



OPEN

Unravelling the genetics of non-random fertilization associated with gametic incompatibility

Audrey A. A. Martin¹, Samir Id-Lahoucine^{1,2}, Pablo A. S. Fonseca¹, Christina M. Rochus¹, Lucas M. Alcantara¹, Dan Tulpan¹, Stephen J. LeBlanc³, Filippo Miglior¹, Joaquim Casellas⁴, Angela Cánovas¹, Christine F. Baes^{1,5} & Flavio S. Schenkel¹✉

In the dairy industry, mate allocation is dependent on the producer's breeding goals and the parents' breeding values. The probability of pregnancy differs among sire-dam combinations, and the compatibility of a pair may vary due to the combination of gametic haplotypes. Under the hypothesis that incomplete incompatibility would reduce the odds of fertilization, and complete incompatibility would lead to a non-fertilizing or lethal combination, deviation from Mendelian inheritance expectations would be observed for incompatible pairs. By adding an interaction to a transmission ratio distortion (TRD) model, which detects departure from the Mendelian expectations, genomic regions linked to gametic incompatibility can be identified. This study aimed to determine the genetic background of gametic incompatibility in Holstein cattle. A total of 283,817 genotyped Holstein trios were used in a TRD analysis, resulting in 422 significant regions, which contained 2075 positional genes further investigated for network, overrepresentation, and guilt-by-association analyses. The identified biological pathways were associated with immunology and cellular communication and a total of 16 functional candidate genes were identified. Further investigation of gametic incompatibility will provide opportunities to improve mate allocation for the dairy cattle industry.

Reproduction is a main concern for the dairy industry, as lactation is dependent on the capacity of a cow to produce a calf. Moreover, the fertility of high genetic merit animals is of utmost importance, as these animals are the parents of the next generation. Nevertheless, while female fertility has been part of national selection indexes for dairy cattle around the world over the past decades, the same is not observed for male fertility^{1,2}. Moreover, there is a low genetic correlation between male and female fertility traits, which along with the low heritability of these traits, indicate that the indirect response to selection on female fertility traits would not be sufficient to improve male fertility¹. The males' contribution needs to be considered for the general improvement of reproductive capacity, as it is an integral component of fertilization. For example, genomic selection for sire conception rate is feasible but complicated³. The true reproduction ability and natural variation of bull's conception rate have been masked by the necessary standardization of semen for the widespread use of artificial insemination (AI). Semen doses are extended and titrated to reach an acceptable probability of pregnancy per AI across the population based on sperm characteristics for young bulls and later adjusted when sire conception rate phenotypes are available, which reduces the observed variation in male fertility¹.

Fertilization success is strongly dependent on the compatibility of gametes. The paired "lock and key" mechanism of the spermatozoa and the oocyte is essential and relies on the proper interaction between the proteins of both gametes, which is also called gametic compatibility^{4,5}. It is most studied in externally fertilizing species, such as fish and sea urchin, where the strong interspecies gametic incompatibility replaces the physical isolating barrier⁴. More recently, the biological patterns of gametic interaction have been investigated in mammals, e.g. gamete-mediated mate choice⁶ and cryptic female choice⁷.

¹Centre for Genetic Improvement of Livestock, University of Guelph, Guelph, ON, Canada. ²Department of Animal and Veterinary Sciences, Scotland's Rural College, Edinburgh, Scotland, UK. ³Ontario Veterinary College, University of Guelph, Guelph, ON, Canada. ⁴Departament de Ciència Animal i dels Aliments, Universitat Autònoma de Barcelona, Bellaterra, Spain. ⁵Institute of Genetics, Vetsuisse Faculty, University of Bern, Bern, Switzerland. ✉email: schenkel@uoguelph.ca

Under Mendel's first law, alleles from the same locus should segregate independently, so fertilization, the union of two gametes, should create an equal genotypic ratio in the offspring generation. However, exceptions due to different mechanisms are known, resulting in a biased ratio of genotypes within the descendants⁸. This bias can be investigated with a transmission ratio distortion (TRD) analysis, which identifies if one of the two parental alleles is over- or under-represented in the descendants. Transmission ratio distortion is therefore defined by a deviation from the Mendelian law of inheritance and is linked to diverse mechanisms, from gametic formation to embryo development⁹. A few recent studies used TRD to investigate gametic compatibility (e.g. in mice¹⁰ and plants¹¹), but to our knowledge, no studies have explored genetic incompatibility in livestock species. To study possible gametic incompatibility, the TRD model from Casellas et al.^{12,13} was adapted by including a gametic interaction term in addition to the direct TRD effect, which is based on the parental allele transmission differential.

The complete or partial genetic incompatibility of gametes produces an unbalanced genotypic frequency in the offspring population. This study aimed to uncover genomic regions and to identify candidate genes associated with incompatible gametic combinations, as well as to apply network, over-representation, and guilt-by-association analyses to discover the underlying pathways of gametic incompatibility. In the long-term, and with further investigation, mate allocation could potentially be improved by avoiding mating that are incompatible or have a lesser chance of successful pregnancy.

Results

TRD regions and positional genes. After running the TRD model with the additional gametic interaction term, a total of 482 genomic regions (5217 allelic combinations) were identified as having one or both significant direct and gametic interaction TRD effects. After applying filtering for strong gametic interaction, 429 unique regions were left and 422 were successfully remapped to the ARS-UCD1.2 assembly. The complete list of genomic regions with the corresponding positional genes is reported in Supplementary Material Table S1 with the chromosomal position of each region and the Ensembl ID of every gene contained in each region. All the supplementary materials (Tables and Figures) are available at <https://aaamartin.shinyapps.io/netview/>. The genomic regions contained 2075 Ensembl gene IDs, including 490 without annotation. The number of positional candidate genes within each *Bos taurus* autosome (BTA) is presented in supplementary material Table S2. Compared to the total gene distribution, there were more genes associated with gametic interaction on BTA 10, and fewer on BTA 18.

Network analysis. Creating a minimum network, i.e. reducing the network to a minimum of interactions by only keeping the shortest pair-wise paths between genes with NetworkAnalyst¹⁴ (www.networkanalyst.ca) resulted in 13 subnetworks, one continent (399 nodes) and 12 islands (3 to 11 nodes each). A supplementary network was created when a positional gene could not be connected with an already existing network. Only the first subnetwork had a substantial number of genes and related pathways, therefore only this network was further reported and discussed. However, the complete results, including each network and related Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways^{15,16} and Gene Ontology (GO) terms are in the supplementary materials Fig. S1 and Table S3 (<https://aaamartin.shinyapps.io/netview/>), respectively.

