Using human genomics to decipher biological mechanisms governing reproductive ageing and fertility in women

Stasa Stankovic

MRC Epidemiology
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Clare Hall

September 2022

Declaration

The work presented in this dissertation was carried out at the MRC Epidemiology Unit, University of Cambridge, and Department of Cellular and Molecular Medicine, University of Copenhagen, between October 2019 and September 2022. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the 'Contributions and Collaborations' sections at the beginning of each chapter or as specified in the text.

No part of this thesis is substantially similar to any work that I have submitted, or, is currently being submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution.

It does not exceed the limit of 60,000 words set out by the Degree Committee of the faculties of Clinical Medicine and Veterinary Medicine.

Stasa Stankovic

September 2022

Abstract: **Using human genomics to decipher biological mechanisms governing reproductive ageing and fertility in women**

Stasa Stankovic

Women are born with a non-renewable ovarian reserve, which is depleted throughout reproductive life. When this reserve is exhausted, they experience menopause and cease ovulating. Importantly, menopause timing is highly variable and can impact health outcomes in later life. One in 100 women experience menopause before the age of 40. As natural fertility begins to decline 10 years prior to menopause, the age of menopause impacts reproductive options for many women, leading to increased demand for fertility treatments, which have low success rate. This is especially important as more women delay childbearing. Endocrine and imaging tests used in the clinical setting only record changes in ovarian function that have already taken place, thus disabling early prediction and timely identification of women with reduced reproductive lifespan. Human genetic studies have attempted to overcome this problem by identifying genetic markers associated with menopause timing and thus providing substantial insight into the biological mechanisms governing ovarian ageing. However, previous approaches have been largely restricted to assessing common genetic variation, leaving many aspects of the trait biology unexplored. This dissertation describes five distinct projects that advance our understanding of the genetic determinants of female reproductive ageing by employing state-of-art genomic and proteomic technologies with robust functional models.

Chapter 3 uses whole exome sequence data to identify rare protein-coding variants associated with menopause timing in ~120K women in the UK Biobank (UKBB), and implicates five novel ANM genes with effect sizes up to ~5 times larger than previously discovered for common variants. Notably, heterozygous loss of *ZNF518A* shortens reproductive lifespan by delaying puberty timing in girls and reducing ANM by nearly 6 years in carriers, an effect larger than any variation currently tested in clinical genetics for premature ovarian ageing. Furthermore, I provide evidence that *ZNF518A* is a master transcriptional regulator of ovarian development and establishment of the ovarian reserve in foetal life, thus highlighting novel mechanisms involved in ANM aetiology. I also identify a new cancer predisposition gene, *SAMHD1*, which has a comparable effect size in women and men to well-established genes such as *CHEK2*, further reinforcing the link between cancer and reproductive ageing. Finally, I show that mothers with genetic susceptibility to earlier ovarian ageing have a higher rate of *de novo* mutations in their offspring. This provides direct evidence that female germline mutation rate is heritable and highlights a mechanism for maternal effects on offspring health. **Chapter 4** extends the exome sequence analysis to an extreme form of early menopause, i.e. POI, which is often considered a monogenic disorder, with pathogenic mutations reported in ~100 genes. However, such reports are based

on small numbers of individuals without independent replication, or/and no functional validation. I systematically evaluate the penetrance of these reported genes in ~120K UKBB women, 2,231 of whom reported ANM before age 40. In this largest study of POI to date, I find limited evidence to support any previously reported autosomal dominant gene. For nearly all these genes I could rule out even modest penetrance, with 97.8% of all identified protein truncating variants found in reproductively healthy women with ANM over 40. In addition, I demonstrate novel haploinsufficiency effects in studied POI genes, including *TWNK* and *SOHLH2*. Collectively my results suggest that most POI cases are likely oligogenic or polygenic in nature, which has major implications for future clinical genetic testing and counselling. **Chapter 5** presents the first proteogenomic study for the ANM targeting 4,775 distinct proteins measured from plasma samples of 10,713 European ancestry individuals in the Fenland study. Although this analysis did not identify robust protein candidates associated with ANM, it demonstrates the potential of such approaches to discover new biomarkers. **Chapter 6** presents the largest genomic meta-analysis for age at menarche on ~566,000 women of European ancestry and 696 genomic loci that contribute to regulation of menarche timing. I use this data to explore biological mechanisms and overlap between genetic architectures of reproductive health outcomes. I provide the first evidence on the enrichment of DDR mechanisms for menarche timing, indicating the involvement of DDR in regulation of both extremes of reproductive lifespan, i.e. menarche and menopause. In addition, I report first gene candidates that I speculate may act via oocyte-specific mechanisms to modify reproductive longevity. I also highlight DDR and other novel mechanisms, including ribosome biogenesis, which impact multiple reproductive health outcomes, such as polycystic ovarian syndrome (PCOS), twinning and number of children (NEB). Finally, I demonstrate the first population genomic evidence on the role of DDR related mechanisms in various anthropometric, metabolic and reproductive health outcomes, indicating that DDR could act as a marker of health outcomes beyond cancer. Combining human genomic evidence with cutting edge CRISPR technology and the *In vitro* gametogenesis system, in **Chapter 7** I investigate the role of *PARP-1* in proliferation of primordial germ cells during the establishment of the ovarian reserve. I demonstrate suggestive evidence on the role of *PARP-1* in decreasing ANM in women and, paradoxically, that deletion of *PARP-1* increases the efficiency of primordial germ cell production *in vitro*. I speculate that, despite the initial increase in primordial germ cells in the *PARP-1* knockout, the quality of these cells could be compromised, thus ultimately limiting the functional ovarian pool. Collectively, these findings provide significant insights into the biological processes of reproductive ageing in women and have the potential to guide future experimental work aimed towards identification of new therapies for enhancing reproductive function and preserving fertility in women, as well as designing intervention strategies to prevent or diminish menopause-related health outcomes.

Acknowledgments

Most of life-changing experiences in my life happened as a collision of a magic of chance, my endless and sometimes ‘childish’ hunger for knowledge and, most importantly, luck to be surrounded by incredible people who tirelessly support and teach me, and who make, all in their own way, every moment the most exciting life ride! I am blessed to have you and I dedicate my work to all of you.

This PhD story started in the same way, over a cup of coffee during the networking event at the University of Cambridge, where I was, at the time, doing my MPhil degree. There, I met incredibly talented Victoria Young, a PhD student at the MRC Epidemiology Unit, to whom I was trying to ‘sell’ my viral passion for reproductive medicine while subconsciously waiting for any ‘sign’ that would help me decide where to conduct my PhD studies. ’Oh, you gotta meet John and Ken, they are amazing guys!’ - Vicky did not know that with this casual reply she will shape my next three (and definitely most exciting) years of life.

John and Ken are indeed amazing!
I will forever be grateful for the opportunity to conduct such cutting-edge research in your group. Thank you for your faith in me throughout these years. I felt supported and guided, while at the same time having freedom to shape my PhD journey the way my passion for science ‘desired’. You have been fascinating mentors. It is interesting how you managed to create this magic atmosphere in our group - a mix of intelligently provoking humour, dedication and love towards good science, and last but definitely not the least a family-like feeling that enabled me to feel like ‘home’. You taught me the importance of critical thinking and statistical rigour and you helped me grow immensely as a research scientist. I am beyond grateful for that.

Being an international student in the UK seeks creative solutions, a lot of hard work, tireless will to succeed backed up by strong knowledge and a vision to convince people that you are the right person to get available opportunities. This whole work would not be possible without the support of Clare Hall Ivan D. Jankovic scholarship and Mr Jankovic, a visionary himself. Thank you for your generous support and for giving me a chance to live this ‘life-changing’ experience. I owe so much to you, and I wish that one day I will be able to give back to the Serbian community and support young talents to live their dreams. You are part of every scientific discovery we make.

Additionally, I am grateful for all the generous support I received from Clare Hall. So many incredible experiences would have not been possible without your input, both emotional and financial - you helped my dreams grow even bigger. I will miss all of Clare Hall's stuff! Clare Hall was a home for three years, both a safe and inspiring zone where I met the most incredible people, who I can today call family. Many thanks to all other sponsors who enabled me to share my scientific and entrepreneurial achievements internationally and recognized the importance of this work for the future of women's health – McKinsey&Company for their award 'Next Generation Women Leader 2020', One Young World for 'Leading Europe 2022' award, the UK Government for the 'STEM for Britain', SRI Reproduction and many others.

I would also like to give a heartfelt, special thanks to Prof Eva Hoffman, who generously opened the door to her laboratory to me and in a short time-frame equipped me with incredible knowledge and skills. Eva, you are a role model and it was inspiring having you by my side. Your passion for science is contagious and combined with your beautiful personality; you really left an important mark on me as a woman in science. Thank you for giving me the best mentors I could ask for - special thanks go to Jason Halliwell. Jason, thank you for selflessly sharing your knowledge with me, being patient with me, making Copenhagen feel like home! You are incredible! I hope that we will continue collaborating in future and continue creating magic with science. Many thanks to Hannah Schorle and Amy Kaucher from whom I learnt so much!

This work would not be possible without collaborators. I feel grateful to have Prof Anna Murray as a mentor! Thank you for the opportunity to work with you, in an exciting and fruitful collaboration. It has been an absolute privilege. Thank you for your guidance, infectious passion for biology, for always putting our work in a bigger and fulfilling context and your amazing humour! This has been a great ride!

Finally, thanks to all other members of our MRC Epidemiology group - you showed me the power of the team work! The work presented in this thesis would not have been possible without many you! I owe my deepest gratitude to Dr Felix Day - Felix thank you for being the best support I could have asked for, a friend and a mentor – your guidance was crucial to my academic success! Many thanks to Dr Eugene Gardner who stepped up my 'science game' in every way. Eugene, I learnt so much from you, thank you for opening new knowledge horizons for me, and for patiently listening (and provoking) to move me away from my comfort zone. I have been tremendously fortunate to have Dr Katherine Kentistou during my PhD journey - thank you for smooth collaborations, friendly talks and advice!

My time in Cambridge was definitely more special because I had Dr Jason Melland and Dr Blagoje Soskic on my side. Thank you for being true friends, for so much love and support.

Most importantly, I thank my family, who has provided unwavering support and limitless pride. I dedicate this piece of work to my best friends...my mum and dad. Thank you for being my zone of comfort, but most importantly, thank you for letting me out of it and being the wind behind my back whenever I needed to spread my wings and be fearless.

I am eternally thankful to God for giving me many challenges and beautiful opportunities, which from day to day build my self-confidence and faith to dream big.

Publications & Awards

General Awards

Ivan D Jankovic scholarship: three years PhD funding; selected as the only recipient in the competitive process within the 4 years timeframe.

STEM for Britain: One of ten finalists in the Biological and Biomedical Sciences category across the UK, where I presented my research to Members of UK Parliament, House of Commons.

McKinsey & Company Next Generation Women Leader 2020: Selected as the most influential young female leader around the globe, recognised based on the remarkable impact through entrepreneurial projects and outstanding academic achievements. I was nominated for 2020 together with 34 women among 6,000 applicants across Europe and The Middle East. I received financial support (£3,000) and mentorship to run my entrepreneurial project.

Top 100 Women in Tech: selected as 1 of 100 top women in tech by Goldman Sachs for my contribution to science and entrepreneurship.

Leader of Tomorrow - GAP Summit 2020: selected as 1 of 100 Leaders of Tomorrow from around the world to attend a biotechnology leadership summit to discuss pressing challenges facing the bio-economy and catalysing innovation to solve these challenges through building ventures.

One Young World Leader 2021 and 2022 (United Nations): received a "Leading Europe" scholarship (£3,500) as a delegate of Serbia to attend the One Young World Summit, which gathers the world's most impactful young leaders with influential political, business and humanitarian leaders to accelerate social impact.

Chapter 1

Publications

Stankovic, S., Murray, A., Hoffmann, E., and Perry, J.R. B. **Reproductive ageing through human genomics lens**. (2022). Trends in Genetics. *Invited Review (in preparation).*

Ruth, K.S., …[29 authors]... , **Stankovic, S**…[244 authors], and Perry, J.R.B. **Genetic insights into biological mechanisms governing human ovarian ageing**. Nature 596, 393–397 (2021). https://doi.org/10.1038/s41586-021-03779-7.

Chapter 3

Publications

Stankovic, S., Shekari, S., Huang, Q., Gardner, E. *et al*. Genetic susceptibility to earlier ovarian ageing increases de novo mutation rate in offspring. (2022). *Under review in Nature*.

Conferences

World Menopause Conference 2022: accepted talk (won the **Robert Greenblatt Award**)
Award: Clare Hall Boak Fund (£500) to support travelling and accommodation costs.
Cambridge Reproduction Early Researchers Seminar: invited talk [January, 2021]
FemTechnology Summit: **'The Future of Fertility'** (invited panel speaker) [June, 2022]
University of Cambridge SRI **Summer School of Reproduction** (invited lecturer) [October, 2022]

Chapter 4

Publications

Shekari S.*, **Stankovic, S**.*, *et al*. Genomic analyses in 104,733 UK women show that previously reported monogenic genes are not common causes of Premature Ovarian Insufficiency. (2022). *Submitted to Nature Medicine*.

Conferences

European Society of Human Reproduction and Embryology "ESHRE" 2022: accepted talk (given by Shekari, S.)
Award: MRC Epidemiology Fund (1,000£) to support travelling and accommodation costs.

Chapter 6

Publications

Two papers in preparation.

Chapter 7

Publications

In preparation.

Awards

The following Awards were received to support my collaborative visit to the University of Copenhagen, where I spent 1 month and 15 days working in the laboratory of Prof Eva Hoffmann.

Clare Hall Boak Fund: double recipient of £1,000
Clare Hall Tutor support (£500)
Cambridge SRI Reproduction: 'Development Fund' (£500)
MRC Epidemiology Unit (£500)

Arising elsewhere

Zhao, Y., **Stankovic, S**., Koprulu, M. *et al*. GIGYF1 loss of function is associated with clonal mosaicism and adverse metabolic health. Nature Communications 12, 4178 (2021). https://doi.org/10.1038/s41467-021-24504-y.

Gardner, E., Kentistou, K., **Stankovic, S**., *et al*. Damaging missense variants in IGF1R implicate a role for IGF-1 resistance in the aetiology of type 2 diabetes. Accepted by Cell Genomics (2022). https://doi.org10.1101/2022.03.26.22272972.

Iain Mathieson, Felix R. Day, … , Melinda C. Mills, John R.B. Perry, on behalf of the Human Reproductive Behaviour Consortium (**Stankovic, S.**). Genome-wide analysis identifies genetic effects on reproductive success and ongoing natural selection at the FADS locus. Accepted by Nature Human Behaviour (2022). doi: https://doi.org/10.1101/2020.05.19.104455

Song, J.*, Li, J.*, … , **Stankovic, S**. *et al*. Evolutionarily conserved FOXO target OSER1 extends lifespan and healthspan. *In preparation*.

Commonly Used Abbreviations

100 kGP . . . 100,000 Genomes Project

AB . . . . . Allele balance

ABC . . . . . Activity-by-contact

AD . . . . . Autosomal dominant

AFC . . . . . Antral follicle count

AMH . . . . . Anti-Müllerian hormone

ANM . . . . . Age at natural menopause

AR . . . . . Autosomal recessive

BCAC . . . . Breast Cancer Association Consortium

BER . . . . . Base excision repair

BMD . . . . . Bone mineral density

BMI . . . . . Body mass index

BVSC . . . . . Blimp1-mVenus & Stella-ECFP

CAD . . . . . Coronary artery disease

CADD . . . . Combined annotation dependent depletion

CHH . . . . . Congenital Hypogonadotropic Hypogonadism

CHR . . . . . Chromosome

CI . . . . . Confidence interval

cMAGMA .. Coding MAGMA

DBP . . . . . Diastolic blood pressure

DDR . . . . . DNA damage response

DNM . . . . . De novo mutation

DSB . . . . . Double strand break

EA . . . . . Effect allele

EAF . . . . . Effect allele frequency

IV . . . . . Instrumental variable

IVF . . . . . In vitro fertilisation

IVG . . . . . In vitro gametogenesis

IVW . . . . Inverse-variance weighted

KEGG . . . Kyoto encyclopaedia of genes and genomes

KO . . . . . Knockout

LD . . . . . Linkage disequilibrium

LDL . . . . . Low-density lipoprotein

LMM . . . . Linear mixed models

LoF . . . . . Loss-of-function

LOFTEE Loss-of-function transcript effect estimator

LOY . . . . Loss of chromosome Y

MAC . . . . Minor allele count

MAF . . . . Minor allele frequency

Mb . . . . . Mega base

mESC . . . Mouse embryonic stem cell

MHC . . . . Major histocompatibility complex

MR . . . . . Mendelian Randomisation

NEB . . . . . Number of children

NGS . . . . . Next generation sequencing

OA . . . . . Other allele

OR . . . . . Odds ratio

PC . . . . . Principal component

PCOS . . . Polycystic Ovarian Syndrome

PCR . . . . . Polymerase chain reaction

EI . . . . . Egger intercept

EM . . . . . Early menopause

EpiLC . . . . Epiblast-like cell

eQTL . . . . . Expression quantitative trait loci

FACS . . . . . Fluorescence activated cell sorting

FDR . . . . . False discovery rate

FSH . . . . . Follicle-stimulating hormone

G2G . . . . . GWAS to Gene

GCTA . . . . . Genome-wide complex trait analysis

GeL . . . . . Genomics England

GLM . . . . . Generalised linear model

GnRH . . . . . Gonadotropin releasing hormone

GO . . . . . Gene Ontologies

GRCh37 . . . Genome reference human build 37

GRM . . . . . Genetic relatedness matrix

GRS . . . . . Genetic risk score

GTEx . . . . . Genotype tissue expression

GWAS . . . . Genome-wide association study

HC . . . . . High confidence

HDL . . . . . High-density lipoprotein

hg38 . . . . . Human genome build 38

HGP . . . . . Human Genome Project

HPA . . . . . Hypothalamic pituitary adrenal axis

HRT . . . . . Hormone replacement therapy

iPSC . . . . . Induced pluripotent stem cell

PGC . . . . . Primordial germ cell

PGCLC . . Primordial germ-like cell

pLI . . . . . Probability of loss-of-function intolerance

POI . . . . . Premature ovarian insufficiency

PoPs . . . . . Polygenic priority score

pQTL . . . Protein quantitative trait loci

PTV . . . . . Protein truncating variant

PWM . . . . Penalised weighted median

QC . . . . . Quality control

RAP . . . . . Research analysis platform

RCT . . . . . Randomised controlled trial

RHR . . . . Resting heart rate

SBP . . . . . Systolic blood pressure

SE . . . . . Standard error

SLDP . . . . Signed linkage disequilibrium profile

SNP . . . . . Single nucleotide polymorphism

SNV . . . . . Single nucleotide variant

T2D . . . . . Type 2 diabetes

TL . . . . . Telomere length

TSS . . . . . Transcription start site

UKBB . . . United Kingdom biobank

VEP . . . . . Variant effect predictor

WES . . . . Whole exome sequencing

WM . . . . . Weighted median

Table of Contents

List of Figures

List of Tables

CHAPTER 1: Introduction

Overview of scientific and clinical advances in reproductive ageing

Summary

This thesis studies reproductive ageing as a common theme. The Introductory chapter provides an overview of the current scientific and clinical knowledge on reproductive ageing and fertility, specifically focusing on the epidemiological perspective, genetic architecture and underlying biological mechanisms governing ovarian function and its relationship to the overall health status in women. In addition, it touches on the clinical perspective to define current challenges for accurate prediction and diagnosis of reproductive health outcomes. Finally, it highlights unexplored questions, gaps in knowledge and opportunities that this thesis aims to address to advance our understanding of reproductive ageing and fertility in women.

1.1 Ovarian ageing and fertility: an understudied phenomenon of the neglected organ

Female reproductive health encompasses a diverse range of traits and health outcomes, including diseases of the reproductive organs and tissues, pregnancy-related outcomes, sexually transmitted infections (STIs) and violence against women. Aspects of reproductive health and fertility are likely to be geographically patterned and population specific, having critical implications for clinical outcomes and individual wellbeing[1]. However, an interesting pattern could be observed in the case of reproductive longevity. Improvements in healthcare, hygiene and the availability of food have significantly contributed to increased human life expectancy over the past 2 centuries, shifting it from 45 to 85 years [2]. On the contrary, the length of the female reproductive lifespan has remained relatively constant around the age of 51, when menopause occurs due to the depletion of functional follicles in the ovaries[3,4]. Menopause is generally defined as the last menstrual period followed by 12 months of amenorrhea, and has a profound impact on fertility and health outcomes in later life. Interestingly, menopause is a process almost unique to humans, with only a few other species found to have post-reproductive lifespans, including gall-forming social aphids, killer whales and short-finned pilot whales[5]. Evolutionary theory suggests that decline in reproductive ability before the end of life would be selected against and so there should be evolutionary advantages to menopause[6]. Gain in fitness for post-reproductive women is believed to be the result of the reduction of mortality risk from pregnancy in later life, thus investing in their children (‘*mother hypothesis*’), grandchildren (‘*grandmother hypothesis*’) and reducing reproductive overlap between generations (‘*reproductive conflict*’)[6,7]. For the most part of human history, cultures had no word to describe menopause, and it was merely recognised as a transition to the status of ‘elder grandmother’. It was around the 1700s that people began to see menopause as a harmful condition due to its link to upsetting symptoms that rendered women weak and vulnerable, as well as its association to average life expectancy in women at the time. During the 19th and 20th centuries the male dominant medical community considered menopause a taboo. Sigmund Freud wrote that “*It is a well-known fact … that after women have lost their genital function their character often undergoes a peculiar alteration*” and they become “*quarrelsome, vexatious and overbearing*”. Psychiatrist David Rueben weighed in…”*Once the ovaries stop, the very essence of being a woman stops…she is no longer a functional woman*.”[8]. This attitude towards female reproductive health significantly impacted our understanding of biology that underlies reproductive longevity and timing of menopause. Consequently, women’s health therapeutics to preserve fertility and prevent associated health outcomes have not seen major advancement in decades. Due to the increased life expectancy, women are now spending a large proportion of their lives in ill

health and disability, thus highlighting further the need to understand the process of reproductive ageing better.

Female reproductive longevity varies substantially between women in the general population (**Figure 1.1**)[9]. Women are born with a non-renewable ovarian reserve[10]. Follicles, consisting of oocytes and surrounding granulosa cells, are formed *in utero* and maintained as resting primordial follicles arrested in the first meiotic division (M1) in the cortex, constituting the ovarian reserve (**Figure 1.2A**). Approximately 7 million oocytes are derived from primordial germ cells by 6 months post conception[11]. The oocytes may stay in this 'resting phase' for many decades to resume meiosis only just before being released at ovulation. During this protracted period, the oocytes are vulnerable as they may be subjected to various endogenous and exogenous insults that cause DNA damage, and thus are highly dependent on efficient DNA damage response and repair (DDR) mechanisms to maintain germ cells' genomic integrity and prevent DNA from possible damage[12,13]. Primordial follicles are activated from the ovarian reserve at a rate of several hundred per month in childhood, peaking at around 900 per month at approximately 15 years of age[14]. The number of oocytes in the reserve significantly drops to about 400,000 (~5%) at puberty due to the elimination of abnormal, damaged or excess cells at every stage of oogenesis[11,15–18]. This follicle loss mainly happens via atresia prompted by apoptosis of the primary oocyte through mechanisms specific to the ovary, in addition to conventional apoptotic pathways[19–22]. Autophagy and necroptosis may also contribute to atresia[19].

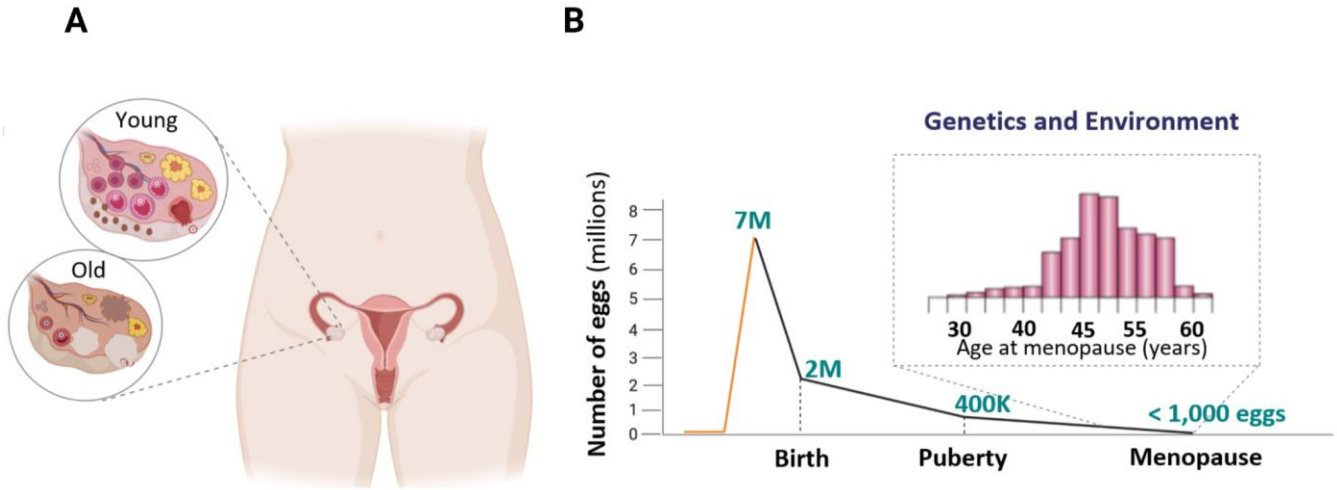

Figure 1.1: Number of ovarian follicles declines during a woman's lifetime. *This schematic figure depicts* ***(A)*** *the appearance of a young and old ovary with the oocyte reserve and* ***(B)*** *ovarian germ cell numbers throughout key life stages. The histogram indicates the normal population distribution of age at natural menopause. Besides DDR mechanisms that have a role to preserve the oocyte pool, other mechanisms might contribute to the rate of oocyte depletion, including the rate of follicular recruitment, for example, by follicle stimulating hormone (FSH).*

Following recruitment, follicles grow by mitotic division of granulosa cells and expansion of oocyte volume for almost six months until meiosis is reinitiated at ovulation triggered by pituitary gonadotropins[23]. Folliculogenesis occurs in waves, with the whole process taking about 120 days, and so the ovary contains follicles at all stages of development. Waves of atresia accompany developmental transitions and growing follicles are continuously induced to undergo cell death such that, typically, only a single follicle matures to ovulate each month[20]. Unlike atresia, only a small proportion of oocytes are lost through ovulation (**Figure 1.2B**)[24]. Only oocytes that are fertilised will complete meiosis II and the remainder degenerate (**Figure 1.2**). As ovarian reserve declines, the rate of follicle recruitment decreases, but the preovulatory follicles continue to produce substantial amounts of oestrogen, while other important hormones such as anti-Müllerian hormone (AMH) and inhibin-B decline, leading to upregulation of the hypothalamus–pituitary gonadal axis, i.e. increase in levels of gonadotropin releasing hormone (GnRH), follicle-stimulating hormone (FSH) and luteinizing hormone (LH)[25]. This gradual decline in the ovarian follicle pool is associated with irregular menstruation in the years prior to menopause, a period called perimenopause, and finally, follicle exhaustion and menopause when follicle numbers have dropped below 1000[23,26–28]. Therefore, the high variability in menopausal timing observed in women can be due to differences in the size of the ovarian reserve at birth, but also due to the differences in the rate of follicle loss.

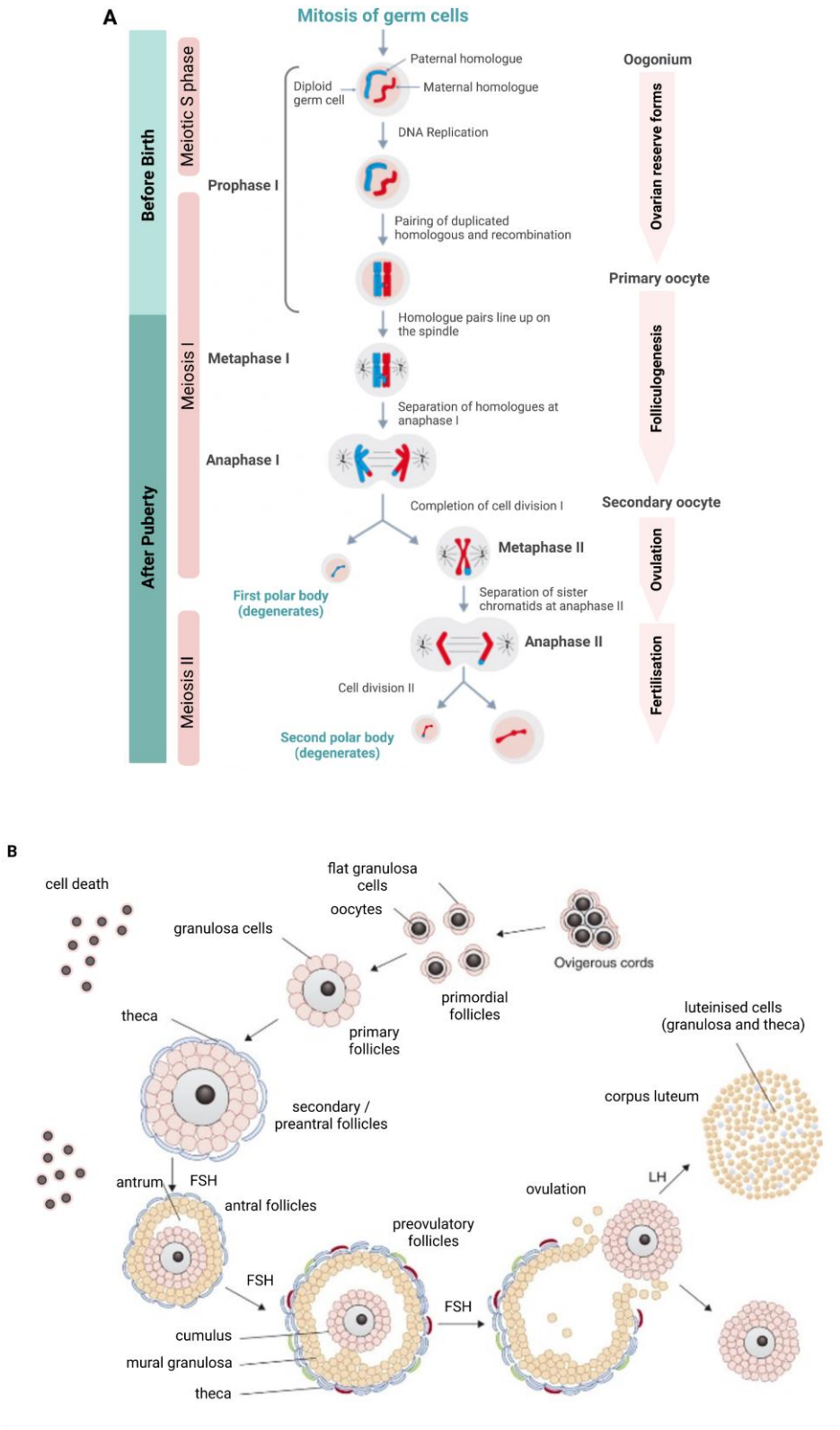

Figure 1.2: Oogenesis and ovarian cycle. (A) *The process of oogenesis creates oocytes by meiosis. Ovarian follicles are produced as a result of mitosis of germ cells and their differentiation to oogonia, which then enter meiosis to form primary oocytes. A significant proportion of oocytes go through atresia during meiosis I, where*

DDR genes play an important role to protect from and repair introduced DNA damage. Specifically, the DDR genes are crucial during recombination that occurs at the time of homologous chromosome pairing. ***(B)*** *The figure demonstrates different stages of the ovarian cycle, including the ovulation and cell death of damaged follicles. Image is adapted from Meiosis, Genomics Education*[29]*.*

Alongside a quantitative decline in the oocyte number, there is also a qualitative decline in oocyte quality, primarily attributed to a loss of genetic integrity and increase in aneuploidy amongst ageing oocytes. Previous research has suggested that the oocytes created first are the first to be ovulated, and that they have more recombination events and a lower risk of non-disjunction - '*the production line hypothesis*'[30]. However, subsequent studies found there is no difference in recombination rates in oocytes from older women compared with those from younger women[31] or in the number of recombination events in oocytes entering meiosis early in foetal life compared to those created later[32]. Other explanations for aneuploidy relate to the extended period of time for which oocytes are arrested in prophase I. These include loss of cohesion between sister chromatids, age-dependant decay of components of the cell machinery required for meiosis and influence of environmental exposures [32,33].

Although both oocyte quantity and quality decline with increasing age, it is unclear whether they are controlled by the same mechanisms and whether they decline in parallel.

1.2 Reproductive longevity in ageing populations

Natural fertility is believed to be closely associated with menopause timing and to start significantly declining around 10 years before menopause in most of the cases[24,34]. On average, menopause occurs at around the age of 51[35]. Notably, epidemiological studies have shown that menopause timing varies across ethnic groups, suggesting that different modifiers might exist in different ethnic backgrounds. More specifically, African and African-American women have earlier, while Japanese later average menopause timing, as compared to women of European descent[183,557]. In addition to the ethnic variation, average age at natural menopause may also vary across time periods, with a secular trend towards later menopause being observed in multiple studied cohorts[562,563,564]. For example, a study in more than 300,000 Norwegian women found that mean age at menopause increased from 50.31 years among women born during 1936–1939 to 52.73 years among women born during 1960–1964[562]. Such an increase could be a result of lifestyle changes. Growth restriction during foetal life may possibly impair ovarian development, and poor nutritional status during early life could increase the rate of follicle atresia and thereby decrease age at menopause[565].Therefore, the increase in birthweight across birth cohorts due to improved childhood nutrition and health could possibly explain part of the increase in age at menopause[566]. On the other hand

this secular trend seems paradoxical because several adult determinants, such as smoking, sedentarity, and nulliparity, associated with early menopause are on the rise in Europe.

Epidemiologically 10% of women undergo early menopause (EM), defined as menopause before the age of 46, while premature ovarian insufficiency (POI), defined as the extreme form of EM under age of 40, affects 1 in 100 (1%) women[36,37]. This means that women with EM will start experiencing accelerated decline in fertility around the age of 30[11]. The effect of EM on infertility, which now affects ~15% of couples in the United Kingdom (UK), and childlessness (~20%), is becoming increasingly relevant due to the secular trend of delaying parenthood to later maternal age at childbirth, especially in Western populations [38–42]. These trends have resulted in increased demand for fertility treatments, such as *in vitro* fertilisation (IVF) and cryopreservation, to prolong reproductive longevity [43–46]. However, often these treatments are unaffordable, invasive and with limited success [34,47,48]. For example, there is a ~6.5% chance of achieving a pregnancy with each mature oocyte thawed, but this decreases dramatically in women of advanced maternal age[4 3,45,46,49,50]. Even with fertility preservation treatments, the live-birth rate in women >40 years is only 3.2% per treatment cycle, and this is accompanied by additional increased risks of miscarriage and adverse foetal and maternal outcomes, including pregnancy-related maternal complications such as pre-eclampsia and gestational diabetes[32,33,51]. The core problem lies in the current clinical practice treating rather than preventing symptoms, and this becomes crucial as oocyte death is irreversible thus requiring early prevention strategies[44]. New methods for preserving natural fertility and/or enhancing the success of IVF would be particularly welcome.

Additionally, identification of women with reduced reproductive lifespan cannot be accurately achieved by any endocrine or imaging tests that are in clinical practice, most often including antral follicle count (AFC) on ultrasound, and levels of AMH and FSH [52–55]. These tests only record changes in ovarian function that have already taken place, disabling the long-term prediction of reproductive expectations [54,56–58]. Preferably, we would like a test that can accurately predict the age at which a woman will become menopausal, which would open up the opportunity for any young woman to be tested for reproductive expectations and counselled on the availability of elective fertility preservation. It is likely that we will arrive at that position by combining endocrine, imaging and, especially, genetic information present from birth, which requires thorough assessment of the regulators and physiological mechanisms involved in reproductive ageing [43,45,46,49,50].

The aetiology behind reproductive ageing appears to be complex, being influenced by a combination of environmental, social and genetic factors. The environmental determinants have been increasingly well

characterised in recent years, with factors such as oral contraceptive use, parity, smoking, and alcohol consumption receiving much attention [59,60]. However, biological pathways underlying ovarian ageing are not yet fully characterised. Despite the great number of genes implicated in reproductive physiology by the study of animal models, only a subset of these genes is associated with human reproductive longevity and infertility. Human genetic studies have attempted to overcome this problem by identifying genetic markers associated with menopause timing and thus providing substantial insight into the functional mechanisms governing ovarian ageing. The identification of these genes was largely achieved through targeted DNA sequencing in individuals with rare disorders of reproductive timing and, more recently, by large-scale array genotyping of single nucleotide polymorphisms (SNPs) in population-based samples[61].

1.3 Human genome: science behind inheritance, genetic variation and causal estimation

The identification and understanding of genetic factors and the ways they influence an individual's susceptibility for a certain trait lie in the centre of human genetics. Sequencing of the reference genome, accomplished by the Human Genome Project (HGP) in 2003, marked a turning point in gene-mapping research revolutionising the way we study health outcomes.

In case of complex diseases, genes containing a genetic variation that increases phenotype predisposition are referred to as "susceptibility genes" (**Figure 1.3A**)[62]. They do not directly cause disease, but rather influence disease risk. In combination with the environment, this multifactorial genetic architecture suggests an additive effect of multiple genes and mostly common frequency alleles on the phenotype[62]. For complex disorders, instead of mapping disease genes by tracing transmission in families, the HGP enabled the creation of high-density polymorphism maps initiated by The International HapMap Consortium[63–65]. This expedited population-based association testing at variant sites throughout the genome, which revealed part of the story behind the role of inheritance and genetic variation in disease aetiology[63,64,66]. These advances gave an insight into the specific patterns of associations among alleles present across the genome, known as linkage disequilibrium (LD) blocks. LD blocks are formed due to a non-random association of alleles at two or more loci resulting from different historical evolutionary forces, including recombination rate, natural selection, mutations etc. (**Figure 1.3 B,C**)[67].

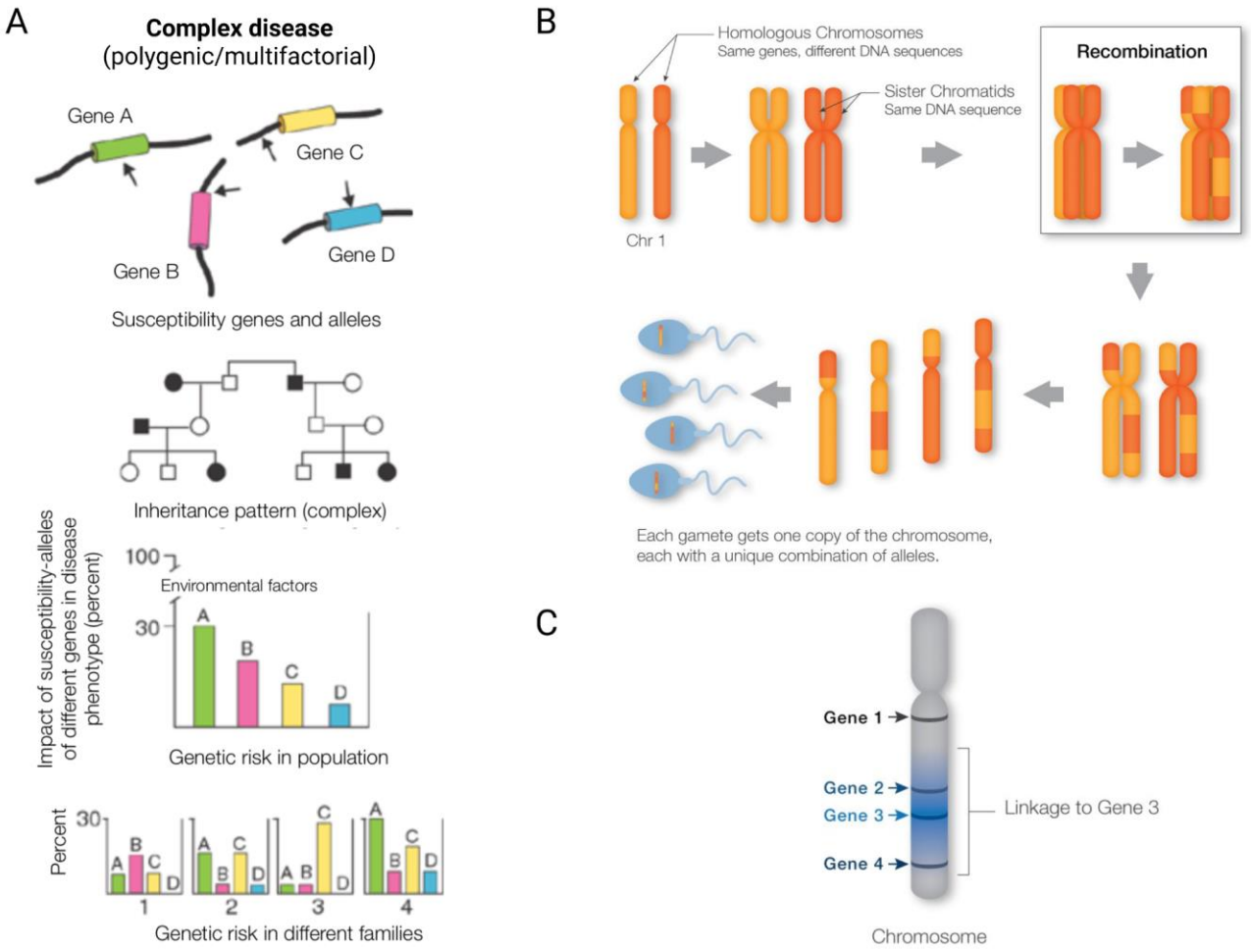

Figure 1.3: Mechanisms behind inheritance of complex diseases. (A) *Inheritance of complex disorders: in complex disorders, several alleles in a number of genes result in a genetic predisposition to a clinical phenotype. Genes containing variation related to complex traits are thus referred to as "susceptibility genes", and variants are neither sufficient nor necessary to explain the disease phenotype. Environment and life-style factors are contributors to the pathogenesis of these disorders.* ***(B)*** *Exchanges of information between chromosomes during homologous recombination resulting in chromosomes inherited together by offspring being different from those parental chromosomes. Recombination indicates that variants located physically close on chromosomes are more likely to be inherited together.* ***(C)*** *Genetic mapping using linkage: Alleles of genes which are physically closer to gene 3 are more likely to be inherited with gene 3; this is defined as a greater degree of linkage. Similarly, proximal alleles are also more likely to be inherited together. Gene 1 and Gene 3 are not linked, but by chance they will still be inherited together 50% of the time, the same as if they were on separate chromosomes. Image adapted from Peltonen et al*[68] *and Genetic linkage*[69].

These LD patterns between SNPs were then used to enable genotyping arrays to tag the majority of common variants by selecting and analysing only a part of the total number of SNPs for association with the phenotype[70]. Genotypic data can be phased and untyped genotypes imputed using information from matched reference populations from repositories such as 1000 Genomes Project or TopMed[71,72]. Together, these achievements paved the way towards the first genome-wide association studies (GWAS), a transformative step for the study of complex disorders (**Figure 1.4**) [73,74].

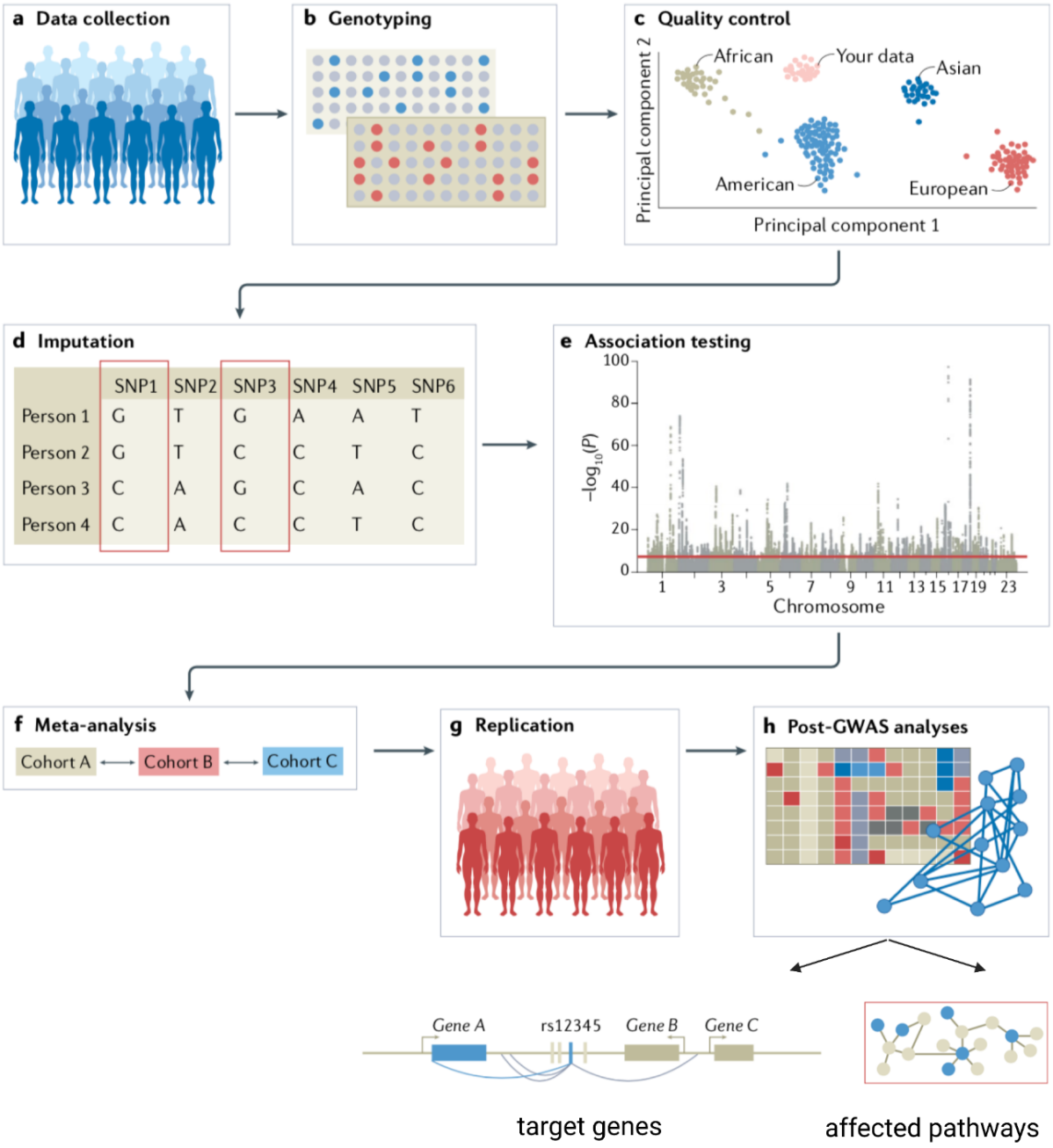

Figure 1.4: Schematic representation of a GWAS study and post-GWAS analyses: (A, B) *Data can be collected from study cohorts or available genetic and phenotypic information can be used from biobanks or repositories. In a case-control GWAS, a large cohort of diseased individuals (cases) and controls is genotyped for hundreds of thousands of SNPs spread throughout the genome.* ***(C)*** *The ancestry of individuals in the cohort of interest is determined through principal components analyses.* ***(D)*** *Genotypic data can be phased, and untyped genotypes imputed using information from matched reference populations from repositories such as 1000 Genomes Project or TopMed.* ***(E)*** *An associated region will often contain dozens of correlated SNPs in high LD with very similar association signals that, together, can span numerous genes.* ***(F, G)*** *GWAS analysis can be performed as part of the meta-analysis to boost the power, and should be replicated in an independent sample.* ***(H)*** *To narrow the multiple correlated signals down to a single or very few causal variants, researchers perform functional follow-up applying*

various post-GWAS strategies. This can include gene prioritisation and pathway enrichment analysis. Figure adapted from Uffelmann et al (2021)[75].

By combining data from multiple large-scale population sources, such as UK Biobank (UKBB) and 23andMe, recent GWAS obtained sample sizes that provide statistical power to identify associations of small effect with great precision at large numbers of genetic loci.

1.4 Insights into common genetic variation underlying ovarian ageing and menopause timing

The variation in timing of menopause reflects a complex mix of genetic and environmental factors that population-based studies have begun to unravel. The contribution of the genetic component derived from twin and family studies is believed to range from 44% to 65%, involving hundreds of rare, low frequency and common variants [76–81]. The number of GWASs on age at natural menopause (ANM) have been growing in size over time, leading to the discovery of ~300 common genetic variants responsible for menopausal timing in ~200,000 women of European ancestry (**Figure 1.5**)[76,82–86]. These reported variants cumulatively explain 10-12% of the variance in ANM and 31-38% of the overall estimated SNP heritability, with individual SNPs having an effect to shift menopause timing from ~3.5 weeks to ~1.5 years. In addition to common and low-frequency coding variants with <5% minor allele frequency (MAF), relatively large effect sizes have been identified in two genes, *BRCA2* and *CHEK2*, by analysing exome sequence data in 45,351 women in UKBB. In aggregate and compared to non-carriers, women carrying loss-of-function (LOF) variants in *BRCA2* and *CHEK2* reported ANM 1.54 years earlier (95% CI 0.73–2.34, P=6.8*10^{-5}) and 3.49 years later (95% CI 2.36–4.63, P=1*10^{-13}), respectively[82].

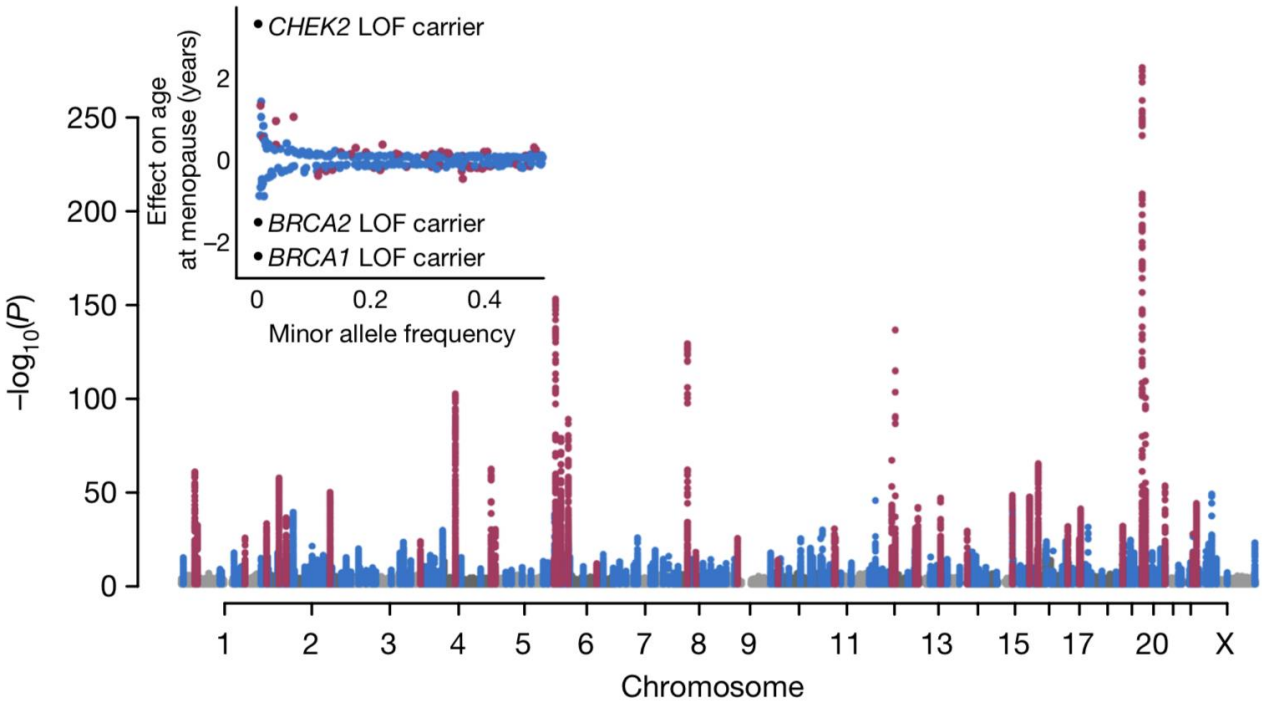

Figure 1.5: Manhattan plot of age at natural menopause GWAS signals in the ReproGen consortium. *The genetic variants coloured in purple represent the loci identified in previous GWASs, while the ones in blue are the variants discovered in the latest ANM meta-analysis in 201,323 women from the ReproGen consortium. Plotted variants have $P < 0.01$ with $P < 1\times10^{-300}$ truncated. Inset, effect sizes and MAFs of the loci, with CHEK2, BRCA1 and BRCA2 LoF variants are highlighted. Figure is obtained from Ruth et al (2021)*[82].

However, besides explaining only a fraction of heritability[87], these common genetic variants are limited in their ability to identify causal genes because the majority of associated common haplotypes contain non-coding variants. This hinders the translation of GWAS findings into mechanistic understanding and effective therapeutic solutions, which further highlights the need of exploring the sources of 'undiscovered heritability' in rare, high impact variants via whole exome sequencing (WES)[88]. To address this gap in knowledge, **Chapter 3** of my thesis will aim to assess how much these rare protein coding variants can further our understanding of the ANM genetic architecture, and whether the mechanisms they reveal can be used to pave the path towards more targeted biomarker identification for prediction and treatment development.

Even though it has been widely believed that extreme forms of early menopause (POI) are the result of rare monogenic alleles, previous research demonstrated that identified common alleles also influence these clinical extremes[82]. More specifically, women in the top 1% of polygenic susceptibility had a six-fold increased risk of POI, which is equivalent to those carrying monogenic *FMR1* premutations, a screened

monogenic cause (**Figure 1.6**)[82,89]. Importantly, this highlighted a shared genetic aetiology between normal ANM, EM and POI, which is at least partly explained by the additive effects of the same polygenic variants[76,90]. Insight into these genetic risk factors started paving the path towards the opportunity for prediction of individuals with early menopause and POI. Current predictive power using GWAS discovered variants is at 64% and 65% chance to distinguish EM and POI from the rest of the population. Even though the genetic risk alone would be a weak predictor, despite its overall low discriminative ability, extremes of the polygenic score were able to identify some individuals at high risk of POI[82]. However, this is yet not sufficient to be used as part of the clinical practice. As we identify a greater proportion of genetics governing reproductive ageing, not only in Europeans but also other ancestries, predictive models will improve.

Finally, this novel insight into the shared genetic architecture between common and rare variants underlying POI, suggests that it would be important to evaluate and better understand the penetrance of ~100 genes that are part of gene panels currently used in clinical and non-clinical environments to identify and diagnose POI. **Chapter 4** of this thesis will use the WES data in UKBB to address this question.

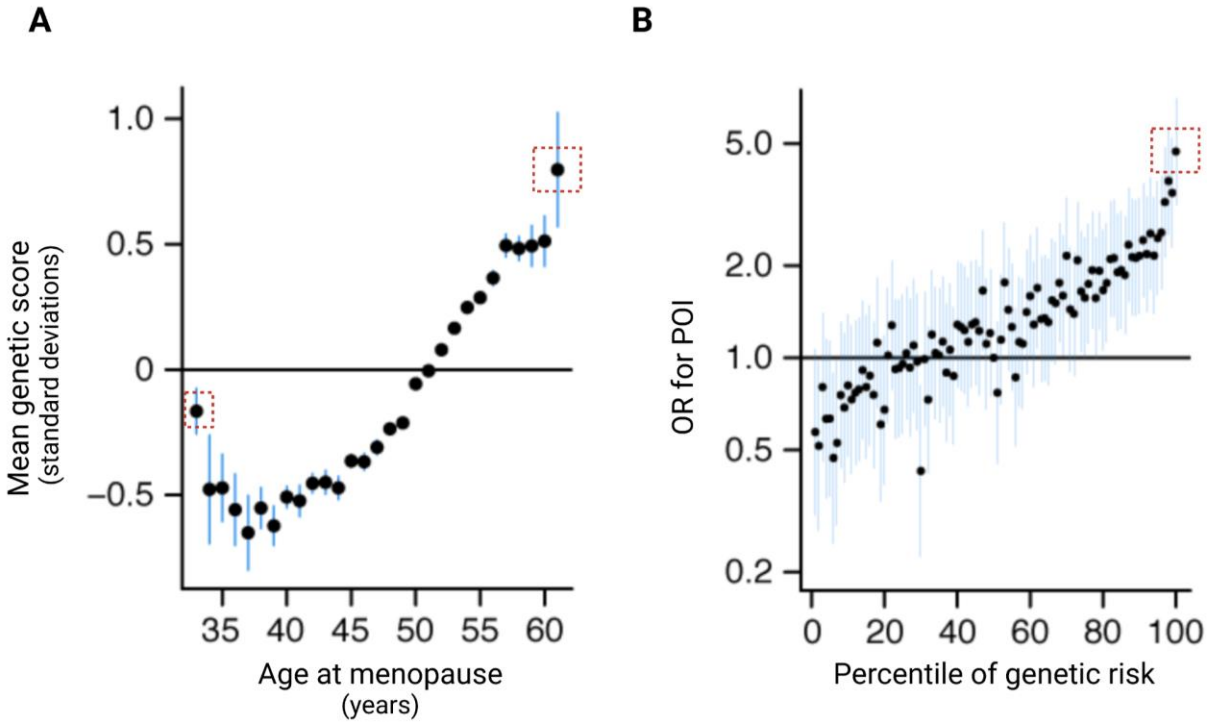

Figure 1.6: Polygenic prediction of age at menopause. (A) *Mean PGS (scaled to have mean = 0, s.d. = 1) for a given ANM. Higher PGS indicates later ANM.* ***(B)*** *Association of each centile of PGS compared with the 50th centile with premature ovarian insufficiency. Error bars indicate 95% CI. Figure is obtained from Ruth et al (2021)*[82].

The majority of discovered ANM loci implicated genes that regulate DDR at different stages of the ovarian cycle, highlighting the particular sensitivity of oocytes to DNA damage due to the prolonged state of cell cycle arrest across the life-course (**Figure 1.7A,B**) [91–93]. This involves various DDR mechanisms, such as homologous recombination, base excision, mismatch, nucleotide excision repair, apoptosis etc. that act across the life-course to shape the ovarian reserve and its rate of depletion[91,93–98] (**Figure 1.7C**). Besides DDR, other biological processes are implicated in ovarian ageing including control of cell cycle, embryonic development, metabolism, gene expression, hormone signalling, immune function, meiosis, protein synthesis and gonad development (**Figure 1.7**) [82].

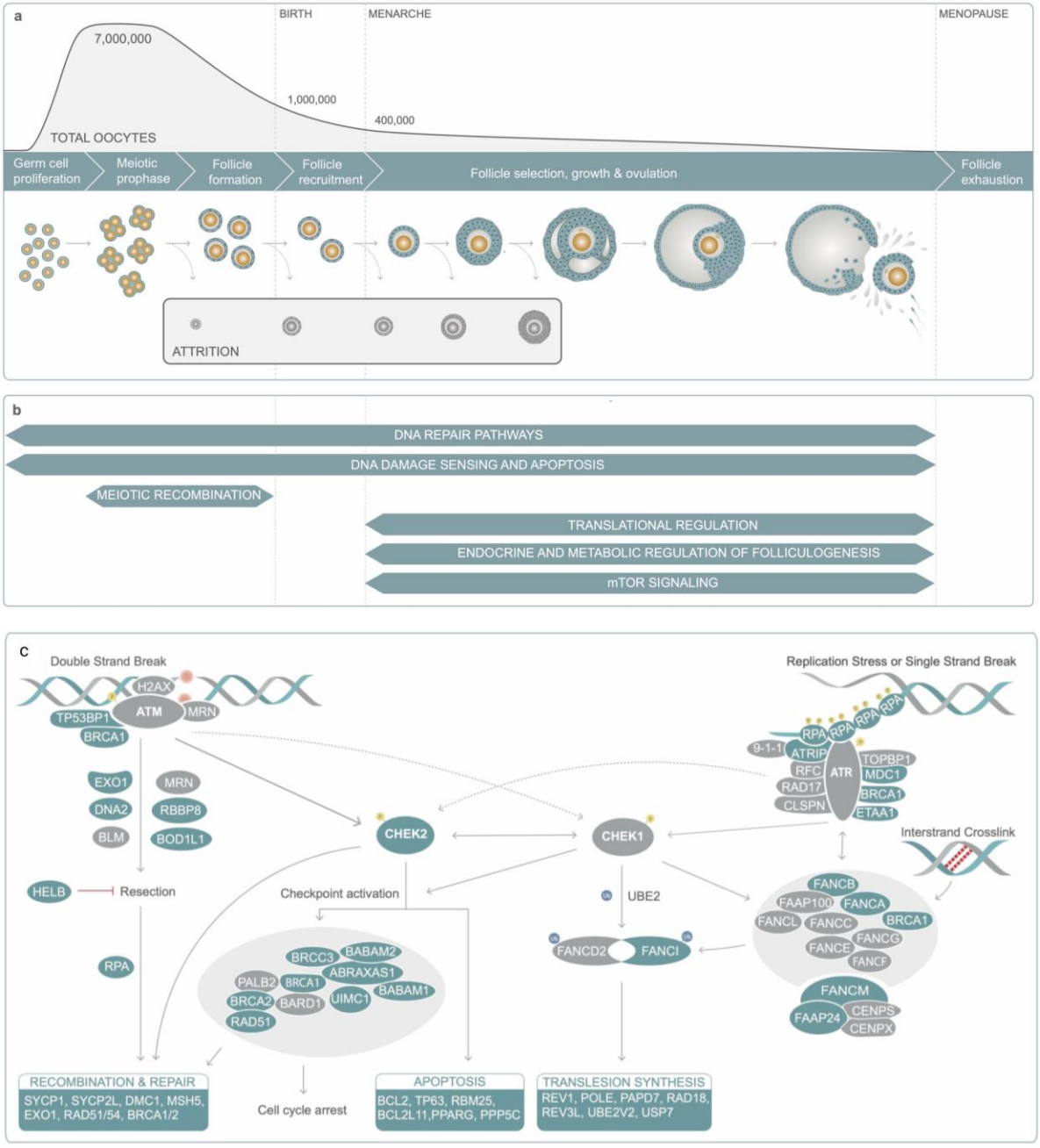

Figure 1.7: Overview of the ovarian reserve and follicular activity across lifespan with underlying mechanisms that regulate reproductive ageing. (A) *Key processes involved in follicular activity from foetal development to menopause showing the numbers of oocytes at each stage* ***(B)*** *Summary of key biological pathways involved in follicular activity and their relationship to stage of reproductive life* ***(C)*** *Genes involved in downstream DNA damage response and repair pathways with those within 300 kb of an ANM signal shown in blue. Figure is adapted from Ruth et al (2021)*[82].

These most likely causal genes are suggested on the basis of their relationship with the strongest association signal in a region, including their physical proximity, association with expression levels of the gene and/or biological plausibility. Therefore, many of the genes identified by GWAS remain to be confirmed as functionally important by *in vitro* studies. These findings are critical as improved knowledge of the underlying mechanisms may also allow their manipulation in humans, more specifically halting or temporising the process of oocyte wastage. Using state-of-the art CRISPR technologies, previous work demonstrated that experimental manipulation of DDR pathways highlighted by human genetics increases fertility and extends reproductive life in mice. This was achieved either by altering the initial size of the ovarian reserve or its decline in checkpoint kinases, *Chek1* and *Chek2* transgenic mice (**Figure 1.8 A**)[82].

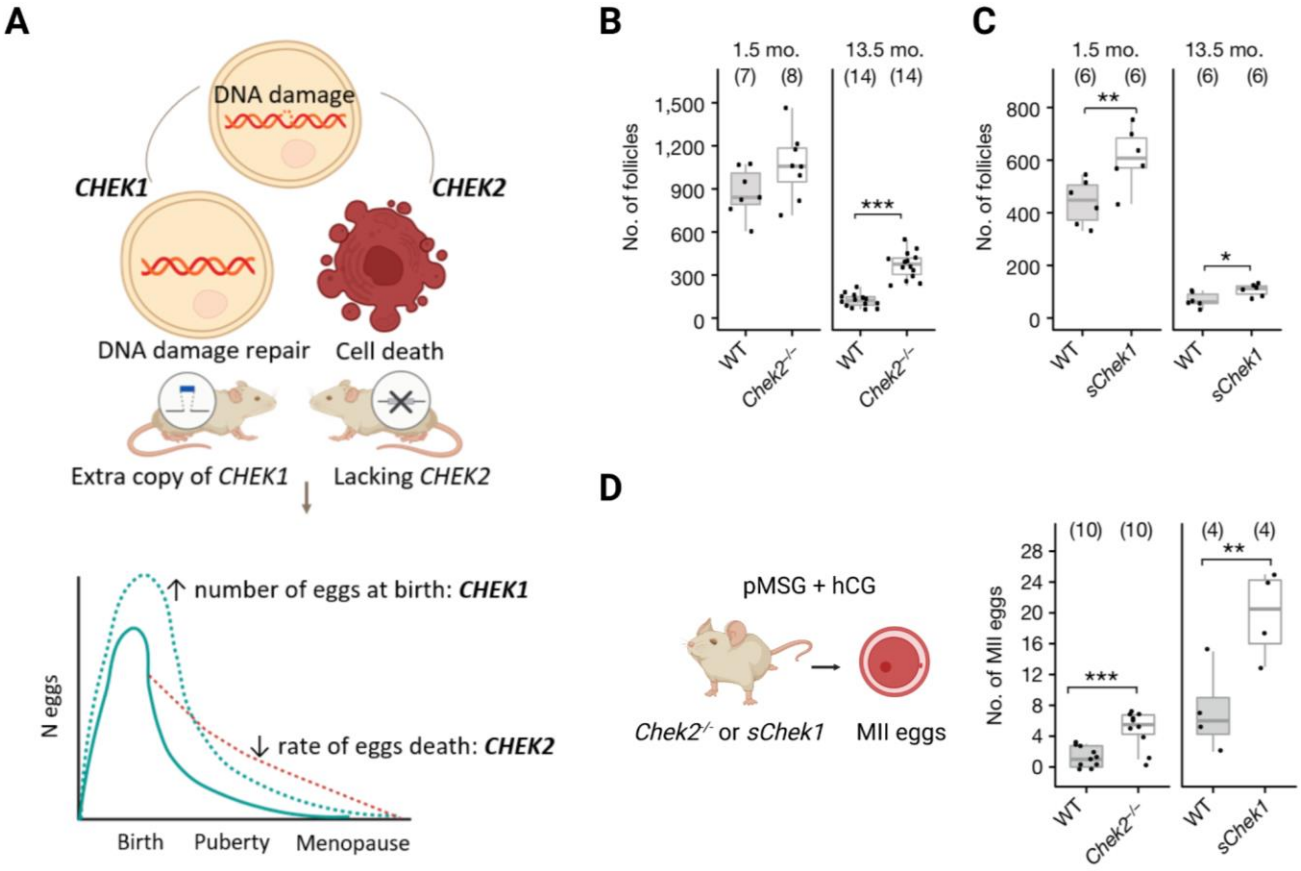

Figure 1.8: Genetic manipulation of checkpoint kinases extends reproductive lifespan in mice by limiting the destruction of egg cells or upregulating the DNA-repair process. (A) *Schematic representation of the experimental strategy and reproductive outcome in Chek2$^{-/-}$ and sChek1 mice.* ***(B)*** *Numbers of follicles in young and aged Chek2$^{-/-}$* ***(C)*** *and sChek1 female mice.* ***(D)*** *Response to gonadotrophin stimulation of 13.5-month-old Chek2$^{-/-}$ and 11- to 13-month-old sChek1 females assessed by the number of MII oocytes retrieved. Figure is adapted from Ruth et al (2021)*[82].

Inactivating *Chek2*, which normally has a crucial role in destruction of eggs compromised by DNA damage, slowed the depletion of the ovarian reserve by disrupting apoptosis, which happened without a significant initial increase in the ovarian reserve at birth (**Figure 1.8 B**) [82,99–104]. In addition, the transgenic *Chek2* mice, when around the age equivalent of menopause in human, had increased ovarian response to hormones used to stimulate the release of immature eggs for IVF procedures[82]. This was the first evidence of a potential therapeutic target for enhancing ovarian stimulation in women undergoing IVF treatment through short-term apoptotic inhibition. Unlike *Chek2, Chek1* is needed for embryo development and helps DNA repair; its inactivation specifically in oocytes leads to female infertility, while the full gene knockout is embryonically lethal[100,105–108]. By contrast, introducing an extra copy of *Chek1* resulted in increased ovarian reserve at birth compared to litter-mate controls leading to prolonged genomic integrity, enhanced follicular activity and delayed reproductive senescence. This was likely a consequence of upregulation of replication-associated DNA repair processes during mitosis/meiosis, and demonstrates that this repair might be limiting for establishing and maintaining the ovarian reserve (**Figure 1.8C**)[82]. Together, these data demonstrated for the first time ever that modulation of key DDR genes can extend reproductive lifespan by ~25% in vivo, and increase fertility potential thus leading to the generation of healthy pups that are fertile over several generations (**Figure 1.8D**)[82]. However, a large proportion of gene candidates remain to be functionally validated to better understand the mechanism they operate via to regulate ovarian ageing and modulate the timing of menopause; a question that would need to be answered to initiate the translation of the genomic findings for clinical purposes[109]. One such candidate, *PARP-1*, I study using human genomics and functional evaluation in **Chapter 7**. I particularly focus on *PARP-1* due to recent evidence in the literature that reported the impact of Olaparib drug, a *PARP-1* inhibitor, on the reduction of the ovarian reserve[110].

Besides genomic data that are extensively described in **Chapter 1** and further studied in my thesis, the human serum proteome represents a valuable resource of potential biomarkers for polygenic disorders as it enables direct assessment of changes in protein levels[111]. Importantly, most pharmaceutical drugs also target proteins, further increasing their actionability. Therefore, studying human proteome could help us identify novel determinants of reproductive ageing and **Chapter 5** will use the largest proteomic data up to date to address this question.

Finally, the recent study involving single nuclei multi-omic analysis of young and reproductively aged ovaries provided high resolution characterisation of the transcriptional regulatory landscape at the single cell level[112]. They demonstrated that ageing significantly remodels the cellular architecture of the human

ovary, and introduces coordinated transcriptomic changes thus causing alterations in gene regulatory networks and cellular communications among oocytes and somatic cell types. This suggests that regulation of transcription also contributes to an age related loss of follicular function, tissue fibrosis and epithelial hyperplasia.

1.5 Reproductive ageing and related health outcomes

Normal variation in reproductive lifespan is causally associated with the risk of a wide range of disease outcomes in women[113]. The most robust evidence on the association with later life diseases was obtained from randomised controlled trials (RCTs) and causal inference analysis via Mendelian Randomisation (MR) frameworks (**Figure 1.9**). MR relies on random assortment of alleles during gamete production and fertilisation, and uses that as a basis of so-called 'naturally randomised trial'[114]. The evidence derived using MR indicated that earlier menopause timing deteriorated bone health, including bone mineral density and fracture, and increased the risk of type 2 diabetes (T2D)[82]. This is in line with evidence from RCTs on oestrogen therapy and bone health[115–118]. Consistent with previous research, each one-year genetically mediated delay in ANM increased the relative risks of several hormone-sensitive cancers by up to 5%[82]. This life stage in women is linked to one of the major hormonal changes, characterised by a decline in oestrogen and progesterone levels and to a lesser degree, testosterone. Notably, the association with cancer outcomes appeared not to be driven by DDR mechanisms, but rather hormonal regulation, including lifetime exposure to oestrogen. This explains the protective effect given that the enrichment of DDR genes is associated with both menopause and cancer. In agreement with the trial data in younger women taking hormone replacement therapy (HRT) and opposite to evidence from observational studies, our causal inference analysis suggested no increased risk of cardiovascular disease[119–121], lipid levels, Alzheimer's disease, body mass or longevity[122–130]. Finally, previous observational studies demonstrated conflicting evidence on the association between earlier menopause timing and increased risk of dementia, Parkinson's disease and depression that suggest that menopause itself may be a dynamic neurological transition.

Unlike observational studies that report conflicting evidence on the relationship between age at menarche and menopause, causal inference analysis indicated that genetically mediated age at menarche was associated with a decrease in ANM of about 8 weeks per year of earlier menarche[82,568]. However, the regulatory mechanisms clearly differed between the two traits overall. Unlike ANM that is mainly driven by DDR mechanisms, age at menarche was enriched in genes expressed in the hypothalamus and pituitary gland[131]. Results from GWAS studies demonstrated how more robust statistical tools and increases in

sample size over time have improved the power to identify more genetic determinants of reproductive traits. Based on this, we were interested to further study the relationship between ANM and menarche timing, and **Chapter 6** will specifically address this question utilising the largest ANM and menarche GWASs up to date as well as novel techniques for gene prioritisation.

Both causal inference analysis and observational studies evaluated putative modifiable determinants of ANM and reported that increase in alcohol consumption and tobacco smoking were associated with earlier ANM[82]. Besides smoking[132,133] and alcohol consumption[132,134,135], a number of other epidemiological risk factors has been found as associated with ANM. This includes socio-economic status[136,137], diet[132,138,139], exercise[132,134,138,140], and exposure to environmental toxins[137,141,142]. Expanding our data both in terms of size and ethnic heterogeneity will allow us to draw better powered conclusions on the association between ANM and later-life health outcomes, as well as give us an insight into how we can utilise the knowledge on the epidemiological risk factors to design effective public health intervention strategies.

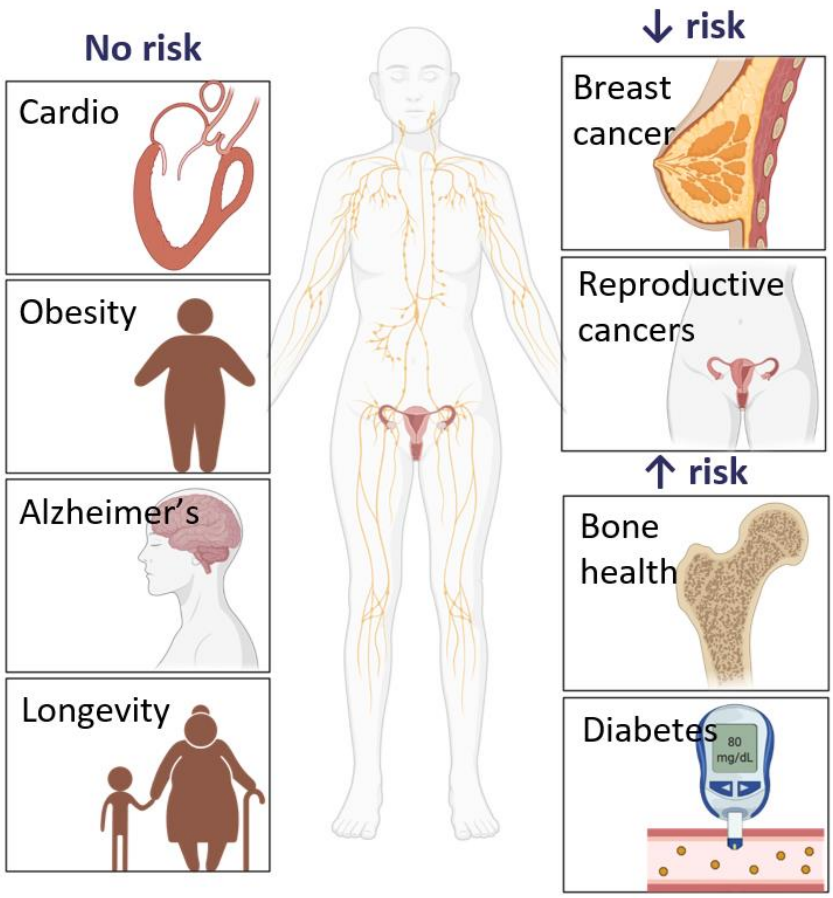

Figure 1.9: Earlier menopause timing and associated health outcomes from Mendelian Randomisation study.

One interpretation of these findings is that premature or early menopause is the first step in a chain of causality leading to tissue or organ dysfunctions and lesions via hormonal mechanisms[143]. Alternative yet complementary hypothesis suggests that premature menopause is the result of an accelerated aging process determined by genetic or non-genetic causes and involving all tissues and organs throughout the body, including the ovaries. Both hypotheses point towards 5 times faster ageing of ovaries than any other organ in the human body, speculating that this 'neglected' organ could be serving as a 'biological marker' of overall health in women [144–146].

The number of menopausal women worldwide is estimated to reach 1.1 billion by 2025, while there are 1.9 million pre-menopausal women diagnosed with cancer and thus at risk from chemotherapy-induced ovarian failure (CIOF) and infertility. These numbers are hard to ignore given the strong link between reproductive longevity and related health outcomes. Current research sheds light on underlying biological mechanisms and interesting interplay between DDR and hormonal regulation, further highlighting the need for better understanding of the regulators and physiological mechanisms involved in reproductive ageing.

Collectively, the current state of reproductive ageing research provides strong insights into the biological processes of reproductive ageing in women, how they can be manipulated to extend reproductive life, and what the consequences of this might be at a population level. These findings have the potential to guide future experimental work aimed towards identification of new therapies for enhancing reproductive function and preserving fertility in women, as well as designing intervention strategies to prevent or diminish menopause-related health outcomes.

1.6 Areas of opportunities and Aims

This thesis aims to advance the understanding of reproductive longevity by combining clinical and biomarker data from large-scale population studies within a multi-omic approach together with the robust functional models to identify novel genetic determinants and mechanisms underlying ovarian ageing in women.

In summary and as previously discussed, I describe five distinct projects to address the following questions:

Chapter 3: What is the role of rare damaging genetic variation in the timing of menopause?
Chapter 4: Are reported monogenic causes of primary ovarian insufficiency valid?
Chapter 5: How does ovarian ageing impact the proteomic profile in women?
Chapter 6: Is there a shared genetic architecture between the beginning and end of female reproductive lifespan?
Chapter 7: What is the role of *PARP-1* in gametogenesis and ovarian function?

A better understanding of how and when molecular processes influence the establishment and decline of the ovarian reserve will inform future strategies for treating infertility and preserving fertility. Ultimately addressing these aims will help women make more informed reproductive choices, reduce the number of women undergoing invasive, painful and expensive IVF treatment, and enable timely management of menopause-related conditions. Addressing the issue of infertility is also of huge socio-economic importance. Many countries are already seeing fertility rates below replacement levels, meaning falling populations will soon become the norm. A global decline in population will have inevitable consequences on the economy, leading to high unemployment, ageing populations and thus a rising tax burden. Finding practical, accessible and non-invasive approaches to overcoming infertility brought about by advanced maternal age or certain medical conditions can provide personalised reproductive approaches and maintain economic growth.

CHAPTER 2

Methods used to analyse human genomic and proteomic data

Summary

Chapter 2 describes the common methods and datasets used across different chapters of this thesis to analyse human genomic and proteomic data coming from large-scale population studies. It gives a methodological overview of the steps undertaken, as well as the rationale behind choosing specific bioinformatic tools to answer biological questions of interest.

2.1 UK Biobank: a resource for deciphering the genetic architecture of reproductive ageing

2.1.1 About the UKBB study

The UKBB study is a large prospective cohort of ~500,000 individuals, out of which 245,820 are female participants, living in the United Kingdom, aged between 40 and 69 at baseline, and registered with an National Health Service (NHS) general practice (GP) service[147,148]. A total of 503,325 participants were recruited in the period between 2006 and 2010 and attended one of 22 assessment centres across the UK for baseline data collection. The data included information on individual's genotypes and phenotypes with the aim to enable identification of genetic and non-genetic determinants of various health outcomes (**Figure 2.1**)[147,149]. Data collection and processing are extensively described elsewhere[147,148], while sections **2.1.1.1** and **2.1.1.2** give a high-level overview of the process.

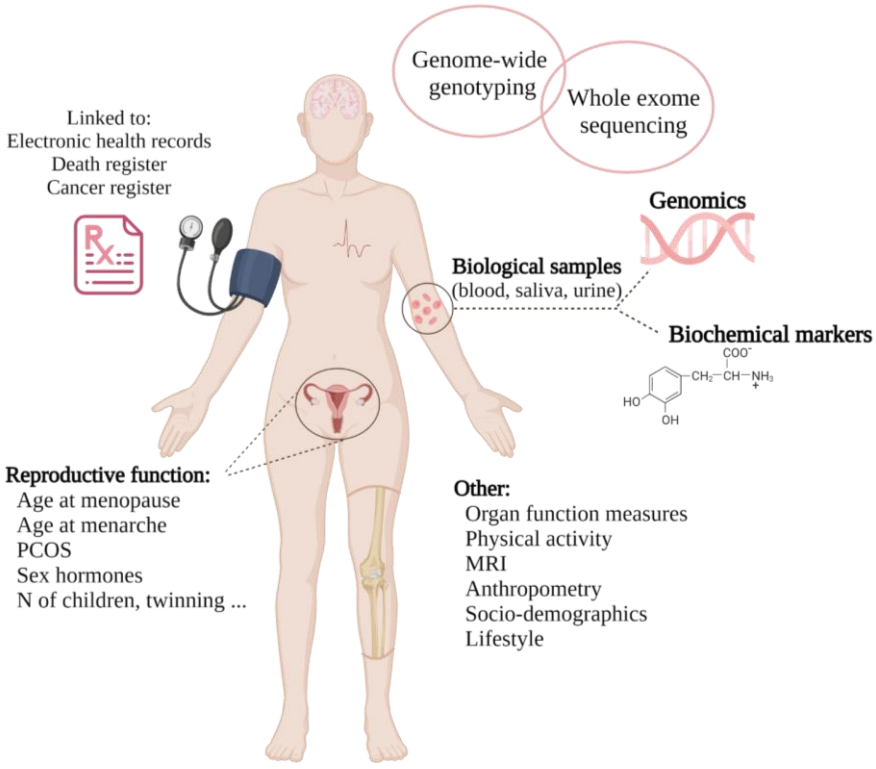

Figure 2.1: UK Biobank - a large-scale prospective epidemiological resource. *A large number of health outcomes and genotype data have been recorded to study relationships between individuals' genetic profiles and their susceptibility to those outcomes. The figure is adapted from Bycroft et al (2018)*[147].

2.1.1.1 Baseline and follow-up assessments

All participants attended an initial 2 to 3-hour research assessment, which included an online touchscreen questionnaire, a face-to-face interview and a collection of physical measurements and biological samples (blood, saliva, urine) (**Figure 2.1**)[148,149]. The information collected via questionnaire and interview included data on current and past health outcomes, both of an individual and family history, lifestyle exposures, psychological well-being, cognitive function and socioeconomic factors[149]. In addition to data collected during the assessment centre visits, the study was linked to NHS electronic health records, death and cancer registry that provided information on incident and prevalent healthcare events. Collected phenotype data also included information on reproductive function, including age at menopause and menarche, sex hormones levels, malignant neoplasms of reproductive organs and many others - health outcomes which are in focus of this thesis. Details on individual measurements relevant to specific studies in this thesis are described in the corresponding chapter.

2.1.1.2 Analysis of common genetic variants: Genotyping and Imputation

Genetic data were obtained from a subset of the cohort (N = 49,950) using the UKBB Lung Exome Variant Evaluation (UK BiLEVE) study array at 807,411 probes[150], while the rest of the cohort (N = 438,427) was genotyped at 825,927 probes using the UKBB Axiom array from Affymetrix. The two arrays share 95% of their probes, yet the Axiom array was designed to assay more variants, in particular insertions and deletions[147]. The selection of probes aimed to assay both common and low frequency variants, as well as variants previously suggested to be important in other phenotypes, such as cancer, autoimmune disease or blood phenotypes[147]. Blood samples were collected from participants during their assessment centre visit and then shipped to Affymetrix for genotyping. Sample retrieval and DNA extraction are described in more detail in Welsh *et al*, 2017[151]. Heterogeneous ancestry characterises the UKBB cohort - of genotyped individuals 94% reported their ancestry as 'White' with the remaining 6% as Asian, Black, Chinese, mixed or unknown ancestry.

The quality control (QC) was performed on the probe/marker- and sample-based level. The exclusion criteria for poor quality markers involved missing rate, heterozygosity adjusted for population ancestry effects, and non-XX and XY sex chromosome karyotypes (samples with mismatch between self-reported and genotypic sex or with potential aneuploidy in sex chromosomes). As part of the UKBB QC, SNPs

that had a missing rate >5% and MAF <1% were excluded. Finally, genotype information was available on 488,377 participants at 805,426 biallelic SNPs and insertions/deletions (indels).

The Haplotype Reference Consortium (HRC) reference panel was used for the imputation of haplotypes[152], with separate imputation done by IMPUTE4 software using the combined UK10K and 1000 Genomes phase 3 reference panels[71,153]. Last two panels were selected as they contain a high percentage of European ancestry individuals and a small subset of individuals with diverse ancestry, thus having a similar ancestry distribution to the UKBB cohort. The final dataset included 93,095,623 autosomal genetic variants and 3,963,705 X chromosome variants in 487,442 individuals. European ancestry individuals were identified by projecting samples on the two major principal components (PCs) from the 1000 Genomes cohort[71] and selecting samples which fall in the European cluster (CEU) identified by sequencing of European ancestry individuals by the 1000 Genomes project. This analysis resulted in identification of 463,844 European ancestry individuals[147].

2.2 Identifying gene variant associations with complex diseases

2.2.1 Genome-wide association study (GWAS)

As introduced in **Section 1.3**, GWAS systematically assesses the association between genetic variants and phenotype of interest across the genome with an aim to identify genetic alterations that determine or modify the susceptibility of that trait at the genome-wide significance level of $P < 5*10^{-8}$ [75]. For this thesis, GWAS analysis was used to test for association of genetic variants in UKBB cohort with various reproductive and metabolic health outcomes using linear mixed models, BOLT-LMM. Depending on the specific case and as described in following chapters, association analysis was performed in sex-combined and/or sex-stratified samples. The trait specific summary statistic results obtained via BOLT-LMM are described in following chapters. It is important to note that GWAS results for these traits were already generated by other members of our research group. Next sections will describe a standard GWAS procedure conducted for all traits of interest.

2.2.1.1 Linear Mixed Models: BOLT-LMM

The BOLT-LMM algorithm uses linear mixed models (LMM) to examine the association between the genetic variants and traits of interest[154]. BOLT-LMM represents the method of choice due to its ability to account for population stratification and cryptic relatedness, which tend to inflate false positive or negative results. More specifically, in case of population stratification the correlated ancestry within a

stratum of population may have common environmental or genetic exposures. This limits our understanding if the variant is associated with an outcome because of a genetically mediated mechanism, or because an allele of that variant happens to be common in a stratum of the population where the presence of the measured phenotypic outcome is common by chance. In addition to population stratification, closely related individuals in the sample population, 'cryptic relatedness' can lead to similar inflation[155].

An increase in statistical power observed in linear mixed models is achieved by jointly modelling all genotyped markers. To further optimise power, in addition to the infinitesimal model used by standard mixed models, BOLT-LMM computes the non-infinitesimal model. These two models differ based on the Gaussian distribution assumption, where unlike the infinitesimal model that assumes that all variants are causal with small effect sizes ('additive per allele effects'), the non-infinitesimal one models SNP effects with non-Gaussian prior distributions, thus allowing better accommodation of the SNPs with small and large effects[154]. In reality, traits usually have a few associated variants with large effect sizes compared to many associations with smaller effect sizes. Therefore empirically, effect sizes are not Gaussian distributed. In this thesis, we mostly used the non-infinitesimal model, unless stated otherwise.

We made the genetic discoveries using BOLT-LMM (v2.3.4), following next steps:

(1a) estimation of variance parameters
(1b) computation of infinitesimal mixed-model association statistics (BOLT-LMM inf)
(2a) estimation of Gaussian mixture-model parameters
(2b) computation of Gaussian mixture model association statistics (BOLT-LMM)

2.2.1.2 Signal selection

The LD between variants makes it challenging to understand how many independent association signals are present in a locus. To address this, we first implemented a distance-based clumping method that allows the selection of the genetic variant most strongly associated with the phenotype within a 1-Mb genomic region. To select statistically significant signals we apply the genome-wide significance *P*-value threshold of $5*10^{-8}$, MAF > 0.1% and an imputation quality score > 0.5. Signals with the lowest *P*-value, which were not correlated with other signals in each LD block (r^2<0.05) were defined as the lead SNPs. The lead SNPs are assumed to explain the maximum amount of trait variation tagged by that region, as the other signals in a region would show the association by being in LD with that leading SNP. However, the total phenotypic variation explained by that LD block might be underestimated due to the potential

existence of multiple independent signals that contribute to the overall heritability. To identify independent secondary association signals within each LD block, we introduced approximate conditional analysis implemented in genome-wide complex trait analysis (GCTA)[156]. GCTA relies on a reference dataset that estimates LD between variants and selects additional signals significantly associated with the phenotype, conditioning on the effect of the primary signal at that locus. Secondary signals had to meet the following criteria: (1) $P \leq 5\times10^{-8}$ in both pre- and post-conditional analyses, (2) uncorrelated with another signal ($r^2 < 0.05$) and (3) its beta estimate changed by < 20% between pre- and post-conditional models[157]. However, it is important to know that conditional analysis cannot determine which variants are casual for the association signal. Causality can only be truly determined with a downstream follow up experiment. The problem of causality could also be addressed via genomic technologies that study the role of rare, protein coding variants with larger impact on the trait of interest, such as whole exome sequencing (WES), which is discussed in future chapters.

2.2.2 ReproGen consortium genome-wide association meta-analysis

When considering common variant GWAS, most of this thesis will implement the ANM signals obtained from the currently largest ANM GWAS conducted in the ReproGen consortium and reported in Ruth *et al*, 2021[82]. The ReproGen consortium is an international collaboration of investigators interested in the genetics of reproductive ageing. Detailed description of this ANM GWAS study can be found in Ruth *et al*, 2021[82], while the following section will give a brief overview of the methodology applied.

2.2.2.1 Phenotype definition

The ANM was derived from self-reported questionnaire data by each study in the ReproGen consortium as described in Ruth *et al*, 2021[82]. The ANM was defined as the age at last naturally occurring menstrual period followed by at least 12 consecutive months of amenorrhea. Women with menopause caused by hysterectomy, bilateral ovariectomy, radiation or chemotherapy, and those using HRT before menopause were excluded from the study. All participants provided written informed consent and the study protocol was approved by the institutional review board at each parent institution.

2.2.2.2 Genome-wide association study meta-analysis

A genome-wide meta-analysis for ANM was performed on summary statistics in women of European ancestry from analyses in three strata. These included:

(1) Meta-analysis of 1000 Genomes imputed studies

(2) Meta-analysis of samples from the Breast Cancer Association Consortium (BCAC: http://bcac.ccge.medschl.cam.ac.uk)

(3) UKBB GWAS

Meta-analysis of **1000 Genomes imputed studies** was carried out including SNPs with imputation quality ≥ 0.4 and MAF ≥ 0.001. GWAS was performed in each individual study using a two-tailed additive linear regression model adjusted for genetic principal components/relationship matrix depending on the software used, without GC correction. Variants present in at least half of datasets for either the autosomes or for chromosome X were taken forward to the overall meta-analysis, resulting in ∼10.9 million variants.

GWAS summary statistics for the **BCAC data** were provided as four datasets, containing breast cancer cases and controls, with each genotyped on the iCOGs and OncoArray genotyping arrays. Summary statistics from the four BCAC datasets were meta-analysed, including variants with imputation quality ≥0.4 and MA F≥ 0.001. Variants in two or more of the four datasets were taken forward to the overall meta-analysis, resulting in ∼14.5 million variants.

GWAS in **UKBB** was carried out by applying a linear mixed model in BOLT-LMM, as described in **Section 2.2.1.1**, to adjust for population structure and relatedness, study centre and data release. UKBB summary statistics taken forward to the overall meta-analyses were for ∼16.6 million variants with imputation quality ≥0.5 and MAF≥0.001. Genome-wide significance was set at $P \leq 5*10^{-8}$ and signal selection was performed using distance-based clumping and approximate conditional analysis, as described in Section **2.2.1.2**.

Variants which were present in at least two of the three strata were included in the final ReproGen meta-analysis. Genome-wide array data, imputed to ∼13.1 million genetic variants with MAF ≥0.1%, were available to 201,323 women of European ancestry. All meta-analyses were performed using inverse-variance transformation models in METAL (https://genome.sph.umich.edu/wiki/METAL_Documentation) without GC correction. In total 38,707 genetic variants were associated with ANM at genome-wide significance ($P \leq 5*10^{-8}$), which were ultimately resolved to 290 statistically independent signals.

2.2.3 Studying rare genetic variation using WES data in UKBB

Human genetic studies that relied on GWAS array genotyping have identified genetic markers associated with menopause timing and thus provided substantial insight into the biological mechanisms governing ovarian ageing. However, these approaches have largely been restricted to assessing common genetic variation which brings previously described challenges related to the identification of responsible genes and biological pathways involved in regulation of the trait of interest. WES provides the opportunity to directly assess protein-altering variants, genetic variations with large and thus more readily interpretable functional consequences, which provide an insight into the biological mechanisms and thus potential therapeutic applications.

Therefore, in addition to GWAS, we utilised WES[158] to explore protein-altering variants and their consequences in 454,787 participants in the UKBB study[147,159]. The study identified 12 million coding variants, including around 1 million loss-of-function (LoF) and around 1.8 million deleterious missense variants, across the coding regions of 18,893 genes, of which 99.6% were rare variants (MAF < 1% across all ancestries). The roles of these genetic alterations were then explored in different reproductive, metabolic phenotypes and cancer.

2.2.3.1 UKBB sample preparation and sequencing

The exome sequencing of DNA samples from the UKBB study was performed by Regeneron Genetics Centre [160]. The detailed description of the procedure can be found elsewhere [159,160]. In brief, genomic DNA samples were transferred to the Regeneron Genetics Centre from the UKBB before sample preparation. The samples were sequenced using 75-base-pair paired-end reads with two 10-base-pair index reads on the Illumina NovaSeq 6000 platform using S2 (first 50,000 samples) or S4 (all other samples) flow cells.

Sample read mapping and variant calling, aggregation and quality control were performed using the SPB protocol described elsewhere[160]. WES reads were mapped with Burrows-Wheeler Aligner Maximum Entropy Method (BWA MEM) to the human genome build 38 (hg38) reference genome. No-call genotypes were defined as SNV genotypes with read depth (DP) < 7 and indel genotypes with DP < 10. Variants that met the following criteria were kept for further analysis: (1) at least one homozygous variant carrier; or (2) at least one heterozygous variant carrier with an allele balance (AB) greater than the cut-off ($AB \geq 0.15$ for SNVs and $AB \geq 0.20$ for indels). The samples were further excluded if they showed: (1) disagreement between genetically determined and reported sex, (2) high rates of heterozygosity or

contamination, (3) low sequence coverage (less than 80% of targeted bases achieving 20× coverage) or genetically identified sample duplicates, and (4) WES variants discordant with GWAS array genotypes, resulting in the exclusion of 1,105 individuals. Finally, 454,787 samples were used to compile a pVCF for downstream analysis, using the GLnexus joint genotyping tool.

2.2.3.2 Data Processing and Quality Control

To conduct rare variant burden analyses, we obtained data on WES for 454,787 individuals from the UKBB study, as described above[159]. Participants were excluded based on excess heterozygosity, autosomal variant missingness on genotyping arrays ≥ 5%, or inclusion in the subset of phased samples as defined in Bycroft *et al*[147]. Analysis was restricted to participants with European genetic ancestry, leaving a total of 421,065 individuals. Variant QC and annotation were performed using the UKBB Research Analysis Platform (RAP; https://ukbiobank.dnanexus.com/), a cloud-based central data repository for UKBB WES and phenotypic data. Besides the QC described briefly in **Section 2.2.3.1** and in detail by Backman *et al.*[159], we performed additional steps using custom applets designed for the RAP. Firstly, we processed provided population-level VCF files by splitting and left-correcting multi-allelic variants into separate alleles using 'bcftools norm'[161]. Secondly, we performed genotype-level filtering applying 'bcftools filter' separately for Single Nucleotide Variants (SNVs) and Insertions/Deletions (InDels) using a missingness-based approach. We set to missing (i.e. ./.) all SNV genotypes with depth < 7 and genotype quality < 20 or InDel genotypes with a depth < 10 and genotype quality < 20. Next, we applied a binomial test to assess an expected alternate allele contribution of 50% for heterozygous SNVs; we set to missing all SNV genotypes with a binomial test *P* value ≤ $1x10^{-3}$. Following genotype-level filtering we recalculated the proportion of individuals with a missing genotype for each variant and filtered all variants with a missingness value > 50%.

The variant annotation was performed using the ENSEMBL Variant Effect Predictor (VEP) v104 (**Figure 2.2**)[162] with the '--everything' flag and plugins for Combined Annotation Dependent Depletion (CADD)[163] and Loss-of-Function Transcript Effect Estimator (LOFTEE)[164] enabled. For each variant we prioritised the highest impact individual consequence as defined by VEP and one ENSEMBL transcript as determined by whether or not the annotated transcript was protein-coding, MANE select v0.97, or the VEP canonical transcript.

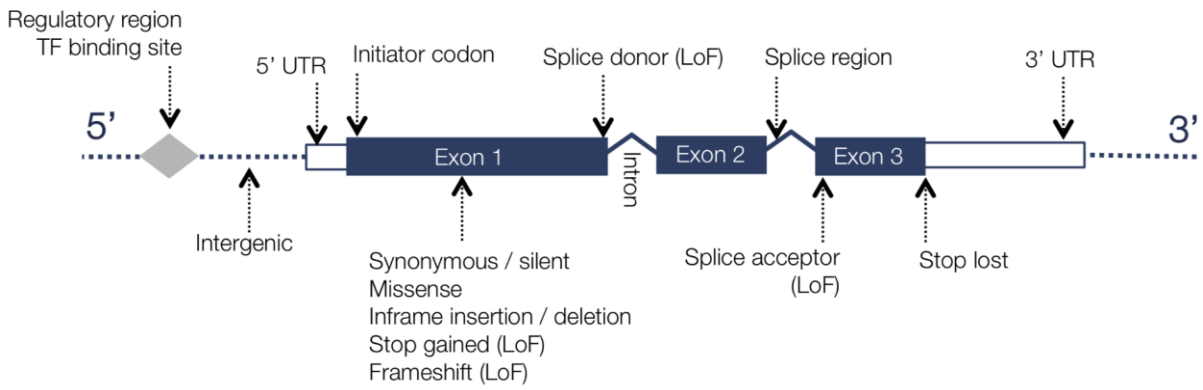

Figure 2.2: Functional consequences of various types of genetic variants. *The diagram illustrates the set of functional consequence terms given by the Ensembl Variant Effect Predictor (VEP) tool. A splice donor is a splice variant that changes the invariable 2-base region at the 5' end of an intron. A splice acceptor is a splice variant that changes the invariable 2-base region at the 3' end of an intron. A splice region is a sequence variant in which a change has occurred within the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron, but not at the donor/acceptor splice sites*

Following annotation, variants were categorised based on their predicted impact on the annotated transcript. Protein Truncating Variants (PTVs) were defined as all variants annotated as stop gained, frameshift, splice acceptor, and splice donor. Missense variant consequences are identical to those defined by VEP. Only autosomal or chrX variants within ENSEMBL protein-coding transcripts and within transcripts included on the UKBB exome-sequencing assay[159] were retained for subsequent burden testing.

2.2.3.3 Exome-wide association analyses in UKBB

In order to perform rare variant gene burden tests, we used a custom implementation of BOLT-LMM v2.3.6[154] for the RAP. Two primary inputs are required by BOLT-LMM: 1) a set of genotypes with minor allele count (MAC) > 100 derived from genotyping arrays to construct a null linear mixed effects model and 2) a larger set of variants collapsed on ENSEMBL transcript to perform association tests. For the former, we queried genotyping data available on the RAP and restricted to an identical set of individuals included for rare variant association tests. For the latter, and as BOLT-LMM expects imputed genotyping data as input rather than per-gene carrier status, we created dummy genotype files where each variant represents one gene and individuals with a qualifying variant within that gene are coded as heterozygous, regardless of the number of variants that individual has in that gene.

To test a range of variant annotation categories with MAF < 0.1%, we created dummy genotype files for high confidence (HC) PTVs as defined by LOFTEE, all missense variants, missense variants with CADD ≥ 25, and damaging variants that included both high confidence PTVs and missense variants with CADD ≥ 25. For each phenotype tested, BOLT-LMM was then run with default parameters other than the inclusion of the 'lmmInfOnly' flag. To derive association statistics for individual markers, we also provided all 26,657,229 individual markers regardless of filtering status as input to BOLT-LMM. All tested phenotypes were run as continuous traits corrected by age, age^2, sex, the first ten genetic PCs as calculated in Bycroft *et al*[147] and study participant exome sequence batch as a categorical covariate (either 50k, 200k, or 450k).

To generate accurate odds ratio (OR) and standard error (SE) estimates for binary traits, we also implemented a generalised linear model using the 'statsmodels'[165] for Python in a three step process. First, a null model was run with the phenotype as a continuous trait, corrected for control covariates as described above. Second, we regressed carrier status for individual genes on the residuals of the null model to obtain a preliminary *P* value. Thirdly, all genes were again tested using a full model to obtain odds ratios and standard errors with the family set to 'binomial'. Generalised linear models used identical input to BOLT-LMM converted to a sparse matrix.

For phenotype definitions and additional details, please refer to specific Chapters.

2.3 Exploring causal estimation through Mendelian Randomisation analyses

2.3.1. Using genetic variants in causal estimation

In order to study patterns of health and disease at the population level, science has implemented different epidemiological study frameworks. Although being recognized as a gold standard methodology, RCTs are characterised by certain limitation factors that restrict their broad application for testing scientific hypotheses, including cost, time-consumption and ethics[166]. Observational studies have been used to establish correlations between numerous exposures and complex diseases in the past, however causality associations drawn from these studies often faced different challenges, including unmeasured confounding and reverse causation that questioned their reliability[167]. To address reverse causation and confounding, Katan proposed the Mendelian Randomisation (MR) method, also called 'naturally randomised trial'[168–170]. MR relies on random assortment of alleles during gamete production and

fertilisation, as described in Mendel's second law. This natural, non-modifiable genetic variation across individuals mimics the rationale behind RCT (**Figure 2.3**) and it is now widely used in genetic epidemiology for inferring causality of an exposure on complex disease outcomes. Previous research on the genetic architecture of complex diseases has revealed that multiple genes and common frequency alleles of small effects act through an additive effect in combination with the environmental factors to define a certain trait. Even though they explain only part of heritability of a certain trait, using multiple variants as genetic instruments will maximise our potential and power to assess causal inference between the exposure and outcome.

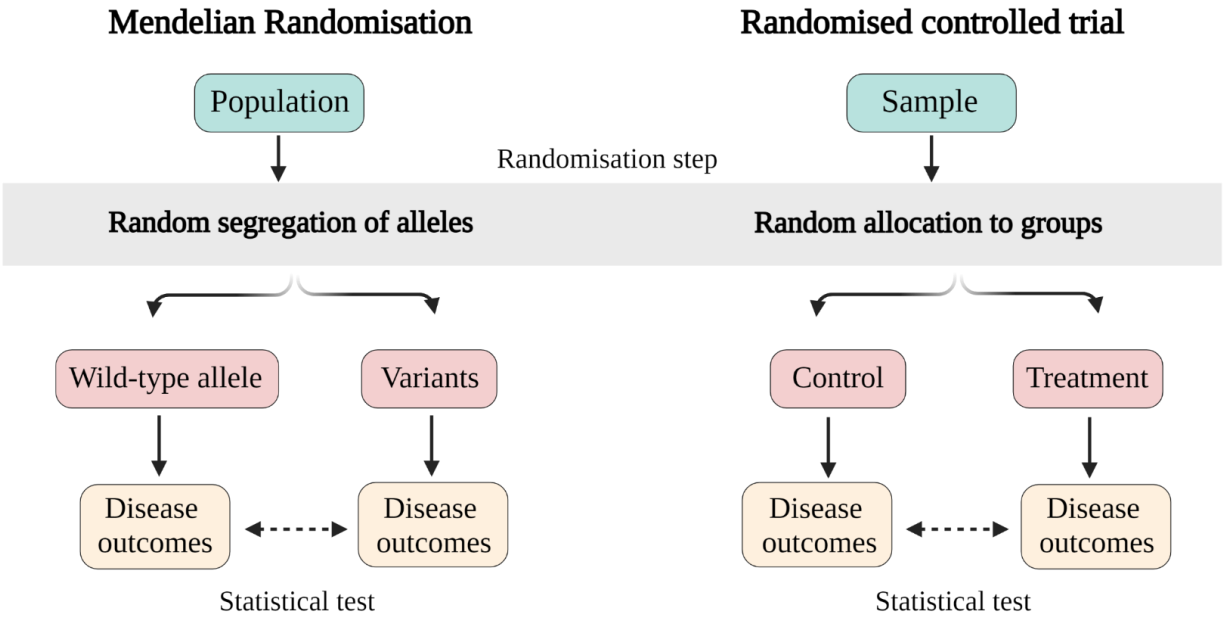

Figure 2.3: Comparison between Mendelian Randomisation study and Randomised controlled trial. *This diagram compares and contrasts the methodological rationale behind MR and RCTs. Unlike RCTs, MR is based on randomization that is dependent on the random allocation of alleles at birth. MR causal inference implies that a change in the exposure caused by genetic variation has the same effect on the outcome as a change in that exposure caused by environmental factors.*

This thesis applied various MR frameworks to address bidirectional causal estimation between reproductive ageing, early-life exposures and later life health outcomes, as described in following chapters.

2.3.2 Instrumental variable selection

MR analysis was applied with an aim to investigate the likelihood of a causal effect of the exposure of interest, i.e. a risk factor, on the outcome of interest. MR uses genetic variants, which are significantly associated with an exposure, as instrumental variables (IVs) to test this causality across the population sample. It measures the effect of these variants in both exposure and outcome. For a genetic variant to be a reliable instrument, following assumptions have to be met (**Figure 2.4**)[168]:

(1) **Relevance assumption:** genetic instrument must be associated with the exposure of interest

(2) **Independence assumption:** genetic instrument must not be associated with any other competing risk factor that is a confounder

(3) **Exclusivity assumption:** genetic instrument must not be associated with the outcome, except via the causal pathway through the exposure of interest

Assumption **(1)** can be tested with a standard GWAS analysis which determines significance of association between a genetic variant and phenotype.

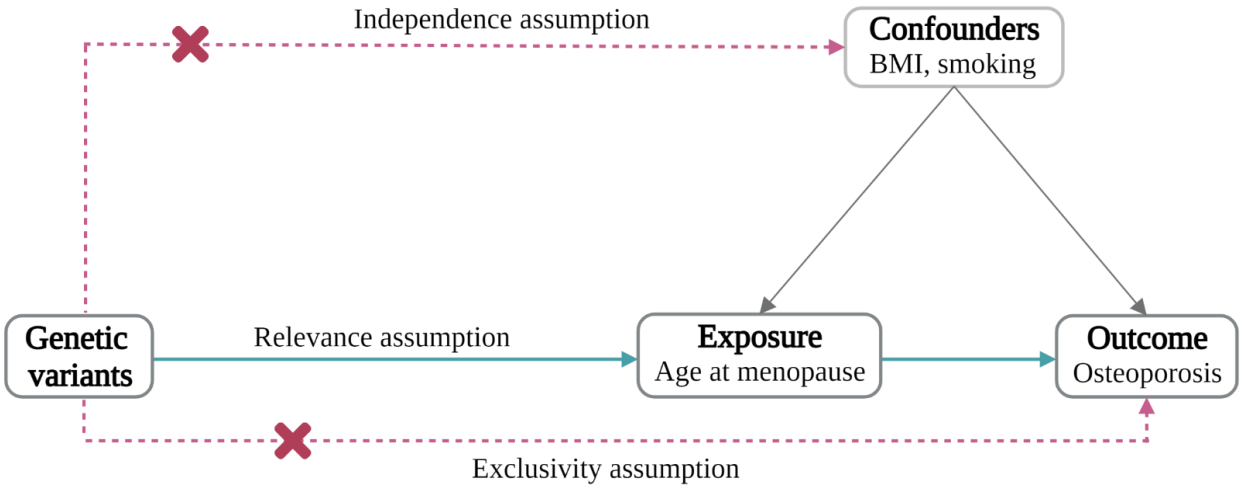

Figure 2.4: Schematic representation of Mendelian Randomisation theory. *Genetic variants which are associated with the exposure may serve as instrumental variables to assess the causal inference between the exposure and the outcome of interest (blue line), assuming that MR assumptions are met (pink dashed lines). The tested association between the exposure and the outcome may be confounded by unmeasured factors (grey line).*

Potential for bias in MR could arise in case IVs are measured in the outcome sample that is the same as discovery one. This thesis addressed causal estimation using an agnostic polygenic two-sample MR approach, where the association between IVs and exposure was estimated in one dataset, and IVs and

outcome in a second, independent dataset[171]. In cases where a particular signal was not present in the outcome GWAS, we searched for proxies in the UKBB white European dataset that are in LD with the target SNP within the LD window of 1 Mb and $r^2 > 0.5$. The proxy with the highest r^2 value was chosen as an alternative genetic instrument.

2.3.3 Variant harmonisation

In order to ensure that genetic associations are expressed per additional copy of the same allele, the effect allele and the effect allele frequency were compared between the exposure and outcome dataset for each SNP and aligned according to exposure-increasing beta value. In case of palindromic SNPs, where the allele frequency was close to 50% and thus it was not possible to verify the allele orientation, the variant was dropped out.

2.3.4 MR frameworks

2.3.4.1 Primary analysis

MR analysis was conducted using the inverse-variance weighted (IVW) model as the primary model due to the highest statistical power[172]. IVW calculates the association estimate between the exposure and the outcome through weighted regression of the effect of genetic variants on the outcome with the effect of the same variance on the exposure[173]. In **formula A**, β_l is the effect estimated using genetic variant l. The calculation will take into account estimates of π_l, the estimated effect of genetic variant *l* on the exposure with variance $\sigma^2_{x,l}$, and Γ_l, the estimated effect of genetic variant *l* on the outcome with variance $\sigma^2_{y,l}$. These individual ratios are weighted by their associated uncertainty in formula B. The IVW estimator (β_{IVW}) was then computed as (**formula B**):

(A) **(B)**

$$\hat{\beta}_l = \frac{\hat{\Gamma}_l}{\hat{\pi}_l} \qquad \hat{\beta}_{\mathrm{IVW}} = \frac{\sum_{l=1}^{L} \hat{\pi}_l \hat{\Gamma}_l \sigma_{y,l}^{-2}}{\sum_{l=1}^{L} \hat{\pi}_l^2 \sigma_{y,l}^{-2}}$$

Even though it has the greatest statistical power when the aforementioned assumptions are satisfied, 0% breakdown level implies that the estimates of causal effect will become biased even if one genetic variant does not satisfy the assumptions due to method's lack of power to correct for heterogeneity in outcome risk estimates between individual variants [174]. The bias most commonly arises due to the presence of

horizontal pleiotropy, where the genetic variant used as an instrument does not influence the outcome exclusively through the exposure of interest (**Figure 2.4**)[172,174]. The only exception where the bias is absent is when the pleiotropy is balanced, i.e. the overall sum of the pleiotropic effects across the genetic variants is summed to zero (**Figure 2.5**).

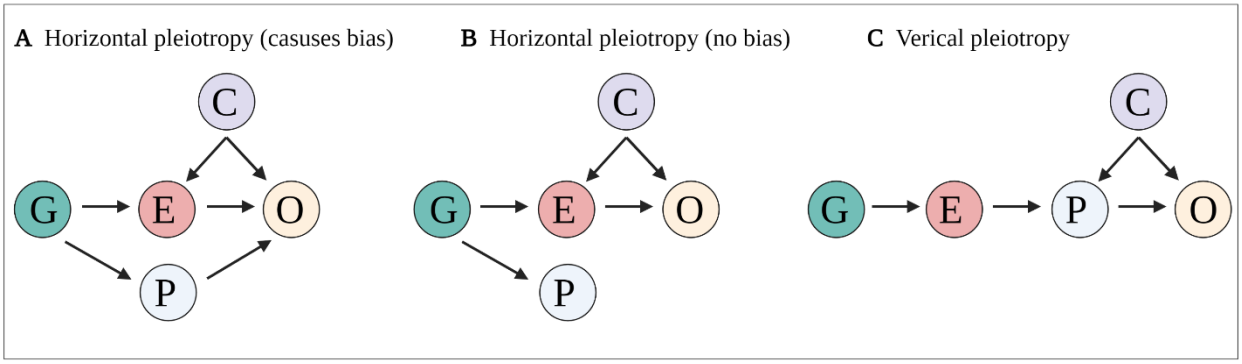

Figure 2.5: Pleiotropy types in Mendelian Randomisation. *Diagram demonstrates different types of pleiotropy in Mendelian randomization (MR), where G is a genetic variant or set of genetic variants associated with the exposure, E is the exposure of interest, O is the outcome of interest, C is an unmeasured confounder and P is another (potentially unmeasured) phenotype that is also associated with the genetic variants.* ***(A)*** *Horizontal pleiotropy with bias occurs when a IVs or 'G' are associated with multiple health outcomes that lie on different biological pathways, thus violating assumption 3;* ***(B)*** *Horizontal pleiotropy with no bias occurs when the genetic variants are not associated with other phenotypes on the pathway to the outcome;* ***(C)*** *Vertical pleiotropy happens when another phenotypes lies on the genetic variant–exposure–outcome pathway, yet it does not bias the MR result for the trait of interest.*

2.3.4.2 Sensitivity analysis

Various methods have been designed to test the robustness of MR associations and account for violations of the MR assumptions[175,176]. One such example is MR Egger used to identify and correct for unbalanced heterogeneity, i.e. 'horizontal pleiotropy'[177]. As we are modelling complex biological systems, variants are generally pleiotropic, meaning they influence multiple traits and phenotypes. The MR Egger approach allows the genetic variants to have pleiotropic effects, as long as they are independent of the variant-exposure association and do not alter the magnitude of association with the outcome, so-called 'balanced pleiotropy[178]. In the formula below, β_{XI} represents the estimated effect on the exposure, β_{YI} is estimated effect on the outcome, γ_E is the estimated causal effect of the exposure on the outcome and γ_0 represents the intercept parameter:

$$\beta_{YI} = \gamma_0 + \gamma_E \beta_{XI}$$

The MR Egger intercept, with $P < 0.05$, was used as an indicator of the overall pleiotropic effect across instrumental variables[177]. However, allowing the intercept term results in MR Egger being less powered in

causal estimation than IVW as the outlying and influential data points have strong effect on the precision of the effect estimate.

In addition, we applied weighted median (WM) and penalised weighted median (PWM) models to correct for balanced heterogeneity[179]. If at least half of the variants are valid IVs in WM, an unbiased causal estimation robust to outliers will be generated by calculating the median value from a distribution of ratio estimates from all IVs[180]. In case of heterogeneity of causal ratios, PWM reduces the effect of outlying variants by downweighting the outlying IVs based on Cochrane's Q statistics[179].

In an ideal, 'no-pleiotropy' scenario each SNP would influence the outcome proportional to its impact on the exposure, generating equal proportional factor across all instrumental variables. Genetic variants that satisfy this assumption will have homogenous causal ratio estimates, and variants with outlying effects would suggest the potential impact of horizontal pleiotropy. Therefore, in this thesis additional variants were excluded if they were recognized as outliers by the MR Radial method based on Rucker's Q statistics[181].

The overall MR result was considered as significant ($P < 0.05$) based on the consistency of all estimated effects across different primary and sensitivity models applied.

CHAPTER 3

Rare damaging variants in *ZNF518A* reduce menopause timing in carriers by six years

Summary

Genome-wide association studies have identified many common polymorphisms that modify the timing of menopause in women and shed light towards underlying biological mechanisms. However, GWAS is limited in the ability to identify causal genes because the majority of associated common haplotypes contain no coding variants, and the closest gene to a noncoding variant is not reliably the causal gene in the absence of functional evidence. In this Chapter, we use whole exome sequence data to study rare protein-coding variants associated with menopause timing in ~120K women in the UKBB, and implicate 5 novel genes with effect sizes up to ~5 times larger than previously discovered in analyses of common variants. Notably, we found that heterozygous loss of *ZNF518A* reduces menopause timing by nearly 6 years in carriers, an effect larger than most of the genetic variation currently tested in clinical genetics for premature ovarian ageing. Furthermore, we provided evidence that *ZNF518A* is a master transcriptional regulator of ovarian development and establishment of the ovarian reserve in foetal life. My results highlight novel mechanisms involved in age at natural menopause aetiology, identify further links between ovarian ageing and cancer susceptibility and demonstrate that genetic susceptibility to earlier ovarian ageing in women increases *de novo* mutation rate in their offspring, providing direct example of a mechanism for the maternal genome influencing child health.

Contributions and Collaborations

Dr Eugene Gardner created a pipeline on the UKBB DNAnexus RAP to process, annotate UKBB WES data and perform variant QC. This pipeline is now used by multiple groups at the MRC Epidemiology Unit when analysing UKBB WES data. Dr Katherine Kentistou prepared the ANM phenotypes. Dr Felix Day and I prepared various cancer phenotypes, Dr Day prepared age at menarche and sex hormones phenotypes, while Yajie Zhao prepared the telomere length phenotype. Both Dr Gardner and I conducted gene burden and variant-level association testing using BOLT-LMM using MAF < 0.1%, while I additionally conducted it for MAF < 1%. All data analysis and interpretation were done by me. I ran logistic regression analysis under the guidance and supervision of Dr Gardner. Gene expression analysis was run by Ajuna Azad, interpreted by Ajuna Azad and me, all under the supervision of Prof Eva Hoffmann. Functional enrichment tests for *ZNF518A* were run by me using fGWAS and SLDP, while functional analysis of *ZNF518A* binding sites were mainly done by Dr Nick Owens with my input on some analysis. De novo mutation rate analyses in Genomics England data were run by Dr Qin Qin Huang supervised by Prof Hilary Martin and Prof Matthew Hurles, while I conducted Mendelian Randomisation analysis. Prof John Perry, Prof Ken Ong, Prof Anna Murray, Prof Eva Hoffmann, Prof Hilary Martin and

Prof Matthew Hurles provided valuable advice on the analyses and writing of the manuscript, currently under review at Nature.

3.1 Exploring 'hidden heritability' of reproductive longevity using whole exome sequencing

The variation in timing of menopause reflects a complex mix of genetic and environmental factors that population-based studies have begun to unravel. As described in **Chapter 1**, GWASs have successfully identified ~300 distinct common genomic loci associated with the timing of menopause[82]. These reported variants cumulatively explain 10%-12% of the variance in ANM and 31-38% of the overall estimated SNP heritability[82,109]. The majority of these loci implicated genes that regulate DDR, highlighting the particular sensitivity of oocytes to DNA damage due to the prolonged state of cell cycle arrest across the life-course[76,78,83,84,91,182–184].

Besides explaining only a fraction of heritability[87], these common genetic variants are limited in their ability to identify relevant genes because the majority of associated common haplotypes contain non-coding variants. Even though they are thought to play a role in gene expression regulation, it is unclear what genes these variants regulate, which further complicates the identification of causal mechanisms[87,185–187]. In addition, the correlation between multiple variants in a locus due to LD would make it challenging to distinguish the causal variants responsible for the association with the trait of interest[187]. This limitation is compounded by the difficulty to experimentally test large numbers of genes or variants using high throughput methods sensitive enough to detect small molecular effects. Consequently, this has hindered the translation of GWAS findings into effective mechanistic understanding and therapeutic solutions. However, part of the undiscovered heritability spans from rare, low frequency variants with higher impact[88]. GWASs are typically underpowered when applied to low-frequency or rare variants, unless sample sizes or effects are very large. These genetic alterations can be captured by evaluating aggregate association over multiple variants in a genomic region, so-called 'gene burden', using WES data that comprises of protein-coding genes[88,188–190]. Sequencing offers a significant advantage over array-based methods, with the potential to detect and genotype all variants present in a sample, not only those present on an array or imputation reference panel.

Genetic studies for ANM to date have largely focused on assessing common genetic variation, with little insight into the role of rarer, protein-coding variants. Initial WES analyses on 132,370 women in UKBB identified gene-based associations with ANM for *CHEK2*, *DCLRE1A*, *HELB*, *TOP3A*, *BRCA2* and *CLPB*[82,191]. In this Chapter, we aimed to explore the role of rare damaging variants in ovarian function and reproductive lifespan with much greater power than previously possible. Through a combination of enhanced phenotype curation, better powered statistical tests and assessment of different types of variant classes at lower allele frequency thresholds, we identified five novel genes harbouring rare variants of

large effect that have not previously been implicated in ovarian ageing. Furthermore, these observations were extended to show that women at increased genetic risk of earlier menopause have higher rates of *de novo* mutations in their offspring. These findings have the potential to contribute towards unlocking the opportunity for novel methods of prediction of the reproductive lifespan and fertility preservation.

3.2 Methods

3.2.1 Exome-wide association analyses in UKBB

The rare variant burden analyses of WES data from the UKBB study was performed following the methodology described in **Section 2.2.3**.

To test a range of variant annotation categories for MAF < 0.1%, we created dummy genotype files for HC PTVs as defined by LOFTEE, missense variants with CADD ≥ 25, and damaging variants that included both HC PTVs and missense variants with CADD ≥ 25. For the phenotype definitions used in this study, please refer to the below section on '**Phenotype derivation**'. The tested phenotypes were run as continuous traits corrected by age, age^2, sex, the first ten genetic PCs as calculated in Bycroft *et al*[147] and study participant ES batch as a categorical covariate (either 50k, 200k, or 450k). For discovery analysis in the primary trait of interest, ANM, we analysed 17,475 protein-coding genes with the minimum of 10 rare allele carriers in at least one of the masks tested using BOLT-LMM (**Figure 3.3**). To reduce the number of false positive results the significant gene-level associations for ANM were identified applying Bonferroni correction for the number of masks with MAC≥10 (N=46,251 masks) in 17,475 protein-coding genes (*P*: 0.05/46,251 = 1.08*10^{-6}) (**Figure 3.3**). Furthermore, in order to compare and explain potential differences between our WES results and the previously published ones[159,191], we ran the above described approach using MAF < 1%, a cutoff applied by other studies (**Appendix Table 3.2**). In addition, for some genes we also run Cox proportional hazard model using R package '*survival*' (https://github.com/therneau/survival).

3.2.2 Phenotype derivation

ANM phenotype was derived for individuals within UKBB, who were deemed to have undergone natural menopause, i.e. not affected by surgical or pharmaceutical interventions, as follows:
Firstly, European female participants (n=245,820) who indicated during any of the attended visits having had a hysterectomy were collated (fields 3591 and 2724) and their reported hysterectomy ages were

extracted (field 2824) and the median age was kept (n=47,218 and 46,260 with reported ages). The same procedure was followed for participants indicating having undergone a bilateral oophorectomy (surgery field 2834 and age field 3882, n=20,495 and 20,001 with reported ages).

For individuals having indicated the use of HRT (field 2814), HRT start and end ages were collated (fields 3536 and 3546, accordingly) across the different attended visits (n=98,104). In cases where the reported chronological HRT age at later attended visits was greater than that at previous visits, the later instances were prioritised, i.e. as they would potentially indicate an updated use of HRT. In cases where different HRT ages were reported, but not in chronologically increasing order, the median age was kept.

Menopausal status was determined using data across instances (field 2724) and prioritising the latest reported data, to account for changes in menopause status. For participants indicating having undergone menopause, their reported ANM were collated (field 3581) using the same procedure as for HRT ages (n=158,264).

Exclusions were then applied to this ANM variable, as follows:

- Participants reporting hysterectomy and/or oophorectomy, but not the age at which this happened (n=958 and 494, respectively)
- Participants reporting multiple hysterectomy and/or oophorectomy ages, which were more than 10 years apart (n=38 and 23, respectively)
- Participants reporting multiple HRT start and/or end ages, which were not in chronologically ascending order and were more than 10 years apart (n=124 and 137, respectively)
- Participants reporting multiple ages at menopause, which were not in chronologically ascending order and were more than 10 years apart (n=73) and participants who reported both having and not having been through menopause and no other interventions (n=98)
- Participants having undergone a hysterectomy/oophorectomy before or during the year they report undergoing menopause
- Participants starting HRT prior to undergoing menopause and participants reporting HRT use, with no accompanying dates

The resulting trait was representative of an age at natural menopause (ANM, n=115,051). Two ANM traits were derived for downstream analyses: the primary trait was winsorised by coding all values of ANM younger than 34, as 34, (n=115,051 total, reduced to 106,973 after covariate-resulting exclusions); and a sensitivity trait was derived by only including participants reporting ANM between 40 and 60, inclusive (n=104,506) (**Figure 3.1, Appendix Table 3.1**).

All manipulations were conducted in R (v4.1.2) on the UKB RAP (https://ukbiobank.dnanexus.com/).

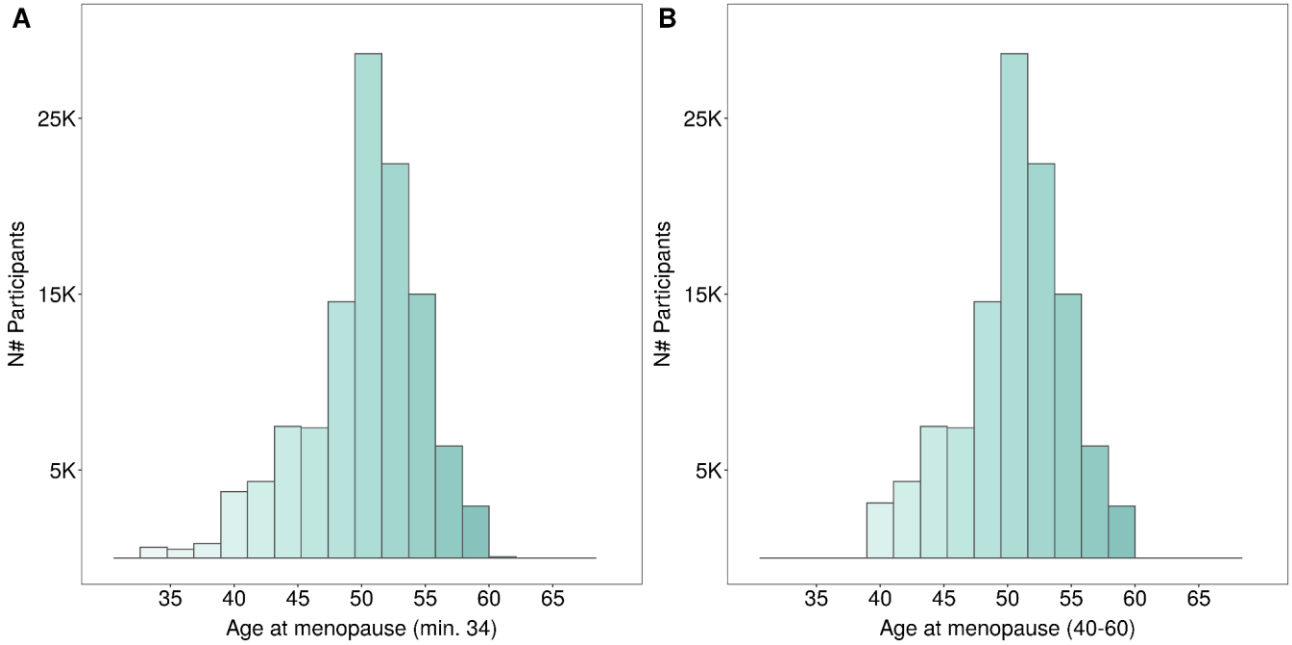

Figure 3.1: Age at menopause distribution in the UKBB study in two traits of interest. (A) *Age at menopause windsored at 34 was used as the primary phenotype, while* ***(B)*** *age at menopause 40-60 was used in the sensitivity analysis.*

3.2.3 Phenome-wide association analysis

In order to test the association of ANM identified genes in other phenotypes, we processed additional reproductive ageing-related phenotypes, including age at menarche, various cancer types, telomere length (TL) and circulating sex hormone concentrations (**Appendix Table 3.8**). All tested phenotypes were run as either continuous (age at menarche, TL and sex hormones) or binary traits (cancer) corrected for age, age^2, sex, the first ten genetic principal components as calculated in Bycroft *et al*[147], and study participant ES batch as a categorical covariate (either 50k, 200k, or 450k). Phenotype definitions and processes are described in **Appendix Table 3.6**. Only the first instance (initial visit) was used for generating all phenotype definitions unless specifically noted in **Appendix Table 3.6**. In case of cancer-specific analysis data from cancer registries, death records, hospital admissions and self-reported were harmonised to International Classification of Disease (ICD10) coding. If a participant had a code for any of the cancers recorded in ICD10 (C00-C97) then they were counted as a case for this phenotype. Minimal filtering was performed on the data, with the only exclusions being those cases where a diagnosis of sex-specific cancer was discordant with the sex data contained in UKBB record 31. The association burden test for quantitative traits was done using BOLT-LMM, while a generalised linear model using the

statsmodels package[165] for python was applied in a three step process to analysed the binary traits, as described in **Section 2.2.3** (**Appendix Table 3.7**).

3.2.4 Cancer PheWAS Associations

To test for an association between the genes we identified as associated with menopause timing (**Figure 3.3**) and 90 individual cancers as included in cancer registries, death records, hospital admissions and self-reported data provided by UKBB (e.g. breast, prostate, etc.) we used a logistic regression model with identical covariates as used during gene burden testing (N = 2430 tests) (**Appendix Table 3.8**). As standard logistic regression can lead to inflated *P* value estimates in cases of severe case/control imbalance[192], we also performed a logistic regression with penalised likelihood estimation as described by Firth[193] (**Figure 3.5**). Models were run as discussed in Kosmidis *et al*.[194] using the 'brglm2' package implemented in R. brglm2 was run via the 'glm' function with default parameters other than "family" set to "binomial", "method" set to "brglmFit", and "type" set to "AS_mean".

3.2.5 Common variant GWAS lookups

Genes within 500kb upstream and downstream of the 290 lead SNPs from the latest GWAS of ANM[82] were extracted for comparison to the exome-wide analysis. There were a total of 2149 genes within the GWAS regions. Burden tests of these genes with a Bonferroni corrected *P* value of $<2.3*10^{-5}$ (0.05/2149) were highlighted.

3.2.6 Analysis of GWAS and WES genes expression profiles in human female germ cells at various stages of development

We studied the mRNA abundance of WES genes during various stages of human female germ cell development using single-cell RNA sequencing data. We used the processed single cell RNA resequencing datasets from two published studies. This included single-cell RNA sequencing data from foetal primordial germ cells of human female embryos (Accession code: GSE86146)[195], and from oocyte and granulosa cell fractions during various stages of follicle development (Accession code: GSE107746)[196]. A pseudo score of 1 was added to all values before log transformation of the dataset. The samples from foetal germ cells (FGCs) were categorised into sub-clusters as defined in the original study. The study by Li *et al*[195] had identified 17 clusters by performing a t-distributed stochastic neighbour embedding (t-SNE) analysis and using expression profiles of known marker genes for various stages of foetal germ cell development. In our analysis we included four clusters of female foetal germ cells

(Mitotic, Retinoic Acid (RA) responsive, Meiotic, Oogenesis) and four clusters containing somatic cells in the foetal gonads (Endothelial, Early_Granulosa, Mural_Granulosa, Late_Granulosa). Software packages for R - tidyverse (https://www.tidyverse.org/), pheatmap, (https://CRAN.R-project.org/package=pheatmap), reshape2 (https://github.com/hadley/reshape), were used in processing and visualising the data.

3.2.7 Functional enrichment tests for *ZNF518A* transcription factor binding sites using fGWAS and SLDP

fGWAS (*v.0.3.6)*, a hierarchical model for joint analysis of GWAS and genomic annotations, was used to test the functional enrichment of ANM GWAS hits in *ZNF518A* transcription factor binding sites[197]. The fGWAS input file contained the ANM GWAS summary statistics derived from the ReproGen study[82] annotated for *ZNF518A* binding sites. The *ZNF518A* annotation file was derived from the ENCODE ChIP-seq data from human HEK293 cell line [ENCSR159GFL][198], where the optimal independent discovery rate peak calling against hg19 [ENCFF415VBF] was used. The ANM GWAS hits were annotated for the presence/absence of the *ZNF518A* transcription factor binding sites in a binary way (0, 1), with '1' if the SNP falls within the transcription factor binding site and '0' otherwise. The fGWAS tool available from https://github.com/joepickrell/fgwas and was run in annotation mode "-w" for the described *ZNF518A* annotation. Detailed description of fGWAS methodology is available in Pickrell *et al,* 2014[197]. In short, the genome is split into independent blocks, which are allowed to contain either a single polymorphism that causally influences the trait or none. fGWAS then models the prior probability that any given block contains an association and the conditional prior probability that any given SNP in the block is the causal one, with probabilities allowed to vary according to functional annotations. The priors are then estimated using an empirical Bayes approach. The fGWAS output contained the maximum likelihood parameter estimates for each parameter in the model, in this case *ZNF518A*, with the lower and upper bound of the 95% confidence interval (CI) on the parameter. The P value was calculated from fGWAS lower and upper CI (stated on the page 84) in 3 following steps: (1) Standard error (SE) calculation: $\boldsymbol{SE = (Upper\ CI - Lower\ CI)/(2*1.96)}$; (2) Test statistics calculation: $\boldsymbol{Z = Estimate / SE}$; and (3) P value calculation: $\boldsymbol{P = exp(-0.717*Z - 0.416*Z^2)}$[567].

Signed LD profile (SLDP) regression was applied to explore the directional effect of a signed functional annotation, *ZNF518A*, on a heritable trait like ANM using GWAS summary statistics. More specifically, we tested whether alleles that are predicted to increase the binding of the transcription factor *ZNF518A* have a genome-wide tendency to increase or decrease ANM. The SLDP tool was installed from

https://github.com/yakirr/sldp, with the comprehensive methodological steps described in Reshef *et al*, 2018[199]. For the analysis to be conducted, SLDP required GWAS summary statistics for ANM, signed LD profiles for *ZNF518A* binding, signed background model and reference panel in a SLDP compatible format. For the reference we used a 1000 Genomes Phase 3 European reference panel in *plink* format, which contained approximately 10M SNPs and 500 people and was available for download at the refpanel page. The ANM GWAS summary statistics, available from our latest Reprogen study, was pre-processed using the '*preprocesspheno*' tool from the SLDP package. To conduct this step, we also obtained the list of regression SNPs along with the LD scores for the reference panel from the refpanel page. The pre-processing step included filtering down to SNPs that are also present in the reference panel, harmonising alleles to the reference, and multiplying the summary statistics by the SLDP regression weights. In addition, we applied the '*preprocessrefpanel*' tool to compute truncated singular value decomposition (SVD) for each LD block in the reference panel. These SVDs were later used to weight the SLDP regression. The *ZNF518A* annotation file was obtained from the ENCODE CHIP-seq analysis, as described above, and preprocessed using the '*preprocessannot*' tool that turns signed functional annotations into signed LD profiles. Prior to running SLDP, we also obtained the signed background LD profiles that enabled us to control for systematic signed effects of minor alleles, which could arise from either population stratification or negative selection. SLDP was then run on our data using the '*sldp*' function.

To explore the relevance of *ZNF518A* for ANM in comparison to other transcription regulators, we tested whether genome-wide sequence changes introduced by SNP alleles identified in ANM GWAS increase or decrease binding of additional 382 transcription factors (TFs). The pre-processed annotation files for 382 TFs derived from ENCODE CHIP-seq experiments, were available for download at the annotation data page. The results are presented in **Figure 3.7**.

3.2.8 Functional analysis of *ZNF518A* binding sites

ZNF518A peaks were derived from unique genomic regions in ENCODE accession ENCFF415VBF described above. Quantification of ChIP-seq signal by aligning paired-end replicates (ENCFF174HBR, ENCFF574GQY, ENCFF808AJP, ENCFF453FDD) to the hg19 genome with Bowtie2 v2.3.5.1[200] with options "-I 0 -X 1000 –no-discordant –no-mixed", reads were filtered for those with MAPQ > 30 with samtools v1.10. Assessment of H3K27ac[201] and chromatin accessibility by ATAC-seq[202] in day 4 human primordial germ cell like cells (hPGCLCs) at *ZNF518A* peaks was performed. For H3K27ac single end reads from accessions GSM4257216, GSM4257217,GSM4257218 were obtained and aligned with

Bowtie2 v2.3.5.1 with default settings and MAPQ > 30 reads retained as above. For ATAC-seq paired-end reads were obtained from accessions GSM3406938, GSM3406939 and mapped and filtered as *ZNF518A* reads above.

Quantification of ChIP-seq and ATAC-seq signals for peak heights, heatmaps was performed with https://github.com/owensnick/GenomeFragments.jl. Peak to transcription start site (TSS) distances were calculated against Gencode v36 release liftover to hg19 using GenomicFeatures.jl and https://github.com/owensnick/ProximityEnrichment.jl. We consider four categories of peaks: TSS intersecting, TSS proximal (TSS < 2000kb, outside gene body), Gene body intersecting, Intergenic and Distal (TSS > 5kb).

To perform *de novo* motif discovery we used Homer v4.11.1[203] using findMotifsGenome.pl with options "hg19 -size 200". We ran this on all *ZNF518A* peaks, distal peaks and those intersecting TSS, we recovered a motif matching JASPAR[204] unvalidated motif UN0199.1 in all peak sets apart from those intersecting TSS. We then used https://github.com/exeter-tfs/MotifScanner.jl to quantify the occurrence of all instances of motif UN0199.1 in *ZNF518A* peaks.

We downloaded the 18-state ChromHMM[205] models for all 833 biosamples in Epimap[206] from http://compbio.mit.edu/epimap/. We calculated the intersection between each state in each biosample and either all *ZNF518A* peaks or distal *ZNF518A* peaks using GenomicFeatures.jl. We calculated ORs from contingency tables using the approximation of bedtools[207] and Giggle[208], by estimating total genomic intervals as hg19 genome size divided by the sum of the mean *ZNF518A* peak size and the chromatin state interval size. The results are presented in **Figure 3.7**.

3.2.9 *De novo* mutation rate analyses

We calculated GWAS-based PGSs in participants from the rare disease programme of the 100,000 Genome Project (100kGP) v14. There were 77,901 individuals in the Aggregated Variant Calls (aggV2) after excluding participants whose genetically inferred sex was inconsistent with their phenotypic sex. We restricted the PGS analysis to individuals of European ancestry, which was predicted by the Genomics England Bioinformatics team using a random forest model based on genetic PCs generated by projecting aggV2 data onto the 1000 Genomes phase 3 PC loadings. We removed one sample in each pair of related probands with kinship coefficient > 1/(2^4.5), i.e. up to and including third degree relationships. Probands with the highest number of relatives were removed first. Similarly, we retained unrelated mothers and fathers of these unrelated probands. It left us with 8,089 mother-offspring duos and 8,029 father-offspring duos.

We used the lead variants (or proxies, as described below) for genome-wide significant GWAS loci previously reported for age at ANM[82] to calculate PGS in the parents. In 100kGP, we removed variants with MAF <0.5% or missing rate >5% from the aggV2 variants prepared by the Genomics England bioinformatics team. For lead variants that did not exist in 100kGP, we used the most significant proxy variants with LD r^2 >0.5 if available in 100kGP. This resulted in a PGS constructed from 287 of the 290 previously reported GWAS loci. We included 20 genetic PCs that were calculated within the European subset from the PGS and scaled the residuals to have mean = 0 and standard deviation = 1. Higher PGS indicates later age at menopause.

De novo mutations (DNMs) were called in 10,478 parent offspring trios by the Genomics England Bioinformatics team. The detailed analysis pipeline is documented at: https://research-help.genomicsengland.co.uk/display/GERE/De+novo+variant+research+dataset. Extensive QC and filtering were applied by Kaplanis *et al*. (2021) as described[209]. De novo SNVs (dnSNVs) were phased using a read-based approach based on heterozygous variants near the DNM that were able to be phased to a parent. About one third of the dnSNVs were phased, of which three quarters were paternally phased (**Figure 3.13**).

In association models, we accounted for parental age, the primary determinant of the number of DNMs, and various data quality metrics as described in Kaplanis *et al* (2021)[209]:

- Mean coverage for the child, mother and father (child_mean_RD, mother_mean_RD, father_mean_RD)
- Proportion of aligned reads for the child, mother and father (child_prop_aligned, mother_prop_aligned, father_prop_aligned)
- Number of SNVs called for child, mother and father (child_SNVs, mother_SNVs, father_SNVs)
- Median variant allele fraction of DNMs called in child (median_VAF)
- Median 'Bayes Factor' as outputted by Platypus for DNMs called in the child. This is a metric of DNM quality (median_BF).

We first tested the association between parental PGSs and total dnSNV count in the offspring in a Poisson regression:

$$dnSNVs_{total} = \beta_0 + \beta_1 paternal\_PGS + \beta_2 maternal\_PGS + \beta_3 paternal\_age + \beta_4 maternal\_age + \beta_5 child\_mean\_RD + \beta_6 mother\_mean\_RD + \beta_7 father\_mean\_RD + \beta_8 child\_prop\_aligned + \beta_9 mother\_prop\_aligned + \beta_{10} father\_prop\_aligned + \beta_{11} child\_snvs + \beta_{12} mother\_snvs + \beta_{13} father\_snvs + \beta_{14} median\_VAF + \beta_{15} median\_BF$$

We also fitted Poisson regression models to test the association between the PGS of one of the parents and the dnSNVs in the offspring that were phased to the relevant parent.
The paternal model included paternal PGS, age, and data quality metrics that are related to the proband and the father:

$$dnSNVs\_paternal = \beta_0 + \beta_1 paternal\_PGS + \beta_2 paternal\_age + \beta_3 child\_mean\_RD + \beta_4 father\_mean\_RD + \beta_5 child\_prop\_aligned + \beta_6 father\_prop\_aligned + \beta_7 child\_snvs + \beta_8 father\_snvs + \beta_9 median\_VAF + \beta_{10} median\_BF$$

Similarly, the maternal model was as follows:

$$dnSNVs\_maternal = \beta_0 + \beta_1 maternal\_PGS + \beta_2 maternal\_age + \beta_3 child\_mean\_RD + \beta_4 mother\_mean\_RD + \beta_5 child\_prop\_aligned + \beta_6 mother\_prop\_aligned + \beta_7 child\_snvs + \beta_8 mother\_snvs + \beta_9 median\_VAF + \beta_{10} median\_BF$$

Finally, as a sanity check, we assessed the association between the maternal PGS and paternally phased dnSNVs, and vice versa:

$$dnSNVs\_paternal = \beta_0 + \beta_1 maternal\_PGS + \beta_2 paternal\_age + \beta_3 child\_mean\_RD + \beta_4 father\_mean\_RD + \beta_5 child\_prop\_aligned + \beta_6 father\_prop\_aligned + \beta_7 child\_snvs + \beta_8 father\_snvs + \beta_9 median\_VAF + \beta_{10} median\_BF$$

$$dnSNVs\_maternal = \beta_0 + \beta_1 paternal\_PGS + \beta_2 maternal\_age + \beta_3 child\_mean\_RD + \beta_4 mother\_mean\_RD + \beta_5 child\_prop\_aligned + \beta_6 mother\_prop\_aligned + \beta_7 child\_snvs + \beta_8 mother\_snvs + \beta_9 median\_VAF + \beta_{10} median\_BF$$

3.2.10 Mendelian Randomisation

MR analysis was applied to examine the likely causal effect of ANM (expressed as a PGS) on the risk of de novo mutation rates in the offspring. The analysis was conducted following the methodology described in **Section 2.3.1**. Genotypes at all variants were aligned to designate the ANM PGS-increasing alleles as the effect alleles as described above and this was used as a genetic instrument of interest. The effect sizes of genetic instruments (genotypes in the mother) on maternally phased dnSNVs in the offspring estimated in 8,089 duos were obtained from Genomics England.

The results were considered as significant based on the *P* value significance consistency across different primary and sensitivity models applied, which are in details described in **Section 2.3.1**. The results are available in **Tables 3.2 and 3.3**. Finally, in order to calculate the effect of ANM on offspring *de novo* mutation rate when comparing women with ANM at two extremes of the ANM distribution curve, we multiplied the effect obtained by MR IVW, i.e. a de novo count beta per 1 year change in ANM, by 20, an arbitrary number that compares women with ANM 20 years apart.

3.3 Results

3.3.1 Exome-wide gene burden associations with ANM

To assess the impact of rare damaging variants on ANM, we analysed WES data available in 106,973 post-menopausal UKBB female participants of European genetic-ancestry[158]. Individual gene burden association tests were conducted by collapsing genetic variants according to their predicted functional categories. We defined three categories of rare exome variants with MAF < 0.1%: HC PTVs, missense variants with CADD score ≥ 25, and 'damaging' variants (defined as combination of HC-PTVs and missense variants with CADD ≥ 25). We analysed 17,475 protein-coding genes with the minimum of 10 rare allele carriers in at least one of the three masks tested. The primary burden association analysis was conducted using linear mixed models BOLT-LMM[154]. The low exome-wide inflation scores (e.g. PTV λ=1.047) and the absence of significant association with the synonymous variant burden for any gene indicate good calibration of our statistical tests (**Figure 3.2**).

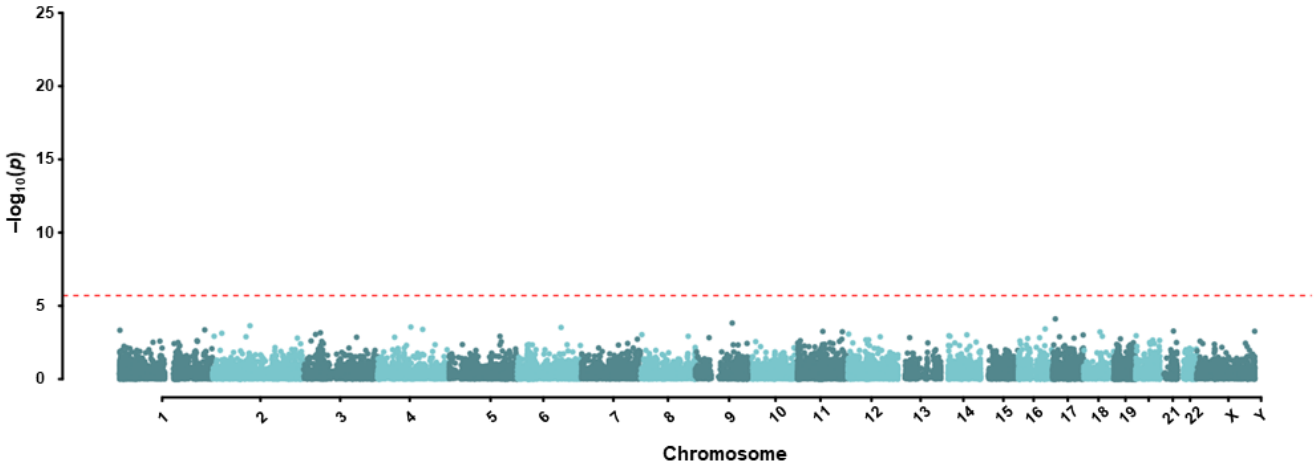

Figure 3.2: Exome-wide association results for synonymous variants. *Plotted are per-gene burden results for synonymous variants. The red line indicates the exome-wide significant P value after Bonferroni correction of $1.08*10^{-6}$. Lack of association was expected for this 'negative control' analysis.*

We identified rare variation in nine genes associated with ANM at corrected exome-wide significance ($P<1.08*10^{-6}$, **Figures 3.3 and 3.4**). Three of these genes have been previously reported in a previous analysis of the same UKBB WES sample[191] - we confirm the associations of *CHEK2* (beta=1.57 years, 95% CI: 1.23-1.92, $P=1.60*10^{-21}$, N=578 damaging allele carriers) and *HELB* (beta=1.84, 95% CI: 1.08-2.60, $P=4.20*10^{-7}$, N=120 HC-PTV carriers) with later ANM and a previously borderline association of *HROB* with earlier ANM (beta= -2.89 years, 95% CI: 1.86-3.92, $P=1.90*10^{-8}$, N=65 HC-PTV carriers). In addition, our previous ANM GWAS analyses[82] identified an individual with low-frequency PTV variant in *BRCA2*, which we now extend to demonstrate that in aggregate *BRCA2* HC-PTV carriers exhibit 1.18 years earlier ANM (beta= -1.18, 95% CI: 0.72-1.65, $P=2.60*10^{-7}$, N=323). Rare variants in the remaining five genes – *ETAA1*, *ZNF518A*, *PNPLA8*, *PALB2* and *SAMHD1* have not been previously implicated in ovarian ageing. Effect sizes of these associations range from 5.61 years earlier ANM for HC-PTV carriers in *ZNF518A*, to 1.35 years later ANM for women carrying damaging alleles in *SAMHD1*. This contrasts with a maximum effect size of 1.06 years (median 0.12 years) for common variants (MAF>1%) identified by previous ANM GWAS[82].

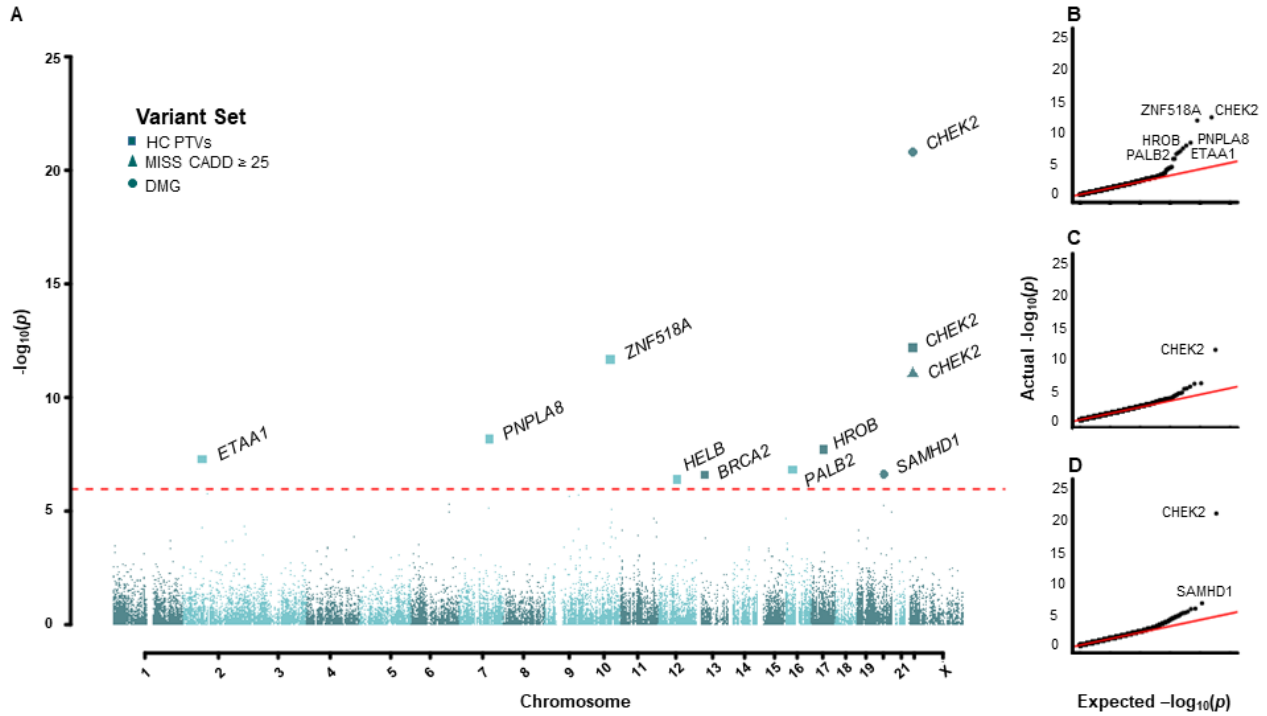

Figure 3.3: Exome-wide associations with age at natural menopause. *(A) Manhattan plot showing gene burden test results for age at natural menopause. Genes passing exome-wide significance ($P<1.08x10^{-6}$) are indicated, with point shape showing the variant class tested. (B-D) QQ plots for (B) high confidence PTVs (C) CADD ≥ 25 missense variants (D) damaging variants.*

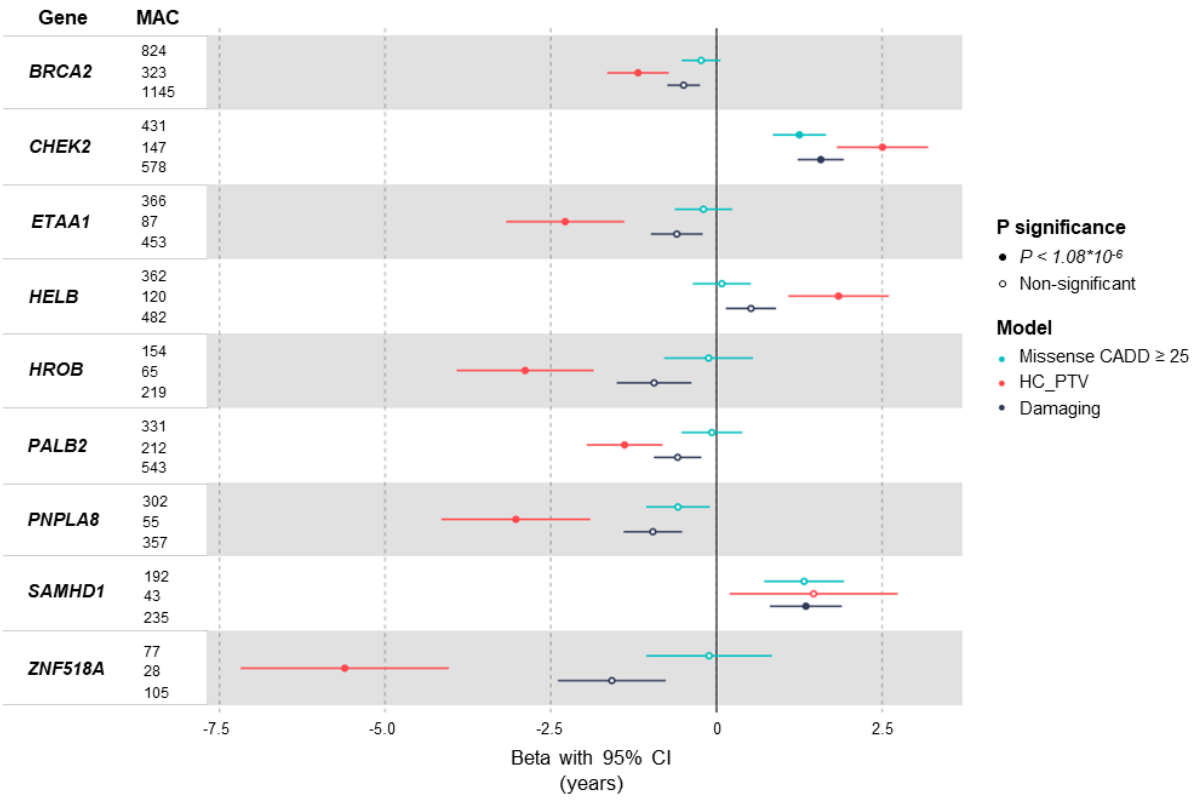

Figure 3.4: Forest plot for gene burden associations with age at natural menopause. *Exome-wide significant ($P < 1.08*10^{-6}$) genes are displayed. Points and bars indicate beta and 95% CI for specific variant categories, MAC and P values derived from BOLT-LMM.*

In order to depict the variants that contribute to the gene-level burden score, we plotted *lolliplot* plots demonstrating variant level associations and indicating the direction of the effect, strength of association with ANM, as well as the number of allele carriers (**Figure 3.5** and **3.6**). This suggests that, consistent with expectation, the significant burden level results for the newly identified genes are driven by multiple variants contributing towards the overall signal, some of which had variants with more dominant effect on that signal.

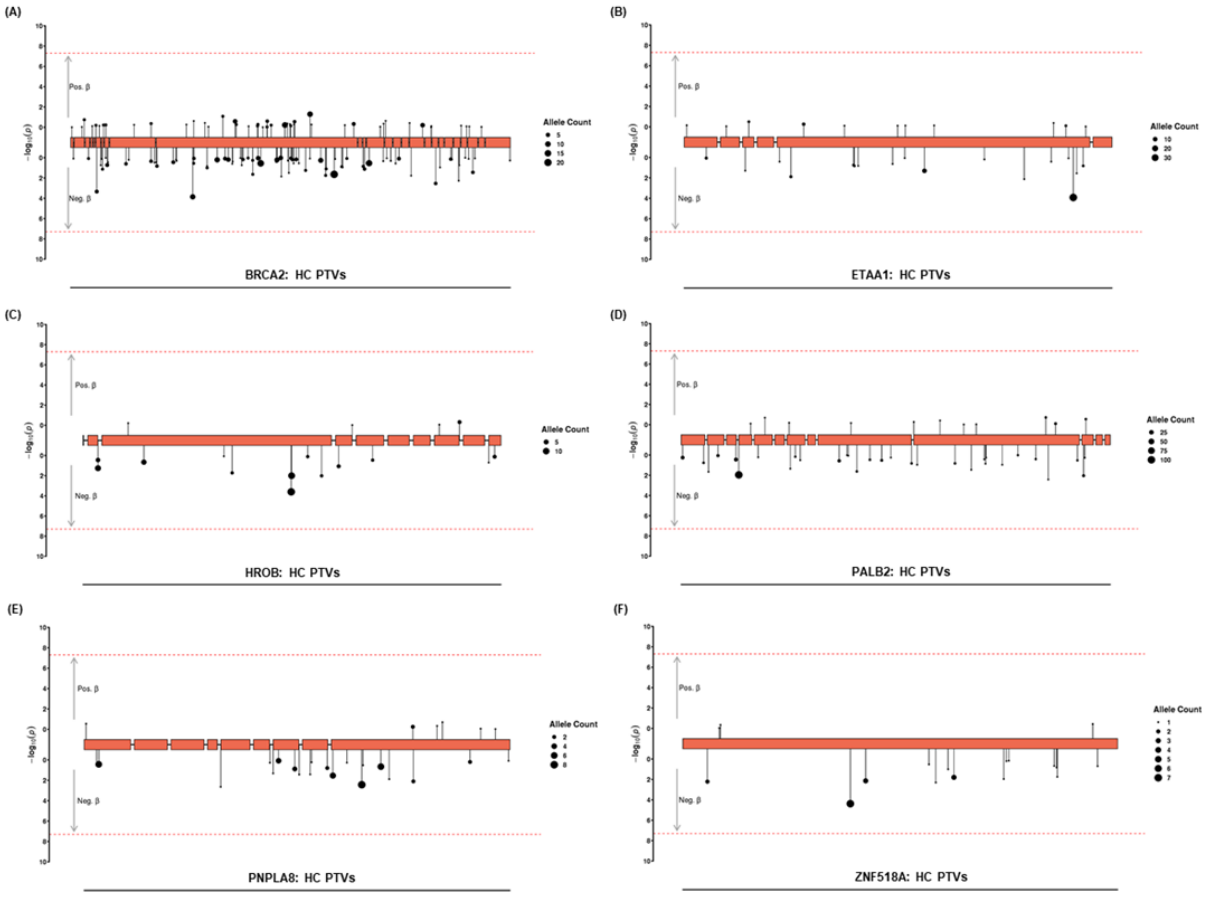

Figure 3.5: Variant level associations for age at natural menopause decreasing WES genes. *Lolliplot plots show the variants clustered for the best performing functional mask per gene from gene burden tests for ANM using BOLT-LMM. These include:* ***(A)*** *BRCA2 HC PTV mask;* ***(B)*** *ETAA1, HC PTV mask;* ***(C)*** *HROB, HC PTV mask;* ***(D)*** *PALB2, HC PTV mask;* ***(E)*** *PNPLA8, HC PTV mask; and* ***(F)*** *ZNF518A, HC PTV mask. The lines pointing upwards represent the variants positively associated with ANM, while the downwards ones show the negatively associated variants. The size of the point indicates the allele count.*

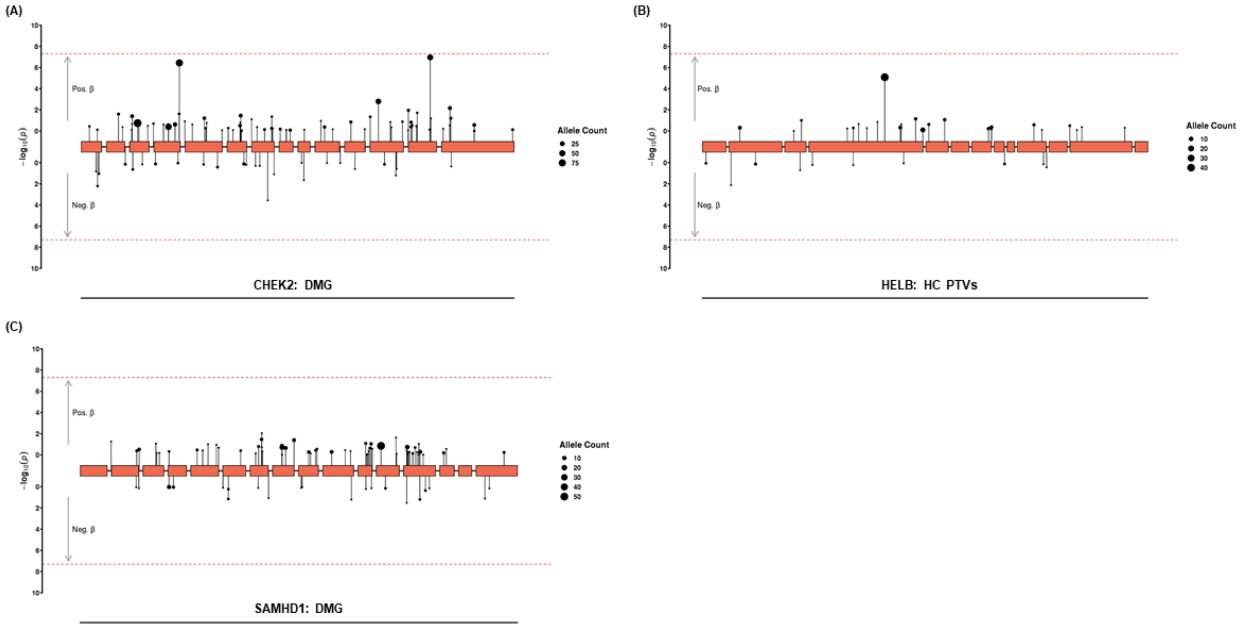

Figure 3.6: Variant level associations for age at natural menopause increasing WES genes. *Lolliplot plots show the variants clustered for the best performing functional mask in a gene that went into the gene burden test for ANM using BOLT-LMM. These include:* ***(A)*** *CHECK2, damaging mask;* ***(B)*** *HELB, HC PTV mask; and* ***(C)*** *SAMHD1, damaging mask. The lines pointing upwards represent the variants positively associated with ANM, while the downwards ones show the negatively associated variants. The size of the point indicates the allele count.*

We next sought to understand why previous analyses of the same UKBB WES data missed the associations we report here, and conversely why we did not identify associations with other previously reported genes. Of the seven genes identified by Ward *et al.*[191], three were also identified by our study (*CHEK2*, *HELB* and *HROB*), three were recovered when we increased our burden test MAF threshold from 0.1% to 1% (*DCLRE1A*, *RAD54L*, *TOP3A*), and an additional gene fell just below our *P* value threshold when considering variants with <1% MAF (*CLPB; P =1.2*10*$^{-5}$). Importantly, our results provide more robust evidence ($P = 1.9*10^{-8}$) for the previously described suggestive *HROB* association ($P = 2.9*10^{-6}$) and we identified six genes, which were not captured by Ward *et al.*: *ZNF518A*, *BRCA2*, *ETAA1*, *PALB2*, *PNPLA8* and *SAMHD1* (**Appendix Table 3.2**). We investigated potential study design differences that could account for variation in the findings as the same data was used in both studies.

3.3.1.1 Associations not captured in current analysis

First we investigated potential analytical parameters that could account for differences in the findings. Associations with *DCLRE1A*, *RAD54L*, *TOP3A* and *CLPB* were not identified in our study, because we restricted our analysis to variants with a MAF <0.1%, rather than <1%. We re-analysed our data with a burden test MAF threshold of <1% and three of the four associations were replicated: *DCLRE1A* (P_{MAF}

$_{1\%}$= 3.8*10^{-8}, N: 1056), *RAD54L* ($P_{\text{MAF 1\%}}$= 6.4*10^{-7}, N: 1892) and *TOP3A* ($P_{\text{MAF 1\%}}$= 1.5*10^{-7}, N: 2001). *RAD54L and TOP3A* were genes highlighted by GWAS and the exome association in *TOP3A* was driven by a single, relatively common variant (rs34001746, MAF=0.7%, P=1.63*10^{-10}). This variant was in LD with the previously reported lead GWAS SNP (rs569145577, r^2=0.92), with little evidence for association after its exclusion (P=0.50 for all other missense and PTVs). *CLPB* just missed our *P*-value threshold, but again a single variant (rs150343959, P=8.22*10^{-6}) was largely driving the association signal - when excluded in leave-one-out analysis, the *CLPB* burden association dropped (P=1.19*10^{-2}). By including relatively common variants in gene burden masks, single variants can dominate the general functional effect being tested, which could be contributed to by LD with non-exomic functional variants. Therefore in order to be able to make a stronger link between genetic variants and individual genes, we chose to restrict our analysis to rarer variants with MAF <0.1%.

3.3.1.2 Associations not captured by Ward *et al*

Differences in MAF thresholds did not explain why our study identified an additional six genes (*BRCA2*, *ETAA1*, *PALB2*, *PNPLA8*, *SAMHD1* and *ZNF518A*) compared with Ward *et al*[191]. We therefore tested differences in the phenotype preparation, tools and variant masks used to test the associations. Four of our six additional gene burden associations (*BRCA2*, *PALB2*, *PNPLA8*, and *SAMHD1*) were relatively near the borderline of the significance threshold in our analyses in the BOLT-LMM pipeline.

We included a ~20% larger sample size (106,973 post-menopausal women) in comparison to Ward *et al*. (78,311 unrelated post-menopausal women), which would have resulted in more statistical power in our analyses (**Appendix Table 3.2**). This was particularly important for *BRCA2*, *ETAA1* and *PALB2* - Ward *et al.* included 63 (19.5%) fewer *BRCA2* and 46 (21.7%) fewer *PALB2* carriers of rare damaging variants (**Appendix Table 3.2**) and identified the ANM association in these genes only at the borderline of exome-wide significance, P=1.55*10^{-6} and P=7.47*10^{-5}, respectively. Similarly for *SAMHD1*, Ward *et al.* captured 57 (24.3%) fewer carriers in their linear regression model compared with our main analysis, which resulted in association P values of 6.38*10^{-4} in the linear regression model and P=8.02*10^{-6} in the time to event analysis. Reasons for the smaller sample size in the previous study included: Ward *et al.* used only unrelated individuals in their primary analyses, whereas we used linear mixed models and were therefore able to include an additional ~19,000 related individuals. Secondly, Ward *et al.* excluded ~2,300 women with ANM <40 and >60 years, while we used the full (winsorised) natural menopause distribution. Finally, we took into account four instances where questions regarding ANM were asked, whereas Ward *et al.* used data from the baseline visit in their main analysis, resulting in an additional ~7,400 women.

As sensitivity analyses and to better replicate the methods of Ward *et al.*, for four of the genes (*CHEK2, DCLRE1A, ZNF518A* and *PNPLA8*) we compared results from linear regression in unrelated individuals using MAF<1% with those from a time-to-event Cox proportional hazards model (**Appendix Table 3.2**). Association statistics from the Cox model (*CHEK2*: P=2.4*10^{-39}, *DCLRE1A*: 6*10^{-8}, *ZNF518A*: 1.4*10^{-9}, *PNPLA8*: 5.7*10^{-10}) were comparable to those from linear regression models based on the full range of ANM (*CHEK2*: P=3.1*10^{-46}, *DCLRE1A*: 2.5*10^{-8}, *ZNF518A*: 1.2*10^{-9}, *PNPLA8*: 1.9*10^{-9}).

Finally, differences in variant annotation may also explain some inconsistencies between studies. *ZNF518A* was not reported by Ward *et al.*, which may be because all variants are in the last and only coding exon of the gene, and in some annotations such variants would inappropriately be excluded from being considered as LoF. We note that another single coding exon gene (*NFIL3*) was not included by Ward *et al.*, but was in our analysis.

Detailed comparisons between our study and Ward *et al.* are available in **Appendix Table 3.2**.

3.3.2 Exploring common variant associations at identified ANM genes

To explore the overlap between common and rare variant association signals for ANM, we integrated our exome-wide results with data generated from the largest reported common variant GWAS of ANM[82]. Five of our nine identified WES genes (*CHEK2*, *BRCA2*, *ETAA1*, *HELB* and *ZNF518A)* mapped within 500kb of a common GWAS signal. Notably, we previously reported a common, predicted benign, missense variant (rs35777125-G439R, MAF=11%) in *ETAA1* associated with 0.26 years earlier ANM. In contrast, our WES analysis identified that rare HC-PTV carriers show a nearly 10-fold earlier ANM (beta= -2.28 years, 95% CI: 1.39-3.17, P=5.30*10^{-8}, N=87). Furthermore, three independent non-coding common GWAS signals ~150kb apart (MAF: 2.8-47.5%, beta: -0.28-0.28 years per minor allele) were reported proximal to *ZNF518A*, whereas our gene burden testing finds that rare HC-PTV carriers show nearly 20-fold earlier ANM than common variant carriers (beta= -5.61 years, 95% CI: 4.04-7.18, P=2.10*10^{-12}, N=28).

In addition there were two genes within 500kb of GWAS loci (*BRCA1* and *SLCO4A1*) that were associated with ANM by gene burden testing at P <1.7*10-5 (Bonferroni correction for the 2910 genes included). Effect sizes for the common variant association ranged from 0.07-0.24 years per allele at these loci, whereas gene burden tests for rarer variants revealed much larger effect sizes: for *BRCA1*, 2.1 years earlier for PTVs ($P=2.4*10^{-6}$) and for *SLCO4A1*, 1.13 years earlier ANM for damaging variants ($P=1.1*10^{-5}$), with non-overlapping 95% confidence intervals between common and rare variant associations for *BRCA1*.

3.3.3 Common ANM associated variants are enriched in *ZNF518A* binding sites

Heterozygous loss of function of *ZNF518A* had the largest effect on ANM of the genes we identified by WES. *ZNF518A* is poorly characterised C2H2 zinc finger transcription factor, which has been shown to physically associate with PRC2 and G9A-GLP repressive complexes along with its paralog *ZNF518B*, suggesting a potential role in transcriptional repression[210]. *ZNF518A* localises robustly to 18,706 sites in the genome, based on ChIP-seq data available from ENCODE[211,212] and binds primarily to gene promoters, with 33.5% (6,263) of *ZNF518A* binding sites within 2kb of a transcription start site (TSS) (**Figure 3.7 A-C**). Common variants associated with ANM[82] were enriched in the transcriptional targets of *ZNF518A* (estimate: 2.016 [95% CI: 0.745-2.797], P=1.32*10^{-4}) using fGWAS[197]. Note that the estimates derived from fGWAS do not indicate the directionality of the effect. We further tested functional enrichment using SLDP regression[199]. This confirmed the enrichment of *ZNF518A* binding sites near to loci associated with ANM and showed that its transcriptional repression is associated with earlier ANM (P=0.02), consistent with results of our rare variant burden tests. Separating *ZNF518A* sites by those proximal (< 2Kb) and distal (>5kb) from a TSS, demonstrated this association was due to *ZNF518A* binding at regulatory regions distal to the TSS (proximal TSS P=0.3, distal *ZNF518A* P=0.002). Notably, these regulatory *ZNF518A* bound loci produce the largest association amongst an SLDP catalogue of 382 transcription factors and regulators (**Figure 3.7 D**). These results suggest a different functional role for *ZNF518A* at TSS and more distal regulatory regions. In order to explore this further we assessed the sequence determinants of *ZNF518A* binding. *De novo* motif discovery identified an AT-rich motif enriched at distal regulatory *ZNF518A* binding sites, but not at TSS bound by *ZNF518A*. This AT-rich motif was centrally enriched within *ZNF518A* ChIP-seq peaks, and matched an unvalidated motif present in the JASPAR transcription factor motif database[204] **(Figure 3.7 E**). We found the number of perfect instances of this AT-rich motif to be strongly associated with *ZNF518A* occupancy as assessed by *ZNF518A* ChIP-seq signal at distal regions but not at TSS (**Figure 3.7 F, G**). At distal regions, the maximal association between peaks greater than the median height was found at least seven motif instances (Hypergeometric right tail $P < 10^{-389}$, OR=7.41). These data suggest that *ZNF518A* is recruited by DNA sequence at distal sites, but at TSS may be recruited to gene promoters by interaction with another DNA binding factor.

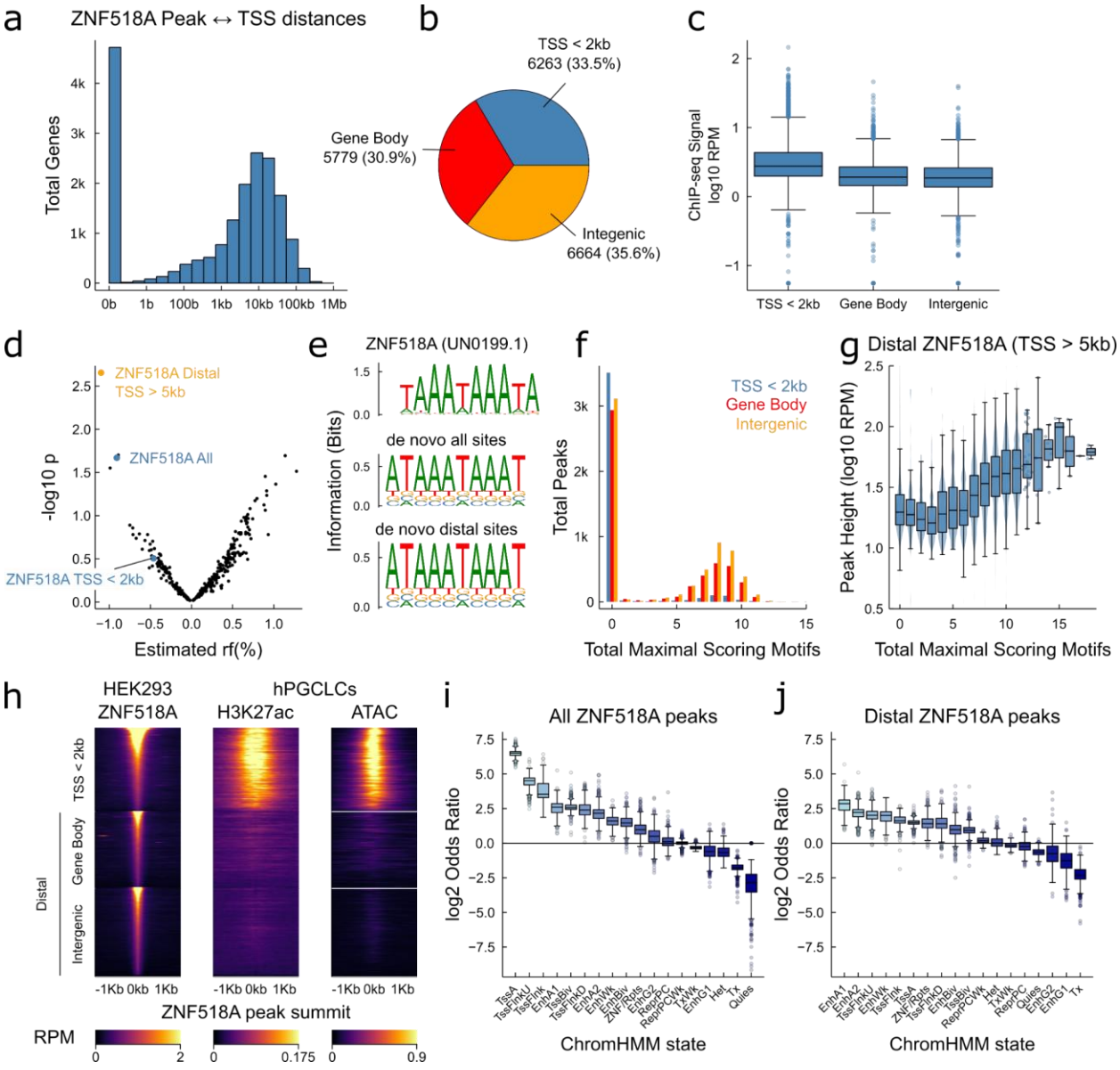

Figure 3.7: Functional analysis of ZNF518A bound loci. (A) *Histogram of log10-scale distances between ZNF518A and nearest gene transcription start site (TSS).* ***(B)*** *Proportion of ZNF18A peaks falling proximal to TSS (TSS < 2kb), within gene bodies and in intergenic regions.* ***(C)*** *Boxplots showing total normalised reads per million (RPM) for every peak for categories TSS < 2kb, gene body and intergenic - ZNF518A peaks have greater signal at proximal to TSS.* ***(D)*** *SLDP association between ANM GWAS variants and ZNF518A peaks, stratified by all peaks, proximal (< 2kb) from a TSS, and distal (> 5kb) from a TSS. The association between ANM variants and ZNF518A peaks appears due to distal ZNF518A peaks (either gene body or intergenic, > 5kb TSS) and not proximal TSS binding.* ***(E)*** *De novo motif discovery recovers unvalidated JASPAR motif for ZNF518A UN0199.1. Homer enrichment statistics: all sites P= 10^{-6451} motif in 31.2% of targets (1.15% background); distal sites P= 10^{-4590} motif in 47.3% of targets (1.81 % background).* ***(F)*** *Proportion of maximal scoring instances of UN0199.1 (sequences that exactly match motif consensus) by ZNF518A peak category. Many distal peaks contain multiple perfect instances of the motif.* ***(G)*** *Boxplots, violin plots and dot plots depicting the relationship between ZNF518A ChIP-seq peak height and number of maximal scoring motifs present in peak. A strong relationship between peak height and number of motif instances can be observed.* ***(H)*** *Heatmaps depicting ZNF518A ChIP-seq, H3K27ac ChIP-seq in hPGCLCs, and chromatin accessibility by ATAC-seq in hPGCLCs. Signal shown over all ZNF518A peaks in RPM +/- 1kb of ZNF518A peak summit. ZNF518A bound promoters (TSS < 2kb) are accessible and are marked with H3K27ac,*

distal regions either in gene bodies or intergenic regions show no H3K27ac or chromatin accessibility, suggestive that ZNF518A represses these regulatory regions. ***(I,J)*** *Association shown in odds ratios of ChromHMM states over 833 tissues/cell types from Epimap, boxplots with outliers shown, each boxplot summarises the distribution of associations over all tissues/cell types for a given chromatin state.* ***(I)*** *All ZNF518A peaks;* ***(J)*** *ZNF518A peaks distal from TSS.*

We next used publically available data on *in vitro* differentiated human primordial germ like-cells[201,202] to assess the chromatin state at *ZNF518A* bound loci, directly comparing distal regions with TSS. *ZNF518A* bound TSS showed chromatin accessibility[202] and were marked with H3K27ac[201]. In contrast, distal regions lacked H3K27ac and showed minimal chromatin accessibility **(Figure 3.7 H)**. Extending this comparison to the Epimap chromatin states[206], we find that overall *ZNF518A* bound loci are enriched in active TSS and that distal *ZNF518A* regions are variously enriched in active and repressed chromatin **(Figure 3.7 I,J)**. Consistent with previous data which has found *ZNF518A* in repressive complexes, these data suggest that *ZNF518A* is recruited by DNA sequence to distal regulatory regions where it acts to repress local chromatin.

While *ZNF518A* is known to have diverse tissue expression including the ovary, we found that it was highly expressed in foetal germ cells at both the mitotic and meiotic stages (**Figures 3.8 and 3.9**). The eight other WES genes identified in this study were expressed at varying levels in foetal gonadal cells, oocytes and granulosa cells across different developmental stages.

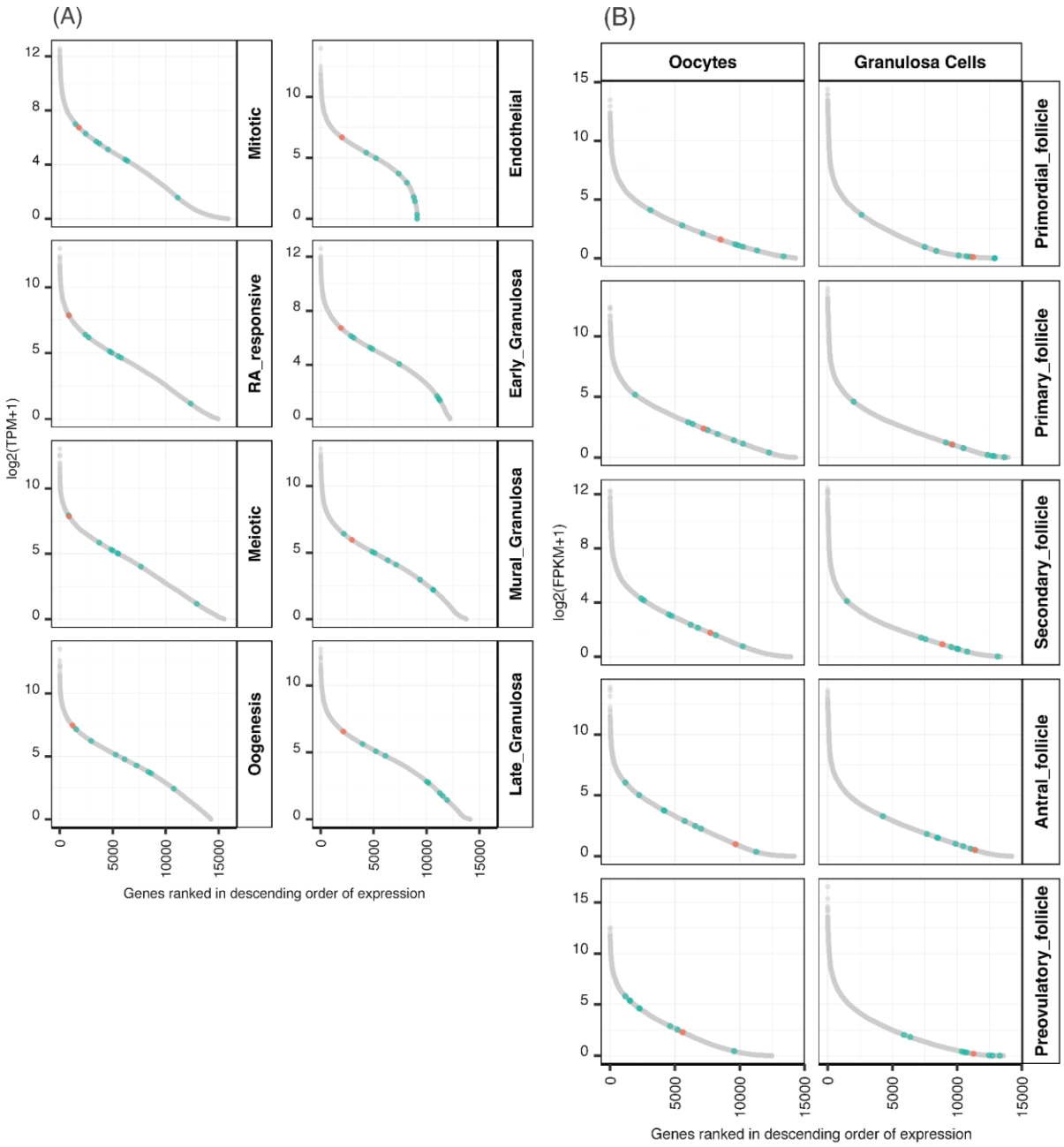

Figure 3.8: Expression levels of genes across various stages of female germ cell development. *In the X-axis, genes are ranked according to their average expression at each stage (Y-axis)* ***(A)*** *in human foetal primordial germ cells and* ***(B)*** *in granulosa cells in adult follicles. Genes identified as novel ANM genes in WES analysis are coloured in green and all other genes in the genome are in grey. ZNF518A is depicted in orange for the ease of comparison with other genes.*

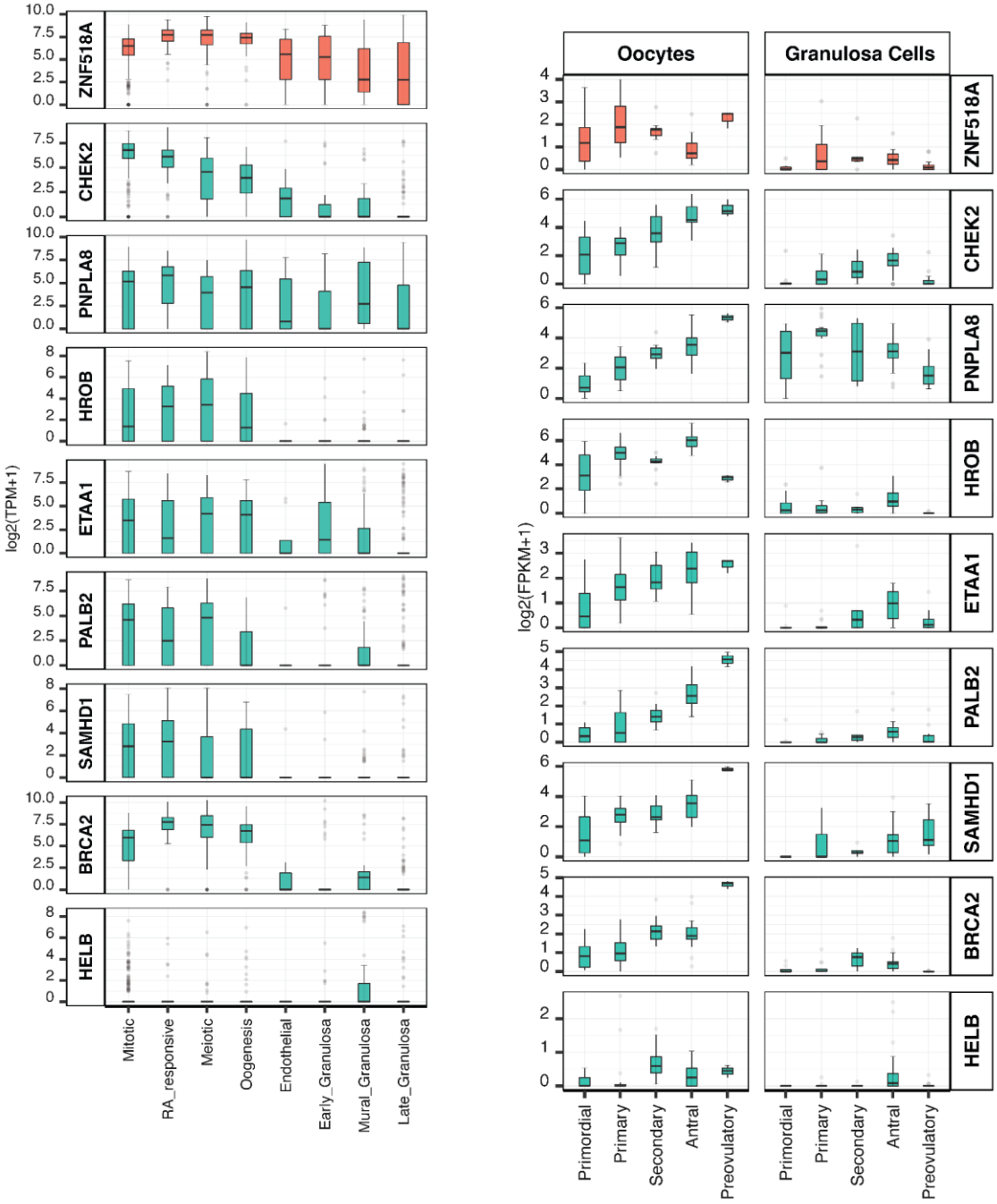

Figure 3.9: mRNA expression of WES genes during foetal stages and folliculogenesis. *Box and whisker plots of mRNA expression of the WES genes at different stages of germ cell development. The plots represent the interquartile range of TPM values, the line at the centre of the box representing the median, error bars indicate the 95% confidence interval and outliers shown as dots.* ***(A)*** *The sub-clusters from single foetal cells from week 5 to 26 post-fertilisation are on the X-axis with the average TPM expression values log2(TPM+1) on the Y-axis.* ***(B)*** *Different stages of folliculogenesis in oocytes and granulosa cells are represented on the X-axis with their average expression values log2(FPKM+1) on the Y-axis.*

3.3.4 Identified genes influence other aspects of health and disease

Deciphering the genetic control of menopause is important for understanding the relationship between menopause and associated disease risk. Our genetic studies have previously shown that the genetic mechanisms regulating the end of reproductive life are largely distinct from those determining its beginning[131]. However, it is noteworthy that the largest reported GWAS for age at menarche identified a common variant signal at the *ZNF518A* locus for later puberty timing in girls (rs1172955, beta= 0.04 years, 95% CI: 0.03-0.05, $P=6.6*10^{-12}$), which appears nominally associated with earlier ANM (beta=-0.04, 95% CI: 0.01-0.06, $P=6.6*10^{-3}$)[131]. To extend this observation, we found that our identified *ZNF518A* PTVs were also associated with later age at menarche (0.56 years, 95% CI: 0.14- 0.98, $P=9.2*10^{-3}$). Furthermore, using fGWAS and SLDP, we discovered that, similar to ANM, common variants that influence puberty in girls were enriched in transcriptional targets of *ZNF518A*. These data suggest that loss of *ZNF518A* shortens reproductive lifespan, by delaying puberty and reducing age at menopause.

We next explored the impact of ANM-associated genes on cancer outcomes and found a novel association of *SAMHD1* damaging variants and HC-PTVs with 'All cancer' in both males (OR=2.12, 95% CI: 1.72-2.62, $P=4.7*10^{-13}$) and females (OR=1.61, 95% CI: 1.31-1.96, $P=4*10^{-6}$; **Figure 3.10, Appendix Table 3.4**). In addition, we replicated previously reported associations with PTVs in *BRCA2*, *CHEK2* and *PALB2* and cancer outcomes in males and females[82,86].

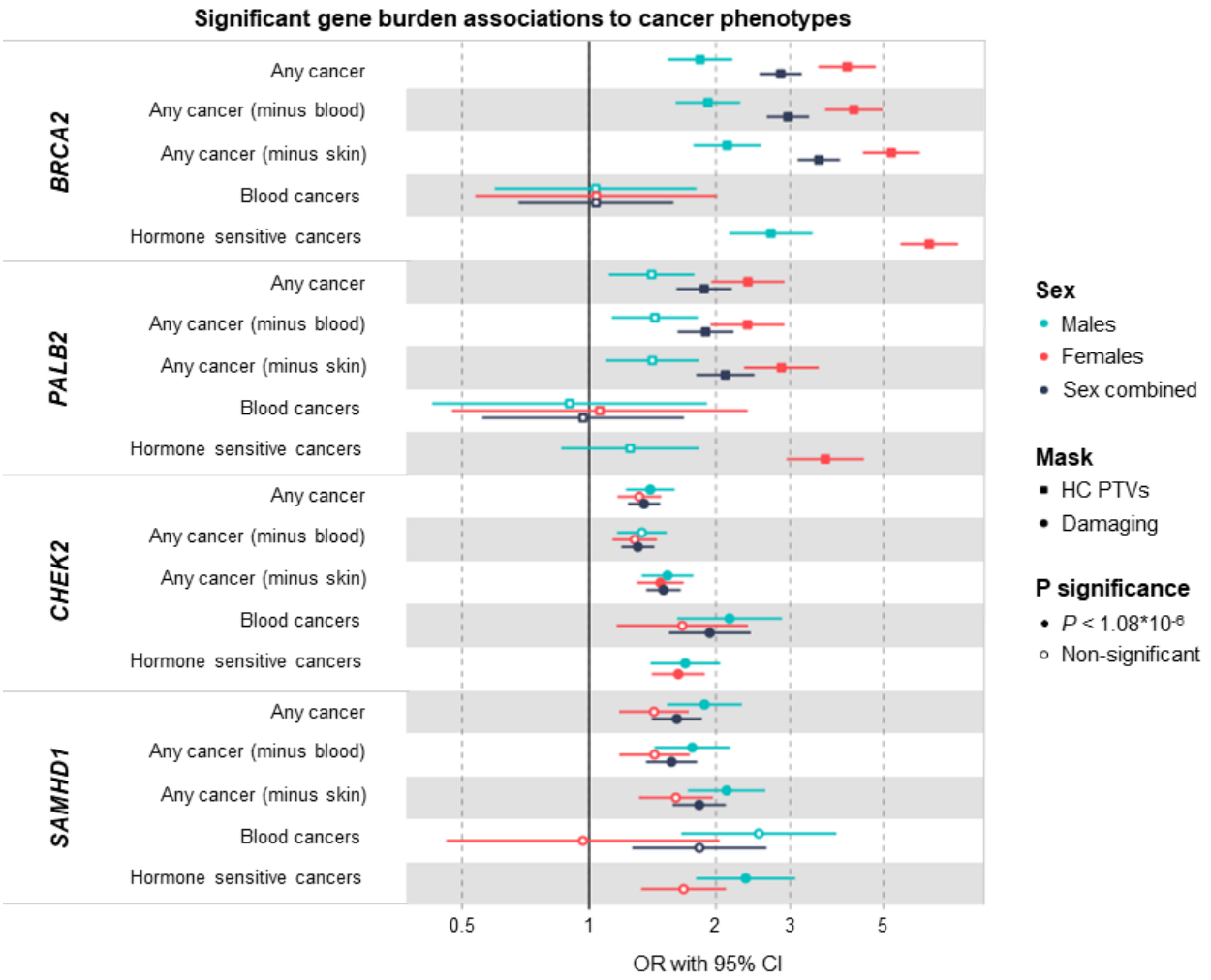

Figure 3.10: Forest plot for age at natural menopause WES genes with significant gene burden associations for cancer phenotypes. *Exome-wide significant ($P < 1.08*10^{-6}$) genes are displayed, showing sex-stratified and combined results. Hormone sensitive cancers were only tested in males and females separately (Methods). The presented masks were selected based on the most significant association per gene and cancer type. Points and bars indicate OR and 95% CI for specific genes and their variant categories in cancer. Filled symbols indicate a result passing a Bonferroni-corrected significance threshold of $P < 1.08*10^{-6}$.*

SAMHD1 associations with cancer appear to be driven by increased risk for multiple site-specific cancers, notably prostate cancer in males and mesothelioma in both males and females, as well as suggestive evidence for higher susceptibility of breast cancer in females (**Figure 3.11, Appendix Table 3.6**). Although the numbers of mutation carriers with each site-specific cancer was small, most of these findings persisted using logistic regression with penalised likelihood estimation, which is robust to extreme case/control imbalance[193].

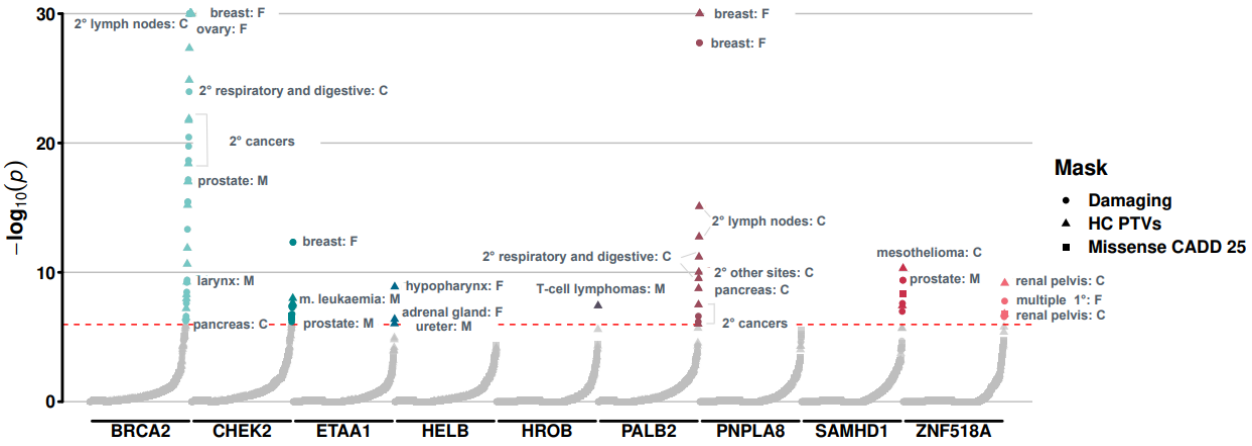

Figure 3.11: Genetic susceptibility to earlier ovarian ageing and increased risk for diverse cancer types. *Plot showing the association between loss of ANM genes identified in this study and risk of 90 site specific cancers among UKBB participants. Summary statistics for cancer associations were obtained using a logistic regression with penalised likelihood estimation that controls for case/control imbalance (Methods)*[193]*. Associations highlighted in text passed exome-wide significance ($P < 1.08*10^{-6}$). The y-axis is capped at $-log_{10}(P) = 30$ for visualisation purposes; un-capped summary statistics can be found in* ***Appendix Table 3.6****.* ***F:*** *females,* ***M:*** *males,* ***C:*** *sex-combined.* ***1°:*** *primary cancer,* ***2°:*** *secondary cancer.*

Cancer risk-increasing alleles in *SAMHD1* were associated with later ANM, which is similar to the pattern demonstrated previously for *CHEK2*. This finding is consistent with a mechanism of disrupted DNA damage sensing and apoptosis, resulting in slowed depletion of the ovarian reserve[82]. This is in contrast to *BRCA2* and *PALB2*, where cancer risk-increasing LoF alleles inhibit DNA repair and lead to earlier ANM. In addition, we provide robust evidence for a previously described rare variant association for *SAMHD1* with telomere length (TL)[213], highlighting that rare damaging variants cause longer TL (P=1.4*10^{-59}) (**Figure 3.12, Appendix Table 3.4**). This further supports its role in tumorigenesis as genetic variants associated with longer TL are considered to be risk factors for various cancer types[214]. Furthermore, a woman's reproductive life span has previously been shown to positively correlate with TL, a known biomarker for biological ageing [215,216].

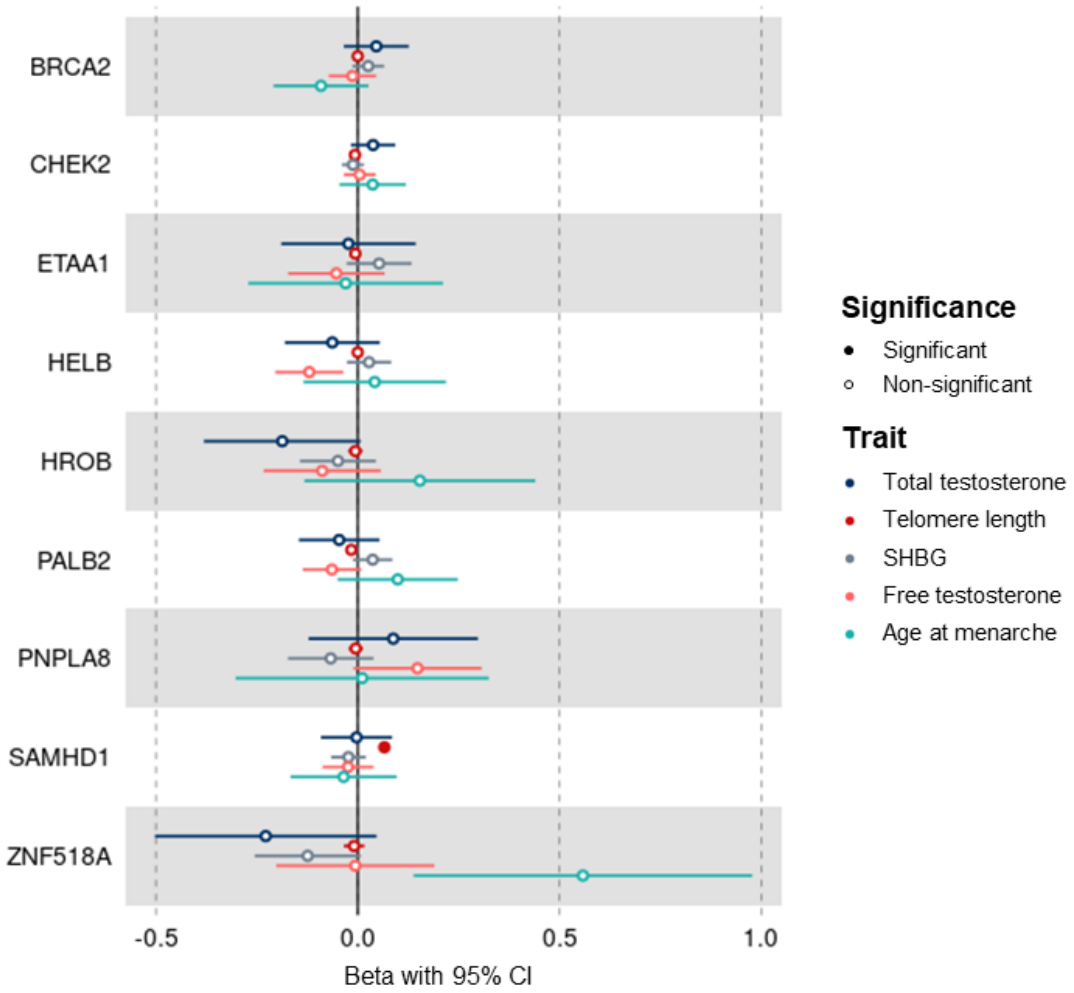

Figure 3.12: Age at natural menopause gene burden associations with reproductive ageing-related traits of interest in females only. *The coefficients and 95% CIs were female-specific and plotted for the quantitative traits only. The association was tested using BOLT-LMM.*

3.3.5 Genetic susceptibility to ANM in mothers influences de novo mutation rate in offspring

Of the nine genes we identified by WES as associated with ANM, seven are involved in DDR, further supporting the role of this mechanism in ovarian ageing. For genes that inhibit DNA double-strand break (DSB) repair, the hypothesis is that they cause premature depletion of the ovarian reserve due to a failure to repair oocytes with DNA damage. This is evidenced by the reported increased numbers of DNA DSBs in the oocytes of *Brca1*-deficient mice and of women with *BRCA1* mutations who underwent elective oophorectomy[93]. Our current study adds further support, with heterozygous *BRCA1* and *BRCA2* LoF alleles causing 2.1 and 1.18 years earlier ANM, respectively.

We sought to build on these observations by testing the hypothesis that inter-individual variation in these DDR processes would influence the mutation rate in germ cells and hence in the offspring. More specifically, we hypothesised that genetic susceptibility to earlier ovarian ageing would be associated with a higher *de novo* mutation (DNM) rate in the offspring. To test this, we analysed 8,089 whole-genome sequenced parent-offspring trios recruited in the rare disease programme of the 100,000 Genome Project (100kGP, **Figure 3.13**, **Table 3.1**).

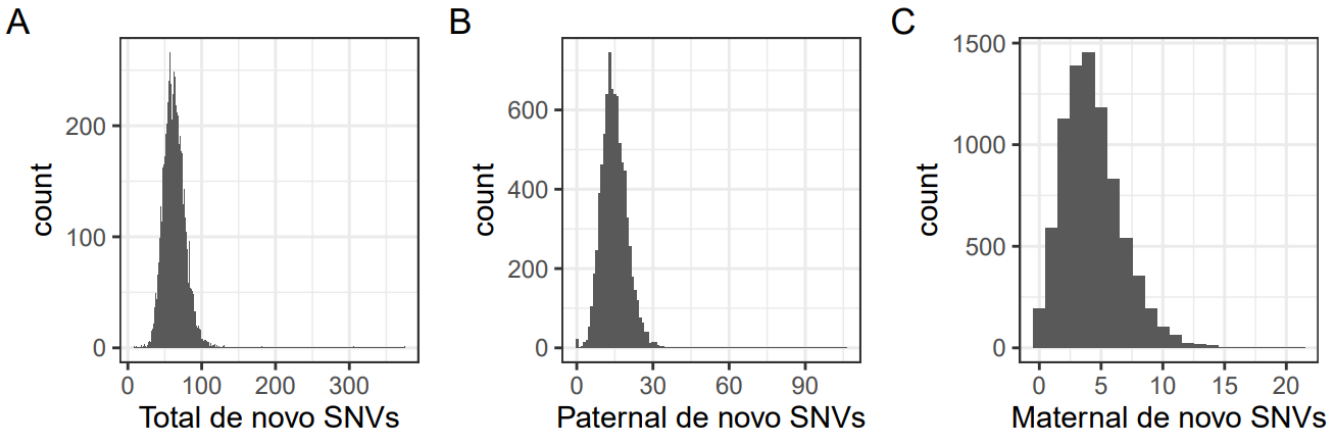

Figure 3.13: Distribution of de novo single nucleotide variants (dnSNVs). *The histogram shows the number of* ***(A)*** *total dnSNVs,* ***(B)*** *paternally derived dnSNVs and* ***(C)*** *maternally derived dnSNVs in unrelated probands with European ancestry from the 100,000 Genomes Project.*

We calculated a polygenic score (PGS) for ANM in the parents based on our previously identified 290 common variants[82] and tested this against the phased DNM rate in the offspring, adjusted for age. We found that maternal genetic susceptibility to earlier ANM was associated with an increased rate of maternally-derived DNMs in the offspring (rate ratio = 1.02 per SD of PGS, P=6.8*10^{-4}, N=8,089 duos with European ancestry; **Table 3.1**).

Table 3.1: The association between parental polygenic scores for age at natural menopause and de novo mutations in offspring. *We tested the association between parental PGS and total dnSNVs as well as phased dnSNVs in a Poisson regression. Details of the association models and covariates are in the Methods. Note that the values are aligned to ANM-increasing PGS.*

De novo mutations	PGS	Sample size	Beta	SE	Rate ratio	Lower CI Rate ratio	Upper CI Rate ratio	P
total dnSNVs	mother	7672	-0.0010	0.0014	0.999	0.996	1.002	$4.83*10^{-1}$
total dnSNVs	father	7672	-0.0007	0.0014	0.999	0.996	1.002	$6.17*10^{-1}$
maternal dnSNVs	mother	8089	-0.0183	0.0054	0.982	0.972	0.992	**$6.81*10^{-4}$**
paternal dnSNVs	father	8029	-0.0019	0.0029	0.998	0.992	1.004	$5.08*10^{-1}$
maternal dnSNVs	father	8029	0.0032	0.0054	1.003	0.993	1.014	$5.48*10^{-1}$
paternal dnSNVs	mother	8089	-0.0025	0.0029	0.997	0.992	1.003	$3.78*10^{-1}$

We confirmed this finding in sensitivity analyses using the same data, in a two-sample MR framework that can better model the dose-response relationship of these variants (**Figure 3.14**, **Table 3.2** and **3.3**). These results were highly concordant, with all models showing a significant result and no heterogeneity (*P*min=$6.3*10^{-5}$). In contrast, the paternal PGS was not associated with paternally-derived DNMs (*P*=0.51, N=8,029) nor was the maternal PGS associated with paternally-derived DNMs (*P*=0.55) (**Table 3.1**).

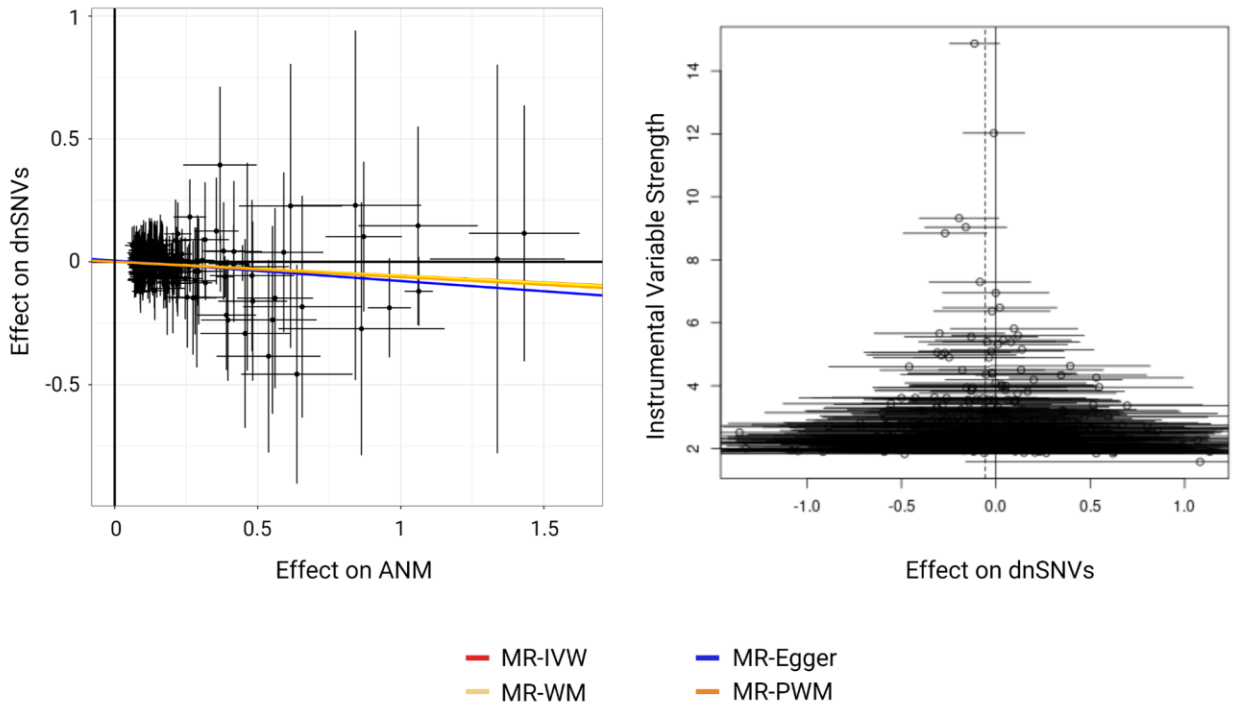

Figure 3.14: Mendelian Randomisation of the effect of age at natural menopause PGS on de novo mutations. (A) *Scatter plot showing the results for primary and sensitivity MR analysis.* ***(B)*** *Funnel plot to test directional pleiotropy*

Table 3.2: Primary Mendelian Randomisation analysis of genetically-mediated age at natural menopause in the mother and the rate of de novo mutations in offspring. *Note that the values are aligned to ANM-increasing alleles. The total number of genetic instruments used per analysis is presented by N SNPs.* ***IVW:*** *Inverse Variance Weighted method,* ***Coch Qp:*** *Cochran's q test P value.*

Model	Exposure	Outcome	N SNPs	Beta IVW	SE IVW	P IVW	Coch Qp	I^2
Pre-Radial filtering	AAM	dnSNVs	287	-0.056	0.018	$\mathbf{1.73*10^{-3}}$	0.551	0
Post-Radial filtering	AAM	dnSNVs	272	-0.074	0.018	$\mathbf{6.34*10^{-5}}$	1.000	0

Table 3.3: Secondary Mendelian Randomisation analysis of genetically-mediated age at natural menopause in the mother and the rate of de novo mutations in offspring. *Note that the values are aligned to ANM-increasing alleles. The total number of genetic instruments used per analysis is presented by N SNPs.* ***EI****: Egger intercept,* ***WM:*** *weighted median,* ***PWM:*** *penalised weighted median.*

	Egger						WM			PWM		
Radial	**Beta**	**SE**	**P**	**EI**	**SE EI**	**P EI**	**Beta**	**SE**	**P**	**Beta**	**SE**	**P**
Pre	-0.083	0.034	**$1.45*10^{-2}$**	0.005	0.005	$3.49*10^{-1}$	-0.057	0.031	$6.65*10^{-2}$	-0.062	0.030	**$4.06*10^{-2}$**
Post	-0.100	0.034	**$3.60*10^{-3}$**	0.005	0.004	$2.87*10^{-1}$	-0.071	0.031	**$2.43*10^{-2}$**	-0.071	0.031	**$2.22*10^{-2}$**

3.4 Discussion

3.4.1 Large effect sizes of rare variants

Previous GWASs have revealed a limited fraction of heritability behind reproductive ageing. They initiated to pave the path towards better understanding of how and when molecular processes influence the establishment and decline of the ovarian reserve. To overcome the challenges related to identification of causal mechanisms and relevant genes in GWAS, we explored the impact of rare protein-coding genetic alternations on menopausal timing by studying exome sequence data in UKBB. Our study extends the number of genes implicated in ovarian ageing - effect sizes ranged from 5.61 years earlier ANM for HC-PTV carriers in *ZNF518A*, to 1.35 years later ANM for women carrying damaging alleles in *SAMHD1* compared to a maximum effect size of 1.06 years (median 0.12 years) reported for common variants (MAF>1%)[82]. Several of these effect estimates were comparable to those conferred by *FMR1* premutations, which are currently used as part of the only routinely applied clinical genetics test for POI[217]. A damaging variant in at least one of our identified nine ANM genes was carried by 1,703 women in UKBB (1.6%), with around 0.7% of women carrying genomic variants that reduce reproductive lifespan by over a year.

3.4.2 New biological mechanisms

Novel biological mechanisms of ovarian ageing were revealed by finding associations with two non-DDR genes, *ZNF518A* and *PNPLA8*. *ZNF518A* belongs to the zinc finger protein family and is likely a transcriptional regulator for a large number of genes[210]. Both common and rare genetic variation in this gene indicate that female carriers have shorter reproductive lifespan due to delayed puberty timing and earlier menopause[82,183,218], highlighting a mechanism for shared aetiology between these traits[197,199,212].

Previous epidemiological and GWASs have identified a modest shared genetic aetiology behind the timing of puberty and menopause, mainly spanning from the discovery of genes involved in regulation of the hypothalamic-pituitary-gonadal axis and sex hormones[77,86,131]. The shared association at the *ZNF518A* loci points to a potentially novel mechanism involved in regulation of the beginning and end of reproductive longevity. Enrichment of GWAS signals at *ZNF518A* binding sites suggests that *ZNF518A* regulates the genes involved in reproductive longevity by repression of elements distal to transcription start sites.

We identified a second novel non-DDR gene (*PNPLA8*) associated with ANM, where rare damaging mutations lead to more than 3 years earlier menopause. *PNPLA8* is a calcium-independent phospholipase[219,220] and a recessive cause of neurodegenerative mitochondrial disease and mitochondrial myopathy[221–223] an association with reproductive phenotypes has not been described previously. Notably, other genes in the same phospholipase family, *PLA2G4A* and *PLA2G6*, are mainly described for reproductive system specific defects in animal models, including reduced reproductive ability in females and impaired fertilisation ability in male mice, indicating a critical role of this biological mechanism in functional regulation of both oocytes and spermatozoa[224,225].

3.4.3 DDR genes newly implicated in ANM

Seven of the nine ANM genes identified in this study have known roles in DDR, and three of these (*PALB2*, *ETAA1* and *HROB*) are linked to ANM for the first time here, shedding further light on the mechanisms involved. We identified PTVs in *PALB*2 associated with a 1.39 years earlier ANM. *PALB2* is involved in *BRCA2* localization and stability, mediating double-strand break repair via homologous recombination. Complete loss of *PALB2* is embryonic lethal[226] and compound heterozygous mutations result in Fanconi anaemia and predispose to childhood malignancies[227]. *ETAA1* accumulates at DNA damage sites in response to replication stress[228–231] and is involved in regulation of DNA damage checkpoint[231,232]. Finally, *HROB* encodes a DNA repair protein that takes part of homologous recombination by recruiting the *MCM8-MCM9* helicase to sites of DNA damage to promote DNA synthesis[233,234]. Homozygous LoF of *HROB* is associated with POI[235] and infertility in both sexes in mouse models[233]. Variants in both *MCM8* and *MCM9* are associated with reproductive ageing in humans[236–238], with MCM8 missense variants being associated with ANM from GWAS, and non-sense variants in *MCM9* being implicated in POI[239], which additionally supports the role of this biological mechanism in reproductive longevity.

3.4.4 Two new DDR genes extend reproductive lifespan

Robust prioritisation of likely causative genes and mechanisms facilitates downstream analyses and identification of potential therapeutic targets. Deleterious variants in three genes (*CHEK2*, *HELB* and *SAMHD1*) were associated with an increase in ANM and therefore represent potential therapeutic targets for enhancing ovarian stimulation in women undergoing *IVF* treatment through short-term apoptotic inhibition. *HELB* was previously identified by GWAS[82,86], but this is the first evidence that LoF of HELB can extend reproductive lifespan, with an effect size of 1.8 years (95% CI: 1.08-2.60, P=4.2*10^{-7}). *HELB* is a DNA helicase essential for DNA replication and inhibits homologous repair of double strand breaks by preventing end resection[240–242]. Loss of HELB results in *PARP* inhibitor resistance in *BRCA1*-deficient cells suggesting loss of HELB could improve double-strand break repair and thus affect oocyte quantity[86,242].

While mutation in *SAMHD1* is a common somatic event in a variety of cancers[243–247], it has not been described as a germline risk factor previously.

Recessive inheritance of *SAMHD1* missense and PTV variants have been associated with Aicardi–Goutieres syndrome, a congenital autoimmune disease[248]. Our identified damaging variants in *SAMHD1* increased risk of 'All cancer' in males and females, as well as in sex-specific cancers, highlighting *SAMHD1* as a novel risk factor for prostate cancer in males and hormone-sensitive cancers in females. Of additional site-specific cancers we tested, suggestive association was identified with mesothelioma, a rare cancer generally caused by exposure to asbestos and affecting the lining of the lungs. Here, *SAMHD1* damaging allele carriers exhibited a seven-fold increased risk relative to non-carriers. Only a single genetic risk factor discovered to date is considered to be a reliable candidate for early prediction - *BAP1*, a *BRCA1* associated protein involved in regulation of transcription, cell cycle, response to DNA damage and chromatin dynamics[249]. Besides *BAP1*, additional high-risk genetic factors that are being studied for mesothelioma all include genes belonging to DNA repair or Fanconi anaemia pathways[250]. This further supports the discovery of *SAMHD1* that could potentially contribute towards screening of people at risk for mesothelioma and thus early diagnosis and treatment, which should be evaluated by future studies.

SAMHD1 has a role in preventing the accumulation of excess deoxynucleotide triphosphates (dNTPs), particularly in non-dividing cells[251]. A regulated dNTP pool is important for the fidelity of DNA repair, thus highlighting additional roles of this gene in facilitation of DNA end resection during DNA replication and repair[252–255]. *SAMHD1* was specifically described to be involved in homologous recombination-mediated double-strand break repair and DNA end joining[256]. *SAMHD1* deficiency leads

to resistance to apoptosis[257,258], suggesting that delayed ANM might originate from slowed depletion of ovarian reserve due to disrupted apoptosis, analogous to the mechanism for *CHEK2* that has been reported previously (**Figure 3.14**).

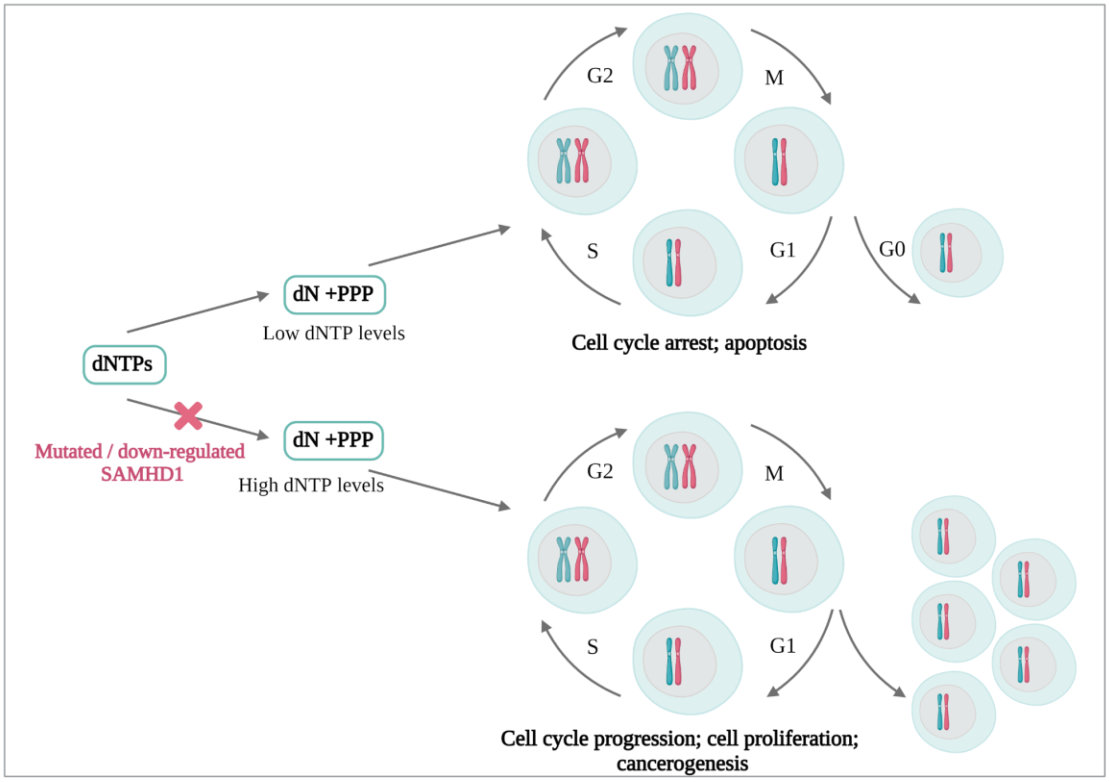

Figure 3.15: Downregulated SAMHD1 or mutated SAMHD1 are involved in cancerogenesis. *SAMHD1 depletes dNTPs to a low level in normal nondividing cells, which induces cell cycle arrest and promotes apoptosis. Downregulated SAMHD1 or mutated SAMHD1 results in high dNTPs levels in cells. In this environment, DNA synthesis is strengthened, cell cycle progression is out of control, and cells are proliferated, thus leading to cancerogenesis.*

Alteration in the size of dNTP pools due to defective *SAMHD1* function was reported to have a specific influence on TL homeostasis, which was confirmed in our and previous WES analysis[259]. Furthermore, shorter leukocyte TL has been previously associated with earlier reproductive ageing, including shorter reproductive lifespan and occult ovarian insufficiency [260,261] or poorer IVF outcomes[261], with both TL and ANM being viewed as 'proxies' for biological ageing[146].

3.4.5 Genetic susceptibility to earlier ovarian ageing increases *de novo* mutation rate in offspring

Previous studies have demonstrated that parental age is strongly associated with the number of *de novo* mutations in offspring[262], with the majority of these mutations arising from the high rate of spermatogonial stem cell divisions that underlie spermatogenesis throughout adult life of males[263]. Our current study provides the first direct evidence that maternal mutation rate is heritable, with women at higher genetic risk of earlier menopause transmitting an increased rate of *de novo* mutations to offspring. This could have direct implications for the health of future generations given the widely reported link between *de novo* mutations and increased risk of psychiatric disease and developmental disorders[264–267]. We speculate that if genetic susceptibility to earlier menopause influences *de novo* mutation rate, it is possible that non-genetic risk factors for earlier ANM, such as smoking and alcohol intake, would likely have the same effect[268]. Our observations make conceptual sense given that menopause timing appears to be primarily driven by the genetic integrity of oocytes and their ability to sustain, detect, repair and respond to acquired DNA damage[82]. These observations also build on earlier work in mice and humans that *BRCA1/2* deficiency increases the rate of double strand breaks in oocytes and reduces ovarian reserve [93,269,270].

An important limitation of our study, shared by many other similar large-scale exome sequencing studies, is that we were unable to replicate our findings in an independent cohort. Instead, we aimed to accumulate additional evidence where possible to support our observations and evaluate the biological plausibility of our findings. For example, the identified rare LoF alleles in *ZNF518A* have the largest effect on ovarian ageing reported to date, which is supported by high expression in foetal germ cells, genome-wide significant common variants at the same locus, and the observation that *ZNF518A* binding sites genome-wide are significantly enriched for common variant ANM association(s). In addition, participants undergoing radiation or chemotherapy before undergoing menopause were not excluded from our analysis. Chemotherapy-induced menopause could be decreasing the power of our conclusions, however we excluded participants who had hysterectomy and oophorectomy so we accounted for the great proportion of these individuals. For all identified genes further experimental studies will ultimately be required to fully understand the biological mechanisms governing the observed effects on ovarian ageing.

CHAPTER 4

Monogenic causes of Premature Ovarian Insufficiency are likely rare and mostly recessive

Summary

The identification of genetic risk factors and underlying mechanisms especially becomes important when we talk about more clinically relevant cases, such as the ones with extreme early menopause timing, i.e. POI. POI impacts 1% of the female population and is a leading cause of infertility. It is often considered to be a monogenic disorder, with pathogenic mutations in ~100 genes described in the literature. However, as the evidence on the reported genes is often based on observations from one or small number of individuals without conclusive replication in an independent cohort, or/and the causative impact of the variants has not been proven functionally, the validity of these genes being causative is questionable. We sought to systematically evaluate the penetrance of these genes using exome sequence data in 104,733 women from the UKBB, 2,569 of whom had menopause under the age of 40. Our results from the largest POI study up to date demonstrate limited evidence to support any previously published autosomal dominant gene, with 97.8% of all identified PTVs found in reproductively healthy women with menopause over 40. Assuming 100% penetrance, we estimated that 1.6 million genes would be required to carry a homozygous or compound heterozygous LoF knockout in order to reach the observed 1% frequency of POI in the population. If all 20,000 genes in the genome carried such a knockout, this would still only result in a population frequency of 0.012%. Although we were unable to fully assess autosomal recessive effects, we found evidence of novel haploinsufficiency effects on menopausal timing in the normal range in several of these genes, including *TWNK* (1.39 years earlier menopause, $P=8.5*10^{-6}$) and *SOHLH2* (3.26 years earlier, $P=1.6*10^{-4}$).

This Chapter indicates that autosomal dominant mutations in genes currently described in the literature or evaluated in clinical diagnostic panels are not common causes of POI, suggesting that the majority of POI cases are likely oligogenic or polygenic in nature. These findings have strong implications for clinicians diagnosing causes of POI, where described genes should be interpreted with caution to avoid misdiagnosis as it seems that they do not have plausible clinical significance.

Contributions and Collaborations

I performed the literature review and the manual curation of POI genes, together with the review of GeL Panel App POI gene panel under the supervision of Prof Anna Murray. Dr Eugene Gardner created a pipeline on the UKBB DNAnexus Research Analysis Platform (RAP) to process, annotate UKBB WES data and perform variant quality control. Dr Katherine Kentistou prepared the age at natural menopause phenotypes. Both Dr Gardner and I conducted gene burden and variant-level association testing using BOLT-LMM using MAF < 0.1%, while I additionally conducted it for MAF < 1%. All data analysis and

interpretation were done by me. I performed the constraint analysis of pathogenicity. In collaboration with Dr Gardner, Dr Alexander Mörseburg prepared the pipeline for the gene-set burden analysis on the RAP, while I conducted all the analysis. I performed so-called ‘misdiagnosis analysis’, while Dr Katherine Ruth calculated the expected frequency of having a gene with homozygous or compound heterozygous LoF knockout. Prof John Perry and Prof Anna Murray provided valuable advice on the analyses and writing of the manuscript, which is submitted at Nature Medicine.

4.1 Complex aetiology of extreme forms of menopause timing

Premature ovarian insufficiency (POI) is the loss of ovarian activity and permanent cessation of menstruation occurring before the age of 40[271]. It represents a major cause of female infertility, affecting 1 in 100 women[37,272,273]. POI can be caused by initially reduced ovarian reserve at the time of birth, accelerated loss of follicles, an inability of the remaining follicles to respond to ovulatory signals, or combination of all[274]. POI patients show a wide range of clinical phenotypes. Some are diagnosed with primary amenorrhea, usually identified at young age in individuals with delayed puberty and absence of menses. Others are presenting with a more common version, secondary amenorrhea, characterised by normal pubertal development and irregular menstrual cycle followed by amenorrhea[271]. POI can also occur in the syndromic form, in which it accompanies other phenotypic features, such as Turner's syndrome. Genetic causes of POI have been reported in 1-10% of cases while other causes include autoimmune and iatrogenic[275–279]. However, 50-90% of cases are idiopathic and likely involve a substantial genetic contribution[92]. Approximately 10-30% of idiopathic cases are familial, with more than one family member affected. Furthermore, heritability estimates of menopausal age between mother-daughter pairs range from 44% to 65%[76,82] and there is a six times increased risk of early menopause in daughters of affected mothers[280,281]. With current endocrine tests available in clinical practice, which only record changes in ovarian function that have already taken place thus limiting long term POI prediction, it is critical to better understand the genetics behind POI aetiology to enable early prediction and accurate diagnosis. A genetic diagnosis can provide important information to families about risks of POI in other family members, as well as provide understanding of the aetiology of the condition.

Over 100 single gene causes have been reported as being involved in POI pathogenesis. They cover a spectrum of biological processes including DNA repair, cell cycle and death, hormonal regulation and metabolism. Part of the genetic origin is explained through the existence of monogenic forms, with genes having an autosomal dominant (AD) inheritance (e.g. *BNC1*, *FANCA* and *NOBOX*). Other genes are described as being inherited in an autosomal recessive (AR) manner, requiring both copies of the gene to be disrupted in order to cause the phenotype (eg. *HFM1*, *LARS2* and *MCM8*). In addition to the autosomal genes, X chromosome abnormalities have long been known to play an essential role in the maintenance of ovarian development and function, representing about 13% of POI cases[282]. More recently, GWASs have identified ~300 common genetic variants associated with timing of menopause in the broader population. There is growing evidence that some POI cases may be polygenic in nature[274,282], where women inherit large numbers of common alleles associated with earlier menopause that, alongside other risk factors, push them into the pathological end of the phenotypic distribution.

With decreasing cost and improved analysis pipelines, WES is increasingly being used in the clinical setting as a powerful prediction and diagnostic tool for disorders of sex development, which encompasses POI[283–287]. However, the evidence on the reported POI genes is often based on small numbers of families or individuals without conclusive replication in an independent cohort, and with variable functional validation. Consequently, the importance of these genes being causative of POI pathogenesis is questionable[283]. In order to support more frequent clinical and non-clinical genetic testing, we need to ensure that we are examining valid candidates, i.e. that databases and literature are not populated by false positives but only well-validated, clearly pathogenic variants. This is especially important if interpreting the consequence of variants in one of these genes in the absence of family or functional evidence of pathogenicity.

To address this issue, we aimed to assess the penetrance of variants in known POI genes in the general population using UKBB study. We focused on the POI genes that are part of Genomics England (GeL) gene panel for POI, an expert reviewed and publicly available virtual panel database, additionally supplemented with manually curated literature-reported POI genes. Our results indicate that AD causes of POI are rare, and haploinsufficiency of recessive genes does not cause POI or early menopause. The reported individual AD variants were also detected in our control group with ANM above 40 years, thus introducing a high misdiagnosis rate of 98.4% if used in the clinical setting. We conclude that menopause under 40 years is likely to be a multifactorial trait in most cases and we highlight five novel genes associated with the reduction in ANM of between 1.5 and 5 years. Accurate estimates of the penetrance of genetic variants could significantly increase patient compliance with appropriate intervention strategies and fertility guidance, preventing misdiagnosis.

4.2 Methodology

4.2.1 Identification of POI gene candidates

In order to identify relevant gene candidates considered to be involved in POI aetiology, we initially focused on the POI gene panel available through GeL Panel App, publicly available virtual panel database (https://panelapp.genomicsengland.co.uk/panels/155/). This panel was selected as the 'gold standard' resource as it is the most thoroughly curated one, reviewed by 12 professional clinical geneticists. We considered the following evidence as part of our gene evaluation:
(1) Selection and categorisation: inheritance and phenotype, and (2) Number of reviews and gene ranking based on their traffic light system. This includes "RED" genes that do not have enough evidence for the association with the disease and should not be used for genome interpretation, "AMBER" genes with moderate evidence that should not be yet used for the interpretation, and "GREEN" genes with high level of evidence, which demonstrates confidence that this gene should be used for genome interpretation (**Appendix Table 4.1**). In total, we identified 67 genes: 28 green, 23 amber and 16 red.

This list was additionally supplemented with manually curated literature-reported POI genes, not provided as part of the GeL POI panel. We refer to these genes as "BLACK" to fit the traffic light categorisation. The search was performed using PubMed and Google Scholar, focusing on original articles published up to June 2022. The key word combinations included 'premature ovarian failure', 'primary ovarian insufficiency', 'premature ovarian insufficiency', 'early menopause', 'premature menopause', 'POI', 'POF', 'infertility', 'hypergonadotropic hypogonadism', 'ovarian dysgenesis', 'genetic variants', 'sequencing', and 'primary amenorrhea'. Studies were also identified by a manual search of original publications described in review articles. Where appropriate, reference lists of identified articles were also searched for further relevant papers. Identified articles were restricted to English language full-text papers.

Studies were included according to following criteria: (1) the phenotype of interest was described as POI, primary or secondary amenorrhea, (2) one or more affected individuals for particular causal variant were identified, (3) focus was on either autosomal or X chromosome, (4) genetic variants were discovered by traditional family segregation studies, consanguineous pedigree analysis, unrelated cohort studies on WES/targeted NGS data, and/or (5) variant discovery was supported by validation in animal models and/or cell based assays. We excluded studies that: (1) described hypothalamic pituitary adrenal axis (HPA) and/or puberty related phenotypes, (2) genes that were discovered via GWAS due to the lack of

statistical power as a result of small sample sizes and the challenge to locate causative genes, and finally (3) genes that were discovered through array analysis due to the high inconsistency of the results coming from varied resolution of arrays across studies and thus uncommon replications. We recorded and analysed genes described for either non-syndromic or syndromic POI, however the main focus of our paper was on genes associated with non-syndromic POI. We also considered papers that exclusively reported the role of candidate genes in animal models, yet these were only used as supporting evidence when assessing the functional evaluation of the gene and to guide our conclusions. Besides the causal genes, we recorded specific genetic variants reported as associated with the phenotype. Following information, available in the literature, were considered when assessing the evidence on the overall gene causality:
(1) Inheritance: AD, AR or X-linked, (2) Individuals' phenotype, (3) Sample size: number of the genetic variant carriers, cases versus controls, if reported, (4) Study type: WES, NGS, Sanger sequencing, or pedigree analysis, (5) POI classification: syndromic/nonsyndromic, (6) Male phenotype, if existent, and (7) Functional evidence: mouse model and/or cell-based assays, if existent. Using this manual curation approach, we identified 38 additional genes.

Overall, we identified 105 unique POI genes that we classified according to their mode of inheritance (**Appendix Table 4.1**, **Figure 4.1**). Genes were considered as inherited through the AD pattern if the reported variants in the heterozygous state were sufficient to cause POI, leading to 39 genes in total (**Figure 4.1A**). Of those, 7 were reported to act through the LoF mechanism only, while in 32 genes both LoF and missense genetic alterations caused the phenotype. If variants in both copies of the gene were necessary for the phenotype development the gene was classified as recessive (N=57). For two genes both dominant and recessive causes were identified, while 7 genes had an X-linked inheritance pattern.

4.2.2 UKBB Data Processing and Quality Control

The UKBB data processing and QC were performed as described in **Chapter 2**, **Section 2.2.3**.

4.2.3 Phenotype derivation

The ANM phenotype preparation is extensively described in **Chapter 3**, **Section 3.2.2**. This winsorized phenotype, so-called 'ANM 34' (N=106,973), where everyone reporting ANM younger than 34 was coded with ANM of 34, was treated in the discovery analysis as the primary one.
In addition, we prepared the binary POI phenotype. Here, the individuals reporting ANM lower than 40 years old were treated as potential POI cases (N=2,569, mean ANM in cases of 35.65 ± 0.03 years).

However, due to the high case-control imbalance in the POI phenotype, the ANM34 one was used for the primary, while the secondary association analysis was conducted on the POI phenotype. All manipulations were performed in R (v4.1.2) on the UKBB RAP (https://ukbiobank.dnanexus.com/).

4.2.4 Exome-wide association analyses in the UKBB

Rare variant burden tests were performed using a custom implementation of BOLT-LMM v2.3.6[154] for the RAP, as described in **Chapter 2**, **Section 2.2.3**. In order to examine a range of variant annotation categories for MAF < 0.1%, we created dummy genotype files for HC PTVs as defined by LOFTEE, missense variants with CADD ≥ 25, and damaging variants that included both HC PTVs and missense variants with CADD ≥ 25. BOLT-LMM was then run with default parameters as previously described. As BOLT-LMM association test statistic is a more powerful method for quantitative traits, yet less powered for unbalanced case-control traits like POI[288], we used 'ANM34' phenotype to derive and analyse primary burden test statistics. 'ANM 34' phenotype was run as a continuous trait corrected by age, age^2, sex, the first ten genetic PCs as calculated in Bycroft *et al.*[147] and study participant exome sequencing batch as a categorical covariate (either 50k, 200k, or 450k). The same methodology was applied when performing the secondary analysis on the POI phenotype.

To assess whether reported heterozygous monogenic causes of POI have full penetrance as previously suggested, we specifically looked at the menopausal age of all women who carry HC PTVs, as these variants are considered to have the highest impact on protein function (**Figure 4.2, Appendix Table 4.2**). Using the derived POI phenotype, we obtained the binary POI status of individuals ('0': ANM > 40 years and '1': ANM < 40 years), and calculated the number of carriers, i.e. number of cases and controls, per variant. In addition, we used quantitative 'ANM34' phenotype to derive the menopausal age range, i.e. minimum, average and maximum ANM across all carriers of HC PTVs. The ANM age range was also calculated at the gene level, collapsing all HC PTVs per gene and identifying the lowest, average and highest ANM reported (**Figure 4.2**). We report the ANM age range and burden level results for the HC PTV mask in the **Appendix Table 4.2**.

Finally, to test the association of less penetrant mutations with the phenotype, we considered variants with both missense and LoF consequences, incorporating them in 3 previously described masks: HC PTV, damaging and missense variants with CADD ≥ 25. The significant gene-level associations for ANM were identified by applying Bonferroni correction for the number of masks in 105 genes of interest (correction term: 3 masks * 105 genes; *P*: $0.05/315 = 1.6*10^{-4}$) (**Figure 4.3, Appendix Table 4.3**).

4.2.5 Constraint metric of pathogenicity

In addition, we recorded the Genome Aggregation Database (gnomAD) v2.1.1 predicted constraint metric of pathogenicity to identify genes that are subject to strong selection against PTV variation[164]. The metric encompassed observed and expected variant counts per gene, observed/expected ratio (O/E) and probability of loss of function intolerance (pLI) (**Appendix Table 4.1**). In short, observed count represents the number of unique SNPs in each gene (MAF < 0.1%), while expected count relies on a depth-corrected probability prediction model that takes into account sequence context, coverage and methylation to predict expected variant count. The O/E is a continuous measurement that takes into account gene and sample size and measures how tolerant a gene is to a certain class of variation. Low O/E value indicates that the gene is under stronger selection for that class of variation. Finally, the pLI score reflects the constraint or intolerance of a given gene to a PTV variation, with a score closer to 1 indicating that the gene cannot tolerate PTV variation.

4.2.6 Identified AD genes as a fictive 'POI Gene panel'

4.2.6.1 Gene-set burden analysis

We ran gene-set burden tests by collapsing the genes of interest and their variants into one unit for analysis. The gene-set burden tests were performed by extending an association testing workflow of applets designed for the UKBB RAP for single genes to gene-sets. The RAP association workflow is described in detail in **Chapter 2**, **Section 2.2.3** and Gardner *et al*, 2022[289]. In total, we conducted four gene-set burden tests, collapsing variants and genes into following categories: (1) AD genes (N=39), (2) AR genes (N=57), (3) genes with both AD and AR inheritance (N=2), and (4) all 105 genes.

Briefly, for each of the gene-sets we included variants that fulfil the following criteria: a) MAF < 0.1% and b) HC PTVs as predicted by the LOFTEE tool[164]. For each gene-set we ran two related approaches. First, we implemented a generalised linear model (GLM) using the Python package 'statsmodels'[165]. For the GLM, the number of variant alleles across the gene-set was summed up into a single score under a simple additive model. This score was used as a predictor of the ANM phenotype in a three-step regression.

Second, we ran the STAAR method, implemented in R package "*STAAR*"[290].
This method corrects for population stratification by including a genetic relatedness matrix (GRM) in the test framework. The GRM used by us was based on pre-computed autosomal kinship coefficients from

Bycroft *et al*[147]. For each STAAR test the genotype information was represented by a single n*p matrix where n was the sample size and p the number of included genetic variants across all genes of interest. For all association tests we corrected for age, age^2, the first ten genetic PCs provided by Bycroft *et al*[147] and study participants WES batch as a categorical covariate.

4.2.6.2 'Misdiagnosis' analysis

We then wanted to investigate how many women of 2,569 POI cases in UKBB are carriers versus non-carriers of HC PTVs in AD genes of interest. We calculated the percentage of carriers and non-carriers and used it as a suggestive metric of the number of individuals that would be misdiagnosed if using reported AD genes as the clinical diagnostic panel.

4.2.6.3 Estimating of frequency of homozygous or compound heterozygous LoF individuals in the population

We estimated the frequency of homozygous or compound heterozygous LOF individuals for each gene as F^2, where F is the frequency of individuals with any HC LOF allele with MAF<0.1% in a gene as estimated from the primary analysis. To find the total frequency of individuals with homozygous or compound heterozygous LOF knockouts, we then summed F^2 for the 105 POI genes reported in the literature.

4.3 Results

4.3.1 Heterozygous loss-of-function is not a common cause of POI

The GeL POI Panel App includes 67 validated genes rated as either 'GREEN' (high level of evidence for disease association), 'AMBER' (moderate evidence) or 'RED' (not enough evidence). We also identified a further 38 genes from the literature with good evidence of being causal for POI that we refer to as 'BLACK' in this case. This gave a total of 105 genes, which we classified according to the reported mode of inheritance (**Figure 4.1, Appendix Table 4.1**).

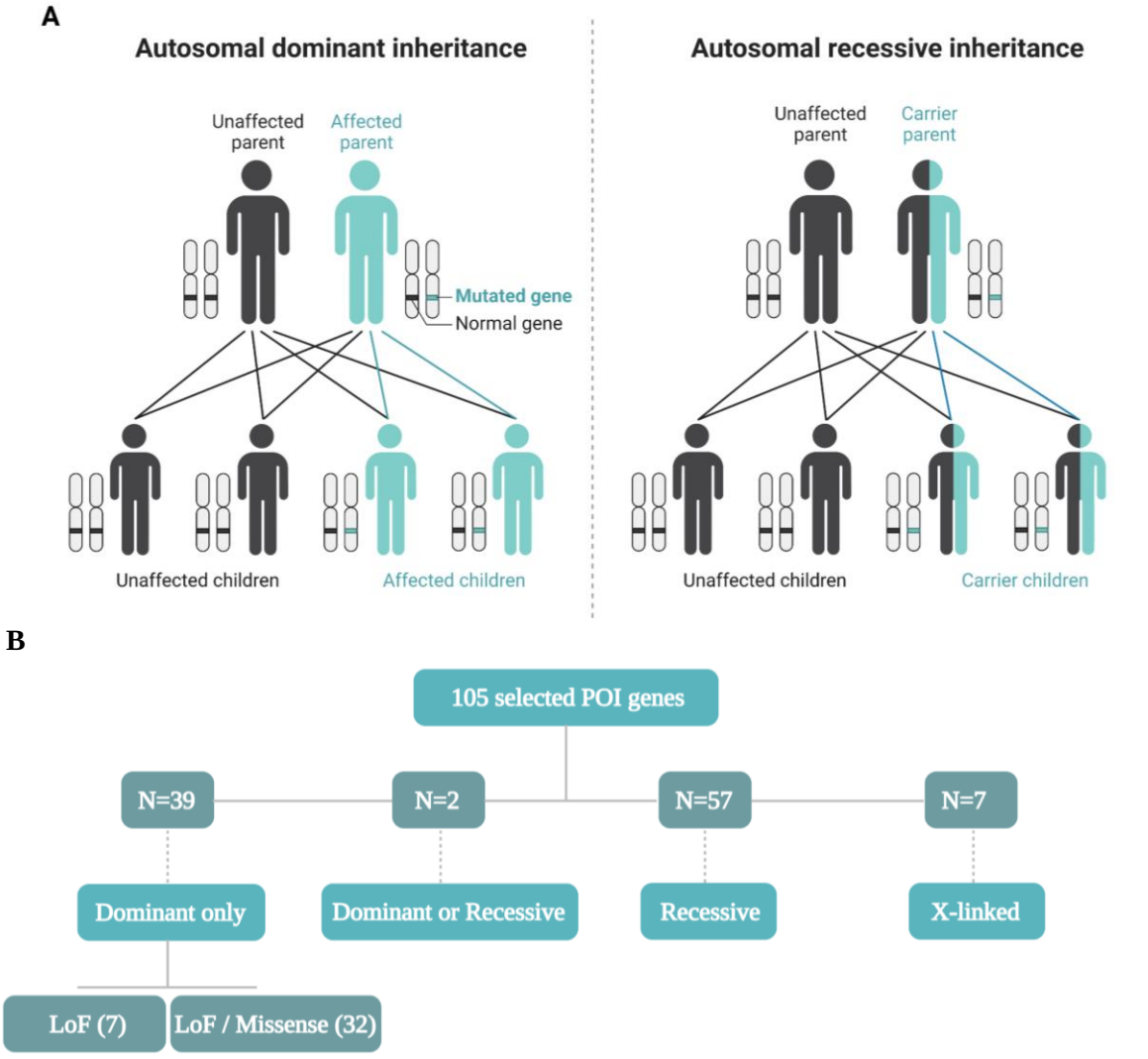

Figure 4.1: Mode of inheritance for POI genes. (A) *The schematic describes the autosomal dominant and recessive inheritance pattern. Genes were considered as inherited through an autosomal dominant (AD) pattern if the reported variants in the heterozygous state were sufficient to cause POI. If variants in both copies of the gene were necessary for the phenotype development the gene was classified as recessive. For some genes we found both AD*

and AR mode of inheritance. ***(B)*** *The inheritance classification and number of POI genes per category.* ***LoF:*** *loss-of-function.*

We used WES data to identify genetic variants in these 105 POI genes, available in 106,973 post-menopausal UKBB female participants of European genetic-ancestry[158], of which 2,569 reported ANM below the age of 40 and were thus considered as POI cases. We initially tested heterozygous loss of the 41 genes with a reported dominant mode of inheritance, by combining all LOFTEE[164] predicted HC-PTVs with MAF < 0.1% in each gene, and assessing their association with menopause timing per gene. This enabled us to explore the most damaging genetic changes, which introduce perturbations that should yield a severe functional defect. We used the 'ANM 34' phenotype for this primary association test based on the insights on shared genetic susceptibility between POI and ANM, and thus the opportunity for more robust identification and validation of genetic mechanisms involved in POI[76,82]. This is especially critical as the high case-control imbalance that characterises the binary POI phenotype limits us to derive well powered conclusions. There were 43 of 2,569 women with menopause under 40 years who carried a HC-PTV in at least one of the 41 genes, but these same variants were also detected in 1,823 (ANM range= 40-63, mean=50.4, SD= 3.9) controls (ANM >40 or still menstruating after 40). For all 41 genes, HC-PTVs were identified in both the case and the control group, with mean ANM for heterozygous LoF carriers in each gene between 45 and 56 years (**Figure 4.2, Appendix Table 4.2**). For three (*BMPR1A, FOXL2* and *NR5A1*) of the 41 genes we did not identify any women carrying HC-PTVs in our study, thus we were not able to assess their association with menopause timing. Similar results were observed when we considered only missense variants with CADD ≥ 25. Finally, we performed the secondary analysis in the POI phenotype and found consistent results as with the primary analysis. Overall, we could not find evidence that any of the assessed AD genes was completely, or even partially penetrant to cause POI.

The high intolerance towards protein truncating variation, so-called constraint, has previously been linked to reduced reproductive success[291]. Our results demonstrate that a significant proportion of AD genes (28/41, 68.3%) have limited evidence on being under strong selective constraint (pLI ≤ 0.9) as assessed by gnomAD, which further supports that these genes might not play an important role for reproductive success (**Appendix Table 4.1**).

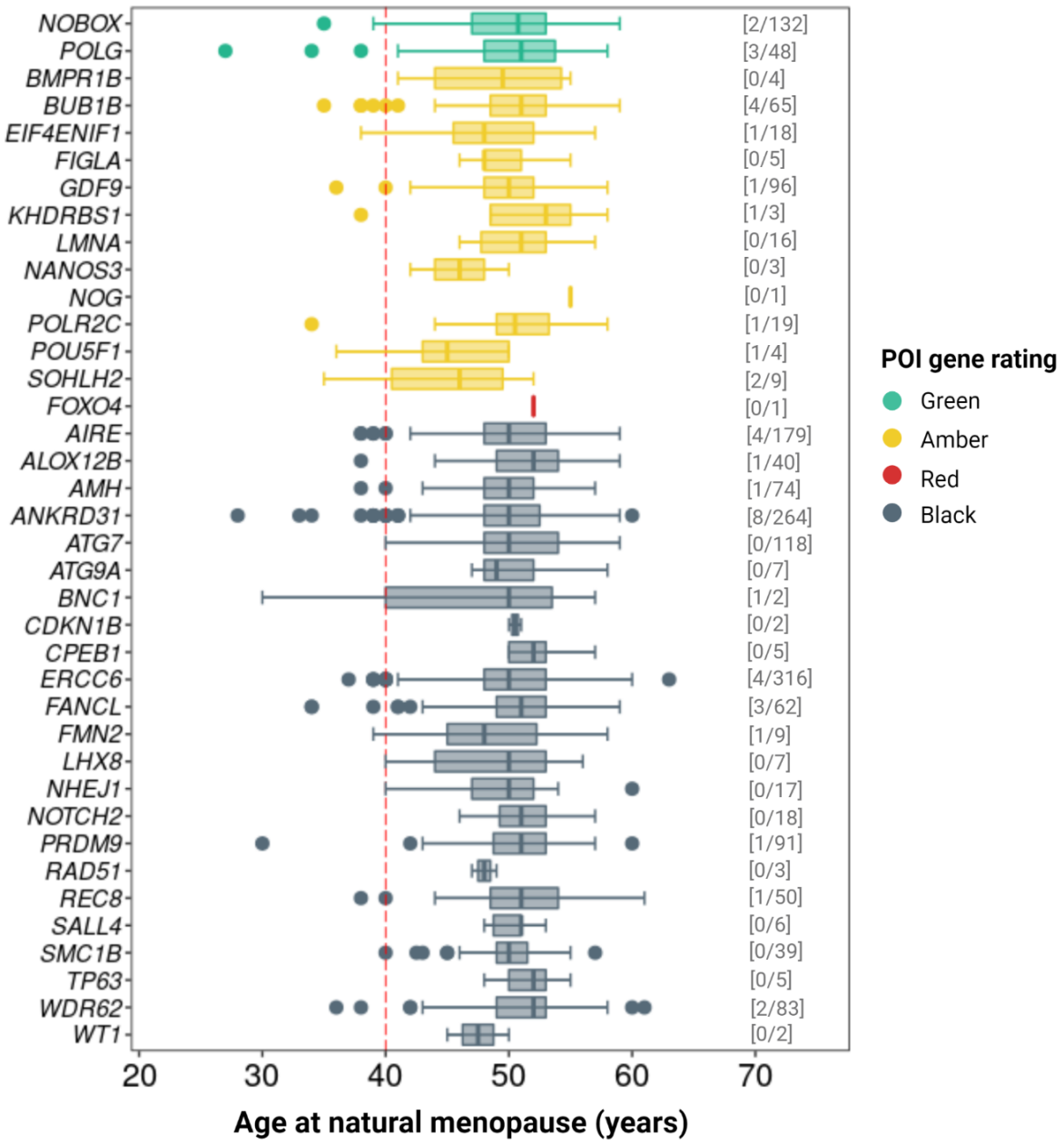

Figure 4.2: Range of age at natural menopause in carriers of HC-PTVs in genes reported to have an autosomal dominant pattern of inheritance. *17 genes were identified as 'monoallelic' in GeL Panel App and are coloured according to the strength of evidence categories: "GREEN", "AMBER" and "RED". In addition 24 genes, here named as 'BLACK', were reported in the literature to be a likely monogenic cause of POI in the heterozygous state, but were not included on the Panel App. The numbers in brackets in the right corner reported as part of each panel represent* ***[N POI cases/N controls]*** *of women carrying HC PTVs in each gene.*

Next we tested individual variants that have been reported previously to be pathogenic for POI in the 41 AD genes (**Appendix Table 4.4**). Of the 179 variants in the literature, 160 were present in post-menopausal women in the UKBB, all in the heterozygous state. For the 31 variants detected in women

with ANM < 40 years, all except one (p.Arg943His in *POLG*) were also present in controls, while the remaining variants were only found in controls. A single predicted deleterious missense variant c.2828G>A (p.Arg943His) in the *POLG* gene was found in a woman who had menopause at 34 years (beta: 16.8 years earlier ANM, 95% CI: 8.5-25.1, P=1.1*10^{-5}), but not in the control group. Our results indicate that previously reported AD POI genes are generally not pathogenic in the heterozygous state and that AD causes of POI are rare in the population.

It is important to note that most of these AD genes are part of the diagnostic gene panels, which are increasingly being used within the clinical setting to assess and detect the genetic cause of POI. We thus imitated a scenario where we predict the rate of potential genetic misdiagnoses of POI using the reported AD genes as a diagnostic gene panel. We examined our cohort of POI cases (N=2,569) and calculated the proportion of women who presented with or without variants in AD POI genes. Only 1.6% of women were carrying the protein truncating variants in these genes, indicating that 98.4% of women would have no diagnosis while 1.6% would receive an incorrect genetic diagnosis if using these genes as part of the POI panel. Finally, to further explore previous observation, we evaluated whether this 1.6% of women had any difference in menopause age relative to the 98.4%. We created gene-set burden scores consisting of all HC-PTVs in sets of AD POI genes, to assess whether collapsing all LoF variants into one score would identify an association with the phenotype. This included a score for: (1) AD genes (N of genes=39), (2) AR genes (N=57), (3) genes with both AD and AR inheritance (N=2), and (4) all 105 POI genes. All scores showed no association with the ANM phenotype, in both GLM and STAAR Omnibus statistical models (**Table 4.1**).

Table 4.1: POI gene-set burden scores.

POI gene-set category	N carriers	N variants	GLM Effect	GLM SE	GLM P	STAAR Omnibus P
[1] AD genes	1744	499	-0.036	0.103	7.25*10^{-1}	7.15*10^{-1}
[2] AR genes	6288	1768	-0.065	0.054	2.26*10^{-1}	9.40*10^{-1}
[3] AD / AR genes	102	47	0.247	0.432	5.67*10^{-1}	2.28*10^{-1}
[4] All genes	8120	2357	-0.053	0.047	2.63*10^{-1}	1.63*10^{-1}

4.3.2 No evidence of haploinsufficiency of recessive genes as a cause of POI

Of the 105 monogenic POI genes we selected to test in our study, 57 were reported to cause POI through a recessive mechanism only and a further 7 were X-linked. We were unable to evaluate compound heterozygotes, but did identify 2 women with homozygous HC-PTV (9:135697628:C:T, hg38) in *SOHLH1* gene with normal ANM, defined as controls. No individuals with ANM < 40 were identified for this variant [N carriers: 34, ANM range (min - mean - max): 42 - 50.3 - 59]. Our results suggest that homozygous variants in *SOHLH1* recessively inherited gene are not true causes of POI.
We hypothesised that there may be a heterozygous effect for carriers of deleterious variants in these recessive genes. There was no association with menopause in the normal range for heterozygous variants in 62 of the 64 genes ($P > 0.05$). In two genes we did identify an evidence for a heterozygous effect in HC-PTVs on menopause timing, *BRCA2* and *HROB*, which have both recently been reported as genetic determinants of ANM[191]. This evidence extends on our previous GWAS analyses of age at natural menopause[82], which identified an individual low-frequency PTV variant in *BRCA2*. There was however no evidence that haploinsufficiency of these recessive genes is sufficient to cause POI (*BRCA2* effect: 1.18 years earlier ANM [95% CI: 0.72-1.65] with $P=2.6*10^{-7}$, and *HROB* effect: 2.89 earlier ANM [95% CI: 1.86-3.92] with $P=1.9*10^{-8}$) (**Figure 4.3**).

The expected frequency of having a gene with a homozygous or compound heterozygous LoF knockout would be $6*10^{-9}$ individuals (median frequency in gnomAD)[292]. Assuming 100% penetrance, 1.6 million genes would be required to carry such a knockout in order to reach the observed 1% frequency of POI in the population. If all 20,000 genes in the genome carried such a knockout, this would still only result in a population frequency of 0.012%.

Finally, the expected frequency of individuals who would be homozygous or compound heterozygous for a HC LoF variant in any of the 105 POI genes was 0.003%, thus we would expect about 3 to 4 such individuals in our study cohort.

4.3.3 Three new genes associated with variation in menopause timing in the normal range

To further assess the impact of rare damaging variants in other POI genes on menopausal timing in the normal range, we conducted individual gene burden association tests by collapsing genetic variants with MAF < 0.1% into three functional categories. These included: HC-PTVs (previously described), missense

variants with CADD score ≥ 25, and 'damaging' variants, defined as combination of HC-PTVs and missense variants with CADD ≥ 25. Besides previously described *BRCA2* and *HROB*, we identified rare variation in three genes associated with earlier menopausal timing after multiple test correction (*P* (0.05/315 masks) ≤1.6*10^{-4}, **Figure 4.3, Appendix Table 4.3**). The novel associations were detected for damaging variants in *TWNK*, a mitochondrial helicase involved in mitochondrial mtDNA replication and repair (beta= -1.39, 95% CI:0.78-2.00, *P*=8.5*10^{-6} for 185 damaging *TWNK* carriers)[293,294], *NR5A1*, a key gene for gonadal function (beta= -1.75, 95% CI: 1.02-2.47, *P*=2*10^{-6} for 131 *NR5A1* carriers)[295], and *SOHLH2*, a transcription factor involved in both male and female germ cell development and differentiation (beta= -3.26, 95% CI: 1.56-4.96, *P*=1.6*10^{-4} for 24 *SOHLH2* carriers)[296,297]. It is interesting to note that we did not identify any carrier of HC-PTVs in *NR5A1*, additionally supporting the evidence of this gene being under extreme constraint for this type of genetic alteration in population (**Appendix Table 4.1**).

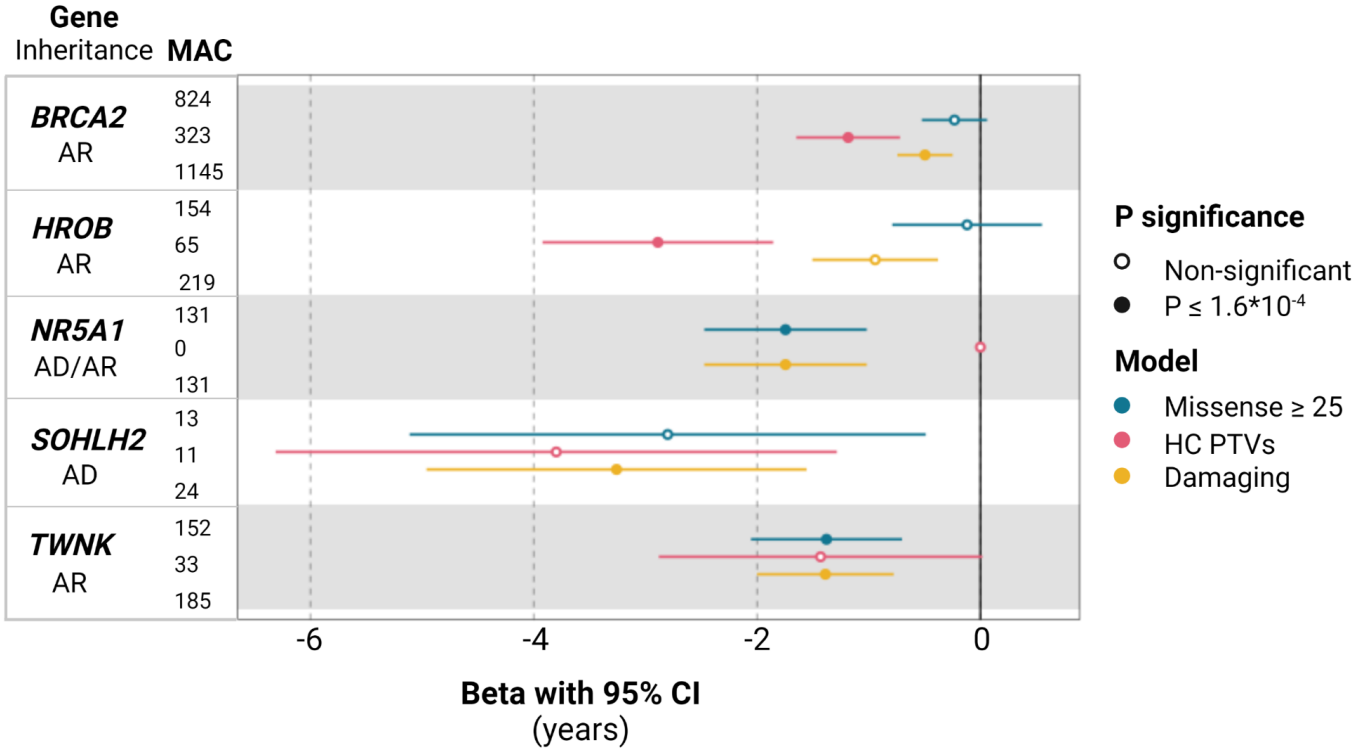

Figure 4.3: Forest plot for significant gene burden associations with age at natural menopause. *Exome-wide significant (P ≤ 1.6*10^{-4}) genes are displayed. Points and error bars indicate beta and 95% CI for the indicated variant category. The gene burden association test was performed using BOLT-LMM.*

These findings have strong implications for clinicians diagnosing causes of POI, where described genes should be interpreted with caution to avoid misdiagnosis as they do not have plausible clinical significance.

4.4 Discussion

Many genes have emerged as monogenic causes of POI, but a majority has been identified as causative in small numbers of families or individuals, with variable functional validation. Our study is the first to demonstrate that AD mutations in genes currently described in the literature or evaluated in clinical diagnostic panels are not common, highly penetrant causes of menopause under 40 years. We tested 105 genes in total. This includes 68 genes with a pathogenicity rating for POI reviewed in the GeL open access PanelAPP resource[298], as well as additional 37 manually curated genes reported in the literature. Of these, 41 are reported to be inherited in an AD fashion, i.e. heterozygous variants are sufficient to cause the phenotype. In order to assess the most damaging genetic changes in these genes that are expected to introduce perturbations with severe functional defects, we initially focused on LoF variants identified via robust and sensitive LOFTEE predictor, the mechanism implied by most studies[164]. For each gene the mean menopause age for carriers of LoF variants in UKBB was over 40, with the distribution of menopause age broadly mirroring that of non-carriers. We were also able to test the effect of 160 of 179 individual variants that have previously been reported to be likely causal variants for POI in the heterozygous state. For each variant there was at least one woman who had gone through menopause over 40 years or was still menstruating over 40. For 17 of these there was an allele frequency > 0.1% in controls. The presence of genetic variants at relatively high frequency in controls is a strong indicator that the variant is not causal and should always be carefully considered when assessing likely pathogenicity within the clinical and non-clinical setting. We also found no evidence that haploinsufficiency of AR genes causes POI. Of the genes we tested, 55 had a purely recessive inheritance mechanism reported and for these genes we would not necessarily expect to see an effect of heterozygous variants. There are examples of genes where a haploinsufficiency effect is seen for recessive genes[299], but this does not seem to be the case for POI.

The evidence to support causality of genes and variants in the literature is variable. Guidelines are available for genomic variant interpretation[300]. However, many publications are historical and published before the guidelines, thus making it difficult for non-genomics specialists to interpret the findings. Databases of clinically relevant variants in disease causing genes, such as ClinVar and Decipher, are also available to aid clinical interpretation, and panels of monogenic POI genes have been proposed. For example, four AD POI genes are rated as ‘GREEN’ on the GeL PanelApp, and we found heterozygous LoF variants in two of them, *NOBOX* and *POLG*. We did not find any LoF variants in *FOXL2* and *NR5A1*, which could be explained by high constraint pLI scores of 0.88 and 0.99, respectively. The relevance of LoF in *FOXL2* and *NR5A1* thus remains to be elucidated in larger datasets in future. LoF

variants in the other two green PanelAPP genes, *NOBOX* and *POLG*, were much more common in women with menopause over 40 years than those with ANM under 40.

Most women with isolated POI only have karyotype and Fragile X (*FMR1*) testing with potential pelvic ultrasonography [89,301]. This screening might be helpful for the identification of POI aetiology, however it is well established that most of the cases ultimately remain without clarified aetiology [302]. Given that POI is accompanied by a high rate of infertility, identifying the causal gene in nonsyndromic POI families has proven difficult as the majority of these cases have small or no family histories [285,303]. WES approaches have also been increasingly used to identify causal genes and variants, but often relying on candidate genes to narrow down the likely variants [184,285], providing circular evidence of causality. Therefore, the robust assessment of the penetrance of gene candidates becomes even more critical with higher accessibility of the genomic data and thus more frequent individual diagnostic endeavours in the absence of family history information [283,304]. Current commercial POI gene panels mostly rely on NGS and claim to make a directed and accurate differential diagnosis of infertility, ultimately leading to a better management of the patient. Using web search, we identified ~10 companies offering gene panel testing for POI, including Igenomix 'Premature Ovarian Insufficiency Precision Panel'. In order to forecast the outcome of the scenario where these gene panels are utilised for diagnostics in the clinical setting, we predicted the rate of potential genetic misdiagnoses of POI using the reported AD genes (N=41) as a diagnostic gene panel within the cohort of POI cases (N=2,569). Only 1.6% of women were carrying the protein truncating variants in these genes, indicating that 98.4% of women would have no diagnosis while 1.6% would receive an incorrect genetic diagnosis if using these genes as part of the POI panel. Most often, post-diagnostic procedure would lead to the risk assessment and genetic counselling of asymptomatic family members, thus introducing additional emotional burden. Providing accurate diagnosis would significantly increase patient compliance and enable female carriers to undertake appropriate intervention and plan their conception before ovarian failure occurs. This possibility is becoming more and more important as women tend to conceive more frequently in their 30s and 40s when the risk of POI is between 1-2%. POI not only interferes with a woman's reproductive potential, but it is also associated with an increased risk of osteoporosis and cardiovascular disease. Accurate and timely diagnosis of POI would allow clinicians to prescribe HRT early on to prevent problems associated with oestrogen, thus optimising bone, cardiovascular and overall health.

Although heterozygous variants were not fully penetrant causes of POI, our study suggests that a higher burden of rare coding variants in five of the POI genes can substantially reduce average menopause age. Two of these genes have already been identified by WES analysis in **Chapter 3**, *BRCA2* and *HROB*. Three other genes have not been described previously as associated with menopause timing in the general

population: *NR5A1*, *SOHLH2* and *TWNK*. The effect ranged from 5.13 years earlier ANM for carriers of rare LoF variants in *SOHLH2* to 1.54 years earlier ANM for damaging variants in *TWNK*. *NR5A1*[295,305] and *SOHLH2* are both on PanelApp, rated green and amber respectively and both have a dominant mechanism. Previous studies have suggested that *NR5A1* could have an oligogenic mode of inheritance, in which multiple variants individually contribute to the phenotype[306–309]. *NR5A1* encodes an orphan nuclear receptor that regulates transcription of an array of genes involved in reproduction, steroidogenesis and male sexual differentiation. The conditional knockout of *NR5A1* in mouse granulosa cells causes infertility, as a result of hypoplastic ovaries and a reduced number of oocytes[310]. Mutations in this gene were associated with the wide spectrum of reproductive phenotypes due to altered folliculogenesis, including gonadal dysgenesis with PA or SA, as well as other disorders of sex development[295,305]. These studies confirm that haploinsufficiency seems not to explain the highly variable phenotype[306]. This is because subjects harbouring identical *NR5A1* disease-causing variants may present with completely different phenotypes, as in the case of heterozygous *NR5A1* p.Arg255Leu/Cys variant that was detected in a 46,XX female with adrenal failure, but intact ovarian function[311] and a 46,XX female with normal adrenal function but POF[305].

SOHLH2 is a transcription factor acting as master regulator of oocyte-specific genes critical for early follicle growth and differentiation, including *NOBOX*, *FIGLA*, *BMP-15*, and *GDF-9*, and expressed exclusively in primordial follicles up until the primary follicle stage. *Sohlh1/2-/-* deficient mice present with infertility and atrophied ovaries characterised by accelerated follicle loss due to defective primordial-to-primary follicle transition [296,312,313]. Finally, *TWNK*, mitochondrial helicase involved in replication and repair of mtDNA, causes syndromic POI (Perrault syndrome) and presents in association with other neurologic findings[314]. The reproductive phenotype is specific to women - in cases where male siblings carry the exact same variant as the affected female sibling no reproductive abnormalities are detected[314]. Rare variants in these genes therefore predispose women to earlier menopause and are likely to add to the genetic burden that predisposes to POI. Previous GWAS studies have suggested that polygenic risk has an effect on risk of POI, ie. women with POI have a higher burden of common genetic variants associated with menopause timing than those with average menopause age[82]. While POI can have a monogenic origin where the mechanism is predominantly recessive LoF, many POI cases may in fact be caused by a higher than average load of common and rare genetic variants that each influence menopause timing by a few months or years, but in combination can result in menopause before 40 years leading to a diagnosis of POI. This suggests highly oligogenic or polygenic POI nature.

Our study reveals how despite functional evidence in some of these genes, their clinical significance should be interpreted with caution as they are not always fully penetrant and not sufficient to cause POI in isolation.

Functional models for POI, including cell lines and animal models, are currently limited and future studies that aim to investigate novel genetic causes of POI should focus on approaches that can more specifically mimic human biology and physiology. Patient-specific induced pluripotent stem cells (iPSCs) lines might offer an individually targeted genetic model for identification, manipulation and better understanding of reproductive biological pathways.

The frequency of some POI variants is extremely low, compatible with the incidence of POI. It could be that some monogenic AD causes of POI are much rarer in the population than the PTVs that we are testing here, thus larger scale studies are necessary to address this limitation. One of the main advantages of our study is the ability to evaluate the joint effect of damaging mutations at the gene level via burden testing in a large number of participants. Although an important limitation is the fact that we have not assessed a clinically-defined cohort of POI cases, the power of our approach is that we can demonstrate these genes are false positives by examining them in a large number of healthy non-POI women, i.e. controls. Clinically defined POI cohort would only be critical if this study was aiming to identify novel POI effects. Notably, the UKBB study is known to be biased towards healthier participants and thus highly penetrant monogenic disease-causing variants may be under-represented, yet it is difficult to imagine how having POI would influence participation in the study. In addition, one could argue that some women report an incorrect ANM due to misremembering, however we took into account four instances where questions regarding ANM were asked thus ensuring consistency in the reported ANM. Secondly, we were only able to only assess the penetrance of heterozygous variants and not homozygous or compound heterozygous carriers. We have also not considered complex structural variants or cytogenetic abnormalities, so we make no statement on the penetrance of those. For five genes we did not identify any heterozygous LOF variants so were unable to assess these, although we can rule them out as common causes of POI given they were not present in over 2000 cases. Given our observed results for genes with a dominant mode of inheritance, we advise caution in interpreting reported recessive effects, although we predict this will be by far the most common cause of monogenic POI. Third, we predominantly focussed on predicted LOF alleles as that is the mechanism implied or demonstrated in most studies. It is however possible that some of the literature reported missense variants may act in a gain of function or dominant negative manner such that they have more severe effects than protein truncated variants. Whilst potentially true of a small number, this is unlikely to be widespread given no highly penetrant effects were seen in the individual literature reported missense variants we were able to assess, or in the burden tests we performed that were restricted to damaging missense variants predicted by CADD.

Finally, our study is specific to individuals of European ancestry, thus we cannot comment on how replicable these results are in other ancestries, although while specific variants are likely to vary between populations[315], we focused on testing the loss of function mechanism, which should be widely applicable.

The dependence of ovarian function on complex gene networks probably explains the poor correlation between the genotype and phenotype. The origin of POI development may not be due to a single mutation in a candidate gene, but an interaction of low frequency polymorphisms or mutations in different genes in the same woman. This suggests that less penetrant mutations may be more frequent in POI individuals as a genetic cause, which aligns with few previous studies that hypothesised an oligogenic aetiology for this disorder[316,317]. Using panels of genes to find causative genetic variants for un-related idiopathic POI might not be a very fruitful endeavour and is unlikely to be cost-effective. Monogenic causes of POI are more likely to be recessive and therefore much rarer in the POI population given 1% of women in the population have it.

CHAPTER 5

Human proteomic analysis of menopause timing

Summary

Chapter 5 describes the first proteo-genomic study performed on ANM phenotype to identify protein candidates that could serve as biomarkers of ovarian ageing in women. The analysis was performed on genome-proteome-wide association data on 4,775 protein targets (4,979 human somamers, SomaLogic) available in the Fenland study on 10,713 participants and data on ANM for ~200,000 women from the ReproGen study. We do not identify any robust protein candidate whose levels are altered due to ovarian ageing. However, we demonstrate the potential of such analysis, and propose that future attempts should focus on the studies with larger sample size, once available.

Contributions and Collaborations

Dr Mike Pietzner and Dr Eleanor Wheeler provided me with genome-proteome-wide association summary statistics. I created the menopause GRS under the supervision of Dr Jian'an Luan, while Dr Erin Oerton performed linear regression analysis using individual level data. I used summary statistics data to perform bidirectional Mendelian Randomisation analysis in the discovery and replication meta-analysis. Meta-analysis data were provided by Dr Luan and Dr Wheeler. Prof John Perry, Prof Ken Ong and Prof Claudia Langenberg provided valuable advice on the analyses of the data.

5.1 From genome to menopause timing via proteome

Most genetic-related research on reproductive ageing has focused on the assessment of common and rare genetic variation and their contribution in regulation of the menopause timing in women[82,189,318]. As previously discussed, one of the major challenges that human genomics faces represents identification of the causal genes and understanding of the mechanisms by which mutations and trait-susceptibility alleles act to modify the phenotype[185]. As a consequence, this limits efficient translation of genomic findings for the development of prediction and treatment strategies. To address these challenges, scientists are implementing robust functional techniques, including tissue-specific gene expression data and CRISPR screens, which can be applied on a variety of *in vitro* and *in vivo* biological models, to functionally characterise identified loci and assign causal genes[319–323]. However, these cellular and animal models might not fully replicate complex human reproductive biology and regulatory processes that are required for DNA to RNA transcription and RNA to protein translation, thus leading to low correlation between transcripts and proteins. In addition, functionally testing large numbers of GWAS candidates in animal models is not cost- and time-efficient and it is ethically questionable. Alternative methods for gene prioritisation are particularly welcome. This chapter explores the proteogenomic approach that focuses on the essential functional units of the human body, the proteins, which are the central layer of information transfer from genome to the phenome [324,325]. This approach offers an opportunity for identification of gene and protein targets that can improve our understanding of novel mechanisms underlying reproductive ageing. Studying the human proteome has many benefits - firstly, proteins represent the largest class of drug targets, indicating that prioritisation and translation of protein candidates from 'bench to bedside' seem to be fruitful and effective[324–326]. Secondly, proteins are measured from readily available biofluids, such as blood. Therefore, they represent an attractive biomarker for prediction of menopause timing due to the straightforward implementation within the clinical setting as their measurements are anyway done as part of the regular checkups [327,328]. Implementing the knowledge from proteogenomic studies could help us address one of the main challenges related to reproductive ageing, i.e. the lack of long-term biomarkers of ovarian reserve. Identification of women with reduced reproductive lifespan cannot be accurately achieved by any endocrine or imaging tests that are in clinical practice, such as AFC on ultrasound, and levels of AMH and FSH[52–55]. These tests only record changes in ovarian function that have already taken place, disabling the long-term prediction of reproductive expectations[54,56–58]. Identifying novel protein biomarkers might help us more accurately predict the age at which a woman will become menopausal, which would open up the opportunity for any young woman to be tested for reproductive expectations and counseled on the availability of elective fertility preservation.

Although major technological advances have enabled studying human proteome at the large scale, most of the studies still focus on bespoke panels[329–331] and proteomic platforms[324–326,332], thus leaving the gap in knowledge about the genetic architecture and relevance of most proteins for human health. To address this gap and provide insights into the biology of various human diseases, Pietzner *et al* (2021) started paving the path towards the creation of first genome-wide proteogenomic maps by undertaking at the time the largest genome-proteome-wide association study that included 4,775 protein targets in the Fenland cohort of ~10,700 European descent individuals (mean age 48.6 years, 53.3% women)[111,333]. They identified 10,674 genetic associations ($P < 1.004 \times 10^{-11}$) for 3,892 plasma proteins and created a cis-anchored gene-protein-disease map of 1,859 connections that points towards strong cross-disease biological convergence. However, the relevance of human proteome in reproductive ageing remains unexplored.

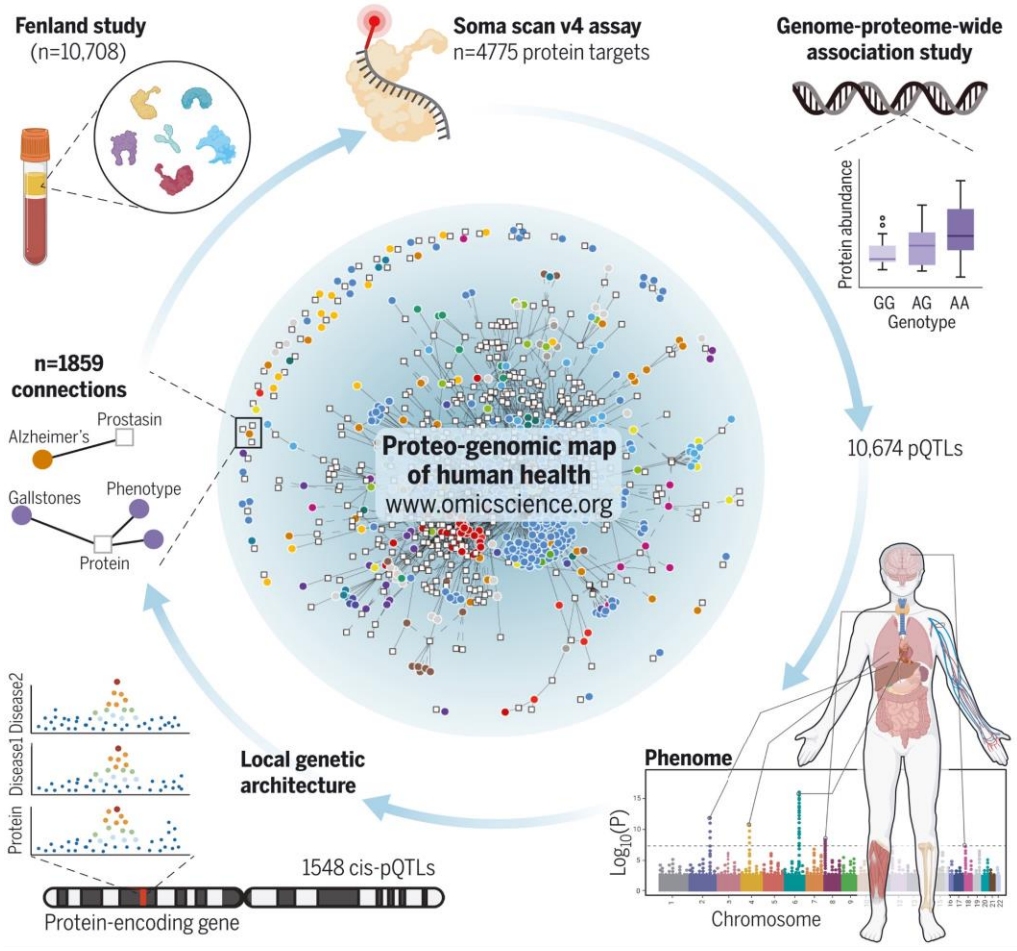

Figure 5.1: Graphic representation of the study design to construct a proteo-genomic map of human health in Fenland study. *The image is obtained from Pietzner et al (2021)*[111].

We rely on above described broad-capture proteomic approach to explore the existence of potential causal genes and genetic signals that alter protein abundance, and which also influence the timing of menopause in women. More specifically, this study aimed to understand whether ovarian ageing impacts the proteomic profile in women, and get a potential mechanistic insight into the type of pathways that are being perturbed due to menopause onset, which ultimately lead to the change in protein levels. Our study is the first to explore the association between ovarian ageing and human proteome - this and future proteomic studies will hopefully pave the path towards detection of novel protein biomarkers of reproductive ageing and enable early prediction of ANM.

5.2 Methods

5.2.1 ReproGen consortium data on age at natural menopause

The ReproGen consortium summary statistics on common genetic variants associated with ANM[82] were used to examine the association between ANM and protein levels. The phenotype preparation and genome-wide association meta-analysis are described in **Chapter 2**, **Section 2.2.2**.
Of 290 independent ANM signals identified in Ruth *et al* (2021)[82], 276 genetic variants (or proxies) were identified in the Fenland genomic-proteomics summary statistics described below. In cases where a particular target signal was not present in the outcome GWAS, we searched the UKBB white European dataset for proxies (within 1 Mb and $r^2 > 0.5$) and chose the variant with the highest r^2 value. Genotypes at all variants were aligned to designate the ANM-increasing alleles as the effect alleles.

5.2.2 Genetic and proteomic data from the Fenland study

5.2.2.1 Study design and recruitment of participants

The Fenland study represents a population-based cohort consisting of 12,435 participants, predominantly of White British ancestry, born between 1950 and 1975[333]. The participants were recruited from general practice surgeries in the Cambridgeshire region in the UK, after which they underwent detailed phenotyping at a baseline visit between 2005 and 2015 at one of three MRC Epidemiology Unit testing centres. Individuals with clinically diagnosed diabetes mellitus, inability to walk unaided, terminal illness, clinically diagnosed psychotic disorder, pregnancy, or lactation were excluded from the study. The study was approved by the Cambridge Local Research Ethics Committee (NRES Committee - East of England Cambridge Central, ref. 04/Q0108/19) and all participants provided written informed consent. Participants in the study were on average 48.6 years old (SD: 7.5 years) and 53.4% of them were female. While at

testing centres, they completed questionnaires on lifestyle and general health, and had clinical, anthropometric and physical health measurements taken. Finally, blood and urine samples were collected from each participant for further metabolic assessment and genotyping.

5.2.2.2 Proteomic measurements

Proteomic profiling of fasting EDTA plasma samples from 12,084 Fenland Study participants, collected at the baseline visit, was performed by SomaLogic Inc. (Boulder, US) using an aptamer-based technology (SomaScan v4 assay). A detailed description of this process can be found elsewhere[111]. Briefly, the SomaScan assay utilises a library of short single-stranded DNA molecules, which are chemically modified to specifically bind to protein targets. DNA microarrays are used to determine the relative amount of aptamer binding to protein targets. All quality control steps undertaken during this process are described in Pietzner *et al* (2021)[111]. Samples were removed if they were deemed by SomaLogic to have failed or did not meet the acceptance criteria of 0.25-4 for all scaling factors. In addition to passing SomaLogic QC, only human protein targets were taken forward for subsequent analysis (4,979 out of 5,284 aptamers). Aptamers' target annotation and mapping to UniProt accession numbers together with the Entrez gene identifiers were provided by SomaLogic.
Rank-based inverse normal transformations were applied to aptamer abundances. Additional details are available in Pietzner *et al* (2021)[111].

5.2.2.3 Genotyping and imputation

Fenland participants were genotyped using one of three genotyping arrays: the Affymetrix UKBB Axiom array (OMICS, N=8994), Illumina Infinium Core Exome 24v1 (Core-Exome, N=1060) and Affymetrix SNP5.0 (GWAS, N=1402). More details on this process can be found in Pietzner *et al* (2021)[111]. Autosomes for the OMICS and GWAS subsets were imputed to the HRC (r1) panel using IMPUTE4, and the Core-Exome subset and the X-chromosome (for all subsets) were imputed to HRC.r1.1 using the Sanger imputation server. Sanger imputation server was also used to impute all three array subsets to the UK10K+1000G phase3 panel to obtain additional variants that do not exist in the HRC reference panel.

5.2.2.4 Sex-combined GWAS and meta-analysis

A total of 10,713 Fenland participants had both phenotypes and genetic data for the analysis (OMICS=8,355, Core-Exome=1,026, GWAS=1,332), after excluding ancestry outliers and related individuals. Within each genotyping subset, aptamer abundances were transformed to follow a normal

distribution using the rank-based inverse normal transformation. Transformed aptamer abundances were then adjusted for age, sex, sample collection site and 10 principal components in STATA v14 and the residuals were used as input for the genetic association analyses. GWAS was performed under an additive model using BGENIE (v1.3)[334]. Results for the three genotyping arrays were combined in a fixed-effects meta-analysis in METAL. Following the meta-analysis, 17,652,797 genetic variants present in the largest subset of the Fenland data (Fenland-OMICS) were taken forward for further analysis. For each protein target, a multiple test corrected genome-wide significance threshold of $1.004*10^{-11}$ was used. For more details on signal selection refer to Pietzner *et al* (2021)[111]. The pQTLs were classified as cis-acting instruments if the variant was within 500kb of the body of the protein encoding gene.
For each identified pQTL we first obtained all SNPs in at least moderate LD (R^2>0.1) using PLINK (version 2.0), and queried comprehensive annotations using the VEP software (version 98.3)[162], applying the pick option as described in Pietzner *et al* (2021)[111].

5.2.2.5 Sex-stratified GWAS

The statistical normalization procedure for aptamer abundances was updated by SomaLogic after the sex-combined GWAS was run. This had little effect on GWAS results, however, a small number of individuals flagged as QC exclusions by the new normalisation were not flagged by the previous normalisation, and vice versa. The sex-stratified GWAS therefore includes the 8,348 individuals in the Fenland-OMICS subset that were included by both the original and updated normalization procedures.

Measurements for 4,979 aptamers targeting 4,775 human protein targets were inverse rank normalized and transformed abundances were adjusted for age, sample collection site and 10 PCs in R v3.6 and the residuals were subsetted for each sex to be used as input for the sex-stratified genetic association analyses. A sex-stratified GWAS was conducted using fastGWA[335] software through GCTA version 1.93.2. A sparse GRM was created through GCTA version 1.93.2 using the default settings with the binary files for autosomes for the Fenland-OMICS subset. Linear regression analysis was performed through fastGWA for the abundance of each aptamer separately in each sex (N female=4,403, N male=3,945). fastGWA also applied further QC where variants with genotype missingness < 0.05 and MAF < 0.0001 were also filtered out in each sex. Results for the sex-stratified GWAS were meta-analysed in a fixed-effects meta-analysis in METAL to assess the heterogeneity between sexes.

5.2.2.6 Meta-analysis data

In addition to primary analysis conducted in the Fenland study, we replicated the results meta-analysing Fenland data with proteomic data available in the deCODE study (N individuals=46,075). The detailed description of deCODE human proteomic data is available in Ferkingstad *et al* (2021)[336]. The total sample sizes included in the analyses were: deCODE (N=35,362), Fenland-OMICS (N=8,355), Fenland-GWAS (N=1,332) and Fenland-Core Exome (N=1,026). The meta-analysis was run using METAL with a random effect model. The replication data were used for the Mendelian Randomisation (MR) analysis of the effect of ANM on *RACGAP1* abundance (**Appendix Table 5.6**). A total of 270 ANM variants (or their proxies) were identified in the meta-analysis dataset. For more details on the MR frameworks, refer to **Section 5.2.5**.

5.2.3 Genetic risk score for the age at natural menopause

Effect estimates from ReproGen ANM GWAS meta-analysis were used to construct a genetic risk score (GRS) for ANM using 276 ANM signals. The GRS was calculated using PLINK v1.90b4.4 in the OMICS subset of Fenland data. Genotypes at all variants were aligned to designate the ANM-increasing alleles as the effect alleles.

5.2.4 Statistical analyses

Linear regression analyses, using the *lm* function in R, were used to model the effect of menopause on measured abundance of the 4,979 protein targets in the 8,355 individuals (4,406 women, 3,949 men) in the Fenland-OMICS subset passing the QC. In addition to protein data and covariate information on the test site, 10PCs, age and sex of participants, we obtained the data on the menopause status, which was determined based on self-reported cessation of menstrual periods and was recorded for 3,792 women included in this analysis, of whom 1,992 were pre-menopause and 1,800 were post-menopause. Menopause status acts here as a non-genetic factor - it represents the binary trait (yes/no for menopause status) that does not only incorporate the effect of ovarian ageing on the onset of menopause, but it could be seen as unifying factor of all symptoms and outcomes that are associated with this transition, including the effect of overall ageing, hormonal status etc.

Five different models were tested:

(1) using the GRS for menopause adjusted for age, sex, test site, and 10 principal components of genetic variation: **protein ~ GRS + age + sex + testSite + 10PCs**;

(2) a sex-stratified version of (1);

(3) additionally adjusting of (1) for **menopause status**, in women only: **protein ~ GRS + age + sex + testSite + 10PCs + menoStatus**;

(4) the effect of menopause status in women on protein levels, adjusted for age and test site: **protein ~ menoStatus + age + testSite**

(5) In order to examine the effect of ageing on protein levels, independent of menopause, we modelled the association between age and protein levels, adjusting only for sex and test site: **protein ~ age + sex + testSite**

We used these models to narrow down the selection of proteins that we will study in detail in next stages. To identify significant associations in each model we applied a Bonferroni-corrected *P*-value threshold ($P < 0.05/4979 = 1.004*10^{-5}$) (**Appendix Table 5.1**).

We finally performed additional linear regression model by taking into account menopause-status stratification, i.e. assessing the effects on pre- and post-menopausal women independently. This analysis was run on linear regression models (2) and (5), which were described above (**Appendix Table 5.2**).

5.2.5 Mendelian Randomisation

Bidirectional MR analysis was conducted to examine the likelihood of a causal effect of ANM on the proteomic profile, as well as the causal effect of the protein levels on ANM. The MR analysis was treated as the secondary analysis. Various MR approaches that were applied here are described in detail in **Chapter 2**, **Section 2.3**. There is a potential for bias in MR analysis where IVs and outcomes are drawn from the same sample. We conducted a two-sample MR analysis using data from two independent studies, ANM data from the ReproGen study[82] and proteomics data available in the Fenland study[111], which enabled us to avoid participant overlap.
The MR analysis was conducted in two stages. **Stage 1** was conducted in the largest sample size available, i.e. *sex-combined* data, with the purpose of narrowing down the number of protein candidates

of interest, i.e. protein prioritization. **Stage 2** was performed to further decipher the effect of ANM in the *sex-stratified* data and to control for the effect of *cis*-acting loci.

5.2.5.1 Protein prioritisation

In **Stage 1** we assessed the causal inference of ANM on the levels of all 4,979 protein targets in the *sex-combined* framework, and identified the ones positively associated for further analysis. We used 276 ANM genome-wide significant signals as exposure IVs. Sex-combined data on 4,979 protein targets from all three genotyping subsets of the Fenland study (OMICS=8,355, core-exome=1,026, GWAS=1,332) was used as the outcome. Genotypes at all variants were aligned to designate the ANM-increasing alleles as the effect alleles. We applied the Bonferroni (*P*: $0.05/4979=1.004*10^{-5}$) correction to identify significant protein associations from both GRS and MR analyses described in **Section 5.2.4** and **Section 5.2.5**. In total, we identified significant associations between the ANM and the levels of 196 protein targets, which were considered for further analysis (**Section 5.2.5.2**).

5.2.5.2 Additional MR analysis for the detection of robust protein candidates for ANM

Stage 2 of MR was conducted in order to assess the causal inference of ANM on 196 selected proteins in the *sex-stratified* framework. We used 276 ANM genome-wide significant signals as exposure IVs, and sex-stratified data on 196 protein targets as the outcome data, coming from 4406 female and 3949 male Fenland participants from the OMICS subset. All primary and sensitivity analysis described in **Chapter 2**, **Section 2.3** were conducted. In order to examine the degree of effect of the cis loci on the observed association, we additionally applied a 10mb window filter to exclude the cis loci, and repeated all MR analysis following the described methodology.

For the analysis of the identified protein candidate, *RACGAP1* (check the Results section), we also performed additional MR for ANM to *RACGAP1* levels but this time excluding SNPs that fall within the HLA region (N=6). This exclusion left us with 270 independent ANM signals as instrumental variables (**Appendix Table 5.5**). Described methodology was conducted for both the primary and replication (meta) analysis (**Appendix Table 5.6**).

Stage 2 also included the bidirectional MR approach, where we additionally assessed the causal inference of 196 pQTLs on the menopause timing (**Appendix Table 5.1**). We used the pQTL signals identified in the Fenland GWAS (**Section 5.2.2.4**) for each of the 196 proteins of interest as genetic instruments. The analysis was conducted in both *sex-stratified* and *sex-combined* framework, applying all primary and

sensitivity MR models. The sex-stratified data were available for 3945 male and 4403 female participants in the OMICS subset of Fenland study, while the sex-combined model included 10,713 participants from all three chips (OMICS=8350, core-exome=1026, GWAS=1332). The reverse MR analysis, protein abundance to ANM, was conducted using two different IVs: the first set of analysis was performed using cis pQTL variants as the genetic instrument, while the second part of the analysis included only primary signals of both cis and trans origin. In cases where only one SNP was available as pQTL, the Wald ratio was calculated instead of default MR primary and sensitivity analysis. In addition to the MR analysis, we performed the lookup of all pQTLs in the ANM dataset, where alleles were harmonised in the protein level increasing direction (**Appendix Table 5.3**).

5.2.6 Categorisation

In order to understand the nature of the effect of ANM genetic variants, we categorized the results into 3 different groups according to: (a) the origin of the overall observed significance in the MR result, i.e. whether the effect is driven by the cis or trans-acting loci, (b) the significance of the effect of age and (c) menopause status. Those 3 categories were:

1) Strong cis effect → the association is driven primarily by the inclusion of cis-acting loci. For the proteins that fall into this category the MR result that includes cis-acting loci is significant while it becomes non-significant after the exclusion of those loci. The cis-acting loci are defined in the **Section 5.2.5.2**.
2) Many trans effects, *P* GRS in females is significant
    a) *P* GRS in females remains significant after controlling for MenoStatus and association with age is significant → the association is driven by both genetic (GRS) and non-genetic (menopause status) factors, yet most likely mediated by ageing process that associates with menopause onset in women
    b) *P* GRS in females is significant but becomes non-significant after controlling for MenoStatus and association with age → the genetic association (GRS) could be entirely explained by the menopause status (non-genetic) and association with age
3) *P* GRS in females is non-significant
    a) only menopause status and association with age are significant
    b) menopause status is significant, while association with age is non-significant

5.3 Results

5.3.1 Genetic associations for protein targets and menopause timing

To study the impact of ovarian ageing on the proteomic profile in women, we used a genome-proteome-wide association data on 4,775 protein targets (4,979 human somamers, SomaLogic) available in the Fenland study on 10,713 participants[111] and data on ANM for ~200,000 women from the ReproGen study[82]. We performed the analysis in two stages. **Stage 1** was conducted with the purpose of narrowing down the number of potential protein candidates that associate with the ANM by considering the significant result in any of the described analyses, MR or GRS linear regression. **Stage 2** was performed to further decipher the effect of ANM by studying in more details the factors driving the observed associations, such as effect of cis-acting loci and menopause status, and taking into account the consistency of the results across different primary and secondary analysis with a final aim to select robust ANM-associated protein candidates of interest (**Figure 5.2**).

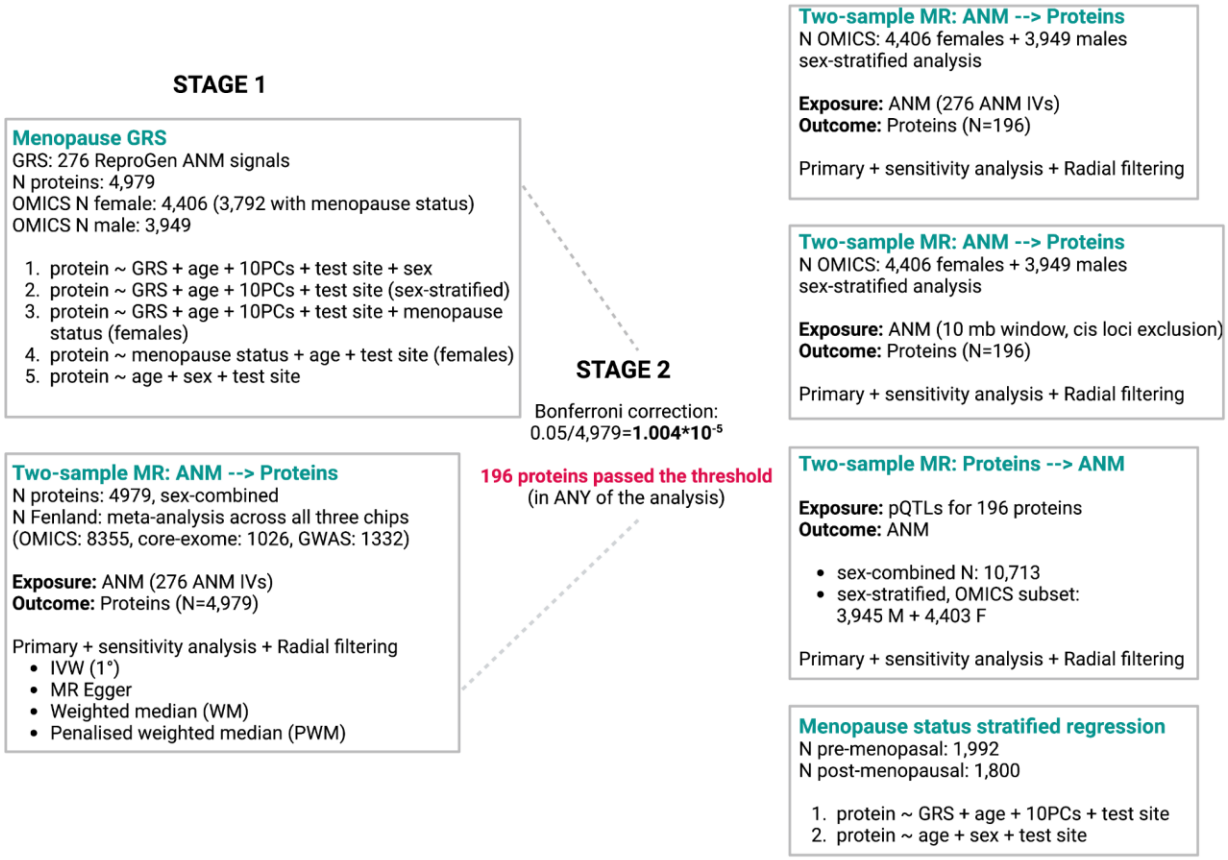

Figure 5.2: A flow chart summarising the study strategy.

To perform **Stage 1**, we combined all evidence obtained from linear regression analysis in both sex-combined and stratified data with the results from the MR analysis in the largest sample size available, i.e. the *sex-combined* data. The linear regression analysis assessed the effect of ANM GRS on 4,979 protein targets, also taking into account the effect of age and menopause status. Two-sample MR analysis used 276 ANM ReproGen signals (or their proxies) as an exposure and sex-combined data on 4,979 proteins in meta-analysis of all three Fenland chips as outcomes (N individuals - OMICS: 8,355, core-exome: 1,026 and GWAS: 1,332) (Methods, **Section 5.2.2.4**). To prioritise potential protein candidates that associate with ANM, we applied a multiple test correction (*P*: $0.05/4{,}979 = 1.004*10^{-5}$) on both types of analyses, and identified 196 out of 4,979 protein associations that passed this significance threshold in any analysis, either linear regression or MR (**Appendix Table 5.1**). These 196 protein candidates were taken forward for further examination.

5.3.2 Proteogenomic analysis did not identify robust protein markers of ovarian ageing

Stage 2 specifically focused on the 196 selected protein candidates from Stage 1.
Firstly, to boost the evidence obtained from the secondary MR analysis and get a better insight into the nature of association between the ANM and protein abundances, we ran additional bi-directional MR models on sex-stratified data. This included two models with ANM signals as instrumental variables (IVs) and 196 proteins as outcomes ('Forward MR': from ANM to proteins) (**Figure 5.2**). Unlike the first model that used all 276 ANM signals on sex-stratified data, the second model excluded the ANM variants that were located within the 10mb window around the protein of interest with an aim to determine whether the observed associations were driven by the cis-acting signals. The protein associations tested using the MR frameworks that became non-significant after the exclusion of the cis-acting signals were defined as non-robust candidates. Finally, using the identified protein pQTLs, as described in **Section 5.2.2.4** and Koprulu, M. *et al* (2022)[337], we tested the effect of these 196 protein levels on ANM as the outcome ('Reverse MR': from proteins to ANM) via MR or pQTL look-ups in cases where only a single pQTL per protein was identified (**Appendix Table 5.1** and **Appendix Table 5.3**).

We also ran additional linear regression models, stratifying the analysis based on menopause status, i.e. in pre- (N=1,992) and post-menopausal women (N=1,800) separately (**Appendix Table 5.2**). Menopause status serves as a non-genetic factor, which not only incorporates the effect of ovarian ageing on the onset of menopause, but it could be seen as unifying factor of all symptoms and outcomes that are associated

with this transition, including an effect of overall ageing, changes in the hormonal status etc. Therefore, the significance in the result within the pre-menopausal group would suggest that the ANM-protein association is driven by genetic factors underlying the process of ovarian ageing rather than menopause status itself.

In order to identify protein candidates as robust biomarkers of ovarian ageing, we looked for statistical consistency across the above described evidence and further explored their biological rationale. We observed 3 patterns and thus classified all candidate proteins (except for *RACGAP1*, see below) into the following categories:

1) ANM variants with a strong cis effect on protein abundance (N = 3 proteins): The effect of ANM on the protein is driven by ANM cis-acting variants, i.e. in or near to the protein-encoding gene. This category was identified based on the MR result that became non-significant after the exclusion of the cis-acting loci. Notably, all 3 proteins in this category were located within a highly pleiotropic HLA region and thus these findings are difficult to interpret.

2) Multiple ANM variants with trans effects on protein abundance (N = 101 proteins). Here, we identified 2 subcategories with associations either persisting or attenuating after controlling for menopause status. Despite these differences, all proteins in this category were strongly associated with age (**Appendix Table 5.1**). This indicates that the ANM GRS association may be confounded by age and other related factors. Some of these proteins are potential markers of menopause status - however, their levels seem not to be solely altered by the change in ovarian function but also other ageing processes, which drive the ANM GRS association.

3) Associated with menopause status but not with ANM GRS (either before or after controlling for menopause status) (N = 91 proteins). These proteins are likely influenced by mechanisms linked to menopause status. They showed reported biological functional links to health outcomes associated with ANM, including bone health, e.g. *MRC2*[338] and *BMP1*, and brain function, e.g. *NTN4*[339] and *LSAMP*[340] involved in neurite growth, migration and axon targeting.

The proteins classified into these 3 categories could not be considered as robust candidates for specific markers of ovarian ageing (prior to menopause).

5.3.3 The effect of ANM on *RACGAP1* abundance

Only a single protein, *RACGAP1*, was outside the above 3 categories. We observed an inverse association between the ANM GRS and *RACGAP1* abundance (P=1.38*10^{-5}), which strengthened (P=2.97*10^{-6}) after controlling for menopause status and age (**Table 5.1**). In addition, *RACGAP1* was the only associated protein in both pre- and post-menopausal women, separatately. More specifically, *RACGAP1* was associated with ANM GRS in pre-menopausal women (P=4.83*10^{-5}) after the multiple test correction (P: 0.05/196 = 2.6*10^{-4}) and showed only nominally significant association with post-menopause status (P=1.11*10^{-2}) (**Appendix Table 5.2**). This suggests a strong genetic involvement of the ovarian ageing impact on *RACGAP1* abundance (**Appendix Table 5.2**). Notably, the effect of ANM on *RACGAP1* levels seemed not to be driven by age per se, as there was little or no association with age in pre-menopausal (P=4.16*10^{-2}) or post-menopausal women (P=1.78*10^{-1}) (**Appendix Table 5.2**). Interestingly, we also observed a significant association between ANM GRS and *RACGAP1* abundance in men (P=4*10^{-5}), which may indicated a wider relevance of this protein for health and disease and further supported the fact that the effect we observed in women was not menopause status driven (**Table 5.1**).

Table 5.1: Linear regression analysis for effect of ANM GRS on RACGAP1 abundance, with covariates. *C: sex-combined, F: females, M: males.*

Model	Beta GRS	SE GRS	P GRS	P Age	P MenoStatus
Protein ~ GRS C + age + 10PCs + Test site	-0.045	0.007	$1.77*10^{-9}$	$2.80*10^{-2}$	-
Protein ~ GRS F + age + 10PCs + Test site	-0.044	0.010	$1.38*10^{-5}$	$9.62*10^{-1}$	-
Protein ~ GRS M + age + 10PCs + Test site	-0.045	0.011	$4*10^{-5}$	$2.30*10^{-3}$	-
Protein ~ GRS F + age + 10PCs + Test site + menopause status	-0.052	0.011	$2.97*10^{-6}$	$5.38*10^{-3}$	$7.01*10^{-4}$
Protein ~ age + Test site + menoStatus	Beta PostMeno: -0.126	SE PostMeno: 0.045	-	$1.32*10^{-2}$	$5.22*10^{-3}$
Protein ~ age + sex + Test site	Beta Age: 0.003	SE Age: 0.001	-	$2.64*10^{-2}$	-
Pre-ANM GRS F Protein ~ GRS + age + 10PCs + Test site	-0.063	0.015	$4.83*10^{-5}$	$2.16*10^{-2}$	-
Post-ANM GRS F Protein ~ GRS + age + 10PCs + Test site	-0.041	0.016	$1.11*10^{-2}$	$1.37*10^{-1}$	-
Pre-ANM Age Protein ~ age + Test site	Beta Age: 0.008	SE Age: 0.004	-	$4.16*10^{-2}$	-
Post-ANM Age Protein~age + Test site	Beta Age: 0.006	SE Age: 0.005	-	$1.78*10^{-1}$	-

Finally, the linear regression results were supported by the MR analysis which suggested a causal effect of later ANM on lower *RACGAP1* abundance (Post Radial IVW: $P=2.22*10^{-2}$). Sensitivity models were inconsistent (MR Egger: $P=1.93*10^{-2}$; MR WM: $P=1.45*10^{-1}$, MR PWM: $P=1.44*10^{-1}$) (**Table 5.2**, **Figure 5.3 A, B**), however the findings were directionally consistent and significant after the exclusion of the cis-acting signals (**Table 5.2**, **Figure 5.3 C, D**), and in both pre- and post-Radial models that aimed to detect and remove the outliers.

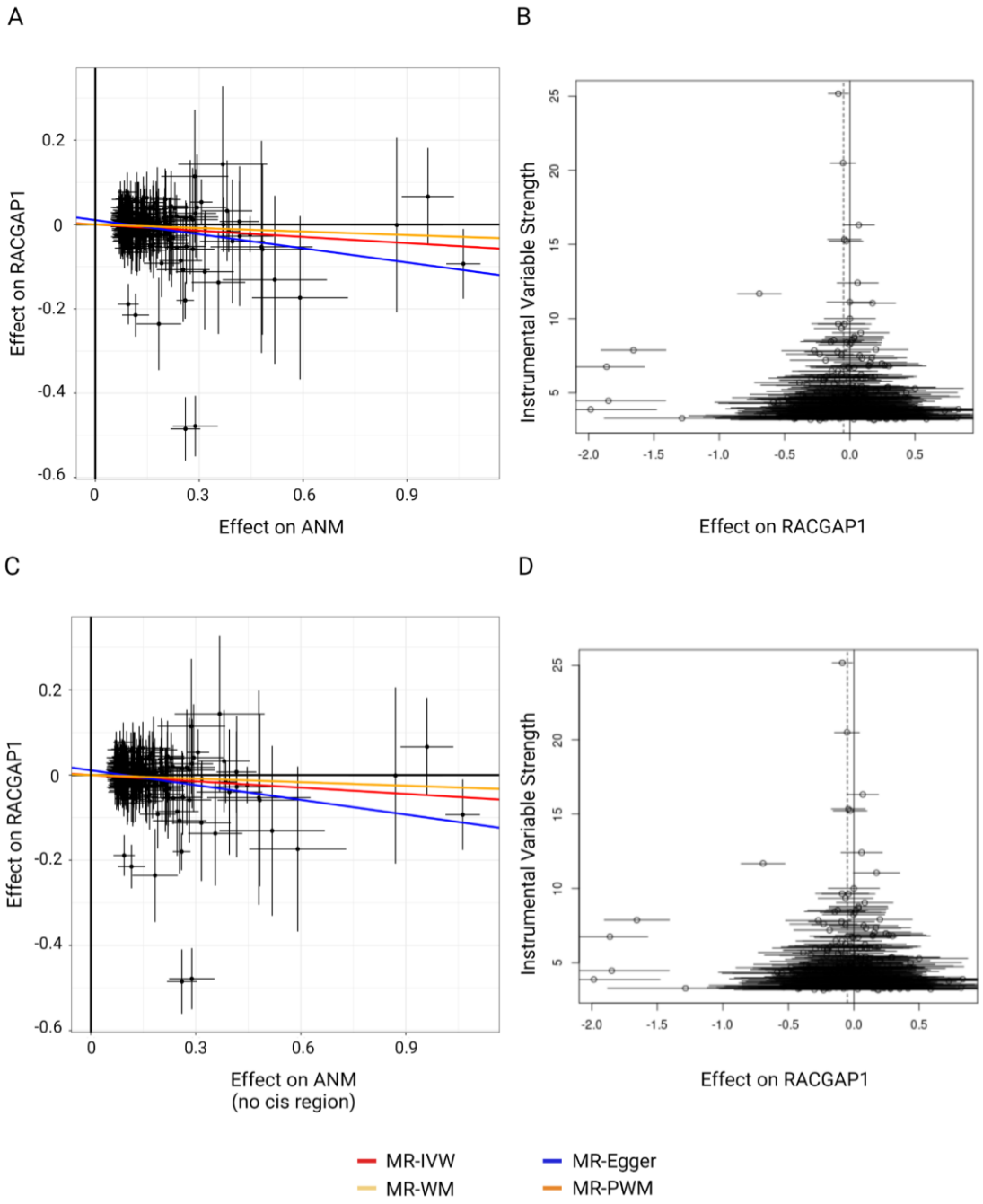

Figure 5.3: Mendelian Randomisation analysis of the effect of ANM on RACGAP1 abundance. *Figures* ***(A)*** *and* ***(C)*** *present the dosage plots when using all ANM instrumental variables (N IVs: 276) (A) and after excluding the cis-acting signals. Figures (C) and (D) show the funnel plots as the measure of heterogeneity in described analysis. IVW: inverse variance weighted, WM: weighted median, PWM: penalised weighted median.*

Table 5.2: Mendelian randomisation models of the effect of ANM on RACGAP1 abundance. *'No HLA' indicates models that excluded SNPs in the HLA region (N=6). 'Lead HLA' indicates that all SNPs in the HLA region were excluded except the leading, i.e. top scoring one (N=5). This was done to test whether this single most significant HLA variant drives the overall significant MR result. For more detailed MR results refer to* ***Appendix Table 5.1****.*

MR model	N SNPs	Beta IVW	SE IVW	P IVW	P Egger	P WM	P PWM
ANM > P Pre Radial	276	-0.049	0.018	$7.31*10^{-3}$	$1.91*10^{-3}$	$1.42*10^{-1}$	$1.27*10^{-1}$
ANM > P Post Radial	250	-0.026	0.011	$2.22*10^{-2}$	$1.93*10^{-2}$	$1.45*10^{-1}$	$1.44*10^{-1}$
ANM > P CIS exluded Pre Radial	273	-0.049	0.018	$7.49*10^{-3}$	$1.56*10^{-3}$	$1.41*10^{-1}$	$1.13*10^{-1}$
ANM > P CIS exluded Post Radial	248	-0.027	0.011	$1.89*10^{-2}$	$1.99*10^{-2}$	$1.29*10^{-1}$	$1.34*10^{-1}$
ANM > P no HLA Pre Radial	270	-0.009	0.011	**$4.22*10^{-1}$**	$1.78*10^{-2}$	$1.69*10^{-1}$	$1.74*10^{-1}$
ANM > P no HLA Post Radial	250	-0.007	0.012	**$5.38*10^{-1}$**	$4.10*10^{-1}$	$4.68*10^{-1}$	$4.98*10^{-1}$
ANM > P lead HLA Pre Radial	271	-0.021	0.014	$1.32*10^{-1}$	$4.23*10^{-3}$	$9.54*10^{-2}$	$1.01*10^{-1}$
ANM > P lead HLA Post Radial	249	-0.023	0.011	$3.64*10^{-2}$	$2.71*10^{-2}$	$9.30*10^{-2}$	$9.43*10^{-2}$

We explored the biological function of the 6 individual *RACGAP1*-associated ANM signals that based on dosage and funnel plots seemed to significantly contribute to the effect we observe (**Appendix Table 5.4**). All of these signals were located in the HLA region, a well-described region for high pleiotropy and thus challenging for robust biological interpretations. Therefore, to test whether the association between ANM and *RACGAP1* is driven by signals in the HLA region, we performed two additional MR analysis: fully excluding HLA SNPs (N=6) (**Figure 5.4 A, B**), and excluding all but the top scoring HLA SNP (N=5) (**Figure 5.4 C,D**). When excluding all HLA SNPs, the ANM-*RACGAP1* association became non-

significant (*P* IVW post-Radial=5.38*10^{-1}), and when keeping the single top scoring HLA SNP, a nominal association remained between ANM and RACGAP1 abundance (*P* IVW post-Radial=3.64*10^{-2}) (**Table 5.2**). These results indicate that the effect of later ANM on lower *RACGAP1* is driven by the highly pleiotropic HLA region, and thus might not be a biologically robust observation.

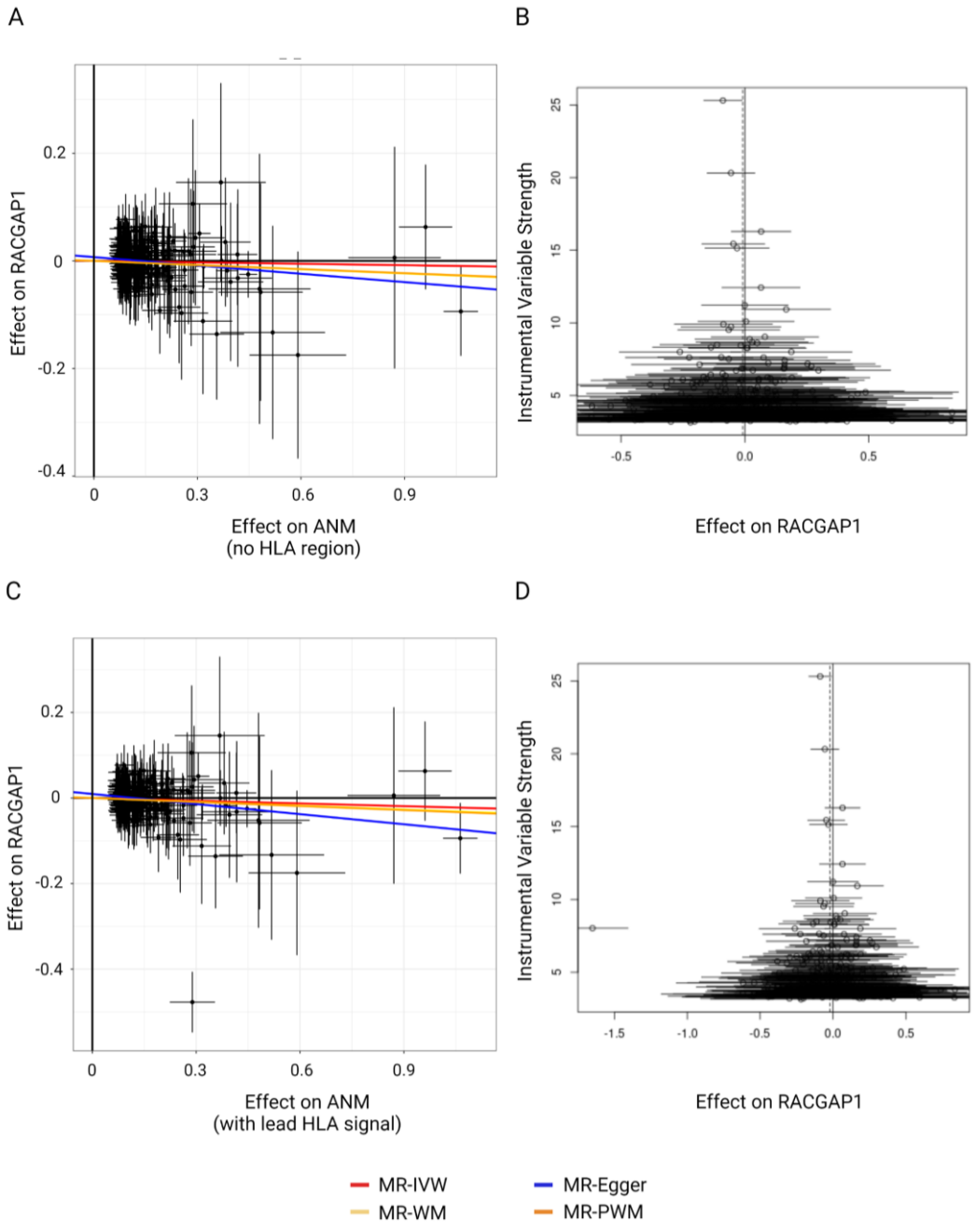

Figure 5.4: Mendelian Randomisation on the effect of ANM on RACGAP1 abundance after excluding the HLA region. *Figures present the dosage plots when excluding all SNPs in the HLA region (A) and when excluding all but keeping the top scoring HLA SNP (C). Figures (C) and (D) show the funnel plots as the measure of heterogeneity in described analysis. IVW: inverse variance weighted, WM: weighted median, PWM: penalised weighted median.*

In order to better estimate the effect of ANM variants on *RACGAP1* abundance, we conducted a meta-analysis of a substantially larger sample, the Fenland and deCODE studies (total N=46,075). No association was seen with or without inclusion of SNPs in the HLA region (**Figure 5.5, Table 5.3**). These results indicate that larger sample sizes might enable future more robust analyses of protein abundance and ovarian function.

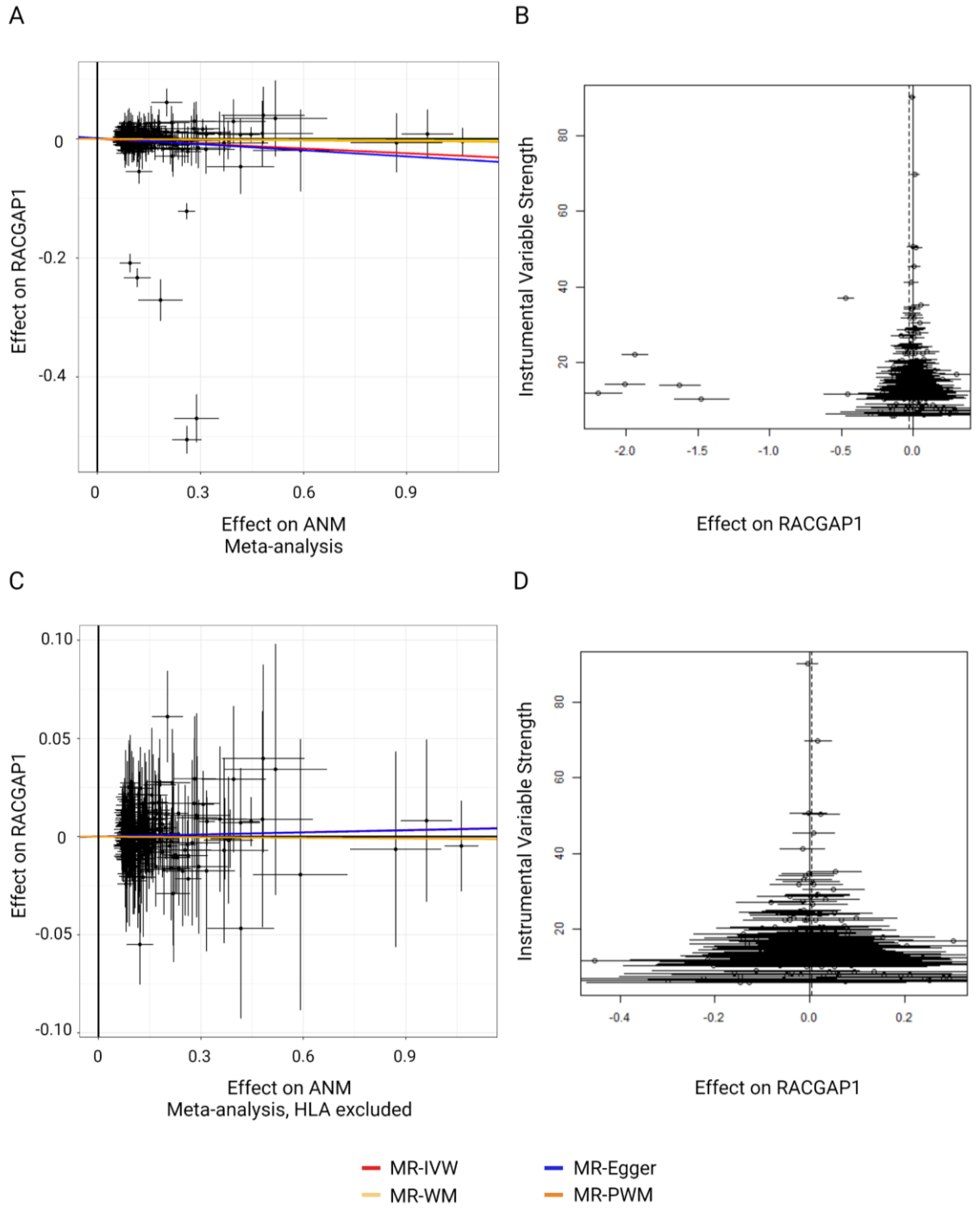

Figure 5.5: Mendelian Randomisation on the effect of ANM on RACGAP1 abundance in the meta-analysis. *Figures present the dosage plots when including all ANM IVs (A) and when excluding all ANM IVs in the HLA region (C). Figures (C) and (D) show the funnel plots as the measure of heterogeneity in described analysis. IVW: inverse variance weighted, WM: weighted median, PWM: penalised weighted median.*

Table 5.3: Summary of the Mendelian Randomisation on the effect of ANM on RACGAP1 abundance in the meta-analysis. *The results with the label 'HLA excluded' represent the MR analysis where the SNPs within the HLA region (N=6) were removed. For more detailed MR results refer to* ***Appendix Table 5.6****.*

MR model	N SNPs	Beta IVW	SE IVW	P IVW	P Egger	P WM	P PWM
ANM > P Pre Radial	270	-0.027	0.014	$5.38*10^{-2}$	$1.98*10^{-1}$	$6.66*10^{-1}$	$5.26*10^{-1}$
ANM > P Post Radial	233	-0.008	0.004	$5.39*10^{-2}$	$3.44*10^{-1}$	$6.30*10^{-1}$	$6.34*10^{-1}$
ANM > P HLA excluded Pre Radial	264	0.003	0.004	$3.45*10^{-1}$	$6.41*10^{-1}$	$8.58*10^{-1}$	$8.58*10^{-1}$
ANM > P HLA excluded Post Radial	242	0.003	0.003	$4.27*10^{-1}$	$3.15*10^{-1}$	$8.64*10^{-1}$	$8.63*10^{-1}$

5.4 Discussion

High-throughput proteomic profiling has the potential to accelerate our understanding of human biology and disease, including reproductive ageing. This Chapter represents the first proteogenomic study performed on ANM, aimed to examine the potential impact of ovarian ageing on the proteomic profile in women. Using linear regression analysis on the individual level data and two-sample MR on the summary statistic data, we examined the association between ANM and 4,979 protein targets available in the Fenland cohort of ~10,700 European descent individuals (53.3% women). We firstly identified 196 unique protein targets that were significantly associated with menopause timing in either of the analysis, and further explored those in additional regression and MR models that tried to decipher potential mechanisms driving the observed associations. For example, this included modelling the linear regression with age and menopause status to understand whether the effect we observe is solely due to changes in ovarian function or it is mediated by other factors that associate with menopause timing. Our analysis identified a potential protein candidate, *RACGAP1*, which demonstrated a strong genetic association with ANM. More specifically, we showed that later ANM was associated with lower levels of *RACGAP1* in the analysis of both individual- and summary-level data. Biologically, RACGAP1 does indeed represent an attractive protein target due to its involvement in the cell cycle cytokinesis and cell growth, mechanisms that were already recognised as critical for menopause timing[82,341]. In addition, *RACGAP1* is

involved in spermatogenesis and regulation of sulfate transport in male germ cells, and deletion of *RACGAP1* in the germ cells causes male sterility in the mouse model, thus indicating its importance for the reproductive function in both males too[342].

However, through a thorough examination of the ANM genetic variants we concluded that the significant association we observed with *RACGAP1* is highly driven by genetic variants located within the HLA region. These variants are highly pleiotropic and are associated with numerous complex human diseases, thus their biological interpretation is not yet straightforward[343]. Even though this analysis did not bring fruitful results, it shows the potential of human proteomic data for the identification of new biomarkers of ovarian ageing and gives a solid basis for the future studies that should further explore the association between human proteome and reproductive longevity. The success of future studies will highly depend on the sample size - our meta-analysis on *RACGAP1* clearly demonstrates how an increase in sample size enables more robust conclusions. To address this, my future analysis will use the proteomic data on ~55,000 individuals in UKBB that will become available to our group in late 2022[344]. Besides the limited sample size, this study has additional limitations, including its predominant European ancestral composition. This limits us to capture the full genetic and phenotypic diversity, therefore future studies should try to be more inclusive and integrate the proteogenomic data from under-represented populations[345]. In addition, future attempts should also incorporate the exome sequencing information, which will enable a multi-dimensional perspective on reproductive health and ageing. Finally, next steps in my proteogenomic research will also try to address whether potential protein candidates that are affected by ovarian ageing have any impact on the susceptibility to later life diseases that associate with menopause timing, as well as investigate whether other reproductive health outcomes, such as menarche timing, have an effect on the same protein targets as ANM.

.

CHAPTER 6

DNA damage repair and insights into shared aetiology between menarche and menopause

Summary

While the heritable determinants of many common reproductive traits have been studied, little effort has focused on understanding the degree of shared genetic architecture underlying different reproductive traits. Here, I describe an enlarged GWAS meta-analysis of age at menarche in ~566,000 women in the ReproGen consortium, which I used to study the shared genetic architecture and biological mechanisms between menopause and menarche timing, as well other reproductive health outcomes. This work provides the first evidence on the involvement of DDR in regulating the beginning of the reproductive lifespan, i.e. menarche timing, and the first gene candidates that we believe act via oocyte-specific mechanism to modify age at menarche. This indicates the relevance of maintaining genomic stability not only for the establishment and maintenance of the ovarian reserve, but also for the initiation of reproductive biological activity. This finding led us to think about DDR as a broader marker of health and disease - we provide the first human genomic evidence on the involvement of DDR genes across multiple health outcomes, demonstrating their impacts across anthropometric, metabolic and reproductive traits.

Contributions and Collaborations

Dr Felix Day performed the age at menarche GWAS meta-analysis and replication. Dr Katherine Kentistou created a G2G pipeline for gene prioritisation. The ‘G2G’ pipeline was run by Dr Kentistou, while some parts of the pipeline, specifically colocalization, MAGMA and gene expression analysis were also run by me. I performed the functional annotation of identified genes via thorough literature review and gene-set and pathway analysis. ‘Expert curated DDR 1’ gene list was curated by Professor Steve Jackson’s group, while ‘Expert curated DDR 2’ was curated through collaboration of internal (Professor John Perry) and external experts (Professor Eva Hoffmann and Professor Anna Murray). I performed GTEx tissue expression analysis, as well as gene-level and gene-set MAGMA analysis. The methodology behind ‘cMAGMA’ was created by Professor John Perry. I performed all the lookup and colocalization analysis between the genetic variants across multiple health outcomes, as well as gene-set enrichment analysis using gProfiler and functional enrichment tests using fGWAS and SLDP. Professor John Perry and Professor Ken Ong provided valuable advice on the analyses and writing of the manuscript.

6.1 Insights into shared aetiology between reproductive traits

Female reproductive health represents an important aspect of overall wellbeing, with increasing evidence highlighting its implications for the risk of later life health outcomes and ageing[130,346–348]. For many of the reproductive traits which have been studied in-depth the aetiology appears to be complex and varies considerably in the population [347,349]. While much progress has been made towards elucidating the genetic factors that contribute to the variation in these traits individually, not many efforts have been focused on understanding the degree of shared genetic architecture. Investigating the links and common biological mechanisms between various reproductive traits can have important benefits[350–352]. Firstly, it gives us an insight into whether and how genetic factors influence reproductive health as a whole. Secondly, it is particularly important for interventions and drug discovery to develop treatments that more specifically target the outcome of interest while reducing the possibility of any unwanted secondary effects.

The first menstrual period, menarche, and onset of menopause are key milestones of female reproductive ageing[30]. They represent the start and end of reproductive capacity and define the length of a woman's reproductive lifespan[351,353]. Menarche occurs with maturation of the reproductive endocrine system, usually between the age of 10 and 15, denoting sexual maturity for women[354]. It is a highly polygenic trait, with both rare and common variants contributing to the phenotype[77,131,355]. The timings of both menarche and menopause vary widely between individuals[347,351]. The distribution spans from extreme forms that include the absence of puberty and hypogonadotropic hypogonadism, early menopause and POI, to normal ranges that were previously described, and also conditions of late puberty timing[347]. With considerable secular change in which age at puberty declines and age at first pregnancy increases, the mechanisms that regulate both ends of reproductive lifespan became increasingly relevant to population health[347,356]. Understanding whether there is a relationship between age at menarche and menopause, as well as the magnitude, direction and factors influencing this relationship might lead to preventive strategies for infertility, associated chronic disease and improvements in quality of life.

Over the past decade, there was an expansion of GWASs that revealed the complex genetic architectures of menarche and menopause, and aimed at deciphering the association between these two ends of reproductive lifespan (**Figure 6.1**)[347,351,353].

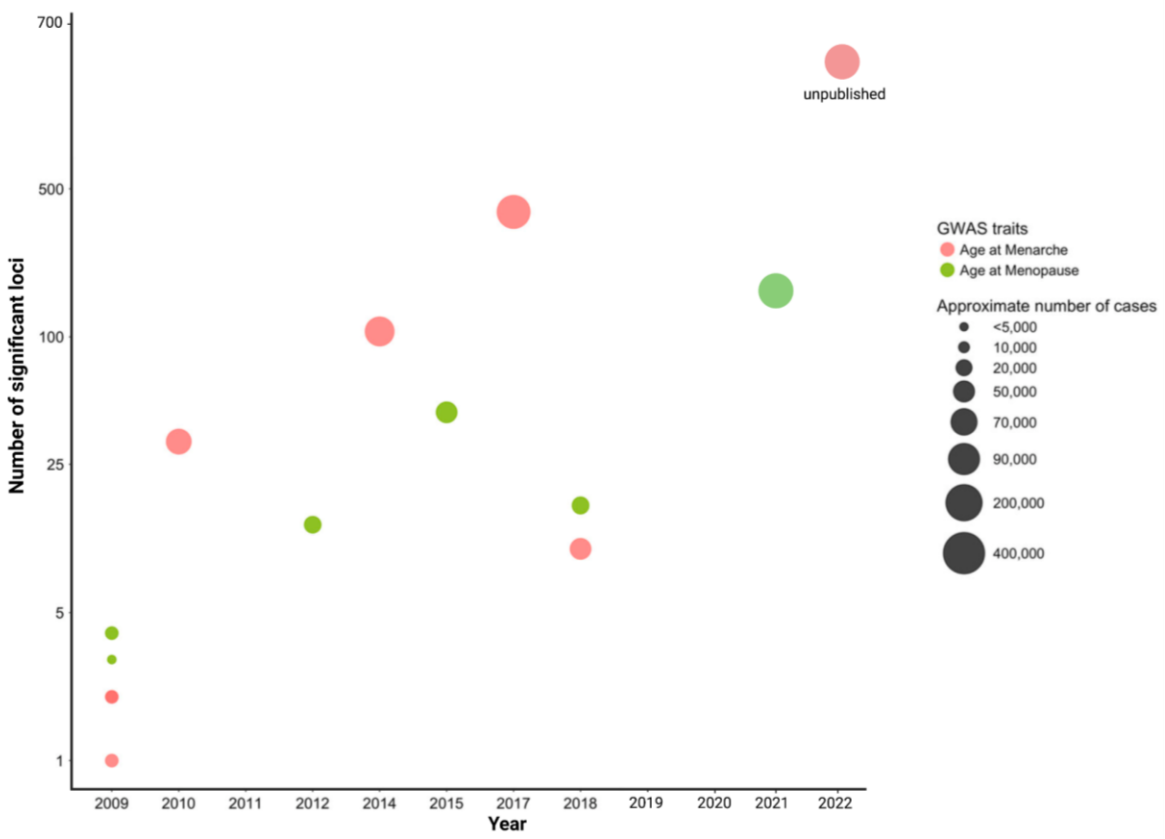

Figure 6.1: The number of identified loci in GWAS for age at menarche and menopause. *The results are plotted as a function of the date of publication and demonstrate the progress in GWAS with enlarged sample sizes and improved genotyping arrays that contributed towards increased power to identify higher numbers of menarche and menopause genetic determinants*[347].

Early studies reported inconclusive evidence on the genetic overlap between identified loci for the timing of menarche and menopause[348,357–360]. However, more recent, better powered GWASs have estimated a modest shared genetic aetiology (genome-wide genetic correlation: $rg = 0.14$; $P = 0.003$)[86], with estimates being supported by new epidemiological evidence linking these two traits[60,83131]. Specifically, causal inference analysis indicated that (genetically mediated) a year earlier age at menarche decreased ANM by about 8 weeks[82]. Significant enrichment (P=0.01) in overlapping signals was found in/near genes that regulate the hypothalamic-pituitary reproductive axis (*CHD7*, *FGFR1*, *SOX10*, *KISS1* and *TAC3*), which were also reportedly mutated in hypogonadotropic hypogonadism[131]. This finding suggested that the same mechanisms may regulate both extremes of reproductive ageing and it initiated discussion around the use of reproductive longevity as a 'proxy' of the general health status[86,351]. However, even with the presence of modest shared genetic aetiology, the regulatory mechanisms clearly differed between the two traits overall. Age at menarche was enriched in genes expressed in the hypothalamus and pituitary gland, thus

highlighting the importance of the central nervous system as the key regulator[131]. On the contrary, menopause gene candidates were mainly expressed in the ovary and other reproductive tissues, and implicated DDR processes that maintain genome stability and hence preserve the ovarian primordial follicle pool[82].

Results from GWAS studies (**Figure 6.1**) demonstrated how increases in sample size over time and more robust statistical tools have improved the power to identify more genetic determinants of reproductive traits[347,351]. These latest insights on the genetics of menarche and menopause were derived from the GWASs conducted on women that are part of the ReproGen consortium[82,131]. The menopause GWAS that identified ~300 genetic variants is described in detail in **Chapter 1** and **2**[82]. Menarche GWAS was conducted in ~370,000 women of European ancestry and detected 389 genetic factors regulating puberty timing. These signals explained ~7.4% of the population variance in age at menarche, corresponding to ~25% of the estimated heritability[131]. Estimates of heritability suggest that 50–70% of variance in age at menarche is due to genetic risk factors[131]. This suggests that many genetic variants contributing to its variation have yet to be identified, some of which might also be responsible for regulation of the other end of reproductive lifespan, i.e. menopause timing. This unrevealed heritability prompted a number of research questions regarding the existence of other potential biological processes that contribute to the wide population variance in reproductive timing, and whether and how these processes are linked to the susceptibility of non-reproductive health outcomes.

To address these questions, we conducted the largest menarche GWAS to date in ~566,000 women from ReproGen consortium and identified 696 independent loci, increasing the total number of menarche associated loci by ~2 fold. A primary challenge in obtaining biological insights from GWAS arises from the inability to directly implicate causal genes and the mechanisms involved. Deciphering how associated variants modulate the outcome risk and severity, and how they impact cellular phenotypes provides a mechanistic insight essential for effective predictive and therapeutic solutions[361]. To improve the approach of identifying the causal genes and addressing the mechanistic interpretation of our findings, we specifically developed 'GWAS2Gene' (G2G), a novel tool that takes advantage of a variety of data sources to provide a solution for gene prioritisation and functional interpretation.
In this Chapter, we combined these novel GWAS findings with developed tools to further investigate the shared aetiology with menopause and other reproductive traits. We provided the first evidence on the involvement of DDR mechanism in regulating the timing of both ends reproductive lifespan, menarche and menopause. This is the first time where the maintenance of the genomic stability was demonstrated to be critical not only for the establishment and maintenance of the ovarian reserve, but also for the initiation

of reproductive activity. Finally, this prompted us to explore the role of DDR as a broader marker of health and disease, and led to the novel human genomic evidence on the involvement of DDR in regulating metabolic and anthropometric traits, such as body mass index (BMI) and height, and suggestive evidence for T2D.

Understanding the functional consequences of genetic association for the same signals in different traits will provide important insights into the similarities and differences in gene regulation underlying risk for various health outcomes.

6.2 Methods

6.2.1 Genome-wide association study for age at menarche

The genetic variants from each individual study coming from ReproGen and BCAC Consortia were tested for the association with age at menarche using an additive linear regression model. Age was included as a covariate, as were any study-specific variables. Insertion and deletion polymorphisms were coded as "I" and "D" to allow harmonisation across all studies. Genetic variants and individuals were filtered based on study-specific quality control metrics. Association statistics for each SNP were then uploaded by study analysts for central processing. Study-level result files were assessed following a standardised quality control pipeline[362]. The results for each variant were meta-analysed using an inverse-variance-weighted model implemented in METAL[363] using a two-stage process. First, for each individual file each of the composite final strata were combined and then filtered such that only variants that appeared in over half of these studies were taken forward. Second, aggregated ReproGen consortium and BCAC results were combined with data from the UK Biobank[147] and 23andMe studies. Variants were only included in the results file if they had combined MAF > 0.1%.

Significant associated loci ($P \leq 5*10^{-8}$) were initially selected using distance-based (1Mb windows) clumping, as explained in **Chapter 2**, **Section 2.2.1**. Then, independent signals were identified using approximate conditional analysis in GCTA[156] with an LD reference panel from the UKBB study. The methodology behind the approximate conditional analysis is described in detail in **Chapter 2**, **Section 2.2.1**. Primary and secondary signals were then checked for LD in 10Mb windows in *plink* (v1.90b6.18)[364]. Only secondary signals that were uncorrelated with previously identified primary signals ($r^2 < 0.05$) were included in the final list. Finally, data was merged with the allele information from

UKBB to provide the full genomic sequence for those alleles that had been designated either "I" or "D". In total, we identified 696 independent genomic signals that regulate menarche timing.

6.2.2 Variance explained

The variance explained by each of the identified variants under the additive model was calculated using the formula $2f(1-f)\beta^2_a$, where f denotes the MAF of the variant and β_a is the effect. Variance explained across multiple variants was calculated by summing these individual variances for all uncorrelated variants. Finally, the percentage of the heritability explained by our top hits was calculated based on the chip heritability for age at menarche obtained from UKBB[131].

6.2.3 Replication

Replication of identified signals was performed in the independent sample in the deCODE study of 39,360 Icelandic women. Given the smaller sample size of the replication cohort, alongside the specific SNP replication, we also performed a global replication test, based on a Binomial sign test.

6.2.4 'GWAS2Gene' (G2G) pipeline for functional annotation and gene prioritisation

The primary challenge in obtaining biological insights from GWAS arises from the inability to directly implicate causal genes and the mechanisms involved. Previous studies showed that the closest gene of a given leading signal is often the causal one[185,365,366], yet it is an imperfect predictor of causality. To prioritise the target genes of identified causal variants our group developed an approach that integrates evidence from multiple individual data sources, which, when combined, provide a more powered source of information on functional links between associated variants and candidate genes. We named this tool 'GWAS2Gene' (G2G).

Leveraging LD information on the sentinel variants within each of 696 signals, we ascertained whether signals could be linked to known enhancers and regulatory elements for, or coding variants within, each of their proximal genes. This would help us assess whether the leading signals directly impact the transcription and/or translation of the target genes. For each GWAS signal we calculated windows of high LD (r^2>0.8), generating a list of proxies with the main signal. These were matched to locations of known enhancers using activity-by-contact (ABC) enhancer maps[367], generated across 131 human cell types and tissues. We then use these data to score individual genes if the leading signal or its proxy fall within one of these identified enhancer regions. If there were coding variants among the list of proxy signals, they

were annotated using SIFT[368] and PolyPhen[369]. We also used gene expression data, in the form of expression quantitative trait loci (eQTL) and protein QTL (pQTL) from Fenland study[111], in tissues specifically enriched in our GWAS, to match the pattern of association we see towards variation in menarche and variation in gene expression. The tissue enrichment analysis was performed via LDSC-SEG[370] and Cell type-specific analysis [https://github.com/bulik/ldsc/wiki/Cell-type-specific-analyses]. The tissues with $P < 0.05$ were highlighted, alongside the data from GTEx tissue fixed-effects meta-analysis (v7)[371], eQTLGen[372] and Brain-eMeta[373].

We then applied the SMR & HEIDI approach (version 0.68) that uses summary-level data from GWAS and eQTL studies to test if a transcript and phenotype are associated due to the shared causal variant, i.e. coincidental overlap of signals due to extended patterns of LD[374]. We used FDR-corrected *P*-SMR <0.05 and *P*-HEIDI >0.001. We supported this analysis by conducting colocalization using Bayes factors (Coloc-ABF) with the R package 'coloc' (Version: 5.1.0)[375] to assess whether two association signals were consistent with a shared causal variant. Coloc uses a user-set prior for the chance of association between a SNP and the phenotype and the variance for this effect size. It relies on a null hypothesis and four alternative hypothesis as described below:

Null Hypothesis: No associated genetic variants with either trait.
Model 1: The locus contains one genetic variant which influences the first phenotype.
Model 2: The locus contains one genetic variant which influences the second phenotype.
Model 3: The locus contains two separate genetic variants which influence the first and second phenotype respectively.
Model 4: The locus contains one genetic variant which influences both phenotypes.

SNPs that follow model 4, i.e. where the posterior probability that both traits are associated and share a single causal variant is >= 0.75, were defined as co-localised causal variants.

At the gene-level association, we applied the Polygenic Priority Score (PoPs) method[376], which uses bulk human and mouse data with information on scRNA, gene pathways and protein interactions and prioritises genes proximal to GWAS signals based on these biological annotations. In addition to PoPs, we integrated the evidence from gene-level MAGMA analysis based on the protein coding variants[377].

Finally, signals were paired to the closest genes based on the 1Mb window of the genes' start or end sites (i.e. 500kb up- and downstream of each signal) using National Center for Biotechnology Information (NCBI) RefSeq gene map for GRCh37 as a reference for the gene location (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/). The gene for intragenic signals was

automatically assigned as the closest. We overlaid all this information to calculate aggregate scores for every gene in the genome, based on these six layers of evidence:

1. The closest gene evidence carried 1.5 points.
2. From the eQTL colocalization analysis, evidence found for either SMR or coloc carried 1 point, while if the evidence was present for both analyses we assigned 1.5 points in total. We added 1 extra point if the eQTL causal variant and GWAS signal had r^2<0.05.
3. The same rules for eQTLs were applied for pQTL colocalization.
4. For coding variants, variants from MAGMA analysis and variants in LD were scored together because of the related input. The variants with FDR-corrected MAGMA *P* value <5% carried 0.5 points. In addition, if the variant was annotated as deleterious or damaging it obtained 1 point, while benign or tolerated coding variants got 0.5 points.
5. The evidence from ABC enhancers carried 1 point
6. The evidence from PoPs carried 1.5 points.

The score was summed for gene-signal pairs across 6 described analyses. Genes with 0 points were removed. The remaining gene-signal scores were adjusted for signal LD window size (r^2 >0.5). For genes with more than one signal we kept the highest scoring signal and added an additional point to the final score, ultimately obtaining the so-called 'unique gene score', which was adjusted for the gene length. Finally, we calculated the genome-wide rank per gene and summarised the number of evidence sources supporting each gene (maximum 6). For downstream interpretation, we focused on scored genes (N=509) that have >=3 evidence sources. The outcome of this pipeline can be found in the **Appendix Table 6.1**.

6.2.5 Functional annotation of G2G genes

Functional annotation of menarche genes highlighted via G2G analysis was performed using integrative databases that provide comprehensive functional and molecular information on all annotated human genes. These included Gene Cards and Open Targets Genetics. In addition, the information on specific genetic disorders related to 509 AAM genes of interest was obtained from the Online Mendelian Inheritance in Man (OMIM) database, an online catalogue of human genes and genetic disorders. Due to the specific focus on the potential role of the DDR mechanism in regulation of menarche timing, which will become relevant in later sections of this Chapter, we built comprehensive evidence that integrates 5 different sources, listing genes with known involvement in DDR. These included expert curated DDR gene lists: (1) 'Broad DDR', the gene list curated in the laboratory of Professor Stephen Jackson and (2) 'Expert curated DDR 2', the DDR gene list curated for our previous study[82] through collaboration of

internal (Professor John Perry) and external experts (Professor Eva Hoffmann and Professor Anna Murray). 'Broad DDR' covered a few angles of DDR biology: DNA repair genes, broader DNA damage response genes (damage-induced chromatin remodelling, transcription regulation and cell cycle checkpoint induction) and general maintenance of genome stability, such as genes involved in DNA replication. In addition, we obtained the DDR-related gene sets from various online enrichment analysis tools, including REACTOME (DNA Repair: R-HSA-73894) and Gene Ontology (DNA Repair: GO:0006281 and Cellular response to DNA damage stimulus: GO:0006974) (**Appendix Table 6.5**).

6.2.6 The Genotype Tissue Expression (GTEx)

We used GTEX, a publicly available resource of tissue-specific gene expression, to lookup the tissue expression of 509 AAM genes highlighted by G2G analysis (**Appendix Table 6.6**)[378].

6.2.7 Exploring biological mechanisms underlying menarche timing using MAGMA

The causative genetic factors usually have small effect sizes in complex diseases, thus their detection by single-variant statistical analyses is challenging. To increase the statistical power of the hypothesis-free GWAS and reduce the burden of multiple testing, it is better to consolidate the effects of multiple variants by providing functional annotations. To address this, we applied a pathway analysis, MAGMA, a gene and gene-set analysis of GWAS genotype data[377]. This enabled us to analyse multiple genetic markers simultaneously to determine their joint effect, detect potential risk genes and better explain mechanisms underlying health outcomes of interest. We used age at menarche GWAS meta-analysis summary statistics, described in **Section 6.2.1**, as an input file.
MAGMA consists of three steps: (1) an annotation step to map SNPs onto genes; (2) a gene analysis step to compute gene p-values; and (3) a gene-level analysis step, i.e. gene-set analysis.

6.2.7.1 Annotation

SNP to gene annotation was performed as part of the pre-processing step using Variant Effect Predictor (VEP)[162]. As the gene names were under the Ensembl stable ID setting, we further converted the gene IDs to Entrez accession numbers using g:Convert (https://biit.cs.ut.ee/gprofiler_beta/convert), a format required by MAGMA. The gene location file was downloaded from the MAGMA website for genome build 37 (https://ctg.cncr.nl/software/magma). The annotation output file consisted of each row corresponding to a gene, containing the gene ID, a specification of the gene's location, and a list SNPs

mapped to that gene. This default version of the MAGMA input file contained all genetic variants across the genome available in the summary statistic of the phenotype of interest. We named the default MAGMA version as 'Unrestricted MAGMA'.

Additionally, we proposed a new approach, namely 'coding MAGMA (cMAGMA)' for gene-wise annotation of protein coding, predicted deleterious variants from GWAS summary statistics. The same methodology, as described above, was used in case of cMAGMA with an exception of the variant file provided, which in this case contained only predicted deleterious variants as defined per VEP. The results from the cMAGMA model were used as the primary ones.
To account for LD between SNPs we used 1,000 Genomes European reference genome, available for download from MAGMA website: https://ctg.cncr.nl/software/magma. The 1000 Genomes reference data file was created from the Phase 3, with SNP locations in reference to human genome build 37. The two main requirements for the reference data were followed: (1) the existence of a strong overlap between the reference data and the input file, since only SNPs that occurred in both files were used in the analysis; and that 2) the general ancestry of the reference data matched the input file data.
The input data coming from GWAS summary statistics have undergone appropriate quality control and filtering prior to running the MAGMA, including imputation quality.

6.2.7.2 MAGMA gene-level analysis

MAGMA gene-level analysis tested the joint association of all markers in the gene with the phenotype of interest using a SNP-wise gene model that computed gene test statistics combining SNP *P*-values through multiple regression. The computed gene level metrics quantified the degree of association each gene has with menarche timing. In addition, the correlations between neighbouring genes were estimated as a prerequisite for gene set analysis, required to compensate for the dependencies between genes.
Computed SNP *P*-values, sample size and reference data were provided to run: (1) 'cMAGMA' analysis, which focused solely on coding variants, and (2) 'Unrestricted' analysis, which focused on all GWAS variants. The *P* and *Z* value statistics were computed to see if tested genes are significantly associated with menarche timing.

6.2.7.3 MAGMA gene-set analysis

The gene-set analysis was performed with an aim to explore the involvement of specific biological pathways in menarche aetiology. In this analysis individual genes were aggregated to groups of genes sharing certain biological, functional or other characteristics based on the 'Canonical pathways'. These

contained the evidence from Kyoto Encyclopaedia of Genes and Genomes (KEGG), Reactome, Gene Ontologies (GO), BioCarta and Pathway Interaction Database (PID). Then, competitive gene-set analysis using regression models was performed to test whether the genes in a specific gene-set are more strongly associated with the trait of interest than other genes, suggesting enrichment. By default, the gene set variable is conditioned on the gene size, gene density, the relative level of LD between SNPs in that gene and the inverse of the MAC to correct for potential power loss in very low MAC SNPs, as well the log value of these three variables. The association that a gene has with the phenotype was quantified as a Z-score, a probit transformation of the gene *P*-value computed during the gene analysis step ($Zg = \Phi(1 - Pg)$, mapping low *P*-values onto high positive Z-scores ($Zg = 0$ corresponds to $Pg = 0.5$). We applied the Bonferroni correction (*P*: $0.05/2233 = 2.24*10^{-5}$) to highlight the pathways that were significantly associated with the age at menarche.

6.2.8 Lookups of age at menarche signals in other reproductive health outcomes

To explore the potential shared genetic architecture between age at menarche and menopause, we performed a lookup of 696 menarche lead signals in the ReproGen menopause summary statistics available for ~250,000 women of European descent[82]. The data were harmonised in the direction of the age at menarche increasing allele. The *P* value significance and direction of the effect were compared between two traits of interest. The signals were highlighted if they passed a Bonferroni corrected ($P \leq 0.05/696 = 7.2*10^{-5}$) or genome-wide significant *P* value ($P \leq 5*10^{-8}$).

The same approach was taken in the opposite direction, i.e. to assess the genetic overlap of 290 menopause signals, identified in our most recent ReproGen GWAS[82], in age at menarche summary statistics. The signals were highlighted if they passed a Bonferroni corrected ($P \leq 0.05/290 = 1.7*10^{-4}$) or genome-wide significant *P* value ($P \leq 5*10^{-8}$).

We were also interested to examine the effect of 696 menarche lead signals in the largest available summary statistics of other reproductive traits, including polycystic ovarian syndrome (PCOS)[379], NEB[380] and twinning[381]. We examined the consistency in the direction of associations between individual signals, as well as their statistical significance. If a particular menarche signal was not present in other datasets, we searched the UKBB white European dataset for proxies. The criteria applied for the proxy selection included the proxy distance that falls within 1 Mb window and the $r2 > 0.5$ with the main signal. The proxy with the highest r^2 value was selected as the proxy of choice.

Finally, in order to test whether 696 menarche signals also associate with BMI, we performed a variant lookup using sex-combined BMI data available in the UKBB. We applied both Bonferroni correction ($P \leq 0.05/696 = 7.2*10^{-5}$) and genome-wide significance threshold ($P \leq 5*10^{-8}$) to identify significant associations.

6.2.9 Understanding genetic associations using colocalization

The lookups described in the **Section 6.2.8** were supported by the evidence from the colocalization analysis, which was used to determine if age at menarche and menopause genetic association in the same locus is being mediated by the same underlying causal variant. The methodological rationale behind colocalization using the *coloc* package was previously described in the Section **6.2.4**. It is important to note that colocalization analysis cannot determine causal relationships between the phenotypes being studied, and neither can it determine the direction of causality between the two phenotypes.

6.2.10 Functional enrichment tests for *ZNF483* transcription factor binding sites using fGWAS and SLDP

We implemented fGWAS (*v.0.3.6)*, a hierarchical model for joint analysis of GWAS and genomic annotations, to test the functional enrichment of age at menarche GWAS hits in *ZNF483* transcription factor binding sites[197]. The fGWAS input file contained the menarche GWAS summary statistics, described in **Section 6.2.1**, annotated for *ZNF483* binding sites. The *ZNF483* annotation file was derived from the ENCODE ChIP-seq data from human HepG2 cell line [ENCSR436PIH] with available start and end transcription binding sites across the genome (GRCh38). As the genomic location of transcription binding sites was in GRCh38 format, we converted it to hg19 assembly using NCBI remapping tool (https://www.ncbi.nlm.nih.gov/genome/tools/remap). Menarche GWAS hits were annotated for the presence/absence of the *ZNF483* transcription factor binding sites in a binary way (0, 1), with '1' if the SNP falls within the transcription factor binding site and '0' otherwise. fGWAS was run using the fGWAS tool available at https://github.com/joepickrell/fgwas. Detailed description of fGWAS methodology is available in Pickrell *et al*, 2014[197] and **Chapter 3**, **Section 3.2.7**. The fGWAS output contained the maximum likelihood parameter estimates for each parameter in the model, in this case *ZNF483*, with the lower and upper bound of the 95% confidence interval on the parameter.

SLDP regression was applied to explore the directional effect of a signed functional annotation, *ZNF483*, on age at menarche using GWAS summary statistics. More specifically, we tested whether alleles that are predicted to increase the binding of the transcription factor *ZNF483* have a genome-wide tendency to

increase or decrease puberty timing in girls. The SLDP tool was installed from https://github.com/yakirr/sldp, with the comprehensive methodological steps described in Reshef *et al*, 2018[199] and **Chapter 3**, **Section 3.2.7**. For the analysis to be conducted, SLDP required age at menarche GWAS summary statistics, signed LD profiles for *ZNF483* binding, signed background model and reference panel in a SLDP compatible format. The *ZNF483* annotation file, obtained from the ENCODE CHIP-seq analysis as described above, was preprocessed using the '*preprocessannot*' tool that turns signed functional annotations into signed LD profiles. SLDP was run on our data using the '*sldp*' function.

6.2.11 Exploring the role of DDR, cell cycle and death across multiple health outcomes using MAGMA

A wide range of phenotypes was tested within the MAGMA gene and pathway analysis, covering reproductive, cardiovascular, metabolic, and neurodevelopmental health outcomes. The analysis was run on 35 health outcomes in total. Age at natural menopause, loss of chromosome Y (LOY) and cancer phenotypes were selected as positive controls due to the previously described role of DDR, cell cycle and death in their aetiology. Since raw genotype data were not available for some traits, we focused on SNP-wise gene analysis model that required summary statistics data. Most of the health outcomes of interest were available in the UKBB study. GWASs of mentioned phenotypes were conducted using linear mixed models, implemented in BOLT-LMM v2.3.4 and described in **Chapter 2**, **Section 2.2**[154]. The regression models included age and genotyping array as covariates, unless stated otherwise. For the phenotypes that were not available in UKBB, we searched for the largest, publicly available GWAS summary statistics. For more details on these datasets, including the studies they were obtained from, please refer to the **Appendix Table 6.7** and original publications.

MAGMA was run following three steps as previously described: (1) an annotation step to map SNPs onto genes; (2) a gene analysis step to compute gene p-values; and (3) a gene-level analysis step, i.e. gene-set analysis. We performed both cMAGMA and unrestricted default version, however in the main text we only describe the results from cMAGMA analysis. To account for LD between SNPs we used UKBB reference genome and we replicate the analysis using 1,000 Genomes European reference genome, available for download from MAGMA website: https://ctg.cncr.nl/software/magma.

Being specifically interested in the role of DDR and cell cycle mechanisms in the phenotypes of interest, we focused on the following biological pathway sources:

(1) **Gene Ontologies (GO)** that describe gene properties from a hierarchical class structure of three main aspects: molecular functions, biological processes, and cellular components. GO annotation was downloaded from the Gene Ontology website (http://geneontology.org/docs/download-go-annotations/). Our main focus were biological processes, including GO5:DNA_Repair (GO:0006281), GO5:Cell_cycle (GO:0007049), and GO5:Cellular_response_to_DNA_damage_stimulus (GO:0006974). The selection of the named GO terms from the GO ancestor tree was based on the most comprehensive description of the pathways of interest that incorporated multiple 'child terms', which more specifically describe the mechanisms involved.
(2) **Expert-curated pathway 'Broad DDR'**: In addition to the publicly available categories, we also obtained a customised pathway 'Broad DDR', specifically curated by the DDR experts from the Jackson's laboratory, University of Cambridge. The content of this pathway is described in detail in the **Section 6.2.5** (**Appendix Table 6.5**). In order to detect reliable gene candidates across phenotypes of interest, we checked for the consistency between the genes identified using GO terms DNA Repair and 'Broad DDR' pathways.

We applied the false discovery rate (FDR) threshold to highlight the pathways that were significantly associated with tested health outcomes, and calculated 'score' per outcome that reflected the number of times a certain pathway was significant (**Appendix Table 6.8**).

In order to understand which genes from our pathways of interest play a significant role across selected health outcomes, we extracted MAGMA gene level statistics for each gene in each pathway, and highlighted the ones significantly associated with the traits based on the Bonferroni corrected *P* value. A final score was computed according to the number of traits a particular gene was significant for.

6.3 Results

6.3.1 GWAS discovery and replication of common genetic variants regulating menarche timing

We performed an expanded GWAS meta-analysis for age at menarche, combining 73 studies with GWAS data imputed to either the Haplotype Reference Consortium (HRC) or 1000 Genomes imputation panels, comprising ~566,000 women of European ancestry. This specifically included data from the ReproGen and BCAC consortiums, combined with the UKBB and 23andMe studies. We detected 655 loci, defined as a 1 Mb window containing variants associated with age at menarche at $P \leq 5.0\times10^{-8}$. Following approximate conditional analyses using GCTA, we identified 696 independent signals associated with age at menarche (**Appendix Table 6.2**). We sought independent replication in the deCODE study (N=39,360 Icelandic women). Of our 696 independent signals, 687 (99%) were present in deCODE of which 616 (90%) showed directionally concordant associations with AAM ($P_{\text{Binomial}} = 1.3\times10^{-109}$), while 281 signals were at $P<0.05$. In deCODE, these signals explained 9.1% of the population variance in menarche timing. In the Danish Blood Donors Study (N= 35,467 Danish women), the variance explained, estimated from summary statistics, in menarche timing increased from 11% by the previously reported 389 signals to 14% by the current 696 signals, corresponding to ~29% of the estimated heritability.

6.3.2 DNA damage response is a novel regulatory mechanism of menarche timing

To implicate potential biological pathways that regulate menarche timing, we applied the MAGMA gene and gene-set analysis approach to our genome-wide common variant associations with age at menarche. We adapted the default MAGMA approach to strengthen the level of inference by considering only coding variants, and termed this 'coding MAGMA' (cMAGMA). Three canonical pathways were enriched for coding variant associations with menarche timing at the multiple test corrected $P \leq 2.2\times10^{-5}$. The significant pathways included: '*DNA repair*' ($P=6.4\times10^{-6}$), '*Acute myeloid leukemia*' ($P=6.4\times10^{-6}$), and '*DNA damage reversal*' ($P=9.8\times10^{-6}$) (**Figure 6.2**). This is the first evidence linking DDR pathways to puberty timing regulation.

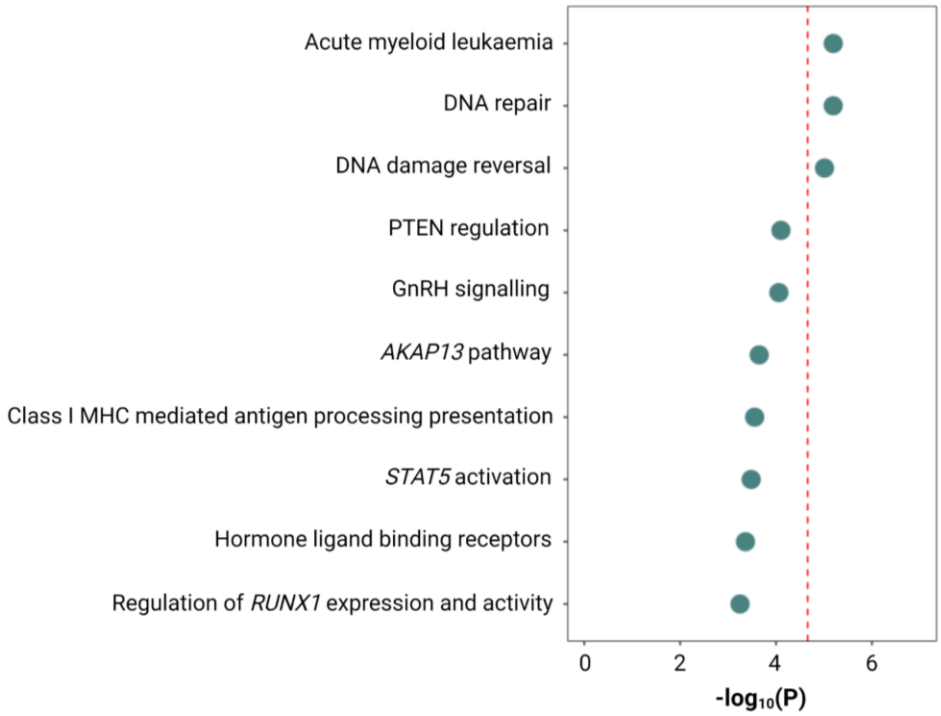

Figure 6.2: Top 10 enriched pathways for age at menarche identified using cMAGMA. *The red vertical dotted line represents Bonferroni P value threshold (0.05/2233=2.24×10^{-5}).*

Other specific DDR-related pathways were enriched at nominal significance: '*PTEN regulation*' (P=7.9×10^{-5})[382], '*DNA damage bypass*' (P=1.2×10^{-3}), '*Sumoylation of DDR proteins*' (P=1.8×10^{-3}), '*Cellular senescence*' (P=6.6×10^{-3}), '*DNA damage telomere stress induced senescence*' (P=1.2×10^{-2}), and '*DNA double strand break repair*' (P=1.4×10^{-2}). Beyond DDR, the well-established central regulator of reproductive hormone axis activity, '*GnRH signalling pathway*', was at the borderline multiple test correction significance (P=8.8×10^{-5}).

6.3.3 Identification of high confidence menarche genes using G2G

To implicate the genes that regulate menarche timing from genome-wide common variant associations, we developed an analytical framework, G2G, which integrates genomic and functional evidence across 6 sources (**Appendix Table 6.1**).

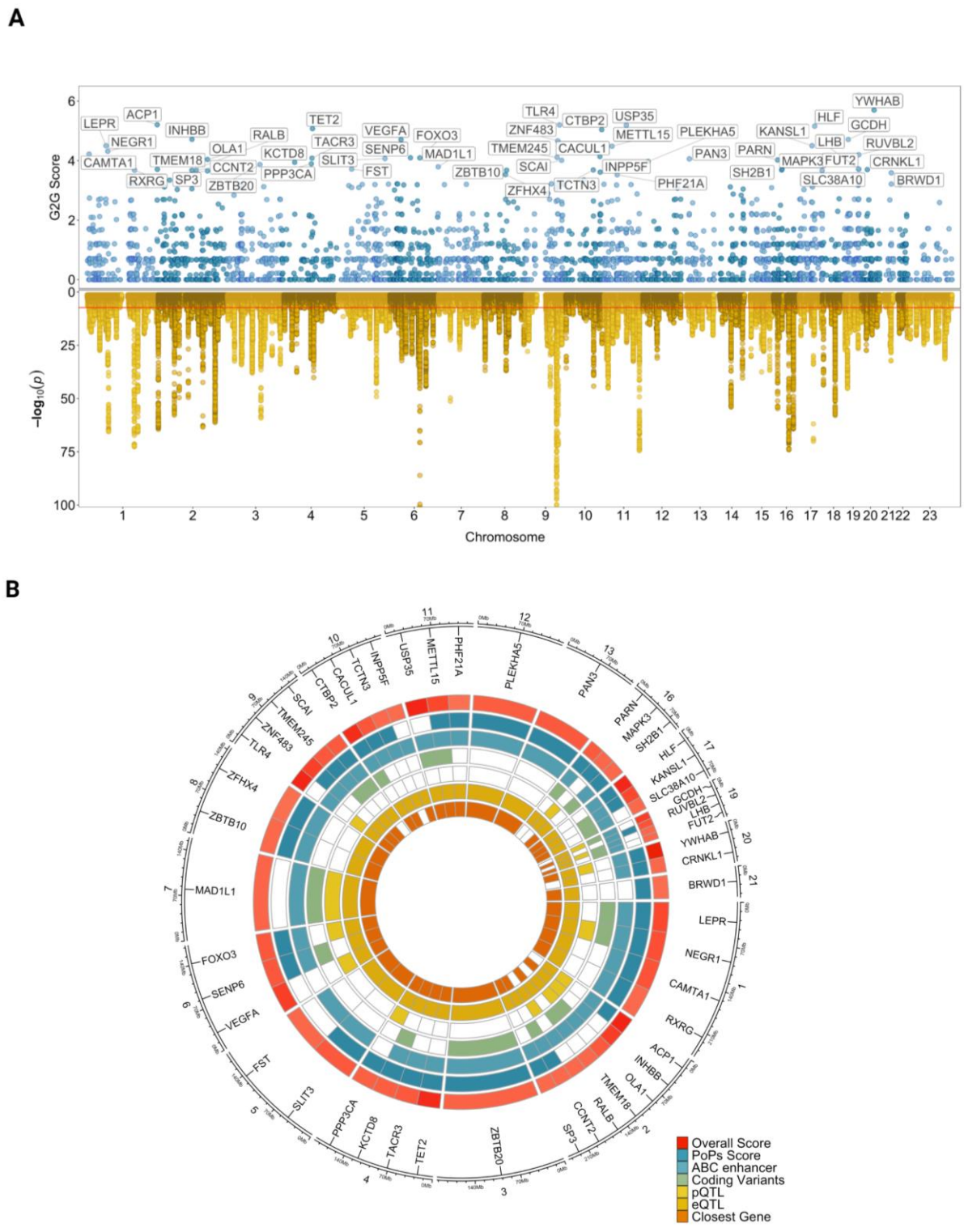

Figure 6.3: Age at menarche gene prioritisation using G2G. (A) *Miami plot showing signals from the GWAS meta-analysis for age at menarche (lower) with genome-wide G2G scores (upper).* ***(B)*** *The 50 highest scoring genes implicated by the G2G pipeline are indicated, demonstrating the evidence for 6 different predictors used for gene prioritisation.*

We identified 509 high-confidence age at menarche genes, proximal to a GWAS signal and implicated by at least three concordant predictors (**Appendix Table 6.1**). We manually curated these 509 top scoring menarche genes using Gene Cards, Open Targets Genetics, and PubMed to get an insight into their biological function. In addition, we combined evidence from the expert curated DDR gene sets and DDR-related gene sets from various online enrichment analysis tools to explore the potential role of individual menarche genes in the DDR mechanism.

Top G2G scoring menarche genes included established components of the hypothalamic-pituitary axis that regulates sex hormone secretion and gametogenesis[383]: *LHB, FSHB, TACR3* and *GNRH1*, as well as genes that encode gonadal secreted hormones or peripheral sex hormone metabolism: *INHBB*[384], *FST* [385] and *POR*[386]. In addition, we detected well-known genes that are reported as disrupted in monogenic disorders of reproduction. This included *CADM1*[387], *SEMA3G*[388], *SRA1*[389] and *TYRO3*[390–392] responsible for Congenital Hypogonadotropic Hypogonadism (CHH), *MKRN3* [393,394] detected in Central Precocious Puberty (CPP), *LEPR*[395–398] in severe obesity with delayed puberty, as well as *THRB* [399] in thyroid hormone resistance.

6.3.4 Common menarche associated variants are enriched in *ZNF483* binding sites

Among the potential novel regulators of the reproductive hormone axis, we highlighted *ZNF483,* which also came to the attention of our group due to the significant association of *ZNF483* damaging variants with age at menarche (1.31 years later menarche, P=4.90*10^{-11}, N female carriers: 59) in UKBB whole exome sequence analysis (unpublished work by PhD student Lena Kaisinger). Several other members of the zinc finger superfamily have previously been linked to menarche timing using GWAS data[400–402] as well as animal models[403], which suggests that ZNFs may be transcriptional regulators involved in the process of epigenetic regulation of puberty timing. *ZNF483* is involved in transcriptional regulation, regulation of neuronal differentiation via interaction with MeCP2[404], and maintenance of the self-renewal of human pluripotent stem cells[405]. In *ZNF483* ChIP-seq data available from the ENCODE project and using fGWAS[197], we found that common variants associated with menarche timing were enriched in the transcriptional targets of *ZNF483* (P=1.2*10^{-6}). This included binding site co-localisation with 4 previously reported menarche GWAS loci – rs1131017 (*RPS26*), rs192330071 (*PRSS57*), rs2659005 (*SLC38A10*) and rs3813321 (*EGR1* – which roles in DDR are starting to be revealed[406]. We extended this result by testing functional enrichment using SLDP regression[199]. This demonstrated that *ZNF483* bound loci are significantly associated with earlier menarche timing (Z=-4.1, P=4.1*10^{-5}). This is directionally

concordant with the results of rare damaging *ZNF483* variants, and together indicate that 'more' *ZNF483* binding promotes earlier age at menarche.

6.3.5 DDR genes with novel evidence on the involvement in the initiation of reproductive activity

We also highlight notable examples of G2G top scoring menarche genes that are implicated in DDR processes. We identified 44 genes with 2 or more evidence sources as being involved in DDR mechanisms and 37 genes with suggestive evidence on DDR, i.e. a single evidence obtained via manual curation of the literature. The highest scoring gene in menarche G2G analysis, *YWHAB*, encodes an adapter protein with roles in metabolism, protein trafficking, DDR, apoptosis and cell cycle regulation[407]. Another interesting example involves a transcription factor, *FOXO3*, which plays an important role in the cell cycle, DDR, apoptosis, autophagy and energy metabolism[408–411]. Recent studies have shown that *FOXO3* is involved in the physiological regulation of follicular development and pathological progression of related ovarian diseases[412]. Other examples with evidence on the involvement of DDR include: *UBE2B* (ubiquitination and post-replicative DNA damage repair)[413], *ALKBH3* (single-stranded DNA repair)[414], *ASCC3* (3'-5' DNA helicase involved in repair of alkylated DNA)[415], *TLR4* (regulation of apoptosis, innate immune system and mediation of the effects of palmitate on GnRH neurons)[416,417], and *TP53BP1* (involved in response to DNA damage, telomere dynamics and class-switch recombination during antibody genesis; expressed in GnRH positive neuron cells)[418,419].

6.3.6 The effect of BMI on the association between DDR and menarche timing

The role of adiposity in regulating menarche timing is supported by epidemiological and genetic studies. These reported a strong genetic correlation between puberty timing and body mass index (BMI) and that many genes involved in the regulation of fat mass are also associated with timing of menarche[420,421]. To explore whether observed menarche associations are potentially mediated by BMI, we looked them up in the UKBB BMI sex-combined dataset (**Appendix Table 6.2**). Specifically, we were interested if observed DDR effects are specific to menarche or might influence menarche by changing BMI. We found that 158 out of 696 genetic variants were influenced by BMI at the Bonferroni corrected *P* value ($P \leq 7.18*10^{-5}$), 93 of which were at the genome-wide significance level ($P \leq 5x10^{-8}$). Notable examples include genes that are well described for their association with BMI and obesity. This includes *SEC16B*[422], *ACP1*[423], and *ADCY3* that causes monogenic form of severe obesity[424,425], which all lead to earlier puberty timing in girls and raise the susceptibility for higher BMI and obesity. Of those 158 genetic variants that influence both age at menarche and BMI, only 17 encoded genes that are involved in

DDR mechanism, 7 of which influenced BMI at the genome-wide significance *P* value ($P=5*10^{-8}$). Those included *MAPK3* (activated by LH in ovarian granulosa cells and also involved in DDR, cell cycle, apoptosis, initiation and regulation of meiosis, mitosis, and postmitotic functions), *TRA2B* (involved in splicing regulation in oogenesis and regulation of DDR)[426,427], *FOXO3*, *RAD52*, *MUS81* (HR repair)[428], *GADD45G* (candidate gene for male infertility and 46,XY sex reversal in humans) and *SPI1*. All of these BMI-associated menarche genes were expressed in ovarian, adipose and brain tissues, besides *SPI1*, which showed very low expression in ovaries and brain yet high in adipose tissue. The remaining menarche loci that tagged DDR genes showed no association with BMI in females, suggesting a direct influence of DDR on menarche.

Some notable examples include DDR genes (or genes with suggestive DDR function) that were lowly expressed in the brain but highly expressed in ovarian tissue, potentially indicating an ovary specific mechanism for regulation of menarche timing (**Figure 6.4**, **Appendix Table 6.6**).

These include *DDIT4* (regulates cell growth, proliferation and survival via inhibition of the activity of mTORC1)[429], *MSH6* (mismatch repair gene involved in reproductive ageing and menopause timing)[184], *GLI2* (zinc-finger transcription factor involved in the Sonic Hedgehog pathway)[430–432] and *DET1* (component of the ubiquitination machinery that mediates the destabilization of key regulators of cell differentiation and proliferation). The underlying variants were not associated with BMI and these genes did not show shared genetic architecture with BMI.

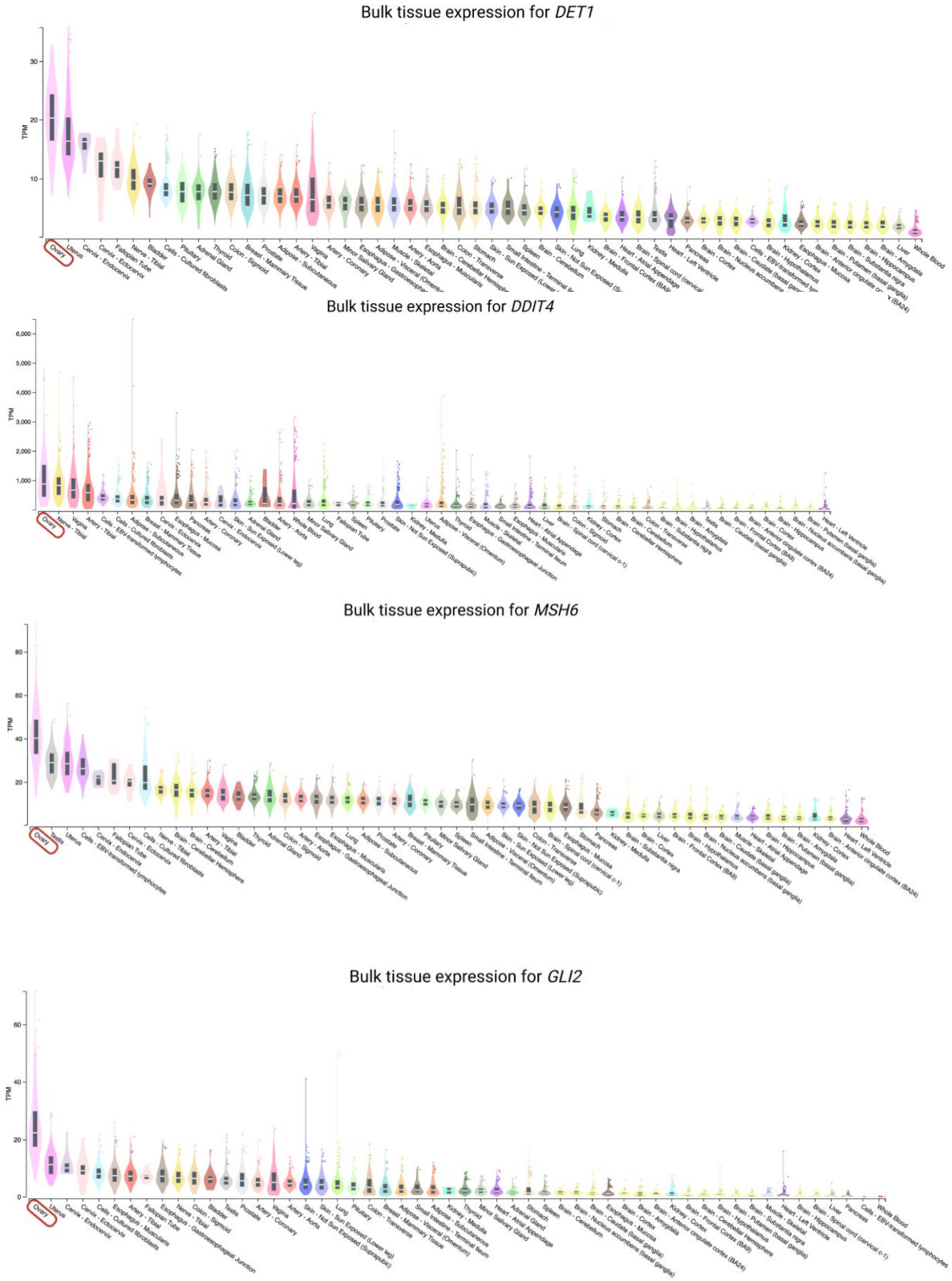

Figure 6.4: GTEx tissue expression of age at menarche genes of interest. *The tissue expression of DDIT4, MSH6, GLI2 and DET1. Ovarian tissue is coloured in pale pink while brain tissue is in yellow.*

6.3.7 Shared genetic architecture between menarche and menopause timing

The role of DDR is well established in menopause timing. Following our novel evidence on the enrichment of DDR mechanisms in the regulation of menarche timing, we hypothesised that DDR-related menarche genes might additionally be associated with menopause timing, thus impacting both beginning and end of reproductive longevity. We performed a lookup and colocalization analysis of the 696 lead menarche variants in the largest menopause GWAS on ~250,000 women of European ancestry (**Appendix Table 6.2**). Of the 696 menarche variants, 13 were also genome-wide significant for ANM (P=5*10^{-8}) and a further 10 passed the multiple test correction threshold ($P \leq 0.05/696$=7.2*10^{-5}). Of these 23 common signals, 7 colocalized with and passed genome-wide significance for menopause timing (P=5*10^{-8}), 5 colocalized and passed the multiple test correction threshold ($P \leq 0.05/696$=7.2*10^{-5}), and an additional 2 variants colocalized and were at the borderline of multiple test corrected significance (**Figure 6.5A**). Of these 14 variants that regulate both ends of the reproductive lifespan, 11 showed the same direction of effect (i.e shifted reproductive lifespan), and the other 3 had directionally-opposing effects on menarche and ANM, hence extending/shortening reproductive lifespan.

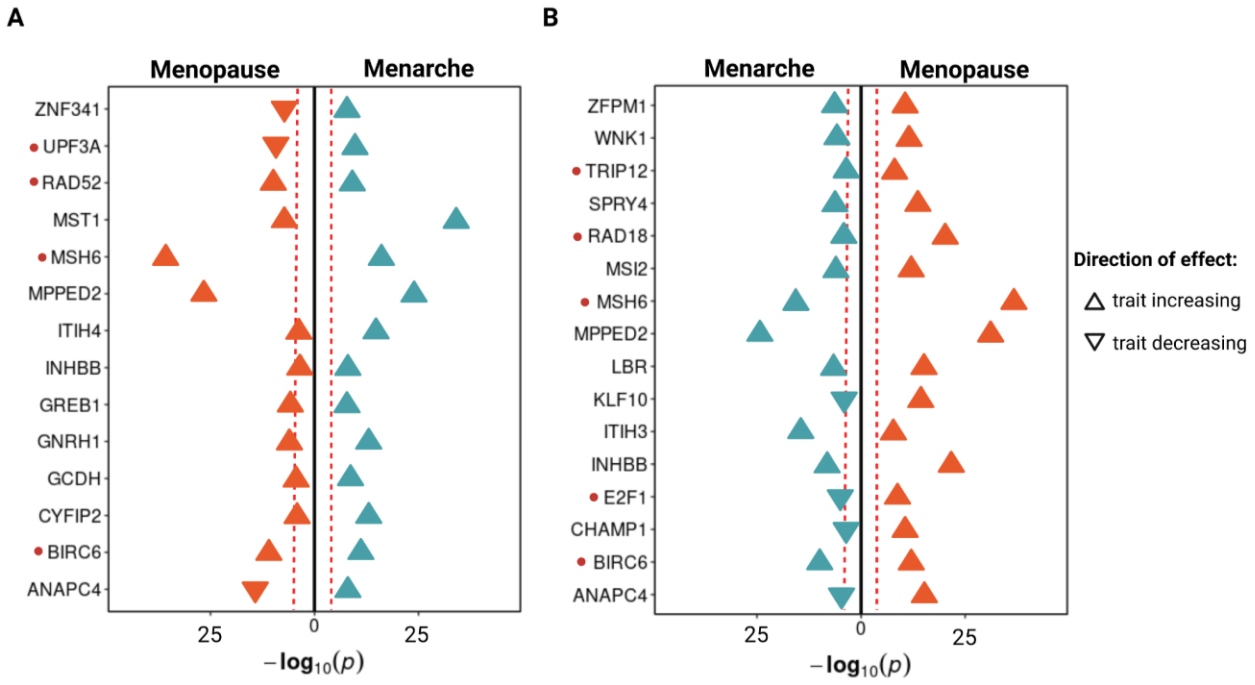

Figure 6.5: Shared genetic loci underlying aetiology of age at menarche and natural menopause. (A) *The shared genetic loci from the lookup and colocalazation analysis on 696 menarche signals in the menopause GWAS.* ***(B)*** *The shared genetic loci from the lookup and colocalazation analysis on 290 menopause signals in the menarche GWAS. The red vertical dotted line represents the Bonferroni corrected significance threshold. Red dots next to the gene name indicate that the gene belongs to the DDR pathway.*

Notably, four of the colocalised signals mapped to genes that belong to DDR mechanisms, including *BIRC6*, *MSH6*, *RAD52* and *UPF3A*. All DDR genes besides *UPF3A* influenced both ends of reproductive lifespan in the same direction, thus shifting the reproductive 'window'. Conversely, *UPF3A* led to later menarche timing and earlier age at natural menopause, thus extending/shortening the length of reproductive lifespan. One of the shared variants (rs3136249; age at menarche $P=7.4 \times 10^{-17}$ and ANM $P=2.2 \times 10^{-36}$) is intronic in *MSH6*, a DNA mismatch repair gene, which is mutated in the cancer predisposing Lynch syndrome and previously described to influence ANM by acting on the ovarian reserve[184]. Additionally, we found that the variant was not associated with puberty timing in boys ($P=0.053$) (**Appendix Table 6.2**) and that *MSH6* expression was the highest in peripheral reproductive tissues, such as ovary and uterus (**Figure 6.4**). This suggests that the DDR mechanism we observe in puberty timing for *MSH6* is potentially the same one that acts via ovarian reserve to modify the timing of menopause. Mice conditionally lacking *UPF3A* were reported to display defects in embryogenesis and gametogenesis[433], while *BIRC6* expression was reported to be essential for embryo survival during preimplantation development[434].

Of these four colocalised signals, only *RAD52* was also associated with BMI ($P=4.8*10^{-22}$). *Rad52* knockout mice are characterised by defective repair of meiotic recombination and subsequent apoptosis of foetal oocytes resulting in decreased primordial follicles from birth and infertility[435].

Besides DDR, other biological pathways involved in the regulation of both menarche and menopause timing belonged to hormonal regulation and development of the hypothalamic-pituitary-adrenal-gonadal axis, regulation of the cell cycle and transcription, and immune response.

A notable gene that demonstrated the highest expression in the ovarian tissue and very low or almost absent in other tissues was *GREB1* (**Figure 6.6**). The shared genetic signal showed no association with BMI ($P=3.3*10^{-1}$) and a nominal association with puberty timing in boys (voice breaking) ($P=7.5*10^{-3}$). We therefore speculate that this gene might act via ovary-specific mechanism to modify menarche and menopause timing. *GREB1* functions as a coactivator of oestrogen receptor alpha (ERα) and it represents the first inducible O-GlcNAc glycotransferase in the cytoplasm to be identified in mammals. Mice lacking *Greb1* exhibit growth and fertility defects reminiscent of phenotypes in ERα-null mice[436,437].

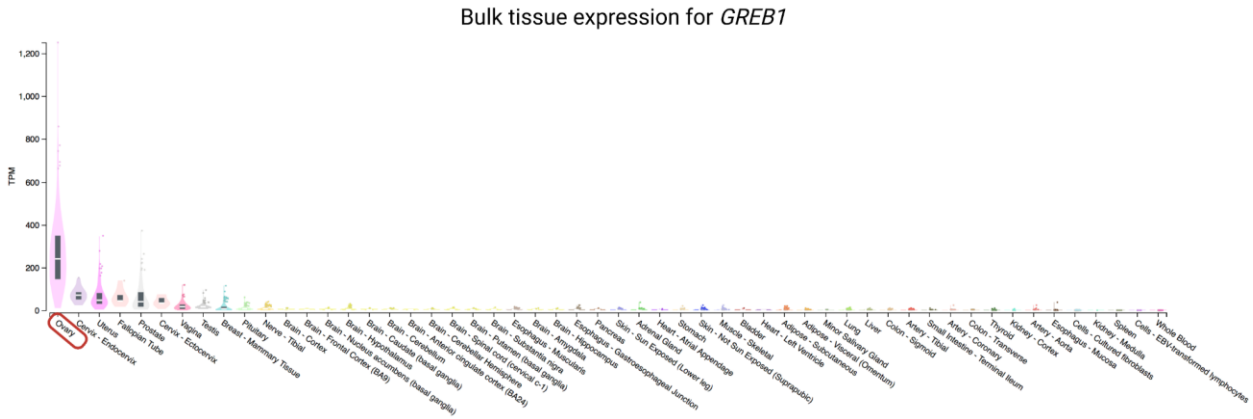

Figure 6.6: GTEx tissue expression of GREB1.

To further explore this shared genetic architecture between menarche and menopause timing, we performed as similar approach, but using the 290 reported lead ANM variants[82] and conducting their lookup and colocalization in the menarche meta-analysis GWAS, described in **Section 6.2.1** (**Appendix Table 6.3**). Of the 290 ANM variants, 9 were genome-wide significant ($P \leq 5*10^{-8}$) and 23 passed the multiple test correction threshold ($P \leq 0.05/290 = 1.7*10^{-4}$) for age at menarche. Of these 32 common signals, 5 colocalized with and passed genome-wide significance for age at menarche ($P \leq 5*10^{-8}$), 9 colocalized and passed the multiple test correction threshold ($P \leq 0.05/290 = 1.7*10^{-4}$), and 2 colocalized and reached borderline multiple test corrected significance (total colocalized signals: 16, of which 5 were also identified by the above approach: *MPPED2*, *MSH6*, *BIRC6*, *INHBB* and *ANAPC4*) (**Figure 6.5**). Of these 16 variants that regulate both ends of the reproductive lifespan, 12 showed the same direction of effect on ANM and menarche timing, and the other 4 had directionally-opposing effects on menarche and ANM, hence extending/shortening reproductive lifespan. Most of the shared signals, which were not highlighted above, belonged to DDR and/or cell cycle mechanisms.

6.3.8 Shared genetic architecture with other reproductive health outcomes

After observing the shared genetic architecture between menarche and menopause timing, we were interested to further explore whether a similar pattern could be identified with other reproductive traits, including polycystic ovarian syndrome (PCOS), number of children and twinning. To do so, we performed a lookup of menarche 696 SNPs in mentioned traits, and identified 17 variants that were associated with 2 or more reproductive health outcomes (**Appendix Table 6.4**).

MPPED2, metallophosphoesterase highly expressed in the foetal brain and involved in the development of the nervous system[438], was the gene mapped to the variant with the highest number of associations with reproductive traits. These include menarche (P=9.4*10^{-25}), PCOS (P=6*10^{-11}), menopause (P=1.8*10^{-27}), and twinning (P=4.9*10^{-21}). All traits had the same direction of the effect besides twinning, where the age at menarche and menopause increasing variant conferred a smaller chance of having twins. *MPPED2* was also associated with the menstrual cycle length and cell proliferation[439,440]. Notably, *MPPED2* is located near the *FSHB* gene on the 11p13 chromosomal region, a well described risk factor for multiple reproductive trait outcomes, which could be mediating the association we observe at *MPPED2*. The GWAS signal at *FSHB* has been previously reported to influence menopause timing, PCOS and dizygotic twinning[441].

Additional shared signal tagged *WDR43* gene, known for its roles in ribosome biogenesis, was associated with later menarche (P=5.2*10^{-10}) and earlier ANM (P=4.6*10^{-7}), thus shortening reproductive lifespan. The same allele was also associated with higher number of children (P=1*10^{-5}). The dynamic regulation of ribosome biogenesis and global protein synthesis represents a relatively new and underexplored theme in the context of germ cell development, so future research should aim to decipher its importance for reproductive ageing.

Finally, a shared signal at rs4871939, which is correlated with a deleterious variant in *GNRH1*, had an effect on shifting overall reproductive window by having positive associations with both age at menarche (P=1*10^{-13}) and ANM (P=7.2*10^{-7}), while reducing chances for dizygotic twinning (P=6.7*10^{-10}). Loss-of-function mutations in *GNRH1* gene have been identified as rare genetic causes of normosmic Idiopathic hypogonadotropic hypogonadism (IHH), having an impact on the development and migration of GnRH neurons, the regulation of GnRH synthesis, secretion and action or gonadotropin cascades[442,443].

6.3.9 DDR regulates the aetiology of broad spectrum of health outcomes

Having observed novel evidence on the enrichment of DDR mechanisms for the regulation of menarche timing, we aimed to explore the implication of DDR for susceptibility of other 35 health outcomes, ranging from metabolic, cardiovascular, neurodevelopmental and reproductive origin (**Appendix Table 6.7**). To increase the statistical power of the hypothesis-free GWAS and reduce the burden of multiple testing, we applied MAGMA, a gene and gene-set analysis of GWAS genotype data, which determines the joint effect of multiple genetic markers. We specifically focused on the evaluation of the DNA repair, cellular response to DNA damage stimulus and cell cycle biological pathways due to their well-described role in the maintenance of genomic stability. We tested the enrichment of these pathways using the

default version of MAGMA, which combines association statistics across all types of genetic variants ('unrestricted'), as well as 'cMAGMA', a new adapted approach developed to strengthen the level of inference by considering only coding variants. The results from cMAGMA were considered as the primary ones. Menopause, LOY and cancer phenotypes were selected as positive controls due to previously described role of DDR in their aetiology (**Figure 6.7**). Our analysis sheds light on the enrichment of pathways of interest in a number of novel health outcomes (**Appendix Table 6.8**). .

Firstly, we confirmed the previously reported enrichment of DDR pathways in age at menarche using alternative DDR gene-sets and data - menarche GWAS associations from UKBB study were associated with *'Gene Ontology DNA Repair*' (P=4.4×10^{-4}) and an expert-customised DDR gene-set, named here as '*Broad DDR*' (P=1.0×10^{-3}). The enrichment of cell cycle (P=3.3×10^{-4}) and cellular response to DNA damage (P=2.0×10^{-3}) further highlighted the importance of genetic stability in regulation of menarche timing. Unlike menarche, it is interesting to note DDR was not associated with age at voice breaking (P=1.5×10^{-1}), a proxy of the puberty timing in boys, thus suggesting that these biological pathways could be acting via female-specific, and possibly ovary-specific mechanism.

We also observed enrichment of genetic associations for BMI and all three pathways of interest (DNA repair P: 4.0×10^{-3}, Cell cycle regulation P: 1.5×10^{-2}, Response to DNA damage P: 3.1×10^{-3}). This is in line with previous research that suggested impaired DNA repair as a potential underlying molecular mechanism of increased BMI[444,445], adipocyte metabolism and senescence[446,447]. Disruption of DDR has also been previously associated with cell growth abnormalities and severe intrauterine growth retardation, ultimately leading to rare phenotypes such as dwarfism, as in Seckel syndrome patients[448]. We also observed enrichment of genetic associations for adult height and DNA repair (P=5.4*10^{-4}), cell cycle (P= 5.7*10^{-7}) and response to DNA damage (P=5.1*10^{-5}) pathways. Finally, telomere length genetic associations were enriched for DNA repair (P=1.3*10^{-4}) and cell cycle (P=3.5*10^{-3}) , which were previously suggested in the literature in relation to ageing and longevity[449].

The enrichment of the cell cycle mechanism was observed in additional cardio-metabolic traits, including birthweight (P=9*10^{-4}), T2D (P=8.8*10^{-7}), waist hip ratio (WHR) adjusted for BMI (P=1.5*10^{-4}), diastolic (DBP) (P=5.2*10^{-3}), systolic blood pressure (SBP) (P=7.1*10^{-5}), and resting heart rate (RHR) (P=2.8*10^{-4}), as well as reproductive (uterine fibroids P=5.6*10^{-5}) and bone health (bone mineral density BMD P=8.3*10^{-6}).

These results demonstrate how the same DDR-related stimulus can yield markedly different responses in different cells and tissues. This opens up opportunities for better understanding and managing human health and disease where DDR could act as a potential marker of overall wellbeing.

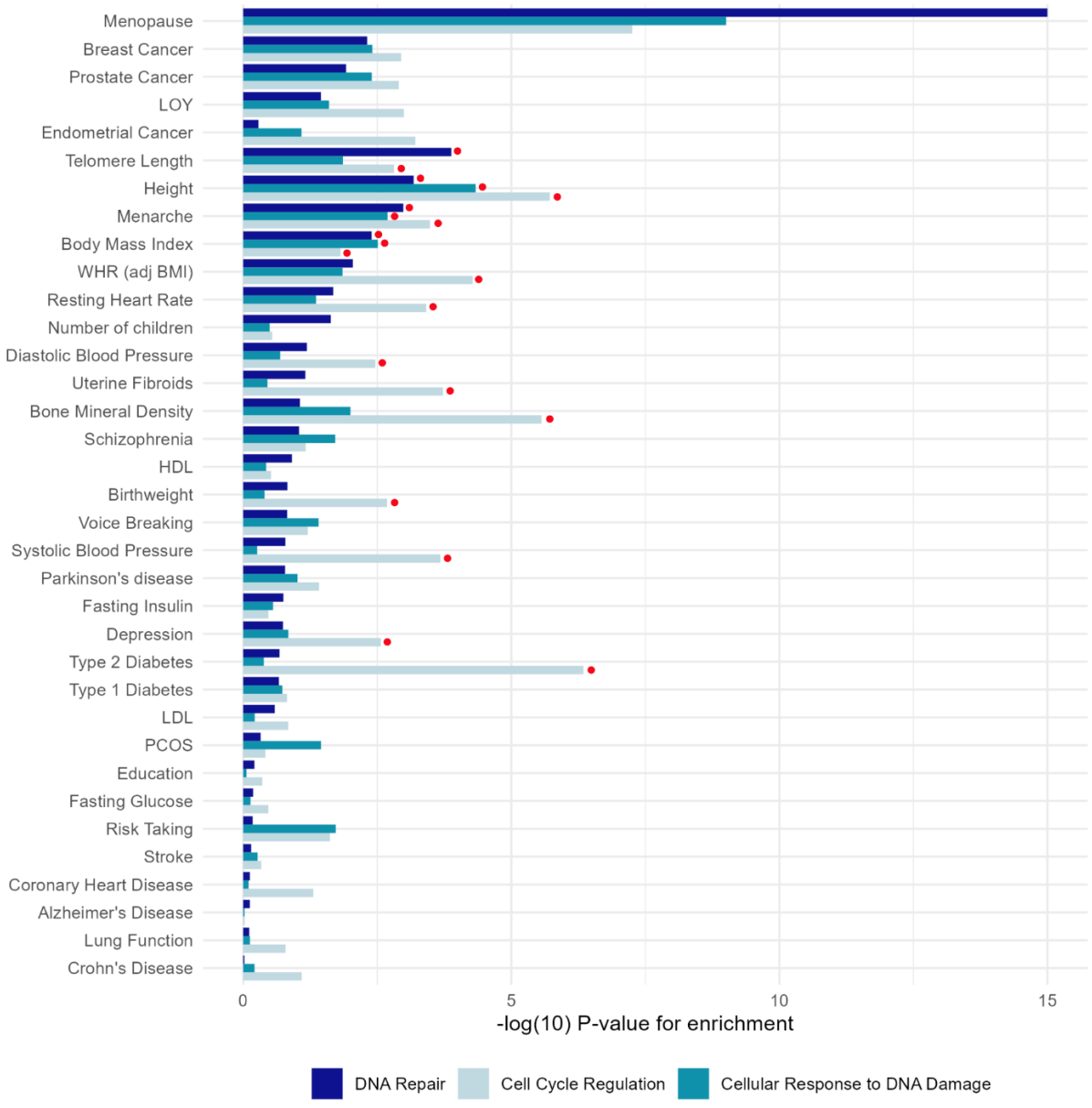

Figure 6.7: The enrichment of DDR-related and cell cycle mechanisms in 35 health outcomes. *The red dot indicates the significant results after the FDR correction. Menopause, LOY and cancer phenotypes were treated as positive controls here.*

Finally, we investigated the significantly associated genes from DDR-related pathways that are shared across multiple traits (**Appendix Tables 6.9 - 6.12**), by conducting MAGMA gene-levels analysis on predicted deleterious variants. The genes were scored according to the number of traits they were associated with. The gene associated with the highest number of health outcomes was *MAPT*, which encodes for Tau, a protein involved in the promotion of microtubule assembly and stability. Mutations in *MAPT* lead to frontotemporal dementia with Parkinsonism, where abnormal phosphorylation and folding cause Tau detachment from microtubules, Tau accumulation, and neuronal dysfunction[450]. Here we confirmed its genetic association with Parkinson's disease and found significant associations with 10 more health outcomes, including: schizophrenia, bone mineral density (BMD), breast and ovarian cancer, height, menarche, NEB, systolic blood pressure (SBP), voice breaking and WHR. The emerging function of Tau in DNA stability offers an alternative role of Tau in neurodegeneration and, importantly and insufficiently investigated, also in the DDR.

Another notable example is *PML*, which was genetically associated with six health outcomes: BMI, height, lung function, menarche, resting heart rate (RHR), and WHR, indicating that this gene is critical for both reproductive and metabolic health (**Appendix Tables 6.9 - 6.12**). *PML* regulates transcription, apoptosis, senescence, and DDR - it negatively affects the phosphoinositide 3-kinase (PI3K) pathway by inhibiting MTOR and activating PTEN, and positively regulates p53/TP53. It is a well-known tumour suppressor, regulated by oestrogen receptor beta (ERβ) signalling. The *PML*-ERβ network acts as a therapeutic axis by suppressing cellular survival and promoting cellular apoptosis in breast carcinoma[451]. Another example is *RTEL1*, a DNA helicase required for proper telomere replication and stability. Biallelic *RTEL1* mutations generate a large clinical spectrum ranging from classical Hoyeraal-Hreidarsson syndrome, a rare and severe telomere biology disorder characterised by intrauterine growth retardation, bone marrow failure, microcephaly and/or cerebellar hypoplasia, and immunodeficiency, to isolated aplastic anaemia[452]. Here we confirm the association with telomere length and provide evidence that *RTEL1* is also involved in susceptibility of Crohn's disease and regulation of high-density lipoprotein (HDL), low-density lipoprotein (LDL), diastolic blood pressure (DBP) and SBP. Significant associations identified for *MSH5* (with 6 health outcomes) and *BAG6* (with 11 health outcomes) could be biased by the location of these genes within the highly polymorphic human major histocompatibility complex (MHC).

6.4 Discussion

In a substantially enlarged genomic meta-analysis using data on ~566,000 women of European ancestry, we have identified 696 independent, genome-wide significant signals for age at menarche, increasing the total number of associated loci by two-fold. In aggregate these signals explained ~14% of the population variance in menarche timing, corresponding to ~29% of the estimated heritability.

To identify mechanisms responsible for menarche timing, we implemented a genome-wide MAGMA pathway analysis approach and provided the first evidence on the enrichment of DDR mechanisms in the regulation of age at menarche. While the role of DDR has not been previously demonstrated as important for menarche timing, these pathways have been clearly highlighted by previous GWASs for age at natural menopause - indeed almost ⅔ of GWAS ANM signals can be mapped to DDR-related genes[82]. It has been believed that the overall enrichment in pathways regulating these two extremes of reproductive lifespan clearly differs. Menarche timing has thought to be driven by loci located in/near genes that regulate the hypothalamic-pituitary reproductive axis, with menopause being mainly determined by the genomic stability of oocytes and hence the size of the ovarian pool. However, it seems that the role of DDR extends further from just the determinant of the onset of menopause within the reproductive axis, with evidence of its involvement in regulation of both beginning and end of reproductive lifespan.

We then aimed to understand what specific genes might be involved in the regulation of this wide spectrum of reproductive longevity. A primary challenge in obtaining biological insights from GWASs arises from the inability to directly implicate causal genes and the mechanisms involved from association data[453]. Deciphering how associated variants modulate disease risk and severity, and how they impact cellular phenotypes provides a mechanistic insight essential for effective predictive and therapeutic solutions. This is especially critical when translating GWAS findings into the functional experimental setting for validation in cellular and animal models. These approaches are rapidly expanding in scale and scope. However, there are substantial limitations to the scale and cost of genome perturbation and cellular phenotyping, the availability of good animal model systems as well as the ethicality for using those to investigate hundreds to thousands of genetic variants discovered through GWAS[185]. Numerous strategies, including statistical methods and genomic functional annotations[185,453], have been extensively applied to prioritise causal variants and their target genes. Accelerating these efforts could be achieved through an integration of evidence from multiple individual data sources, which, when combined, could potentially provide a better powered source of information on functional links between associated variants and candidate genes. Here, I applied a novel tool developed in our group, G2G, which takes advantage of a variety of such data, encompassing information on coding variants, the nearest gene, expression and

protein quantitative trait loci, and provides a solution for the gene prioritisation. Using this approach, we prioritised 509 genes from the 696 menarche genomic loci. As the top scoring novel menarche gene we identified *YWHAB*, a DDR gene that was previously associated with multiple cancer outcomes, such as melanoma, cervical, prostate and lung cancer, as well as with recurrent fertilization failure[454], idiopathic pulmonary arterial hypertension[455], and postmenopausal osteoporosis[456]. Other top scoring genes were involved in hormonal regulation, such as LH and FSH, which secretion represents the hallmark of sexual maturity and functioning[457,458]. Besides hormonal regulation, we also identified genes involved in metabolism, immunity, ribosomal biogenesis, growth hormone axis and notably DDR. Some of these DDR-related genes, including *MSH6*, *DDIT4*, *GLI2* and *DET1*, had the highest or exclusive expression in the ovaries, thus providing the first suggestive evidence on DDR regulating menarche timing via the effect on ovaries - an effect which is potentially independent of hypothalamic axis. We speculate that the body has a sense of the level of depletion of oocytes prior to puberty, and perhaps starts the reproductive window earlier on the basis that it will end earlier. A notable example includes *FOXO3*, which was identified to act as a master regulator and suppressor of primordial follicle activation, controlled via PI3K signalling pathway, the first such factor to be defined[459]. It seems, surprisingly, that the only essential role of a negative regulator of PI3K, *Pten*, within the oocyte is to regulate *Foxo3*[459], suggesting the high importance of this gene. In *Foxo3* knockout mice, primordial follicles are assembled normally[408] but then immediately undergo global activation, resulting in a distinctive syndrome of ovarian hyperplasia, follicle depletion before puberty, premature ovarian failure, and infertility[411]. Moniruzzaman *et al* (2010)[460] suggested that the mechanism regulating the activation of primordial oocytes is different in the pre-pubertal stage compared to infancy and observed a differential expression of *FOXO3* in infancy and pre-puberty. In infant pigs *FOXO3 was* detected in 42±7% primordial oocytes, while almost all (94±2%) primordial oocyte nuclei were *FOXO3* positive during the pre-pubertal stage[460]. Primordial oocytes in prepubertal pigs took much longer time to initiate growth than did those in infants, suggesting a mechanism that aims to preserve the oocyte pool as much as possible before the start of reproductive activity via inhibition of the primordial oocyte activation. The observed differential expression and the role of this DDR gene in regulating menarche timing might be the first concrete evidence of the importance of the genomic stability of oocytes not only for the reproductive longevity, i.e. menopause timing, but also for the initiation of the reproductive activity. We speculate that *FOXO3* senses the depletion and thus might serve as a mechanism that sends signals to the ovaries when the size of the pool is reducing significantly, thus initiating puberty earlier to ensure long-term fertility before the pool is fully depleted. Finally, the mouse studies demonstrated that overexpression of constitutively active *FOXO3* can increase ovarian reproductive capacity by 31-49% via the increase in follicle numbers[461]. Importantly, due to the significant association of *FOXO3* with BMI that we identified in this Chapter, it would be

necessary to better understand mechanisms that *FOXO3* is using to modify the susceptibility to mentioned phenotypes and decipher if its function in menarche and menopause timing takes place via ovary-specific mechanism or it is mediated by BMI.

Although heritable determinants of many common reproductive traits have been studied, not many efforts have been focused on understanding the degree of shared genetic architecture underlying reproductive health. The knowledge on shared biological mechanisms between various reproductive traits could have important benefits, especially related to development of public health interventions and/or fertility treatments and preservation. Here, we implemented the largest GWAS studies for age at menarche, menopause, PCOS, NEB and twinning to explore common regulatory mechanisms. Some of these shared loci were located in/near the genes involved in DDR, supporting our previous observations and further extending it to other reproductive health outcomes. In addition to DDR, we highlight ribosome biogenesis, which represents a relatively new and underexplored mechanism in the context of reproductive ageing. Emerging evidence points towards the global regulation of mRNA translation as an important mechanism for germ cell development, infertility and reproductive ageing. For example, ribosome biogenesis is often regulated in a stage specific manner during gametogenesis. Moreover, oocytes need to produce and store a sufficient number of ribosomes to support the development of the early embryo until the initiation of zygotic transcription. In addition, specific translation initiation and elongation factors are also enriched and regulated in the germline[462]. These findings highlight the importance of gaining further insights into how ribosomes and the translation machinery work together during the development and ageing of the oocyte.

Having observed novel evidence on the enrichment of DDR mechanisms for the regulation of menarche timing, we explored the implication of DDR and related mechanisms for susceptibility of other 35 health outcomes, ranging from metabolic, cardiovascular, neurodevelopmental and reproductive origin. Maintenance of genome stability is essential to healthy human physiology, with unrepaired cellular DNA damage being implicated in the aetiology and progression of different types of human pathologies[463,464]. Much of the current research in DDR is devoted towards understanding the mechanisms and its biological implications in cancerogenesis, immunodeficiencies, longevity and rare hereditary diseases with severe developmental problems[464–466]. However, not many efforts have been focused on deciphering the role of DDR across the wide spectrum of human phenotypes. In addition to reproductive ageing that was previously discussed, research conducted in our group suggests that loss of chromosome Y (LOY), the most frequent age-related somatic change in leukocytes associated with various later life diseases, represents a powerful marker of DDR[467]. However, the plethora of DNA damage, response and repair processes, along with their profound and complex interactions across a broad spectrum of prevalent

human health outcomes, have not yet been fully elucidated. In addition, there is a lack of understanding whether disturbed DDR capacity associated with reproductive ageing and LOY is oocyte and leukocyte-specific, or if it is present in other cell types and tissues as a potential marker of broader genomic instability and disease susceptibility. Here, we provided the first population genomic insights into the role of DDR-related mechanisms in various anthropometric, metabolic and reproductive health outcomes. Notable examples include BMI, T2D, and height associations. Previous functional studies suggested that DDR is involved in regulation of metabolic homeostasis. DNA damage could impair metabolic organ functions by causing cell death or senescence. Accumulation of senescent cells could impair tissue regeneration and homeostasis, leading to metabolic dysfunction. In addition, accumulation of senescent cells in the tissues leads to chronic inflammation mediated by various proinflammatory cytokines and chemokines, accelerating disease progression[468–470]. There is accumulating evidence that impaired DNA repair and accumulative DNA damage together with chronic inflammation associated with senescence have a pivotal role in the progression of age-related diseases such as diabetes and cardiovascular disease[471–473].

T2D is a complex metabolic disease characterised by an insulin resistance, i.e. deficient insulin secretion by pancreatic β-cells[474]. The adult β-cells maintain their numbers through self-replication, a low frequency process in these cells[475]. As a consequence β-cells become more vulnerable to a variety of stressors, such as DNA damage, oxidative stress and glucolipotoxicity, which leads to a reduction in β-cell mass and impaired cell function[476–479]. Genome-wide association studies, which contributed to the identification of more than 500 T2D-associated genetic variants, allowed insight into the genetic architecture of this disease, and highlighted the importance of DDR mechanisms[480]. More specifically, recent findings from our group gave evidence on the association between LOY and T2D. Shared risk loci in these two traits were involved in cell cycle regulation and DNA repair. They tagged genes that encode cyclins and cyclin-dependent kinases, known to be involved in pancreatic β cell growth and maturation (such as *CCND2*, *CDKN1B*), and genes regulating apoptosis (*TP53INP1*)[481]. This suggests that altered cell cycle regulation and genomic instability, which lead to increased clonal mosaicism, likely modify risk of T2D through higher cell death, i.e. reduced number of pancreatic β cells. Preliminary genetic findings are also supported by evidence from *in vivo* experiments in NHEJ-p53R172P mutant mouse model characterised by deficient non-homologous end-joining (NHEJ), which result in DNA double-strand breaks (DSB), and *p53* deficiency. Combined DSBs with an absence of p53-dependent apoptosis activate p53-dependent senescence, causing a diminished cell self-replication, decrease in pancreatic islet mass, and severe diabetes[482].

Similar findings are observed in case of obesity where DNA damage contributes to adipose tissue inflammation[444]. Recent studies shed light on the impact of adipose tissue depot-specific regulation of DDR genes linking DNA methylation to lipid metabolism and fat distribution[445]. The mechanisms that cause senescence in adipose depots and the roles of senescent cells in obesity are unclear. Lee *et al* (2022) made first steps toward deciphering these mechanisms by analysing transcriptomes of adipocytes from mice fed either a normal chow or a high-fat diet, and of senescence models. The authors identified that sterol regulatory element-binding proteins (SREBPs) protect adipocytes from genome instability and senescence, which contribute to inflammation and insulin resistance in obesity. SREBPs are master regulators of cholesterol and fatty acid metabolism and promote the activity of the DNA repair enzyme *PARP1*. Specific phenotypic analyses showed that, without SREBP1c, adipocytes accumulate DNA lesions and are prone to senescence, which contributes to white adipose tissue inflammation in obesity, while the elimination of senescent cells in SREBP1c-null mice improved insulin resistance[446]. Finally, previous research suggested that a common molecular feature shared by the prenatal growth retardation phenotypes is that they are caused by mutations in genes that are necessary for an appropriate response to DSBs during the S phase of the cell cycle - here, our evidence suggest that DDR also regulates the normal variation in height[448].

Confirmation of these findings requires a significant amount of future work, however they give the first indication of DDR acting as a marker of broader health outcomes. If this is true, the effect of DDR on ovarian ageing and reproductive longevity could really act as a proxy of overall wellbeing of a woman. Future studies should systematically explore the potential influence of identified DDR genes and associated variants on different health outcomes. This specifically involves assessing the direction of the effect on different traits and identifying biomarkers that could serve as basis for the development of potential intervention and treatment strategies. In summary, our findings highlight a novel evidence on the role of DDR mechanisms in regulating both menarche and menopause timing, and provide an insight into DDR as potential marker of health and disease.

CHAPTER 7

Novel functional insights into the role of *PARP1* in gametogenesis and reproductive ageing

Summary

The thesis aims to further explore the relevance of DDR mechanisms in reproductive ageing, taking a gene-centric approach in this Chapter. Previous studies in mice demonstrated that inhibition of *PARP-1* is associated with the reduction of ovarian reserve, highlighting *PARP-1* codon 762 variant (V762A) that is well-known to reduce the *PARP-1* catalytic activity by 30-40% and is present in about 5-30% of the general population. In addition, human genomic evidence demonstrated the protective role of this variant against LOY in men, suggesting its importance also for human health and disease and making this gene an attractive candidate for further exploration. Our human genomic meta-analysis of ReproGen Consortium and 23andMe data provide the first indicative evidence on association between a *PARP-1* V762A variant and reduced age of natural menopause (beta: -0.06 years per allele [0.04-0.09], $P=5.3*10^{-8}$) in women. As the concept of *PARP-1* inhibition represents the basis of the treatment of breast and ovarian cancer patients, better understanding of its impact on reproductive longevity is crucial. Combining human genomic evidence with cutting edge CRISPR technology and the in *vitro* gametogenesis (IVG) system, we investigated the role of *PARP-1* in ovarian function, more specifically in proliferation of primordial germ cells during the establishment of the ovarian reserve. We demonstrate for the first time that deletion of *PARP-1* increases the efficiency of primordial germ cell production *in vitro* via upregulation of *Oct4*, which could be driving their self-renewal. We speculate that, even though there is an initial increase in primordial germ cells, the quality of these cells could be compromised. This could expose them to substantial 'clearance' via cell death mechanisms at later stages of gametogenesis, ultimately leading to the creation of smaller ovarian reserve.

Contributions and Collaborations

I performed the computational analysis of the *PARP-1* V762A variant in the ReproGen consortium and 23andMe study. When it comes to the functional work, BVSCH18 mouse embryonic stem cell line bearing the *Blimp1-mVenus* & *Stella-ECFP* (BVSC) reporter construct was generated by Mitinori Saitou, and was gifted to the Hoffmann lab by Bernard De-Massey. Hannah R. Schorle generated the *PARP-1* knockout cell line. I performed the cell culture establishment, discovery and the first replicate analysis of the wild type and *PARP-1* KO differentiation, from mESCs to PGCLCs, using *In vitro* gametogenesis system. I conducted PGCLC purification and FACS sorting. Both Dr Halliwell and I conducted RNA extraction, reverse transcription and RT-qPCR independently. However, the reported results for the qPCR section in this Chapter come from the experiment where I conducted differentiation, purification and FACS sorting, while Dr Halliwell conducted qPCR analysis.

This project was supervised by Professor Eva Hoffmann and Professor John Perry. My day-to-day laboratory work was supervised by Dr Jason A. Halliwell and Amy V. Kaucher. Professor Perry, Dr Halliwell and Ms Schorle provided valuable advice on the analyses and writing.

7.1 Introduction

The past decade has seen a great progress in identification of genetic determinants of reproductive longevity and timing of menopause through human genomics. The GWAS approach has shed light towards biological mechanisms that act across stages of gametogenesis to determine an individual's reproductive lifespan[481]. However, the functional role of unique genes and underlying molecular mechanisms through which they operate remain to be elucidated. This becomes critical as early detection of these pathological changes and their thorough functional understanding allow us to pave the path towards prevention and delivery of therapeutic solutions. Similar issues could be observed in case of variants and genes that were initially discovered in cell and animal models, which lack confirmatory evidence on their role in human health and disease. Combining an 'omic' approach with functional models therefore represents unique way towards deciphering the genetics underlying phenotypes of interest and enables us to deliver robust findings by addressing potential differences in physiology between functional models and humans.

Being driven by the functional evidence from the literature, this Chapter aims to follow a gene-centric approach to study the role of Poly(ADP-Ribose) Polymerase 1 (*PARP-1*) in gamete formation and ovarian function. *PARP-1* is known to facilitate the maintenance of genomic integrity. This involves roles in DNA replication and repair, transcription control, chromatin organisation, as well as cell proliferation and death (**Figure 7.1**)[483–487]. *PARP-1* catalytic activity remains at low basal levels until it is strongly stimulated as a response to single- or double-strand breaks, thereby recruiting the repair machinery. When DNA damage is extensive, activation of *PARP-1* can lead to either necrotic or apoptotic cell death[484,488–495].

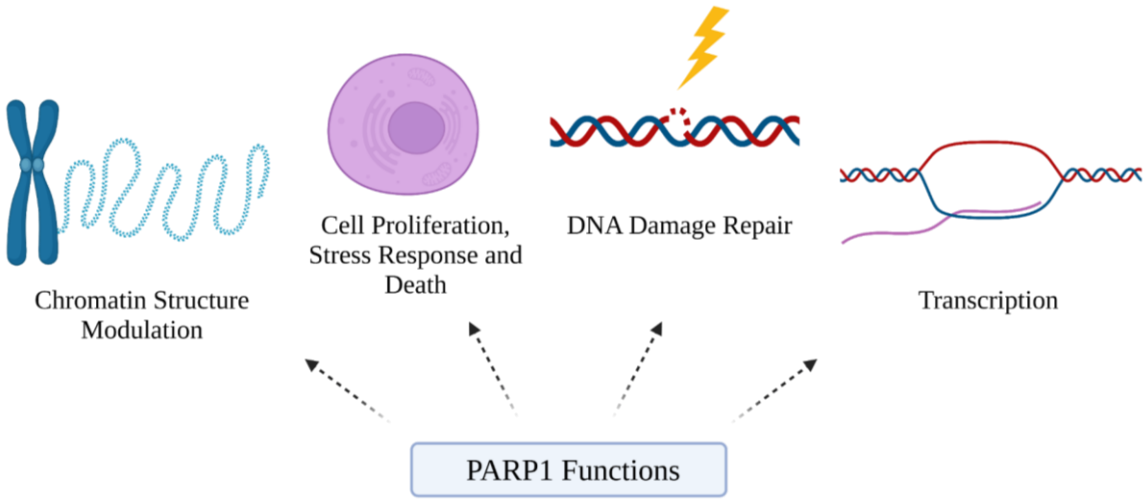

Figure 7.1: Schematic representation of various roles of PARP-1.

Previous clinical and epidemiological studies have shown large inter-individual differences in *PARP-1* activity within both healthy and cancer patient cohorts, yet the potential genetic mechanisms that modify this activity are still not completely understood[496]. Most of the SNPs discovered so far in the *PARP-1* region are rarely found in the general population, having an allele frequency less than 1%[497,498]. On the contrary, missense variant rs1136410 introduces an amino acid change V762A in the catalytic domain and is present in about 5-40% of the general population. It is the rarest in African/Americans (~4.9%), more common in Europeans (~23%) and Asians (~20.7%), and present in almost half of the Latino/Admixed American cohort (~42%)[164,489].

Existing evidence from various case-control studies point to the association between the minor allele (A>G: G *PARP-1*-Ala) and the reduction in the catalytic activity of *PARP-1*[489]. More specifically, *in vitro* experiments on purified *PARP-1* enzyme showed that rs1136410 minor allele reduced *PARP-1* activity by 30-40% in comparison to *PARP-1*-Val[496], the effect which was consistent both in the presence or absence of *PARP* inhibitor, 3-AB[489]. The reduced catalytic activity was explained through an increase in *PARP-1*-Ala polymorphism kinetic activity (Km), which was 1.2 fold higher compared to the Km of *PARP-1*-Val[489]. *PARP-1* V762A genetic alteration may confer protection against coronary artery disease (CAD) with up to 84% decreased CAD risk[499], yet the evidence on the association with cancer phenotypes have been largely contradictory, suggesting high variability across different ethnicities. The association with higher susceptibility of cancer risk has been observed for prostate, oesophageal, lung, thyroid, brain and cervical cancer[496,500–507]. The suppression of *PARP* activity can also have a profound effect on chemotherapy-induced toxicity, as well as the efficacy of chemotherapy, and this has been a base for the development of Olaparib, a first-in-class oral *PARP* inhibitor used to treat patients with breast and ovarian cancer caused by *BRCA* mutations. Olaparib inhibits *PARP-1/2* enzymatic activity and traps *PARP1* on DNA at single-strand breaks. This leads to the replication-induced DNA damage that requires *BRCA1/2*-dependent homologous recombination repair, which is deficient in these patients, and ultimately cell death[508–511]. However, cancer therapy can cause off-target effects, including ovarian damage accompanied by accelerated loss of follicles. This may result in impaired fertility leading to premature ovarian failure in girls and premenopausal women. Indeed, recent studies based on the mouse model have indicated a potential role of *PARP-1* in late gametogenesis. More specifically, it has been shown that Olaparib destroys a significant proportion (36%) of the immature eggs that are contained within primordial follicles but not other follicle classes, suggesting an impact of the inhibition on the ovarian reserve rather than growing oocytes (**Figure 7.2 A,B**)[110]. There is a contradictory evidence that demonstrates that Olaparib does not only reduce the ovarian reserve but additionally affects the growing follicles, including total

primordial, primary, early secondary and late secondary follicles, augmenting the number of atretic follicles (**Figure 7.2 C,D**)[512].

Finally, besides the experimental evidence in the mouse model on the importance of the V762A *PARP-1*, our recent study demonstrated the first human genomic evidence on the protective role of this variant against LOY in men[481], suggesting its importance also for human health and disease.

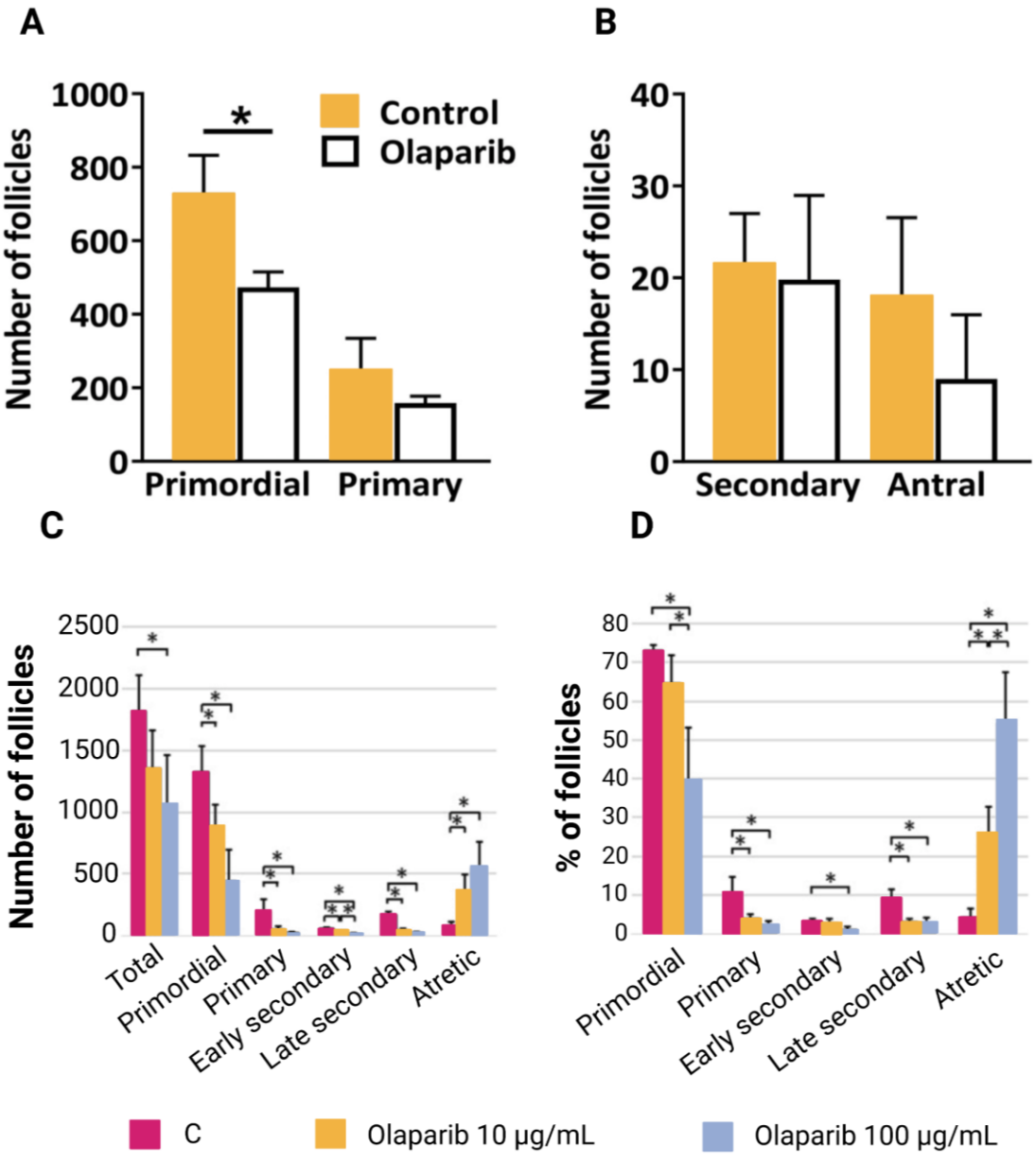

Figure 7.2: PAPR-1 inhibitor Olaparib depletes the primordial and growing follicle pool in mice. *Panel* ***(A)*** *and* ***(B)*** *belong to Winship et al (2020) study*[110] *and show that Olaparib treatment significantly decreases only the primordial oocyte pool* ***(A)****, but not the primary, secondary and antral follicles* ***(A and B)****. Panel* ***(C)*** *and* ***(D)*** *belong to Nakamura et al (2020) study*[512] *and demonstrate the in vitro follicle dynamics with or without Olaparib treatment, including the quantification data on total, primordial, primary, secondary and atretic follicles. The data were*

*obtained from three ovaries per each group. Panel **(C)** represents the number of follicles, while distribution of follicles (%) is shown in panel **(D)**.*

Despite the evidence obtained from animal models on *PARP-1* inhibition being toxic to ovaries, there is no preclinical or clinical information regarding potential impact on female fertility. It may be many years before clinical data on fertility outcomes for women treated with *PARP* inhibitors become available. This highlights the importance and urge for using robust human genomics data and conducting rigorous research on animal and cell models to understand potential reproductive outcomes. Importantly, better understanding of *PARP-1* mechanism in ovarian function does not only relate to the inhibition induced by external cancer therapeutics, but also to the germline mutations that naturally decrease *PARP-1* activity and potentially affect female fertility, as in the case of V762A. Available evidence points towards inefficient DDR mechanism as a potential cause of follicle depletion due to the fact that *PARP-1* is required by the base excision repair (BER) for the primordial germ cell (PGC) survival and proliferation (**Figure 7.3 A**). PGCs undergo significant global demethylation (**Figure 7.3 B**), also orchestrated by the BER pathway, to erase parental imprints ensuring faithful transmission of the genome between generations, which additionally highlights the importance of DDR in PGCs[513–519]. Therefore, we hypothesise that this inefficient repair of DNA damage due to *PARP-1* inhibition drives damaged PGCs into apoptosis, ultimately diminishing the ovarian reserve (**Figure 7.3 A**)[94].

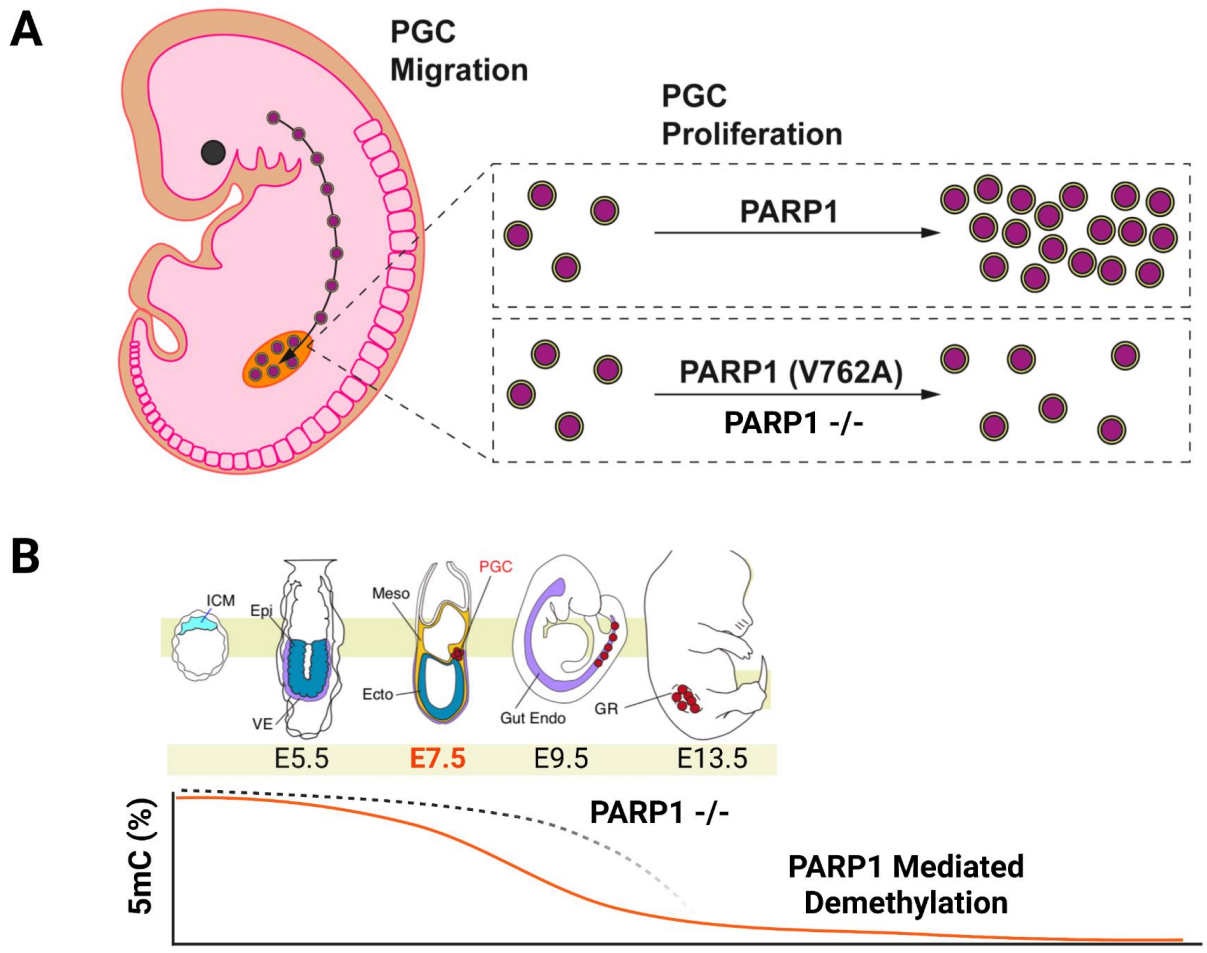

Figure 7.3: Base Excision Repair and germ cell reprogramming and survival. *The figure summarises scientific hypotheses that will be explored in this Chapter. Experiments related to panel* ***(A)*** *will aim to decipher the process of PGC proliferation affected by the inefficient PARP1-1 function, while panel* ***(B)*** *will investigate the role of PARP-1 in global demethylation. In germline, a progressive dilution of DNA methylation can be observed at around embryonic day 7.5 (E7.5) during PGC specification, while at the PGC expansion (from E9.5) a global methylation takes place, reaching the lowest DNA methylation levels at E13.5 through the germ-line cycle. Panel* ***(B)*** *is adapted from Kurimoto and Saitou (2018)*[520].

The female germ line undergoes a unique sequence of differentiation processes during gametogenesis that we can replicate *in vitro* to reconstitute potent mature oocytes from pluripotent stem cells[517,521]. This state-of-the-art technology, named *In Vitro* Gametogenesis (IVG), successfully mimics the whole gametogenesis cycle. This cycle could be divided into four stages *in vitro*: (1) formation and migration of PGCs, (2) *in vitro* differentiation (IVDi), (3) *in vitro* growth and (4) *in vitro* maturation (IVM), in which oogenesis would proceed to primary oocytes in the secondary follicle, fully grown germinal vesicle oocytes and metaphase II (MII) oocytes, respectively (**Figure 7.4**)[521,522]. The experiments described in this Chapter are restricted to only stage (1). Due to the scarcity of the reproductive tissue that challenged

our functional understanding, this culture system provides a unique platform for studying *PARP-1* and mechanisms that govern gametogenesis and ovarian function[523–525].

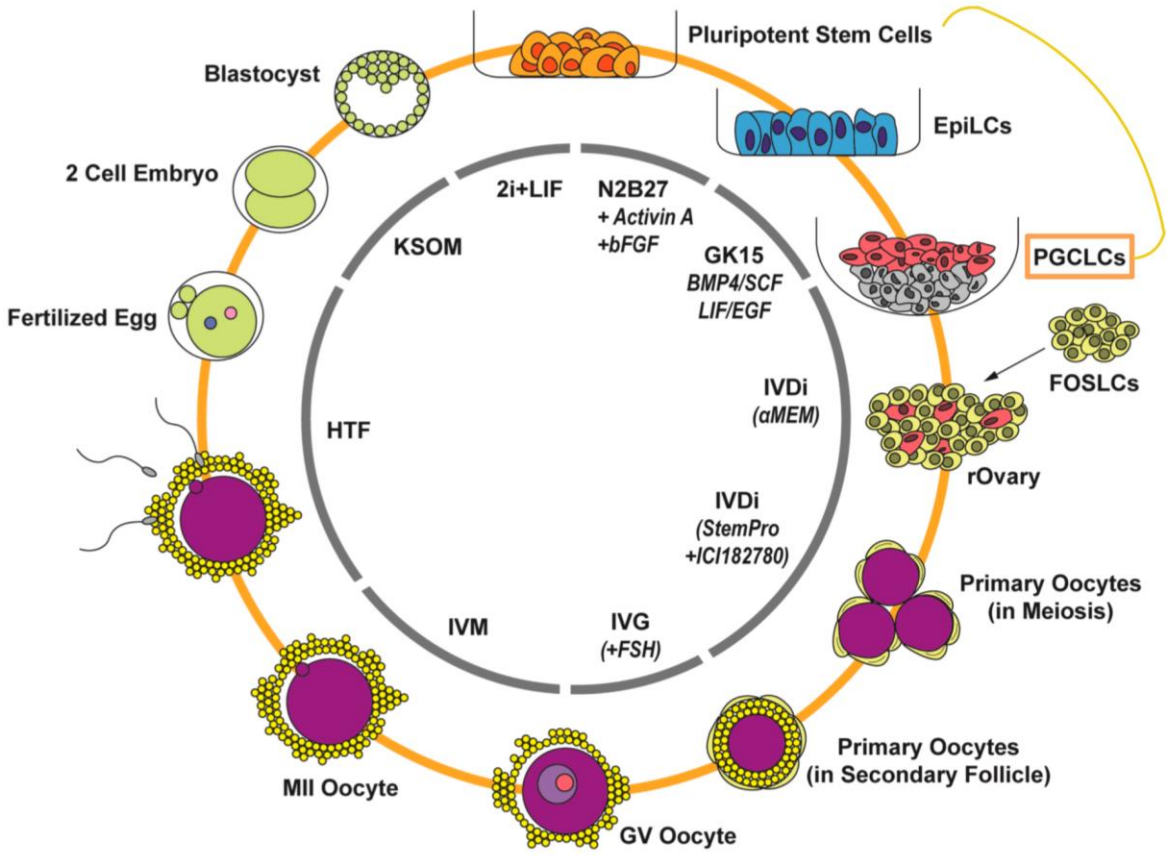

Figure 7.4: A schematic showing the reconstitution of the entire female germ line in vitro. *In the mammalian embryo pluripotency is established from the epiblast in the inner cell mass (ICM) of the preimplantation blastocyst. Eggs originate from primordial germ cells (PGCs), which are specified at around embryonic day 6.5 in mice*[526]*. PGCs then migrate into the gonads, enter meiosis in female embryos*[527] *and therefore become primary oocytes. Following puberty, primary oocytes begin to grow to mature oocytes that are fully ready for fertilisation. This Figure depicts the equivalent of this process in vitro, consisting of 4 stages as described above: PGC formation and migration, IVDi, IVG and IVM. This thesis will only address the first stage where mouse embryonic stem cells (mESCs) are first differentiated into epiblast-like cells (EpiLCs), and then induced to form primordial germ-like cells (PGCLCs).*

Using the largest, population-scale human genomic data, this Chapter will for the first time ever explore the evidence on the role of *PARP-1* V762A missense variant in reproductive ageing and menopause timing in women. Human genomic work will be followed up by thorough functional analysis that will specifically model and investigate the role of *PARP-1* during early stages of the establishment of ovarian reserve, i.e. at the PGC level. Using CRISPR technology to generate a *Parp-1* knock-out in conjunction

with the IVG culture system, I will explore the efficiency of PGC proliferation during ovarian reserve establishment in the absence of *PARP-1*.

7.2 Methods

7.2.1 Human genomic evidence on the role of *PARP-1* in ovarian ageing

In order to investigate the role of the *PARP-1* V762A missense variant 1:226555302:A:G (rs1136410) in ovarian function in humans, we utilised large-scale population human genomic data that were available through the ReproGen Consortium and 23andMe Inc. study. Details of the individual studies are provided below.

The ReproGen GWAS on ANM in ~200,000 women of European ancestry is described in detail in **Chapter 2**, **Section 2.2.2**.

The 23andMe study: The data were collected as part of the customer base of 23andMe Inc. (Mountain View, CA, USA), a personal genomics company, which provides direct-to-consumer genetic testing. All participants provided informed consent and answered online surveys following 23andMe's human subjects protocol, which was reviewed and approved by the external AAHRPP-accredited IRB, Ethical & Independent Review Services (E&I Review), a private institutional review board (http://www.eandireview.com). Direct-to-consumer process involves participants providing saliva samples, which are then processed by the 23andMe research team. Derived genetic reports are made available to consumers informing them on their personal ancestry and health profile. The DNA processing was performed by the National Genetics Institute (NGI), a Clinical Laboratory Improvement Amendments (CLIA)-licensed clinical laboratory and a subsidiary of Laboratory Corporation of America. The methodology behind the DNA extraction and genotyping has been previously described in detail[528]. The variant-level data for the 23andMe dataset is disclosed in this Chapter. Individual-level data are not publicly available because of participants' confidentiality, and in accordance with the IRB-approved protocol. To record women's age at menopause multiple surveys were conducted and responses were used to estimate ANM:

- "About how old were you when you had your last menstrual period? (under 30/ 30-34/ 45-49 / 40-44 / 55+ / 50-54 / 35-39 / declined / not sure)"
- "How old were you when you had your last menstrual period?"

As menopause age was ascertained in 4-year bins, the effect estimates were appropriately rescaled to be on the same 1-year as our ReproGen discovery analyses described in **Chapter 2**. We applied a linear model (Gaussian) to perform the genetic association analyses, and controlled for age (in years), the top five genetic principal components and genotyping platform. The information on ANM was obtained on 294,828 women of European ancestry.

We then meta-analysed two studies using METAL (https://genome.sph.umich.edu/wiki/METAL_Documentation). The variant lookups were performed based on the chr, variant position and minor and major alleles to extract the variant association summary statistics with the ANM phenotype.

7.2.2 Using IVG to study ovarian function in *PARP-1* mutants

7.2.2.1 Animals and derivation of ESCs

To conduct *In Vitro* Gametogenesis (IVG) experiments we obtained the mouse embryonic stem cells (mESCs) from Hikabe *et al*[521]. The protocol for the mESCs derivation and animals used is described in detail elsewhere[521]. In short, the BVSCH18 ESC line bearing the *Blimp1-mVenus* & *Stella-ECFP* (BVSC) reporter construct was established from the blastocysts collected from independent pairs of 129+Ter/svj (agouti) females and C57BL/6 BVSC males[529]. The mESCs used in this study were derived from the female mouse to specifically study female germ cell development. The reporter *Blimp1* was used as a marker of a PGC, and *Stella* as a PGC and germline marker. The maintenance of the mESC line was done under 2i condition, prepared in batches of 50 mL and stored at 4°C: 5uL PD0325901 (10 mM in 1 mL DMS), 500uL L-Glutamine, 0.63uL Monothioglycerol, 50uL LIF and 50uL CHIR99021 [3mM] were added to a 50mL aliquot of SFES basal media containing Neurobasal, DMEM F12, N-2 supplement, B27, 7.55% BSA and Penicillin-streptomycin solution[530].

7.2.2.2 Generation of the *PARP1* knockout

PARP1 knockout (KO) cell lines were generated using CRISPR-Cas9 gene editing (**Figure 7.5**). Predesigned guides were bought from IDTDNA, targeting exon 3 and exon 4 (**Table 7.1**). Cells were transfected with the Cas9 and guide complex using the Neon Transfection Kit as described in: IDTDNA Perform gene knockout in your research[531] and Alt-R CRISPR-Cas9 System[532].

Figure 7.5: Gene knockout workflow. *The figure is obtained from [IDTDNA Perform gene knockout in your research]* [531].

Table 7.1: Guides used to target PARP-1.

Guides	Sequence
Guide 1	CGGAGTACGCCAAGTCCAAC
Guide 3	CCGGCAGCCTGATGTTGAGG

7.2.2.3 PCR and Sanger sequencing

The individual clones were picked for genotype analysis by polymerase chain reaction (PCR) and Sanger sequencing. DNA was isolated using the DNeasy Blood and Tissue Kit (QIAGEN). The real time-qPCR was carried out in triplicate on a CFX96 machine (Bio-Rad) using the SYBR Green Mix, with the following conditions: initial denaturation at 95 °C for 10 min, then 40 cycles of 95 °C for 10 s, 60 °C for 10 s and 72 °C for 20 s. The primers are listed in **Table 7.2.**

Table 7.2: PCR primers. *The table shows the forward and reverse sequence of PCR primers with the primer melting temperature (Tm) stated.*

Exon	Primer	Sequence	TM
3	Forward	CATTATCTAGGTCCCGTTCCTTATT	64
3	Reverse	GCTCTCGTGTTTCTCTCAGTT	
4	Forward	CTAGGCTGTGCAGTGGAAATTA	65
4	Reverse	GCTCTTTCTTGGGAGGTAGTAG	

After amplifying exon 3 and 4 using the primer sequences above, the PCR products were purified using the Qiagen PCR purification kit (28104) and sent to Macrogen for Sanger sequencing to confirm the precise mutant sequence of each allele. The forward primers were used as sequencing primers.

7.2.2.4 Western blot analysis

After confirming mutation at the genome level, western blot was used to analyse gene KO at the protein level. To perform protein extraction, cells were lysed by RIPA (Sigma Aldrich) with 10x Phosphastop and Proteinase Inhibitor (500 000 cells in 100µl Cell lysis buffer). After that, 4x sample buffer (1800 µl LDS + 200µl DTT) was added to the sample and further incubated at 95℃ for 5 minutes. Sodium dodecyl-sulfate polyacrylamide gel electrophoresis (SDS-PAGE) Precast Gel (Invitrogen) was used for electrophoresis. *PARP-1* antibody (ab191217, rabbit, Invitrogen) was used to detect the presence of *PARP-1* protein. The antibody was diluted in Primary Antibody Dilution Buffer (Sigma Aldrich) at a density of 2 µg/mL, and incubated with the sample overnight at 4℃. After washing with TBST three times, samples were incubated with a Secondary antibody (Goat anti-rabbit), which was diluted in Secondary Antibody Dilution Buffer (Sigma Aldrich) at a ratio of 1:1000, at room temperature for 1 h.

7.2.3 Primordial Germ Cell-like Cell (PGCLC) differentiation from mESC

PGCLCs were differentiated from mESCs as previously stated[521,530]. A brief methodology description is available in the following sections.

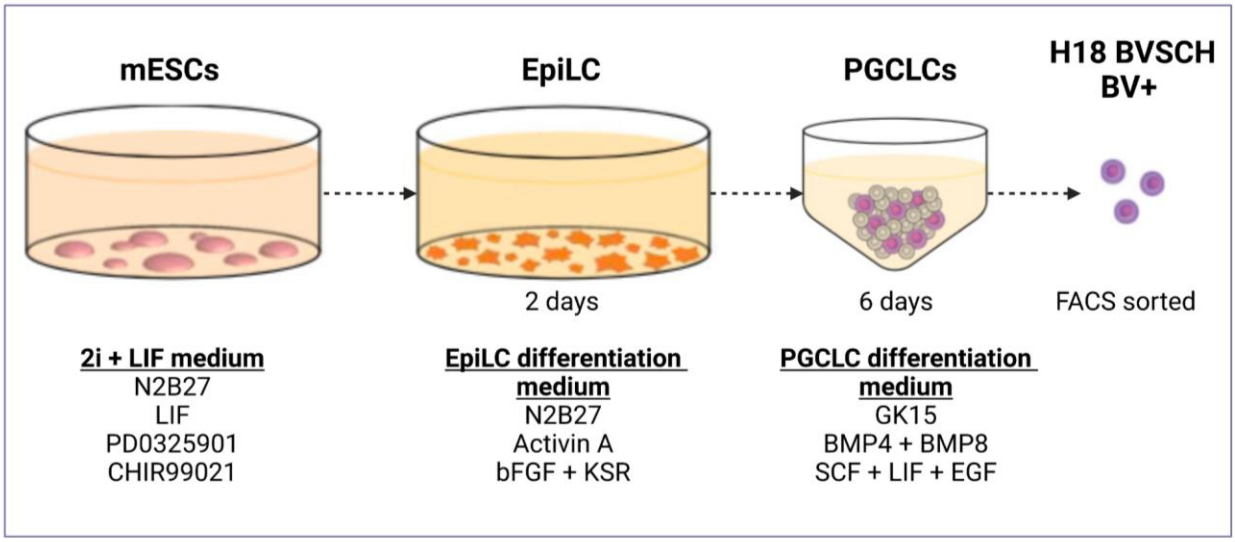

Figure 7.6: Overview of the In vitro gametogenesis protocol. *A schematic representation of the protocol to generate PGCLCs from mSECs. mESCs are induced into EpiLCs (Steps 2: 2 days) and then into mPGCLCs (Step 3: 6 days), which are sorted using FACS to isolate BV+ cell population.*

7.2.4 The cell culture establishment and mESC maintenance

The following steps were undertaken to establish and maintain the cell line. The growing surface of a T25 flask was prepared for the cell culture by coating it with 0.01% Poly-L-Orthinine solution, required to enhance cell attachment and adhesion. After the flasks were aspirated and washed with PBS, the growing surface was coated with laminin (150 ng/mL), an extracellular matrix multi-domain trimeric glycoprotein that supports cell adhesion, proliferation and differentiation. The cells were resuspended in 1 mL 2i maintenance media described above, and passaged onto the laminin coated culture plate when reaching the confluence similar to the one shown in **Figure 7.7**. The cell culture incubation was performed at 37°C in the 5% CO2 incubator, and the media was being changed daily following the protocol described below. At the pictured confluence, cells were usually split at a ratio of 1:100, however the split ratio was adjusted to account for any differences in observed confluence.

To maintain the mESC lines, the 2i media was aspirated, the surface of the flask was coated with TryplE and incubated for 4 minutes at room temperature to detach the cells. The cells were washed with TryplE wash medium (DMEM F12 with 75.5% BSA) and transferred to the Falcon tube and centrifuged at 300

xg for 3 minutes. After aspirating the supernatant, the cells were resuspended in 1 mL 2i maintenance medium and following a 1:100 ratio the cell solution was split to each new flask. The cytokine leukaemia inhibitory factor (LIF), which activates STAT3 signalling that is central to ESC renewal, and CHIR99021, which enhances survival at low density, restores viability and allows efficient expansion of undifferentiated ESCs, were key for successful maintenance of the mESC cell line.

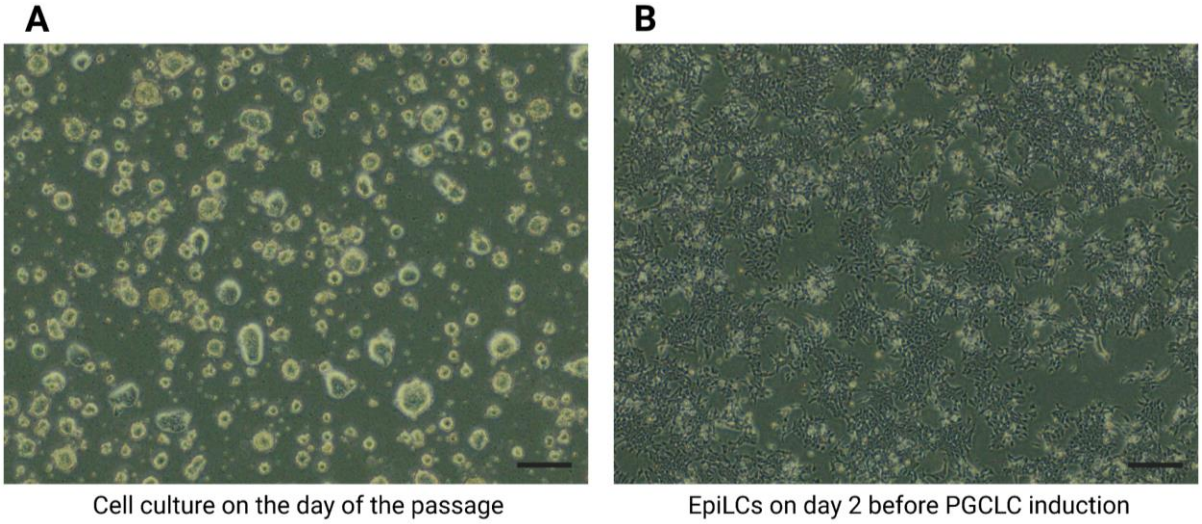

Figure 7.7: BVSCH18 cell line culture. (A) *mESC culture on the day of the passage that should be split at the ratio of 1:100.* ***(B)*** *The representative image of EpiLCs on day 2, just before PGCLC induction. Both images were captured at 4X magnification. The images were taken by J.A. Halliwell.*

7.2.5 EpiLC differentiation

The EpiLC induction was performed as stated in Hikabe *et al* (2016)[521]. Briefly, mESCs cultured in 2i/LIF were dissociated as described in the previous section and $1x10^5$ cells were seeded onto fibronectin (16.7 mg/ml) coated plates with EpiLC differentiation medium (N2B27, 1% KnockOut Serum Replacement (KSR), bFGF (10 ng/ml), Activin A (50 ng/ml)). The cells were incubated at 37°C for 42 hours and the media was changed every day. The density of EpiLCs was monitored as it was crucial for efficient differentiation of PGCLCs. The desirable density is depicted in the above **Figure 7.7**.

7.2.6 PGCLC differentiation

EpiLC were gently dissociated using TryplE for 2 minutes, which was then blocked using TryplE wash medium. The total cell solution was transferred to a Falcon tube, centrifuged at 300 xg for 3 minutes and resuspended in 1 mL of GK15 (GMEM, 15% KSR, 1% NEAA, 1% 100 mM sodium pyruvate, 0.1mM 2-mercaptoethanol, 1% penicillin/streptomycin and 1x 2mM L-glutamine). The cells were counted using a

hemocytometer and diluted in PGCLC induction medium, GK15 supplemented with cytokines: BMP4 (50 ug per mL of 4mM HCl 0.1% BSA solution), LIF (1000 U/ml), SCF (50ug per mL of PBS 0.1% BSA solution), BMP8a (50 ug per mL of 4mM HCl 0.1% BSA solution) and EGF (500 ug per mL of PBS 0.1% BSA solution) to a concentration of $1.5x10^4$ cells per mL. It is important to note that PGCLC differentiation is performed in aggregates. We plated 100 uL of the EpiLC cell solution into each well of a ultra low-cell binding U-bottom 96-well Lipidcure-coated plate, which was incubated for 6 days at 37°C to reach the greatest yield of PGCLCs. The cells were observed under the microscope on each day around the same time to monitor their development. After 6 days, PGCLCs were collected for FACS analysis.

7.2.7 PGCLC purification on H18 cells

To prepare PGCLCs for FACS sorting, cultured cells were collected from the plate, washed with PBS and dissociated in TryplE for 6 minutes at 37°C. TryplE was neutralised using MEF medium (DMEM, 10% FBS, 2mM L-Glutamine and 1% Penicillin-streptomycin solution), and dissociated cells were passed through 70 µM strainer to remove large clumps of cells. The cells were counted using the automatic cell counter, centrifuged at 300 xg for 5 minutes and resuspended in FACS buffer (0.1% BSA in PBS) before being taken to FACS sorting.

7.2.8 Fluorescence activated cell sorting (FACS)

The characterisation and analysis of different cell populations were based on the cell's size, granularity and fluorescence. These were detected by FACS (**Figure 7.8**)[533]. H18 Blimp1-mVenus PGCLC (BV+ cells) were sorted from the rest of the sample using a SONY SH800Z cell sorter. The sort was performed on WT and KO cells separately, yet following the same parameters. The cell suspension was run through the cytometer where the cells were gated using forward (FSC-A) and back scatter (BSC-A) to remove the debris, while doublet or multiplet cell exclusion was performed based on FSC-A and FSC-W. The BV+ fluorescent cells were recognised by the laser beam as explained in **Figure 7.8** using Venus (yellow channel) and phycoerythrin (PE) (orange channel) fluorochromes, and sorted into a FACS buffer. For each sort the maximum number of cells was collected, while for flow analysis at least 10,000 cell data points were captured to achieve a representative sample.

Finally, it is important to note that due to the limited timeframe, no statistical methods were used to predetermine sample size required to draw statistically significant conclusions. However, the protocol was independently performed by two scientists (S. Stankovic and J. A. Halliwell) in two different cell clones, replicating the final observations.

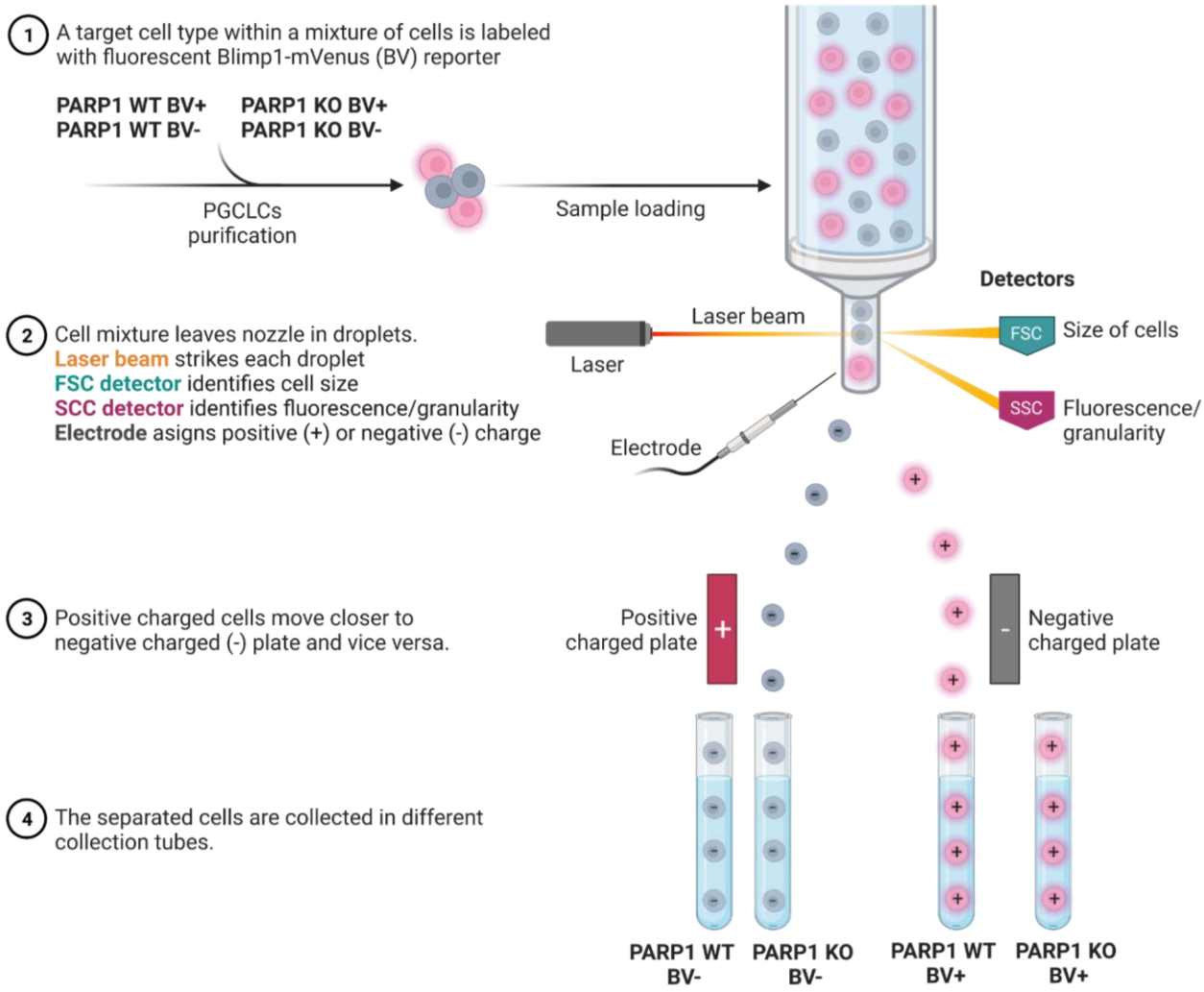

Figure 7.8: A schematic representation of the methodology behind FACS. *Cells that generate the negative charge while passing through the electric field are defined as fluorescent, while the ones with the positive charge do not fluoresce. Based on the charge they carry, the cells will be removed from the electric field and collected in the homogenous solutions, including fluorescent cells, non-fluorescent cells and waste. The output of the fluorescent signal is received in the form of histogram, recorded on the computer connected to the Influx sorter, reporting the percentage of the cell population of interest.*

7.2.9 RNA extraction, reverse transcription, and quantitative real-time polymerase chain reaction (qPCR)

In order to assess in what way inactivation of *PARP-1* affects the expression of other genes with an important role in maintenance of pluripotency and self-renewal, we specifically focused on transcription factors: *Sox2*, *Oct4*, *Nanog*, *Stella* and *Blimp1*. We focused on three scenarios: the gene expression in (1) control wild type (WT) PGCLCs, (2) *PARP-1* KO PGCLCs, as well as (3) control WT PGCLCs incubated for 24 hours in 3 μM of mitogen-activated protein kinase (MEK) inhibitor (PD0325901) that

inhibits FGF/ERK pathway, thus decreasing *Parp-1* enzymatic activity and *Parp1-Sox2* interactions and promoting self-renewal of cells[534].

Cells were rinsed twice with PBS. Total RNA was extracted and purified using RNeasy Plus kits (QIAGEN) from each biological sample according to the manufacturer's instructions. cDNA synthesis was performed with 1 μg of total RNA using oligo (dT) primers and reverse transcribed using Superscript III (Invitrogen) and diluted ten-fold. Real time PCR analysis of the pluripotency markers was carried out with SYBR Green Supermix (Bio-Rad) and gene specific primers (sequences available upon request) using the MJ Research Opticon 2 Real-Time System. Thermocycler program consisted of an initial hot start cycle at 95 °C for 3 min, followed by 32 cycles at 95 °C for 10 sec and 59 °C for 30 sec. The analysis was performed in triplicate. Mouse glyceraldehyde-3-phosphate dehydrogenase (*GADPH*) was used for normalisation of the qRT-PCR results. Results were then log transformed and the fold (relative) expression to the WT control was calculated using the $2^{-\Delta\Delta CT}$ method[535].

7.3 Results

7.3.1 Common *PARP-1* V762A variant leads to earlier ANM in women

Driven by the evidence from studies in the mouse model on the role of *PARP-1* inhibition in gametogenesis and ovarian function, we investigated the role of the most common *PARP-1* genetic alteration, V762A. This variant represents an interesting candidate as it reduces *PARP-1* activity up to 40% in carriers, thus acting as partial *PARP-1* inhibitor[496]. To assess its impact we focused on the ANM phenotype, which informs on the genetic regulation of menopause timing and acts as a proxy for extreme cases of early menopause, such as POI. GWAS array data on variants with MAF ≥ 0.1%, were available for 201,323 women of European ancestry from ReproGen Consortium and 294,828 women from 23andMe study. The genetic association analysis was performed using a linear model in each strata and then combined by meta-analysis. We specifically investigated *PARP-1* V762A and identified suggestive human genomic evidence on the role of the V762A minor allele, which reduces *PARP-1* function, in decreasing menopausal age in women when analysing the two strata independently, ReproGen (Beta: -0.075 years per allele [95% CI: 0.04-0.11], $P=4.6*10^{-6}$) and 23andMe (Beta: -0.05 years per allele [95% CI: 0.02-0.08], $P=1.77*10^{-3}$). We replicate these findings by meta-analysing the data and confirm a significant association with earlier ANM (Beta: -0.06 years per allele [95% CI: 0.04-0.09], $P=5.3*10^{-8}$).

Having demonstrated the impact of *PARP-1* inhibition on reproductive longevity, we aimed to explore the biological mechanisms that underlie the effect of *PARP-1* on gametogenesis and ovarian function using CRISPR technology in conjunction with the IVG culture system.

7.3.2 Generation of the *PARP-1* knockout mESC

In the mammalian embryo, pluripotency is established from the epiblast in the inner cell mass (ICM) of the preimplantation blastocyst. We relied on this pluripotent feature to establish a mESC line, and monitor its potency to differentiate into PGC, the founding germ cell population, using the IVG system. We specifically used the BVSCH18 ESC line, which bears the BVSC reporter construct, to study female germ cell development *in vitro*[536]. This reported construct enabled us to trace the differentiation of PGCLCs - early PGC specifications are *Blimp1*-positive while lineage restricted germ cells, i.e. migrating PGCs, are *Blimp1*- and *Stella*-positive[523,537–539].

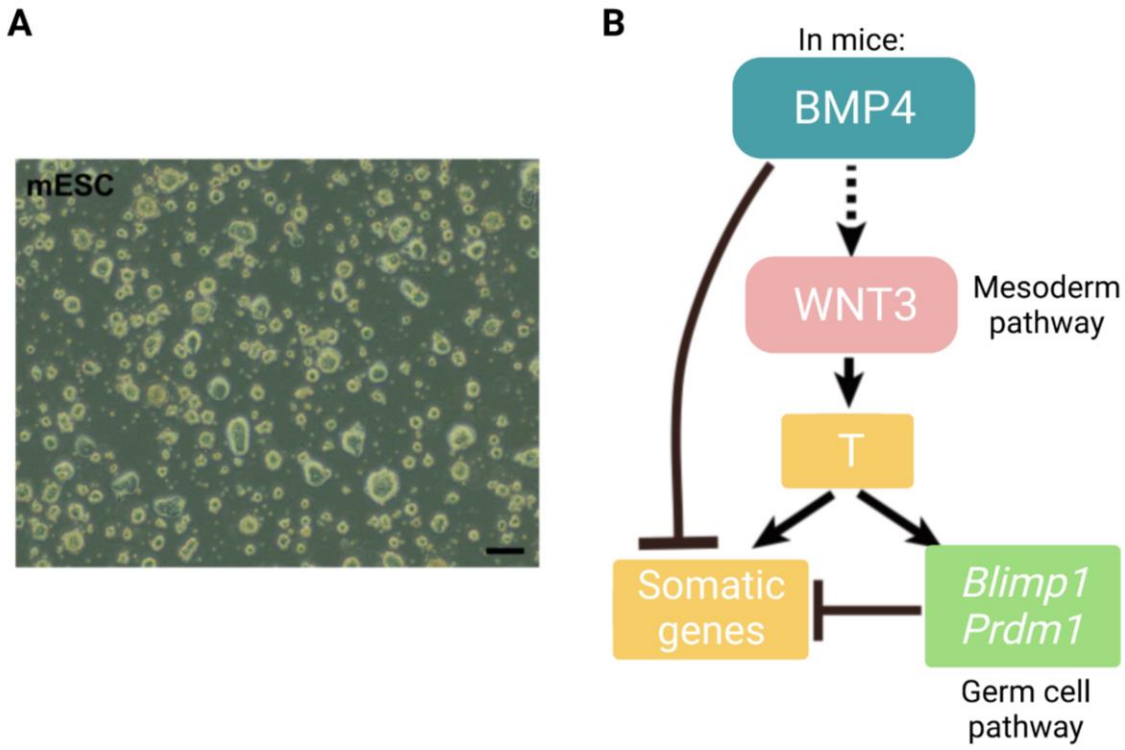

Figure 7.9: Germ cell specification. (A) *The representative sample of BVSCH18 ESC line in the cell culture containing 2i medium.* ***(B)*** *The role of Blimp1 in lineage restricted germ cell specification. Blimp1 is necessary for successful repression of the somatic mesodermal program, re-acquisition of pluripotent potential, and genome-wide epigenetic reprogramming. Figure is adapted from Saitou et al (2016)*[540].

In order to explore the role of *PARP-1* in establishing ovarian reserve and controlling sequential cell fate decisions during germ cell-specific differentiation, we generated the *PARP-1* knockout mESCs by disrupting exons 3 and 4 using CRISPR/Cas9 targeting. Potential candidates of mutant cell clones were confirmed by Sanger sequencing and Western Blot, thus enabling us to monitor the role of *PARP-1* in ovarian function and development (**Figure 7.10**). The clones used for the discovery analysis carried a 3 base (AGG) deletion in exon 4.

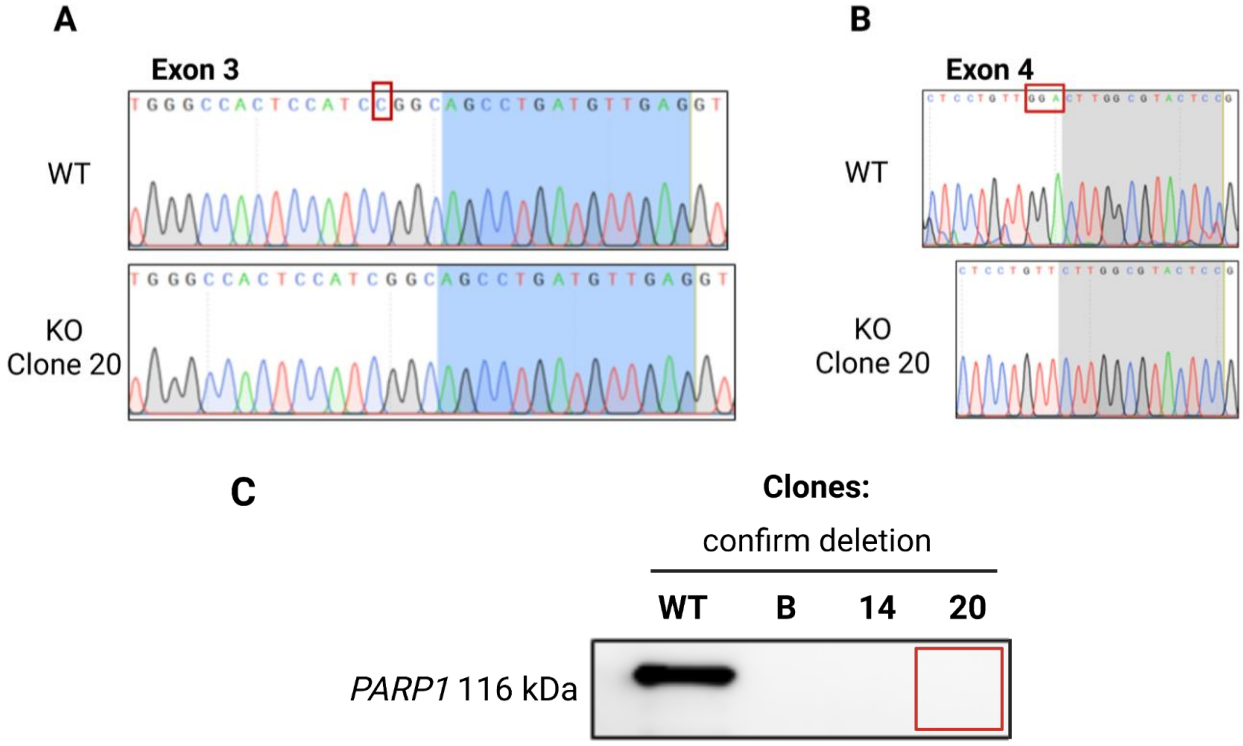

Figure 7.10: Evidence on the successful PARP-1 knockout in mESCs. *Panels* ***(A)*** *and* ***(B)*** *show results from Sanger sequencing.* ***(C)*** *Western blot and evaluation of the PARP-1 expression. B = blank, WT = wild type, Clones: 14 and 20. Clone 20 with AGG deletion in exon 4 was used in the discovery analysis, while clone 14 with C deletion was used in an independent replication by J.A.Halliwell.*

Off-target mutations from the CRISPR could cause unusual behaviour in a single clone[541]. By using two clones with AGG deletion in exon 4, we can be fairly certain that potential off target mutations are not responsible for the phenotype we observe. Furthermore, the KO in the third clonal line is slightly different, i.e. carrying a C deletion, which adds weight to our findings as 2 independent mutation induced KOs presented the same phenotype. Finally, to avoid bias the experiments were performed independently by two scientists, S. Stankovic and J.A. Halliwell.

Using the IVG model, we embarked on investigation of the efficacy of WT and *PARP-1* KO mESCs in (1) the transition to the germline competent state (EpiLC) and (2) specification and differentiation potential of the PGC lineage, thus elucidating the impact of *PARP-1* on the very beginning of the establishment of ovarian reserve. We observed no growth or morphological differences between the WT and *PARP-1* mESCs.

7.3.3 Differentiation of mESCs into PGCLCs

The postimplantation epiblast has a unique cell context with the capability of evoking PGC fate, in contrast to ESCs, which do not have this capacity and rather promote self-renewal. Therefore, to be able to reconstitute PGC specification in vitro, it is necessary to construct an epiblast-like state with PGC competence in ESCs. We induced the transient differentiation of naive mESCs under a defined set of conditions, including cytokines basic fibroblast growth factor (bFGF) and activin A, which were necessary for successful epiblast induction. Under these culture conditions, the mESCs exhibited a rapid change in cell morphology during the 2-days long induction. This involved the transition of round into flat colonies, assuming a more epithelium-like structure and demonstrating a successful transition in EpiLC state (**Figure 7.11**). *PARP-1* KOs exhibited indistinguishable morphology under the microscope from matched WT EpiLCs derived following the same methodology (**Figure 7.11**).

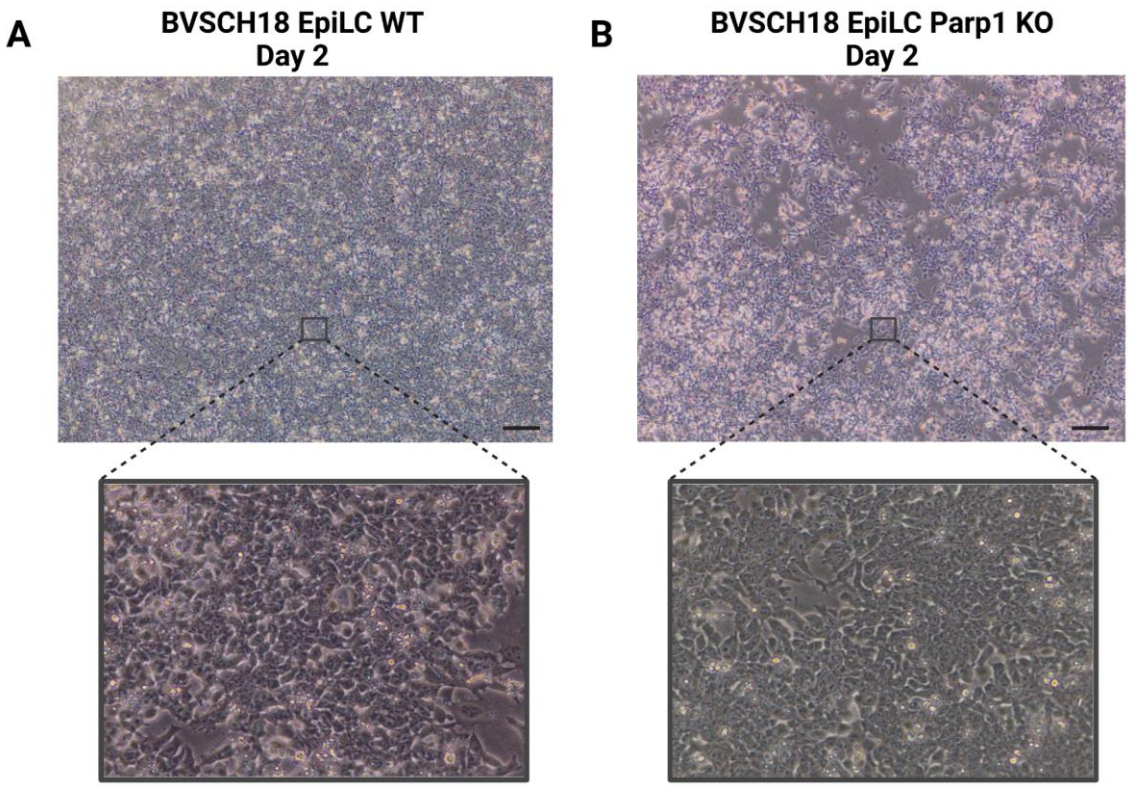

Figure 7.11: Induction of mESCs into EpiLCs. *The representative figure demonstrates EpiLCs on day 2 just before the PGCLC induction, including the WT in* ***(A)*** *and PARP-1 KO in* ***(B)****. The images were taken under the microscope using 4x and 20x magnification (for zoomed-in sample). Scale bars are 100 µm.*

We next examined whether both WT and *PARP-1* KO EpiLCs have the potential to be induced into PGCLCs. The PGCLC differentiation was performed in aggregates. Over the course of 6 days, we successfully stimulated EpiLCs by BMP4 and WNT signalling to form PGCLCs, the founding germ cell population[540,542,543]. This induction results in Blimp1/Prdm1 mediated transcriptional regulation of epiblast cells which promotes the expression of PGC-specific genes, such as *Stella*, and represses the expression of somatic cell genes such as members of the Hox gene family[526,544]. Even though it has been shown that BVSC-positive foci appeared from day 4 to day 6 of culture under full induction conditions, previous experiments in the same laboratory (by J.A. Halliwell) demonstrated that the highest PGCLC yield is at day 6. These specified PGCLC are equivalent to migratory PGCs *in vivo* (E8.5-E10.5), and have the potential to develop to mature functional gametes. The morphology of the WT and PARP-1 KO cells was monitored across the 6 day differentiation course. No major differences were observed in the size of the cell aggregates, yet we could not observe the morphology of cells at the single level (**Figure 7.12**).

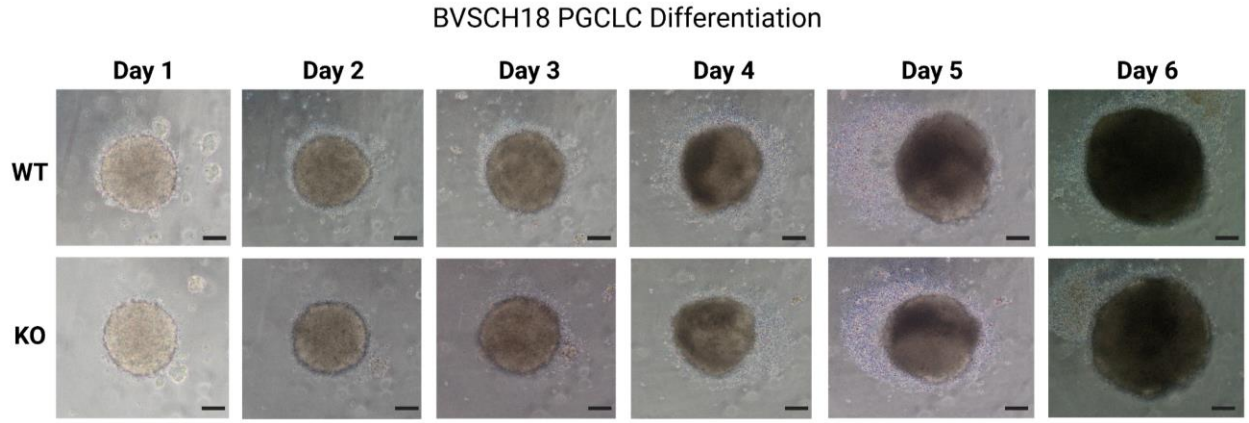

Figure 7.12: PGCLC induction over the course of 6 days. *Each image shows a representative sample of PGCLC aggregate. The same plate well was imaged during 6 days for both WT and the PARP-1 KO. The magnification used was 10x. Scale bars are 100 µm.*

BV-positive (BV+) PGCLCs are induced in many but not in all EpiLCs. The non-PGC population includes undefined somatic cell lineages and an undifferentiated ESC-like cell lineage. Therefore, in order to explore the dynamics of PGCLC induction and proliferation in the WT and *PARP-1* KO BV+ PGCLCs, we sorted our samples into BV+ and BV- cells. This was done using FACS to isolate the BV+ population of interest, with fluorescence indicative of successful transition to PGCLC state. Acquisition of PGCLC fate indicated by BV+ cells was markedly increased in *PARP-1* KO, which is more than double in comparison to the WT PGCLCs (**Figure 7.13**). The average percentage of BV+ cells across three *PARP-1* KO clones was 30.77%, whereas that of WT cells was 10.68% (**Figures 7.13 - 7.15**). This result revealed a drastic increase in the proliferation rate in the KO PGCLCs, with a 3.37 times (absolute difference 20.09%) higher number of BV+ cells on average, which is opposite to what we hypothesised regarding an expected decrease in the number of BV+ cells.

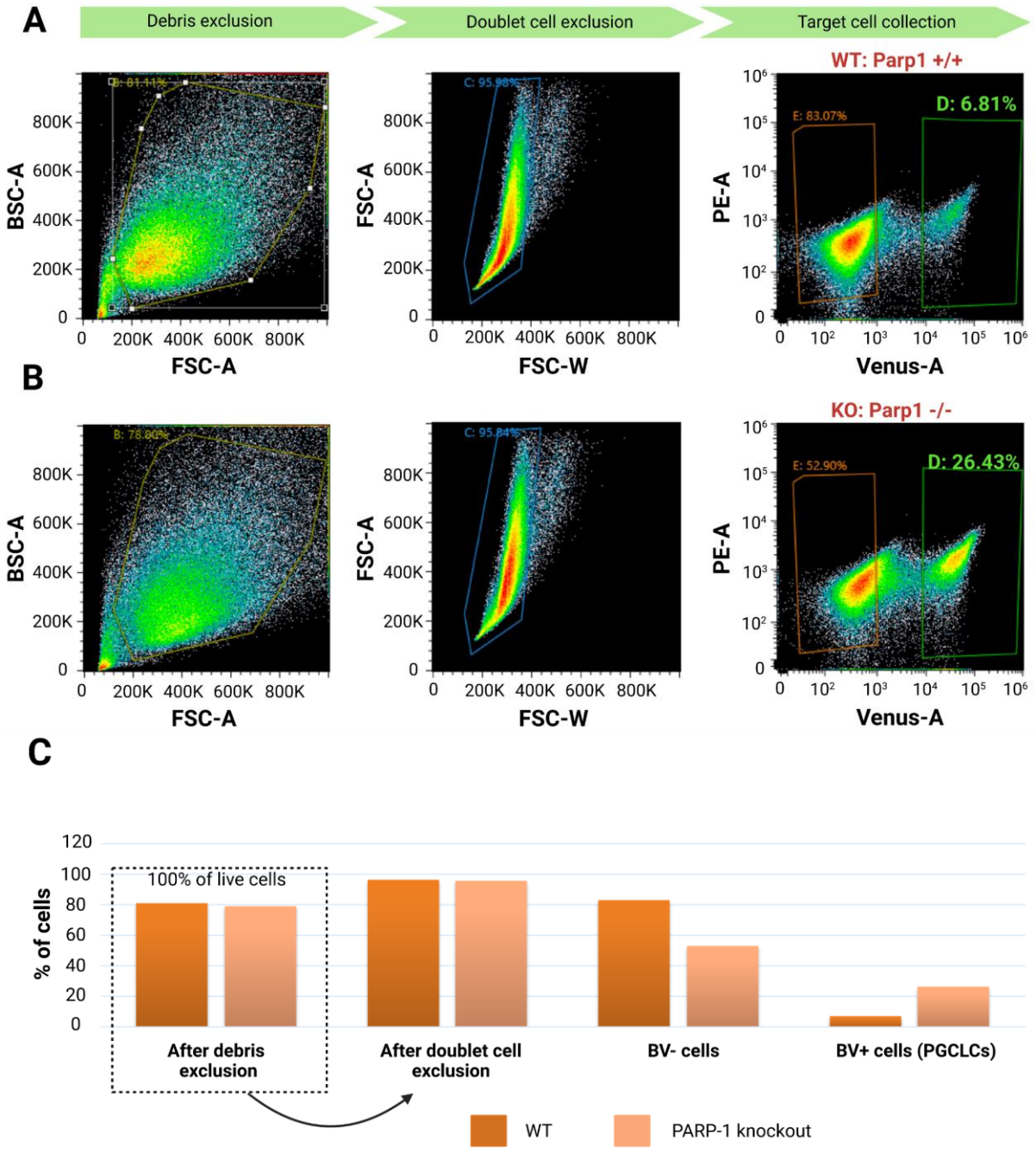

Figure 7.13: FACS workflow for isolation of BV+ PGCLCs in the replicate 1 of clone 20. *The figure represents FACS gatings for sorting PGCLCs. Panel* ***(A)*** *is for the WT and panel* ***(B)*** *is for the PARP-1 KO. The cell debris and doublet cells were sorted using an appropriate gating of the forward scatter (FSC) and the back scatter (BSC) as shown in the top and bottom two panels, while cells with autofluorescence were sorted by Venus-A and the PE-A channels as shown in the third panel. A and W indicate area and width, respectively. The results are shown for the replicate 1 of the clone 20 with AGG deletion in exon 4. There was a 3.9 times (absolute difference 19.62%) higher number of PGCLCs in the KO compared to the WT.*

We obtained consistent results in the replication analysis of clone 20 (absolute difference between the WT and KO 25.58%) (**Figure 7.14**), as well as in the replication using clone 14 with a C deletion in exon 3 (absolute difference 15.08%) (**Figure 7.15**).

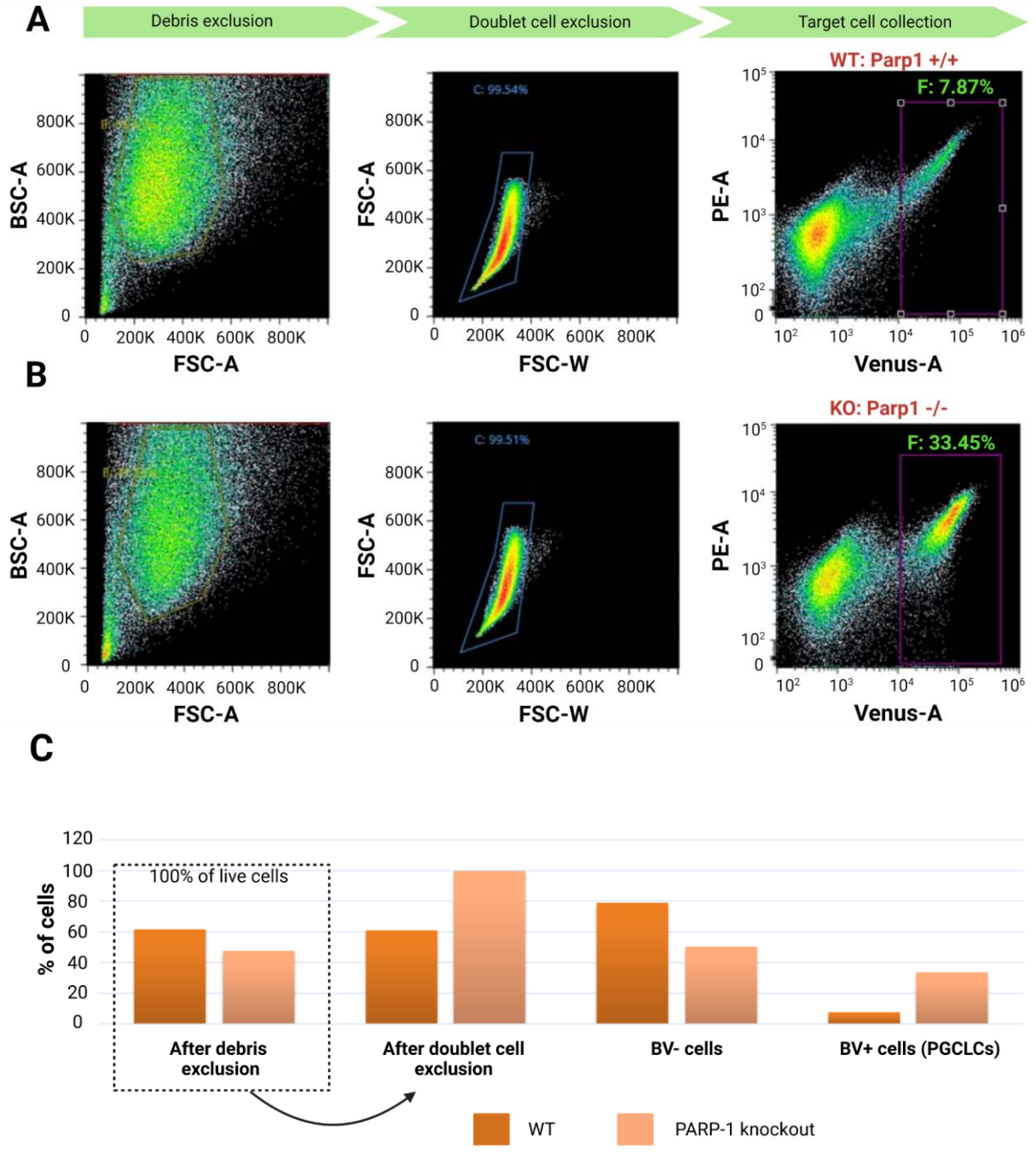

Figure 7.14: FACS workflow for isolation of BV+ PGCLCs in the replicate 2 of clone 20. *The figure represents FACS gatings for sorting PGCLCs. Panel (A) is for the WT and panel (B) is for the PARP-1 KO. The cell debris and doublet cells were sorted by an appropriate gating of the forward scatter (FSC) and the back scatter (BSC) as shown in the top and bottom two panels, while cells with autofluorescence were sorted by Venus-A and the PE-A channels as shown in the third panel. A and W indicate area and width, respectively. The results are shown for the*

replicate 2 of clone 20. There is a 4.3 times (absolute difference 25.58%) higher number of PGCLCs in the KO compared to the WT.

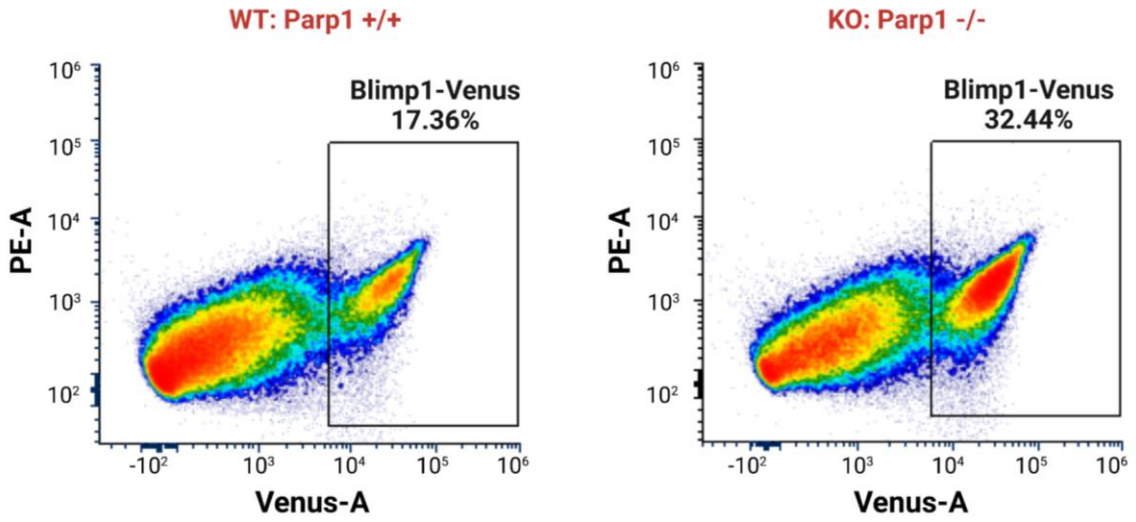

Figure 7.15: FACS workflow for isolation of BV+ PGCLCs in the replicate 1 of clone 14. *The figure represents the ultimate FACS gating for isolating PGCLCs. BV+ cells with autofluorescence were sorted by Venus-A and the PE-A channels. A indicates area. There is a 1.9 times (absolute difference 15.08%) higher number of PGCLCs in the KO compared to the WT.*

The results suggested that the effect of *PARP-1* on ovarian function starts early, at the very beginning of the establishment of ovarian reserve, having a dramatic influence on limiting PGCLC development and proliferation.

7.3.4 Elevated levels of Oct4 drive PGCLC self-renewal

To further assess the functional relevance of *PARP-1* in PGCLCs, we performed a gene expression analysis focusing on the transcription factors involved in the maintenance of pluripotency and self-renewal, including *Sox2*, *Oct4*, *Nanog*, *Stella* and *Blimp1*. All genes showed a significant increase in relative expression in *PARP-1* KOs, the most significant one being *Oct4* with a 2.9 fold increase (**Figure 7.16A**). *Oct4*, *Sox2* and *Nanog* are highly expressed in pluripotent cells and become silenced upon differentiation. Significant increase in the relative expression in *PARP-1* KO cells suggests accelerated self-renewal potential in these cells. The high expression was also observed in *Stella* and *Blimp1*, providing a confirmation that these cells are indeed at the PGCLC state and that *PARP-1* deletion increases the efficiency of their production *in vitro*.

Previous experiments by Lai *et al.* (2012)[534] have demonstrated that suppression of the fibroblast growth factor (FGF)/extracellular signal-regulated kinase (ERK) signalling sustains ESCs in the ground, self-

renewal state by mediating *Parp1-Sox2* interactions and *Parp1* enzymatic activity (**Figure 7.16B**). Using mitogen-activated protein kinase (MEK) inhibitor (PD0325901) of the FGF/ERK pathway, we demonstrate directionally consistent results with the ones observed in the *PARP-1* KO cells. This demonstrates that *Parp1* plays important roles in the control of genes regulated by *Oct-Sox* enhancers and involved in maintenance of self-renewal and pluripotency.

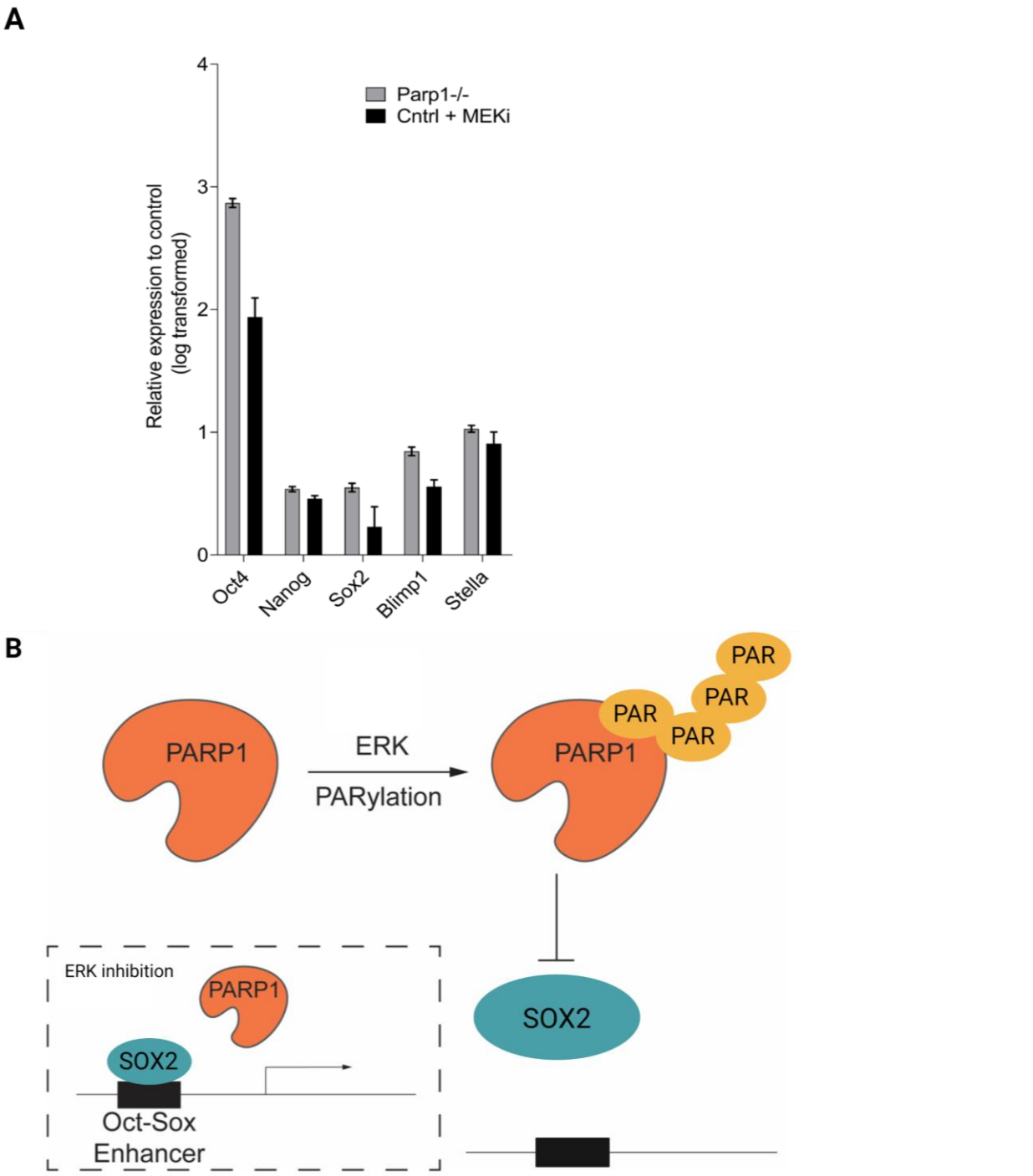

Figure 7.16: Controlled primordial germ cell expansion via regulation of self-renewal and pluripotency transcription factors. (A) *qPCR relative expression of Oct4, Nanog, Sox2, Blimp1 and Stella genes in Parp1 KO and WT MEKi treated cells. MEK inhibition represses FGF/ERK pathway activity.* ***(B)*** *The mechanism behind the role of Parp1 in the control of genes regulated by Oct-Sox enhancers. FGF/ERK signalling induces Parp1 enzymatic activity (auto-PARylation) and Sox2-Parp1 interaction. Via this interaction, functional Parp1 suppresses Sox2 activity at Oct-Sox target genes thus stimulating cell differentiation. On the contrary, inhibition of ERK signalling or Parp1 deletion enables binding of Sox2 to Oct-Sox enhancers, leading to increased levels of Oct4 and thus enhances self-renewal of cells. Figure, part B, is adapted from Lai et al. (2012)*[534].

7.4 Discussion

A significant progress has been made in the past decade towards identification of novel gene candidates regulating ovarian ageing. Most of the genes in which the knowledge about the functional mechanism has advanced, act during later stages of gametogenesis, such as in the case of *CHEK2*[481]. Our knowledge remains obscure when it comes to the genes involved in establishing the ovarian pool. Combining the evidence from human genomics with the robust functional tools, this study aimed to elucidate potential mechanisms behind the role of *PARP-1* in ovarian ageing and the creation of the germ cell line.

Using genome-wide association data derived from UKBB and 23andMe cohorts, we provided suggestive human evidence on the role of common *PARP-1* V762A variant as a risk factor for earlier ANM in women. This is in line with previous evidence that experimentally showed reduction of *PARP-1* catalytic activity by 30-40% due to the minor allele (the alanine substitution) of a missense variant (V762A)[496] and its association with the reduction of ovarian reserve [110,512]. In addition, this variant has been previously shown as important for other health outcomes in human genomic data by being protective for LOY in men[481].

Considering the effects of *PARP-1* inhibition on reproductive function is especially important as Olaparib, a *PARP* inhibitor, is already used in clinics to treat breast and ovarian cancer. These types of cancer often occur in reproductively active women, so better understanding of the consequences on ovarian function could allow more aggressive fertility preservation or personalised treatment approach. Serum AMH concentrations and menstrual cycling are used clinically as surrogate indicators of ovarian damage in response to cancer therapies [545,546]. However, AMH is primarily secreted from granulosa cells of growing follicles and not the primordial germ cells[546]. If *PARP-1* inhibition affects the ovarian reserve and thus primordial follicles, we would not be able to quantify changes to the ovarian function in the clinical setting, and thus would be exposing reproductive active women to misdiagnosis and risk of premature menopause.

Following up the evidence from human genomics, we investigated the functional mechanism through which *PARP-1* acts to shape the ovarian reserve. Using CRISPR/Cas9 in combination with the IVG system as an *in vitro* model of ovarian establishment, we traced the pathway from pluripotency to germ cell fate in the BVSCH18 mESC line. We demonstrated that deletion of *Parp-1* in mESCs leads to an increased efficiency of PGCLC differentiation with an average absolute difference of ~20% compared to the WT. Consequently, this leads to the creation of a larger ovarian reserve at the very beginning of the germ line formation. This contrasts with our initial hypothesis that predicted a greater germ cell loss in the

Parp-1 KO due to the faulty DDR mechanism and thus higher cell death rate. Instead, we speculate that *PARP-1* activation in PGCs is likely to be independent of DDR. Previous research indicated the non-catalytic role of *PARP-1* in the regulation of pluripotency transcription factor, *SOX2*[534]. In response to FGF/ERK signalling, *PARP-1* acts as an inhibitor by restricting *Sox2* binding to *Oct4*/Sox2 enhancers and thus balancing ESC pluripotency and differentiation. Given *Oct4*'s importance in promoting self-renewal[547], elevated expression in the PGCLC state in *PARP-1* KO cells, as observed in our study, could promote the efficiency of PGCLC production[534]. These results suggest a significant role of *PARP-1/MEK/ERK* in governing reproductive lifespan in women.

Even though we observe an increased efficiency of PGCLC induction in the *PARP-1* KO compared to WT cells, this does not tell us anything about the cell quality. We speculate that this early acceleration of proliferation might expose the cells to various insults, which would most likely be eliminated through apoptosis at later stages when the effect of DDR is manifested. This would then ultimately lead to greater ovarian reserve depletion. If the loss of function increases the PGC pool size, this could alter the size of the territory occupied by the PGCs and perhaps also the density of these cells at this stage. Previous research demonstrates that signals diffuse across these cellular structures making a signalling gradient [548,549], which is key to the eventual transformation of these cells into oocytes. If the size of this cell population and their density is altered then it would make sense that this signalling will be impeded, thus lowering the overall ovarian pool. Significant questions remain about the temporality, source of damage and nature of repair transactions that are active in the germline.

This study has opened up a 'black box' behind the role of *PARP-1* in gametogenesis and ovarian ageing. Extensive future research is needed to understand its precise mechanism. This includes better understanding of the involvement of the *PARP-1/ERK* mechanisms on the regulation of pluripotency and self-renewal balance of PGCLCs, as well as the oocyte quality by assessing the levels of DNA damage via immunofluorescence. In addition, looking at the Chip-Seq and RNA-Seq analysis for the overview of the global chromatin and transcriptomic landscape could advance our understanding of epigenetic reprogramming in these cells. Reprogramming during gametogenesis ensures genetic totipotency in normal development and is essential for the imprinting mechanism that regulates the differential expression of paternally and maternally derived genes [516,550].
This study only explored the most detrimental effect of *PARP-1* through its deletion. Future research should specifically address the disease model of the V762A mutation. In addition, further advances in human IVG will create possibilities to replicate these experiments and thus provide a more applicable model, addressing potential differences in reproductive physiology between mouse and human[551]. The

effect of Olaparib was not explored as it is not a *PARP-1* specific inhibitor, but it also affects the activity of *PARP-2*[552]. To address the concerns coming from the clinical setting, it would be crucial to consider the effect of both of these genes in combination. If a drug is detrimental to growing follicles, ovulation and fertility might be temporarily impacted, but more eggs can be activated from the immature primordial reserve and ovulation will resume as normal. In contrast, if the primordial follicles are affected, this would lead to infertility and early menopause.

One of the limitations of this study is that it recapitulates a quite short window of gametogenesis. There thus remains a long period of time, perhaps over a month, which must be reconstituted to produce mature oocytes in culture to assess other factors and stages of gametogenesis that could be affected by this genetic alterations in *PARP-1*. This is especially interesting from the point of non- versus DDR-dependent mechanisms.

Some studies suggest that the *PARP-1* V762A polymorphism is involved in spermatogenesis impairments, increasing the risk of oligospermia in men[496,553–555], thus highlighting the importance of understanding this mechanism in relation to male fertility.

Finally, our experimental strategy outlines a principle for using CRISPR screens to deconvolve the genetic basis of successive cell fate decisions. We complement it with the IVG platform that showed successful reconstitution of the entire process of oogenesis from mouse pluripotent stem cells and the capability to develop viable and fertile offsprings from *in vitro* generate MII oocytes via *in vitro* fertilisation[521]. We demonstrate here the use of this powerful functional tool that will enable analysis to not only further our basic mechanistic understanding of the process of ovarian ageing, but also to diagnose and model infertility, explore new remedies and improve assisted reproductive technologies[556].

CHAPTER 8

Conclusions and Future prospects

8.1 Summary of my research

This dissertation described five distinct projects in which state-of-art genomic technologies with robust functional models were employed to identify genetic determinants of female reproductive ageing, focusing on aspects of the trait biology that have been poorly studied thus far. In this section I discuss how effective these studies have been, how these results have improved our knowledge and what opportunities and questions they raise to direct future research.

Chapter 3 presents a comprehensive WES analysis to study rare protein-coding variants associated with menopause timing in ~120K women in the UKBB. I describe significant advances in our understanding of the biology of human female reproductive ageing by identifying novel associations for five genes where heterozygous LoF has an effect on menopause timing substantially larger than previously reported by GWAS. Notably, I find that heterozygous loss of *ZNF518A* reduces menopause timing by nearly 6 years in carriers, an effect larger than anything currently tested in clinical genetics for premature ovarian ageing. Furthermore, I provide the first evidence of *ZNF518A* acting as a master transcriptional regulator of ovarian development and establishment of the ovarian reserve in foetal life. I also identify a new cancer predisposition gene, *SAMHD1*, which has a comparable effect size in women and men to well-established genes such as *CHEK2*, further reinforcing the link between cancer and reproductive ageing. Finally, I show that mothers with earlier ovarian ageing have a higher rate of *de novo* mutations in their offspring. This provides direct evidence that female mutation rate is heritable and highlights a mechanism for maternal effects on offspring health. Our study offers a robust start for future research, especially related to the discovery of potential candidates for the treatment development. Unlike GWAS, WES focuses on the coding part of the genome assessing the most damaging genetic changes, which introduce perturbations that should yield a severe functional defect. Consequently, this enables more straightforward translation of our findings into the clinical or drug discovery settings.

Chapter 4 expands the WES analysis to study the genetic architecture of extreme cases of early menopause, i.e. POI. Many genes have emerged as monogenic causes of POI, but a majority have been identified as causative in small numbers of families or individuals, with variable functional validation. This study is the first to demonstrate that autosomal dominant mutations in genes currently described in the literature or evaluated in clinical diagnostic panels are not common, highly penetrant causes of menopause under 40 years. I suggest that the origin of POI development may not be due to a single mutation in a candidate gene, but an interaction of low frequency polymorphisms or mutations in different genes in the same woman, indicating oligo- or polygenic aetiology for this disorder. This has an important impact on WES approaches that are increasingly being used within the clinical setting to make diagnostic

decisions. I indicate that using panels of genes to find causative genetic variants for un-related idiopathic POI is not a very fruitful endeavour and is unlikely to be cost-effective. Monogenic causes of POI are more likely to be recessive - this introduces disturbing changes that will require strong collaboration between the scientific and clinical world, and points towards detailed robust future research to further decipher genetic architecture of this condition.

Chapter 5 presents the first proteogenomic study for the age at menopause targeting 4,775 distinct proteins measured from plasma samples of 10,713 European descent individuals who were participants in the Fenland study. Even though this analysis did not bring fruitful and robust findings in terms of the protein candidates associated with menopause timing, it clearly demonstrates the potential of this type of analysis for the discovery of proteomic markers of reproductive ageing. Future enlarged population studies should further explore the association between the protein levels and menopause status.

Chapter 6 presents the largest genomic meta-analysis for age at menarche on ~566,000 women of European ancestry and 696 genomic loci that contribute to regulation of menarche timing. Using these data, I demonstrate the continued value of large genomic studies and how improvements in their size can greatly increase the number of genetic signals identified. I use this enlarged genomic dataset to get an insight into the biological mechanisms involved as well as potential shared genetic architecture between menarche and other reproductive health outcomes. Here, I provide the first ever evidence on the enrichment of DDR mechanisms for menarche timing, suggesting the involvement of DDR in regulation of both ends of the reproductive lifespan, i.e. menarche and menopause. In addition, I also point to some DDR gene candidates that could exclusively act in oocytes to modify the beginning of reproductive activity - this is the first evidence on the ovary specific mechanism, as it has been widely believed that menarche timing is being driven by mechanisms that act via hypothalamic-pituitary axis. I also highlight other novel mechanisms that impact both menarche and menopause timing, such as ribosome biogenesis, and suggest that future research should decipher this unexplored mechanism and its involvement in reproductive health. Finally, I demonstrate first human genomic evidence on the role of DDR and its related mechanisms in various anthropometric, metabolic and reproductive health outcomes. I show how different types of variation affecting the same gene can result in different phenotypes. These findings provide the first indication of DDR acting as a marker of broader health outcomes, and pave the path towards development of intervention strategies that could impact the outcome of multiple traits simultaneously. This study has just opened the ‘black box’ regarding the involvement of DDR in health and disease - future research should be directed towards better understanding of the relationship between menarche and menopause, and interventions that could be applied in early life and puberty to impact the reproductive longevity.

Finally, **Chapter 7**, follows a gene-centric approach to study the role of *PARP-1* in gametogenesis and reproductive ageing. I apply a unique methodology by combining human genomic evidence with cutting edge CRISPR technology and the IVG system to investigate the role of *PARP-1* in proliferation of primordial germ cells during the establishment of the ovarian reserve. The findings demonstrate for the first time that deletion of *PARP-1* increases the efficiency of primordial germ cell production *in vitro* via upregulation of *Oct4*, which could be driving their self-renewal. I speculate that, even though there is an initial increase in primordial germ cells, the quality of these cells could be compromised. This could expose them to substantial 'clearance' via cell death mechanisms at later stages of gametogenesis, ultimately leading to the creation of smaller ovarian reserve. This would be in line with the first human genomic evidence that I provide on the role of *PARP-1* inhibition on earlier menopause timing in women. These findings present an important basis for future research that relates to better mechanistic understanding of *PARP-1*, its effective translation into potential target for drug discovery as well as the effect of *PARP-1* inhibitors on ovarian ageing and fertility.

Several themes have emerged from presented studies as relevant and unexplored, yet critical for future advances in biological and mechanistic knowledge behind reproductive ageing. In the following pages, I will discuss the types of studies that are necessary to shape reproductive ageing research over the coming years and comment on how they will provide important clues in the road towards more personalised treatment of reproductive health outcomes.

8.2 Diverse ethnic origin

Although this work made significant progress in understanding the genetic architecture underlying reproductive ageing, one important limiting factor is that our insights have been restricted to women of European descent. One of the main reasons causing this is the lack of available large-scale studies in non-Europeans. Notably, epidemiological studies have shown that menopause timing varies across ethnic groups, suggesting that different modifiers might exist in different ethnic backgrounds. More specifically, African and African-American women have earlier, while Japanese later average menopause timing, as compared to women of European descent[183,557]. In addition, our previous study in the Japanese cohort, considerably smaller than the European one (43,861 and ~200,000 respectively), demonstrated the benefits of conducting multi-ethnic studies by identifying 8 new loci implicating novel genes and pathways involved in human reproductive ageing[183]. Due to potential distinct reproductive profiles and thus different risk for important health outcomes it is important to be inclusive when deciphering the genetic architecture of this universal reproductive event. Focusing only on studies in Europeans restricts the generalizability and translation of the findings to other ethnicities, and potentially limits the detection

of key genes and pathways that are poorly represented in the European population, thus urging for more inclusive research to enable effective and tailored prediction and prevention strategies. To unlock the true potential of this work, understanding how trajectories of growth and development relate to fertility and later disease risk in all ethnicities is an important question, as it may allow more informed reproductive choices for women and increased understanding of aetiology of accompanied health outcomes. In addition, the multi-ancestry analysis could also increase the power to fine map the causal variants due to the reduced linkage disequilibrium windows. Furthermore, better understanding of the process that determines ovarian development and function may well lead to new methods of contraception, assisted fertility and fertility preservation. Accruing large GWAS sample sizes for age at menopause is far more difficult than many other complex traits as the phenotype can by definition only be measured in older women. However, through academic and industrial collaboration, our future work will try to address this question - the proposed study aims to provide insights into the genomic architecture of reproductive ageing in less-well-studied populations, boosting their representation on global reference panels and improving our understanding of their population genetics characteristics to enable powerful study design as well as effective prediction and interventions. Diseases that are influenced by the timing of menopause, including cancer, cardiovascular diseases and T2D, are highly prevalent in underrepresented populations. Importantly, in some of these ethnic groups, such as African one, cardiovascular disease and cancer kill twice as many women aged 60 and above in low- and middle-income countries compared with high-income countries, predisposing women to a higher risk of poor health due to limited access to screening, late diagnosis and inadequate access to effective treatment[561]. Therefore, better definition of genetic architecture of reproductive ageing as well as of the complex interplay between genetic and environmental factors in a population-specific context is necessary, given an already demonstrated association between these traits and menopause timing. Our findings could provide insight into the mechanisms governing ovarian ageing, when they act across the life-course, and how they might be targeted by therapeutic approaches to extend fertility and prevent disease in women of diverse ethnic origins.

8.3 Investigating human metabolome and its relevance for reproductive ageing

Rare and common sequence variation across the genome identified up to date demonstrated a significant contribution to the regulation of the menopause timing. However, the translation of many established and emerging genome-to-phenome links is limited due to the challenges of assigning the causal gene driving the identified associations. This presents a major limitation for experimental follow-up, mechanistic

understanding, and use of the emerging genomic evidence in drug development. **Chapter 5** demonstrated how the assessment of human proteome has the potential to contribute to better understanding of regulation of reproductive ageing and identification of novel biomarkers. In addition to proteome, circulating levels of small molecules or metabolites and the impact of genetic differences in metabolism on human health represent an unexplored area. Future studies should focus on better understanding of the relationship of metabolites levels and menopausal timing as well as their utilisation as potential biomarkers of reproductive ageing. Metabolites are attractive biomarkers as they are widely measured in clinical medicine for diagnosis, prognosis or treatment response for other health outcomes. Previous studies have demonstrated that blood levels of metabolites are highly heritable, therefore representing an opportunity to integrate genetic association for metabolites and assess how genetics influences their levels and whether these changes modify the susceptibility to earlier vs later menopause timing. This study will be conducted within meta-analysis of genetic effects on levels of 174 blood metabolites measured in large-scale population-based studies on the Biocrates (AbsoluteIDQ™ p180, Fenland Study), Nightingale (1H-NMR, Interval Study) or Metabolon (Discovery HD4™, EPIC-Norfolk and Interval Studies) platforms.

8.4 From variant discovery to disease mechanisms

Previously described and future studies that encompass both common and rare genetic determinants of reproductive ageing will yield numerous candidate-disease causing mutations that have the potential to pave the path towards novel prediction and treatment strategies, as well as personalised medicine opportunities. Although variant discovery represents an important breakthrough towards this vision, it is only the first step. Understanding the biological mechanisms by which mutations and disease-susceptibility alleles contribute to observed outcomes is certainly a greater challenge and requires intense functional research that relies on robust and high-throughput models. Emerging genome-editing tools, including the CRISPR-Cas9 system, enable highly specific genome modifications at single-nucleotide resolution, thus significantly contributing towards better understanding of variant effects, gene function and disease mechanisms[319]. Notably, these editing tools can be used in a variety of *in vitro* and *in vivo* biological models, including induced pluripotent stem cells (iPSC), enabling a broad spectrum of experimental strategies to decipher the underlying biological mechanisms[320,321]. Most importantly these approaches are relatively quick and allow simultaneous assessment of multiple genetic mutations. **Chapter 7** already demonstrated the usefulness of this approach in studying the effect of a specific gene of interest and suggest that discussed IVG model in combination with CRISPR knockouts should be widely applied to investigate the functional effects of mutations, place putative disease-associated genes into a biological context, and to elucidate the mechanistic basis of reproductive ageing. In addition, the

simultaneous study of patient-specific iPSC lines with different risk-variants can aid in our understanding of how various disease-associated loci interact to produce a phenotype. Finally, investigating the functional relevance of non-coding variants that overlap with known (or putative) regulatory elements and epigenetic marks is also becoming feasible with the availability of novel Cas9-based techniques. These epigenome-editing proteins can therefore be targeted to candidate regulatory elements in order to modify local chromatin structure and determine the role of these elements in influencing gene expression and pathological mechanisms of disease.

8.5 Non-genetic risk factors and menopause timing

Previous studies have suggested a wide-range of non-genetic risk factors and their association with menopause timing. These non-genetic causes of menopause account for the remaining population variation in menopause age. As previously discussed, odds of early menopause were shown to be increased by smoking, alcohol consumption, having decreased levels of education and being nulliparous. Future studies should focus on assessing the ways in which genetic risk factors interact with non-genetic risk factors to modify the menopause timing - i.e. studying 'nature versus nurture' effect. In addition, in the wider context of female reproduction, we also have limited understanding of the role of genetics in determining sex hormone levels, which are vital for the normal function of the female reproductive cycle and which change markedly around menopause.

In order to understand the complex diseases that arise in later life researchers have been looking for answers in early life exposures. This association of early life determinants with the risk of adult complex diseases was captured by the Developmental Origins of Health and Disease (DOHaD) model[558]. DOHaD suggests that adaptations made by a developing foetus in utero, as a response to adverse environmental factors, together with postnatal exposures during early childhood and puberty do indeed have long lasting impact on health and modify the susceptibility of diseases[559]. Previous research in mice indeed demonstrated that a maternal obesogenic diet during pregnancy decreases the ovarian reserve in offspring[350,560] and that DDR mechanisms that act in utero to influence reproductive lifespan might be modified by maternal exposure[82]. Further deciphering the ways in which external factors regulate reproductive longevity is also critical for the development of public health interventions. By combining the results of these complementary areas of research, we should be able to develop a more complete understanding of female reproductive ageing.

8.6 Concluding remarks

Human genomics has made a remarkable impact on our knowledge of the genetic determinants of reproductive ageing. NGS is now significantly altering our ability to conduct gene-mapping studies and is yielding unprecedented biological insights that are truly driving a revolution in women's health and healthcare more broadly. Utilisation of NGS within the clinical setting is becoming increasingly attractive due to the lower cost and more robust bioinformatic approaches for the interpretation of identified genetic variants. Consequently, NGS might perhaps soon be the universal diagnostic and public health tool, allowing us to more rapidly diagnose disease and predict its onset. Our analyses contributed to the knowledge of the genetic architecture of reproductive ageing and the biological processes that underpin it, while future suggested studies are necessary to drive translation of these findings forward. Finally, we aim to utilise the knowledge obtained from this work to help achieve three main goals:

(1) To identify putative therapeutic targets for preserving and enhancing fertility. Our recent work demonstrates our ability to identify causal genes and experimentally characterise them in ovarian models.

(2) To identify causal, modifiable, risk factors for premature ovarian ageing using human genetic approaches.

(3) Developing tests (based on genetics and other traits) to predict which women are most at risk of infertility due to premature ovarian ageing.

Appendix

All Appendix files can be found as an electronic version with detailed description available at: https://universityofcambridgecloud-my.sharepoint.com/personal/ss2472_cam_ac_uk/_layouts/15/onedrive.aspx?id=%2Fpersonal%2Fss2472%5Fcam%5Fac%5Fuk%2FDocuments%2FStasa%20Stankovic%20PhD%20Appendix%20Tables

Bibliography

1. World Health Organisation. Sexual and Reproductive Health and Research. https://www.who.int/teams/sexual-and-reproductive-health-and-research-(srh)/overview (2022).
2. Christensen, K., Doblhammer, G., Rau, R. & Vaupel, J. W. Ageing populations: the challenges ahead. *The Lancet* vol. 374 at https://doi.org/10.1016/S0140-6736(09)61460-4 (2009).
3. Team, I. S. Variations in reproductive events across life: a pooled analysis of data from 505 147 women across 10 countries. *Hum. Reprod.* **34**, (2019).
4. Faddy, M. J., Gosden, R. G., Gougeon, A., Richardson, S. J. & Nelson, J. F. Accelerated disappearance of ovarian follicles in mid-life: Implications for forecasting menopause. *Hum. Reprod.* **7**, (1992).
5. Jones, O. R. *et al.* Diversity of ageing across the tree of life. *Nature* **505**, (2014).
6. Croft, D. P., Brent, L. J. N., Franks, D. W. & Cant, M. A. The evolution of prolonged life after reproduction. *Trends in Ecology and Evolution* vol. 30 at https://doi.org/10.1016/j.tree.2015.04.011 (2015).
7. Hawkes, K. Grandmothers and the evolution of human longevity. *American Journal of Human Biology* vol. 15 at https://doi.org/10.1002/ajhb.10156 (2003).
8. Mattern, S. *The slow moon climbs*. (2019).
9. Broekmans, F. J., Faddy, M. J., Scheffer, G. & Te Velde, E. R. Antral follicle counts are related to age at natural fertility loss and age at menopause. *Menopause* **11**, (2004).
10. Pelosi, E., Simonsick, E., Forabosco, A., Garcia-Ortiz, J. E. & Schlessinger, D. Dynamics of the ovarian reserve and impact of genetic and epidemiological factors on age of menopause. *Biology of Reproduction* vol. 92 at https://doi.org/10.1095/biolreprod.114.127381 (2015).
11. Broekmans, F. J., Soules, M. R. & Fauser, B. C. Ovarian aging: Mechanisms and clinical consequences. *Endocrine Reviews* vol. 30 at https://doi.org/10.1210/er.2009-0006 (2009).
12. Roos, W. P. & Kaina, B. DNA damage-induced cell death: From specific DNA lesions to the DNA damage response and apoptosis. *Cancer Letters* vol. 332 at https://doi.org/10.1016/j.canlet.2012.01.007 (2013).
13. Lips, J. & Kaina, B. DNA double-strand breaks trigger apoptosis in p53-deficient fibroblasts. *Carcinogenesis* **22**, (2001).
14. Wallace, W. H. B. & Kelsey, T. W. Human ovarian reserve from conception to the menopause. *PLoS One* **5**, (2010).
15. Guigon, C. J. & Magre, S. Contribution of germ cells to the differentiation and maturation of the ovary: Insights from models of germ cell depletion. *Biology of Reproduction* vol. 74 at https://doi.org/10.1095/biolreprod.105.047134 (2006).
16. De Felici, M. *et al.* Establishment of oocyte population in the fetal ovary: Primordial germ cell proliferation and oocyte programmed cell death. *Reprod. Biomed. Online* **10**, (2005).
17. Wood, M. A. & Rajkovic, A. Genomic markers of ovarian reserve. *Semin. Reprod. Med.* **31**, (2013).
18. Lee, K. Y. *et al.* MCM8-9 complex promotes resection of double-strand break ends by MRE11-RAD50-NBS1 complex. *Nat. Commun.* **6**, (2015).
19. Kerr, J. B., Myers, M. & Anderson, R. A. The dynamics of the primordial follicle reserve. *Reproduction* vol. 146 at https://doi.org/10.1530/REP-13-0181 (2013).
20. Matsuda, F., Inoue, N., Manabe, N. & Ohkura, S. Follicular growth and atresia in mammalian ovaries: Regulation by survival and death of granulosa cells. *Journal of Reproduction and Development* vol. 58 at https://doi.org/10.1262/jrd.2011-012 (2012).
21. Vaskivuo, T. E. *et al.* Survival of human ovarian follicles from fetal to adult life: Apoptosis, apoptosis-related proteins, and transcription factor GATA-4. *J. Clin. Endocrinol. Metab.* **86**, (2001).
22. Monniaux, D. *et al.* The ovarian reserve of primordial follicles and the dynamic reserve

of antral growing follicles: What is the link? *Biology of Reproduction* vol. 90 at https://doi.org/10.1095/biolreprod.113.117077 (2014).
23. Hale, G. E., Robertson, D. M. & Burger, H. G. The perimenopausal woman: Endocrinology and management. *Journal of Steroid Biochemistry and Molecular Biology* vol. 142 at https://doi.org/10.1016/j.jsbmb.2013.08.015 (2014).
24. De Bruin, J. P. *et al.* The role of genetic factors in age at natural menopause. *Hum. Reprod.* **16**, (2001).
25. Klein, N. A. *et al.* Age-related analysis of inhibin A, inhibin B, and activin A relative to the intercycle monotropic follicle-stimulating hormone rise in normal ovulatory women. in *Journal of Clinical Endocrinology and Metabolism* vol. 89 (2004).
26. Female subfertility. *Nature Reviews Disease Primers* vol. 5 at https://doi.org/10.1038/s41572-019-0062-7 (2019).
27. Faddy, M. J. & Gosden, R. G. A model conforming the decline in follicle numbers to the age of menopause in women. *Hum. Reprod.* **11**, (1996).
28. Treloar, A. E., Boynton, R. E., Behn, B. G. & Brown, B. W. Variation of the human menstrual cycle through reproductive life. *Int. J. Fertil.* **12**, (1970).
29. NHS, G. E. P. Meiosis. https://www.genomicseducation.hee.nhs.uk/glossary/meiosis/ (2022).
30. Te Velde, E. R. & Pearson, P. L. The variability of female reproductive ageing. *Hum. Reprod. Update* **8**, 141–154 (2002).
31. Kong, A. *et al.* Recombination rate and reproductive success in humans. *Nat. Genet.* **36**, (2004).
32. Rowsey, R., Gruhn, J., Broman, K. W., Hunt, P. A. & Hassold, T. Examining variation in recombination levels in the human female: A test of the production-line hypothesis. *Am. J. Hum. Genet.* **95**, (2014).
33. Nagaoka, S. I., Hassold, T. J. & Hunt, P. A. Human aneuploidy: Mechanisms and new insights into an age-old problem. *Nature Reviews Genetics* vol. 13 at https://doi.org/10.1038/nrg3245 (2012).
34. Lambalk, C. B., van Disseldorp, J., de Koning, C. H. & Broekmans, F. J. Testing ovarian reserve to predict age at menopause. *Maturitas* **63**, 280–291 (2009).
35. Luoto, R., Kaprio, J. & Uutela, A. Age at natural menopause and sociodemographic status in Finland. *Am. J. Epidemiol.* **139**, (1994).
36. Tsiligiannis, S., Panay, N. & Stevenson, J. C. Premature Ovarian Insufficiency and Long-Term Health Consequences. *Curr. Vasc. Pharmacol.* **17**, (2019).
37. Rudnicka, E. *et al.* Premature ovarian insufficiency - aetiopathology, epidemiology, and diagnostic evaluation. *Przeglad Menopauzalny* vol. 17 at https://doi.org/10.5114/pm.2018.78550 (2018).
38. Bahamondes, L. & Makuch, M. Y. Infertility care and the introduction of new reproductive technologies in poor resource settings. *Reproductive Biology and Endocrinology* vol. 12 at https://doi.org/10.1186/1477-7827-12-87 (2014).
39. Inhorn, M. C. & Patrizio, P. Infertility around the globe: New thinking on gender, reproductive technologies and global movements in the 21st century. *Hum. Reprod. Update* **21**, (2014).
40. Boivin, J., Bunting, L., Collins, J. A. & Nygren, K. G. International estimates of infertility prevalence and treatment-seeking: Potential need and demand for infertility medical care. *Hum. Reprod.* **22**, (2007).
41. Breart, G. Delayed childbearing. *Eur. J. Obstet. Gynecol. Reprod. Biol.* **75**, (1997).
42. Balasch, J. & Gratacós, E. Delayed childbearing: Effects on fertility and the outcome of pregnancy. *Current Opinion in Obstetrics and Gynecology* vol. 24 at https://doi.org/10.1097/GCO.0b013e3283517908 (2012).
43. Fritz, R. & Jindal, S. Reproductive aging and elective fertility preservation. *Journal of*

Ovarian Research vol. 11 at https://doi.org/10.1186/s13048-018-0438-4 (2018).
44. Broekmans, F. J., Knauff, E. A. H., te Velde, E. R., Macklon, N. S. & Fauser, B. C. Female reproductive ageing: current knowledge and future trends. *Trends in Endocrinology and Metabolism* vol. 18 at https://doi.org/10.1016/j.tem.2007.01.004 (2007).
45. Donnez, J. & Dolmans, M.-M. Fertility Preservation in Women. *N. Engl. J. Med.* **377**, (2017).
46. Yding Andersen, C., Mamsen, L. S. & Kristensen, S. G. Freezing of ovarian tissue and clinical opportunities. *Reproduction* vol. 158 at https://doi.org/10.1530/REP-18-0635 (2019).
47. Henderson, K. D. L., Bernstein, L., Henderson, B., Kolonel, L. & Pike, M. C. Predictors of the timing of natural menopause in the multiethnic cohort study. *Am. J. Epidemiol.* **167**, (2008).
48. Leridon, H. Can assisted reproduction technology compensate for the natural decline in fertility with age? A model assessment. *Hum. Reprod.* **19**, (2004).
49. Argyle, C. E., Harper, J. C. & Davies, M. C. Oocyte cryopreservation: Where are we now? *Hum. Reprod. Update* **22**, (2016).
50. Tan, T. Y., Lau, M. S. K., Loh, S. F. & Tan, H. H. Female ageing and reproductive outcome in assisted reproduction cycles. *Singapore Med. J.* **55**, (2014).
51. Hassold, T. & Chiu, D. Maternal age-specific rates of numerical chromosome abnormalities with special reference to trisomy. *Hum. Genet.* **70**, (1985).
52. Hendriks, D. J. *et al.* Single and repeated GnRH agonist stimulation tests compared with basal markers of ovarian reserve in the prediction of outcome in IVF. *J. Assist. Reprod. Genet.* **22**, (2005).
53. Kwee, J. *et al.* Comparison of endocrine tests with respect to their predictive value on the outcome of ovarian hyperstimulation in IVF treatment: Results of a prospective randomized study. *Hum. Reprod.* **18**, (2003).
54. Broekmans, F. J., Kwee, J., Hendriks, D. J., Mol, B. W. & Lambalk, C. B. A systematic review of tests predicting ovarian reserve and IVF outcome. *Human Reproduction Update* vol. 12 at https://doi.org/10.1093/humupd/dml034 (2006).
55. Seifer, D. B., MacLaughlin, D. T., Christian, B. P., Feng, B. & Shelden, R. M. Early follicular serum müllerian-inhibiting substance levels are associated with ovarian response during assisted reproductive technology cycles. *Fertil. Steril.* **77**, (2002).
56. De Boer, E. J., Den Tonkelaar, I., Te Velde, E. R., Burger, C. W. & Van Leeuwen, F. E. Increased risk of early menopausal transition and natural menopause after poor response at first IVF treatment. *Hum. Reprod.* **18**, (2003).
57. Lawson, R. *et al.* Poor response to ovulation induction is a stronger predictor of early menopause than elevated basal FSH: A life table analysis. *Hum. Reprod.* **18**, (2003).
58. Klinkert, E. R., Broekmans, F. J. M., Looman, C. W. N. & Te Velde, E. R. A poor response in the first in vitro fertilization cycle is not necessarily related to a poor prognosis in subsequent cycles. *Fertil. Steril.* **81**, (2004).
59. Whitcomb, B. W. *et al.* Cigarette Smoking and Risk of Early Natural Menopause. *Am. J. Epidemiol.* **187**, (2018).
60. Mishra, G. D. *et al.* Early menarche, nulliparity and the risk for premature and early natural menopause. *Hum. Reprod.* **32**, (2017).
61. Chang, M., He, L. & Cai, L. An overview of genome-wide association studies. in *Methods in Molecular Biology* vol. 1754 (2018).
62. Mitchell, K. J. What is complex about complex disorders? *Genome Biology* vol. 13 at https://doi.org/10.1186/gb-2012-13-1-237 (2012).
63. Sherry, S. T., Ward, M. & Sirotkin, K. dbSNP - database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Research* vol. 9 at https://doi.org/10.1101/gr.9.8.677 (1999).
64. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, (2001).

65. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, (2007).
66. Claussnitzer, M. *et al.* A brief history of human disease genetics. *Nature* vol. 577 at https://doi.org/10.1038/s41586-019-1879-7 (2020).
67. McVean, G. A. T. *et al.* The Fine-Scale Structure of Recombination Rate Variation in the Human Genome. *Science (80-.).* **304**, (2004).
68. Peltonen, L. & McKusick, V. A. Genomics and medicine: Dissecting human disease in the postgenomic era. *Science* vol. 291 at https://doi.org/10.1126/science.291.5507.1224 (2001).
69. Genetic Science Learning Center. Genetic linkage. https://learn.genetics.utah.edu/content/pigeons/geneticlinkage/ (2022).
70. Barrett, J. C. & Cardon, L. R. Evaluating coverage of genome-wide association studies. *Nat. Genet.* **38**, (2006).
71. Auton, A. *et al.* A global reference for human genetic variation. *Nature* vol. 526 at https://doi.org/10.1038/nature15393 (2015).
72. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, (2021).
73. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics* vol. 101 at https://doi.org/10.1016/j.ajhg.2017.06.005 (2017).
74. Pickrell, J. K. *et al.* Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.* **48**, (2016).
75. Uffelmann, E. *et al.* Genome-wide association studies. *Nat. Rev. Methods Prim.* **1**, (2021).
76. Perry, J. R. B. *et al.* A genome-wide association study of early menopause and the combined impact of identified variants. *Hum. Mol. Genet.* **22**, 1465–1472 (2013).
77. Perry, J. R. B., Murray, A., Day, F. R. & Ong, K. K. Molecular insights into the aetiology of female reproductive ageing. *Nat. Rev. Endocrinol.* **11**, 725–734 (2015).
78. Hartge, P. Genetics of reproductive lifespan. *Nat. Genet.* **41**, 637–638 (2009).
79. Murabito, J. M., Yang, Q., Fox, C., Wilson, P. W. F. & Cupples, L. A. Heritability of age at natural menopause in the framingham heart study. *J. Clin. Endocrinol. Metab.* **90**, (2005).
80. Snieder, H., MacGregor, A. J. & Spector, T. D. Genes Control the Cessation of a Woman's Reproductive Life: A Twin Study of Hysterectomy and Age at Menopause 1 . *J. Clin. Endocrinol. Metab.* **83**, (1998).
81. Van Asselt, K. M. *et al.* Heritability of menopausal age in mothers and daughters. *Fertil. Steril.* **82**, (2004).
82. Ruth, K. S. *et al.* Genetic insights into biological mechanisms governing human ovarian ageing. *Nature* **596**, 393–397 (2021).
83. He, C. *et al.* Genome-wide association studies identify loci associated with age at menarche and age at natural menopause. *Nat. Genet.* **41**, 724–728 (2009).
84. Stolk, L. *et al.* Loci at chromosomes 13, 19 and 20 influence age at natural menopause. *Nat. Genet.* **41**, 645–647 (2009).
85. Chen, C. T. L. *et al.* Replication of loci influencing ages at menarche and menopause in Hispanic women: The Women's Health initiative SHARe study. *Hum. Mol. Genet.* **21**, (2012).
86. Day, F. R. *et al.* Large-scale genomic analyses link reproductive aging to hypothalamic signaling, breast cancer susceptibility and BRCA1-mediated DNA repair. *Nat. Genet.* **47**, 1294–1303 (2015).
87. Frazer, K. A., Murray, S. S., Schork, N. J. & Topol, E. J. Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics* vol. 10 at https://doi.org/10.1038/nrg2554 (2009).
88. Eichler, E. E. *et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics* vol. 11 at https://doi.org/10.1038/nrg2809 (2010).

89. Murray, A. *et al.* Population-based estimates of the prevalence of FMR1 expansion mutations in women with early menopause and primary ovarian insufficiency. *Genet. Med.* **16**, (2014).
90. Murray, A. *et al.* Common genetic variants are significant risk factors for early menopause: results from the Breakthrough Generations Study. *Hum. Mol. Genet.* **20**, 186–192 (2011).
91. Stolk, L. *et al.* Meta-analyses identify 13 loci associated with age at menopause and highlight DNA repair and immune pathways. *Nat. Genet.* **44**, 260–268 (2012).
92. Venturella, R. *et al.* The genetics of non-syndromic primary ovarian insufficiency: A systematic review. *International Journal of Fertility and Sterility* vol. 13 at https://doi.org/10.22074/ijfs.2019.5599 (2019).
93. Titus, S. *et al.* Impairment of BRCA1-related DNA double-strand break repair leads to ovarian aging in mice and humans. *Sci. Transl. Med.* **5**, (2013).
94. Hill, R. J. & Crossan, G. P. DNA cross-link repair safeguards genomic stability during premeiotic germ cell development. *Nat. Genet.* **51**, (2019).
95. Sale, J. E., Lehmann, A. R. & Woodgate, R. Y-family DNA polymerases and their role in tolerance of cellular DNA damage. *Nature Reviews Molecular Cell Biology* vol. 13 at https://doi.org/10.1038/nrm3289 (2012).
96. Wei, K. *et al.* Inactivation of exonuclease I in mice results in DNA mismatch repair defects, increased cancer susceptibility, and male and female sterility. *Genes Dev.* **17**, (2003).
97. Abreu, C. M. *et al.* Shu complex SWS1-SWSAP1 promotes early steps in mouse meiotic recombination. *Nat. Commun.* **9**, (2018).
98. Yoshida, K. *et al.* The mouse RecA-like gene Dmc1 is required for homologous chromosome synapsis during meiosis. *Mol. Cell* **1**, (1998).
99. Tuppi, M. *et al.* Oocyte DNA damage quality control requires consecutive interplay of CHK2 and CK1 to activate p63. *Nat. Struct. Mol. Biol.* **25**, (2018).
100. Rinaldi, V. D., Bloom, J. C. & Schimenti, J. C. Oocyte elimination through DNA damage signaling from CHK1/CHK2 to p53 and p63. *Genetics* **215**, (2020).
101. Bolcun-Filas, E., Rinaldi, V. D., White, M. E. & Schimenti, J. C. Reversal of female infertility by Chk2 ablation reveals the oocyte DNA damage checkpoint pathway. *Science (80-.).* **343**, (2014).
102. Adhikari, D. *et al.* Inhibitory phosphorylation of Cdk1 mediates prolonged prophase I arrest in female germ cells and is essential for female reproductive lifespan. *Cell Res.* **26**, (2016).
103. Rinaldi, V. D., Bolcun-Filas, E., Kogo, H., Kurahashi, H. & Schimenti, J. C. The DNA Damage Checkpoint Eliminates Mouse Oocytes with Chromosome Synapsis Failure. *Mol. Cell* **67**, (2017).
104. Tharp, M. E., Malki, S. & Bortvin, A. Maximizing the ovarian reserve in mice by evading LINE-1 genotoxicity. *Nat. Commun.* **11**, (2020).
105. Liu, Q. *et al.* Chk1 is an essential kinase that is regulated by Atr and required for the G2/M DNA damage checkpoint. *Genes Dev.* **14**, (2000).
106. Chen, L. *et al.* Checkpoint kinase 1 is essential for meiotic cell cycle regulation in mouse oocytes. *Cell Cycle* **11**, (2012).
107. Pacheco, S. *et al.* ATR is required to complete meiotic recombination in mice. *Nat. Commun.* **9**, (2018).
108. Pacheco, S., Maldonado-Linares, A., Garcia-Caldés, M. & Roig, I. ATR function is indispensable to allow proper mammalian follicle development. *Chromosoma* **128**, (2019).
109. Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **20**, 467–484 (2019).
110. Winship, A. L. *et al.* The PARP inhibitor, olaparib, depletes the ovarian reserve in mice: Implications for fertility preservation. *Hum. Reprod.* **35**, (2020).
111. Pietzner, M. *et al.* Mapping the proteo-genomic convergence of human diseases. *Science*

(80-.). **374**, (2021).
112. Jin, C., Wang, X., Hudgins, A., Gamliel, A. & Pei, M. The regulatory landscapes of human ovarian ageing. *BioRxiv* (2022).
113. Shuster, L. T., Rhodes, D. J., Gostout, B. S., Grossardt, B. R. & Rocca, W. A. Premature menopause or early menopause: Long-term health consequences. *Maturitas* vol. 65 at https://doi.org/10.1016/j.maturitas.2009.08.003 (2010).
114. Davies, N. M., Holmes, M. V. & Davey Smith, G. Reading Mendelian randomisation studies: A guide, glossary, and checklist for clinicians. *BMJ* **362**, (2018).
115. Kritz-Silverstein, D. & Barrett-Connor, E. Early menopause, number of reproductive years, and bone mineral density in postmenopausal women. *Am. J. Public Health* **83**, (1993).
116. Svejme, O., Ahlborg, H. G., Nilsson, J. Å. & Karlsson, M. K. Early menopause and risk of osteoporosis, fracture and mortality: A 34-year prospective observational study in 390 women. *BJOG An Int. J. Obstet. Gynaecol.* **119**, (2012).
117. Demir, B. *et al.* Identification of the risk factors for osteoporosis among postmenopausal women. *Maturitas* **60**, (2008).
118. Svejme, O., Ahlborg, H. G., Nilsson, J. Å. & Karlsson, M. K. Low BMD is an independent predictor of fracture and early menopause of mortality in post-menopausal women-A 34-year prospective study. *Maturitas* **74**, (2013).
119. Wellons, M., Ouyang, P., Schreiner, P. J., Herrington, D. M. & Vaidya, D. Early menopause predicts future coronary heart disease and stroke: The Multi-Ethnic Study of Atherosclerosis. *Menopause* **19**, (2012).
120. Løkkegaard, E. *et al.* The association between early menopause and risk of ischaemic heart disease: Influence of Hormone Therapy. *Maturitas* **53**, (2006).
121. Hu, F. B. *et al.* Age at natural menopause and risk of cardiovascular disease. *Arch. Intern. Med.* **159**, (1999).
122. Jacobsen, B. K., Heuch, I. & Kvåle, G. Age at natural menopause and all-cause mortality: A 37-year follow-up of 19,731 Norwegian women. *Am. J. Epidemiol.* **157**, (2003).
123. Manson, J. A. E. *et al.* Menopausal hormone therapy and health outcomes during the intervention and extended poststopping phases of the women's health initiative randomized trials. *JAMA - J. Am. Med. Assoc.* **310**, (2013).
124. Dam, V. *et al.* Association of menopausal characteristics and risk of coronary heart disease: A pan-European case-cohort analysis. *Int. J. Epidemiol.* **48**, (2019).
125. de Kat, A. C. *et al.* Unraveling the associations of age and menopause with cardiovascular risk factors in a large population-based study. *BMC Med.* **15**, (2017).
126. Atsma, F., Bartelink, M. L. E. L., Grobbee, D. E. & Van Der Schouw, Y. T. Postmenopausal status and early menopause as independent risk factors for cardiovascular disease: A meta-analysis. *Menopause* **13**, (2006).
127. Ambikairajah, A., Walsh, E. & Cherbuin, N. Lipid profile differences during menopause: A review with meta-analysis. *Menopause* vol. 26 at https://doi.org/10.1097/GME.0000000000001403 (2019).
128. Pike, C. J. Sex and the development of Alzheimer's disease. *Journal of Neuroscience Research* vol. 95 at https://doi.org/10.1002/jnr.23827 (2017).
129. Tao, X. *et al.* Body mass index and age at natural menopause: A meta-analysis. *Menopause* **22**, (2015).
130. Shadyab, A. H. *et al.* Ages at menarche and menopause and reproductive lifespan as predictors of exceptional longevity in women: The Women's Health Initiative. *Menopause* **24**, (2017).
131. Day, F. R. *et al.* Genomic analyses identify hundreds of variants associated with age at menarche and support a role for puberty timing in cancer risk. *Nat. Genet.* **49**, 834–841 (2017).
132. Morris, D. H. *et al.* Body mass index, exercise, and other lifestyle factors in relation to age at natural menopause: Analyses from the breakthrough generations study. *Am. J. Epidemiol.*

175, (2012).
133. McKinlay, S. M., Bifano, N. L. & McKinlay, J. B. Smoking and age at menopause in women. *Ann. Intern. Med.* **103**, (1985).
134. Stepaniak, U. *et al.* Age at natural menopause in three Central and Eastern European urban populations: The HAPIEE study. *Maturitas* **75**, (2013).
135. Gold, E. B. *et al.* Factors related to age at natural menopause: Longitudinal analyses from SWAN. *Am. J. Epidemiol.* **178**, (2013).
136. Li, L. *et al.* Factors associated with the age of natural menopause and menopausal symptoms in Chinese women. *Maturitas* **73**, (2012).
137. Steiner, A. Z., D'Aloisio, A. A., Deroo, L. A., Sandler, D. P. & Baird, D. D. Association of intrauterine and early-life exposures with age at menopause in the sister study. *Am. J. Epidemiol.* **172**, (2010).
138. Dorjgochoo, T. *et al.* Dietary and lifestyle predictors of age at natural menopause and reproductive span in the Shanghai Women's Health Study. *Menopause* **15**, (2008).
139. Nagel, G., Altenburg, H. P., Nieters, A., Boffetta, P. & Linseisen, J. Reproductive and dietary determinants of the age at menopause in EPIC-Heidelberg. *Maturitas* **52**, (2005).
140. Gudmundsdottir, S. L., Flanders, W. D. & Augestad, L. B. Physical activity and age at menopause: The Nord-Trondelag population-based health study. *Climacteric* **16**, (2013).
141. Hatch, E. E. *et al.* Age at natural menopause in women exposed to diethylstilbestrol in utero. *Am. J. Epidemiol.* **164**, (2006).
142. Sakata, R. *et al.* Effect of radiation on age at menopause among atomic bomb survivors. *Radiat. Res.* **176**, (2011).
143. Wise, P. M., Krajnak, K. M. & Kashon, M. L. Menopause: The aging of multiple pacemakers. *Science (80-.).* **273**, (1996).
144. Zhang, J. *et al.* Can ovarian aging be delayed by pharmacological strategies. *Aging* vol. 11 at https://doi.org/10.18632/aging.101784 (2019).
145. Snowdon, D. A. *et al.* Is early natural menopause a biologic marker of health and aging? *Am. J. Public Health* **79**, (1989).
146. Sievert, L. L. Menopause as a measure of population health: An overview. *Am. J. Hum. Biol.* **13**, (2001).
147. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
148. Elliott, P. & Peakman, T. C. The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. *Int. J. Epidemiol.* **37**, (2008).
149. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med.* **12**, (2015).
150. Wain, L. V. *et al.* Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): A genetic association study in UK Biobank. *Lancet Respir. Med.* **3**, (2015).
151. Welsh, S., Peakman, T., Sheard, S. & Almond, R. Comparison of DNA quantification methodology used in the DNA extraction protocol for the UK Biobank cohort. *BMC Genomics* **18**, (2017).
152. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, (2016).
153. Walter, K. *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, (2015).
154. Loh, P. R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
155. Sul, J. H., Martin, L. S. & Eskin, E. Population structure in genetic studies: Confounding factors and mixed models. *PLoS Genetics* vol. 14 at https://doi.org/10.1371/journal.pgen.1007309 (2018).

156. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, (2011).
157. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, (2012).
158. Szustakowski, J. D. *et al.* Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. *Nat. Genet.* **53**, 942–948 (2021).
159. Backman, J. D. *et al.* Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628–634 (2021).
160. Van Hout, C. V. *et al.* Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* **586**, (2020).
161. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, (2021).
162. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, (2016).
163. Rentzsch, P., Schubach, M., Shendure, J. & Kircher, M. CADD-Splice-improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Med.* **13**, (2021).
164. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
165. Seabold, S. P. J. Statsmodels: Econometric and Statistical Modeling with Python. *SCIPY* (2010).
166. Ference, B. A., Holmes, M. V. & Smith, G. D. Using Mendelian randomization to improve the design of randomized trials. *Cold Spring Harb. Perspect. Biol.* **13**, (2021).
167. Greenland, S. An introduction to instrumental variables for epidemiologists. *Int. J. Epidemiol.* **29**, (2000).
168. Smith, G. D. & Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* **32**, 1–22 (2003).
169. Katan, M. B. Apolipoprotein E isoforms, serum cholesterol, and cancer. *International Journal of Epidemiology* vol. 33 at https://doi.org/10.1093/ije/dyh312 (2004).
170. Thanassoulis, G. & O'Donnell, C. J. Mendelian randomization: Nature's randomized trial in the post-genome era. *JAMA - Journal of the American Medical Association* vol. 301 at https://doi.org/10.1001/jama.2009.812 (2009).
171. Pierce, B. L. & Burgess, S. Efficient design for mendelian randomization studies: Subsample and 2-sample instrumental variable estimators. *Am. J. Epidemiol.* **178**, (2013).
172. Slob, E. A. W. & Burgess, S. A comparison of robust Mendelian randomization methods using summary data. *Genet. Epidemiol.* **44**, 313–329 (2020).
173. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet. Epidemiol.* **37**, (2013).
174. Bowden, J. *et al.* A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Stat. Med.* **36**, 1783–1802 (2017).
175. Burgess, S., Bowden, J., Fall, T., Ingelsson, E. & Thompson, S. G. Sensitivity Analyses for Robust Causal Inference from Mendelian Randomization Analyses with Multiple Genetic Variants. *Epidemiology* **28**, 30–42 (2017).
176. Verbanck, M., Chen, C. Y., Neale, B. & Do, R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat. Genet.* **50**, (2018).
177. Bowden, J., Smith, G. D. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44**, 512–525 (2015).
178. Burgess, S. & Thompson, S. G. Interpreting findings from Mendelian randomization using the MR-Egger method. *Eur. J. Epidemiol.* **32**, (2017).
179. Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent Estimation in

Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet. Epidemiol.* **40**, 304–314 (2016).
180. Cho, Y. *et al.* Exploiting horizontal pleiotropy to search for causal pathways within a Mendelian randomization framework. *Nat. Commun.* **11**, (2020).
181. Bowden, J. *et al.* Improving the visualization, interpretation and analysis of two-sample summary data Mendelian randomization via the Radial plot and Radial regression. *Int. J. Epidemiol.* **47**, 1264–1278 (2018).
182. Lunetta, K. L. *et al.* Genetic correlates of longevity and selected age-related phenotypes: a genome-wide association study in the Framingham Study. *BMC Med. Genet.* **8 Suppl 1**, (2007).
183. Horikoshi, M. *et al.* Elucidating the genetic architecture of reproductive ageing in the Japanese population. *Nat. Commun.* **9**, (2018).
184. Perry, J. R. B. *et al.* DNA mismatch repair gene MSH6 implicated in determining age at natural menopause. *Hum. Mol. Genet.* **23**, 2490–2497 (2014).
185. Cano-Gamez, E. & Trynka, G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Frontiers in Genetics* vol. 11 at https://doi.org/10.3389/fgene.2020.00424 (2020).
186. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science (80-.).* **337**, (2012).
187. Slatkin, M. Linkage disequilibrium - Understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics* vol. 9 at https://doi.org/10.1038/nrg2361 (2008).
188. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: Study designs and statistical tests. *American Journal of Human Genetics* vol. 95 at https://doi.org/10.1016/j.ajhg.2014.06.009 (2014).
189. McMahon, A. *et al.* Sequencing-based genome-wide association studies reporting standards. *Cell Genomics* **1**, (2021).
190. Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci. U. S. A.* **109**, (2012).
191. Ward, L. D. *et al.* Rare coding variants in DNA damage repair genes associated with timing of natural menopause. *HGG Adv.* **3**, (2021).
192. Ma, C., Blackwell, T., Boehnke, M. & Scott, L. J. Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genet. Epidemiol.* **37**, 539–550 (2013).
193. Firth, D. Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38 (1993).
194. Kosmidis, I. K. P. E. C. S. N. Mean and median bias reduction in generalized linear models. *Stat. Comput.* **30**, 43–59 (2019).
195. Li, L. *et al.* Single-Cell RNA-Seq Analysis Maps Development of Human Germline Cells and Gonadal Niche Interactions. *Cell Stem Cell* **20**, 858-873.e4 (2017).
196. Zhang, Y. *et al.* Transcriptome Landscape of Human Folliculogenesis Reveals Oocyte and Granulosa Cell Interactions. *Mol. Cell* **72**, 1021-1034.e4 (2018).
197. Pickrell, J. K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* **94**, 559–573 (2014).
198. Lou, S. *et al.* TopicNet: a framework for measuring transcriptional regulatory network change. *Bioinformatics* **36**, I474–I481 (2020).
199. Reshef, Y. A. *et al.* Detecting genome-wide directional effects of transcription factor binding on polygenic disease risk. *Nat. Genet.* **50**, 1483–1493 (2018).
200. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
201. Chen, D. *et al.* Human Primordial Germ Cells Are Specified from Lineage-Primed Progenitors. *Cell Rep.* **29**, 4568-4582.e5 (2019).
202. Chen, D. *et al.* The TFAP2C-Regulated OCT4 Naive Enhancer Is Involved in Human Germline Formation. *Cell Rep.* **25**, 3591-3602.e5 (2018).

203. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
204. Castro-Mondragon, J. A. *et al.* JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **50**, D165–D173 (2022).
205. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
206. Boix, C. A., James, B. T., Park, Y. P., Meuleman, W. & Kellis, M. Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature* **590**, 300–307 (2021).
207. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
208. Layer, R. M. *et al.* GIGGLE: a search engine for large-scale integrated genome analysis. *Nat. Methods* **15**, 123–126 (2018).
209. Kaplanis, J. *et al.* Genetic and chemotherapeutic causes of germline hypermutation. *bioRxiv* (2021).
210. Maier, V. K. *et al.* Functional Proteomic Analysis of Repressive Histone Methyltransferase Complexes Reveals ZNF518B as a G9A Regulator. *Mol. Cell. Proteomics* **14**, 1435–1446 (2015).
211. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
212. Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* **46**, D794–D801 (2018).
213. Codd, V. *et al.* Polygenic basis and biomedical consequences of telomere length variation. *Nat. Genet.* **53**, 1425–1433 (2021).
214. McNally, E. J., Luncsford, P. J. & Armanios, M. Long telomeres and cancer risk: The price of cellular immortality. *J. Clin. Invest.* **129**, (2019).
215. Aydos, S. E., Elhan, A. H. & Tükün, A. Is telomere length one of the determinants of reproductive life span? *Arch. Gynecol. Obstet.* **272**, (2005).
216. Olsen, K. W. *et al.* Identification of a unique epigenetic profile in women with diminished ovarian reserve. *Fertil. Steril.* **115**, (2021).
217. Sherman, S. L. Premature ovarian failure in the fragile X syndrome. *Am. J. Med. Genet.* **97**, 189–194 (2000).
218. Kichaev, G. *et al.* Leveraging Polygenic Functional Enrichment to Improve GWAS Power. *Am. J. Hum. Genet.* **104**, (2019).
219. Hara, S., Yoda, E., Sasaki, Y., Nakatani, Y. & Kuwata, H. Calcium-independent phospholipase A 2 γ (iPLA 2 γ) and its roles in cellular functions and diseases. *Biochim. Biophys. acta. Mol. cell Biol. lipids* **1864**, 861–868 (2019).
220. Mancuso, D. J. *et al.* Genetic ablation of calcium-independent phospholipase A2gamma leads to alterations in mitochondrial lipid metabolism and function resulting in a deficient mitochondrial bioenergetic phenotype. *J. Biol. Chem.* **282**, 34611–34622 (2007).
221. Shukla, A., Saneto, R. P., Hebbar, M., Mirzaa, G. & Girisha, K. M. A neurodegenerative mitochondrial disease phenotype due to biallelic loss-of-function variants in PNPLA8 encoding calcium-independent phospholipase A2γ. *Am. J. Med. Genet. A* **176**, 1232–1237 (2018).
222. Saunders, C. J. *et al.* Loss of function variants in human PNPLA8 encoding calcium-independent phospholipase A2 γ recapitulate the mitochondriopathy of the homologous null mouse. *Hum. Mutat.* **36**, 301–306 (2015).
223. Masih, S., Moirangthem, A. & Phadke, S. R. Homozygous Missense Variation in PNPLA8 Causes Prenatal-Onset Severe Neurodegeneration. *Mol. Syndromol.* **12**, 174–178 (2021).
224. Bonventre, J. V. *et al.* Reduced fertility and postischaemic brain injury in mice deficient in cytosolic phospholipase a2. *Nature* **390**, (1997).
225. Bao, S. *et al.* Male mice that do not express Group VIA Phospholipase A2 produce

spermatozoa with impaired motility and have greatly reduced fertility. *J. Biol. Chem.* **279**, (2004).
226. Rantakari, P. *et al.* Inactivation of Palb2 gene leads to mesoderm differentiation defect and early embryonic lethality in mice. *Hum. Mol. Genet.* **19**, (2010).
227. Reid, S. *et al.* Biallelic mutations in PALB2 cause Fanconi anemia subtype FA-N and predispose to childhood cancer. *Nat. Genet.* **39**, 162–164 (2007).
228. Bass, T. E. *et al.* ETAA1 acts at stalled replication forks to maintain genome integrity. *Nat. Cell Biol.* **18**, 1185–1195 (2016).
229. Feng, S. *et al.* Ewing Tumor-associated Antigen 1 Interacts with Replication Protein A to Promote Restart of Stalled Replication Forks. *J. Biol. Chem.* **291**, 21956–21962 (2016).
230. Haahr, P. *et al.* Activation of the ATR kinase by the RPA-binding protein ETAA1. *Nat. Cell Biol.* **18**, 1196–1207 (2016).
231. Saldivar, J. C. *et al.* An intrinsic S/G 2 checkpoint enforced by ATR. *Science* **361**, 806–810 (2018).
232. Michelena, J., Gatti, M., Teloni, F., Imhof, R. & Altmeyer, M. Basal CHK1 activity safeguards its stability to maintain intrinsic S-phase checkpoint functions. *J. Cell Biol.* **218**, (2019).
233. Hustedt, N. *et al.* Control of homologous recombination by the HROB-MCM8-MCM9 pathway. *Genes Dev.* **33**, 1397–1415 (2019).
234. Huang, J. W. *et al.* MCM8IP activates the MCM8-9 helicase to promote DNA synthesis and homologous recombination upon DNA damage. *Nat. Commun.* **11**, (2020).
235. Tucker, E. J. *et al.* Meiotic genes in premature ovarian insufficiency: variants in HROB and REC8 as likely genetic causes. *Eur. J. Hum. Genet.* **30**, 219–228 (2022).
236. Fauchereau, F. *et al.* A non-sense MCM9 mutation in a familial case of primary ovarian insufficiency. *Clin. Genet.* **89**, (2016).
237. Wood-Trageser, M. A. *et al.* MCM9 mutations are associated with ovarian failure, short stature, and chromosomal instability. *Am. J. Hum. Genet.* **95**, (2014).
238. AlAsiri, S. *et al.* Exome sequencing reveals MCM8 mutation underlies ovarian failure and chromosomal instability. *J. Clin. Invest.* **125**, (2015).
239. Imielinski, M. *et al.* Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150**, 1107–1120 (2012).
240. Hazeslip, L., Zafar, M. K., Chauhan, M. Z. & Byrd, A. K. Genome maintenance by DNA helicase B. *Genes* vol. 11 at https://doi.org/10.3390/genes11050578 (2020).
241. Taneja, P. *et al.* A dominant-negative mutant of human DNA helicase B blocks the onset of chromosomal DNA replication. *J. Biol. Chem.* **277**, (2002).
242. Tkáč, J. *et al.* HELB Is a Feedback Inhibitor of DNA End Resection. *Mol. Cell* **61**, (2016).
243. Clifford, R. *et al.* SAMHD1 is mutated recurrently in chronic lymphocytic leukemia and is involved in response to DNA damage. *Blood* **123**, 1021–1031 (2014).
244. Schott, K. *et al.* SAMHD1 in cancer: curse or cure? *J. Mol. Med. (Berl).* **100**, 351–372 (2022).
245. Sjöblom, T. *et al.* The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268–274 (2006).
246. Parsons, D. W. *et al.* The genetic landscape of the childhood cancer medulloblastoma. *Science* **331**, 435–439 (2011).
247. Coggins, S. A., Mahboubi, B., Schinazi, R. F. & Kim, B. SAMHD1 Functions and Human Diseases. *Viruses* **12**, (2020).
248. Rice, G. I. *et al.* Mutations involved in Aicardi-Goutières syndrome implicate SAMHD1 as regulator of the innate immune response. *Nat. Genet.* **41**, 829–832 (2009).
249. Bononi, A. *et al.* BAP1 regulates IP3R3-mediated Ca 2+ flux to mitochondria suppressing cell transformation. *Nature* **546**, (2017).
250. Betti, M. *et al.* Genetic predisposition for malignant mesothelioma: A concise review.

Mutation Research - Reviews in Mutation Research vol. 781 at https://doi.org/10.1016/j.mrrev.2019.03.001 (2019).
251. Franzolin, E. *et al.* The deoxynucleotide triphosphohydrolase SAMHD1 is a major regulator of DNA precursor pools in mammalian cells. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 14272–14277 (2013).
252. Kumar, D. *et al.* Mechanisms of mutagenesis in vivo due to imbalanced dNTP pools. *Nucleic Acids Res.* **39**, 1360–1371 (2011).
253. Coquel, F. *et al.* SAMHD1 acts at stalled replication forks to prevent interferon induction. *Nature* **557**, 57–61 (2018).
254. Mathews, C. K. Deoxyribonucleotide metabolism, mutagenesis and cancer. *Nat. Rev. Cancer* **15**, 528–539 (2015).
255. Mao, Z., Bozzella, M., Seluanov, A. & Gorbunova, V. DNA repair by nonhomologous end joining and homologous recombination during cell cycle in human cells. *Cell Cycle* **7**, 2902–2906 (2008).
256. Daddacha, W. *et al.* SAMHD1 Promotes DNA End Resection to Facilitate DNA Repair by Homologous Recombination. *Cell Rep.* **20**, 1921–1935 (2017).
257. Bonifati, S. *et al.* SAMHD1 controls cell cycle status, apoptosis and HIV-1 infection in monocytic THP-1 cells. *Virology* **495**, 92–100 (2016).
258. Kodigepalli, K. M., Li, M., Liu, S. L. & Wu, L. Exogenous expression of SAMHD1 inhibits proliferation and induces apoptosis in cutaneous T-cell lymphoma-derived HuT78 cells. *Cell Cycle* **16**, 179–188 (2017).
259. Taub, M. A. *et al.* Genetic determinants of telomere length from 109,122 ancestrally diverse whole-genome sequences in TOPMed. *Cell Genomics* **2**, (2022).
260. Butts, S. *et al.* Correlation of telomere length and telomerase activity with occult ovarian insufficiency. *J. Clin. Endocrinol. Metab.* **94**, (2009).
261. Keefe, D. L., Marquard, K. & Liu, L. The telomere theory of reproductive senescence in women. *Current Opinion in Obstetrics and Gynecology* vol. 18 at https://doi.org/10.1097/01.gco.0000193019.05686.49 (2006).
262. Rahbari, R. *et al.* Timing, rates and spectra of human germline mutation. *Nat. Genet.* **48**, 126–133 (2016).
263. Crow, J. F. The origins, patterns and implications of human spontaneous mutation. *Nat. Rev. Genet.* **1**, 40–47 (2000).
264. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014).
265. Fromer, M. *et al.* De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179–184 (2014).
266. Wang, W., Corominas, R. & Lin, G. N. De novo Mutations From Whole Exome Sequencing in Neurodevelopmental and Psychiatric Disorders: From Discovery to Application. *Front. Genet.* **10**, (2019).
267. Coe, B. P. *et al.* Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. *Nat. Genet.* **51**, 106–116 (2019).
268. Linschooten, J. O. *et al.* Paternal lifestyle as a potential source of germline mutations transmitted to offspring. *FASEB J.* **27**, 2873–2879 (2013).
269. Miao, Y. *et al.* BRCA2 deficiency is a potential driver for human primary ovarian insufficiency. *Cell Death Dis.* **10**, (2019).
270. Lin, W., Titus, S., Moy, F., Ginsburg, E. S. & Oktay, K. Ovarian Aging in Women With BRCA Germline Mutations. *J. Clin. Endocrinol. Metab.* **102**, 3839–3847 (2017).
271. Wesevich, V., Kellen, A. N. & Pal, L. Recent advances in understanding primary ovarian insufficiency. *F1000Research* vol. 9 at https://doi.org/10.12688/f1000research.26423.1 (2020).
272. Coulam, C. B., Adamson, S. C. & Annegers, J. F. Incidence of premature ovarian failure. *Obstet. Gynecol.* **67**, (1986).

273. Golezar, S., Ramezani Tehrani, F., Khazaei, S., Ebadi, A. & Keshavarz, Z. The global prevalence of primary ovarian insufficiency and early menopause: a meta-analysis. *Climacteric* **22**, (2019).
274. Shelling, A. N. Premature ovarian failure. *Reproduction* vol. 140 at https://doi.org/10.1530/REP-09-0567 (2010).
275. Harlow, B. L. & Signorello, L. B. Factors associated with early menopause. *Maturitas* vol. 35 at https://doi.org/10.1016/S0378-5122(00)00092-X (2000).
276. Goswami, D. & Conway, G. S. Premature ovarian failure. *Hormone Research* vol. 68 at https://doi.org/10.1159/000102537 (2007).
277. Szeliga, A. *et al.* Autoimmune diseases in patients with premature ovarian insufficiency—our current state of knowledge. *International Journal of Molecular Sciences* vol. 22 at https://doi.org/10.3390/ijms22052594 (2021).
278. Kirshenbaum, M. & Orvieto, R. Premature ovarian insufficiency (POI) and autoimmunity-an update appraisal. *Journal of Assisted Reproduction and Genetics* vol. 36 at https://doi.org/10.1007/s10815-019-01572-0 (2019).
279. Domniz, N. & Meirow, D. Premature ovarian insufficiency and autoimmune diseases. *Best Practice and Research: Clinical Obstetrics and Gynaecology* vol. 60 at https://doi.org/10.1016/j.bpobgyn.2019.07.008 (2019).
280. Qin, Y. *et al.* ESR1, HK3 and BRSK1 gene variants are associated with both age at natural menopause and premature ovarian failure. *Orphanet J. Rare Dis.* **7**, (2012).
281. Van Kasteren, Y. M. *et al.* Familial idiopathic premature ovarian failure: An overrated and underestimated genetic disease? *Hum. Reprod.* **14**, (1999).
282. Pu, D., Xing, Y., Gao, Y., Gu, L. & Wu, J. Gene variation and premature ovarian failure: A meta-analysis. *European Journal of Obstetrics and Gynecology and Reproductive Biology* vol. 182 at https://doi.org/10.1016/j.ejogrb.2014.09.036 (2014).
283. Chapman, C., Cree, L. & Shelling, A. N. The genetics of premature ovarian failure: Current perspectives. *International Journal of Women's Health* vol. 7 at https://doi.org/10.2147/IJWH.S64024 (2015).
284. Liu, H. *et al.* Whole-exome sequencing in patients with premature ovarian insufficiency: Early detection and early intervention. *J. Ovarian Res.* **13**, (2020).
285. França, M. M. & Mendonca, B. B. Genetics of primary ovarian insufficiency in the next-generation sequencing era. *J. Endocr. Soc.* **4**, (2020).
286. Jin, H. *et al.* Identification of potential causal variants for premature ovarian failure by whole exome sequencing. *BMC Med. Genomics* **13**, (2020).
287. Patiño, L. C. *et al.* New mutations in non-syndromic primary ovarian insufficiency patients identified via whole-exome sequencing. *Hum. Reprod.* **32**, (2017).
288. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, (2018).
289. Gardner, E. *et al.* Damaging missense variants in IGF1R implicate a role for IGF1 resistance in the aetiology of type 2 diabetes. *MedRxiv* (2022).
290. Li, X. *et al.* Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nat. Genet.* **52**, (2020).
291. Gardner, E. J. *et al.* Reduced reproductive success is associated with selective constraint on human genes. *Nature* **603**, (2022).
292. Minikel, E. V. *et al.* Evaluating drug targets through human loss-of-function genetic variation. *Nature* **581**, (2020).
293. Korhonen, J. A., Gaspari, M. & Falkenberg, M. TWINKLE has 5′ → 3′ DNA helicase activity and is specifically stimulated by mitochondrial single-stranded DNA-binding protein. *J. Biol. Chem.* **278**, (2003).
294. Khan, I. *et al.* Biochemical Characterization of the Human Mitochondrial Replicative

Twinkle Helicase. *J. Biol. Chem.* **291**, (2016).
295. Bashamboo, A. & McElreavey, K. NR5A1/SF-1 and development and function of the ovary. *Ann. Endocrinol. (Paris).* **71**, (2010).
296. Choi, Y., Yuan, D. & Rajkovic, A. Germ cell-specific transcriptional regulator Sohlh2 is essential for early mouse folliculogenesis and oocyte-specific gene expression. *Biol. Reprod.* **79**, (2008).
297. Cerván-Martín, M. *et al.* Intronic variation of the SOHLH2 gene confers risk to male reproductive impairment. *Fertil. Steril.* **114**, (2020).
298. Stark, Z. *et al.* Scaling national and international improvement in virtual gene panel curation via a collaborative approach to discordance resolution. *American Journal of Human Genetics* vol. 108 at https://doi.org/10.1016/j.ajhg.2021.06.020 (2021).
299. Serpieri, V. *et al.* SUFU haploinsufficiency causes a recognisable neurodevelopmental phenotype at the mild end of the Joubert syndrome spectrum . *J. Med. Genet.* (2021) doi:10.1136/jmedgenet-2021-108114.
300. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, (2015).
301. Shim, J. Y., Laufer, M. R. & Grimstad, F. W. Dysmenorrhea and Endometriosis in Transgender Adolescents. *J. Pediatr. Adolesc. Gynecol.* **33**, (2020).
302. Webber, L. *et al.* ESHRE Guideline: Management of women with premature ovarian insufficiency. *Human Reproduction* vol. 31 at https://doi.org/10.1093/humrep/dew027 (2016).
303. Woad, K. J., Watkins, W. J., Prendergast, D. & Shelling, A. N. The genetic basis of premature ovarian failure. *Aust. New Zeal. J. Obstet. Gynaecol.* **46**, (2006).
304. Qin, Y., Jiao, X., Simpson, J. L. & Chen, Z. J. Genetics of primary ovarian insufficiency: New developments and opportunities. *Hum. Reprod. Update* **21**, (2015).
305. Lourenço, D. *et al.* Mutations in NR5A1 Associated with Ovarian Insufficiency . *N. Engl. J. Med.* **360**, (2009).
306. Ferraz-de-Souza, B., Lin, L. & Achermann, J. C. Steroidogenic factor-1 (SF-1, NR5A1) and human disease. *Molecular and Cellular Endocrinology* vol. 336 at https://doi.org/10.1016/j.mce.2010.11.006 (2011).
307. Boehm, U. *et al.* Expert consensus document: European Consensus Statement on congenital hypogonadotropic hypogonadism-pathogenesis, diagnosis and treatment. *Nature Reviews Endocrinology* vol. 11 at https://doi.org/10.1038/nrendo.2015.112 (2015).
308. Camats, N., Fernández-Cancio, M., Audí, L., Schaller, A. & Flück, C. E. Broad phenotypes in heterozygous NR5A1 46,XY patients with a disorder of sex development: an oligogenic origin? *Eur. J. Hum. Genet.* **26**, (2018).
309. Domenice, S. *et al.* Wide spectrum of NR5A1-related phenotypes in 46,XY and 46,XX individuals. *Birth Defects Research Part C - Embryo Today: Reviews* vol. 108 at https://doi.org/10.1002/bdrc.21145 (2016).
310. Jeyasuria, P. *et al.* Cell-specific knockout of steriodogenic factor 1 reveals its essential roles in gonadal function. *Mol. Endocrinol.* **18**, (2004).
311. Biason-Lauber, S. & Schoenle, E. Apparently normal ovarian differentiation in a prepubertal girl with transcriptionally inactive steroidogenic factor 1 (NR5A1/SF-1) and adrenocortical insufficiency. *Am. J. Hum. Genet.* **67**, (2000).
312. Choi, Y. *et al.* Microarray analyses of newborn mouse ovaries lacking Nobox. *Biol. Reprod.* **77**, (2007).
313. Pangas, S. A. *et al.* Oogenesis requires germ cell-specific transcriptional regulators Sohlh1 and Lhx8. *Proc. Natl. Acad. Sci. U. S. A.* **103**, (2006).
314. Lerat, J. *et al.* An Application of NGS for Molecular Investigations in Perrault Syndrome: Study of 14 Families and Review of the Literature. *Hum. Mutat.* **37**, (2016).
315. Luborsky, J. L., Meyer, P., Sowers, M. F., Gold, E. B. & Santoro, N. Premature

menopause in a multi-ethnic population study of the menopause transition. *Hum. Reprod.* **18**, (2003).
316. França, M. M. *et al.* Screening of targeted panel genes in Brazilian patients with primary ovarian insufficiency. *PLoS One* **15**, (2020).
317. Fonseca, D. J. *et al.* Next generation sequencing in women affected by nonsyndromic premature ovarian failure displays new potential causative genes and mutations. *Fertil. Steril.* **104**, (2015).
318. Stankovic, S. *et al.* Genetic susceptibility to earlier ovarian ageing increases de novo mutation rate in offspring. *MedRxiv* (2022).
319. Torres-Ruiz, R. & Rodriguez-Perales, S. CRISPR-Cas9 technology: Applications and human disease modelling. *Brief. Funct. Genomics* **16**, (2017).
320. Seah, Y. F. S., El Farran, C. A., Warrier, T., Xu, J. & Loh, Y. H. Induced pluripotency and gene editing in disease modelling: Perspectives and challenges. *Int. J. Mol. Sci.* **16**, (2015).
321. Sterneckert, J. L., Reinhardt, P. & Schöler, H. R. Investigating human disease using stem cell models. *Nature Reviews Genetics* vol. 15 at https://doi.org/10.1038/nrg3764 (2014).
322. Barbeira, A. N. *et al.* Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. *Genome Biol.* **22**, (2021).
323. Abell, N. S. *et al.* Multiple causal variants underlie genetic associations in humans. *Science (80-.).* **375**, (2022).
324. Emilsson, V. *et al.* Co-regulatory networks of human serum proteins link genetics to disease. *Science (80-.).* **361**, (2018).
325. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, (2018).
326. Suhre, K. *et al.* Connecting genetic risk to disease end points through the human blood plasma proteome. *Nat. Commun.* **8**, (2017).
327. Boschetti, E., D'Amato, A., Candiano, G. & Righetti, P. G. Protein biomarkers for early detection of diseases: The decisive contribution of combinatorial peptide ligand libraries. *Journal of Proteomics* vol. 188 at https://doi.org/10.1016/j.jprot.2017.08.009 (2018).
328. Rifai, N., Gillette, M. A. & Carr, S. A. Protein biomarker discovery and validation: The long and uncertain path to clinical utility. *Nature Biotechnology* vol. 24 at https://doi.org/10.1038/nbt1235 (2006).
329. Folkersen, L. *et al.* Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease. *PLoS Genet.* **13**, (2017).
330. Gilly, A. *et al.* Whole-genome sequencing analysis of the cardiometabolic proteome. *Nat. Commun.* **11**, (2020).
331. Folkersen, L. *et al.* Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. *Nat. Metab.* **2**, (2020).
332. Yao, C. *et al.* Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nat. Commun.* **9**, (2018).
333. Lindsay, T. *et al.* Descriptive epidemiology of physical activity energy expenditure in UK adults (The Fenland study). *Int. J. Behav. Nutr. Phys. Act.* **16**, (2019).
334. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **53**, 1097–1103 (2021).
335. Jiang, L. *et al.* A resource-efficient tool for mixed model association analysis of large-scale data. *Nat. Genet.* **51**, (2019).
336. Ferkingstad, E. *et al.* Large-scale integration of the plasma proteome with genetics and disease. *Nat. Genet.* **53**, (2021).
337. Koprulu, M., Zanini, J., Wheeler, E. & Langenberg, C. From genome to phenome via the proteome: broad capture, antibody-based proteomics to explore disease mechanisms.
338. Dong, Y. *et al.* Mannose receptor C type 2 mediates 1,25(OH) 2 D 3 /vitamin D receptor-regulated collagen metabolism through collagen type 5, alpha 2 chain and matrix metalloproteinase 13 in murine MC3T3-E1 cells. *Mol. Cell. Endocrinol.* **483**, (2019).

339. Serafini, T. *et al.* The netrins define a family of axon outgrowth-promoting proteins homologous to C. elegans UNC-6. *Cell* **78**, (1994).
340. Bregin, A. *et al.* Expression and impact of Lsamp neural adhesion molecule in the serotonergic neurotransmission system. *Pharmacol. Biochem. Behav.* **198**, (2020).
341. Ruan, X. & Jiang, J. RACGAP1 promotes proliferation and cell cycle progression by regulating CDC25C in cervical cancer cells. (2022).
342. Lorès, P. *et al.* Deletion of MgcRacGAP in the male germ cells impairs spermatogenesis and causes male sterility in the mouse. *Dev. Biol.* **386**, (2014).
343. Ritari, J., Koskela, S., Hyvärinen, K., FinnGen & Partanen, J. HLA-disease association and pleiotropy landscape in over 235,000 Finns. *Hum. Immunol.* **83**, (2022).
344. Sun, B. et al. Genetic regulation of the human plasma proteome in 54,306 UK Biobank participants. *BioRxiv* (2022).
345. Zhang, J. *et al.* Plasma proteome analyses in individuals of European and African ancestry identify cis-pQTLs and models for proteome-wide association studies. *Nat. Genet.* **54**, (2022).
346. Elks, C. E. *et al.* Age at menarche and type 2 diabetes risk: The EPIC-InterAct study. *Diabetes Care* **36**, (2013).
347. Gajbhiye, R., Fung, J. N. & Montgomery, G. W. Complex genetics of female fertility. *npj Genomic Medicine* vol. 3 at https://doi.org/10.1038/s41525-018-0068-1 (2018).
348. Day, F. R., Elks, C. E., Murray, A., Ong, K. K. & Perry, J. R. B. Puberty timing associated with diabetes, cardiovascular disease and also diverse health outcomes in men and women: The UK Biobank study. *Sci. Rep.* **5**, (2015).
349. Yatsenko, S. A. & Rajkovic, A. Reproductive aging and MCM8/9. *Oncotarget* vol. 6 at https://doi.org/10.18632/oncotarget.4589 (2015).
350. Ruth, K. S. *et al.* Events in Early Life are Associated with Female Reproductive Ageing: A UK Biobank Study. *Sci. Rep.* **6**, (2016).
351. McGrath, I. M., Mortlock, S. & Montgomery, G. W. Genetic regulation of physiological reproductive lifespan and female fertility. *Int. J. Mol. Sci.* **22**, (2021).
352. Day, F. R. *et al.* Shared genetic aetiology of puberty timing between sexes and with health-related outcomes. *Nat. Commun.* **6**, (2015).
353. Yatsenko, S. A. & Rajkovic, A. Genetics of human female infertility. *Biology of Reproduction* vol. 101 at https://doi.org/10.1093/biolre/ioz084 (2019).
354. Parent, A. S. *et al.* The Timing of Normal Puberty and the Age Limits of Sexual Precocity: Variations around the World, Secular Trends, and Changes after Migration. *Endocrine Reviews* vol. 24 at https://doi.org/10.1210/er.2002-0019 (2003).
355. Lunetta, K. L. *et al.* Rare coding variants and X-linked loci associated with age at menarche. *Nat. Commun.* **6**, (2015).
356. Ruth, K. S. & Murray, A. Lessons from Genome-Wide Association Studies in Reproductive Medicine: Menopause. *Semin. Reprod. Med.* **34**, (2016).
357. Hardy, R. & Kuh, D. Reproductive characteristics and the age at inception of the perimenopause in a British national cohort. *Am. J. Epidemiol.* **149**, (1999).
358. Farahmand, M., Tehrani, F. R., Pourrajabi, L., Najafi, M. & Azizi, F. Factors associated with menopausal age in Iranian women: Tehran Lipid and Glucose Study. *J. Obstet. Gynaecol. Res.* **39**, (2013).
359. Snieder, H., Macgregor, A. J. & Spector, T. D. Genes control the cessation of a woman's reproductive life: A twin study of hysterectomy and age at menopause. *J. Clin. Endocrinol. Metab.* **83**, (1998).
360. He, L. N. *et al.* Association study of the oestrogen signalling pathway genes in relation to age at natural menopause. *J. Genet.* **86**, (2007).
361. Lappalainen, T. & MacArthur, D. G. From variant to function in human disease genetics. *Science* vol. 373 at https://doi.org/10.1126/science.abi8207 (2021).

362. Winkler, T. W. *et al.* Quality control and conduct of genome-wide association meta-analyses. *Nat. Protoc.* **9**, (2014).
363. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, (2010).
364. Purcell, S. *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, (2007).
365. Stacey, D. *et al.* ProGeM: A framework for the prioritization of candidate causal genes at molecular quantitative trait loci. *Nucleic Acids Res.* **47**, (2019).
366. Aragam, K. et al. Discovery and systematic characterization of risk variants and genes for coronary artery disease in over a million participants. *MedRxiv* (2021).
367. Nasser, J. *et al.* Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, (2021).
368. Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M. & Ng, P. C. SIFT missense predictions for genomes. *Nat. Protoc.* **11**, (2016).
369. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* (2013) doi:10.1002/0471142905.hg0720s76.
370. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, (2018).
371. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science (80-.).* **348**, (2015).
372. Vosa, U. et al. Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. *BioRxiv* (2018).
373. Qi, T. *et al.* Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. *Nat. Commun.* **9**, (2018).
374. Meng, X. H. *et al.* Integration of summary data from GWAS and eQTL studies identified novel causal BMD genes with functional predictions. *Bone* **113**, (2018).
375. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet.* **10**, (2014).
376. Weeks, E. et al. Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. *MedRxiv* (2020).
377. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLoS Comput. Biol.* **11**, (2015).
378. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nature Genetics* vol. 45 at https://doi.org/10.1038/ng.2653 (2013).
379. Day, F. *et al.* Large-scale genome-wide meta-analysis of polycystic ovary syndrome suggests shared genetic architecture for different diagnosis criteria. *PLoS Genet.* **14**, (2018).
380. Mathieson, I. Genome-wide analysis identifies genetic effects on reproductive success and ongoing natural selection at the FADS locus. (2020).
381. Mbarek, H. *et al.* Identification of Common Genetic Variants Influencing Spontaneous Dizygotic Twinning and Female Fertility. *Am. J. Hum. Genet.* **98**, (2016).
382. Ming, M. & He, Y. Y. PTEN in DNA damage repair. *Cancer Letters* vol. 319 at https://doi.org/10.1016/j.canlet.2012.01.003 (2012).
383. Herbison, A. E. Control of puberty onset and fertility by gonadotropin-releasing hormone neurons. *Nature Reviews Endocrinology* vol. 12 at https://doi.org/10.1038/nrendo.2016.70 (2016).
384. Chaudhary, S. *et al.* FSH-stimulated Inhibin B (FSH-iB): A Novel Marker for the Accurate Prediction of Pubertal Outcome in Delayed Puberty. *J. Clin. Endocrinol. Metab.* **106**, (2021).
385. Das, N. & Kumar, T. R. Molecular regulation of follicle-stimulating hormone synthesis, secretion and action. *Journal of Molecular Endocrinology* vol. 60 at https://doi.org/10.1530/JME-17-0308 (2018).

386. Arlt, W. *et al.* Congenital adrenal hyperplasia caused by mutant P450 oxidoreductase and human androgen synthesis: Analytical study. *Lancet* **363**, (2004).
387. Tang, R. *et al.* Nectin-like molecule 2, a necessary sexual maturation regulator, participates in congenital hypogonadotropic hypogonadism. *Gene* **754**, (2020).
388. Oleari, R. *et al.* A Novel SEMA3G Mutation in Two Siblings Affected by Syndromic GnRH Deficiency. *Neuroendocrinology* **111**, (2021).
389. Kotan, L. D. *et al.* Idiopathic hypogonadotropic hypogonadism caused by inactivating mutations in SRA1. *JCRPE J. Clin. Res. Pediatr. Endocrinol.* **8**, (2016).
390. Salian-Mehta, S. *et al.* Functional consequences of AXL sequence variants in hypogonadotropic hypogonadism. *J. Clin. Endocrinol. Metab.* **99**, (2014).
391. Pierce, A. *et al.* Hypothalamic but not pituitary or ovarian defects underlie the reproductive abnormalities in Axl/Tyro3 null mice. *Mol. Cell. Endocrinol.* **339**, (2011).
392. Pierce, A. *et al.* Axl and Tyro3 modulate female reproduction by influencing gonadotropin-releasing hormone neuron survival and migration. *Mol. Endocrinol.* **22**, (2008).
393. Ramos, C. D. O. *et al.* Outcomes of Patients with Central Precocious Puberty Due to Loss-of-Function Mutations in the MKRN3 Gene after Treatment with Gonadotropin-Releasing Hormone Analog. *Neuroendocrinology* **110**, (2020).
394. Valadares, L. P. *et al.* MKRN3 mutations in central precocious puberty: A systematic review and meta-analysis. *J. Endocr. Soc.* **3**, (2019).
395. Niazi, R. K. *et al.* Identification of novel LEPR mutations in Pakistani families with morbid childhood obesity. *BMC Med. Genet.* **19**, (2018).
396. Serra-Juhé, C. *et al.* Heterozygous rare genetic variants in non-syndromic early-onset obesity. *Int. J. Obes.* **44**, (2020).
397. Elias, C. F. Leptin action in pubertal development: Recent advances and unanswered questions. *Trends in Endocrinology and Metabolism* vol. 23 at https://doi.org/10.1016/j.tem.2011.09.002 (2012).
398. Kleinendorst, L., Van Haelst, M. M. & Van Den Akker, E. L. T. Young girl with severe early-onset obesity and hyperphagia. *BMJ Case Rep.* **2017**, (2017).
399. Mourouzis, I., Lavecchia, A. M. & Xinaris, C. Thyroid Hormone Signalling: From the Dawn of Life to the Bedside. *Journal of Molecular Evolution* vol. 88 at https://doi.org/10.1007/s00239-019-09908-1 (2020).
400. Perry, J. R. B. *et al.* Meta-analysis of genome-wide association data identifies two loci influencing age at menarche. *Nat. Genet.* **41**, (2009).
401. Elks, C. E. *et al.* Thirty new loci for age at menarche identified by a meta-analysis of genome-wide association studies. *Nat. Genet.* **42**, (2010).
402. Perry, J. R. B. *et al.* Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature* **514**, (2014).
403. Lomniczi, A. *et al.* Epigenetic regulation of puberty via Zinc finger protein-mediated transcriptional repression. *Nat. Commun.* **6**, (2015).
404. Yasui, G. *et al.* Zinc finger protein 483 (ZNF483) regulates neuronal differentiation and methyl-CpG-binding protein 2 (MeCP2) intracellular localization. *Biochem. Biophys. Res. Commun.* **568**, (2021).
405. Oleksiewicz, U. *et al.* TRIM28 and Interacting KRAB-ZNFs Control Self-Renewal of Human Pluripotent Stem Cells through Epigenetic Repression of Pro-differentiation Genes. *Stem Cell Reports* **9**, (2017).
406. Quiñones, A., Dobberstein, K. U. & Rainov, N. G. The egr-1 gene is induced by DNA-damaging agents and non-genotoxic drugs in both normal and neoplastic human cells. *Life Sci.* **72**, (2003).
407. Knijnenburg, T. A. *et al.* Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. *Cell Rep.* **23**, (2018).
408. John, G. B., Shirley, L. J., Gallardo, T. D. & Castrillon, D. G. Specificity of the

requirement for Foxo3 in primordial follicle activation. *Reproduction* **133**, (2007).
409. Chung, Y. M. *et al.* FOXO3 signalling links ATM to the p53 apoptotic pathway following DNA damage. *Nat. Commun.* **3**, (2012).
410. Tran, H. *et al.* DNA repair pathway stimulated by the forkhead transcription factor FOXO3a through the Gadd45 protein. *Science (80-.).* **296**, (2002).
411. Castrillon, D. H., Miao, L., Kollipara, R., Horner, J. W. & DePinho, R. A. Suppression of ovarian follicle activation in mice by the transcription factor Foxo3a. *Science (80-.).* **301**, (2003).
412. Zhang, H., Lin, F., Zhao, J. & Wang, Z. Expression Regulation and Physiological Role of Transcription Factor FOXO3a During Ovarian Follicular Development. *Frontiers in Physiology* vol. 11 at https://doi.org/10.3389/fphys.2020.595086 (2020).
413. Wang, D. *et al.* DNA Damage-Induced Foci of E2 Ubiquitin-Conjugating Enzyme are Detectable upon Co-transfection with an Interacting E3 Ubiquitin Ligase. *Biochem. Genet.* **54**, (2016).
414. Liefke, R. *et al.* The oxidative demethylase ALKBH3 marks hyperactive gene promoters in human cancer cells. *Genome Med.* **7**, (2015).
415. Dango, S. *et al.* DNA unwinding by ASCC3 helicase is coupled to ALKBH3-dependent DNA alkylation repair and cancer cell proliferation. *Mol. Cell* **44**, (2011).
416. Ahmad, I. *et al.* Toll-like receptor-4 deficiency enhances repair of UVR-induced cutaneous DNA damage by nucleotide excision repair mechanism. *J. Invest. Dermatol.* **134**, (2014).
417. Haziak, K. *et al.* Effect of CD14/TLR4 antagonist on GnRH/LH secretion in ewe during central inflammation induced by intracerebroventricular administration of LPS. *J. Anim. Sci. Biotechnol.* **9**, (2018).
418. Gupta, A. *et al.* Role of 53BP1 in the regulation of DNA double-strand break repair pathway choice. *Radiation Research* vol. 181 at https://doi.org/10.1667/RR13572.1 (2014).
419. Dimitrova, N., Chen, Y. C. M., Spector, D. L. & De Lange, T. 53BP1 promotes non-homologous end joining of telomeres by increasing chromatin mobility. *Nature* **456**, (2008).
420. Wang, W., Zhao, L. J., Liu, Y. Z., Recker, R. R. & Deng, H. W. Genetic and environmental correlations between obesity phenotypes and age at menarche. *Int. J. Obes.* **30**, (2006).
421. Speliotes, E. K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* **42**, (2010).
422. Hotta, K. *et al.* Association between obesity and polymorphisms in SEC16B, TMEM18, GNPDA2, BDNF, FAIM2 and MC4R in a Japanese population. *J. Hum. Genet.* **54**, (2009).
423. Lucarini, N., Antonacci, E., Bottini, N. & Bottini, F. G. Low-molecular-weight acid phosphatase (ACP1), obesity, and blood lipid levels in subjects with non-insulin-dependent diabetes mellitus. *Hum. Biol.* **69**, (1997).
424. Saeed, S. *et al.* Loss-of-function mutations in ADCY3 cause monogenic severe obesity. *Nat. Genet.* **50**, (2018).
425. Toumba, M. *et al.* Molecular modelling of novel ADCY3 variant predicts a molecular target for tackling obesity. *Int. J. Mol. Med.* **49**, (2022).
426. Li, J. *et al.* Aberrant spliceosome expression and altered alternative splicing events correlate with maturation deficiency in human oocytes. *Cell Cycle* **19**, (2020).
427. Best, A. *et al.* Human Tra2 proteins jointly control a CHEK1 splicing switch among alternative and constitutive target exons. *Nat. Commun.* **5**, (2014).
428. Fu, H. *et al.* The DNA repair endonuclease Mus81 facilitates fast DNA replication in the absence of exogenous damage. *Nat. Commun.* **6**, (2015).
429. Yang, J. *et al.* Dysfunction of DNA damage-inducible transcript 4 in the decidua is relevant to the pathogenesis of preeclampsia. *Biol. Reprod.* **98**, (2018).
430. Roessler, E. *et al.* Loss-of-function mutations in the human GLI2 gene are associated with pituitary anomalies and holoprosencephaly-like features. *Proc. Natl. Acad. Sci. U. S. A.* **100**,

(2003).
431. Arnhold, I. J. P., França, M. M., Carvalho, L. R., Mendonca, B. B. & Jorge, A. A. L. Role of GLI2 in hypopituitarism phenotype. *Journal of Molecular Endocrinology* vol. 54 at https://doi.org/10.1530/JME-15-0009 (2015).
432. Agyeman, A., Mazumdar, T. & Houghton, J. A. Regulation of DNA damage following termination of hedgehog (HH) survival signaling at the level of the GLI genes in human colon cancer. *Oncotarget* **3**, (2012).
433. Shum, E. Y. *et al.* The Antagonistic Gene Paralogs Upf3a and Upf3b Govern Nonsense-Mediated RNA Decay. *Cell* **165**, (2016).
434. Salilew-Wondim, D. *et al.* Depletion of BIRC6 leads to retarded bovine early embryonic development and blastocyst formation in vitro. *Reprod. Fertil. Dev.* **22**, (2010).
435. Winship, A. L., Stringer, J. M., Liew, S. H. & Hutt, K. J. The importance of DNA repair for maintaining oocyte quality in response to anti-cancer treatments, environmental toxins and maternal ageing. *Hum. Reprod. Update* **24**, (2018).
436. Shin, E. M. *et al.* GREB1: An evolutionarily conserved protein with a glycosyltransferase domain links ERα glycosylation and stability to cancer. *Sci. Adv.* **7**, (2021).
437. Hodgkinson, K. *et al.* GREB1 is an estrogen receptor-regulated tumour promoter that is frequently expressed in ovarian cancer. *Oncogene* **37**, (2018).
438. Schwartz, F., Neve, R., Eisenman, R., Gessler, M. & Bruns, G. A WAGR region gene between PAX-6 and FSHB expressed in fetal brain. *Hum. Genet.* **94**, (1994).
439. Liguori, L. *et al.* The metallophosphodiesterase Mpped2 impairs tumorigenesis in neuroblastoma. *Cell Cycle* **11**, (2012).
440. Laisk, T. *et al.* Large-scale meta-analysis highlights the hypothalamic-pituitary-gonadal axis in the genetic regulation of menstrual cycle length. *Hum. Mol. Genet.* **27**, (2018).
441. Ruth, K. S. *et al.* Genetic evidence that lower circulating FSH levels lengthen menstrual cycle, increase age at menopause and impact female reproductive health. *Hum. Reprod.* **31**, (2016).
442. Chan, Y. M. *et al.* GNRH1 mutations in patients with idiopathic hypogonadotropic hypogonadism. *Proc. Natl. Acad. Sci. U. S. A.* **106**, (2009).
443. Skrapits, K. *et al.* The cryptic gonadotropin-releasing hormone neuronal system of human basal ganglia. *Elife* **10**, (2021).
444. Włodarczyk, M. & Nowicka, G. Obesity, DNA damage, and development of obesity-related diseases. *International Journal of Molecular Sciences* vol. 20 at https://doi.org/10.3390/ijms20051146 (2019).
445. Rohde, K. *et al.* Role of the DNA repair genes H2AX and HMGB1 in human fat distribution and lipid profiles. *BMJ Open Diabetes Res. Care* **8**, (2020).
446. Lee, G. SREBP1c-PARP1 axis tunes anti-senescence activity of adipocytes and ameliorates metabolic imbalance in obesity. *Cell Metab.* **34**, (2022).
447. Vergoni, B. *et al.* DNA damage and the activation of the p53 pathway mediate alterations in metabolic and secretory functions of adipocytes. *Diabetes* **65**, (2016).
448. García-De Teresa, B., Hernández-Gómez, M. & Frías, S. DNA Damage as a Driver for Growth Delay: Chromosome Instability Syndromes with Intrauterine Growth Retardation. *BioMed Research International* vol. 2017 at https://doi.org/10.1155/2017/8193892 (2017).
449. Hewitt, G. *et al.* Telomeres are favoured targets of a persistent DNA damage response in ageing and stress-induced senescence. *Nat. Commun.* **3**, (2012).
450. Sola, M. *et al.* Tau affects P53 function and cell fate during the DNA damage response. *Commun. Biol.* **3**, (2020).
451. Datta, N. *et al.* Promyelocytic Leukemia (PML) gene regulation: implication towards curbing oncogenesis. *Cell Death Dis.* **10**, (2019).
452. Touzot, F. *et al.* Extended clinical and genetic spectrum associated with biallelic RTEL1 mutations. *Blood Adv.* **1**, (2016).

453. Gazal, S. *et al.* Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity. doi:10.1038/s41588-022-01087-y.
454. Suo, L. *et al.* Transcriptome profiling of human oocytes experiencing recurrent total fertilization failure. *Sci. Rep.* **8**, (2018).
455. Wang, T. *et al.* Integrated bioinformatic analysis reveals YWHAB as a novel diagnostic biomarker for idiopathic pulmonary arterial hypertension. *J. Cell. Physiol.* **234**, (2019).
456. