VARIAÇÃO INTERINDIVIDUAL PERVASIVA NA EXPRESSÃO ALELO-ESPECÍFICA EM GÊMEOS MONOZIGÓTICOS

CRISTINA DOS SANTOS FERREIRA

UNIVERSIDADE ESTADUAL DO NORTE FLUMINENSE DARCY RIBEIRO

CAMPOS DOS GOYTACAZES-RJ

MARÇO DE 2020

VARIAÇÃO INTERINDIVIDUAL PERVASIVA NA EXPRESSÃO ALELO-ESPECÍFICA EM GÊMEOS MONOZIGÓTICOS

CRISTINA DOS SANTOS FERREIRA

Orientador: Prof. Dr. Enrique Medina-Acosta

Tese de Doutorado apresentada programa de pós-graduação em biociências e biotecnologia do centro de Biociências e Biotecnologia da Universidade Estadual do Norte Fluminense Darcy Ribeiro como parte das exigências para obtenção do título de Doutora em Biociências e Biotecnologia.

UNIVERSIDADE ESTADUAL DO NORTE FLUMINENSE DARCY RIBEIRO

CAMPOS DOS GOYTACAZES-RJ

MARÇO DE 2020

FICHA CATALOGRÁFICA
UENF - Bibliotecas
Elaborada com os dados fornecidos pela autora.

F383 Ferreira, Cristina dos Santos.

VARIAÇÃO INTERINDIVIDUAL PERVASIVA NA EXPRESSÃO ALELO-ESPECÍFICA EM GÊMEOS MONOZIGÓTICOS. / Cristina dos Santos Ferreira. - Campos dos Goytacazes, RJ, 2020.

81 f. : il.
Bibliografia: 50 - 64.

Tese (Doutorado em Biociências e Biotecnologia) - Universidade Estadual do Norte Fluminense Darcy Ribeiro, Centro de Biociências e Biotecnologia, 2020.
Orientador: Enrique Medina- Acosta.

1. expressão alelo-específica. 2. síndrome de Down. 3. co-gêmeos monozigóticos heterocariotípicos. 4. desbalanço da expressão alélica. 5. imprinting genômico. I. Universidade Estadual do Norte Fluminense Darcy Ribeiro. II. Título.

CDD - 570

VARIAÇÃO INTERINDIVIDUAL PERVASIVA NA EXPRESSÃO ALELO-ESPECÍFICA EM GÊMEOS MONOZIGÓTICOS

CRISTINA DOS SANTOS FERREIRA

Tese de Doutorado apresentada ao programa de pós-graduação em biociências e biotecnologia do centro de Biociências e Biotecnologia da Universidade Estadual do Norte Fluminense Darcy Ribeiro como parte das exigências para obtenção do título de Doutora em Biociências e Biotecnologia.

Aprovada em 21 de março de 2020

Comissão Examinadora:

Claudia Caixeta F. Andrade Coléte

Dra. Claudia Caixeta Franco Andrade Coléte (doutora em Genética) – FAMESC e UNIFAMINAS

Victor Martin Quintana Flores

Dr. Victor Martin Quintana Flores (doutor em Biociências e Biotecnologia) – UENF

Antônio Jesus Dorighetto Cogo

Dr. Antônio Jesus Dorighetto Cogo (doutor em Biociências e Biotecnologia) – UENF

Enrique Medina

Dr. Enrique Medina-Acosta (doutor em Parasitologia Médica e Molecular) – UENF
(orientador)

Agradecimentos

Agradeço a minha mãe por todo carinho e dedicação, por me ensinar as primeiras palavras e me mostrar como é bom adquirir conhecimento, por sempre acreditar e confiar em mim.

Aos meus irmãos (Beatriz, Giseli, Marcelo e Ítalo) por todo apoio, carinho e amizade, por sempre me acompanhar em todas as fases de minha vida.

Aos meus amigos presentes em todos os momentos, me ajudando a superar dificuldades.

Ao meu marido por estar sempre presente, me compreender e me fazer entender os outros.

A todas as pessoas que conheci durante esses anos de curso.

Ao meu Orientador, Prof. Dr. Enrique Medina-Acosta pelos ensinamentos, amizade e horas de discussão.

Aos colegas e amigos do NUDIM, LBR e LNCC, Thais, Alan, Rebeka, Juan, Amanda, Douglas, Ronaldo

À CAPES e CNPq pelo financiamento.

Ao revisor de minha Tese Dr. Fabrício Moreira Almeida

Aos membros da banca avaliadora Dra. Cláudia Caixeta, Dr. Victor Flores e Dr. Antônio Cogo.

RESUMO

Apesar de terem sido desenvolvidos a partir de um zigoto, os co-gêmeos monozigóticos (MZ) heterocariotípicos exibem cariótipos discordantes. Estudos epigenômicos em amostras biológicas de co-gêmeos MZ heterocariotípicos são de valor significativo para avaliar os efeitos na expressão alelo/gene específica de uma cópia cromossômica extranumerária ou disparidades cromossômicas estruturais em contribuições genéticas de linha germinativa quase idênticas. Aqui, usamos dados de RNA-Seq de repositórios existentes para estabelecer correlações dentro do par de co-gêmeos para a amplitude e magnitude da expressão alelo-específica (ASE) em co-gêmeos MZ heterocariotípicos discordantes para trissomia 21 e herança materna de 21q, bem como co-gêmeos MZ homocariotípicos. Mostramos que existe uma disparidade em todo o genoma nos sítios de ASE entre os co-gêmeos MZ heterocariotípicos. Embora a maior parte da disparidade corresponda a alterações na magnitude do desbalanço da expressão bialélica, os sítios de ASE que alternam de desbalanço estritamente monoalélico para bialélico ou o inverso ocorrem em poucos genes conhecidos ou previstos para serem imprintados, sujeitos à inativação do cromossomo X ou edição de RNA A-para-I(G). Também descobrimos diferenças comparáveis de ASE entre gêmeos MZ homocariotípicos. A extensão da discordância de ASE em gêmeos MZ (2,7%) foi cerca de 10 vezes menor do que o esperado entre pares de machos ou fêmeas não relacionados, não gêmeos. Os resultados indicam que as dissimilaridades entre pares observadas na amplitude e magnitude dos sítios de ASE nos co-gêmeos MZ heterocariotípicos não poderiam ser atribuídas apenas à aneuploidia e à herdabilidade alélica ausente na região 21q.

Palavras-chave: expressão alelo-específica, desbalanço da expressão alélica, síndrome de Down, *imprinting* genômico, co-gêmeos monozigóticos heterocariotípicos, heteroplasmia mitocondrial, expressão monoalélica aleatória, trissomia 21

ABSTRACT

Despite being developed from one zygote, heterokaryotypic monozygotic (MZ) co-twins exhibit discordant karyotypes. Epigenomic studies in biological samples from heterokaryotypic MZ co-twins are of the most significant value for assessing the effects on gene- and allele-specific expression of an extranumerary chromosomal copy or structural chromosomal disparities in otherwise nearly identical germline genetic contributions. Here, we use RNA-Seq data from existing repositories to establish within-pair correlations for the breadth and magnitude of allele-specific expression (ASE) in heterokaryotypic MZ co-twins discordant for trisomy 21 and maternal 21q inheritance, as well as homokaryotypic co-twins. We show that there is a genome-wide disparity at ASE sites between the heterokaryotypic MZ co-twins. Although most of the disparity corresponds to changes in the magnitude of biallelic imbalance, ASE sites switching from either strictly monoallelic to biallelic imbalance or the reverse occur in few genes that are known or predicted to be imprinted, subject to X-chromosome inactivation or A-to-I(G) RNA edited. We also uncovered comparable ASE differences between homokaryotypic MZ twins. The extent of ASE discordance in MZ twins (2.7%) was about 10-fold lower than the expected between pairs of unrelated, non-twin males or females. The results indicate that the observed within-pair dissimilarities in breadth and magnitude of ASE sites in the heterokaryotypic MZ co-twins could not solely be attributable to the aneuploidy and the missing allelic heritability at 21q.

Keywords: allele-specific expression, allele imbalance, Down syndrome, genomic imprinting, heterokaryotypic monozygotic co-twins, mitochondrial heteroplasmy, random monoallelic expression, trisomy 21

ABREVIATURAS

ASE - expressão alelo-específica

CNV - variação do número de cópias genômicas

DNA-Seq – Sequenciamento de DNA

eQTLs - *Expression quantitative trait loci*

eSNPs ou eSNVs - SNPs expressos

GATK - *Genome Analysis Toolkit*

GEO - *Gene Expression Omnibus*

GTEx - *Genotype-Tissue Expression*

iCpG - ilhas CpG

iPSC - Célula-tronco pluripotente induzida

lncRNA - RNA longo não codificante

MAF - Frequência do alelo menor

MAPQ - Mapeamento de qualidade

mtDNA - DNA mitocondrial individual

MZ - Geminação monozigótica

ncRNA - RNA não codificante

RNA-Seq - sequenciamento de transcriptoma

SNPs - polimorfismo de base única

SRA - *Sequence Read Archive*

STAR - *Spliced Transcripts Alignment to a Reference*

T1DS – co-gêmeo MZ heterocariotípico com a trissomia do 21

T21 - trissomia do cromossomo 21

T2N - co-gêmeo MZ heterocariotípico normal

TEA - Transtorno do espectro do autismo

UPD - dissomia uniparental

UTRs -regiões não traduzidas

Xa - cromossomo X ativo

XCI - inativação do cromossomo X

Xi - cromossomo X inativo

ÍNDICE

1. INTRODUÇÃO

As tecnologias atuais de sequenciamento de DNA genômico e RNA total (transcriptoma) geram uma grande quantidade de dados de excelente qualidade e enorme potencial de mineração, oferecendo uma visão inédita da complexidade de qualquer organismo. A velocidade com que estas novas tecnologias disponibilizam dados biológicos em larga escala cria um ambiente que favorece o desenvolvimento de estratégias necessárias para o armazenamento, acessibilidade, manejo e análises dos dados. Tais estratégias oferecem a oportunidade para viabilizar uma melhor compreensão dos sistemas biológicos em múltiplas amostras via a biocuração, isto é, a atividade de organizar, representar e fazer acessível as informações biológicas tanto aos pesquisadores quanto aos computadores para assim permitir cruzar referências com dados publicados. Entretanto, um dos maiores desafios está na interpretação de dados biológicos “ômicos” como: genoma, transcriptoma, proteoma, epigenoma, metaboloma, lipidoma, microbioma, entre outros (HASIN *et al.*, 2017; KIM e TAGKOPOULOS, 2017).

A abordagem de análise ômica, pode fornecer informações sobre os processos biológicos que são ativos entre determinado grupo de doenças, em comparação ao grupo normal (JOYCE e PALSSON, 2006; YAN *et al.*, 2017). Adicionalmente, o genoma humano é complexo e regulado em múltiplos níveis, que podem ser analisados por vários ensaios ômicos. Embora cada um desses ensaios forneça uma visão particionada deste sistema complexo, esses eventos são interdependentes e interativos (YAN *et al.*, 2017).

Neste trabalho a abordagem ômica tem como objetivo utilizar estudos de genômica e transcriptômica, a fim de identificar sítios que estejam associados a desregulação biológica que possam explicar as diferenças existentes entre gêmeos monozigóticos discordantes para trissomia do cromossomo 21 (T21).

As etapas de pré-processamento variam para dados ômicos diferentes, uma vez que os dados são obtidos a partir de diferentes plataformas/banco de dados (NARDINI *et al.*, 2015; HASIN *et al.*, 2017). Dessa forma, a integração multiômica requer a criação de um *pipeline* que integre dados gerados de diferentes plataformas, com utilização de metodologias próprias que permita a

análise de grande volume de dados, de forma adaptada a analise requerida e a tecnologia disponível (GREENE *et al.*, 2014; KIM e TAGKOPOULOS, 2017).

Neste contexto, este projeto trouxe como objetivo adicional o desenvolvimento de novos algoritmos e métodos para integração de dados ômicos. De acordo com NARDINI *et al.* (2015) tais abordagens têm sido necessárias, visto que apesar do crescente desenvolvimento computacional na área, existe uma solicitação urgente por novas metodologias, ferramentas e estrutura que permitam a interpretação das dados ômicos como um todo (WITZANY e BALUSKA, 2012). Permitindo assim, verificar como características fenotípicas se relacionam com a sequência gênica, expressão de proteínas, mRNAs e miRNAs, além da caracterização epigenômica.

Os bancos de dados quando estudados com metodologias integrativas de análise de dados oferecem uma abordagem completa, intuitiva e versátil para a representação de sistemas complexos. Tal estratégia é explorada para representar os aspectos das doenças autoimunes complexas (WHITAKER *et al.*, 2015); desordens psiquiátricas (DEUSSING e JAKOVCEVSKI, 2017); a fim de identificar moléculas de doença que podem ser tanto alvos terapêuticos eficazes, como marcadores relevantes de progressão de tumores (TORDINI *et al.*, 2016; SEREEWATTANAWOOT *et al.*, 2018); entender sistemas de regulação vegetal complexos (WANG *et al.*, 2016); além da identificação de genes imprintados e/ou doenças relacionadas ao *imprinting* genômico (PEREZ *et al.*, 2015; HILL *et al.*, 2017; VAN BAAK *et al.*, 2018).

Considerando os aspectos descritos sobre a importância e as estratégias necessárias para investigação ômica, assim como as várias aplicações que tal abordagem oferece, o presente trabalho descreverá o emprego de análises ômicas, computacionais integrativas, na investigação das diferenças transcriptômicas de gêmeos monozigóticos discordantes para T21. Para tal serão utilizados dados públicos disponíveis em repositórios de banco de dados, que serão analisados a partir de ferramentas e algoritmos que foram desenvolvidos durante este trabalho. Desta forma, dados de genômica, e transcriptômica, serão integrados.

Portanto, realizamos uma análise computacional comparativa dos dados de RNA-Seq de co-gêmeos MZ heterocariotípicos discordantes para a trissomia 21 e co-gêmeos MZ homocariotípicos. Cruzamos os sítios de ASE com repositórios de dados públicos como ClinVar (LANDRUM *et al.*, 2016), dados de heteroplasmia (SMIGRODZKI e KHAN, 2005) e análise de Gene *Ontology* (CONSORTIUM, 2008) para exemplificar as fontes e as consequências de disparidades dentro do par de co-gêmeos MZ além de anotar os efeitos de ASE em genes que estão sujeitos aos processos (epi)genéticos de *imprinting* genômico, XCI e edição RNA. Identificamos disparidade considerável de ASE entre co-gêmeos MZ heterocariotípicos ou homocariotípicos.

Em princípio, o perfil de expressão alélica em co-gêmeos MZ deve ser o mesmo entre eles devido a sua identidade genômica. Nossa hipótese de trabalho é que em indivíduos com genomas idênticos existe discordância alélica à nível de RNA que gera fenótipos diferentes. Parte da base dessa discordância alélica é devida a diferenças de *imprinting* genômico, inativação do cromossomo X e edição de RNA.

2. REVISÃO BIBLIOGRÁFICA

2.1. Gêmeos monozigóticos discordantes

A geminação monozigótica (MZ) envolve a partição de células progenitoras derivadas de um zigoto em colapso em dois conjuntos que formam dois fetos separados (co-gêmeos) de genótipos quase idênticos. Os co-gêmeos MZ se desenvolvem através da placentação monocoriônica ou dicoriônica como resultado da divisão dos conjuntos de células progenitoras. Os mecanismos exatos que desencadeiam a geminação MZ são vagos, mas fatores genéticos (LIU *et al.*, 2018), epigenéticos e ambientais têm sido implicados (KNOPMAN *et al.*, 2014).

Uma considerável quantidade de evidências experimentais demonstram que a maioria dos co-gêmeos MZ não são idênticos, mas discordante para marcas (epi)genéticas (BENNETT *et al.*, 2008; BARANZINI *et al.*, 2010; FURUKAWA *et al.*, 2013; SOUREN *et al.*, 2016) e doenças congênitas (CHAIYASAP *et al.*, 2014; HUANG *et al.*, 2019). Tipicamente, um par de co-gêmeos MZ heterocariotípicos exibe cariótipos discordantes para aneuploidias autossômicas ou gonossômicas (isto é, trissomia 21, trissomia 13, XO ou XXY) com surgimento provavelmente pós-zigoto o que leva ao mosaicismo em vários graus (GILBERT *et al.*, 2002; TACHON *et al.*, 2014). Existe um forte contraste entre os co-gêmeos MZ homocariotípicos e os co-gêmeos MZ heterocariotípicos que diferem por anomalias cromossômicas constitutivas (SCOTT e FERGUSON-SMITH, 1973; NIEUWINT *et al.*, 1999). Os co-gêmeos MZ heterocariotípicos podem ser discordantes para rearranjos cromossômicos estruturais (LEUNG *et al.*, 2009; ESSAOUI *et al.*, 2013), incluindo a variação do número de cópias genômicas (CNV) que também é comum em gêmeos MZ homocariotípicos (ABDELLAOUI *et al.*, 2015; HUANG *et al.*, 2019).

Outras causas prováveis de discordância genotípica em co-gêmeos MZ monocoriônicas incluem alterações na expressão gênica (BUIL *et al.*, 2015), efeitos origem parental dependente associados a inativação anormal não aleatória (distorcida) do cromossomo X (XCI) (ORSTAVIK *et al.*, 1995) e *imprinting* genômico (WEKSBERG *et al.*, 2002; BEGEMANN *et al.*, 2018). Existem 43 casos bem documentados de co-gêmeos MZ heterocariotípicos em

humanos (Tabela S1). A maioria dos casos relatados são gestações espontâneas, e não associadas à tecnologia de reprodução assistida. Um considerável corpo de evidências experimentais demonstra que estudos epigenômicos em co-gêmeos MZ heterocariotípicos possuem valor mais significativo para avaliar os efeitos na expressão alelo/gene específica de uma cópia cromossômica extranumerária ou disparidades cromossômicas estruturais em genomas quase idênticos.

2.2. Epigenética, o *Imprinting* genômico e a inativação do Cromossomo X

O termo epigenética refere-se ao estudo de mudanças hereditárias na expressão de genes sem que haja mudanças na sequência de DNA (EGGER *et al.*, 2004). A regulação epigenética da expressão gênica tem um papel crítico no desenvolvimento normal do indivíduo, assim como nas funções da célula. Os tipos de regulação epigenética incluem *imprinting* genômico, inativação do cromossomo X (XCI - *X-chromosome inactivation*) e expressão gênica tecido específica. (CSANKOVSZKI *et al.*, 2001; JONES e TAKAI, 2001). Uma série de processos levam a regulação epigenética, incluindo metilação do DNA; modificação na estrutura de cromatina; e RNAs não traduzidos (lncRNA, miRNA, entre outros).

A inativação do cromossomo X (ICX) é um fenômeno epigenético de compensação gênica (VAN DEN BERG *et al.*, 2009; MOREIRA DE MELLO *et al.*, 2017) que permite que em um mesmo tecido, coexistem duas populações celulares distintas em relação à origem parental cromossômica dos alelos originados do X ativo (Xa), evento que ocorre em todos os tecidos de um mesmo indivíduo (LYON, 1962; AUGUI *et al.*, 2011; WUTZ, 2011). Em cada célula comumente observa-se um perfil monoalélico de expressão dos genes no Xa, enquanto que nos tecidos o padrão de expressão é bialélico (PESSIA *et al.*, 2012), sendo raros os casos em que a expressão tecidual é monoalélica. Isso ocorre devido a inativação não aleatória dos cromossomos X em que uma cópia, materna ou paterna, é inativa na maioria das células em todos os tecidos (AMOS-LANDGRAF *et al.*, 2006).

Em humanos, cerca de 15% dos genes do cromossomo X escapam do silenciamento e são transcritos no Xa assim como no Xi (COTTON *et al.*, 2013;

VALLOT e ROUGEULLE, 2013; COTTON *et al.*, 2015). Há ainda estudos que indicam que há uma discordância no escape de modo tecido-específico entre as mulheres, uma vez que além desses 15% mais 10% escapam em algumas mulheres em determinados tecidos, porém não escapam em outras nos mesmos tecidos (BERLETCH *et al.*, 2010; COTTON *et al.*, 2013; COTTON *et al.*, 2015).

O silenciamento alélico na ICX está relacionado com a presença de três condições ou marcas epigenéticas: modificações pós-traducionais das caudas de histonas, associadas com a ativação ou a repressão da transcrição; a expressão em *cis* de um RNA não codificante (ncRNA) que recruta complexos de repressão para o Xi e a metilação em regiões ricas no dinucleotídeo CpG, as ilhas CpG (iCpG), localizadas geralmente na região promotora dos genes, impedindo que complexos de transcrição se liguem nessa região e os transcreva (COTTON *et al.*, 2015).

De modo semelhante as mesmas marcas epigenéticas são associadas ao *Imprinting* genômico, um fenômeno estabelecido de forma parental que resulta na expressão monoalélica de certos genes, de acordo com a origem parental do *imprinting* (REIK e DEAN, 2001; REIK e WALTER, 2001; KHATIB, 2007; HANNA e KELSEY, 2014). Os genes sujeitos a *imprinting* são essenciais para o desenvolvimento embrionário. Contudo a alteração na expressão de genes imprintados contribui para ocorrência de várias doenças humanas (REIK *et al.*, 2001; CONERLY e GRADY, 2010), uma vez que a perda do *imprinting* genômico pode levar a uma variedade de distúrbios (EGGERMANN *et al.*, 2015; MACKAY *et al.*, 2015).

2.3. Expressão monoalélica parental-dependente

Geralmente, ambos os alelos parentais são expressos igualmente em organismos diploides, caracterizando a expressão bialélica. A equidade da expressão parental minimiza a ocorrência de doenças genéticas recessivas (MASSAH *et al.*, 2015). No entanto, nem todos os genes apresentam perfil de expressão bialélica. Em um grupo de genes, apenas a cópia do cromossomo materno ou paterno é expressa, enquanto o outro é silenciado, caracterizando a expressão monoalélica (ZAKHAROVA *et al.*, 2009; TARUTANI e TAKAYAMA, 2011; SHIBA e TAKAYAMA, 2012). Levando em consideração a possibilidade da herança de doenças genéticas recessivas, a expressão monoalélica parece

desvantajosa, por causa do alto risco de distúrbios ou anormalidades (KANEKO-ISHINO *et al.*, 2006; HA *et al.*, 2012).

Polimorfismo e variações na expressão gênica fornecem a base genética para a variação fenotípica humana. A herança mendeliana assume que os genes dos cromossomos maternos e paternos contribuem igualmente para o desenvolvimento humano. Exceções notáveis às leis da herança mendeliana devido à expressão alelo-específica incluem a *imprinting* genômico, XCI e UPD (BJORNSSON *et al.*, 2008; TYCKO, 2010; KORIR e SEOIGHE, 2014).

A inativação do cromossomo X silencia a expressão gênica de um dos dois cromossomos do par, fornecendo uma exceção à herança mendeliana (LO *et al.*, 2003). O *imprinting* genômico especificamente, induz o silenciamento monoalélico, resultando na expressão parental-dependente do *loci*. Desta forma, a expressão gênica é influenciada pela variação genética e epigenética, além de fatores ambientais (MASSAH *et al.*, 2015).

Há evidência que as síndromes genéticas como trissomias autossômicas e dissomia uniparental (UPD), resultam do desequilíbrio de dosagem transcricional dos alelos parentais localizados nos cromossomos envolvidos. Tal desequilíbrio pode afetar diretamente ou indiretamente genes localizados ora nos mesmos cromossomos afetados ou em outros autossomos. (FITZPATRICK *et al.*, 2002; MAO *et al.*, 2005; AIT YAHYA-GRAISON *et al.*, 2007; LETOURNEAU *et al.*, 2014).

Estudos que envolvem a análise alelo-específica da expressão dos genes acabam propiciando uma maior sensibilidade para descoberta da influência direta da variação epigenética. No entanto, em heterozigotos, os estudos alelo-específico podem facilmente detectar diferenças na expressão entre os alelos e, portanto, revelar o efeito da variação genética (GIMELBRANT *et al.*, 2007; BJORNSSON *et al.*, 2008; MAYNARD *et al.*, 2008)

Geralmente, estudos da expressão alelo-específica baseiam-se no estudo de polimorfismo de base única (SNPs - *single-nucleotide polymorphisms*) para heterozigóticos dentro de mRNAs maduros (PANT *et al.*, 2006; PANOUSIS *et al.*, 2014). SNPs usados nestes estudos são marcadores frequentes para

distinguir a origem parental da expressão (MAYNARD *et al.*, 2008; SERRE *et al.*, 2008; PANOUSIS *et al.*, 2014).

2.4. Expressão alelo-específica

Os estudos de *Oligo microarray* (YAN *et al.*, 2002; LO *et al.*, 2003; MORLEY *et al.*, 2004) e estudos de sequenciamento de transcriptoma em todo o genoma (RNA-Seq) em várias amostras biológicas revelaram que muitos genes estão sujeitos a expressão transcricional diferencial de um alelo de um par de alelos (DIXON *et al.*, 2015; PIRINEN *et al.*, 2015; WEISSBEIN *et al.*, 2016). A expressão alelo-específica (ASE – *allele-specific expression*) refere-se ao afastamento da razão de expressão alélica mendeliana 1:1. Tipicamente, os padrões de expressão de alelos incluem simetricamente (estritamente) bialélicos, assimétricos bialélicos (desbalanço de expressão bialélica ou viés alélico) e estritamente monoalélicos (DIXON *et al.*, 2015; PIRINEN *et al.*, 2015; WEISSBEIN *et al.*, 2016).

A análise de RNA-Seq permite determinar a amplitude e magnitude dos sítios de ASE simultaneamente. Em uma dada condição experimental, cada tipo de célula deve exibir uma matriz de sítios de ASE, uma assinatura de ASE ou impressão digital do transcriptoma, que se espera que seja notavelmente particular para a amostra biológica individual. As assinaturas de ASE podem ser alteradas pelas condições ambientais, de saúde e de doença (MOYERBRAILEAN *et al.*, 2016; WEISSBEIN *et al.*, 2016).

Em essência, a mesma fonte de células dos co-gêmeos MZ deve exibir assinaturas de ASE idênticas. No entanto, estudos baseados na análise da sequência do transcriptoma revelaram discordância generalizada nos sítios de ASE em amostras biológicas de gêmeos MZ homocariotípicos aparentemente saudáveis (CHEUNG *et al.*, 2008; BUIL *et al.*, 2015). Portanto, no nível do RNA, a ocorrência de discordância de ASE constitui uma forma de herdabilidade enigmática, inexplicável/inexistente em indivíduos que compartilham, em princípio, genomas “idênticos”. Por outro lado, a discordância de ASE em todo o genoma implica que os mecanismos para transferência ou fluxo confiável de informações genéticas do DNA para o RNA dentro dos seres humanos são frouxos, com implicações profundas para a saúde e as doenças humanas (CHAKRAVARTI, 2011).

As causas da discordância de ASE estão associadas a fatores (epi)genéticos, interação gene-gene e ambiente-gene (Figura 1, Dataset S1). Para genes que não estão sujeitos a mecanismos reguladores epigenéticos, como *imprinting* genômico (BARAN *et al.*, 2015) e XCI (TUKIAINEN *et al.*, 2017), a ASE se relaciona principalmente aos efeitos da expressão associados aos *loci* de características quantitativas (eQTLs), que pode ser atribuído a variantes de sequência de ambos os alelos (efeito *cis*), enquanto a extensão do efeito de ASE depende de variantes genéticas *trans* e fatores ambientais que interagem com as variantes genéticas *cis* (BUIL *et al.*, 2015).

Além disso, sabe-se que mais de 2,6 milhões de sítios de ribonucleotídeos são submetidos pós-transcricionalmente à edição alelos-específica em extensões variadas em diversos tecidos humanos, contribuindo assim, em um grau muito mais alto, para a expressão fenotípica de prováveis sítios variantes na forma de epitranscriptomas diferenciais (LI *et al.*, 2011; RAMASWAMI e LI, 2014; ZHOU *et al.*, 2018). Entre os fatores genéticos, também existem diferenças na recombinação meiótica e aberrações cromossômicas (WEISSBEIN *et al.*, 2016).

2.5. Edição pós-transcricional alelo-específica

Os processos que regulam o transcriptoma humano são um dos responsáveis pelo aumento da complexidade funcional de seres eucarióticos superiores (BLENCOWE, 2006) e são determinados em parte por eventos de processamento de RNA, incluindo modificações de RNA. Um exemplo é a edição pós-transcricional alelo-específica (Edição de RNA) que altera a sequência do RNA mensageiro. Os sítios editados são amplamente distribuídos, sendo um dos responsáveis pela diversidade transcriptômica dos indivíduos (EISENBERG e LEVANON, 2018). Segundo bancos de dados de edição de RNA, mais de 2.6 milhões de sítios alvos já foram identificados no transcriptoma humano (RAMASWAMI e LI, 2014).

A edição de RNA de adenosina para inosina (A-para-I(G)) é um mecanismo pós-transcricional de regulação da expressão gênica mediado por enzimas codificadas pelos genes da família adenosina desaminase de RNA (ADAR) que catalisa a desaminação de ribonucleotídeos de adenosina convertendo-os em inosina (TAN *et al.*, 2017). Devido à similaridade entre as

bases nitrogenadas, as técnicas de sequenciamento e a maquinaria celular reconhecem a inosina como a guanina. Portanto, essas duas bases se tornam indistinguíveis nos estudos de RNA-Seq.

Em sequências codificadoras de proteínas, as edições de RNA podem causar substituições não-sinônimas de aminoácidos que podem gerar diferentes variantes proteicas. Essas edições são conhecidas como recodificações (EISENBERG e LEVANON, 2018). Embora, as recodificações sejam uma grande fonte de variações proteicas e implicações funcionais, as edições de RNA em regiões não codificantes como íntrons e regiões não traduzidas (UTRs - *untranslated regions*) são igualmente importantes para o estudo do transcriptoma. Uma vez que, edições nessas regiões podem contribuir para alteração em padrões de *splicing* gerando novas isoformas proteicas (ZHOU *et al.*, 2018).

Nos estudos de RNA-Seq, a edição de RNA é uma das variações que podem ser encontradas a nível de RNA, fortalecendo o argumento de que a edição é um dos fatores que causam variabilidade entre os indivíduos (ZHOU *et al.*, 2018). Em indivíduos não relacionados, os perfis de expressão, bem como os níveis de edição de RNA tendem a ser mais diferentes do que aqueles que se relacionam (OUYANG *et al.*, 2018). Em princípio, o perfil de expressão alélica em co-gêmeos MZ deveria ser o mesmo entre eles devido às similaridades genéticas. Contudo, existem diversas discordâncias à nível de RNA que geram fenótipos diferentes mesmo em indivíduos com genomas idênticos (BUIL *et al.*, 2015).

As edições podem configurar o perfil de expressão monoalélica, caso todos os transcritos de um determinado locus seja editado, e ainda bialélica, caso 50% dos transcritos sejam editados (Figura 1). Hipoteticamente, um sítio de edição com diferentes níveis entre os co-gêmeos MZ (por exemplo: 30% de edição para o primeiro gêmeo e 70% de edição para o segundo gêmeo) pode ser suficiente para gerar diferenças funcionais. Dessa forma, edições diferenciais podem aumentar ou diminuir o sucesso em um contexto evolutivo (LEVANON *et al.*, 2005).