There were 120 significantly enriched KEGG pathways (false discovery rate; FDR < 0.05) in the network. Most pathways were related to immunological and signalling functions: 44 KEGG pathways directly related to disease (due to perturbations of biological networks), e.g. "Hepatitis B" and "Pancreatic cancer"; 14 to different actors or processes of the immune system, such as "Th17 cell differentiation" and "Leukocyte transendothelial migration"; and 40 related to diverse signalling processes, e.g. "Rap1 signalling pathway" and "Thyroid hormone signalling pathway". There were a few pathways associated with reproduction mechanisms including: "GnRH signalling pathway", "Prolactin signalling pathway", "Progesterone-mediated oocyte maturation", "Oocyte meiosis", "Oestrogen signalling pathway", and the "Oxytocin signalling pathway". The other significant KEGG terms usually referred to central cellular mechanisms, for example the MAPK (mitogen-activated protein kinases) signalling pathway that participates in the regulation of cell proliferation, differentiation, motility, and survival¹⁷.

For the GO terms analysis of the network, 55 GO-BP, 51 GO-MF and 38 GO-CC were enriched. Like the KEGG pathway analysis, terms related to immunologic functions, such as "Inflammatory response" and "Positive regulation of cytokine secretion", were present. Additionally, terms relating to different mechanisms associated with the regulation of genetic materials, such as "DNA replication initiation" (GO-BP), "Nucleotide binding" (GO-MF) and "Chromosome" (GO-CC), were abundant.

Overrepresentation analysis. Similar to the network analysis results, the overrepresentation analysis (ORA) pointed to communication, immunological, and metabolic functions but not directly to reproduction processes. It can be noted that only the analysis using GO-CC terms, presented in Table 1, resulted in significant enrichment (FDR < 0.05). Most terms referred to entities carrying or processing the genetic material within the cell, such as chromosome and endoplasmic reticulum. Based on the enrichment ratio, processes involving the aggresome were the most overrepresented. An aggresome is an inclusion body that can transport aggregated proteins on the cytoskeleton to have them "recycled". It is linked to the autophagy process where the cell recycles one of its components, which is also found in the network analysis (Networks 1 and 12). This process is involved in cellular structural changes.

Table 2 shows the ten most enriched KEGG pathways, in which most of them were related to immunology and diseases. Specifically, the process of self and non-self recognition appears to be important, because the highest ratios of overrepresentation were associated with pathways relating to autoimmune diseases, for which the recognition process is defective. However, these results were not statistically significant at a FDR of 5%,

GO-CC	Total	Expected	Hits	Ratio	FDR
Aggresome	18	1	6	5.71	0.03
Nuclear chromosome	302	18	34	1.93	0.03
Nuclear chromosome part	287	17	34	1.91	0.03
Chromosomal part	525	31	52	1.70	0.03
Chromosome	573	33	55	1.64	0.03
Ribonucleoprotein complex	573	33	54	1.61	0.03
Endoplasmic reticulum	875	51	76	1.49	0.03
Nucleoplasm	1444	84	120	1.42	0.01
Cytosol	1795	105	146	1.39	0.01
Nuclear lumen	1937	113	148	1.31	0.03

Table 1. Significantly overrepresented GO-CC terms for a list of 2075 Ensembl gene ID positioned in transmission ratio distortion genomic regions associated with strong gametic interaction. *GO-CC* Gene Ontology-Cellular Component, *Total* total number of genes in the genome associated with each GO term, *Expected* expected number of genes within the network dataset associated with each GO term, *Hits* observed number of genes within the network dataset associated with each GO term, *Ratio* Ratio of overrepresentation, *FDR* false discovery rate.

KEGG Pathway	Total	Expected	Hits	Ratio	FDR
Glycosphingolipid biosynthesis	16	1	5	5.37	0.24
Graft-versus-host disease	47	3	8	2.92	0.26
Type I diabetes mellitus	58	3	9	2.66	0.26
Th17 cell differentiation	112	7	14	2.15	0.26
Gastric cancer	153	9	17	1.91	0.26
Endocytosis	246	14	26	1.81	0.24
Systemic lupus erythematosus	181	11	19	1.80	0.26
Human T-cell leukemia virus 1 infection	268	16	27	1.73	0.26
Human papillomavirus infection	354	21	32	1.55	0.26
Metabolic pathways	1329	77	104	1.34	0.20

Table 2. First 10 most overrepresented KEGG pathways for a list of 2075 Ensembl gene ID from genes positioned in transmission ratio distortion genomic regions associated with strong gametic interaction. *KEGG* Kyoto Encyclopedia of Genes and Genomes, *Total* total number of genes associated with the KEGG pathway, *Expected* expected number of genes within the network dataset associated with each KEGG pathway, *Hits* observed number of genes within the network dataset associated with each KEGG pathway, *Ratio* Ratio of overrepresentation, *FDR* false discovery rate.

therefore they should not be used on their own, but rather as an additional information for the interpretation of the significant results from other analyses in combination with the available literature.”

Guilt-by-association analysis. The identification of the most likely functional candidate genes with the guilt-by-association analysis resulted in 12 genes associated with fertilization, gametic interaction, and recognition. The genes are presented in Table 3. An important note was the presence of *CD9*, which is involved in the fusion of gametes, and was already proposed as a candidate gene for gametic incompatibility⁴. From the gene prioritization based on the functional profile similarity with the 12 candidate genes mentioned above, there were four significant genes ($FDR < 0.05$), which are presented in Table 4.

