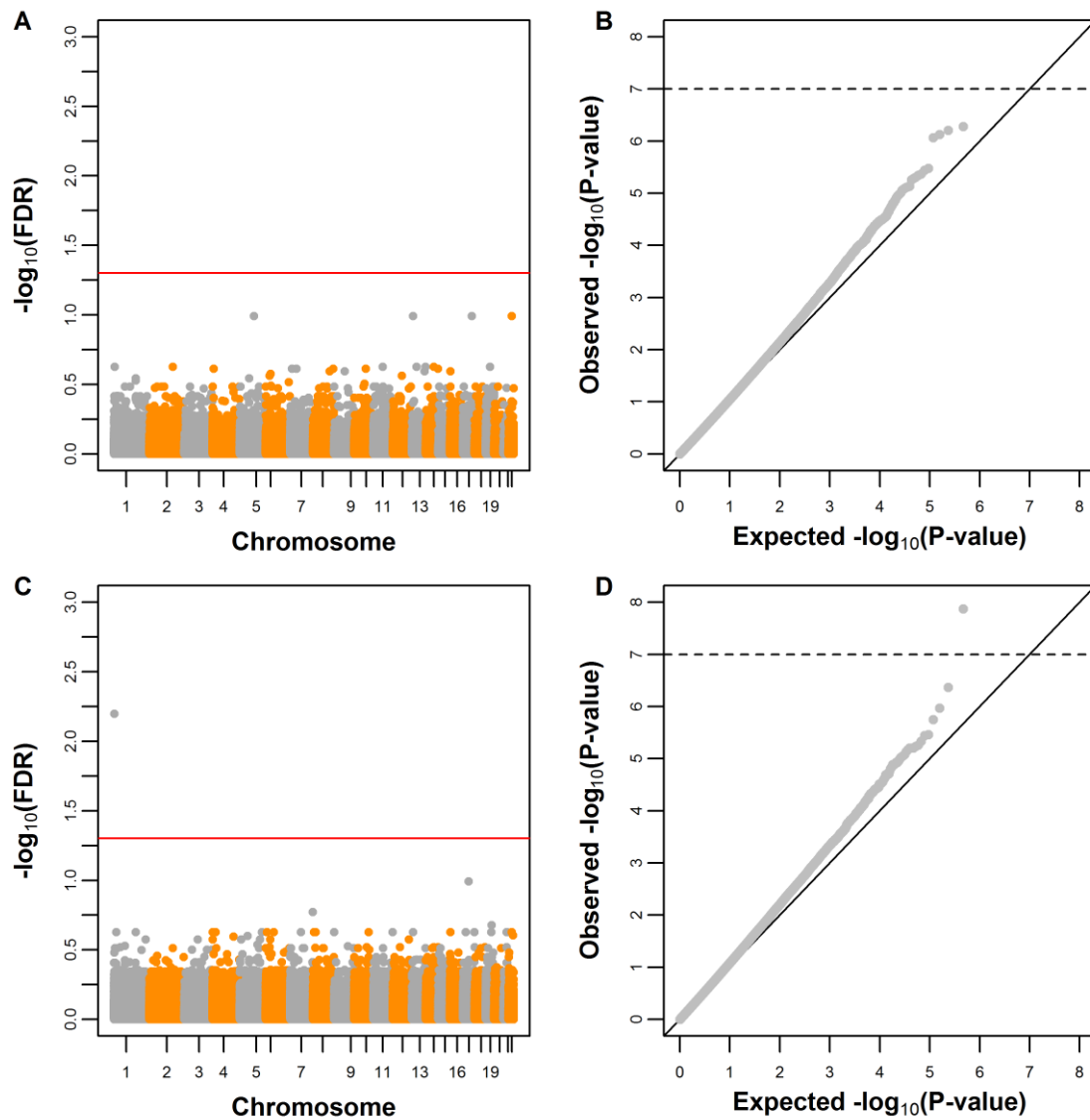
DMR (Chr:Start-End ^a)	Number of CpGs	At birth and 7 years		7 years and 15-17 years	
		Concordance	P-value	Concordance	P-value
5:97,867-98,797	275	66.2	8.7×10^{-8}	69.1	2.2×10^{-10}
19:365,914-366,989	205	47.8	0.576	54.1	0.264
18:106,178-106,850	18	72.2	0.096	83.3	0.008
1:425,524-426,297	64	68.8	0.004	56.3	0.382
9:91,296-92,146	185	54.1	0.303	58.4	0.027
17:222,498-222,991	140	55.7	0.205	49.3	0.933
4:136,643-137,027	13	69.2	0.267	61.5	0.581
22:255,590-256,045	30	63.3	0.200	83.3	3.3×10^{-4}
4:33,482-33,808	5	60.0	0.999	60.0	0.999
8:409,905-410,098	7	85.7	0.125	100.0	0.016
1:224,191-225,190	129	57.4	0.113	47.3	0.597
9:61,093-61,964	91	57.1	0.208	56.0	0.294

^aHuman Genome Assembly GRCh37.

Chr: Chromosome.

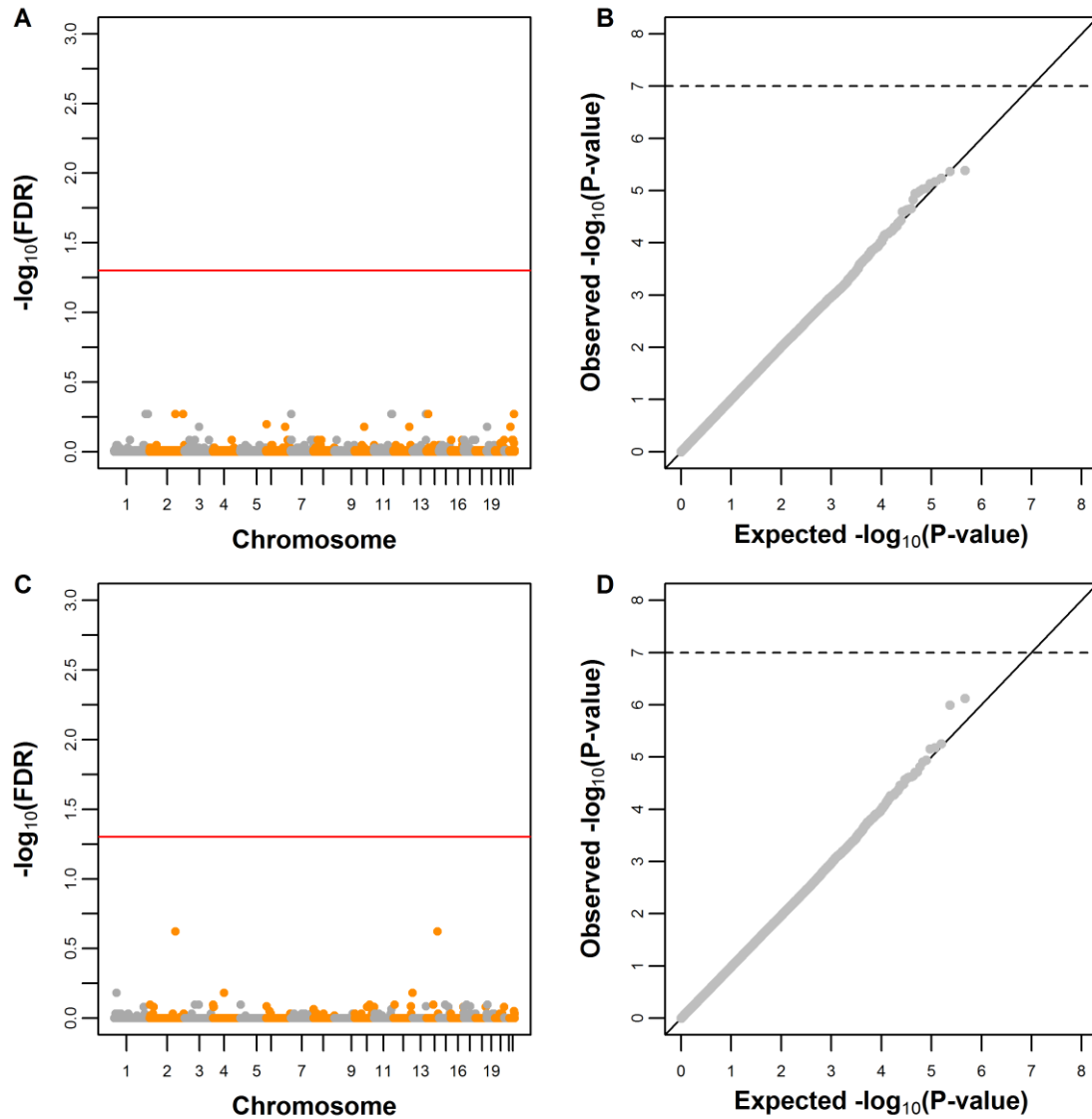
Figures

Figure 1. Manhattan and Q-Q plots of the breastfeeding EWAS, comparing peripheral blood methylation at age 7 between never vs. ever breastfed individuals.



A,C: Manhattan plots. B,D: Q-Q plots. A,B: Minimally-adjusted model. C,D: Fully-adjusted model.

Figure 2. Manhattan and Q-Q plots of the breastfeeding EWAS, comparing peripheral blood methylation at age 7 according to breastfeeding duration (in categories, assuming a linear trend).



A,C: Manhattan plots. B,D: Q-Q plots. A,B: Minimally-adjusted model. C,D: Fully-adjusted model.

Supplementary Material

Supplementary Tables

Supplementary Table 1. Description of the individuals included in the main analysis, compared to all ARIES participants, restricting to those with age 7 methylation data available.

Variable	Statistic/category ^a	All ARIES participants (n=995)	Participants included in this study (n=702)
Maternal education at birth	CSE	8.9%	7.2%
	Vocational education	7.4%	6.0%
	GCE Ordinary level	34.3%	33.8%
	GCE Advanced level	29.1%	29.9%
	Degree	20.3%	23.1%
Maternal age at birth (years)	Mean (SD)	29.5 (4.4)	30.0 (4.4)
Parity	0	46.5%	45.7%
	1	36.9%	37.5%
	2	12.7%	13.4%
	≥3	3.9%	3.4%
Maternal smoking in relation to Pregnancy	Never	86.3%	87.7%
	Before	3.7%	4.0%
	During	10.0%	8.3%
Folic acid Supplementation	No	75.9%	75.9%
	Yes	24.1%	24.1%
Caesarean section	No	90.4%	90.2%
	Yes	9.6%	9.8%
Birthweight (g)	Mean (SD)	3487 (486)	3490 (476)
Sex	Male	48.9%	49.1%
	Female	51.1%	50.9%
Ethnicity	European	97.0%	99.9%
	Other	3.0%	0.1%
Breastfeeding duration (months)	0	11.1%	10.4%
	0.1-3	32.0%	31.0%
	3.1-6	16.2%	16.2%
	6.1-12	27.6%	28.2%
	>12	13.1%	14.2%

^aMean and SD for continuous variables, and each category (for which proportions are shown) for categorical variables.

CSE: Certificate of Secondary Education. GCE: General Certificate of Education.

SD: standard deviation.

Supplementary Table 2. Description of the CpGs that presented at least suggestive evidence of association with ever breastfeeding in the fully-adjusted analysis at age 7.

CpG	Chromosome: position (bp)^a	Nearest gene	Distance (bp) to nearest gene
cg11414913	1:2,799,662	<i>TTC34</i>	93,432
cg00234095	17:39,440,474	<i>KRTAP9-7</i>	8,015
cg04722177	19:39,737,768	<i>IFNL4</i>	Intragenic
cg03945777	7:157,514,049	<i>PTPRN2</i>	Intragenic
cg17052885	17:78,896,012	<i>RPTOR</i>	Intragenic
cg05800082	6:56,508,429	<i>DST</i>	Intragenic
cg24134845	10:100,992,149	<i>HPSE2</i>	Intragenic

^aHuman Genome Assembly GRCh37.

bp: base pairs.

Supplementary Table 3. Differentially methylated regions (DMR) in peripheral blood at age 7 according to ever breastfeeding.

DMR (Chr:Start-End ^a)	Number of CpGs	P-value	Nearest gene	Distance (bp) to nearest gene
5:97,867-98,797	275	3.2×10^{-6}	<i>PLEKHG4B</i>	Intragenic
19:365,914-366,989	205	9.7×10^{-5}	<i>THEG</i>	Intragenic
18:106,178-106,850	18	0.001	<i>DUX4</i>	Intragenic
1:425,524-426,297	64	0.002	<i>BC036251</i>	4,458
9:91,296-92,146	185	0.003	<i>PGM5P3-AS1</i>	1,719
17:222,498-222,991	140	0.003	<i>RPH3AL</i>	19,865
4:136,643-137,027	13	0.007	<i>ZNF595/ZNF718</i>	Intragenic
22:255,590-256,045	30	0.012	<i>AK022914</i>	15,894,215
4:33,482-33,808	5	0.019	<i>ZNF595/ZNF718</i>	19,419
8:409,905-410,098	7	0.025	<i>FBXO25</i>	Intragenic
1:224,191-225,190	129	0.045	<i>LOC729737</i>	83,625
9:61,093-61,964	91	0.046	<i>AY343892</i>	10,734

^aHuman Genome Assembly GRCh37.

^bNo gene within a 100,000 bp window centred at this region.

Chr: Chromosome. bp: base pairs.

Supplementary Table 4. Directional concordance (in %) between time points for each individual CpG belonging to the same differentially methylated region (DMR). Only CpGs that achieved $P < 0.05$ in at least one time point were considered.

DMR (Chr:Start-End ^a)	Number of CpGs	At birth and 7 years		7 years and 15-17 years	
		Concordance	P-value	Concordance	P-value
5:97,867-98,797	69	72.5	2.4×10^{-4}	85.5	1.4×10^{-9}
19:365,914-366,989	38	52.6	0.871	68.4	0.034
18:106,178-106,850	8	75.0	0.289	100.0	0.008
1:425,524-426,297	15	80.0	0.035	73.3	0.118
9:91,296-92,146	38	63.2	0.143	71.1	0.014
17:222,498-222,991	22	45.5	0.832	50.0	0.999
4:136,643-137,027	3	100.0	0.250	66.7	0.999
22:255,590-256,045	16	68.8	0.210	93.8	0.001
4:33,482-33,808	2	50.0	0.999	0.0	0.500
8:409,905-410,098	3	100.0	0.250	100.0	0.250
1:224,191-225,190	23	39.1	0.405	56.5	0.678
9:61,093-61,964	24	75.0	0.023	66.7	0.152

^aHuman Genome Assembly GRCh37.

Chr: Chromosome.

Effect modification of *FADS2* polymorphisms on the association between breastfeeding and intelligence: results from a collaborative meta-analysis

Fernando Pires Hartwig^{1,2*}, Neil Martin Davies^{2,3}, Bernardo Lessa Horta¹, Tarunveer S. Ahluwalia⁴, Hans Bisgaard⁴, Klaus Bønnelykke⁴, Avshalom Caspi^{5,6}, Terrie E. Moffitt^{5,6}, Richie Poulton⁷, Ayesha Sajjad⁸, Henning W Tiemeier^{8,9}, Albert Dalmau Bueno^{10,11,12}, Mònica Guxens^{9,10,11,12}, Mariona Bustamante Pineda^{10,11,12,13}, Loreto Santa-Marina^{12,14,15}, Nadine Parker^{16,17}, Tomáš Paus^{16,18}, Zdenka Pausova^{19,20,21}, Lotte Lauritzen²², Theresia M. Schnurr²³, Kim F. Michaelsen²², Torben Hansen²³, Wendy Oddy²⁴, Craig E. Pennell²⁵, Nicole M. Warrington^{25,26}, George Davey Smith^{2,3†} and Cesar Gomes Victora^{3†}

¹Postgraduate Program in Epidemiology, Federal University of Pelotas, Pelotas, Brazil.

²Medical Research Council Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom.

³School of Social and Community Medicine, University of Bristol, Bristol, United Kingdom.

⁴COPSAC, Copenhagen Prospective Studies on Asthma in Childhood, Herlev and Gentofte Hospital, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark.

⁵Duke University, Durham, USA.

⁶Institute of Psychiatry, Psychology, and Neuroscience, King's College London, London, United Kingdom.

⁷Department of Psychology, University of Otago, Dunedin, New Zealand.

⁸Department of Epidemiology, Erasmus University Medical Centre, Rotterdam, The Netherlands.

⁹Department of Child and Adolescent Psychiatry/Psychology, Erasmus University Medical Centre, Rotterdam, The Netherlands.

¹⁰ISGlobal, Centre for Research in Environmental Epidemiology (CREAL), Barcelona, Spain.

¹¹Universitat Pompeu Fabra (UPF), Barcelona, Spain.

¹²CIBER Epidemiología y Salud Pública (CIBERESP), Madrid, Spain.

¹³Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain.

¹⁴BIODONOSTIA Health Research Institute, San Sebastian, Spain.

¹⁵Public Health Division of Gipuzkoa, San Sebastian, Spain.

¹⁶Rotman Research Institute, Baycrest Centre for Geriatric Care, Toronto, Canada.

¹⁷Institute of Medical Science, University of Toronto, Toronto, Canada.

¹⁸Departments of Psychiatry and Psychology, University of Toronto, Toronto, Canada.

¹⁹Hospital for Sick Children Research Institute, Peter Gilgan Centre for Research and Learning, Toronto, Canada.

²⁰Department of Nutritional Sciences, University of Toronto, Toronto, Canada.

²¹Department of Physiology, University of Toronto, Toronto, Canada.

²²Department of Nutrition, Exercise and Sports, Faculty of Science, University of Copenhagen, Copenhagen, Denmark.

²³Novo Nordisk Foundation Centre for Basic Metabolic Research, Section of Metabolic Genetics, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.

²⁴Menzies Institute for Medical Research, University of Tasmania, Hobart, Australia.

²⁵School of Women's and Infants' Health, The University of Western Australia, Perth, Australia.

²⁶The University of Queensland Diamantina Institute, The University of Queensland, Translational Research Institute, Brisbane, Australia.

†Joint senior authors.

*Corresponding author. Postgraduate Program in Epidemiology, Federal University of Pelotas, Pelotas (Brazil) 96020-220. Phone: 55 53 981068670. E-mail: fernandophartwig@gmail.com; fh15144@bristol.ac.uk.

Abstract

Background: Accumulating evidence suggests that breastfeeding benefits the children's intelligence, possibly due to long-chain polyunsaturated fatty acids (LC-PUFAs) present in breast milk. Under a nutritional adequacy hypothesis, an interaction between breastfeeding and genetic variants associated with endogenous LC-PUFAs synthesis might be expected. However, the literature on this topic is controversial.

Methods: We investigated this GenexEnvironment interaction through a collaborative effort. The primary analysis involved >12,000 individuals and used ever breastfeeding, *FADS2* polymorphisms rs174575 and rs1535 coded assuming a recessive effect of the G allele, and intelligence quotient (IQ) in Z scores.

Results: There was no strong evidence of interaction, with pooled covariate-adjusted interaction coefficients (i.e., difference between genetic groups of the difference in IQ Z scores comparing ever with never breastfed individuals) of 0.12 (95% CI: -0.19; 0.43) and 0.06 (95% CI: -0.16; 0.27) for the rs174575 and rs1535 variants, respectively. Secondary analyses corroborated these results. In studies with ≥ 5.85 and < 5.85 months of breastfeeding duration, pooled estimates for the rs174575 variant were 0.50 (95% CI: -0.06; 1.06) and 0.14 (95% CI: -0.10; 0.38), respectively, and 0.27 (95% CI: -0.28; 0.82) and -0.01 (95% CI: -0.19; 0.16) for the rs1535 variant.

Conclusions: Our findings do not support an interaction between ever breastfeeding and *FADS2* polymorphisms. However, subgroup analysis raised the possibility that breastfeeding supplies LC-PUFAs requirements for cognitive development if it lasts for some (currently unknown) time. Future studies in large individual-level datasets would allow properly powered subgroup analyses and further improve our understanding on the breastfeedingx*FADS2* interaction.

Keywords: Breastfeeding; Intelligence; *FADS2*; Fatty acids; Effect modification; Meta-analysis.

Key messages

- Breastfeeding is suggested to improve children's intelligence, possibly due to long-chain polyunsaturated fatty acids (LC-PUFAs).
- The literature on the interaction between breastfeeding and variants in the *FADS2* on intelligence quotient (IQ) is controversial.
- Our *de novo* collaborative meta-analysis did not support this interaction when comparing ever vs. never breastfed individuals.
- Subgroup analyses, although underpowered, were compatible with a role of breastfeeding duration in this interaction.

