Characterization of Recent Evolutionary Signatures of Genetic Elements involved in Human Pregnancy

By

Jiyun Michelle Moon

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Biological Sciences

May 31, 2019

Nashville, Tennessee

Approved:

Julián F. Hillyer, Ph.D.
Antonis Rokas, Ph.D.
Douglas K. Abbot, Ph.D.
John A. Capra, Ph.D.
Maulik Patel, Ph.D.
David M. Aronoff, M.D.

To Kevin Kelly, my loving and supportive fiancé,

and

Eevee, our lovely corgi girl, who has always been there for licks and snuggles

throughout most of my graduate training

ACKNOWLEDGMENTS

I would first like to thank the Vanderbilt University Department of Biological Sciences and the administrative staff, including Leslie Maxwell, Angela Titus, Christopher Patterson, LaDonna Smith and Alicia Goostree for providing me with great administrative help whenever I needed it.

My research has greatly benefited from the input of my committee members and advisors. Thanks so much to Julián Hillyer, David Aronoff, Tony Capra, and Maulik Patel for serving as members in my committee (and Julián for serving as my chair), and Antonis Rokas and Patrick Abbot for providing valuable guidance as my mentors, both in terms of science and life in general.

I would also like to thank the past and present lab members of the Rokas lab, especially Juan Felipé Ortiz, for being both a very helpful lab mate as well as a wonderful friend. Juan, thanks so much for our discussions of science and life over coffee.

I would also like to thank friends from my supportive cohort, Stephen Ferguson, Ling Chen, Juan Felipé Ortiz, and especially Stephanie Birnbaum, who’s been a wonderful lab mate, great friend, as well as an inspiring role model for me as a scientist. Stephanie, thanks for all the hugs and words of encouragement whenever I was

feeling down.

My friends and family members also have been there during the whole graduate training, and they have been of great support. My close friends Ji-in and Hyeon-suk from South Korea, who have always been there for cheer me up and make me laugh, despite the 15-hour time difference. You guys are the best and I miss you guys a lot. I couldn't have done this without you.

Thank you, mom and dad, grandmother, and my younger sister Diane for always praying for my sake.

I am also very blessed to have a loving and supportive fiancé and future in-laws, who have accepted me into the family enthusiastically and have emotionally supported my endeavors. Thank you so much Barbara and Tom, for being amazing soon-to-be parents-in-laws, and Chris, my soon-to-be brother-in-law who always makes me smile.

Finally, Kevin, thanks so much for being a stalwart supporting tree during the past four years. You have always been by my side no matter what, and I could not have done this without your unwavering support and love. I can't wait to be your supporting tree when you're going through your dissertation soon! And of course, Eevee, our baby corgi girl for always making me smile even when I'm having a bad day.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER IV

LIST OF FIGURES

CHAPTER I

Figure Page

CHAPTER II

CHAPTER III

CHAPTER IV

CHAPTER I

Introduction[1]

Preface

Evolution of mammalian pregnancy is characterized by several novelties. For instance, approximately 200 million years ago, “proto” trophoblasts, or the cells of outer layer of the mammalian blastocyst that develops into fetal membranes, evolved in the ancestors of mammals (Chuong *et al.* 2013; Renfree and Shaw 2001)In addition, evolution of viviparity or live-bearing in occurred in therian mammals (Killian *et al.* 2001; Baker *et al.* 2004; Griffith and Wagner 2017), as well as, emergence of a novel organ - the placenta - in therian mammals and further elaboration in eutherian mammals. Eutherian mammals are further distinguished from other groups of mammals by the occurrence of direct maternal-fetal contact early on in pregnancy (Griffith *et al.* 2017). This contact is further exaggerated in mammals that form invasive placentas (i.e., hemochorial mammals), such as rodents and primates, in which a population of trophoblasts invade into the maternal decidua (Carter and Enders 2004; Robbins and Bakardjiev 2012; Furukawa *et al.* 2014). Such formation of the invasive placentas in some eutherian mammals likely required novel mechanisms by which to regulate the extent of trophoblast invasion into the

[1] The Contributing authors are: Jiyun M. Moon, John A. Capra, Patrick Abbot, and Antonis Rokas.

maternal decidua, as both excessive and insufficient invasion are associated with adverse pregnancy outcomes (Hiby *et al.* 2004; Hiby *et al.* 2010; Hiby *et al.* 2014).

Mammalian pregnancy entails co-existence of a genetically distinct entity (i.e., the fetus) within the mother, and therefore, is an intrinsically immunological process. The relatively extended gestational period, as well as direct maternal-fetal contact starting early in gestation creates an even more difficult challenge to eutherian mothers (Griffith *et al.* 2017) - to sustain a genetically distinct fetus without inducing an immunological reaction and at the same time, maintain optimal host defense against pathogens. Therefore, it is reasonable to assume that some of the novel mechanisms that enabled successful establishment of pregnancy in eutherian mammals involves immune modulation (Griffith *et al.* 2017). In this regard, while it was previously believed that the maternal immune system experiences global suppression to prevent immunological rejection of the fetus (Billington 2003), recent studies have shown that distinct aspects of the maternal immune system are differentially regulated both spatially (i.e., locally versus systematically) (Veenstra van Nieuwenhoven *et al.* 2003; Svensson-Arvelund *et al.* 2013) and temporally (i.e., in different trimesters) (Luppi 2003; Somerset *et al.* 2004; Germain *et al.* 2007; Bartmann *et al.* 2013; Aghaeepour *et al.* 2017), resulting in a functionally active, yet carefully regulated immune system throughout the duration of pregnancy.

In this review, we will discuss the evolution of mammalian, and more specifically, eutherian pregnancy. We will first briefly discuss what sets apart eutherian pregnancy by drawing comparisons with pregnancy in a closely related lineage, marsupials. We will follow this up with a short summary describing the diversity that exists within eutherians. Next, we will outline the global mechanisms underlying the evolutionary innovations that occurred with the emergence of eutherian mammals. Next, we focus on the evolution of two immunity gene families in particular: killer immunoglobulin-like receptors (KIRs) and sialic-acid binding immunoglobulin-like lectins (SIGLECs). We will briefly discuss the global processes that led to the emergence of these two gene families and conclude with the discussion of how differential advantages in terms of reproductive success and host defense activities have likely influenced their patterns of genetic variation in modern humans. In short, based on this evidence, we emphasize the importance of regulatory re-wiring, and birth of novel genes that function in immune modulation in evolution of eutherian pregnancy.

Overview of pregnancy in eutherian mammals

While both marsupials and eutherians are viviparous, several major differences exist. Marsupial pregnancies are generally shorter than eutherian pregnancies (mean of 25 days versus 131 days) (Bainbridge 2000; Renfree and Shaw 2001), and for the bulk of the gestational period, the embryo is separated from the maternal uterine epithelial cells by a thick but permeable egg shell (Figure 1.1a): during this period, there is no direct maternal-fetal contact, and the developing embryo receives nutrients from uterine secretions (Zeller and Freyer 2001; Freyer *et al.* 2002; Griffith *et al.* 2017). This egg shell is only breached shortly before labor, followed by apposition and attachment of the embryo and formation of a short-lived yolk sac placenta (Figure 1.1a) (Selwood 2000; Griffith *et al.* 2017). In contrast, eutherian pregnancy involves a prolonged contact between the mother and the fetus, beginning at implantation that occurs early on in pregnancy (Figure 1.1b): the embryo is apposed to the uterine wall, which subsequently attaches, and in some species, invades, the uterus (Enders and Schlafke 1975; Cross *et al.* 1994). This is followed by the formation of the placenta (Figure 1.b), which serves nutritive and immune-modulatory roles for the remainder of pregnancy. It has been recently found that conserved sets of inflammatory markers and immune pathways are involved in both eutherian implantation and marsupial attachment and labor (Griffith *et al.* 2017; Hansen *et al.* 2017) In contrast to marsupials, eutherians experience major switches in the overall activities of the immune system in the uterus during pregnancy: the up-regulation of inflammation in the uterus during implantation is followed by an extended anti-inflammatory period (Figure 1.b) (Mor *et al.* 2011; 2017; Griffith *et al.* 2017) which likely stems from the need to protect the developing fetus from detrimental immune responses. Towards the end of pregnancy, the eutherian uterus experiences a spike in inflammation yet again, resulting in initiation of labor (Mor *et al.* 2011; Leimert *et al.* 2018). Therefore, mechanisms to regulate the immune system in the uterus was likely required for establishing prolonged pregnancies in eutherians (Chavan *et al.* 2017). In this regard, it has been hypothesized that the absence of such immune-modulation mechanisms (due to the lack of alternation of the immune system during pregnancy) in marsupials has prevented longer gestation periods in this lineage (Hansen *et al.* 2017).

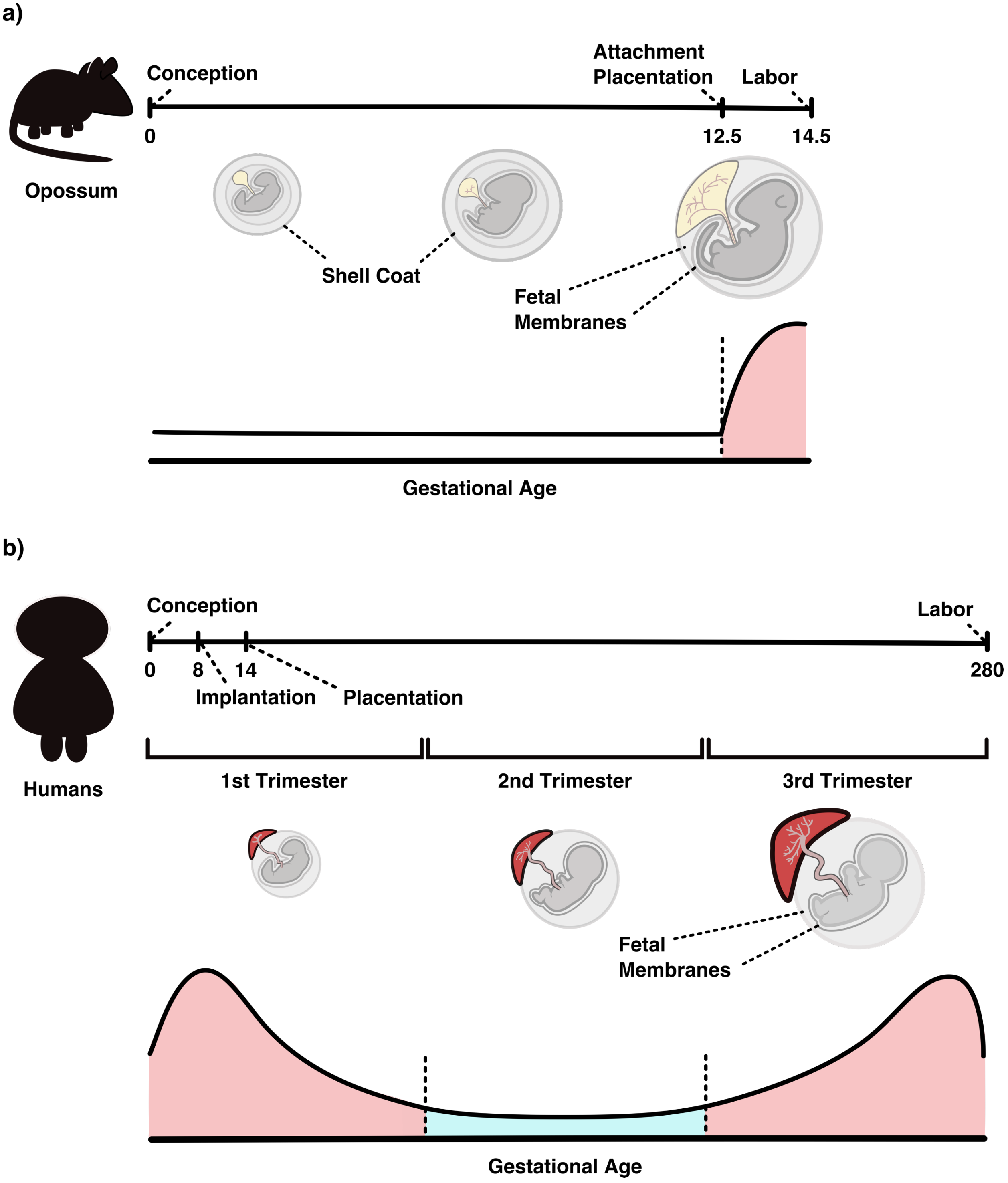

Figure 1.1. Comparison of pregnancy in eutherians and marsupials. A schematic illustration of the differences in the pregnancy between two lineages of mammals. a) Marsupial pregnancy is relatively short (mean gestation length of 25 days), and for the majority of the gestational period, the fetus is separated from maternal tissues by a shell

coat. This shell coat is only breached shortly before the initiation of labor, and is accompanied the formation of a transient yolk sac placenta. Formation of the placenta is associated with a spike in inflammation, which persists until labor. b) In contrast, eutherian mammals exhibit longer periods of pregnancy (mean gestation length of 131 days). Implantation and subsequent placentation occur relatively early on in pregnancy, establishing direct maternal-fetal contact that lasts for the remainder of pregnancy. Eutherian pregnancy is further distinguished from marsupial pregnancy in the occurrence of immune modulation throughout the entire gestational period: after the pro-inflammatory state following implantation, the immune state switches to a prolonged anti-inflammatory state, likely allowing optimal growth of the fetus. Towards the end of pregnancy, there is another surge in inflammation, which likely leads to initiation of labor.

In order to support extended growth of the fetus, the uterus in eutherians undergo several changes to create a receptive environment. Most noteworthy of these changes involve the transformation of the endometrial stromal fibroblasts (ESFs) into plump, secretory cells, or decidualized stromal cells (DSCs) that produce cytokines, growth factors and extracellular matrix proteins in response to “pregnancy signals” (e.g., progesterone, cAMP, and in some species, fetal signals) (Gellersen and Brosens 2003; Gellersen *et al.* 2007). Other changes include vascular remodeling to form spiral arteries that can supply large amounts of blood and oxygen to the fetus (Craven *et al.* 1998) and recruitment of immune cells to the site of maternal-fetal contact to engage in immune modulation required for preventing rejection of the genetically distinct fetus.

Furthermore, extensive variation exists in terms of the nature of pregnancies even within eutherians. Eutherian placentas have diversified from the ancestral invasive placenta (i.e., hemochorial placenta) (Wildman *et al.* 2006; Elliot and Crespi 2009), resulting in three major modes of placentation, differentiated by the extent of invasiveness of the placenta (Mess and Carter 2007). The most invasive, hemochorial placentation, results in invasion of the fetal trophoblast into the uterus and erosion of all maternal tissue layers, leading to direct contact between the fetus and the maternal blood (Furukawa *et al.* 2014). This mode of placentation is observed in rodents and in some primates, including humans (Carter and Enders 2004; Carter *et al.* 2015). Endothelichorial placentation involves invasion of the trophoblast into the uterus but without direct contact with the maternal circulation. Endothelichorial placentation is seen in groups such as canines and bats (Carter and Enders 2004; Enders and Carter 2012). The least invasive is the epitheliochorial placentation, in which the trophoblast is in superficial contact with the uterine epithelium without any invasion or destruction of maternal tissues (Furukawa *et al.* 2014). Ruminants such as cows and sheep exhibit this form of placentation (Carter and Enders 2004). Given this diversity, it is likely that different groups of eutherian mammals are characterized by lineage-specific adaptations that result from modifications of the global innovations that occurred with the emergence of eutherians. Furthermore, it is reasonable to hypothesize that hemochorial mammals would have acquired additional mechanisms by which to carefully regulate the extent of placentation to prevent both excessive and insufficient placental invasion.

Numerous studies have focused on understanding the molecular mechanisms that underlie the evolution of pregnancy in mammals, identifying three major trends. First, extensive re-wiring of regulatory networks involving pre-existing genes have happened to enable appropriate transcriptional responses (Figure 1.2a-b). Second, birth of novel genes that serve purposes specific to pregnancy in eutherian mammals have occurred in this lineage (Figure 1.2c-d). Third, species-specific changes as well as population-specific adaptations on top of such general and global changes within eutherian mammals have likely resulted in diverse strategies of pregnancies within eutherian mammals.

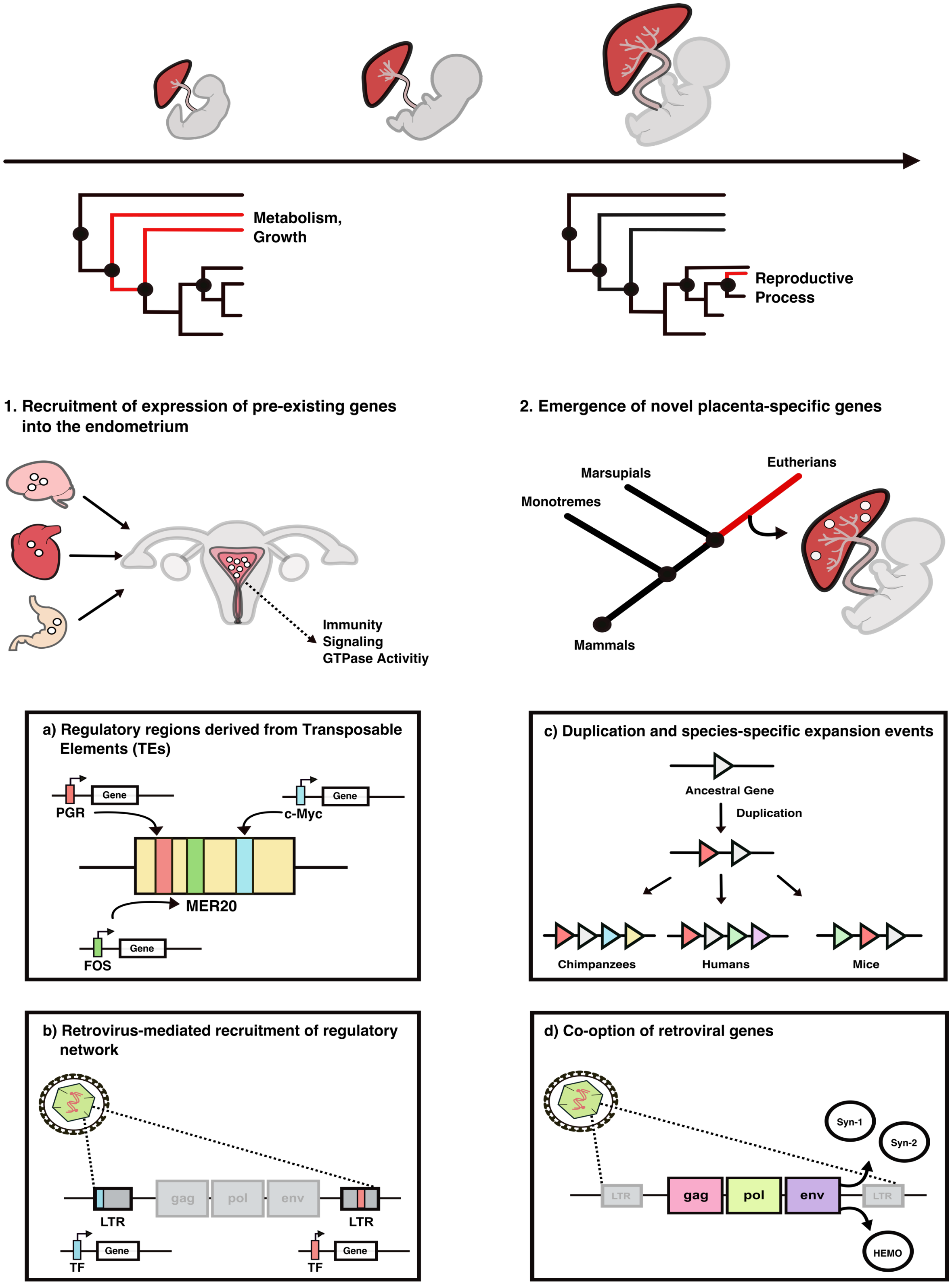
Metabolism,
Growth
Reproductive
Process
1. Recruitment of expression of pre-existing genes into the endometrium
2. Emergence of novel placenta-specific genes
Immunity
Signaling
GTPase Activitiy
Marsupials
Eutherians
Monotremes
Mammals
a) Regulatory regions derived from Transposable Elements (TEs)
Gene
PGR
Gene
c-Myc
MER20
Gene
FOS
c) Duplication and species-specific expansion events
Ancestral Gene
Duplication
Chimpanzees
Humans
Mice
b) Retrovirus-mediated recruitment of regulatory network
gag
pol
env
LTR
LTR
Gene
TF
Gene
TF
d) Co-option of retroviral genes
Syn-1
Syn-2
LTR
gag
pol
env
LTR
HEMO

Figure 1.2. Molecular mechanisms underlying evolution of eutherian pregnancy. A schematic illustration of the proposed molecular mechanisms that have contributed to the

evolution of eutherian pregnancy. According to Knox and Baker (Knox and Baker 2008), genes of different function and evolutionary age are involved in the early and late placental development process, respectively. More specifically, ancient genes with functions involved in metabolism and growth have been co-opted to act in early development of the placenta. One mechanism by which pre-existing ancient genes have been co-opted to participate in reproductive processes could be recruitment of such genes into expression within reproductive tissues. This could have been facilitated by a) contribution of regulatory regions by transposable elements (TEs) as well as b) retroviruses. In contrast, evolutionarily younger (often species-specific) genes associated with pregnancy-specific processes participate in the later developments of the placenta. c) Duplication events followed by subsequent species-specific expansions and d) co-option of retroviral genes could have contributed to the emergence of novel genes involved in pregnancy.

Regulatory evolution within eutherian mammals

Emergence of novel phenotypes could be facilitated by changes in gene regulation, and several studies have shown this to be the case for eutherian pregnancy as well. For instance, genes highly expressed in the developing placenta of mice are enriched for ancient genes (i.e., genes sharing orthologs with eukaryotic organisms) (Knox and Baker 2008). These ancient genes are enriched for roles in growth and metabolic processes, which is in contrast to enrichment of functions more specific to pregnancy/reproductive processes and negative regulation of physiological and cellular processes among genes highly expressed in the mature placenta, which are often rodent-specific (Figure 1.2) (Knox and Baker 2008). Such ancient, pre-existing genes could have been co-opted for usage in pregnancy by re-wiring of regulatory networks and some studies have provided evidence that such changes in gene regulation could underlie the extensive differences we see in marsupial and eutherian pregnancies.

For example, eutherian DSCs were likely formed by reprogramming of the pregnancy-related stress response in marsupial (grey short-tailed opossum) ESFs (Erkenbrack *et al.* 2018). Interestingly, when stimulated with decidualization signals (i.e., progesterone and cAMP), several core regulatory genes known to be involved in eutherian decidualization are also up-regulated in opossum ESFs, some of which were transcription factors (TFs) with well-known roles in decidualization: these include *PGR*, *CEBPB*, *HOXA11* and *STAT3*. However, some TFs that are up-regulated in human DSCs (i.e., *PRL*, *IGFBP1*) are eutherian-specific and are not found in the opossum ESF, suggesting that recruitment of new TFs contributed to evolution of DSC. Subsequent functional enrichment analyses revealed that genes up-regulated in the opossum were enriched for stress responses and apoptosis, indicating that the decidualization signals trigger a different biological process in marsupials compared to eutherians. Interestingly, *FOXO1*, a TF that is up-regulated in opossum ESF upon stimulation with decidualization signals, regulates different sets of genes in human DSC and opossum ESF, suggesting that decidualization in eutherians evolved through co-option of the marsupial stress pathway by rewiring downstream targets genes of the highly conserved regulatory network.

In addition, striking differences exist in the reconstructed ancestral transcriptomes of ESFs of eutherian mammals and transcriptomes of ESFs of opossums (Kin *et al.* 2016): more specifically, genes that gained uterine expression in eutherian mammals were enriched for functions involved in cell division and proliferation, while genes with expression in opossum ESFs but not in ancestral eutherians were enriched for inflammation, cell movement and adhesion processes. Other examples include Siglec-5, 6 and 14, which gained placental expression only in humans (Brinkman-Van der Linden *et al.* 2007; Ali *et al.* 2014). While the mechanism by which Siglec-5 and 14 gained expression in the human placenta remains unknown, one study suggests that Siglec-6 likely gained human-specific placental expression via changes in the 5' UTR region (Brinkman-Van der Linden *et al.* 2007): these include loss of GATA-binding site and gain of E-box and Oct-1 sites. More recently, extensive differences in species-specific TFs exist between mice and humans, possibly accounting for the large differences that exist between these two species in terms of placental development (Soncin *et al.* 2018). For instance, *VGLL1*, a potential human-specific regulator of trophoblast differentiation of embryonic stem cells, is not detected in mouse placenta at any point in pregnancy.

More broadly, the evolution of eutherian pregnancy is associated with loss and, more interestingly, gain of uterine expression of numerous genes (Lynch *et al.* 2015). These genes are often involved in regulation of immune responses, metabolic processes, and cell divisions and this may stem from the need for carefully regulating such processes in the uterine environment during the prolonged gestational periods. We will discuss two mechanisms that have contributed to regulatory evolution within eutherian mammals (Figure 1.2a-b). The first mechanism by which these genes have gained novel uterine expression may involve transposable elements (TEs), DNA sequences that can "jump" to new sites within genomes (Figure 1.2a). For instance, a significant proportion of genes that undergo changes in gene expression upon decidualization are in proximity to MER20, a hAT-Charlie family DNA transposon (Lynch *et al.*). In addition, MER20s are able to bind to TFs important for hormone responsiveness and pregnancy, such as PGR, FOXO1A and HOXA11, and other more general TFs such as CTCF, p53 and p300 (Lynch *et al.*). In addition, depending on the combination of bound TFs, such MER20s can act as either insulators or cis-regulatory elements (i.e., enhancers, repressors). Similarly, the proximal part of an alternative prolactin (PRL) that results in extrapituitary expression occurs within MER20 and another long terminal repeat (LTR), MER39 (Gerlo *et al.* 2006). More broadly, locations of regulatory elements active in DSCs often overlap with ancient mammalian TEs (AncMam-TEs) and genes associated with regulatory regions derived from such TEs exhibit significant changes in expression levels upon decidualization, which is especially true for genes that have evolved novel uterine expression (Lynch *et al.* 2015). In addition, these AncMam-TEs are enriched for binding sites for TFs with roles in pregnancy, such as those that mediate hormone responses (e.g., PGR, NR4A1, ERRA) or have functions in endometrial cells or immune regulation (e.g., ELK4, ARNT, c-Myc, E2F1) (Lynch *et al.* 2015). In short, TEs have provided novel regulatory elements in eutherian mammals to enable appropriate transcriptional response to decidualization.

Another source of regulatory evolution involves endogenous retroviruses (ERVs), retroviruses that infected and have been integrated into the genomes of host germline cells and passed onto the next generation. A typical ERV consists of three core retroviral genes (*env*, *gag*, *pol*), which are flanked by LTRs (Chuong 2013). While ERV LTRs normally function to promote transcription of the viral genome, it has been hypothesized that they could also act as novel promoters or enhancers for nearby genes if they contain appropriate binding sites for trophoblast-specific TFs (Figure 1.2b) (Chuong 2013). Such LTR-derived regulatory elements could potentially co-opt entire gene regulatory networks, resulting in extensive changes in the placental transcriptome. While less investigated than the co-option of ERV genes (which will be discussed in the following sections), some studies have uncovered instances of LTR-derived regulatory elements. For example, a recently discovered anthropoid primate-specific LTR-derived THE1B element regulates expression of corticotropin-releasing hormone (CRH), a hormone that is involved in controlling gestation lengths in humans, and other placental genes (Dunn-Fletcher *et al.* 2018). This THE1B element was also found to bind to DLX3, a TF that contributes to trophoblast differentiation. Another example involves a recently discovered novel trophoblast-specific enhancer that is required for expression of HLA-G in human extravillous trophoblasts (EVTs), which lies within a LTR region associated with ERV1 (Ferreira *et al.* 2016). Similar to the THE1B element described above, this enhancer exhibits binding activities for CEBP (CEBPB) and GATA family TFs (GATA2, GATA3) that are highly expressed in the placenta. Other examples of ERV-driven regulatory evolution include leptin (Bi *et al.* 1997), pleiotropin (Schulte *et al.* 1996), endothelin B receptor (EBR) (Medstrand *et al.* 2001; Landry and Mager 2003), Midline 1 gene (MID1) (Landry and Mager 2002), which have gained placental expression via LTR-derived promoters.

In summary, novel regulatory elements arising from TEs and LTRs of ERVs could have facilitated the recruitment of gene expression in reproductive tissues- uterus and trophoblast, respectively- resulting in appropriate transcriptional responses to promote successful pregnancy.

Emergence of novel genes in eutherian mammals

As discussed above, eutherian pregnancy involves activities of both pre-existing genes that were originally involved in other biological processes (Figure 1.2a-b), as well as novel, evolutionarily young genes (Figure 1.2c-d) (Knox and Baker 2008). For instance, a large-scale classification of proteins on the basis of sequence similarity has identified numerous genes that have emerged in the stem lineage of eutherian mammals and have either been conserved in all species or lost across different subgroups of eutherian mammals (Dunwell *et al.* 2017). Subsequent functional enrichment analyses revealed that these genes are often involved in immune responses, development, and regulation of transcription. Two mechanisms- gene duplication (Figure 1.2c) and co-option of ERV genes (Figure 1.2d)- could result in emergence of novel genes in eutherians.

Several phylogenetic analyses suggest gene duplication, which is followed by species-specific expansion or contraction events, to have played a role in emergence of several novel gene families (Figure 1.2c), including pregnancy-specific glycoproteins (PSGs), KIRs, and SIGLECs. Part of the carcinoembryonic antigen (CEA) family, PSGs are the most abundant secreted trophoblast-specific proteins detected in the maternal blood during pregnancy and may play roles in modulating the maternal immune responses. These include inducing monocytes and dendritic cells to produce anti-inflammatory cytokines(Wessells *et al.* 2000; Blois *et al.* 2012; Martínez *et al.* 2012), promotion of alternative macrophage activation that results in suppression of T cell activation and proliferation (Motrán *et al.* 2002). To date, PSGs have been found exclusively in hemochorial mammals (e.g., humans, rodents, and some primates) or mammals that possess a population of trophoblasts with invasive properties (e.g., horses) (Kammerer and Zimmermann 2010). Phylogenetic studies have shown that PSGs have likely risen from duplication of an ancestral CECAM (CECAM1) gene, followed by additional subsequent expansion events (Kammerer and Zimmermann 2010). In primates, PSGs have undergone a major expansion after the separation of wet-nosed and dry-nosed primates, followed by independent expansion events in apes and rhesus macaques (Chang *et al.* 2013), respectively.

Another example involves KIRs, a family of genes that is part of the leukocyte receptor complex (LRC) on chromosome 19 in humans (Parham 2005). KIRs are expressed on uterine natural killer cells (uNKs) that come into contact with EVTs invading into the decidua (Moffett and Colucci 2014). The interaction between uNK cells and EVTs regulates the degree of trophoblast invasion into the decidua and subsequent remodeling of the spiral arteries (Lash 2006; Fraser *et al.* 2015). A subset of KIRs expressed on uNK cells binds to human leukocyte antigen (HLA) type C on the EVTs (Varla-Leftherioti 2004; Male *et al.* 2011): this KIR-HLA-C interaction results in the activation of uNK cells, leading to secretion of cytokines and angiogenic factors that facilitate placentation (Xiong *et al.* 2013). Duplication event of an ancestral KIR gene that occurred approximately 140 million years ago resulted in two founder KIR genes: *KIR3DX* for cattle and *KIR3DL* for simian-primates (Guethlein *et al.* 2007). During the past 40 - 58 million years of simian-primate evolution, KIR genes have expanded from the single *KIR3DL* gene (Sambrook *et al.* 2006; Guethlein *et al.* 2007). For instance, a subset of KIRs that can bind to MHC-C (HLA-C in humans) first expanded in orangutans (Guethlein *et al.* 2007), followed by additional species-specific expansions in chimpanzees and humans (Abi-Rached *et al.* 2010). In contrast, KIRs have experienced intensive deletion and mutations in gibbons, resulting in contraction of the KIR locus (Abi-Rached *et al.* 2010).

A similar scenario unfolded in Siglecs, immunoglobulin-like lectins expressed on immune cells that bind sialic acids ubiquitously found on host cells. Some genes belonging to the subgroup CD33r-related (CD33r) Siglecs exhibit expression in reproductive tissues (Brinkman-Van der Linden *et al.* 2007; Ali *et al.* 2014): While the exact role other CD33r Siglecs play in pregnancy is less well known, one mechanism may involve regulation of the peripheral maternal immune system during pregnancy, as discussed in more detail in the following sections. In mammals, the primordial CD33r Siglecs cluster (which formed via tandem duplications of the ancient Siglecs cluster) has been shown to have undergone a large-scale inverse gene duplication event approximately 180 million years ago (i.e., before the eutherian/marsupial split), followed by species-specific expansions and contraction events (Cao *et al.* 2009): for instance, while the early post-duplication cluster underwent additional duplication events in primates and dogs, this cluster experienced significant contraction in rodents.

Other examples of pregnancy-related genes arising from gene duplications include hormones such as growth hormones (GH) in primates (Chen *et al.* 1989; González *et al.* 2006; Wallis and Wallis 2006; Mendoza *et al.*), prolactins (PRLs) in rodents (Li and Zhang 2006), and chorionic gonadotropin (CG) subunit genes (Maston and Ruvolo 2002; Fortna *et al.* 2004). Additional examples consist of the homeodomain-containing TFs or HOX genes, that are known to play important roles in pregnancy (Ruddle *et al.* 1997; Soshnikova *et al.* 2013); galectins, a highly conserved family of β-galactosidase binding lectins expressed on immune cells (Houzelstein *et al.* 2004; Than *et al.* 2009); and placentally-expressed cathepsins (PECs) in rodents (Sol-Church *et al.* 2002). In all cases, there is considerable diversity in the repertoire of such gene families among species, resulting from species-specific expansion/contract events that occurred after major duplication event of a single ancestral gene. In this regard, it is interesting to note that Knox and Baker also found that the majority of the rodent-specific genes highly expressed in the mature placenta are members of three gene families that experienced major rodent or mouse-specific expansions: PRLs, PSGs/CECAMs and mouse PECs (Knox and Baker 2008).

A rather unusual mechanism by which novel genes potentially arose in eutherian mammals involves genes of the ERVs (Figure 1.2d). While most of the ERV genes no longer code for functional proteins due to accumulation of mutations and deletions, some proteins have been retained to serve important functions for their hosts. Syncytins-1 (*Syn-*

1) and 2 (*Syn*-2), envelope genes of HERV-W (Mi *et al.* 2000) and HERV-FRD (Vargas *et al.* 2009), respectively, are such genes. Early experimental works have elucidated direct role of these genes in mediating the fusion of the mononucleate cytotrophoblasts to form the multinucleate syncytiotrophoblasts (SYNs), the outer layer of trophoblasts that comes into direct contact with maternal blood and involved in maternal-fetal exchange (Mi *et al.* 2000). Furthermore, reduced expression of syncytins has been shown to be associated with adverse outcomes of pregnancy, such as pre-eclampsia in humans and embryonic lethality in mice (Chen *et al.* 2008; Dupressoir *et al.* 2009). Furthermore, ERV *env* genes have been integrated and co-opted independently in multiple lineages throughout mammalian evolution (Lavialle *et al.* 2013). In addition to its involvement in formation of the SYN, *Syn*-1 may also play a role in suppressing detrimental antiviral immune responses, thereby creating an immunologically tolerant environment for the fetus (Tolosa *et al.* 2015): *Syn*-1 treated immune cells from non-pregnant women exhibited reduced production of interferons (e.g., IFN-λ, IFN-α), whereas levels of the anti-inflammatory cytokines (e.g., IL-10) were increased, and changes in levels of such cytokines were similar to those seen in pregnant women without *Syn*-1 treatment (Tolosa *et al.* 2015).

Apart from syncytins, genome-wide searches for ERV Env proteins have uncovered 45 Env-encoding ORFs, one of them encoding an unusual Env protein, HEMO of the MER34 family, which is missing several characteristics that are shared among other Env proteins (Heidmann *et al.* 2017). Unlike other Env proteins, HEMO does not possess fusogenic properties and is shed into the local and peripheral blood in pregnant women. While the exact role of HEMO in pregnancy is yet unknown, it is speculated that the shed HEMO proteins may act to sequester receptors that could be used by other retroviruses, thereby aiding host defense activities (Heidmann *et al.* 2017). While the exact roles in reproduction are yet unknown, there are other examples of ERV-derived genes (and proteins) expressed in the placenta. These include the transmembrane protein (TM; resulting from cleavage of the Env in the Golgi) of HERV-K (Kämmerer *et al.* 2011), *env* gene of HERV-E (Yi and Kim 2007) and *env-3* (Andersson *et al.* 2005; Mangeney *et al.* 2007). These ERV-derived elements are hypothesized to play immunosuppressive roles via their immunosuppressive domain (ISD).

In summary, the evolution of eutherian pregnancy was likely facilitated by both co-option of pre-existing ancient genes to promote early growth of the placenta and the emergence of novel evolutionarily younger (often species-specific) genes via gene duplication and/or ERV-co-option events, which act in later stages of placental development to regulate more pregnancy-specific processes.

Trade-offs between reproductive processes and host defense have influenced the patterns of genetic variation of genes involved in pregnancy within humans

Prolonged direct maternal-fetal contact in eutherian pregnancy (and in some species, exposure of the fetus to the maternal circulation) likely requires careful regulation of the maternal immune responses to prevent immunological rejection of the fetus. This refined immune-regulation must occur while maintaining efficient host defense abilities and it is commonly believed that the resulting selective pressures on these processes can be conflicting. For this discussion, we will focus on two families of immunity-related genes- KIRs and Siglecs (Figure 1.3) - that have been found to play roles in both reproduction and host defense.

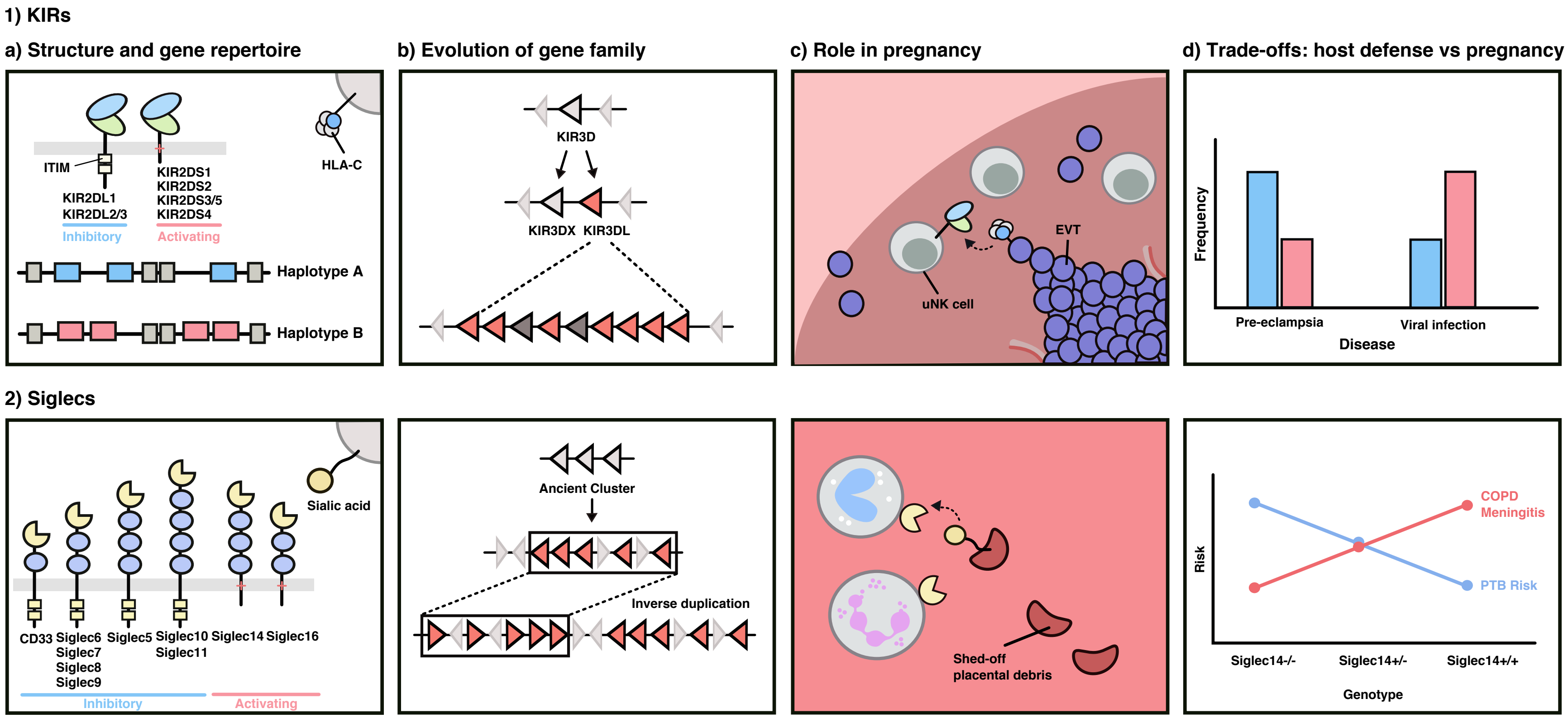

Figure 1.3. Examples of trade-offs between reproductive processes and host defense: KIRs and Siglecs. A summary of the key features of two gene families that have been shown to play roles in both reproduction and host defense. This figure illustrates what is known about the a) general structure of the gene and repertoire of the gene families b) the evolutionary process by which these gene families have emerged c) their known (and in the case of CD33r Siglecs, predicted) roles in pregnancy and d) trade-offs between pregnancy and immune activities.

KIR haplotypes can be divided into two functionally distinct groups – A and B – which are found only within humans. KIR A haplotypes are characterized by fixed gene content which mostly consists of inhibitory receptors with strong binding affinity (Figure 1.3) (Abi-Rached *et al.* 2010). The inhibitory KIR A haplotype, due to insufficient placental growth resulting from suppressed uNK activities, is associated with lower birth weight and adverse pregnancy outcomes such as pre-eclampsia, especially in combination with fetal HLA-C with the C2 epitope of paternal origin (Hiby *et al.* 2004; Hiby *et al.* 2010; Hiby *et al.* 2014; Farrell *et al.* 2014; Nakimuli *et al.* 2015). Interestingly, KIR A haplotypes are associated with more effective host defenses against acute viral infections such as Ebola and Hepatitis than B haplotypes (Khakoo *et al.* 2004; Wauquier *et al.* 2010), likely due to the attenuated KIRs of the B haplotypes and increased polymorphism of the KIR A haplotypes (Figure 1.3). In contrast, KIR B haplotypes are more variable in gene content and consist of less polymorphic and attenuated KIRs, some of which are activatory (Abi-Rached, Moesta, *et al.* 2010). Contrary to KIR A haplotypes, combination of maternal KIR B haplotype with fetal HLA-C1 is associated with the lowest risk for pre-eclampsia (Hiby *et al.* 2010; Nakimuli *et al.* 2015). KIR B haplotypes are also associated with KIRs that exhibit attenuated activities or even lost binding affinity for MHC-C ligands (Abi-Rached *et al.* 2010), and therefore, are likely to confer weaker host defense against pathogens (Figure 1.3). Frequencies of the KIR haplotypes, individual KIR genes, as well as diversity and identify of profiles of KIR haplotypes, vary among human populations (Nakimuli *et al.* 2013): for example, there are more KIR haplotype profiles in Ugandans than in the Europeans, homozygosity for the KIR A haplotype being the most common in Ugandans.

In addition, the frequency of HLA-C2 epitope is higher in Ugandans (and other sub-Saharan Africans) compared to elsewhere in the world, likely because of its protective role against pathogens such as malaria. Furthermore, *KIR2DS5*, which is present in higher frequencies in Ugandans compared to Europeans, confers protection against pre-eclampsia only in Ugandans when present in the centromeric region of KIR B haplotypes (Hiby *et al.* 2010; Nakimuli *et al.* 2015). In contrast, *KIR2DS1*, which is the protective gene in Europeans and not in Ugandans (Nakimuli *et al.* 2015), is present in higher frequencies in Europeans (Nakimuli *et al.* 2013). More broadly, it has been suggested that within a population, the frequencies of the KIR haplotypes (and HLA-C epitopes) fluctuate over time in response to changing selective pressures (Parham and Moffett 2013): for example, emergence of a novel pathogen might result in higher frequencies of KIR A haplotypes (and HLA-C1 epitope) but following successful eradication of the pathogen, increased selective pressures on reproduction could result in increase in frequencies of KIR B haplotypes (and HLA-C2 epitope) (Parham and Moffett 2013). In this regard, it is interesting to note that the risk of pre-eclampsia is high among sub-Saharan Africans (Nakimuli and Moffett 2017), which is likely due to the high frequencies of KIR A haplotypes and HLA-C2 epitope (Nakimuli *et al.* 2013). In summary, patterns of genetic variation of KIRs within any population reflect the actions of selection on both well-regulated extent of placentation and optimal host defense against pathogens.

Some CD33r Siglecs occur as paired receptors - receptors with binding affinity to almost identical ligands but with opposing signaling motifs (Figure 1.3). Siglec-5 and Siglec-14 are such paired receptors: Siglec-5, via the immunoreceptor tyrosine-based inhibitory motif (ITIM) on its cytoplasmic tail, acts as an inhibitory receptor, while Siglec-14 associates with the immunoreceptor tyrosine-based activation motif (ITAM)- bearing adaptor DAP12 and therefore acts as an activating receptor (Angata *et al.* 2006). Interestingly, some individuals lack a functional Siglec-14 (Yamanaka *et al.* 2009): this is due to fusion between Siglec-5 and 14 into a single gene (Siglec14/5) that is under the control of the Siglec-14 promoter but is functionally equivalent to Siglec-5 (Yamanaka *et al.* 2009). Interestingly, such Siglec-14 null allele in infants is associated with incidences of preterm birth only in the context of maternal Group B *Streptococcus* (GBS) rectovaginal colonization, speculated to be associated with an suppressed innate immune response (Figure 1.3) (Ali *et al.* 2014). This possibly protective role of wild type Siglec-14 allele against preterm birth is in contrast with its association with exacerbation of disease symptoms (Figure 1.3): for instance, patients with increased Siglec-14 allele dosage are more likely to experience exacerbation of chronic obstructive pulmonary disease (COPD) symptoms, such as tightening of airways, increased mucus production, inflammation, and reduced amount of airflow, when infected with nontypeable *Haemophilus influenzae* (NTHi) (Angata *et al.* 2013): this is thought to be due to the strong pro-inflammatory immune reactions mounted by the Siglec-14 wildtype allele. Similarly, the Siglec-14 null allele is associated with increased protection against *Mycobacterium tuberculosis* in a Vietnamese patient cohort (Graustein *et al.* 2017). While the exact nature of the protective effect against this disease is yet unclear, one possible mechanism may involve IL-2: Siglec-14 null allele is associated with increased secretion of pro-inflammatory IL-2 in response to BCG stimulation (Graustein *et al.* 2017). As IL-2 is associated with Th1 type adaptive immune responses (Zhou *et al.* 2003), this could partly explain the increased efficiency with which the causative pathogen is cleared from individuals with the Siglec-14 null allele. In this regard, it is interesting to note that Siglec-14 null allele frequency varies among different populations, with the frequencies being highest among East and South Asians, followed by middle Eastern populations, sub-Saharan Africans, and lowest in Northern Europeans (Yamanaka *et al.* 2009).

As pregnancy requires careful immune modulation both at the local and systemic level, it is possible that other immunity genes have experienced selective pressures arising from both reproductive processes and host defense activities. For instance, galectins are known to be involved in both modulation of immune responses during infection (Ideo *et al.* 2009; Yang *et al.* 2011) and suppression of detrimental immune responses at the maternal-fetal interface (Kopcow *et al.* 2008; Tirado-González *et al.* 2013). More broadly, a recent study uncovered several precisely timed events of systemic immune modulations that occur over the course of a normal pregnancy (Aghaeepour *et al.* 2017). While studies comprehensively examining the recent evolution of genes involved in such immune pathways are yet lacking, we hypothesize that the patterns of genetic variation of these genes would reflect selection acting on both reproductive success and efficient host defense.

In conclusion, these examples suggest that selection on reproductive success and effective host defense often acts in opposing directions, and that the patterns of genetic variation of these genes observed in present populations often reflect the action of selection on both factors in the context of environmental factors.

Summary

The evolution of mammalian, and more specifically, eutherian pregnancy is characterized by several innovations, which have been enabled by extensive re-wiring of regulatory networks to incorporate pre-existing genes to participate in pregnancy, and the emergence of novel genes that are involved in more pregnancy-specific roles. Regulatory evolution has been facilitated by contribution of regulatory regions by TEs that provided binding sites for TFs that play critical roles in pregnancy and similarly, LTRs of ERVs that can act as novel *cis*-regulatory regions. Gene duplication events followed by species-specific expansions, as well as co-option ERV-derived genes, underlie the birth of new genes in eutherians. Interestingly, while there are exceptions, genes involved in immune regulation were recruited into expression within reproductive tissues (Lynch *et al.* 2015) and also newly emerged within eutherians (Cao *et al.* 2009; Kammerer and Zimmermann 2010; Dunwell *et al.* 2017). This could partly be explained by the need for careful regulation of the maternal immune system during eutherian pregnancy, due to the extended direct maternal-fetal contact that happens throughout the majority of the gestational period. For instance, in humans, T cell activities are strongly suppressed at the maternal-fetal interface by induction of apoptosis via Fas-FasL signaling, starvation of these cells by tryptophan depletion by indolamine (IDO) 2,3-dioxygenase, and expression of PD-L1 on T cells (Jerzak *et al.* 2011). In addition, differentiation of regulatory T cells with anti-inflammatory properties (Somerset *et al.* 2004), alternative activation of macrophages leading to suppression of cytotoxic T cell activities (Svensson *et al.* 2011), reduced production of costimulatory molecules and increased secretion of Th2 cytokines by dendritic cells occur at the maternal-fetal interface (Martínez *et al.* 2012) as well, creating an immunologically tolerant environment for the growth of the fetus.

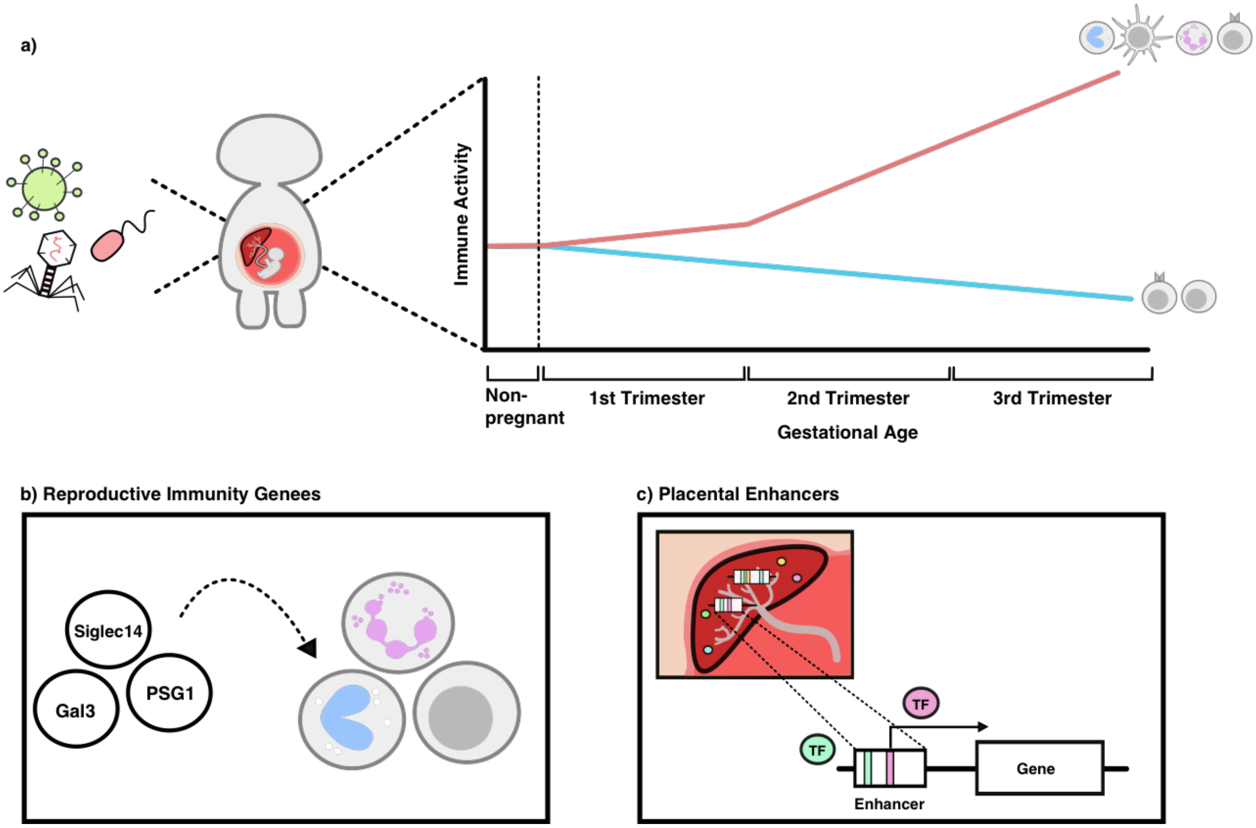

Figure 1.4. Changes in the maternal immune system that occur during pregnancy and the genetic elements that are involved in this process. This figure summarizes the changes in the maternal immune system that happen during pregnancy and the genetic elements that are potentially associated with such changes. a) Different components of the maternal immune system experience alternations in their activities

over the course of gestational period: more specifically, activities of monocytes, dendritic cells, granulocytes, and regulatory T cells (T_{reg} cells) increase over the course of pregnancy, especially starting in 2nd trimester. In contrast, activities of CD4+ and CD8+ T cells, as well as natural killer cells, decrease with gestational age. Involved in this immune modulation process are b) reproductive immunity genes and potentially, c) enhancers, especially the ones expressed in the placenta.

In humans, starting from second trimester and onwards, shed off SYN microparticles (likely resulting from trophoblastic apoptosis that is involved in continuous renewal of the SYN) can be detected in the maternal circulation (Germain *et al.* 2007). These floating microparticles come into close contact with maternal immune cells and the peripheral immune system is therefore modulated to down-regulate detrimental cell-mediated immunity (Figure 1.4a) (Sacks *et al.* 1999; Kanai *et al.* 2001; Somerset *et al.* 2004). Conversely, it has been shown that certain components of the innate immune system are activated to compensate for this relative suppression of cell-mediated immune activities: circulating innate immune cells exhibit activated phenotypes and are primed to produce pro-inflammatory cytokines upon stimulation (Figure 1.4a) (Sacks *et al.* 1999; Sacks *et al.* 2003). These alert innate immune components must be carefully regulated and loss of control (and subsequent excessive activation) has been linked to pregnancy complications (Sacks *et al.* 1998; Redman *et al.* 1999). At the same time, infectious diseases are a cause of considerable mortality, and therefore, pathogens represent one of the most important factors shaping human genetic variation among different populations (Fumagalli *et al.* 2011). As such, genes involved in host defense are often proposed targets of strong positive selection (Figure 1.4b). Thus, the patterns of genetic variation of such genes is likely a product of adaptation to local pathogen threats as well.

It is widely accepted that changes in gene regulation play a major role in the evolution of novel traits (King and Wilson 1975; Carroll 2005): this is because regulatory elements such as enhancers can alter expression patterns of a given gene in a particular context without affecting expression in others: such modular organization can therefore facilitate phenotypic evolution while minimizing pleiotropic effects (Carroll 2005; Wray 2007; Sholtis and Noonan 2010). Indeed, emergence of eutherian pregnancy is associated with novel regulatory elements, which are often involved in modulating the transcriptional activities of genes that are involved in immune responses, metabolic processes, and other pregnancy-specific processes (Lynch *et al.* 2015; Ferreira *et al.* 2016; Lynch *et al.*). In addition, recent studies suggest that selection on regulatory regions such as enhancers has contributed to recent human evolution: for instance, one study found population-level differences in the transcriptional responses to pathogens (Nédélec *et al.* 2016). In this regard, it is possible that the need for refined regulation of processes involved in pregnancy, such as immune activities, has influenced the recent evolution of these regulatory regions (Figure 1.4c). However, studies investigating patterns of genetic variation of regulatory regions involved in pregnancy are lacking. As such, future studies on the evolution of genetic elements involved in pregnancy could also incorporate regulatory regions such as promoters and enhancers.

Chapter previews

In this dissertation, I studied the patterns of recent evolution of genetic elements that are associated with human pregnancy using various population genetics metrics that can detect distinct genomic signatures of selection events that happened at different time ranges in recent human history, as well as under different modes (i.e., hard versus soft selective sweeps).

In Chapter II, I outline the methods I used in my studies (Chapter III and IV). In short, I use three metrics that are known to detect incidences of hard selective sweeps (Tajima's D, F_{ST}, and nS_L) and one metric that exhibits increased power towards soft selective sweeps (H12) into my studies. As non-adaptive processes such as past demographic events, as well as random processes like genetic drift can produce false signatures of selection, I also incorporate simulations of neutral evolution that explicitly accounts for these confounding factors into my studies.

In Chapter III, I study the patterns of recent evolution of 55 genes that are collectively involved in sialic acid biology genes, to test the hypothesis that reproductive immunity genes have experienced recent positive selection in modern humans. I use the methods discussed in Chapter II: in short, I calculate the various metrics of recent positive selection and compare the empirical values to those calculated on the neutrally simulated sequences. I found that the majority of the sialic acid biology genes do not exhibit significant evidence of recent positive selection. However, I found that the identity of the genes that do exhibit evidence of recent positive selection are different among the metrics calculated, suggesting that different genes have experienced selective events at different time ranges in recent human history. In addition, I also found that more genes exhibit evidence of having undergone soft selective sweeps than hard selective sweeps, suggesting the prevalence of soft selective sweeps in recent human evolution.

In Chapter IV, I study the patterns of recent evolution of a different type of genetic element- enhancers. More specifically, I calculate metrics of recent positive selection on enhancers active in 41 human tissues and compare them to those calculated on the “neutrally simulated” enhancers. I first calculate the proportions of enhancers exhibiting significant evidence of recent positive selection. Next, I carry out functional enrichment analyses on the putative target genes of enhancers with significant evidence of recent positive selection. Third, I compare the patterns of recent evolution in enhancers among the different tissues, between tissue-specific and tissue-broad enhancers. I found that on average, 5.90% of enhancers show evidence of recent positive selection across all tissues and metrics studied. In addition, putative target genes of these enhancers were enriched for functions related to immunity. Furthermore, enhancers active in brain and testis were found to exhibit significant differences in the patterns of recent evolution compared to enhancers of other tissues: I also found that tissue-specific and tissue-broad enhancers show significant differences in brain and testis.

Together, my dissertation aims to understand how genetic elements involved in pregnancy have evolved in recent human history. In doing so, my research also sheds light on the challenge associated with interpreting the results of such studies.

References

A Guethlein, L., L. Abi-Rached, J. Hammond, and P. Parham, 2007 The expanded cattle KIR genes are orthologous to the conserved single-copy KIR3DX1 gene of primates. Immunogenetics. 59:517-22.

Abi-Rached, L., H. Kuhl, C. Roos, B. ten Hallers, B. Zhu *et al.*, 2010 A Small, Variable, and Irregular Killer Cell Ig-Like Receptor Locus Accompanies the Absence of MHC-C and MHC-G in Gibbons. J. Immunol. 184: 1379.

Abi-Rached, L., A. K. Moesta, R. Rajalingam, L. A. Guethlein, and P. Parham, 2010 Human-Specific Evolution and Adaptation Led to Major Qualitative Differences in the Variable Receptors of Human and Chimpanzee Natural Killer Cells. PLoS Genet 6: e1001192.

Aghaeepour, N., E. A. Ganio, D. Mcilwain, A. S. Tsai, M. Tingle *et al.*, 2017 An immune clock of human pregnancy. Sci. Immunol. 2: eaan2946.

Ali, S. R., J. J. Fong, A. F. Carlin, T. D. Busch, R. Linden *et al.*, 2014 Siglec-5 and Siglec-14 are polymorphic paired receptors that modulate neutrophil and amnion signaling responses to group B Streptococcus. J Exp Med 211: 1231.

Amodio, G., A. Mugione, A. Sanchez, P. Viganò, M. Candiani *et al.*, 2013 HLA-G expressing DC-10 and CD4+ T cells accumulate in human decidua during pregnancy. Hum. Immunol. 74:406-411

Andersson, A.-C., Z. Yun, G. O. Sperber, E. Larsson, and J. Blomberg, 2005 ERV3 and Related Sequences in Humans: Structure and RNA Expression. J. Virol. 79: 9270.

Angata, T., T. Hayakawa, M. Yamanaka, A. Varki, and M. Nakamura, 2006 Discovery of Siglec-14, a novel sialic acid receptor undergoing concerted evolution with Siglec-5 in primates. The FASEB Journal 20: 1964–1973.

Angata, T., T. Ishii, T. Motegi, R. Oka, R. E. Taylor *et al.*, 2013 Loss of Siglec-14 reduces the risk of chronic obstructive pulmonary disease exacerbation. Cellular and molecular life sciences 70: 3199–3210.

Bainbridge, D. R., 2000 Evolution of mammalian pregnancy in the presence of the maternal immune system. Rev. Reprod. 5: 67–74.

Baker, M. L., J. P. Wares, G. A. Harrison, and R. D. Miller, 2004 Relationships Among the Families and Orders of Marsupials and the Major Mammalian Lineages Based on Recombination Activating Gene-1. Journal of Mammalian Evolution 11: 1–16.

Bartmann, C., S. E. Segerer, L. Rieger, M. Kapp, M. Sütterlin *et al.*, 2013 Quantification of the Predominant Immune Cell Populations in Decidua Throughout Human Pregnancy.

Am J Reprod Immunol 71: 109–119.

Bi, S., O. Gavrilova, D.-W. Gong, M. M. Mason, and M. Reitman, 1997 Identification of a Placental Enhancer for the Human Leptin Gene. J. Biol. Chem. 272: 30583–30588.

Billington, W. D., 2003 The immunological problem of pregnancy: 50 years with the hope of progress. A tribute to Peter Medawar. Journal of Reproductive Immunology 60: 1–11.

Blois, S. M., I. Tirado-González, J. Wu, G. Barrientos, B. Johnson *et al.*, 2012 Early expression of pregnancy-specific glycoprotein 22 (PSG22) by trophoblast cells modulates angiogenesis in mice. biolreprod 86: 191–191.

Brinkman-Van der Linden, E. C. M., N. Hurtado-Ziola, T. Hayakawa, L. Wiggleton, K. Benirschke *et al.*, 2007 Human-specific expression of Siglec-6 in the placenta. Glycobiology 17: 922–931.

Cao, H., B. de Bono, K. Belov, E. Wong, J. Trowsdale *et al.*, 2009 Comparative genomics indicates the mammalian CD33rSiglec locus evolved by an ancient large-scale inverse duplication and suggests all Siglecs share a common ancestral region. Immunogenetics 61: 401-417

Carroll, S. B., 2005 Evolution at Two Levels: On Genes and Form. PLoS Biol 3: e245.

Carter, A., and A. Enders, 2013 The Evolution of Epitheliochorial Placentation. Annu. Rev. Anim. Biosci. 1: 443-467

Carter, A. M., and A. C. Enders, 2004 Comparative aspects of trophoblast development and placentation. Reproductive Biology and Endocrinology 2: 46.

Carter, A. M., A. C. Enders, and R. Pijnenborg, 2015 The role of invasive trophoblast in implantation and placentation of primates. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 370: 20140070–20140070.

Caryl Wallis, O., and M. Wallis, 2006 Evolution of Growth Hormone in Primates: The GH Gene Clusters of the New World Monkeys Marmoset (Callithrix jacchus) and White-Fronted Capuchin (Cebus albifrons). J. Mol. Evol. 63: 591-601.

Chang, C. L., J. Semyonov, P. J. Cheng, S. Y. Huang, J. I. Park *et al.*, 2013 Widespread Divergence of the CEACAM/PSG Genes in Vertebrates and Humans Suggests Sensitivity to Selection. PLoS ONE 8: e61701.

Chavan, A., O. Griffith, and G. Wagner, 2017 The inflammation paradox in the evolution of mammalian pregnancy: turning a foe into a friend. Curr. Opin. Genet. Dev. 47: 24-32.

Chen, C. P., C. Y. Chen, C. C. Ko, G. D. Chang, L. F. Chen *et al.*, 2008 Functional Characterization of the Human Placental Fusogenic Membrane Protein Syncytin 21. biolreprod 79: 815–823.

Chen, E. Y., Y. C. Liao, D. H. Smith, H. A. Barrera Saldaña, R. E. Gelinas *et al.*, 1989 The human growth hormone locus: Nucleotide sequence, biology, and evolution. Genomics 4: 479–497.

Chuong, E. B., 2013 Retroviruses facilitate the rapid evolution of the mammalian placenta. BioEssays 35: 853-861.

Chuong, E. B., R. L. Hannibal, S. L. Green, and J. C. Baker, 2013 Evolutionary perspectives into placental biology and disease. Applied & Translational Genomics. 2: 64–69.

Craven, C. M., T. Morgan, and K. Ward, 1998 Decidual spiral artery remodelling begins before cellular interaction with cytotrophoblasts. Placenta 19: 241–252.

Cross, J. C., Z. Werb, and S. J. Fisher, 1994 Implantation and the placenta: key pieces of the development puzzle. Science 266: 1508.

Dunn-Fletcher, C. E., L. M. Muglia, M. Pavličev, G. Wolf, M.A. Sun *et al.*, 2018 Anthropoid primate–specific retroviral element THE1B controls expression of CRH in placenta and alters gestation length. PLoS Biol 16: e2006337.

Dunwell, T. L., J. Paps, and P. W. H. Holland, 2017 Novel and divergent genes in the evolution of placental mammals. Proc. R. Soc. B 284: 20171357.

Dupressoir, A., C. Vernochet, O. Bawa, F. Harper, G. Pierron *et al.*, 2009 Syncytin-A knockout mice demonstrate the critical role in placentation of a fusogenic, endogenous retrovirus-derived, envelope gene. Proc Natl Acad Sci USA 106: 12127.

Elliot, M., and B. Crespi, 2009 Phylogenetic Evidence for Early Hemochorial Placentation in Eutheria.Placenta 30: 949-967

Enders, A. C., and A. M. Carter, 2012 The evolving placenta: Convergent evolution of variations in the endotheliochorial relationship. Placenta 33: 319–326.

Enders, A. C., and S. Schlafke, 1975 Cellular Basis of Interaction Between Trophoblast and Uterus at Implantation. biolreprod 12: 41–65.

Erkenbrack, E. M., J. D. Maziarz, O. W. Griffith, C. Liang, A. R. Chavan *et al.*, 2018 The mammalian decidual cell evolved from a cellular stress response. PLoS Biol 16: e2005594.

Farrell, L., S. E Hiby, R. Apps, O. Chazara, L. Trogstad *et al.*, 2014 KIR and HLA-C: Immunogenetic regulation of human birth weight. Norsk. Epidemiologi. 24: 107-110.

Ferreira, L. M. R., T. B. Meissner, T. S. Mikkelsen, W. Mallard, C. W. O'Donnell *et al.*, 2016 A distant trophoblast-specific enhancer controls HLA-G expression at the maternal–fetal interface. Proc Natl Acad Sci USA 113: 5364.

Fortna, A., Y. Kim, E. MacLaren, K. Marshall, G. Hahn *et al.*, 2004 Lineage-Specific Gene Duplication and Loss in Human and Great Ape Evolution. PLoS Biol 2: e207.

Fraser, R., G. S. J. Whitley, B. Thilaganathan, and J. E. Cartwright, 2015 Decidual natural killer cells regulate vessel stability: implications for impaired spiral artery remodelling. Journal of Reproductive Immunology 110: 54–60.

Freyer, C., U. Zeller, and M. B. Renfree, 2002 Ultrastructure of the placenta of the tammar wallaby, Macropus eugenii: comparison with the grey short-tailed opossum, Monodelphis domestica. Journal of anatomy 201: 101–119.

Fumagalli, M., M. Sironi, U. Pozzoli, A. Ferrer-Admetlla, L. Pattini *et al.,* 2011. Signatures of Environmental Genetic Adaptation Pinpoint Pathogens as the Main Selective Pressure through Human Evolution. PLoS Genet. 7: e1002355.

Furukawa, S., Y. Kuroda, and A. Sugiyama, 2014 A comparison of the histological structure of the placenta in experimental animals. Journal of toxicologic pathology 27: 11–18.

Gellersen, B., and J. Brosens, 2003 Cyclic AMP and progesterone receptor cross-talk in human endometrium: A decidualizing affair. J. Endocrinol. 178: 357- 372.

Gellersen, B., I. Brosens, and J. Brosens, 2007 Decidualization of the Human Endometrium: Mechanisms, Functions, and Clinical Perspectives. Semin. Reprod. Med. 25: 445-453

Gerlo, S., J. Davis, D. L Mager, and R. Kooijman, 2006 Prolactin in man: A tale of two promoters. Bioessays. 28: 1051-1055.

Germain, S. J., G. P. Sacks, S. R. Soorana, I. L. Sargent, and C. W. Redman, 2007 Systemic Inflammatory Priming in Normal Pregnancy and Preeclampsia: The Role of Circulating Syncytiotrophoblast Microparticles. J. Immunol. 178: 5949.

González Alvarez, R., A. Revol de Mendoza, D. Esquivel Escobedo, G. Corrales Félix, I. Rodríguez Sánchez *et al.*, 2006 Growth hormone locus expands and diverges after the separation of New and Old World Monkeys. Gene 380: 38–45.

Graustein, A. D., D. J. Horne, J. J. Fong, F. Schwarz, H. C. Mefford *et al.*, 2017 The SIGLEC14 null allele is associated with Mycobacterium tuberculosis- and BCG-induced clinical and immunologic outcomes. Tuberculosis 104: 38–45.

Griffith, O. W., and G. P. Wagner The placenta as a model for understanding the origin and evolution of vertebrate organs. Nature Ecology & Evolution 1: 0072.

Griffith, O. W., A. R. Chavan, S. Protopapas, J. Maziarz, R. Romero *et al.*, 2017 Embryo implantation evolved from an ancestral inflammatory attachment reaction. Proc Natl Acad

Sci USA 114: E6566.

Guethlein, L. A., A. M. Older Aguilar, L. Abi-Rached, and P. Parham, 2007 Evolution of Killer Cell Ig-Like Receptor Genes: Definition of an Orangutan KIR Haplotype Reveals Expansion of Lineage III KIR Associated with the Emergence of MHC-C. J. Immunol. 179: 491.

Hansen, V. L., L. S. Faber, A. A. Salehpoor, and R. D. Miller, 2017 A pronounced uterine pro-inflammatory response at parturition is an ancient feature in mammals. Proc. R. Soc. B 284: 20171694.

Heidmann, O., A. Béguin, J. Paternina, R. Berthier, M. Deloger *et al.*, 2017 HEMO, an ancestral endogenous retroviral envelope protein shed in the blood of pregnant women and expressed in pluripotent stem cells and tumors. Proc Natl Acad Sci USA 114: E6642.

Hiby, S. E., R. Apps, O. Chazara, L. E. Farrell, P. Magnus *et al.*, 2014 Maternal KIR in Combination with Paternal HLA-C2 Regulate Human Birth Weight. J. Immunol. 192: 5069.

Hiby, S. E., R. Apps, A. M. Sharkey, L. E. Farrell, L. Gardner *et al.*, 2010 Maternal activating KIRs protect against human reproductive failure mediated by fetal HLA-C2. J Clin Invest 120: 4102–4110.

Hiby, S. E., J. J. Walker, K. M. O 039 Shaughnessy, C. W. G. Redman, M. Carrington *et al.*, 2004 Combinations of Maternal KIR and Fetal HLA-C Genes Influence the Risk of Preeclampsia and Reproductive Success. J Exp Med 200: 957.

Houzelstein, D., I. R. Gonçalves, A. J. Fadden, S. S. Sidhu, D. N. W. Cooper *et al.*, 2004 Phylogenetic Analysis of the Vertebrate Galectin Family. Molecular Biology and Evolution 21: 1177–1187.

Ideo, H., K. Fukushima, K. Gengyo-Ando, S. Mitani, K. Dejima *et al.*, 2009 A Caenorhabditis elegans Glycolipid-binding Galectin Functions in Host Defense against Bacterial Infection. J. Biol. Chem. 284: 26493–26501.

Jerzak, M., M. Kasprzycka, P. Wierbicki, J. Kotarski, and A. górski, 2011 Apoptosis of T Cells in the First Trimester Human Decidua. Am J Reprod Immunol 40: 130–135.

Kammerer, R., and W. Zimmermann, 2010 Coevolution of activating and inhibitory receptors within mammalian carcinoembryonic antigen families. BMC Biology 8: 12.

Kanai, T., T. Fujii, N. Unno, T. Yamashita, H. Hyodo *et al.*, 2001 Human Leukocyte Antigen-G-Expressing Cells Differently Modulate the Release of Cytokines from Mononuclear Cells Present in the Decidua Versus Peripheral Blood. Am J Reprod Immunol 45: 94–99.

Kämmerer, U., A. Germeyer, S. Stengel, M. Kapp, and J. Denner, 2011 Human endogenous retrovirus K (HERV-K) is expressed in villous and extravillous cytotrophoblast cells of the human placenta. Journal of Reproductive Immunology 91: 1–8.

Khakoo, S. I., C. L. Thio, M. P. Martin, C. R. Brooks, X. Gao *et al.*, 2004 HLA and NK Cell Inhibitory Receptor Genes in Resolving Hepatitis C Virus Infection. Science 305: 872.

Killian, J. K., T. R. Buckley, N. Stewart, B. L. Munday, and R. L. Jirtle, 2001 Marsupials and Eutherians reunited: genetic evidence for the Theria hypothesis of mammalian evolution. Mammalian Genome 12: 513–517.

Kin, K., J. Maziarz, A. R. Chavan, M. Kamat, S. Vasudevan *et al.*, 2016 The Transcriptomic Evolution of Mammalian Pregnancy: Gene Expression Innovations in Endometrial Stromal Fibroblasts. Genome Biology and Evolution 8: 2459–2473.

King, M. C., and A. C. Wilson, 1975 Evolution at two levels in humans and chimpanzees. Science 188: 107–116.

Knox, K., and J. C. Baker, 2008 Genomic evolution of the placenta using co-option and duplication and divergence. Genome Research 18: 695–705.

Kopcow, H. D., F. Rosetti, Y. Leung, D. S. J. Allan, J. L. Kutok *et al.*, 2008 T cell apoptosis at the maternal–fetal interface in early human pregnancy, involvement of galectin-1. Proc Natl Acad Sci USA 105: 18472.

Landry, J. R., and D. L. Mager, 2003 Functional Analysis of the Endogenous Retroviral Promoter of the Human Endothelin B Receptor Gene. J. Virol. 77: 7459.

Landry, J. R., and D. L. Mager, 2002 Widely Spaced Alternative Promoters, Conserved between Human and Rodent, Control Expression of the Opitz Syndrome Gene MID1. Genomics 80: 499–508.

Lash, G. E., 2006 Expression of angiogenic growth factors by uterine natural killer cells during early pregnancy. Journal of Leukocyte Biology 80: 572–580.

Lavialle, C., G. Cornelis, A. Dupressoir, C. Esnault, O. Heidmann *et al.*, 2013 Paleovirology of “syncytins,” retroviral env genes exapted for a role in placentation. Philosophical Transactions of the Royal Society B: Biological Sciences 368: 20120507–20120507.

Leimert, K. B., D. M. Olson, A. Messer, T. Gray, X. Fang *et al.*, 2018 Maternal and fetal intrauterine tissue crosstalk promotes proinflammatory amplification and uterine transition. Biol. Reprod. ioy232, https://doi.org/10.1093/biolre/ioy232

LI, Y., and Y.-P. Zhang, 2006 Molecular Evolution of Prolactin Gene Family in Rodents. Acta Genetica Sinica 33: 590–597.

Liu, S., L. Diao, C. Huang, Y. Li, Y. Zeng *et al.*, 2017 The role of decidual immune cells on human pregnancy. Journal of Reproductive Immunology 124: 44–53.

Luppi, P., 2003 How immune mechanisms are affected by pregnancy. Vaccine 21: 3352–3357.

Lynch, V. J., R. D. Leclerc, G. May, and G. P. Wagner, 2011 Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. Nature Genetics 43: 1154-1159.

Lynch, V. J., M. C. Nnamani, A. Kapusta, K. Brayer, S. L. Plaza *et al.*, 2015 Ancient Transposable Elements Transformed the Uterine Regulatory Landscape and Transcriptome during the Evolution of Mammalian Pregnancy. CellReports 10: 551–561.

Male, V., A. Sharkey, L. Masters, P. R. Kennedy, L. E. Farrell *et al.*, 2011 The effect of pregnancy on the uterine NK cell KIR repertoire. Eur. J. Immunol. 41: 3017–3027.

Mangeney, M., M. Renard, G. Schlecht-Louf, I. Bouallaga, O. Heidmann *et al.*, 2007 Placental syncytins: Genetic disjunction between the fusogenic and immunosuppressive activity of retroviral envelope proteins. Proc Natl Acad Sci USA 104: 20534.

Martínez, F. F., C. P. Knubel, M. C. Sánchez, L. Cervi, and C. C. Motrán, 2012 Pregnancy-specific glycoprotein 1a activates dendritic cells to provide signals for Th17-, Th2-, and Treg-cell polarization. Eur. J. Immunol. 42: 1573–1584.

Maston, G. A., and M. Ruvolo, 2002 Chorionic Gonadotropin Has a Recent Origin Within Primates and an Evolutionary History of Selection. Molecular Biology and Evolution 19: 320–335.

Medstrand, P., J.-R. Landry, and D. L. Mager, 2001 Long Terminal Repeats Are Used as Alternative Promoters for the Endothelin B Receptor and Apolipoprotein C-I Genes in Humans. J. Biol. Chem. 276: 1896–1903.

Mess, A., and A. M. Carter, 2007 Evolution of the placenta during the early radiation of placental mammals. Comp. Biochem. Physiol. A. Mol. Inter. Physiol.148: 769–779.

Mi, S., X. Lee, X.-P. Li, G. M. Veldman, H. Finnerty *et al.,* 2000 Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature 403: 785 -789.

Moffett, A., and F. Colucci, 2014 Uterine NK cells: active regulators at the maternal-fetal interface. J Clin Invest 124: 1872–1879.

Mor, G., P. Aldo, and A. B. Alvero, 2017 The unique immunological and microbial aspects of pregnancy. Nat. Rev. Immunol. 17: 469–482.

Mor, G., I. Cardenas, V. Abrahams, and S. Guller, 2011 Inflammation and pregnancy: the role of the immune system at the implantation site. Annals of the New York Academy of Sciences 1221: 80–87.

Motrán, C. C., F. L. Díaz, A. Gruppi, D. Slavin, B. Chatton *et al.*, 2002 Human pregnancy-specific glycoprotein 1a (PSG1a) induces alternative activation in human and mouse monocytes and suppresses the accessory cell-dependent T cell proliferation. Journal of Leukocyte Biology 72: 512–521.

Nakimuli, A., and A. Moffett, 2017 Pregnancy disorders in Africa and the obstetric dilemma. Trans. R. Soc. Trop. Med. Hyg. 110: 681–683.

Nakimuli, A., O. Chazara, L. Farrell, S. E. Hiby, S. Tukwasibwe *et al.*, 2013 Killer cell immunoglobulin-like receptor (KIR) genes and their HLA-C ligands in a Ugandan population. Immunogenetics 65: 765–775.

Nakimuli, A., O. Chazara, S. E. Hiby, L. Farrell, S. Tukwasibwe *et al.*, 2015 A KIR B centromeric region present in Africans but not Europeans protects pregnant women from pre-eclampsia. Proc Natl Acad Sci USA 112: 845.

Nédélec, Y., J. Sanz, G. Baharian, Z. A. Szpiech, A. Pacis *et al.*, 2016 Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens. Cell 167: 657–664.e21.

Parham, P., 2005 Immunogenetics of killer cell immunoglobulin-like receptors. Mol. Immunol. 42: 459–462.

Parham, P., and A. Moffett, 2013 Variable NK cell receptors and their MHC class I ligands in immunity, reproduction and human evolution. Nat. Rev. Immunol. 13: 133-144.

Redman, C. W. G., G. P. Sacks, and I. L. Sargent, 1999 Preeclampsia: An excessive maternal inflammatory response to pregnancy. American Journal of Obstetrics and Gynecology 180: 499–506.

Renfree, M., and G. Shaw, 2001 Reproduction in Monotremes and Marsupials. eLS. doi:10.1038/npg.els.0001856.

Revol de Mendoza, A., D. Escobedo, D. Santiago-Alarcon, and H. Barrera-Saldaña, 2011 Independent Duplication of the Growth Hormone Gene in Three Anthropoidean Lineages. Int. J. Disabil. Hum. Dev. 2: 151-160.

Robbins, J. R., and A. I. Bakardjiev, 2012 Pathogens and the placental fortress. Current opinion in microbiology 15: 36–43.

Ruddle, F. H., G. P. Wagner, J. Kim, and W. J. Bailey, 1997 Phylogenetic reconstruction of vertebrate Hox cluster duplications. Molecular Biology and Evolution 14: 843–853.

Sacks, G., I. Sargent, and C. Redman, 1999 An innate view of human pregnancy.

Immunology Today 20: 114–118.

Sacks, G. P., C. W. G. Redman, and I. L. Sargent, 2003 Monocytes are primed to produce the Th1 type cytokine IL-12 in normal human pregnancy: an intracellular flow cytometric analysis of peripheral blood mononuclear cells. Clin Exp Immunol 131: 490–497.

Sacks, G. P., K. Studena, I. L. Sargent, and C. W. G. Redman, 1998 Normal pregnancy and preeclampsia both produce inflammatory changes in peripheral blood leukocytes akin to those of sepsis. YMOB 179: 80–86.

Sambrook, J. G., A. Bashirova, H. Andersen, M. Piatak, G. S. Vernikos *et al.*, 2006 Identification of the ancestral killer immunoglobulin-like receptor gene in primates. BMC Genomics 7: 209.

Schulte, A. M., S. Lai, A. Kurtz, F. Czubayko, A. T. Riegel *et al.*, 1996 Human trophoblast and choriocarcinoma expression of the growth factor pleiotrophin attributable to germ-line insertion of an endogenous retrovirus. Proc Natl Acad Sci USA 93: 14759.

Selwood, L., 2000 Marsupial Egg and Embryo Coats. Cell Tissues Organs. 166: 208-2019.

Sholtis, S. J., and J. P. Noonan, 2010 Gene regulation and the origins of human biological uniqueness. Trends in Genetics 26: 110–118.

Sol-Church, K., G. N. Picerno, D. L. Stabley, J. Frenck, S. Xing *et al.*, 2002 Evolution of placentally expressed cathepsins. Biochemical and Biophysical Research Communications 293: 23–29.

Somerset, D. A., Y. Zheng, M. D. Kilby, D. M. Sansom, and M. T. Drayson, 2004 Normal human pregnancy is associated with an elevation in the immune suppressive CD25+ CD4+ regulatory T-cell subset. Immunology 112: 38–43.

Soncin, F., M. Khater, C. To, D. Pizzo, O. Farah *et al.*, 2018 Comparative analysis of mouse and human placentae across gestation reveals species-specific regulators of placental development. Development 145: dev156273.

Soshnikova, N., R. Dewaele, P. Janvier, R. Krumlauf, and D. Duboule, 2013 Duplications of hox gene clusters and the emergence of vertebrates. Developmental Biology 378: 194–199.

Svensson, J., M. C. Jenmalm, A. Matussek, R. Geffers, G. Berg *et al.*, 2011 Macrophages at the Fetal–Maternal Interface Express Markers of Alternative Activation and Are Induced by M-CSF and IL-10. J. Immunol. 187: 3671.

Svensson-Arvelund, J., J. Ernerudh, E. Buse, J. M. Cline, J.-D. Haeger *et al.*, 2013 The Placenta in Toxicology. Part II. Toxicol Pathol 42: 327–338.

Than, N. G., R. Romero, M. Goodman, A. Weckle, J. Xing *et al.*, 2009 A primate subfamily of galectins expressed at the maternal–fetal interface that promote immune cell death. Proc Natl Acad Sci USA 106: 9731.

Tirado-González, I., N. Freitag, G. Barrientos, V. Shaikly, O. Nagaeva *et al.*, 2013 Galectin-I influences trophoblast immune evasion and emerges as a predictive factor for the outcome of pregnancy.Mol. Hum. Reprod. 19: 43-53.

Tolosa, J. M., K. S. Parsons, P. M. Hansbro, R. Smith, and P. A. B. Wark, 2015 The Placental Protein Syncytin-1 Impairs Antiviral Responses and Exaggerates Inflammatory Responses to Influenza. PLoS ONE 10: e0118629.

Vargas, A., J. Moreau, S. Landry, F. LeBellego, C. Toufaily *et al.*, 2009 Syncytin-2 Plays an Important Role in the Fusion of Human Trophoblast Cells. J. Mol. Biol. 392: 301-318.

Varla-Leftherioti, M., 2004 Role of a KIR/HLA-C allorecognition system in pregnancy. Journal of Reproductive Immunology 62: 19–27.

Veenstra van Nieuwenhoven, A. L., M. J Heineman, and M. Faas, 2003 The immunology of successful pregnancy. Hum Reprod. Update. 9: 347-357.

Wauquier, N., C. Padilla, P. Becquart, E. Leroy, and V. Vieillard, 2010 Association of KIR2DS1 and KIR2DS3 with fatal outcome in Ebola virus infection. Immunogenetics. 63: 767-771.

Wessells, J., D. Wessner, R. Parsells, K. White, D. Finkenzeller *et al.*, 2000 Pregnancy specific glycoprotein 18 induces IL-10 expression in murine macrophages. Eur. J. Immunol. 30: 1830–1840.

Wildman, D. E., C. Chen, O. Erez, L. I. Grossman, M. Goodman *et al.*, 2006 Evolution of the mammalian placenta revealed by phylogenetic analysis. Proc. Natl. Acad. Sci. U.S.A. 103: 3203.

Wray, G. A., 2007 The evolutionary significance of cis-regulatory mutations. Nature Reviews Genetics 8: 206–216.

Xiong, S., A. M. Sharkey, P. R. Kennedy, L. Gardner, L. E. Farrell *et al.*, 2013 Maternal uterine NK cell–activating receptor KIR2DS1 enhances placentation. J Clin Invest 123: 4264–4272.

Yamanaka, M., Y. Kato, T. Angata, and H. Narimatsu, 2009 Deletion polymorphism of SIGLEC14 and its functional implications. Glycobiology 19: 841–846.

Yang, M. L., Y. H. Chen, S. W. Wang, Y. J. Huang, C. H. Leu *et al.*, 2011 Galectin-1 Binds to Influenza Virus and Ameliorates Influenza Virus Pathogenesis. J. Virol. 85: 10010.

Yi, J. M., and H. S. Kim, 2007 Molecular Phylogenetic Analysis of the Human Endogenous

Retrovirus E (HERV-E) Family in Human Tissues and Human Cancers. Genes & Genetic Systems 82: 89–98.

Zeller, and Freyer, 2001 Early ontogeny and placentation of the grey short-tailed opossum, Monodelphis domestica (Didelphidae: Marsupialia): contribution to the reconstruction of the marsupial morphotype. J Zoological System 39: 137–158.

Zhou, W., F. Zhang, and T. M. Aune, 2003 Either IL-2 or IL-12 Is Sufficient to Direct Th1 Differentiation by Nonobese Diabetic T Cells. J. Immunol. 170: 735.

CHAPTER II

Summary of the population genetics methods used in this dissertation

Abstract

Since their origin in Africa approximately 250,000 - 200,000 years ago, modern humans have undergone several massive changes, which have exposed them to novel selective pressures. Adaptation to such changes has likely left signatures on the genomes of modern humans, in the form of changes in both allele and haplotype frequencies. The availability of high quality whole genome sequence data that spans numerous populations, coupled with various population genetics metrics that can capture distinct aspects of genetic variation present within extant humans, has enabled detecting episodes of past recent positive selection that occurred at different time periods in human history. In this chapter, I briefly discuss the population genetics metrics I have used in my research and how these metrics can be utilized in detecting selection events that took place throughout different time periods in recent human history. As these metrics are affected by past demographic events, I also describe how simulations of neutral evolution that explicitly account for past demographic events can be used to avoid false positive instances of recent positive selection.

Introduction

Since the emergence of the anatomically modern humans in Africa approximately 200,000 - 250,000 years, several distinctive events have happened (Figure 2.1): approximately 50,000 to 75,000 years ago, modern humans left Africa and spread out to novel environments worldwide; the Eurasian population split into modern Europeans and Asians approximately 36,000 to 45,000 years ago; the Agricultural Revolution occurred around 10,000 to 20,000 years ago, resulting in the emergence of farming practices and establishment of communal living (Sabeti *et al.* 2006; Karlsson *et al.* 2014; Page *et al.* 2016). All of these major events have exposed human populations to novel selective pressures such as new pathogens and dietary changes. The resulting selection events have would have likely left genomic signatures, including changes in allele and haplotype frequencies. Therefore, examining patterns of genetic variation in present day humans could shed light on the actions of past selection events. In this regard, a number of studies have been carried out to uncover the consequences of past selection events on the human genome (Voight *et al.* 2006; Sabeti *et al.* 2007; Pickrell *et al.* 2009).

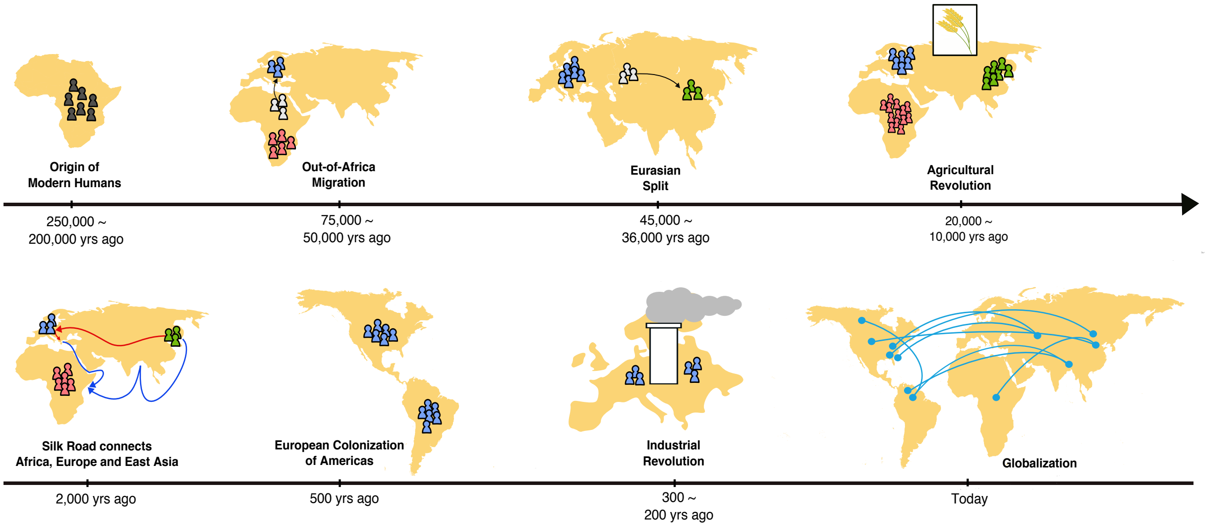

Figure 2.1. Several major events happened in recent human history in the past 200,000 - 250,000 years.

Studies of recent evolution require large-scale genetic variation data. Throughout the years, improvements in sequencing technology have made high quality whole genome sequencing data readily available, facilitating large-scale population genetics analyses (Pool *et al.* 2010; Fumagalli *et al.* 2013; Wall *et al.* 2017). For example, the efforts of the final phase of the 1000 Genome Project (Phase III), aided by improvement in sequence data accuracy, genotype calling, and phasing algorithms, has resulted in the collection of whole genome sequence data from over 2,500 individuals belonging to 26 different subpopulations, which are further classified into five super groups: Africans, Europeans, East Asians, South Asians, and Americans (The 1000 Genomes Project Consortium *et al.* 2015). Furthermore, new metrics that are sensitive to various genomic signatures (Biswas and Akey 2006; Vitti *et al.* 2013) have expanded the inferences of past selection that can be made (Figure 2.2). In addition, software that are capable of simultaneous analysis of the sequence data from large number of individuals have been developed (Danecek *et al.* 2011; Pfeifer *et al.* 2014; Szpiech and Hernandez 2014), further facilitating large scale studies of recent selection (Figure 2.2). In the following sections, I will discuss in more detail the metrics of recent positive selection I used for my research and how each metric can shed light on the actions of past selection that have acted on genomic regions of interest in modern humans.

Hard selective sweep model

Also referred to as to classical selective sweeps, hard selective sweeps occur when a single advantageous variant arises after the onset of selection and sweeps (i.e., increases in frequency) within the population, eventually becoming fixed in the population (i.e., reaches a frequency of 100%) (Messer and Petrov 2013). As the selected variant rises in frequency within the population, linked neutral variants also become more prevalent, drastically reducing genetic diversity in the region surrounding the selected locus (Figure 2.2). This results in several distinctive genomic signatures, such as excess of rare alleles, increased population differentiation, and the presence of a single long, dominant haplotype (Sabeti *et al.* 2006; Messer and Petrov 2013). In addition, different metrics have been shown to be capable of detecting genomic signatures of selection that happened over different time ranges throughout recent human history (Sabeti *et al.* 2006; Voight *et al.* 2006) (Figure 2.2). For instance, Tajima's *D*, a metric that compares the average number of pairwise differences (π) with the number of segregating sites (θ), can detect excess of rare alleles (Tajima 1989) that have resulted from selective sweeps. More specifically, negative values of Tajima's *D*, which reflect presence of more segregating sites than pairwise differences, are indicative of selective sweeps. This is because selective sweeps often result in the presence of one dominant haplotype within a population and therefore, any new mutation is likely to be rare. As mutagenesis is rare in humans, this signature persists for a relatively long period of time and therefore, Tajima's *D* has been proposed to be able to detect selection events that happened approximately 200,000 - 250,000 years ago, or around the origin of modern humans (Figure 2.2) (Sabeti *et al.* 2006).

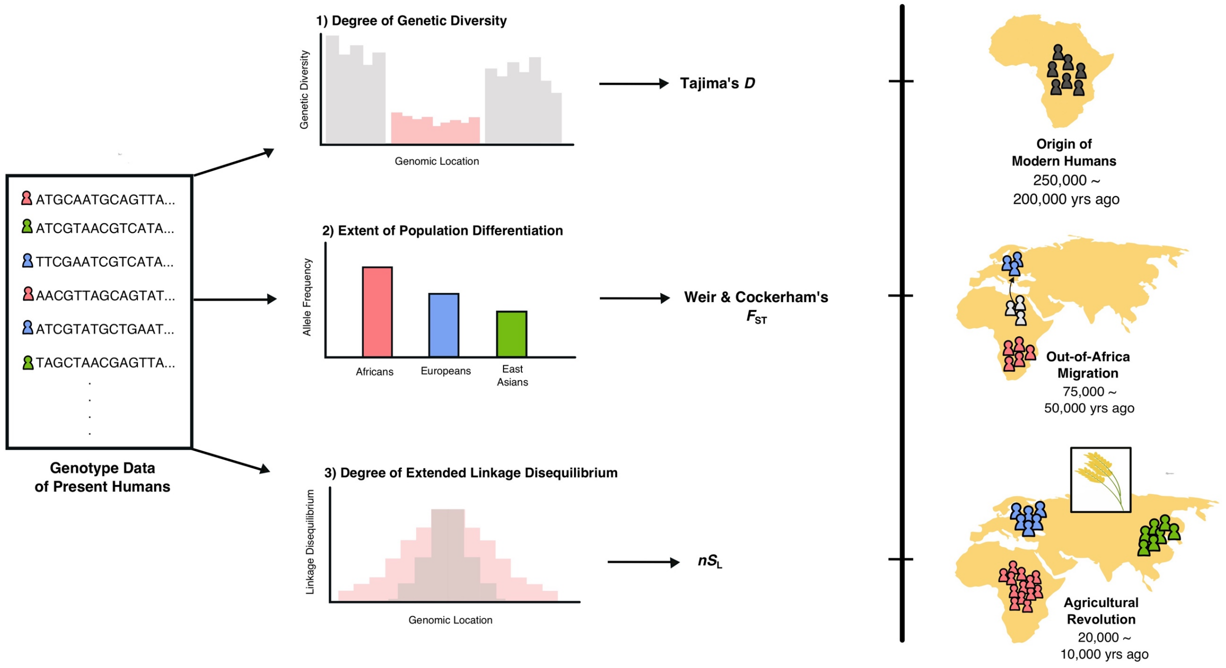

Figure 2.2. Actions of selection at different time ranges in human history leave distinct types of genomic signatures that can be detected by various population genetics metrics. This figure illustrates how calculating population genetics metrics using the genotype data of modern human populations can shed light on selection events that happened in the past. Different population genetics metrics are sensitive to distinct types of genomic signatures resulting from the action of recent selection. In addition, as these genomic signatures persist for different lengths of time, these metrics can be used

to infer selection events that happened in different time ranges in human history. For instance, Tajima's *D* detects the excess of rare alleles, which persist for the longest period of time because mutagenesis is rare in humans. Therefore, Tajima's *D* can be used to infer selection events that happened approximately 200,000 to 250,000 years ago. In contrast, the presence of extended linkage disequilibrium (or haplotype) breaks down relatively quickly by recombination and therefore, this signature remains for the shortest period of time. nS_L is sensitive to such signatures and therefore, can be used to detect selection events that happened approximately 10,000 to 20,000 years ago.

Following the out-of-Africa migration that occurred approximately 50,000 - 75,000 years ago, humans that settled in new environments worldwide would have been exposed to a suite of novel selective agents. Local adaptations to different environments are expected to result in differences in allele frequencies among geographically distinct populations (i.e., increased population differentiation), which can be detected by a metric called fixation index (F_{ST}) (Figure 2.2) (Wright 1951). Multiple estimators have been developed for this metric (Nei 1973) (Nei 1986; Weir and Cockerham 1984), and for my studies, I used the estimator developed by Weir and Cockerham (Weir and Cockerham 1984), which is unbiased in terms of sample sizes of the populations being examined.

As briefly mentioned above, one characteristic signature of hard selective sweeps is the presence of a single, extended haplotype within the population, which results from hitch-hiking of linked neutral variants along with the selected variant (Figure 2.2) (Biswas and Akey 2006; Sabeti *et al.* 2006; Voight *et al.* 2006; Messer and Petrov 2013). Metrics such as extended haplotype homozygosity (EHH) (Sabeti *et al.* 2007) that 1) measure the decay of linkage disequilibrium (LD) of the selected variant with others over varying distances, or 2) integrate EHH values over a specified distance surrounding a variant of interest (e.g., integrated haplotype score or iHS) (Voight *et al.* 2006) can be used to detect such genomic signatures. As recombination can rapidly break down the haplotype, this signature persists for the shortest period of time, and can be used to infer selection events that happened approximately 10,000 to 20,000 years ago, or around the Agricultural Revolution (Figure 2.2). In addition, haplotype-based metrics also exhibit increased sensitivity to incidences of incomplete or partial selective sweeps, in which the selected variant does not reach complete fixation (i.e., reaches an intermediate frequency of 65 - 85%) (Sabeti *et al.* 2006; Voight *et al.* 2006). I used nS_L (Number of Segregating Sites by Length; (Ferrer-Admetlla *et al.* 2014)), a metric that is very similar to iHS but more robust to variations in recombination rates (Figure 2.2) (Ferrer-Admetlla *et al.* 2014). More specifically, nS_L integrates EHH values calculated over various intervals from the specified variant by using number of segregating sites as proxy for distance, while iHS instead uses recombination distance for integration. Similar to iHS, higher absolute values of nS_L indicate the presence of long haplotypes, and therefore, the action of selective sweeps on the variant of interest.

Alternative mode of selection: soft selective sweep model

While hard selective sweeps have been the most widely studied mode of recent selection in humans, it has been recently argued that this may not be the dominant mode of selection in modern humans (Pennings and Hermisson 2006a; 2006b; Hernandez *et al.* 2011; Messer and Petrov 2013). Recently proposed as an alternative mode of selection, soft selective sweeps occur when standing variation or multiple *de novo* mutations introduce several variants into the population (Pennings and Hermisson 2006a; Messer and Petrov 2013). In either scenario, soft selective sweeps result in the presence of more than one, independent haplotypes within a population (Figure 2.3). As no single variant becomes completely fixed, there is no drastic reduction in genetic diversity as seen in the case of hard selective sweeps. Therefore, metrics designed to detect hard selective sweeps are known to have very low power to detect signatures resulting from such soft selective sweeps (Messer and Petrov 2013). For instance, one study found that while Tajima' *D* retains power to detect incidences of selective sweeps resulting in allele frequencies close to 100% (reflective of hard selective sweeps), it quickly loses power as the allele frequency falls to more moderate levels (65-85%) (Ferrer-Admetlla *et al.* 2014). In this regard, Garud et al. devised a novel metric, the H12, that combines the frequencies of the most common haplotype (H1) and the second most common haplotype (H2) (Figure 2.3) (Garud *et al.* 2015). While still biased towards detecting hard selective sweeps, Garud et al. showed that H12 exhibits increased power to detect soft selective sweeps compared to other metrics such as iHS, as long as the sweeps are not "too soft" (i.e., selection has acted on multiple variants, resulting in a high number of independent haplotypes being present within a population) (Garud *et al.* 2015).

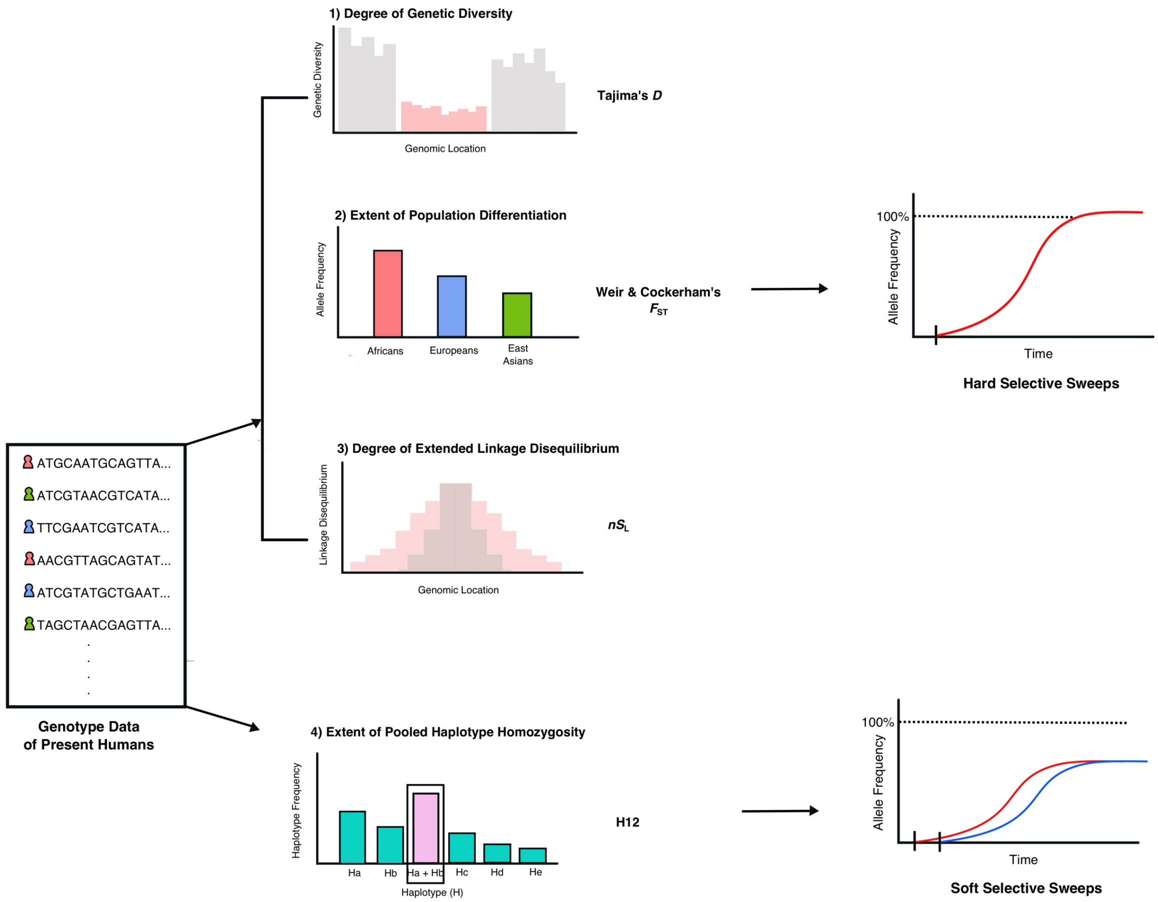

Figure 2.3. Different population genetics metrics have power to distinguish signatures left by distinct modes of selection. This figure compares two models of selective sweeps: hard and soft selective sweeps. Under a hard selective sweep model, a single advantageous variant rises in frequency within the population after the onset of selection, resulting in drastic reduction in genetic diversity in the region surrounding the selected variant. In contrast, soft selective sweeps occur when selection acts on multiple variants (arising from either standing variation or multiple *de novo* mutations), resulting in the presence of multiple, independent haplotypes within the population. Metrics such as Tajima's D, Weir & Cockerham's F_{ST}, and nS_L are known to detect incidences of hard selective sweeps, whereas H12, which is a metric of pooled haplotype homozygosity, has increased power to detect soft selective sweeps.

Simulations of neutral evolution

Actions of past demographic events, such as population expansion and/or contraction can result in similar changes in genetic variation as selection (Sabeti *et al.* 2006; Nielsen *et al.* 2007). For instance, during a bottleneck event, rare alleles are likely to get removed from a population due to sampling effect, resulting in positive values of Tajima's *D* (i.e., excess of alleles with intermediate frequencies) (Figure 2.4). Immediately after such a bottleneck event, Tajima's *D* values will drop to negative values, as genetic diversity has been dramatically reduced and new mutations emerging afterwards will initially be rare (Figure 2.4). As the population expands, more rare mutations will arise via *de novo* mutations, and as a result, Tajima's *D* values will further decrease (Figure 2.4). This example demonstrates how non-adaptive processes alone can change the values of metrics of recent positive selection. Therefore, if past demographic events are not accounted for, one could incorrectly conclude that a region has undergone selection (Nielsen *et al.* 2007) in the absence of selection.

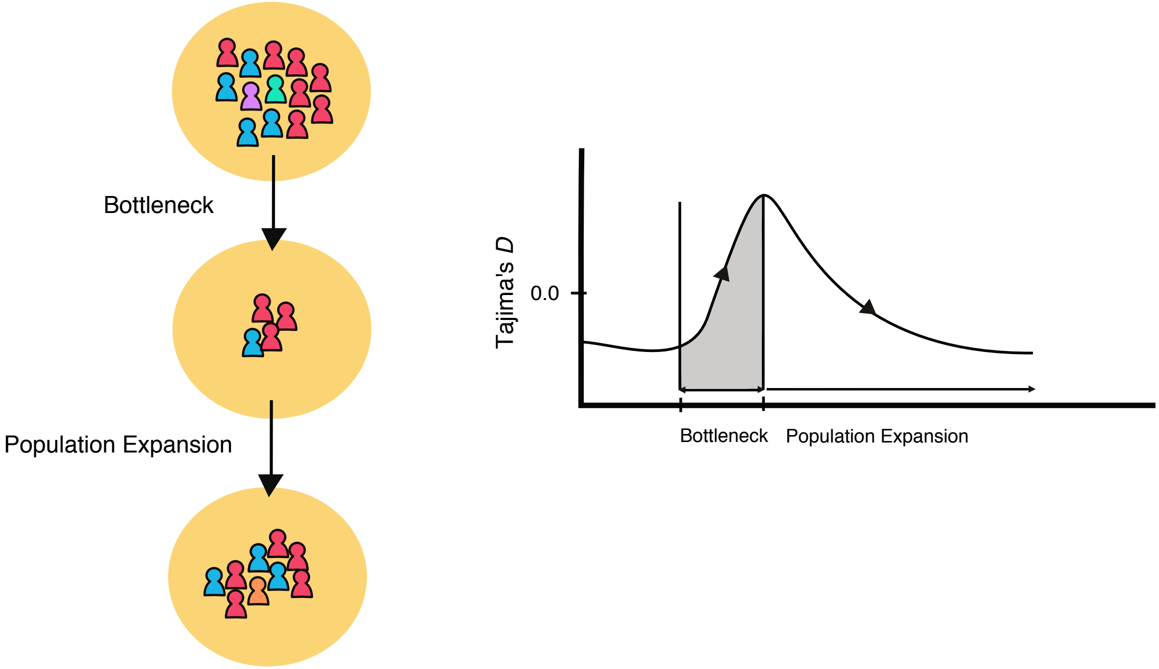

Figure 2.4. Confounding effects of past demographic events in studies of recent positive selection. This example illustrates how non-adaptive processes such as changes in population size can alone alter the values of the metrics of recent positive

selection in the absence of selection. An extreme contraction in population size (i.e., bottleneck) would likely result in deficiency of rare alleles due to sampling effect, resulting in positive values of Tajima's *D*. As the population recovers from such a bottleneck event, novel rare mutations will emerge via de novo, lowering the values of Tajima's *D*.

One way to account for the confounding effects of non-adaptive processes is to generate sequences that evolved neutrally while experiencing similar demographic events as modern humans (Figure 2.5). Metrics of recent positive selection then can be calculated on such neutrally simulated sequences and compared to the observed values. Several attempts, including that of Gravel and his colleagues (Plagnol and Wall 2006; Nielsen *et al.* 2009; Gravel *et al.* 2011), have been made to infer parameters of past demographic events. For instance, Gravel et al. estimated maximum likelihood parameters for three super-populations (Africans, Europeans, East Asians) by combining low-coverage whole genome data and high-coverage targeted exon data of the pilot phase 1000 Genomes Project (Gravel *et al.* 2011). Inferred parameters include effective population sizes, migration rates, mutation rates, magnitudes of changes in the population, and timing of major splitting events.

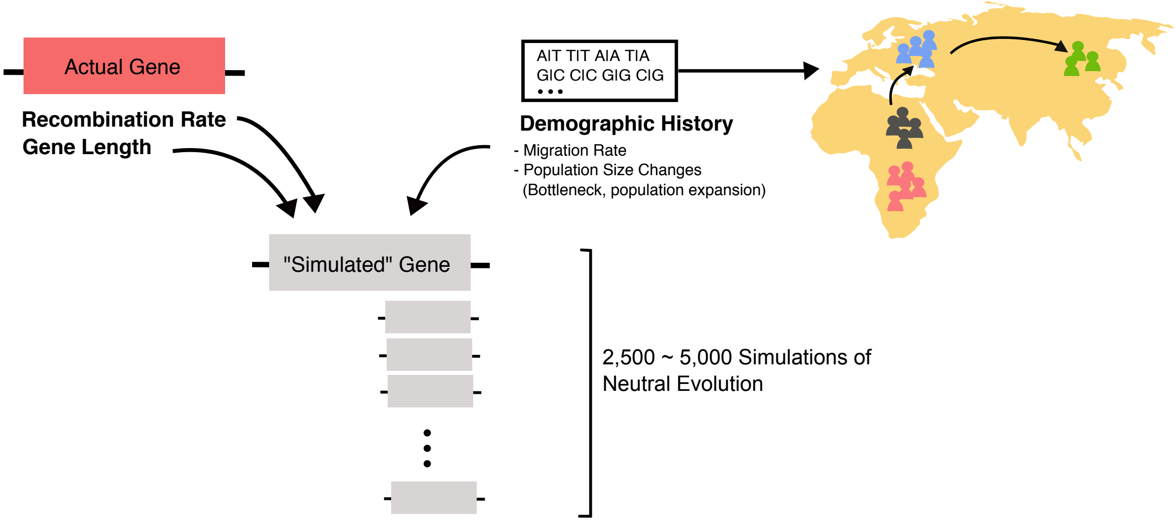

Figure 2.5. Simulations of neutral evolution. A schematic illustration of simulations of neutral evolution. Pre-estimated parameters of past demographic events including migration rates and magnitude of population size changes, as well as other attributes associated with the "actual" region of interest such as recombination rates and length of the region can be incorporated into simulations of neutral evolution to create "simulated" regions that closely match the "actual" region. In my studies, I carried out 2,500 (for generating regions used for nS_L calculations) to 10,000 (for other metrics) simulations of neutral evolution.

Simulation programs such as *SLiM* allow execution of complicated simulations of neutral evolution with explicit incorporation of the estimated demographic parameters (Messer 2013; Haller and Messer 2017). To create a neutrally "simulated" region that matches the "actual" region of interest as closely as possible, other attributes (derived from the empirical data) such as recombination rate and length of the genetic region can be specified as well (Figure 2.5). To assess the likelihood that the observed metrics of recent positive selection have resulted from actions of selection, these values can be compared to the distribution of same metrics calculated on the neutrally simulated sequences (Figure 2.6). An empirical *p*-value, calculated as the proportion of simulated sequences that exhibit values equal to or more extreme than the observed value, can be calculated and compared to a pre-set significance threshold. For my studies, I used a *p*-value cutoff of 0.05. In short, a *p*-value lower than 0.05 indicates significant deviations from expectations of neutral evolution and therefore, significant evidence of recent positive selection (Figure 2.6).

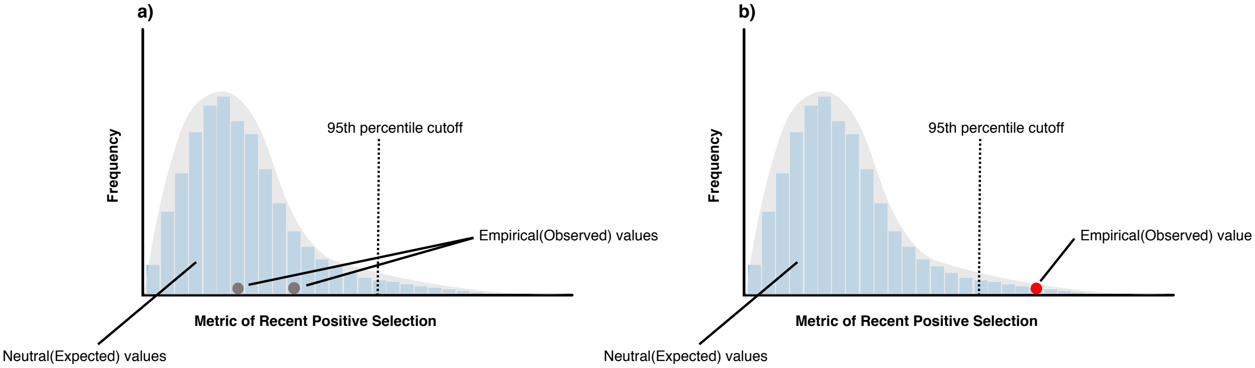

Figure 2.6. Assessing significance of the empirical metrics of recent positive selection. A schematic illustration of assessing whether a region of interest exhibits significant deviations from neutral expectations (i.e., exhibits evidence of recent positive selection). The empirical value of the calculated metric can be compared to the distribution of values calculated using the neutrally simulated sequences. Empirical values that are "more extreme" than the 95% percentile of distribution of neutrally simulated values are considered significant.

Software and commands

More detailed and specific commands for the software used in each of my studies, as well as the location of the data files used, can be found in Additional Files 3.1 and 4.1. In this section, I will briefly outline the software and the generalized commands that were used to calculate each metric of recent positive selection.

For calculations of Weir & Cockerham's F_{ST}, I used *VCFtools* (v.0.1.13) (Danecek *et al.* 2011). I calculated the global F_{ST} among three super populations (Africans, Europeans, East Asians) using the following command:

```
./vcftools --gzvcf input.vcf.gz --chr chr_id --from-bp start_pos --to-bp end_pos --remove-indels --weir-fst-pop AFR_ --weir-fst-pop EUR --weir-fst-pop EAS --out output    (1)
```

'input.vcf.gz' refers to the compressed VCF file, sorted by chromosome, containing the genotype data, created by the 1000 Genomes Project. 'chr_id' indicates the chromosome on which the region of interest in located on. 'start_pos' and 'end_pos' refer to start and end positions of the region of interest, in base pairs, respectively. 'AFR', 'EUR', and 'EAS' refer to files that list the individuals labeled as Africans, Europeans, and East Asians in the 1000 Genomes Project data, respectively.

Tajima's *D* was calculated using the R package PopGenome (version 2.1.6) (Pfeifer *et al.* 2014), via the "neutrality.stats" function as follows:

```
Input_vcf <- readVCF(input_vcf, 1000, tid = chromosome_id, from = start_pos, to = end_pos, approx. = FALSE)    (2)
all_pops <- as.character(read.table(3pops_samples_list))[[1]]    (3)
Input_neut_all_pops <- neutrality.stats(input_vcf_data, list(all_pops))    (4)
Input_TajimaD_all_pops <- input_neut_all_pops@Tajima.D[[1]]    (5)
```

The first line reads the input VCF file containing the genotypes in the region of interest (‘input_vcf’), from ‘start_pos’ to ‘end_pos’. The second line reads the file listing the IDs of all individuals belonging to the 3 populations (Africans, Europeans, East Asians) (‘3pops_samples_list’) into R. The third line computes a set of neutrality statistics, including Tajima’s *D* Fu and Li’s *D** and *F**, using the genotype file read from (2) of individuals belonging to the 3 populations of interest that have been called using (3). The last line extracts the value of Tajima’s *D* from the list of neutrality statistics calculated in the previous step.

Next, for the calculation of nS_L, I used *Selscan* (version 1.2.0) (Szpiech and Hernandez 2014) using the following command:

```
./selscan --nsl --vcf input_vcf --maf 0.01 --threads 16 --out output                    (6)
```

As with other metrics, ‘input_vcf’ refers to the VCF file that contains genotype data spanning the region of interest. Apart from the minor allele frequency cutoff value (0.01), I used the default settings of *Selscan*. The maximum absolute value of the unstandardized nS_L calculated over the entire window was used to represent the entire region of interest being examined.

Finally, I used *SLIM* (Messer 2013; Haller and Messer 2017) to generate VCF files of variants resulting from the action of neutral evolution alone, given past demographic events. I incorporated *SLiM*'s implementation of Gravel et al's pre-computed parameters of human demographic history (Gravel *et al.* 2011) into my neutral simulations. More specifically, at the beginning of the simulation (i.e. generation 1), the ancestral African effective population size was set to 7,310, which next expanded to 14,474 approximately 148,000 years ago (i.e. 5,920 generations ago). Approximately 51,000 years ago (i.e. 2,040 generations ago), the non-Africans split from Africans; the initial effective population size of these non-Africans was set to 1,861. The migration rates between Africans and non-Africans were set to 15 x 10^{-5}. Next, approximately 23,000 years ago (i.e. 920 generations ago), the above-mentioned non-African population split into European and East Asian populations, with the initial effective population size for East Asians set to 554. In the same generation, the European effective population size was reduced to 1,032. The following migration rates were established for the remainder of the simulation: 2.5 x 10^{-5} for between Africans and Europeans, 0.78 x 10^{-5} for between Africans and Asians, and 3.11 x 10^{-5} for between Europeans and East Asians. Between generations 57,080 and 58,000, the European and East Asian populations were set to experience increase in their effective population sizes: for Europeans, the exponential coefficient was 0.0038 and 0.0048 for East Asians. For each region being studied, I set the length of the genomic element being simulated equal to the length of the region of interest. For instance, to simulate a region that is 1,000bp long, the following argument was used:

```
initializeGenomicElement(g1, 0, 999)                                        (7)
```

After 58,000 generations (i.e. end of the simulation), I sampled 661, 503, and 504 individuals from the “simulated” African, European, and East Asian population, respectively, to match the number of individuals included in the 1000 Genomes Project dataset. I used the following arguments to ensure that pairs of genomes being sampled belonged to the same individual:

```
p1_sample = p1.individuals;                                   (8)
sampled_p1 = sample(p1_sample, 661);                          (9)
p2_sample = p2.individuals;                                   (10)
sampled_p2 = sample(p2_sample, 503);                          (11)
pe_sample = p3.individuals;                                   (12)
sampled_p3 = sample(p3_sample, 504);                          (13)
```

p1, p2, and p3, correspond to the simulated African, European, and East Asian populations. Finally, to specify the output format as VCF, I used the commands shown below:

```
sampled_individuals = c(sampled_p1, sampled_p2, sampled_p3);  (14)
sampled_individuals.genomes.outputVCF();                      (15)
```

References

Biswas, S., and J. M. Akey, 2006 Genomic insights into positive selection. Trends in Genetics 22: 437–446.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks *et al.*, 2011 The variant call format and VCFtools. Bioinformatics 27: 2156–2158.

Ferrer-Admetlla, A., M. Liang, T. Korneliussen, and R. Nielsen, 2014 On Detecting Incomplete Soft or Hard Selective Sweeps Using Haplotype Structure. Molecular Biology and Evolution 31: 1275–1291.

Fumagalli, M., F. G. Vieira, T. S. Korneliussen, T. Linderoth, E. Huerta-Sánchez *et al.*, 2013 Quantifying Population Genetic Differentiation from Next-Generation Sequencing Data. Genetics 195: 979.

Garud, N. R., P. W. Messer, E. O. Buzbas, and D. A. Petrov, 2015 Recent Selective Sweeps in North American Drosophila melanogaster Show Signatures of Soft Sweeps (G. P. Copenhaver, Ed.). PLoS Genet 11: e1005004.

Gravel, S., B. M. Henn, R. N. Gutenkunst, A. R. Indap, G. T. Marth *et al.*, 2011 Demographic history and rare allele sharing among human populations. Proceedings of the National Academy of Sciences 108: 11983–11988.

Haller, B. C., and P. W. Messer, 2017 SLiM 2: Flexible, Interactive Forward Genetic Simulations. Molecular Biology and Evolution 34: 230–240.

Hernandez, R. D., J. L. Kelley, E. Elyashiv, S. C. Melton, A. Auton *et al.*, 2011 Classic selective sweeps were rare in recent human evolution. Science 331: 920–924.

Karlsson, E. K., D. P. Kwiatkowski, and P. C. Sabeti, 2014 Natural selection and infectious disease in human populations. Nature Reviews Genetics 15: 379–393.

Messer, P. W., 2013 SLiM: Simulating Evolution with Selection and Linkage. Genetics 194: 1037–.

Messer, P. W., and D. A. Petrov, 2013 Population genomics of rapidadaptation by soft selective sweeps. Trends in Ecology & Evolution 1–11.

Nei, M., 1973 Analysis of gene diversity in subdivided populations. Proceedings of the National Academy of Sciences 70: 3321–3323.

Nei, M., and T. Gojobori, 1986 Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Molecular Biology and Evolution 3: 418–426.

Nielsen, R., I. Hellmann, M. Hubisz, C. Bustamante, and A. G. Clark, 2007 Recent and ongoing selection in the human genome. Nature Reviews Genetics 8: 857–868.

Nielsen, R., M. J. Hubisz, I. Hellmann, D. Torgerson, A. M. Andrés *et al.*, 2009 Darwinian and demographic forces affecting human protein coding genes. Genome Research 19: 838–849.

Page, A. E., S. Viguier, M. Dyble, D. Smith, N. Chaudhary *et al.*, 2016 Reproductive trade-offs in extant hunter-gatherers suggest adaptive mechanism for the Neolithic expansion. Proceedings of the National Academy of Sciences 113: 4694–4699.

Pennings, P. S., and J. Hermisson, 2006a Soft Sweeps III: The Signature of Positive Selection from Recurrent Mutation. PLoS Genet 2: e186.

Pennings, P. S., and J. Hermisson, 2006b Soft Sweeps II—Molecular Population Genetics of Adaptation from Recurrent Mutation or Migration. Molecular Biology and Evolution 23: 1076–1084.

Pfeifer, B., U. Wittelsbürger, S. E. Ramos-Onsins, and M. J. Lercher, 2014 PopGenome: An Efficient Swiss Army Knife for Population Genomic Analyses in R. Molecular Biology and Evolution 31: 1929–1936.

Pickrell, J. K., G. Coop, J. Novembre, S. Kudaravalli, J. Z. Li *et al.*, 2009 Signals of recent positive selection in a worldwide sample of human populations. Genome Research 19: 826–837.

Plagnol, V., and J. D. Wall, 2006 Possible Ancestral Structure in Human Populations. PLoS Genet 2: e105.

Pool, J. E., I. Hellmann, J. D. Jensen, and R. Nielsen, 2010 Population genetic inference from genomic sequence variation. Genome Research 20: 291–300.

Sabeti, P. C., S. F. Schaffner, B. Fry, J. Lohmueller, P. Varilly *et al.*, 2006 Positive natural selection in the human lineage. Science 312: 1614–1620.

Sabeti, P. C., P. Varilly, B. Fry, J. Lohmueller, E. Hostetter *et al.*, 2007 Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913–918.

Szpiech, Z. A., and R. D. Hernandez, 2014 selscan: An Efficient Multithreaded Program to Perform EHH-Based Scans for Positive Selection. Molecular Biology and Evolution 31: 2824–2827.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 123: 585-595.

The 1000 Genomes Project Consortium, R. A. Gibbs, E. Boerwinkle, H. Doddapaneni, Y.

Han *et al.*, 2015 A global reference for human genetic variation. Nature 526: 68–74.

Vitti, J. J., S. R. Grossman, and P. C. Sabeti, 2013 Detecting Natural Selection in Genomic Data. Annu. Rev. Genet. 47: 97–120.

Voight, B. F., S. Kudaravalli, X. Wen, and J. K. Pritchard, 2006 A Map of Recent Positive Selection in the Human Genome (L. Hurst, Ed.). PLoS Biol 4: e72.

Wall, D. P., J. P. A. Ioannidis, M. J. Khoury, E. A. Ashley, and R. L. Goldfeder, 2017 Human Genome Sequencing at the Population Scale: A Primer on High-Throughput DNA Sequencing and Analysis. aje 186: 1000–1009.

Weir, B. S., and C. C. Cockerham, 1984 Estimating F-Statistics for the Analysis of Population Structure. Evol. 38: 1358–1370.

Wright, S, 1951. The Genetical Structure of Populations. Annals of Eugenics, 15, 323-354. http://dx.doi.org/10.1111/j.1469-1809.1949.tb02451.x

CHAPTER III

Examination of signatures of recent positive selection on genes involved in human sialic acid biology[2,3]

Abstract

Sialic acids are nine carbon sugars ubiquitously found on the surfaces of vertebrate cells and are involved in various immune response-related processes. In humans, at least 58 genes spanning diverse functions, from biosynthesis and activation to recycling and degradation, are involved in sialic acid biology. Because of their role in immunity, sialic acid biology genes have been hypothesized to exhibit elevated rates of evolutionary change. Consistent with this hypothesis, several genes involved in sialic acid biology have experienced higher rates of non-synonymous substitutions in the human lineage than their counterparts in other great apes, perhaps in response to ancient pathogens that infected hominins millions of years ago (paleopathogens). To test whether sialic acid biology genes have also experienced more recent positive selection during the evolution of the modern human lineage, reflecting adaptation to contemporary cosmopolitan or geographically-restricted pathogens, we examined whether their protein-coding regions showed evidence of recent hard and soft selective sweeps. This examination involved the calculation of four measures that quantify changes in allele frequency spectra, extent of population differentiation, and haplotype homozygosity caused by recent hard and soft

[2] This chapter was published in G3. 2018 April 1;8:4. doi.org/10.1534/g3.118.200035

[3] The contributing authors are: Jiyun M. Moon, David M. Aronoff, John A. Capra, Patrick Abbot, and Antonis Rokas.

selective sweeps for 55 sialic acid biology genes using publicly available whole genome sequencing data from 1,668 humans from three ethnic groups. To disentangle evidence for selection from confounding demographic effects, we compared the observed patterns in sialic acid biology genes to simulated sequences of the same length under a model of neutral evolution that takes into account human demographic history. We found that the patterns of genetic variation of most sialic acid biology genes did not significantly deviate from neutral expectations and were not significantly different among genes belonging to different functional categories. Those few sialic acid biology genes that significantly deviated from neutrality either experienced soft sweeps or population-specific hard sweeps. Interestingly, while most hard sweeps occurred on genes involved in sialic acid recognition, most soft sweeps involved genes associated with recycling, degradation and activation, transport, and transfer functions. We propose that the lack of signatures of recent positive selection for the majority of the sialic acid biology genes is consistent with the view that these genes regulate immune responses against ancient rather than contemporary cosmopolitan or geographically restricted pathogens.

Introduction

Sialic acids are nine carbon sugars that are commonly found on the ends of glycoconjugates in deuterostome animals (Schauer 1982; Varki 2007). These molecules are involved in several biological processes, such as intercellular adhesion and signaling (Kelm and Schauer 1997; Varki and Varki 2007), and play important roles in the modulation of various aspects of the host immune system (Pilatte *et al*. 1993; Varki and Gagneux 2012) including activation of immune responses, leukocyte trafficking, complement pathway activation, and microbial attachment. Modification of sialic acid molecules during their biosynthesis and subsequent attachments to underlying sugars via different linkages give rise to a diverse repertoire of sialic acids (Angata and Varki 2002; Varki and Varki 2007; Cohen and Varki 2010; Hayakawa and Varki 2011).

More than 50 genes are known to be involved in various aspects of sialic acid biology in humans, and they fall into five broad functional categories (Altheide *et al*. 2006): 1) biosynthesis, 2) activation, transport, and transfer, 3) modification, 4) recognition, and 5) recycling and degradation (Figure 3.1). Biosynthesis genes are involved in assembling sialic acids from precursor molecules in the cytosol, while genes involved in the second category activate the sialic acids by attaching cytidine monophosphate to them, transport the activated sialic acids to the Golgi for attachment to glycoconjugates via multiple forms of α linkages, and transfer them to the cell surface. Activated sialic acids that are linked to the underlying sugar can be further modified by incorporation of additional molecules before the sialylated glycoconjugate is transferred to the cell surface. At the cell surface, sialic acids are recognized by specialized receptors and finally, the used sialic acids end up in the lysosome for recycling and degradation (Altheide *et al*. 2006).

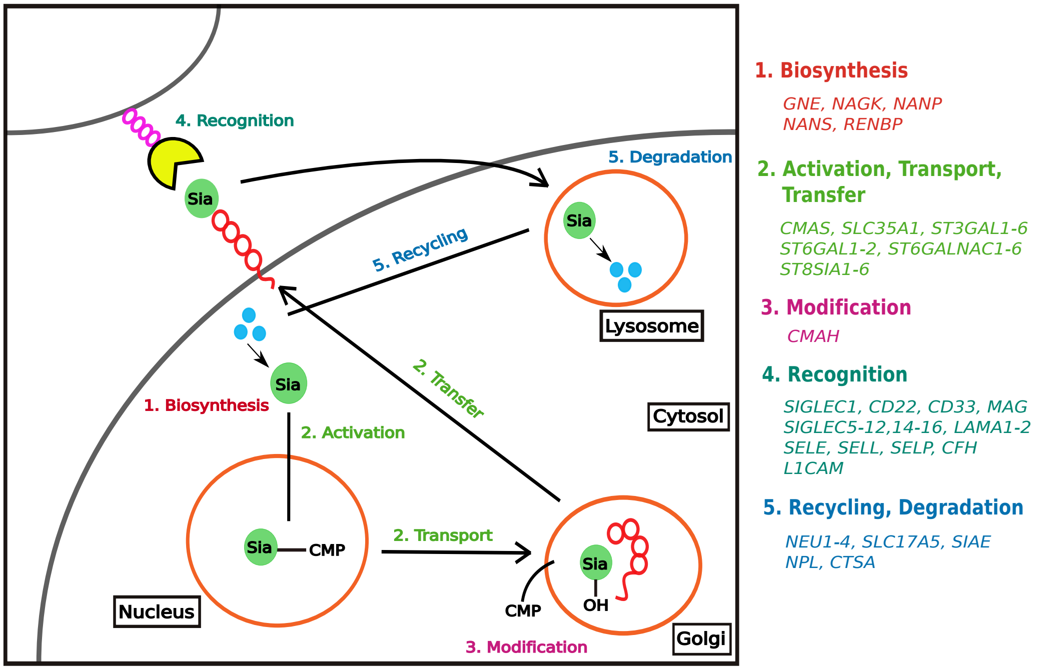

Figure 3.1. Summary of the biochemical processes involved in human sialic acid biology and the genes known to be associated with them

Both the large number of genes involved in human sialic acid biology and their high sequence diversity have been hypothesized to be a result of ancient selective pressures from paleopathogens that infected hominins millions of years ago and left their signatures on genes involved in sialic acid biology (Varki 2009). The most striking example involves the *Alu*-mediated inactivation of the *CMAH* gene specifically in the human lineage (Chou *et al*. 1998; Irie *et al*. 1998; Varki 2001): the enzyme encoded by this gene is responsible for the conversion of N-acetylneuraminic acid (Neu5Ac) to N-glycolylneuraminic acid (Neu5Gc), and Neu5Ac is the main form of sialic acid in humans. This change in sialic acid composition has been suggested to act as a means to escape infection by *Plasmodium reichenowi*, a parasite that preferentially binds to Neu5Gc and is known to cause malaria in chimpanzees. Conversely, adoption of Neu5Ac likely made humans susceptible to infection by the malaria parasite *Plasmodium falciparum*, which instead binds to Neu5Ac (Martin *et al*. 2005; Varki and Gagneux 2009). Additionally, in contrast to humans, the inhibitory receptor Siglec-5 and its immune-stimulatory counterpart Siglec-14 in great apes lack the essential arginine residue required for optimal binding with sialic acids, likely as a consequence of the above-mentioned switch from Neu5Gc to Neu5Ac in the human lineage (Angata *et al*. 2006). Finally, molecular evolutionary analyses in primates and rodents have shown that genes involved in sialic acid recognition exhibit a considerable degree of sequence divergence, even between closely related species, suggestive of ancient positive selection in response to paleopathogens (Altheide *et al*. 2006).

The examples above suggest that sialic acid biology genes experienced adaptive changes early in the evolutionary radiation of primates, including hominids and hominins. However, whether these same genes also experienced positive selection during recent human evolution (i.e., in the last 250,000 years), reflective of adaptation to contemporary pathogens, remains an open question. To test this hypothesis, we estimated the population-level genetic variation of the protein-coding regions of sialic acid biology genes, and carried out tests aimed to detect hard selective sweeps, including tests of deviation from neutrality (measured by Tajima's *D*; Tajima 1989), population differentiation (measured by F_{ST}; Weir and Cockerham 1984), and extended haplotype homozygosity (measured by nS_{L}; Ferrer-Admetlla *et al*. 2014) using the 1000 Genomes Project sequencing data from 1,668 humans belonging to three ethnic groups (Auton *et al*. 2015). To also test for the possible action of soft selective sweeps, we employed H12 (Garud *et al*. 2015), a measure of modified pooled haplotype homozygosity. To determine whether sialic acid biology genes exhibit patterns of genetic variation that deviate from neutral expectations, we compared the values of each of these four measures for sialic acid biology genes against those from sequences of the same length simulated under neutral evolution assuming a realistic model of human demographic history (Gravel *et al*. 2011; Messer 2013; Haller and Messer 2017). Examination of patterns of genetic variation of sialic acid biology genes showed that, irrespective of their functional category, most of them conform to neutral expectations. Sialic acid biology genes that deviate from neutrality exhibit evidence of either temporally or spatially varying positive selection. For example, both recognition genes *SIGLEC5* and *SIGLEC12* show nS_{L} values in Europeans that are consistent with the occurrence of a selective sweep approximately 20,000 years ago, whereas the Tajima's *D* value of the recycling and degradation gene *SIAE* suggests that it experienced a hard selective sweep long before the out-of-Africa migration, approximately 250,000 years ago. Combined with previous work showing the widespread occurrence of ancient positive selection on genes involved in sialic acid biology prior to the emergence of modern humans (Altheide *et al*. 2006), our results suggest that the evolution of human sialic acid biology genes has been more strongly influenced by ancient primate pathogens rather than by contemporary cosmopolitan or geographically-restricted human pathogens.

Methods

Genotype dataset

To examine whether sialic acid biology genes have experienced positive selection during recent human evolution, for all analyses we used the genotype data for 1,668 individuals from three ethnic backgrounds (661 Africans, 503 Europeans, and 504 East Asians) from Phase 3 of the 1000 Genomes Project (for information on data sources, see Additional File 3.1) (Auton *et al*. 2015).

Sialic acid biology genes

To identify all the genes involved in human sialic acid biology, we started from a previously published list of 55 loci (Altheide *et al*. 2006) and added *SIGLEC14* and *SIGLEC16*, which were discovered more recently, as well as *SIGLEC15*, which was not included in the previous study. As sex-linked genes tend to exhibit different patterns of genetic variation than genes located on autosomal chromosomes (Schaffner 2004), we excluded the two genes that are located on the X chromosome, *L1CAM* and *RENBP*; in addition, we removed *CMAH*, the sole gene involved in modification of newly synthesized sialic acids because it is a non-functional pseudogene in humans (Chou *et al*. 1998; Irie *et al*. 1998). Thus, 55 genes were retained for subsequent analyses (Figure 3.1).

To extract genotypes for these 55 genes, we used *VCFtools*, version 0.1.13 *(*Danecek *et al*. 2011) (for specific commands used, see Additional File 3.1), with the gene coordinates (GRCh37p.13) obtained from Ensembl (release 75) via BioMart as input; indels were excluded from our analyses. Compressed VCF files and index files required for subsequent analyses were created using *tabix*, version 0.2.6 (Li 2011).

Measuring signatures of recent positive selection in human sialic acid biology genes

To determine the levels of genetic polymorphism of sialic acid biology genes, we first calculated pairwise nucleotide diversity (π) (Nei and Li 1979) using *PopGenome*, version 2.1.6 (Pfeifer *et al*. 2014). As there may have been variation in when and where human sialic acid biology genes experienced positive selection, we next used four different measures aimed to capture genomic signatures left from the action of selection at different time ranges during recent human evolution (Figure 3.2) (Sabeti *et al*. 2006).

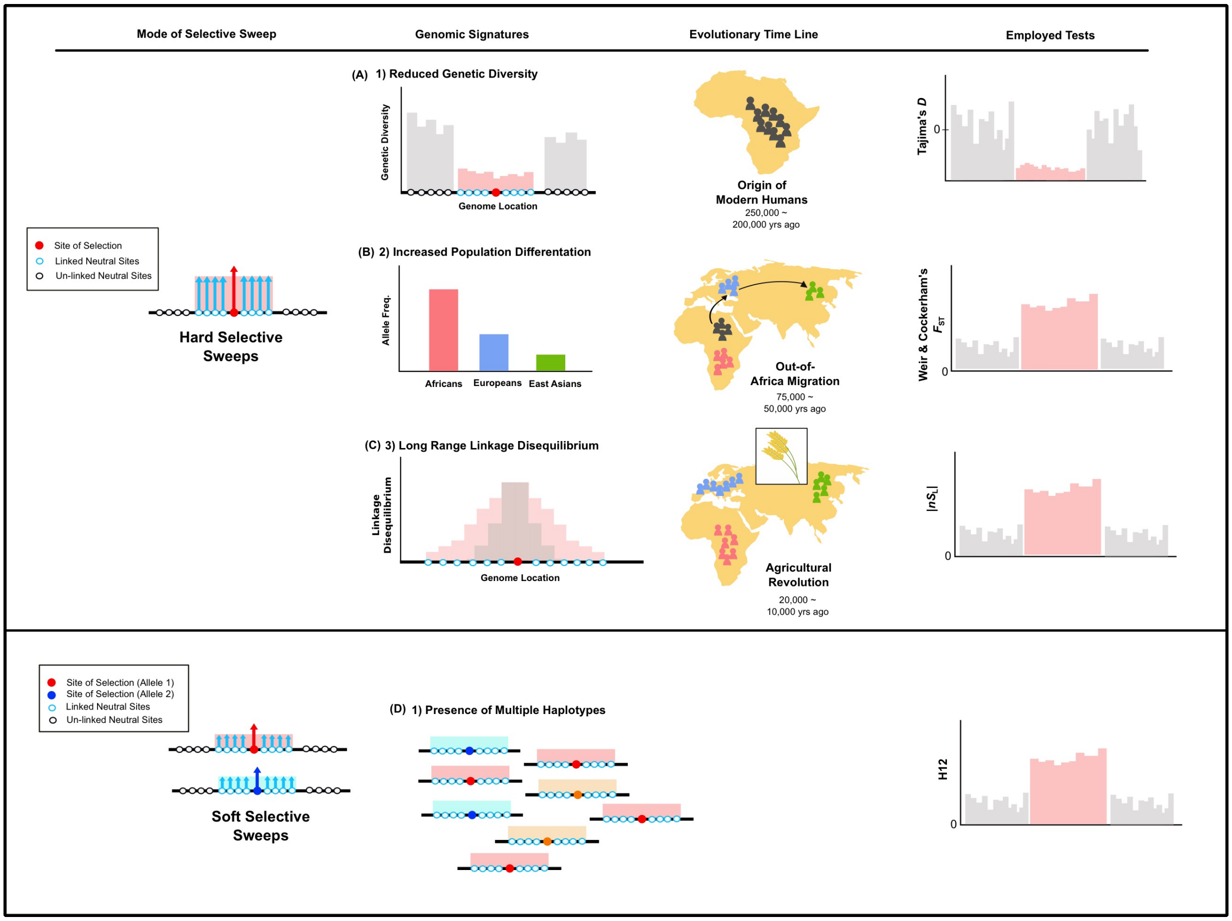

Figure 3.2. Summary of the population genetics measures used in this study. This figure depicts the nature of the genomic signatures left by varying modes of selective sweeps occurring at different evolutionary time points, the measures that we use to detect such signatures, and the expected values for regions that have experienced selective sweeps (depicted in pink) compared to regions that have not (depicted in gray). (A) In a hard selective sweep model, a single, advantageous new variant rises to high frequency ('sweeps') within the population, dramatically reducing variation in the nearby (linked) neutral variants. This results in a signature of reduced levels of genetic diversity in the genomic region surrounding the selected variant. The Tajima's *D* neutrality index assesses the extent of the abundance of rare variants resulting from such selective sweep events. As mutagenesis is rare, this signature persists for a relatively long period of time, and has been proposed to detect selective events that happened around the origin of modern humans approximately 250,000 years ago. (B) Approximately around 75,000 – 50,000 years ago, modern humans migrated out of Africa to inhabit different parts of the world, resulting in exposure to new selective pressures (e.g., pathogens). Subsequent local adaptation events would have resulted in differences in allele frequencies among distinct populations (i.e., population differentiation), which can be quantified by Weir & Cockerham's F_{ST} index. (C) Another genomic signature of hard selective sweeps is the presence of one dominant extended haplotype within the population, which stems from the hitch-hiking of linked neutral variants alongside the beneficial variant. As

recombination events eventually break down the haplotype structure, this signature persists for a relatively short period of time (i.e. 20,000 years). This signature can be quantified by measures of extended haplotype homozygosity such as $nS_{L.}$ (D) In contrast to the hard selective sweeps, soft selective sweeps occur when standing variation or multiple *de novo* mutations introduce several beneficial alleles into the population, resulting in an increase in the frequency of more than one, independent haplotypes. Measures of pooled haplotype homozygosity that combine the frequencies of the most common and second-most-common haplotypes (i.e., H12) have been shown to possess increased power to capture the signatures resulting from such soft selective sweeps.

To detect signatures left by hard selective sweeps that occurred approximately 250,000 years ago, we calculated Tajima's *D* (Figure 3.2a) (Tajima 1989) using the R package *PopGenome*, version 2.1.6 for each of the 55 genic regions (i.e., regions defined by the gene coordinates obtained via Ensembl) of the sialic acid biology genes (Pfeifer *et al*. 2014)*.* To detect local selection that occurred after the out-of-Africa migration approximately 75,000 years ago, we calculated both weighted and mean Weir & Cockerham's fixation index (F_{ST}) (Figure 3.2b) (Weir and Cockerham 1984) between all three ethnic groups for the genic regions of sialic acid biology genes using *VCFtools* (Danecek *et al*. 2011). To capture signatures of selective sweeps that occurred approximately 20,000 years ago (Sabeti *et al*. 2006; Voight *et al*. 2007), we calculated the nS_L statistic (number of segregating sites by length; Ferrer-Admetlla *et al*. 2014) (Figure 3.2c) (Sabeti *et al*. 2006) using *Selscan*, version 1.2.0 (Szpiech and Hernandez 2014). For each gene, we calculated nS_L for the region extending 100kb upstream and downstream using the default settings of *Selscan*; the only deviation from the default settings was that we used 0.01, instead of 0.05, for the minor allele frequency (MAF) cut-off value. As extended haplotype methods are thought to detect selection events that happened after the out-of-Africa migration, we calculated this measure for each of the three ethnic groups, rather than globally. We used the maximum absolute nS_L values calculated over the entire window to represent each sialic acid biology gene and report un-standardized nS_L values in this paper.

Tajima's *D*, Weir & Cockerham's F_{ST}, and the nS_L statistic are designed to capture signatures of completed or ongoing hard selective sweeps; therefore, they have limited power to detect incidences of soft selective sweeps, in which standing variation or multiple *de novo* mutations introduce several beneficial alleles into a population (Pennings and Hermisson 2006a; Messer and Petrov 2013). To detect soft sweeps, we used the H12 index (Figure 3.2d) (Garud *et al.* 2015), a modified haplotype homozygosity index that combines the frequency of the most common and second most common haplotypes in a given sample (Pennings and Hermisson 2006b). We modified the original python script written by Garud to carry out the calculations of H12 (Garud *et al.* 2015) on the genic regions of sialome genes.

Simulations of neutral evolution

The distribution of each of the four measures in the absence of selection can be influenced by demographic events and random processes, such as genetic drift. Therefore, to assess the probability that observed values in sialic acid biology genes reflect the action of selection, we compared them to those "expected" from a realistic neutral model that accounts for human demography. To identify any sialic acid biology genes that significantly deviate from selective neutrality in any of the four measures, we conducted simulations of neutral evolution using *SLiM*, version 2.4.1 (Messer 2013; Haller and Messer 2017). To account for the confounding effects of past demographic events, we incorporated previously calculated demographic parameters for the three ethnic groups in our study (Gravel *et al*. 2011), which were estimated by combining low-coverage whole genome data and high-coverage targeted exon data of the pilot phase 1000 Genomes Project. More specifically, the simulation model assumes that: a) approximately 148,000 years ago, the ancestral African population experienced a population expansion from an initial effective population size of 7,310 to 14,474, b) the out-of-Africa migration occurred approximately 51,000 years ago, c) the subsequent Eurasian split happened approximately 23,000 years ago, d) after the Eurasian split, a bottleneck event in the European population reduced its effective size to 1,032 individuals, e) for the last 23,000 years, both the European and the East Asian populations each experienced an exponential growth in effective population size (exponential quotients for Europeans: 0.0038; exponential quotients for East Asians: 0.0048), and f) both the recombination rate (1.0 x10^{-8} recombination events per bp) and the mutation rate (2.36 x 10^{-8} mutations per bp) were fixed. More detailed information on the parameters used for the simulations can be found in the Additional File 3.1.

Using this model and parameters, we simulated the genotypes of 661, 503, and 504 individuals corresponding to the number of individuals that we have genotype data from the African, European, and East Asian groups, respectively. For each gene, we carried out 5,000 simulations to create neutrally-evolved sequences of the same length. We next used these simulated sequences to calculate Tajima's *D*, Weir & Cockerham's F_{ST}, and H12 values as described above. To calculate nS_L values, we used *SLiM* to simulate 2,500 genomic regions that extend 100kb upstream and downstream a "simulated sialic acid biology gene". nS_L values were calculated on these regions using *Selscan* as described above. For all tests, we calculated the *p*-value for a given gene as the proportion of values on simulated sequences that were equal to or more extreme than the observed value. We used a *p*-value of 0.05 as cutoff for significance: a gene associated with a *p*-value lower than 0.05 likely indicates 1) deviation from neutrality, and 2) patterns of variation are due to the action of positive selection and not due to demographic events.

Statistical Analysis

To compare the patterns of genetic variation across different functional categories of sialic acid biology genes, we conducted pairwise Mann-Whitney *U* (MW*U*) tests (two-sided): the exact *p*-values were calculated for each comparison via the *coin* package, version 1.1-3, in the R programming environment (Hothorn *et al*. 2008). *P*-values calculated for comparisons among functional categories were adjusted *post hoc* to correct for the testing of multiple hypotheses via the Bonferroni method using R.

Results

Most sialic acid biology genes do not exhibit signatures of recent positive selection

To test the hypothesis that human sialic acid biology genes experienced recent positive selection resulting from hard selective sweeps approximately 250,000 years ago (i.e., around the origin of modern humans), we calculated Tajima's *D* for each of the 55 sialic acid biology genes across our sample of 1,668 individuals (Figure 3.2a & Additional Table 3.1). Comparison of each gene's Tajima's *D* value against the distribution of Tajima's *D* values estimated from 5,000 simulated sequences under neutrality showed that only 2 / 55 sialic acid biology genes exhibited significantly extreme values (*SIAE*, involved in recycling, degradation functions: *p*-value = 0.024; and *ST3GAL2*, involved in activation, transport, and transfer: *p*-value = 0.003; Figure 3.3 & Additional Table 3.1).

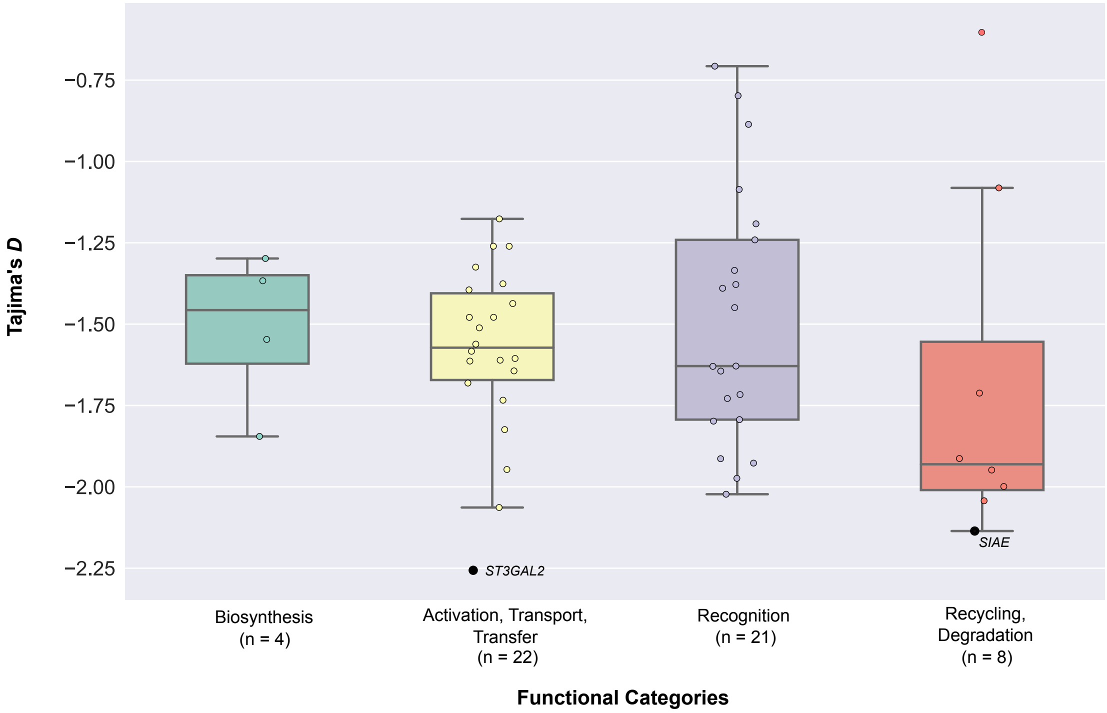

Figure 3.3. The distribution of Tajima's *D* values across the four functional categories of sialic acid biology genes. The graph presents the values of Tajima's *D*

calculated for all 1,668 individuals of the 1000 Genomes Project. The 55 sialic acid biology genes are displayed according to the functional categories they belong to. Genes exhibiting significant deviation from neutral expectations (i.e., genes with *p*-values less than 0.05) are highlighted in black and their names are shown next to the data points; the names of all other (non-significant) genes have been omitted. The values of Tajima's *D* for each of the 55 genes can be found in Additional Table 3.1. The number of genes belonging to each category is shown below the name of each functional category.

To test the hypothesis that human sialic acid biology genes experienced recent local positive selection after the major human migration out of Africa approximately 75,000 years ago, we calculated Weir & Cockerham's F_{ST} between our three human populations for each of the sialic acid biology genes (Figure 3.2b & Additional Table 3.1). We found that 53 / 55 sialic acid biology genes did not exhibit significantly higher levels of population differentiation compared to the values calculated for 5,000 neutrally simulated sequences (Figure 3.4 & Additional Table 3.1); the two exceptions were the biosynthesis-involved *NANP* (*p*-value = 0.045) and the recognition-involved *SELP* (*p*-value = 0.035).

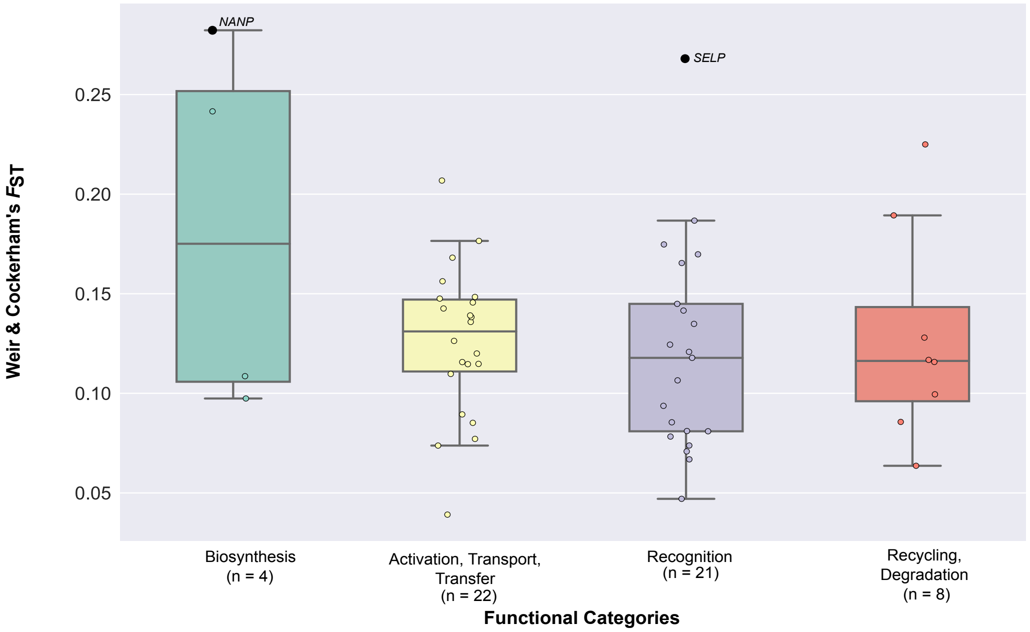

Figure 3.4. The distribution of Weir & Cockerham's F_{ST} values across the four functional categories of sialic acid biology genes. The graph presents the values of

weighted Weir & Cockerham's F_{ST} values calculated by pairwise comparisons of the three ethnic groups. The 55 sialic acid biology genes are displayed according to the functional categories they belong to. Genes exhibiting significant deviation from neutral expectations (i.e., genes with *p*-values less than 0.05) are highlighted in black, their names shown next to the data points; the names of all other (non-significant) genes have been omitted. The values of the F_{ST} for each gene, along with the values of the mean Weir & Cockerham's F_{ST}, can be found in Additional Table 3.1. The number of genes belonging to each category is shown below the name of each functional category.

To test the hypothesis that human sialic acid biology genes experienced selective sweeps in one or more human populations approximately 20,000 years ago (i.e., around or shortly prior to the Agricultural Revolution), we calculated nS_L values for the regions flanking 100kb upstream and downstream of sialic acid biology genes within each ethic group (Figure 3.2c & Additional Table 3.2). For all three ethnic groups, we found that most sialic acid biology genes did not deviate from neutral expectations: only five genes (four involved in recognition and one in recycling, degradation) in Europeans (*SIGLEC5*: *p*-value = 0.0004; *SIGLEC6*: *p*-value = 0.0004; *SIGLEC12*: *p*-value = 0.0004; *SIGLEC14*: *p*-value = 0.0004; and *NEU2*: *p*-value = 0.036; Figure 3.5b & Additional Table 3.2), five genes (three involved in recognition and two in activation, transport, and transfer) in Africans (*LAMA2*: *p*-value = 0.018; *ST6GALNAC1*: *p*-value = 0.002; *ST6GALNAC2*: *p*-value = 0.001; *SIGLEC8*: *p*-value = 0.002; and *SIGLEC10*: *p*-value = 0.002; Figure 3.5a & Additional Table 3.2), and four genes (two involved in recognition and two in activation, transport, and transfer) in East Asians (*CD22*: *p*-value = 0.046; *MAG*: *p*-value = 0.046; *ST6GAL1*: *p*-value = 0.001; and *ST6GALNAC5*: *p*-value = 0.030; Figure 3.5c & Additional Table 3.2) exhibited significant *p*-values. Notably, the genes exhibiting significant deviations from neutral expectations were different for each ethnic group.

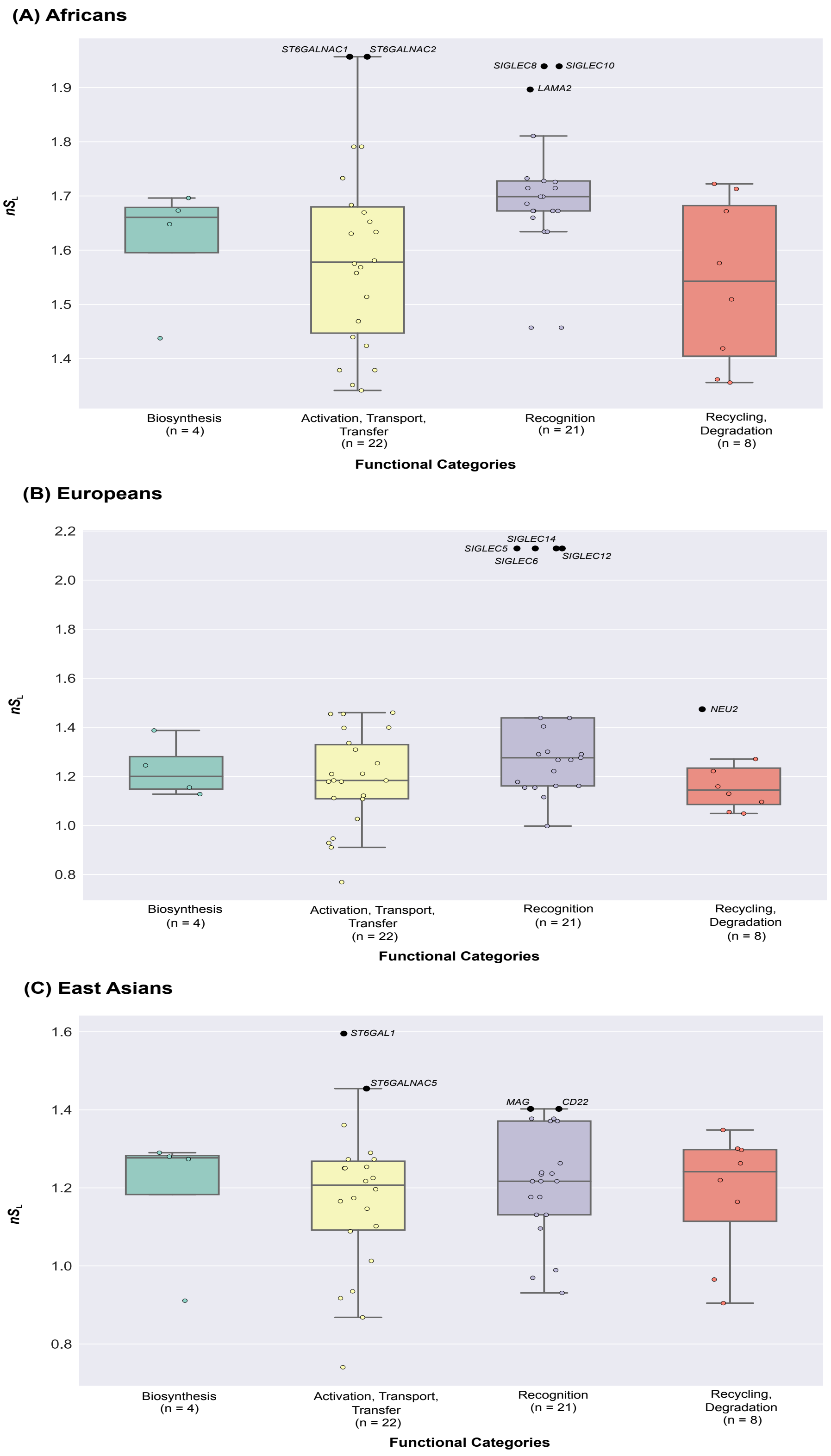

Figure 3.5. The distribution of nS_L values across the four functional categories of sialic acid biology genes. The graph presents the values of nS_L values calculated for

each ethnic group (Africans: (A), Europeans: (B), East Asians: (C)). Unstandardized nS_L values were calculated in windows that expand 100kb up and downstream of a given sialic acid biology gene: the maximum value of all the absolute nS_L values in a given window was used to represent each gene. The 55 sialic acid biology genes are displayed according to the functional categories they belong to. Genes exhibiting significant deviation from neutral expectations (i.e., genes with *p*-values less than 0.05) are highlighted in black, their names shown next to the data points; the names of all other (non-significant) genes have been omitted. The values of the nS_L for each gene can be found in Additional Table 3.2. The number of genes belonging to each category is shown below the name of each functional category.

Finally, to test the hypothesis that the sialic acid biology genes experienced soft selective sweeps, we calculated the modified pooled haplotype homozygosity index H12 for each of the 55 sialic acid biology genes (Figure 3.2d & Additional Table 3.1). 10 / 55 sialic acid biology genes showed H12 values that significantly deviated from neutral expectations (*NANS*: *p*-value = 0.0004; *NEU1*: *p*-value = 0.027; *NEU3*: *p*-value = 0.001; *NPL*: *p*-value = 0.0004; *SIAE*: *p*-value = 0.001; *SIGLEC6*: *p*-value = 0.023; *SLC35A1*: *p*-value = 0.001; *ST3GAL5*: *p*-value = 0.002; *ST6GALNAC1*: *p*-value = 0.032; *ST8SIA4*: *p*-value = 0.015; Figure 3.6 & Additional Table 3.1). Of these 10 genes, four function in sialic acid activation, transport, and transfer (*SLC35A1*, *ST3GAL5*, *ST6GALNAC1*, and *ST8SIA4*), four in recycling and degradation (*NEU1*, *NEU3*, *NPL*, and *SIAE*), one in recognition (*SIGLEC6*) and one in biosynthesis (*NANS*).

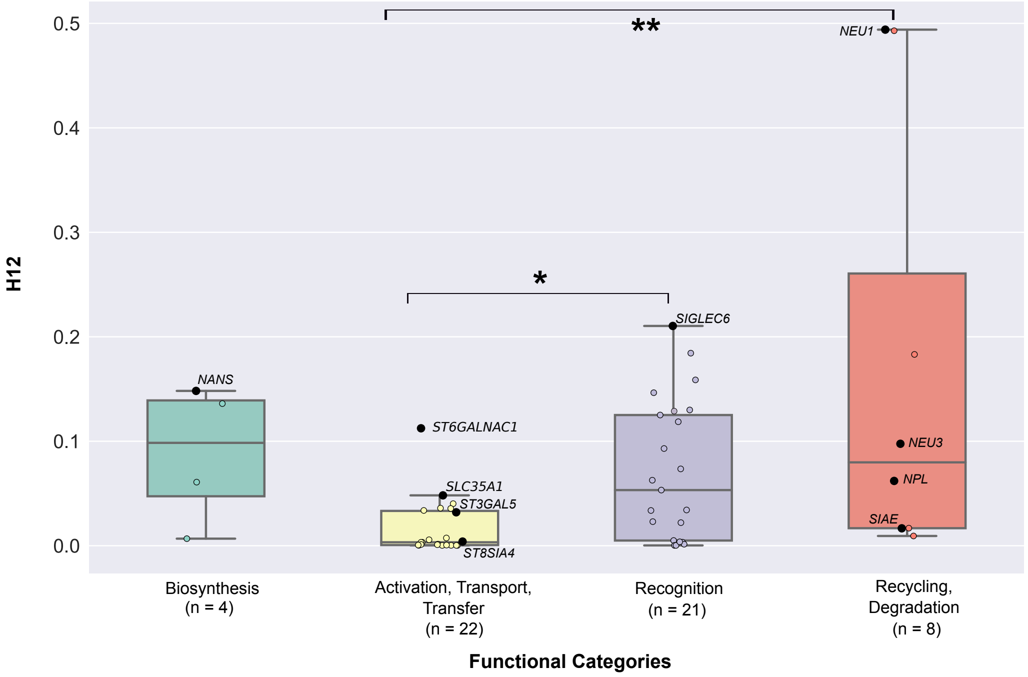

Figure 3.6. The distribution of H12 values across the four functional categories of sialic acid biology genes. The graph presents the values of H12 values calculated across all three ethnics groups. The 55 sialic acid biology genes are displayed according to the functional categories they belong to. Significant pairwise functional comparisons are indicated with asterisks (*: *p*-value < 0.05; **: *p*-value < 0.01). Genes exhibiting significant deviation from neutral expectations (i.e., genes with *p*-values less than 0.05) are highlighted in black, their names shown next to the data points; the names of all other (non-significant) genes have been omitted. The values of the H12 for each gene can be found in Additional Table 3.1. The number of genes belonging to each category is shown below the name of each functional category.

The four functional categories of sialic acid biology genes do not significantly differ in their signatures of recent positive selection

We next examined whether the four distinct functional categories of sialic acid biology genes (Figure 3.1) exhibit statistically significant differences in their extent of polymorphism, allele frequency spectra, extent of population differentiation, and haplotype homozygosity by comparing the patterns of nucleotide diversity (π), Tajima's

D, F_{ST}, nS_L, and H12 values across the four functional categories. 4 / 5 measures (π: Figure 3.7 & Additional Table 3.3a; Tajima's *D*: Figure 3.3 & Additional Table 3.3b; F_{ST}: Figure 3.4 & Additional Table 3.3c; nS_L: Figure 3.5a-c & Additional Table 3.3d-f) did not show significant differences among the four functional categories. In contrast, comparison of H12 values across the four functional categories showed significant differences between activation, transport, transfer and recognition (*U*-value = 114, adjusted *p*-value = 0.023; Figure 3.6 & Additional Table 3.3g) and between activation, transport, transfer and recycling, degradation (*U*-value = 23, adjusted *p*-value = 0.008; Figure 3.6 & Additional Table 3.3g).

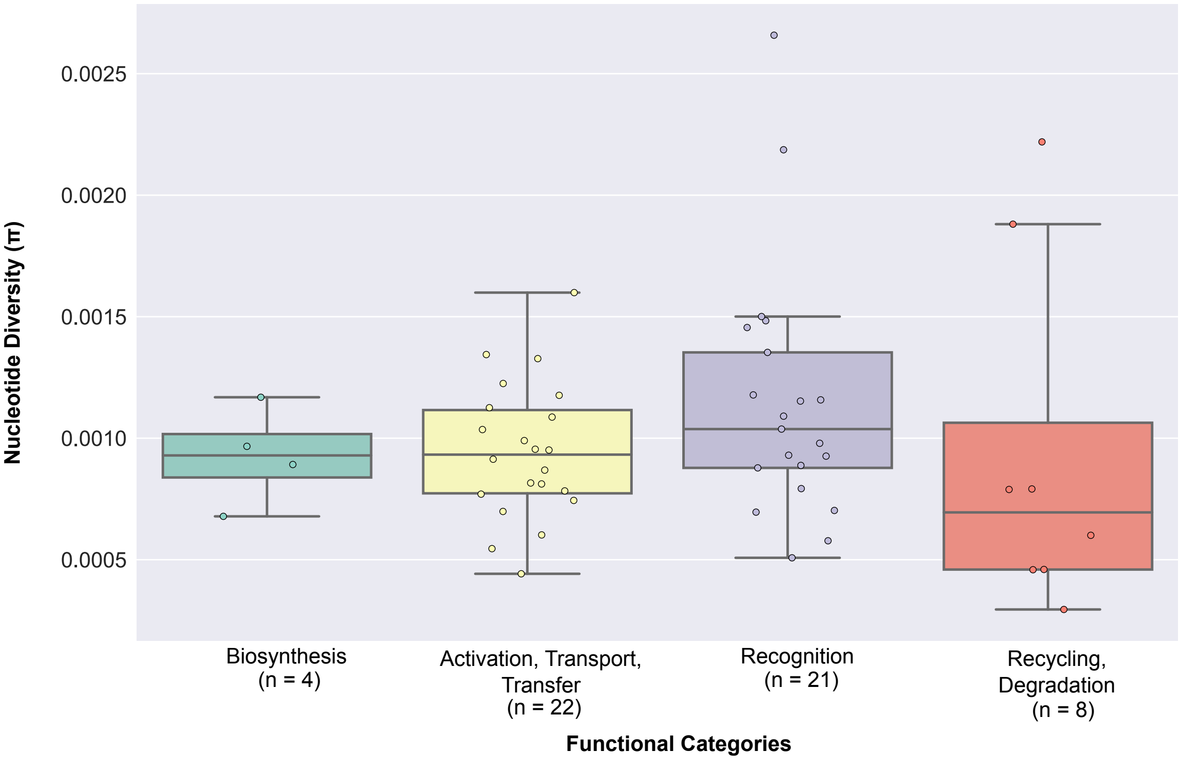

Figure 3.7. The distribution of nucleotide diversity values across the four functional categories of sialic acid biology genes. The graph presents the nucleotide diversity values calculated across all three ethnic groups. The 55 sialic acid biology genes are

displayed according to the functional categories they belong to. The nucleotide diversity value for each gene can be found in Additional Table 3.1. The number of genes belonging to each category is shown below the name of each functional category.

Discussion

In this study, we tested the hypothesis that loci involved in human sialic acid biology experienced recent positive selection by employing four population genetic tests designed to detect selective signatures at different time points during the evolution of modern humans. In all of our analyses, the majority of sialic acid biology genes did not significantly deviate from neutral expectations.

One likely explanation of our results is that, for the most part, genes involved in sialic acid biology have not been targets of positive selection in the last 250,000 years of human evolution. Previous studies have shown that sialic acid biology genes were subject to strong ancient selective pressures that have resulted in species-specific adaptations (Varki 2001; Angata *et al*. 2004; Altheide *et al*. 2006; Varki and Gagneux 2009), suggesting that they likely evolved in response to ancient pathogens. The function of sialic acid biology genes, i.e., the generation of host-specific sialic acids and recognition of those self-associated molecular patterns (SAMPs) by cognate receptors, contributes to the detection of missing self (Medzhitov and Janeway 2002). It is possible that evolution of such markers of normal self, and the ability to recognize them correctly, have been influenced by ancient, but not, recent pathogens. For instance, the human-specific *Alu*-mediated inactivation of *CMAH,* the protein product of which is responsible for generating Neu5Gc from its precursor Neu5Ac, has been suggested as a means of escaping infection by *Plasmodium reichenowi*, a malaria-causing parasite that infects other great apes (Martin *et al*. 2005; Varki and Gagneux 2009). In addition, a human-specific mutation of an essential arginine residue required for sialic acid binding in Siglec-12 has been suggested to be evolutionarily related to the loss of Neu5Gc in humans (Varki 2009). Several other human-specific changes in sialic acid biology (e.g., deletion of *SIGLEC13* in humans or increased expression of *SIGLEC1* on human macrophages compared to other primates) could have resulted from selective pressures exerted by paleopathogens (Varki 2009; Wang *et al*. 2012). It should be noted that several other innate-immunity genes have been shown to lack any signatures of positive selection in more recent evolutionary time scales (Mukherjee *et al*. 2009; Siddle and Quintana-Murci 2014).

An alternative, but not mutually exclusive, possibility is that sialic acid biology genes may have experienced recent positive selection but – as the strength of selection can fluctuate over time and space – any signatures left from such selective episodes were not strong enough to alter the allele frequency spectra and extent of haplotype homozygosity of sialic acid biology genes within a single population or across multiple populations to detectable levels (Bell 2010; Pritchard *et al*. 2010). For example, even our most sensitive measure (the H12 index, which can discern past soft selective sweeps) is still biased towards detecting hard selective sweeps and loses power when sweeps are too soft (i.e., when selection has acted on numerous variants, resulting in a high number of haplotypes being present in the population) (Garud *et al*. 2015). Therefore, it is possible that signatures of selection left by very soft selective sweeps are not being detected by the methods employed in this study. In addition, host defense is a complex biological process that involves interactions between numerous genes belonging to diverse biological pathways. As such, it is feasible that selection has operated not on individual sialic acid biology genes, but on the set of pathways that comprise sialic acid biology, resulting in only subtle changes in allele frequency spectra and haplotype homozygosity patterns of individual genes that could not be detected by the gene-specific tests employed in our study. The occurrence of such polygenic adaptation (Daub *et al*. 2013; Berg and Coup 2014; Daub *et al*. 2015) could similarly result in non-significant deviations from neutral expectations at the level of individual loci.

Most of the sialic acid biology genes that did show evidence of recent positive selection were recovered by the haplotype-based H12 index (10 genes), with many fewer recovered by either the haplotype-based nS_L (4-5 per population) or the two frequency spectra-based measures (Tajima's D and F_{ST}; 2 each) (Figure 3.8). The key difference among these measures is that H12 has increased power to detect soft selective sweeps whereas the other three are designed to detect hard selective sweeps, suggesting that sialic acid biology genes likely experienced more soft than hard sweeps. This is consistent with a previous study suggesting that hard selective sweeps were likely rare during modern human evolution (Hernandez *et al*. 2011). Interestingly, genes that underwent hard selective sweeps were often involved in recognition of sialic acids, whereas genes that underwent soft sweeps were more often in the activation, transport, transfer and recycling, degradation categories (Figures 3.2 and 3.8).

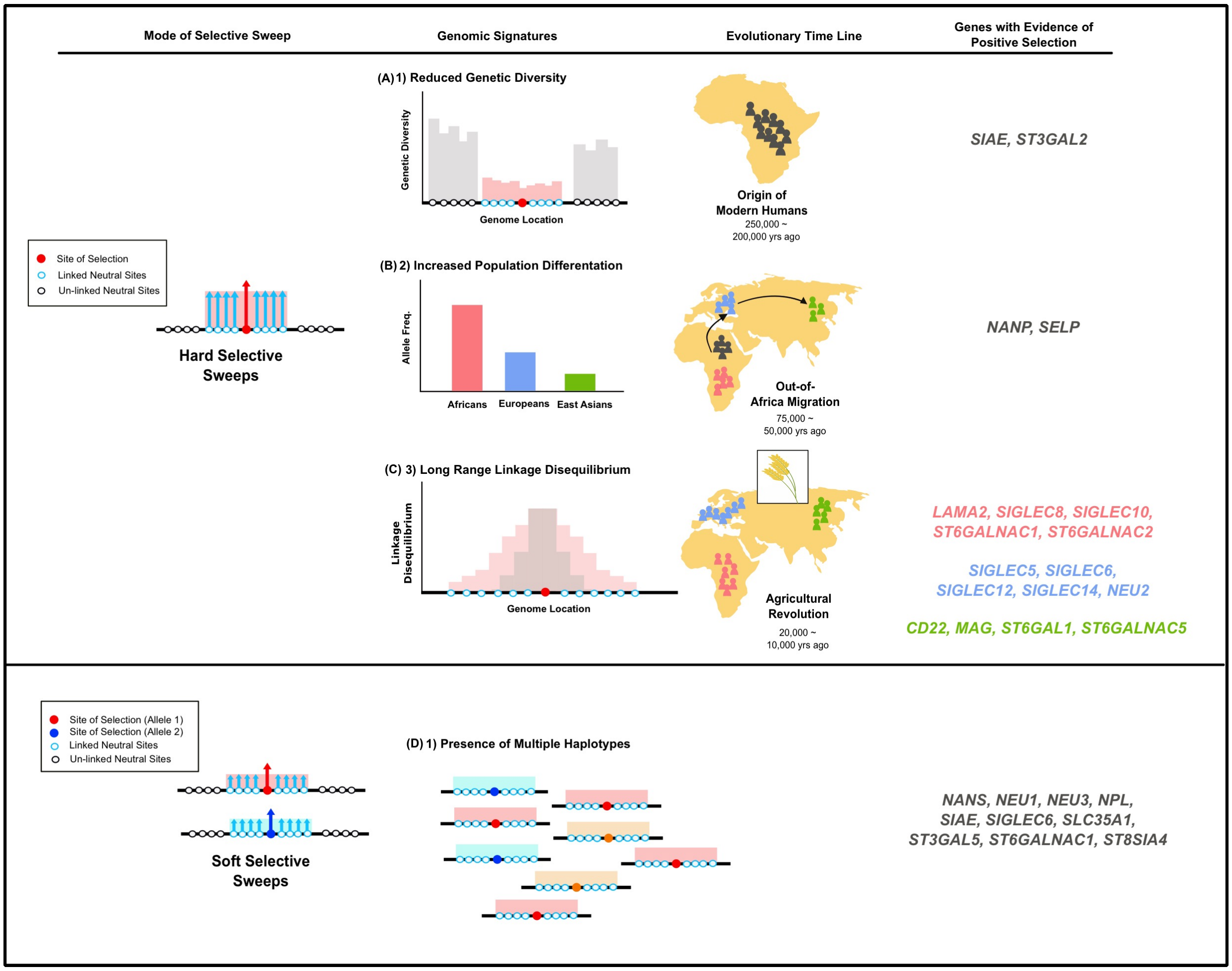
Mode of Selective Sweep
Genomic Signatures
Evolutionary Time Line
Genes with Evidence of Positive Selection
(A) 1) Reduced Genetic Diversity
Genetic Diversity
Genome Location
Origin of Modern Humans
250,000 ~ 200,000 yrs ago
SIAE, ST3GAL2
● Site of Selection
○ Linked Neutral Sites
○ Un-linked Neutral Sites
Hard Selective Sweeps
(B) 2) Increased Population Differentation
Allele Freq.
Africans
Europeans
East Asians
Out-of-Africa Migration
75,000 ~ 50,000 yrs ago
NANP, SELP
(C) 3) Long Range Linkage Disequilibrium
Linkage Disequilibrium
Genome Location
Agricultural Revolution
20,000 ~ 10,000 yrs ago
LAMA2, SIGLEC8, SIGLEC10, ST6GALNAC1, ST6GALNAC2
SIGLEC5, SIGLEC6, SIGLEC12, SIGLEC14, NEU2
CD22, MAG, ST6GAL1, ST6GALNAC5
● Site of Selection (Allele 1)
● Site of Selection (Allele 2)
○ Linked Neutral Sites
○ Un-linked Neutral Sites
Soft Selective Sweeps
(D) 1) Presence of Multiple Haplotypes
NANS, NEU1, NEU3, NPL, SIAE, SIGLEC6, SLC35A1, ST3GAL5, ST6GALNAC1, ST8SIA4

Figure 3.8. Few sialic acid biology genes have experienced both temporally and spatially varying modes of selective sweeps in recent human history

We also found that the identity of the genes exhibiting significant deviations from neutral expectations was different among the various tests of selection employed in this study, and, in the case of nS_L, among the three ethnic groups studied. The only overlap was for *SIGLEC6*, which was shown to exhibit significant values of both nS_L (in Europeans) and H12. The lack of overlap likely reflects the differences in sensitivities of the four measures (see preceding paragraph), and hence the occurrence of spatially and/or temporally varying selective events. For instance, it has been demonstrated that tests that quantify the extent of excess or depletion of rare alleles (e.g., Tajima's *D*) have the most power to detect hard selective sweeps that happened at earlier time points, since the low rate of mutation in humans does not erase such signatures (Sabeti *et al*. 2006). Similarly, nS_L has been shown to exhibit high power to detect signatures of selective sweeps that happened more recently and have resulted in only incomplete or partial fixation of alleles (Sabeti *et al*. 2006; Voight *et al*. 2007; Ferrer-Admetalla *et al*. 2014).

In summary, considered jointly with previous evidence on the widespread occurrence of ancient positive selection on genes involved in sialic acid biology in hominids and hominins (Altheide *et al*. 2006; Angata *et al*. 2006; Varki and Gagneux 2009; Hayakawa and Varki 2011), these results support the hypothesis that human sialic acid biology genes bear many and strong evolutionary signatures of adaptation in response to ancient primate pathogens, but rather few and weak adaptive signatures in response to contemporary cosmopolitan or geographically-restricted human pathogens.

Acknowledgments

This work was conducted in part using the resources of the Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University. J.M.M. was partially supported by the Graduate Program in Biological Sciences at Vanderbilt University. This work was supported in part by the March of Dimes through the March of Dimes Prematurity Research Center Ohio Collaborative, a Burroughs Wellcome Fund Preterm Birth Initiative Award (to J.A.C. and A.R.), and by the National Science Foundation (DEB-1442113 to A.R.).

References

Altheide, T. K., T. Hayakawa, T. S. Mikkelsen, S. Diaz, N. Varki *et al.*, 2006 System-wide Genomic and Biochemical Comparisons of Sialic Acid Biology Among Primates and Rodents: EVIDENCE FOR TWO MODES OF RAPID EVOLUTION. J Biol Chem. 281: 25689–25702.

Angata, T., T. Hayakawa, M. Yamanaka, A. Varki, and M. Nakamura, 2006 Discovery of Siglec-14, a novel sialic acid receptor undergoing concerted evolution with Siglec-5 in primates. FASEB J. 20: 1964–1973.

Angata, T., E. H. Margulies, E. D. Green, and A. Varki, 2004 Large-scale sequencing of the CD33-related Siglec gene cluster in five mammalian species reveals rapid evolution by multiple mechanisms. Proc. Natl. Acad. Sci. USA. 101: 13251–13256.

Angata, T., and A. Varki, 2002 Chemical Diversity in the Sialic Acids and Related α-Keto Acids: An Evolutionary Perspective. Chem. Rev. 102: 439–470.

Bell, G., 2010 Fluctuating selection: the perpetual renewal of adaptation in variable environments. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 365: 87-97.

Auton, A., G. R. Abecasis, D. M. Altshuler, R. M. Durbin, D. R. Bentley *et al.*, 2015 A global reference for human genetic variation. Nature. 526: 68–74.

Berg J. J., and G. Coop, 2014 A Population Genetic Signal of Polygenic Adaptation. PLoS Genet.10(8): e1004412.

Chou, H. H., H. Takematsu, S. Diaz, J. Iber, E. Nickerson *et al.*, 1998 A mutation in human CMP-sialic acid hydroxylase occurred after the Homo-Pan divergence. Proc. Natl. Acad. Sci. USA. 95: 11751–11756.

Cohen, M., and A. Varki, 2010 The Sialome—Far More Than the Sum of Its Parts. OMICS. 14: 455–464.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks *et al.*, 2011 The variant call format and VCFtools. Bioinformatics. 27: 2156–2158.

Daub, J. T., T. Hofer, E. Cutivet, I. Dupanloup, L. Quintana-Murci *et al.*, 2013 Evidence for polygenic adaptation to pathogens in the human genome. Mol. Biol. Evol. 30: 1544-1558.

Daub, J. T., I. Dupanloup, M. Robinson-Rechavi, and L. Excoffier, 2015 Inference of Evolutionary Forces Acting on Human Biological Pathways. Genome Biol. Evol. 7: 1546-1558.

Ferrer-Admetalla, A., M. Liang, T. Korneliussen, and R. Nielsen, 2014 On Detecting Incomplete Soft or Hard Selective Sweeps Using Haplotype Structure. Mol. Biol. Evol. 31: 1275-1291.

Garud, N. R., P. W. Messer, E. O. Buzbas, and D. A. Petrov, 2015 Recent Selective Sweeps in North American *Drosophila melanogaster* Show Signatures of Soft Selective. PLoS Genet. 11(2): e1005004.

Gravel, S., B. M. Henn, R. N. Gutenkunst, A. R. Indap, G. T. Marth *et al*., 2011 Demographic history and rare allele sharing among human populations. Proc. Natl. Acad. Sci. USA. 108: 11983-11988

Haller, B. C., and P. W. Messer, 2017 SLiM 2: Flexible, interactive forward genetic simulations. Mol. Biol. Evol. 34: 230–240

Hayakawa, T., and A. Varki, 2011 Human-Specific Changes in Sialic Acid Biology, pp.123-148 in *Post-Genome Biology of Primates*, edited by H. Hirai, H. Imai, and Y. Go, Springer.

Hernandez, R. D., J. L. Kelley, E. Elyashiv, S. C. Melton, A. Auton *et al*., 2011 Classic Selective Sweeps Were Rare in Recent Human Evolution. Science. 331: 920-924.

Hothorn, T., K. Hornik, M. A. van de Wiel, and A. Zeileis, 2008 Implementing a Class of Permutation Tests: The coin Package. Journal of Statistical Software. 28: 1-23.

Irie, A., S. Koyama, Y. Kozutsumi, T. Kawasaki, and A. Suzuki, 1998 The Molecular Basis for the Absence of N-Glycolylneuraminic Acid in Humans. J Biol Chem. 273: 15866–15871.

Kelm, S., and R. Schauer, 1997 Sialic Acids in Molecular and Cellular Interactions. Int Rev of Cytol. 175: 137–240.

Li, H., 2011 Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics. 27: 718–719.

Martin, M. J., J. C. Rayner, P. Gagneux, J. W. Barnwell, and A. Varki, 2005 Evolution of human-chimpanzee differences in malaria susceptibility: Relationship to human genetic loss of *N*-glycolylneuraminic acid. Proc. Natl. Acad. Sci. USA. 102: 12819–12824.

Medzhitov, R., and C. A. Janeway, 2002 Decoding the Patterns of Self and Nonself by the Innate Immune System. Science. 296: 298–300.

Messer, P. W., 2013 SLiM: Simulating Evolution with Selection and Linkage. Genetics. 194: 1037-1039.

Messer, P. W., and D. A. Petrov, 2013 Population genomics of rapid adaptation by soft selective sweeps. Trends in Ecology & Evolution. 28: 1–11.

Mukherjee, S., N. Sarkar-Roy, D. K. Wagener, and P. P. Majumder, 2009 Signatures of natural selection are not uniform across genes of innate immune system, but purifying selection is the dominant signature. Proc. Natl. Acad. Sci. USA. 106: 7073–7078.

Nei, M., and W. H. Li, 1979 Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA. 76: 5269–5273.

Pennings, P. S., and J. Hermisson, 2006a. Soft Sweeps II--Molecular Population Genetics of Adaptation from Recurrent Mutation or Migration. Mol Biol Evol. 23: 1076–1084.

Pennings, P. S., and J. Hermisson, 2006b. Soft Sweeps III: The Signature of Positive Selection from Recurrent Mutation. PLoS Genet. 2:e186.

Pfeifer, B., U. Wittelsbürger, S. E. Ramos-Onsins, and M. J. Lercher, 2014 PopGenome: An Efficient Swiss Army Knife for Population Genomic Analyses in R. Mol Biol Evol. 31: 1929–1936.

Pilatte, Y., J. Bignon, and C. R. Lambré, 1993 Sialic acids as important molecules in the regulation of the immune system: pathophysiological implications of sialidases in immunity. Glycobiology. 3: 201–218.

Pritchard, J. K., J. K. Pickrell, and G. Coop, 2010 The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation. Current Biology. 20: R208–R215.

Sabeti, P. C., S. F. Schaffner, B. Fry, J. Lohmueller, P. Varilly *et al.*, 2006 Positive Natural Selection in the Human Lineage. Science. 312: 1614–1620.

Schaffner S. F., 2004 The X chromosome in population genetics. Nature Rev Genet. 5: 43–51.

Schauer, R., 1982 Chemistry, Metabolism, and Biological Functions of Sialic Acids. Adv. Carbohydr. Chem. Biochem. 40: 31–234.

Siddle, K. J., and L. Quintana-Murci, 2014 The Red Queen's long race: human adaptation to pathogen pressure. Curr. Opin. Genet. Dev. 29: 31–38.

Szpiech, Z. A., and R. D. Hernandez, 2014 selscan: An Efficient Multithreaded Program to Perform EHH-Based Scans for Positive Selection. Mol. Biol. Evol. 31: 2824-2827

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 123: 585-595.

Varki, A., 2007 Glycan-based interactions involving vertebrate sialic-acid-recognizing proteins. Nature. 446: 1023–1029.

Varki, A., 2001 Loss of N-glycolylneuraminic acid in humans: Mechanisms, consequences, and implications for hominid evolution. Am. J. Phys. Anthropol. 116: 54–69.

Varki, A., 2009 Multiple changes in sialic acid biology during human evolution. Glycoconj J. 26: 231–245.

Varki, A., and P. Gagneux, 2009 Human-specific evolution of sialic acid targets: Explaining the malignant malaria mystery? Proc. Natl. Acad. Sci. USA. 106: 14739–14740.

Varki, A., and P. Gagneux, 2012 Multifarious roles of sialic acids in immunity. Ann. N.Y. Aca. Sci. 1253: 16–36.

Varki, N. M., and A. Varki, 2007 Diversity in cell surface sialic acid presentations: implications for biology and disease. Lab Invest. 87: 851–857.

Voight, B.F., S. Kudaravalli, X. Wen, and J. K. Pritchard, 2007 A Map of Recent Positive Selection in the Human Genome. PLoS. Biol. 4(3): e72.

Wang, X., N. Mitra, I. Secundino, K. Banda, P. Cruz *et al*., 2012 Specific inactivation of two immunomodulatory *SIGLEC* genes during human evolution. Proc. Natl. Acad. Sci. USA. 109: 9935–9940.

Weir, B. S., and C. C. Cockerham, 1984 Estimating F-Statistics for the Analysis of Population Structure. Evol. 38: 1358–1370.

CHAPTER IV

Signatures of recent positive selection in enhancers across 41 human tissues[4,5]

Abstract

Evolutionary changes in enhancers are widely associated with variation in human traits and diseases. However, studies comprehensively quantifying levels of selection on enhancers at multiple evolutionary time points during recent human evolution and how enhancer evolution varies across human tissues are lacking. To address these questions, we integrated a dataset of 41,561 transcribed enhancers active in 41 different human tissues (FANTOM Consortium) with whole genome sequences of 1,668 individuals from the African, Asian, and European populations (1000 Genomes Project). Our analyses based on four different metrics (Tajima's D, F_{ST}, H12, nS_{L}) showed that ~5.90% of enhancers considered showed evidence of recent positive selection and that genes associated with enhancers under positive selection are enriched for diverse immune-related functions. The distributions of these metrics for brain and testis enhancers were often statistically significantly different compared to those of other tissues; the same was true for brain and testis enhancers that are tissue-specific compared to those that are tissue-broad and for testis enhancers associated with tissue-enriched and non-tissue-enriched genes. These differences varied considerably across metrics and tissues and

[4] This work is uploaded in https://www.biorxiv.org/content/10.1101/534461v1.

[5] The contributing authors are: Jiyun M. Moon, John A. Capra, Patrick Abbot, and Antonis Rokas.

were generally due to changes in distributions' shapes rather than shifts in their values. These results suggest that many human enhancers experienced recent positive selection throughout multiple time periods in human evolutionary history, that this selection occurred in a tissue-dependent and immune-related functional context, and that much like the evolution of their coding counterparts, the evolution of brain and testis enhancers has been markedly different from that of enhancers in other tissues.

Introduction

Enhancers are *cis*-acting DNA segments that, either independently of or in concert with other regulatory elements, control spatial, temporal and quantitative aspects of gene expression (Ong and Corces 2011; Rubinstein and de Souza 2013; Long et al. 2016). While the precise architecture of enhancers is still debated (Long et al. 2016), a typical enhancer contains multiple transcription factor binding sites (TFBS) arranged in specific order and distance from one another. Enhancers facilitate initiation of gene transcription by helping to recruit RNA Polymerase II, general transcription factors, and additional components of the transcriptional machinery to the gene's promoter (Pennacchio et al. 2013; Rubinstein and de Souza 2013). Multiple enhancers, each with its own repertoire of TFBS, can regulate the activities of a gene in a tissue-specific manner or across distinct developmental stages, enabling enhancers to alter its expression patterns in a particular context without affecting expression of other genes (Wray 2007; Sholtis and Noonan 2010). Enhancers in the human genome are more numerous than protein-coding genes (Pennacchio et al. 2013), facilitating the induction of diverse gene expression programs in different spatial and temporal contexts (Long et al. 2016).

Changes in gene regulation have long been thought to play a major role in the adaptive evolution of human traits (King and Wilson 1975; Carroll 2005). One reason for regulatory regions in general, and enhancers in particular, as preferential targets of selection is that, compared to protein-coding regions, they tend to be modularly organized (Sholtis and Noonan 2010). This modular organization means that mutations in enhancers are less likely to have pleiotropic effects and more likely to contribute to phenotypic evolution (Carroll 2005; Wray 2007; Rebeiz and Tsiantis 2017). In the context of human evolution, several studies suggest that evolutionary changes in enhancers might have played a major role in the acquisition of human-specific traits. For example, a considerable proportion of regions that has experienced accelerated evolution in the human lineage are developmental enhancers active in the brain (Capra et al. 2013), and *cis*-regulatory regions of genes with neurological and nutritional roles have been shown to exhibit evidence of accelerated evolution in the human lineage (Rockman et al. 2005; Haygood et al. 2007). Accelerated evolution is one signature of positive selection, and several studies suggest that recent selection might also have preferentially acted on human *cis*-regulatory regions (Wray 2007). For instance, Rockman et al. found that the promoter region of a gene that encodes the precursor of an opioid neuropeptide (*PDYN*) exhibits a significant degree of population differentiation between human populations, especially between Europeans and East Asians, which is suggestive of local adaptation (Rockman et al. 2005). Similarly, patterns of variation in the promoter region of the *LCT* gene that confers lactase persistence in Africans are consistent with the action of selective sweeps (Tishkoff et al. 2006). Finally, SNPs with evidence of recent positive selection are more likely to be associated with expression of nearby genes than random SNPs, and this enrichment is strongest for Yorubans (Kudaravalli et al. 2008).

More broadly, genome-wide studies have found that, compared to protein-coding regions, enhancers are enriched for variants that are statistically associated with various human diseases (e.g., inflammatory diseases, metabolic diseases) (Andersson et al. 2014; Karnuta and Scacheri 2018). Moreover, recent studies have argued that human population-level differences in transcriptional responses to infection likely resulted from local adaptation, further supporting the idea that selection on enhancers has contributed to recent human evolution (Nédélec et al. 2016). In addition, multiple events, such as migrations and shifts in cultural practices, have occurred in different time periods during recent human evolution (Karlsson et al. 2014), likely introducing novel selective agents. Given the role of enhancers in contributing to phenotypic differences, it is likely that enhancers experienced selection events in response to such selective pressures.

Each human tissue serves a particular physiological function so it is reasonable to hypothesize that the genetic elements active in each tissue are influenced by distinct selective pressures that shape their evolutionary rates (Wray 2007; Gu and Su 2007). For example, both gene expression and protein-coding sequence divergence patterns might be expected to be more conserved in developmentally constrained tissues (e.g., nervous tissues) than in developmentally relaxed tissues (e.g., testis or endocrine tissues, such as pancreas). Alternatively, certain functions of tissues (e.g., reproductive processes for testis) might have experienced increased levels of positive selection (Khaitovich 2005). Early studies examining patterns of divergence in gene expression and protein-coding sequence evolution among mammals found that both were lowest for nervous tissues (e.g., brain, cerebellum) and greatest for testis (Wray 2007; Khaitovich 2005; Brawand et al. 2011). More recent examination of a larger number of human tissues has provided further support for this hypothesis. For example, the correlation between expression levels of genes and the strength of purifying selection (as assessed by *d*N/*d*S) is strongest for the brain, while this correlation is lowest for liver, placenta and testis (Kryuchkova-Mostacci and Robinson-Rechavi 2015).

Hitherto, studies investigating the variation of evolutionary patterns among tissues have mainly focused on inter-species divergence of gene expression and of protein-coding sequences. However, studies comprehensively quantifying levels of selection on enhancers at multiple evolutionary time periods during recent human evolution and how enhancer evolution varies across human tissues are lacking. Furthermore, it is plausible, if not likely, that distinct selective pressures are also acting on regulatory elements in different tissues, further contributing to the global differences in gene expression patterns among the tissues described above (Ong and Corces 2011; Rubinstein and de Souza 2013).

To examine the influence of selection on enhancers at multiple evolutionary time periods during recent human evolution and how enhancer evolution varies across human tissues, we used a dataset of 41,561 enhancers active in 41 human tissues and genotype data of 1,668 individuals from three different super-populations from the 1000 Genomes Project to calculate signatures of recent selection that happened at different time points and via different modes (i.e., hard vs soft selective sweeps). We found that on average, 5.90% of enhancers exhibit evidence of recent positive selection and that their putative target genes are enriched for immunity-related functions. Furthermore, we found that enhancers expressed in the testis and brain exhibit statistically different patterns of recent evolution compared to numerous other tissues. Examination of patterns of recent evolution between enhancers that are active in only one tissue and those active in two or more tissues revealed statistically significant differences for several tissues, including brain and testis; we found similar results when we studied enhancers associated with tissue-enriched and non-tissue-enriched genes in these two tissues. Our results suggest that human enhancers, in particular ones associated with immune-related genes, have experienced different modes of recent positive selection at different periods of recent human evolution. Furthermore, patterns of selection on human enhancers differ between tissues, including for brain and testis, a pattern reminiscent of their protein-coding counterparts.

Methods

Genetic variation data

To examine signatures of natural selection on human enhancers, we used the whole genome sequence data for 1,668 individuals from Phase 3 of the 1000 Genomes Project (The 1000 Genomes Project Consortium et al. 2015) (for information on data sources, see Additional File 4.1). The 1,668 individuals represent three major human populations (661 Africans, 503 Europeans, and 504 East Asians). We used these three populations because their demographic histories have been most confidently estimated and excluded the other two (Admixed Americans and South Asians) because their demographic histories are more complex.

FANTOM enhancer data

To study patterns of recent evolution on human enhancers active in different tissues, we used the enhancers detected in 41 human tissues compiled by the Functional Annotation of the Mammalian Genomes (FANTOM) project (Figure 4.1 & Additional File 4.1) (Andersson et al. 2014) based bidirectional transcription identified via cap analysis of gene expression (CAGE) in diverse human cell and tissue types (The FANTOM Consortium et al. 2014). We further classified the enhancers into 'tissue-specific' enhancers that are detected in only a single tissue, and 'tissue-broad' enhancers that are detected in two or more tissues. As genomic regions located on sex chromosomes exhibit patterns of genetic variation that differ from those observed in autosomal genomic regions (Schaffner 2004), we restricted our analyses to only autosomal enhancers.

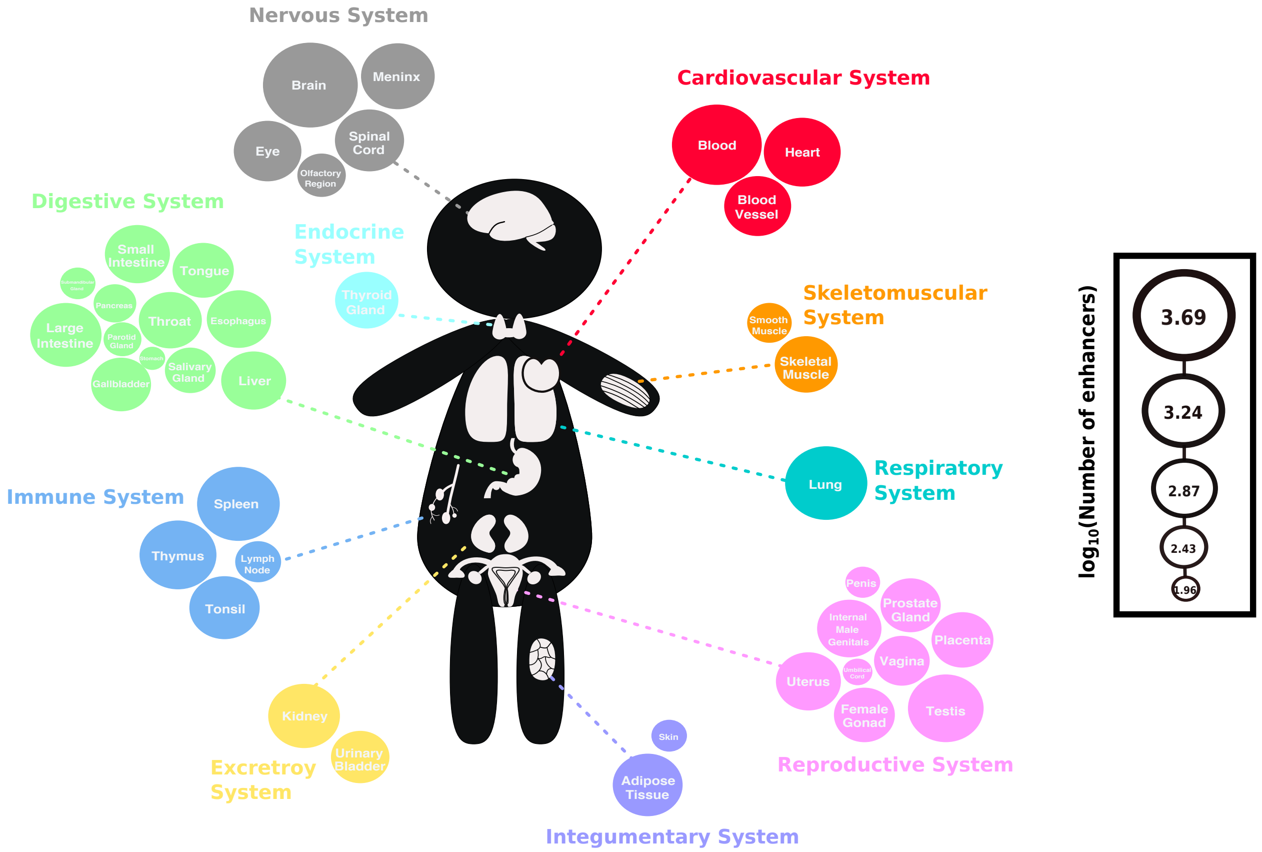

Figure 4.1. Visual summary of the FANTOM5 enhancer data set used in this study. Each circle represents a tissue and all 41 tissues are grouped into non-overlapping organ systems and are color-coded accordingly. The size of the circle is proportional to the $\log_{10}$ of the number of enhancers active within each tissue (Additional Table 4.1: from highest to lowest: brain: 4,883, blood: 3,657, spleen: 2,429, lung: 2,366, thymus: 1,741, heart: 1,737, testis: 1,621, meninx: 1,447, kidney: 1,294, large-intestine: 1,266, tonsil: 1,193, adipose-tissue: 1,161, spinal-cord: 1,115, eye: 1,019, blood-vessel: 982, small-intestine: 880, uterus: 877, liver: 875, internal-male-genitalia: 846, esophagus: 845, thyroid-gland: 795, skeletal-muscle: 784, throat: 782, placenta: 735, tongue: 708, female-gonad: 693, prostate-gland: 667, gallbladder: 654, urinary-bladder: 602, vagina: 512, salivary-gland: 387, olfactory-region: 338, lymph-node: 292, smooth-muscle: 270, pancreas: 243, parotid-gland: 184, skin: 161, submandibular-gland: 159, penis: 154, umbilical-cord: 115, stomach: 92).

Human Protein Atlas data

To compare the patterns of recent evolution between enhancers associated with genes that exhibit higher expression levels in one tissue compared to others and those that are not in our tissues of interest, the brain and testis, we downloaded the list of tissue-enriched genes for the brain and testis tissues from the Human Protein Atlas database (Uhlen et al. 2015) (version 88; download date: January 17th, 2018): the Human Protein Atlas Database defines a gene as tissue-enriched if its mRNA levels in a given tissue are at least five-fold higher compared to its levels of expression in all other tissues. We defined 'non-tissue-enriched' genes as those genes that are not labeled as 'tissue-enriched' according the to the Protein Atlas database.

Identifying signatures of recent positive selection

To identify signatures of recent positive selection on human enhancers, we calculated four metrics aimed to capture genetic signatures left from the action of selection at different time ranges during recent human evolution (Sabeti et al. 2006; Vitti et al. 2013; Moon et al. 2018) using previously described methodology (Moon et al. 2018). Briefly, the four metrics that we used can be divided into two categories: those that are designed to detect hard selective sweeps (Tajima's D, Weir & Cockerham's F_{ST} and nS_{L}) and those designed to detect soft selective sweeps (H12). Furthermore, the three tests for hard selective sweeps are most sensitive to selection events that occurred at different time points in human history: Tajima's D can detect selection that happened approximately 250,000 - 200,000 years ago; F_{ST} can identify signatures of local selection that occurred following the out-of-African migration approximately 75,000 - 50,000 years ago; finally, nS_L has power to detect signatures left by selection events that occurred approximately 20,000 - 10,000 years ago (Sabeti et al. 2006; Moon et al. 2018). Following the methods of Moon et al. (Moon et al. 2018) for each enhancer region in every tissue we calculated Tajima's *D* using the R package *PopGenome*, version 2.1.6 (Pfeifer et al. 2014) and the weighted Weir & Cockerham's fixation index (F_{ST}) (Weir & Cockerham 1984) among all three populations using *VCFtools*, version 0.1.13 (Danecek et al. 2011). In addition, we calculated the nS_L statistic (number of segregating sites by length; (Ferrer-Admetlla et al. 2014)) using *Selscan*, version 1.2.0 (Szpiech and Hernandez 2014). For each enhancer, we calculated nS_L for the region extending 50 kilobases (kb) upstream and downstream using the default settings of *Selscan*, except for the minor allele frequency (MAF) cut-off value, for which we used 0.01. We calculated this metric for each of the three populations, as haplotype-based methods are known to detect selection events that occurred after the out-of-African migration (Voight et al. 2006; Sabeti et al. 2006). We used the maximum absolute un-standardized nS_L values calculated over the entire window to represent each enhancer region. Finally, for each enhancer region we also calculated the H12 index using a custom script, based on Garud *et al.*'s original script (Garud et al. 2015). The values of all enhancers in the 41 human tissues for all metrics can be found in Additional File 4.4.

To test for statistical differences in the distributions of each of the four metrics between any two tissues, we also carried out pairwise Kolmogorov-Smirnov tests (two-sided).

Visualization of the data (Additional Figure 4.1) revealed that assumptions of normality and equal variance could not be upheld, and therefore, we chose the non-parametric Kolmogorov-Smirnov test which makes no assumptions regarding the shape of the distributions and carried out this test in the R programming environment. We also conducted this test to examine whether there are any significant differences between the distributions of patterns of recent evolution between tissue-specific and tissue-broad enhancers, as well as between enhancers that are associated with tissue-enriched and non-tissue-enriched genes. All statistical tests were followed by *post hoc* Bonferroni corrections in R. The results of the statistical tests for all metrics can be found in Additional Tables 4.22-27, 28-33.

Simulations of neutral evolution

The action of non-adaptive processes and demographic changes, such as genetic drift and population expansion, can produce patterns of variation that these tests may mistake for positive selection (Sabeti et al. 2006). To determine the likelihood of the empirical values being generated by the action of selection, we carried out simulations of neutral evolution based on a model of recent human demographic history and compared the observed values to expected values under neutrality. Simulations of neutrality were conducted using *SLiM*, version 2.4.1 (Messer 2013; Haller and Messer 2017). Following Gravel et al., we used previously calculated demographic parameters for the three populations included in our study (Gravel et al. 2011; Messer 2013). Detailed information on the parameters used for the simulations can be found in the Additional File 4.1.

We simulated the genotypes of 661, 503, and 504 individuals corresponding to the numbers of individuals from each of the three populations analyzed. Since there are 41,561 autosomal enhancers in the FANTOM dataset and anywhere from 92 (stomach) to 4,883 (brain) enhancers in any given tissue (Additional Table 4.1), carrying out individual simulations for each enhancer was computationally prohibitive. To reduce the computational load of our simulations, we focused on simulating the mutational profile of an average enhancer for a given tissue under a model of neutral evolution. Specifically, to carry out the neutral simulations for enhancers in a given tissue, we used the average length of all the autosomal enhancers found in that tissue (Additional File 4.2) and a fixed recombination rate of 1×10^{-8}. For each tissue, we carried out 10,000 simulations. We next used these simulated sequences to calculate Tajima's *D* values, Weir & Cockerham's F_{ST}, and H12 as described above. We calculated an empirical *p*-value for a given enhancer for any metric (including for nS_L, the simulation process of which is described below) as the proportion of simulated sequences that obtained scores equal to or more extreme than the observed value. We used a *p*-value of 0.05 as cutoff for significance; enhancers with *p*-values lower than 0.05 were considered to significantly deviate from neutral evolution and to have experienced selection.

Recombination rate interpolation and simulation of nS_L values

As variation in recombination rates can affect the values of nS_L (Ferrer-Admetlla et al. 2014), we carried out a separate set of neutral simulations for the calculation of nS_L values by incorporating average recombination rates for each tissue. To do this, we used the genome-wide genetic map curated by the HapMap II consortium (Frazer et al. 2007), which provides pre-computed recombination rates (cM/Mb) for the variants included in the HapMap II project. More specifically, we carried out linear interpolation to infer the recombination rates of the variants in the 1000 Genomes Project dataset that are not included in the HapMap Phase II dataset. Next, we calculated the average recombination rate (cM/Mb) for the region spanning 50 kb up and downstream of each enhancer in a given tissue, and then used the average of all recombination rates of the enhancer regions in a given tissue as the recombination rate parameter (converted to probability of crossovers per bp) for the neutral simulations. The average recombination rates, as used in our *SLiM* simulations, for each tissue can be found in Additional File 4.3. Using these calculated recombination rates, we then created 2,500 enhancers that span 50 kb upstream and downstream of the average length of all autosomal enhancers found in that given tissue and calculated nS_L values on these simulated enhancers using *Selscan* as described above. As with the actual enhancers, we only included variants with MAF greater than 0.01 in the calculations and used the maximum absolute un-standardized values to represent a given simulated window.

Functional enrichment analyses and semantic similarity calculations

To gain insight into the functions of genes associated with enhancers that show evidence of selection, we next carried out functional enrichment analyses. To associate enhancers with putative target genes, we used the transcription start site (TSS)-enhancer mapping file generated by Andersson et al. (Andersson et al. 2014) (file location information can be found in Additional File 4.1), which was compiled by measuring the pairwise correlation between enhancer activity and transcription level of putative target genes. We carried out functional enrichment analyses on the list of these putative target genes using the R package *TopGO*, version 2.32.0 (Alexa and Rahnenfuhrer 2016). Detailed description of the files and commands used can be found in Additional File 4.1. In short, we used the 'weight' algorithm that compares the significance scores of the connected nodes to explicitly account for the hierarchy of the gene ontology tree and carried out analyses for the three general ontologies: Biological Process (BP), Molecular Function (MF), and Cellular Component (CC). Subsequent corrections for multiple comparisons were carried out by calculating the FDR-adjusted *p*-values in the R statistical environment; only those GO terms with an FDR-adjusted *p*-value less than 0.05 were retained for further analyses. All *TopGO* enrichment analysis results for all the metrics can be found in Additional Files 4.5-21.

To quantify the overall similarity of the patterns of gene functional enrichment among different tissues, we calculated pairwise semantic similarity of the GO terms between any two tissues using a graph-based method (Wang et al. 2007), which takes into consideration the topology of the GO graph structure (i.e., the location of the GO terms in the graph and the relationship of the terms to their ancestor terms), as implemented in the R package *GOSemSim*, version 2.4.1 (Yu et al. 2010): this analysis was carried out for only those tissues that have 5 or more significant GO terms. To determine if the semantic similarity scores thus calculated significantly deviated from random expectations, we randomly sampled the same number of GO terms for each tissue from the pool of all GO terms associated with a given tissue 1,000 times and calculated the pairwise semantic similarity score; scores equal or greater than the 95th percentile value of the distribution of semantic similarity scores were considered significant. The calculated semantic similarity values, along with the 95th percentile values obtained from the 1,000 randomly sampled GO terms for each tissue pair, can be found in Additional Tables 4.7-15.

Transposable element (TE) data and analyses

TE-derived sequences often contribute to the origin of new enhancers and the modification of regulatory networks (Lynch et al. 2015; Simonti et al. 2017). To determine if the enhancers with evidence of recent positive selection harbor higher proportions of TEs than those with nonsignificant deviations from neutral expectations, we used BEDOPS, version 2.4.35 (Neph et al. 2012) to overlap the enhancer regions with the RepeatMasker-annotated regions. We downloaded the RepeatMasker annotation track for the hg19 assembly from the UCSC Genome Browser (Kent et al. 2002; Casper et al. 2018; Raney et al. 2014). The location of the RepeatMasker annotation file and the specific commands used for BEDOPS can be found in Additional File 4.1. To test for statistical differences in the proportions of TEs between enhancers with evidence of recent selection and those that do not exhibit such significant deviation from neutral expectations, we carried out a 2x2 chi-squared test with Yate's continuity correction using R. The results of the chi-squared tests can be found in Additional Tables 4.16-21.

GWAS catalog data annotation

Previous studies have shown that many variants in human *cis*-regulatory regions are significantly associated with a broad range of complex traits and diseases (Lee and Young 2013; Andersson et al. 2014; Karnuta and Scacheri 2018). To investigate whether brain and testis enhancers that exhibit evidence of recent positive selection show enrichment of variants in the Genome-Wide Association Study (GWAS) catalog, we downloaded the NHGRI-EBI GWAS catalog, version1.0.2 (MacArthur et al. 2017) (download date: June 2nd, 2018) and first compared the SNPs that reside in the enhancer regions with GWAS catalog SNPs. We also included SNPs that are outside the enhancer regions but in complete linkage disequilibrium (LD) with the SNPs within the enhancer regions. We used Plink, version 1.0.9 (Chang et al. 2015) to obtain a list of SNPs that are in complete LD ($r^2 = 1.0$) with those that lie within the enhancers of interest ('complete LD SNPs') and carried out the same analyses as described above. Detailed description of the Plink analyses can be found in Additional File 4.1. For subsequent analyses, we combined the GWAS hits of the SNPs that reside within the enhancers and those that are in complete LD with SNPs inside the enhancers and considered them together. We also carried out the same analyses for brain and testis enhancers that exhibit nonsignificant deviations from neutral expectations. To determine if there is a statistical difference in the enrichment of variants in the GWAS hits between enhancers that exhibit evidence of recent positive selection and those that do not, we carried out 2x2 chi-squared tests with Yate's continuity correction using R. The full list of all the GWAS hits that overlap with all the variants considered, as well as the chi-squared test results, can be found in Additional Tables 4.36-37 and 4.38-39.

Results

Enhancers experienced positive selection at different time ranges during recent human history

Calculation of four metrics of selection (Weir & Cockerham's F_{ST}, Tajima's D, H12 and nS_L) for all 41,561 enhancers in 41 tissues (Additional File 4.4) revealed that an average 5.90% of enhancers have experienced recent positive selection considering all metrics and tissues (Figure 4.2). The proportions of enhancers that exhibit significant deviations from neutral expectations for each metric varied (Figure 4.2; Additional Table 4.2). Specifically, greater fractions of enhancers show evidence of recent positive selection according to H12 and Tajima's D metrics (H12: from 6.57% to 16.67%; Tajima's D: 6.75% to 13.92%) than other metrics (F_{ST}: 1.91% to 5.56%; nS_L in Africans: 1.12% to 7.08%; nS_L in Europeans: 2.25% to 5.39%; nS_L in East Asians: 2.25% to 5.93%). Furthermore, for 33/41 tissues examined, the H12 metric had the highest proportion of enhancers exhibiting evidence of recent selection. These results imply that human enhancers experienced selection at different time points in human history, with substantial evidence for soft selective sweeps in which multiple haplotypes increase in frequency within a population (Pennings and Hermisson 2006; Messer and Petrov 2013). However, we caution against overly interpreting differences in the proportions of enhancers across metrics as differences in the action of selection across different time periods since they likely vary in power to detect selection.

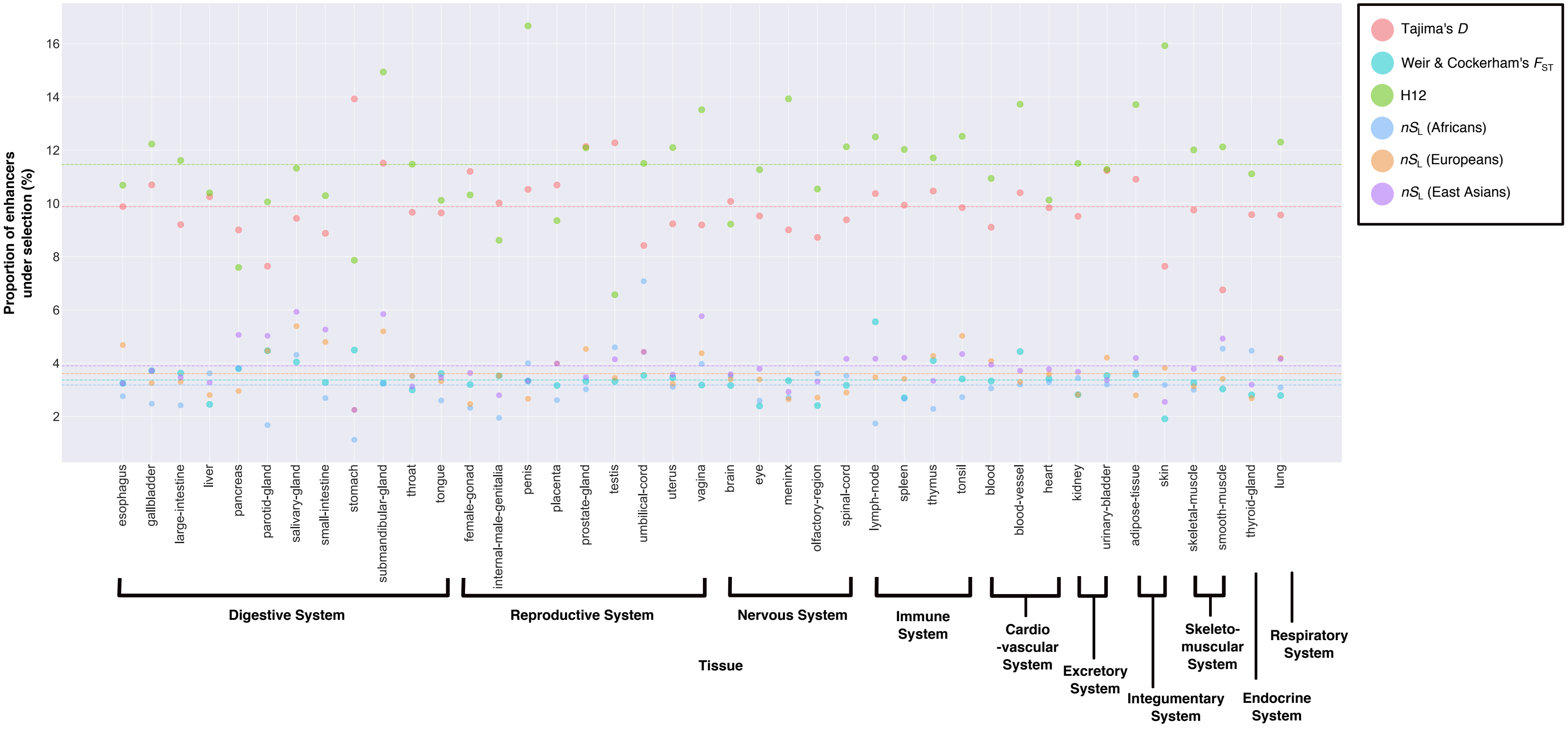

Figure 4.2. Proportions of enhancers exhibiting significant deviations from neutral expectations for different recent positive selection metrics across 41 tissues. Each dot represents the proportion of enhancers (Y-axis) that exhibited significant deviation from the neutral model (i.e., *p*-value < 0.05) in a given tissue (X-axis). Differently colored dots correspond to the different selection metrics used (Salmon: Tajima's *D*; Teal: F_{ST}; Green: H12; Blue Gray: nS_L (Africans); Orange: nS_L (Europeans); Purple: nS_L (East Asians)). For each metric, the *p*-value for the observed value for an enhancer (i.e., the likelihood under neutral expectations of obtaining a value as or more extreme as the observed

value) was assessed by comparing to 10,000 simulated values calculated on sequences generated from the neutral simulations. Differently colored horizontal lines correspond to the average of all proportions calculated for all 41 tissues according to the different selection metrics used (Salmon: Tajima's *D*; Teal: F_{ST}; Green: H12; Blue Gray: nS_{L} (Africans); Orange: nS_{L} (Europeans); Purple: nS_{L} (East Asians)). Note the 1) differences in the proportion of enhancers exhibiting significant deviations from neutral expectations across tissues in any given metric, and 2) similarly, differences in the proportion of enhancers with significant deviations from the neutral model across all metrics in any given tissue.

Enhancers that exhibit evidence of recent positive selection putatively regulate the activities of genes with immunity-related functions

To study the functions of the enhancers that show evidence of recent positive selection, we carried out enrichment analysis on the putative target genes associated with them (these results can be found in the Additional Files 4.5-21). We found that there was little functional enrichment among the enhancers identified by F_{ST}, Tajima's *D* and H12 (Table 4.1). However, many tissues had multiple significant enriched GO terms among the nS_{L} hits (Table 4.1). For the F_{ST}, Tajima's *D* and H12 metrics, we also collapsed all enhancers irrespective of their tissues and carried out the same analyses: we found similar patterns as described above, with no significant GO terms for the F_{ST} metric and very few GO terms for Tajima's *D* and H12 metrics (Additional Tables 4.3-6). The majority of top 10 most frequent significant GO terms among all tissues for the nS_{L} metric were immunity-related (Tables 4.2-4.4). In addition, we also carried out the same analyses on the putative target genes of enhancers with no evidence of recent positive selection according to the nS_{L} metric and found that none of the tissues possessed any significantly enriched GO terms: the only exception was a single Cellular Component term in the parotid-gland for nS_{L} metric in Europeans (GO:0044444: cytoplasmic part, adjusted *p*-value = 0.047).

These results suggest that approximately 20,000–10,000 years ago, enhancers that putatively regulate the activities of immunity-related genes in multiple tissues underwent selection, a finding consistent with the hypothesis that human populations faced novel local selective pressures as they moved into new environments (Balaresque et al. 2007; Fumagalli et al. 2011).

Table 4.1 Number of tissues with one or more significantly enriched GO terms

Metric	**# of Biological Process GO terms**	**# of Molecular Function GO terms**	**# of Cellular Compartment GO terms**
Tajima's *D*	4	2	4
F_{ST}	2	2	2
H12	0	1	1
nS_L (AFR)	18	18	18
nS_L (EUR)	20	20	19
nS_L (EAS)	19	19	19

Table 4.2. Most frequently-occurring GO terms among enhancers with evidence of recent positive selection by nS_L across tissues in Africans

BP GO Terms	# of tissues	MF Go Terms	# tissues	CC GO Terms	# of tissues
Antigen processing and presentation of exogenous peptide antigen	18	MHC class II receptor activity	18	trans-Golgi network membrane	18
T cell co-stimulation	18	MHC class II protein complex binding	18	Integral component of luminal side of endoplasmic reticulum membrane	18
interferon-gamma-mediated signaling pathway	17	Peptide antigen binding	15	Endosome membrane	18
T cell receptor signaling pathway	16	TAP1 binding	13	ER to Golgi transport vesicle membrane	17
Antigen processing and presentation of exogenous peptide antigen via MHC class I TAP dependent	15	Peptide antigen-transporting ATPase activity	12	Lysosomal membrane	16
MHC class II protein complex assembly	14	TAP2-binding	10	TAP complex	14
				MHC protein complex	11

Table 4.3. Most frequently-occurring GO terms among enhancers with evidence of recent positive selection by nS_L across tissues in Europeans

BP GO Terms	# of tissues	MF GO Terms	# of tissues	CC GO Terms	# of tissues
Interferon gamma-mediated signaling pathway	18	MHC class II receptor activity	18	Integral component of luminal side of endoplasmic reticulum membrane	19
Antigen processing and presentation of exogenous peptide antigen	17	MHC class II protein complex binding	18	ER to Golgi transport vesicle membrane	18
T cell co-stimulation	17	Peptide antigen binding	15	trans-Golgi network membrane	18
T cell receptor signaling pathway	14	TAP1 binding	14	Endosome membrane	17
MHC class II protein complex assembly	13			Lysosomal membrane	15
				Endocytic vesicle membrane	12
				TAP complex	12
				MHC protein complex	12

Table 4.4. Most frequently-occurring GO terms among enhancers with evidence of recent positive selection by nS_L across tissues in East Asians

BP GO Terms	# of tissues	MF GO Terms	# of tissues	CC GO Terms	# of tissues
Antigen processing and presentation of exogenous peptide antigen	17	Peptide antigen binding	19	Integral component of luminal side of endoplasmic reticulum membrane	19
T cell co-stimulation	17	MHC class II receptor activity	18	trans-Golgi network membrane	18
interferon-gamma-mediated signaling pathway	16	MHC class II protein complex binding	18	Lysosomal membrane	18
T cell receptor signaling pathway	15	TAP1 binding	12	ER to Golgi transport vesicle membrane	17
MHC class II protein complex assembly	14			Endocytic vesicle membrane	16
Antigen processing and presentation of exogenous peptide antigen via MHC class I TAP dependent	11			TAP complex	12
				Endosome membrane	10
				MHC II protein complex	10

We also quantified the semantic similarity (SS) of the functionally enriched terms among the tissues to determine if the pattern of enrichment for immunity-related functions was shared across tissues (Additional Figures 4.2a-i). The pairwise SS values calculated using the *TopGO* GO terms associated with the nS_L metrics were consistently very high (BP: Africans: 0.687 to 1.000; Europeans: 0.786 to 1.000; MF: Africans: 0.779 to 1.000; Europeans: 0.715 to 1.000; East Asians: 0.902 to 1.000; CC: Africans: 0.873 to 1.000; Europeans: 0.864 to 1.000; East Asians: 0.853 to 1.000): one exception to this general trend were the SS scores calculated on the Biological Process terms in East Asians (from 0.208 to 1.000) (Additional Figure 4.2c). To determine if the range of SS values we observed were unusually high, we compared each pairwise SS value to the 95th percentile value of the SS values calculated on the 1,000 randomly sampled GO terms for the same pair of tissues (these values, as well as the empirical SS values, can be found in Additional Tables 4.7-15). We found that most of the empirical SS values were higher than the 95th percentile values of the SS values calculated on the randomly sampled GO terms, with the exception of pairwise comparisons involving liver for the Biological Process terms in East Asians (Additional Figure 4.2c). In other words, most pairs of tissues were associated with semantically similar GO terms, suggesting that patterns of functional enrichment were very similar across tissues. The sole exception to this were 13 comparisons involving the liver, which had three additional biological terms in East Asians (GO:0071294: cellular response to zinc ion, adjusted *p*-value = 0.005; GO:0046597: negative regulation of viral entry into host cell, adjusted *p*-value = 0.005; GO:007126: cellular response to cadmium ion, adjusted *p*-value = 0.014), which were found only in the liver. Overall, our analyses suggest that the enrichment for immunity-related functions found in putative target genes of enhancers with evidence of recent positive selection is not confined to specific tissues, but represents a general trend.

Enhancers that exhibit evidence of recent positive selection are not enriched for transposable element origins

To examine whether enhancers that exhibit statistically significant signatures of recent positive selection tend to have arisen from TEs, we compared the proportions of TE overlap between enhancers that exhibit evidence of recent positive selection versus enhancers with no evidence of recent positive selection across all 41 tissues. Of the 41 tissues examined, only the lung ($\chi^2 = 12.859$, adjusted *p*-value = 0.002) and the female gonad ($\chi^2 = 8.632$, adjusted *p*-value = 0.020) displayed statistically significant differences in terms of the proportions of TE-overlapping regions between the two groups of enhancers for the H12 metric (Additional Figures 4.3a-f). Moreover, in these two tissues, enhancers with evidence of recent positive selection showed lower proportions of TE overlap than enhancers with no such evidence. Overall, these results suggest that enhancers that have arisen from TEs were not preferential targets of recent positive selection.

Enhancers active in testis and brain show different patterns of recent evolution

To test the hypothesis that signatures of selection during recent human history differed across enhancers active in different tissues, we compared the distributions of the four metrics across all 820 pairs of the 41 tissues (these results can be found in Additional Tables 4.22-27). The number of tissue pairs exhibiting different distributions for Tajima's *D* was 22 / 820, for F_{ST} was 5 / 820, for H12 was 56 / 820, and for nS_L was 2 / 820 in Africans, and 0 / 820 in Europeans and East Asians (Figures 4.3a-f). All tissue pairs exhibiting significantly different distributions involved either brain or testis, with the exception of two nS_L comparisons (Africans: Adipose-tissue and Large-intestine: *p*-value = 0.014; Adipose-tissue and Thymus: *p*-value = 0.007). Specifically, brain enhancers exhibited significantly different distributions for Tajima's *D*, F_{ST}, and H12 compared to enhancers from 5, 2, and 23 other tissues (Figures 4.3a-c); similarly, testis enhancers exhibited significantly different distributions for Tajima's *D*, F_{ST} and H12 compared to enhancers from fifteen, three, and thirty other tissues (Figures 4.3a-c). We note that the high fractions of enhancers under selection in brain and testis are unlikely to be solely due to the high number of enhancers found in these tissues; specifically, there are several other tissues that have high number of enhancers (Additional Table 4.1) and yet fail to exhibit statistically significant pairwise comparisons with other tissues.

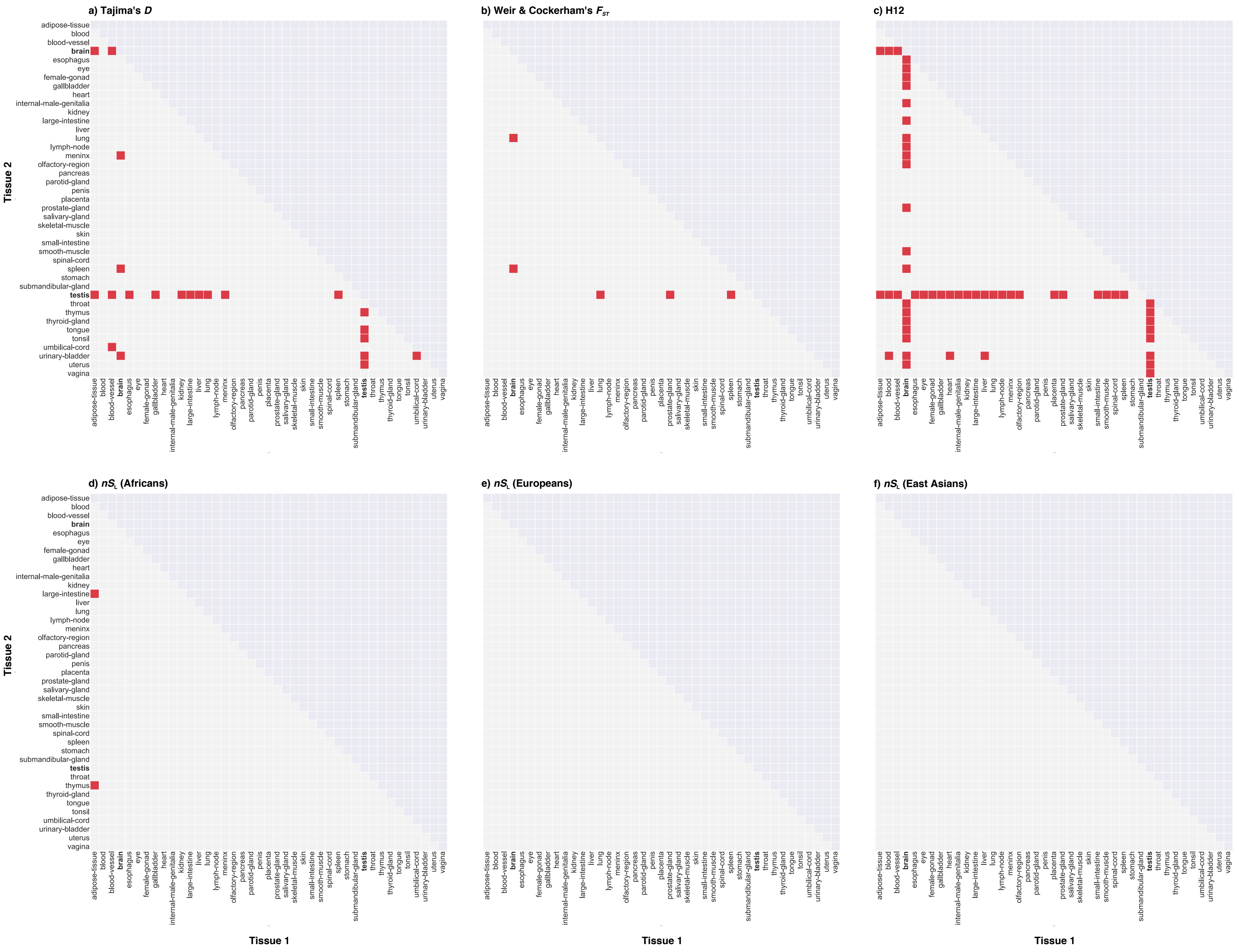
a) Tajima's D
b) Weir & Cockerham's F_{ST}
c) H12
d) nS_L (Africans)
e) nS_L (Europeans)
f) nS_L (East Asians)
Tissue 2
Tissue 1
adipose-tissue
blood
blood-vessel
brain
esophagus
eye
female-gonad
gallbladder
heart
internal-male-genitalia
kidney
large-intestine
liver
lung
lymph-node
meninx
olfactory-region
pancreas
parotid-gland
penis
placenta
prostate-gland
salivary-gland
skeletal-muscle
skin
small-intestine
smooth-muscle
spinal-cord
spleen
stomach
submandibular-gland
testis
throat
thymus
thyroid-gland
tongue
tonsil
umbilical-cord
urinary-bladder
uterus
vagina

Figure 4.3. Pairwise comparisons of the patterns of recent evolution among enhancers from 41 tissues. Each graph shows the results of the pairwise Kolmogorov-Smirnov tests carried out between all pairs of 41 tissues (shown on the X- and Y-axes) for all recent positive selection metrics. Each cell represents a pairwise comparison between two specific tissues: red-filled cells represent pairwise comparisons that are statistically significant (i.e., adjusted *p*-value < 0.05) and empty cells represent non-significant pairwise comparisons.

In general, the significant differences in the distributions of these metrics between testis or brain and other tissues were due to differences in the magnitude of the peaks of the distributions and/or differences in shapes of the distributions rather than from shifts in the range of the distributions. For all metrics, most pairs of tissues with significant differences had very subtle shifts in the distributions, and in cases where the shifts were more noticeable, brain and testis enhancers' distributions tended to be shifted to the left (for F_{ST}; Additional Figure 4.5) or the right (for Tajima's *D*; Additional Figure 4.4), both in the direction suggestive of lower levels of positive selection. The nature of the differences in the peaks of the magnitude varied for each metric: for F_{ST}, the peaks for both brain and testis were consistently higher compared to those of the other tissues. In contrast, for Tajima's *D*, the peaks for brain and testis were consistently lower than the other tissues being compared; the sole exception was the comparison between brain and meninx and urinary-bladder, in which the peaks were almost equal or slightly higher for the brain, respectively. In addition, for Tajima's *D*, there were differences in the shapes of the distribution (Additional Figure 4.4). For H12, differences in the shape of the distributions also contributed to the statistical differences we observed between brain or testis and other tissues (Additional Figures 4.6-7). These results imply that enhancers active in the brain and testis experienced different selective pressures during recent human history compared to enhancers active in other tissues.

We also ranked the tissues according to the median values of each metric (Additional Figure 4.8) and the proportion of enhancers that exhibit statistically significant evidence of recent positive selection for each metric (Additional Figure 4.9). Overall, we did not find any general patterns regarding the recent evolution of brain and testis, in terms of either the median values of the metrics of recent selection (Additional Figure 4.8) or the proportion of enhancers that have putatively been under selection (Additional Figure 4.9). Exceptions to the general trend were that testis and brain ranked 1st and 2nd, respectively, for median values of the H12 metric (Additional Figure 4.8c) and that testis ranked 2nd in the proportion of enhancers with evidence of recent selection according to Tajima's *D* and nS_L in Africans (Additional Figures 4.9a & d).

Evolution of tissue-specific enhancers differs from that of tissue-broad enhancers in the brain and testis

We next compared the patterns of recent evolution between tissue-specific and tissue-broad enhancers for each tissue. We found that the number of tissues with statistically significant differences in the distributions of the metrics between enhancers active in a single tissue and those active in multiple tissues was as follows, sorted from highest to lowest: 20/41 for H12, 8/41 for Tajima's *D*, 8/41 for F_{ST} and 1/41 for nS_L in Africans and East Asians (Figure 4.4a & Additional Figures 4.10-15). Tissue-specific and tissue-broad enhancers in the brain and testis show statistically significant differences in the patterns of recent evolution for all metrics except nS_L (Tajima's D: brain: 1.882e-07; testis: 9.523e-11; F_{ST}: brain: 6.074e-07; testis: 1.639e-07; H12: brain: < 2.2e-16; testis: < 2.2e-16 (Figures 4.4b-d). Regarding the nature of the differences between the two groups of enhancers in these tissues, we found contrasting patterns for H12 compared to Tajima's D and F_{ST}: for Tajima's D and F_{ST}, we found that the interquartile values of the tissue-specific enhancers were lower (or higher for Tajima's D) than those of the tissue-broad enhancers in brain and testis, whereas those values were consistently higher for the tissue-specific enhancers compared to the tissue-broad enhancers in the same tissues for the H12 metric. These results suggest that, in general, enhancers with different breadth of tissue activity did not differ in their patterns of recent evolution. Nevertheless, enhancers active only in the brain do show significant differences compared to those active in multiple tissues (including brain) and the same is true for testis, raising the possibility that the recent evolution of brain- and testis-specific enhancers may have been different from that of other tissue-specific enhancers.

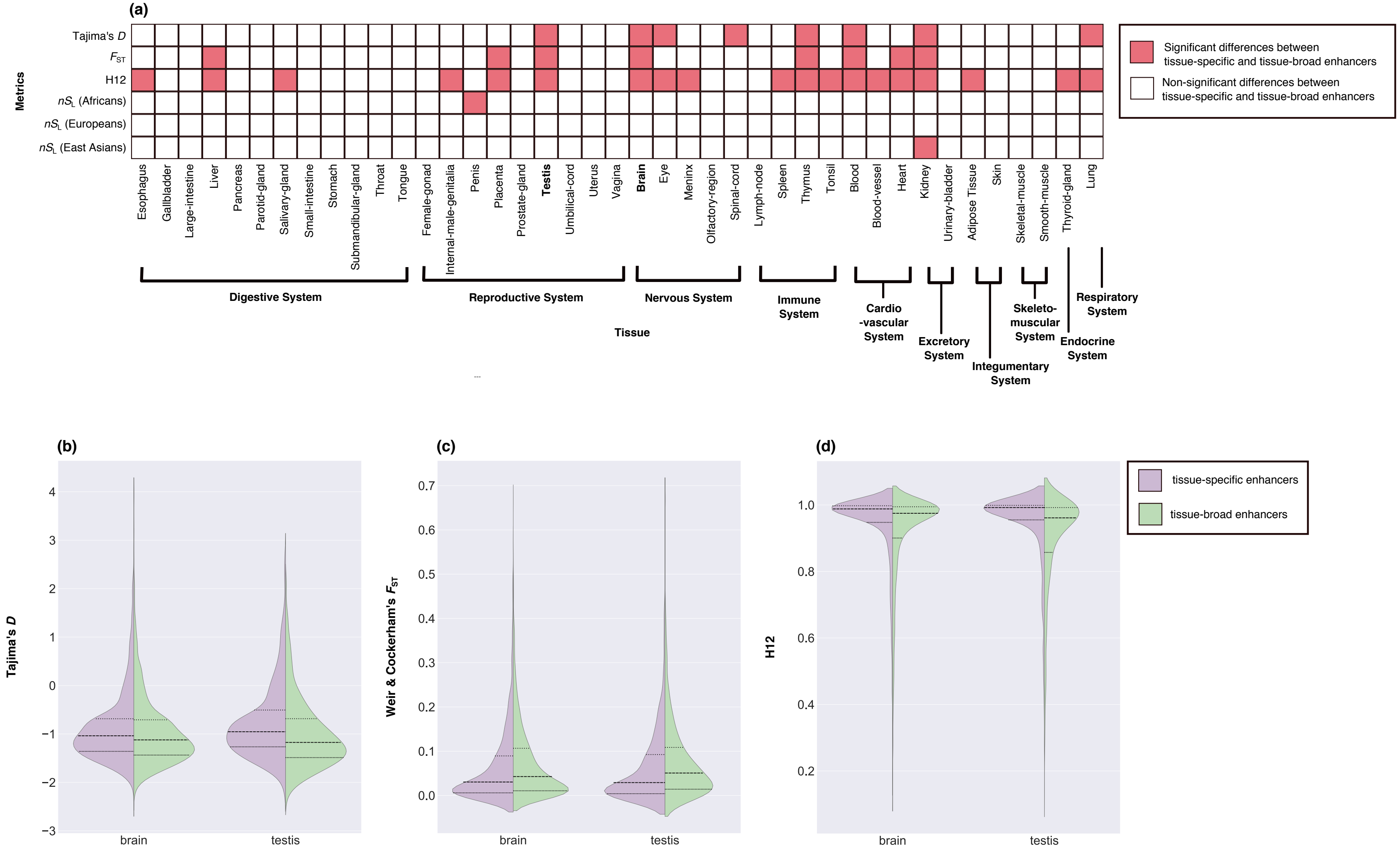

Figure 4.4. The recent evolution of testis enhancers associated with tissue-enriched genes differs from the evolution of enhancers associated with non-tissue-enriched genes. (a) The grid panel depicts which tissues exhibit significantly different distributions of metrics between tissue-specific and tissue-broad enhancers. Each cell represents a

comparison between tissue-specific and tissue-broad enhancers in a given tissue: pink-filled cells represent the comparisons that are statistically significant (i.e., adjusted *p*-value < 0.05) and empty cells represent non-significant comparisons. (b–d) Violin plots depicting the distributions of the metrics for tissue-specific and tissue-broad enhancers for brain and testis for (b) Tajima's *D*, (c) Weir & Cockerham's F_{ST}, and (d) H12. In all cases, the distributions of tissue-specific and tissue-broad enhancers were significantly different (see also panel a).

The recent evolution of testis enhancers associated with tissue-enriched genes differs from the evolution of enhancers associated with non-tissue-enriched genes

To determine whether there are differences in the patterns of recent evolution between enhancers associated with tissue-enriched genes versus those associated with non-tissue-enriched genes, we compared the distributions of the four metrics between brain and testis enhancers stratified by the breadth of expression of their target genes (these results can be found in Additional Tables 4.34-35). There were significant differences for H12 (adjusted *p*-value = 3.667e-06) and Tajima's *D* (adjusted *p*-value = 0.004) in the testis (Figure 4.5a-b). More specifically, we found that the distribution of Tajima's *D* values of the enhancers associated with testis-enriched genes were shifted to the left (i.e., more negative values) compared to the enhancers associated with non-testis-enriched genes (Figure 4.5a). The difference between the same groups of enhancers for the H12 metric was comparatively subtle, with the main difference being the magnitude of the peak (Figure 4.5b); the peak of the distribution of the enhancers associated with the non-testis-enriched genes was higher than that of the enhancers associated with testis-enriched genes. In contrast, there were no significant differences in the distributions of F_{ST} and nS_{L} values between these two categories of enhancers for either tissue and the distributions of H12 and Tajima's *D* between the same categories of enhancers for the brain (Figure

4.5). These results imply that enhancers associated with genes with enriched expression in the testis have experienced different degrees of selection in the form of both hard and soft selective sweeps compared to enhancers associated with genes that do not show enriched expression in the same tissue.

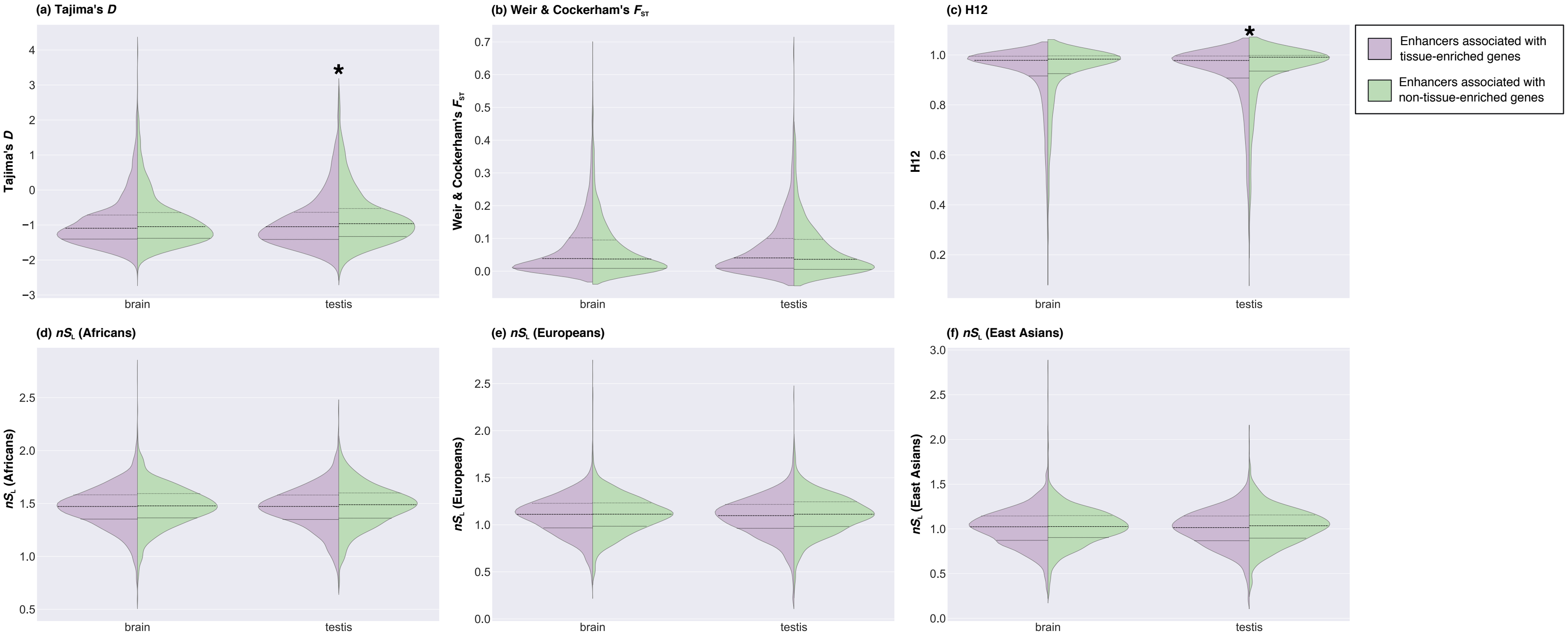

Figure 4.5. Comparisons of patterns of recent evolution between enhancers associated with tissue-enriched genes and enhancers associated with non-tissue-enriched genes in brain and testis. The violin plots depicting the distributions of the metrics for enhancers associated with tissue-enriched and non-tissue-enriched genes for brain and testis for: (a) Tajima's *D* (b) Weir & Cockerham's F_{ST} (c) H12 (d) nS_L (Africans) (e) nS_L (Europeans) (f) nS_L (East Asians). Plots with asterisks above them indicate significant pairwise comparisons.

Brain and testis enhancers under recent positive selection are not significantly enriched for variants associated with complex human traits and diseases

To examine if genetic variants in brain and testis enhancers showing signatures of recent selection are associated with particular complex human traits and diseases, we queried the NHGRI-EBI GWAS catalog (MacArthur et al. 2017). We found that for all metrics, most or all enhancers did not harbor SNPs that are associated with any human traits (brain: Tajima's *D*: 372/421; F_{ST}: 131/150; H12: 430/437; nS_L(Africans): 150/167; nS_L (Europeans): 144/161; nS_L (East Asians): 152/170; testis: Tajima's *D*: 143/163; F_{ST}: 47/52; H12: 102/103; nS_L (Africans): 64/72; nS_L (Europeans): 48/54; nS_L (East Asians): 60/65). In the few instances in which SNPs within enhancers showing evidence of selection were associated with human traits, these associations were with traits such as Alzheimer's disease (brain: Tajima's *D*), obesity-related traits (brain: Tajima's *D*), and type 2 diabetes (brain: nS_L (Africans)) (Tables 4.5 & 6; the full list of overlapping GWAS hits, including variants in LD with those inside enhancers, can be found in Additional Files 4.36-37). We next compared the proportions of enhancers that overlapped with GWAS hits between enhancers that exhibit evidence of recent positive selection and enhancers that do not and did not find any significant difference between them. More specifically, we saw no statistically significant differences in the proportions of overlap with GWAS hits between enhancers with or without evidence of recent positive selection: the only exception was the H12 metric in the brain (χ^2= 30.552, adjusted *p*-value = 1.950e-07), which showed depletion of GWAS hits among enhancers with evidence of recent positive selection (These results can be found in Additional Tables 4.38-39). These results show that overall, enhancer SNPs under recent positive selection are not preferentially associated with specific human traits or complex human diseases.

Table 4.5. Complex human traits and diseases associated with variants within the brain enhancers that exhibit evidence of recent selection

Enhancer ID	rsID	Traits	Metrics
chr2:219271996-219272374	rs921968	Mean corpuscular hemoglobin concentration	F_{ST}
chr3:181418145-181418802	rs34308817	Ankle injury	Tajima’s D
chr6:30069810-30070038	rs1111180	Eosinophil percentage of granulocytes, Eosinophil percentage of white cells	nSL (EUR)
chr6:30923441-30923743	rs17189763	Conotruncal heart defects (maternal effects)	nS_L (EUR)
chr6:32427743-32428120	rs9268831	Response to hepatitis B vaccine	nS_L (AFR, EUR, EAS*)*
chr6:32427743-32428120	rs9268835	Type 2 diabetes	nS_L (AFR, EUR, EAS)
chr6:32577297-32577935	rs660895	IgA nephropathy, Rheumatoid arthritis, Rheumatoid arthritis (ACPA-negative)	nS_L (AFR, EUR, EAS*)*
chr7:130720133-130720826	rs10265693	Lung cancer	F_{ST}
chr7:29217860-29218383	rs245914	Psychosis (atypical), Obesity-related traits	Tajima’s D
chr13:110790027-110791241	rs641862	Obesity-related traits	Tajima’s D
chr16:22201170-22202123	rs145049847	Alzheimer disease and age of onset	Tajima’s D
chr16:87856343-87856555	rs76069656	Triglyceride change in response to fenofibrate in statin-treated type 2 diabetes	nS_L (EUR)
chr16:87886258-87886670	rs68149176	Mean corpuscular volume, Mean corpuscular hemoglobin	nS_L (EUR), F_{ST}
chr21:45616099-45616530	rs4456788	Chronic inflammatory diseases (ankylosing spondylitis, Crohn's disease, psoriasis, primary sclerosing cholangitis, ulcerative colitis) (pleiotropy)	nS_L (EAS)

Table 4.6. Complex human traits and diseases associated with variants within the testis enhancers that exhibit evidence of recent selection

Enhancer ID	rsID	Traits	Metrics
Chr6:30069810-30070038	rs1111180	Eosinophil percentage of granulocytes, Eosinophil percentage of white cells	nS_L (Europeans)
Chr6:30923441-30923743	rs17189763	Conotruncal heart defects (maternal effects)	nS_L (Europeans)

Discussion

In this study, we calculated four different metrics that detect distinct genomic signatures of recent selection on the enhancers active in 41 human tissues and compared the empirical values of these metrics to those calculated on neutrally simulated sequences. We found that across all tissues and metrics, approximately 5.90% of enhancers exhibit significant evidence of recent positive selection. We also found that the putative target genes of such enhancers are enriched for immunity-related functions, and that this enrichment pattern is shared across multiple tissues. Furthermore, enhancers active in the brain and testis exhibited significantly different patterns of recent evolution compared to enhancers in other tissues.

Across tissues, we observed variation in the proportions of enhancers that exhibited signatures of selection across the different metrics (Figure 4.2). Since the origin of modern humans approximately 250,000 - 200,000 years ago, there have been multiple distinctive events in human history, including the out-of-Africa migration approximately 75,000 - 50,000 years ago, the Eurasian split that occurred approximately 45,000 - 36,000 years ago, and the Agricultural Revolution approximately 20,000 - 10,000 years ago, all of which would have likely exposed the human populations to novel selective pressures (e.g., pathogens, dietary changes) (Sabeti et al. 2006; Karlsson et al. 2014). Adaptations in response to such changes would have likely altered the allele frequencies in both the genic and regulatory regions. Furthermore, it has been previously shown that different metrics of selection are sensitive to distinct types of genomic signatures (Sabeti et al.

2006; Voight et al. 2006). For example, Tajima's *D* is most suitable for detecting events that occurred approximately 250,000 - 200,000 years ago, F_{ST} is most suitable for identifying selection events that occurred as humans left Africa and were exposed to novel environments around the world (i.e., 75,000 - 50,000 years ago), whereas nS_L is most suitable for detecting selection events that happened approximately 20,000 - 10,000 years ago (Sabeti et al. 2006; Voight et al. 2006). Therefore, one possible interpretation is that within any given tissue, the differences in percentages of enhancers under selection between metrics reflect temporal variation in the action of selection during human history.

While the variation in the proportion of enhancers with significant evidence of recent selection across different metrics could represent temporal variation in the action of selection, it is also likely that these metrics have different power to detect selection. For instance, Tajima's *D*, a metric that detects excess of rare alleles in a region of interest, has been shown to be most sensitive to selective sweeps that have resulted in almost complete fixation (i.e., allele frequencies = 100%) of the target locus and conversely has limited power to detect actions of selection that have resulted in incomplete fixation of the allele (Sabeti et al. 2006; Ferrer-Admetlla et al. 2014). In contrast, nS_L, a haplotype-based metric, best detects ongoing or incomplete hard selective sweeps resulting in intermediate allele frequencies (i.e., allele frequencies = 60 - 80%) and rapidly loses power as allele frequencies increase to 100% (Ferrer-Admetlla et al. 2014). Thus, the observed differences in the proportions of enhancers with evidence of recent positive selection across the metrics also reflect the disparity in power of the metrics to successfully detect incidences of selection that happened at a particular time period. Consequently, the proportions of enhancers with evidence of recent selection across metrics cannot be directly compared to each other. Distinguishing whether our results are best explained by temporal variation in selection or by differences in the power of our metrics could be achieved via simulations in which selective events are introduced at specific time points and result in a fixed proportion of regions being selected. This is an important future research direction that has potential to shed light into the tempo of selection in the course of recent human evolution.

An additional result of our study is that the putative target genes of enhancers exhibiting significant evidence of recent positive selection according to the nS_L metric are enriched for immunity-related functions (Tables 4.2-4). Haplotype-based metrics such as nS_L are known to be sensitive to signatures resulting from selection that occurred approximately 10,000 - 20,000 years ago, which corresponds to the incidence of the Agricultural Revolution (Voight et al. 2006). The advent of farming practices, communal living in settlements, and migrations of farmers across the globe resulted in an increase of numbers and densities of humans in any given location, likely facilitating the spread of pathogens (Varki and Gagneux 2009; Page et al. 2016; Nielsen et al. 2017). In addition, recent studies suggest that selection on *cis*-regulatory regions, such as enhancers, might have been important in driving adaptation of modern human populations to distinct environments, due to their modular organization: change of expression pattern in one temporal or spatial context can often occur without affecting others, which could contribute to phenotypic changes without incurring negative pleiotropic effects (Carroll 2005; Wray 2007). Therefore, it is possible that around 10,000 - 20,000 years ago, enhancers regulating the activities of immunity-related genes underwent selection to allow refined fine-tuning of host defense processes in response to the stronger pressure from pathogens resulting from increased human population sizes.

Perhaps the most striking result of our study is that brain and testis enhancers exhibited different patterns of recent evolution compared to enhancers of other tissues (Figure 4.3 & Additional Figures 4.4-7). For both brain and testis, the high number of enhancers included in these tissues could partly explain our results. Brain, with 4,883 enhancers, has the most enhancers of all the tissues examined, while testis, with 1,621 enhancers, is ranked 7th. The effect of the large number of enhancers included in these tissues is reflected in the differences in the magnitudes of peaks of the distributions of the metrics between brain and testis and other tissues (Additional Figures 4.4-7). However, we also observed other types of differences in comparisons of the distributions of these metrics between brain and testis and other tissues: while subtle, some metrics (Tajima's D, F_{ST}; Additional Figures 4.4-5) showed shifts of the distributions of brain and testis enhancers towards weaker signatures of recent positive selection compared to other tissues, suggesting that the significant pairwise differences we see are not solely caused by disparities in the number of enhancers within tissues. In addition, there are several other tissues (e.g., lung, spleen, blood) whose numbers of enhancers are nearly on par with those of the brain and more than the testis; however, these tissues did not show significant differences in the distribution of these metrics compared to other tissues, further suggesting that the observed differences in selection are likely biological.

Why are patterns of selection different for brain and testis enhancers? Answering these questions is challenging without additional functional experiments that shed light on the phenotypic effects of the selected variants. In the case of the brain, existing evidence suggests that the brain size of modern humans has not changed substantially in the last 250,000 - 300,000 years (Neubauer et al. 2018). Similarly, expression patterns of genes expressed in neural tissues (e.g., brain, cerebellum) show low levels of divergence across both species and within humans, suggesting that the overall structure of the neural network is highly conserved (Khaitovich 2005; Brawand et al. 2011). In contrast, brain shape has gradually changed in modern humans, reaching the present-day variation human variation about 100,000 - 35,000 years ago, but whether this has anything to do with positive selection being relatively weaker in the brain compared to other tissues is pure speculation.

In the case of testis, our results suggest that enhancers active in the testis often show weaker signatures of positive selection compared to enhancers in other tissues. Previous studies have shown that testis exhibits the highest degree of expression pattern divergence between species, likely due to either strong action of positive selection on reproductive processes (Khaitovich 2005; Brawand et al. 2011), relaxation of purifying selection in the testis compared to other tissues (Gershoni and Pietrokovski 2014), or both. Unfortunately, our results are not directly comparable with studies measuring gene expression divergence between species due to the difference in the features being compared and the temporal window examined; previous studies have looked at the divergence of the overall transcriptome, which is influenced not just by changes in enhancer sequences but by other types of changes and other regulatory elements as well, whereas we have specifically looked at changes in allele frequency and haplotypes in enhancer regions. In addition, previous studies have looked at divergence of the expression patterns between species that diverged several millions of years ago, whereas we have examined selection events that have occurred in more recent evolutionary times, on the order of hundreds or tens of thousands of years. Therefore, it is possible that there has been temporal variation (i.e., ancient vs. recent) in the occurrence of selection events on these enhancer regions. It is worth noting that Khaitovich et al. (Khaitovich et al. 2005) also found that testis had the most significant reductions in diversity of expression (i.e., low inter-individual variation in expression patterns) compared to other tissues examined, which points to differences in the action of selection over various temporal windows within the testis. In short, differences between our results and those of previous studies can be explained by differences in actions of ancient versus recent selection.

Acknowledgments

This research was supported by the March of Dimes through the March of Dimes Prematurity Research Center Ohio Collaborative, the Burroughs Wellcome Fund, the Guggenheim Foundation, and the National Institutes of Health (R01GM115836, R35GM127087). This work was conducted, in part, using the Advanced Computing Center for Research and Education at Vanderbilt University.

References

Alexa A, Rahnenfuhrer J. (2016). Gene set enrichment analysis with topGO, R package version 2.30.1

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, et al. 2014. An atlas of active enhancers across human cell types and tissues. *Nature* 507: 455–461.

Balaresque PL, Ballereau SJ, Jobling MA. 2007. Challenges in human genetic diversity: demographic history and adaptation. *Human Molecular Genetics* 16: R134–R139.

Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, et al. 2011. The evolution of gene expression levels in mammalian organs. *Nature* 478: 343–348.

Capra JA, Erwin GD, McKinsey G, Rubenstein JLR, Pollard KS. 2013. Many human accelerated regions are developmental enhancers. *Philosophical Transactions of the Royal Society B: Biological Sciences* 368: 20130025–20130025.

Carroll SB. 2005. Evolution at Two Levels: On Genes and Form. *PLoS Biol* 3: e245.

Casper J, Zweig AS, Villarreal C, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, Lee CM, Lee BT, Karolchik D, et al. 2018. The UCSC Genome Browser database: 2018 update. *Nucleic Acids Research* 46: D762–D769.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaSci* 4: 559.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. *Bioinformatics* 27: 2156–2158.

Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R. 2014. On Detecting Incomplete Soft or Hard Selective Sweeps Using Haplotype Structure. *Molecular Biology and Evolution* 31: 1275–1291.

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449: 851–861.

Fumagalli M, Sironi M, Pozzoli U, Ferrer-Admettla A, Pattini L, Nielsen R. 2011. Signatures of Environmental Genetic Adaptation Pinpoint Pathogens as the Main Selective Pressure through Human Evolution ed. J.M. Akey. *PLoS Genet* 7: e1002355.
Garud NR, Messer PW, Buzbas EO, Petrov DA. 2015. Recent Selective Sweeps in North American Drosophila melanogaster Show Signatures of Soft Sweeps ed. G.P. Copenhaver. *PLoS Genet* 11: e1005004.

Gershoni M, Pietrokovski S. 2017. Reduced selection and accumulation of deleterious mutations in genes exclusively expressed in men. *Nature Communications* 5: 4438.

Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, 1000 Genomes Project, Bustamante CD. 2011. Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences* 108: 11983–11988.

Gu X, Su Z. 2007. Tissue-driven hypothesis of genomic evolution and sequence-expression correlations. *Proc Natl Acad Sci USA* 104: 2779–2784.

Haller BC, Messer PW. 2017. SLiM 2: Flexible, Interactive Forward Genetic Simulations. *Molecular Biology and Evolution* 34: 230–240.

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

Karlsson EK, Kwiatkowski DP, Sabeti PC. 2014. Natural selection and infectious disease in human populations. *Nature Reviews Genetics* 15: 379–393.

Karnuta JM, Scacheri PC. 2018. Enhancers: bridging the gap between gene control and human disease. *Human Molecular Genetics* 27: R219–R227.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Research* 12: 996–1006.

Khaitovich P. 2005. Parallel Patterns of Evolution in the Genomes and Transcriptomes of Humans and Chimpanzees. *Science* 309: 1850–1854.

King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. *Science* 188: 107–116.

Kryuchkova-Mostacci N, Robinson-Rechavi M. 2015. Tissue-Specific Evolution of Protein Coding Genes in Human and Mouse ed. M. Anisimova. *PLoS ONE* 10: e0131673.

Kudaravalli S, Veyrieras JB, Stranger BE, Dermitzakis ET, Pritchard JK. 2008. Gene Expression Levels Are a Target of Recent Natural Selection in the Human Genome. *Molecular Biology and Evolution* 26: 649–658.

Lee TI, Young RA. 2013. Transcriptional Regulation and Its Misregulation in Disease. *Cell* 152: 1237–1251.

Long HK, Prescott SL, Wysocka J. 2016. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell* 167: 1170–1187.

Lynch VJ, Nnamani MC, Kapusta A, Brayer K, Plaza SL, Mazur EC, Emera D, Sheikh SZ, Grützner F, Bauersachs S, et al. 2015. Ancient Transposable Elements Transformed the Uterine Regulatory Landscape and Transcriptome during the Evolution of Mammalian Pregnancy. *CellReports* 10: 551–561.

MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, et al. 2017. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research* 45: D896–D901.

Messer PW. 2013. SLiM: Simulating Evolution with Selection and Linkage. *Genetics* 194: 1037–1039.

Messer PW, Petrov DA. 2013. Population genomics of rapidadaptation by soft selective sweeps. *Trends in Ecology & Evolution* 1–11.

Moon JM, Aronoff DM, Capra JA, Abbot P, Rokas A. 2018. Examination of Signatures of Recent Positive Selection on Genes Involved in Human Sialic Acid Biology. *G3* g3.200035.2018.

Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, Rynes E, Maurano MT, Vierstra J, Thomas S, et al. 2012. BEDOPS: high-performance genomic feature operations. *Bioinformatics* 28: 1919–1920.

Neubauer S, Hublin J-J, Gunz P. 2018. The evolution of modern human brain shape. *Sci Adv* 4: eaao5961.

Nédélec Y, Sanz J, Baharian G, Szpiech ZA, Pacis A, Dumaine A, Grenier J-C, Freiman A, Sams AJ, Hebert S, et al. 2016. Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens. *Cell* 167: 657–664.e21.

Nielsen R, Akey JM, Jakobsson M, Pritchard JK, Tishkoff S, Willerslev E. Tracing the peopling of the world through genomics. *Nature* 541: 302-310.

Ong C-T, Corces VG. 2011. Enhancer function: new insights into the regulation of tissue-specificgene expression. *Nature Reviews Genetics* 12: 284–293.

Page AE, Viguier S, Dyble M, Smith D, Chaudhary N, Salali GD, Thompson J, Vinicius L, Mace R, Migliano AB. 2016. Reproductive trade-offs in extant hunter-gatherers suggest adaptive mechanism for the Neolithic expansion. *Proceedings of the National Academy of Sciences* 113: 4694–4699.

Pennacchio LA, Bickmore W, Dean A, Nobrega MA, Bejerano G. 2013. Enhancers: five essential questions. *Nature Reviews Genetics* 14: 288–295.

Pennings PS, Hermisson J. 2006. Soft Sweeps II—Molecular Population Genetics of Adaptation from Recurrent Mutation or Migration. *Molecular Biology and Evolution* 23: 1076–1084.

Pfeifer B, Wittelsbürger U, Ramos-Onsins SE, Lercher MJ. 2014. PopGenome: An Efficient Swiss Army Knife for Population Genomic Analyses in R. *Molecular Biology and Evolution* 31: 1929–1936.

Raney BJ, Dreszer TR, Barber GP, Clawson H, Fujita PA, Wang T, Nguyen N, Paten B, Zweig AS, Karolchik D, et al. 2014. Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics* 30: 1003–1005.

Rebeiz M, Tsiantis M. 2017. Enhancer evolution and the origins of morphological novelty. *Current Opinion in Genetics & Development* 45: 115–123.

Rockman MV, Hahn MW, Soranzo N, Zimprich F, Goldstein DB, Wray GA. 2005. Ancient and Recent Positive Selection Transformed Opioid cis-Regulation in Humans. *PLoS Biol* 3: e387.

Rubinstein M, de Souza FSJ. 2013. Evolution of transcriptional enhancers and animal diversity. *Philosophical Transactions of the Royal Society B: Biological Sciences* 368: 20130017–20130017.

Sabeti, P. C., Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES. 2006. Positive natural selection in the human lineage. *Science* 312: 1614–1620.

Schaffner SF. 2004. The X chromosome in population genetics. *Nature Reviews Genetics* 5: 43–51.

Sholtis SJ, Noonan JP. 2010. Gene regulation and the origins of human biological uniqueness. *Trends in Genetics* 26: 110–118.

Simonti CN, Pavličev M, Capra JA. 2017. Transposable Element Exaptation into Regulatory Regions Is Rare, Influenced by Evolutionary Age, and Subject to Pleiotropic Constraints. *Molecular Biology and Evolution* 34: 2856–2869.

Szpiech ZA, Hernandez RD. 2014. selscan: An Efficient Multithreaded Program to Perform EHH-Based Scans for Positive Selection. *Molecular Biology and Evolution* 31: 2824–2827.

The 1000 Genomes Project Consortium, Gibbs RA, Boerwinkle E, Doddapaneni H, Han Y, Korchina V, Kovar C, Lee S, Muzny D, Reid JG, et al. 2015. A global reference for human genetic variation. *Nature* 526: 68–74.

The FANTOM Consortium and the RIKEN PMI and CLST (DGT). 2014. A promoter-level mammalian expression atlas. *Nature* 507: 462–470.

Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M, et al. 2006. Convergent adaptation of human lactase persistence in Africa and Europe. *Nature Genetics* 39: 31–40.

Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjostedt E, Asplund A, et al. 2015. Tissue-based map of the human proteome. *Science* 347: 1260419–1260419.

Varki A, Gagneux P. 2009. Human-specific evolution of sialic acid targets: Explaining the malignant malaria mystery? *Proc Natl Acad Sci USA* 106: 14739.

Vitti JJ, Grossman SR, Sabeti PC. 2013. Detecting Natural Selection in Genomic Data. *Annu Rev Genet* 47: 97–120.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A Map of Recent Positive Selection in the Human Genome ed. L. Hurst. *PLoS Biol* 4: e72.

Wang JZ, Du Z, Payattakool R, Yu PS, Chen CF. 2007. A new method to measure the semantic similarity of GO terms. *Bioinformatics* 23: 1274–1281.

Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. *Nature Reviews Genetics* 8: 206–216.

Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S. 2010. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* 26: 976–978.

CHAPTER V

Conclusion

Recent evolution of sialic acid biology genes in modern humans

In Chapter III, I examined the patterns of recent evolution of a set of genes that are involved in sialic acid biology. I used several metrics that are sensitive to distinct types of genomic signatures resulting from recent positive selection. In addition, I carried out simulations of neutral evolution that incorporated pre-estimated parameters of past demographic events to assess the probability that the observed metrics resulted from actions of selection. In short, I showed that the majority of sialic acid biology genes do not exhibit evidence of recent positive selection across all time ranges examined. Furthermore, even when I calculated H12, a novel metric shown to exhibit increased sensitivity towards incidences of soft selective sweeps (Garud *et al.* 2015), the majority of the genes did not show significant deviations from neutral expectations.

It has been widely hypothesized that the sialic acid biology pathway has evolved via rapid evolution due to strong selective pressures from pathogens (Varki 2002; Angata and Varki 2002; Varki and Gagneux 2009; Hayakawa and Varki 2011; Wang *et al.* 2012) and previous work has focused on ancient selection events (Altheide *et al.* 2006) and found evidence of positive selection in response to paleopathogens. The contrasting results of my study with the prevailing hypothesis and past findings suggest the action of temporally varying (i.e., ancient versus recent) selection on the pathway of sialic acid biology: more specifically, it is possible that sialic acid biology genes experienced adaptation in response to ancient but not more recent pathogens. More broadly, my study shows that the same gene (or set of genes) could experience temporally varying selection events and emphasize that selective pressures are not necessarily consistent across time.

Despite the overall lack of significant evidence of recent positive selection in sialic acid biology genes, this approach could still serve as a first step towards understanding the evolution of genetic elements that are associated with human pregnancy. Given the importance of refined regulation of the maternal immune system during pregnancy, it is possible that the patterns of recent evolution of immunity-related genes reflect the action of pregnancy-related selective pressures as well. In this regard, a recent study identified a set of changes in immune system that occur over the course of a term pregnancy (Aghaeepour *et al.* 2017): these include endogenous pSTAT5ab signaling in naive CD4+ T cells, pSTAT5ab response to IL in neutrophils, and pSTAT1 response to IFN-α in NK cells. Genes involved in these immune system alternations could serve as potential candidates for studies of recent positive selection of genetic elements associated with pregnancy.

Recent evolution of enhancers across 41 human tissues

In Chapter IV, I studied the recent evolution of enhancers active in 41 human tissues, which include reproductive tissues such as placenta, female gonad, and testis, as well as somatic tissues such as brain, heart, and liver. As in Chapter III, I used several metrics that can detect distinct genomic signatures of recent positive selection. These metrics were then compared to those calculated on sequences created by simulations of neutral evolution that incorporated pre-estimated parameters of past demographic events to determine the likelihood of recent positive selection. First, I found that on average, approximately 5.90% of enhancers exhibit evidence of significant recent positive selection across all metrics and tissues. Second, I found that the putative target genes of enhancers with significant deviations from neutral expectations according to the nS_L metric exhibit enrichment for immunity-related functions. In addition, this pattern of functional enrichment was shared across tissues, and thus represents a more global trend. Next, I found that the patterns of recent evolution differed between enhancers active in brain and testis and several other tissues: while this difference could be due to non-adaptive processes (as majority of the enhancers exhibit non-significant deviations from neutral expectations), one possibility is differential actions of positive selection on enhancers active in the brain and testis and other tissues.

One possible explanation for the enrichment for immunity functions among putative target genes of enhancers showing evidence of recent positive selection according to nS_L is the incidence of the Agricultural Revolution, which occurred approximately 10,000 to 20,000 years ago. During this period, the advent of farming practices, communal living in settlements, and migrations of farmers across the globe resulted in an increase of numbers and densities of humans in any given location, likely facilitating the spread of pathogens (Varki and Gagneux 2009; Pool *et al.* 2010; Nielsen *et al.*). Such an increase in pathogen burden would have likely acted as a strong selective pressure on immunity genes, as well as their *cis*-regulatory regions. Another selective pressure that could have acted simultaneously on these enhancers is the need to regulate immune responses during pregnancy to prevent inappropriate and excessive activation of detrimental immune responses. Therefore, it is possible that around 10,000 - 20,000 years ago, enhancers regulating the activities of immunity-related genes underwent selection to allow refined fine-tuning of immune activities to enable effective host defenses against the increased threat from pathogens resulting from increased human population sizes as well as, establishment and maintenance of successful pregnancy.

As more research suggests the importance of *cis*-regulatory regions in driving local adaptations of modern humans (Wray 2007; Nédélec *et al.* 2016) to different environments, enhancers could be incorporated into future studies of recent evolution of genetic elements involved in human pregnancy. For instance, a recent study successfully predicted a large number of placental enhancers using a machine learning approach (Zhang *et al.* 2018): these enhancers were likely to be near genes associated with placental development and birth disorders, validating their roles in human pregnancy.

Finally, additional functional information of enhancers – for instance, the gestational period at which such enhancers are active – could provide additional interesting insights.

Future direction: the need for quantitative assessment of the relative contributions of various selective pressures

I have described in Chapter II how various population genetics metrics can be used to infer past selection events that occurred in recent human history. In addition, programs that can execute sophisticated simulations of neutral evolution by incorporating pre-estimated demographic parameters can greatly reduce false positives. However, the metric of selection by itself cannot provide information regarding the specific selective pressures that have acted on the region of interest: the metric summarizes the simultaneous actions of multiple evolutionary forces, including neutral and random processes as well as several selective forces. Furthermore, significant deviation from neutral expectations indicates that the region of interest has likely experienced selection, but again, cannot identify the exact causative selective agent. In this regard, one caution in interpreting the results of my studies in Chapters III and IV is not to definitively conclude that the patterns of genetic variation have resulted from selective pressures associated with pregnancy, host defenses, or any single selective pressure.

In this regard, a previous study calculated the relative contributions of various environmental variables related to climatic and dietary factors, as well as pathogen loads within each population to the allele frequency spectrum of variants and determined that once corrected for demography, pathogen accounted for most of the genetic variance among populations (Fumagalli *et al.* 2011). However, quantifying the relative contribution of selection pressures associated with reproductive processes on the observed patterns of genetic variation is even more challenging, especially because of 1) the uncertainty associated with defining pregnancy as an environmental variable and 2) the lack of in-depth phenotypic data that spans multiple populations. Specific aspects of pregnancy (e.g., average birth weight, gestational length, incidences of pregnancy disorders within a population) could serve as proxies for selective pressures that pregnancy imposes on genetic regions of interest within modern humans. As in Fumagali and his colleagues' work (Fumagalli *et al.* 2011), these variables could be incorporated into the matrix of environmental variables to calculate the relative contribution of selective pressures associated with pregnancy to the observed metrics of recent selection.

Summary

Collectively, my dissertation research has investigated the patterns of recent evolution of genetic elements involved in pregnancy within modern humans. More specifically, I tested the hypothesis that sialic acid biology genes (Chapter III) and enhancers expressed in different tissues (Chapter IV) have experienced recent positive selection in modern humans. My studies show the utility of incorporating different metrics that can capture distinct genomic signatures resulting from selection events that occurred at different time ranges (and under different modes of selection), allowing a more comprehensive investigation of the action of recent selection in a region of interest (Chapter II). The discovery of new genes or pathways that are associated with term pregnancy and pregnancy disorders (Aghaeepour *et al.* 2017; Zhang *et al.* 2017) could provide interesting candidate regions for future studies. In addition, new metrics with increased sensitivity to novel genomic signatures could detect selection events that have been previously been missed due to lack of power (Field *et al.* 2016). Finally, quantitatively comparing the relative contributions of selective pressures stemming from host defense and pregnancy could shed light into how trade-offs between immunity and reproduction have shaped the patterns of genetic variation in modern humans.

References

Aghaeepour, N., E. A. Ganio, D. Mcilwain, A. S. Tsai, M. Tingle *et al.*, 2017 An immune clock of human pregnancy. Sci. Immunol. 2: eaan2946.

Altheide, T. K., T. Hayakawa, T. S. Mikkelsen, S. Diaz, N. Varki *et al.*, 2006 System-wide Genomic and Biochemical Comparisons of Sialic Acid Biology Among Primates and Rodents. J. Biol. Chem. 281: 25689–25702.

Angata, T., and A. Varki, 2002 Chemical Diversity in the Sialic Acids and Related α-Keto Acids: An Evolutionary Perspective. Chem. Rev. 102: 439–470.

Field, Y., E. A. Boyle, N. Telis, Z. Gao, K. J. Gaulton *et al.*, 2016 Detection of human adaptation during the past 2000 years. Science 354: 760.

Fumagalli, M., M. Sironi, U. Pozzoli, A. Ferrer-Admettla, L. Pattini *et al.*, 2011 Signatures of Environmental Genetic Adaptation Pinpoint Pathogens as the Main Selective Pressure through Human Evolution. PLoS Genet 7: e1002355.

Garud, N. R., P. W. Messer, E. O. Buzbas, and D. A. Petrov, 2015 Recent Selective Sweeps in North American Drosophila melanogaster Show Signatures of Soft Sweeps. PLoS Genet 11: e1005004.

Hayakawa, T., and A. Varki, 2011 Human-Specific Changes in Sialic Acid Biology, pp.123-148 in *Post-Genome Biology of Primates*, edited by H. Hirai, H. Imai, and Y. Go, Springer.

Nédélec, Y., J. Sanz, G. Baharian, Z. A. Szpiech, A. Pacis *et al.*, 2016 Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens. Cell 167: 657–664.e21.

Nielsen, R., J. M. Akey, M. Jakobsson, J. K. Pritchard, S. Tishkoff *et al.* Tracing the peopling of the world through genomics. Nature 541: 302-310.

Pool, J. E., I. Hellmann, J. D. Jensen, and R. Nielsen, 2010 Population genetic inference from genomic sequence variation. Genome Research 20: 291–300.

Varki, A., 2002 Loss of N-glycolylneuraminic acid in humans: Mechanisms, consequences, and implications for hominid evolution. Am. J. Phys. Anthropol. 116: 54–69.

Varki, A., and P. Gagneux, 2009 Human-specific evolution of sialic acid targets: Explaining the malignant malaria mystery? Proc Natl Acad Sci USA 106: 14739.

Wang, X., N. Mitra, I. Secundino, K. Banda, P. Cruz *et al.*, 2012 Specific inactivation of two immunomodulatory SIGLEC genes during human evolution. Proc Natl Acad Sci USA 109: 9935.

Wray, G. A., 2007 The evolutionary significance of cis-regulatory mutations. Nature Reviews Genetics 8: 206–216.

Zhang, G., B. Feenstra, J. Bacelis, X. Liu, L. M. Muglia *et al.*, 2017 Genetic Associations with Gestational Duration and Spontaneous Preterm Birth. N Engl J Med 377: 1156–1167.

Zhang, J., C. N. Simonti, and J. A. Capra, 2018 Genome-wide maps of distal gene regulatory enhancers active in the human placenta. PLoS ONE 13: e0209611.

APPENDIX

All additional files are available at https://figshare.com/s/241fb6cd6b6be6a2ec9a