Discussion

Little is known about gametic incompatibility in mammals and few associated genes and mechanisms have been identified in the last decades⁴. The functional analyses in this study revealed that genomic regions associated with gametic interaction TRD and gametic incompatibility are mostly linked to immunology and communication pathways. It was expected because reproduction and immunology are often intertwined, relying both on cellular communication. Modulation of immunity in the uterus is necessary to protect against pathogens yet allow the survival of allogeneic cells, such as spermatozoa and foetal tissue¹⁸. Immunology and fertility also share a similar genetic architecture. They are both complex traits, meaning that they result from the expression of numerous genes with small effects. These traits are influenced by the environment and are strongly affected by natural selection pressure. The genetic structure of complex traits allows for the maintenance of mechanisms that are central to animal fitness, while permitting important polymorphisms necessary for adaptation. The

Gene Name	BTA	Start (bp)	End (bp)	Description
CATSPERD	7	18,552,688	18,590,398	Cation channel sperm associated auxiliary subunit delta
CD9	5	104,123,604	104,159,285	CD9 molecule
DMC1	5	110,195,281	110,232,976	DNA meiotic recombinase 1
GNPDA1	7	53,066,435	53,077,719	Glucosamine-6-phosphate deaminase 1
MAEL	3	1,829,695	1,901,654	Maelstrom spermatogenic transposon silencer
MCM9	9	32,100,137	32,225,167	Mini-chromosome maintenance 9 homologous recombination repair factor
MFGES	21	20,467,445	20,503,982	Milk fat globule EGF and factor V/VIII domain containing
PCSK4	7	43,888,435	43,897,837	Proprotein convertase subtilisin/kexin type 4
PRND	13	47,086,007	47,091,045	Prion like protein doppel
SLC9B1	6	21,861,338	21,928,497	Solute carrier family 9 member B1
SMAD4	24	50,524,995	50,577,277	SMAD family member 4
UBE2Q1	3	16,022,438	16,032,162	Ubiquitin conjugating enzyme E2 Q1

Table 3. List of candidate genes related to gametic incompatibility based on the keywords “fertilization”, “gamete interaction”, “single fertilization” and “sperm-egg recognition”, present in a list of 2075 Ensembl gene ID from genes positioned in transmission ratio distortion genomic regions associated with strong gametic interaction. *BTA* Bos taurus autosome, *bp* base pair.

Gene Name	BTA	Start (bp)	End (bp)	Description	FDR
ANXA1	8	49,321,701	49,339,582	Annexin A1	0.01
IFNG	5	45,624,513	45,629,336	Interferon gamma	0.01
MET	4	51,620,793	51,735,647	MET proto-oncogene, receptor tyrosine kinase	0.02
ATP2B4	16	1,398,176	1,500,871	ATPase plasma membrane Ca ²⁺ transporting 4	0.03

Table 4. Prioritized candidate genes relating to gametic incompatibility identified by guilt-by-association analysis in a list of 2075 Ensembl gene ID from genes positioned in transmission ratio distortion genomic regions associated with strong gametic interaction. *BTA* Bos taurus autosome, *bp* base pair, *FDR* false discovery rate.

main function of the traits remains unchanged, but it is regulated by a tight gene network, which allows for a lot of potential variation¹⁹.

Interestingly, purely reproductive mechanisms seemed not to be of primary importance for gametic compatibility variation. Some sexual hormone pathways and oocyte development processes were significant in this study, but they are more likely to affect fertilization by producing a sub-optimal environment for the gametes or the embryo. Likewise, the significance of many architectural processes necessary for the general function of the organism was more closely linked to the viability of the embryo than the combination of the gametes. Attributing TRD to specific biological mechanisms is complicated. First, the effects of a disease or phenotypic preselection of the studied animals could create a false TRD signal and confound the interpretation⁹. Preselection bias was mostly avoided with the chosen method for collecting genotypes (described in Materials and Methods, Dataset). Therefore, observed TRD could be attributed to diverse biological mechanisms such as germline selection, meiotic drive, gametic competition, imprinting errors, and embryo lethality⁹. By considering the interaction between the gametic haplotypes, focus was given to mechanisms taking effect in a diploid cell, i.e. at or after fertilization. This specific bias is usually attributed to lethal mutations or lower fitness genotypes in the embryo or young animals, but biased genotypic distributions with normal litter sizes and the absence of dead embryos support non-random fertilization rather than lethality after fertilization⁸. In cattle, there is usually only one calf per pregnancy and detecting embryo losses before approximately 28 days of gestation is difficult. Although data before day 28 are sparse, in dairy cows, fertilization failure appears to occur in 10–20% of AI, whereas approximately 35% of embryos are lost between days 8 and 60 of after AI, skewed to earlier times²⁰. A similar study in pigs or mice would more easily distinguish between failure of fertilization and embryo lethality.

Insights into important processes related to gametic incompatibility can be gained from the overrepresentation analyses even though the results were not significant after FDR correction for multiple tests. The number of identified processes did not increase with the consideration of the overrepresentation analysis based on the KEGG pathways, i.e. the resulting pathways were subsets or on a par with mechanisms of the other analyses with significant results. Although the results from the overrepresentation analysis based on the KEGG pathways should be interpreted with caution, they can add additional information for the interpretation of the significant results from the other analyses in combination with the available literature for deciphering important gametic incompatibility mechanisms. No conclusions should be drawn solely from the non-statistically significant results of the overrepresentation analysis based on the KEGG pathways.

The pathway with the highest enrichment ratio was “Glycosphingolipid biosynthesis” (FDR = 0.24, Ratio = 5.37). Glycosphingolipids are important for cell membrane structure, the modulation of membrane

protein function and external cell communication (host–pathogen interaction and cell–cell recognition)²¹. Genes associated with their biosynthesis are highly conserved in humans and mutated genes are associated with excessive apoptosis during embryo development or are correlated with diverse cancers later in life²¹. Another interesting term is “Endocytosis” (FDR = 0.24, Ratio = 1.81), which relates to functions similar to the aggresome and the autophagy processes, which were significant in the network analysis. After gametic fusion, there is an important release of cortical granules as compensatory endocytosis to maintain the cell surface and avoid polyspermy^{22,23}. Without proper function of this mechanism, egg activation is compromised²².