Introduction

Breastfeeding has well-established short term benefits on children's health. There is also accumulating evidence that breastfeeding may also benefit cognitive development [1]. A recent meta-analysis of observational studies reported that breastfed subjects scored higher on intelligence quotient (IQ) tests [mean difference 3.4 (95% CI: 2.3; 4.6)] than non-breastfed subjects [2]. Although issues such as residual confounding [3] and publication bias [4] may have affected this estimate, randomised controlled trials of breastfeeding promotion reported benefits in motor development in the first year of life [5] and in IQ at 6.5 years of age [6]. Additional studies corroborate the notion that breastfeeding has a causal effect on IQ. These include comparisons between cohorts with different confounding structures [7], and between mothers who tried, but could not breastfeed their child, and mothers who had formula feeding as their first choice [8].

One of the possible biological mechanisms underlying the effect of breastfeeding on IQ is through long-chain polyunsaturated fatty acids (LC-PUFAs), such as docosahexaenoic acid (DHA). Meta-analyses of randomised controlled trials of supplementation of DHA and other LC-PUFAs in infants reported improved cognitive development [9] and visual acuity [10]. Indeed, DHA is an important component of the membrane of brain cells and retina cells [11,12]. Studies in animal models and humans suggest that adequate levels of DHA are important for cognitive development through influencing several processes, such as biogenesis and fluidity of cellular membranes, neurogenesis, neurotransmission and protection against oxidative stress [12,13].

The role of LC-PUFAs in the association between breastfeeding and IQ can be investigated through a Gene×Environment (G×E) interaction analysis. For example, it is possible that there is an upper limit for the benefits of increasing DHA levels and such requirements are met by pre-formed DHA available in breast milk. In this case, inter-individual variation in IQ due to genetically determined differences in DHA endogenous synthesis from metabolic precursors would only be observable in individuals who were not breastfed [14]. This G×E interaction has been investigated using single nucleotide

polymorphisms (SNPs) in the *FADS2* gene [14-18]. This gene encodes a desaturase enzyme that catalyses a rate-limiting reaction in the LC-PUFAs pathway [19,20]. Candidate gene and genome-wide approaches reported that minor alleles of SNPs in the *FADS2* gene were associated with lower levels of PUFAs in plasma and erythrocyte phospholipids [21-24].

Caspi et al. were the first to evaluate the interaction between genetic variation in *FADS2* and breastfeeding, with IQ in children as the outcome. Two SNPs were evaluated: rs1535 (major/minor alleles: A/G) and rs174575 (major/minor alleles: C/G). For both SNPs, having ever being breastfed was positively associated with IQ in all genetic groups, except in G-allele homozygotes, where there was no association [15]. Although there was evidence for a G×E interaction, it was not consistent with the nutritional adequacy hypothesis outlined above. However, in a replication study, Steer et al. results were consistent with the nutritional adequacy hypothesis (and therefore inconsistent with Caspi et al.'s findings), with breastfed individuals presenting similar mean values of IQ across *FADS2* genotypes. Such values were higher than those observed in never breastfed individuals, with the lowest value (and thus the greatest effect of breastfeeding) being in GG individuals [14]. Morales et al. reported that a negative association between genotypes in other genetic variants related to lower activity of enzymes involved in elongation and desaturation processes and cognition was only evident in non-breastfed individuals [25]. Three studies in twins (but not twin studies, in the sense that they did not aim at estimating heritability) did not detect strong evidence supporting this G×E interaction [16-18].

The controversial results observed in the literature may be due to lack of power (in the case smaller studies) and/or to contextual differences that lead to heterogeneity between studies, as discussed in detail elsewhere [26]. In this study, we aimed at improving the current understanding on this G×E interaction and gaining insights into the sources of heterogeneity between studies through a consortium-based initiative [26].

Methods

Overview of the study protocol

The protocol of this study has been published elsewhere [26]. Briefly, studies that were known by the coordinating team to have at least some of the data required available, as well as other studies suggested by collaborators, were invited to participate. All studies that were contacted (and were eligible) accepted to participate.

All of the following criteria were required for eligibility: i) availability of at least a binary breastfeeding variable (i.e., whether or not the study individuals were ever breastfed), intelligence measured using standard tests, and at least rs174575 or rs1535 SNPs (either genotyped or imputed); and ii) European-ancestry studies, or multi-ethnic studies if possible to define a subsample of European ancestry individuals. Exclusion criteria were: i) only poorly imputed genetic data were available (metrics of imputation such as r^2 or INFO quality below 0.3); ii) twin studies; iii) lack of appropriate ethical approval.

Data analysis was performed locally by data analysts of the collaborating studies. Standardised analysis scripts written in R (<http://www.r-project.org/>) were prepared centrally and distributed to the analysts, along with a detailed analysis plan and instructions to format the data. The scripts automatically generated files containing summary descriptive and association statistics, which were centrally meta-analysed.

As the analyses progressed, some modifications in the original protocol were required. These are described in the Supplementary Methods.

Participating studies

A total of 10 eligible studies were identified, all of which were included in the meta-analysis: the 1982 Pelotas Birth Cohort Study [27,28], Dunedin Multidisciplinary Health and Development Study [15], Avon Longitudinal Study of Parents and Children (ALSPAC) [29], Copenhagen Prospective Study on Asthma in Childhood (COPSAC) 2010 [30,31], Generation R Study [32-34], *Infancia y Medio Ambiente* (INMA) Project [35],

Western Australian Pregnancy Cohort (Raine) Study [36-38], *Småbørn Kost Og Trivsel-I* (SKOT-I) [39,40], SKOT-II [41,42] and Saguenay Youth Study (SYS) [43,44].

In addition, a subsample of 32,842 individuals from the UK Biobank [45] was included. However, this subsample did not fulfil the pre-established eligibility criteria because IQ was not measured using a standard test. Therefore, these data were used in secondary analyses only.

Information about the participating studies is shown in Supplementary Tables 1-3.

Statistical analyses

The main outcome variable was IQ. IQ tests varied between studies (Supplementary Table 1), so IQ measures were converted to Z scores (mean=0 and variance=1) within each participating study. The primary analysis involved breastfeeding (coded as never=0 and ever=1), *FADS2* polymorphism assuming a recessive genetic effect of the G allele (i.e., GG individuals=1; heterozygotes and non-G allele homozygotes=0) and an interaction term between them. Different genetic effects, different categorizations of breastfeeding, and exclusive breastfeeding (defined as receiving only breast milk and no other food or drink, including water) were evaluated in pre-planned secondary analyses. Unless explicitly stated, all analyses refer to any quality of breastfeeding (i.e., combining exclusive and non-exclusive).

Three analysis models were performed: (i) unadjusted (i.e., no covariates); (ii) adjusted 1: controlling for sex and age (linear and quadratic terms) when IQ was measured, ancestry-informative principal components (when available) and genotyping centre (for studies involving multiple laboratories); (iii) adjusted 2: same covariates in “adjusted 1” model, as well as maternal education (linear and quadratic terms) and maternal cognition (linear and quadratic terms); if only one of the maternal variables was available, adjusted model 2 controlled only for that variable. Continuous covariates, as well as sex (which was coded as male=0 and female=1), were mean-centred before analysis, and squaring was performed before mean centring. Covariate

adjustment was performed by including not only a “main effect” term, but also (*FADS2*×Covariate) and (Breastfeeding×Covariate) interaction terms [46].

As a sensitivity analysis, the role of gene-environment correlation was evaluated by repeating models i) and ii), but having maternal cognition (in Z scores) or maternal schooling (in years) as outcome variables rather than the participant’s IQ. Maternal cognition or schooling are important predictors of an individual’s IQ, and cannot be consequences of the participant’s genotype. Therefore, any evidence of breastfeeding×*FADS2* interaction in this analysis is indicative that those maternal variables may confound the main breastfeeding×*FADS2* interaction analysis (i.e., having participant’s IQ as the outcome variable).

Analyses were performed using linear regression with heteroskedasticity-robust standard errors. Results from all studies were pooled using fixed and random effects meta-analysis. Stratified meta-analysis and random effects meta-regression were used to evaluate the potential moderating role of the following variables (one meta-regression model per moderator): IQ test; adjustment for ancestry-informative principal components; age at IQ measurement; timing of breastfeeding measurement; continental region; mean year of birth; prevalence of having ever being breastfed; mean breastfeeding duration; and sample size. Adjusted R^2 values, which can be interpreted as the amount of between-study heterogeneity explained by the moderator, were obtained from the meta-regression models.

Results

Characteristics of participating studies

As shown in Supplementary Table 1, seven out of the 10 eligible studies were conducted in Europe, four were population-based and two were multi-ethnic. The average year at birth ranged from 1972 to 2011. Three studies measured breastfeeding prospectively, and four measured IQ using the Wechsler Intelligence Scale (two for children and two for adults).

Supplementary Table 2 provides a description of the two *FADS2* SNPs in each study. The SNPs rs174575 and rs1535 were directly genotyped in three and five studies, respectively. The minimum value of imputation quality was 0.984. The frequency of the G allele ranged from 20.5% to 30.8% for the rs174575 variant, and from 28.5% to 39.1% for the rs1535 variant. There was no strong statistical evidence against Hardy-Weinberg Equilibrium, with the smallest P-values being 0.058 (Generation R), 0.074 (SKOTI-II) for rs174575, and 0.085 (1982 Pelotas Birth Cohort), 0.044 (Raine) and 0.089 (SKOTI-II) for rs1535. Although these results may be suggestive of some population substructure (especially in Generation R and in the 1982 Pelotas Birth Cohort, which are multi-ethnic studies) or batch effects (especially in SKOTI-II, which is a combination of two independent studies), it is unlikely that such phenomena substantially influenced the results because ancestry-informative principal components computed using genome-wide genotyping data were available and adjusted for in these four studies.

Additional study characteristics are displayed in Supplementary Table 3. Among eligible studies (i.e., excluding the UK Biobank), the mean age, maternal education, and breastfeeding duration ranged from 2.5 to 30.2 years, 11 to 19 years, and 2.3 to 8.2 months, respectively. All IQ measures produced a variable with mean close to 100 and similar standard deviations (median: 12.2; range: 9.6 to 16.3). The exception was the one used in SKOT-I and SKOT-II (i.e., third edition of the Ages and Stages Questionnaire), which produced a variable with mean close to 50.

Primary analysis

In analyses without stratification according to genotype, ever breastfeeding was associated with increases of 0.37 (95% CI: 0.32; 0.42) and 0.30 (95% CI: 0.20; 0.40) Z scores in IQ in fixed and random effects meta-analysis, respectively. Assuming that a Z score corresponds to 12.2 points (the median of the standard deviation of IQ measures among participating studies), these coefficients correspond to 4.5 and 3.7 points in IQ. In the fully adjusted model (adjusted 2), the respective coefficients were 0.26 (95% CI: 0.21; 0.32) and 0.17 (95% CI: 0.03; 0.32), or 3.2 and 2.1 points in IQ.

Table 1 and Figure 1 display the results of the primary analysis. There was considerable between-study heterogeneity. Among non-G carriers for the rs174575 SNP, pooled random effects estimates of IQ Z scores according to breastfeeding (ever=1; never=0) were 0.29 (95% CI: 0.17; 0.40) and 0.15 (95% CI: 0.00; 0.31) in the unadjusted and fully-adjusted models, respectively. Among GG individuals, the respective estimates were 0.43 (95% CI: 0.16; 0.70) and 0.31 (95% CI: 0.05; 0.58). There was no strong evidence of interaction, with pooled estimates of the breastfeeding×*FADS2* interaction term of 0.18 (95% CI: -0.18; 0.54) and 0.12 (95% CI: -0.19; 0.43), respectively. These coefficients can be interpreted as the difference between genetic groups of the difference in IQ Z scores comparing ever with never breastfed individuals. Similar results were obtained when using fixed effects meta-analysis.

(Table 1 here)

Results for the rs1535 variant presented a similar trend, but were even less suggestive of interaction. When using random effects meta-analysis, the estimates of the interaction term were -0.04 (95% CI: -0.24; 0.15) and 0.06 (95% CI: -0.16; 0.27) in the unadjusted and fully-adjusted models, respectively. Using fixed effects meta-analysis yielded similar results.

Secondary analysis

As shown in Table 2 and Supplementary Tables 4-6, there was no strong indication of interaction when analysing other categorisations of breastfeeding duration and *FADS2* SNPs coded assuming a recessive effect. This was also the case when *FADS2* variants were coded assuming additive (Supplementary Table 7), dominant (Supplementary Table 8) and overdominant (Supplementary Table 9) effects. The same was observed for exclusive breastfeeding (Supplementary Tables 10-13).

(Table 2 here)

Supplementary Table 14 displays the results obtained when including the UK Biobank, which was analysed as two independent samples according to the genotyping platform (Biobank_Axiom and Biobank_BiLEVE). Its inclusion resulted in a combined sample size

of more than 45,000 individuals. When *FADS2* variants were coded assuming recessive effects, the pooled estimates from the unadjusted model -0.02 (95% CI: -0.10; 0.06) and 0.08 (95% CI: -0.13; 0.29) for fixed and random-effects meta-analysis, respectively. The corresponding estimates from the adjusted (1) model were -0.04 (95% CI: -0.13; 0.04) and 0.00 (95% CI: -0.21; 0.20), respectively. There was also no strong statistical evidence supporting an interaction when other genetic effects were assumed.

Sensitivity analysis

Table 3 displays the results of random-effects meta-regression. Neither type of IQ test, timing of breastfeeding measurement, continental region nor mean year of birth explained a substantial amount of between-study heterogeneity. For rs174575, the adjusted R^2 of ancestry-informative principal components was 88.0%, with pooled estimates of 0.28 (95% CI: 0.02; 0.54) and -0.38 (95% CI: -0.72; -0.04) Z scores in IQ from studies that did and did not adjust for principal components, respectively, which would be suggestive of confounding due to population stratification towards a negative association. Age at IQ measurement was inversely associated with the magnitude of the interaction term, with pooled estimates of 0.06 (95% CI: -0.46; 0.58) and 0.20 (95% CI: -0.18; 0.58) when IQ was measured at 10 years of age or more, or before that age (respectively), possibly suggesting an attenuation of the effect over time. The adjusted R^2 was 10.4% when entering age as a continuous variable, but 0% when dichotomised. When stratifying studies according to prevalence of ever breastfeeding, the pooled estimate among studies with a prevalence $\geq 90\%$ was 0.36 (95% CI: -0.19; 0.90), and -0.04 (95% CI: -0.38; 0.29) when pooling the remaining studies. Adjusted R^2 estimates were 16.4% and 72.3% when prevalence of ever breastfeeding was analysed as a binary and as a continuous variable, respectively. Among studies with breastfeeding duration equal to or greater than the median among studies (i.e., 5.85 months), the pooled estimate was 0.50 (95% CI: -0.06; 1.06), compared to 0.14 (95% CI: -0.10; 0.38) when pooling the remaining studies. The adjusted R^2 was 45.5% when breastfeeding duration was dichotomised at the median, but 0% when analysed continuously. When stratifying studies into larger (≥ 1000 individuals) and smaller (< 1000 individuals), the pooled estimates were 0.26 (95% CI:

0.00; 0.52) and -0.03 (95% CI: -0.63; 0.56), with an adjusted R^2 of 33.8% when sample size was dichotomised, and of 0% when analysed in continuous form.

(Table 3 here)

Regarding the rs1535 variant, the respective subgroup-specific estimates were consistent with those of the rs174575 SNP: adjustment for principal components, with pooled estimates of 0.09 (95% CI: -0.19; 0.37) and -0.03 (95% CI: -0.32; 0.25) among studies that did and did not perform this adjustment, respectively; age at IQ measurement, with pooled estimates of 0.04 (95% CI: -0.19; 0.37) and 0.07 (95% CI: -0.31; 0.45) among studies that measured IQ when individuals were ≥ 10 and < 10 years-old, respectively; and sample size, with pooled estimates of 0.11 (95% CI: -0.12; 0.34) and 0.01 (95% CI: -0.43 and 0.45) among larger and smaller studies, respectively. However, in all those cases the adjusted R^2 values were 0%. Prevalence of ever breastfeeding presented adjusted R^2 values of 0% and 8.3% when dichotomised and analysed continuously, respectively. The pooled estimates for the rs1535 variant were 0.15 (95% CI: -0.31; 0.62) and 0.01 (95% CI: -0.15; 0.18) among studies with prevalence of ever breastfeeding of $\geq 90\%$ and $< 90\%$, respectively. The most consistent moderator between SNPs was breastfeeding duration, with pooled estimates for the rs1535 SNP of 0.27 (95% CI: -0.28; 0.82) and -0.01 (95% CI: -0.19; 0.16) among studies with ≥ 5.85 and < 5.85 months of duration, respectively; adjusted R^2 values were 22.2% and 4.9% when breastfeeding duration was dichotomised and analysed continuously, respectively (Figure 2).

There was no strong evidence in support of gene-environment correlation involving maternal education or maternal cognition (Table 4). Regarding the rs174575 variant, random effects meta-analytical estimates from the adjusted model were 0.16 (95% CI: -0.45; 0.78) for maternal education, and -0.02 (95% CI: -0.25; 0.21) for maternal cognition, respectively. The corresponding estimates for the rs1535 SNP were -0.12 (95% CI: -0.51; 0.27) and 0.14 (95% CI: -0.04; 0.33).

(Table 4 here)

Discussion

Our primary analyses were not supportive of the hypothesis that the *FADS2* polymorphisms rs174575 and rs1535 and breastfeeding interact to affect IQ. This was also the case in a priori secondary analyses using different categorisations of breastfeeding, exclusive rather than any quality of breastfeeding, assuming different genetic effects and including a large study that did not meet all eligibility criteria. Sensitivity analyses were not supportive that gene-environment correlation involving maternal education or maternal cognition substantially influenced the results. Random effects meta-regression suggested that breastfeeding duration was an important moderator.

Results from our primary and secondary analyses were not supportive of the nutritional adequacy hypothesis, according to which a positive interaction coefficient would be expected [14]. In other words, there might be no upper limit (or it may be very high) of the effects of LC-PUFAs on IQ, so that supplementing infants with LC-PUFAs could be beneficial for cognition for both lactating and non-lactating infants alike. Importantly, this does not imply that LC-PUFAs supplementation completely replaces the benefits of breastfeeding, since the latter may act through diverse mechanisms, and also provide benefits other than for intelligence [1,47].

On the other hand, in our random effects meta-regression analysis, studies with longer average breastfeeding duration generally presented interaction coefficients that were positive and stronger in magnitude than studies with shorter breastfeeding duration. Moreover, average breastfeeding duration was the most consistent moderator between polymorphisms. Considering that positive interaction coefficients are expected under the nutritional adequacy hypothesis, this result raises the possibility that there may be an upper limit of the benefits of LC-PUFAs, but achieving such limits from breast milk requires that breastfeeding lasts for some (currently unknown) time. Given that breastfeeding practices in the participating studies were generally well below international recommendations [48,49], it is possible that the amount of LC-PUFA received from breast milk were, on average, lower than this threshold.

The strengths of our study include: appropriate sample size for the primary analysis [26]; publication of study protocol I[26], which helps to avoid biased reporting; analyses performed using standardised analysis scripts and harmonised (as much as possible) datasets; inclusion of published and unpublished reports, thus minimising publication bias; several *a priori* defined secondary and sensitivity analyses; proper adjustment for covariates in the G×E setting; and IQ measures with similar variances, which reduces heterogeneity that could arise due to Z score conversion [50,51].

Our study also had limitations. Some of them were related to the small numbers of individuals in some categories, which we tried to resolve by changing the protocol, such as in the case of the definition of never being breastfed and exclusion of some categorisations of breastfeeding from the analysis. Indeed, had the latter been maintained, the hypothesis above regarding breastfeeding duration and nutritional adequacy could have been studied. However, due to statistical issues, we opted for excluding this variable. Other limitations were: small sample size for some analyses, such as those involving exclusive breastfeeding; heterogeneity in important study characteristics, such as age, IQ test, timing of breastfeeding measurement, etc.; and small number of studies for meta-regression analyses. Another potential limitation is lack of adjustment for maternal genotypes, which may confound the association between participant's genotype and IQ by influencing fatty acid composition in breast milk [25]. However, although there is evidence that this may be the case for some genetic variants implicated in LC-PUFA metabolism [25], there is no strong evidence that maternal genotypes with regards to the particular SNPs that we studied are associated with offspring's IQ or that they interact with breastfeeding [14]. It is also possible that there are epistatic relationships between genes implicated in this pathway, so that focusing only on two variants in a single gene may not capture the whole complexity of the interplay between genetic influences in LC-PUFA levels, breastfeeding and cognitive development.