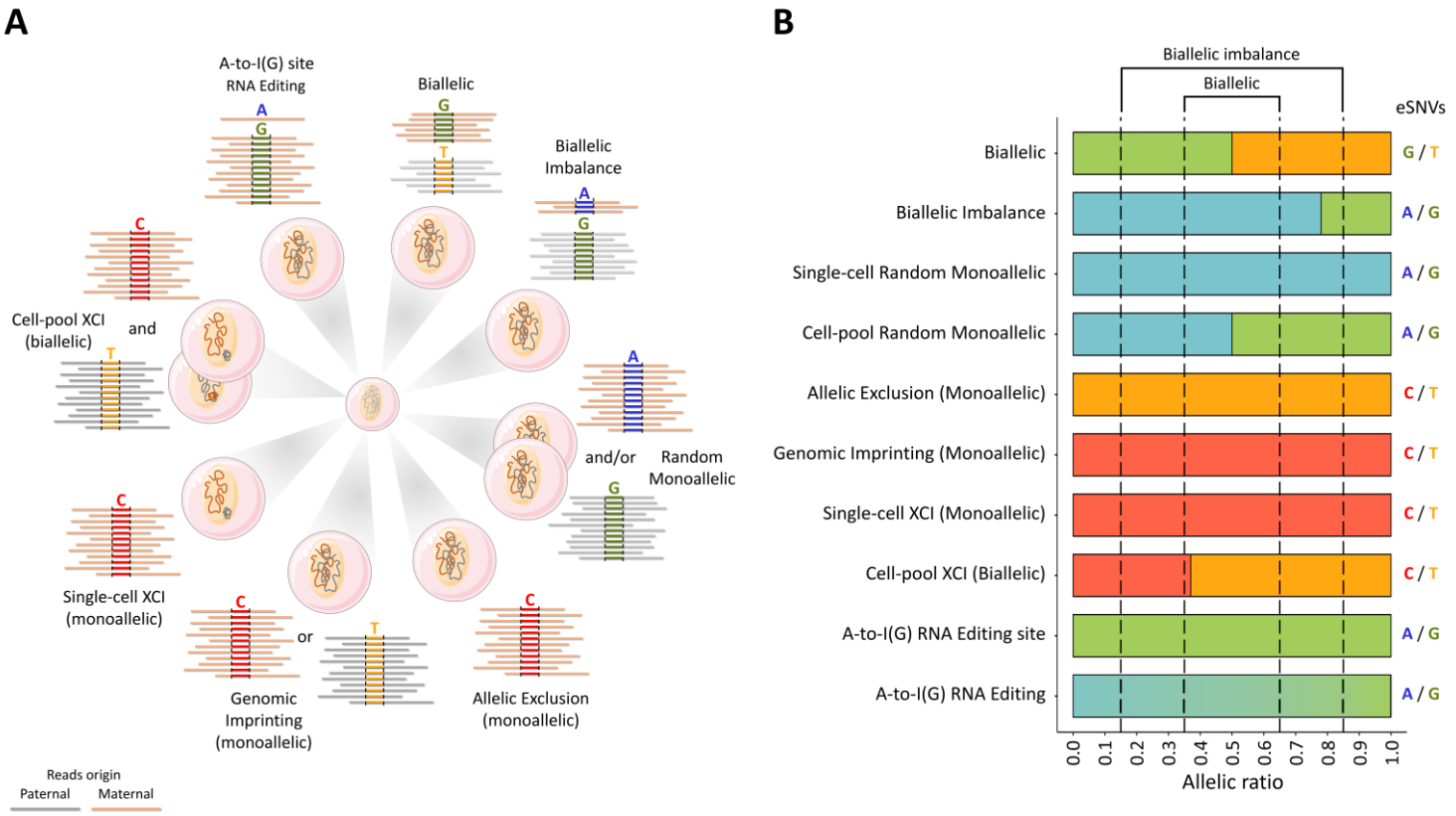


FIGURA 1. Processos epigenéticos envolvidos na expressão de RNA alelo-específico. (A) A expressão diferencial de alelos em genes reflete melhor os processos de regulação dinâmica consistentes com um alelo sendo preferencialmente silenciado ou um alelo inativo sendo restaurado. Os cenários são para RNA total no estado estacionário, para o qual são representadas por no mínimo 12 *reads* nos alelos de referência e alternativos em sítios hipoteticamente heterozigotos ou de edição de RNA A-para-I (G). A amplitude e a magnitude do desvio da razão de leitura estritamente bialélica 6:6 esperada pode ser atribuída a um dos vários processos reguladores epigenéticos que envolvem a variação compensatória e não compensatória de ação *cis*, epistática à variação de ação *trans*. Os cenários são organizados no sentido horário: estritamente bialélico, desbalanço da expressão bialélica, monoalélico aleatório, exclusão alélica, *imprinting* genômico, inativação do

cromossomo X de célula única (XCI), XCI de *pool* de células e edição de RNA. Até 30% de todos os genes autossômicos codificadores de proteínas testados estão sujeitos a expressão monoalélica aleatória clonal (mitoticamente) estável, que pode ser coordenada ou descoordenada (GIMELBRANT *et al.*, 2007; SAVOVA *et al.*, 2016a; SAVOVA *et al.*, 2016b; VIGNEAU *et al.*, 2018). Até 23% dos genes ligados ao cromossomo X são expressos a partir do X inativo (isto é, genes que escapam da XCI) e, portanto, são expressos bialelicamente em cada célula somática feminina (TUKIAINEN *et al.*, 2017). Sabe-se que cerca de 2,6 milhões de sítios de ribonucleotídeos em todo o genoma estão sujeitos à edição de RNA A-para-I (G) (RAMASWAMI e LI, 2014). Assim, os tecidos humanos são, em essência, mosaicos de expressão devido a covariáveis de ação epigenética, *cis* e *trans*. (B) A extensão da expressão alelo-específica para os cenários ilustrados no painel (A) usando RNA-Seq pode ser lido através de SNVs em genes que se sabe serem submetidos aos processos reguladores indicados. *WRB* (bialélico) (ALVES DA SILVA *et al.*, 2016; DE SA MACHADO ARAUJO *et al.*, 2018), *SH3BP5L* (desbalanço da expressão bialélica) (BARAN *et al.*, 2015), *EVC* (monoalélico aleatório) (GIMELBRANT *et al.*, 2007), *SNURF* (imprintado materno) (GRAY *et al.*, 1999; DE SA MACHADO ARAUJO *et al.*, 2018), *OR2L13* (exclusão alélica) (DE SA MACHADO ARAUJO *et al.*, 2018), *DGKZP1* e *AL391244*.3 (Edição de RNA; presente estudo), *FMR1* (sujeito a XCI) (TUKIAINEN *et al.*, 2017). Os dados que suportam as relações alélicas representadas nos histogramas são apresentados no Dataset S1.

2.6. Heteroplasmia mitocondrial

A heteroplasmia mitocondrial representa uma co-expressão determinada dinamicamente de polimorfismos herdados e patologia somática em proporções variáveis nos genomas de DNA mitocondrial individual (mtDNA) com padrões repetitivos de especificidade tecidual (MELTON, 2004).

A presença de heteroplasmia em mitocôndrias de tipos celulares distintos pode ser compreendida tendo como base a presença de padrões distintos e repetíveis da heteroplasmia do mtDNA, variando entre tipos de células diferentes do mesmo indivíduo (STEFANO *et al.*, 2017).

Nos padrões de tecido heteroplasmático, o mtDNA aparece mecanicamente mantido por trans-complementação intra-mitocondrial de nucleoides heteroplásmicos e transcritos derivados do mtDNA, assim como, a troca intercelular de mtDNA. A natureza policistrônica de transcritos codificados por mtDNA heteroplásmicos introduzem um nível adicional de complexidade que pode ser empregado para avaliar os papéis facilitadores dos padrões da heteroplasmia do mtDNA nos processos metabólicos homeostáticos (MELTON, 2004; STEFANO *et al.*, 2017).

3. OBJETIVOS

Utilizar dados de RNA-Seq obtidos a partir de repositórios existentes, para estabelecer correlações entre pares de co-gêmeos, a fim de definir a amplitude e magnitude da expressão alelo-específica em co-gêmeos MZ heterocariotípicos discordantes para a trissomia 21 e herança materna 21q, bem como co-gêmeos MZ homocariotípicos.

3.1. Objetivos específicos

Desenvolver e implementar ferramentas de análise de dados de RNA-Seq primário, a fim de obter os sítios com expressão alelo-específica;

Analisar os sítios de ASE nos pares de co-gêmeos MZ a fim de comparar a expressão alelo-específica e definir a amplitude e magnitude de ASE;

Definir se a trissomia é responsável pela discordância na amplitude e magnitude de ASE;

Cruzar os sítios de ASE discrepantes entre os co-gêmeos MZ com repositórios de dados públicos para exemplificar as fontes e as consequências de disparidades dentro do par de co-gêmeos MZ;

Verificar a influência do *imprinting* genômico, edição de RNA, inativação de Cromossomo X e microheteroplasmia mitocondrial na amplitude e magnitude da discrepância ASE entre os gêmeos.

4. METODOLOGIA

4.1. BioProjects

Utilizamos dados primários (não processados) de sequência de RNA dos experimentos públicos de SRA (*Sequence Read Archive*) em 10 pares gêmeos, sendo um par de gêmeos heterocariotípicos e nove pares de gêmeos homocariotípicos. As amostras biológicas incluíram: fibroblastos fetais primários (GEO BioProject PRJNA239814) do estudo de LETOURNEAU *et al.* (2014), células-tronco pluripotentes induzidas (iPSC) do estudo de (HIBAOUI *et al.*, 2014) (GEO BioProject PRJNA227902) e células B cultivadas (linhas celulares linfoblastóides transformadas pelo vírus *Epstein-Barr* a partir de linfócito B de sangue adulto periférico; GEO BioProject PRJNA170210) (Dataset S2). Selecionamos o estudo do transcriptoma de LETOURNEAU *et al.* (2014) em um par de co-gêmeos MZ que eram cariotipicamente discordantes para a trissomia 21 (T21) de origem materna (DAHOUN *et al.*, 2008) e, portanto, são gêmeos heterocariotípicos (isto é, co-gêmeos que diferem em relação a anomalias cromossômicas constitutivas), tais gêmeos apresentaram desenvolvimento embrionário monocoriônico e diamniótico (DAHOUN *et al.*, 2008).

A transcriptômica comparativa nesses gêmeos heterocariotípicos leva à proposta dos chamados domínios de desregulação da expressão gênica em todo o genoma na síndrome de Down (LETOURNEAU *et al.*, 2014). O caso é emblemático porque, além da aneuploidia materna discordante T21, os fibroblastos fetais primários dos gêmeos MZ exibiram herdabilidade alélica ausente no 21º trimestre como resultado de evento(s) de recombinação (DAHOUN *et al.*, 2008). Um diagrama da herança 21q materna discordante no par de co-gêmeos heterocariotípicos para trissomia 21 está representado na Figura S1. Para estimar a extensão e magnitude da discordância de ASE em indivíduos não relacionados e não gêmeos, incluímos os experimentos de RNA-Seq do BioProject PRJNA316578 (Dataset S2), que inclui amostras de sangue total de dois homens e duas mulheres, com idade média de 34 anos controles saudáveis.

4.2. Identificação, quantificação e classificação de sítios de expressão alelo-específicos em dados do transcriptoma

Implementamos o PipASE, um pipeline computacional para identificar, quantificar e classificar sítios de ASE nos dados de transcriptoma (Figura S2). O PipASE varre todo o genoma em busca de variantes expressas de nucleotídeo único (eSNVs) em leituras alinhadas de alta qualidade. Reconhecemos que as contagens de leitura de RNA-Seq, talvez sejam artefatos discordantes entre co-gêmeos como resultado da química de sequenciamento e de viés de cadeia reversa/frente na taxa de erro da tecnologia de sequenciamento de alto rendimento (HEAP *et al.*, 2010; PICKRELL *et al.*, 2012; LIU *et al.*, 2014; SODERLUND *et al.*, 2014; HU *et al.*, 2015; WOOD *et al.*, 2015; RAGHUPATHY *et al.*, 2018; RICHARD ALBERT *et al.*, 2018). Portanto, fontes primárias de artefatos técnicos, como erros sistemáticos no sequenciamento e na leitura de sequências de mapeamento para um genoma haplóide de referência, foram restringidos ao incluir no PipASE os seguintes algoritmos específicos que reduzem ou controlam o viés de mapeamento: i) relaxando o número de desencontros admitidos por *string*, mas excluindo leituras com incompatibilidades espúrias nas últimas bases de leituras alinhadas apenas a uma fita de DNA; ii) excluir leituras alinhadas em torno de inserções, exclusões e repetições simples em tandem; iii) excluir o mapeamento de leituras para regiões parálogas (isto é, duplicações segmentares); iv) exigir ≥ 12 profundidade de leitura de alta qualidade para chamar um sítio informativo candidato; e v) priorizar a classificação dos sítios de ASE por vários padrões de expressão consistentes.

As leituras brutas foram cortadas com *Trimmomatic* (BOLGER *et al.*, 2014) e alinhadas ao genoma de referência hg38 usando o software *Spliced Transcripts Alignment to a Reference* (STAR, v3.5a) (DOBIN *et al.*, 2013). Exigimos leituras mapeadas exclusivas e de alta qualidade (MAPQ ≥ 30), filtrando-as usando as ferramentas de alinhamento/mapa de sequência (SAMtools) (LI *et al.*, 2009). Processamos os dados de RNA-Seq de acordo com as diretrizes de melhores práticas, usando a ferramenta *ASEReadCounter* do *Genome Analysis Toolkit* de código aberto (GATK, v3.8), instrumentada para a descoberta de variantes em dados de sequenciamento de alto rendimento

(MCKENNA *et al.*, 2010; DEPRISTO *et al.*, 2011; VAN DER AUWERA *et al.*, 2013). Polimorfismos de nucleotídeo único anotado (SNP) e SNVs privados foram identificados usando o *HaplotypeCaller* da GATK em cada posição heterozigótica hipotética de acordo com *HapMap* (International HapMap, 2003) e banco de dados de SNPs (SHERRY *et al.*, 2001). A anotação das posições do sítio da variante de ASE no genoma de referência hg38 foi realizada usando o pacote R / Bioconductor biomaRt (DURINCK *et al.*, 2005; DURINCK *et al.*, 2009). Os dados de SNP da população (MAF, alelo ancestral) foram integrados usando o pacote rsnps versão 0.3.0 (CHAMBERLAIN *et al.*, 2018). Para a avaliação de ASE, as contagens de leitura das réplicas foram amalgamadas e os valores de Q1 em cada sítio informativo de eSNV foram calculados para todas as amostras por gêmeo.

Para sítios de ASE que ocorreram apenas uma vez em cada conjunto de amostras, o valor de ASE foi fornecido pela execução informativa. Assim, os sítios de ASE são suportados por pelo menos uma execução informativa. Por exemplo, o BioProject PRJNA239814, que se refere as amostras de fibroblastos fetais coletados a partir do par duplo MZ discordante para trissomia 21, que compreende 12 experimentos com RNA-Seq, sendo seis por gêmeo. O projeto inclui quatro amostras para cada gêmeo, duas das quais são réplicas. Para esse projeto, a distribuição de sítios informativos de ASE é de 51,7; 18,2; 18,5 e 11,6% de sítios suportados por pelo menos 1, 2, 3 e 4 amostras, respectivamente.

O ASE nos sítios de SNP heterozigotos imputados foi calculado como a diferença das contagens de leitura de RNA-Seq entre os dois alelos, usando a equação ASE = | 0.5 - Ref _allele_read count / (Ref _allele_read count / (Alt _allele _read count + Alt _allele _read count) | O valor do desbalanço da expressão alélica por sítio (variando entre 0 e 0,5) é, portanto, uma medida de afastamento da razão de expressão alélica mendeliana 1:1 esperada (BABAK *et al.*, 2015; BARAN *et al.*, 2015). Anotamos os dados de ASE calculando as referências nulas/razões alternativas esperadas e os valores *p* do teste binomial (WANG e CLARK, 2014) usando o teste binomial em função do código R (CORE TEAM, 2019) e de acordo com o contexto de sequência da estrutura gênica (exon, intron, 5´ UTR, 3´ UTR e intergênico) usando o GRCh38.92 Ensembl

release 96 na extensão *gtf (Gene transfer format)* e o *GenomicFeatures* pacote de anotação no código R (LAWRENCE *et al.*, 2013).

O teste estatístico I-quadrado foi utilizado para avaliar o grau de heterogeneidade nos perfis de ASE de genes suportados por vários eSNVs. O teste é baseado nos valores do qui-quadrado e do grau de liberdade e foi usado para medir a inconsistência dos perfis de ASE em cada gene. Os genes foram classificados de acordo com os seguintes critérios: homogeneidade (I-quadrado <30%), heterogeneidade moderada (entre 30 e 50%), heterogeneidade substancial (entre 50 e 75%) e considerável heterogeneidade (> 75%). Os valores negativos do I-quadrado foram considerados como 0% (WANG e CLARK, 2014; VON HIPPEL, 2015). Um fluxograma para o PipASE usado para digitalizar e classificar diferenças específicas de alelos em todo o genoma entre os co-gêmeos MZ é mostrado na Figura S2.

4.3. Referência cruzada com repositórios de dados públicos

Para cada sítio de ASE observado em cada amostra de RNA-Seq, extraímos informações funcionais por referência cruzada computacional com bancos de dados públicos sobre alelos variantes patogênicos de alteração de expressão ou de perda de função patogênica (ADZHUBEI *et al.*, 2010; LANDRUM *et al.*, 2016; VASER *et al.*, 2016), genes imprintados (JIRTLE e MURPHY, 2012; WEI *et al.*, 2014; BARAN *et al.*, 2015; PIRINEN *et al.*, 2015), sítios de edição de RNA A-para-I (G) (RAMASWAMI e LI, 2014), sítios discordantes de ASE da linha germinativa em gêmeos MZ (CHEUNG *et al.*, 2008) e genes que escapam e não escapam da XCI (CARREL e WILLARD, 2005; COTTON *et al.*, 2013; BALATON *et al.*, 2015; COTTON *et al.*, 2015; TUKIAINEN *et al.*, 2017; GARIERI *et al.*, 2018; SHVETSOVA *et al.*, 2019; WAINER KATSIR e LINIAL, 2019). Os perfis de expressão alélica foram validados computacionalmente pela integração de dados com os perfis de ASE observados em vários tecidos humanos do projeto *Genotype-Tissue Expression* (GTEx) (CONSORTIUM, 2015), usando a ferramenta Data Integrator disponível no UCSC Genome Browser, que contém *track hubs* para os dados da segunda fonte GTEx (release V6, outubro de 2015), principalmente conforme relatado anteriormente (DE SA MACHADO ARAUJO *et al.*, 2018).

4.4. Edição canônica de ácido ribonucleico A para I (G)

Os sítios de ASE foram consultados no banco de dados RADAR, que inclui uma lista de cerca de 2,6 milhões de bancos de dados rigorosamente anotados de sítios de edição de RNA A-para-I (G). Para referência cruzada dos sítios de ASE, mesclamos a versão 1 dos dados RADAR (disponível on-line no navegador RADAR) e a versão 2, que é baseada no conjunto de dados GTEx RNA-Seq de 30 tecidos (hg19; versão 6p) e relata o RNA níveis de edição para sites com ≥20 leituras (TAN *et al.*, 2017), gentilmente fornecidos como um banco de dados simples pelo Dr. Jin Billy Li na Universidade de Stanford (RAMASWAMI e LI, 2014). As coordenadas hg19 foram elevadas para hg38 usando "hg19ToHg38. função over.chain" e scripts R com base nas bibliotecas *AnnotationHub* (MORGAN, 2017) e *rtracklayer* (LAWRENCE *et al.*, 2009). Limitamos a análise às posições de base correspondentes às variantes canônicas de A-para-I (G), excluindo todos os SNVs que mapeiam dentro de duplicações segmentais ou repetições simples no genoma de referência hg38, usando a ferramenta *Short Match* com sequências de caracteres de 50 bases de comprimento contendo a variante na posição 26º.

A etapa de seleção do filtro acima seguiu as diretrizes de qualidade publicadas (LI *et al.*, 2011; PISKOL *et al.*, 2013; RAMASWAMI e LI, 2014). Para cada sítio de ASE que corresponde a um local de sítio de edição de referência RADAR, calculamos os níveis de edição de RNA A-para-I (G) como a proporção de leituras contendo G dividida pela soma de leituras contendo A e G nos experimentos de RNA-Seq de cada par de gêmeos. A força da coassociação entre os níveis de edição de RNA nos sítios de ASE em pares duplos foi medida usando modelos lineares em R.

5. RESULTADOS

5.1. Diferenças alelo-específicas em todo o transcriptoma observadas em co-gêmeos monozigóticos discordantes para a trissomia 21 e a recombinação

Recombinação e variação de sequência são as principais fontes evolutivas de diversidade no genoma humano. Portanto, desejamos primeiro avaliar como essas duas forças impactaram a ASE em co-gêmeos "idênticos". Entre os co-gêmeos MZ discordantes para T21, identificamos 1.227 (3,8%) sítios de ASE cujos padrões alélicos eram discordantes (isto é, monoalélicos versus bialélicos) em fibroblastos e 3.295 (6%) desses locais em iPSC (Figura 2A, Dataset S3).

Estimamos a magnitude da mudança de expressão entre as condições para as variantes chamadas (Figura 2B). A maior parte dos sítios de ASE exibiu um valor de LogASE próximo a zero, o que significa que a maioria dos sítios de ASE não foi alterada na condição de trissomia 21. É importante ressaltar que 19 eSNVs foram significativamente alteradas nos fibroblastos dos gêmeos afetados pela trissomia 21, sendo 16 sítios com LogASE ≥ 0,8 e três sítios com LogASE ≤ -0,8.

Vale ressaltar que 11 genes implicados mapeados para a região 21q discordam quanto à herança materna devido a um evento de recombinação. Entre esses genes, *CASP6, FAM86GP* e *PDXDC1/PKD1P6* foram expressos monoalelicamente, enquanto o gene *IL17RA* foi expresso bialelicamente em fibroblastos do gêmeo T21. Em iPSC, observamos 260 eSNVs com LogASE ≥0,8 e 214 ≤ −0,8 anotado em 274 genes (Figura 2B). Dos 19 sítios de ASE com valores de LogASE ≤ -0,8 ou ≥ 0,8 em fibroblastos, 14 também foram chamados em iPSC. No entanto, apenas 10 sítios foram alterados em ambos os tipos de células com valores de LogASE ≥ 0,8 (Dataset S3) e estão localizados na região 21q, abrangendo o evento de recombinação.

A distribuição geral dos genes pelo número de sítios de ASE observados em fibroblastos e iPSC é mostrada na Figura 2C. As discrepâncias de ASE entre os co-gêmeos MZ discordantes para T21 observadas nos dois fibroblastos (Figura 3A) e iPSC (Figura S3A) foram generalizadas no genoma (média de 20

sítios de ASE por Mb). Validamos o status heterocariotípico dos gêmeos MZ discordantes para T21 comparando as razões alélicas globais dentro do par e as plotamos como cariótipos de expressão (e-cariótipos) (Figura 3B e Figura S3B).

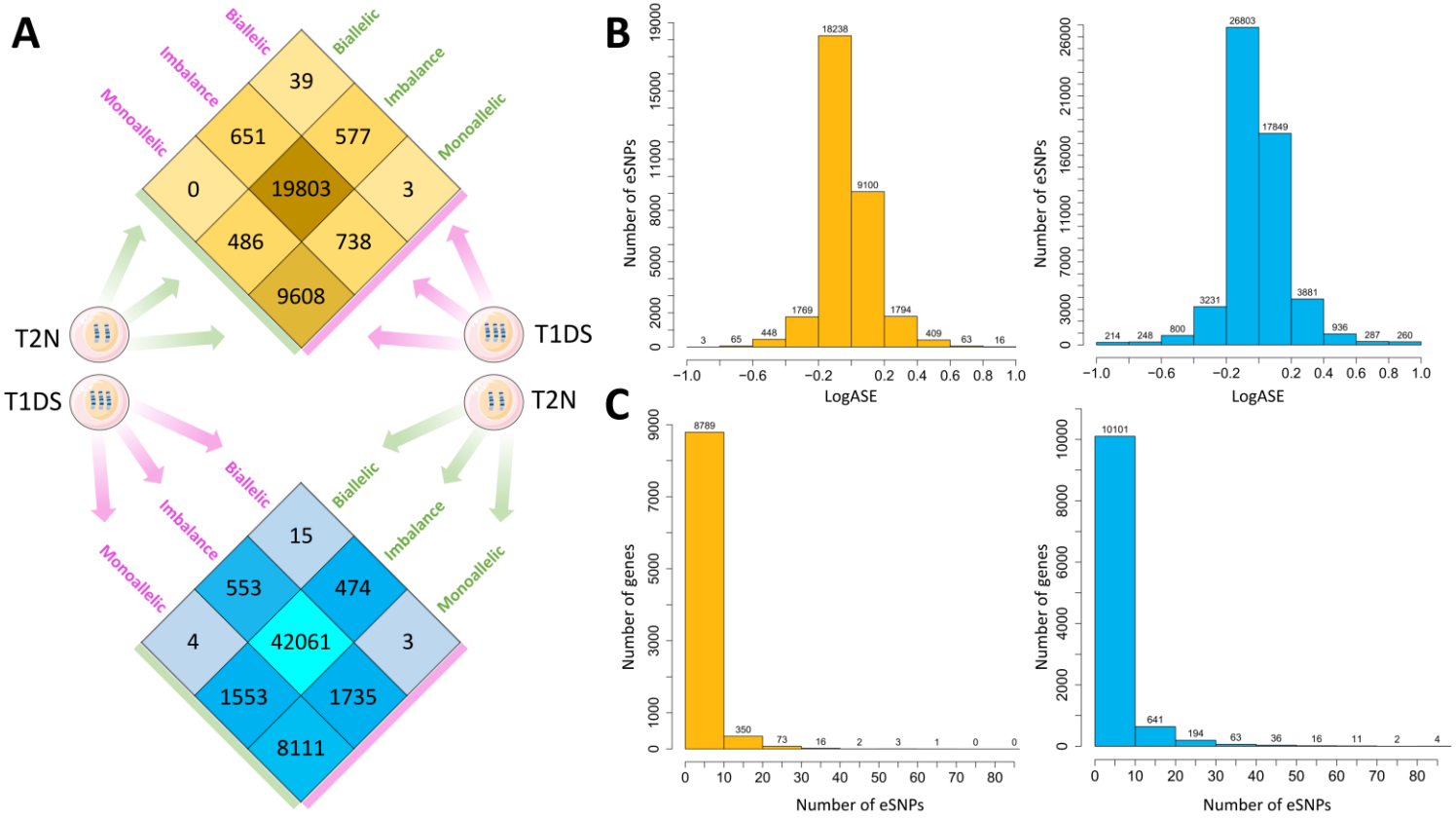
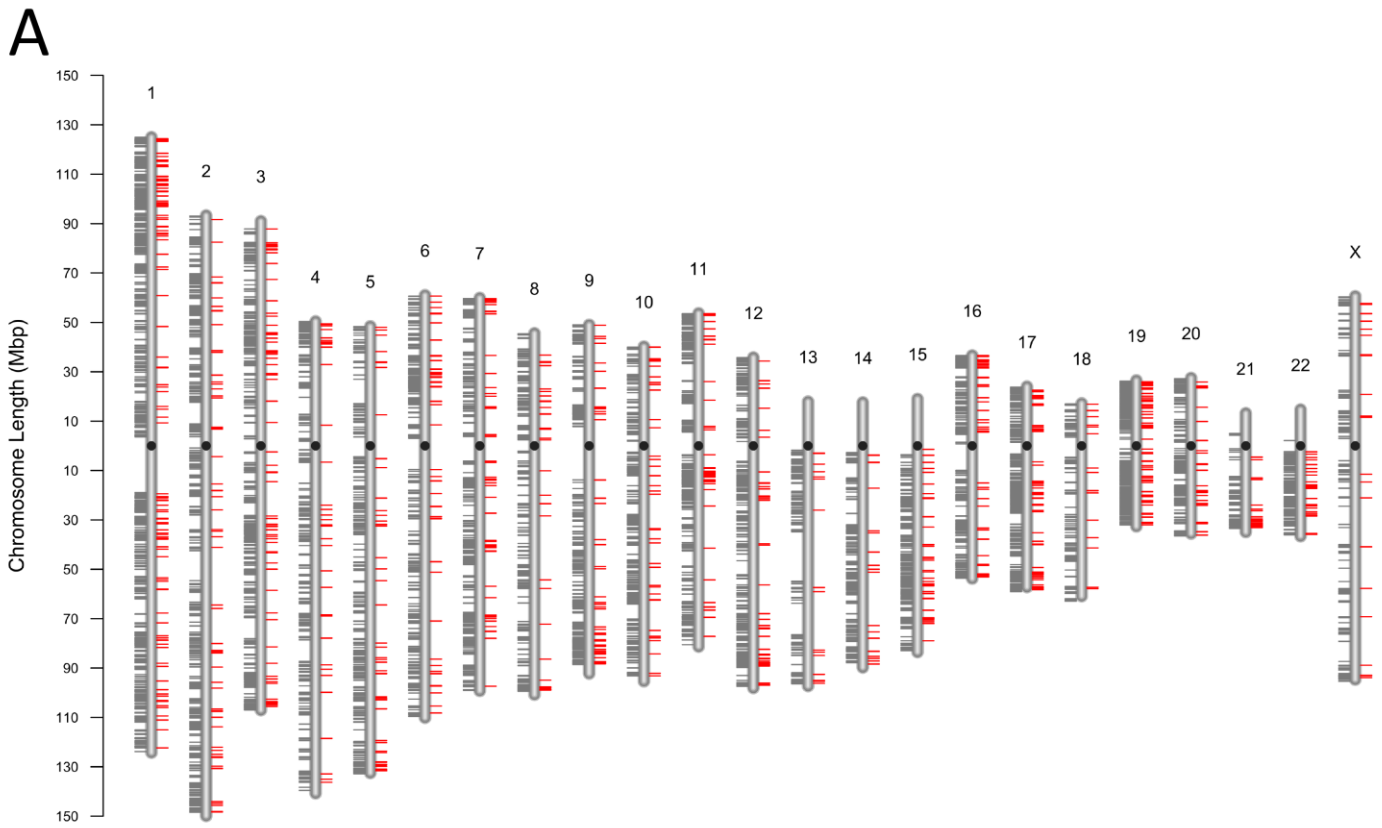


Figura 2 Visão geral da amplitude e magnitude da disparidade de expressão alelo-específica entre gêmeos monozigóticos (MZ) heterocariotípicos. (A) Número de sítios de expressão alelo-específica (ASE) distribuídos pelo status dentro do par de concordância ou discordância em gêmeos MZ heterocatotípicos para trissomia 21 e discordantes para herança materna 21q testada em fibroblastos primários

(painel superior no gráfico de calor laranja) e iPSC (painel inferior no gráfico de calor azul). Em cada tipo de célula, a maioria dos sítios de ASE é concordante com o status de desbalanço da expressão bialélica nos co-gêmeos com trissomia 21 (T1DS) e normal (T2N). Em média, os co-gêmeos são discordantes em 2.261 ± 1.462,3 sítios de ASE. (B) Comparação do tamanho do efeito do LogASE entre fibroblastos e iPSC, respectivamente. Calculamos o log2 da alteração da dobra de expressão específica do alelo usando a equação LogASE = log2 (T1DS _ ASE / T2N _ ASE) para cada variante de nucleotídeo única expressa em cada tecido. O LogASE estima a magnitude da mudança de expressão entre as condições da variante. (C) Distribuição dos genes pelo número de sítios de ASE observados em fibroblastos (barras laranja) e iPSC (barras azuis).