More generally, the enriched pathways were related directly to immunologic functions or their dysfunctions linked to diseases. The two KEGG pathways with a highest enrichment ratio were “Graft-versus-host disease” (FDR = 0.26, Ratio = 2.92) and “Type I diabetes mellitus” (FDR = 0.26, Ratio = 2.66). These pathways shared similarities with mechanisms important for reproduction. The pathway “Graft-versus-host disease” shares functions with foetal acceptance by the dam. Seven of the eight genes (*BOLA*, *BOLA-DOA*, *BOLA-DMB*, *BOLA-DMA*, *BOLA-NC1*, *LOC512672*, *JSP1*) are part of the major histocompatibility complex (MHC), a set of highly polymorphic genes present on BTA 23 coding for self- and non-self-recognition^{24,25}. The last of the eight genes codes for interferon-gamma (IFN- γ), which plays a role in the foetal immuno-acceptance and male fertility^{26,27}. Type I diabetes mellitus, which involves auto-immunity, also involves the same MHC and IFN- γ genes. Furthermore, five out of ten overrepresented KEGG pathways included a combination of the MHC genes above. Recent studies have concluded that MHC genes may mediate mate selection at both the individual and molecular levels in different ways (reviewed in Ref.⁵). In dairy cattle, the use of artificial insemination placed mating choices in human hands, removing the pre-copulatory sexual selection from the individual animals. However, mate selection also occurs post-copulation and pre-fertilization with the interaction of the male gamete with both the female reproductive tract and gamete. The most relevant example of the importance of gametic compatibility in mammals is the *Izumo* and *Juno* gene pair, coding for sperm surface protein and complementary egg receptor, respectively⁴. The two genes affect the sperm-egg fusion capacity, but not other aspects of fertilization, meaning that current sperm and oocyte quality evaluations done in the dairy industry would not detect anomalies. Discovering genes and mechanisms of gametic incompatibility would allow better prediction of fertility and mating outcomes. Moreover, further studies may allow for the identification of haplotypes with their specific compatibilities to others. This information could be later used as a tool for decision making in mate allocation and, in longer term, to select for individuals with higher general compatibility in the population.

The intent of the guilt-by-association analysis was to focus on the mechanisms of the mate selection within the reproductive tract. Because there are few genes associated with gametic incompatibility in the literature⁴, a thorough discussion of both candidate and prioritized genes follows. Most genes could be divided into three groups based on which fertilization mechanisms they influence: Gain of fertilization ability of the sperm (Group 1), Gametic interaction (Group 2), and Female immuno-acceptance (Group 3). The first mechanism relates to the sperm maturation and activation (Group 1) in the female reproductive tract and in proximity to the oocyte. Calcium signalling is important for sperm motility as it mediates the increased intracellular calcium necessary for capacitation, acrosome reaction, and hypermotility²⁸. Both *CATSPERD* and *ATP2B4*, also called Plasma membrane calcium ATPase 4 (*PMCA4*), participate in this process and are associated with sperm fertilization success in mammals^{28,29}. *SLC9B1* regulates intracellular pH, which is linked to the intracellular calcium concentration^{30,31}. It is important to note that the different steps of sperm activation must be triggered at specific times in the female reproductive tract for the sperm cells to gain their fertilizing ability. For example, the acrosome reaction and hypermotility should be triggered only when the sperm reaches the zona pellucida (ZP), a protein layer protecting the oocyte, for fertilization³². In males, *PCSK4* protects against such premature acrosome reaction associated with lesser binding to the ZP³³. Additionally, different *PRND* genotypes have been significantly associated with different acrosome reaction patterns, leading to differences in fertilization rates in rams³⁴. The expression of *PRND* is also associated with the cryoresistance of sperm in sheep³⁴.

After the proper activation and maturation of the spermatozoon, the interaction of the female and male gametes (Group 2) could reveal pair incompatibility. *CD9* is a well-known gene expressed on the oocyte membrane that defines the fusion ability of the gametes^{35,36}, while *MFGE8* mediates sperm adhesion to the ZP³⁵. Early embryo development and attachment is the next step where a gametic incompatibility may be expressed. By sequence similarity, *GNPDA1* is thought to play a role in calcium oscillations necessary for the egg activation³⁷ and *UBE2Q1* is involved in female hormonal homeostasis and embryoid body formation³⁸. In addition to its gamete interaction function, *MFGE8* is also involved in the regulation and maintenance of the endometrium to prepare for embryo attachment³⁹. *MAEL* is mostly known for its link with spermatogenesis, but it is also associated with germ-cell differentiation within the embryo⁴⁰. In females, *PCSK4* contributes to the development of follicles and may promote placentation³³. The *MET* protein, also called hepatocyte growth factor receptor, has a PCSK-mediated activation in the male germ-line³³ and is associated with abnormal mitosis in endometriosis, which may affect the receptivity of the endometrium to an embryo⁴¹.

Lastly, the immuno-acceptance of the allogenic cells (Group 3), i.e. male gamete and embryo, by the female reproductive tract is critical. Consistent with the overrepresentation analysis, *IFNG* was a prioritized candidate gene for gametic incompatibility. Levels of IFN- γ in the maternal serum can be used to predict the success of the early pregnancy^{27,42}. *ANXA1* is an anti-inflammatory protein that regulates the secretion of steroids. Its knockout (KO) in female mice revealed a female-skewed sex ratio and larger litter size, but no functional alterations were observed in KO males⁴³. Knockout of *ANXA1* changed the uterine inflammatory profile to promote early maternal–fetal interactions and implantation⁴³.

The implications of other genes in gametic incompatibility were more complex to determine. *SMAD4* does not have a known major direct effect on fertility, but it takes part in the regulation of gametic and embryo development^{44–46}. *MCM9* is associated with premature ovarian failure due to a deficit of the germ cell renewal⁴⁷, and this deficiency results in genomic instability due to a reduction in the replication quality check, which

promotes cancer in adults and negatively affects germ-line stem cells⁴⁷. *DMCI* is a meiotic gene that has been associated with infertility in humans and mice^{48,49} but to our knowledge, the mechanism behind this infertility has yet to be linked to any system dysfunctions. Allelic incompatibility, within or between loci, would be a promising hypothesis to describe this unexplained infertility. It is important to note that the choice of keywords used in the analysis is limited to current knowledge of the investigated phenotype. Therefore, the results of the analysis can only pinpoint genes that have already been associated with or are suspected to be associated with mechanisms of gametic incompatibility. As gametic incompatibility is a novel phenotype in livestock, and even in humans, little is known about its genetic background and only a few candidate genes have been identified⁴. With further investigation into the mechanisms, more keywords will be identified, and the candidate gene list will become more comprehensive.

Conclusion

Gametic incompatibility in dairy cattle partially explains unequal genotype ratios in the offspring generation. Based on recognition and immunological functions within the reproductive tract, fertilization is not always a random event as commonly assumed. An animal's probability of fertilization is therefore not only predicted by its innate fertility, but also by its mate compatibility. Currently, the industry does not account for gametic incompatibility, reducing the accuracy of prediction of fertility at the individual and gametic levels. With further investigation, the identification of compatible haplotype pairs and prediction of mating success could become possible, improving mate allocation in dairy cattle.

Materials and methods

Dataset. To study TRD effects linked to gametic incompatibility, a dataset from a study that identified TRD regions associated with reproduction defects was used⁵⁰. This dataset consisted of 436,651 genotyped Canadian Holstein cattle (5976 sires, 132,282 dams and 283,817 offspring) provided by Lactanet (Guelph, Canada), which composed of 283,817 dam-sire-offspring trios. Both parents could be part of different trios with an average of 57.07 and 2.57 trios per sire and dam, respectively. Only the first offspring for a specific mating was considered. Only animals with offspring genotyped within 90 days of birth were selected to avoid a bias associated with the phenotypic preselection of the individuals to be genotyped within a family. The animals were genotyped with different single nucleotide polymorphism (SNP) genotyping arrays and low-density genotypes were imputed with Fimpute⁵¹ to 47,910 SNPs. The genotypes and imputed genotypes were based on the genome assembly UMD3.1.

Transmission ratio distortion model. Under the hypothesis that some allelic TRD signals could be caused by gametic incompatibility, regions previously identified with direct TRD⁵⁰ could be analysed with an alternative model considering the interaction between gametes in the offspring generation. Thus, the inheritance of alleles from parent to offspring could be parametrized including a direct TRD effect (α_{ij}) and an interaction between offspring alleles (β_{ij}). In this case, the probability of an offspring (P_{off}) in a locus with n alleles must be parameterized for each specific mating. Assuming a locus with four alleles and a mating of heterozygous parents with different alleles ($A_1A_2 \times A_3A_4$),

$$P_{off}(A_1A_3) = [(0.5 + \alpha_{12}) \times (0.5 + \alpha_{34})] \times (1 + \beta_{13}) \times (1 - (\beta_{14}/3)) \times (1 - (\beta_{23}/3)) \times (1 - (\beta_{24}/3))$$

$$P_{off}(A_1A_4) = [(0.5 + \alpha_{12}) \times (0.5 - \alpha_{34})] \times (1 - (\beta_{13}/3)) \times (1 + \beta_{14}) \times (1 - (\beta_{23}/3)) \times (1 - (\beta_{24}/3))$$

$$P_{off}(A_2A_3) = [(0.5 - \alpha_{12}) \times (0.5 + \alpha_{34})] \times (1 - (\beta_{13}/3)) \times (1 - (\beta_{14}/3)) \times (1 + \beta_{23}) \times (1 - (\beta_{24}/3))$$

$$P_{off}(A_2A_4) = [(0.5 - \alpha_{12}) \times (0.5 - \alpha_{34})] \times (1 - (\beta_{13}/3)) \times (1 - (\beta_{14}/3)) \times (1 - (\beta_{23}/3)) \times (1 + \beta_{24})$$

where α_{12} and α_{34} were the heterozygous pairwise combinations of direct TRD effects of the implicated parents; β_{ij} was the interaction of alleles i and j in the offspring generation. Note that 1, 2, 3 and 4 were the alleles implicated in this mating.

The direct effects described the probability of transmission of one allele at the expense of the opposite allele in the heterozygous pairwise combination. Flat priors were assumed within a parametric space ranging from -0.5 to 0.5 , based on the principles of Mendelian inheritance. The probability of transmission of one specific allele ranged from 0 ($\alpha_{ij} = -0.5$) to 1 ($\alpha_{ij} = 0.5$), where 0.5 ($\alpha_{ij} = 0$) corresponds to no TRD.

The TRD parameters of each specific alleles' interaction in the offspring genotype were included to model the incompatibility of the offspring genotype. For gametic incompatibility, certain alleles' interactions should be underrepresented and display a genotype interaction TRD pattern.

The gametic interaction TRD parameter indicated the probability of an offspring to be produced and viable. This parameter was assumed to range from -1 (decreased probability) to $+1$ (increased probability). A magnitude of 0 indicated no gametic interaction TRD. Note that when one offspring was over- or under-represented, the change in the frequency was compensated by one or several other offspring genotypes of the same mating.

For the gametic interaction TRD parameters, flat priors were also assumed. The probabilities of offspring genotypes must be adapted for each specific mating, but could be generalized as:

$$P_{off}(A_{s_1}A_{d_1}) = [(0.5 + \alpha_{s_1,s_2}) \times (0.5 + \alpha_{d_1,d_2})] \times (1 + \beta_{s_1,d_1}) \times (1 - (\beta_{s_1,d_2}/3)) \times (1 - (\beta_{s_2,d_1}/3)) \times (1 - (\beta_{s_2,d_2}/3))$$

$$P_{off}(A_{s_1}A_{d_2}) = [(0.5 + \alpha_{s_1,s_2}) \times (0.5 - \alpha_{d_1,d_2})] \times (1 - (\beta_{s_1,d_1}/3)) \times (1 + \beta_{s_1,d_2}) \times (1 - (\beta_{s_2,d_1}/3)) \times (1 - (\beta_{s_2,d_2}/3))$$

$$P_{\text{off}}(A_{s_2}A_{d_1}) = [(0.5 - \alpha_{s_1,s_2}) \times (0.5 + \alpha_{d_1,d_2})] \times (1 - (\beta_{s_1,d_1}/3)) \times (1 - (\beta_{s_1,d_2}/3)) \times (1 + \beta_{s_2,d_1}) \times (1 - (\beta_{s_2,d_2}/3))$$

$$P_{\text{off}}(A_{s_2}A_{d_2}) = [(0.5 - \alpha_{s_1,s_2}) \times (0.5 - \alpha_{d_1,d_2})] \times (1 - (\beta_{s_1,d_1}/3)) \times (1 - (\beta_{s_1,d_2}/3)) \times (1 - (\beta_{s_2,d_1}/3)) \times (1 + \beta_{s_2,d_2})$$

where α_{ij} was the heterozygous pairwise combination of direct TRD effect; β_{ij} was the gametic interaction TRD parameter; s_1 and s_2 were the alleles from the sire; and d_1 and d_2 were the alleles from the dam.

For a specific mating, a maximum of 4 gametic interaction TRD parameters were involved. When all TRD effects were null, the probability of each allelic combination for a genotype was 0.25 and summed to 1 for the 4 possible combinations of genotypes. It is important to remember that the same offspring's genotype may be generated in more than one of these 4 possible combinations and must be added to obtain the probability of each offspring's genotype. Under a Bayesian implementation, the conditional posterior probabilities of the TRD parameters were defined as:

$$p(\alpha_{12}, \alpha_{13}, \dots, \alpha_{(n-1)n}, \beta_{11}, \beta_{12}, \dots, \beta_{nn} | y) \\ \propto p(y | \alpha_{12}, \alpha_{13}, \dots, \alpha_{(n-1)n}, \beta_{11}, \beta_{12}, \dots, \beta_{nn}) p(\alpha_{12}) p(\alpha_{13}) \dots p(\alpha_{(n-1)n}) p(\beta_{11}) p(\beta_{12}) \dots p(\beta_{nn})$$

where y was the vector of genotypes of the offspring generation.

As an exhaustive search for gametic interaction TRD is computationally intractable, the estimation was based on 602 regions previously identified with significant direct TRD effect⁵⁰. The analysis was performed within a Bayesian framework with the metropolis-Hastings⁵² sampling technique using an adapted version of TRDscan v.1.0 software⁵³. Each parameter was sampled separately, assuming the other parameters to be known a priori. A unique Monte Carlo Markov chain of 110,000 iterations was used, with the first 10,000 discarded as burn-in. The statistical significance of TRD was evaluated using a Bayes factor (BF)⁵⁴. A threshold of $\text{BF} \geq 100$ was used to determine decisive evidence for TRD.

Functional analysis. The regions were filtered based on the magnitude of the gametic interaction term. This value is specific to the haplotype combination within each region. Therefore, multiple allelic combinations within one region could be detected with significant gametic interaction TRD. Only regions with at least one combination with a magnitude less than -0.5 or greater than 0.5 were retained.

As the TRD analysis was performed with genotypes based on the previous genome assembly, UMD3.1, the TRD regions were remapped to the current genome assembly, ARS-UCD1.2, using the NCBI remap tool (www.ncbi.nlm.nih.gov/genome/tools/remap) before performing the functional analyses. Regions with large insertions or low coverage after remapping were removed from the list. To ensure proper conversion of the regions, a comparison of the extracted annotations from the two assemblies was performed using the GALLO package⁵⁵. The two gene extractions overlapped by 85% of the annotations, and the remaining 15% were composed of repetitive elements, such as pseudogenes, or annotations without assigned gene symbols.

The positional genes were then extracted based on the chromosomal positions of the remapped regions using the GALLO package⁵⁵. An additional 50,000 bp were added up and downstream to each region. The extracted list was used to perform a network analysis with the NetworkAnalyst software¹⁴. The study was performed with the Ensembl gene IDs, using a protein-protein interactions (PPI) analysis. This approach used predefined networks from the STRING interactome database⁵⁶ to build a network from the input gene list and other proteins. This PPI analysis created a broad view of the biological mechanisms linked to the investigated phenotype, in this case, gametic incompatibility.

A minimum interaction network that kept the essential proteins needed to maintain the network connections and minimize the use of enriched proteins was created. A biological processes analysis was performed to identify significantly overrepresented biological pathways within the gene list through the NetworkAnalyst software¹⁴ based on GO terms and the KEGG pathways.

To further the investigation of gametic incompatibility, a "guilt-by-association"-based prioritization analysis was performed on the positional gene list using GUILDify 2.0⁵⁷ and ToppGene⁵⁸. First, GUILDify was used to create a trained list of the positional genes associated with specific keywords relating to the phenotype, i.e. "fertilization", "gamete interaction", "single fertilization" and "sperm-egg recognition". The definition and potential implication of the genes in gametic incompatibility were investigated with the GeneCards database⁵⁹ to confirm their relevance within the trained list. This trained list was then fitted into ToppGene against the full list of positional genes, for an annotation-based prioritization analysis. This is a multivariate approach that scores the functional profile similarity of each positional gene with the trained list profile, based on Gene Ontology terms for molecular function (MF), biological process (BP), and cellular component (CC); human and mouse phenotypes; metabolic pathways; Pubmed publications; Coexpression atlas; ToppCell atlas; and diseases. Intermediate p-values for each of the above functional terms and pathways, were obtained and combined into an overall p-value for each gene. This information was calculated by comparing the gene list with a random sample of 5000 genes from the whole genome for each annotation information. Then, the overall p-value was corrected for multiple testing using an FDR of 5%. Significantly prioritized genes shared a similar functional profile with the genes from the trained list. As the genes from the trained list were known to be associated with the phenotype, the prioritized genes are likely to also be associated with gametic incompatibility under the principle of guilt-by-association.

Additionally, an over-representation analysis (ORA) was performed to determine if any biological pathways were over-represented in the gene list. This was done by uploading the list of genes identified by their Ensembl gene ID to the WebGestalt software⁶⁰. The significance of each pathway is defined by the difference between the number of genes observed and the expected number of genes that would have been obtained from a random set of genes of the same size given the whole genome annotation. The ratio of over-representation was also obtained and reported.

Data availability

The genotypic data analysed for the current study were obtained from Lactanet (Guelph, Canada). This dataset was used under agreement, and therefore, not publicly available. However, data can be made available from F.M. on reasonable request and with permission of Lactanet.

Received: 22 July 2022; Accepted: 21 December 2022

Published online: 24 December 2022

References

- Taylor, J., Schnabel, R. & Sutovsky, P. Review: Genomics of bull fertility. *Animal* **12**, s172–s183 (2018).
- Miglior, F. *et al.* A 100-year review: Identification and genetic selection of economically important traits in dairy cattle. *J. Dairy Sci.* **100**, 10251–10271 (2017).
- Abdollahi-Arpanahi, R., Morota, G. & Peñagaricano, F. Predicting bull fertility using genomic data and biological information. *J. Dairy Sci.* **100**, 9656–9666 (2017).
- Springate, L. & Frasier, T. R. Gamete compatibility genes in mammals: Candidates, applications and a potential path forward. *R. Soc. Open Sci.* **4**, 170577 (2017).
- Jokiniemi, A. *et al.* Gamete-level immunogenetic incompatibility in humans—towards deeper understanding of fertilization and infertility?. *Heredity* **125**, 281–289 (2020).
- Kekäläinen, J. & Evans, J. P. Gamete-mediated mate choice: Towards a more inclusive view of sexual selection. *Proc. R. Soc. B Biol. Sci.* **285**, 20180836 (2018).
- Firman, R. C., Gasparini, C., Manier, M. K. & Pizzari, T. Postmating female control: 20 years of cryptic female choice. *Trends Ecol. Evol.* **32**, 368–382 (2017).
- Nadeau, J. H. Do gametes woo? Evidence for their nonrandom union at fertilization. *Genetics* **207**, 369–387 (2017).
- Huang, L. O., Labbe, A. & Infante-Rivard, C. Transmission ratio distortion: Review of concept and implications for genetic association studies. *Hum. Genet.* **132**, 245–263 (2013).
- Arends, D., Heise, S., Korkuc, P., Hesse, D. & Brockmann, G. A. Transmission distortion and genetic incompatibilities between alleles in a multigenerational mouse advanced intercross line. *Genetics* <https://doi.org/10.1101/2021.06.09.447720> (2021).
- Kerwin, R. E. & Sweigart, A. L. Mechanisms of transmission ratio distortion at hybrid sterility loci within and between *mimulus* species. *G3 Genes Genomes Genet.* **7**, 3719–3730 (2017).
- Casellas, J., Id-Lahoucine, S. & Cánovas, A. Discriminating between allele- and genotype-specific transmission ratio distortion. *Anim. Genet.* **51**, 847–854 (2020).
- Casellas, J., Manunza, A., Mercader, A., Quintanilla, R. & Amills, M. A flexible Bayesian model for testing for transmission ratio distortion. *Genetics* **198**, 1357–1367 (2014).
- Zhou, G. *et al.* NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. *Nucl. Acids Res.* **47**, W234–W241 (2019).
- Kanehisa, M. KEGG: Kyoto encyclopedia of genes and genomes. *Nucl. Acids Res.* **28**, 27–30 (2000).
- Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucl. Acids Res.* **44**, D457–D462 (2016).
- Cargnello, M. & Roux, P. P. Activation and function of the MAPKs and their substrates, the MAPK-activated protein kinases. *Microbiol. Mol. Biol. Rev.* **75**, 50–83 (2011).
- Sheldon, I. M., Owens, S.-E. & Turner, M. L. Innate immunity and the sensing of infection, damage and danger in the female genital tract. *J. Reprod. Immunol.* **119**, 67–73 (2017).
- Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: From polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).
- Wiltbank, M. C. *et al.* Pivotal periods for pregnancy loss during the first trimester of gestation in lactating dairy cows. *Theriogenology* **86**, 239–253 (2016).
- Schnaar, R. L. & Kinoshita, T. *Glycosphingolipids. Essent. Glycobiol.* <https://doi.org/10.1101/glycobiology.3e.011> (2015).
- Gómez-Eliás, M. D., Fissore, R. A., Cuasnicú, P. S. & Cohen, D. J. Compensatory endocytosis occurs after cortical granule exocytosis in mouse eggs. *J. Cell. Physiol.* **235**, 4351–4360 (2020).
- Betz, W. J. & Richards, D. A. What goes out must come in. *Nat. Neurosci.* **3**, 636–637 (2000).
- Janeway Jr, C. A., Travers, P., Walport, M. & Shlomchik, M. J. The major histocompatibility complex and its functions. In *Immunobiology: The Immune System in Health and Disease* (Garland Science, 2001).
- Behl, J. D. *et al.* The major histocompatibility complex in bovines: A review. *ISRN Vet. Sci.* **2012**, 1–12 (2012).
- Paradisi, R., Capelli, M., Mandini, M., Bellavia, E. & Flamigni, C. Increased levels of interferon-gamma in seminal plasma of infertile men. *Andrologia* **28**, 157–161 (2009).
- Murphy, S. P. *et al.* Interferon gamma in successful pregnancies. *Biol. Reprod.* **80**, 848–859 (2009).
- Chung, J.-J., Navarro, B., Krapivinsky, G., Krapivinsky, L. & Clapham, D. E. A novel gene required for male fertility and functional CATSPER channel formation in spermatozoa. *Nat. Commun.* **2**, 153 (2011).
- Olli, K. E., Li, K., Galileo, D. S. & Martin-DeLeon, P. A. Plasma membrane calcium ATPase 4 (PMCA4) co-ordinates calcium and nitric oxide signaling in regulating murine sperm functional activity. *J. Cell. Physiol.* **233**, 11–22 (2018).
- Liu, T. *et al.* A novel testis-specific Na/H exchanger is involved in sperm motility and fertility. *Front. Biosci.* **E2**, 115 (2010).
- Chen, S.-R. *et al.* Sodium–hydrogen exchanger NHA1 and NHA2 control sperm motility and male fertility. *Cell Death Dis.* **7**, e2152–e2152 (2016).
- Georgadaki, K., Khoury, N., Spandidos, D. A. & Zoumpourlis, V. The molecular basis of fertilization (Review). *Int. J. Mol. Med.* **38**, 979–986 (2016).
- Gyamera-Acheampong, C. & Mbikay, M. Proprotein convertase subtilisin/kexin type 4 in mammalian fertility: A review. *Hum. Reprod. Update* **15**, 237–247 (2009).
- Ferreira, L. M. *et al.* Prion protein 2 (duplet) gene (PRND): role in ovine semen capacitation, cryopreservation and fertility. *Reprod. Fertil. Dev.* **29**, 985 (2017).
- Ensslin, M. A. & Shur, B. D. Identification of mouse sperm SED1, a Bimotif EGF repeat and discoidin-domain protein involved in sperm-egg binding. *Cell* **114**, 405–417 (2003).
- Miyado, K. *et al.* The fusing ability of sperm is bestowed by CD9-containing vesicles released from eggs in mice. *Proc. Natl. Acad. Sci.* **105**, 12921–12926 (2008).
- UniProt. Glucosamine-6-phosphate isomerase 1. (2021). Available at: <https://www.uniprot.org/uniprot/P46926#function>. Accessed 7 Dec 2021.
- UniProt. Ubiquitin-conjugating enzyme E2 Q1. (2021). Available at: <https://www.uniprot.org/uniprot/Q7Z7E8>. Accessed 7 Dec 2021.
- Yu, L. *et al.* MFGE8 regulates TGF- β -induced epithelial mesenchymal transition in endometrial epithelial cells in vitro. *Reproduction* **152**, 225–233 (2016).

40. Bahena, I. *et al.* Role of Mael in early oogenesis and during germ-cell differentiation from embryonic stem cells in mice in vitro. *Zygote* **22**, 513–520 (2014).
41. Osuga, Y. *et al.* Hepatocyte growth factor concentrations are elevated in peritoneal fluid of women with endometriosis. *Hum. Reprod.* **14**, 1611–1613 (1999).
42. Fasouliotis, S. J. Maternal serum levels of interferon- and interleukin-2 soluble receptor- predict the outcome of early IVF pregnancies. *Hum. Reprod.* **19**, 1357–1363 (2004).
43. Hebeda, C. B. *et al.* Endogenous annexin A1 (AnxA1) modulates early-phase gestation and offspring sex-ratio skewing. *J. Cell. Physiol.* **233**, 6591–6603 (2018).
44. GeneCards. SMAD Family Member 4. Available at: <https://www.genecards.org/cgi-bin/carddisp.pl?gene=SMAD4&keywords=smad4>. Accessed 7 Dec 2021.
45. Li, X., Tripurani, S. K., James, R. & Pangas, S. A. Minimal fertility defects in mice deficient in oocyte-expressed Smad41. *Biol. Reprod.* **86**, 1–6 (2012).
46. Zhang, L., Du, X., Wei, S., Li, D. & Li, Q. A comprehensive transcriptomic view on the role of SMAD4 gene by RNAi-mediated knockdown in porcine follicular granulosa cells. *Reproduction* **152**, 81–89 (2016).
47. Hartford, S. A. *et al.* Minichromosome maintenance helicase paralog MCM9 is dispensible for DNA replication but functions in germ-line stem cells and tumor suppression. *Proc. Natl. Acad. Sci.* **108**, 20271–20271 (2011).
48. Tran, T. N. & Schimenti, J. C. A putative human infertility allele of the meiotic recombinase DMC1 does not affect fertility in mice. *Hum. Mol. Genet.* **27**, 3911–3918 (2018).
49. Mandon-Pépin, B. *et al.* Genetic investigation of four meiotic genes in women with premature ovarian failure. *Eur. J. Endocrinol.* **158**, 107–115 (2008).
50. Id-Lahoucine, S. *Mendelian Inheritance in the Genomics and Big Data Era: Transmission Ratio Distortion Phenomenon in Cattle Breeds* (University of Guelph, 2020).
51. Sargolzaei, M., Chesnais, J. P. & Schenkel, F. S. A new approach for efficient genotype imputation using information from relatives. *BMC Genomics* **15**, 478 (2014).
52. Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109 (1970).
53. Id-Lahoucine, S. *et al.* Implementation of Bayesian methods to identify SNP and haplotype regions with transmission ratio distortion across the whole genome: TRDscan v.1.0. *J. Dairy Sci.* **102**, 3175–3188 (2019).
54. Kass, R. E. & Raftery, A. E. Bayes Factors. *J. Am. Stat. Assoc.* **90**, 773–795 (1995).
55. Fonseca, P. A. S., Suárez-Vega, A., Marras, G. & Cánovas, Á. GALLO: An R package for genomic annotation and integration of multiple data sources in livestock for positional candidate loci. *GigaScience* **9**, 1–9 (2020).
56. Szklarczyk, D. *et al.* STRING v10: Protein–protein interaction networks, integrated over the tree of life. *Nucl. Acids Res.* **43**, D447–D452 (2015).
57. Aguirre-Plans, J. *et al.* GUILDify v2.0: A tool to identify molecular networks underlying human diseases, their comorbidities and their druggable targets. *J. Mol. Biol.* **431**, 2477–2484 (2019).
58. Chen, J., Bardes, E. E., Aronow, B. J. & Jegga, A. G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucl. Acids Res.* **37**, W305–W311 (2009).
59. Stelzer, G. *et al.* The GeneCards suite: From gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinform.* **54**(1), 1–30 (2016).
60. Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z. & Zhang, B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucl. Acids Res.* **47**, W199–W205 (2019).

Acknowledgements

This research was financially supported by Agriculture and Agri-Food Canada, and by additional contributions from Dairy Farmers of Canada, the Canadian Dairy Network and the Canadian Dairy Commission under the Agri-Science Clusters Initiative. As per the research agreement, aside from providing financial support, the funders have no role in the design and conduct of the studies, data collection and analysis or interpretation of the data. Researchers maintain independence in conducting their studies, own their data, and report the outcomes regardless of the results. The decision to publish the findings rests solely with the researchers. Drs Tulpan, Cánovas, Baes and Schenkel gratefully acknowledge the funding and support from the Resilient Dairy Genome Project as administered through Genome Canada (Canada), Genome Alberta (Canada), Genome British Columbia (Canada), Genome Ontario (Canada) and Genome Québec (Canada). Dr Baes acknowledges support from the NSERC CRC program.

Author contributions

A.M., S.I.L., J.C., A.C. and F.S. conceived and designed the TRD model. F.M. facilitated the transfer of data from Lactanet. S.I.L. performed the TRD analysis. A.M. and P.F. performed the functional analyses. A.M., P.F., D.T., S.L. and F.S. discussed, analysed, and interpreted the results. C.R. and C.B. helped the first author to finalize the writing of the manuscript. A.M., D.T., and L.A. created the interactive supplementary materials. F.S. and C.B. supervised A.M.'s research work. All the authors reviewed the manuscript.

Funding

Funding was provided by Dairy Research Cluster grant number Dairy Research Cluster 3 - Activity 3, Resilient Dairy Genome Project (Genome Canada LSARP) grant number L18DCR and NSERC CRC grant number 950-232238.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-022-26910-8>.

Correspondence and requests for materials should be addressed to F.S.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022