Although our primary findings were not supportive of an interaction between breastfeeding and *FADS2* polymorphisms, random effects meta-regression results suggest that such interaction exist, with studies with longer average breastfeeding

duration generally presenting estimates in accordance with the nutritional adequacy hypothesis. This should be investigated in future studies comparing different categories of breastfeeding duration, rather than simply never vs. ever comparisons (or other categorisations used here). Since such analysis would involve many subgroupings, the best alternative is likely to perform such analysis in a large dataset of individual-level data, which may be achieved by a consortium-based effort such as this collaborative meta-analysis. This and other future investigations will be important to further refine our understanding on the role of LC-PUFAs on the association between breastfeeding and intelligence. This will also have more practical implications, such as identifying whether current breastfeeding recommendations allow achieving the upper limit of cognitive benefits related to LC-PUFAs intake (if such limit exists), and the potential benefits (if any) of supplementing a lactating infant with LC-PUFAs.

Funding

This work was supported by several funding agencies – see the Supplementary Material for study-specific funders and grant numbers. This work was coordinated by researchers working within the Medical Research Council (MRC) Integrative Epidemiology Unit, which is fund by the MRC and the University of Bristol [MC_UU_12013/1, MC_UU_12013/9].

Acknowledgements

We are thankful to all participants in and funders of the studies included in this meta-analysis. See the Supplementary Material for study-specific acknowledgements.

References

1. Victora CG, Bahl R, Barros AJ *et al.* Breastfeeding in the 21st century: epidemiology, mechanisms, and lifelong effect. *Lancet* 2016; 387(10017):475-490.
2. Horta BL, Loret de Mola C, Victora CG. Breastfeeding and intelligence: a systematic review and meta-analysis. *Acta Paediatr* 2015; 104(467):14-19.

3. Walfisch A, Sermer C, Cressman A, Koren G. Breast milk and cognitive development--the role of confounders: a systematic review. *BMJ Open* 2013; 3(8):e003259.
4. Ritchie SJ. Publication bias in a recent meta-analysis on breastfeeding and IQ. *Acta Paediatr* 2016.
5. Dewey KG, Cohen RJ, Brown KH, Rivera LL. Effects of exclusive breastfeeding for four versus six months on maternal nutritional status and infant motor development: results of two randomized trials in Honduras. *J Nutr* 2001; 131(2):262-267.
6. Kramer MS, Aboud F, Mironova E *et al*. Breastfeeding and child cognitive development: new evidence from a large randomized trial. *Arch Gen Psychiatry* 2008; 65(5):578-584.
7. Brion MJ, Lawlor DA, Matijasevich A *et al*. What are the causal effects of breastfeeding on IQ, obesity and blood pressure? Evidence from comparing high-income with middle-income cohorts. *Int J Epidemiol* 2011; 40(3):670-680.
8. Lucas A, Morley R, Cole TJ, Lister G, Leeson-Payne C. Breast milk and subsequent intelligence quotient in children born preterm. *Lancet* 1992; 339(8788):261-264.
9. Jiao J, Li Q, Chu J, Zeng W, Yang M, Zhu S. Effect of n-3 PUFA supplementation on cognitive function throughout the life span from infancy to old age: a systematic review and meta-analysis of randomized controlled trials. *Am J Clin Nutr* 2014; 100(6):1422-1436.
10. Qawasmi A, Landeros-Weisenberger A, Bloch MH. Meta-analysis of LCPUFA supplementation of infant formula and visual acuity. *Pediatrics* 2013; 131(1):e262-272.
11. Cetin I, Koletzko B. Long-chain omega-3 fatty acid supply in pregnancy and lactation. *Curr Opin Clin Nutr Metab Care* 2008; 11(3):297-302.

12. Innis SM. Dietary (n-3) fatty acids and brain development. *J Nutr* 2007; 137(4):855-859.
13. Innis SM. Dietary omega 3 fatty acids and the developing brain. *Brain Res* 2008; 1237:35-43.
14. Steer CD, Davey Smith G, Emmett PM, Hibbeln JR, Golding J. FADS2 polymorphisms modify the effect of breastfeeding on child IQ. *PLoS One* 2010; 5(7):e11570.
15. Caspi A, Williams B, Kim-Cohen J *et al.* Moderation of breastfeeding effects on the IQ by genetic variation in fatty acid metabolism. *Proc Natl Acad Sci U S A* 2007; 104(47):18860-18865.
16. Martin NW, Benyamin B, Hansell NK *et al.* Cognitive function in adolescence: testing for interactions between breast-feeding and FADS2 polymorphisms. *J Am Acad Child Adolesc Psychiatry* 2011; 50(1):55-62 e54.
17. Groen-Blokhuis MM, Franic S, van Beijsterveldt CE *et al.* A prospective study of the effects of breastfeeding and FADS2 polymorphisms on cognition and hyperactivity/attention problems. *Am J Med Genet B Neuropsychiatr Genet* 2013; 162B(5):457-465.
18. Rizzi TS, van der Sluis S, Derom C *et al.* FADS2 Genetic Variance in Combination with Fatty Acid Intake Might Alter Composition of the Fatty Acids in Brain. *PLoS One* 2013; 8(6):e68000.
19. Sprecher H. Metabolism of highly unsaturated n-3 and n-6 fatty acids. *Biochim Biophys Acta* 2000; 1486(2-3):219-231.
20. Nakamura MT, Nara TY. Structure, function, and dietary regulation of delta6, delta5, and delta9 desaturases. *Annu Rev Nutr* 2004; 24:345-376.
21. Schaeffer L, Gohlke H, Muller M *et al.* Common genetic variants of the FADS1 FADS2 gene cluster and their reconstructed haplotypes are associated with the fatty acid composition in phospholipids. *Hum Mol Genet* 2006; 15(11):1745-1756.

22. Tanaka T, Shen J, Abecasis GR *et al.* Genome-wide association study of plasma polyunsaturated fatty acids in the INCHIANTI Study. *PLoS Genet* 2009; 5(1):e1000338.
23. Bisgaard H, Stokholm J, Chawes BL *et al.* Fish Oil-Derived Fatty Acids in Pregnancy and Wheeze and Asthma in Offspring. *N Engl J Med* 2016; 375(26):2530-2539.
24. Steer CD, Hibbeln JR, Golding J, Davey Smith G. Polyunsaturated fatty acid levels in blood during pregnancy, at birth and at 7 years: their associations with two common FADS2 polymorphisms. *Hum Mol Genet* 2012; 21(7):1504-1512.
25. Morales E, Bustamante M, Gonzalez JR *et al.* Genetic variants of the FADS gene cluster and ELOVL gene family, colostrums LC-PUFA levels, breastfeeding, and child cognition. *PLoS One* 2011; 6(2):e17181.
26. Hartwig FP, Davies NM, Horta BL, Victora CG, Davey Smith G. Effect modification of FADS2 polymorphisms on the association between breastfeeding and intelligence: protocol for a collaborative meta-analysis. *BMJ Open* 2016; 6(6):e010067.
27. Victora CG, Barros FC. Cohort profile: the 1982 Pelotas (Brazil) birth cohort study. *Int J Epidemiol* 2006; 35(2):237-242.
28. Horta BL, Gigante DP, Goncalves H *et al.* Cohort Profile Update: The 1982 Pelotas (Brazil) Birth Cohort Study. *Int J Epidemiol* 2015; 44(2):441, 441a-441e.
29. Fraser A, Macdonald-Wallis C, Tilling K *et al.* Cohort Profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *Int J Epidemiol* 2013; 42(1):97-110.
30. Bisgaard H, Vissing NH, Carson CG *et al.* Deep phenotyping of the unselected COPSAC2010 birth cohort study. *Clin Exp Allergy* 2013; 43(12):1384-1394.
31. Thyssen AH, Rasmussen MA, Kreiner-Moller E *et al.* Season of birth shapes neonatal immune function. *J Allergy Clin Immunol* 2016; 137(4):1238-1246 e1231-1213.

32. Jaddoe VW, van Duijn CM, van der Heijden AJ *et al.* The Generation R Study: design and cohort update 2010. *Eur J Epidemiol* 2010; 25(11):823-841.
33. Jaddoe VW, van Duijn CM, Franco OH *et al.* The Generation R Study: design and cohort update 2012. *Eur J Epidemiol* 2012; 27(9):739-756.
34. Kruithof CJ, Kooijman MN, van Duijn CM *et al.* The Generation R Study: Biobank update 2015. *Eur J Epidemiol* 2014; 29(12):911-927.
35. Guxens M, Ballester F, Espada M *et al.* Cohort Profile: the INMA--Infancia y Medio Ambiente--(Environment and Childhood) Project. *Int J Epidemiol* 2012; 41(4):930-940.
36. Newnham JP, Evans SF, Michael CA, Stanley FJ, Landau LI. Effects of frequent ultrasound during pregnancy: a randomised controlled trial. *Lancet* 1993; 342(8876):887-891.
37. Williams LA, Evans SF, Newnham JP. Prospective cohort study of factors influencing the relative weights of the placenta and the newborn infant. *BMJ* 1997; 314(7098):1864-1868.
38. Evans S, Newnham J, MacDonald W, Hall C. Characterisation of the possible effect on birthweight following frequent prenatal ultrasound examinations. *Early Hum Dev* 1996; 45(3):203-214.
39. Madsen AL, Schack-Nielsen L, Larnkjaer A, Molgaard C, Michaelsen KF. Determinants of blood glucose and insulin in healthy 9-month-old term Danish infants; the SKOT cohort. *Diabet Med* 2010; 27(12):1350-1357.
40. Jensen SM, Ritz C, Ejlerskov KT, Molgaard C, Michaelsen KF. Infant BMI peak, breastfeeding, and body composition at age 3 y. *Am J Clin Nutr* 2015; 101(2):319-325.
41. Andersen LB, Pipper CB, Trolle E *et al.* Maternal obesity and offspring dietary patterns at 9 months of age. *Eur J Clin Nutr* 2015; 69(6):668-675.

42. Andersen LB, Molgaard C, Michaelsen KF, Carlsen EM, Bro R, Pipper CB. Indicators of dietary patterns in Danish infants at 9 months of age. *Food Nutr Res* 2015; 59:27665.
43. Pausova Z, Paus T, Abrahamowicz M *et al.* Genes, maternal smoking, and the offspring brain and body during adolescence: design of the Saguenay Youth Study. *Hum Brain Mapp* 2007; 28(6):502-518.
44. Paus T, Pausova Z, Abrahamowicz M *et al.* Saguenay Youth Study: a multi-generational approach to studying virtual trajectories of the brain and cardio-metabolic health. *Dev Cogn Neurosci* 2015; 11:129-144.
45. Sudlow C, Gallacher J, Allen N *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015; 12(3):e1001779.
46. Keller MC. Gene x environment interaction studies have not properly controlled for potential confounders: the problem and the (simple) solution. *Biol Psychiatry* 2014; 75(1):18-24.
47. Hoddinott P, Tappin D, Wright C. Breast feeding. *BMJ* 2008; 336(7649):881-887.
48. World Health Organization and UNICEF. Protecting, Promoting and Supporting Breastfeeding: The Special Role of Maternity Services. Geneva, Switzerland 1989.
49. World Health Organization. The Optimal Duration of Exclusive Breastfeeding. Geneva, Switzerland: World Health Organization; 2001.
50. Greenland S, Schlesselman JJ, Criqui MH. The fallacy of employing standardized regression coefficients and correlations as measures of effect. *Am J Epidemiol* 1986; 123(2):203-208.
51. Greenland S, Maclure M, Schlesselman JJ, Poole C, Morgenstern H. Standardized regression coefficients: a further critique and review of some alternatives. *Epidemiology* 1991; 2(5):387-392.

Tables

Table 1. Meta-analytical linear regression coefficients (β) of cognitive measures (in standard deviation units) according to breastfeeding (never=0; ever=1), within strata of *FADS2* rs174575 or rs1513 genotypes (recessive effect).

Model	Statistic	Fixed effects			Random effects		
		FADS2		G×E	FADS2		G×E
		Other genotypes	GG		Other genotypes	GG	
rs174575 (CC or CG=0; GG=1)							
Unadjusted	<i>I</i> ²	-	-	-	76.4	64.4	77.6
N _{estimates} =8	P-value	8.6×10 ⁻⁵⁰	3.8×10 ⁻⁸	0.188	7.6×10 ⁻⁷	0.002	0.323
N _{subjects} =12,614	β	0.37	0.43	0.11	0.29	0.43	0.18
	95% CI	0.32; 0.41	0.28; 0.58	-0.05; 0.27	0.17; 0.40	0.16; 0.70	-0.18; 0.54
Adjusted (1) ^a	<i>I</i> ²	-	-	-	74.2	67.2	75.5
N _{estimates} =8	P-value	7.7×10 ⁻⁴⁸	9.3×10 ⁻⁷	0.603	7.9×10 ⁻⁷	0.024	0.705
N _{subjects} =12,590	β	0.37	0.39	0.04	0.29	0.35	0.07
	95% CI	0.32; 0.42	0.23; 0.54	-0.12; 0.21	0.18; 0.41	0.04; 0.65	-0.29; 0.43
Adjusted (2) ^b	<i>I</i> ²	-	-	-	84.1	47.4	59.5
N _{estimates} =8	P-value	6.4×10 ⁻²⁰	6.4×10 ⁻⁵	0.244	0.055	0.020	0.445
N _{subjects} =12,077	β	0.25	0.34	0.10	0.15	0.31	0.12
	95% CI	0.20; 0.31	0.17; 0.51	-0.07; 0.28	0.00; 0.31	0.05; 0.58	-0.19; 0.43
rs1535 (AA or AG=0; GG=1)							
Unadjusted	<i>I</i> ²	-	-	-	73.5	54.1	42.6
N _{estimates} =9	P-value	9.2×10 ⁻⁴⁹	2.2×10 ⁻⁶	0.663	4.6×10 ⁻⁷	0.013	0.646
N _{subjects} =13,202	β	0.37	0.29	-0.03	0.29	0.24	-0.04
	95% CI	0.32; 0.42	0.17; 0.41	-0.16; 0.10	0.18; 0.40	0.05; 0.43	-0.24; 0.15
Adjusted (1) ^a	<i>I</i> ²	-	-	-	76.0	47.7	60.9
N _{estimates} =9	P-value	9.9×10 ⁻⁴⁷	2.2×10 ⁻⁷	0.720	7.1×10 ⁻⁶	5.4×10 ⁻³	0.778
N _{subjects} =13,175	β	0.37	0.33	-0.02	0.29	0.27	-0.03
	95% CI	0.32; 0.42	0.20; 0.45	-0.16; 0.11	0.16; 0.42	0.08; 0.47	-0.28; 0.21
Adjusted (2) ^b	<i>I</i> ²	-	-	-	84.0	25.9	49.6
N _{estimates} =9	P-value	1.9×10 ⁻¹⁹	1.2×10 ⁻⁵	0.277	0.065	0.003	0.592
N _{subjects} =12,633	β	0.26	0.28	0.07	0.15	0.25	0.06
	95% CI	0.20; 0.31	0.16; 0.41	-0.06; 0.21	-0.01; 0.32	0.09; 0.41	-0.16; 0.27

^aCovariates were sex, age (linear and quadratic terms), ancestry-informative principal components (if available) and genotyping centre (if necessary).

^bSame covariates than in the Adjusted (1) model, in addition to maternal education (linear and quadratic terms) and/or maternal cognition (linear and quadratic terms).

G×E: interaction between breastfeeding and polymorphisms in the *FADS2* gene.

$N_{\text{estimates}}$: number of estimates being pooled. N_{subjects} : pooled sample size.

Table 2. Meta-analytical linear regression coefficients (β) of the interaction term between *FADS2* rs174575 or rs1535 genotypes (recessive effect) with breastfeeding (<6 months vs. ≥ 6 months, in ordinal categories or in months), having cognitive measures (in standard deviation units) as the outcome.

Model	Statistic	Fixed effects			Random effects		
		<6 months=0 ≥6 months=1	Numerically- coded categories	Months	<6 months=0 ≥6 months=1	Numerically- coded categories	Months
rs174575 (CC or CG=0; GG=1)							
Unadjusted	I^2	-	-	-	23.1	57.1	13.9
N _{estimates} =8	P-value	0.515	0.104	0.371	0.647	0.150	0.335
N _{subjects} =11,733	β	0.05	0.04	0.01	0.04	0.06	0.01
	95% CI	-0.10; 0.20	-0.01; 0.09	-0.01; 0.02	-0.14; 0.22	-0.02; 0.15	-0.01; 0.03
Adjusted (1) ^a	I^2	-	-	-	53.6	58.7	63.3
N _{estimates} =8	P-value	0.378	0.189	0.608	0.546	0.282	0.635
N _{subjects} =11,706	β	0.07	0.04	0.00	0.08	0.06	0.01
	95% CI	-0.09; 0.23	-0.02; 0.09	-0.01; 0.02	-0.18; 0.35	-0.05; 0.16	-0.02; 0.04
Adjusted (2) ^b	I^2	-	-	-	82.6	84.6	85.3
N _{estimates} =8	P-value	0.244	0.132	0.782	0.496	0.346	0.602
N _{subjects} =11,242	β	0.10	0.04	0.00	0.17	0.09	0.01
	95% CI	-0.07; 0.26	-0.01; 0.10	-0.01; 0.02	-0.32; 0.65	-0.09; 0.26	-0.04; 0.07
rs1535 (AA or AG=0; GG=1)							
Unadjusted	I^2	-	-	-	0.0	0.0	0.0
N _{estimates} =8	P-value	0.460	0.966	0.805	0.460	0.966	0.805
N _{subjects} =12,018	β	-0.05	0.00	0.00	-0.05	0.00	0.00
	95% CI	-0.17; 0.08	-0.04; 0.04	-0.01; 0.01	-0.17; 0.08	-0.04; 0.04	-0.01; 0.01
Adjusted (1) ^a	I^2	-	-	-	8.0	54.3	59.6
N _{estimates} =8	P-value	0.248	0.508	0.538	0.302	0.635	0.330
N _{subjects} =11,991	β	-0.07	-0.01	0.00	-0.07	-0.02	-0.01
	95% CI	-0.20; 0.05	-0.06; 0.03	-0.01; 0.01	-0.20; 0.06	-0.09; 0.05	-0.03; 0.01
Adjusted (2) ^b	I^2	-	-	-	3.9	29.9	35.5
N _{estimates} =8	P-value	0.194	0.675	0.320	0.216	0.728	0.344
N _{subjects} =11,499	β	-0.08	-0.01	-0.01	-0.08	-0.01	-0.01
	95% CI	-0.21; 0.04	-0.05; 0.03	-0.02; 0.01	-0.21; 0.05	-0.07; 0.05	-0.02; 0.01

^aCovariates were sex, age (linear and quadratic terms), ancestry-informative principal components (if available) and genotyping centre (if necessary).

^bSame covariates than in the Adjusted (1) model, in addition to maternal education (linear and quadratic terms) and/or maternal cognition (linear and quadratic terms).

$N_{\text{estimates}}$: number of estimates being pooled. N_{subjects} : pooled sample size.

Table 3. Stratified random effects meta-analytical linear regression coefficients (β) of the interaction term between *FADS2* rs174575 or rs1535 genotypes (recessive effect) with breastfeeding (never=0; ever=1), having cognitive measures (in standard deviation units) as the outcome. Estimates from the fully adjusted model were used.

Variable	Categories	rs174575 (CC or CG=0; GG=1)				rs1535 (AA or AG=0; GG=1)			
		N _{estimates} (N _{subjects})	β (95% CI)	P-value	Adjusted R ² (%)	N _{estimates} (N _{subjects})	β (95% CI)	P-value	Adjusted R ² (%)
IQ test	Wechsler ^a	8055 (4)	0.12 (-0.32; 0.56)	0.591	0.0	8070 (4)	0.09 (-0.14; 0.32)	0.452	0.0
	Other	4022 (4)	0.12 (-0.37; 0.61)	0.631		4563 (5)	0.02 (-0.45; 0.49)	0.932	
Adjustment for PCs	Yes	10,441 (6)	0.28 (0.02; 0.54)	0.036	88.0	10753 (7)	0.09 (-0.19; 0.37)	0.531	0.0
	No	1636 (2)	-0.38 (-0.72; -0.04)	0.028		1880 (2)	-0.03 (-0.32; 0.25)	0.814	
Age at IQ Measurement	≥10 years	4373 (4)	0.06 (-0.46; 0.58)	0.825	0.0 ^b ; 10.4 ^c	4374 (4)	0.04 (-0.25; 0.34)	0.773	0.0 ^b ; 0.0 ^c
	<10 years	7704 (4)	0.20 (-0.18; 0.58)	0.304		8259 (5)	0.07 (-0.31; 0.45)	0.700	
BF measurement	Prospective	6912 (3)	0.27 (-0.10; 0.63)	0.155	0.0	6926 (3)	0.20 (-0.25; 0.64)	0.383	0.0
	Retrospective	5165 (5)	-0.01 (-0.48; 0.47)	0.979		5707 (6)	-0.01 (-0.28; 0.27)	0.951	
Continental Region	Europe	7704 (4)	0.20 (-0.18; 0.58)	0.304	0.0	8259 (5)	0.07 (-0.31; 0.45)	0.700	0.0
	Other	4373 (4)	0.06 (-0.46; 0.58)	0.825		4374 (4)	0.04 (-0.25; 0.34)	0.773	
Mean year of Birth	≥2000	3002 (3)	0.20 (-0.58; 0.98)	0.616	0.0 ^b ; 2.9 ^c	3543 (4)	0.03 (-0.62; 0.69)	0.917	0.0 ^b ; 0.0 ^c
	<2000	9075 (5)	0.10 (-0.27; 0.46)	0.601		9090 (5)	0.07 (-0.13; 0.27)	0.469	
Prevalence of any BF (%)	≥90	4798 (4)	0.36 (-0.19; 0.90)	0.200	16.4 ^b ; 72.3 ^c	5339 (5)	0.15 (-0.31; 0.62)	0.519	0.0 ^b ; 8.3 ^c
	<90	7279 (4)	-0.04 (-0.38; 0.29)	0.803		7294 (4)	0.01 (-0.15; 0.18)	0.869	
Duration of any BF (months)	≥5.85	3367 (3)	0.50 (-0.06; 1.06)	0.081	45.5 ^b ; 0.0 ^c	3665 (4)	0.27 (-0.28; 0.82)	0.333	22.2 ^b ; 4.9 ^c
	<5.85	7866 (4)	0.14 (-0.10; 0.38)	0.255		8123 (4)	-0.01 (-0.19; 0.16)	0.882	
Sample size (N)	≥1000	9177 (4)	0.26 (0.00; 0.52)	0.052	33.8 ^b ; 0.0 ^c	9191 (4)	0.11 (-0.12; 0.34)	0.365	0.0 ^b ; 0.0 ^c
	<1000	2900 (4)	-0.03 (-0.63; 0.56)	0.910		3442 (5)	0.01 (-0.43; 0.45)	0.974	

^aIncludes both Wechsler Adult Intelligence Scale (ALSPAC and Dunedin Multidisciplinary Health and Development Study) and Wechsler Intelligence Scale for Children (1982 Pelotas Birth Cohort and Saguenay Youth Study). ^bVariable categorised as shown in the table. ^cVariable entered in continuous form (e.g., age at outcome measurement modelled in years, as a continuous variable). PCs: ancestry-informative genetic principal components. BF: breastfeeding. N: number of. CI: confidence interval. N_{estimates}: number of estimates being pooled. N_{subjects}: pooled sample size.

Table 4. Meta-analytical linear regression coefficients (β) of the interaction term between *FADS2* rs174575 or rs1535 genotypes (recessive effect) with breastfeeding (never=0; ever=1), having maternal education (in complete years) or maternal cognitive measures (in standard deviation units) as the outcome.

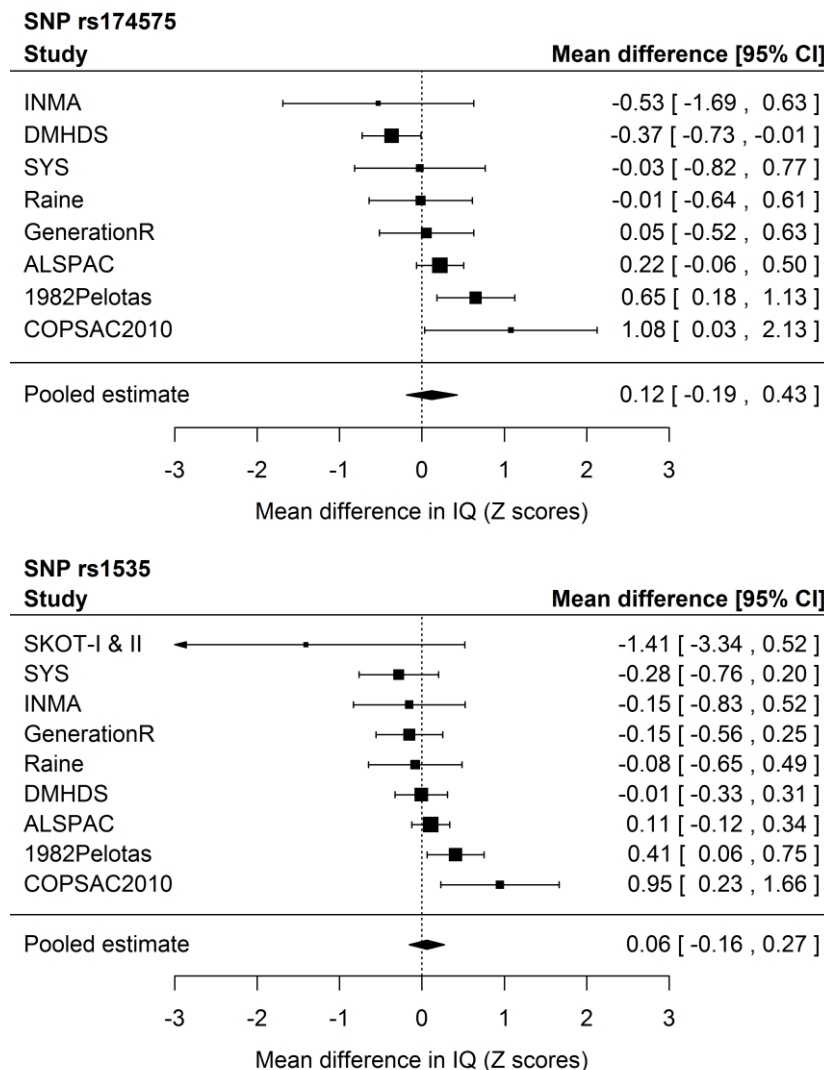
Model	Statistic	Fixed effects		Random effects	
		Maternal education	Maternal cognition	Maternal education	Maternal cognition
rs174575 (CC or CG=0; GG=1)					
Unadjusted	N _{estimates}	7	5	7	5
	N _{subjects}	14,671	6299	14671	6299
	I ²	-	-	81.1	18.1
	P-value	0.159	0.326	0.375	0.389
	β	0.28	0.10	0.59	0.10
	95% CI	-0.11; 0.66	-0.10; 0.31	-0.72; 1.91	-0.13; 0.33
Adjusted (1) ^a	N _{estimates}	7	5	7	5
	N _{subjects}	12,113	6126	12113	6126
	I ²	-	-	14.1	0.0
	P-value	0.509	0.854	0.607	0.854
	β	0.16	-0.02	0.16	-0.02
	95% CI	-0.31; 0.62	-0.25; 0.21	-0.45; 0.78	-0.25; 0.21
rs1535 (AA or AG=0; GG=1)					
Unadjusted	N _{estimates}	8	5	8	5
	N _{subjects}	15,447	6556	15447	6556
	I ²	-	-	1.4	0.0
	P-value	0.784	0.272	0.814	0.272
	β	-0.05	0.10	-0.04	0.10
	95% CI	-0.38; 0.28	-0.08; 0.28	-0.39; 0.31	-0.08; 0.28
Adjusted (1) ^a	N _{estimates}	8	5	8	5
	N _{subjects}	12,743	6378	12743	6378
	I ²	-	-	0.0	0.0
	P-value	0.540	0.160	0.540	0.160
	β	-0.12	0.14	-0.12	0.14
	95% CI	-0.51; 0.27	-0.05; 0.33	-0.51; 0.27	-0.05; 0.33

^aCovariates were sex, age (linear and quadratic terms), ancestry-informative principal components (if available) and genotyping centre (if necessary).

N_{estimates}: number of estimates being pooled. N_{subjects}: pooled sample size.

Figures

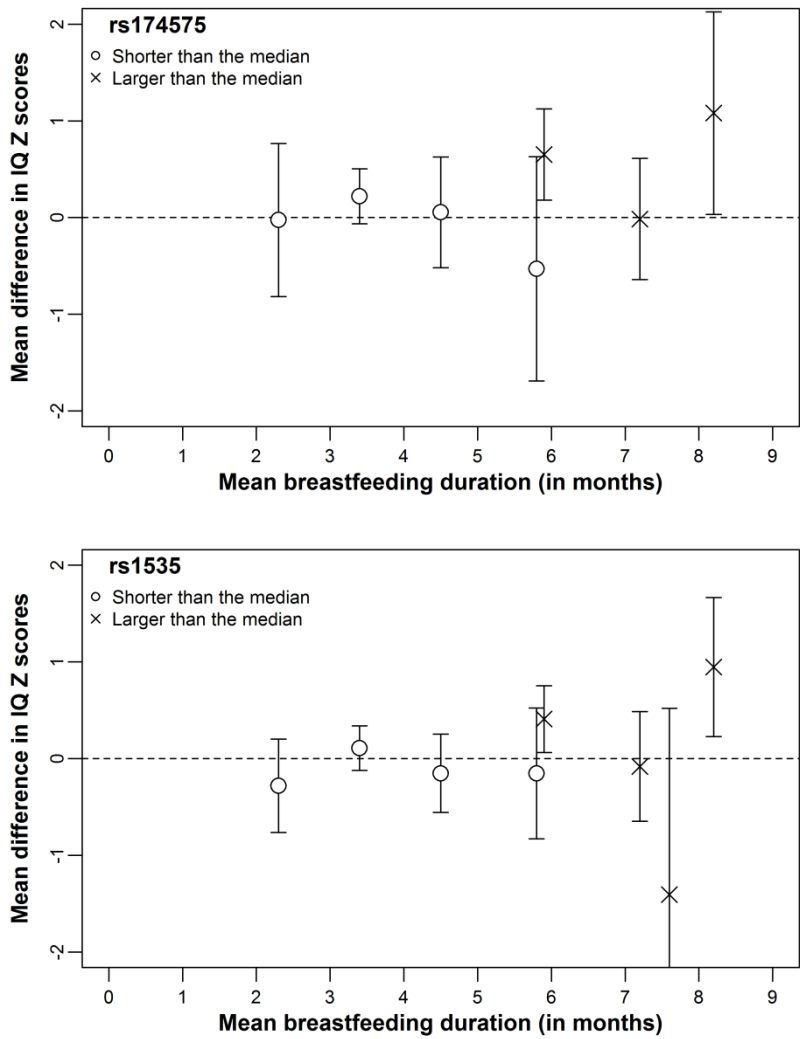
Figure 1. Forest plots of mean differences in IQ Z scores from the fully adjusted model comparing ever with never breastfed individuals based on random effects meta-analysis.



SKOT-I and SKOT-II were excluded from the analyses for the rs174575 polymorphism because the model did not fit (due to a combination of modest sample size, high prevalence of breastfeeding and assuming a recessive genetic effect of the rarest allele).

1982Pelotas: 1982 Pelotas Birth Cohort. ALSPAC: Avon Longitudinal Study of Parents and Children. COPSAC2010: Copenhagen Prospective Study on Asthma in Childhood 2010. DMHDS: Dunedin Multidisciplinary Health and Development Study. GenerationR: Generation R Study. INMA: *Infancia y Medio Ambiente* - Environment and Childhood. Raine: Western Australian Pregnancy Cohort (Raine) Study. SKOT-I & II: *Småbørn Kost Og Trivsel* (I and II). SYS: Saguenay Youth Study.

Figure 2. Scatter plots of mean differences (with 95% confidence intervals) in IQ Z scores from the fully adjusted model comparing ever with never breastfed individuals according to prevalence (%) of ever breastfeeding and average breastfeeding duration in months.



Supplementary Material

Summary

Supplementary Methods..... 235

Study Acknowledgements 237

Supplementary Tables 243

Supplementary Methods

Modifications in the study protocol

After the publication of the protocol, some revisions in the analysis plan were necessary. These were performed after evaluating descriptive statistics, but before pooling study-level regression coefficients. In addition to the inclusion of a study (UK Biobank) that did not achieve all eligibility criteria (as explained in the main text), the revisions were:

i) Combination of SKOT-I and SKOT-II into a single study.

SKOT-I and SKOT-II were the studies with the smallest number of participants. Their main difference is that SKOT-II included only obese (pre-pregnancy BMI > 30 kg/m²) mothers. Due to the small number of participants (likely accentuated by the very high prevalence of breastfeeding), in some analysis the model failed to converge, thus preventing these studies from contributing. To overcome this, both studies were combined into a single sample.

ii) Re-definition of never being breastfed.

In two studies (COPSAC 2010 and SKOT-I & II) the prevalence of never being breastfed was <1% (Supplementary Table 3). This was an issue especially because these studies were not large (551 and 299 individuals, respectively), the analyses involve fitting an interaction term, and the primary analysis assumes a recessive effect of the rarest allele. Therefore, in those studies, the binary variables of never vs. ever breastfeeding (for both any quality and exclusive) were re-defined as follows: 0: never breastfed or breastfed for less than 1 month; 1: breastfed for at least 1 month.

iii) Re-definition of exclusive breastfeeding.

Data on exclusive breastfeeding was unavailable in the 1982 Pelotas Birth Cohort, INMA, RAINE and SKOT-I & II studies. Those used predominant breastfeeding instead.

iv) Exclusion of the ordinal breastfeeding (for both any quality and exclusive) variable.

In the study protocol, one of the breastfeeding variables was an ordinal variable coded as follows: 0: none; 1: 0.01-1.00 months; 2: 1.01-3.00 months; 3: 3.01-6.00 months; 4: >6.00 months. After evaluating descriptive statistics, we noted that some categories (especially regarding exclusive breastfeeding) had very few individuals (Supplementary Table 3). Information on breastfeeding duration was not available for one eligible study (nor for the UK Biobank subsample), and in only two of the remaining studies the median of breastfeeding duration was at least six months. This was an issue due to the same reasons explained above, so we opted for removing this variable. However, this same variable coded numerically (i.e., assuming a linear trend) was maintained.

v) Exclusion of the exclusive breastfeeding dichotomised into <6 months vs. ≥6 months.

The pre-planned analyses described in the protocol included analyses of exclusive breastfeeding in four different categorisations: never vs. ever; ordinal variable of breastfeeding duration (coded assuming a linear effect); exclusive breastfeeding duration, in months; and <6 vs. ≥6 months. However, in studies with information on exclusive breastfeeding duration, fewer than 2% of all the children (Supplementary Table 3) were breastfed exclusively for more than 6 months. This was an issue due to the same reasons explained above, so we opted for removing this variable.

vi) Additional moderators in meta-regression analysis.

In the study protocol, it was specified that the following variables would be studied as moderators in meta-regression analyses: IQ test, adjustment for ancestry-informative principal components, age when IQ was measured, timing of breastfeeding measurement, continental region, prevalence of having ever being breastfed and mean breastfeeding duration. After publishing the protocol, we decided to also include average year of birth of study participants and sample size of each study.

vii) Not adjusting for maternal cognition in the ALSPAC study.

Apart from the UK Biobank, ALSPAC was the largest study included in this meta-analysis, with >4700 individuals in the unadjusted model of the primary analysis. Both

maternal education and maternal cognition were available, but the latter was measured in less than 2000 of the individuals included in the primary analysis. To avoid such substantial sample size loss and given that education is highly correlated with cognitive measures, we opted for adjusting ALSPAC estimates only for maternal education in the “adjusted 2” model (in addition to the covariates adjusted for in the “adjusted 1” model). However, ALSPAC still contributed to the sensitivity analysis that had maternal cognition as the outcome variable.

Study Acknowledgements

1982 Pelotas Birth Cohort Study

The 1982 Pelotas Birth Cohort Study is conducted by the Postgraduate Program in Epidemiology at Federal University of Pelotas (*Universidade Federal de Pelotas*) in collaboration with the Brazilian Public Health Association (ABRASCO). From 2004 to 2013, the Wellcome Trust supported the study. The International Development Research Center, World Health Organization, Overseas Development Administration, European Union, National Support Program for Centers of Excellence (PRONEX), the Brazilian National Research Council (CNPq), and the Brazilian Ministry of Health supported previous phases of the study.

Genotyping was supported by the Department of Science and Technology (DECIT, Ministry of Health) and National Fund for Scientific and Technological Development (FNDCT, Ministry of Science and Technology), Funding of Studies and Projects (FINEP, Ministry of Science and Technology, Brazil), Coordination of Improvement of Higher Education Personnel (CAPES, Ministry of Education, Brazil).

More information about the 1982 Pelotas Birth Cohort Study are available in cohort profile papers by Victora and Barros (PMID: 16373375), and by Horta et al. (PMID: 25733577).

Avon Longitudinal Study of Parents and Children (ALSPAC)

We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. The UK Medical Research Council and Wellcome (Grant ref: 102215/2/13/2) and the University of Bristol provide core support for ALSPAC. This publication is the work of the authors and Fernando Pires Hartwig will serve as guarantors for the contents of this paper. Genetic data was generated by Sample Logistics and Genotyping Facilities at Wellcome Sanger Institute and LabCorp (Laboratory Corporation of America) using support from 23andMe.

Copenhagen Prospective Study on Asthma in Childhood (COPSAC) 2010

We greatly acknowledge the private and public research funding allocated to COPSAC and listed on www.copsac.com, with special thanks to The Lundbeck Foundation (Grant nr. R16-A1694); Ministry of Health (Grant nr. 903516); Danish Council for Strategic Research (Grant nr.: 0603-00280B); The Danish Council for Independent Research and The Capital Region Research Foundation as core supporters. The funding agencies did not have any influence on study design, data collection and analysis, decision to publish or preparation of the manuscript. No pharmaceutical company was involved in the study. We gratefully express our gratitude to the participants of the COPSAC 2010 study for all their support and commitment. We also acknowledge and appreciate the unique efforts of the COPSAC research team.

Dunedin Multidisciplinary Health and Development Study

The Dunedin Multidisciplinary Health and Development Research Unit is funded by the New Zealand Health Research Council and the New Zealand Ministry of Business, Innovation and Employment (MBIE). Research was supported by grants from the National Institute on Aging (AG032282), National Institute of Child Health and Development (HD077482), and Medical Research Council (MR/P005918/1). We thank the Dunedin Study founder Phil Silva. More information about the Dunedin Study is available in a cohort profile paper by Poulton, Moffitt, and Silva (PMID: 25835958).

Generation R Study

The Generation R Study is conducted by researchers at the Erasmus Medical Center in close collaboration with the School of Law and Faculty of Social Sciences of the Erasmus University Rotterdam; the Municipal Health Service for the Rotterdam area; the Rotterdam Homecare Foundation; and the Stichting Trombosedienst and Artsenlaboratorium Rotterdam. We gratefully acknowledge the contributions of general practitioners, hospitals, midwives and pharmacies in Rotterdam. The Generation R Study is made possible by financial support from the Erasmus Medical Center, Rotterdam, the Erasmus University Rotterdam, the Netherlands Organization for Health Research and Development (ZonMw), the Netherlands Organisation for Scientific Research (NWO), and the Ministry of Health, Welfare and Sport. H.T. received additional grants from the Netherlands Organization for Health Research and Development (ZonMw VIDI 017.106.370).

More information about The Generation R Study is available in cohort profile paper by Jaddoe et al. (PMID: 20967563).

INMA (*Infancia y Medio Ambiente* – Environment and Childhood)

Population-based birth cohorts were established as part of the *Infancia y Medio Ambiente* (INMA) Project in several regions of Spain following a common protocol. The present analysis uses the INMA subcohorts of Menorca, Valencia, Sabadell, and Gipuzkoa. More information about the INMA project is available in a cohort profile paper (PMID: 21471022), and in the INMA webpage (<http://www.proyectoinma.org/>).

This study was funded by grants from Instituto de Salud Carlos III [G03/176, CB06/02/0041, 97/0588, 00/0021-2, FIS PI041436, PI06/0867, PI061756, PI081151, PI041705, and PS09/00432, PS09/00090, PS0901958, FIS-FEDER 03/1615, 04/1509, 04/1112, 04/1931, 05/1079, 05/1052, 06/1213, 07/0314, 09/02647, 11/0178, 11/02591, 11/02038, 13/1944, 13/2032, 14/0891, and 14/1687, PI14/00677 incl. FEDER funds], Spanish Ministry of Science and Innovation [SAF2008-00357], European Commission [ENGAGE project and grant agreement HEALTH-F4-2007-201413, FP7-ENV-2011 cod 282957 and HEALTH.2010.2.4.5-1], CIBERESP, Fundació La Marató de TV3 (090430), Generalitat de Catalunya-CIRIT 1999SGR 00241, Beca de la IV

convocatoria de Ayudas a la Investigación en Enfermedades Neurodegenerativas de La Caixa, and EC Contract No. QLK4-CT-2000-00263, Conselleria de Sanitat Generalitat Valenciana, Department of Health of the Basque Government (2005111093 and 2009111069) and the Provincial Government of Gipuzkoa (DFG06/004 and DFG08/001), and Fundación Roger Torné.