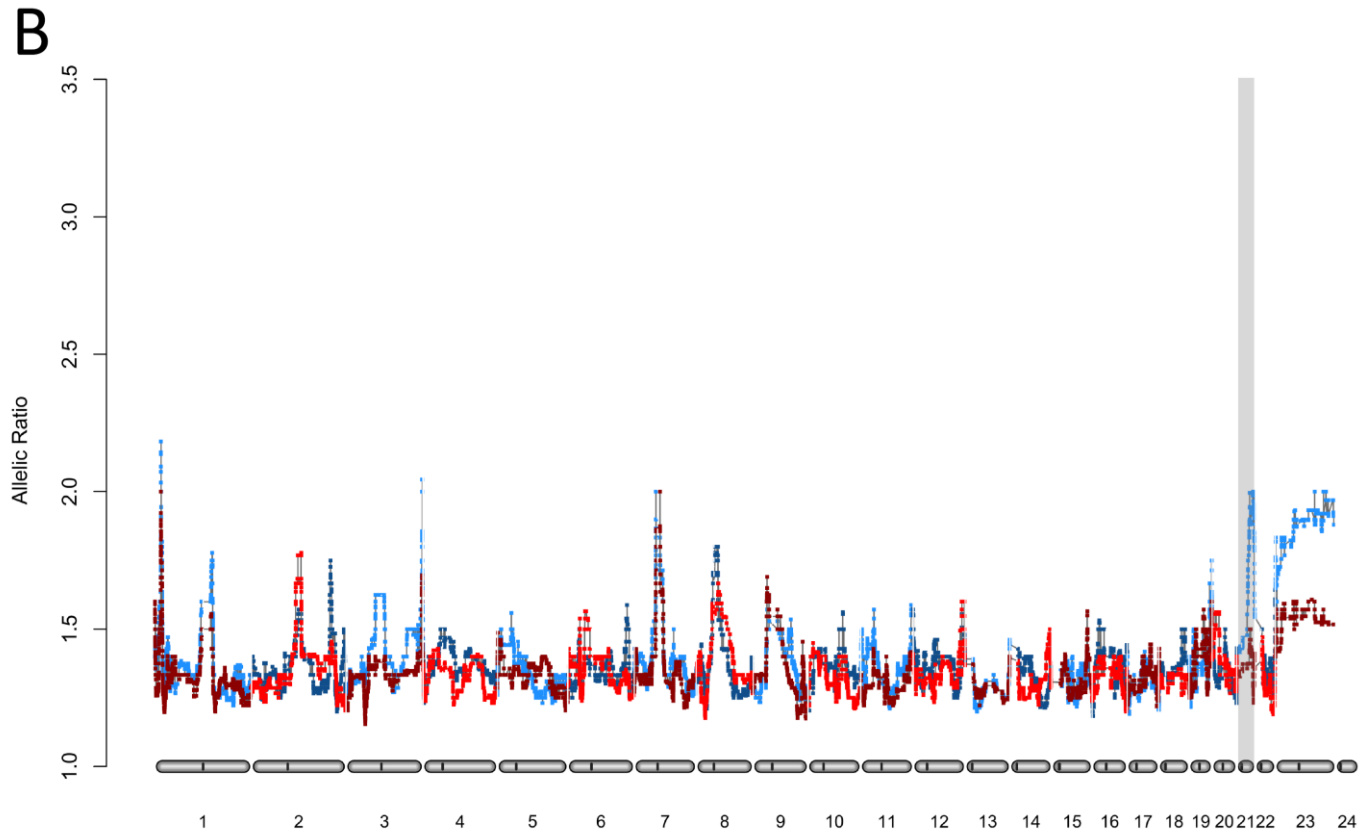


Figura 3. Distribuição cromossômica de variantes de nucleotídeo único expressas. (A) e-cariotipagem em todo o genoma para os SNPs e variantes que exibem expressão alelo-específica em fibroblastos fetais primários dos co-gêmeos discordantes para T21 e recombinação materna em 21q. É mostrada a distribuição de todos os sítios de ASE que foram concordantes (marcações cinza na esquerda de cada ideograma cromossômico) ou discordantes (marcações vermelhas no lado direito). (B) Detecção de trissomia 21 por e-cariotipagem de viés alélico usando dados de RNA-Seq de fibroblastos primários em (A). O sombreamento cinza destaca a ocorrência de uma terceira cópia discordante do cromossomo 21 em um gêmeo.

5.2. Disparidade de expressão alelo-específica observada em pares de gêmeos homocariotípicos

Para começar a resolver as causas prováveis da discordância generalizada de ASE encontrada em co-gêmeos, examinamos a amplitude e magnitude dos sítios discordantes de ASE em nove pares de co-gêmeos não discordantes para aneuploidia e recombinação. Surpreendentemente, a amplitude e magnitude da concordância e discordância de ASE nos pares gêmeos controle foram comparáveis àquelas observadas no par gêmeo heterocariotípico, com uma média de 1.074 sítios discordantes (2,7%) por par gêmeo (Figura S4). Os sítios de ASE discordantes também foram distribuídos em todo o genoma (Figura S5). Apesar de suas origens parentais diversas, havia, em média, 19.488 sítios de ASE comuns nos nove pares MZ homocariotípicos; 90 (0,46%) sítios foram discordantes em todo o conjunto de pares de gêmeos. No entanto, havia, em média, 571 sítios de ASE discordantes em um dado par gêmeo, mas concordantes em outro. Os sítios recorrentes em todos os nove pares refletem melhor a identidade por estado (Dataset S2).

Para todo o conjunto de pares de gêmeos MZ, a distribuição média de eSNVs por gene foi a seguinte: 34,3% (n = 3.162) dos genes foram chamados por um eSNV; 57,8% (n = 5.333) foram apoiados por 2 a 10 eSNVs; 7,9% (n = 729) foram chamados por 11 a 200 eSNVs; 0,02% (n = 2,4) foram chamados por 201 a 500 eSNVs e 0,01% (n = 1,1) exibiram> 500 eSNVs (Dataset S4A e S4B). Nós carregamos um teste estatístico de heterogeneidade para pesquisar efeitos de intervenção (variação nas estimativas de efeito além do acaso) em uma determinada região genômica. Para todo o conjunto de amostras, descobrimos, em média, que 43,6% (n = 2.619) dos genes suportados por vários eSNVs exibiram considerável homogeneidade entre os perfis de eSNV; 3,8% (n = 225) apresentaram heterogeneidade moderada; 6,1% (n = 366) apresentaram heterogeneidade substancial; e 46,5% (n = 2.795) apresentaram considerável heterogeneidade (Dataset S4C e S4D). Observamos que os genes que exibem considerável heterogeneidade são grandes (em média 130 Kbp, isto é, *CD226*) e são suportados em média por 7,2 (intervalo de 2 a 318) eSNVs. Por outro lado, os perfis mais homogêneos estão em genes com um tamanho médio de 12 Kbp (ou seja, *JRK*), que são suportados em média por 3,7 eSNVs (intervalo de 2 a

38 sítios). Além disso, comparando genes suportados pelo mesmo número de eSNV (ou seja, 30 sítios), notamos que os eSNVs são distribuídos de maneira diferente, em direção ao 3´UTR em genes classificados como homogêneos (por exemplo, *LGALS8* e *PLEC*) e espalhados pelo corpo do gene naqueles classificados como heterogêneos (isto é, *CD226* e *GLEC17A*).

Também validamos o status homocariotípico dos nove pares gêmeos MZ controle comparando as razões alélicas globais dentro do par e os plotamos como e-cariótipos (Figura S6). Juntas, as análises de cariotipagem de expressão demonstram que há herdabilidade alélica difusa em falta entre o transcriptoma de co-gêmeos MZ e que a maior parte das disparidades dentro do par de ASE nos co-gêmeos heterocariotípicos não pode ser atribuída apenas à ocorrência diferencial de aneuploidia e herdabilidade alélica ausente em 21q.

5.3. Disparidade de expressão alelo-específica observada em homens e mulheres não relacionados, não gêmeos

Homens e mulheres não relacionados, não gêmeos, exibiram extensões comparáveis de discordância em ASE em todo o genoma: 24,8% (6.546/26.371 sítios de eSNV) nos machos (Dataset S5A) e 25.57% (5.992/23.431 sítios de eSNV) em fêmeas (Dataset S5B). Portanto, a extensão da discordância de ASE em homens e mulheres não gêmeos não relacionados é cerca de 10 vezes maior do que a observada entre pares de gêmeos MZ (2,7%). No conjunto masculino e feminino não relacionado, 47,4 e 45,4% dos genes suportados por ≥ 2 eSNVs, respectivamente, exibiram perfis de ASE consideravelmente heterogêneos, enquanto 43,4 e 45,5% dos genes foram classificados como consideravelmente homogêneos (Dataset 5C). Semelhante à descoberta em gêmeos MZ, os genes que exibem considerável heterogeneidade são grandes (em média 83 Kbp, isto é, *GAK* em homens e 221 Kbp em mulheres, isto é, *SAMD3*) e são suportados, em média, por 8,2 (intervalo 2 a 104) eSNVs em machos e 7,7 (intervalo de 2 a 106) eSNVs em fêmeas. Por outro lado, os perfis mais homogêneos estão em genes com tamanho médio de 18 Kbp (ou seja, *UBE2I* nos machos e *EEF1D* nas fêmeas), que são suportados em média por 3,7 (variação de 2 a 39) eSNVs nos machos e 3,5 (2 a 36) eSNVs em mulheres. Novamente, comparando genes suportados pelo mesmo número de eSNVs (ou seja, 25 sítios), notamos que os eSNVs estão distribuídos de maneira diferente, em direção ao 3´UTR em genes

classificados como homogêneos (exemplo, *HCP5* e *PRRC2B* em homens e *AC004151*.1 e *NOTCH1* no sexo feminino) ou espalhado ao longo do corpo do gene naqueles classificados como heterogêneos (isto é, *FCGBP* e *GAK* no sexo masculino e *SAMD3* e *SYNE3* no sexo feminino).

5.4. Avaliação das causas subjacentes da herança alélica desaparecida e pervasiva observada

As causas subjacentes da falta de herança alélica generalizada observada podem incluir i) variações da sequência de DNA em todo o genoma dentro de pares de co-gêmeos MZ, como suportado por descobertas recentes em gêmeos MZ discordantes para transtorno do espectro do autismo (TEA) usando sequenciamento de genoma inteiro (HUANG *et al.*, 2019) e ii) expressão diferencial de alelos. Dado que nenhum dos dez pares gêmeos MZ aqui mencionados possui sequências genômicas disponíveis em repositórios públicos, primeiro cruzamos os sítios de ASE observados com dados sobre a distribuição de eSNVs relatados entre os co-gêmeos MZ discordantes para TEA (HUANG *et al.*, 2019).

Em média, os co-gêmeos MZ discordantes para TEA exibiram 54 disparidades de eSNVs anotados em exons, 3.912 em íntrons, 13 em 5´ UTR e 74 em 3´ UTR para 2.786 genes (Tabela S2). Surpreendentemente, entre os co-gêmeos MZ heterocariotípicos para T21 ou os co-gêmeos MZ homocariotípicos, identificamos, em média, 10.111 sítios discordantes de ASE em exons anotados, 8.037 em íntrons, 2.066 em 5´ UTR e 18.032 em 3´UTR de 18.432 genes. Assim, um aumento médio de 120 vezes em sítios de ASE discordantes por categoria de anotação. Essa diferença não pode ser atribuída apenas à taxa média de distribuição de eSNVs discordantes de $1{,}1x10^{-4}$ por local exônico relatados em genes humanos entre os genomas de co-gêmeos MZ (HUANG *et al.*, 2019).

Além disso, há um déficit de 20 vezes nos locais ASE anotados em regiões intergênicas, em comparação com o número de locais eSNV discordantes pelo sequenciamento genômico completo, o que apoia a visão de que a distribuição tendenciosa das discordâncias dos sítios de ASE nos genes pode ser biologicamente relevante. Também validamos alguns dos sítios discordantes de ASE por referência cruzada com os conjuntos de sítios de ASE em gêmeos MZ do estudo de CHEUNG *et al.* (2008) (Dataset S3).

Em seguida, cruzamos os sítios de ASE com dados sobre genes conhecidos ou previstos para serem expressos a partir de um alelo por vez através do *imprinting* genômico, XCI e edição de RNA de A-para-I (G). No geral, identificamos sítios de ASE discordantes em 205 genes imprintados conhecidos ou candidatos (Dataset S3), 12 genes ligados ao X (Tabela S3) e 3.955 sítios provavelmente sujeitos à edição de RNA de A-para-I (G) (Dataset S3).

5.5. Alternância de expressão alelo-específica em genes imprintados

Observamos que, em média, 4.574 sítios de ASE eram concordantes monoalelicamente entre co-gêmeos. A anotação desses sítios revelou que 8.867 genes exibiram vários eSNVs monoalélicos sem sítios expressos bialelicamente (Dataset S3, S6-S14). Entre esses genes, anotamos cinco genes imprintados conhecidos (*DR1, BRD2, VARS2, MEG3* e *H19*), cada um classificado com ≥8 eSNVs. A referência cruzada desses genes com dados secundários do projeto GTEx validou sua expressão monoalélica em vários tecidos (Dataset S15) e, portanto, seu status imprintado (JIRTLE e MURPHY, 2012; WEI *et al.*, 2014; BARAN *et al.*, 2015; PIRINEN *et al.*, 2015). Infelizmente, o projeto GTEx não inclui amostras de fibroblastos embrionários, iPSC ou células B.

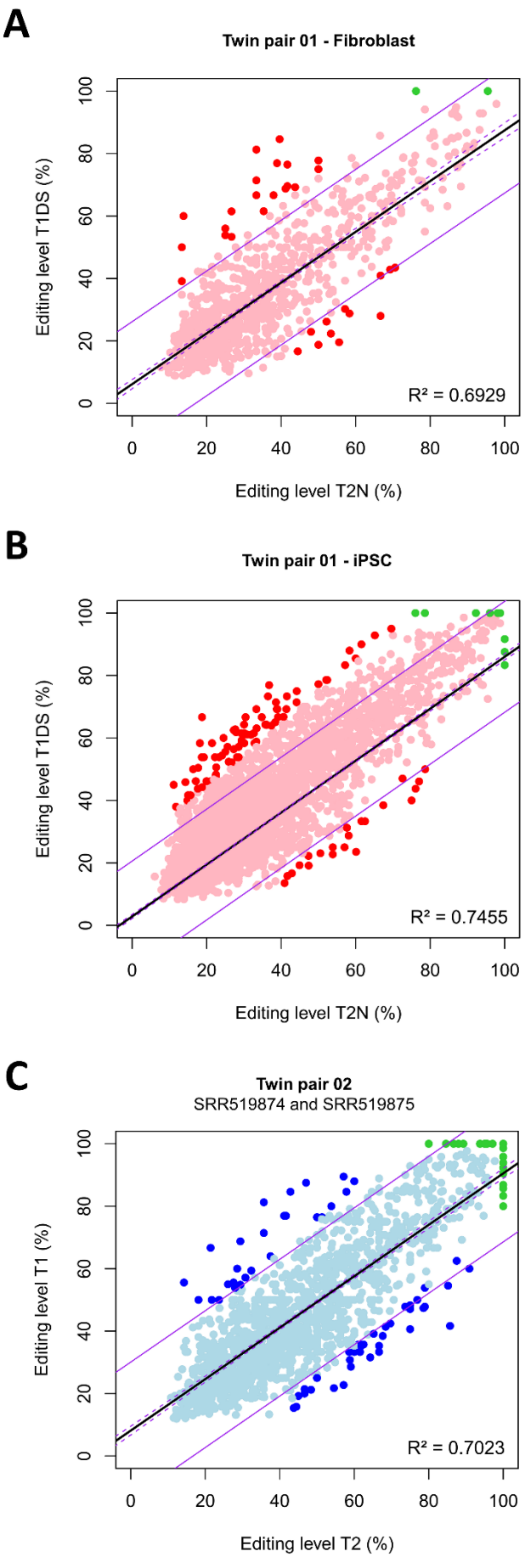
Por outro lado, a maioria dos outros genes classificados com ≥8 eSNVs monoalélicos foram expressos bialelicamente em múltiplos tecidos no banco de dados do GTEx (Dataset S15). Especulamos que a concordância monoalélica em vários sítios observados nos co-gêmeos reflete homozigose prolongada, em vez de efeitos de origem parental. Notamos cinco exceções genéticas. Primeiro, o gene *AC091729*.3, que exibiu uma média de 12 eSNVs monoalelicamente em dois dos dez pares co-gêmeos MZ, também é expresso monoalelicamente de uma maneira isoforma específica, com quatro SNPs em 35 tecidos nas amostras do GTEx (Dataset S15) Segundo, o gene *SPINK5* com 11 eSNVs monoalelicamente, expresso exclusivamente na bexiga nos dados do GTEx. Enquanto o perfil ASE no gene *AC091729*.3 era homogêneo, o gene *SPINK5* era moderadamente heterogêneo. Vemos esses dois genes como potenciais derivações e sugerimos que o gene *AC091729*.3 seja submetido ao *imprinting* genômico isoforma específica, enquanto o gene *SPINK5* é imprintado de maneira tecido específica.

Terceiro, os genes impressos conhecidos *SNURF*, *SNHG14* e *ZNF264*, que são expressos monoalelicamente em vários tecidos nas amostras do GTEx, exibiram ≥ 10 eSNVs com desbalanço da expressão bialélica em iPSC (Dataset S15). Curiosamente, vários eSNVs de desbalanço da expressão bialélica nesses três genes estão listados no banco de dados RADAR e provavelmente estão sujeitos à edição de RNA A-para-I (G): 5 dos 16 eSNVs (*SNURF*), 19/35 (*SNHG14*), 4/14 (*ZNF264*) (Dataset S3). Sugerimos, portanto, que a modificação do epitranscriptoma desses produtos gênicos pela edição do RNA altera seu fenótipo de *imprinting* esperado (pelo menos *in vitro*) em iPSCs.

5.6. Impacto estimado da edição de ácido ribonucleico canônico A-para-I (G) na disparidade da expressão alelo-específica

O número de sítios de ASE que correspondem posicionalmente aos sítios canônicos de A-para-I (G) foi, em média, 2.012 ± 786 por par de gêmeos (Dataset S3, S6 – S14), e todos os sítios abrangem 419 ± 116 genes. A grande maioria dos sítios exibia um perfil de desbalanço da expressão bialélica concordante (pontos rosa e azul claro na Figura 4 e Figura S7). Assim, entre os gêmeos, houve uma concordância geral no estado de desbalanço da expressão bialélica. O número de sítios que exibiram perfis alélicos discordantes, sendo bialélicos em um gêmeo e monoalélicos no outro, foi, no entanto, mínimo, embora mais abundante nos co-gêmeos homocariotípicos do que nos heterocariotípicos (pontos verdes na Figura 4 e Figura S7). Esta observação indicou que nas células analisadas, poucos eSNVs foram editados 100% (isto é, expressos estritamente monoalélico) no conjunto completo de 10 pares de gêmeos. Entre os co-gêmeos heterocariotípicos, havia apenas sete sítios discordantes, sendo expressos monoalelicamente o co-gêmeo trissômico e com desbalanço da expressão bialélica no co-gêmeo normal (Figura 4). Os sete sítios discordantes ocorrem em sete genes codificadores de proteínas, incluindo *CD46* (um receptor imune do tipo I) e *ING5* (um Supressor de tumor).

Figura 4. Disparidades de pares de gêmeos nas proporções de expressão de alelos em variantes de nucleotídeo único expressas (eSNVs) que são coincidentes com os sítios de edição de RNA canônico A-para-I (G). É mostrada a distribuição de eSNVs que correspondem positivamente aos sítios de edição de RNA canônico entre co-gêmeos heterocariotípicos, analisados em fibroblastos fetais (A), iPSC (B) derivado de fibroblastos fetais ou entre co-gêmeos homocariotípicos testados em células-cultura B (C) Cada ponto corresponde a uma eSNV. A grande maioria dos sítios exibia um perfil de desbalanço da expressão bialélica concordante (pontos rosa e azul claro). Os pontos vermelhos representam eSNVs discordantes entre os gêmeos, pois apresentaram diferenças nas proporções alélicas superiores a 25%, independentemente da discordância ou concordância no cariótipo. Pontos verdes representam eSNVs que exibiram perfis alélicos discordantes, sendo bialélicos em um gêmeo e monoalélicos no outro. Os modelos lineares (linhas pretas sólidas), o intervalo de confiança dos modelos (linhas roxas quebradas) e as previsões (linhas roxas sólidas) foram construídos usando R. As equações do modelo: (A) Y = 6,15718 + 0,81302X; (B) Y = 2,832681 + 0,831631X; (C) Y = 8,15439 + 0,82362X. Para todos os pares, P <2.2e-16.



Nos gêmeos heterocariotípicos, 117 genes expressos exibiram altas proporções de sítios de ASE (≥ 4 sítios por gene) coincidentes com sítios de edição de RNA. Notavelmente, para os genes *CYP20A1* e *ZNF621*, 74 e 77% de todos os sítios de ASE são sítios canônicos de A-para-I (G) validados no banco de dados RADAR. A extensão do desbalanço da expressão alélica variou de 6 a 98%. No entanto, cerca de 5% de todos os sítios exibiram níveis discordantes de edição de RNA acima de 25% entre co-gêmeos, independentemente de condições hetero- ou homocariotípicas (pontos vermelhos e azuis na Figura 4, Figura S5 e Dataset S3). Por exemplo, o gene *CYP20A1* apresentou a maior porcentagem de desbalanço da expressão alélica entre os co-gêmeos heterocariotípicos nos fibroblastos (T1DS = 81,25% / T2N = 33,33%) enquanto no iPSC o gene *GCFC2*/*MRPL19* exibiu a maior discrepância (T1DS = 66,66% / T2N = 18,75%).

Como a edição de RNAs A-para-I (G) de mRNAs pode criar códons de parada (variantes de efeito de truncamento de proteínas) ou resultar em mutações não sinônimas, era importante anotar os sítios com proporções alélicas discordantes. Nos co-gêmeos heterocariotípicos, nenhum dos sítios anotados criou códons de parada, mas prevê-se que nove sítios causem mutações não-sinônimos (Dataset S16). O gene *CDK13* da cinase 13 dependente de ciclina apresentou dois sítios de edição não-sinônimos que alteram a lisina (Lys; chr7_39950928) e glutamina (Gln; chr7_39950949) para arginina. Dentro dos 5.372 eSNVs anotados como sítios de edição de RNA canônico nos transcriptomas dos pares de gêmeos homocariotípicos, nenhum cria códons de parada e 21 sítios correspondem a mutações não-sinônimas (Dataset S16).

5.7. Co-gêmeos MZ são discordantes na expressão alelo-específica de genes que não escapam da inativação do cromossomo X

Analisamos sítios de ASE discordantes em genes ligados ao X nos sete pares de gêmeos femininos e integramos dados para os sítios cruzados sobre o status de classificação XCI de repositórios públicos (CARREL e WILLARD, 2005; COTTON *et al.*, 2013; BALATON *et al.*, 2015; COTTON *et al.*, 2015; TUKIAINEN *et al.*, 2017; GARIERI *et al.*, 2018; SHVETSOVA *et al.*, 2019; WAINER KATSIR e LINIAL, 2019). A análise foi restrita a genes que não escapam da inativação

porque, em conjunto de células, esses genes devem exibir perfis de expressão bialélicos. Para esta análise específica, aceitamos apenas produtos gênicos que exibiam pelo menos dois eSNVs discordantes (ou seja, monoalélicos em uma gêmea versus bialélico na irmã gêmea). Para as co-gêmeas discordantes heterocariotípicos, houve disparidade de ASE nos produtos do gene *UBL4A* nos fibroblastos, enquanto, em iPSC, houve disparidade de ASE nos produtos dos genes *FANCB* e *FTX* (Tabela S3). Três pares de gêmeas de controle expressaram genes com pelo menos dois eSNVs: genes *TAB3, WDR44 e XIAP* no par de gêmea 05; *IDS, MAP7D3, RLIM*, Genes *RPL10, SLC9A7, TBC1D25, TLR7, XIAP e ZNF275* no par de gêmeas 08; e o gene *ZNF275* no par de gêmeas 09 (Tabela S3). Nenhuma das disparidades ASE acima são sítios de edição de RNA A-para-I (G) no banco de dados RADAR.

5.8. O impacto geral da expressão alelo-específica de variantes patogênicas

Anotamos 32 eSNVs associados a 131 patologias humanas nos transcriptomas dos gêmeos heterocariotípicos (Dataset S17A). A maioria dos eSNVs patogênicos está ligada a fenótipos autossômicos recessivos e foi coexpressa com o alelo do tipo selvagem, provavelmente superando os efeitos deletérios previstos. Quatro alelos patogênicos (rs1799990 * G>A, rs1800562 * G>A, rs200855215 * A>G e rs4784677 * A>G) foram expressos monoalelicamente e estão associados à doença de *Jakob-Creutzfeldt* (OMIM # 123400), hemocromatose (OMIM # 235200), atrofia óptica de Leber (OMIM # 535000) e síndrome de *Bardet-Biedl* 2 (OMIM # 615981), respectivamente. Um alelo patogênico (rs11583680 * C>A), associado à hipercolesterolemia familiar autossômica dominante (OMIM # 603776), também foi co-expresso com o alelo do tipo selvagem. Nos conjuntos de gêmeos homocariotípicos, 23 eSNVs, previstos como patogênicos no banco de dados ClinVar, predominantemente co-expressos com os alelos do tipo selvagem (Dataset S17B). Por exemplo, rs1799958 * G>A, associado à deficiência de butiril-coenzima A desidrogenase (OMIM # 201470), foi co-expresso com o alelo do tipo selvagem no co-gêmeo 03.

5.9. Evidências para Microheteroplasmia Mitocondrial Expressa

Também identificamos uma forma de microheteroplasmia mitocondrial (SOUREN *et al.*, 2016), embora em limites mais baixos, em todos os 10 pares de gêmeos MZ, demonstrados pela presença de 237 eSNVs (número médio de 25 eSNVs por conjunto de dados, Tabela S4). O número limitado observado de eSNVs mitocondriais não se relaciona exclusivamente com a idade embrionária inicial na amostragem, porque a idade dos pares gêmeos controle variou de 19 a 65 anos (idade média 26 anos) (Dataset S2). Assim, para o conjunto de doadores investigados, não observamos o acúmulo de eSNVs mitocondriais com a idade (SMIGRODZKI e KHAN, 2005).

Por fim, consultamos os bancos de dados públicos ClinVar, PolyPhen e SIFT em busca de evidências sobre a previsão de patogenicidade para os eSNVs mitocondriais para avaliar as mutações de ponto mitocondriais funcionalmente cruciais. Prevê-se que quinze eSNVs sejam patogênicos em pelo menos um banco de dados (Tabela S4). Por exemplo, rs28358569 * A>G, expresso monoalelicamente no co-gêmeo 09, está relacionado à perda auditiva neurossensorial mitocondrial não sindrômica (OMIM # 500008) e surdez induzida por aminoglicosídeo (OMIM # 580000); rs193302980 * C>T e rs2853508 * A>G estão relacionados ao câncer de mama familiar (OMIM # 114480).

5.10. Análise de Ontologia Genética de Locais de Expressão Discordantes para Alelos Específicos

Nos fibroblastos, os genes *CASP6* e *PDXDC1*, representados por sítios de ASE exibindo troca bialélica para monoalélica (LogASE ≥ 0,8) nos co-gêmeos heterocariotípicos, estavam relacionados com processos metabólicos de compostos nitrogenados e substâncias orgânicas (Dataset S18A). Por outro lado, *IL17RA*, o único gene com LogASE ≤ -0,8 e mapeamento fora da região de recombinação 21q bem caracterizada, é enriquecido em processos imunológicos como migração de leucócitos, transdução de sinal, produção de citocinas e ativação celular (Dataset S18B). Em iPSCs, a maioria dos genes (72 de 100 genes) com sítios de ASE discordantes (comutadores bi-para-mono ou mono-bialélicos) está relacionada à regulação do processo biológico (Dataset S9C e Dataset S9D).

6. DISCUSSÃO

Nosso objetivo foi compilar sítios variantes com perfis de expressão diferentes entre os co-gêmeos MZ que são discordantes ou não para uma condição específica. Nossa estratégia de varredura permitiu a identificação, quantificação e classificação da expressão alélica diferencial por meio de sítios discordantes de ASE (isto é, SNPs expressos (eSNPs) ou eSNVs) ocorrendo em todo o genoma entre co-gêmeos que são discordantes ou não para T21. Notavelmente, a amplitude e magnitude dos sítios discordantes de ASE foram altos e comparáveis entre os gêmeos heterocariotípicos ou homocariotípicos. Em média, identificamos cerca de 1.342 sítios discordantes de ASE nos 10 pares de co-gêmeos MZ.

A extensão dos sítios de ASE nos co-gêmeos discordantes T21 foi comparável entre os co-gêmeos não discordantes, analisados em três tipos de células (fibroblastos, iPSC e células B). No geral, as análises indicam que a discordância de ASE entre os co-gêmeos MZ pode decorrer de aneuploidia, recombinação, *imprinting* genômico e edição de RNA. Os sítios de ASE discordantes observados entre os co-gêmeos refletem melhor um efeito combinado de processos genéticos e epigenéticos na expressão diferencial de alelos.

Observamos que a cariotipagem da expressão revela matrizes dinâmicas de sítios ASE que podem ser considerados assinaturas que exibem singularidade notável para a amostra biológica individual. Por exemplo, para os co-gêmeos heterocariotípicos, os conjuntos de eSNVs observados em fibroblastos ou iPSC não se sobrepõem completamente. No geral, 38,67% (n = 24.103) dos sítios de ASE foram chamados em amostras de fibroblastos e iPSC, 804 (3,3%) dos quais exibiram perfis discrepantes de expressão de alelo em fibroblastos, mas concordantes em iPSC. Da mesma forma, 1.318 sítios (5,4%) eram concordantes em fibroblastos, mas discordantes em iPSC. Além disso, 187 sítios (0,7%) foram discordantes em ambos os tipos de amostra. A relativa falta de sobreposição entre os experimentos provavelmente é explicada pela expressão diferencial de genes nesses tipos de células. Portanto, as assinaturas de cariotipagem da expressão podem ter valor forense e poder de resolução para

discriminar co-gêmeos clinicamente não discordantes. Tais assinaturas podem ser específicas ao nível de cada condição experimental para a mesma fonte de uma amostra biológica.

A observação do viés alélico está se tornando comum nas análises de transcriptoma de alto rendimento seja em conjunto de células ou célula única, o estudo do viés alélico tem respondido perguntas sobre a descoberta de genes imprintados, RNA *splicing* variantes e isoformas, além de genes com expressão monoalélica randômica ou aberrações cromossômicas (DEVEALE *et al.*, 2012; MARINOV *et al.*, 2014; METSALU *et al.*, 2014; WOOD *et al.*, 2015; WEISSBEIN *et al.*, 2016). É aceitável que a expressão da maioria dos genes pode ser alterada entre réplicas de amostras biológicas e que o RNA celular total não seja constante. O viés alélico no RNA-Seq pode ser, em parte, atribuído ao impacto diferencial das condições de cultivo *in vitro* resultante de padrões de recombinação meiótica (GIMELBRANT *et al.*, 2007; WEISSBEIN *et al.*, 2016). Assim, parte da discordância da ASE observada entre fibroblastos e iPSC em co-gêmeos em nossa análise pode ser devida a anormalidades cromossômicas adquiridas durante a derivação da iPSC e sua propagação em cultura a partir dos fibroblastos.

6.1. Papel da Inativação do cromossomo X e *Imprinting* genômico

Com o início da geminação MZ, a XCI e o *imprinting* genômico podem ocorrer aproximadamente ao mesmo tempo durante o desenvolvimento embrionário (MACHIN, 1996), entretanto, seu entrelaçamento pode afetar a distribuição de células portadoras do cromossomo X inativado ou marcas epigenéticas anormais de *imprinting* e, portanto, a manifestação variável de diferenças alélicas desses processos. Surpreendentemente, o efeito ocorre principalmente em mulheres gêmeas, em vez de homens gêmeos, e, portanto, provavelmente devido à presença de mais de um cromossomo X em mulheres. Essa discrepância pode ser valiosa na análise do efeito de uma doença na expressão gênica ou na variação fenotípica e ter implicações a longo prazo (LUBINSKY e HALL, 1991; MATIAS *et al.*, 2014). Neste contexto, ORSTAVIK *et al.* (1995) relata um excesso de gêmeas monozigóticas discordantes para a *Wiedemann-Beckwith syndrome*, uma doença de *imprinting*, tendo proposto que esse excesso poderia estar relacionado a XCI.

Adicionalmente, há casos de co-gêmeas MZ discordantes por XCI enviesada com distúrbios de *imprinting* na qual a gêmea afetada teve a inativação X completamente enviesada, com o alelo paterno no cromossomo X ativo em todas as células, enquanto que a gêmea não afetada teve uma inativação do X enviesada moderadamente na mesma direção, de forma contrária a mãe teve um padrão aleatório de XCI (ORSTAVIK *et al.*, 1995). Outros relatos também demostram XCI enviesada em gêmeas discordantes para doenças não relacionadas ao *imprinting* como Hemofilia A na qual a gêmea com hemofilia mostrou inativação não enviesada em direção ao X paterno, enquanto a gêmea saudável mostrou inativação aleatória do X (BENNETT *et al.*, 2008).

Curiosamente, neste estudo, 1.050 sítios de ASE são mapeados para 205 genes conhecidos e candidatos ao *imprinting*. A discordância de ASE, ainda que em uma extensão consideravelmente menor do que a descrita aqui, foi relatada entre um par de co-gêmeos MZ “idênticos”, clinicamente discordantes para esclerose múltipla em estudos de transcriptoma de amostras de sangue (SOUREN *et al.*, 2016). Neste estudo, a expressão alélica alterada de dois genes imprintados (*ZNF331* e *GNAS*) e cinco genes não imprintados (*ABLIM1, UBE2I, KIAA1267, CD6* e *ATHL1*) foram detectados entre os gêmeos discordantes para esclerose múltipla.

Quatorze genes ligados ao X sujeitos a XCI (genes que não escapam da inativação como *UBL4A, FANCB, FTX, TAB3, WDR44, XIAP, IDS, MAP7D3, RLIM, RPL10, SLC9A7, TBC1D25, TLR7 e ZNF275*) mostrados no presente estudo, exibiram disparidades de ASE em que um co-gêmeo apresentava perfil bialélico e a irmã mostrava padrão monoalélico. Os experimentos de RNA-Seq foram obtidos a partir de conjunto de células em vez de células únicas e, portanto, um modelo bialélico é o perfil de expressão esperado para genes que não escapam da inativação (CARREL e WILLARD, 2005; COTTON *et al.*, 2013; BALATON *et al.*, 2015; COTTON *et al.*, 2015; TUKIAINEN *et al.*, 2017; GARIERI *et al.*, 2018; SHVETSOVA *et al.*, 2019; WAINER KATSIR e LINIAL, 2019).

A discordância de ASE em genes ligados ao X que são submetidos ao XCI foi relatada entre co-gêmeas MZ em humanos mostrando que a extensão da expressão alélica diferencial é altamente semelhante nos pares de gêmeas monozigóticas para muitos *loci*, o que pode significar que as diferenças alélicas

na expressão gênica estão sob controle genético, mas para outros *loci* a discordância de ASE não pode ser desconsiderada (CHEUNG *et al.*, 2008; GARIERI *et al.*, 2018) e em camundongos (WANG *et al.*, 2010). Razões de expressão alélica iguais são frequentemente determinadas geneticamente em *cis* (exemplo, eQTLs) e *trans*, mas parte da disparidade também pode ser atribuída ao efeito de amostragem aleatória da inativação do X (CARREL e WILLARD, 2005; CHEUNG *et al.*, 2008; COTTON *et al.*, 2013; BALATON *et al.*, 2015; COTTON *et al.*, 2015; GARIERI *et al.*, 2018; SHVETSOVA *et al.*, 2019; WAINER KATSIR e LINIAL, 2019). Além disso, o desbalanço da expressão alélica no cromossomo X também pode afetar a expressão alélica autossômica. Não obstante, observamos que a extensão da discordância de ASE nos pares de gêmeos homocariotípicos do sexo masculino (em média 3,2%, n = 1.478 sítios discordantes) é comparável em todo o genoma à dos pares gêmeos femininos (2,5%, n = 958 sítios discordantes).

6.2. Discordância observada de ASE em co-gêmeos MZ

Dados de estudos anteriores de RNA-Seq em co-gêmeos (BARANZINI *et al.*, 2010; LIN *et al.*, 2012; BROWN *et al.*, 2014; HIBAOUI *et al.*, 2014; BUIL *et al.*, 2015; DIXON *et al.*, 2015; DING *et al.*, 2017; SANTONI *et al.*, 2017) indicam que a expressão diferencial de alelos de genes autossômicos reflete melhor os processos de regulação dinâmica consistentes com um alelo sendo preferencialmente silenciado ou um alelo inativo sendo restaurado.

A expressão bialélica dos genes é um mecanismo regulador que desequilibra os efeitos prejudiciais dos alelos variantes patogênicos da expressos ou da perda de função (ADZHUBEI *et al.*, 2010; LANDRUM *et al.*, 2016; VASER *et al.*, 2016). Em cada célula somática euplóide humana, prevê-se que genes autossômicos sejam simetricamente expressos a partir de ambos os alelos parentais, de maneira específica ao tipo de célula, durante todo o desenvolvimento. No entanto, o padrão de expressão do RNA bialélico não é uma marca fenotípica de todos os genes, já que 10 a 30% dos genes autossômicos humanos analisados para sítios variantes polimórficos [isto é, eSNVs] são submetidos dinamicamente aos fenômenos epigenéticos, isto e, expressão monoalélica aleatória mitoticamente estável. Os genes com expressão monoalélica exibem uma taxa de recombinação elevada e densidade

aumentada de contextos de sequência hiper mutável. Tais genes por apresentar tal perfil de expressão podem gerar heterogeneidade célula a célula e aumento da variação genética contribuindo para heterogeneidade. O que pode explicar, em parte, o aumento da diversidade relacionada a expressão monoalélica. (GIMELBRANT *et al.*, 2007; SAVOVA *et al.*, 2016a; SAVOVA *et al.*, 2016b; SAVOVA *et al.*, 2017) ou viés alélico (DIXON *et al.*, 2015).

Os genes mais enigmáticos que são expressos bialelicamente em uma célula podem ser regulados em uma célula vizinha para mudar aleatoriamente sua expressão de RNA de bialélico para monoalélico (CHESS, 2013; ECKERSLEY-MASLIN e SPECTOR, 2014; ECKERSLEY-MASLIN *et al.*, 2014). Além disso, subconjuntos distintos de genes autossômicos e ligados ao X são submetidos ao silenciamento epigenético de um alelo, de maneira dependente da origem dos pais por *imprinting* genômico autossômico (BARAN *et al.*, 2015) ou de forma aleatória por XCI (exemplo, no sexo feminino) (TUKIAINEN *et al.*, 2017).

Existem consequências genéticas e funcionais dos locais variantes autossômicos nos genes que são expressos a partir de um único alelo em uma célula por vez. Principalmente, i) se eles conferem diversidade genética mais extensa em humanos (SAVOVA *et al.*, 2016a); ii) se frequentemente são variantes de risco de alteração de expressão ou perda de função patogênica (isto é, para distúrbios do desenvolvimento neurológico), e influenciam a variação da expressão em *cis*; iii) quando a faixa de nível de expressão dos genes expressos monoalelicamente é maior que os genes expressos bialelicamente (SAVOVA *et al.*, 2017); iv) finalmente, aumentam a variabilidade da expressão célula a célula com um impacto benéfico de evitar fenótipos de doenças genéticas (SAVOVA *et al.*, 2016a).

6.3. Efeito do cromossomo extranumerário 21

Sabe-se que o cromossomo extranumerário 21 em células trissômicas de pacientes com síndrome de Down resulta em desregulação da expressão gênica em todo o genoma representada por domínios cromossômicos com genes cujos níveis de expressão são compensados por cópia, ou aumentada, em

comparação com as células euplóides (LETOURNEAU *et al.*, 2014). Nos co-gêmeos discordantes para recombinação 21q e T21, os perfis discrepantes de ASE podem ser vistos, em última análise, como herdabilidade inexplicável ou falta de herdabilidade devido à discordância na trissomia 21, recombinação em 21q, *imprinting* genômico alterado, expressão monoalélica aleatória e edição de RNA.

O desbalanço da expressão alélica (heterogeneidade alélica específica), em genes sensíveis à dosagem, pode surgir por uma regulação adaptativa estocástica em células euplóides e aneuplóides como consequência da abundância de mRNA e aumento da frequência de transcrição (DENG e DISTECHE, 2019; LARSSON *et al.*, 2019; SYMMONS *et al.*, 2019). A extensão do desbalanço da expressão bialélica entre os eSNVs provavelmente reflete um efeito de expressão da rede gênica que opera na forma de eQTLs (MOTT *et al.*, 2014; PETTIGREW *et al.*, 2016). Assim, parte da herdabilidade inexplicável ou da falta de herdabilidade pode ser explicada por diferenças nas interações genéticas específicas das células.

Além disso, as discrepâncias do perfil de ASE entre fibroblastos e iPSC podem ser devidas a efeitos de origem parental não imprintados em cada tipo de célula associado à aneuploidia. Por exemplo, o eSNP rs93366794 exibia um perfil bialélico concordante em iPSC, mas discordante nos fibroblastos, sendo monoalélico no gêmeo T21 e bialélico no co-gêmeo normal. Curiosamente, em um estudo de RNA-Seq de um cérebro saudável, foi relatado que o gene *WRD4* portador do sítio rs93366794 era expresso monoalelicamente de origem paterna, mas ocorreu uma troca mono-bialélicos na prole com autismo versus sem autismo (LIN *et al.*, 2018).

6.4. Significado biológico das diferenças na magnitude e amplitude de ASE

Embora a análise apresentada tenha permitido a identificação de uma disparidade generalizada nos perfis de ASE entre co-gêmeos, o significado biológico da extensão e amplitude das diferenças observadas na expressão de alelos deve ser avaliado apenas por experimentos independentes. No entanto, os seguintes resultados são dignos de nota: i) entre os eSNVs, havia vários alelos que se sabe estarem associados às condições da doença ou que são

patogênicos; ii) o gene canônico impresso *SNURF*, que é expresso monoalelicamente em mais de 50 tecidos no conjunto de dados do GTEx, foi expresso bialelicamente em iPCS; iii) em todos os 10 pares de gêmeos, houve microheteroplasmia mitocondrial expressa; iv) entre todos os genes expressos nos 10 pares de gêmeos, havia 55±17 genes que exibiram proporções elevadas (variando de 50 a 100%) dos sítios de ASE coincidentes com os sítios de edição de RNA.

A amplitude e magnitude da discordância de ASE apresentam variações interindividuais epigenômicas sem precedentes que ocorrem em co-gêmeos MZ. Estudos prévios de ASE em gêmeos MZ são restritos a um gene ou conjuntos de genes específicos e, portanto, não descobrem o estado aparente de herdabilidade alélica generalizada ausente em co-gêmeos MZ mostrados no presente estudo. Embora a validação independente por meio de experimentação por via úmida (ou seja, transcrição reversa quantitativa específica de alelo-PCR, hibridização de fluorescência de RNA *in situ* ou piro sequenciamento alelo-específico) seja necessária para os sítios de ASE discordantes biologicamente relevantes, implicações críticas emergem da variação interindividual em todo o epigenoma observada nos co-gêmeos MZ: i) como no caso da variação interindividual na metilação do DNA (MAUNAKEA *et al.*, 2010; BELL *et al.*, 2012; YOUNG *et al.*, 2017; GARG *et al.*, 2018), a discordância de ASE pode ter que ser analisada calculando o impacto da variação fenotípica na susceptibilidade diferencial às condições humanas específicas e doenças (SKIPPER, 2008; SUN *et al.*, 2013); ii) A discordância da ASE também pode ser considerada fundamental para o desenvolvimento de assinaturas de biomarcadores de RNA para identificação forense de fluidos corporais e análise de parentesco (BLAY *et al.*, 2019).

6.5. Limitações importantes desta análise integrativa

A presente metanálise sistemática e integrativa tem três limitações importantes: tamanho da amostra, a certeza de chamar corretamente um local de eSNV positivo para uma posição teórica de heterozigotos, e comparações feitas em três tipos diferentes de células. Primeiro, o cenário experimental deve ser visto como um estudo de caso sobre uma discordância heterocariotípica de dois pares MZ para recombinação distal no cromossomo 21 e trissomia do 21.

Casos relatados de pares gêmeos heterocariotípicos de MZ são raros. Na Tabela S1, listamos todos os casos relevantes estudados na literatura.

Não obstante, existe apenas um estudo público de RNA-Seq em um par de gêmeos MZ heterocariotípicos, ou seja, o caso de índice discordante selecionado. Utilizamos o caso de índice como um caso de referência para investigar se a discordância subjacente no cariótipo e na recombinação afetam os perfis de ASE. Inicialmente, foi levantada a hipótese de que qualquer discordância provável nas diferenças de ASE seria restrita ao cromossomo 21 e que as diferenças seriam mais significativas além do evento de recombinação. No entanto, a análise inicial indicou a ocorrência de discordância de ASE em todo o genoma, em vez daquela restrita ao chr21.

Para investigar se a discordância de ASE observada em todo o genoma foi limitada ao caso índice único, investigamos nove pares de gêmeos MZ homocariotípicos. Surpreendentemente, observamos discordância de ASE em todo o genoma, semelhante em amplitude e magnitude à observada no caso-índice, embora em diferentes linhagens celulares.

Segundo, para diminuir as chances de chamadas de ASE falso-positivas, chamamos sítios de eSNV usando o controle de qualidade de base Q30 e profundidade de leitura ≥12, que selecionam critérios que se destacam em relatórios rigorosos publicados (leituras Q20 e ≥8) (CHEUNG *et al.*, 2008; BARAN *et al.*, 2015; TAN *et al.*, 2017; TUKIAINEN *et al.*, 2017). Como a probabilidade de chamadas SNV corretas aumenta em níveis de cobertura mais altos para uma posição teórica de heterozigotos, fornecemos resultados para três coberturas de leitura (12, 20 e 40). A cobertura de 40 fornece uma probabilidade de 99,9% de chamada SNV correta (conjuntos de dados S3-S14).

Terceiro, a observação de discordância de ASE em todo o genoma no mesmo tipo de linhagem celular (nove pares de gêmeos MZ) e diferentes linhas celulares (caso índice) é uma observação muito tranquilizadora. CHEUNG *et al.* (2008) usaram o arranjo *Affymetrix GeneChip Human Mapping 100K* SNP no mesmo conjunto de amostras MZ homocariotípicas e identificaram 201 SNPs com evidência significativa de expressão alélica diferencial. Desses, confirmamos 137 eSNVs como discordantes, dos quais 38 eram comuns aos

nove pares de gêmeos (Dataset S4 – S14). Infelizmente, nenhum dado público de sequenciamento de próxima geração está disponível para as amostras compatíveis com DNA-Seq e RNA-Seq de gêmeos MZ. Portanto, não podemos abordar atualmente a questão de quanto da variação genética contribui para a porcentagem total de ASE em gêmeos MZ.

7. Considerações finais

Nossas varreduras genômicas para discordância de expressão alélica revelam um estado aparente de herdabilidade alélica generalizada ausente em co-gêmeos MZ. De forma interessante, a extensão e a amplitude dos sítios discordantes ASE não estão exclusivamente associadas a aberrações cromossômicas e recombinação, mas também com os fenômenos de expressão diferencial do alelo em todo o epigenoma da expressão monoalélica aleatória, *imprinting* genômico e edição de RNA. A maioria dos sítios de ASE discordantes observados dentro de todos os pares de co-gêmeos MZ não pode ser atribuída apenas às incongruências estimadas no DNA ou correspondem ao ruído alélico transcricional aleatório variando através de experimentos. Neste caso, a discordância da expressão alelo específica do epigenoma pode ter efeitos essenciais na fisiologia, fenótipo ou herança, e implicações na abordagem das Origens do Desenvolvimento da Saúde e Doenças em co-gêmeos.

8. REFERÊNCIAS BIBLIOGRÁFICAS

ABDELLAOUI, A. *et al.* CNV Concordance in 1,097 MZ Twin Pairs. **Twin Res Hum Genet,** v. 18, n. 1, p. 1-12, Feb 2015.

ADZHUBEI, I. A. *et al.* A method and server for predicting damaging missense mutations. **Nat Methods,** v. 7, n. 4, p. 248-249, Apr 2010.

AIT YAHYA-GRAISON, E. *et al.* Classification of human chromosome 21 gene-expression variations in Down syndrome: impact on disease phenotypes. **Am J Hum Genet,** v. 81, n. 3, p. 475-491, Sep 2007.

ALVES DA SILVA, A. F. *et al.* Trisomy 21 Alters DNA Methylation in Parent-of-Origin-Dependent and -Independent Manners. **PLoS One,** v. 11, n. 4, p. e0154108, 2016.

AMOS-LANDGRAF, J. M. *et al.* X chromosome-inactivation patterns of 1,005 phenotypically unaffected females. **Am J Hum Genet,** v. 79, n. 3, p. 493-499, Sep 2006.

AUGUI, S.; NORA, E. P.; HEARD, E. Regulation of X-chromosome inactivation by the X-inactivation centre. **Nat Rev Genet,** v. 12, n. 6, p. 429-442, Jun 2011.

BABAK, T. *et al.* Genetic conflict reflected in tissue-specific maps of genomic imprinting in human and mouse. **Nat Genet,** v. 47, n. 5, p. 544-549, May 2015.

BALATON, B. P.; COTTON, A. M.; BROWN, C. J. Derivation of consensus inactivation status for X-linked genes from genome-wide studies. **Biol Sex Differ,** v. 6, p. 35, 2015.

BARAN, Y. *et al.* The landscape of genomic imprinting across diverse adult human tissues. **Genome Res,** v. 25, n. 7, p. 927-936, Jul 2015.

BARANZINI, S. E. *et al.* Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis. **Nature,** v. 464, n. 7293, p. 1351-1356, Apr 29 2010.

BEGEMANN, M. *et al.* Maternal variants in NLRP and other maternal effect proteins are associated with multilocus imprinting disturbance in offspring. **J Med Genet**, Mar 24 2018.

BELL, J. T. *et al.* Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. **PLoS Genet,** v. 8, n. 4, p. e1002629, 2012.

BENNETT, C. M.; BOYE, E.; NEUFELD, E. J. Female monozygotic twins discordant for hemophilia A due to nonrandom X-chromosome inactivation. **Am J Hematol,** v. 83, n. 10, p. 778-780, Oct 2008.

BERLETCH, J. B.; YANG, F.; DISTECHE, C. M. Escape from X inactivation in mice and humans. **Genome Biol,** v. 11, n. 6, p. 213, 2010.

BJORNSSON, H. T. *et al.* SNP-specific array-based allele-specific expression analysis. **Genome Res,** v. 18, n. 5, p. 771-779, May 2008.

BLAY, N. *et al.* Assessment of kinship detection using RNA-seq data. **Nucleic Acids Res,** v. 47, n. 21, p. e136, Dec 2 2019.

BLENCOWE, B. J. Alternative splicing: new insights from global analyses. **Cell,** v. 126, n. 1, p. 37-47, Jul 14 2006.

BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: a flexible trimmer for Illumina sequence data. **Bioinformatics,** v. 30, n. 15, p. 2114-2120, Aug 1 2014.

BROWN, A. A. *et al.* Genetic interactions affecting human gene expression identified by variance association mapping. **Elife,** v. 3, p. e01381, Apr 25 2014.

BUIL, A. *et al.* Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. **Nat Genet,** v. 47, n. 1, p. 88-91, Jan 2015.

CARREL, L.; WILLARD, H. F. X-inactivation profile reveals extensive variability in X-linked gene expression in females. **Nature,** v. 434, n. 7031, p. 400-404, Mar 17 2005.

CHAIYASAP, P. *et al.* Whole genome and exome sequencing of monozygotic twins with trisomy 21, discordant for a congenital heart defect and epilepsy. **PLoS One,** v. 9, n. 6, p. e100191, 2014.

CHAKRAVARTI, A. Widespread promiscuous genetic information transfer from DNA to RNA. **Circ Res,** v. 109, n. 11, p. 1202-1203, Nov 11 2011.

CHAMBERLAIN, S.; USHEY, K.; ZHU, H. rsnps: Get 'SNP' ('Single-Nucleotide' 'Polymorphism'). p. https://cran.r-project.org/web/packages/rsnps/, 2018.

CHESS, A. Random and non-random monoallelic expression. **Neuropsychopharmacology,** v. 38, n. 1, p. 55-61, Jan 2013.

CHEUNG, V. G. *et al.* Monozygotic twins reveal germline contribution to allelic expression differences. **Am J Hum Genet,** v. 82, n. 6, p. 1357-1360, Jun 2008.

CONERLY, M.; GRADY, W. M. Insights into the role of DNA methylation in disease through the use of mouse models. **Dis Model Mech,** v. 3, n. 5-6, p. 290-297, May-Jun 2010.

CONSORTIUM, G. The Gene Ontology project in 2008. **Nucleic Acids Res,** v. 36, n. Database issue, p. D440-444, Jan 2008.

CONSORTIUM, G. T. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. **Science,** v. 348, n. 6235, p. 648-660, May 8 2015.

CORE TEAM, R. A language and environment for statistical computing. 2019.

COTTON, A. M. *et al.* Analysis of expressed SNPs identifies variable extents of expression from the human inactive X chromosome. **Genome Biol,** v. 14, n. 11, p. R122, Nov 1 2013.

COTTON, A. M. *et al.* Landscape of DNA methylation on the X chromosome reflects CpG density, functional chromatin state and X-chromosome inactivation. **Hum Mol Genet,** v. 24, n. 6, p. 1528-1539, Mar 15 2015.

CSANKOVSZKI, G.; NAGY, A.; JAENISCH, R. Synergism of Xist RNA, DNA methylation, and histone hypoacetylation in maintaining X chromosome inactivation. **J Cell Biol,** v. 153, n. 4, p. 773-784, May 14 2001.

DAHOUN, S. *et al.* Monozygotic twins discordant for trisomy 21 and maternal 21q inheritance: a complex series of events. **Am J Med Genet A,** v. 146A, n. 16, p. 2086-2093, Aug 15 2008.

DE SA MACHADO ARAUJO, G. *et al.* Maternal 5(m)CpG Imprints at the PARD6G-AS1 and GCSAML Differentially Methylated Regions Are Decoupled From Parent-of-Origin Expression Effects in Multiple Human Tissues. **Front Genet,** v. 9, p. 36, 2018.

DENG, X.; DISTECHE, C. M. Rapid transcriptional bursts upregulate the X chromosome. **Nat Struct Mol Biol,** v. 26, n. 10, p. 851-853, Oct 2019.

DEPRISTO, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. **Nat Genet,** v. 43, n. 5, p. 491-498, May 2011.

DEUSSING, J. M.; JAKOVCEVSKI, M. Histone Modifications in Major Depressive Disorder and Related Rodent Models. **Adv Exp Med Biol,** v. 978, p. 169-183, 2017.

DEVEALE, B.; VAN DER KOOY, D.; BABAK, T. Critical evaluation of imprinted gene expression by RNA-Seq: a new perspective. **PLoS Genet,** v. 8, n. 3, p. e1002600, 2012.

DING, N. *et al.* Transcriptome Analysis of Monozygotic Twin Brothers with Childhood Primary Myelofibrosis. **Genomics Proteomics Bioinformatics,** v. 15, n. 1, p. 37-48, Feb 2017.

DIXON, J. R. *et al.* Chromatin architecture reorganization during stem cell differentiation. **Nature,** v. 518, n. 7539, p. 331-336, Feb 19 2015.

DOBIN, A. *et al.* STAR: ultrafast universal RNA-seq aligner. **Bioinformatics,** v. 29, n. 1, p. 15-21, Jan 1 2013.

DURINCK, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. **Bioinformatics,** v. 21, n. 16, p. 3439-3440, Aug 15 2005.

DURINCK, S. *et al.* Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. **Nat Protoc,** v. 4, n. 8, p. 1184-1191, 2009.

ECKERSLEY-MASLIN, M. A.; SPECTOR, D. L. Random monoallelic expression: regulating gene expression one allele at a time. **Trends Genet,** v. 30, n. 6, p. 237-244, Jun 2014.

ECKERSLEY-MASLIN, M. A. *et al.* Random monoallelic gene expression increases upon embryonic stem cell differentiation. **Dev Cell,** v. 28, n. 4, p. 351-365, Feb 24 2014.

EGGER, G. *et al.* Epigenetics in human disease and prospects for epigenetic therapy. **Nature,** v. 429, n. 6990, p. 457-463, May 27 2004.

EGGERMANN, T. *et al.* Imprinting disorders: a group of congenital disorders with overlapping patterns of molecular changes affecting imprinted loci. **Clin Epigenetics,** v. 7, p. 123, 2015.

EISENBERG, E.; LEVANON, E. Y. A-to-I RNA editing - immune protector and transcriptome diversifier. **Nat Rev Genet,** v. 19, n. 8, p. 473-490, Aug 2018.

ESSAOUI, M. *et al.* Monozygotic twins discordant for 18q21.2qter deletion detected by array CGH in amniotic fluid. **Eur J Med Genet,** v. 56, n. 9, p. 502-505, Sep 2013.

FITZPATRICK, D. R. *et al.* Transcriptome analysis of human autosomal trisomy. **Hum Mol Genet,** v. 11, n. 26, p. 3249-3256, Dec 15 2002.

FURUKAWA, H. *et al.* Genome, epigenome and transcriptome analyses of a pair of monozygotic twins discordant for systemic lupus erythematosus. **Hum Immunol,** v. 74, n. 2, p. 170-175, Feb 2013.

GARG, P. *et al.* A survey of inter-individual variation in DNA methylation identifies environmentally responsive co-regulated networks of epigenetic variation in the human genome. **PLoS Genet,** v. 14, n. 10, p. e1007707, Oct 2018.

GARIERI, M. *et al.* Extensive cellular heterogeneity of X inactivation revealed by single-cell allele-specific expression in human fibroblasts. **Proc Natl Acad Sci U S A,** v. 115, n. 51, p. 13015-13020, Dec 18 2018.

GILBERT, B. *et al.* Prenatal diagnosis of female monozygotic twins discordant for Turner syndrome: implications for prenatal genetic counselling. **Prenat Diagn,** v. 22, n. 8, p. 697-702, Aug 2002.

GIMELBRANT, A. *et al.* Widespread monoallelic expression on human autosomes. **Science,** v. 318, n. 5853, p. 1136-1140, Nov 16 2007.

GRAY, T. A.; SAITOH, S.; NICHOLLS, R. D. An imprinted, mammalian bicistronic transcript encodes two independent proteins. **Proc Natl Acad Sci U S A,** v. 96, n. 10, p. 5616-5621, May 11 1999.

GREENE, C. S. *et al.* Big data bioinformatics. **J Cell Physiol,** v. 229, n. 12, p. 1896-1900, Dec 2014.

HA, G. *et al.* Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. **Genome Res,** v. 22, n. 10, p. 1995-2007, Oct 2012.

HANNA, C. W.; KELSEY, G. The specification of imprints in mammals. **Heredity (Edinb),** v. 113, n. 2, p. 176-183, Aug 2014.

HASIN, Y.; SELDIN, M.; LUSIS, A. Multi-omics approaches to disease. **Genome Biol,** v. 18, n. 1, p. 83, May 5 2017.

HEAP, G. A. *et al.* Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. **Hum Mol Genet,** v. 19, n. 1, p. 122-134, Jan 1 2010.

HIBAOUI, Y. *et al.* Modelling and rescuing neurodevelopmental defect of Down syndrome using induced pluripotent stem cells from monozygotic twins discordant for trisomy 21. **EMBO Mol Med,** v. 6, n. 2, p. 259-277, Feb 2014.

HILL, K. E. *et al.* An imprinted non-coding genomic cluster at 14q32 defines clinically relevant molecular subtypes in osteosarcoma across multiple independent datasets. **J Hematol Oncol,** v. 10, n. 1, p. 107, May 15 2017.

HU, Y. J. *et al.* Proper Use of Allele-Specific Expression Improves Statistical Power for cis-eQTL Mapping with RNA-Seq Data. **J Am Stat Assoc,** v. 110, n. 511, p. 962-974, 2015.

HUANG, Y. *et al.* Identifying Genomic Variations in Monozygotic Twins Discordant for Autism Spectrum Disorder Using Whole-Genome Sequencing. **Mol Ther Nucleic Acids,** v. 14, p. 204-211, Mar 1 2019.

JIRTLE, J.; MURPHY, C. Geneimprint database. 2012.

JONES, P. A.; TAKAI, D. The role of DNA methylation in mammalian epigenetics. **Science,** v. 293, n. 5532, p. 1068-1070, Aug 10 2001.

JOYCE, A. R.; PALSSON, B. O. The model organism as a system: integrating 'omics' data sets. **Nat Rev Mol Cell Biol,** v. 7, n. 3, p. 198-210, Mar 2006.

KANEKO-ISHINO, T. *et al.* Complementation hypothesis: the necessity of a monoallelic gene expression mechanism in mammalian development. **Cytogenet Genome Res,** v. 113, n. 1-4, p. 24-30, 2006.

KHATIB, H. Is it genomic imprinting or preferential expression? **Bioessays,** v. 29, n. 10, p. 1022-1028, Oct 2007.

KIM, M.; TAGKOPOULOS, I. Data integration and predictive modeling methods for multi-omics datasets. **Molecular Omics**, p. 1-18, 2017.

KNOPMAN, J. M. *et al.* What makes them split? Identifying risk factors that lead to monozygotic twins after in vitro fertilization. **Fertil Steril,** v. 102, n. 1, p. 82-89, Jul 2014.

KORIR, P. K.; SEOIGHE, C. Inference of allele-specific expression from RNA-seq data. **Methods Mol Biol,** v. 1112, p. 49-69, 2014.

LANDRUM, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. **Nucleic Acids Res,** v. 44, n. D1, p. D862-868, Jan 4 2016.

LARSSON, A. J. M. *et al.* X-chromosome upregulation is driven by increased burst frequency. **Nat Struct Mol Biol,** v. 26, n. 10, p. 963-969, Oct 2019.

LAWRENCE, M.; GENTLEMAN, R.; CAREY, V. rtracklayer: an R package for interfacing with genome browsers. **Bioinformatics,** v. 25, n. 14, p. 1841-1842, Jul 15 2009.

LAWRENCE, M. *et al.* Software for computing and annotating genomic ranges. **PLoS Comput Biol,** v. 9, n. 8, p. e1003118, 2013.

LETOURNEAU, A. *et al.* Domains of genome-wide gene expression dysregulation in Down's syndrome. **Nature,** v. 508, n. 7496, p. 345-350, Apr 17 2014.

LEUNG, W. C. *et al.* Monozygotic dichorionic twins heterokaryotypic for duplication chromosome 2q13-q23.3. **Fetal Diagn Ther,** v. 25, n. 4, p. 397-399, 2009.

LEVANON, E. Y. *et al.* Evolutionarily conserved human targets of adenosine to inosine RNA editing. **Nucleic Acids Res,** v. 33, n. 4, p. 1162-1168, 2005.

LI, H. *et al.* The Sequence Alignment/Map format and SAMtools. **Bioinformatics,** v. 25, n. 16, p. 2078-2079, Aug 15 2009.

LI, M. *et al.* Widespread RNA and DNA sequence differences in the human transcriptome. **Science,** v. 333, n. 6038, p. 53-58, Jul 1 2011.

LIN, C. Y. *et al.* Allele-specific expression in a family quartet with autism reveals mono-to-biallelic switch and novel transcriptional processes of autism susceptibility genes. **Sci Rep,** v. 8, n. 1, p. 4277, Mar 9 2018.

LIN, M. *et al.* Allele-biased expression in differentiating human neurons: implications for neuropsychiatric disorders. **PLoS One,** v. 7, n. 8, p. e44017, 2012.

LIU, S. *et al.* Four-Generation Pedigree of Monozygotic Female Twins Reveals Genetic Factors in Twinning Process by Whole-Genome Sequencing. **Twin Res Hum Genet,** v. 21, n. 5, p. 361-368, Oct 2018.

LIU, Z. *et al.* Comparing computational methods for identification of allele-specific expression based on next generation sequencing data. **Genet Epidemiol,** v. 38, n. 7, p. 591-598, Nov 2014.

LO, H. S. *et al.* Allelic variation in gene expression is common in the human genome. **Genome Res,** v. 13, n. 8, p. 1855-1862, Aug 2003.

LUBINSKY, M. S.; HALL, J. G. Genomic imprinting, monozygous twinning, and X inactivation. **Lancet,** v. 337, n. 8752, p. 1288, May 25 1991.

LYON, M. F. Sex chromatin and gene action in the mammalian X-chromosome. **Am J Hum Genet,** v. 14, p. 135-148, Jun 1962.

MACHIN, G. A. Some causes of genotypic and phenotypic discordance in monozygotic twin pairs. **Am J Med Genet,** v. 61, n. 3, p. 216-228, Jan 22 1996.

MACKAY, D. J. *et al.* Multilocus methylation defects in imprinting disorders. **Biomol Concepts,** v. 6, n. 1, p. 47-57, Mar 2015.

MAO, R. *et al.* Primary and secondary transcriptional effects in the developing human Down syndrome brain and heart. **Genome Biol,** v. 6, n. 13, p. R107, 2005.

MARINOV, G. K. *et al.* From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. **Genome Res,** v. 24, n. 3, p. 496-510, Mar 2014.

MASSAH, S.; BEISCHLAG, T. V.; PREFONTAINE, G. G. Epigenetic events regulating monoallelic gene expression. **Crit Rev Biochem Mol Biol,** v. 50, n. 4, p. 337-358, 2015.

MATIAS, A. *et al.* Monozygotic twins: ten reasons to be different. **Diagnóstico Prenatal,** v. 25, p. 53-57, 2014.

MAUNAKEA, A. K.; CHEPELEV, I.; ZHAO, K. Epigenome mapping in normal and disease States. **Circ Res,** v. 107, n. 3, p. 327-339, Aug 6 2010.

MAYNARD, N. D. *et al.* Genome-wide mapping of allele-specific protein-DNA interactions in human cells. **Nat Methods,** v. 5, n. 4, p. 307-309, Apr 2008.

MCKENNA, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. **Genome Res,** v. 20, n. 9, p. 1297-1303, Sep 2010.

MELTON, T. Mitochondrial DNA Heteroplasmy. **Forensic Sci Rev,** v. 16, n. 1, p. 1-20, Jan 2004.

METSALU, T. *et al.* Using RNA sequencing for identifying gene imprinting and random monoallelic expression in human placenta. **Epigenetics,** v. 9, n. 10, p. 1397-1409, Oct 2014.

MOREIRA DE MELLO, J. C. *et al.* Early X chromosome inactivation during human preimplantation development revealed by single-cell RNA-sequencing. **Sci Rep,** v. 7, n. 1, p. 10794, Sep 7 2017.

MORGAN, M. AnnotationHub: Client to access AnnotationHub resources. R package version 2.14.5. 2017.

MORLEY, M. *et al.* Genetic analysis of genome-wide variation in human gene expression. **Nature,** v. 430, n. 7001, p. 743-747, Aug 12 2004.

MOTT, R. *et al.* The architecture of parent-of-origin effects in mice. **Cell,** v. 156, n. 1-2, p. 332-342, Jan 16 2014.

MOYERBRAILEAN, G. A. *et al.* High-throughput allele-specific expression across 250 environmental conditions. **Genome Res,** v. 26, n. 12, p. 1627-1638, Dec 2016.

NARDINI, C.; DENT, J.; TIERI, P. Editorial: Multi-omic data integration. **Front Cell Dev Biol,** v. 3, p. 46, 2015.

NIEUWINT, A. *et al.* 'Identical' twins with discordant karyotypes. **Prenat Diagn,** v. 19, n. 1, p. 72-76, Jan 1999.

ORSTAVIK, R. E. *et al.* Non-random X chromosome inactivation in an affected twin in a monozygotic twin pair discordant for Wiedemann-Beckwith syndrome. **Am. J. Med. Genet.,** v. 56, p. 210-214, 1995.

OUYANG, Z. *et al.* The landscape of the A-to-I RNA editome from 462 human genomes. **Sci Rep,** v. 8, n. 1, p. 12069, Aug 13 2018.

PANOUSIS, N. I. *et al.* Allelic mapping bias in RNA-sequencing is not a major confounder in eQTL studies. **Genome Biol,** v. 15, n. 9, p. 467, Sep 20 2014.

PANT, P. V. *et al.* Analysis of allelic differential expression in human white blood cells. **Genome Res,** v. 16, n. 3, p. 331-339, Mar 2006.

PEREZ, J. D. *et al.* Quantitative and functional interrogation of parent-of-origin allelic expression biases in the brain. **Elife,** v. 4, p. e07860, Jul 3 2015.

PESSIA, E. *et al.* Mammalian X chromosome inactivation evolved as a dosage-compensation mechanism for dosage-sensitive genes on the X chromosome. **Proc Natl Acad Sci U S A,** v. 109, n. 14, p. 5346-5351, Apr 03 2012.

PETTIGREW, K. A. *et al.* Further evidence for a parent-of-origin effect at the NOP9 locus on language-related phenotypes. **J Neurodev Disord,** v. 8, p. 24, 2016.

PICKRELL, J. K.; GILAD, Y.; PRITCHARD, J. K. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". **Science,** v. 335, n. 6074, p. 1302; author reply 1302, Mar 16 2012.

PIRINEN, M. *et al.* Assessing allele-specific expression across multiple tissues from RNA-seq read data. **Bioinformatics,** v. 31, n. 15, p. 2497-2504, Aug 1 2015.

PISKOL, R. *et al.* Lack of evidence for existence of noncanonical RNA editing. **Nat Biotechnol,** v. 31, n. 1, p. 19-20, Jan 2013.

RAGHUPATHY, N. *et al.* Hierarchical analysis of RNA-seq reads improves the accuracy of allele-specific expression. **Bioinformatics,** v. 34, n. 13, p. 2177-2184, Jul 1 2018.

RAMASWAMI, G.; LI, J. B. RADAR: a rigorously annotated database of A-to-I RNA editing. **Nucleic Acids Res,** v. 42, n. Database issue, p. D109-113, Jan 2014.

REIK, W. *et al.* Imprinted genes and the coordination of fetal and postnatal growth in mammals. **Novartis Found Symp,** v. 237, p. 19-31; discussion 31-42, 2001.

REIK, W.; DEAN, W. DNA methylation and mammalian epigenetics. **Electrophoresis,** v. 22, n. 14, p. 2838-2843, Aug 2001.

REIK, W.; WALTER, J. Genomic imprinting: parental influence on the genome. **Nat Rev Genet,** v. 2, n. 1, p. 21-32, Jan 2001.

RICHARD ALBERT, J. *et al.* Development and application of an integrated allele-specific pipeline for methylomic and epigenomic analysis (MEA). **BMC Genomics,** v. 19, n. 1, p. 463, Jun 15 2018.

SANTONI, F. A. *et al.* Detection of Imprinted Genes by Single-Cell Allele-Specific Gene Expression. **Am J Hum Genet,** v. 100, n. 3, p. 444-453, Mar 2 2017.

SAVOVA, V. *et al.* Genes with monoallelic expression contribute disproportionately to genetic diversity in humans. **Nat Genet,** v. 48, n. 3, p. 231-237, Mar 2016a.

SAVOVA, V. *et al.* dbMAE: the database of autosomal monoallelic expression. **Nucleic Acids Res,** v. 44, n. D1, p. D753-756, Jan 4 2016b.

SAVOVA, V. *et al.* Risk alleles of genes with monoallelic expression are enriched in gain-of-function variants and depleted in loss-of-function variants for neurodevelopmental disorders. **Mol Psychiatry,** v. 22, n. 12, p. 1785-1794, Dec 2017.

SCOTT, J. M.; FERGUSON-SMITH, M. A. Heterokaryotypic monozygotic twins and the acardiac monster. **J Obstet Gynaecol Br Commonw,** v. 80, n. 1, p. 52-59, Jan 1973.

SEREEWATTANAWOOT, S. *et al.* Identification of potential regulatory mutations using multi-omics analysis and haplotyping of lung adenocarcinoma cell lines. **Sci Rep,** v. 8, n. 1, p. 4926, Mar 21 2018.

SERRE, D. *et al.* Differential allelic expression in the human genome: a robust approach to identify genetic and epigenetic cis-acting mechanisms regulating gene expression. **PLoS Genet,** v. 4, n. 2, p. e1000006, Feb 29 2008.

SHERRY, S. T. *et al.* dbSNP: the NCBI database of genetic variation. **Nucleic Acids Res,** v. 29, n. 1, p. 308-311, Jan 1 2001.

SHIBA, H.; TAKAYAMA, S. Epigenetic regulation of monoallelic gene expression. **Dev Growth Differ,** v. 54, n. 1, p. 120-128, Jan 2012.

SHVETSOVA, E. *et al.* Skewed X-inactivation is common in the general female population. **Eur J Hum Genet,** v. 27, n. 3, p. 455-465, Mar 2019.

SKIPPER, M. Gene expression - One allele or two? **Nat. Rev. Genet.,** v. 9, p. 4-5, 2008.

SMIGRODZKI, R. M.; KHAN, S. M. Mitochondrial microheteroplasmy and a theory of aging and age-related disease. **Rejuvenation Res,** v. 8, n. 3, p. 172-198, Fall 2005.

SODERLUND, C. A.; NELSON, W. M.; GOFF, S. A. Allele Workbench: transcriptome pipeline and interactive graphics for allele-specific expression. **PLoS One,** v. 9, n. 12, p. e115740, 2014.

SOUREN, N. Y. *et al.* Mitochondrial DNA Variation and Heteroplasmy in Monozygotic Twins Clinically Discordant for Multiple Sclerosis. **Hum Mutat,** v. 37, n. 8, p. 765-775, Aug 2016.

STEFANO, G. B. *et al.* Mitochondrial Heteroplasmy. **Adv Exp Med Biol,** v. 982, p. 577-594, 2017.

SUN, C. *et al.* Effects of early-life environment and epigenetics on cardiovascular disease risk in children: highlighting the role of twin studies. **Pediatr Res,** v. 73, n. 4 Pt 2, p. 523-530, Apr 2013.

SYMMONS, O. *et al.* Allele-specific RNA imaging shows that allelic imbalances can arise in tissues through transcriptional bursting. **PLoS Genet,** v. 15, n. 1, p. e1007874, Jan 2019.

TACHON, G. *et al.* Discordant sex in monozygotic XXY/XX twins: a case report. **Hum Reprod,** v. 29, n. 12, p. 2814-2820, Dec 2014.

TAN, M. H. *et al.* Dynamic landscape and regulation of RNA editing in mammals. **Nature,** v. 550, n. 7675, p. 249-254, Oct 11 2017.

TARUTANI, Y.; TAKAYAMA, S. Monoallelic gene expression and its mechanisms. **Curr Opin Plant Biol,** v. 14, n. 5, p. 608-613, Oct 2011.

TORDINI, F. *et al.* The Genome Conformation As an Integrator of Multi-Omic Data: The Example of Damage Spreading in Cancer. **Front Genet,** v. 7, p. 194, 2016.

TUKIAINEN, T. *et al.* Landscape of X chromosome inactivation across human tissues. **Nature,** v. 550, n. 7675, p. 244-248, Oct 11 2017.

TYCKO, B. Allele-specific DNA methylation: beyond imprinting. **Hum Mol Genet,** v. 19, n. R2, p. R210-220, Oct 15 2010.

VALLOT, C.; ROUGEULLE, C. Long non-coding RNAs and human X-chromosome regulation: a coat for the active X chromosome. **RNA Biol,** v. 10, n. 8, p. 1262-1265, Aug 2013.

VAN BAAK, T. E. *et al.* Epigenetic supersimilarity of monozygotic twin pairs. **Genome Biol,** v. 19, n. 1, p. 2, Jan 9 2018.

VAN DEN BERG, I. M. *et al.* X chromosome inactivation is initiated in human preimplantation embryos. **Am J Hum Genet,** v. 84, n. 6, p. 771-779, Jun 2009.

VAN DER AUWERA, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. **Curr Protoc Bioinformatics,** v. 43, p. 11 10 11-11 10 33, 2013.

VASER, R. *et al.* SIFT missense predictions for genomes. **Nat Protoc,** v. 11, n. 1, p. 1-9, Jan 2016.

VIGNEAU, S. *et al.* High prevalence of clonal monoallelic expression. **Nat Genet,** v. 50, n. 9, p. 1198-1199, Sep 2018.

VON HIPPEL, P. T. The heterogeneity statistic I(2) can be biased in small meta-analyses. **BMC Med Res Methodol,** v. 15, p. 35, Apr 14 2015.

WAINER KATSIR, K.; LINIAL, M. Human genes escaping X-inactivation revealed by single cell expression data. **BMC Genomics,** v. 20, n. 1, p. 201, Mar 12 2019.

WANG, M. *et al.* Multi-omics maps of cotton fibre reveal epigenetic basis for staged single-cell differentiation. **Nucleic Acids Res,** v. 44, n. 9, p. 4067-4079, May 19 2016.

WANG, X.; CLARK, A. G. Using next-generation RNA sequencing to identify imprinted genes. **Heredity (Edinb),** v. 113, n. 2, p. 156-166, Aug 2014.

WANG, X.; SOLOWAY, P. D.; CLARK, A. G. Paternally biased X inactivation in mouse neonatal brain. **Genome Biol,** v. 11, n. 7, p. R79, 2010.

WEI, Y. *et al.* MetaImprint: an information repository of mammalian imprinted genes. **Development,** v. 141, n. 12, p. 2516-2523, Jun 2014.

WEISSBEIN, U. *et al.* Analysis of chromosomal aberrations and recombination by allelic bias in RNA-Seq. **Nat Commun,** v. 7, p. 12144, Jul 7 2016.

WEKSBERG, R. *et al.* Discordant KCNQ1OT1 imprinting in sets of monozygotic twins discordant for Beckwith-Wiedemann syndrome. **Hum Mol Genet,** v. 11, n. 11, p. 1317-1325, May 15 2002.

WHITAKER, J. W. *et al.* Integrative omics analysis of rheumatoid arthritis identifies non-obvious therapeutic targets. **PLoS One,** v. 10, n. 4, p. e0124254, 2015.

WITZANY, G.; BALUSKA, F. Life's code script does not code itself. The machine metaphor for living organisms is outdated. **EMBO Rep,** v. 13, n. 12, p. 1054-1056, Dec 2012.

WOOD, D. L. *et al.* Recommendations for Accurate Resolution of Gene and Isoform Allele-Specific Expression in RNA-Seq Data. **PLoS One,** v. 10, n. 5, p. e0126911, 2015.

WUTZ, A. Gene silencing in X-chromosome inactivation: advances in understanding facultative heterochromatin formation. **Nat Rev Genet,** v. 12, n. 8, p. 542-553, Jul 18 2011.

YAN, H. *et al.* Allelic variation in human gene expression. **Science,** v. 297, n. 5584, p. 1143, Aug 16 2002.

YAN, J. *et al.* Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. **Brief Bioinform**, Jun 30 2017.

YOUNG, P. E. *et al.* Epigenetic differences between monozygotic twins discordant for amyotrophic lateral sclerosis (ALS) provide clues to disease pathogenesis. **PLoS One,** v. 12, n. 8, p. e0182638, 2017.

ZAKHAROVA, I. S.; SHEVCHENKO, A. I.; ZAKIAN, S. M. Monoallelic gene expression in mammals. **Chromosoma,** v. 118, n. 3, p. 279-290, Jun 2009.

ZHOU, Z. Y. *et al.* Genome wide analyses uncover allele-specific RNA editing in human and mouse. **Nucleic Acids Res,** v. 46, n. 17, p. 8888-8897, Sep 28 2018.

frontiers in Genetics

ORIGINAL RESEARCH
published: 26 November 2019
doi: 10.3389/fgene.2019.01178

Pervasive Inter-Individual Variation in Allele-Specific Expression in Monozygotic Twins

Ronaldo da Silva Francisco Junior[1†‡], Cristina dos Santos Ferreira[2‡], Juan Carlo Santos e Silva[2], Douglas Terra Machado[2], Yasmmin Côrtes Martins [1], Victor Ramos[3†], Gustavo Simões Carnivali[4], Ana Beatriz Garcia[2] and Enrique Medina-Acosta[2*]*

[1] Laboratório de Bioinformática, Laboratório Nacional de Computação Científica, Petrópolis, Brazil, [2] Laboratório de Biotecnologia, Núcleo de Diagnóstico e Investigação Molecular, Universidade Estadual do Norte Fluminense, Campos dos Goytacazes, Brazil, [3] Department of Genetics, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, Brazil, [4] Department of Computational Science, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

Despite being developed from one zygote, heterokaryotypic monozygotic (MZ) co-twins exhibit discordant karyotypes. Epigenomic studies in biological samples from heterokaryotypic MZ co-twins are of the most significant value for assessing the effects on gene- and allele-specific expression of an extranumerary chromosomal copy or structural chromosomal disparities in otherwise nearly identical germline genetic contributions. Here, we use RNA-Seq data from existing repositories to establish within-pair correlations for the breadth and magnitude of allele-specific expression (ASE) in heterokaryotypic MZ co-twins discordant for trisomy 21 and maternal 21q inheritance, as well as homokaryotypic co-twins. We show that there is a genome-wide disparity at ASE sites between the heterokaryotypic MZ co-twins. Although most of the disparity corresponds to changes in the magnitude of biallelic imbalance, ASE sites switching from either strictly monoallelic to biallelic imbalance or the reverse occur in few genes that are known or predicted to be imprinted, subject to X-chromosome inactivation or A-to-I(G) RNA edited. We also uncovered comparable ASE differences between homokaryotypic MZ twins. The extent of ASE discordance in MZ twins (2.7%) was about 10-fold lower than the expected between pairs of unrelated, non-twin males or females. The results indicate that the observed within-pair dissimilarities in breadth and magnitude of ASE sites in the heterokaryotypic MZ co-twins could not solely be attributable to the aneuploidy and the missing allelic heritability at 21q.

Keywords: allele-specific expression, allele imbalance, Down syndrome, genomic imprinting, heterokaryotypic monozygotic co-twins, mitochondrial heteroplasmy, random monoallelic expression, trisomy 21

OPEN ACCESS

Edited by:
Kazuhiko Nakabayashi, National Center for Child Health and Development (NCCHD), Japan

Reviewed by:
Xu Wang, Auburn University, United States
Miho Ishida, University College London, United Kingdom

***Correspondence:**
Ronaldo da Silva Francisco Junior
ronaldoj@lncc.br
Enrique Medina-Acosta
quique@uenf.br

†Present address:
Ronaldo da Silva Francisco Junior, Laboratório de Bioinformática, Laboratório Nacional de Computação Científica, Petrópolis Brazil
Victor Ramos, Laboratory of Molecular Immunology, The Rockefeller University, New York, NY, United States

‡*These authors have contributed equally to this work*

Specialty section:
This article was submitted to Epigenomics and Epigenetics, a section of the journal Frontiers in Genetics

Received: *11 July 2019*
Accepted: *24 October 2019*
Published: *26 November 2019*

Citation:
da Silva Francisco Junior R, dos Santos Ferreira C, Santos e Silva JC, Terra Machado D, Côrtes Martins Y, Ramos V, Simões Carnivali G, Garcia AB and Medina-Acosta E (2019) Pervasive Inter-Individual Variation in Allele-Specific Expression in Monozygotic Twins. Front. Genet. 10:1178. doi: 10.3389/fgene.2019.01178

INTRODUCTION

Monozygotic (MZ) twinning entails the partitioning of progenitor cells derived from one zygote collapsing into two sets that form two separate fetuses (co-twins) of nearly identical genotypes. MZ co-twins develop through monochorionic or dichorionic placentation as a result of when the sets of progenitor cells are split. The exact mechanisms that trigger MZ twinning are vague but genetic

(Liu et al., 2018), epigenetic, and environmental factors have been implicated (Knopman et al., 2014).

A considerable body of experimental evidence demonstrates that most MZ co-twins are not identical but discordant for (epi)genetic traits (Bennett et al., 2008; Baranzini et al., 2010; Furukawa et al., 2013; Souren et al., 2016) and congenital diseases (Chaiyasap et al., 2014; Huang et al., 2019). In stark contrast to homokaryotypic MZ co-twins, the heterokaryotypic MZ co-twins differ for constitutive chromosomal anomalies (Scott and Ferguson-Smith, 1973; Nieuwint et al., 1999). Typically, a pair of heterokaryotypic MZ co-twins exhibits discordant karyotypes for autosomal or gonosomal aneuploidies (i.e., trisomy 21, trisomy 13, XO or XXY) arising most likely post-zygotically and leading to mosaicism at various degrees (Gilbert et al., 2002; Tachon et al., 2014). Heterokaryotypic MZ co-twins may be discordant for structural chromosomal rearrangements (Leung et al., 2009; Essaoui et al., 2013), including genome-wide copy number variation (CNV) that is also commonplace in homokaryotypic MZ twins (Abdellaoui et al., 2015; Huang et al., 2019). Other likely causes for genotypic discordance in MZ monochorionic co-twins include alterations in gene expression (Buil et al., 2015), parent-of-origin effects associated to abnormal non-random (skewed) X-chromosome inactivation (XCI) (Orstavik et al., 1995), and genomic imprinting (Weksberg et al., 2002; Begemann et al., 2018). There are 43 well-documented cases of heterokaryotypic MZ co-twins in humans (**Table S1**). Most of the reported cases are spontaneous pregnancies, rather than associated with assisted reproductive technology.

Epigenomic studies in heterokaryotypic MZ co-twins are of the most significant value for assessing the effects on gene- and allele-specific expression of an extranumerary chromosomal copy or structural chromosomal disparities in otherwise nearly identical genomes.

Oligo microarray (Yan et al., 2002; Lo et al., 2003; Morley et al., 2004) and genome-wide transcriptome shotgun sequencing (RNA-Seq) studies in multiple biological samples have unveiled that many genes are subjected to the differential transcriptional expression of one allele of a pair of alleles (Dixon et al., 2015; Pirinen et al., 2015; Weissbein et al., 2016). Allele-specific expression (ASE) refers to the departure from the Mendelian 1:1 allelic expression ratio assumption. Typically, the patterns of allele expression include symmetrically (strictly) biallelic, asymmetrically biallelic (biallelic imbalance or allelic bias), and strictly monoallelic (Dixon et al., 2015; Pirinen et al., 2015; Weissbein et al., 2016).

RNA-Seq analysis allows determining the breadth and magnitude of ASE sites simultaneously. At a given experimental condition, each cell type should exhibit an array of ASE sites, an ASE signature, or transcriptome fingerprint, which is expected to be remarkably particular to the individual biological sample. The ASE signatures may be altered by environmental, health, and disease conditions (Moyerbrailean et al., 2016; Weissbein et al., 2016). In essence, the same source of cells from MZ co-twins should exhibit identical ASE signatures. However, studies based on transcriptome sequence analysis disclosed widespread discordance in ASE sites in biological samples from apparently healthy homokaryotypic MZ twins (Cheung et al., 2008; Buil et al., 2015). Therefore, at the RNA level, the occurrence of ASE discordance constitutes a form of a cryptic, unexplained/missing heritability in individuals who share, in principle, "identical" genomes. On the other hand, genome-wide ASE discordance implies that the mechanisms for reliable transfer or flow of genetic information from DNA to RNA within humans are loose, with profound implication(s) for human health and disease (Chakravarti, 2011).

The causes of ASE discordance are associated with (epi) genetic factors, gene-gene, and gene-environment interactions (**Figure 1**, **Dataset S1**). For genes that are not subjected to either epigenetic regulatory mechanisms such as genomic imprinting (Baran et al., 2015), and XCI (Tukiainen et al., 2017), ASE mostly relates to the expression effects associated to quantitative trait loci (eQTLs), which can be ascribed to sequence variants of both alleles (cis effect), whereas the extent of the ASE effect relies on trans genetic variants and environmental factors interacting with the cis genetic variants (Buil et al., 2015).

Furthermore, over 2.6 million ribonucleotide sites are known to be post-transcriptionally subjected to allele-specific editing at varying extents in several human tissues, thus contributing, at a much higher degree, to the phenotypic expression of likely mutational sites in the form of differential epitranscriptomes (Li et al., 2009; Ramaswami and Li, 2014; Zhou et al., 2018). Among the genetic factors, there are also differences in meiotic recombination and chromosomal aberrations (Weissbein et al., 2016).

Here, we carried out a comparative computation analysis of RNA-Seq data from heterokaryotypic MZ co-twins discordant for trisomy 21 and homokaryotypic MZ co-twins. We cross-referenced the ASE sites with public data repositories to exemplify the sources and consequences of within-pair disparities to annotate ASE effects in genes that are subjected to the (epi)genetic processes of genomic imprinting, XCI, and RNA editing. We identified considerable ASE disparity between either heterokaryotypic or homokaryotypic co-twins.

MATERIALS AND METHODS

Bioprojects

We used primary (unprocessed) RNA sequence filed data from the Sequence Read Archive (SRA) public experiments in 10 twin pairs, being one pair of heterokaryotypic co-twins and nine pairs of homokaryotypic co-twins. The biological samples included: primary fetal fibroblasts (GEO BioProject PRJNA239814) from the study by Letourneau and collaborators (2014), induced pluripotent stem cells (iPSC) from the study by (Hibaoui et al., 2014) (GEO BioProject PRJNA227902), and cultured B-cells (Epstein-Barr virus transformed lymphoblastoid cell lines from peripheral adult blood B-lymphocyte; GEO BioProject PRJNA170210) (**Dataset S2**). We selected the transcriptome study by Letourneau and collaborators (2014) on a pair of MZ co-twins who were karyotypically discordant for trisomy 21 (T21) of maternal origin (Dahoun et al., 2008), and therefore are heterokaryotypic twins (i.e., co-twins that differ concerning constitutive chromosomal anomalies).

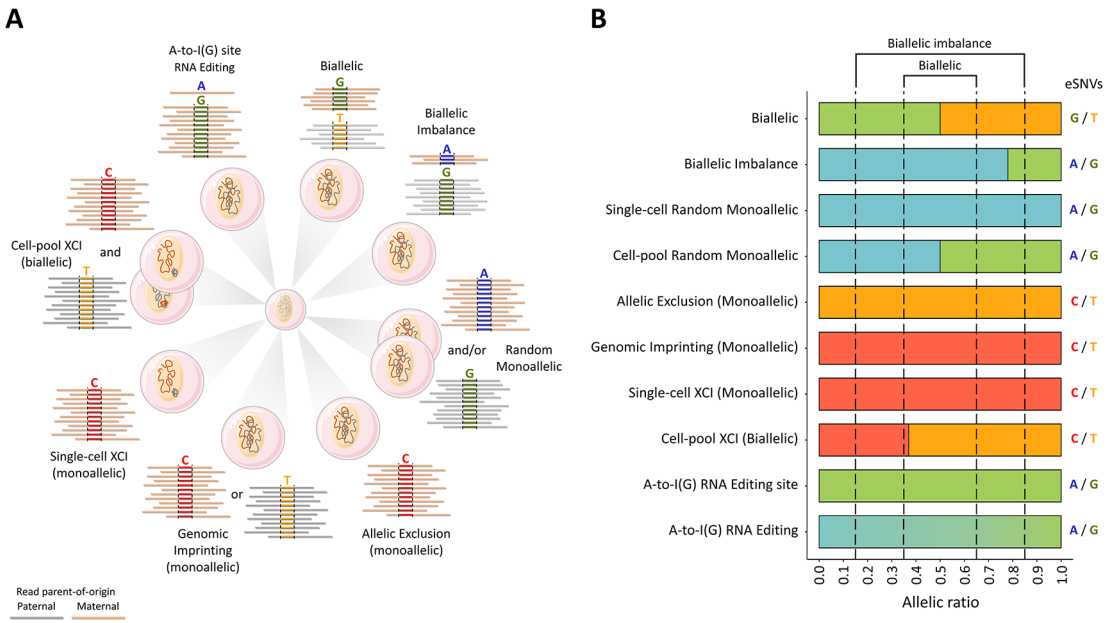


FIGURE 1 | Epigenetic processes involved in allele-specific RNA expression. **(A)** The differential allele expression of genes best reflects dynamic regulation processes consistent with either an allele being preferentially silenced or an inactive allele being restored. The scenarios are for total steady-state RNA, for which a minimum of 12 reads are depicted across the reference and alternative alleles at hypothetic heterozygote or A-to-I(G) RNA editing sites. The breadth and magnitude of the deviation from the expected strictly biallelic 6:6 read ratio may be ascribed to one of several epigenetic regulatory processes involving compensatory and non-compensatory cis-acting variation epistatic to trans-acting variation. The scenarios are organized clockwise: strictly biallelic, biallelic imbalance, random monoallelic, allelic exclusion, genomic imprinting, single-cell X-chromosome inactivation (XCI), cell-pool XCI, and RNA editing. Up to 30% of all tested protein-coding autosomal genes are subjected to clonal (mitotically) stable, random monoallelic expression, which can be either coordinated or uncoordinated (Gimelbrant et al., 2007; Savova et al., 2016a; Savova et al., 2016b; Vigneau et al., 2018). Up to 23% of genes linked to the X-chromosome are expressed from the inactive X (i.e., XCI escapee genes) and, therefore, are biallelically expressed in each female somatic cell (Tukiainen et al., 2017). About 2.6 million ribonucleotide sites genome-wide are known to be subjected to A-to-I(G) RNA editing (Ramaswami and Li, 2014). Thus, the human tissues are, in essence, expression mosaics due to epigenetic-, cis-, and trans-acting covariates. **(B)** The extent of the allele-specific expression for the scenarios illustrated in panel **(A)** using RNA-Seq reads across SNVs in genes known to be subjected to the indicated regulatory processes. *WRB* (biallelic) (Alves Da Silva et al., 2016; De Sa Machado Araujo et al., 2018), *SH3BP5L* (biallelic imbalance) (Baran et al., 2015), *EVC* (random monoallelic) (Gimelbrant et al., 2007), *SNURF* (maternally imprinted) (Gray et al., 1999; De Sa Machado Araujo et al., 2018), *OR2L13* (allelic exclusion) (De Sa Machado Araujo et al., 2018), *DGKZP1* and *AL391244.3* (RNA editing; the present study), *FMR1* (subject to XCI) (Tukiainen et al., 2017).The data supporting the allele ratios depicted in the histograms are presented in **Dataset S1**.

Comparative transcriptomics in these heterokaryotypic twins lead to the proposal of the so-called domains of genome-wide gene expression dysregulation in Down syndrome (Letourneau et al., 2014). The case is emblematic because, in addition to the discordant maternal T21 aneuploidy, primary fetal fibroblasts from the MZ twins exhibited missing allelic heritability at 21qter as a result of recombination event(s) (Dahoun et al., 2008). A diagram of the discordant maternal 21q inheritance in the pair of co-twins heterokaryotypic for trisomy 21 is represented in **Figure S1**. To estimate the extent and magnitude of ASE discordance in unrelated, non-twin individuals, we included the RNA-Seq run experiments from BioProject PRJNA316578 (**Dataset S2**), which comprises whole blood samples from two males and two females, mean age 34-year-old, healthy controls.

Identification, Quantification, and Sorting Out Allele-Specific Expression Sites in Transcriptome Data

We implemented PipASE, an in-house computational pipeline to identify, quantify, and sort out ASE sites in the transcriptome data (**Figure S2**). PipASE scans genome-wide for expressed single nucleotide variants (eSNVs) in high quality aligned reads. We recognize that RNA-Seq read counts and, therefore, expressed allele rates, maybe artifactually made discordant between co-twins as a result from sequencing chemistry and forward/reverse strand biases in the error rate of the high-throughput sequencing technology (Heap et al., 2010; Pickrell et al., 2012; Liu et al., 2014; Soderlund et al., 2014; Hu et al., 2015; Wood et al., 2015; Raghupathy et al., 2018; Richard Albert et al., 2018). Therefore, primary sources of technical artifacts such as systematic errors in sequencing and mapping sequence reads to a haploid reference genome were curbed by including in the PipASE the following specific algorithms that reduce or control the mapping bias: i) relaxing the number of mismatches admitted per string, yet excluding reads with spurious mismatches at the last bases of reads aligning just to one DNA strand; ii) excluding reads aligning around insertions, deletions, and simple tandem repeats; iii) excluding reads mapping to paralogous genomic regions (i.e., segmental duplications); iv) requiring ≥ 12 high-quality read depth to call a candidate informative site, and v) prioritizing the ranking of ASE sites by multiple consistent expression patterns.

Raw reads were trimmed with Trimmomatic (Bolger et al., 2014), and aligned to the hg38 reference genome using the Spliced Transcripts Alignment to a Reference (STAR, v3.5a) software (Dobin et al., 2013). We required uniquely and high-quality mapped reads (MAPQ ≥ 30) by filtering them using the sequence alignment/map tools (SAMtools) (Li et al., 2009). We processed the RNA-Seq data according to the best practice guidance using the ASEReadCounter tool from the open-source Genome Analysis Toolkit (GATK, v3.8), instrumented for variant discovery in high-throughput sequencing data (McKenna et al., 2010; Depristo et al., 2011; Van Der Auwera et al., 2013). Annotated single nucleotide polymorphisms (SNP) and private SNVs were identified using HaplotypeCaller from GATK at each hypothetical heterozygous position according to HapMap (International HapMap, 2003) and database of SNP (Sherry et al., 2001). The annotation of ASE variant site positions to the hg38 reference genome was performed using the R/Bioconductor biomaRt package (Durinck et al., 2005; Durinck et al., 2009). SNP population data (MAF, ancestral allele) were integrated using rsnps package version 0.3.0 (Chamberlain et al., 2018). For the assessment of ASE, the read counts from the replicas were amalgamated, and Q1 values across each informative eSNV site were calculated for all biosamples on a per twin basis. For ASE sites that occurred only once in each set of biosamples, the ASE value was given by the informative run. Thus, ASE sites are supported by at least one informative run. For example, BioProject PRJNA239814, which refers to fetal fibroblasts biosamples collected from the MZ twin pair discordant for trisomy 21, comprises 12 RNA-Seq run experiments, being six per twin. The project includes four biosamples for each twin, and two of which are replicas. For that project, the distribution of informative ASE sites is 51.7, 18.2, 18.5, and 11.6% sites supported by at least 1, 2, 3, and 4 biosamples, respectively. ASE across imputed heterozygous SNP sites was calculated as the difference of RNA-Seq read counts between the two alleles, using the equation ASE =|0.5 — Ref _allele_read count / (Ref _allele _read count + Alt _allele _read count)| The allelic expression imbalance value per site (ranging between 0 and 0.5) is, therefore, a measure of departure from the expected Mendelian 1:1 allelic expression ratio (Babak et al., 2015; Baran et al., 2015). We annotated the ASE data by calculating the expected null reference/alternative ratios and binomial test P-values (Wang and Clark, 2014) using the *binom.test* R code function (R Core Team, 2019), and according to their gene structure sequence context (exon, intron, 5′ UTR, 3′ UTR, and intergenic) using the GRCh38.92 Ensembl release 96 in gtf format and the *GenomicFeatures* annotation package in R code (Lawrence et al., 2013). I-square statistical test was used to assess the degree of heterogeneity in the ASE profiles of genes supported by multiple eSNVs. The test is based on the chi-square and degree of freedom values, and it was used to measure the inconsistency of ASE profiles in each gene. We ranked genes according to the following criteria: homogeneity (I-square <30%), moderate heterogeneity (between 30 and 50%), substantial heterogeneity (between 50 and 75%), and considerable heterogeneity (> 75%). The negative I-square values were considered as 0% (Wang and Clark, 2014; Von Hippel, 2015). A flowchart for the PipASE used for scanning and sorting out genome-wide, allele-specific differences between MZ co-twins is shown in **Figure S2**.

Cross-Referencing With Public Data Repositories

For every ASE site observed in each RNA-Seq sample, we extracted functional information by computational cross-referencing with public databases regarding pathogenic expression-altering or loss-of-function risk variant alleles (Adzhubei et al., 2010; Landrum et al., 2016; Vaser et al., 2016), genomic imprinted genes (Jirtle and Murphy, 2012; Wei et al., 2014; Baran et al., 2015; Pirinen et al., 2015), A-to-I(G) RNA editing sites (Ramaswami and Li, 2014), germline ASE discordant sites in MZ twins (Cheung et al., 2008), and XCI escapee and non-escapee genes (Carrel and Willard, 2005; Cotton et al., 2013; Balaton et al., 2015; Cotton et al., 2015; Tukiainen et al., 2017; Garieri et al., 2018; Shvetsova et al., 2019; Wainer Katsir and Linial, 2019). Allelic expression profiles were validated computationally by data integration with the ASE profiles observed in multiple human tissues from the Genotype-Tissue Expression (GTEx) project (The GTEX Project, 2015), using the *Data Integrator* tool available at the UCSC Genome Browser, that contains track hubs for the second source GTEx data (release V6, October 2015), mainly as previously reported (De Sa Machado Araujo et al., 2018).

Canonical A-to-I(G) Ribonucleic Acid Editing

ASE sites were queried in the RADAR database, which comprises a list of about 2.6 million rigorously annotated database of A-to-I(G) RNA editing sites. For cross-referencing of the ASE sites, we merged RADAR data version 1 (available online from the RADAR browser) and version 2, which is based on the GTEx RNA-Seq dataset from 30 tissues (hg19; version 6p), and reports RNA editing levels for sites with ≥20 reads (Tan et al., 2017), kindly provided as a flat database by Dr. Jin Billy Li at Stanford University (Ramaswami and Li, 2014). The hg19 coordinates were lifted over to hg38 using "hg19ToHg38.over.chain" file and R scripts based on *AnnotationHub* (Morgan, 2017) and rtracklayer libraries (Lawrence et al., 2009). We limited the analysis to base positions corresponding to canonical A-to-I(G) variants, excluding all SNVs that map within segmental duplications or simple repeats in the hg38 reference genome, using the ShortMatch tool with query strings of 50 bases in length containing the variant at position 26th. The filter-selection step above followed published quality guidelines (Lin et al., 2012b; Ramaswami et al., 2012; Piskol et al., 2013). For every ASE site matching a RADAR reference editing site location, we calculated the A-to-I(G) RNA editing levels as the ratio of G-containing reads divided by the sum of A- and G-containing reads in RNA-Seq experiments of each pair of co-twins. The strength of the co-association between the levels of RNA editing at ASE sites within twin-pairs was measured using linear models in R.

RESULTS

Transcriptome-Wide, Allele-Specific Differences Observed in Monozygotic Co-Twins Discordant for Both Trisomy 21 and Recombination

Recombination and sequence variation are major evolutionary sources of diversity in the human genome. We, therefore, wished first to evaluate how these two forces impacted on ASE in "identical" co-twins. Between the MZ co-twins discordant for T21, we identified 1,227 (3.8%) ASE sites whose allelic patterns were discordant (i.e., monoallelic *versus* biallelic) in fibroblasts and 3,295 (6%) such sites in iPSC (**Figure 2A, Dataset S3**). We estimated the magnitude of expression change between conditions for the variants called (**Figure 2B**). The bulk of the ASE sites exhibited a LogASE value close to zero, which means that the majority of the ASE sites were not altered in trisomy 21 condition. Importantly, 19 eSNVs were significantly altered in fibroblasts of the trisomy 21 (T1DS) affected twin, being 16 sites with LogASE ≥ 0.8 and three sites with LogASE ≤ −0.8. Noteworthy, 11 implicated genes mapped to the 21q region discordant for maternal inheritance due to a recombination event. Among those genes, *CASP6*, *FAM86GP,* and *PDXDC1*/*PKD1P6* were expressed monoallelically, whereas the *IL17RA* gene was expressed biallelically in fibroblasts from the T1DS twin. In iPSC, we observed 260 eSNVs with LogASE ≥0.8 and 214 ≤ −0.8 annotated in 274 genes (**Figure 2B**). Of the 19 ASE sites with ASE values ≤ −0.8 or ≥ 0.8 in fibroblasts, 14 were also called in iPSC. However, only 10 sites were altered in both cell types with values of ASE ≥ 0.8 (**Dataset S3**), and are located within the 21q region spanning the recombination event. The overall distribution of genes by the numbers of ASE sites observed in fibroblasts and iPSC is shown in **Figure 2C**.

The discrepancies in ASE between the MZ co-twins discordant for T21 observed in both fibroblasts (**Figure 3A**), and iPSC (**Figure S3A**) were widespread in the genome (average of 20 ASE sites per Mb). We validated the heterokaryotypic status of the MZ twins discordant for T21 by comparing the within-pair global allele-ratios and plotted them as expression karyotypes (e-karyotypes) (**Figure 3B** and **Figure S3B**).

Allele-Specific Expression Disparity Observed in Homokaryotypic Twin-Pairs

To begin to sort out the likely causes of the widespread ASE discordance found in co-twins, we examined the breadth and magnitude of discordant ASE sites in nine pairs of co-twins not discordant for aneuploidy and recombination. Surprisingly, the breadth and magnitude of ASE concordance and discordance in the control twin pairs were comparable to those observed in the heterokaryotypic twin pair, with an average 1,074 discordant sites (2.7%) per twin pair (**Figure S4**). The discordant ASE sites were also distributed genome-wide (**Figure S5**). Despite their

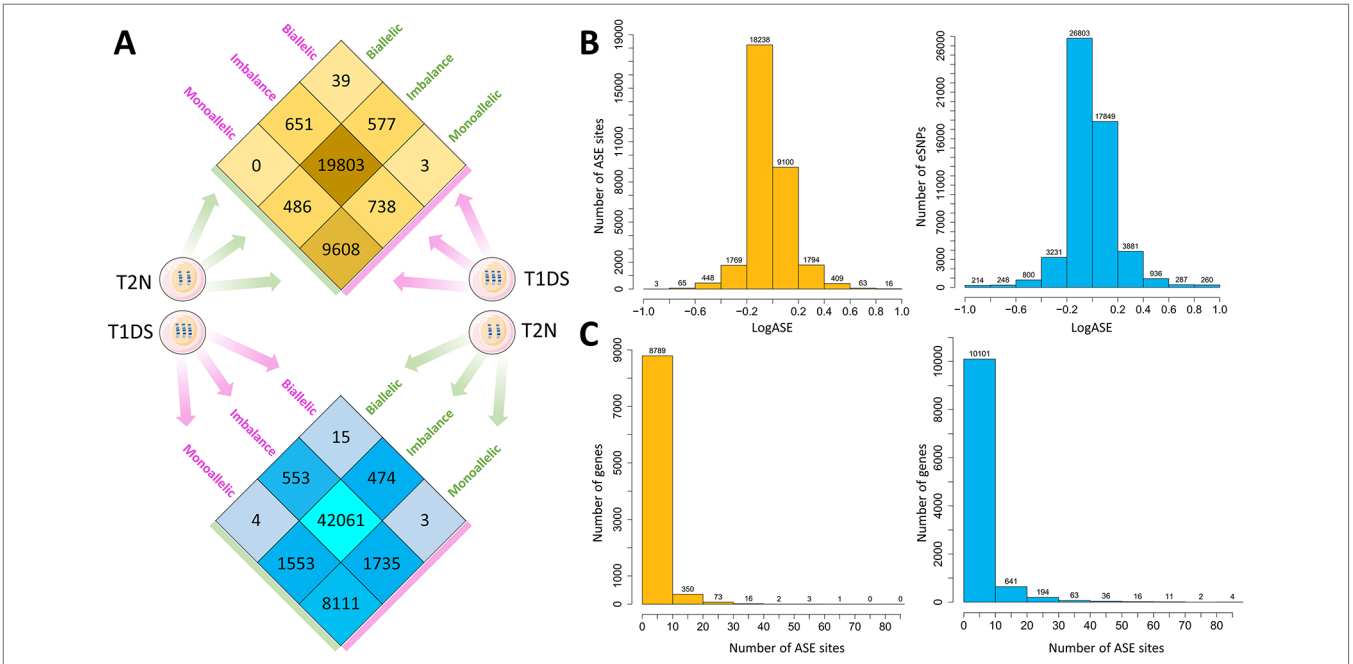


FIGURE 2 | Overview of the breadth and magnitude of allele-specific expression disparity between heterokaryotypic monozygotic (MZ) twins. **(A)** Number of allele-specific expression (ASE) sites distributed by the within-pair status of concordance or discordance in MZ twins heterokatyotypic for trisomy 21 and discordant for maternal 21q inheritance tested in primary fibroblasts (upper panel in orange heat plot) and iPSC (lower panel in blue heat plot). In each cell type, the majority of ASE sites are concordant by biallelic imbalance status in both the trisomy 21 (T1DS) and the normal (T2N) co-twins. On average, the co-twins are discordant in 2,261 ± 1,462.3 ASE sites. **(B)** Comparison of the effect size of the LogASE between fibroblasts and iPSC, respectively. We calculated the log2 of allele-specific expression fold change using the equation *LogASE = log2(T1DS _ ASE / T2N _ ASE)* for each expressed single nucleotide variant in each tissue. LogASE estimates the magnitude of expression change between conditions for the variant. **(C)** Distribution of genes by numbers of ASE sites observed in fibroblasts (orange bars) and iPSC (blue bars).

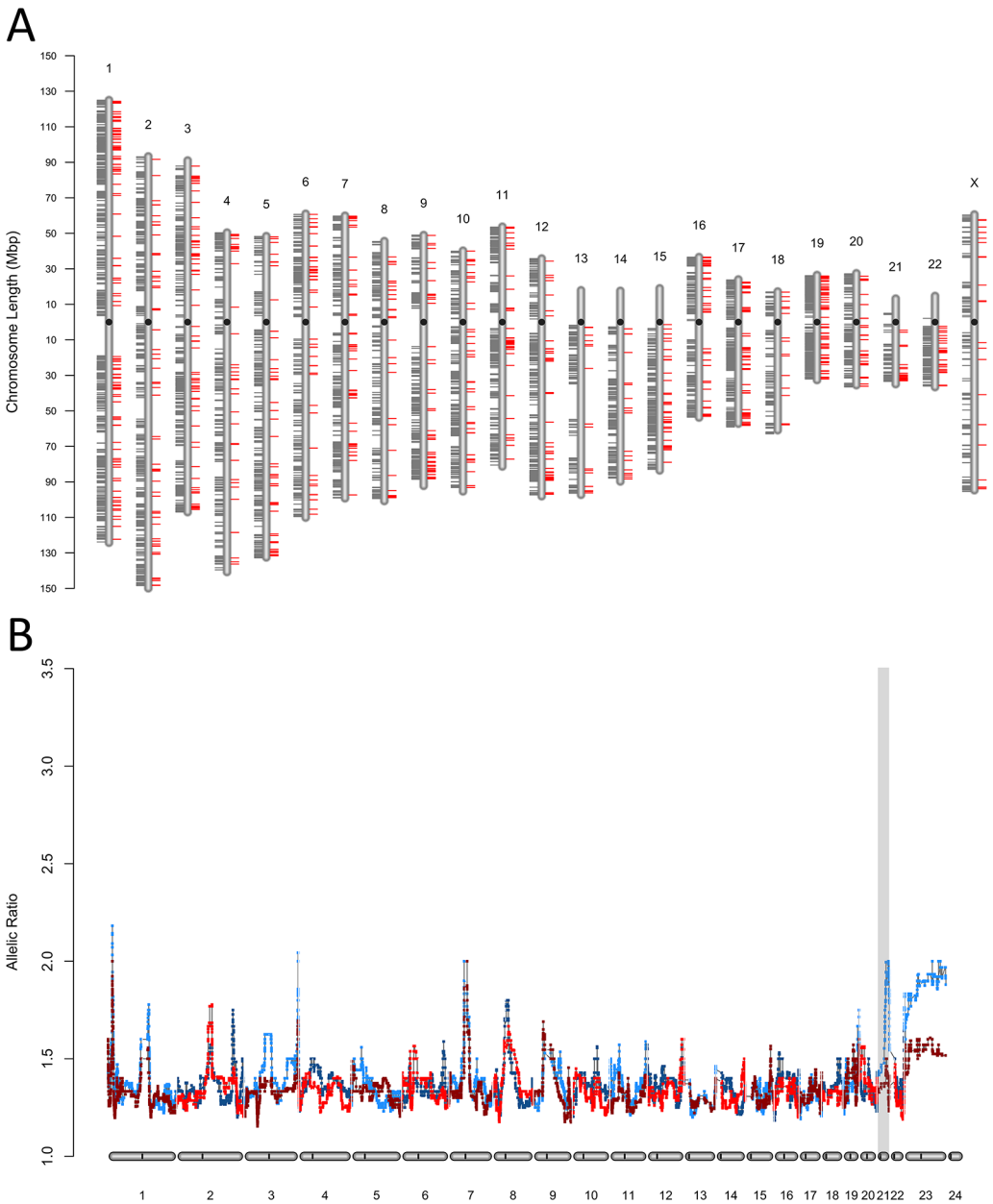


FIGURE 3 | Chromosomal distribution of expressed single nucleotide variants. **(A)** Genome-wide e-karyotyping for the SNPs and variants exhibiting allele-specific expression in primary fetal fibroblasts from the co-twins discordant for T21 and maternal recombination at 21q. Shown is the distribution of all ASE sites that were concordant (gray ticks toward the left of each chromosome ideogram) or discordant (red ticks toward the right side). **(B)** Detection of trisomy 21 by e-karyotyping allelic bias using RNA-Seq data from primary fibroblasts in **(A)**. The gray shading highlights the occurrence of a discordant third copy of chromosome 21 in one twin (T1DS).

diverse parental origins, there were, on average, 19,488 ASE sites common within the nine pairs of homokaryotypic MZ twin pairs; 90 (0.46%) sites were discordant in the entire set of twin pairs. Nevertheless, there were, on average, 571 ASE sites discordant in a given twin pair, but concordant in another. The recurrent sites in all nine pairs best reflect identity by state. We note, however, that monochorionic twin developed with a shared circulation, and therefore, the ASE profiles assessed in cultured transformed B-cells isolated at an early age will tend to be similar. Unfortunately, we could not trace the chorion type of the nine homokaryotypic twin-pairs, which were sampled at the age ranging 19 to 65 (**Dataset S2**).

For the entire set of MZ twin pairs, the average distribution of eSNVs per gene was the following: 34.3% (n = 3,162) of genes were called by one eSNV; 57.8% (n = 5,333) were supported by 2 to 10 eSNVs; 7.9% (n = 729) were called by 11 to 200 eSNVs; 0.02% (n = 2.4) were called by 201 to 500 eSNVs, and 0,01% (n = 1.1) exhibited >500 eSNVs (**Dataset S4A** and **S4B**). We carried

a statistical test for heterogeneity to query for intervention effects (variation in effect estimates beyond chance) across a given genomic region. For the entire set of biosamples, we found, on average, that 43.6% (n = 2,619) of genes supported by multiple eSNVs exhibited considerable homogeneity across the eSNV profiles; 3.8% (n = 225) had moderate heterogeneity; 6.1% (n = 366) had substantial heterogeneity; and 46.5% (n = 2,795) had considerable heterogeneity (**Dataset S4C** and **S4D**). We note that genes exhibiting considerable heterogeneity are large (on average 130 Kbp, i.e., *CD226*) and are supported on average by 7.2 (range 2 to 318) eSNVs. Conversely, the most homogeneous profiles are in genes with an average size of 12 Kbp (i.e., *JRK*), which are supported on average by 3.7 eSNVs (range 2 to 38 sites). Moreover, comparing genes supported by the same number of eSNV (i.e., 30 sites), we note that the eSNVs are distributed differently, toward the 3′UTR in genes ranked as homogeneous (i.e., *LGALS8* and *PLEC*) and spread along the gene body in those ranked as heterogeneous (i.e., *CD226* and *GLEC17A*).

We also validated the homokaryotypic status of the nine MZ control twin pair by comparing the within-pair global allele-ratios and plotted them as e-karyotypes (**Figure S6**). Jointly, the e-karyotyping analyses demonstrate that there is pervasive missing allelic heritability between the transcriptome of MZ co-twins and that the bulk of the ASE site within-pair disparities in the heterokaryotypic co-twins cannot be solely attributed to the differential occurrence of aneuploidy and the missing allelic heritability at 21q.

Allele-Specific Expression Disparity Observed in Unrelated, Non-Twin Males and Females

Unrelated, non-twin males and females exhibited comparable extents of ASE discordance genome-wide: 24.8% (6,546/26,371 eSNV sites) in males (**Dataset S5A**) and 25.57% (5,992/23,431 eSNV sites) in females (**Dataset S5B**). Therefore, the extent of ASE discordance in unrelated, non-twin males and females is about 10-fold higher than the observed between pairs of MZ twin-pairs (2.7%). In the unrelated male and female set, 47.4 and 45.4% of genes supported by ≥ 2 eSNVs, respectively, exhibited considerably heterogeneous ASE profiles, whereas 43.4 and 45.5% of genes were ranked as considerably homogeneous (**Dataset 5C**). Similar to the finding in MZ twins, genes exhibiting considerable heterogeneity are large (on average 83 Kbp, i.e., *GAK* in males and 221 Kbp in females, i.e., *SAMD3*) and are supported on average by 8.2 (range 2 to 104) eSNVs in males and 7.7 (range 2 to 106) eSNVs in females. Conversely, the most homogeneous profiles are in genes with an average size of 18 Kbp (i.e., *UBE2I* in males and *EEF1D* in females), which are supported on average by 3.7 (range 2 to 39) eSNVs in males and 3.5 (2 to 36) eSNVs in females. Again, comparing genes supported by the same number of eSNVs (i.e., 25 sites), we note that the eSNVs are distributed differently, toward the 3′UTR in genes ranked as homogeneous (i.e., *HCP5* and *PRRC2B* in males and *AC004151.1* and *NOTCH1* in females) or spread along the gene body in those ranked as heterogeneous (i.e., *FCGBP* and *GAK* in males and *SAMD3* and *SYNE3* in females).

Assessment of the Underlying Causes of the Observed Pervasive Missing Allelic Inheritability

The underlying causes of the observed pervasive missing allelic inheritability can include i) genome-wide DNA sequence variations within pairs of MZ co-twins, as supported by recent findings in MZ twins discordant for autism spectrum disorder (ASD) using whole-genome sequencing (Huang et al., 2019) and ii) differential expression of alleles. Given that none of the ten MZ twin pairs referred here has genomic sequences available in public repositories, we first cross-referenced the observed ASE sites with data about the distribution of eSNVs reported between the MZ co-twins discordant for ASD (Huang et al., 2019). On average, the MZ co-twins discordant for ASD exhibited 54 eSNVs disparities annotated in exons, 3,912 in introns, 13 in 5′ UTR, and 74 in 3′ UTR for 2,786 genes (**Table S2**). Remarkably, between either the MZ co-twins heterokaryotypic for T21 or the homokaryotypic MZ co-twins, we identified, on average, 10,111 ASE discordant sites in annotated exons, 8,037 in introns, 2,066 in 5′ UTR, and 18,032 in 3 UTR for 8,495 genes. Thus, a mean 120-fold increase in discordant ASE sites per annotation category. This fold difference cannot be attributed solely to the average distribution rate of discordant eSNVs of $1.1x10^{-4}$ per exonic site reported across human genes between the genomes of MZ co-twins (Huang et al., 2019). Furthermore, there is a 20-fold deficit in ASE sites annotated in intergenic regions as compared with the number of eSNV sites discordant by whole genome sequencing, which supports the view that the biased distribution of ASE sites discordances within genes may be biologically relevant. We also validated some of the ASE discordant sites by cross-referencing with the sets ASE sites in MZ twins from the study by Cheung et al. (2008) (**Dataset S3**).

Next, we cross-referenced the ASE sites with data about genes known or predicted to be expressed from one allele at a time through genomic imprinting, XCI, and A-to-I(G) RNA editing. Overall, we identified discordant ASE sites in either 205 known or candidate imprinted genes (**Dataset S3**), 12 X-linked genes (**Table S3**), and 3,955 sites likely subjected to A-to-I(G) RNA editing (**Dataset S3**).

Allele-Specific Expression Switching in Imprinted Genes

We note that, on average, 4,574 ASE sites were monoallelic concordant within co-twins. Annotation of those sites revealed that 8,867 genes exhibited multiple monoallelically eSNVs with no biallelically expressed sites (**Datasets S3, S6–S14**). Among those genes, we annotated five known imprinted genes (*DR1*, *BRD2*, *VARS2*, *MEG3,* and *H19*), each one ranked with ≥8 eSNVs. Cross-reference of those genes with secondary data from the GTEx project validated their monoallelic expression in multiple tissues (**Dataset S15**), and therefore their imprinted status (Jirtle and Murphy, 2012; Wei et al., 2014; Baran et al., 2015; Pirinen et al., 2015). Unfortunately, the GTEx project does not include samples of embryonic fibroblasts, iPSC, or B-cells. In contrast, most other genes ranked with ≥8 monoallelic eSNVs were expressed biallelically in multiple tissues in the GTEx

database (**Dataset S15**). We speculate that the monoallelic concordance at multiple sites observed in the co-twins reflects extended homozygosis, rather than parent-of-origin effects. We note five gene exceptions. First, the *AC091729.3* gene, which exhibited an average of 12 monoallelically eSNVs in two of the ten MZ co-twin pairs, is also expressed monoallelically in an isoform-specific fashion with four SNPs in 35 tissues in the GTEx samples (**Dataset S15**). Second, the *SPINK5* gene with 11 monoallelically eSNVs, expressed monoallelically exclusively in the bladder in the GTEx data. While the ASE profile across the *AC091729.3* gene was homogeneous, the *SPINK5* gene ranked moderately heterogeneous. We view these two genes as potential leads and suggest that the *AC091729.3* gene is subjected to isoform-specific genomic imprinting, whereas the *SPINK5* gene is imprinted in a tissue-specific manner. Third, the known imprinted genes *SNURF*, *SNHG14*, and *ZNF264,* which are expressed monoallelically in multiple tissues in the GTEx samples, exhibited ≥ 10 biallelically imbalance eSNVs in iPSC (**Dataset S15**). Interestingly, various biallelically imbalance eSNVs in these three genes are listed in the RADAR database and are likely subjected to A-to-I(G) RNA-editing: 5 out of 16 eSNVs (*SNURF*), 19/35 (*SNHG14*), 4/14 (*ZNF264*) (**Dataset S3**). We, therefore, suggest that the epitranscriptome modification of these gene products by RNA editing alters their expected imprinting phenotype (at least *in vitro*) in iPSCs.

Estimated Impact of Canonical A-to-I(G) Ribonucleic Acid-Editing on Allele-Specific Expression Disparity

The number of ASE sites that positionally correspond to canonical A-to-I(G) sites was, on average, 2,012 ± 786 per twin pair (**Datasets S3, S6–S14**), and all the sites cover 419 ± 116 genes. The vast majority of sites exhibited a concordant biallelic imbalance profile (pink and light blue dots in **Figure 4** and **Figure S7**). Thus, within the co-twins, there was an overall concordance in the biallelic imbalance state. The number of sites that exhibited discordant allelic profiles, being biallelic in one twin and monoallelic in the other, was however minimal, albeit more abundant in the homokaryotypic than in the heterokaryotypic co-twins (green dots in **Figure 4** and **Figure S7**). This observation indicated that in the cells analyzed, few eSNVs were 100% edited (i.e., expressed strictly monoallelically) in the complete set of 10 pairs of twins. Between the heterokaryotypic co-twins, there were only seven such discordant sites, being monoallelically expressed in T1DS and biallelically imbalance in the normal co-twin (**Figure 4**). The seven discordant sites occur in seven protein-coding genes, including *CD46* (an immune type I receptor) and *ING5* (a tumor suppressor).

In the heterokaryotypic twins, 117 expressed genes exhibited high proportions of ASE sites (≥ 4 sites per gene) coincident with RNA editing sites. Notably, for the *CYP20A1* and *ZNF621* genes, 74 and 77% of all ASE sites are canonical A-to-I(G) sites validated in the RADAR database. The extent of allele imbalance ranged from 6 to 98%. However, about 5% of all sites exhibited discordant RNA editing levels higher than 25% between co-twins,

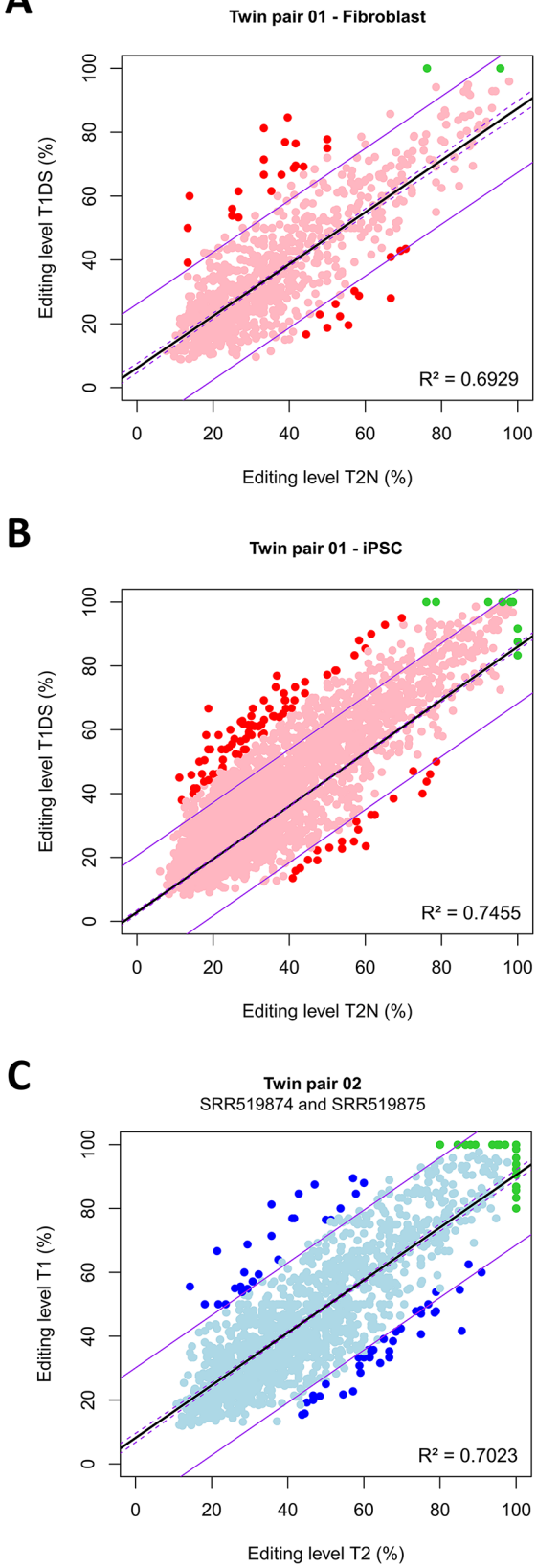


FIGURE 4 | Continued

FIGURE 4 | Within twin-pair disparities in allele expression proportions at expressed single nucleotide variants (eSNVs) that are coincident with canonical A-to-I(G) RNA editing sites. Shown is the distribution of eSNVs that positionally match canonical RNA editing sites between heterokaryotypic co-twins, assayed either in fetal fibroblasts **(A)**, fetal fibroblast-derived iPSC **(B)**, or between homokaryotypic co-twins tested in culture-B-cells **(C)**. Each dot corresponds to an eSNV. The vast majority of sites exhibited a concordant biallelic imbalance profile (pink and light blue dots). Red dots represent eSNVs that were discordant between co-twins in that they showed allelic proportions differences higher than 25%, regardless of the discordance or concordance in the karyotype. Green dots represent eSNVs that exhibited discordant allelic profiles, being biallelic in one twin and monoallelic in the other. The linear models (solid black lines), the confidence interval of the models (broken purple lines), and the predictions (solid purple lines) were constructed using R. Model equations: (A) Y = 6.15718 + 0.81302X; (B) Y = 2.832681 + 0.831631X; (C) Y = 8.15439 + 0.82362X. For all pairs, $P < 2.2e{-}16$.

regardless of whether hetero- or homokaryotypic conditions (red and blue dots in **Figure 4**, **Figure S5** and **Dataset S3**). For example, the gene *CYP20A1* presented the highest percentage of allele imbalance between the heterokaryotypic co-twins in fibroblasts (T1DS = 81.25%/T2N = 33.33%) whereas in iPSC the gene *GCFC2*/*MRPL19* exhibited the highest discrepancy (T1DS = 66.66%/T2N = 18.75%).

Since A-to-I(G) RNA-editing of mRNAs can create stop codons (protein-truncating effect variants) or result in non-synonymous mutations, it was important to annotate the sites with discordant allelic proportions. In the heterokaryotypic co-twins, none of the annotated sites created stop codons, but nine sites are predicted to cause non-synonymous mutations (**Dataset S16**). The cyclin-dependent kinase 13 gene *CDK13*, presented two non-synonymous editing sites that change lysine (Lys; chr7_39950928) and glutamine (Gln; chr7_39950949) to arginine. Within the 5,372 eSNVs annotated as canonical RNA editing sites in the transcriptomes of the homokaryotypic twin pairs, none creates stop codons, and 21 sites correspond to non-synonymous mutations (**Dataset S16).**

MZ Co-Twins Are Discordant in the Allele-Specific Expression of X-Chromosome Inactivation Non-Escapee Genes

We scanned for discordant ASE sites in X-linked genes within the seven female twin pairs, and integrated data for the intersected sites about the XCI classification status from public repositories (Carrel and Willard, 2005; Cotton et al., 2013; Balaton et al., 2015; Cotton et al., 2015; Tukiainen et al., 2017; Garieri et al., 2018; Shvetsova et al., 2019; Wainer Katsir and Linial, 2019). The analysis was restricted to non-escapee genes because, in pooled cells, those genes must exhibit biallelic expression profiles. For this specific analysis, we only accepted gene products that displayed at least two discordant eSNVs (i.e., monoallelic in one twin *versus* biallelic in the sister twin). For the heterokaryotypic discordant co-twins, there was ASE disparity in the *UBL4A* gene products in fibroblasts, whereas, in iPSC, there was ASE disparity in the *FANCB* and *FTX* gene products (**Table S3**). Three control twin pairs expressed genes with at least two eSNVs: *TAB3*, *WDR44,* and *XIAP* genes in twin pair 05; *IDS*, *MAP7D3*, *RLIM*, *RPL10*, *SLC9A7*, *TBC1D25*, *TLR7*, *XIAP,* and *ZNF275* genes in twin pair 08; and the *ZNF275* gene in twin pair 09 (**Table S3**). None of the ASE disparities above are A-to-I(G) RNA editing sites in the RADAR database.

The Overall Impact of Allele-Specific Expression of Pathogenic Variants

We annotated 32 eSNVs associated with 131 human pathologies in the transcriptomes of the heterokaryotypic twins (**Dataset S17A**). Most pathogenic eSNVs are linked to autosomal recessive phenotypes and were coexpressed with the wild type allele, likely outbalancing the predicted deleterious effects. Four pathogenic alleles (rs1799990*G > A, rs1800562*G > A, rs200855215*A > G, and rs4784677*A > G) were expressed monoallelically, and are associated with Jakob-Creutzfeldt disease (OMIM #123400), hemochromatosis (OMIM #235200), Leber optic atrophy (OMIM #535000), and Bardet-Biedl syndrome 2 (OMIM #615981), respectively. One pathogenic allele (rs11583680*C > A), associated with autosomal dominant familial hypercholesterolemia (OMIM #603776), was also coexpressed with the wild type allele. In the homokaryotypic twin sets, 23 eSNVs, predicted to be pathogenic in the ClinVar database, predominantly coexpressed with the wild type alleles (**Dataset S17B**). For example, rs1799958*G > A, associated with deficiency of butyryl-coenzyme A dehydrogenase (OMIM #201470), was coexpressed with the wild-type allele in MZT3.

Evidence for Expressed Mitochondrial Microheteroplasmy

We also identified an ASE form of mitochondrial microheteroplasmy (Souren et al., 2016), albeit at lower limits, in all 10 MZ twin pairs, demonstrated by the presence of 237 eSNVs (median number of 25 eSNVs per dataset, **Table S4**). The observed limited number of mitochondrial eSNVs does not relate exclusively with the early embryonic age at sampling because the age of the control twin pairs ranged from 19- to 65-year-old (median age 26 years) (**Dataset S2**). Thus, for the set of donors investigated, we did not observe the accumulation of mitochondrial eSNVs with age (Smigrodzki and Khan, 2005).

Lastly, we queried ClinVar, PolyPhen, and SIFT public databases for evidence about the pathogenicity prediction for the mitochondrial eSNVs to assess the most functionally crucial mitochondrial point mutations. Fifteen eSNVs are predicted to be likely pathogenic in at least one database (**Table S4**). For example, rs28358569*A > G, monoallelically expressed in MZT9, is related to mitochondrial non-syndromic sensorineural hearing loss (OMIM #500008) and aminoglycoside-induced deafness (OMIM #580000); rs193302980*C > T and rs2853508*A > G are related to familial breast cancer (OMIM #114480).

Gene Ontology Analysis of Discordant Allele-Specific Expression Sites

In fibroblasts, the *CASP6* and *PDXDC1* genes, represented by ASE sites exhibiting biallelic to monoallelic switch (LogASE ≥ 0.8) within the heterokaryotypic co-twins were related with

nitrogen compound and organic substance metabolic processes (**Dataset S18A**). On the other hand, *IL17RA*, the only gene with LogASE ≤ −0.8 and mapping outside the well-characterized 21q recombination region, is enriched in immunological processes such as leukocyte migration, signal transduction, cytokine production, and cell activation (**Dataset S18B**). In iPSCs, most of the genes (72 of 100 genes) with discordant ASE sites (either bi-to-mono or mono-to-biallelic switches) are related to the regulation of biological process (**Datasets S9C** and **DatasetsS9D**).

DISCUSSION

We aimed to compile variant sites with expression profiles that are dissimilar between MZ co-twins who are discordant or not for a specific condition. Our scanning strategy permitted the identification, quantification, and classification of differential allelic expression by way of ASE discordant sites (i.e., eSNV) occurring genome-wide between co-twins who are either discordant or not for T21. Remarkably, the breadth and magnitude of ASE discordant sites were high and comparable between either heterokaryotypic or homokaryotypic co-twins. On average, we identified about 1,342 ASE discordant sites in the 10 pairs of MZ co-twins.

The extent of the ASE sites in T21 discordant co-twins was comparable between the non-discordant co-twins, assayed in three cell types (fibroblasts, iPSC, and B-cells). Overall, the analyses indicate that ASE discordance between MZ co-twins stems from aneuploidy, recombination, genomic imprinting, and RNA editing. We interpret the widespread occurrence of ASE discordance between MZ co-twins as being the result of sister chromatid-specific alterations in transcription. The discordant ASE sites observed between co-twins best reflect a combined effect of genetic and epigenetic processes on differential allele expression.

We note that e-karyotyping unveils dynamic arrays of ASE sites that can be considered as signatures that exhibit remarkable singularity to the individual biological sample. For example, for the heterokaryotypic co-twins, the sets of eSNVs observed in fibroblasts or iPSC do not overlap entirely. Overall, 38.67% (n = 24,103) of ASE sites were called in both fibroblasts and iPSC samples, 804 (3.3%) of which exhibited discordant allele-expression profiles in fibroblasts, but concordant in iPSC. Similarly, 1,318 sites (5.4%) were concordant in fibroblasts but discordant in iPSC. Moreover, 187 sites (0.7%) were discordant in both sample types. The relative lack of overlap among the experiments is likely explained by the differential expression of genes in these cell types. Therefore, e-karyotyping signatures might have forensic value and resolution power to discriminate clinically non-discordant co-twins. The e-karyotyping signatures may be specific to the level of each experimental condition for the same source of a biological sample. In principle, no sharing of e-karyotyping signatures is expected to occur within co-twins.

The observation of allelic bias is becoming commonplace in high-throughput transcriptome analyses (Deveale et al., 2012; Marinov et al., 2014; Metsalu et al., 2014; Wood et al., 2015; Weissbein et al., 2016). It is acceptable that the expression of most genes can be altered among biological sample replicas and that the total cellular RNA is not constant. Allelic bias in RNA-Seq can be, in part, attributed to the differential impact of the *in vitro* culture conditions (Gimelbrant et al., 2007; Weissbein et al., 2014). Thus, part of the ASE discordance observed between fibroblasts and iPSC in co-twins in our analysis may be due to acquired chromosomal abnormalities during the iPSC derivation and their propagation in culture.

Because the onset of MZ twinning, XCI, and genomic imprinting may occur at about the same time of embryological development (Machin, 1996), twining may affect the distribution of cells bearing the inactivated X-chromosome or abnormal epigenetic marks of imprinting, and therefore, the varying manifestation of allelic differences from these processes. Surprisingly, the effect primarily occurs in female co-twins rather than male co-twins, and, thus, it is likely due to the presence of more than one X-chromosome in females (Lubinsky and Hall, 1991; Matias et al., 2014). Furthermore, there are cases of MZ female co-twins discordant for skewed XCI and imprinted disorders (Orstavik et al., 1995) and non-imprinted diseases (Bennett et al., 2008). Interestingly, 1,050 ASE sites map to 205 known and candidate imprinted genes. ASE discordance, yet at a considerably lower extent than the described here, has been reported between at a pair of MZ "identical" co-twins clinically discordant for multiple sclerosis (Souren et al., 2016). Importantly, altered allelic expression of two imprinted genes (*ZNF331* and *GNAS*) and five non-imprinted genes (*ABLIM1*, *UBE2I*, *KIAA1267*, *CD6*, and *ATHL1*) were detected between the multiple sclerosis discordant co-twins.

Fourteen X-linked genes subjected to XCI (the non-escapee genes *UBL4A*, *FANCB*, *FTX, TAB3*, *WDR44*, *XIAP*, *IDS*, *MAP7D3*, *RLIM*, *RPL10*, *SLC9A7*, *TBC1D25*, *TLR7*, and *ZNF275*) in four female twin pairs exhibited ASE disparities in which one co-twin presented a biallelic profile and the sister female showed a monoallelic pattern. The RNA-Seq experiments were from pooled cells rather than single-cells and, thus, a biallelic model is the anticipated expression profile for non-escapee genes. The observed ASE disparities cannot be attributed to differences in cell culture conditions that result in decreased percentage of XCI mosaicism, which are expected to affect the expression profiles of all the 457 genes that are subjected to XCI (Carrel and Willard, 2005; Cotton et al., 2013; Balaton et al., 2015; Cotton et al., 2015; Tukiainen et al., 2017; Garieri et al., 2018; Shvetsova et al., 2019; Wainer Katsir and Linial, 2019).

What mechanisms might explain the observed ASE discordance in MZ co-twins? Data from prior RNA-Seq studies in co-twins (Baranzini et al., 2010; Lin et al., 2012a; Brown et al., 2014; Hibaoui et al., 2014; Buil et al., 2015; Dixon et al., 2015; Ding et al., 2017; Santoni et al., 2017) indicate that the differential allele expression of autosomal genes best reflects dynamic regulation processes consistent with either an allele being preferentially silenced or an inactive allele being restored. The biallelic expression of genes is a regulatory mechanism that outbalances the harmful effects of pathogenic expression-altering or loss-of-function risk variant alleles (Adzhubei et al., 2010; Landrum et al., 2016; Vaser et al., 2016). In each human euploid somatic cell, autosomal genes are anticipated to be

symmetrically expressed from both the parental alleles in a cell type-specific manner throughout development. However, the biallelic RNA expression pattern is not a phenotypic hallmark of all genes since 10–30% of human autosomal genes assayed for polymorphic variant sites [i.e., expressed SNPs (eSNPs) or eSNVs] are dynamically subjected to the epigenetic phenomena of clonal (mitotically) stable, random monoallelic expression (Gimelbrant et al., 2007; Savova et al., 2016a; Savova et al., 2016b; Savova et al., 2017), or allelic bias (Dixon et al., 2015). Most enigmatic, genes that are biallelically expressed in a cell can be regulated in a neighboring cell to randomly switch their RNA expression from biallelic to monoallelic at a time (Chess, 2013; Eckersley-Maslin and Spector, 2014; Eckersley-Maslin et al., 2014). Also, distinct subsets of autosomal and X-linked genes are subjected to epigenetic silencing of one allele, in a parent-of-origin dependent manner by autosomal genomic imprinting (Baran et al., 2015) or in a random fashion by XCI (i.e., in females) (Tukiainen et al., 2017).

ASE discordance in X-linked genes that are subjected to XCI has been reported between MZ female co-twins in humans (Cheung et al., 2008; Antonarakis et al., 2018) and in mice (Wang et al., 2010). Subtle departure from equal allelic expression ratios is often genetically determined in cis (i.e., eQTLs) and trans, but part of the disparity can also be ascribed to the random sampling effect of X inactivation (Carrel and Willard, 2005; Cheung et al., 2008; Cotton et al., 2013; Balaton et al., 2015; Cotton et al., 2015; Tukiainen et al., 2017; Garieri et al., 2018; Shvetsova et al., 2019; Wainer Katsir and Linial, 2019). Moreover, allelic imbalance on the X-chromosome could also affect autosomal allelic expression. Notwithstanding, we note that the extent of ASE discordance in homokaryotypic male twin-pairs (on average, 3.2%, n = 1,478 discordant sites) is comparable genome-wide to that in female twin pairs (2.5%, n = 958 discordant sites).

There are genetic and functional consequences of the autosomal variant sites in genes that are expressed from a single allele in one cell at a time. Mainly, i) they bestow more extensive genetic diversity in humans (Savova et al., 2016a); ii) they often are gain-of-function rather than pathogenic expression-altering or loss-of-function risk variants (i.e., for neurodevelopmental disorders), and influence expression variance in cis; iii) the range of expression level of monoallelically expressed genes is higher than biallelically expressed genes (Savova et al., 2017); iv) ultimately, they increase cell-to-cell expression variability with a beneficial impact of avoiding genetic disease phenotypes (Savova et al., 2016a).

The extranumerary chromosome 21 in trisomic cells of Down syndrome patients is well known to result in genome-wide dysregulation of gene expression represented by chromosomal domains with genes whose expression levels are copy-dosage compensated, upregulated, or downregulated as compared with euploid cells (Letourneau et al., 2014). In the co-twins discordant for T21 and 21q recombination, the ASE discrepant profiles can be viewed ultimately as unexplained heritability or missing heritability due to the discordance in trisomy 21, recombination at 21q, altered genomic imprinting, random monoallelic expression, and RNA editing. Allelic imbalance (allelic-specific heterogeneity) in dosage-sensitive genes can arise by an stochastic adaptive regulation in both euploid and aneuploid cells as a consequence of low-level mRNA abundance and increased transcriptional burst frequency, rather than burst size (Deng and Disteche, 2019; Larsson et al., 2019; Symmons et al., 2019). The extent of the biallelic imbalance across eSNVs most likely reflects a gene network expression effect operating in the form of eQTLs (Mott et al., 2014; Pettigrew et al., 2016). Thus, part of the unexplained heritability or missing heritability could be explained by differences in cell-specific gene interactions. Moreover, the ASE profile discrepancies between fibroblasts and iPSC could be due to non-imprinted parental origin effects in each cell type associated with the aneuploidy. For example, the rS93366794 eSNP exhibited a concordant biallelic profile in iPSC but discordant in fibroblasts, being monoallelic in the T21 twin and biallelic in the normal co-twin. Interestingly, in an RNA-Seq study of a healthy brain, the *WRD4* gene bearing the rS93366794 site was reported to be expressed monoallelically from the paternal allele, but a mono-to-biallelic switch occurred in the offspring with *versus* without ASD (Lin et al., 2018).

Although the analysis presented here allowed the identification of a pervasive disparity in ASE profiles between co-twins, the biological significance of the extent and breadth of the observed differences in allele expression must only be assessed by independent experiments. However, the following results are of worth noting: i) among the eSNVs, there were several alleles known to be associated with disease conditions or predicted to be pathogenic; ii) the canonical imprinted gene *SNURF*, which is expressed monoallelically in over 50 tissues in the GTEx dataset, was expressed biallelically in iPCS; iii) in all the 10 twin sets, there was expressed mitochondrial microheteroplasmy; iv) among all the genes expressed in the 10 twin pairs, there were 55 ± 17 genes that exhibited elevated proportions (ranging from 50 to 100%) of ASE sites coincident with RNA editing sites.

The breadth and magnitude of ASE discordance disclose unprecedented epigenomic-wide inter-individual variation occurring in MZ co-twins. Prior ASE studies in MZ twins are restricted to a specific gene or gene sets and, therefore, do not uncover the apparent state of pervasive missing allelic heritability in MZ co-twins shown in the present study. Although independent validation through wet experimentation (i.e., allele-specific quantitative reverse transcription-PCR, RNA-fluorescence *in situ* hybridization, or allele-specific pyrosequencing) is required for the biologically relevant candidate ASE discordant sites (here regarded as potential leads), two critical implications emerge from the epigenomic-wide inter-individual variation observed in MZ co-twins: i) as in the case of inter-individual variation in DNA methylation (Maunakea et al., 2010; Bell et al., 2012; Young et al., 2017; Garg et al., 2018), ASE discordance may have to be looked at when assessing and calculating the impact of phenotypic variation in the differential susceptibility to specific human conditions and diseases (Skipper, 2008; Sun et al., 2013); ii) ASE discordance also might be considered instrumental for developing RNA biomarker signatures for forensic body fluid identification and kinship analysis (Blay et al., 2019).

The present systematic and integrative meta-analysis has three important limitations: sample size, the certainty of correct calling a positive eSNV site for a theoretical heterozygote position,

and comparisons made in three different cell types. First, the experimental setting must be viewed as a case-study regarding a heterokaryotypic MZ twin pair discordant for trisomic 21 and chromosome 21 distal recombination. Reported cases of heterokaryotypic MZ twin pairs are rare. In **Table S1**, we listed all relevant cases studied in the literature. Notwithstanding, there is only one RNA-Seq public (i.e., not controlled) study in a pair of heterokaryotypic MZ twins, namely, the selected discordant index case. We used the index case as a reference case to investigate whether the underlying discordance in karyotype and recombination affect the ASE profiles. We initially hypothesized that any likely discordance in ASE differences will be restricted to chromosome 21 and that the differences will be more significant across and beyond the recombination event. However, the initial analysis indicated the occurrence of genome-wide rather that chr21-restricted ASE discordance. To investigate whether the observed genome-wide ASE discordance was limited to the unique index case, we investigated nine pairs of homokaryotypic MZ twins. Surprisingly, we observed genome-wide discordance in ASE, similar in breadth and magnitude to that observed in the index case, albeit in different cell lines. Second, to decrease the chances of false-positive ASE calls, we called eSNV sites using base quality control Q30 and ≥12 read depth, which are selecting criteria that excel in stringency published reports of the kind (Q20 and ≥8 reads) (Cheung et al., 2008; Baran et al., 2015; Tan et al., 2017; Tukiainen et al., 2017). Because the probability of correct SNV calling increases at higher coverage levels for a theoretical heterozygote position, we provided results for three read coverages (12, 20, and 40 reads). Coverage of 40 reads provides a 99,9% probability of correct SNV call (**Datasets S3-S14**). Third, the observation of genome-wide ASE discordance in the same type of cell lines (nine MZ twin pairs) and different cell lines (index case) is a very reassuring remark. Cheung et al. (2008) used 100K Affymetrix SNP array on the same set of homokaryotypic MZ twin biosamples and identified 201 SNPs with significant evidence of differential allelic expression. Of those, we confirmed 137 eSNVs as discordant, 38 sites of which were common to the nine twin-pairs (**Datasets S4–S14**). Unfortunately, no public next-generation sequencing data are available for DNA-Seq and RNA-Seq matched biosamples from MZ twins. Thus, we cannot address at present the question of how much of the genetic variation does contribute to the total percentage of ASE in MZ twins.

CONCLUDING REMARKS

Our genome-wide scans for allelic expression discordance reveal an apparent state of pervasive missing allelic heritability in MZ co-twins. The extent and breadth of the ASE discordant sites are not exclusively associated with differences due to chromosomal aberrations and recombination, but also relate to the epigenome-wide differential allele expression phenomena of genomic imprinting and RNA editing. We conclude that most of ASE discordant sites observed within MZ pairs (either homokaryotypic or heterokaryotypic co-twins) cannot be attributed solely to the estimated within-pair incongruencies in DNA (Huang et al., 2019) or correspond to random transcriptional allelic noise varying across experiments (Chakravarti, 2011). The epigenome-wide ASE discordance may have essential effects on physiology, phenotype, or inheritance, and implications for the Developmental Origins of Health and Disease (DOHaD) approach in co-twins (Yamada and Chong, 2017).

WEB RESOURCES

The URLs for public data used herein are as follows:

UCSC Genome Browser, https://genome.ucsc.edu/
GTEx portal, https://www.gtexportal.org/
NCBI SRA, https://www.ncbi.nlm.nih.gov/sra/
NCBI GEO, http://www.ncbi.nlm.nih.gov/geo/
dbGaP, http://www.ncbi.nlm.nih.gov/gap
PhenoScanner, http://www.phenoscanner.medschl.cam.ac.uk/phenoscanner
e-GRASP, http://www.mypeg.info/egrasp
HaploReg, http://compbio.mit.edu/HaploReg
PheGenI, https://www.ncbi.nlm.nih.gov/gap/phegeni
Geneimprint, http://www.geneimprint.com/
Metaimprint, http://bioinfo.hrbmu.edu.cn/MetaImprint/
dbMAE, https://mae.hms.harvard.edu/
OMIM, http://www.omim.org
RADAR, http://rnaedit.com/
Otago´s Catalogue of Imprinted Genes, http://igc.otago.ac.nz/
R software package, http://www.R-project.org
GATK, https://software.broadinstitute.org/gatk/

DATA AVAILABILITY STATEMENT

RNA-Seq data experiments used in this manuscript are from publicly available BioProjects. The primary data can be accessed from the NCBI SRA repository (https://www.ncbi.nlm.nih.gov/sra) under the following run entries: SRR1182244, SRR1182246, SRR1182249, SRR1182248, SRR1182252, SRR1182253, SRR1182245, SRR1182247, SRR1182250, SRR1182251, SRR1182254, SRR1182255, SRR1028343, SRR1028344, SRR1028345, SRR1028346, SRR1028347, SRR1028348, SRR1028349, SRR519874, SRR519875, SRR519876, SRR519877, SRR519878, SRR519879, SRR519880, SRR519881, SRR519882, SRR519883, SRR519884, SRR519885, SRR519886, SRR519887, SRR519888, SRR519889, SRR519890, SRR519891, SRR3390461, SRR3390473, SRR3389246, SRR3390437.

AUTHOR CONTRIBUTIONS

RJ, CF, JS, DM, EM-A: Conceived and designed experiments. RJ, CF, JS, YC, VR, GC, EM-A: Develop scripts and carried computational analyses. JS, DM, AG: performed SRA RNA-Seq analysis for control SNPs queries. RJ, CF, JS, DM, AG, EM-A: Performed cross-reference of ASE sites. DM, RJ, CF, JS, EM-A: Prepared figures. EM-A: Supervised the work and drafted the manuscript. All authors provided substantial

contributions to the interpretation of data,revised and approved the manuscript.

FUNDING

This study was supported by grants from the Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro - FAPERJ (BR) (http://www.faperj.br/) [grant number E26/010.001036/2015 to EM-A] and from the Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq (BR) (http://cnpq.br/) [grant number 308780/2015-9 to EM-A]. JS and DM received undergraduate PIBIC/CNPq fellowships from the Universidade Estadual do Norte Fluminense Darcy Ribeiro UENF (BR) (http://www.uenf.br/). CF is a recipient of a graduate fellowship from the Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES (http://www.capes.gov.br/). The agencies had no role in the study design, data collection, and analysis, decision to publish, or preparation of the manuscript.

ACKNOWLEDGMENTS

The authors are thankful to the high-performance platform of the MultiUser Equipment (EMU) of the Center for Genomic Medicine at the Clinics Hospital of Ribeirão Preto (FMRP-USP) for the support, particularly to the Drs. Wilson Araújo da Silva Junior, Raul Torrieri, and Marcelo Gomes. Special thanks to Dr. Thiago Motta Venancio from the Universidade Estadual do Norte Fluminense for the support with access to his computer processing facility. The authors are grateful to the members of the Molecular Identification and Diagnostics Unit of the Laboratory of Biotechnology, Universidade Estadual do Norte Fluminense, for their insights and productive brainstorm lab meetings.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01178/full#supplementary-material

DATASET S1 | RNA-Seq read counts and allele ratios supporting **Figure 1**.

DATASET S2 | BioProjects and RNA-Seq experiments used in this study.

DATASET S3 | ASE concordant and discordant sites identified in heterokaryotypic MZ twins assayed in fibroblasts and induced pluripotent stem cells (iPSC) derived from the co-twins discordant for trisomy 21 and maternal recombination at 21q.

DATASET S4 | Summary of the average distribution of eSNVs per gene in the entire set of biosamples and heterogeneity scores across genes. **(A)** Summary of informative of eSNVs, **(B)** Summary of eSNVs by ASE profile. Ranking of genes by ASE heterogeneity/homogeneity scores in either heterokaryotypic **(C)** and homokaryotypic **(D)** twins.

DATASET S5 | Concordant and discordant ASE sites identified in unrelated, non-twin males **(A)** and female **(B)** pairs assayed in peripheral mononuclear cells. Ranking of genes by ASE heterogeneity/homogeneity scores in unrelated, non-twin males and females **(C)**.

DATASETS S6–S14 | Concordant and discordant ASE sites identified in homokaryotypic MZ twins assayed in cultured B-cells derived from nine pairs of co-twins not discordant for a specific health condition.

DATASET S15 | Cross-reference of ASE sites from the ten monozygotic twin pairs with allele expression profiles from secondary data of the GTEx project in multiple tissues.

DATASET S16 | Summary of A-to-I(G) RNA-editing sites resulting in synonymous and non-synonymous substitutions in the ten twin pairs set.

DATASET S17 | Protein-coding genetic eSNVs associated with disease risk or disease pathogenesis in the ten twin pairs set.

DATASET S18 | Gene ontology enrichment annotations.

TABLE S1 | Review of cases of heterokaryotypic co-twins reported in the literature.

TABLE S2 | Distribution of ASE sites by gene structure sequence context. Cross-reference of ASE sites disparities observed in the present study with the annotation features of the eSNV disparities found in the whole-genome sequencing study by Huang et al. 2019.

TABLE S3 | ASE disparities in X-linked genes subject to XCI between MZ twins.

TABLE S4 | Evidence for allele-specific expression form of mitochondrial microheteroplasmy in MZ twin pairs.

FIGURE S1 | Representation of the discordant maternal 21q inheritance in the pair of co-twins heterokaryotypic for trisomy 21 reported by Dahoun et al. (2008). The MZ co-twins are discordant for trisomy 21 of maternal origin in twin 1 (T1DS) and maternal allelic disparity at 21qter likely carried by disomic twin 2 (T2N). The discordant inheritance of 21q is probably due to meiosis I subtelomeric recombination event likely occurring between the maternal chromosomes 21 within the 1.7Mb interval (hg38) delimited by the short tandem repeat marker D21S1445, where alleles were identical in both twins and the short tandem repeat marker D21S1611, where different alleles were inherited.

FIGURE S2 | Flowchart of analysis. The in-house computational pipeline, PipASE, used for scanning and sorting out genome-wide, allele-specific differences between MZ co-twins.

FIGURE S3 | Chromosomal distribution of eSNVs in iPSC. **(A)** Genome-wide e-karyotyping for the SNVs exhibiting allele-specific expression in iPSC derived from the co-twins discordant for T21 and maternal recombination at 21q. Shown is the distribution of all ASE sites that were concordant (gray ticks towards the left of each chromosome ideogram) or discordant (blue ticks towards the right side). **(B)** Assessment of chromosomal aberrations by e-karyotyping allelic bias using RNA-Seq data from iPSC in **(A)**.

FIGURE S4 | Overview of the breadth and magnitude of allele-specific expression disparity between nine MZ control twin pairs. The RNA-Seq SRA entries for the nine twin pairs are SRR519874, SRR519875, SRR519876, SRR519877, SRR519878, SRR519879, SRR519880, SRR519881, SRR519882, SRR519883, SRR519884, SRR519885, SRR519886, SRR519887, SRR519888, SRR519889, SRR519890, and SRR519891. For each SRA entry above, the panels are **(A)** numbers of ASE sites distributed by the within-pair status of concordance or discordance in control homokaryotypic MZ control twins tested in cultured B-cells. The majority of ASE sites are concordant for a biallelic imbalance status. On average, the co-twins are discordant in 1074 ± 252.03 ASE sites. **(B)** Comparison of the effect size of LogASE. We calculated the log2 of allele-specific expression fold change using the equation LogASE = log2(T1_ASE / T2_ASE) for each eSNV in each tissue. LogASE estimates the magnitude of expression change between conditions for the variant. **(C)** Distribution of genes by numbers of ASE sites observed in cultured B-cells.

FIGURE S5 | Chromosomal distribution of eSNVs in the homokaryotypic twin pairs. Genome-wide e-karyotyping for the SNVs exhibiting allele-specific expression in cultured B-cells from nine control twin pairs. Shown in each panel, **(A)** through **(I)**, is the distribution of all ASE sites that were concordant (gray ticks towards the left

of each chromosome ideogram) or discordant (red ticks towards the right side). The RNA-Seq SRA entries for the nine twin pairs in are SRR519874, SRR519875, SRR519876, SRR519877, SRR519878, SRR519879, SRR519880, SRR519881, SRR519882, SRR519883, SRR519884, SRR519885, SRR519886, SRR519887, SRR519888, SRR519889, SRR519890, and SRR519891, respectively.

FIGURE S6 | Assessment of chromosomal aberrations by e-karyotyping allelic bias using RNA-Seq data from control twin pairs. (A) through (I). The RNA-Seq SRA entries for the nine twin pairs used as controls are SRR519874, SRR519875, SRR519876, SRR519877, SRR519878, SRR519879, SRR519880, SRR519881, SRR519882, SRR519883, SRR519884, SRR519885, SRR519886, SRR519887, SRR519888, SRR519889, SRR519890, and SRR519891, respectively. For each SRA entry above, a plot is shown, which represents the distribution of allele ratios in cultured B-cells from nine pairs of co-twins who are not discordant for a specific health condition. None of the control twin pairs present detectable chromosomal aberrations.

FIGURE S7 | Within twin-pair disparities in allele expression proportions at eSNVs that are coincident with canonical A-to-I(G) RNA editing sites in homokaryotypic twins. Shown is the distribution of eSNVs that positionally match canonical RNA editing sites between nine homokaryotypic co-twins, **(A)** through **(I)**, assayed in culture-B-cells. Each dot corresponds to an eSNV. The vast majority of sites exhibited a concordant biallelic imbalance profile (pink and light blue dots). Red dots represent eSNPs that were discordant between co-twins in that they exhibited allelic proportions differences higher than 25%, regardless of the discordance or concordance in the karyotype. Green dots represent eSNVs that exhibited discordant allelic profiles, being biallelic in one twin and monoallelic in the other. The linear models (solid black lines), the confidence interval of the models (broken purple lines) and of the prediction (solid purple lines) were constructed using R. Model equations: **(A)** Y = 8.15439 + 0.82362X; **(B)** Y = 7.20067 + 0.80502X; **(C)** Y = 6.0758 + 0.8080X; **(D)** Y = 12.74607 + 0.78267X; **(E)** Y = 14.78510 + 0.79019X; **(F)** Y = 5.64298 + 0.81255X; **(G)** Y = 7.97718 + 0.82152X; **(H)** Y = 8.36441 + 0.81231X; **(I)** Y = 6.60496 + 0.72912X. For all pairs, P < 2.2e-16. The RNA-Seq SRA entries for the nine twin pairs used as controls are SRR519874, SRR519875, SRR519876, SRR519877, SRR519878, SRR519879, SRR519880, SRR519881, SRR519882, SRR519883, SRR519884, SRR519885, SRR519886, SRR519887, SRR519888, SRR519889, SRR519890, and SRR519891.

REFERENCES

Abdellaoui, A., Ehli, E. A., Hottenga, J. J., Weber, Z., Mbarek, H., Willemsen, G., et al. (2015). CNV Concordance in 1,097 MZ Twin Pairs. *Twin Res. Hum. Genet.* 18, 1–12. doi: 10.1017/thg.2014.86

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., et al. (2010). A method and server for predicting damaging missense mutations. *Nat. Methods* 7, 248–249. doi: 10.1038/nmeth0410-248

Alves Da Silva, A. F., Machado, F. B., Pavarino, E. C., Biselli-Perico, J. M., Zampieri, B. L., Da Silva Francisco Junior, R., et al. (2016). Trisomy 21 Alters DNA Methylation in Parent-of-Origin-Dependent and -Independent Manners. *PloS One* 11, e0154108. doi: 10.1371/journal.pone.0154108

Antonarakis, M. G., Georgios, S., Xavier, B., Emilie, F., Pascale, R., Christelle, B., et al. (2018). Extensive cellular heterogeneity of X inactivation revealed by single-cell allele-specific expression in human fibroblasts. doi: 10.1073/pnas.1806811115

Babak, T., Deveale, B., Tsang, E. K., Zhou, Y., Li, X., Smith, K. S., et al. (2015). Genetic conflict reflected in tissue-specific maps of genomic imprinting in human and mouse. *Nat. Genet.* 47, 544–549. doi: 10.1038/ng.3274

Balaton, B. P., Cotton, A. M., and Brown, C. J. (2015). Derivation of consensus inactivation status for X-linked genes from genome-wide studies. *Biol. Sex Differ.* 6, 35. doi: 10.1186/s13293-015-0053-7

Baran, Y., Subramaniam, M., Biton, A., Tukiainen, T., Tsang, E. K., Rivas, M. A., et al. (2015). The landscape of genomic imprinting across diverse adult human tissues. *Genome Res.* 25, 927–936. doi: 10.1101/gr.192278.115

Baranzini, S. E., Mudge, J., Van Velkinburgh, J. C., Khankhanian, P., Khrebtukova, I., Miller, N. A., et al. (2010). Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis. *Nature* 464, 1351–1356. doi: 10.1038/nature08990

Begemann, M., Rezwan, F. I., Beygo, J., Docherty, L. E., Kolarova, J., Schroeder, C., et al. (2018). Maternal variants in NLRP and other maternal effect proteins are associated with multilocus imprinting disturbance in offspring. *J. Med. Genet.* 55, 497–504. doi: 10.1136/jmedgenet-2017-105190

Bell, J. T., Tsai, P. C., Yang, T. P., Pidsley, R., Nisbet, J., Glass, D., et al. (2012). Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. *PloS Genet.* 8, e1002629. doi: 10.1371/journal.pgen.1002629

Bennett, C. M., Boye, E., and Neufeld, E. J. (2008). Female monozygotic twins discordant for hemophilia A due to nonrandom X-chromosome inactivation. *Am. J. Hematol.* 83, 778–780. doi: 10.1002/ajh.21219

Blay, N., Casas, E., Galvan-Femenia, I., Graffelman, J., De Cid, R., and Vavouri, T. (2019). Assessment of kinship detection using RNA-seq data. *Nucleic Acids Res.* gkz776. doi: 10.1093/nar/gkz776 10.1101/546937.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Brown, A. A., Buil, A., Vinuela, A., Lappalainen, T., Zheng, H. F., Richards, J. B., et al. (2014). Genetic interactions affecting human gene expression identified by variance association mapping. *Elife* 3, e01381. doi: 10.7554/eLife.01381

Buil, A., Brown, A. A., Lappalainen, T., Vinuela, A., Davies, M. N., Zheng, H. F., et al. (2015). Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. *Nat. Genet.* 47, 88–91. doi: 10.1038/ng.3162

Carrel, L., and Willard, H. F. (2005). X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* 434, 400–404. doi: 10.1038/nature03479

Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E., and Lappalainen, T. (2015). Tools and best practices for data processing in allelic expression analysis. *Genome Biol.* 16, 195. doi: 10.1186/s13059-015-0762-6

Chaiyasap, P., Kulawonganunchai, S., Srichomthong, C., Tongsima, S., Suphapeetiporn, K., and Shotelersuk, V. (2014). Whole genome and exome sequencing of monozygotic twins with trisomy 21, discordant for a congenital heart defect and epilepsy. *PloS One* 9, e100191. doi: 10.1371/journal.pone.0100191

Chakravarti, A. (2011). Widespread promiscuous genetic information transfer from DNA to RNA. *Circ. Res.* 109, 1202–1203. doi: 10.1161/RES.0b013e31823c4992

Chamberlain, S., Ushey, K., and Zhu, H. (2018). rsnps: Get 'SNP' ('Single-Nucleotide' 'Polymorphism') Data on the Web [Online]. Available at URL: https://cran.r-project.org/web/packages/rsnps/.

Chess, A. (2013). Random and non-random monoallelic expression. *Neuropsychopharmacology* 38, 55–61. doi: 10.1038/npp.2012.85

Cheung, V. G., Bruzel, A., Burdick, J. T., Morley, M., Devlin, J. L., and Spielman, R. S. (2008). Monozygotic twins reveal germline contribution to allelic expression differences. *Am. J. Hum. Genet.* 82, 1357–1360. doi: 10.1016/j.ajhg.2008.05.003

Core Team, R. (2019). R: A language and environment for statistical computing [Online]. Available at URL: https://www.R-project.org/.

Cotton, A. M., Ge, B., Light, N., Adoue, V., Pastinen, T., and Brown, C. J. (2013). Analysis of expressed SNPs identifies variable extents of expression from the human inactive X chromosome. *Genome Biol.* 14, R122. doi: 10.1186/gb-2013-14-11-r122

Cotton, A. M., Price, E. M., Jones, M. J., Balaton, B. P., Kobor, M. S., and Brown, C. J. (2015). Landscape of DNA methylation on the X chromosome reflects CpG density, functional chromatin state and X-chromosome inactivation. *Hum. Mol. Genet.* 24, 1528–1539. doi: 10.1093/hmg/ddu564

Dahoun, S., Gagos, S., Gagnebin, M., Gehrig, C., Burgi, C., Simon, F., et al. (2008). Monozygotic twins discordant for trisomy 21 and maternal 21q inheritance: a complex series of events. *Am. J. Med. Genet. A* 146A, 2086–2093. doi: 10.1002/ajmg.a.32431

De Sa Machado Araujo, G., Da Silva Francisco Junior, R., Dos Santos Ferreira, C., Mozer Rodrigues, P. T., Terra Machado, D., Louvain De Souza, T., et al. (2018). Maternal 5(m)CpG Imprints at the PARD6G-AS1 and GCSAML Differentially Methylated Regions Are Decoupled From Parent-of-Origin Expression Effects in Multiple Human Tissues. *Front. Genet.* 9, 36. doi: 10.3389/fgene.2018.00036

Deng, X., and Disteche, C. M. (2019). Rapid transcriptional bursts upregulate the X chromosome. *Nat. Struct. Mol. Biol.* 26, 851–853. doi: 10.1038/s41594-019-0314-y

Depristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–498. doi: 10.1038/ng.806

Deveale, B., Van Der Kooy, D., and Babak, T. (2012). Critical evaluation of imprinted gene expression by RNA-Seq: a new perspective. *PloS Genet.* 8, e1002600. doi: 10.1371/journal.pgen.1002600

Ding, N., Zhang, Z., Yang, W., Ren, L., Zhang, Y., Zhang, J., et al. (2017). Transcriptome analysis of monozygotic twin brothers with childhood primary myelofibrosis. *Genomics Proteomics Bioinf.* 15, 37–48. doi: 10.1016/j.gpb.2016.12.002

Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., et al. (2015). Chromatin architecture reorganization during stem cell differentiation. *Nature* 518, 331–336. doi: 10.1038/nature14222

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi: 10.1093/bioinformatics/bts635

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., et al. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* 21, 3439–3440. doi: 10.1093/bioinformatics/bti525

Durinck, S., Spellman, P. T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* 4, 1184–1191. doi: 10.1038/nprot.2009.97

Eckersley-Maslin, M. A., and Spector, D. L. (2014). Random monoallelic expression: regulating gene expression one allele at a time. *Trends Genet.* 30, 237–244. doi: 10.1016/j.tig.2014.03.003

Eckersley-Maslin, M. A., Thybert, D., Bergmann, J. H., Marioni, J. C., Flicek, P., and Spector, D. L. (2014). Random monoallelic gene expression increases upon embryonic stem cell differentiation. *Dev. Cell* 28, 351–365. doi: 10.1016/j.devcel.2014.01.017

Essaoui, M., Nizon, M., Beaujard, M. P., Carrier, A., Tantau, J., De Blois, M. C., et al. (2013). Monozygotic twins discordant for 18q21.2qter deletion detected by array CGH in amniotic fluid. *Eur. J. Med. Genet.* 56, 502–505. doi: 10.1016/j.ejmg.2013.06.007

Furukawa, H., Oka, S., Matsui, T., Hashimoto, A., Arinuma, Y., Komiya, A., et al. (2013). Genome, epigenome and transcriptome analyses of a pair of monozygotic twins discordant for systemic lupus erythematosus. *Hum. Immunol.* 74, 170–175. doi: 10.1016/j.humimm.2012.11.007

Garg, P., Joshi, R. S., Watson, C., and Sharp, A. J. (2018). A survey of inter-individual variation in DNA methylation identifies environmentally responsive co-regulated networks of epigenetic variation in the human genome. *PloS Genet.* 14, e1007707. doi: 10.1371/journal.pgen.1007707

Garieri, M., Stamoulis, G., Blanc, X., Falconnet, E., Ribaux, P., Borel, C., et al. (2018). Extensive cellular heterogeneity of X inactivation revealed by single-cell allele-specific expression in human fibroblasts. *Proc. Natl. Acad. Sci. U. S. A.* 115, 13015–13020. doi: 10.1073/pnas.1806811115

Gilbert, B., Yardin, C., Briault, S., Belin, V., Lienhardt, A., Aubard, Y., et al. (2002). Prenatal diagnosis of female monozygotic twins discordant for Turner syndrome: implications for prenatal genetic counselling. *Prenat. Diagn.* 22, 697–702. doi: 10.1002/pd.383

Gimelbrant, A., Hutchinson, J. N., Thompson, B. R., and Chess, A. (2007). Widespread monoallelic expression on human autosomes. *Science* 318, 1136–1140. doi: 10.1126/science.1148910

Gray, T. A., Saitoh, S., and Nicholls, R. D. (1999). An imprinted, mammalian bicistronic transcript encodes two independent proteins. *Proc. Natl. Acad. Sci. U.S.A.* 96, 5616–5621. doi: 10.1073/pnas.96.10.5616

Heap, G. A., Yang, J. H., Downes, K., Healy, B. C., Hunt, K. A., Bockett, N., et al. (2010). Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. *Hum. Mol. Genet.* 19, 122–134. doi: 10.1093/hmg/ddp473

Hibaoui, Y., Grad, I., Letourneau, A., Sailani, M. R., Dahoun, S., Santoni, F. A., et al. (2014). Modelling and rescuing neurodevelopmental defect of Down syndrome using induced pluripotent stem cells from monozygotic twins discordant for trisomy 21. *EMBO Mol. Med.* 6, 259–277. doi: 10.1002/emmm.201302848

Hu, Y. J., Sun, W., Tzeng, J. Y., and Perou, C. M. (2015). Proper Use of Allele-Specific Expression Improves Statistical Power for cis-eQTL Mapping with RNA-Seq Data. *J. Am. Stat. Assoc.* 110, 962–974. doi: 10.1080/01621459.2015.1038449

Huang, Y., Zhao, Y., Ren, Y., Yi, Y., Li, X., Gao, Z., et al. (2019). Identifying genomic variations in monozygotic twins discordant for autism spectrum disorder using whole-genome sequencing. *Mol. Ther. Nucleic Acids* 14, 204–211. doi: 10.1016/j.omtn.2018.11.015

International Hapmap, C. (2003). The International HapMap Project. *Nature* 426, 789–796. doi: 10.1038/nature02168

Jirtle, J., and Murphy, C. (2012). Geneimprint database [Online]. Available: http://www.geneimprint.com [Accessed September 2017]. .

Knopman, J. M., Krey, L. C., Oh, C., Lee, J., Mccaffrey, C., and Noyes, N. (2014). What makes them split?Identifying risk factors that lead to monozygotic twins after in vitro fertilization. *Fertil. Steri.l* 102, 82–89. doi: 10.1016/j.fertnstert.2014.03.039

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., et al. (2016). ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* 44, D862–D868. doi: 10.1093/nar/gkv1222

Larsson, A. J. M., Coucoravas, C., Sandberg, R., and Reinius, B. (2019). X-chromosome upregulation is driven by increased burst frequency. *Nat. Struct. Mol. Biol.* 26, 963–969. doi: 10.1038/s41594-019-0306-y

Lawrence, M., Gentleman, R., and Carey, V. (2009). rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* 25, 1841–1842. doi: 10.1093/bioinformatics/btp328

Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., Gentleman, R., et al. (2013). Software for computing and annotating genomic ranges. *PloS Comput. Biol.* 9, e1003118. doi: 10.1371/journal.pcbi.1003118

Letourneau, A., Santoni, F. A., Bonilla, X., Sailani, M. R., Gonzalez, D., Kind, J., et al. (2014). Domains of genome-wide gene expression dysregulation in Down's syndrome. *Nature* 508, 345–350. doi: 10.1038/nature13200

Leung, W. C., Choi, H., Lau, W. L., Ng, L. K., Lau, E. T., Lo, F. M., et al. (2009). Monozygotic dichorionic twins heterokaryotypic for duplication chromosome 2q13-q23.3. *Fetal Diagn. Ther.* 25, 397–399. doi: 10.1159/000236153

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, M., Wang, I. X., Li, Y., Bruzel, A., Richards, A. L., Toung, J. M., et al. (2011). Widespread RNA and DNA sequence differences in the human transcriptome. *Science* 333, 53–58. doi: 10.1126/science.1207018

Lin, M., Hrabovsky, A., Pedrosa, E., Wang, T., Zheng, D., and Lachman, H. M. (2012a). Allele-biased expression in differentiating human neurons: implications for neuropsychiatric disorders. *PloS One* 7, e44017. doi: 10.1371/journal.pone.0044017

Lin, W., Piskol, R., Tan, M. H., and Li, J. B. (2012b). Comment on widespread RNA and DNA sequence differences in the human transcriptome. *Science* 335, 1302. doi: 10.1126/science.1210624 author reply 1302.

Lin, C. Y., Chang, K. W., Lin, C. Y., Wu, J. Y., Coon, H., Huang, P. H., et al. (2018). Allele-specific expression in a family quartet with autism reveals mono-to-biallelic switch and novel transcriptional processes of autism susceptibility genes. *Sci. Rep.* 8, 4277. doi: 10.1038/s41598-018-22753-4

Liu, Z., Yang, J., Xu, H., Li, C., Wang, Z., Li, Y., et al. (2014). Comparing computational methods for identification of allele-specific expression based on next generation sequencing data. *Genet. Epidemiol.* 38, 591–598. doi: 10.1002/gepi.21846

Liu, S., Hong, Y., Cui, K., Guan, J., Han, L., Chen, W., et al. (2018). Four-generation pedigree of monozygotic female twins reveals genetic factors in twinning process by whole-genome sequencing. *Twin Res. Hum. Genet.* 21, 361–368. doi: 10.1017/thg.2018.41

Lo, H. S., Wang, Z., Hu, Y., Yang, H. H., Gere, S., Buetow, K. H., et al. (2003). Allelic variation in gene expression is common in the human genome. *Genome Res.* 13, 1855–1862. doi: 10.1101/gr.1006603

Lubinsky, M. S., and Hall, J. G. (1991). Genomic imprinting, monozygous twinning, and X inactivation. *Lancet* 337, 1288. doi: 10.1016/0140-6736(91)92956-3

Machin, G. A. (1996). Some causes of genotypic and phenotypic discordance in monozygotic twin pairs. *Am. J. Med. Genet.* 61, 216–228. doi: 10.1002/(SICI)1096-8628(19960122)61:3<216::AID-AJMG5>3.0.CO;2-S

Marinov, G. K., Williams, B. A., Mccue, K., Schroth, G. P., Gertz, J., Myers, R. M., et al. (2014). From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome Res.* 24, 496–510. doi: 10.1101/gr.161034.113

Matias, A., Silva, S., Martins, Y., and Blickstein, I. (2014). Monozygotic twins: ten reasons to be different. *Diagnóstico Prenatal.* 25, 53–57. doi: 10.1016/j.diapre.2013.09.003

Maunakea, A. K., Chepelev, I., and Zhao, K. (2010). Epigenome mapping in normal and disease States. *Circ. Res.* 107, 327–339. doi: 10.1161/CIRCRESAHA.110.222463

Mckenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

Metsalu, T., Viltrop, T., Tiirats, A., Rajashekar, B., Reimann, E., Koks, S., et al. (2014). Using RNA sequencing for identifying gene imprinting and random monoallelic expression in human placenta. *Epigenetics* 9, 1397–1409. doi: 10.4161/15592294.2014.970052

Morgan, M. (2017). AnnotationHub: Client to access AnnotationHub resources. R package version 2.14.5 [Online]. Available at URL: https://bioconductor.org/packages/release/bioc/html/AnnotationHub.html

Morley, M., Molony, C. M., Weber, T. M., Devlin, J. L., Ewens, K. G., Spielman, R. S., et al. (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature* 430, 743–747. doi: 10.1038/nature02797

Mott, R., Yuan, W., Kaisaki, P., Gan, X., Cleak, J., Edwards, A., et al. (2014). The architecture of parent-of-origin effects in mice. *Cell* 156, 332–342. doi: 10.1016/j.cell.2013.11.043

Moyerbrailean, G. A., Richards, A. L., Kurtz, D., Kalita, C. A., Davis, G. O., Harvey, C. T., et al. (2016). High-throughput allele-specific expression across 250 environmental conditions. *Genome Res.* 26, 1627–1638. doi: 10.1101/gr.209759.116

Nieuwint, A., Van Zalen-Sprock, R., Hummel, P., Pals, G., Van Vugt, J., Van Der Harten, H., et al. (1999). 'Identical' twins with discordant karyotypes. *Prenat. Diagn.* 19, 72–76. doi: 10.1002/(SICI)1097-0223(199901)19:1<72::AID-PD465>3.0.CO;2-V

Orstavik, R. E., Tommerup, N., Eiklid, K., and Orstavik, K. H. (1995). Non-random X chromosome inactivation in an affected twin in a monozygotic twin pair discordant for Wiedemann-Beckwith syndrome. *Am. J. Med. Genet.* 56, 210–214. doi: 10.1002/ajmg.1320560219

Pettigrew, K. A., Frinton, E., Nudel, R., Chan, M. T. M., Thompson, P., Hayiou-Thomas, M. E., et al. (2016). Further evidence for a parent-of-origin effect at the NOP9 locus on language-related phenotypes. *J. Neurodev. Disord.* 8, 24. doi: 10.1186/s11689-016-9157-6

Pickrell, J. K., Gilad, Y., and Pritchard, J. K. (2012). Comment on widespread RNA and DNA sequence differences in the human transcriptome. *Science* 335, 1302. doi: 10.1126/science.1210484 author reply 1302.

Pirinen, M., Lappalainen, T., Zaitlen, N. A., Consortium, G. T., Dermitzakis, E. T., Donnelly, P., et al. (2015). Assessing allele-specific expression across multiple tissues from RNA-seq read data. *Bioinformatics* 31, 2497–2504. doi: 10.1093/bioinformatics/btv074

Piskol, R., Peng, Z., Wang, J., and Li, J. B. (2013). Lack of evidence for existence of noncanonical RNA editing. *Nat. Biotechnol.* 31, 19–20. doi: 10.1038/nbt.2472

Raghupathy, N., Choi, K., Vincent, M. J., Beane, G. L., Sheppard, K., Munger, S. C., et al. (2018). Hierarchical analysis of rNA-seq reads improves the accuracy of allele-specific expression. *Bioinformatics* 34, 2177–2184. doi: 10.1093/bioinformatics/bty078

Ramaswami, G., and Li, J. B. (2014). RADAR: a rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Res.* 42, D109–D113. doi: 10.1093/nar/gkt996

Ramaswami, G., Lin, W., Piskol, R., Tan, M. H., Davis, C., and Li, J. B. (2012). Accurate identification of human Alu and non-Alu RNA editing sites. *Nat. Methods* 9, 579–581. doi: 10.1038/nmeth.1982

Richard Albert, J., Koike, T., Younesy, H., Thompson, R., Bogutz, A. B., Karimi, M. M., et al. (2018). Development and application of an integrated allele-specific pipeline for methylomic and epigenomic analysis (MEA). *BMC Genomics* 19, 463. doi: 10.1186/s12864-018-4835-2

Santoni, F. A., Stamoulis, G., Garieri, M., Falconnet, E., Ribaux, P., Borel, C., et al. (2017). Detection of Imprinted Genes by Single-Cell Allele-Specific Gene Expression. *Am. J. Hum. Genet.* 100, 444–453. doi: 10.1016/j.ajhg.2017.01.028

Savova, V., Chun, S., Sohail, M., Mccole, R. B., Witwicki, R., Gai, L., et al. (2016a). Genes with monoallelic expression contribute disproportionately to genetic diversity in humans. *Nat. Genet.* 48, 231–237. doi: 10.1038/ng.3493

Savova, V., Patsenker, J., Vigneau, S., and Gimelbrant, A. A. (2016b). dbMAE: the database of autosomal monoallelic expression. *Nucleic Acids Res.* 44, D753–D756. doi: 10.1093/nar/gkv1106

Savova, V., Vinogradova, S., Pruss, D., Gimelbrant, A. A., and Weiss, L. A. (2017). Risk alleles of genes with monoallelic expression are enriched in gain-of-function variants and depleted in loss-of-function variants for neurodevelopmental disorders. *Mol. Psychiatry* 22, 1785–1794. doi: 10.1038/mp.2017.13

Scott, J. M., and Ferguson-Smith, M. A. (1973). Heterokaryotypic monozygotic twins and the acardiac monster. *J. Obstet. Gynaecol. Br. Commonw.* 80, 52–59. doi: 10.1111/j.1471-0528.1973.tb02131.x

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., et al. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29, 308–311. doi: 10.1093/nar/29.1.308

Shvetsova, E., Sofronova, A., Monajemi, R., Gagalova, K., Draisma, H. H. M., White, S. J., et al. (2019). Skewed X-inactivation is common in the general female population. *Eur. J. Hum. Genet.* 27, 455–465. doi: 10.1038/s41431-018-0291-3

Skipper, M. (2008). Gene expression - One allele or two?. *Nat. Rev. Genet.* 9, 4–5. doi: 10.1038/nrg2287

Smigrodzki, R. M., and Khan, S. M. (2005). Mitochondrial microheteroplasmy and a theory of aging and age-related disease. *Rejuvenation. Res.* 8, 172–198. doi: 10.1089/rej.2005.8.172

Soderlund, C. A., Nelson, W. M., and Goff, S. A. (2014). Allele Workbench: transcriptome pipeline and interactive graphics for allele-specific expression. *PloS One* 9, e115740. doi: 10.1371/journal.pone.0115740

Souren, N. Y., Gerdes, L. A., Kumpfel, T., Lutsik, P., Klopstock, T., Hohlfeld, R., et al. (2016). Mitochondrial DNA Variation and Heteroplasmy in Monozygotic Twins Clinically Discordant for Multiple Sclerosis. *Hum. Mutat.* 37, 765–775. doi: 10.1002/humu.23003

Sun, C., Burgner, D. R., Ponsonby, A. L., Saffery, R., Huang, R. C., Vuillemin, P. J., et al. (2013). Effects of early-life environment and epigenetics on cardiovascular disease risk in children: highlighting the role of twin studies. *Pediatr. Res.* 73, 523–530. doi: 10.1038/pr.2013.6

Symmons, O., Chang, M., Mellis, I. A., Kalish, J. M., Park, J., Susztak, K., et al. (2019). Allele-specific RNA imaging shows that allelic imbalances can arise in tissues through transcriptional bursting. *PloS Genet.* 15, e1007874. doi: 10.1371/journal.pgen.1007874

Tachon, G., Lefort, G., Puechberty, J., Schneider, A., Jeandel, C., Boulot, P., et al. (2014). Discordant sex in monozygotic XXY/XX twins: a case report. *Hum. Reprod.* 29, 2814–2820. doi: 10.1093/humrep/deu275

Tan, M. H., Li, Q., Shanmugam, R., Piskol, R., Kohler, J., Young, A. N., et al. (2017). Dynamic landscape and regulation of RNA editing in mammals. *Nature* 550, 249–254. doi: 10.1038/nature24041

The GTEX Project, C. (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 348, 648–660. doi: 10.1126/science.1262110

Tukiainen, T., Villani, A. C., Yen, A., Rivas, M. A., Marshall, J. L., Satija, R., et al. (2017). Landscape of X chromosome inactivation across human tissues. *Nature* 550, 244–248. doi: 10.1038/nature24265

Van Der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., et al. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinf.* 43, 11 10 11–33. doi: 10.1002/0471250953.bi1110s43

Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M., and Ng, P. C. (2016). SIFT missense predictions for genomes. *Nat. Protoc.* 11, 1–9. doi: 10.1038/nprot.2015.123

Vigneau, S., Vinogradova, S., Savova, V., and Gimelbrant, A. (2018). High prevalence of clonal monoallelic expression. *Nat. Genet.* 50, 1198–1199. doi: 10.1038/s41588-018-0188-7

Von Hippel, P. T. (2015). The heterogeneity statistic I(2) can be biased in small meta-analyses. *BMC Med. Res. Methodol.* 15, 35. doi: 10.1186/s12874-015-0024-z

Wainer Katsir, K., and Linial, M. (2019). Human genes escaping X-inactivation revealed by single cell expression data. *BMC Genomics* 20, 201. doi: 10.1186/s12864-019-5507-6

Wang, X., and Clark, A. G. (2014). Using next-generation RNA sequencing to identify imprinted genes. *Heredity (Edinb)* 113, 156–166. doi: 10.1038/hdy.2014.18

Wang, X., Soloway, P. D., and Clark, A. G. (2010). Paternally biased X inactivation in mouse neonatal brain. *Genome Biol.* 11, R79. doi: 10.1186/gb-2010-11-7-r79

Wei, Y., Su, J., Liu, H., Lv, J., Wang, F., Yan, H., et al. (2014). MetaImprint: an information repository of mammalian imprinted genes. *Development* 141, 2516–2523. doi: 10.1242/dev.105320

Weissbein, U., Benvenisty, N., and Ben-David, U. (2014). Quality control: genome maintenance in pluripotent stem cells. *J. Cell Biol.* 204, 153–163. doi: 10.1083/jcb.201310135

Weissbein, U., Schachter, M., Egli, D., and Benvenisty, N. (2016). Analysis of chromosomal aberrations and recombination by allelic bias in RNA-Seq. *Nat. Commun.* 7, 12144. doi: 10.1038/ncomms12144

Weksberg, R., Shuman, C., Caluseriu, O., Smith, A. C., Fei, Y. L., Nishikawa, J., et al. (2002). Discordant KCNQ1OT1 imprinting in sets of monozygotic twins discordant for Beckwith-Wiedemann syndrome. *Hum. Mol. Genet.* 11, 1317–1325. doi: 10.1093/hmg/11.11.1317

Wood, D. L., Nones, K., Steptoe, A., Christ, A., Harliwong, I., Newell, F., et al. (2015). Recommendations for accurate resolution of gene and isoform allele-specific expression in RNA-Seq data. *PloS One* 10, e0126911. doi: 10.1371/journal.pone.0126911

Yamada, L., and Chong, S. (2017). Epigenetic studies in developmental origins of health and disease: pitfalls and key considerations for study design and interpretation. *J. Dev. Orig. Health Dis.* 8, 30–43. doi: 10.1017/S2040174416000507

Yan, H., Yuan, W., Velculescu, V. E., Vogelstein, B., and Kinzler, K. W. (2002). Allelic variation in human gene expression. *Science* 297, 1143. doi: 10.1126/science.1072545

Young, P. E., Kum Jew, S., Buckland, M. E., Pamphlett, R., and Suter, C. M. (2017). Epigenetic differences between monozygotic twins discordant for amyotrophic lateral sclerosis (ALS) provide clues to disease pathogenesis. *PloS One* 12, e0182638. doi: 10.1371/journal.pone.0182638

Zhou, Z. Y., Hu, Y., Li, A., Li, Y. J., Zhao, H., Wang, S. Q., et al. (2018). Genome wide analyses uncover allele-specific RNA editing in human and mouse. *Nucleic Acids Res.* 46, 8888–8897. doi: 10.1093/nar/gky613

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.