THE UNIVERSITY OF CHICAGO

MOLECULAR EVOLUTION OF PREGNANCY

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES

AND THE PRITZKER SCHOOL OF MEDICINE

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

DEPARTMENT OF HUMAN GENETICS

BY

KATELYN MARIE MIKA

CHICAGO, ILLINOIS

JUNE 2018

To my family who love me.

To the scientists who inspire me.

To my friends who push me and who tell me I’m good enough as is.

Thank you all.

Table of Contents

Table of Contents (cont.)

List of Figures

List of Tables

Acknowledgements

I have so many people to thank for supporting, encouraging, listening, goofing off with, and just being there while I have been on this crazy grad school journey. Very top of the list is my advisor Vinny Lynch. I always cite Vinny's excitement for science as one of the greatest things about being in his lab, but the amount of support I have felt both scientifically and personally throughout my time here cannot be understated. I'm honored and proud that I get to hold claim to being his first graduate student; I wouldn't be the scientist I am today if I had been in any other lab. I also need to thank Yoav Gilad for the mentoring I received through joint lab meetings and chats over the fish tank in his office, the guidance was always appreciated. I would also be remiss to not thank my undergraduate mentor Kristi Montooth as well for teaching me what research actually is, laying the foundation for my love of evolution, and still encouraging me to this day.

My current and past labmates all deserve a special thank you. In order of appearance- Sravanthi Chigurupati, Mike Sulak, Erin Fry, Marcus Solail, Mirna Marinic, Manny Vazquez- you all are what made every day in lab an adventure. I'm still not sure how we ever got work done, yet here we are. Thank you for all the protocol and coding help, the life discussions, and general fun times. To my "extended lab" of the third floor- the Gilad Lab, Nobrega Lab, and Basu Lab members- thank you as well for the help at both conventional and unconventional hours, the feedback at joint lab meeting, and the fun discussions about anything at all.

There are so many friends who have been physically with me on this journey as other graduate students, understanding both the pain and elation that comes during our program, as well as those looking in from elsewhere who have given me so much love and support. First, Alex Advani, Andrei Anghel, Diedre Reitz. You all have been my closest companions through

this time and there really aren't enough words to say what that has meant to me. Bryan Pavlovic, Aarti Venkat, Bill Richter, Alex Gileta, Darcy Ross, Phil Ross- you all have also meant a lot to me and I'm glad we have gone through this process (and so much life) together. To the many other HG and MolBio students who have made my UofC community so great, thank you. I have to give a shoutout to the Learn Scuba Chicago community- particularly Rafael Vescovi and Sara Ther. I love you giant bunch of weirdos and I'm glad I found my way to your doors. Finally, to my friends who have been there for what seems like forever and still check in to make sure I'm alright and send me encouragement, you all know how much you mean to me more than I can write here- Destinee Metoyer, Andi Kessler, Ajay Major, Kate Sanders, Kate Neff, Zach Hallberg, Erica Anderson.

Last but far from least, I need to thank my family. Mom and Dad, thank you for always trying to understand what I'm doing, pushing me to do better no matter what, and being a sounding board when I need to work out my problems. Tom and Nate, you're buttheads but I love you anyway and you've definitely chipped in to help get me where I am over the years as well. I don't say it enough to you but thank you for that. And a thank you to my dog Finn. You may cover everything I own in hair but your cuddles were a necessary part of surviving this thesis writing process.

Chapter 1:

Introduction

Understanding phenotypic diversity and origins has long been a driver of scientific inquiry. Within this broad category of scientific questioning, one lingering unknown is how do phenotypic novelties arise? Previous investigations have explained a number of examples where a novel phenotype originated from a loss of an existing trait [1-3], but little is known about the reverse, how a novel phenotype is gained with no pre-existing related character. Among the problems with identifying the genetic and molecular mechanisms that underlie the origin of evolutionary novelties is a lack of species that span the origin of novel characters. For example, feathers are a novel character developmentally homologous to other epidermal appendages such as scales but no extant species preserve the intermediate stages in the evolutionary transformation of scales to feathers. In contrast, extant mammals preserve intermediate stages in the origins of viviparity, thus pregnancy is an excellent model system to explore the mechanisms that underlie the origin and diversification of novelty.

Oviparous (egg laying) monotremes, whose only extant members are the echidna and platypus, exemplify the first stage in the stepwise evolution of mammalian pregnancy (**Fig 1**). Despite being oviparous, monotremes have under a month of maternal provisioning of the embryo prior to egg shell formation [4-6]. Viviparity (live birth) evolved in the Therian mammals, including both marsupials and "placental" mammals (**Fig 1**). These two groups, however, have very different reproductive life histories. Marsupial pregnancy occurs within an uninterrupted estrus cycle, except in swamp wallaby (*Wallabia biocolor*) whose gestation is two days longer than the estrus cycle, and gestational length mostly falls within 12.5 to 36.4 days (mean 22.4

days) [6-8]. In contrast, Eutherian (or "placental") mammals evolved interrupted reproductive cycles and decidualization, which underlies a suite of derived traits such as extended gestations (between 16 and 679 days, mean 83 days) [9], placental invasion, maternal recognition of pregnancy, and fetal immunotolerance [5,10]. This stepwise evolution of pregnancy in mammals allows us to use comparative techniques to pull apart the underlying genetic structure and begin to determine what functional molecular changes occurred along pregnancy's evolutionary path.

One molecular mechanism of phenotypic change is the alteration via mutation of transcription factor binding sites (TFBSs) located in cis-regulatory regions. Cis-regulatory regions include enhancers, promoters, insulators, and repressors that modify the expression of a gene. Changes to these regulatory structures have long been hypothesized to be a mechanism of phenotypic change since it was originally proposed in 1961 by Monod and Jacob [11], and argued further in 1975 by King and Wilson [12]. One particular model of regulatory evolution was also proposed at this time by Britten and Davidson [13] which built upon the Nobel prize winning work of Barbara McClintock [14]. Their model proposed that repetitive elements, such as transposable elements (TEs), could integrate into the genome and affect the regulation of nearby genes and contribute to the evolution of novelty. Since then, studies on the loss of traits has highlighted that changes to regulatory sequences could lead to loss or change of gene expression and therefore loss or change of the phenotype [1-3]. However, while there are a few examples of the formation of novel regulatory sequences resulting in a gain in gene expression [15], no study has conclusively shown that the novel expression has led to the gain of a novel phenotype, but rather the altering of an existing one. During the evolution of mammalian pregnancy, the expression of many genes has shifted (**Fig 1**). This dissertation aims to illuminate 1) how TEs underlie these expression shifts and 2) what affect these shifts have had on pregnancy phenotypes.

Maternal-fetal immunotolerance is a critical pregnancy phenotype [16]. The maternal immune system should recognize the developing fetus as foreign and mount an immune response, effectively terminating the pregnancy. However, this does not occur in a successful pregnancy and extensive changes to immune system functions prevent the maternal immune system from recognizing the semi-allogenic fetus as foreign and mounting an immune response [17]. This modulation of the immune system is mediated by differentiation (decidualization) of endometrial stromal fibroblasts (ESFs) to decidual stromal cells (DSCs) in response to

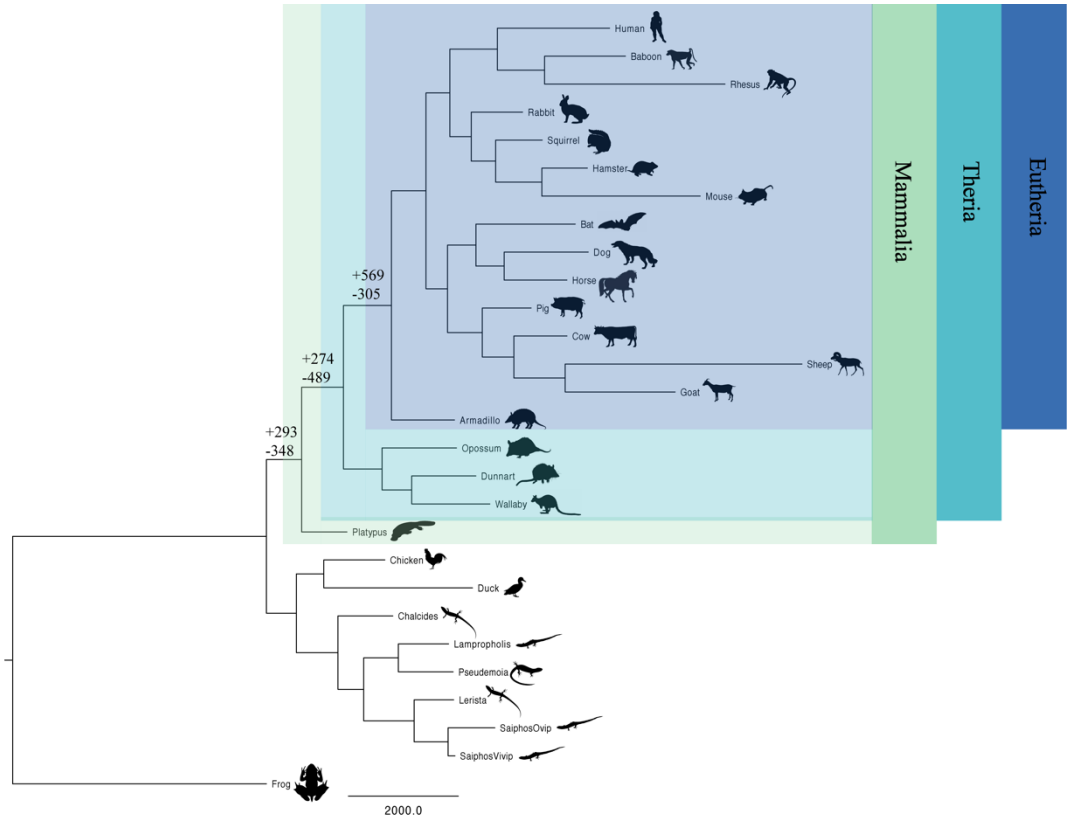

Fig. 1 Mammalian pregnancy evolved stepwise along the lineage. Monotremes, egg-laying mammals like the platypus, are representative of the first evolutionary step. Therian mammals then evolved viviparity, or live birth. Next, Eutherian (or "placental") mammals evolved interrupted reproductive cycles and decidualization. Branch lengths of the tree are proportional to the number of gene expression shifts that occur on that branch. The most parsimonious number of gene expression gains and losses are written on the Mammalian, Therian, and Eutherian branches.

progesterone, cyclic-AMP (cAMP), and, in some species, fetal signals [18-20], which then allows for the 'window of implantation' to occur in the mid-secretory phase of the menstrual cycle [21]. While there are many components to the immune response, one region of interest in both immunology and pregnancy research is the major histocompatability complex (MHC).

The MHC is located on chromosome 6 in humans and divided into three subregions – class I, class II, and class III. The genes in this region, also known as the Human Leukocyte Antigen (HLA) region, play an important role in the rejection of non-self tissues, but also contribute to maternal tolerance of the fetus [22-41]. Canonically the 'classical' MHC genes are expressed on the cell-surface of adult somatic cells and facilitate the identification of self by the immune system. During pregnancy, matching classical HLA antigens leads to decreased fitness with higher miscarriage rates among couples matching for class I HLA-B antigens [26] and longer intervals to pregnancy among couples matching for class II HLA-DR antigens [25]. During pregnancy, 'non-classical' HLA genes, such as HLA-F and HLA-G, have been associated with fecundability, miscarriage, recurrent pregnancy loss, and preeclampsia [30-42]. Also necessary for proper MHC function are the *antigen peptide transporters 1* and *2* (*TAP1* and *TAP2)* gene within the HLA region which together form a heterodimer and translocate peptides from the cytosol to awaiting MHC class I molecules in the endoplasmic reticulum, resulting in cell surface presentation of the trimeric MHC complex to immune cells such as T lymphocytes and natural killer cells [43].

Burrows *et al* [42] identified single nucleotide polymorphisms that are expression quantitative trait loci (eQTL) for *TAP2* (rs2071473) and *HLA-F* (rs2523393). These eQTLs are independently associated with fecundability, or the probability of getting pregnant in one reproductive cycle. The C allele of rs2071473 was associated with longer intervals to pregnancy and higher expression of the *TAP2* gene in mid-secretory phase (receptive) endometrium, while the T allele

was associated with the reverse. Conversely, the G allele of rs2523393 was associated with longer intervals to pregnancy and lower expression of the HLA-F in mid-secretory phase endometrium, whereas the A allele was associated with shorter intervals to pregnancy and higher HLA-F expression. As explored in chapters 1 and 2 of this thesis, we found that these SNPs destroy a DDIT3 (CHOP10) repressor element and create a new GATA2 enhancer element for *TAP2*[44] and *HLA-F* respectively by altering the nucleotides present in the TFBSs. These are further examples of how modification to a single regulatory region can affect gene expression and the downstream phenotype.

When considering a larger-scale remodeling or origin of the entire gene regulatory networks that underlie a complex trait, such as mammalian pregnancy, one prominent hypothesis is that sequences present in transposable elements (TEs) may have been exapted by the genome into a regulatory role (reviewed by Feschotte [45] and Rebollo [46]). Because TEs have a high copy number throughout the genome, multiple exaptations of the same TE could lead to novel regulatory network formation or modification of an existing regulatory pathway. This exaptation could occur in one of two points in time in the TE's evolutionary history. Either 1) the TE enters the genome with a transcription factor binding site (TFBS) or through mutation gains a TFBS early on and then copies this TFBS with the TE throughout the genome. Or 2) after a TE has transposed throughout the genome, individual instances of the TE convergently evolve the same TFBS. Regardless of evolutionary path, TEs have been found to be capable of functioning as regulatory elements by harboring TFBSs [47-49], and a number of these TFBSs have been shown to be bound [50-52].

We have previously used DNaseI-seq, FAIRE-seq, H3K4me3 ChIP-seq, and H3K27ac ChIP-seq to map the regulatory landscape in differentiated human endometrial stromal cells (DSCs)

and found that 42.2% of FAIRE-seq peaks, 67.2% of DNaseI-seq peaks, 58.7% of H3K27ac, and 53.0% of H3K4me3 peaks overlapped annotated transposable elements from diverse age classes [53]. This includes 352 TE families that are enriched within regulatory elements in ESFs compared to their genomic abundances. These results suggest TEs have played a large-scale and continuous role in the origination and turnover of regulatory elements in human endometrium, however, whether these TEs integrated into the genome with existing regulatory function, and therefore immediately affected the expression of host genes, or acquired regulatory functions after integration through additional mutations is unknown. The role of these TEs in the evolution of the novel gene regulatory networks underlying mammalian pregnancy is explored in the third chapter of this dissertation.

This dissertation explores the effects that changes in gene regulation have on the evolution of phenotypic novelties by combining evolutionary and molecular genetic techniques to functionally characterize regulatory variants in our DSC *in vitro* model of pregnancy. The first chapters address how single nucleotide polymorphisms can alter a regulatory region and affect the downstream phenotype. We found that rs2071473 and rs2523393 alter the expression of *TAP2* and *HLA-F* respectively, which in turn alters fecundability. The C allele of rs2071473 interrupts the repressor binding site for DDIT3 and switches the ancestral repressor into a *TAP2* enhancer. The derived A allele of rs2523393 forms a perfect GATA2 binding site with the surrounding nucleotides and increases the expression of *HLA-F*, leading to a faster time to pregnancy. The last chapter begins to pull apart the contribution of TEs to the gene regulatory networks that arose during the evolution of mammals and allowed for the novel gene expression in DSCs underlying many pregnancy traits. We tested the hypothesis that TEs integrate with functional TFBSs and found that nearly all of the TE consensus sequences had regulatory

abilities, the majority of which were cell-type independent repressors in mammalian cells. This led to our development of a new model of TE cooption by the genome where active TEs are first recognized and silenced by KRAB Zinc Finger Proteins (ZFPs) and latent enhancer functions are later revealed once TEs lose KRAB-ZFP binding sites. All three studies contribute to the growing body of knowledge on the genetics of pregnancy, though many questions still linger.

Chapter 2:

An ancient fecundability-associated polymorphism switches a repressor into an enhancer of endometrial *TAP2* expression [44]

Abstract

Variation in female reproductive traits such as fertility, fecundity, and fecundability are heritable in humans, but identifying and functionally characterizing genetic variants associated with these traits has been challenging. Here we explore the functional significance and evolutionary history of a C/T polymorphism of SNP rs2071473, which we have previously shown is an eQTL for the *antigen peptide transporter 2* (*TAP2)* gene and significantly associated with fecundability. We replicated the association between rs2071473 genotype and *TAP2* expression using GTEx data and demonstrate that *TAP2* is expressed by decidual stromal cells at the maternal-fetal interface. Next, we show that rs2071473 is located within a progesterone responsive cis-regulatory element that functions as a repressor with the T allele and an enhancer with the C allele. Remarkably, we found this polymorphism arose before the divergence of modern and archaic humans, is segregating at intermediate to high frequencies across human populations, and has genetic signatures of long-term balancing selection. This variant has also previously been identified in GWA studies of immune related disease, suggesting both alleles are maintained due to antagonistic pleiotropy.

Introduction

Variation in female reproductive traits, such as age of menarche and menopause, fertility, age at first and last birth, fecundity, and fecundability are heritable in humans [22,23]. However, identifying the genetic bases for variation in most of these traits has proven challenging because of limited sample sizes, strong gene-environment interactions [22-24], widespread contraceptive use, and significant clinical heterogeneity among infertile couples. For example, while genome-wide association studies (GWAS) have identified loci associated with age of menarche and menopause [54-59], and age at first and last birth [60], few studies have successfully identified genetic variants associated with fertility, fecundity, and fecundability [24,60-62].

An integrated expression quantitative trait locus (eQTL) mapping and association study was recently performed to identify eQTLs in mid-secretory endometrium that influence pregnancy outcomes in a prospective study of Hutterite women [42]. Among the 189 eQTLs identified, two were also associated with fecundability (probability of pregnancy in one reproductive cycle) [42]. The most significant association was rs2071473 ($P = 1.3x10^{-4}$), an eQTL associated with expression of the *antigen peptide transporter 2* (*TAP2)* gene (MIM: 170261) in the HLA class II region. The C allele of rs2071473 was associated with longer intervals to pregnancy and higher expression of the *TAP2* gene in mid-secretory phase (receptive) endometrium. The median time to pregnancy, for example, was 2.0, 3.1, and 4.0 months among women with the TT, CT, and CC genotypes, respectively [42].

An essential step in implantation is the establishment of receptivity by the hormone-primed endometrium. This 'window of implantation' occurs in the mid-secretory phase of the menstrual cycle [21], after endometrial stromal fibroblasts (ESFs) have differentiated (decidualized) into decidual stromal cells (DSCs) in response to progesterone and cyclic-AMP (cAMP) [18,19].

Decidualization underlies a suite of molecular, cellular, and physiological responses that support pregnancy including maternal immunotolerance of the fetal allograft [19,20]. Although function of TAP2 in the decidualized endometrium and the process of implantation have not been elucidated, TAP2 plays an integral role in translocating peptides from the cytosol to MHC class I molecules in the endoplasmic reticulum [43]. These data suggest that TAP2-dependent antigen processing and presentation by DSCs plays a role in establishing receptivity to implantation and immunotolerance at the maternal-fetal interface, likely by modifying the interactions between DSCs and immune cells in the decidualized endometrium [27].

Here we explore the functional significance and evolutionary history of the rs2071473 C/T polymorphism. We first replicate the association between rs2071473 genotype and *TAP2* expression using GTEx data and demonstrate that *TAP2* is expressed by DSCs at the maternal-fetal interface. Next, we show that rs2071473 is located within a cAMP/progesterone responsive regulatory element and disrupts a putative DDIT3 (CHOP10) binding site that functions as a repressor with the rs2071473 T allele and an enhancer with the C allele. Remarkably, we found that the C/T polymorphism arose before the divergence of modern and archaic humans, is segregating at intermediate to high frequencies across human populations, and has genetic signatures of long-term balancing selection.

Materials and Methods

rs2071473 is a multi-tissue eQTL for TAP2, HLA-DOB, and HLA-DRB6

I replicated the association between the T/C polymorphism at rs2071473 and *TAP2* expression levels using GTEx Analysis Release V6 (dbGaP Accession phs000424.v6.p1) data for

35 tissues including the uterus [63,64]. I also used GTEx data to identify other genes for which rs2071473 was an eQTL. Briefly, I queried the GTEx database using the 'Single tissue eQTLs search form' for SNP rs2071473.

TAP2 is produced by decidual stromal cells at the maternal-fetal interface

To determine the cell-type localization of TAP2 in the endometrium, data from the Human Protein Atlas immunohistochemistry collection [65] was used for endometrium, placenta, and decidua. I examined the expression of *TAP2* in the endometrium across the menstrual cycle [66], in mid-secretory phase endometrial biopsies from women not taking hormonal contraceptives (n=11), and women using either the progestin-based contraceptives depot medroxyprogesterone acetate (DMPA) or levonorgestrel intrauterine system (LNG-IUS) for at least 6 months [67], and in DSCs treated with a PGR-specific siRNA or a non-targeted siRNA [68] using previously generated microarray expression data. These microarray datasets were analyzed with the GEO2R analysis package, which implements the GEOquery [69] and limma R packages [70,71] from the Bioconductor project to quantify differential gene expression. I also examined *TAP2* expression in RNA-Seq data previously generated from ESFs treated with control media or differentiated (decidualized) with cAMP/MPA into DSCs [53,72].

rs2071473 is located within a cAMP/progesterone responsive regulatory element and disrupts putative DDIT3 binding

A 1000bp region spanning 50bp upstream of rs2071473 to the 3'-end of the PGR ChIP-Seq peak was cloned into the pGL3-Basic luciferase vector (Promega), once with the T allele and once with the C allele (Genscript). A pGL3-Basic plasmid without the 1kb rs2071473 insert was

used as a negative control. DDIT3 and PGR expression vectors were also obtained from Genscript. Endometrial stromal fibroblasts (ATCC CRL-4003) immortalized with telomerase were maintained in phenol red free DMEM (Gibco) supplemented with 10% charcoal stripped fetal bovine serum (CSFBS; Gibco), 1x ITS (Gibco), 1% sodium pyruvate (Gibco), and 1% L-glutamine (Gibco). Confluent ESFs in 96 well plates in 80μl of Opti-MEM (Gibco) were transfected with 100ng of the luciferase plasmid, 100ng of DDIT3 and/or PGR as needed, and 10ng of pRL-null with 0.1μl PLUS reagent (Invitrogen) and 0.25μl of Lipofectamine LTX (Invitrogen) in 20μl Opti-MEM. The cells incubated in the transfection mixture for 6hrs and the media was replaced with the phenol red free maintenance media overnight. Decidualization was then induced by incubating the cells in the decidualization media (DMEM with phenol red (Gibco), 2% CSFBS (Gibco), 1% sodium pyruvate (Gibco), 0.5mM 8-Br-cAMP (Sigma), and 1μM medroxyprogesterone acetate (Sigma)) for 48hrs. Decidualization controls were incubated in the decidualization control media (phenol red free DMEM (Gibco), 2% CSFBS (Gibco), and 1% sodium pyruvate (Gibco)) instead for 48hrs. After decidualization, Dual Luciferase Reporter Assays (Promega) were started by incubating the cells for 15mins in 20μl of 1x passive lysis buffer. Luciferase and renilla activity were then measured using the Glomax multi+ detection system (Promega). Luciferase activity values were standardized by the renilla activity values and background activity values as determined by measuring luminescence from the pGL3-Basic plasmid with no insert.

The rs2071473 C allele is derived in humans and segregating at intermediate to high frequencies across multiple human populations

To reconstruct the evolutionary history of the T/C polymorphism, a region spanning 50bp

upstream and downstream of rs2071473 from hg19 (chr6:32814778-32814878) was used as a query sequence to BLAT search the chimpanzee (CHIMP2.1.4), gorilla (gorGor3.1), orangutan (PPYG2), gibbon, (Nleu1.0), rhesus monkey (MMUL_1), hamadryas baboon (Pham_1.0), olive baboon (Panu_2.0), vervet monkey (ChlSab1.0), marmoset (C_jacchus3.2.1), Bolivian squirrel monkey (SalBol1.0), tarsier (tarSyr1), mouse lemur (micMur1), and galago (OtoGar3) genomes. For all other non-human species, the same 101bp region was used as a query for SRA-BLAST against high-throughput sequencing reads deposited in SRA. The top scoring 100 reads were assembled into contigs using the 'Map to reference' option in Geneious v6.1.2 and the human sequence as a reference. Sequences for the Altai Neanderthal, Denisovan, Ust-Ishim, and two aboriginal Australians were obtained from the 'Ancient Genome Browser. The frequency of the C/T allele across the Human Genome Diversity Project (HGDP) populations was obtained from the 'Geography of Genetic Variants Browser'.

Ancestral sequences of the 101bp region were inferred using the ancestral sequence reconstruction (ASR) module of Datamonkey [73] which implements joint, marginal, and sampled reconstruction methods [74], the nucleotide alignment of the 101bp, the best fitting nucleotide substitution model (HKY85), a general discrete model of site-to-site rate variation with 3 rate classes, and the phylogeny shown in **Fig 5A**. All three ASR methods reconstructed the same sequence for the ancestral human sequence at 1.0 support

rs2071473 has signatures of balancing selection in humans and diversifying selection in primates

To infer if there was evidence for positive selection acting on the derived T allele, an improved version of the EvoNC method [75] was used that has been modified into a branch-site

model [76,77], and implemented in HyPhy v2.22 [77,78]. This method utilizes the MG94HKY85 nucleotide substitution model, which was also the best-fitting nucleotide substitution model for our target non-coding region and neutral rate proxy, and includes 10 replicate likelihood searches. Although this implementation of the EvoNC method that has been modified into a branch-site model capable of identifying positively selected sites in a priori defined lineages using Naive Empirical Bayes (NEB) and Bayes Empirical Bayes (BEB), previous studies have shown that these methods have high type-I and type-II error rates. Therefore, evidence was inferred for without reference to NEB or BEB identified sites and instead relied on a significant likelihood ratio test between the null and alternate models. An alignment of the 101bp region described above was analyzed, the phylogeny shown in **Fig 5A**, and synonymous sites from the flanking *HLA-DOB* and *TAP2* genes which were identified from the same primate species using the method described above. F_{st}, Tajima's D, and Fay and Wu's H data for CEU and YRI populations were obtained from the 1000 Genomes Selection Browser 1.0 [79].