The authors would particularly like to thank all the participants for their generous collaboration. The authors are grateful to Silvia Fochs, Anna Sànchez, Maribel López, Nuria Pey, Muriel Ferrer, Amparo Quiles, Sandra Pérez, Gemma León, Elena Romero, and Amparo Cases for their assistance in contacting the families and administering the questionnaires. A full roster of the INMA Project Investigators can be found at <http://www.proyecto-inma.org/presentacion-inma/listado-investigadores/enlistado-investigadores.html>.

Saguenay Youth Study

We thank all families who took part in the Saguenay Youth Study and the following individuals for their contributions in designing the protocol, acquiring and analyzing the data: psychometricians (Chantale Belleau, Mélanie Drolet, Catherine Harvey, Stéphane Jean, Hélène Simard, Mélanie Tremblay, Patrick Vachon), ÉCOBES team (Nadine Arbour, Julie Auclair, Marie-Ève Blackburn, Marie-Ève Bouchard, Annie Gautier, Annie Houde, Catherine Lavoie), laboratory technicians (Denise Morin and Nadia Mior), nutritionists (Caroline Benoit and Henriette Langlais), MRI team (Sylvie Masson, Suzanne Castonguay, Marie-Josée Morin, Caroline Mérette), and cardio nurses (Jessica Blackburn, Mélanie Gagné, Jeannine Landry, Catherine Lavoie, Lisa Pageau, Réjean Savard, France Tremblay, Jacynthe Tremblay). We thank Dr. Jean Mathieu for the medical follow up of participants in who we detected any medically relevant abnormalities. We thank Manon Bernard for designing and managing our online database. We thank Dr. Jean Shin for her statistical advice.

The Saguenay Youth Study has been funded by the Canadian Institutes of Health Research (TP, ZP), Heart and Stroke Foundation of Canada (ZP), and the Canadian Foundation for Innovation (ZP). Computations were performed on the GPC

supercomputer at the SciNet HPC Consortium. SciNet is funded by: the Canada Foundation for Innovation under the auspices of Compute Canada; the Government of Ontario; Ontario Research Fund - Research Excellence; and the University of Toronto.

Småbørn Kost Og Trivsel (SKOT)-I and SKOT-II

We gratefully acknowledge the contribution of all the families and children who participate in the study. SKOT-I was funded by the Danish Directorate for Food, Fisheries and Agricultural Business as part of the project Complementary and Young Child Feeding (CYCF) – Impact on Short- and Long-Term Development and Health. SKOT-II was partially funded by grants from Aase and Ejnar Danielsens Foundation and Augustinus Foundation and further funding was provided by the research program “Governing Obesity” funded by the University of Copenhagen Excellence Program for Interdisciplinary Research (<http://www.go.ku.dk>). The Novo Nordisk Foundation Center for Basic Metabolic Research is an independent research center at the University of Copenhagen partially funded by an unrestricted donation from the Novo Nordisk Foundation (www.metabol.ku.dk).

The SKOT-I and SKOT-II cohorts were initiated by Kim F. Michaelsen and Lotte Lauritzen initiated genotyping of *FADS2* polymorphisms in these studies. The actual genotyping was performed by Theresia M. Schnurr under supervision of Torben Hansen.

More information about the SKOT-I and SKOT-II cohorts are available in previously published papers from the cohorts (PMID: 21059086 and 25646329) and (PMID: 25469467 and 26111966), respectively.

Western Australian Pregnancy Cohort (Raine) Study

The authors are grateful to the Raine Study participants and their families, and to the Raine Study Team for cohort coordination and data collection. The authors gratefully acknowledge the NH&MRC for their long term contribution to funding the study over the last 25 years and also the following Institutions for providing funding for Core Management of the Raine Study: The University of Western Australia (UWA), Raine Medical Research Foundation, UWA Faculty of Medicine, Dentistry and Health

Sciences, Telethon Kids Institute, Women and Infants Research Foundation, Curtin University and Edith Cowan University.

The authors gratefully acknowledge the assistance of the Western Australian DNA Bank (National Health and Medical Research Council of Australia National Enabling Facility). This study was supported by the National Health and Medical Research Council of Australia [grant numbers 572613 and 403981] and the Canadian Institutes of Health Research [grant number MOP-82893]. Nicole M. Warrington is supported by a National Health and Medical Research Council Early Career Fellowship (APP1104818).

This work was supported by resources provided by the Pawsey Supercomputing Centre with funding from the Australian Government and the Government of Western Australia.

Supplementary Tables

Supplementary Table 1 (Columns 1-10). Studies included in the meta-analysis.

Name	Short name	PubMed ID(s) of key paper(s) describing the study	Design	Population-based study	Country of study participants	Continental region of study participants	Mean year of birth of study participants	Multi-ethnic study	If multi-ethnic: definition of ancestry groups
1982 Pelotas Birth Cohort	1982Pelotas	16373375; 25733577	Prospective cohort	<u>Yes</u>	Brazil	South America	1982	Yes	Based on genome-wide genotyping data and reference panels from HapMap and HGDP, the software ADMIXTURE was used to estimate individual-level proportions of European, African and Native American ancestries. Individuals were classified as European if presenting at least 85% of European ancestry.
Avon Longitudinal Study of Parents and Children	ALSPAC	22507742	Prospective cohort	<u>Yes</u>	UK	Europe	1991	No	Not applicable

UK Biobank (Genotyping platform: Axiom)	Biobank_Axiom	25826379	Prospective cohort	<u>No</u> Volunteer sample.	UK	Europe	1951	No	Not applicable
UK Biobank (Genotyping platform: BiLEVE)	Biobank_BiLEVE	25826379	Prospective cohort	<u>No</u> Volunteer sample.	UK	Europe	1951	No	Not applicable
Copenhagen Prospective Study on Asthma in Childhood 2010	COPSAC2010	24118234; 26581916	Prospective cohort	<u>No</u> The study catchment area was Zealand, an island in the eastern part of Denmark, including the capital Copenhagen. Pregnant women were recruited by a monthly surveillance of reimbursement to general practitioners for the mandatory pregnancy visit. They received an invitation by posted mail to contact the clinic during 2008– 2010. Exclusion	Denmark	Europe	2010	No	Not applicable

				criteria were gestational age (at recruitment) above week 26; daily intake of more than 600 IU vitamin D during pregnancy; or having any endocrine, heart, or kidney disorders. Women who contacted the COPSAC clinic by phone received detailed verbal information. Those who were still interested and qualifying for the study received comprehensive study information by posted mail. Finally, the women attended the clinical research unit within pregnancy weeks 22–26 for a visit in the research clinic with detailed information and				

				enrolment into the pregnancy cohort.					
Dunedin Multidisciplinary Health and Development Study	DMHDS	17984066	Prospective cohort	<u>Yes</u>	New Zealand	Pacific	1972	No	Not applicable
Generation R Study	GenerationR	20967563; 23086283; 25527369	Prospective cohort	<u>Yes</u>	Netherlands	Europe	2004	Yes	To characterize the genetic ancestry of the children in the Generation R Study, all samples passing QC procedures were merged with the three genotyped panels from the HapMap Phase II release 22 build 36 including: Northwestern Europeans (CEPH collection or CEU), Sub-saharan West Africans (Yoruba or YRI) and Asians (Han Chinese from Beijing or CHB, and Japanese from Tokyo or JPT) using only independent autosomal SNPs ($r^2 > 0.05$). In the merged dataset, pairwise

INMA (Infancia y Medio Ambiente - Environment and Childhood)	INMA	21471022	Prospective cohort	<u>No</u> Criteria for inclusion of the mothers were: (i) to be resident in one of the study areas, (ii) to be at least 16 years old, (iii) to have a singleton pregnancy, (iv) to not have followed any programme of assisted reproduction, (v) to wish to deliver in the reference hospital and (vi) to have no communication problems.	Spain (from 4 regions: Menorca (Balearic islands), Valencia, Sabadell (Catalonia), Gipuzkoa)	Europe	2003	No	Not applicable
Western Australian Pregnancy Cohort (Raine) Study	Raine	8105165; 9224128; 8855394	Prospective cohort	<u>No</u> Between 1989 and 1991, 2900 pregnant women volunteered to be part of the study at King Edward Memorial Hospital looking a prenatal ultrasound scans when they were 18	Australia	Pacific	1990	No	Only individuals with at least one caucasian parent (based on response to questionnaire) were genotyped. Principal components were generated for those genotyped and individuals and the top three were included in

			weeks pregnant. Some of the mothers were followed up at 24, 28 and 38 weeks gestation. The families then continued with follow-up assessments of their babies. 2868 babies remained with the study and were examined on the first or second day after birth by a child health nurse in King Edward Memorial Hospital.					the analysis.

Småbørn Kost Og Trivsel-I	SKOT-I	21059086; 25646329	Prospective cohort	<u>No</u> Infants in SKOT-I were recruited by postal invitations to randomly selected parents of infants on the basis of extractions from the National Civil Registration System. SKOT1 required participants to be healthy singletons, born at term, with an age of 9 months ± 2 weeks at the first examination and having Danish- speaking parents.	Denmark	Europe	2007	No	Not applicable
Småbørn Kost Og Trivsel-II	SKOT-II	25469467; 26111966	Prospective cohort	<u>No</u> Participants for SKOT-II were recruited among offspring of obese pregnant women participating in the intervention study 'Treatment of Obese Pregnant Women' (TOP) at Hvidovre Hospital	Denmark	Europe	2011	No	Not applicable

				(Hvidovre, Denmark) with dietetic and physical activity counseling, followed by breastfeeding counseling for a subgroup of the participants. The inclusion criteria for SKOT II were equal to SKOT I, except that all participants were required to be offspring of women with a prepregnancy BMI>30 kg/m ² .				

Saguenay Youth Study	SYS	17469173; 25454417	Prospective cohort	<u>No</u> The SYS cohort includes 1029 adolescents and their 962 parents. The cohort was recruited via adolescents attending high schools in the Saguenay–Lac-Saint-Jean region of Quebec, Canada. The region is home to the largest genetic founder population in North America. Both maternal and paternal grandparents of the adolescents were required to be of French-Canadian ancestry and born in the region; as such, all adolescents and their parents are of a single ethnicity [European (French) ancestry]. Half of	Canada	North America	1992	No	Not applicable

the adolescents were exposed prenatally to maternal cigarette smoking. The cohort is family based (481 families), including only adolescents who have one or more siblings of similar age (12 to 18 years) and both biological parents of the FrenchCanadian origin born in the region.

Supplementary Table 1 (Columns 2, 11-17). Studies included in the meta-analysis.

Short name	Collection of breastfeeding information	If retrospectively: mean age of the study participants (offspring)	Cognitive measure	Subtests included	Description of maternal education	Description of maternal cognitive measure	Generation of the genetic data
1982Pelotas	Retrospectively	1.6 years for 95% of the sample, and 3.5 years for 5% of the sample. Overall mean: 1.7 years.	Wechsler Adult Intelligence Scale (3rd version).	Arithmetic, digit symbol, similarities and picture completion	Complete years of education. Offspring age: 0 (measured at offspring birth).	Not available	Illumina HumanOmni2.5-8v1 array. QC: SNPs excluded if call rate <95%, Hardy–Weinberg $P < 1E-7$ or monomorphic. Samples excluded if there were sex mismatches (heterozygosity threshold: 0.02), heterozygosity rate outside the range of median $\pm 1.5 \times$ IQR, missingness >3% and cryptic relatedness (kinship>0.1, as described elsewhere). Imputation: pre-phasing using SHAPEIT and imputation using IMPUTE2 (reference panel 1000 Genomes Phase I integrated haplotypes - December 2013 release).

ALSPAC	Prospectively	Not applicable	Wechsler Intelligence Scale for Children.	Information, similarities, arithmetic, vocabulary, comprehension, picture completion, coding, picture arrangement, block design, object assembly.	Highest level of education coded using ISCED Offspring age: 0 (measured at offspring birth).	Verbal fluency test score included in analysis. The following tests are available: logic memory, digit backwards, digit symbol coding, verbal fluency, spot the word available	Illumina HumanHap550 quad array. QC: SNPs excluded if call rate<95%, Hardy-Weinberg $P < E-7$, or monomorphic. Samples excluded if there were sex mismatches, minimal or excessive heterozygosity, and cryptic relatedness ($IBD > 0.1$, as described elsewhere). Imputation: pre-phasing using SHAPEIT and imputation using IMPUTE2 v2.2.2 (reference panel: Haplotype Reference Consortium (HRC) panel pre-release 2015).

Biobank_Axiom	Retrospectively	56.86 years	This is a simple unweighted sum of the number of correct answers given to the 13 fluid intelligence questions. Participants who did not answer all of the questions within the allotted 2 minute limit are scored as zero for each of the unattempted questions.	Numeric addition test, arithmetic sequence recognition, antonym, square sequence recognition, subset inclusion logic, identify largest number, word interpolation, positional arithmetic, family relationship calculation, conditional arithmetic, synonym, chained arithmetic, concept interpolation.	Not available	Not available	UK Biobank Axiom (Affymetrix) genotyping array (800k markers) QC: SNPs excluded if call rate<99%, Hardy-Weinberg P<E-7, or monomorphis. Imputation: pre-phasing using SHAPEIT3 and imputation using IMPUTE2 (reference panel: Haplotype Reference Consortium (HRC) panel pre-release 2015). Restrctied to those of European genetic ancestry.

Biobank_BiLEVE	Retrospectively	57.04 years	This is a simple unweighted sum of the number of correct answers given to the 13 fluid intelligence questions. Participants who did not answer all of the questions within the allotted 2 minute limit are scored as zero for each of the unattempted questions.	Numeric addition test, arithmetic sequence recognition, antonym, square sequence recognition, subset inclusion logic, identify largest number, word interpolation, positional arithmetic, family relationship calculation, conditional arithmetic, synonym, chained arithmetic, concept interpolation.	Not available	Not available	UK BiLEVE array. QC: SNPs excluded if call rate<99%, Hardy-Weinberg P<E-7, or monomorphic. Imputation: pre-phasing using SHAPEIT3 and imputation using IMPUTE2 (reference panel: Haplotype Reference Consortium (HRC) panel pre-release 2015). Restrict to those of European genetic ancestry.

COPSAC2010	Prospectively	Not applicable	Bayley Scales of Infant and Toddler Development (BSID-III).	Sensorimotor development, exploration and manipulation, object relatedness, concept formation, memory	Complete years of education. Offspring age: 2-3 years.	Not available	Illumina HumanOmniExpressExome bead chip array. QC: SNPs excluded if call rate <95%, Hardy–Weinberg $P < 1E-6$ or monomorphic. Samples excluded if there were sex mismatches, heterogeneity (<0.28 or >0.38) and sample relatedness (duplicates and monozygotic twins). Imputation: pre-phasing using SHAPEIT and imputation using IMPUTE2 (reference panel 1000 Genomes Phase I Version 3 -June 2014 release).
DMHDS	Retrospectively	3 years	Wechsler Intelligence Scale for Children (Revised).	Information, Similarities, Arithmetic, Picture Completion, Block Design, Object Assembly and Digit Symbol	Not available	Mother's IQ was assessed with the SRA verbal test (Thurstone & Thurstone, 1973) administered to the sample mothers when the children were 3 years old.	The two <i>FADS2</i> polymorphisms were genotyped using manufacturer recommended protocols on the AB7900 TaqMan platform. The following functionally tested, made-to-order SNP genotyping assays from Appliedbiosystems were used: rs174575 C__2575522_20, rs1535 C__2575527_10.

GenerationR	Prospectively	Not applicable	Snijders-Oomen Non-verbal intelligence test—Revised (SON-R 2.5–7).	Mosaics and Categories	Completed levels of education according to Dutch educational system. Level of maternal education was established by questionnaire at enrollment	Maternal intelligence was assessed when she accompanied the child in the visit to the research centre, at the age of 6 years, using a computerised Ravens Advanced Progressive Matrices Test, set I. This set consists of 12 items and has been shown to be a reliable and valid short form of the Raven's Progressive Matrices to assess non-	Illumina HumanHap 610 or 660 Quad chips (Illumina Inc., San Diego, USA). If SNPs were not directly genotyped, we used MACH (version 1.0.15) software to impute genotypes using the HapMap II CEU (release 22) as reference set or SNPs were genotyped using the same method as the parents. Samples were excluded in case of low sample call rate (<97.5 %).

						verbal cognitive ability parallel to child non-verbal IQ.	
INMA	Retrospectively	Questionnaire at 6 months and at 14 months	McCarthy Scales of Children's Abilities.	General Cognitive Index which includes verbal scale (pictorial memory, word knowledge, verbal memoery, verbal fluency, opposite analogies test), perceptual- performance scale (block building, puzzle solving, tapping sequence, tigh- left orientation, draw-a-desing, draw-a-child, conceptual grouping tests), and quantitative scale (number questions, numerical	Level of education completed at the beginning of the pregnancy: i) illiterally or primary school unfinished; ii) primary school; iii) secondary school; iv) university or higher	INMA SABADELL, VALENCIA, GIPUZKOA: Similarities test (verbal sub-test) from the Wechsler Adult Intelligence Scale III (WAIS-III) when children where 4-5 years dol. INMA MENORCA: 2 subtest of the Cattell III A test (non- verbal sub- tests) when children where 9-11	<u>INMA SABADELL, VALENCIA AND MENORCA:</u> HumanOmni1-Quad v1.0 Beadchip (Illumina). Genotype calling was done using the GeneTrain2.0 algorithm based on HapMap clusters implemented in the GenomeStudio software. We applied the following initial quality control thresholds: sample call rate>98% and/or LRR SD<0.3. Then, we checked sex, relatedness (excluded: one duplicated sample and one sibling), heterozygosity and population stratification (no stratification was found). Genetic variants were filtered for SNP call rate>95%, MAF>1% and HWE p value>1.10E-6. Imputation: pre-phasing and imputation using IMPUTE2 (reference

				memory, and counting and sorting tests).		years old	panel 1000 Genomes -March 2012 release). <u>INMA</u> <u>GIPUZKOA:</u> HumanExome BeadChip Kit v.1.1 (Illumina). An initial genotype calling was done with the GeneTrain2.0 algorithm (GenomeStudio software) based on CHARGE clusters, and then we applied the zCall algorithm to improve the calling of low frequency variants (Goldstein et al 2012). While we performed a standard sample quality control that included the steps mentioned above, the genetic variant quality control was more strict, in order to filter low quality variants, and included additional filtering for clustering parameters.

Raine	Retrospectively	Information on the duration of breast feeding was collected at the 1, 2 and 3 year follow-ups.	Peabody Picture Vocabulary Test (PPVT-III).	Not applicable	Three questions asked at recruitment during pregnancy: 1. How old were you when you left school? 2. What was the last class at school that you completed? 3. Since leaving school, have you completed any further education?	Not available	Illumina Human660W Quad Array at the Centre for Applied Genomics (Toronto, Ontario, Canada). Individual QC: sex mismatches, one of each pair of individuals where $IBD > 0.1875$, low call rate ($< 97\%$), high heterozygosity (< 0.3). Genotype QC: Hardy-Weinberg $P > 5.7 \times 10^{-7}$, Call rate $> 95\%$, Minor allele frequency $> 1\%$. Imputation: MACH (V1.0.16) using the CEU samples from HapMap phase 2 as a reference panel.
SKOT-I	Retrospectively	9 months for exclusive breastfeeding and 18 months for duration of any breastfeeding. All breastfeeding beyond 18 months is coded as 18.5 months and if lack of	Ages and Stages Questionnaire (3 rd edition).	Gross motor, fine motor, personal/social, communication and problem solving scores at 36 month of age	Data collected at offspring age 9 mo. Two line of questions: basic school education 7-12 years of school and then further education (an number of question with cross-reference) ending in a coding of (none,	Not available	Genotyping array: Illumina HumanCoreExome Beadchip platform. Genotyping Center: The Novo Nordisk Foundation-Center for Basic Metabolic Research, Section of Metabolic Genetics, Copenhagen, Denmark. Genotype calling algorithm: Genotyping module (version 1.9.4) of GenomeStudio software (version 2011.1, Illumina).

		information at 18 months then any breastfeeding is truncated at 9.5 months (but only for very few participants)			vocational, short education (<3 years) tertiary education (3-4 years) and university education (bachelor, master or candidate). We did not ask for PhD and have therefore truncated at ISCED level 5. Short tertiary education was coded as ISCED 4 as this would include also educations shorter than 2 years.		QC: Call rate: 95% (for individuals and SNPs) Heterozygosity: for the inbreeding QC we used the following cut-offs: rare alleles -0.5 to 0.5; common alleles: -0.05 to 0.05. Ethnic outliers/ other exclusions: For the PCA CQ we used the following cut-offs: PCA1 -0.1 to 0.1; PCA2 -0.1 to 0.1. MAF: 0.01 HWE p value: 0.0001 Other: in case of siblings, only the sibling with the best overall genotype call-rate was retained in the study (regardless of gender) Imputation software: IMPUTE2 Imputation panel: 1000 genomes phase 1
SKOT-II	Retrospectively	9 months for exclusive breastfeeding and 18 months for duration of any breastfeeding. All	Ages and Stages Questionnaire (3 rd edition).	Gross motor, fine motor, personal/social, communication and problem solving scores at 36 month of age	Data collected at offspring age 9 mo. Two line of questions: basic school education 7-12 years of school and then further	Not available	Genotyping array: Illumina HumanCoreExome Beadchip platform. Genotyping Center: The Novo Nordisk Foundation-Center for Basic Metabolic Research, Section of Metabolic Genetics, Copenhagen,

		breastfeeding beyond 18 months is coded as 18.5 months and if lack of information at 18 months then any breastfeeding is truncated at 9.5 months (but only for very few participants).			education (an number of question with cross-reference) ending in a coding of (none, vocational, short education (<3 years) tertiary education (3-4 years) and university education (bachelor, master or candidate). We did not ask for PhD and have therefore truncated at ISCED level 5. Short tertiary education was coded as ISCED 4 as this would include also educations shorter than 2 years.		Denmark. Genotype calling algorithm: Genotyping module (version 1.9.4) of GenomeStudio software (version 2011.1, Illumina). QC: Call rate: 95% (for individuals and SNPs) Heterozygosity: for the inbreeding QC we used the following cut-offs: rare alleles -0.5 to 0.5; common alleles: -0.05 to 0.05. Ethnic outliers/ other exclusions: For the PCA CQ we used the following cut-offs: PCA1 -0.1 to 0.1; PCA2 -0.1 to 0.1. MAF: 0.01 HWE p value: 0.0001 Other: in case of siblings, only the sibling with the best overall genotype call-rate was retained in the study (regardless of gender) Imputation software: IMPUTE2 Imputation panel: 1000 genomes phase 1

SYS	Retrospectively	15.02 years	Wechsler Adult Intelligence Scale (3rd version).	digit span, picture completion, information, coding, similarities, picture arrangement, arithmetic, block design, vocabulary, object assembly, comprehension, verbal comprehension, perceptual organization, processing speed, and symbol . Also verbal, perceptual, full, freedom distraction, and processing IQ.	Level of schooling complete and incomplete measured at time of testing (offspring age 12-18): primary not completed, primary completed, high school not completed, high school completed, college not completed, college completed, university not completed, bachelor completed, master or doctorate.	Computerized cognitive battery of 12 tasks only for a subset (n=470) mothers: visuospatial working memory, grammatical reasoning, Stroop, odd out, spatial span, spatial rotate, feature, digit span, spatial planning, paired associate learning, polygons, and self-order.	592 adolescents genotyped with the Illumina Human610-Quad BeadChip (610K SNPs) the remaining adolescents were genotyped with the Illumina HumanOmniExpress BeadChip (700k SNPs). In both genotyping cases SNPs were excluded if call rate <95% and minor allele frequency <0.01 and not in Hardy–Weinberg equilibrium ($P < 1 \times 10^{-6}$) were excluded. Imputation: pre-phasing using SHAPEIT and imputation using IMPUTE2. Markers with low imputation quality (information score <0.5) or low minor allele frequency (<0.01) were removed.

Supplementary Table 2. Characteristics of the studies included in the meta-analysis regarding *FADS2* polymorphisms.

SNP	Characteristic	Study										
		1982Pelotas	ALSPAC	Biobank_Axiom	Biobank_BiLEVE	COPSAC2010	DMHDS	GenerationR	INMA	Raine	SKOT-I & II	SYS
rs174575	HWE P-value ^a	0.999	0.602	0.801	0.392	0.653	0.518	0.058	0.351	0.815	0.074	0.811
	Genotyped or imputed	Imputed	Imputed	Imputed	Imputed	Genotyped	Genotyped	Imputed	Genotyped	Imputed	Imputed	Imputed
	Imputation quality ^b	0.995	0.998	0.997	0.997	NA	NA	0.990	NA	0.984	0.994	0.996
	MAF	26.1%	26.3%	27.9%	27.5%	25.4%	30.7%	27.6%	30.8%	27.0%	20.5%	26.9%
rs1535	HWE P-value ^a	0.085	0.231	0.624	0.451	0.773	0.999	0.500	0.446	0.044	0.089	0.297
	Genotyped or imputed	Imputed	Imputed	Imputed	Imputed	Genotyped	Genotyped	Imputed	Genotyped	Imputed	Genotyped	Genotyped
	Imputation quality ^b	0.999	0.999	0.999	0.999	NA	NA	0.990	NA	0.999	NA	NA
	MAF	34.4%	33.4%	35.2%	34.9%	33.0%	39.1%	35.1%	31.2%	36.1%	28.5%	34.5%

^aComputed using the Fisher's exact test. For imputed variants, best-guess genotypes were used.

^bMetrics such as r² (MACH software) and INFO (IMPUTE2 software).

HWE: Hardy-Weinberg Equilibrium.

MAF: Minor allele frequency. For both genetic variants, G is the minor (ie, rarest) allele.

NA: Not applicable.

Supplementary Table 3. Characteristics of the studies included in the meta analysis regarding sex, age, maternal education, breastfeeding and *FADS2* polymorphisms.

Variable	Study										
	1982Pelotas	ALSPAC	Biobank_Axiom	Biobank_BiLEVE	COPSAC2010	DMHDS	GenerationR	INMA	Raine	SKOT-I & II	SYS
Number of individuals	1799	4809	21774	11068	551	859	1786	1131	1047	299	1011
Sex											
Female (%)	52.3	50.1	57.0	51.5	48.5	48.9	50.5	47.8	47.9	48.5	51.8
Male (%)	47.7	49.9	43.0	48.5	51.5	51.1	49.5	52.2	52.1	51.5	48.2
Age (years)											
Minimum	29.4	7.5	40.0	40.0	2.0	10.0	4.9	3.4	9.4	2.9	11.0
Maximum	31.1	10.5	73.0	70.0	2.8	10.0	9.0	6.9	12.4	3.3	19.0
Mean	30.2	8.6	56.5	56.8	2.5	10.0	6.0	4.8	10.6	3.0	14.5
Standard deviation	0.3	0.3	8.0	7.9	0.1	0.0	0.3	0.6	0.2	0.1	1.8
Median	30.2	8.6	58.0	58.0	2.5	10.0	6.0	4.5	10.5	3.0	14.0
Interquartile range	0.5	0.2	13.0	12.0	0.1	0.0	0.3	0.5	0.1	0.1	3.0
Intelligence measure (points)											
Minimum	67.0	45.0	0.0 ^a	0.0 ^a	85.0	45.5	50.0	35.0	58.0	15.0 ^b	58.0
Maximum	133.0	151.0	13.0 ^a	13.0 ^a	145.0	141.5	150.0	147.9	125.0	60.0 ^b	138.0
Mean	100.4	105.5	6.3 ^a	6.2 ^a	104.5	100.7	105.7	100.7	104.9	51.5 ^b	104.4
Standard deviation	12.2	16.3	2.1 ^a	2.1 ^a	9.6	14.0	14.4	14.4	11.9	10.0 ^b	12.1
Median	100.0	105.0	6.0 ^a	6.0 ^a	105.0	100.8	105.5	100.9	106.0	55.0 ^b	105.0
Interquartile range	16.0	23.0	3.0 ^a	3.0 ^a	10.0	18.5	18.0	18.1	19.0	11.5 ^b	15.0
Maternal education (years)											
Minimum	1.0	10.0	NA	NA	10.0	NA	7.0	1.0	7.0	10.0	7.0

n	22.0	19.0	NA	NA	19.0	NA	22.0	19.0	19.0	19.0	19.0
	11.0	14.7	NA	NA	17.4	NA	19.0	13.0	13.3	17.2	15.8
	4.4	2.3	NA	NA	2.8	NA	3.5	4.8	3.9	2.7	3.1
	10.0	15.0	NA	NA	19.0	NA	19.0	13.0	10.0	19.0	13.0
	6.0	2.0	NA	NA	6.0	NA	3.0	12.0	9.0	4.0	6.0
e	6.8	17.2	27.1	27.7	5.4	43.2	9.0	8.7	10.2	5.0	50.0
	93.2	82.8	72.9	72.3	94.6	56.8	91.0	91.3	89.8	95.0	50.0
ation)											
(%)	6.8	17.2	27.1	27.7	0.9	43.2 ^d	10.5	8.7	10.8	1.0	50.6
	25.1	15.4	NA	NA	4.9	NA	14.5	6.9	10.1	4.0	10.6
(%)	30.4	16.0	NA	NA	7.3	NA	17.1	14.2	15.5	4.6	12.6
(%)	15.0	32.1	NA	NA	18.9	NA	26.8	28.3	17.3	25.2	14.5
	22.7	19.3	NA	NA	68.0	NA	31.1	41.9	46.3	65.2	11.7
months)											
n	0.0	0.0	NA	NA	0.0	NA	0.0	0.0	0.0	0.0	0.0
	49.0	18.0	NA	NA	46.7	NA	14.0	16.0	38.0	18.5	30.0
	5.9	3.4	NA	NA	8.2	NA	4.5	5.8	7.2	7.6	2.3
	9.4	2.7	NA	NA	4.9	NA	3.8	4.3	6.4	3.8	3.6
	3.0	3.5	NA	NA	7.9	NA	3.5	5.1	6.0	7.5	0.0
e	5.0	5.5	NA	NA	5.4	NA	5.1	5.9	9.0	6.0	4.0
eeding ^c											
eeding ation)	7.3	39.2	NA	NA	20.0	NA	54.5	18.4	13.5	14.6	68.3
	92.7	60.8	NA	NA	80.0	NA	45.5	81.6	86.5	85.4	31.7

None (%)	7.3	39.3	NA	NA	3.8	NA	54.5	18.4	13.5	5.0	68.3
0.01-1.00 months (%)	30.6	8.7	NA	NA	16.7	NA	0.0	9.7	15.3	9.6	15.0
1.01-3.00 months (%)	50.2	39.2	NA	NA	10.7	NA	25.0	14.9	28.7	9.6	15.3
3.01-6.00 months (%)	11.2	12.8	NA	NA	64.2	NA	20.5	52.1	39.8	64.6	1.4
>6.00 months (%)	0.7	0.0	NA	NA	4.6	NA	0.0	4.9	2.7	11.2	0.0
Exclusive breastfeeding (continuous, in months)											
Minimum	0.0	0.0	NA	NA	0.0	NA	0.0	0.0	0.0	0.0	0.0
Maximum	12.0	6.9	NA	NA	8.5	NA	6.0	10.0	10.0	7.0	6.0
Mean	2.0	1.6	NA	NA	3.5	NA	1.3	3.0	2.9	3.7	0.4
Standard deviation	1.5	1.6	NA	NA	1.9	NA	1.6	2.2	1.9	1.9	0.9
Median	2.0	1.8	NA	NA	4.1	NA	0.0	3.6	3.0	4.0	0.0
Interquartile range	2.3	3.0	NA	NA	3.1	NA	2.0	4.5	3.0	2.0	0.5
rs174575^e											
CC (%)	54.6	54.2	52.0	52.3	55.2	49.1	51.6	51.3	52.7	61.6	53.1
CG (%)	38.6	39.0	40.2	40.3	38.8	41.3	41.7	41.6	40.0	36.1	39.8
GG (%)	6.8	6.8	7.8	7.4	6.0	9.6	6.7	7.1	7.3	2.3	7.1
rs1535^e											
AA (%)	43.9	43.9	41.9	42.2	44.5	37.1	41.8	46.9	39.4	49.0	43.6
AG (%)	43.3	45.3	45.8	45.8	45.0	47.6	46.3	43.9	49.1	45.0	43.7
GG (%)	12.8	10.8	12.3	12.0	10.5	15.3	11.9	9.2	11.5	6.0	12.7

^aNon-standardised test (as described in Supplementary Table 1).

^bThis study used the Ages and Stages Questionnaire, which is standardised to have mean=50 and standard deviation=10.

^cIn COPSAC2010 and SKOT1-2, the prevalence of having never being breastfed was extremely low. Therefore, in these three studies the following definition was used: "never" if <1 month; "ever" if at least 1 month of duration.

^dCorresponds to the prevalence of ever being breastfed (data on breastfeeding duration was not available for this study).

^eFor imputed variants, best-guess genotypes are shown used.

NA: Not available.

Supplementary Table 4. Meta-analytical linear regression coefficients (β) of cognitive measures (in standard deviation units) according to breastfeeding (0: <6 months; 1: ≥ 6 months), within strata of *FADS2* rs174575 or rs1535 genotypes (recessive effect).

Model	Statistic	Fixed effects			Random effects		
		FADS2		G×E	FADS2		G×E
		Other genotypes	GG		Other genotypes	GG	
rs174575 (CC or CG=0; GG=1)							
Unadjusted	<i>I</i> ²	-	-	-	80.1	23.5	23.1
N _{estimates} =8	P-value	4.2×10 ⁻⁴⁴	1.3×10 ⁻⁵	0.515	1.1×10 ⁻⁵	9.2×10 ⁻⁴	0.646
N _{subjects} =11,733	β	0.28	0.32	0.05	0.22	0.29	0.04
	95% CI	0.24; 0.32	0.18; 0.46	-0.10; 0.20	0.12; 0.32	0.12; 0.47	-0.14; 0.22
Adjusted (1) ^a	<i>I</i> ²	-	-	-	71.5	45.8	53.6
N _{estimates} =8	P-value	2.3×10 ⁻⁴¹	3.4×10 ⁻⁶	0.378	5.6×10 ⁻⁷	4.7×10 ⁻³	0.546
N _{subjects} =11,706	β	0.29	0.37	0.07	0.24	0.35	0.08
	95% CI	0.24; 0.33	0.21; 0.52	-0.09; 0.23	0.14; 0.33	0.11; 0.59	-0.18; 0.35
Adjusted (2) ^b	<i>I</i> ²	-	-	-	69.9	81.2	82.6
N _{estimates} =8	P-value	3.8×10 ⁻¹⁹	1.9×10 ⁻⁴	0.244	1.8×10 ⁻³	0.166	0.496
N _{subjects} =11,242	β	0.20	0.31	0.10	0.15	0.33	0.17
	95% CI	0.15; 0.24	0.15; 0.47	-0.07; 0.26	0.06; 0.25	-0.14; 0.79	-0.32; 0.65
rs1535 (AA or AG=0; GG=1)							
Unadjusted	<i>I</i> ²	-	-	-	81.6	8.3	0.0
N _{estimates} =8	P-value	8.6×10 ⁻⁴⁵	7.3×10 ⁻⁵	0.460	2.0×10 ⁻⁵	3.4×10 ⁻⁴	0.460
N _{subjects} =12,018	β	0.28	0.23	-0.05	0.22	0.23	-0.05
	95% CI	0.24; 0.32	0.12; 0.35	-0.17; 0.08	0.12; 0.33	0.10; 0.35	-0.17; 0.08
Adjusted (1) ^a	<i>I</i> ²	-	-	-	74.2	9.1	8.0
N _{estimates} =8	P-value	2.4×10 ⁻⁴¹	3.3×10 ⁻⁴	0.248	6.7×10 ⁻⁶	0.001	0.302
N _{subjects} =11,991	β	0.29	0.22	-0.07	0.23	0.21	-0.07
	95% CI	0.25; 0.33	0.10; 0.34	-0.20; 0.05	0.13; 0.33	0.08; 0.34	-0.20; 0.06
Adjusted (2) ^b	<i>I</i> ²	-	-	-	71.3	0.1	3.9
N _{estimates} =8	P-value	8.7×10 ⁻²⁰	0.056	0.194	3.5×10 ⁻³	0.057	0.216
N _{subjects} =11,499	β	0.20	0.12	-0.08	0.15	0.12	-0.08
	95% CI	0.16; 0.25	0.00; 0.24	-0.21; 0.04	0.05; 0.24	0.00; 0.24	-0.21; 0.05

^aCovariates were sex, age (linear and quadratic terms), ancestry-informative principal components (if available) and genotyping centre (if necessary).

^bSame covariates than in the Adjusted (1) model, in addition to maternal education (linear and quadratic terms) and/or maternal cognition (linear and quadratic terms).

G×E: interaction between breastfeeding and polymorphisms in the *FADS2* gene.

$N_{\text{estimates}}$: number of estimates being pooled. N_{subjects} : pooled sample size.

Supplementary Table 5. Meta-analytical linear regression coefficients (β) of cognitive measures (in standard deviation units) according to breastfeeding (0: none; 1: 0.01-1.00 months; 2: 1.01-3.00 months; 3: 3.01-6.00 months; 4: >6.00 months), within strata of *FADS2* rs174575 or rs1535 genotypes (recessive effect).

Model	Statistic	Fixed effects			Random effects		
		FADS2		G×E	FADS2		G×E
		Other genotypes	GG		Other genotypes	GG	
rs174575 (CC or CG=0; GG=1)							
Unadjusted	<i>I</i> ²	-	-	-	81.9	44.2	57.1
N _{estimates} =8	P-value	2.8×10 ⁻⁷³	2.5×10 ⁻¹¹	0.104	1.5×10 ⁻⁷	3.5×10 ⁻⁶	0.150
N _{subjects} =11733	β	0.13	0.16	0.04	0.10	0.17	0.06
	95% CI	0.11; 0.14	0.12; 0.21	-0.01; 0.09	0.06; 0.14	0.10; 0.24	-0.02; 0.15
Adjusted (1) ^a	<i>I</i> ²	-	-	-	67.7	41.4	58.7
N _{estimates} =8	P-value	4.8×10 ⁻⁷⁴	2.7×10 ⁻¹⁰	0.189	2.8×10 ⁻¹³	5.9×10 ⁻⁵	0.282
N _{subjects} =11706	β	0.13	0.17	0.04	0.12	0.17	0.06
	95% CI	0.12; 0.15	0.12; 0.22	-0.02; 0.09	0.09; 0.15	0.09; 0.25	-0.05; 0.16
Adjusted (2) ^b	<i>I</i> ²	-	-	-	73.8	83.0	84.6
N _{estimates} =8	P-value	2.3×10 ⁻³⁷	8.3×10 ⁻⁷	0.132	2.8×10 ⁻⁵	0.070	0.346
N _{subjects} =11242	β	0.10	0.14	0.04	0.08	0.15	0.09
	95% CI	0.08; 0.11	0.08; 0.19	-0.01; 0.10	0.04; 0.12	-0.01; 0.32	-0.09; 0.26
rs1535 (AA or AG=0; GG=1)							
Unadjusted	<i>I</i> ²	-	-	-	82.3	11.9	0.0
N _{estimates} =8	P-value	1.9×10 ⁻⁷²	6.3×10 ⁻¹⁰	0.966	1.7×10 ⁻⁷	8.2×10 ⁻⁸	0.966
N _{subjects} =12018	β	0.13	0.12	0.00	0.10	0.12	0.00
	95% CI	0.11; 0.14	0.08; 0.16	-0.04; 0.04	0.06; 0.14	0.07; 0.16	-0.04; 0.04
Adjusted (1) ^a	<i>I</i> ²	-	-	-	71.5	58.0	54.3
N _{estimates} =8	P-value	4.9×10 ⁻⁷²	1.5×10 ⁻⁸	0.508	5.9×10 ⁻¹¹	0.011	0.635
N _{subjects} =11991	β	0.13	0.11	-0.01	0.11	0.09	-0.02
	95% CI	0.12; 0.15	0.07; 0.15	-0.06; 0.03	0.08; 0.15	0.02; 0.16	-0.09; 0.05
Adjusted (2) ^b	<i>I</i> ²	-	-	-	74.1	37.4	29.9
N _{estimates} =8	P-value	3.2×10 ⁻³⁷	1.1×10 ⁻⁴	0.675	5.2×10 ⁻⁵	0.025	0.728
N _{subjects} =11499	β	0.10	0.08	-0.01	0.08	0.07	-0.01
	95% CI	0.08; 0.11	0.04; 0.12	-0.05; 0.03	0.04; 0.12	0.01; 0.13	-0.07; 0.05

^aCovariates were sex, age (linear and quadratic terms), ancestry-informative principal components (if available) and genotyping centre (if necessary).

^bSame covariates than in the Adjusted (1) model, in addition to maternal education (linear and quadratic terms) and/or maternal cognition (linear and quadratic terms).

G×E: interaction between breastfeeding and polymorphisms in the *FADS2* gene.

$N_{\text{estimates}}$: number of estimates being pooled. N_{subjects} : pooled sample size.

Supplementary Table 6. Meta-analytical linear regression coefficients (β) of cognitive measures (in standard deviation units) according to breastfeeding (in months of duration), within strata of *FADS2* rs174575 or rs1535 genotypes (recessive effect).

Model	Statistic	Fixed effects			Random effects		
		FADS2		G×E	FADS2		G×E
		Other genotypes	GG		Other genotypes	GG	
rs174575 (CC or CG=0; GG=1)							
Unadjusted	<i>I</i> ²	-	-	-	95.8	62.2	13.9
N _{estimates} =8	P-value	5.8×10 ⁻²⁹	1.3×10 ⁻⁵	0.371	0.007	0.002	0.335
N _{subjects} =11733	β	0.02	0.03	0.01	0.03	0.04	0.01
	95% CI	0.02; 0.02	0.02; 0.05	-0.01; 0.02	0.01; 0.05	0.02; 0.07	-0.01; 0.03
Adjusted (1) ^a	<i>I</i> ²	-	-	-	95.3	75.1	63.3
N _{estimates} =8	P-value	2.3×10 ⁻³⁰	3.6×10 ⁻⁵	0.608	0.002	0.027	0.635
N _{subjects} =11706	β	0.02	0.03	0.00	0.03	0.04	0.01
	95% CI	0.02; 0.03	0.02; 0.05	-0.01; 0.02	0.01; 0.06	0.01; 0.08	-0.02; 0.04
Adjusted (2) ^b	<i>I</i> ²	-	-	-	91.0	86.8	85.3
N _{estimates} =8	P-value	3.8×10 ⁻¹⁸	0.004	0.782	0.007	0.165	0.602
N _{subjects} =11242	β	0.02	0.02	0.00	0.02	0.04	0.01
	95% CI	0.01; 0.02	0.01; 0.04	-0.01; 0.02	0.01; 0.04	-0.02; 0.09	-0.04; 0.07
rs1535 (AA or AG=0; GG=1)							
Unadjusted	<i>I</i> ²	-	-	-	95.8	66.4	0.0
N _{estimates} =8	P-value	9.8×10 ⁻³⁰	2.6×10 ⁻⁴	0.805	0.006	0.014	0.805
N _{subjects} =12018	β	0.02	0.02	0.00	0.03	0.02	0.00
	95% CI	0.02; 0.03	0.01; 0.03	-0.01; 0.01	0.01; 0.05	0.01; 0.04	-0.01; 0.01
Adjusted (1) ^a	<i>I</i> ²	-	-	-	95.3	72.1	59.6
N _{estimates} =8	P-value	7.7×10 ⁻²⁹	3.1×10 ⁻⁴	0.538	0.006	0.133	0.330
N _{subjects} =11991	β	0.02	0.02	0.00	0.03	0.02	-0.01
	95% CI	0.02; 0.03	0.01; 0.03	-0.01; 0.01	0.01; 0.05	-0.01; 0.04	-0.03; 0.01
Adjusted (2) ^b	<i>I</i> ²	-	-	-	91.1	45.3	35.5
N _{estimates} =8	P-value	6.0×10 ⁻¹⁸	0.035	0.319	0.013	0.190	0.344
N _{subjects} =11499	β	0.02	0.01	-0.01	0.02	0.01	-0.01
	95% CI	0.01; 0.02	0.00; 0.02	-0.02; 0.01	0.00; 0.04	-0.01; 0.03	-0.02; 0.01

^aCovariates were sex, age (linear and quadratic terms), ancestry-informative principal components (if available) and genotyping centre (if necessary).

^bSame covariates than in the Adjusted (1) model, in addition to maternal education (linear and quadratic terms) and/or maternal cognition (linear and quadratic terms).

G×E: interaction between breastfeeding and polymorphisms in the *FADS2* gene.

$N_{\text{estimates}}$: number of estimates being pooled. N_{subjects} : pooled sample size.

Supplementary Table 7. Meta-analytical linear regression coefficients (β) of the interaction between *FADS2* rs174575 or rs1535 genotypes (additive effect) with different categorisations of breastfeeding, having cognitive measures (in standard deviation units) as the outcome.

Model	Statistic	Fixed effects				Random effects			
		Never=0 Ever=1	<6 months=0 ≥6 months=1	Numerically- coded categories	Months	Never=0 Ever=1	<6 months=0 ≥6 months=1	Numerically- coded categories	Months
rs174575 (CC=0; CG=1; GG=2)									
Unadjusted	N _{estimates}	9	8	8	8	9	8	8	8
	N _{subjects}	12916	11733	11733	11733	12916	11733	11733	11733
	I ²	-	-	-	-	51.7	10.3	29.1	0.0
	P-value	0.606	0.893	0.272	0.951	0.719	0.928	0.359	0.951
	β	0.02	0.00	0.01	0.00	0.02	0.00	0.01	0.00
	95% CI	-0.05; 0.09	-0.07; 0.06	-0.01; 0.03	-0.01; 0.01	-0.10; 0.15	-0.07; 0.07	-0.02; 0.04	-0.01; 0.01
Adjusted (1) ^a	N _{estimates}	9	8	8	8	9	8	8	8
	N _{subjects}	12889	11706	11706	11706	12889	11706	11706	11706
	I ²	-	-	-	-	65.6	30.0	32.0	0.0
	P-value	0.547	0.963	0.228	0.897	0.651	0.872	0.240	0.897
	β	0.02	0.00	0.01	0.00	0.03	0.01	0.02	0.00
	95% CI	-0.05; 0.10	-0.06; 0.06	-0.01; 0.03	-0.01; 0.01	-0.11; 0.18	-0.08; 0.09	-0.01; 0.05	-0.01; 0.01
Adjusted (2) ^b	N _{estimates}	9	8	8	8	9	8	8	8
	N _{subjects}	12375	11242	11242	11242	12375	11242	11242	11242
	I ²	-	-	-	-	44.7	26.3	32.8	0.0
	P-value	0.183	0.861	0.132	0.970	0.256	0.761	0.155	0.970
	β	0.05	0.01	0.02	0.00	0.07	0.01	0.02	0.00
	95% CI	-0.02; 0.12	-0.06; 0.07	-0.01; 0.04	-0.01; 0.01	-0.05; 0.18	-0.07; 0.09	-0.01; 0.05	-0.01; 0.01
rs1535 (AA=0; AG=1; GG=2)									
Unadjusted	N _{estimates}	9	8	8	8	9	8	8	8
	N _{subjects}	13202	12018	12018	12018	13202	12018	12018	12018
	I ²	-	-	-	-	14.7	0.0	24.4	0.0

	P-value	0.588	0.254	0.694	0.570	0.594	0.254	0.709	0.570
	β	0.02	-0.03	0.00	0.00	0.02	-0.03	0.00	0.00
	95% CI	-0.05; 0.09	-0.09; 0.02	-0.02; 0.02	-0.01; 0.00	-0.06; 0.10	-0.09; 0.02	-0.02; 0.03	-0.01; 0.00
Adjusted (1) ^a	N _{estimates}	9	8	8	8	9	8	8	8
	N _{subjects}	13175	11991	11991	11991	13175	11991	11991	11991
	I^2	-	-	-	-	47.4	13.7	31.1	16.2
	P-value	0.458	0.223	0.787	0.648	0.403	0.329	0.695	0.504
	β	0.03	-0.04	0.00	0.00	0.05	-0.03	0.01	0.00
	95% CI	-0.04; 0.09	-0.09; 0.02	-0.02; 0.02	-0.01; 0.00	-0.06; 0.15	-0.10; 0.03	-0.02; 0.03	-0.01; 0.00
Adjusted (2) ^b	N _{estimates}	9	8	8	8	9	8	8	8
	N _{subjects}	12633	11499	11499	11499	12633	11499	11499	11499
	I^2	-	-	-	-	8.5	6.7	18.4	0.0
	P-value	0.150	0.413	0.429	0.467	0.157	0.477	0.415	0.467
	β	0.05	-0.02	0.01	0.00	0.05	-0.02	0.01	0.00
	95% CI	-0.02; 0.12	-0.08; 0.03	-0.01; 0.03	-0.01; 0.00	-0.02; 0.13	-0.09; 0.04	-0.01; 0.04	-0.01; 0.00

^aCovariates were sex, age (linear and quadratic terms), ancestry-informative principal components (if available) and genotyping centre (if necessary).

^bSame covariates than in the Adjusted (1) model, in addition to maternal education (linear and quadratic terms) and/or maternal cognition (linear and quadratic terms).

N_{estimates}: number of estimates being pooled. N_{subjects}: pooled sample size.

Supplementary Table 8. Meta-analytical linear regression coefficients (β) of the interaction between *FADS2* rs174575 or rs1535 genotypes (dominant effect) with breastfeeding, having cognitive measures (in standard deviation units) as the outcome.

Model	Statistic	Fixed effects				Random effects			
		Never=0 Ever=1	<6 months=0 ≥6 months=1	Numerically- coded categories	Months	Never=0 Ever=1	<6 months=0 ≥6 months=1	Numerically- coded categories	Months
rs174575 (CC=0; CG=1; GG=2)									
Unadjusted	N _{estimates}	9	8	8	8	9	8	8	8
	N _{subjects}	12916	11733	11733	11733	12916	11733	11733	11733
	I ²	-	-	-	-	43.0	18.3	23.9	0.0
	P-value	0.733	0.511	0.471	0.807	0.780	0.671	0.512	0.807
	β	0.02	-0.03	0.01	0.00	0.02	-0.02	0.01	0.00
	95% CI	-0.08; 0.11	-0.10; 0.05	-0.02; 0.04	-0.01; 0.01	-0.12; 0.16	-0.11; 0.07	-0.02; 0.05	-0.01; 0.01
Adjusted (1) ^a	N _{estimates}	9	8	8	8	9	8	8	8
	N _{subjects}	12889	11706	11706	11706	12889	11706	11706	11706
	I ²	-	-	-	-	56.2	24.0	12.4	0.0
	P-value	0.653	0.593	0.348	0.998	0.704	0.835	0.332	0.998
	β	0.02	-0.02	0.01	0.00	0.03	-0.01	0.02	0.00
	95% CI	-0.07; 0.11	-0.10; 0.06	-0.01; 0.04	-0.01; 0.01	-0.13; 0.20	-0.11; 0.09	-0.02; 0.05	-0.01; 0.01
Adjusted (2) ^b	N _{estimates}	9	8	8	8	9	8	8	8
	N _{subjects}	12375	11242	11242	11242	12375	11242	11242	11242
	I ²	-	-	-	-	29.2	16.7	5.8	0.0
	P-value	0.219	0.748	0.215	0.909	0.230	0.914	0.213	0.909
	β	0.06	-0.01	0.02	0.00	0.08	-0.01	0.02	0.00
	95% CI	-0.04; 0.15	-0.09; 0.06	-0.01; 0.04	-0.01; 0.01	-0.05; 0.21	-0.10; 0.09	-0.01; 0.05	-0.01; 0.01
rs1535 (AA=0; AG=1; GG=2)									
Unadjusted	N _{estimates}	9	8	8	8	9	8	8	8
	N _{subjects}	13202	12018	12018	12018	13202	12018	12018	12018
	I ²	-	-	-	-	0.0	0.0	11.6	0.0

	P-value	0.439	0.289	0.591	0.619	0.439	0.289	0.579	0.619
	β	0.04	-0.04	0.01	0.00	0.04	-0.04	0.01	0.00
	95% CI	-0.06; 0.13	-0.12; 0.03	-0.02; 0.03	-0.01; 0.01	-0.06; 0.13	-0.12; 0.03	-0.02; 0.04	-0.01; 0.01
Adjusted (1) ^a	N _{estimates}	9	8	8	8	9	8	8	8
	N _{subjects}	13175	11991	11991	11991	13175	11991	11991	11991
	I^2	-	-	-	-	24.1	8.6	14.0	0.0
	P-value	0.399	0.315	0.562	0.723	0.295	0.377	0.518	0.723
	β	0.04	-0.04	0.01	0.00	0.07	-0.04	0.01	0.00
	95% CI	-0.05; 0.14	-0.12; 0.04	-0.02; 0.03	-0.01; 0.01	-0.06; 0.19	-0.12; 0.05	-0.02; 0.04	-0.01; 0.01
Adjusted (2) ^b	N _{estimates}	9	8	8	8	9	8	8	8
	N _{subjects}	12633	11499	11499	11499	12633	11499	11499	11499
	I^2	-	-	-	-	0.0	0.0	11.9	0.0
	P-value	0.226	0.681	0.230	0.712	0.226	0.681	0.226	0.712
	β	0.06	-0.02	0.02	0.00	0.06	-0.02	0.02	0.00
	95% CI	-0.04; 0.16	-0.09; 0.06	-0.01; 0.04	-0.01; 0.01	-0.04; 0.16	-0.09; 0.06	-0.01; 0.05	-0.01; 0.01

^aCovariates were sex, age (linear and quadratic terms), ancestry-informative principal components (if available) and genotyping centre (if necessary).

^bSame covariates than in the Adjusted (1) model, in addition to maternal education (linear and quadratic terms) and/or maternal cognition (linear and quadratic terms).

N_{estimates}: number of estimates being pooled. N_{subjects}: pooled sample size.

Supplementary Table 9. Meta-analytical linear regression coefficients (β) of the interaction between *FADS2* rs174575 or rs1535 genotypes (overdominant effect) with breastfeeding, having cognitive measures (in standard deviation units) as the outcome.

Model	Statistic	Fixed effects				Random effects			
		Never=0 Ever=1	<6 months=0 ≥6 months=1	Numerically- coded categories	Months	Never=0 Ever=1	<6 months=0 ≥6 months=1	Numerically- coded categories	Months
rs174575 (CC=0; CG=1; GG=2)									
Unadjusted	N _{estimates}	9	8	8	8	9	8	8	8
	N _{subjects}	12916	11733	11733	11733	12916	11733	11733	11733
	I ²	-	-	-	-	53.1	19.5	13.8	0.0
	P-value	0.965	0.253	0.980	0.567	0.926	0.458	0.915	0.567
	β	0.00	-0.05	0.00	0.00	0.01	-0.04	0.00	0.00
	95% CI	-0.09; 0.10	-0.12; 0.03	-0.03; 0.03	-0.01; 0.01	-0.15; 0.17	-0.13; 0.06	-0.03; 0.03	-0.01; 0.01
Adjusted (1) ^a	N _{estimates}	9	8	8	8	9	8	8	8
	N _{subjects}	12889	11706	11706	11706	12889	11706	11706	11706
	I ²	-	-	-	-	57.0	19.1	0.0	0.0
	P-value	0.862	0.342	0.716	0.867	0.871	0.580	0.716	0.867
	β	0.01	-0.04	0.01	0.00	0.01	-0.03	0.01	0.00
	95% CI	-0.09; 0.10	-0.12; 0.04	-0.02; 0.03	-0.01; 0.01	-0.16; 0.18	-0.12; 0.07	-0.02; 0.03	-0.01; 0.01
Adjusted (2) ^b	N _{estimates}	9	8	8	8	9	8	8	8
	N _{subjects}	12375	11242	11242	11242	12375	11242	11242	11242
	I ²	-	-	-	-	30.8	17.7	0.0	0.0
	P-value	0.353	0.441	0.430	0.900	0.314	0.662	0.430	0.900
	β	0.05	-0.03	0.01	0.00	0.07	-0.02	0.01	0.00
	95% CI	-0.05; 0.14	-0.11; 0.05	-0.02; 0.04	-0.01; 0.01	-0.06; 0.20	-0.12; 0.07	-0.02; 0.04	-0.01; 0.01
rs1535 (AA=0; AG=1; GG=2)									
Unadjusted	N _{estimates}	9	8	8	8	9	8	8	8
	N _{subjects}	13202	12018	12018	12018	13202	12018	12018	12018
	I ²	-	-	-	-	6.7	0.0	0.0	0.0
	P-value	0.469	0.578	0.655	0.738	0.399	0.578	0.655	0.738

	β	0.03	-0.02	0.01	0.00	0.04	-0.02	0.01	0.00
	95% CI	-0.06; 0.13	-0.10; 0.05	-0.02; 0.03	-0.01; 0.01	-0.06; 0.14	-0.10; 0.05	-0.02; 0.03	-0.01; 0.01
Adjusted (1) ^a	N _{estimates}	9	8	8	8	9	8	8	8
	N _{subjects}	13175	11991	11991	11991	13175	11991	11991	11991
	I^2	-	-	-	-	24.4	0.0	0.0	0.0
	P-value	0.522	0.582	0.520	0.756	0.368	0.582	0.520	0.756
	β	0.03	-0.02	0.01	0.00	0.05	-0.02	0.01	0.00
	95% CI	-0.06; 0.12	-0.10; 0.05	-0.02; 0.04	-0.01; 0.01	-0.06; 0.17	-0.10; 0.05	-0.02; 0.04	-0.01; 0.01
Adjusted (2) ^b	N _{estimates}	9	8	8	8				
	N _{subjects}	12633	11499	11499	11499				
	I^2	-	-	-	-				
	P-value	0.658	0.856	0.164	0.863				
	β	0.02	0.01	0.02	0.00				
	95% CI	-0.07; 0.12	-0.07; 0.08	-0.01; 0.05	-0.01; 0.01				

^aCovariates were sex, age (linear and quadratic terms), ancestry-informative principal components (if available) and genotyping centre (if necessary).

^bSame covariates than in the Adjusted (1) model, in addition to maternal education (linear and quadratic terms) and/or maternal cognition (linear and quadratic terms).

N_{estimates}: number of estimates being pooled. N_{subjects}: pooled sample size.

Supplementary Table 10. Meta-analytical linear regression coefficients (β) of the interaction between *FADS2* rs174575 or rs1535 genotypes (recessive effect) with exclusive breastfeeding, having cognitive measures (in standard deviation units) as the outcome.

Model	Statistic	Fixed effects			Random effects		
		Never=0 Ever=1	Numerically- coded categories	Months	Never=0 Ever=1	Numerically- coded categories	Months
rs174575 (CC=0; CG=1; GG=2)							
Unadjusted	N _{estimates}	8	8	8	8	8	8
	N _{subjects}	11388	11386	11386	11388	11386	11386
	I ²	-	-	-	51.8	1.4	0.0
	P-value	0.814	0.944	0.706	0.993	0.957	0.706
	β	0.02	0.00	0.01	0.00	0.00	0.01
	95% CI	-0.14; 0.18	-0.06; 0.07	-0.04; 0.05	-0.26; 0.26	-0.06; 0.07	-0.04; 0.05
Adjusted (1) ^a	N _{estimates}	8	8	8	8	8	8
	N _{subjects}	11363	11361	11361	11363	11361	11361
	I ²	-	-	-	47.5	68.3	70.3
	P-value	0.329	0.647	0.411	0.695	0.891	0.812
	β	0.08	0.02	0.02	0.06	0.01	0.01
	95% CI	-0.08; 0.25	-0.05; 0.09	-0.03; 0.07	-0.22; 0.34	-0.14; 0.16	-0.09; 0.11
Adjusted (2) ^b	N _{estimates}	8	8	8	8	8	8
	N _{subjects}	11000	10998	10998	11000	10998	10998
	I ²	-	-	-	85.5	84.9	86.7
	P-value	0.227	0.552	0.520	0.652	0.682	0.886
	β	0.11	0.02	0.02	0.13	0.05	0.01
	95% CI	-0.07; 0.28	-0.05; 0.09	-0.03; 0.06	-0.43; 0.68	-0.17; 0.26	-0.14; 0.17
rs1535 (AA=0; AG=1; GG=2)							
Unadjusted	N _{estimates}	8	8	8	8	8	8
	N _{subjects}	11671	11669	11669	11671	11669	11669
	I ²	-	-	-	26.8	0.0	0.0
	P-value	0.788	0.633	0.465	0.656	0.633	0.465
	β	-0.02	-0.01	-0.01	-0.04	-0.01	-0.01
	95% CI	-0.15; 0.11	-0.07; 0.04	-0.05; 0.02	-0.21; 0.13	-0.07; 0.04	-0.05; 0.02
Adjusted (1) ^a	N _{estimates}	8	8	8	8	8	8
	N _{subjects}	11646	11644	11644	11646	11644	11644
	I ²	-	-	-	13.2	11.4	62.8
	P-value	0.669	0.349	0.137	0.596	0.339	0.143
	β	-0.03	-0.03	-0.03	-0.04	-0.03	-0.05
	95% CI	-0.16; 0.10	-0.08; 0.03	-0.06; 0.01	-0.19; 0.11	-0.09; 0.03	-0.12; 0.02
Adjusted (2) ^b	N _{estimates}	8	8	8	8	8	8
	N _{subjects}	11255	11253	11253	11255	11253	11253
	I ²	-	-	-	29.2	0.0	20.9
	P-value	0.962	0.369	0.094	0.959	0.369	0.120
	β	0.00	-0.02	-0.03	0.00	-0.02	-0.04
	95% CI	-0.13; 0.14	-0.08; 0.03	-0.07; 0.01	-0.19; 0.18	-0.08; 0.03	-0.08; 0.01

^aCovariates were sex, age (linear and quadratic terms), ancestry-informative principal components (if available) and genotyping centre (if necessary). ^bSame covariates than in the Adjusted (1) model, in addition to maternal education (linear and quadratic terms) and/or maternal cognition (linear and quadratic terms).

N_{estimates}: number of estimates being pooled. N_{subjects}: pooled sample size.

Supplementary Table 11. Meta-analytical linear regression coefficients (β) of the interaction between *FADS2* rs174575 or rs1535 genotypes (additive effect) with exclusive breastfeeding, having cognitive measures (in standard deviation units) as the outcome.

Model	Statistic	Fixed effects			Random effects		
		Never=0 Ever=1	Numerically- coded categories	Months	Never=0 Ever=1	Numerically- coded categories	Months
rs174575 (CC=0; CG=1; GG=2)							
Unadjusted	N _{estimates}	8	8	8	8	8	8
	N _{subjects}	11388	11386	11386	11388	11386	11386
	I ²	-	-	-	66.8	8.0	2.5
	P-value	0.937	0.924	0.938	0.563	0.986	0.977
	β	0.00	0.00	0.00	-0.04	0.00	0.00
	95% CI	-0.07; 0.07	-0.03; 0.03	-0.02; 0.02	-0.19; 0.10	-0.03; 0.03	-0.02; 0.02
Adjusted (1) ^a	N _{estimates}	8	8	8	8	8	8
	N _{subjects}	11363	11361	11361	11363	11361	11361
	I ²	-	-	-	48.0	0.0	0.0
	P-value	0.852	0.726	0.587	0.910	0.726	0.587
	β	0.01	0.01	0.01	-0.01	0.01	0.01
	95% CI	-0.06; 0.07	-0.02; 0.03	-0.01; 0.02	-0.12; 0.11	-0.02; 0.03	-0.01; 0.02
Adjusted (2) ^b	N _{estimates}	8	8	8	8	8	8
	N _{subjects}	11000	10998	10998	11000	10998	10998
	I ²	-	-	-	48.4	0.0	0.0
	P-value	0.572	0.385	0.335	0.805	0.385	0.335
	β	0.02	0.01	0.01	0.02	0.01	0.01
	95% CI	-0.05; 0.09	-0.02; 0.04	-0.01; 0.03	-0.10; 0.13	-0.02; 0.04	-0.01; 0.03
rs1535 (AA=0; AG=1; GG=2)							
Unadjusted	N _{estimates}	8	8	8	8	8	8
	N _{subjects}	11671	11669	11669	11671	11669	11669
	I ²	-	-	-	34.9	5.7	14.5
	P-value	0.772	0.896	0.772	0.747	0.883	0.699
	β	-0.01	0.00	0.00	-0.02	0.00	0.00
	95% CI	-0.07; 0.05	-0.03; 0.02	-0.02; 0.01	-0.11; 0.08	-0.03; 0.03	-0.02; 0.02
Adjusted (1) ^a	N _{estimates}	8	8	8	8	8	8
	N _{subjects}	11646	11644	11644	11646	11644	11644
	I ²	-	-	-	12.9	0.0	9.9
	P-value	0.858	0.949	0.938	0.900	0.949	0.945
	β	-0.01	0.00	0.00	0.00	0.00	0.00
	95% CI	-0.07; 0.06	-0.03; 0.02	-0.02; 0.02	-0.08; 0.07	-0.03; 0.02	-0.02; 0.02
Adjusted (2) ^b	N _{estimates}	8	8	8	8	8	8
	N _{subjects}	11255	11253	11253	11255	11253	11253
	I ²	-	-	-	26.1	0.0	0.0
	P-value	0.697	0.592	0.666	0.636	0.592	0.666
	β	0.01	0.01	0.00	0.02	0.01	0.00
	95% CI	-0.05; 0.08	-0.02; 0.03	-0.01; 0.02	-0.07; 0.11	-0.02; 0.03	-0.01; 0.02

^aCovariates were sex, age (linear and quadratic terms), ancestry-informative principal components (if available) and genotyping centre (if necessary).

^bSame covariates than in the Adjusted (1) model, in addition to maternal education (linear and quadratic terms) and/or maternal cognition (linear and quadratic terms). $N_{\text{estimates}}$: number of estimates being pooled. N_{subjects} : pooled sample size.

Supplementary Table 12. Meta-analytical linear regression coefficients (β) of the interaction

between *FADS2* rs174575 or rs1535 genotypes (dominant effect) with exclusive

breastfeeding, having cognitive measures (in standard deviation units) as the outcome.

Model	Statistic	Fixed effects			Random effects		
		Never=0 Ever=1	Numerically- coded categories	Months	Never=0 Ever=1	Numerically- coded categories	Months
rs174575 (CC=0; CG=1; GG=2)							
Unadjusted	N _{estimates}	8	8	8	8	8	8
	N _{subjects}	11388	11386	11386	11388	11386	11386
	I ²	-	-	-	59.7	0.0	0.0
	P-value	0.653	0.955	0.930	0.453	0.955	0.930
	β	-0.02	0.00	0.00	-0.06	0.00	0.00
	95% CI	-0.10; 0.06	-0.04; 0.03	-0.02; 0.02	-0.22; 0.10	-0.04; 0.03	-0.02; 0.02
Adjusted (1) ^a	N _{estimates}	8	8	8	8	8	8
	N _{subjects}	11363	11361	11361	11363	11361	11361
	I ²	-	-	-	31.2	0.0	0.0
	P-value	0.909	0.806	0.683	0.833	0.806	0.683
	β	0.00	0.00	0.00	-0.01	0.00	0.00
	95% CI	-0.09; 0.08	-0.03; 0.04	-0.02; 0.03	-0.14; 0.11	-0.03; 0.04	-0.02; 0.03
Adjusted (2) ^b	N _{estimates}	8	8	8	8	8	8
	N _{subjects}	11000	10998	10998	11000	10998	10998
	I ²	-	-	-	34.5	0.0	0.0
	P-value	0.757	0.385	0.339	0.862	0.385	0.339
	β	0.01	0.02	0.01	0.01	0.02	0.01
	95% CI	-0.07; 0.10	-0.02; 0.05	-0.01; 0.03	-0.12; 0.14	-0.02; 0.05	-0.01; 0.03
rs1535 (AA=0; AG=1; GG=2)							
Unadjusted	N _{estimates}	8	8	8	8	8	8
	N _{subjects}	11671	11669	11669	11671	11669	11669
	I ²	-	-	-	33.7	26.2	32.4
	P-value	0.714	0.903	0.809	0.815	0.902	0.949
	β	-0.02	0.00	0.00	-0.01	0.00	0.00
	95% CI	-0.10; 0.07	-0.03; 0.04	-0.02; 0.03	-0.14; 0.11	-0.04; 0.05	-0.03; 0.03
Adjusted (1) ^a	N _{estimates}	8	8	8	8	8	8
	N _{subjects}	11646	11644	11644	11646	11644	11644
	I ²	-	-	-	25.9	23.0	30.6
	P-value	0.840	0.814	0.617	0.954	0.714	0.659
	β	-0.01	0.00	0.01	0.00	0.01	0.01
	95% CI	-0.09; 0.08	-0.03; 0.04	-0.02; 0.03	-0.11; 0.12	-0.04; 0.05	-0.02; 0.04
Adjusted (2) ^b	N _{estimates}	8	8	8	8	8	8
	N _{subjects}	11255	11253	11253	11255	11253	11253
	I ²	-	-	-	36.2	20.4	18.0
	P-value	0.684	0.263	0.179	0.530	0.246	0.210
	β	0.02	0.02	0.02	0.04	0.03	0.02
	95% CI	-0.07; 0.10	-0.01; 0.05	-0.01; 0.04	-0.09; 0.18	-0.02; 0.07	-0.01; 0.04

^aCovariates were sex, age (linear and quadratic terms), ancestry-informative principal components (if available) and genotyping centre (if necessary). ^bSame covariates than in the Adjusted (1) model, in addition to maternal education (linear and quadratic terms) and/or maternal cognition (linear and quadratic terms).

N_{estimates}: number of estimates being pooled. N_{subjects}: pooled sample size.

Supplementary Table 13. Meta-analytical linear regression coefficients (β) of the interaction between *FADS2* rs174575 or rs1535 genotypes (overdominant effect) with exclusive breastfeeding, having cognitive measures (in standard deviation units) as the outcome.

Model	Statistic	Fixed effects			Random effects		
		Never=0 Ever=1	Numerically- coded categories	Months	Never=0 Ever=1	Numerically- coded categories	Months
rs174575 (CC=0; CG=1; GG=2)							
Unadjusted	N _{estimates}	8	8	8	8	8	8
	N _{subjects}	11388	11386	11386	11388	11386	11386
	I ²	-	-	-	41.1	0.0	0.0
	P-value	0.320	0.690	0.659	0.325	0.690	0.659
	β	-0.04	-0.01	-0.01	-0.07	-0.01	-0.01
	95% CI	-0.13; 0.04	-0.04; 0.03	-0.03; 0.02	-0.20; 0.07	-0.04; 0.03	-0.03; 0.02
Adjusted (1) ^a	N _{estimates}	8	8	8	8	8	8
	N _{subjects}	11363	11361	11361	11363	11361	11361
	I ²	-	-	-	0.0	0.0	0.0
	P-value	0.586	0.998	0.935	0.586	0.998	0.935
	β	-0.02	0.00	0.00	-0.02	0.00	0.00
	95% CI	-0.11; 0.06	-0.03; 0.03	-0.02; 0.02	-0.11; 0.06	-0.03; 0.03	-0.02; 0.02
Adjusted (2) ^b	N _{estimates}	8	8	8	8	8	8
	N _{subjects}	11000	10998	10998	11000	10998	10998
	I ²	-	-	-	0.0	0.0	0.0
	P-value	0.953	0.454	0.406	0.953	0.454	0.406
	β	0.00	0.01	0.01	0.00	0.01	0.01
	95% CI	-0.09; 0.08	-0.02; 0.05	-0.01; 0.03	-0.09; 0.08	-0.02; 0.05	-0.01; 0.03
rs1535 (AA=0; AG=1; GG=2)							
Unadjusted	N _{estimates}	8	8	8	8	8	8
	N _{subjects}	11671	11669	11669	11671	11669	11669
	I ²	-	-	-	19.6	28.5	45.0
	P-value	0.667	0.770	0.493	0.850	0.711	0.634
	β	-0.02	0.00	0.01	-0.01	0.01	0.01
	95% CI	-0.10; 0.06	-0.03; 0.04	-0.01; 0.03	-0.12; 0.10	-0.04; 0.05	-0.03; 0.04
Adjusted (1) ^a	N _{estimates}	8	8	8	8	8	8
	N _{subjects}	11646	11644	11644	11646	11644	11644
	I ²	-	-	-	27.1	21.9	46.5
	P-value	0.805	0.644	0.372	0.892	0.556	0.498
	β	-0.01	0.01	0.01	0.01	0.01	0.01
	95% CI	-0.09; 0.07	-0.03; 0.04	-0.01; 0.03	-0.11; 0.12	-0.03; 0.06	-0.02; 0.05
Adjusted (2) ^b	N _{estimates}	8	8	8	8	8	8
	N _{subjects}	11255	11253	11253	11255	11253	11253
	I ²	-	-	-	32.4	7.5	34.5
	P-value	0.728	0.110	0.030	0.556	0.109	0.072
	β	0.01	0.03	0.02	0.04	0.03	0.03
	95% CI	-0.07; 0.10	-0.01; 0.06	0.00; 0.05	-0.09; 0.16	-0.01; 0.07	0.00; 0.06

^aCovariates were sex, age (linear and quadratic terms), ancestry-informative principal components (if available) and genotyping centre (if necessary). ^bSame covariates than in the Adjusted (1) model, in addition to maternal education (linear and quadratic terms) and/or maternal cognition (linear and quadratic terms).

N_{estimates}: number of estimates being pooled. N_{subjects}: pooled sample size.

Supplementary Table 14. Meta-analytical linear regression coefficients (β) of the interaction between FADS2 rs174575 or rs1535 genotypes with breastfeeding (never vs. ever), having cognitive measures (in standard deviation units) as the outcome, including the UK Biobank.

Model	Study	Statistic	rs174575				rs1535			
			Recessive AA=0 AG=0 GG=1	Additive AA=0 AG=1 GG=2	Dominant AA=0 AG=1 GG=1	Overdominant AA=0 AG=1 GG=0	Recessive CC=0 CG=0 GG=1	Additive CC=0 CG=1 GG=2	Dominant CC=0 CG=1 GG=1	Overdominant CC=0 CG=1 GG=0
Unadjusted	Biobank_Axiom	N _{subjects}	21774	21774	21774	21774	21774	21774	21774	21774
		P-value	0.639	0.671	0.781	0.986	0.879	0.710	0.538	0.474
		β	-0.03	-0.01	-0.01	0.00	-0.01	0.01	0.02	0.02
		95% CI	-0.14; 0.09	-0.06; 0.04	-0.07; 0.05	-0.06; 0.06	-0.10; 0.08	-0.04; 0.05	-0.04; 0.08	-0.04; 0.08
	Biobank_BiLEVE	N _{subjects}	11068	11068	11068	11068	11068	11068	11068	11068
		P-value	0.081	0.448	0.992	0.325	0.342	0.541	0.841	0.665
		β	-0.15	-0.03	0.00	0.04	-0.06	-0.02	-0.01	0.02
		95% CI	-0.32; 0.02	-0.09; 0.04	-0.08; 0.08	-0.04; 0.13	-0.19; 0.07	-0.08; 0.04	-0.09; 0.08	-0.07; 0.10
	All (fixed effects)	N _{estimates}	10	11	11	11	11	11	11	11
		N _{subjects}	45456	45758	45758	45758	46044	46044	46044	46044
		P-value	0.600	0.650	0.962	0.607	0.425	0.836	0.486	0.282
		β	-0.02	-0.01	0.00	0.01	-0.03	0.00	0.02	0.02
		95% CI	-0.10; 0.06	-0.04; 0.03	-0.04; 0.04	-0.03; 0.05	-0.09; 0.04	-0.03; 0.04	-0.03; 0.06	-0.02; 0.07
	All (random effects)	N _{estimates}	10	11	11	11	11	11	11	11
		N _{subjects}	45456	45758	45758	45758	46044	46044	46044	46044
		I^2	75.0	42.5	29.7	43.7	30.7	1.4	0.0	0.0
		P-value	0.458	0.963	0.874	0.706	0.510	0.840	0.486	0.282
		β	0.08	0.00	0.01	0.02	-0.03	0.00	0.02	0.02
		95% CI	-0.13; 0.29	-0.06; 0.07	-0.06; 0.08	-0.07; 0.10	-0.14; 0.07	-0.03; 0.04	-0.03; 0.06	-0.02; 0.07
Adjusted (1) ^a	Biobank_Axiom	N _{subjects}	21774	21774	21774	21774	21774	21774	21774	21774
		P-value	0.547	0.549	0.672	0.949	0.933	0.700	0.511	0.446
		β	-0.04	-0.01	-0.01	0.00	0.00	0.01	0.02	0.02

	95% CI	-0.15; 0.08	-0.06; 0.03	-0.07; 0.05	-0.06; 0.06	-0.10; 0.09	-0.04; 0.05	-0.04; 0.08	-0.04; 0.08
Biobank_BiLEVE	N _{subjects}	11068	11068	11068	11068	11068	11068	11068	11068
	P-value	0.067	0.225	0.602	0.604	0.429	0.390	0.591	0.931
	β	-0.16	-0.04	-0.02	0.02	-0.05	-0.03	-0.02	0.00
	95% CI	-0.33; 0.01	-0.11; 0.03	-0.11; 0.06	-0.06; 0.11	-0.19; 0.08	-0.09; 0.04	-0.11; 0.06	-0.08; 0.09
All (fixed effects)	N _{estimates}	10	11	11	11	11	11	11	11
	N _{subjects}	45432	45731	45731	45731	46017	46017	46017	46017
	P-value	0.296	0.450	0.719	0.764	0.525	0.847	0.554	0.371
	β	-0.04	-0.01	-0.01	0.01	-0.02	0.00	0.01	0.02
	95% CI	-0.13; 0.04	-0.05; 0.02	-0.05; 0.04	-0.04; 0.05	-0.09; 0.04	-0.03; 0.04	-0.03; 0.06	-0.02; 0.06
All (random effects)	N _{estimates}	10	11	11	11	11	11	11	11
	N _{subjects}	45432	45731	45731	45731	46017	46017	46017	46017
	I^2	71.4	59.8	46.8	46.9	51.9	39.8	13.9	7.2
	P-value	0.979	0.881	0.858	0.721	0.632	0.658	0.558	0.416
	β	0.00	0.01	0.01	0.02	-0.03	0.01	0.02	0.02
	95% CI	-0.21; 0.20	-0.07; 0.08	-0.08; 0.09	-0.07; 0.10	-0.16; 0.10	-0.04; 0.07	-0.04; 0.08	-0.03; 0.07

^aCovariates were sex, age (linear and quadratic terms), ancestry-informative principal components (if available) and genotyping centre (if necessary).

N_{estimates}: number of estimates being pooled. N_{subjects}: pooled sample size.

6 – Comunicado para a imprensa

ESTUDO SUGERE QUE O ALEITAMENTO MATERNO PODE REGULAR O FUNCIONAMENTO DOS GENES

Os benefícios do aleitamento materno para a saúde da criança e da mãe são bem claros. Estudos recentes também têm sugerido uma relação entre o aleitamento e condições de saúde (como obesidade) e de capital humano (como inteligência e escolaridade) na fase adulta. Ou seja, além dos seus conhecidos benefícios para a saúde infantil, o aleitamento também parece ter efeitos mais duradouros. As recomendações internacionais recomendam que crianças devem consumir somente o leite materno até os seis meses de idade.

Apesar disso, não se sabe exatamente como o aleitamento pode trazer benefícios para o indivíduo adulto. Uma possibilidade é que o aleitamento possa alterar um fator biológico bem específico, chamado de marcações epigenéticas. Estas marcações ocorrem no material genético – os genes – que todos os seres humanos têm. Dependendo de estarem ou não “marcados”, genes podem ser ligados ou desligados. Se o aleitamento influencia em quais genes e com que frequência estas marcações ocorrem, também influenciaria o funcionamento desses genes. “O interessante é que muitas marcações epigenéticas que ocorrem após o nascimento persistem ao longo da vida. Inclusive, fatores precoces, como tabagismo materno, já foram relacionados com marcações epigenéticas duradoras. Portanto, é possível que marcações epigenéticas estejam relacionadas com os efeitos duradouros do aleitamento”, explica o biotecnologista Fernando Pires Hartwig, autor da pesquisa publicada em tese de doutorado do Programa de Pós-Graduação em Epidemiologia do UFPel, sob orientação do professor Cesar Gomes Victora.

O estudo utilizou dados sobre amamentação e marcações epigenéticas do estudo britânico *Avon Longitudinal Study of Parents and Children (ASLPAC)*, similar aos estudos de nascimentos de Pelotas. Observou-se que crianças que foram amamentadas apresentaram algumas marcações epigenéticas aos 7 anos de idade que não foram observadas em crianças que não foram amamentadas. Além disso, algumas dessas diferenças epigenéticas também foram observadas aos 15 anos de idade, sugerindo que são persistentes pelo menos até a adolescência. De acordo com os autores da pesquisa, este foi o primeiro estudo a avaliar de forma abrangente a relação entre aleitamento e epigenética. “Estudos anteriores eram escassos e de baixa qualidade. Nosso estudo foi o primeiro a investigar a relação entre aleitamento e mais de 450 mil marcas epigenéticas”, comenta o autor.

O pesquisador também aponta algumas limitações do estudo: “Nossos resultados apenas indicam que é possível que o aleitamento influencie marcas epigenéticas. Mais estudos são necessários para avaliar se nossos resultados são reproduzidos em outros grupos de crianças, bem como para investigar se as marcas epigenéticas relacionadas com aleitamento têm alguma relevância em características de saúde ou de capital humano, como a inteligência. Além disso, é possível que os efeitos duradouros da amamentação sejam explicados por fatores além da epigenética.”