Results

rs2071473 is a multi-tissue eQTL for *TAP2*, *HLA-DOB*, and *HLA-DRB6*

A T/C polymorphism at rs2071473 has been previously shown to be an eQTL for *TAP2* in mid-secretory phase endometrium [42]. To replicate this observation in an independent cohort and in additional tissues, I tested if rs2071473 was correlated with *TAP2* expression using GTEx data. Similarly to the previous observation, I found that rs2071473 was an eQTL for *TAP2* in GTEx uterus samples (n=70, β=-1.1, P=9.1×10^{-11}; **Fig 2A/B**) as well as 34 other tissues (**Fig 2C**); the largest effect size was observed in the uterus (**Fig 2C**). I also used GTEx data to

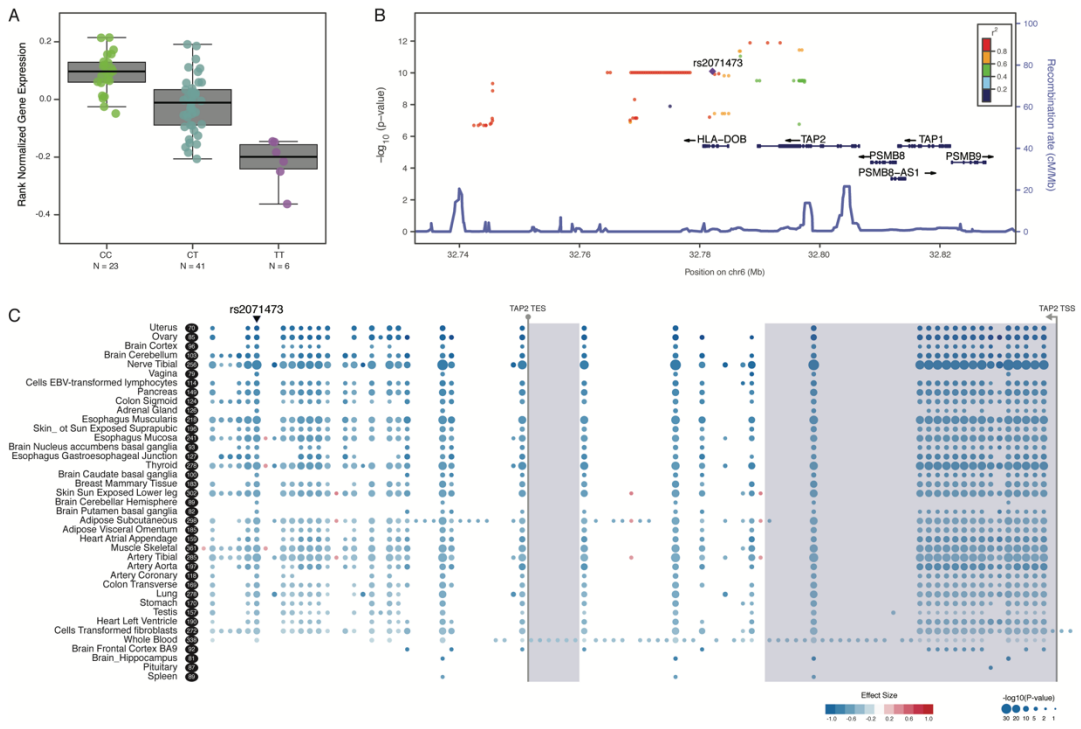

Fig. 2 Replication of the rs2071473 C/T polymorphism as an eQTL for *TAP2*. (**A**) Uterine *TAP2* expression boxplots from GTEx data. Numbers below each boxplot are the number of individuals in each genotype. (**B**) Region association plot showing GTEx SNPs that are significant eQTLs for *TAP2* expression ($-\log_{10}$ P-value, left y-axis), color-coded based on their LD with rs2071473 (purple diamond). Local recombination rates are shown on the right y-axis. (**C**) Gene-eQTL Visualizer plot showing significant eQTLs for *TAP2* (see inset key for ($-\log_{10}$ P-value and effect size coding) across 39 GTEx tissues. Tissues are sorted by largest effect size top to bottom. Numbers in black ovals indicate sample sizes. The location of *TAP2* coding exons are shown with light blue shading, and the start site (TSS) and end site (TES) are indicated.

identify other genes for which rs2071473 was an eQTL and found that it was an eQTL for HLA-DOB in 25 tissues (β=-0.27 – -0.91, P=5.4×10^{-9} - 9.1×10^{-20}) and for HLA-DRB6 in four tissues (β=0.37 - 0.46, P=6.7×10^{-6} – 2.1×10^{-8}). However, rs2071473 was not identified as a uterine eQTL for HLA-DOB or HLA-DRB6 because neither gene is expressed in GTEx uterine tissues.

***TAP2* is expressed by decidual stromal cells at the maternal-fetal interface**

Although *TAP2* is expressed in uterine tissues, the mid-secretory phase endometrium is a complex tissue composed of numerous cell-types including perivascular mesenchymal stem-like cells [20,80], endometrial stromal fibroblasts (ESFs) [19], decidual stromal cells (DSCs) [18,19], luminal and glandular epithelial cells, endothelial cells lining blood vessels, uterine natural killer cells (uNK) [81], uterine macrophage (uMP) [82,83], multiple populations of T-cells [84-87], and dendritic cells [88,89], among many others. To determine which cell-types produce TAP2 we examined its localization in endometrial biopsies from the Human Protein Atlas immunohistochemistry collection. TAP2 staining was found to be localized primarily to the luminal and glandular epithelium and ESFs in non-pregnant endometrium (**Fig 3A**), was particularly intense in DSCs in pregnant endometrium (**Fig 3B**), and was absent from trophoblast cells (**Fig 3B/C**). Thus, TAP2 is produced by maternal cells, particularly DSCs, at the maternal-fetal interface.

Next I tracked the expression of *TAP2* in the endometrium across the menstrual cycle using previously generated microarray expression data [66]. I found that *TAP2* expression reached its peak during the early secretory phase of the menstrual cycle and rapidly decreased in the mid-secretory phase (**Fig 3D**), suggesting that down-regulation of *TAP2* is regulated by progesterone and associated with endometrial receptivity to implantation. To infer if *TAP2* expression in the endometrium is regulated by progesterone I took advantage of an existing gene expression dataset of mid-secretory phase endometrial biopsies from women not taking hormonal contraceptives (n=11), and women using either the progestin-based contraceptives depot medroxyprogesterone acetate (DMPA) or levonorgestrel intrauterine system (LNG-IUS) for at least 6 months [67]. Consistent with regulation by progesterone, I found that *TAP2* expression was significantly lower in the endometria of women taking DMPA (P=0.002) or LNG-IUS (P=0.012)

compared to controls (**Fig 3E**).

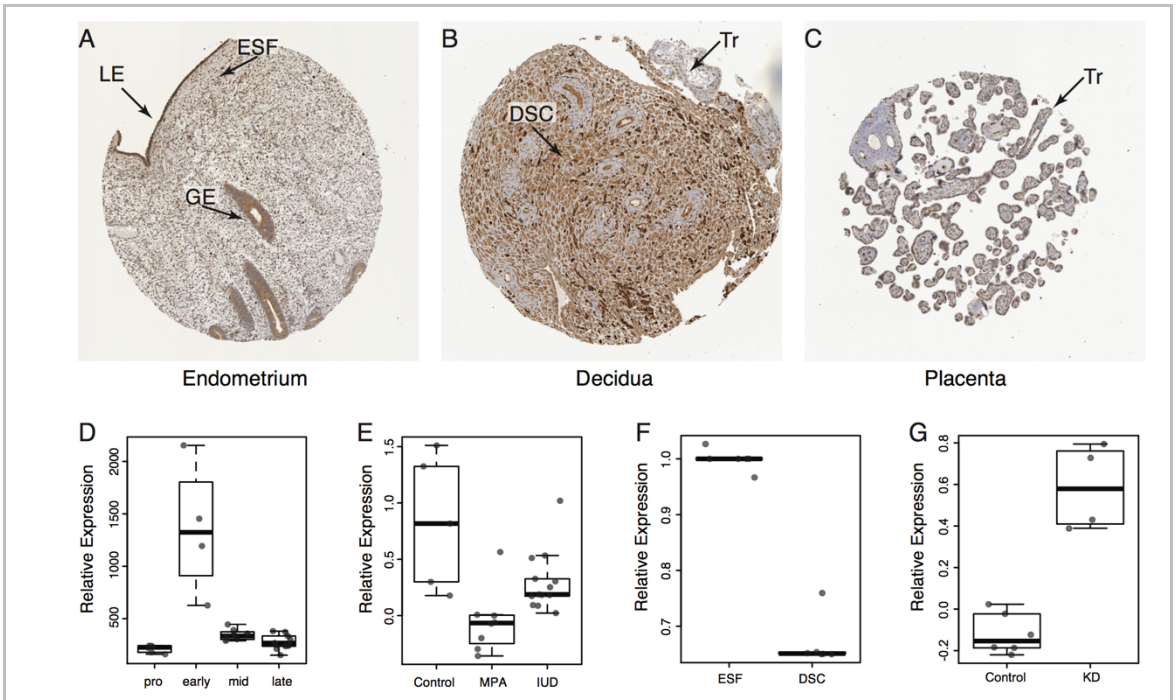

Fig. 3 *TAP2* is regulated by progesterone and is expressed by decidual stromal cells at the maternal-fetal interface. (**A**) Immunohistochemistry staining showing TAP2 localization in the non-pregnant endometrium. LE, luminal epithelium. GE, glandular epithelium. ESF, endometrial stromal fibroblast. (**B**) Immunohistochemistry staining showing TAP2 localization in the pregnant endometrium (decidua). DSC, decidual stromal cell. Tr, trophoblast (placenta). (**C**) Immunohistochemistry staining showing TAP2 localization in the placenta. Tr, trophoblast. (**D**) Relative expression of *TAP2* across the menstrual cycle. (**E**) Relative expression of *TAP2* in the endometrium of women using either the progestin-based contraceptives depot medroxyprogesterone acetate (MPA) or levonorgestrel intrauterine system (IUD). (**F**) Relative expression of *TAP2* in endometrial stromal fibroblasts treated with control media (ESFs) or differentiated into DSCs with cAMP/MPA for 48 hours. (**G**) Relative expression of *TAP2* in cAMP/MPA differentiated DSCs treated with a control siRNA or a PGR-specific siRNA (KD).

To directly test if *TAP2* is regulated by progesterone, I examined its expression in RNA-Seq data from ESFs treated with control media or differentiated (decidualized) with cAMP/MPA into DSCs [53,72] and found that *TAP2* was down-regulated ~33% by cAMP/MPA treatment (**Fig 3F**).

To test if these effects were mediated by the progesterone receptor (PGR), I used a previously published dataset to compare *TAP2* expression in DSCs treated with a PGR-specific siRNA or a non-targeted siRNA [68] and found that knockdown of PGR significantly up-regulated *TAP2* expression in DSCs (P=0.01; **Fig. 3G**). Thus I conclude that *TAP2* is down-regulated by progesterone during the differentiation of ESFs into DSCs and in the endometrium during the period of endometrial receptivity to implantation.

The rs2071473 variant switches a repressor into an activator

The observations that rs2071473 is an eQTL for *TAP2* and that *TAP2* expression is down-regulated by progesterone suggests rs2071473 may be located within or linked to a progesterone responsive enhancer. To identify such a regulatory element I used previously obtained ChIP-Seq data from DSCs for the transcription factors PGR [53,90,91] and NR2F2 (COUP-TFII) [92], which regulate the transcriptional response to progesterone and immune genes, respectively, H3K27ac which marks active enhancers, H3K4me3 which marks active promoters [53], and DNaseI-Seq and FAIRE-Seq to identify regions of open chromatin [53]. I found that rs2071473 was located 260bp upstream of PGR and NR2F2 binding sites, within local DNaseI and FAIRE peaks, and in a region of elevated H3K4me3 and H3K27ac signal (**Fig 4A**) suggesting this region is a progesterone responsive *cis*-regulatory element.

To test if this locus has regulatory potential I synthesized a 1000bp region spanning 50bp upstream of rs2071473 to the 3'-end of the PGR ChIP-Seq peak (**Fig 4A**) with either the reference C allele or the alternate T allele and cloned them into the pGL3-Basic luciferase reporter vector, which lacks both an endogenous promoter and enhancer. Next I transiently transfected either the pGL3Basic-rs2071473T or pGL3Basic-rs2071473C luciferase reporter

along with the pRL-null internal control vector into ESFs and DSCs and quantified luciferase and renilla activity using a dual luciferase assay.

I found that luciferase activity was significantly lower in ESFs (Wilcoxon test $P=1.75\times10^{-11}$) and DSCs (Wilcoxon test $P=2.87\times10^{-9}$) transfected with the pGL3Basic-rs2071473T reporter compared to empty pGL3Basic vector controls (**Fig 4B**). In stark contrast, luciferase activity was significantly higher in ESFs (Wilcoxon test $P=1.42\times10^{-9}$) and DSCs (Wilcoxon test $P=2.53\times10^{-13}$) transfected with the pGL3Basic-rs2071473C reporter compared to controls (**Fig 4B**). The difference in luciferase activity between the pGL3Basic-rs2071473T and pGL3Basic-rs2071473C was also significant in ESFs (Wilcoxon test $P=2.50\times10^{-12}$) and DSCs (Wilcoxon test $P=1.76\times10^{-11}$). Luciferase activity was significantly induced upon differentiation of ESFs to DSCs by cAMP/MPA treatment in both the pGL3Basic-rs2071473T (Wilcoxon test $P=0.02$) and pGL3Basic-rs2071473C (Wilcoxon test $P=2.53\times10^{-13}$) transfected cells (**Fig 4B**). Thus I conclude that the locus in which rs2071473 resides is a progesterone responsive *cis*-regulatory element and that the T/C polymorphism switches an enhancer into a repressor.

The rs2071473 C allele likely disrupts a DDIT3 binding site

The T/C polymorphism at rs2071473 is several hundred base pairs away from the PGR and NR2F2 binding sites and is therefore unlikely to directly affect their binding (**Fig 4A**). However, my luciferase assay results indicate that the T/C polymorphism has regulatory effects, suggesting this polymorphism disrupts a binding site for a transcriptional repressor or creates a binding site for a transcriptional activator. To infer which of these scenarios was most likely I identified putative transcription factor binding sites in a 25bp window upstream and downstream of rs2071473 using ConSite [93] and JASPAR transcription factor binding site profiles [94]. I found that

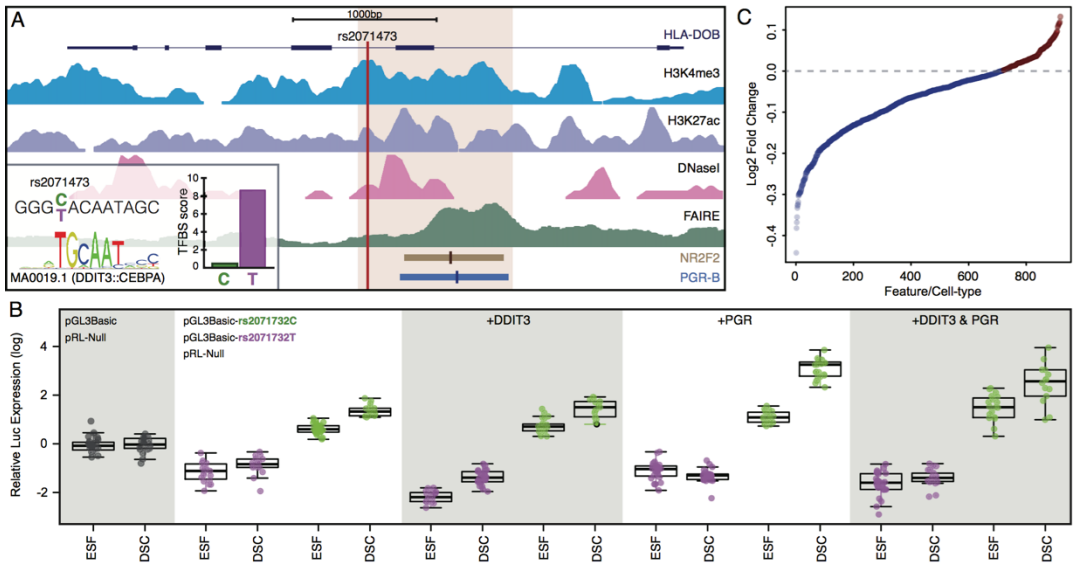

Fig. 4 The C/T polymorphism at rs2071473 is located within a progesterone responsive *cis*-regulatory element in decidual stromal cells. (**A**) Location of rs2071473 with respect to histone modifications that characterize promoters (H3K4me4 ChIP-Seq), enhancers (H3K27ac ChIP-Seq), open chromatin (DNaseI-Seq, FAIRE-Seq), as well as PGR and NR2F2 ChIP-Seq binding sites. Inset, the rs2071473-C allele is predicted to disrupt a DDIT3 binding site. The 1kb region shown in light brown was **Fig 4 (cont.)** cloned into the pGL3Basic luciferase reporter vector for functional characterization. (**B**) Luciferase assay results testing the regulatory potential of the T (pGL3Basic-rs2071473T) and C (pGL3Basic-rs2071473C) alleles in endometrial stromal fibroblasts treated with control media (ESFs) or differentiated into DSCs with cAMP/MPA for 48 hours. Data are shown as luciferase activity from the pGL3Basic-rs2071473C or pGL3Basic-rs2071473T reporter relative to renilla activity (pRL-null and empty vector (pGL3Basic) controls. +DDIT3, Relative luciferase activity from the pGL3Basic-rs2071473C or pGL3Basic-rs2071473T in ESFs and DSCs co-transfected with DDIT3. +PGR, Relative luciferase activity from the pGL3Basic-rs2071473C or pGL3Basic-rs2071473T in ESFs and DSCs co-transfected with PGR. +DDIT3 & PGR, Relative luciferase activity from the pGL3Basic-rs2071473C or pGL3Basic-rs2071473T in ESFs and DSCs co-transfected with DDIT3 and PGR. (**C**) DeepSea predicted effects (log_2 fold change) of rs2071473 on regulatory features across cell-types.

the T/C polymorphism occurs at an invariant T site in the DDIT3 motif (**T**GCAAT), which was predicted to abolish DDIT3 binding (**Fig 4A, inset**). Similarly, the T/C polymorphism was predicted by DeepSea [95], a deep learning-based algorithm that infers the effects of single nucleotide substitutions on chromatin features such as transcription factors binding, DNaseI

sensitivity, and histone marks, to generally have a negative effects on regulatory functions across multiple cell-types (**Fig 4C**).

While DDIT3 (also known as CHOP10 and GADD153) was initially characterized as a dominant negative inhibitor of CEBP family transcription factors [96], it can also function as a transcription factor [97,98] and is transcriptionally regulated by cAMP [99]. Indeed I found that *CHOP10* expression was down-regulated by cAMP/MPA in our RNA-Seq data from ESFs (TPM=187.68) and DSCs (TPM=61.20). These data suggest that the T/C polymorphism may disrupt a DDIT3 binding site that mediates transcriptional repression of *TAP2*, unmasking a secondary enhancer function. To test this hypothesis I co-transfected ESFs and DSCs with a DDIT3 expression vector, the pRL-null internal control vector, and either the pGL3Basic-rs2071473T or pGL3Basic-rs2071473C luciferase reporter, and compared luciferase activity to control ESFs and DSCs. If DDIT3 mediates repression by binding the T allele, then co-transfection of DDIT3 with the pGL3Basic-rs2071473T reporter should augment repression whereas co-transfection with the pGL3Basic-rs2071473C reporter should be unaffected. Indeed, the addition of DDIT3 significantly reduced luciferase activity from the pGL3Basic-rs2071473T reporter in ESFs (Wilcoxon test $P=3.48\times10^{-10}$) and DSCs (Wilcoxon test P=1.591e-05) but had no effect on luciferase activity from the pGL3Basic-rs2071473C reporter in either ESFs or DSCs ($P>0.21$; **Fig 4B**).

To test if this regulatory element is progesterone-responsive, I co-transfected ESFs and DSCs with a PGR rather than a DDIT3 expression vector and repeated the luciferase assays described above. The addition of PGR did not affect luciferase activity from the pGL3Basic-rs2071473T reporter in ESFs (Wilcoxon test $P=0.67$) but did enhance repression in DSCs (Wilcox test $P=9.56\times10^{-5}$; **Fig 4B**). Co-transfection of PGR elevated luciferase activity from the pGL3Basic-

rs2071473C reporter in both ESFs (Wilcoxon test P=2.82×10^{-8}) and DSCs (Wilcoxon test P=2.53×10^{-13}; **Fig 4B**). Finally I tested whether DDIT3 and PGR acted co-operatively or antagonistically by co-transfecting both the DDIT3 and PGR expression vectors with either the pGL3Basic-rs2071473T or pGL3Basic-rs2071473C reporters into ESFs and DSCs. Consistent with a weakly antagonistic functional interaction, luciferase activity in DDIT3 and PGR transfected ESFs and DSCs was intermediate between luciferase activity in DDIT3 or PGR transfected cells (**Fig 4B**). These data suggest that the reference C allele likely disrupts a DDIT3 biding site, which unmasks a progesterone responsive enhancer.

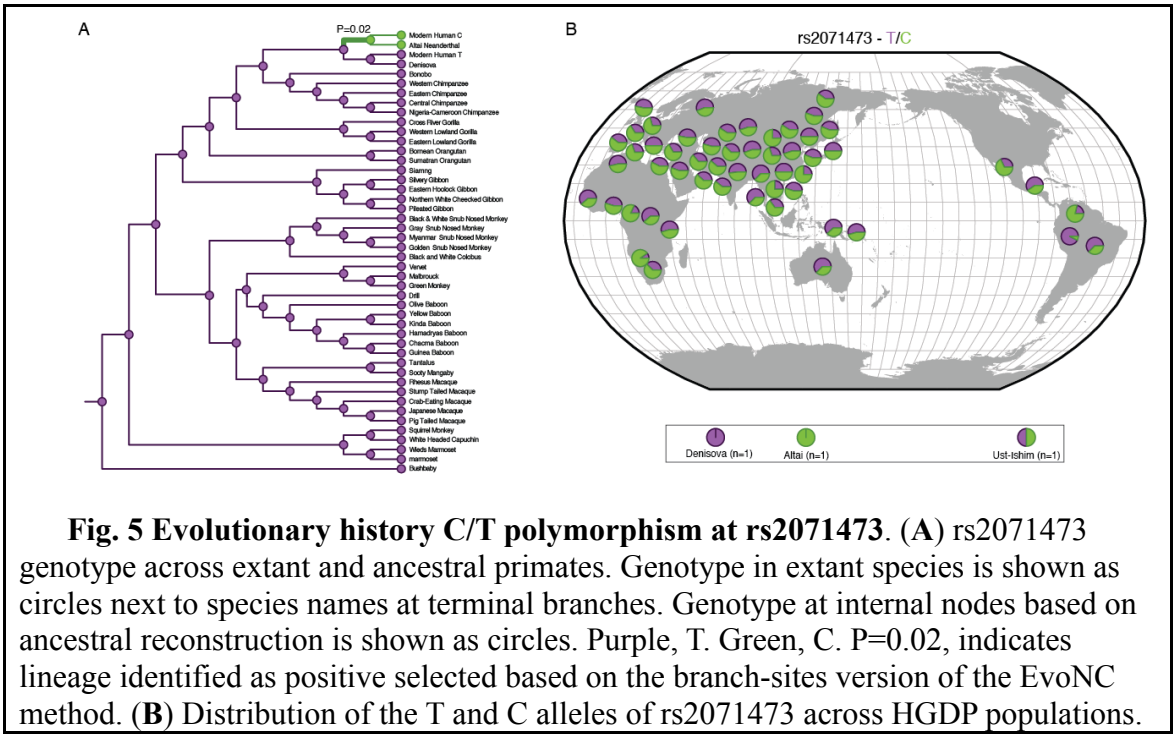

Fig. 5 Evolutionary history C/T polymorphism at rs2071473. (**A**) rs2071473 genotype across extant and ancestral primates. Genotype in extant species is shown as circles next to species names at terminal branches. Genotype at internal nodes based on ancestral reconstruction is shown as circles. Purple, T. Green, C. P=0.02, indicates lineage identified as positive selected based on the branch-sites version of the EvoNC method. (**B**) Distribution of the T and C alleles of rs2071473 across HGDP populations.

The rs2071473 C allele is derived in humans and has evidence of positive selection

To reconstruct the evolutionary history of the T/C polymorphism we identified a region spanning 50bp upstream and downstream of rs2071473 from 45 primates, including species from

each of the major primate lineages, multiple sub-species of African apes (Homininae), as well as modern and archaic (Altai Neanderthal and Denisova) humans. Next we used maximum likelihood methods to reconstruct ancestral sequences for this 101bp region. We found that the T allele was ancestral in primates and that the C variant was only found in modern human populations as well as the Altai Neanderthal genome (**Fig 5A**). Next we examined the frequency of these alleles across the Human Genome Diversity Project (HGDP) populations as well as two aboriginal Australians and found that the derived and ancestral alleles were segregating at intermediate to high frequencies in nearly all human populations (**Fig 5B**). These results indicate that the derived C variant at rs2071473 arose before the divergence of modern and archaic

Fig. 6 The rs2071473 C/T polymorphism has genetic signatures of balancing selection. Fay and Wu's H statistic, Tajima's D, Fst, and nucleotide diversity (Pi) for HGDP CEU (green) and YRI (red) populations across the HLA-region of chromosome 6. The location of rs2071473 is shown with a vertical red line. Linkage disequilibrium across this region is shown, the log odds score (LOD) with white diamonds indicate pairwise D' values less than 1 with no statistically significant evidence of LD (LOD < 2), light blue diamonds indicate high D' values (>0.99) with low statistical significance (LOD < 2), and light pink diamonds indicate high statistical significance (LOD >= 2) but the low D' (less than 0.5).

human lineages.

To test if the derived C allele may have been positively selected a branch-sites version of the EvoNC method [75-77] was used. In this analysis, the nucleotide substitution rate in noncoding regions (d_{nc}) is compared with a neutral rate proxy from either introns or synonymous substitutions (d_S) in nearby coding genes. The strength and direction of selection acting on non-coding regions is given by d_{nc}/d_S or ζ, which is analogous to the d_N/d_S rate or ω, with $\zeta = 1$ indicating neutral evolution, $\zeta < 1$ indicating negative (purifying) selection, and $\zeta > 1$ indicating positive selection. The 101bp region was analyzed as described above and synonymous sites from the flanking HLA-DOB and TAP2 genes, and fit three models to the data: 1) A null model that constrains $\zeta \leq 1$ across all sites and lineages; 2) An alternate model that allows for $\zeta \leq 1$ in foreground and background branches, $\zeta \leq 1$ in background branches and $\zeta > 1$ in the foreground branch; and 3) An alternate model that allows for $\zeta < 1$ in foreground and background branches, $\zeta = 1$ in foreground and background branches, $\zeta < 1$ in background and $\zeta > 1$ in foreground branches, and $\zeta = 1$ in background and $\zeta > 1$ in foreground branches. The null model was rejected in favor of both alternate model 1 (alternate model 1 LRT=5.37; P=0.02) and alternate model 2 (alternate model 2 LRT=5.38; P=0.02), suggesting the T to C substitution may have been positively selected.

rs2071473 has signatures of balancing selection in humans and diversifying selection in primates

The observation that the C allele originated before the divergence of modern and archaic humans and is segregating at intermediate frequencies across modern human populations suggests that the ancestral and derived variants may be maintained by balancing selection.

Previous studies have shown that balancing selection is common in the HLA region in which *TAP2* and rs2071473 are located [100-103]. DeGiorgio et al, for example, developed a model-based approach to identify signatures of ancient balancing selection and found that the *HLA-DOB* lcous, in which rs2071473 is located, was an outlier (top 0.5% of all scores across the genome) in their scan for balancing selection [104]. Consistent with the action of long-term balancing selection, rs2071473 has essentially no differentiation between populations (F_{st} = 0 - 0.016), and is located in a region with a relative excess of common polymorphisms across CEU and YRI populations as measured by Tajima's D (1.16-2.07), Fay and Wu's H (-23.16 – -56.83), and Pi (86.41-107.71) (**Fig 6**).

To test if the rs2071473 enhancer region evolved under positive diversifying selection across primates, the sites version of the EvoNC method [76,77,105] was used. Three models were fit to the alignment described above: 1) A null model that constrains $\zeta \leq 1$ across all sites; 2) An alternate model that allows for categories of sites with $\zeta \geq 1$ and $\zeta < 1$; and 3) An alternate model that allows for categories of sites with $\zeta < 1$, $\zeta = 1$, or $\zeta > 1$. The null model was not rejected in favor of alternate model 1 (LRT=0; *P*=1), however, alternative model 2 was a better fit to the data than either the null model LRT=11.76; *P*=0.019) or alternative model 1 (LRT=11.78; *P*=0.0005). These data indicate that including distinct rate classes for sites with $\zeta < 1$ or $\zeta = 1$ significantly improves the alternate model, and suggest that positive diversifying selection acts on sites in this regulatory element across primates.

Discussion

The mechanisms that promote maternal tolerance of the antigenically distinct fetus are complex [106] and have been the subject of intense study since Medawar formulated the immunological paradox posed by pregnancy and proposed tolerance was achieved by physical separation of maternal and fetal tissues, maternal immunosuppression, and immaturity of fetal antigens [16]. It is now clear that rather than being a site of maternal immunosuppression [107], the maternal immune system in the endometrium plays an active role in establishing a permissive environment for implantation, placentation, and gestation. For example, maternal regulatory T cells (Tregs) [85,108], shifts in the Th1/Th2/Th17 balance [86], uterine natural killer cells [81,109], uterine dendritic cells [88], uterine macrophage [110], and signaling by DSCs [111-113] all contribute to establishing immunotolerance and even promote placental invasion into maternal tissues.

It has also become clear that the major histocompatibility complex (MHC) genes, which play an important role in the rejection of non-self tissues, contribute to maternal tolerance of the fetus [27]. Matching of HLA antigens between couples, for example, is associated with longer intervals from marriage to birth compared to couples not matching for HLA [28,29]; these longer intervals result from both higher miscarriage rates among couples matching for class I HLA-B antigens [26] and longer intervals to pregnancy among couples matching for class II HLA-DR antigens [25]. Similarly, the non-classical HLA class I genes HLA-F are associated with fecundability whereas overwhelming evidence indicates maternal and fetal HLA-G genotypes are associated with miscarriage, recurrent pregnancy loss, and preeclampsia [30-41]. Collectively these data implicate antigen presentation by MHC as one of the molecular mechanisms that underlie successful implantation, maternal immunotolerance, and the establishment and maintenance of pregnancy.

The ATP-binding cassette transporter TAP, a heterodimer composed of TAP1 and TAP2,

translocates peptides from the cytosol to awaiting MHC class I molecules in the endoplasmic reticulum, which results in cell surface presentation of the trimeric MHC complex to immune cells such as T lymphocytes and natural killer cells [43]. Lost and reduced TAP accumulation leads to lost and reduced surface HLA accumulation [114-116], altered surface HLA repertoires [117], and the surface presentation of distinct antigenic peptides that are recognized by cytotoxic T lymphocytes [118]. Our observation that *TAP2* is highly expressed by DSCs at the maternal-fetal interface and that *TAP2* expression levels are associated with fecundability suggests that changes in TAP2 stoichiometry may alter MHC processing and thus interactions between DSCs and maternal immune cells. Indeed DSCs produce numerous HLA class I molecules, including HLA-G and HLA-F [119,120], which are transcriptionally up-regulated by progesterone during decidualization [113].

While the connection between TAP2 levels in DSCs and immune signaling are obvious, our observation that the ancestral and derived alleles of rs2071473 arose before the divergence of modern and archaic humans and has signatures of long-term balancing selection is unexpected. The HLA region has long been recognized to be under balancing selection [100,101,104,121,122]. These signals, however, are usually attributed to polymorphisms within the classical HLA class I genes rather than regulatory regions (cf. [123]). *TAP2* and rs2071473 are also located near the distal end of the HLA region and bounded by regions of high recombination, including a recombination hotspot within *TAP2* [124], and in a LD block that includes few other HLA genes. These data suggest that the signal of balancing selection at rs2071473 is distinct from other signals of balancing selection in the HLA region, but it is difficult to disentangle these signals given the relatively strong linkage across the HLA region. Thus it is possible that the C/T polymorphism at rs2071473 is not itself under balancing selection and is linked to the balanced site.

If the target of balancing selection is rs2071473, what selective forces are acting to maintain the ancestral and derived alleles? Balancing selection is generally attributed to heterozygote advantage (overdominance), frequency dependent selection, or antagonistic pleiotropy [125], which are all probable evolutionary scenarios to explain maintenance of the ancestral and derived alleles of rs2071473. Intriguingly, GWA studies have found rs2071473 genotype is associated with ulcerative colitis (MIM: 266600) [126], Crohn's disease (MIM: 266600) [127], and sarcoidosis (MIM: 609464) [128,129] in addition to fecundability. For example, the ancestral T allele is significantly associated with ulcerative colitis [126] and shorter time to pregnancy whereas the derived C allele is significantly associated with Crohn's disease [127] and longer time to pregnancy suggesting theses alleles may be maintained by antagonistic pleiotropy. These data are strong example of a reproduction-health tradeoff in human evolution.

Author Contributions

Katelyn Mika (KM) and Vincent Lynch (VL) collaboratively conceived, designed, interpreted, and wrote this paper, as well as conducted and analyzed the data on the evolutionary history of the rs2071473 101bp region. KM conducted and analyzed the GTEx replication, *TAP*2 expression across the menstrual cycle, *TAP2* expression in women with and without contraceptives, *TAP2* expression in DSCs treated with PGR siRNAs, *TAP2* expression in cultured ESFs and DSCs, and prepped and executed the luciferase assays. VL conducted and analyzed the localization of TAP2, the DeepSea analysis and all of the selection work.

Acknowledgments

This work was funded by a Burroughs Wellcome Preterm Birth Initiative grant (1013760), an NIH National Institute of General Medical Sciences Graduate Training Grant (T32GM007197), and by the March of Dimes Prematurity Research Center at UChicago-Northwestern-Duke. The authors thank C. Ober for comments on an earlier version of this manuscript.

Web Resources

GTEx database (http://www.gtexportal.org/home/eqtls/bySnp), Human Protein Atlas immunohistochemistry collection (www.proteinatlas.org), GEO2R analysis package (http://www.ncbi.nlm.nih.gov/geo/geo2r/), Ancient Genome Browser (http://www.eva.mpg.de/neandertal/draft-neandertal-genome.html), Geography of Genetic Variants Browser (http://popgen.uchicago.edu/ggv/), Datamonkey web-sever (http://www.datamonkey.org), Online Mendelian Inheritance in Man (http://www.omim.org)

Chapter 3:

An ancient fecundability-associated polymorphism creates a new GATA2 binding site in a distal enhancer of *HLA-F*

Abstract

Variation in female reproductive traits such as fertility, fecundity, and fecundability are heritable in humans, but identifying and functionally characterizing genetic variants associated with these traits has been challenging. Here we explore the functional significance and evolutionary history of a G/A polymorphism of SNP rs2523393, which we have previously shown is an eQTL for the *HLA-F* gene and significantly associated with fecundability. We replicated the association between rs2523393 genotype and *HLA-F* expression using GTEx data and demonstrate that *HLA-F* is up-regulated in the endometrium during the window of implantation and by progesterone in decidual stromal cells. Next, we show that the rs2523393 A allele creates a new GATA2 binding site in a progesterone responsive distal enhancer that loops to the *HLA-F* promoter. Remarkably, we found that the A allele is derived in the human lineage, that G/A polymorphism arose before the divergence of modern and archaic humans, and is segregating at intermediate to high frequencies across human populations. Remarkably, the derived A is also has been identified in a GWAS as a risk allele for multiple sclerosis. These data suggest that the polymorphism is maintained by antagonistic pleiotropy and a reproduction-health tradeoff in human evolution.

Introduction

Female reproductive traits such as fertility, fecundity, and fecundability are heritable in humans [22,23], however, identifying the genetic bases for these traits has been challenging [22-24]. Burrows *et al* [42] previously performed an integrated expression quantitative trait locus (eQTL) mapping and association study to identify eQTLs in mid-secretory endometrium that influence pregnancy outcomes. Among the eQTLs identified was a G/A polymorphism (rs2523393) that was significantly associated with *HLA-F* expression and fecundability [42]. Specifically, the G allele of rs2523393 was associated with longer intervals to pregnancy and lower expression of the *HLA-F* in mid-secretory phase (receptive) endometrium, whereas the A allele was associated with shorter intervals to pregnancy and higher *HLA-F* expression [42]. The median time to pregnancy, for example, was 2.3, 2.6, and 4.9 months among women with the AA, GA, and GG genotypes, respectively [42].

While the functions of HLA-F are enigmatic, it is thought to regulate immune responses and may play an important role in regulating maternal-fetal immunotolerance during pregnancy [130,131]. *HLA-F* is highly expressed, for example, in placental villi, on the surface of invasive and migratory extra-villus trophoblasts (EVTs), and decidual stromal cells (DSCs) [132-134]. Furthermore *HLA-F* expression increases during gestation, reaching a peak at term [132,135]. Like other MHC-I molecules, HLA-F binds natural killer (NK) cell receptors from the LIR and KIR families, including LIR1 and LIR2 [136,137], KIR2DS4 and KIR3DL2 [138], and KIR3DL1 and KIR3DS1 [137,139,140]. Uterine natural killer (uNK) cells, which are essential for the establishment and maintenance of maternal immunotolerance and spiral artery remodeling, express KIR3DL1 and LIR2 suggesting HLA-F expressed by EVTs and DSCs mediate interactions with uNK during implantation, trophoblast invasion, and establishment of the uteroplacental circulation.

HLA-F expression level is also positively correlated with uNK abundance in mid-luteal endometria and is predictive of achieving pregnancy [42,141] consistent an important role for *HLA-F* in the establishment of pregnancy.

Here we explore the functional and evolutionary history of the rs2523393 G/A polymorphism. We first replicate the association between rs2523393 genotype and *HLA-F* expression using GTEx data. We demonstrate that *HLA-F* expression increases during the menstrual cycle and is up-regulated by progesterone in decidual stromal cells (DSCs). Next, we show that rs2523393 is located within a cAMP/progesterone responsive enhancer that makes long range regulatory interactions with the promoters of *HLA-F*, and that the A allele creates a new binding site for the progesterone receptor (PGR) co-factor GATA2. Remarkably the G/A polymorphism arose before the divergence of modern and archaic humans and is segregating at intermediate to high frequencies across human populations and associated with a predisposition to several diseases. These data suggest the G/A polymorphism is maintained by antagonistic pleiotropy and strongly suggest there is a reproduction-health tradeoff in human evolution.

Materials and Methods

HLA-F expression in previously described datasets

Using previously generated microarray expression data, I examined the expression of *HLA-F* in the endometrium across the menstrual cycle (GSE4888 [66]) and in endometrial samples from fertile women (n=5), woman had implantation failure following IVF (n=5), and woman with recurrent spontaneous abortion (n=5) (GSE26787 [142]). These microarray datasets were analyzed with the GEO2R analysis package, which implements the GEOquery [69] and limma R packages

[70,71] from the Bioconductor project to quantify differential gene expression. I also examined *HLA-F* expression in RNA-Seq data previously generated from ESFs treated with control media or differentiated (decidualized) with 100nM medroxyprogesterone acetate, and 1mM 8-bromo-cAMP (all from Sigma-Aldrich Co., St. Louis, MO) into DSCs [53,72].

GATA2 siRNA knockdown and subsequent microarray analysis

Three ESF subcultures were transfected with a GATA2 specific siRNA and treated with decidual media (see above) for 3 days. Total RNA was extracted using the Qiagen RNeasy RNA isolation kit (Qiagen). The RNA from 3 replicates (wells) was pooled for each treatment per cell line. The integrity of all RNA samples was tested with the Bioanalyzer 2100 (Agilent Technologies). The concentration of RNA was quantified on the Nanodrop Spectrophotometer (Nanodrop Technologies). The samples with 260/280 greater than 1.8 were used for microarray hybridization. Microarrays were performed by the Genomic and RNA Profiling Core of Baylor College of Medicine using Affymetrix human genome U133 Plus 2.0 arrays (Affymetrix). Microarray CEL files were analyzed using dChip using the PM-MM model and quantile normalization. Combat was used to normalize differences and for batch correction [143]. Two-side t test followed by a Benjamini and Hochberg adjustment [144] and fold changes were used to define differentially expressed genes. Genes with an adjusted p-value≤0.1, and an absolute fold change 1.4 were considered significant. Raw and processed data are available in SRA and GEO: GSE108409. The probesets used for GATA2 and PGR were 209710_at and 221978_at respectively.

GATA2 ChIP-seq

ESFs isolated from 6 women were cultured separately in 150-mm culture dishes with ESF growth media and allowed to reach approximately 90% confluency before being treated with decidual media. Decidual media was changed every 48 hours. After 72 hours of treatment, cells were fixed for 15 minutes with 1/10 volume of freshly prepared formaldehyde solution (11% formaldehyde, 0.1 M NaCl, 1 mM EDTA, and 50 mM HEPES). The fixation was stopped by adding 1/20 volume 2.5 M glycine for 5 minutes. Fixed HESCs were collected and pelleted at 800 g for 10 minutes at 4°C. Cell pellets were washed 2 times with cold PBS, ESFs from 6 women were pooled before genomic DNA isolation. GATA2 immunoprecipitation and DNA library generation were performed by Active Motif as previously described [145]. DNA libraries were sequenced by Illumina's HiSeq Sequencing Service. 50-nucleotide sequence reads were mapped to the human genome (GRCh Build 37; February 2009) using the Burrows-Wheeler Aligner algorithm with default settings. Alignment information for each read was stored in the Sequence Alignment/Map or Binary version of the Sequence Alignment/Map format. Sequence alignments were extended in silico (using Active Motif software) at their 3 -ends to a length of 150 –250 bp and assigned to 32-nucleotide bins along the genome. The resulting histogram of fragment densities was stored in a binary analysis results file. Raw and processed data are available in SRA and GEO: GSE108409.

rs2523393 luciferase assays

A 1000bp region centered on rs2523393 (chr6:29705520-29706519) was synthesized with either the A or the G allele (Genscript) was cloned into the pGL3-Basic luciferase vector (Promega). A pGL3-Basic plasmid without the 1kb rs2523393 insert was used as a negative

control. GATA2 and PGR expression vectors were also obtained from Genscript. Endometrial stromal fibroblasts (ATCC CRL-4003) immortalized with telomerase were maintained in phenol red free DMEM (Gibco) supplemented with 10% charcoal stripped fetal bovine serum (CSFBS; Gibco), 1x ITS (Gibco), 1% sodium pyruvate (Gibco), and 1% L-glutamine (Gibco). Confluent ESFs in 96 well plates in 80μl of Opti-MEM (Gibco) were transfected with 100ng of the luciferase plasmid, 100ng of GATA2 and/or PGR as needed, and 10ng of pRL-null with 0.1μl PLUS reagent (Invitrogen) and 0.25μl of Lipofectamine LTX (Invitrogen) in 20μl Opti-MEM. The cells incubated in the transfection mixture for 6hrs and the media was replaced with the phenol red free maintenance media overnight. Decidualization was then induced by incubating the cells in the decidualization media: DMEM with phenol red (Gibco), 2% CSFBS (Gibco), 1% sodium pyruvate (Gibco), 0.5mM 8-Br-cAMP (Sigma), and 1μM of the progesterone analog medroxyprogesterone acetate (Sigma) for 48hrs. Decidualization controls were incubated in the decidualization control media (phenol red free DMEM (Gibco), 2% CSFBS (Gibco), and 1% sodium pyruvate (Gibco) instead for 48hrs. After decidualization, Dual Luciferase Reporter Assays (Promega) were started by incubating the cells for 15mins in 20μl of 1x passive lysis buffer. Luciferase and renilla activity were then measured using the Glomax multi+ detection system (Promega). Luciferase activity values were standardized by the renilla activity values and background activity values as determined by measuring luminescence from the pGL3-Basic plasmid with no insert. Each luciferase experiment was replicated in at least 4 independent experiments.

Allele conservation

To reconstruct the evolutionary history of the G/A polymorphism we used a region spanning

50bp upstream and downstream of rs2523393 from hg19 (chr6:32814778-32814878) as a query sequence to BLAT search the chimpanzee (CHIMP2.1.4), gorilla (gorGor3.1), orangutan (PPYG2), gibbon, (Nleu1.0), rhesus monkey (MMUL_1), hamadryas baboon (Pham_1.0), olive baboon (Panu_2.0), vervet monkey (ChlSab1.0), marmoset (C_jacchus3.2.1), Bolivian squirrel monkey (SalBol1.0), tarsier (tarSyr1), mouse lemur (micMur1), and galago (OtoGar3) genomes. For all other non-human species, we used the same 101bp region as a query for SRA-BLAST against high-throughput sequencing reads deposited in SRA. The top scoring 100 reads were assembled into contigs using the 'Map to reference' option in Geneious v6.1.2 and the human sequence as a reference. Sequences for the Altai Neanderthal, Denisovan, and Ust-Ishim were obtained from the 'Ancient Genome Browser. The frequency of the G/A allele across the Human Genome Diversity Project (HGDP) populations was obtained from the 'Geography of Genetic Variants Browser'. Ancestral sequences of the 101bp region were inferred using the ancestral sequence reconstruction (ASR) module of Datamonkey [73] which implements joint, marginal, and sampled reconstruction methods [74], the nucleotide alignment of the 101bp, the best fitting nucleotide substitution model (HKY85), a general discrete model of site-to-site rate variation with 3 rate classes, and the phylogeny shown in **Fig 12A**. All three ASR methods reconstructed the same sequence for the ancestral human sequence at 1.0 support.

Results

rs2523393 is a eQTL for *HLA-F*

It has been previously shown that the rs2523393 A/G polymorphism is an eQTL for *HLA-F* in mid-secretory phase endometrium [42]. To replicate this observation in an independent cohort

and in additional tissues, I tested if rs2523393 was correlated with *HLA-F* expression using GTEx Analysis Release V6 (dbGaP Accession phs000424.v7.p2) data for 35 tissues, including the 101 uterus samples [63,64]. I also used GTEx data to identify other genes for which rs2523393 was an eQTL. Briefly, I queried the GTEx database using the 'Single tissue eQTLs search form' for SNP rs2523393. Confirming our previous observation, we found that rs2523393 was an eQTL for *HLA-F* in the uterus (P=0.028) as well as 22 other tissues (**Fig 7**). I also used GTEx data to identify other genes for which rs2523393 was an eQTL and found that it was an eQTL

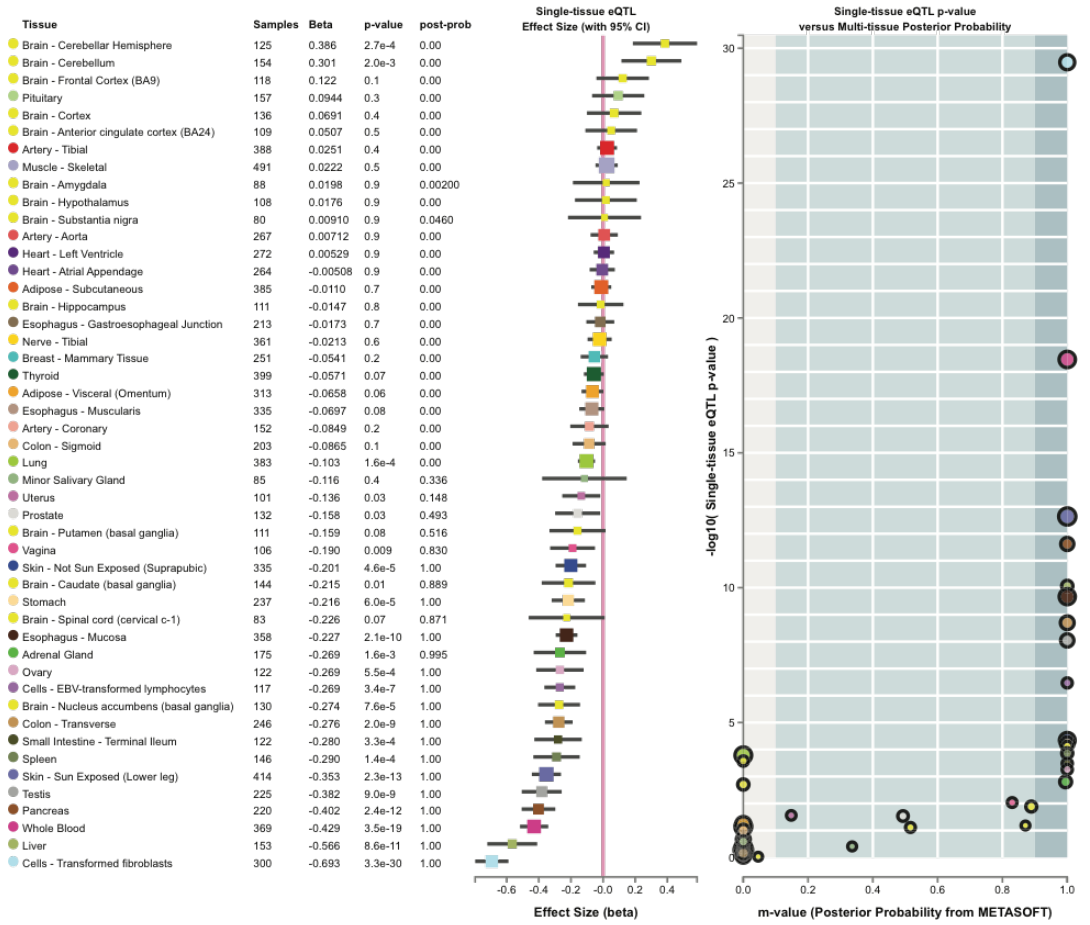

Fig. 7 GTEx multi-tissue eQTL plot for rs2523393. Left, Tissue type, sample size, beta, p-value, and posterior probability that rs2523393 in each tissue tested is also an eQTL across other tissues. Center, single-tissue effect size and 95% confidence interval. Right, single-tissue eQTL p-value versus multi-tissue posterior probability.

for 15 other genes, including *HLA-A* and *HLA-G*. The observation that rs2523393 was an eQTL for *HLA-A* and *HLA-G* in GTEx data prompted us to explore whether it was an eQTL for these genes in the mid-secretory endometrium data from the original study [42]. Indeed, rs2523393 was also found to be an eQTL for *HLA-G* (P=0.041), but not *HLA-A* in mid-secretory endometrium[42].

The rs2523393 G/A polymorphism occurs within a GATA2 motif

To infer the functional consequences of the G/A polymorphism DeepSea [95], a deep learning-based algorithm that infers the effects of single nucleotide substitutions on chromatin features such as transcription factors binding, DNase I sensitivity, and histone marks were used. DeepSea predicted the G allele would have a negative effect on the binding of GATA2 (log2 fold change effect: -2.07, E-value: 0.035) and GATA3 (log2 fold change effect: -1.00, E-value: 0.003) (**Fig 8A**). JASPAR transcription factor binding site (TFBS) motifs [94] were next used to identify

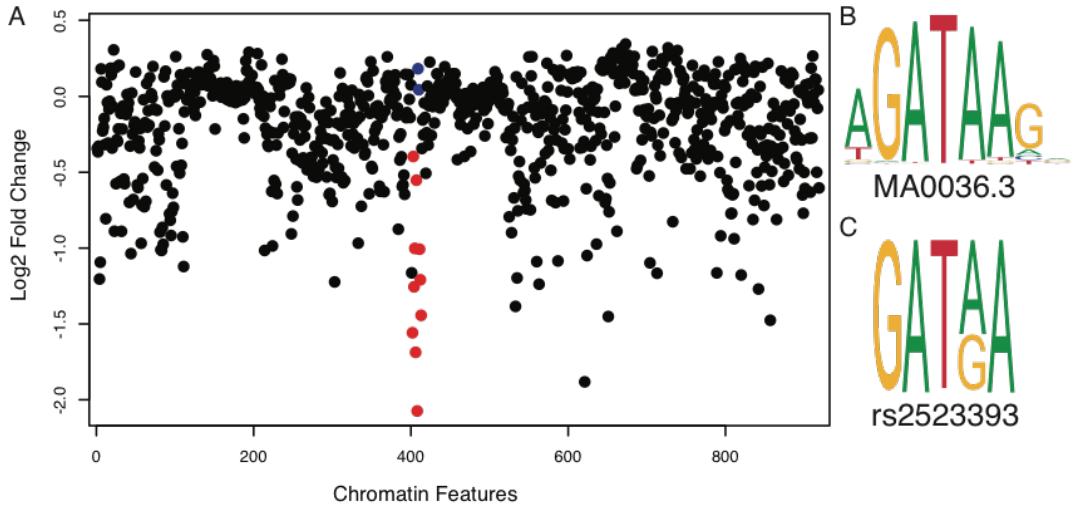

Fig. 8 The rs2523393 G allele is predicted to abolish a GATA2 binding site. (**A**) DeepSea plot showing predicted effects of the G allele on chromatin features. GATA1 binding, blue. GATA2 and GATA3 binding, red. (**B**) JASPAR GATA2 motif. (**C**) GATA2 motif at rs2523393.

putative TFBSs in a 36bp window upstream and downstream of rs2523393. A single TFBS was identified in this window, a GATA motif (matrix ID: MA0036.3, score: 8.70) and found that the G/A polymorphism occurs at an invariant A site in the motif (GATAA). Similar to the DeepSea results, substituting the reference A allele with the alternative G allele is predicted to abolish GATA2 binding (**Fig 8B/C**). These results suggest that the G/A polymorphism alters binding of GATA2, a transcription factor that plays an essential role in mediating the transcriptional response to progesterone in decidual stromal cells (DSCs) [146-148].

HLA-F is up-regulated by progesterone and GATA2 in decidual stromal cells

Previous studies have shown that HLA-F is expressed in placental villi, extra-villus trophoblasts (EVT), and DSCs [119,132-134]. To determine if the expression of *HLA-F* varies in the endometrium across the menstrual cycle, I used previously generated microarray expression data [66]. I found that *HLA-F* expression reached peak expression in the late secretory phase of the menstrual cycle, and was significantly differentially expressed throughout the menstrual cycle (**Fig 9A**). To test if *HLA-F* is up-regulated by progesterone, I examined its expression in RNA-Seq data from ESFs treated with control media or differentiated with cAMP/progesterone into DSCs treated with either a PGR specific siRNA or a scrambled control siRNA [53,68,72]. I found that *HLA-F* was up-regulated 2.11-fold ($P=3.59\times10^{-5}$; **Fig 9B**) in DSCs and that knockdown of PGR down-regulated *HLA-F* 1.67-fold ($P=6.15\times10^{-8}$; **Fig 9B**). Next I used an siRNA to knockdown the expression of GATA2 in DSCs and assayed global gene expression using an Agilent Human Gene Expression 8x60K microarray. Consentient with our previous results, I observed a 2.20-fold down-regulation of *HLA-F* (P=0.02, P adjusted = 0.25) upon GATA2 knockdown (**Fig 9C**). I also observed that *HLA-F* expression was lower in the endometria of

women with implantation failure (1.38-fold, P=1.54×10^{-3}) or recurrent spontaneous abortion (1.35-fold, 0.011) compared to fertile controls (**Fig 9D**). Thus I conclude that *HLA-F* expression increases as the menstrual cycle progresses, is up-regulated as by progesterone, PGR, and GATA2 during the differentiation of ESFs into DSCs, and dysregulation is associated with implantation failure and recurrent spontaneous abortion.

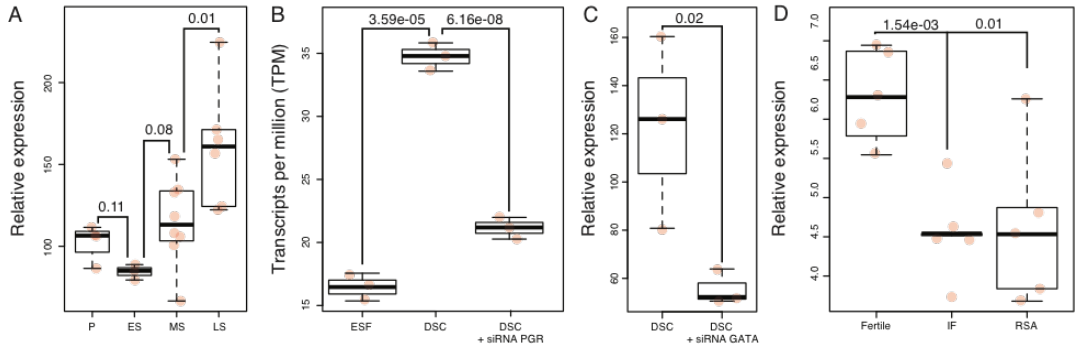

Fig. 9 *HLA-F* is regulated by upregulated progesterone and GATA2 during the menstrual cycle and in human decidual stromal cells. (**A**) Relative expression of *HLA-F* throughout the menstrual cycle. P, proliferative phase. ES, early secretory phase. MS, mid-secretory phase. LS, late secretory phase. N=3-8, T-test. (**B**) Expression level of *HLA-F* in endometrial stromal fibroblasts (ESF), decidual stromal cells (DSCs), and DSCs treated with an siRNA to knockdown PGR expression. (**C**) Relative expression level of *HLA-F* in DSCs and DSCs treated with an siRNA to knockdown GATA2 expression. N=3, T-test. (**D**) Relative expression level of *HLA-F* in the endometrium at the mid-luteal phase of a non-conceptual cycle for fertile women, and women with implantation failure (IF) or recurrent spontaneous abortion (RSA) N=5, T-test.

The rs2523393 A allele creates a new enhancer that loops to the *HLA-F* promoter

The observations that rs2523393 is an eQTL for *HLA-F*[42] and that progesterone and GATA2 up-regulate *HLA-F* in DSCs suggests rs2523393 may be located within or linked to a progesterone responsive enhancer. To identify such a regulatory element, our collaborators generated new GATA2 ChIP-Seq data from DSCs and we used previously published ChIP-Seq data from DSCs for the transcription factors PGR [53,90,91], FOXO1 [90], FOSL2 [90], and NR2F2

(COUP-TFII) [92], H3K27ac which marks active enhancers [53], H3K4me3 which marks active promoters [53], and DNaseI-Seq and FAIRE-Seq to identify regions of open chromatin [53]. Because rs2523393 is an eQTL for *HLA-F* in multiple tissues, including EBV transformed lymphocytes (LCLs), additional transcription factor binding sites were identified using ChIP-Seq data generated by ENCODE and promoter capture Hi-C (PCHi-C) data generated from LCLs. rs2523393 was found to be located in a region of open chromatin and within a GATA2 ChIP-Seq peak in DSCs, 170bp upstream of a FOSL2 ChIP-Seq peak in DSCs, and nearby a cluster of transcription factor binding sites in ENCODE data including a binding site for GATA2 (**Fig 10**). Finally, this region was found to loop to the *HLA-F* promoter in LCLs (as well as the *HLA-G* promoter) located ~14.5kb upstream of rs2523393, suggesting this region is an enhancer for *HLA-F*.

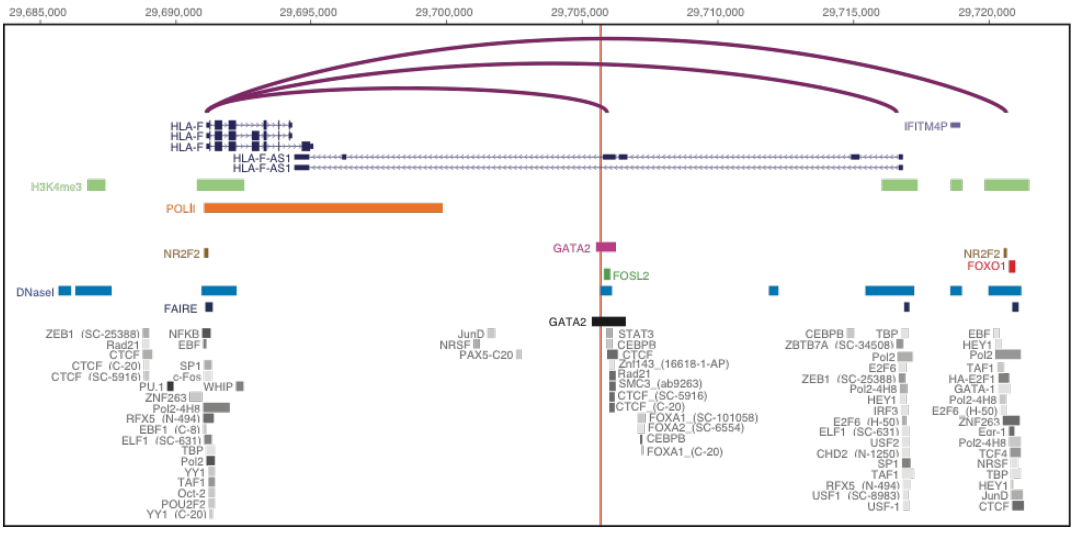

Fig. 10 The rs2523393 G/A polymorphism is located in a distal enhancer of *HLA-F*. Location of rs2523393 polymorphism (vertical red line) with respect to histone modifications that characterize promoters (H3K4me4 ChIP-Seq), enhancers (H3K27ac ChIP-Seq), open chromatin (DNaseI-Seq, FAIRE-Seq), as well as GATA2 and NR2F2 ChIP-Seq binding sites in human DSCs. ENCODE transcription factor binding sites are shown in grey and intrachromosomal loop (PCHiC) data from LCLs is shown in purple.

To test if this locus has regulatory potential I synthesized a 1000bp region spanning 500bp upstream and downstream of rs2523393 (**Fig 10**) with either the reference A allele or the alternate G allele and cloned them into the pGL3-Basic[minP] luciferase reporter vector, which lacks an enhancer but encodes a minimal promoter. Next I transiently transfected either the pGL3-Basic[minP]-rs2523393A or pGL3-Basic[minP]-rs2523393G luciferase reporter along

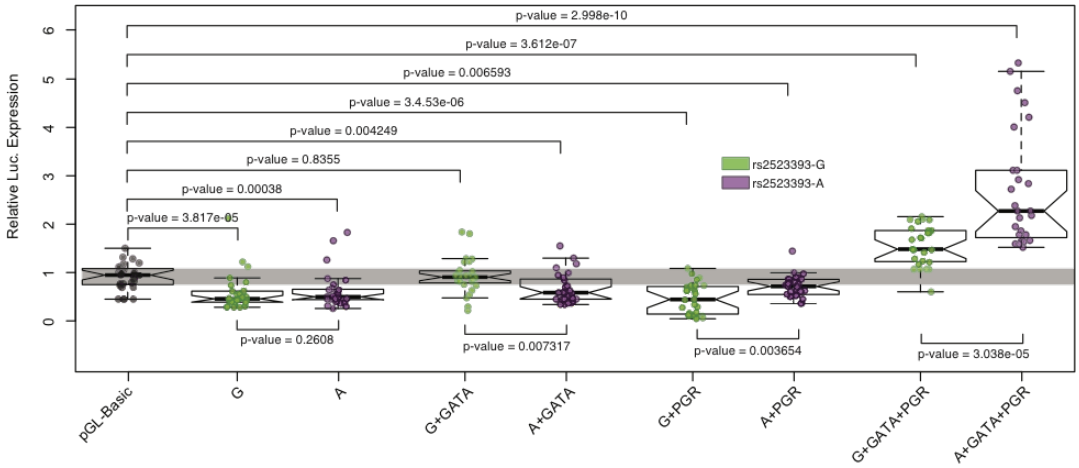

Fig. 11 The rs2523393 G/A polymorphism alters the function of a progesterone response enhancer. Luciferase assay results testing the regulatory potential of the G (pGL3Basic-rs2523393G) and A (pGL3Basic-rs2523393A) alleles in DSCs treated with cAMP/MPA for 48 hours. Data are shown as luciferase activity from the pGL3Basic-rs2523393G or pGL3Basic-rs2523393A reporter relative to renilla activity (pRL-null and empty vector (pGL3Basic) controls. +GATA, Relative luciferase activity from the pGL3Basic-rs2523393G or pGL3Basic-rs2523393A in DSCs co-transfected with GATA. +PGR, Relative luciferase activity from the pGL3Basic-rs2523393G or pGL3Basic-rs2523393A in DSCs co-transfected with PGR. +GATA & PGR, Relative luciferase activity from the pGL3Basic-rs2523393G or pGL3Basic-rs2523393A in DSCs co-transfected with GATA and PGR. n=14, p=Wilcoxon test.

with the pRL-null internal control vector into DSCs and quantified luciferase and renilla activity using a dual luciferase assay. I found that luciferase activity was significantly lower in DSCs transfected either pGL3-Basic[minP]-rs2523393A (Wilcoxon test $P=3.80\times10^{-4}$) or pGL3-Basic[minP]-rs2523393G (Wilcoxon test $P=3.82\times10^{-5}$) compared to empty vector pGL3Basic[minP] controls (**Fig 11**). Next, I co-transfected pGL3-Basic[minP]-rs2523393A or

pGL3-Basic[minP]-rs2523393G along either a GATA2 expression vector or a PGR expression vector. While co-transfection of neither the GATA2 expression vector nor the PGR expression vector enhanced luciferase expression above empty vector controls, there were significant differences between alleles (**Fig 11**). Finally I co-transfected pGL3-Basic[minP]-rs2523393A or pGL3-Basic[minP]-rs2523393G with both GATA2 and PGR expression vectors. Consistent with cooperative interaction between GATA2 and PGR, I observed a significant increase in luciferase expression from both the pGL3-Basic[minP]-rs2523393A (Wilcoxon test $P=3.00\times10^{-10}$) and pGL3-Basic[minP]-rs2523393G (Wilcoxon test $P=3.61\times10^{-7}$) vectors compared to empty vector controls. I also observed that the pGL3-Basic[minP]-rs2523393A vector drove significantly higher luciferase expression than the pGL3-Basic[minP]-rs2523393G vector (Wilcoxon test $P=3.65\times10^{-3}$). Thus I conclude that rs2523393 is located within a GATA2 binding site in a progesterone responsive enhancer, and that the G/A polymorphism affects enhancer function.

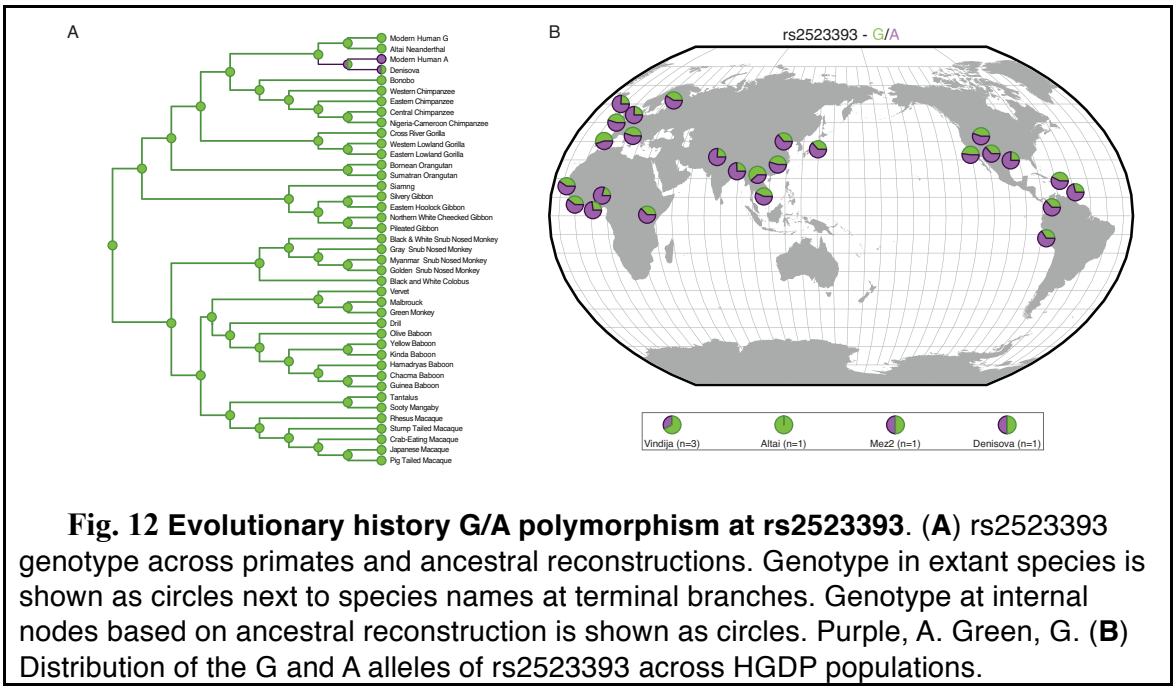

Fig. 12 Evolutionary history G/A polymorphism at rs2523393. (**A**) rs2523393 genotype across primates and ancestral reconstructions. Genotype in extant species is shown as circles next to species names at terminal branches. Genotype at internal nodes based on ancestral reconstruction is shown as circles. Purple, A. Green, G. (**B**) Distribution of the G and A alleles of rs2523393 across HGDP populations.

Fig. 13 PheWAS plot of phenotypes associated with rs2523393 in the UK Biobank. Associations are shown for 3870 ICD10 codes, associations with P<0.05 are shown are triangles and those with P≥0.05 are shown as circles. The most statistical significant phenotype associations are labeled.

The rs2523393 A allele is derived in humans

To reconstruct the evolutionary history of the G/A polymorphism we identified a region spanning 50bp upstream and downstream of rs2523393 from 37 primates, including species from each of the major primate lineages, multiple sub-species of African apes (Homininae), as well as modern and archaic (Altai Neanderthal and Denisova) humans. The *HLA-F* locus is present in all samples as well as it arose early in primates and homologs are found across the primate lineage through New World Monkeys [149,150]. Next, maximum likelihood methods were used to reconstruct ancestral sequences for the 101bp region. The G allele was found to be ancestral in primates and the A variant was only found in Neanderthal, Densiovan, and modern human populations (**Fig 12A**). The frequency of these alleles was then examined across the Human Genome Diversity Project (HGDP) populations and found that the derived and ancestral alleles were segregating at intermediate to high frequencies in nearly all human populations (**Fig 12B**). These results indicate that the derived A variant at rs2523393, which creates a new GATA2 binding site in an enhancer of *HLA-F* expression, arose before the population split between

archaic and modern humans between 550,000 and 765,000 years ago [151].

Discussion

The mechanisms that underlie maternal tolerance of the antigenically distinct fetus are complex [106] and have been the subject of intense study since Medawar proposed tolerance was achieved by maternal immunosuppression and immaturity of fetal antigens [16]. It is now clear from studies in mice that rather than being suppressed, the immune system plays an active role in establishing a permissive environment for implantation, placentation, and gestation, while in humans this has not yet been conclusively demonstrated. Among the diverse immune cells that contribute to the successful establishment and maintenance of pregnancy are maternal regulatory T cells (Tregs) [85,107,108], uterine natural killer cells [81,109], uterine dendritic cells [88], uterine macrophage [110], as well as DSCs themselves [111-113]. DSCs, for example, recruit uterine natural killer cells to the endometrium via IL-15 expression [152-160], stimulate the migration of macrophage into the endometrium via CSF1 expression [17,82,152-154] and suppress cytokine secretion by allogenic CD4+ T cells via PD-1 ligand interactions [112], which are all important for establishing maternal-fetal immunotolerance. These data indicate that DSCs directly regulate local immune responses through multiple mechanisms, potentially also through HLA-F.

Major histocompatibility complex (MHC) genes play an important role in the rejection of non-self tissues, but also contribute to maternal tolerance of the fetus [22-41]. While the precise functions of HLA-F are unclear, its expression level in the endometrium during the window of implantation is associated with fecundability [42], and it is up-regulated by progesterone, implicating it in the establishment of pregnancy. Intriguingly HLA-F binds LIR and KIR natural

killer cell receptors [136-140], suggesting *HLA-F* expressed by DSCs directly signal to uterine natural killer (uNK) cells which are essential for the establishment of maternal immunotolerance and spiral artery remodeling. Consistent with this hypothesis, *HLA-F* expression level is positively correlated with uNK abundance in mid-luteal endometria and is predictive of achieving pregnancy [42,141].

The observation that the G/A polymorphism is shared between modern humans, Neanderthals, and Denisovans indicates it is relatively ancient and may be maintained by balancing selection, frequency dependent selection, or antagonistic pleiotropy [125]. We found that while the derived A allele creates a new GATA2 binding site and augments the function of a progesterone responsive enhancer of *HLA-F*, a previous GWAS found it was also associated with multiple sclerosis (MIM: 126200) with an odds ratio of 1.28 [1.18-1.39] ($P=1.04\times10^{-17}$) [161] and the G polymorphism is also associated with malabsorption/coeliac disease ($P=8.8\times10^{-24}$, beta =-0.0016), multiple sclerosis ($P=2.7\times10^{-10}$, beta =-0.0009), and hyperthyroidism/thyrotoxicosis ($P=1.2\times10^{-8}$, beta =-0.0012) in the UK Biobank GWAS results of ~2,000 phenotypes (**Fig 13**). However, due to strong LD in the HLA region, it would be unsurprising to find these associations with other SNPs as well. These conflicting data are consistent with maintenance through antagonistic pleiotropy, though more evidence is needed to thoroughly conclude this. We have previously shown that another fecundability-associated variant, which switched a repressor into an enhancer of endometrial *TAP2* expression, was also shared between modern humans, Neanderthals, and Denisovans [44]. Remarkably, while the ancestral T allele in the *TAP2* cis-regulatory element was associated with shorter time to pregnancy and ulcerative colitis [126], the derived C allele was associated with longer time to pregnancy and Crohn's disease (MIM: 266600) [127] suggesting these alleles are also maintained by antagonistic pleiotropy. Collectively

these data strongly suggest there is a reproduction-health tradeoff in human evolution.

Author Contributions

Katelyn Mika (KM) and Vincent Lynch (VL) collaboratively conceived, designed, interpreted, and wrote this paper, as well as conducted and analyzed the data on the evolutionary history of the rs2523393 101bp region. KM conducted and analyzed the GTEx replication, determined *HLA-F* expression across the menstrual cycle, *HLA-F* expression across fertile women, women suffering from implantation failure, and women suffering from spontaneous reoccurring abortion, determined *HLA-F* expression in cultured samples of ESFs and DSCs, and prepped and executed the luciferase assays. Xilong Li (XL) and Francesco DeMayo (FD) conceived and executed the GATA2 ChIP-seq and GATA2 knockdown and expression profiling. VL used the CHiCP browser to analyze the Hi-C data.

Acknowledgements

This work was funded by a Burroughs Wellcome Preterm Birth Initiative grant (1013760), an NIH National Institute of General Medical Sciences Graduate Training Grant (T32GM007197), and by the March of Dimes Transdisciplinary Center (TDC) at UChicago-Northwestern-Duke.

Web Resources

GTEx database (http://www.gtexportal.org/home/eqtls/bySnp), GEO2R analysis package (http://www.ncbi.nlm.nih.gov/geo/geo2r/), Geography of Genetic Variants Browser (http://popgen.uchicago.edu/ggv/), Datamonkey web-sever (http://www.datamonkey.org), Online Mendelian Inheritance in Man (http://www.omim.org), Oxford Brain Imaging Genetics (BIG) Server (http://big.stats.ox.ac.uk).

Chapter 4:

Transposable elements continuously remodel the regulatory landscape of decidual stromal cells

Abstract

A long-standing hypothesis of gene expression evolution proposes that genes gain and lose expression domains through a step-wise accumulation of mutations that create new or destroy old cis-regulatory elements, particularly enhancers, repressors, and insulators. Alternatively, a gene may evolve a cis-regulatory element in a single step through the integration and cooption of a transposable element (TE) that harbors functional transcription factor binding sites. While there is ample data to support both models of gene regulatory evolution, important questions with the TE cooption model remain. For example, how are TEs domesticated by the host genome into regulatory elements? Here we trace the evolution of gene expression in the pregnant mammalian uterus and show that TEs have continuously remodeled the regulatory landscape that mediates progesterone responsiveness. We find that genes nearby TE-derived regulatory elements are among the most progesterone responsive in the genome. An assumption of the TE-cooption model is that TEs integrate as functional regulatory elements. We tested this assumption with nearly 100 TE consensus sequences, and surprisingly nearly all were found to be very strong repressors across multiple mammalian cell types and species. However, treatment of mammalian cells with histone deacetylase inhibitors unmasked latent enhancer functions for consensus TEs, while the majority of TEs we tested were strong enhancers in chicken cells, which do not express the TE silencing KRAB-ZFPs. These data suggest a two-step model of TE cooption whereby

active TEs are first recognized and silenced by KRAB-ZFPs and latent enhancer functions are later revealed once TEs lose KRAB-ZFP binding sites.

Introduction

A major challenge in biology is to explain the mechanisms that underlie gene regulatory evolution. A long-standing hypothesis of gene expression evolution proposes that genes gain and lose expression domains through a step-wise accumulation of small-scale mutations that create new or destroy old cis-regulatory elements [162], particularly enhancers, repressors, and insulators. Alternatively, a gene may evolve a cis-regulatory element in a single step through the integration and cooption of a transposable element (TE) that harbors functional transcription factor binding sites [47-53,163]. While there is ample data to support both models of gene regulatory evolution (reviewed in Feschotte [45] and Wray [164]), important questions with the TE cooption model remain. For example: Do TEs exert large or small effects on the genes they regulate? Do those effect sizes change over evolutionary time? Do TEs integrate with regulatory abilities, or as 'pre-regulatory elements' that need additional mutations to acquire regulatory functions? Answers to these (and other) questions are essential for understanding the contribution of TEs in gene regulatory evolution.

Another important component of a TE's ability to be coopted into a regulatory element is the regulation of the TE itself. One well-supported hypothesis in germline and embryonic stem cells is that TEs are recognized by KRAB Zinc Finger Proteins (KRAB-ZFPs), which then recruit silencing machinery, such as KAP1 and the NURD/HDAC repression complex [165], via protein-protein interactions to that location in the genome [165-172]. Of the 467 human KRAB-ZFPs that

have been identified thus far [173], ~2/3 have been found to bind TEs and are associated with their repression [174]. However, the role of KRAB-ZFPs on TE repression in somatic cells is debated [171,175-177]. This host cell response to TEs adds another layer of complexity to the investigation of TEs as regulatory elements because the TE would first need to escape KRAB-ZFP, or any other, silencing before it's function can be coopted by the genome to function as an enhancer or insulator. Are TEs in somatic tissues bound and silenced by KRAB-ZFPs and how do these TEs escape silencing are two questions which remain to be answered and may contribute to the discussion on the role of TE derived *cis*-regulators as well.

Extant mammals span several major evolutionary transitions during the origins of pregnancy and are therefore an ideal system in which to explore the mechanisms of gene regulatory evolution that are associated with origin of novelties. An essential step in the establishment and maintenance of pregnancy in Eutherian mammals is the differentiation (decidualization) of endometrial stromal fibroblasts (ESFs) into decidual stromal cells (DSCs) in response to progesterone acting through the progesterone receptor (PGR), the second messenger cyclic AMP (cAMP), and, in some species, to fetal signals [19]. Decidualization induces large-scale gene regulatory, cellular, and physiological reprogramming in the endometrium, leading to dramatic gene expression changes, the influx of immunosuppressive immune cells, vascular remodeling, and secretory transformation of uterine glands [19,178]. Decidualization evolved in the stem lineage of Eutherian mammals [179-181] and underlies the suite of traits that support prolonged pregnancy in Eutherians, including direct implantation of the blastocyst and trophoblast into maternal endometrium, maternal recognition of pregnancy, maternal-fetal communication, and maternal immunotolerance of the antigenically distinct fetus.

We have previously shown that thousands of genes gained and lost endometrial expression

during the origins of pregnancy and decidualization in early mammals, and that ancient mammalian TEs were co-opted during this process to function as progesterone-responsive cis-regulatory elements [53]. Here we show that successive waves of TEs have been coopted into progesterone-responsive cis-regulatory elements, that genes associated with TEs are among the most strongly differentially regulated by progesterone during decidualization, and suggest that TEs with latent enhancer functions may be coopted by the host genome to function as regulatory elements once the latent function is unmasked.

Materials and Methods

Identification of TE containing regulatory elements

To identify regulatory elements derived from transposable elements (TEs), FAIRE-seq, DNase-seq, H3K27ac ChIP-seq and H3K4me3 ChIP-seq peaks from the analyses in Lynch *et al* (2015)[53] were intersected with the RepMask 3.2.7 track at the U.C.S.C genome browser (repbase libraries release 20050112). Non-transposable element annotations were removed and corrected for fragmented annotations. To identify TEs that were significantly enriched within TE-derived FAIRE-seq, DNase-seq, H3K27ac ChIP-seq and H3K4me3 ChIP-seq peaks the TEanalysis pipeline (https://github.com/4ureliek/TEanalysis) was used with 10,000 replicates.

Tissue collection and processing

The basic structure of the adult female reproductive tract is a hollow tube that includes a ciliated funnel that captures eggs released from the ovary at the anterior end and a muscular tube posteriorly, which may have regional modifications for egg provisioning, shell deposition, egg

storage or retention during embryonic development [182]. In Amniotes, the majority of the female reproductive tract (the oviduct itself) is a broad muscular tube with specialized glands that produce albumin. Distal to the oviduct is a section walled by smooth muscle that contains a hormone-responsive and richly vascular mesenchyme with specialized glands and fibroblasts that provides nutrients to the developing embryo (the endometrium). In oviparous species this region of the uterus lays down the shell; in viviparous species this region is the site of maternal-fetal interactions and placental attachment [183-185]. In viviparous species, the region of the female reproductive tract that is specialized for maternal provisioning of the embryo is termed the uterus, whereas the developmentally and structurally homologous region is termed the shell-gland in oviparous species.

To identify gene expression changes at the maternal side of the maternal-fetal interface in Amniotes, I collected pregnant tissue samples from the endometrium from the pregnant uterus or gravid shell gland portion of the uterus. Tissue samples were dissected to remove myometrium, luminal epithelium, and extraembryonic tissues, and washed 3x in ice cold PBS to remove unattached cell debris and red blood cells. However, tissue samples were not purified into cell types. Thus RNA-Seq derived from these tissues can inform us about what genes are expressed at the interface and how those gene expression patterns changed but not the specific cell-type in which those genes were expressed. Regardless of specific cell type, we are interested in gene expression changes at the maternal fetal interface. For example, I found that transcripts for CD56, GZMa, and GZMb are present in all Eutherian mammal RNA-seq datasets, but not in any of the RNA-seq datasets from non-Eutherian mammals. CD56, GZMa, and GZMb are markers of natural killer (NK) cells [186-188]. Therefore, I can infer that NK cells are present in the uterus of Eutherians, but not outgroups, during pregnancy.

Generation and analyses of RNA-Seq data

RNA-seq data was obtained through the SRA database for samples found in **Table 1**. Additional samples for baboon (x3), mouse (x3), hamster (x3), bat (x2), and squirrel (x2) were prepared according to the tissue collection and processing methods above. Total RNA was extracted using the Qiagen RNeasy plus Minikit per manufacturers instructions. Total RNA was determined by Nanodrop 2000 (Thermo Scientific). A total amount of 2500ng of total RNA per sample was submitted to the University of Chicago Genomics Facility for Illumina Next Gen RNA sequencing. Quality was assessed with the Bioanalyzer 2100 (Agilent). A total RNA library was generated using the TruSEQ stranded mRNA with RiboZero depletion (Illumina) for each sample. The samples were fitted with one of six different adapters with a different 6 base barcode for multiplexing. Completed libraries were run on an Illumina HiSEQ2500 with v4 chemistry on 2 replicate lanes for hamster and 1 lane for everything else of an 8 lane flow cell, generating 30-50 million high quality 50bp single-end reads per sample.

Kallisto version 0.42.4 [189] was used to pseudo align the raw RNA-seq reads. Default parameters were used, bias correction was added, as were 100 bootstraps. Kallisto outputs in TPM, which I then used to determine gene expression.

Table 1 SRA, GEO, or BioProject Accession Numbers for additional RNA-seq datasets used to generate the gene expression gains and losses across the Amniote phylogeny.

Common Name	Scientific Name	Accession Numbers
Human	*Homo sapiens*	SRR1818493, SRR1818521, SRR1818577
Rhesus	*Macaca mulatta*	SSRR324707, SRR324708

Table 1 (cont.)

Rabbit	***Oryctolagus cuniculus***	**SRR3029220, SRR3029221, SRR3029222**
Goat	*Capra hircus*	SRR1917587
Sheep	*Ovis aries*	ERR489180, ERR489181
Cow	*Bos taurus*	Submitted, awaiting accession numbers
Pig	*Sus scrofa*	SRR651716, SRR651717, SRR651718, SRR651719, SRR651720
Dog	*Canis lupus familiaris*	SRR3222429
Horse	*Equus caballus*	GSM525523, GSM525525, GSM525527, GSM525529, GSM525531, GSM525533, GSM525535, GSM525537, GSM525539, GSM525541, GSM525543, GSM731676, GSM731678, GSM731680, GSM731682
Armadillo	*Dasypus novemcinctus*	SRR1289524, SRR1289523
Wallaby	*Notamacropus eugenii*	DRX012238, DRX012249, DRX012257, DRX012259
Dunnart	*Sminthopsis crassicaudata*	PRJNA399240
Opossum	*Monodelphis domestica*	SRX877987

Table 1 (cont.)

Platypus	***Ornithorhynchus anatinus***	**SRR1289525**
Viviparous three-toed skink	*Saiphos equalis*	Submitted, awaiting accession numbers
Oviparous three-toed skink	*Saiphos equalis*	Submitted, awaiting accession numbers
Oviparous skink	*Lerista bougainvillii*	SRR4293357, SRR4293356, SRR4293355
Viviparous skink	*Pseudemoia entrecasteauxii*	SRR3099552, SRR3099550, SRR3099532, SRR3099522, SRR3099520
Oviparous skink	*Lampropholis guichenoti*	SRR4293362, SRR4293354, SRR4293353
Ocellated Skink	*Chalcides ocellatus*	Submitted, awaiting accession numbers
Duck	*Anas platyrhynchos*	SRR5204376, SRR5204377, SRR5204378, SRR5204379, SRR5204380, SRR5204381, SRR5204382, SRR5204383, SRR5204384, SRR5204386, SRR5204387, SRR5204388
Chicken	*Gallus gallus*	SRR546496
Frog	*Xenopus tropicalis*	GSE12975

Gene expression calling

I used a model-based method to classify genes as expressed based on the number of transcripts per million (TPM) in the total RNA-seq dataset. Read counts were normalized by total estimated transcript number, expressed in transcripts per million transcripts or TPM [190]. This normalization is invariant with respect to Trimmed Mean of M-values normalization [191] since the correction factor affects both the numerator as well as the denominator of the TPM value. The distribution of transcript abundances was fitted to a model consisting of a discretized exponential distribution, represented transcripts from repressed genes, and a negative binomial distribution, representing the distribution of abundance values of actively transcribed genes as marked by chromatin modifications. The model suggests that genes with a TPM≤2 are likely from transcriptionally suppressed genes [192]. This threshold is consistent with one obtained by comparing the transcript abundance with the chromatin state of the respective gene [193] and I thus classified genes with TPM>2 as expressed and those with TPM≤2 as not expressed for all species except platypus for which we classified genes with TPM≤1 as not expressed.

Parsimony reconstruction of gene expression gain/loss

I used Mesquite (v2.75) and parsimony optimization to reconstruct ancestral gene expression states, and identify genes that gained and lost endometrial expression in Amniotes. Expression was classified as an unambiguous gain if a gene was not inferred as expressed at the ancestral node (state 0) but inferred as expressed in a descendent node (state 1) and vice versa for the classification of a loss from endometrial expression.

Transcription factor binding site enrichment analysis

A custom bioinformamtic pipeline was used to determine enrichment of transcription factor binding sites in TEs that intersect with the DSC FAIRE-seq, DNase-seq, H3K27ac ChIP-seq and H3K4me3 ChIP-seq peaks [53] versus the genomic abundance of ChIP-Seq peaks; 10,000 bootstrap reshufflings were used to assess statistical significance. Scripts are publically available and archived at https://github.com/4ureliek/TEanalysis. The location of ChIP-Seq peaks in hg19 ENCODE data was downloaded from the USCS genome browser (Txn Fac ChIP V2 - Transcription Factor ChIP-Seq from ENCODE (V2)). The location of PGR ChIP-Seq peaks was obtained from GEO (GSE94036) and is available [194].

Cell culture and luciferase Assays

Endometrial stromal fibroblasts (ATCC CRL-4003) immortalized with telomerase, mouse embryonic fibroblast (MEF) KAP1 knockout, and MEF Flox/Flox control cells (the latter two obtained from the Trono Lab at the Ecole Polytechnique Fédérale de Lausanne) were maintained in phenol red free DMEM (Gibco) supplemented with 10% charcoal stripped fetal bovine serum (CSFBS; Gibco), 1x ITS (Gibco), 1% sodium pyruvate (Gibco), and 1% L-glutamine (Gibco). Chicken embryonic fibroblast cells (ATCC CRL-12203) and HEPG2 cells from the Brown Lab at the University of Pennsylvania were maintained in phenol red containing DMEM + Glutamax (Gibco) supplemented with 10% fetal bovine serum (FBS, Gibco) and Normocin (InviviGen). Elephant fibroblasts were maintained in 1:1 MEM (Corning cellgro) supplemented with 10% FBS, 1% penstrep (Gibco), 1x sodium pyruvate (Gibco), and 1x L-glutamine (Gibco) to FGM-2 (Lonza), made per manufacturer's instructions.

To generate the vectors for functional testing using luciferase assays, I selected 79 of the 352

TE families enriched within DSC regulatory regions as marked by the previously generated H3K4me3 ChIP-seq, H3K27ac ChIP-seq, DNaseI-seq, and FAIRE-seq datasets[53]. These 79 TEs also met the following criteria: 1) at least 1 element from every lineage from Mammalia to Eutheria and 2) represented all 4 classes of TEs (LTR, LINE, SINE, and DNA). Consensus sequences for these elements were taken from the database Dfam[195]. 10 additional TEs also found in the enriched 352 were then chosen that were unique to Old World Monkeys or younger and their consensus sequences were also obtained from Dfam[195]. The total set of elements were biased towards, but not limited to, the DNA class of transposable elements. These 89 consensus sequences were then synthesized by Genscript, who also cloned them into the pGL3 Basic vector (Promega) with an added minimal promoter (pGL3Basic[minP]).

Confluent cells in 96 well plates in 80µl of Opti-MEM (Gibco) were transfected with 100ng of the TE containing luciferase plasmid and 10ng of the pRL-null renilla vector (Promega) with 0.1µl PLUS reagent (Invitrogen) and 0.25µl of Lipofectamine LTX (Invitrogen) in 20µl Opti-MEM. The cells incubated in the transfection mixture for 6hrs and the media was replaced with the maintenance media overnight. Decidualization of ESFs was then induced by incubating the cells in the decidualization media: DMEM with phenol red (Gibco), 2% CSFBS (Gibco), 1% sodium pyruvate (Gibco), 0.5mM 8-Br-cAMP (Sigma), and 1µM of the progesterone analog medroxyprogesterone acetate (Sigma) for 48hrs. ESFs (decidualization controls) were incubated in the decidualization control media (phenol red free DMEM (Gibco), 2% CSFBS (Gibco), and 1% sodium pyruvate (Gibco) instead for 48hrs. For trichostatin A (TSA; Tocris Bioscience) trials, 1µM TSA was added to all the medias from plating through decidualization.

After decidualization for ESFs and DSCs or after 48hrs from transfection for other cell types, Dual Luciferase Reporter Assays (Promega) were started by incubating the cells for 15mins in

20µl of 1x passive lysis buffer. Luciferase and renilla activity were then measured using the Glomax multi+ detection system (Promega). Luciferase activity values were standardized by the renilla activity values and background activity values as determined by measuring luminescence from the pGL3-Basic[minP] plasmid with no insert. Each luciferase experiment was replicated in at 4-6 independent experiments. To identify significant expression shifts, I performed Wilcoxon tests on the data, followed by a Benjamini and Hochberg [144] p-value multiple testing adjustment; significance was determined by having an adjusted p-value ≤ 0.05.

Results

Transposable elements are major contributors to regulatory elements in decidual stromal cells (DSCs)

It has been previously shown that ancient mammalian TEs that invaded the mammalian genome in the Mammalian, Therian, and Eutherian stem-lineages and played an important role in the origin of new cis-regulatory elements in DSCs during the evolution of pregnancy [53]. To expand these studies to all classes of TEs I used ChIP-seq data from human DSCs to identify regions of the genome marked with the enhancer-associated histone modification H3K27ac and the promoter-associated histone modification H3K4me3, and FAIRE-seq and DNaseI-seq to identify regions of open chromatin. I observed that 58.7% of H3K27ac and 53.0% of H3K4me3 ChIP-seq peaks, 42.2% of FAIRE-seq peaks, and 67.2% of DNaseI-seq peaks overlapped annotated transposable elements (**Fig 14A**). DNaseI-seq and FAIRE-seq are biased towards different regions of the genome, for example DNaseI sites have been shown to be enriched within 2kb of transcription start sites while FAIRE-seq sites are enriched in non-promoter intergenic regions [196], which may explain the differences in content seen between the two

different marks of open chromatin. This is interesting with regards to the interpretation of the TE enrichment as we see more TEs enriched in the FAIRE-seq data, suggesting that we may find more TEs in the regions FAIRE is biased towards- such as non-promoter intergenic regions

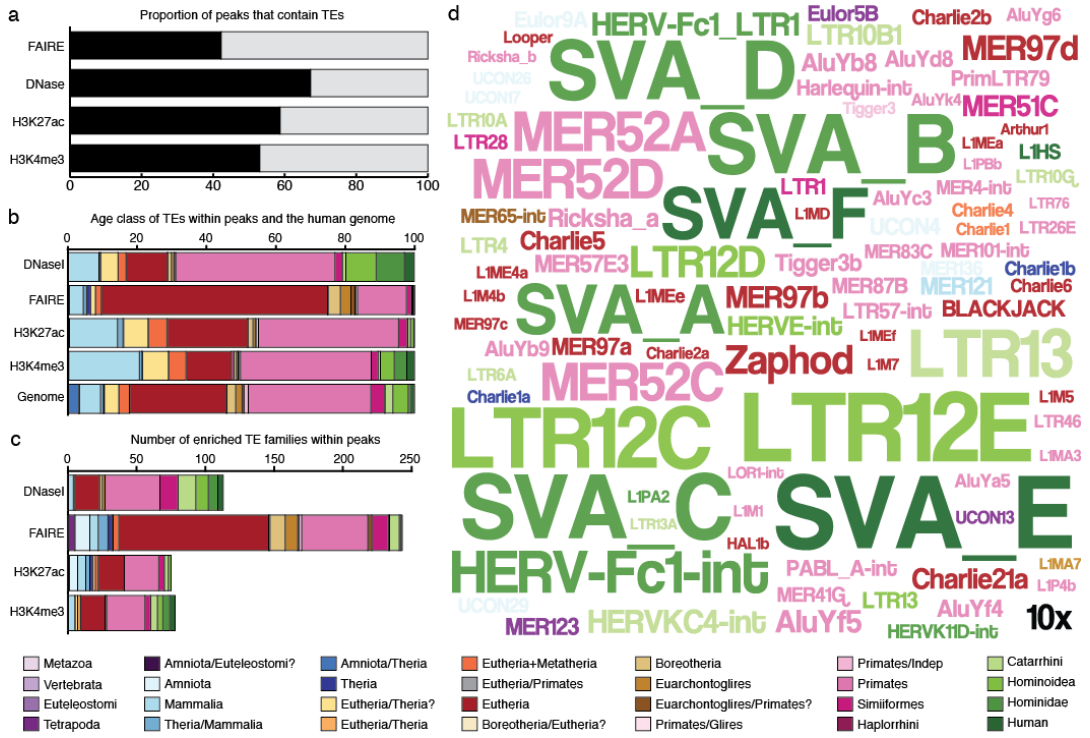

Fig. 14 Transposable elements are major contributors to regulatory elements in endometrial stromal cells. (**A**) Proportion of DNaseI-seq, FAIRE-seq, H3K27ac ChIP-seq, and H3K4me3 ChIP-seq peaks that contain transposable elements (black) or no annotated transposable elements (gray). (**B**) Proportion of transposable elements of different ages (color coded by lineage specificity) within DNaseI-seq, FAIRE-seq, H3K27ac ChIP-seq, and H3K4me3 ChIP-seq peaks and the human genome. See legend for approximate age of transposable element family, i.e. lineage in which the transposable element appears in the genome. (**C**) Number of transposable element families enriched within DNaseI-seq, FAIRE-seq, H3K27ac ChIP-seq, and H3K4me3 ChIP-seq peaks (color coded by lineage specificity) compared to the human genome. See legend for approximate age of transposable element family, i.e. lineage in which the transposable element appears in the genome. (**D**) Word cloud of the 100 most enriched transposable elements under DNaseI-seq, FAIRE-seq, H3K27ac ChIP-seq, and H3K4me3 ChIP-seq peaks compared to the human genome. The size of the transposable element's name corresponds to its enrichment (see inset 10-fold scale). See legend for approximate age of transposable element family, i.e. lineage in which the transposable element appears in the genome.

(possible enhancers), or associated with GATA binding sites (an obligate co-factor of the progesterone receptor). Next I annotated these TEs by their lineage specificity and found that TEs from different age classes differentially contributed to each kind of regulatory element: relatively recent TEs (i.e. primate lineage) dominated the DNaseI and H3K4me3 datasets, whereas relatively ancient TEs (i.e. Eutherian-specific and older) were more common in the FAIRE and H3K27ac datasets (**Fig 14B**). Finally, I identified TEs that were enriched within H3K27ac and H3K4me3 ChIP-seq, and FAIRE-, DNase-seq peaks. 352 TE families (eTE) were significantly enriched (>1.5-fold, P≤0.05, binomial test) (**Fig 14C**), most of which were Eutherian- and Primate-specific (**Fig 14D**).

TE-derived regulatory elements are enriched in motifs for master regulators of ESC identity

To determine if TEs donated motifs for specific transcription factors, I performed *de novo* motif discovery in TE-derived regions of DNaseI-seq, FAIRE-seq, H3K27ac ChIP-seq, and H3K4me3 ChIP-seq peaks. MEME identified 368 enriched motifs (E-value<0.05), including motifs that match core regulators of ESF/DSC cell-type identity, mediate progesterone responsiveness and the early events leading to implantation such as FOSL2 (E-value=3.70-251), HOXA10 (E-value=4.40-211), NR2F2 (E-value=3.00-107), AHR/ARNT (E-value=2.40-80), and CEBPB (E-value=1.10-64). Other enriched motifs include those previously shown to be derived from TEs such as CTCF (E-value=1.40-33) and TP53 (E-value=2.50-19), as well as the KRAB-ZFPs ZNF263 (E-value=7.00-145) and ZNF354C (E-value=2.70-111), which recruit the NuRD and CoREST repressor complexes to silence TEs.

Next, over-represented transcription factor binding sites (TFBS) within TE-derived regions of FAIRE-seq, DNase-seq, H3K27ac ChIP-seq and H3K4me3 ChIP-seq peaks were identified using previously published ENCODE ChIP-seq data for 132 transcription factors and co-factors and PGR ChIP-seq data from human DSCs. 53 TFBSs were enriched within putatively

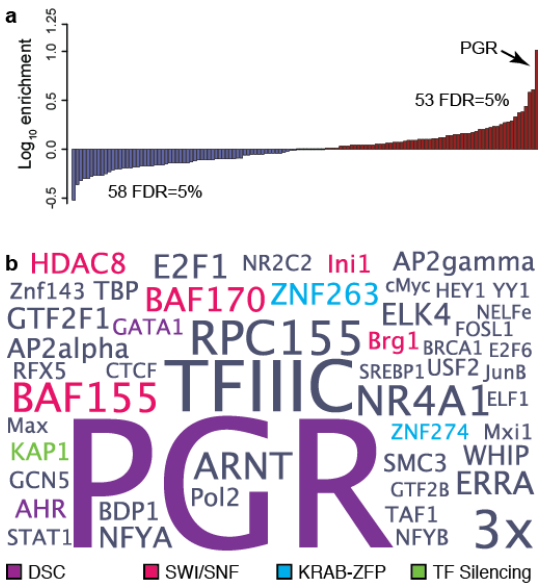

Fig. 15 Transposable elements are enriched in binding sites for transcription factors that mediate hormone responsiveness and endometrial cell identity. (**A**) Distribution of enriched (red) and depleted (blue) transcription factor binding sites in transposable element derived segments of FAIRE-seq, DNaseI-seq, H3K27ac ChIP-seq, and H3K4me3 ChIP-seq peaks compared to transcription factor ChIP-seq peaks with no transposable element overlap. 53 and 58 transcription factor binding sites were significantly enriched and depleted, respectively, at FDR=5%. (**B**) Word cloud of transcription factor binding site enrichment in transposable elements. Colors indicate transcription factors that mediate hormone responses (purple), remodel chromatin (pink), KRAB-ZFPs (blue), or involved in TE silencing (green). Data shown for ≥3-fold enriched transcription factors at FDR=5%.

regulatory TEs relative to genomic TFBS abundances (FDR=5%; **Fig 15A**), most notably PGR (enrichment=10.31, FDR≤1.00×10-250), AHR (enrichment=2.00, FDR=1.00×10-5), and GATA (enrichment=1.25, FDR=1.30×10-20). Enrichment for the KRAB-ZFPs ZNF263

(enrichment=1.71, FDR=7.20×10-150) and ZNF274 (enrichment=1.22, FDR=8.90×10-3), KAP1 (also known as TRIM18; enrichment=2.00, FDR=7.20×10-41), which binds KRAB-ZFPs and functions as a scaffold for the recruitment of histone modifying co-repressor complexes, and parts of the SWI/SNF chromatin remodeling complex such as BAF155 (enrichment=2.36, FDR=1.00×10-76), BAF170 (enrichment=1.93, FDR=4.20×10-14), INI1 (enrichment=1.43, FDR=9.10×10-18), and BRG1 (enrichment=1.38, FDR=2.70×10-5) was also observed. These data suggest that TEs have donated binding sites for transcription factors that mediate decidualization (PGR, CEBPB, HOXA10) as well as more general transcriptional repressor (KRAB-ZFPs, KAP1, SWI/SNF complex members).

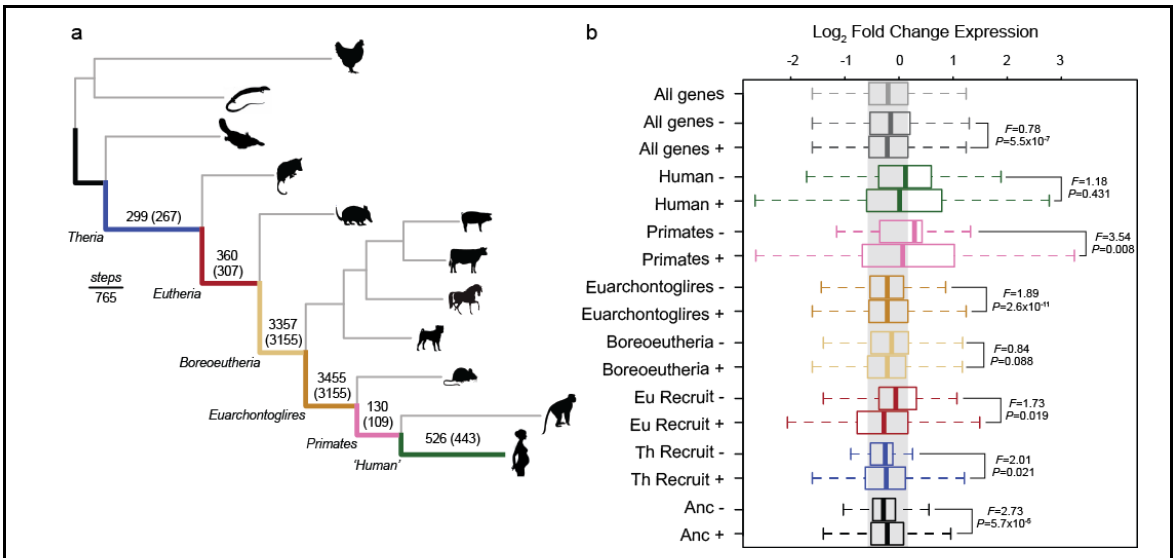

Fig. 16 Genes associated with TE-derived regulatory elements are more strongly differentially regulated during decidualization than genes without TE-derived regulatory elements. (**A**) Parsimony reconstruction of gene expression gains and losses in the endometrium of amniotes. Numbers above branches indicate the average number of genes that gained and lost endometrial expression in that stem-lineage as inferred by Wagner parsimony. Branch lengths are drawn proportional to gene expression gain and loss events for all lineages. (**B**) Recruited and ancestrally expressed genes associated with ancient mammalian TE-derived regulatory elements (+) are more strongly differentially regulated upon cAMP/MPA-induced decidualization than recruited or anciently expressed genes without TE-derived regulatory elements (-). F is the ratio of variances from a two-sample F test.

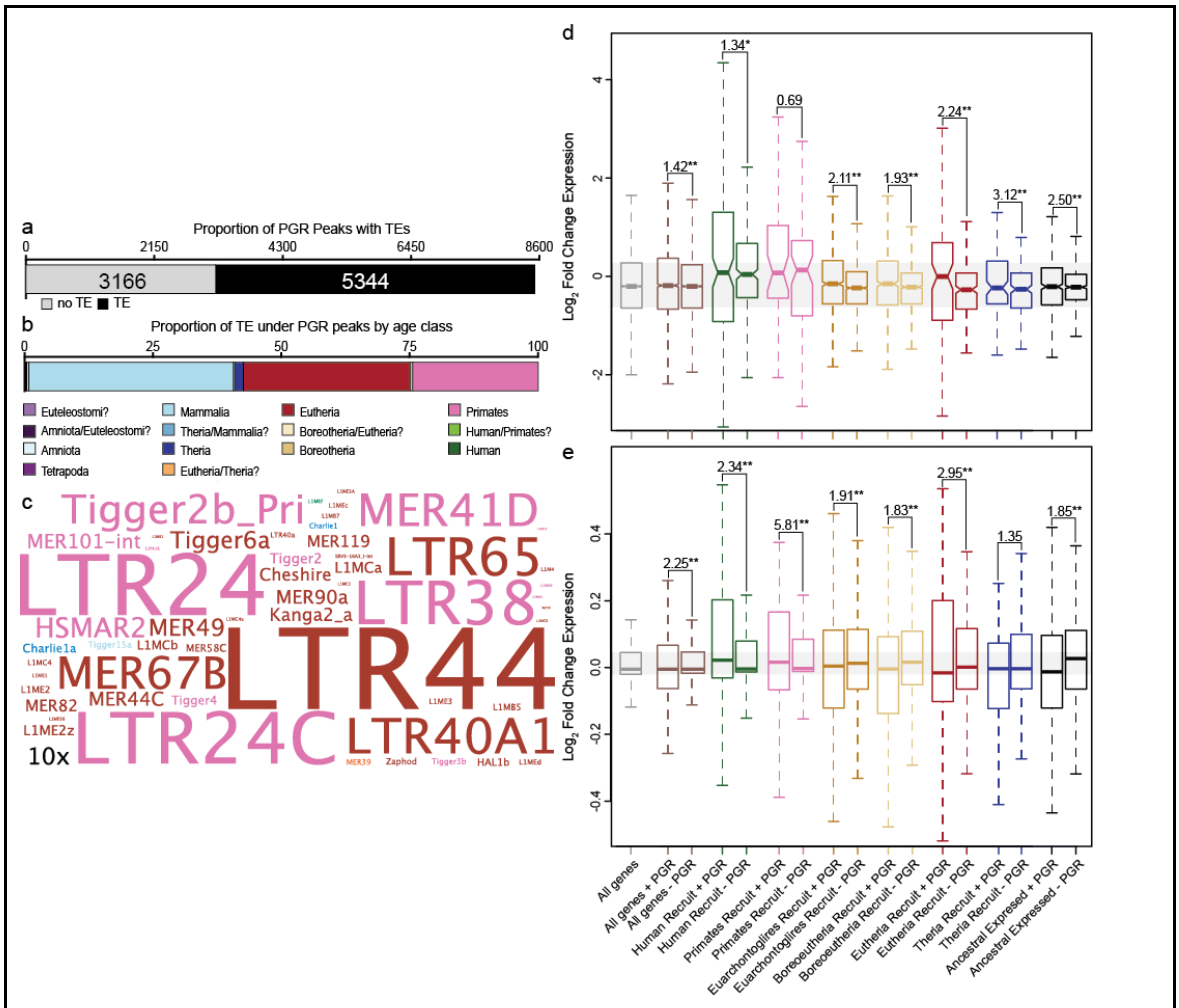

Fig. 17 Lineage specific transposable elements remodeled progesterone receptor binding site architecture across the genome. (**A**) The number of PGR ChIP-seq peaks that contain transposable elements in primary cultures of hESCs treated with 10nM 17β-estradiol, 100nM medroxyprogesterone acetate, and 1mM 8-bromo-cAMP. (**B**) Proportion of transposable elements in PGR ChIP-seq peaks by age class. (**C**) Word cloud of TEs enriched in PGR ChIP-seq peaks. Colors indicate age class. Inset scale (10x) shows 10-fold enrichment. (**D**) Ancestrally expressed and recruited genes associated with TE-derived PGR binding sites (+ PGR) are more strongly differentially regulated in hESCs upon cAMP/MPA-induced decidualization than genes without TE-

Fig 17 (cont.) derived PGR binding sites (-PGR). F is the ratio of variances from a two-sample F test. Genes are grouped according to when they evolved endometrial expression. * = P<0.01, ** = P<1.0×10-5. (**E**) Ancestrally expressed and recruited genes associated with TE-derived PGR binding sites (+ PGR) are more strongly dysregulated by PGR knockdown in cAMP/MPA treated hESCs than genes not associated with ancient mammalian TE-derived PGR biding sites (- PGR). F is the ratio of variances from a two-sample F test. Genes are grouped according to when they evolved endometrial expression. * = P<0.01, ** = P<1.0×10-5.

TE-derived regulatory elements augment progesterone responsiveness

The observation that specific TE families are enriched in DSC regulatory elements suggests that they may contribute to gene expression changes that occur during progesterone- induced decidualization. To test this hypothesis I used parsimony to reconstruct the evolutionary history of gene expression in the pregnant uterus of mammals, GREAT to associate genes with putative regulatory elements, and RNA-seq data from human ESFs and DSCs to quantify gene expression changes induced by decidualization. Next I inferred when each gene evolved to be expressed in the pregnant uterus and homologous structures in outgroups (**Fig 16A**), identified genes with eTE-derived regulatory elements, and inferred if they were expressed differently than genes without eTE-derived regulatory elements. Next, I used an F-test to compare the variance in shifts in gene expression in response to cAMP/MPA treatment between groups of genes that evolved to be expressed in the endometrium in different Eutherian lineages that were either associated with an eTE or not associated with TEs. In response to cAMP/MPA, gene expression may go up, go down, or stay the same. If there are many gene expression shifts in the population we are analyzing, variance will increase but the mean may not shift as the different responses balance out. The F-test tests the null hypothesis that the ratio of the variances of two populations is equal and therefore can test if genes with eTEs are more strongly up- and down-regulated during decidualization (have greater variance) than genes without eTEs. The p-values allow us to determine if the differences in the variance between the populations are significant, the F statistic the direction (>1, + has larger variance; <1, - has larger variance). Indeed, genes with eTE-derived regulatory elements were generally more strongly differentially regulated by decidualization than genes without eTE-derived regulatory elements (**Fig 16B**).

The observations that TEs are enriched in PGR binding sites and are associated with genes that are strongly differentially expressed upon decidualization prompted me to explore contribution of TEs to PGR binding sites in greater detail. I found that 62.8% (5344/8510) of PGR ChIP-seq peaks contained TEs (**Fig 17A**), the vast majority of which are Mammalian-, Eutherian-, and Primate-specific (**Fig 17B**); PGR ChIP-seq peaks, however, are almost exclusively enriched in Eutherian- and Primate-specific TEs (**Fig 17C**). Consistent with a functional role for TE-derived PGR binding sites in orchestrating progesterone responsiveness, genes associated with TE-derived PGR binding sites were more strongly differentially expressed during decidualization than genes not associated with TE-derived PGR binding sites (**Fig 17D**). This trend tended to be exaggerated for more recently recruited genes, such that more recently recruited genes with TE-derived PGR binding sites were more strongly differentially expressed by cAMP/MPA treatment than more anciently recruited genes with TE-derived PGR binding sites (**Fig 17D**). Genes associated with TE-derived PGR binding site are also significantly more dysregulated by siRNA-mediated PGR knockdown in DSCs than genes not associated with TE-derived PGR binding sites (**Fig 17E**).

Consensus TEs are repressors with hidden enhancer potential

Numerous studies have shown that transposable elements have donated binding sites for transcription factors to the genome and function as regulatory elements, however, it is not clear if TEs integrate into the genome with regulatory functions and therefore immediately function as regulatory elements or if they integrate as ‘pre-regulatory elements’ that are weakly- or non-functional and require additional mutations to acquire bona fide regulatory functions. To test these scenarios, I selected 89 of the enriched TEs, synthesized their consensus sequences

(conTE), and cloned them into the pGL3-Basic[minP] luciferase reporter vector; unlike the pGL4 series luciferase reporter vectors, pGL3 is not responsive to progesterone allowing us to

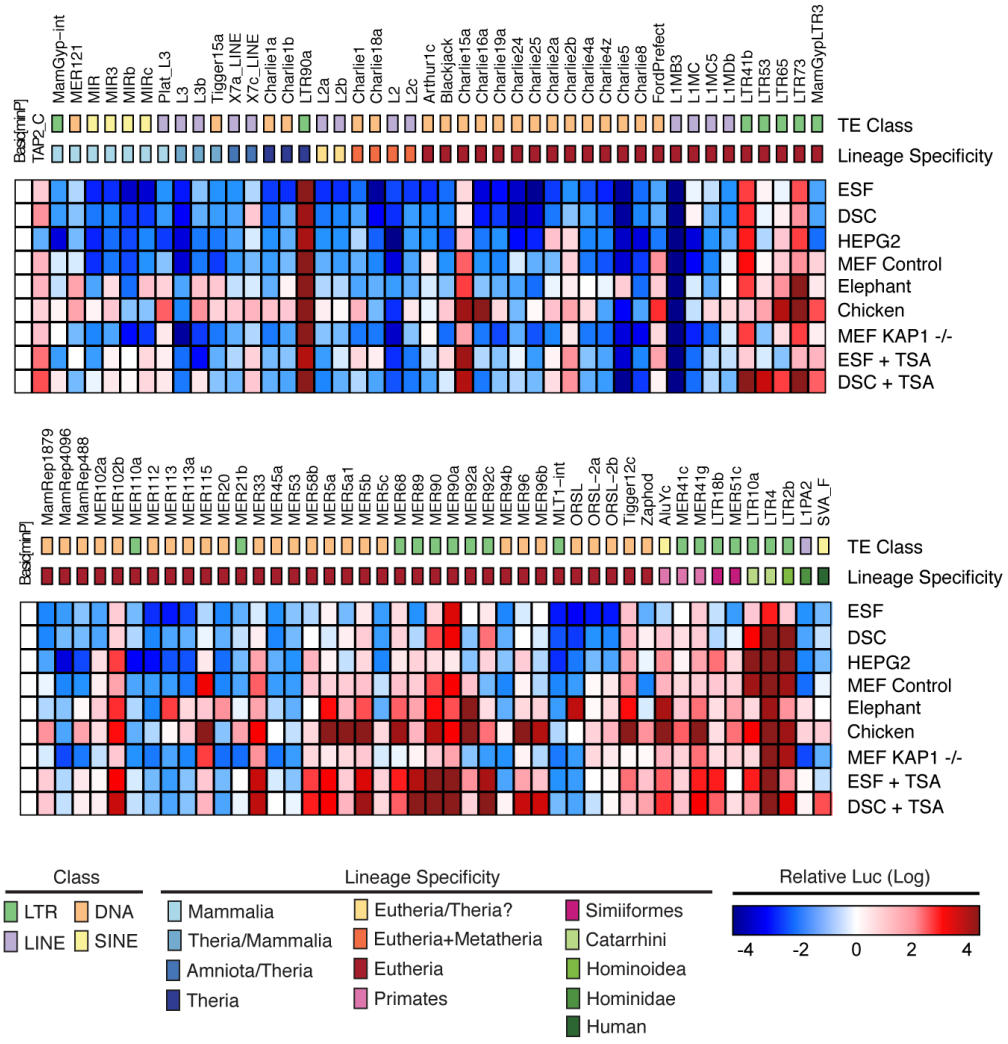

Fig. 18 Consensus TEs are repressors with hidden enhancer potential. Each box represents the mean of 4-6 replicate dual luciferase reporter assays standardized to a renilla vector and empty vector (Basic[minP]) reference. TAP2_C is a known progesterone responsive enhancer [44] used as a positive control for differentiation in DSC cells. The 89 TE constructs are then in order of decreasing age as determined by lineage specificity of the introgression. Darker colors indicate more extreme expression shifts from the reference. Red indicates enhancer function, Blue indicates repressor function. ESF - Endometrial Stromal Fibroblasts, DSC- Decidual Stromal Cells, HEPG2- Liver Cancer, MEF - Mouse Embryonic Fibroblast, Elephant- Elephant Dermal Fibroblast, Chicken- Chicken Embryonic Fibroblast, +TSA - Trichostatin A added to media. See **Table 2** for significance tests.

test whether conTEs can function as progesterone responsive regulatory elements.

To test each conTEs ability to regulate gene expression, I transiently transfected human ESFs and DSCs with each conTE reporter, along with a Renilla control vector, and determine their regulatory ability using a dual luciferase reporter assay. I expected that conTE would generally be evenly divided into enhancer, repressor, and non-functional classes, I found that 35 (40%) and 55 (62%) functioned as repressors in ESFs and DSCs (Padj≤0.05), respectively, with an additional 20 (22%) TEs in ESFs and 3 (3%) TEs in DSCs having repressor-like function with slightly less stringent statistical significance (0.05 < Padj < 0.1); in contrast only 13 and 21 had enhancer functions in ESFs and DSCs (**Fig 18**; Padj ≤ 0.05). To test whether these effects were cell-type specific, I repeated the luciferase reporter assay in the human hepatocellular carcinoma cell line HepG2 and again observed that 59/89 (66%) were strong repressors (Padj≤0.05) whereas 28 were enhancers (**Fig 18**; Padj ≤ 0.05). To determine if these results were species specific we repeated the luciferase assay in mouse embryonic fibroblasts (MEFs) and observed that 51 (57%; Padj ≤ 0.05) of conTEs were repressors (34 were enhancers; Padj ≤ 0.05) whereas in elephant dermal fibroblasts (EDFs) 21 (24%; Padj ≤ 0.05) were repressors (34 were enhancers; Padj ≤ 0.05) (**Fig 18**). While some conTEs, mostly LTR elements which have strong promoters, had enhancer functions in all cell-types, significantly more were repressors than expected- ESF (P=0.027), DSC (P=1.00×10-6), HepG2 (P<1.00×10-6), and MEF (P=0.027).

Our observation that TEs enriched within regulatory elements and open chromatin are also enriched in binding sites for KRAB-ZFPs, KAP1, and histone modifying co-repressor complexes suggest that the conTEs may be recognized by the host cell anti-TE machinery and silenced [166,171,175]. As an initial test of this hypothesis I repeated the dual luciferase reporter assay in ESFs and DSCs but treated cells with trichostatin A (TSA), which selectively inhibits mammalian class

I and II histone deacetylases (HDACs). TSA treatment de-repressed 30/35 (86%; Padj ≤ 0.05) and 39/55 (71%; Padj ≤ 0.05) of the conTEs that were significant repressors in untreated ESFs and DSCs, respectively, including unmasking latent enhancer functions for 29 conTEs in ESFs and 26 conTEs in DSCs (for example: MIR3, MIRb, and MIRc, MamRep1879, and MER96b; **Fig 18**). This de-repression should not be limited to the conTEs but seen genome-wide as I inhibited a global silencing mechanism. However, it gives us insight into the mechanism of repression affecting these constructs.

These data suggest that conTEs may be recognized by the cell's anti-TE surveillance system (KRAB-ZFPs/KAP1) and silenced by histone modifying co-repressor complexes (NuRD, CoREST, SWI/SNF) [166,171,173,175]. To test this hypothesis I repeated the dual luciferase reporter assay in KAP1 knockout mouse embryonic fibroblasts (MEF KAP1-/-), but again observed 53 conTEs (59%; Padj ≤ 0.05) were repressors and relief of repression for only 8 (9%; Padj ≤ 0.05) conTEs (**Fig 18**). Collectively, these results suggest that the repressive effects of conTEs is neither cell-type nor species-specific, and is independent of KAP1. To infer if repression may be mediated by KRAB-ZFPs I took advantage of the restricted lineage specificity of KRAB-ZFPs and tested the regulatory abilities of conTEs in chicken embryonic fibroblasts (CEFs); none of the chicken KRAB-ZFPs are expressed in CEFs [197], compared to the high numbers expressed in mouse and human cell lines [197]. In stark contrast to the different mammalian cell-types we tested, only 18 (20%; Padj ≤ 0.05) of the conTEs functioned as repressors whereas 69 (78%; Padj ≤ 0.05) were strong enhancers in CEFs (**Fig 18**). While these results show that conTEs are generally enhancers in chicken embryonic fibroblasts, which do not express KRAB-ZFPs, we cannot conclusively demonstrate that other mechanisms specific to the chicken lineage are not responsible for these effects. Additional experiments are necessary to demonstrate that conTEs

are bound by KRAP-ZFPs to further support the model that repression is mediated by these proteins, and that the relief of repression seen in chickens is because the CEFs do not express KRAB-ZFPs. However, I can infer that the repression I observe in mammalian cells is consistent with our model that the host cell's TE silencing machinery, in particular its HDAC containing chromatin remodeling factors such as NURD, recognizes and silences conTEs, and this process is likely KAP1 independent as there is no significant difference in luciferase expression in mouse embryonic fibroblasts of the conTEs that are found along that organismal lineage (Wilcoxon test P=0.25).

Discussion

Cooption of transposable elements (TEs) that harbor functional transcription factor binding sites is an alternative route of cis-regulatory evolution compared to the step-wise accumulation of small scale mutations [47-53,162,163]. In previous work on the regulatory systems underlying mammalian pregnancy, Lynch et al [53] showed that thousands of genes gained and lost endometrial expression and ancient mammalian TEs were co-opted during this process to function as progesterone-responsive cis-regulatory elements. Here I expanded upon that study to consider all ages of TEs along the human lineage as well as gene expression changes within the Eutherian mammals. I found mostly Eutherian and Primate specific elements within the 352 TE families enriched within regions marked as having active regulatory function in humans, suggesting multiple waves of TE cooption occurred in decidual tissues, particularly in primates.

How changes to regulatory regions, especially by TEs, contribute to gene regulatory evolution is an open question. The establishment and maintenance of the evolutionary novelty

mammalian pregnancy relies upon progesterone acting through the progesterone receptor (PGR), the second messenger cyclic AMP (cAMP), and, in some species, to fetal signals to differentiate (decidualize) ESFs into DSCs [19]. Corroborating previous results [53], I have shown that genes with TE derived regulatory elements are more responsive to the progesterone pregnancy signal than genes without a TE derived cis-regulatory element. Furthermore, genes near young TEs with PGR binding sites are more progesterone responsive than genes near ancient TEs harboring PGR binding sites or genes without TE derived PGR binding sites. While this suggests that TEs may have played a role in rewiring the progesterone responsive gene regulatory network underlying the evolution of pregnancy, more work needs to be done. In particular, TEs need to be evaluated to assess their responsiveness to progesterone, but if our model is correct this a challenging study to complete as most conTEs are silenced and therefore their underlying progesterone response is masked.

While numerous studies have shown that transposable elements have donated binding sites for transcription factors to the genome, can be bound by transcription factors, and have been coopted into cis-regulatory elements [47-52], it has not been demonstrated if TEs integrate into the genome with regulatory functions and therefore immediately function as regulatory elements or if they integrate as 'pre-regulatory elements' that are not immediately functional and require additional mutations to acquire regulatory functions. We addressed this question using luciferase assay functional tests of conTEs.

Ideally, we would use an ancestral sequence reconstruction (ASR) of each TE to determine if TEs had ancestral regulatory abilities. However, reconstructing the earliest ancestral sequence for most TEs is not possible because most TEs do not have an outgroup to root their phylogeny and therefore we cannot identify which node is the deepest ancestor. In place of an ASR, we and

others who have explored similar questions used conTEs. This introduces an obvious limitation to our inferences: the conTE may not be an accurate representation of the deepest ancestral sequence. Indeed, there is no guarantee that a consensus sequence represents a sequence that ever existed. While these limitations cause us to carefully interpret our results, we believe that the consilience of evidence supports an inference that ancestral TEs, as best they can be approximated by consensus sequences of TEs, may have had regulatory abilities. In contrast, if none of our conTEs had regulatory abilities in any cell type tested we may have inferred that ancestral TEs did not possess regulatory abilities.

Another possible limitation to this approach is that the choice of reference sequence, the luciferase activity driven by an empty vector backbone into which our conTEs are inserted, limits our interpretations to the comparison of the effects of different TEs in the various conditions on luciferase expression relative to each other. The backbone vector does have some residual luciferase expression with no insert, therefore, it is possible that any insert could affect whatever in the backbone is driving this residual expression and it is more than the content of the insert but rather that anything is inserted at all that is altering the luciferase expression in our experiment. If only presence or length of the insertion and not the content of the insertion mattered, then I would not expect to see significant shifts in regulatory ability within the same cell type for sequences of the same length. For example, MER45a and MER5b are both Eutherian specific, DNA elements with 55% and 49% GC content respectively that are both 178bp long. Across all samples MER45a has repressor like function, while MER5b has enhancer like function (**Fig 18**). If function were solely the result of having an insertion and was agnostic of insertion content then I would have expected these two constructs to have the same ability to drive luciferase expression. We also have an ESF and DSC positive control enhancer insert

(TAP2_C) which functions as an enhancer in every cell type but HEPG2. Given the data, I would have expected repressor function if the luciferase expression was altered by insertion and not the content of the insertion. However, since we can't establish that the empty backbone reference drives expression at true background levels, our interpretations remain a comparison of the effects of different TEs in the various conditions on luciferase expression relative to each other.

Despite these caveats, my data suggests that the majority of consensus TEs have enhancer ability but this ability is silenced in mammalian cells, perhaps by KRAB-ZFPs and the NURD HDAC inhibitory complex. We do see a mild enrichment within the dataset of TEs with regulatory marks in DSCs. This enrichment is likely low because 1) the elements analyzed have likely escaped repression by KRAB-ZFPs in order to be coopted into enhancers , and 2) the datasets used to identify KRAB-ZFP, KAP1, and the majority of other transcription factors, were from ENCODE and not DSCs. However, the relief of repression seen in chicken cells, which do not express KRAB-ZFPs [197], is at least coincidental evidence that KRAB-ZFPs may play a role in silencing conTEs. In embryonic stem cells, KAP1 binds the KRAB-ZFPs and coordinates the silencing of the bound TE [166,171,175]. However, in adult somatic tissues the role of KAP1 is unclear [171,175,176]. Here, silencing of the conTEs in MEFs appears to be KAP1 independent. These results need to be confirmed, for example, by ChIP to demonstrate conTEs are indeed bound by KRAB-ZFPs. We hypothesize that only when TEs escape this silencing regulation are they coopted by the genome to play a regulatory role.

Author Contributions

Katelyn Mika (KM) and Vincent Lynch (VL) collaboratively conceived, designed, interpreted, and wrote this paper, as well as analyzed the expression shifts of genes associated with TE derived regulators or TE derived PGR binding sites in response to progesterone signaling and did the binding site enrichment analyses. KM conducted the tissue collection and processing, generated and analyzed the RNA-seq data, performed the gene expression calling, completed the parsimony reconstruction, and prepped, conducted, and analyzed the luciferase assays and associated cell culture. VL identified the TE containing regulatory elements.

Acknowledgements

The authors thank Aurelie Kapusta for assistance with running the TE enrichment program.

Web Resources

UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgGateway), TEanalysis pipeline (https://github.com/4ureliek/TEanalysis)

Table 2 Significance results (Benjamini and Hochberg adjusted p-values) from Wilcoxon tests between each TE and the Basic[minP] empty vector reference for luciferase assays in each cell type. ESF - Endometrial Stromal Fibroblasts, DSC- Decidual Stromal Cells, HEPG2- Liver Cancer, MEF - Mouse Embryonic Fibroblast, Elephant- Elephant Dermal Fibroblast, Chicken- Chicken Embryonic Fibroblast, +TSA - Trichostatin A added to media.

	ESF	DSC	HEPG2	MEF	ELEPHANT	CHICKEN	MEF KAP1 $^{-/-}$	ESF + TSA	DSC + TSA
MAMGYP-INT	0.08231	0.00309	0.00232	0.00481	0.11567	0.00232	0.60213	0.05195	0.07898
MER121	0.15433	0.01921	0.00232	0.00247	0.12692	0.00232	0.00278	0.91429	0.00398
MIR	0.00974	0.00309	0.00232	0.00247	0.00433	0.00448	0.00278	0.01824	0.84639
MIR3	0.00974	0.00309	0.00232	0.00247	0.94776	0.00232	0.00278	0.66093	0.07898
MIRB	0.00590	0.00309	0.00232	0.00247	0.00433	0.00232	0.00278	0.91429	0.41711
MIRC	0.00590	0.00309	0.00232	0.00247	0.75810	0.00232	0.00278	0.05530	0.01321
PLAT_L3	0.08231	0.00309	0.00232	0.00481	0.02525	0.00232	0.41226	0.21730	0.11022
L3	0.00590	0.00309	0.00232	0.00247	0.00433	0.00232	0.00278	0.01824	0.00398
L3B	0.08231	0.00573	0.00232	0.00247	0.00433	0.00232	0.00278	0.01824	0.05364
TIGGER15A	0.02964	0.00309	0.00232	0.00247	0.38690	0.00232	0.00278	0.01824	0.02165
X7A_LINE	0.04870	0.00309	0.00232	0.00247	0.23778	0.00232	0.00278	0.15126	0.00696
X7C_LINE	0.15433	0.00309	0.00906	0.00247	0.17736	0.00232	0.00278	0.03235	0.00398
CHARLIE1A	0.00590	0.00309	0.00232	0.00247	0.00433	0.00232	0.00278	0.01824	0.00398
CHARLIE1B	0.00590	0.00309	0.00232	0.00247	0.31338	0.00232	0.00278	0.09231	0.00696
LTR90A	0.00590	0.00309	0.00232	0.00247	0.00433	0.00232	0.00278	0.01824	0.00398
L2A	0.00590	0.00309	0.00232	0.00247	0.06273	0.00232	0.00278	0.02857	0.00398
L2B	0.00590	0.00309	0.00232	0.00247	0.09276	0.09519	0.00278	0.32143	0.00696
CHARLIE1	0.01855	0.00309	0.00232	0.00247	0.00433	0.00232	0.00278	0.05195	0.05364
CHARLIE18A	0.00590	0.00309	0.00232	0.00247	0.00433	0.00232	0.00278	0.91429	0.00398
L2	0.00974	0.00309	0.00232	0.00247	0.00433	0.00232	0.00278	0.01824	0.00398
L2C	0.00974	0.00309	0.00232	0.00247	0.06273	0.00232	0.00278	0.01824	0.00398
ARTHUR1C	0.00590	0.00309	0.00232	0.01643	0.38690	0.00448	0.50157	0.21730	0.00398
BLACKJACK	0.02964	0.00309	0.00232	0.00247	0.01528	0.00232	0.00278	0.05530	0.50740
CHARLIE15A	0.00590	0.03159	0.00232	0.00247	0.00433	0.00232	0.00278	0.02857	0.00398
CHARLIE16A	0.00590	0.00309	0.00232	0.00247	0.00433	0.00232	0.00278	0.41699	0.00398
CHARLIE19A	0.00590	0.00309	0.00232	0.00247	0.00433	0.00232	0.00278	0.05530	0.05364
CHARLIE24	0.00590	0.00309	0.00232	0.59536	0.00433	0.00232	0.00278	0.66093	0.07898
CHARLIE25	0.00590	0.00309	0.00232	0.00247	0.00433	0.00448	0.00278	0.01824	0.07898
CHARLIE2A	0.00974	0.00309	0.00232	0.00247	0.44879	0.00232	0.00278	0.54945	0.03542
CHARLIE2B	0.06984	0.00573	0.00232	0.00247	0.53872	0.00232	0.00506	0.01824	0.00398
CHARLIE4A	0.01855	0.00309	0.00232	0.00247	0.44879	0.00232	0.00278	0.05530	0.00696
CHARLIE4Z	0.00590	0.00309	0.00232	0.00247	0.38690	0.00232	0.00278	0.01824	0.00398

Table 2 (cont.)

CHARLIE5	0.00590	0.00309	0.00232	0.00247	0.44879	0.00232	0.00278	0.01824	0.00398
CHARLIE8	0.04870	0.00309	0.00232	0.00247	0.00433	0.00232	0.00278	0.01824	0.00398
FORD PREFECT	0.08231	0.00309	0.00232	0.00247	0.00829	0.00232	0.00506	0.05530	0.20211
L1MB3	0.00590	0.00309	0.00232	0.00247	0.00433	0.00232	0.00278	0.01824	0.00398
L1MC	0.71502	1.00000	0.00232	0.00247	0.44879	0.00232	0.00278	0.01824	0.00398
L1MC5	0.33972	0.00573	0.00232	0.00247	0.00433	0.00232	0.00278	0.21730	0.11022
L1MDB	0.08231	0.00309	0.00232	0.04354	0.86631	0.00232	0.01726	0.35225	0.00398
LTR41B	0.00590	0.00309	0.00232	0.00247	0.00433	0.00232	0.00278	0.02857	0.00398
LTR53	0.71502	0.72324	0.00232	0.69913	0.01528	0.00232	0.00278	0.63291	0.00398
LTR65	0.71502	0.04870	0.00232	0.00247	0.00433	0.00232	0.00278	0.03235	0.00398
LTR73	0.00590	0.00309	0.00232	0.00247	0.00433	0.00232	0.00278	0.15126	0.00398
MAMGYPLTR3	0.08231	0.00309	0.00232	0.00247	0.02525	0.00232	0.00506	0.01824	0.00398
MAMREP1879	0.08231	0.00309	0.00232	0.09740	0.06273	0.00232	0.00506	0.01824	0.00398
MAMREP4096	0.06984	0.01129	0.00232	0.00247	0.09276	0.00232	0.00278	0.09231	0.15235
MAMREP488	0.20729	0.03159	0.00232	0.00247	0.64618	0.00232	0.00278	0.15126	0.33563
MER102A	0.08231	0.00309	0.00232	0.00247	0.06273	0.00232	0.00278	0.79734	0.07898
MER102B	0.00590	0.00309	0.00232	0.00247	0.00433	0.00232	0.00278	0.01824	0.00398
MER110A	0.08231	0.07305	0.00232	0.00247	0.09276	0.00232	0.00278	0.66093	0.20211
MER112	0.00590	0.00309	0.00232	0.00247	0.00433	0.00232	0.00278	0.01824	0.03542
MER113	0.00590	0.00309	0.00232	0.00247	0.44879	0.00232	0.00278	0.01824	0.02165
MER113A	0.00974	0.00309	0.00232	0.00247	0.44879	0.00232	0.00278	0.05530	0.02165
MER115	0.33972	0.25742	0.00232	0.00247	0.00433	0.00232	0.00278	0.01824	0.00398
MER20	0.06984	0.00309	0.00232	0.00247	0.00829	0.00232	0.00278	0.03235	0.00398
MER21B	0.15433	0.07305	0.06566	0.00247	0.86631	0.00232	0.00278	0.79734	0.94776
MER33	0.62338	0.01921	0.00232	0.00247	0.00433	0.00232	0.00278	0.01824	0.00398
MER45A	0.06984	0.00309	0.00232	0.00247	0.04250	0.59536	0.00278	0.41729	1.00000
MER53	0.08231	0.00573	0.00232	0.00247	0.00433	0.00232	0.00278	0.09231	0.94776
MER58B	0.15433	1.00000	0.00232	0.00247	0.53872	0.00232	0.00278	0.01824	0.00398
MER5A	0.42208	0.25742	0.06566	0.00247	0.00433	0.00232	0.00506	0.01824	0.00398
MER5A1	0.08231	0.07305	0.00458	0.00247	0.00433	0.00232	0.00278	0.01824	0.00398
MER5B	0.08231	0.00309	0.00232	0.00247	0.00433	0.00232	0.00278	0.01824	0.00398
MER5C	0.08231	0.00309	0.00232	0.02783	0.00433	0.00232	0.07041	0.01824	0.01321
MER68	0.02964	0.00309	0.00232	0.00247	0.00433	0.00232	0.19249	0.01824	0.00696
MER89	0.08231	0.04870	0.00232	0.00247	0.00433	0.00232	0.70699	0.01824	0.00398
MER90	0.00590	0.00309	0.00232	0.00247	0.00433	0.00232	0.00278	0.01824	0.00398
MER90A	0.00590	0.00309	0.00232	0.00247	0.00433	0.00232	0.00278	0.01824	0.00398
MER92A	0.33972	0.10342	0.00232	0.00247	0.00433	0.00232	0.00278	0.01824	0.03542

Table 2 (cont.)

MER92C	0.08231	0.00309	0.00232	0.00247	1.00000	0.00232	0.00278	0.01824	0.00398
MER94B	0.02964	0.00309	0.00232	0.00247	0.00433	0.00232	0.00278	0.01824	0.41711
MER96	0.42208	1.00000	0.00232	0.01643	0.02525	0.00232	0.02922	0.01824	0.00398
MER96B	0.81818	0.19718	0.00232	0.00247	0.01528	0.00232	0.04570	0.01824	0.00398
MLT1-INT	0.00590	0.00309	0.00232	0.00247	0.00433	0.00232	0.00278	0.01824	0.00398
ORSL	0.00590	0.00309	0.00232	0.00247	0.00433	0.81818	0.00278	0.01824	0.00696
ORSL-2A	0.00974	0.00573	0.00232	0.24854	0.94776	0.00232	0.32773	0.54945	0.15235
ORSL-2B	0.00590	0.00309	0.00232	0.00247	0.12692	0.00232	0.00506	0.79734	0.26370
TIGGER12C	0.00590	0.00309	0.00232	0.00247	0.00433	0.00232	0.81818	0.05530	0.00398
ZAPHOD	0.06984	0.41226	0.13203	0.00247	0.44879	0.00232	0.07041	0.01824	0.00398
ALUYC	0.15433	0.00309	0.00232	0.00247	0.00433	0.00232	0.00278	0.01824	0.00398
MER41C	0.81818	0.00309	0.02687	0.00247	0.00433	0.00232	0.00278	0.01824	0.01321
MER41G	0.08231	0.00309	0.00232	0.00247	0.00433	0.00232	0.00278	0.01824	0.00398
LTR18B	0.33972	0.03159	0.00232	0.00247	0.94776	0.00232	0.00999	0.01824	0.00398
MER51C	0.08231	0.32773	0.00232	0.00247	0.00433	0.00232	0.00278	0.91429	0.02165
LTR10A	0.00590	0.00309	0.00232	0.00247	0.01528	0.00232	0.00506	0.01824	0.00398
LTR4	0.00590	0.00309	0.00232	0.00247	0.00433	0.00232	0.00278	0.01824	0.00398
LTR2B	0.00590	0.00309	0.00232	0.00247	0.00433	0.00232	0.00278	0.01824	0.00398
L1PA2	0.08231	0.00309	0.00232	0.00247	0.09276	0.00232	0.00278	0.66093	0.26370
SVA-F	0.11634	0.07305	0.00232	0.49587	0.94776	0.00232	0.00278	0.41729	0.00696

Chapter 5:

Discussion and Summary

Investigating how phenotypic novelties arise has driven many research programs and discoveries. One novelty of particular interest because it preserves step-wise evolutionary stages in extant species is mammalian pregnancy. Oviparous monotremes represent the early egg laying steps of pregnancy development [4-6], while marsupials have live birth with short gestations within a single estrus cycle [6-8], and viviparous eutherian mammals have many derived traits to support an extended gestational length that interrupts reproductive cycles [5,10]. Prior work has shown that changes in gene regulation has played a role in modulating pregnancy phenotypes, such as fecundability [42], and likely in the evolution of the trait itself [53]. My work has built upon these previous studies and highlights how changes to gene regulation can drive phenotypic change.

Modulation of the maternal immune system to tolerate the developing fetus is critical to a successful pregnancy [16,17]. Genes from the major histocompatability complex (MHC) are known to play an important role in the rejection of non-self tissues and also contribute to maternal fetal immunotolerance [22-41]. Burrows *et al* [42] identified single nucleotide polymorphisms that are expression quantitative trait loci (eQTL) for *TAP2* (rs2071473) and *HLA-F* (rs2523393), located in the MHC. These eQTLs are independently associated with fecundability, or the probability of getting pregnant in one reproductive cycle. However, how these SNPs altered gene expression and modified fecundability remain unclear.

The most significant association in Burrows *et al* [42] was rs2071473 ($P = 1.3\text{x}10^{-4}$), an eQTL associated with expression of the *antigen peptide transporter 2* (*TAP2)* gene (MIM: 170261) in the HLA class II region. TAP2, along with TAP1, forms a heterodimer and translocate peptides

from the cytosol to awaiting MHC class I molecules in the endoplasmic reticulum, resulting in cell surface presentation of the trimeric MHC complex to immune cells such as T lymphocytes and natural killer cells [43]. The C allele of rs2071473 was associated with longer intervals to pregnancy and higher expression of the *TAP2* gene in mid-secretory phase (receptive) endometrium. The median time to pregnancy, for example, was 2.0, 3.1, and 4.0 months among women with the TT, CT, and CC genotypes, respectively [42]. Driving the expression change seen to underlie this shift in phenotype, we found that the rs2071473 C allele abolishes binding by DDIT3, a dominant negative inhibitor of CEBP family transcription factors [96]. We also found that *Tap2* is highly expressed by DSCs at the maternal-fetal interface. Together, these data suggest that changes in TAP2 stoichiometry may alter MHC processing and thus interactions between DSCs and maternal immune cells. Furthermore, evidence of balancing selection at this locus suggests it is an example of a reproduction-health tradeoff in human evolution.

Similar to rs2071473, rs2523393 was also identified as an eQTL significantly associated with *HLA-F* expression and fecundability [42]. While thought to regulate immune responses and play an important role in maternal-fetal immunotolerance during pregnancy [130,131], the exact role of HLA-F is still unknown. HLA-F binds natural killer (NK) cell receptors from the LIR and KIR families [136-140], and given uterine natural killer (uNK) cells, which are essential for the establishment and maintenance of maternal immunotolerance and spiral artery remodeling, express KIR3DL1 and LIR2 it suggests uterine HLA-F mediates interactions with uNK during implantation, trophoblast invasion, and establishment of the uteroplacental circulation. Lower expression of *HLA-F* in mid-secretory phase (receptive) endometrium was associated with the G allele of rs2523393, as was a longer interval to pregnancy, whereas the A allele was associated with shorter intervals to pregnancy and higher *HLA-F* expression. The median time to pregnancy,

for example, was 2.3, 2.6, and 4.9 months among women with the AA, GA, and GG genotypes, respectively [22-24,42]. Underlying these shifts in expression and differences in fecundability, we observed that the rs2523393 A allele creates a new GATA2 responsive enhancer that loops to the *HLA-F* promoter. Similar to *TAP2*, evidence of balancing selection at this locus further strengthens the argument for a reproduction-health tradeoff in human evolution.

The work on rs2071473 and rs2523393 highlight how gene expression can change as genes gain and lose regulatory structures such as enhancers and repressors through a step-wise accumulation of small-scale mutations that create new or destroy old cis-regulatory elements. However, the integration and cooption of a transposable element (TE) that harbors functional transcription factor binding sites is another method by which cis-regulatory evolution could occur, and, due to the propagation of TEs throughout the genome, could lead to rapid evolution of novel gene regulatory networks. This hypothesis was originally proposed in the mid-20th century first by Barbara McClintock, then expanded on by Britten and Davidson [13,14]. Since then numerous studies have shown that transposable elements have donated binding sites for transcription factors to the genome, can be bound by transcription factors, and have been coopted into cis-regulatory elements [47-52] but many questions still remain. For example: Do TEs integrate with regulatory abilities, or as 'pre-regulatory elements' that need additional mutations to acquire regulatory functions?

Our work here builds upon the work done in Lynch *et al* [53], which explored the role of ancient mammalian TEs (Mammalian, Therian, and Eutherian specific) in the evolution of pregnancy, by expanding to look at all ages of TEs and assess their regulatory ability. We identified 352 TE families that are enriched within regions marked as having active regulatory function by DNaseI-seq, FAIRE-seq, H3K27ac ChIP-seq, and H3K4me3 ChIP-seq. Of these 352

families, most were Eutherian or Primate specific elements, suggesting multiple waves of TE cooption occurred in decidual tissues, particularly in primates. Furthermore, we found that genes with TE derived regulatory elements are more responsive to progesterone than genes without a TE derived cis-regulatory element, and that genes near young TEs with PGR binding sites are more progesterone responsive than genes near ancient TEs harboring PGR binding sites or genes without TE derived PGR binding sites. This suggests that TEs may have played a role in rewiring the progesterone responsive gene regulatory network underlying the evolution of pregnancy, but more work needs to be done.

From these 352 TE families, we selected 89 to explore if TEs integrate into the genome with regulatory functions and therefore immediately function as regulatory elements or if they integrate as 'pre-regulatory elements' that are weakly- or non-functional and require additional mutations to acquire regulatory ability. Our data suggests that the majority of TEs do integrate with regulatory ability but this ability is silenced, perhaps by KRAB-ZFPs and the NURD HDAC co-repressor complex. One caveat to this study, however, is that TE consensus sequences were used instead of true ancestral reconstructions because most TEs lack out-group representatives and we are unable to root TE phylogenies.

This dissertation aimed to address the questions 1) how TEs underlie expression shifts seen during the evolution of mammalian pregnancy and 2) what affect these shifts have had on pregnancy phenotypes. We demonstrated that there were likely two waves of TE cooption by the genome during the evolution of mammalian pregnancy, specifically in primates, and genes regulated by these TEs are highly progesterone responsive. Furthermore, the majority of the TEs tested likely integrated into the genome with regulatory ability that was silenced by KRAB-ZFPs and the NURD HDAC complex and only those that escaped said regulation drive the gene

expression patterns seen. Upon closer look at how modifications to gene regulatory elements affect pregnancy phenotypes, we showed that two eQTLs independently associated with fecundability create new regulatory structures that influence the expression of immune genes, which affect the pregnancy outcome. Despite these findings, there are still many lingering questions on the role of regulatory changes in the evolution of mammalian pregnancy. For example: Do TEs exert large or small effects on the genes they regulate? Do those effect sizes change over evolutionary time? What affect does shifting immune gene expression in uterine cells have on other pregnancy phenotypes, such as maternal recognition of pregnancy or placental invasion? Answers to these (and other) questions are essential for understanding how the evolution of gene regulation contributed to the evolution of phenotypic novelties.

References

1. McLean, C.Y. *et al.* Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature* **471**, 216-219 (2011).

2. Rausher, M.D. Evolutionary transitions in floral color. *International Journal of Plant Sciences* **169**, 7-21 (2008).

3. Chan, Y.F. *et al.* Adaptive Evolution of Pelvic Reduction in Sticklebacks by Recurrent Deletion of a Pitx1 Enhancer. *Science* **327**, 302-305 (2010).

4. Hughes, R. Monotreme development with particular reference to the extraembryonic membranes. *Journal of Experimental Zoology Part A: Ecological Genetics and Physiology* **266**, 480-494 (1993).

5. Hughes, R.L. & Hall, L.S. Early development and embryology of the platypus. *Philos Trans R Soc Lond B Biol Sci* **353**, 1101-14 (1998).

6. Renfree, M. & Shaw, G. Reproduction in Monotremes and Marsupials. (John Wiley & Sons, Ltd, 2001).

7. Renfree, M. Embryonic diapause in marsupials. *Journal of reproduction and fertility. Supplement* **29**, 67-78 (1981).

8. Renfree, M.B. Maternal recognition of pregnancy in marsupials. *Reviews of Reproduction* **5**, 6-11 (2000).

9. Frazer, J. & Huggett, A.S.G. Species variations in the foetal growth rates of eutherian mammals. *Journal of Zoology* **174**, 481-509 (1974).

10. Chavan, A.R., Bhullar, B.-A.S. & Wagner, G.P. What was the ancestral function of decidual stromal cells? A model for the evolution of eutherian pregnancy. *Placenta* **40**, 40-51 (2016).

11. Monod, J. & Jacob, F. General conclusions: teleonomic mechanisms in cellular metabolism, growth, and differentiation. in *Cold Spring Harbor symposia on quantitative biology* Vol. 26 389-401 (Cold Spring Harbor Laboratory Press, 1961).

12. King, M.C. & Wilson, A.C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107-116 (1975).

13. Britten, R.J. & Davidson, E.H. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *The Quarterly review of biology* **46**, 111-138 (1971).

14. McClintock, B. The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences* **36**, 344-355 (1950).

15. Rebeiz, M., Jikomes, N., Kassner, V.A. & Carroll, S.B. Evolutionary origin of a novel gene expression pattern through co-option of the latent activities of existing regulatory sequences. *Proceedings of the National Academy of Sciences* **108**, 10036-10043 (2011).

16. Medawar, P.B. Some immunological and endocrinological problems raised by the evolution of viviparity in vertebrates. *Symp. Soc. Exp. Biol.* **7**, 320-338 (1953).

17. Erlebacher, A. Immunology of the maternal-fetal interface. *Annual review of immunology* **31**, 387-411 (2013).

18. Gellersen, B. & Brosens, J. Cyclic AMP and progesterone receptor cross-talk in human endometrium: a decidualizing affair. *J Endocrinol* **178**, 357-72 (2003).

19. Gellersen, B., Brosens, I.A. & Brosens, J.J. Decidualization of the human endometrium: mechanisms, functions, and clinical perspectives. *Seminars in reproductive medicine* **25**, 445-453 (2007).

20. Murakami, K. *et al.* Decidualization induces a secretome switch in perivascular niche cells of the human endometrium. *Endocrinology* **155**, 4542-4553 (2014).

21. Dimitriadis, E., Sharkey, A.M., Tan, Y.L., Salamonsen, L.A. & Sherwin, J.R.A. Immunolocalisation of phosphorylated STAT3, interleukin 11 and leukaemia inhibitory factor in endometrium of women with unexplained infertility during the implantation window. *Reproductive Biology and Endocrinology* **5**, 44 (2007).

22. Kosova, G., Abney, M. & Ober, C. Colloquium papers: Heritability of reproductive fitness traits in a human population. *Proceedings of the National Academy of Sciences USA* **107**, 1772-1778 (2010).

23. Christensen, K., Kohler, H.P., Basso, O., Olsen, J., Vaupel, J.W., Rodgers, J.L. . The correlation of fecundability among twins: evidence of a genetic effect on fertility? *Epidemiology* **14**, 60-64 (2003).

24. Tropf, F.C., Verweij, R.M., van der Most, P.J. & Stulp, G. Mega-analysis of 31,396 individuals from 6 countries uncovers strong gene-environment interaction for human fertility. *bioRxiv* (2016).

25. Ober, C., Elias, S., Kostyu, D.D. & Hauck, W.W. Decreased fecundability in Hutterite couples sharing HLA-DR. *Am J Hum Genet* **50**, 6-14 (1992).

26. Ober, C., Hyslop, T., Elias, S., Weitkamp, L.R. & Hauck, W.W. Human leukocyte antigen matching and fetal loss: results of a 10 year prospective study. *Human reproduction* **13**,

33-38 (1998).

27. Fernandez, N., Cooper, J., Sprinks, M., AbdElrahman, M., Fiszer, D., Kurpisz, M., Dealtry, G. . A critical review of the role of the major histocompatibility complex in fertilization, preimplantation development and feto-maternal interactions. *Human Reproduction Update* **5**, 234-248 (1999).

28. Ober, C. *et al.* HLA sharing and fertility in Hutterite couples: evidence for prenatal selection against compatible fetuses. *American Journal of Reproductive Immunology* **18**, 111-115 (1988).

29. Ober, C.L. *et al.* Shared HLA antigens and reproductive performance among Hutterites. *Am J Hum Genet* **35**, 994-1004 (1983).

30. Ober, C. *et al.* Variation in the HLA-G promoter region influences miscarriage rates. *The American Journal of Human Genetics* **72**, 1425-1435 (2003).

31. Tan, C.Y. *et al.* Paternal contribution of HLA-G*0106 significantly increases risk for pre-eclampsia in multigravid pregnancies. *Molecular Human Reproduction* **14**, 317-324 (2008).

32. Loisel, D.A., Billstrand, C., Murray, K., Patterson, K., Chaiworapongsa, T., Romero, R., Ober, C. The maternal HLA-G 1597ΔC null mutation is associated with increased risk of pre-eclampsia and reduced HLA-G expression during pregnancy in African-American women. . *Molecular human reproduction* **19**, 144-152 (2012).

33. O'Brien, M. *et al.* Altered HLA-G transcription in pre-eclampsia is associated with allele specific inheritance: possible role of the HLA-G gene in susceptibility to the disease. *Cellular and Molecular Life Sciences* **58**, 1943-1949 (2001).

34. Larsen, M.H., Hylenius, S., Andersen, A.M. & Hviid, T.V. The 3'-untranslated region of the HLA-G gene in relation to pre-eclampsia: revisited. *Tissue Antigens* **75**, 253-61 (2010).

35. Moreau, P. *et al.* HLA-G gene polymorphism in human placentas: possible association of G*0106 allele with preeclampsia and miscarriage. *Biol Reprod* **79**, 459-67 (2008).

36. Yie, S., Li, L., Xiao, R. & Librach, C.L. A single base-pair mutation in the 3’-untranslated region of HLA-G mRNA is associated with pre-eclampsia. *Molecular human reproduction* **14**, 649-653 (2008).

37. Aldrich, C.L. *et al.* HLA-G genotypes and pregnancy outcome in couples with unexplained recurrent miscarriage. *Mol Hum Reprod* **7**, 1167-72 (2001).

38. Pfeiffer, K.A., Fimmers, R., Engels, G., van der Ven, H. & van der Ven, K. The HLA-G genotype is potentially associated with idiopathic recurrent spontaneous abortion.

Molecular Human Reproduction **7**, 373-378 (2001).

39. Suryanarayana, V. *et al.* Association Between Novel HLA-G Genotypes and Risk of Recurrent Miscarriages: A Case-Control Study in a South Indian Population. *Reproductive Sciences* **15**, 817-824 (2008).

40. Fan, W., Li, S.W., Huang, Z.Y. & Chen, Q. Relationship between HLA-G polymorphism and susceptibility to recurrent miscarriage: A meta-analysis of non-family-based studies. *Journal of Assisted Reproduction and Genetics* **31**, 173-184 (2014).

41. Hylenius, S., Nybo Andersen, A.M., Melbye, M. & Hviid, T.V.F. Association between HLA-G genotype and risk of pre-eclampsia: a case–control study using family triads. *Molecular human reproduction* **10**, 237-246 (2004).

42. Burrows, C.K. *et al.* Expression Quantitative Trait Locus Mapping Studies in Mid-Secretory Phase Endometrial Cells Identifies HLA-F and TAP2 as Fecundability-Associated Genes. *PLoS Genetics* **12**, e1005858 (2016).

43. Karttunen, J.T., Lehner, P.J., Gupta, S.S., Hewitt, E.W. & Cresswell, P. Distinct functions and cooperative interaction of the subunits of the transporter associated with antigen processing (TAP). *Proceedings of the National Academy of Sciences of the United States of America* **98**, 7431-7436 (2001).

44. Mika, K.M. & Lynch, V.J. An Ancient Fecundability-Associated Polymorphism Switches a Repressor into an Enhancer of Endometrial TAP2 Expression. *Am J Hum Genet* **99**, 1059-1071 (2016).

45. Feschotte, C. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* **9**, 397-405 (2008).

46. Rebollo, R., Romanish, M.T. & Mager, D.L. Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annu. Rev. Genet.* **46**, 21-42 (2012).

47. Jordan, I.K., Rogozin, I.B., Glazko, G.V. & Koonin, E.V. Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends in Genetics* **19**, 68-72 (2003).

48. El-Deiry, W., Kern, S. & Pietenpol, J. p53 binding sites in transposons. *Nature Genet.* **1**, 45-49 (1992).

49. Polak, P. & Domany, E. Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. *BMC genomics* **7**, 133 (2006).

50. Lynch, V.J., Leclerc, R.D., May, G. & Wagner, G.P. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nature genetics* **43**, 1154-1159 (2011).

51. Kunarso, G. *et al.* Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet* **42**, 631-4 (2010).

52. Wang, T. *et al.* Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proceedings of the National Academy of Sciences* **104**, 18613-18618 (2007).

53. Lynch, V.J. *et al.* Ancient Transposable Elements Transformed the Uterine Regulatory Landscape and Transcriptome during the Evolution of Mammalian Pregnancy. *Cell reports* **10**, 551-61 (2015).

54. Ruth, K.S. *et al.* Genetic evidence that lower circulating FSH levels lengthen menstrual cycle, increase age at menopause and impact female reproductive health. *Human reproduction (Oxford, England)* **31**, 473-481 (2016).

55. Elks, C.E. *et al.* Thirty new loci for age at menarche identified by a meta-analysis of genome-wide association studies. *Nat Genet* **42**, 1077-85 (2010).

56. He, C., Kraft, P., Chen, C., Buring, J.E., Paré, G., Hankinson, S.E., Chanock, S.J., Ridker, P.M., Hunter, D.J., and Chasman, D.I. . Genome-wide association studies identify loci associated with age at menarche and age at natural menopause. *Nature Genetics* **41**, 724-728 (2009).

57. Perry, J.R.B. *et al.* Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature* **514**, 92-97 (2014).

58. Stolk, L. *et al.* Meta-analyses identify 13 loci associated with age at menopause and highlight DNA repair and immune pathways. *Nature genetics* **44**, 260-268 (2012).

59. Chen, C.T.L. *et al.* Meta-analysis of loci associated with age at natural menopause in African-American women. *Human Molecular Genetics* **23**, 3327-3342 (2014).

60. Mbarek, H. *et al.* Identification of Common Genetic Variants Influencing Spontaneous Dizygotic Twinning and Female Fertility. *Am J Hum Genet* **98**, 898-908 (2016).

61. Aschebrook-Kilfoy, B. *et al.* Genome-wide association study of parity in Bangladeshi women. *PloS one* **10**, e0118488 (2015).

62. Schuh-Huerta, S.M. *et al.* Genetic variants and environmental factors associated with hormonal markers of ovarian reserve in Caucasian and African American women. *Human reproduction* **27**, 594-608 (2012).

63. Consortium, G. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648-660 (2015).

64. Melé, M. *et al.* Human genomics. The human transcriptome across tissues and individuals. *Science* **348**, 660-665 (2015).

65. Uhlen, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).

66. Talbi, S. *et al.* Molecular phenotyping of human endometrium distinguishes menstrual cycle phases and underlying biological processes in normo-ovulatory women. *Endocrinology* **147**, 1097-1121 (2006).

67. Goldfien, G.A. *et al.* Progestin-Containing Contraceptives Alter Expression of Host Defense-Related Genes of the Endometrium and Cervix. *Reproductive Sciences* **22**, 814-828 (2015).

68. Pabona, J.M.P. *et al.* Krüppel-like factor 9 and progesterone receptor coregulation of decidualizing endometrial stromal cells: implications for the pathogenesis of endometriosis. *The Journal of clinical endocrinology and metabolism* **97**, E376-92 (2012).

69. Davis, S. & Meltzer, P.S. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* **23**, 1846-1847 (2007).

70. Smyth, G.K. Limma: linear models for microarray data. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor. (Springer New York)*, 397-420 (2005).

71. Gentleman, R., Carey, V., Huber, W., Irizarry, R. & Dudoit, S. *Bioinformatics and computational biology solutions using R and Bioconductor*, (Springer New York, New York, NY, 2006).

72. Tamura, I. *et al.* Genome-wide analysis of histone modifications in human endometrial stromal cells. *Mol. Endocrinol.* **28**, 1656-1669 (2014).

73. Delport, W., Poon, A.F.Y., Frost, S.D.W. & Kosakovsky Pond, S.L. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics* **26**, 2455-2457 (2010).

74. Kosakovsky Pond, S. & Frost, S. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Molecular biology and evolution* **22**, 1208-1222 (2005).

75. Wong, W.S.W. & Nielsen, R. Detecting selection in noncoding regions of nucleotide sequences. *Genetics* **167**, 949-958 (2004).

76. Zhang, J., Nielsen, R. & Yang, Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular biology and evolution* **22**, 2472-2479 (2005).

77. Haygood, R., Fedrigo, O., Hanson, B., Yokoyama, K.D., Wray, G.A. . Promoter regions of many neural-and nutrition-related genes have experienced positive selection during human evolution. *Nature Genetics* **39**, 1140-1144 (2007).

78. Kosakovsky Pond, S.L., Frost, S.D.W. & Muse, S.V. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21**, 676-679 (2005).

79. Pybus, M., Dall'Olio, G.M., Luisi, P., Uzkudun, M., Carreño-Torres, A., Pavlidis, P., Laayouni, H., Bertranpetit, J., Engelken, J. 1000 Genomes Selection Browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans. *Nucleic Acids Research* **42**, D903-D909 (2014).

80. Barragan, F. *et al.* Human Endometrial Fibroblasts Derived from Mesenchymal Progenitors Inherit Progesterone Resistance and Acquire an Inflammatory Phenotype in the Endometrial Niche in Endometriosis. *Biology of reproduction* **94**, 118 (2016).

81. Felker, A.M. & Croy, B.A. Uterine natural killer cell partnerships in early mouse decidua basalis. *Journal of leukocyte biology* **100**, 645-655 (2016).

82. Svensson, J. *et al.* Macrophages at the fetal-maternal interface express markers of alternative activation and are induced by M-CSF and IL-10. *Journal of immunology (Baltimore, Md. : 1950)* **187**, 3671-3682 (2011).

83. De, M., Sanford, T. & Wood, G.W. Relationship between macrophage colony-stimulating factor production by uterine epithelial cells and accumulation and distribution of macrophages in the uterus of pregnant mice. *Journal of leukocyte biology* **53**, 240-248 (1993).

84. Exley, M.A. & Boyson, J.E. Protective role of regulatory decidual γδ T cells in pregnancy. *Clinical immunology* **141**, 236-239 (2011).

85. Erlebacher, A. Mechanisms of T cell tolerance towards the allogeneic fetus. *Nature reviews. Immunology* **13**, 23-33 (2013).

86. Saito, S., Nakashima, A., Shima, T. & Ito, M. Th1/Th2/Th17 and Regulatory T-Cell Paradigm in Pregnancy. *American Journal of Reproductive Immunology* **63**, 601-610 (2010).

87. Reinhard, G., Noll, A., Schlebusch, H., Mallmann, P. & Ruecker, A.V. Shifts in the TH1/TH2 balance during human pregnancy correlate with apoptotic changes. *Biochemical and Biophysical Research Communications* **245**, 933-938 (1998).

88. Tagliani, E. & Erlebacher, A. Dendritic cell function at the maternal-fetal interface. *Expert review of clinical immunology* **7**, 593-602 (2011).

89. Collins, M.K., Tay, C.S., and Erlebacher, A. Dendritic cell entrapment within the pregnant uterus inhibits immune surveillance of the maternal/fetal interface in mice. *J Clin Invest* **119**, 2062-73 (2009).

90. Mazur, E.C. *et al.* Progesterone receptor transcriptome and cistrome in decidualized human endometrial stromal cells. *Endocrinology* **156**, 2239-2253 (2015).

91. Kaya, H.S. *et al.* Roles of progesterone receptor A and B isoforms during human endometrial decidualization. *Mol. Endocrinol.* **29**, 882-895 (2015).

92. Li, X. *et al.* COUP-TFII regulates human endometrial stromal genes involved in inflammation. *Mol. Endocrinol.* **27**, 2041-2054 (2013).

93. Sandelin, A., Wasserman, W.W. & Lenhard, B. ConSite: web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Research* **32**, W249-52 (2004).

94. Mathelier, A. *et al.* JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **44**, D110-5 (2016).

95. Zhou, J. & Troyanskaya, O.G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature methods* **12**, 931-934 (2015).

96. Ron, D. & Habener, J.F. CHOP, a novel developmentally regulated nuclear protein that dimerizes with transcription factors C/EBP and LAP and functions as a dominant-negative inhibitor of gene transcription. *Genes and Development* **6**, 439-453 (1992).

97. Jauhiainen, A. *et al.* Distinct cytoplasmic and nuclear functions of the stress induced protein DDIT3/CHOP/GADD153. *PloS one* **7**, e33208 (2012).

98. Gao, H. & Schwartz, R.C. C/EBPzeta (CHOP/Gadd153) is a negative regulator of LPS-induced IL-6 expression in B cells. *Molecular Immunology* **47**, 390-397 (2009).

99. Pomerance, M. *et al.* CCAAT/enhancer-binding protein-homologous protein expression and transcriptional activity are regulated by 3',5'-cyclic adenosine monophosphate in thyroid cells. *Mol. Endocrinol.* **17**, 2283-2294 (2003).

100. Hedrick, P.W. & Thomson, G. Evidence for balancing selection at HLA. *Genetics* **104**, 449-456 (1983).

101. Bubb, K.L. *et al.* Scan of human genome reveals no new Loci under ancient balancing selection. *Genetics* **173**, 2165-2177 (2006).

102. Andrés, A.M. *et al.* Targets of balancing selection in the human genome. *Molecular biology and evolution* **26**, 2755-2764 (2009).

103. Cagliani, R. *et al.* Balancing selection is common in the extended MHC region but most alleles with opposite risk profile for autoimmune diseases are neutrally evolving. *BMC evolutionary biology* **11**, 171 (2011).

104. DeGiorgio, M., Lohmueller, K.E. & Nielsen, R. A model-based approach for identifying signatures of ancient balancing selection in genetic data. *PLoS genetics* **10**, e1004561 (2014).

105. Wong, W.S.W. & Nielsen, R. Detecting selection in non-coding regions of nucleotide sequences. *Genetics* **167**, 949-958 (2004).

106. Erlebacher, A. Why isn't the fetus rejected? *Curr Opin Immunol* **13**, 590-3 (2001).

107. Moffett, A. & Loke, Y.W. The immunological paradox of pregnancy: a reappraisal. *Placenta* **25**, 1-8 (2004).

108. Guerin, L.R., Prins, J.R. & Robertson, S.A. Regulatory T-cells and immune tolerance in pregnancy: a new target for infertility treatment? *Hum Reprod Update* **15**, 517-35 (2009).

109. Tirado-González, I., Barrientos, G., Freitag, N., Otto, T., Thijssen, V.L., Moschansky, P., von Kwiatkowski, P., Klapp, B.F., Winterhager, E., Bauersachs, S., Blois, S.M. . Uterine NK cells are critical in shaping DC immunogenic functions compatible with pregnancy progression. *PloS one* **7**, e46755 (2012).

110. Nagamatsu, T. & Schust, D.J. Review: The Immunomodulatory Roles of Macrophages at the Maternal—Fetal Interface. *Reproductive Sciences* (2010).

111. Nancy, P. *et al.* Chemokine Gene Silencing in Decidual Stromal Cells Limits T Cell Access to the Maternal-Fetal Interface. *Science* **336**, 1317-1321 (2012).

112. Nagamatsu, T., Schust, D.J., Sugimoto, J. & Barrier, B.F. Human decidual stromal cells suppress cytokine secretion by allogenic CD4+ T cells via PD-1 ligand interactions. *Human reproduction* **24**, 3160-3171 (2009).

113. Komatsu, T. *et al.* Expression of class I human leukocyte antigen (HLA) and beta2-microglobulin is associated with decidualization of human endometrial stromal cells. *Human reproduction* **13**, 2246-2251 (1998).

114. Cromme, F.V. *et al.* Loss of transporter protein, encoded by the TAP-1 gene, is highly correlated with loss of HLA expression in cervical carcinomas. *J Exp Med* **179**, 335-40 (1994).

115. Van Kaer, L., Ashton-Rickardt, P.G., Ploegh, H.L. & Tonegawa, S. TAP1 mutant mice are deficient in antigen presentation, surface class I molecules, and CD4-8+ T cells. *Cell* **71**, 1205-1214 (1992).

116. Raghavan, M. Immunodeficiency due to defective antigen processing: the molecular basis for type 1 bare lymphocyte syndrome. *J. Clin. Invest.* **103**, 595-596 (1999).

117. Oliveira, C.C. *et al.* Peptide transporter TAP mediates between competing antigen sources generating distinct surface MHC class I peptide repertoires. *European journal of immunology* **41**, 3114-3124 (2011).

118. Durgeau, A. *et al.* Different expression levels of the TAP peptide transporter lead to recognition of different antigenic peptides by tumor-specific CTL. *Journal of immunology* **187**, 5532-5539 (2011).

119. Blanco, O. *et al.* Human decidual stromal cells express HLA-G: Effects of cytokines and decidualization. *Hum Reprod* **23**, 144-52 (2008).

120. Reyes-Perdomo, C., Torres, K., Olivares, E.G. & Silva, P. Frontiers | Immunoregulatory activity of Decidual/Endometrial Stromal Cells in humans. *frontiersin.org*.

121. Black, F.L. & Hedrick, P.W. Strong balancing selection at HLA loci: evidence from segregation in South Amerindian families. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 12452-12456 (1997).

122. Markow, T. *et al.* HLA polymorphism in the Havasupai: evidence for balancing selection. *American journal of Human Genetics* **53**, 943 (1993).

123. Tan, Z., Shon, A.M. & Ober, C. Evidence of balancing selection at the HLA-G promoter region. *Human Molecular Genetics* (2005).

124. Jeffreys, A.J., Ritchie, A. & Neumann, R. High resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hotspot. *Human Molecular Genetics* **9**, 725-733 (2000).

125. Carter, A.J.R. & Nguyen, A.Q. Antagonistic pleiotropy as a widespread mechanism for the maintenance of polymorphic disease alleles. *BMC medical genetics* **12**, 160 (2011).

126. Anderson, C.A. *et al.* Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nature genetics* **43**, 246-252 (2011).

127. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature genetics* **42**, 1118-1125 (2010).

128. Hofmann, S. *et al.* Genome-wide association analysis reveals 12q13.3-q14.1 as new risk locus for sarcoidosis. *The European respiratory journal* **41**, 888-900 (2013).

129. Fischer, A. *et al.* Identification of Immune-Relevant Factors Conferring Sarcoidosis Genetic Risk. *American journal of respiratory and critical care medicine* **192**, 727-736 (2015).

130. Ishitani, A., Sageshima, N. & Hatake, K. The involvement of HLA-E and -F in pregnancy. *Journal of reproductive immunology* **69**, 101-113 (2006).

131. Persson, G., Melsted, W.N., Nilsson, L.L. & Hviid, T.V.F. HLA class Ib in pregnancy and pregnancy-related disorders. *Immunogenetics* **69**, 581-595 (2017).

132. Hackmon, R. *et al.* Definitive class I human leukocyte antigen expression in gestational placentation: HLA-F, HLA-E, HLA-C, and HLA-G in extravillous trophoblast invasion on placentation, pregnancy, and parturition. *American Journal of Reproductive Immunology* **77**, e12643-n/a (2017).

133. Ishitani, A. *et al.* Protein Expression and Peptide Binding Suggest Unique and Interacting Functional Roles for HLA-E, F, and G in Maternal-Placental Immune Recognition. *The Journal of Immunology* **171**, 1376-1384 (2003).

134. Nagamatsu, T. *et al.* Human leukocyte antigen F protein is expressed in the extra-villous trophoblasts but not on the cell surface of them. *American journal of reproductive immunology* **56**, 172-177 (2006).

135. Shobu, T. *et al.* The surface expression of HLA-F on decidual trophoblasts increases from mid to term gestation. *Journal of reproductive immunology* **72**, 18-32 (2006).

136. Lepin, E.J. *et al.* Functional characterization of HLA-F and binding of HLA-F tetramers to ILT2 and ILT4 receptors. *Eur J Immunol* **30**, 3552-61 (2000).

137. Dulberger, C.L. *et al.* Human Leukocyte Antigen F Presents Peptides and Regulates Immunity through Interactions with NK Cell Receptors. *Immunity* **46**, 1018-1029.e7 (2017).

138. Goodridge, J.P., Burian, A., Lee, N. & Geraghty, D.E. HLA-F and MHC class I open conformers are ligands for NK cell Ig-like receptors. *The Journal of Immunology* **191**, 3553-3562 (2013).

139. Garcia-Beltran, W.F. *et al.* Open conformers of HLA-F are high-affinity ligands of the activating NK-cell receptor KIR3DS1. *Nature immunology* **17**, 1067-1074 (2016).

140. Burian, A. *et al.* HLA-F and MHC-I Open Conformers Bind Natural Killer Cell Ig-Like Receptor KIR3DS1. *PloS one* **11**, e0163297 (2016).

141. Kofod, L. *et al.* Endometrial immune markers are potential predictors of normal fertility and pregnancy after in vitro fertilization. *American Journal of Reproductive Immunology* **78**(2017).

142. Lédée, N. *et al.* Specific and extensive endometrial deregulation is present before conception in IVF/ICSI repeated implantation failures (IF) or recurrent miscarriages. *The Journal of Pathology* **225**, 554-564 (2011).

143. Johnson, W.E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118-27 (2007).

144. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B-Methodological* **57**, 289-300 (1995).

145. Kurihara, I. *et al.* COUP-TFII mediates progesterone regulation of uterine implantation by controlling ER activity. *PLoS genetics* **3**, e102 (2007).

146. He, B. *et al.* GATA2 facilitates steroid receptor coactivator recruitment to the androgen receptor complex. *Proc Natl Acad Sci U S A* **111**, 18261-6 (2014).

147. Rubel, C.A., Franco, H.L., Jeong, J.W., Lydon, J.P. & DeMayo, F.J. GATA2 is expressed at critical times in the mouse uterus during pregnancy. *Gene Expr Patterns* **12**, 196-203 (2012).

148. Rubel, C.A. *et al.* Gata2 Is a Master Regulator of Endometrial Function and Progesterone Signaling. *Biology of Reproduction* **85**(2011).

149. Adams, E.J. & Parham, P. Species-specific evolution of MHC class I genes in the higher primates. *Immunological reviews* **183**, 41-64 (2001).

150. Shiina, T. *et al.* Molecular dynamics of MHC genesis unraveled by sequence analysis of the 1,796,938-bp HLA class I region. *Proceedings of the National Academy of Sciences* **96**, 13282-13287 (1999).

151. Meyer, M. *et al.* Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins. *Nature* **531**, 504-507 (2016).

152. Dai, X.-M., Zong, X.-H., Sylvestre, V. & Stanley, E.R. Incomplete restoration of colony-stimulating factor 1 (CSF-1) function in CSF-1-deficient Csf1op/Csf1op mice by transgenic expression of cell surface CSF-1. *Blood* **103**, 1114-1123 (2004).

153. Ryan, G.R. *et al.* Rescue of the colony-stimulating factor 1 (CSF-1)-nullizygous mouse (Csf1(op)/Csf1(op)) phenotype with a CSF-1 transgene and identification of sites of local CSF-1 synthesis. *Blood* **98**, 74-84 (2001).

154. Ovadia, S., Insogna, K. & Yao, G.-Q. The cell-surface isoform of colony stimulating factor 1 (CSF1) restores but does not completely normalize fecundity in CSF1-deficient mice. *Biology of reproduction* **74**, 331-336 (2006).

155. Ashkar, A.A. *et al.* Assessment of requirements for IL-15 and IFN regulatory factors in uterine NK cell differentiation and function during pregnancy. *Journal of immunology* **171**, 2937-2944 (2003).

156. Allavena, P., Giardina, G., Bianchi, G. & Mantovani, A. IL-15 is chemotactic for natural killer cells and stimulates their adhesion to vascular endothelium. *Journal of leukocyte biology* **61**, 729-735 (1997).

157. Laskarin, G. *et al.* Physiological role of IL-15 and IL-18 at the maternal-fetal interface. *Chemical immunology and allergy* **89**, 10-25 (2005).

158. Barber, E.M. & Pollard, J.W. The uterine NK cell population requires IL-15 but these cells are not required for pregnancy nor the resolution of a Listeria monocytogenes infection. (2003).

159. Kitaya, K. Central Role of Interleukin-15 in Postovulatory Recruitment of Peripheral Blood CD16(-) Natural Killer Cells into Human Endometrium. *The Journal of clinical endocrinology and metabolism* **90**, 2932-2940 (2005).

160. Gu, L. *et al.* Control of TH2 polarization by the chemokine monocyte chemoattractant protein-1. *Nature* **404**, 407-411 (2000).

161. De Jager, P.L. *et al.* Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. *Nature genetics* **41**, 776-782 (2009).

162. Stone, J.R. & Wray, G.A. Rapid evolution of cis-regulatory sequences via local point mutations. *Molecular biology and evolution* **18**, 1764-1770 (2001).

163. Bourque, G. *et al.* Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Research* **18**, 1752-1762 (2008).

164. Wray, G.A. The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet* **8**, 206-216 (2007).

165. Schultz, D.C., Friedman, J.R. & Rauscher, F.J. Targeting histone deacetylase complexes via KRAB-zinc finger proteins: the PHD and bromodomains of KAP-1 form a cooperative unit that recruits a novel isoform of the Mi-2α subunit of NuRD. *Genes & development* **15**, 428-443 (2001).

166. Feschotte, C. & Gilbert, C. Endogenous viruses: insights into viral evolution and impact

on host biology. *Nature Reviews Genetics* **13**, 283 (2012).

167. Wolf, D. & Goff, S.P. Embryonic stem cells use ZFP809 to silence retroviral DNAs. *Nature* **458**, 1201 (2009).

168. Castro-Diaz, N. *et al.* Evolutionally dynamic L1 regulation in embryonic stem cells. *Genes & development* **28**, 1397-1409 (2014).

169. Jacobs, F.M. *et al.* An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature* **516**, 242 (2014).

170. Schultz, D.C., Ayyanathan, K., Negorev, D., Maul, G.G. & Rauscher, F.J. SETDB1: a novel KAP-1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated silencing of euchromatic genes by KRAB zinc-finger proteins. *Genes & development* **16**, 919-932 (2002).

171. Matsui, T. *et al.* Proviral silencing in embryonic stem cells requires the histone methyltransferase ESET. *Nature* **464**, 927 (2010).

172. Quenneville, S. *et al.* The KRAB-ZFP/KAP1 system contributes to the early embryonic establishment of site-specific DNA methylation patterns maintained during development. *Cell reports* **2**, 766-773 (2012).

173. Imbeault, M., Helleboid, P.-Y. & Trono, D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* **543**, 550 (2017).

174. Yang, P., Wang, Y. & Macfarlan, T.S. The Role of KRAB-ZFPs in Transposable Element Repression and Mammalian Evolution. *Trends in Genetics* (2017).

175. Rowe, H.M. *et al.* KAP1 controls endogenous retroviruses in embryonic stem cells. *Nature* **463**, 237 (2010).

176. Ecco, G. *et al.* Transposable elements and their KRAB-ZFP controllers regulate gene expression in adult tissues. *Developmental cell* **36**, 611-623 (2016).

177. Wolf, G. *et al.* The KRAB zinc finger protein ZFP809 is required to initiate epigenetic silencing of endogenous retroviruses. *Genes & development* **29**, 538-554 (2015).

178. Giudice, L.C. Elucidating endometrial function in the post-genomic era. *Hum Reprod Update* **9**, 223-35 (2003).

179. Mess, A. & Carter, A.M. Evolutionary transformation of fetal membrane characters in Eutheria with special reference to Afrotheria. *J Exp Zool Part B (Mol Dev Evol)* **306B**, 140-163 (2006).

180. Kin, K., Nnamani, M.C., Lynch, V.J., Michaelides, E. & Wagner, G.P. Cell-type Phylogenetics and the Origin of Endometrial Stromal Cells. *Cell reports* **10**, 1398-409 (2015).

181. Kin, K., Maziarz, J. & Wagner, G.P. Immunohistological Study of the Endometrial Stromal Fibroblasts in the Opossum, Monodelphis domestica: Evidence for Homology with Eutherian Stromal Fibroblasts. *Biology of reproduction* **90**, 111 (2014).

182. Romer, A.S. & Parsons, T.S. *The Vertebrate Body*, (Harcourt Brace Jovanovich College Publishers, New York, 1997).

183. Szalay, F.S., Novacek, M.J. & McKenna, M.C. *Mammal phylogeny: Mesozoic differentiation, multituberculates, monotremes, early therians, and marsupials*, (Springer Science & Business Media, 2012).

184. Wagner, G.P. & Lynch, V.J. Molecular evolution of evolutionary novelties: the vagina and uterus of therian mammals. *Journal of experimental zoology. Part B, Molecular and developmental evolution* **304**, 580-592 (2005).

185. Renfree, M.B. Influence of the embryo on the marsupial uterus. *Nature* **240**, 475 (1972).

186. Moffett-King, A. Natural killer cells and pregnancy. *Nature Reviews Immunology* **2**, 656 (2002).

187. Wang, F., Tian, Z. & Wei, H. Genomic expression profiling of NK cells in health and disease. *European journal of immunology* **45**, 661-678 (2015).

188. Pardo, J. *et al.* The biology of cytotoxic cell granule exocytosis pathway: granzymes have evolved to induce cell death and inflammation. *Microbes and infection* **11**, 452-459 (2009).

189. Bray, N.L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology* **34**, 525-527 (2016).

190. Wagner, G.P., Kin, K. & Lynch, V.J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci* **131**, 281-5 (2012).

191. Robinson, M.D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology* **11**, R25 (2010).

192. Wagner, G.P., Kin, K. & Lynch, V.J. A model based criterion for gene expression calls using RNA-seq data. *Theory in Biosciences* **132**, 159-164 (2013).

193. Hebenstreit, D. *et al.* RNA sequencing reveals two major classes of gene expression levels in metazoan cells. *Mol Syst Biol* **7**, 497 (2011).

194. Mazur, E.C. *et al.* Progesterone receptor transcriptome and cistrome in decidualized human endometrial stromal cells. *Endocrinology* **156**, 2239-2253 (2015).

195. Wheeler, T.J. *et al.* Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Research* **41**, D70-82 (2013).

196. Song, L. *et al.* Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome research* **21**, 1757-1767 (2011).

197. Addison, J.B. *et al.* KAP1 promotes proliferation and metastatic progression of breast cancer cells. *Cancer research* **75**, 344-355 (2015).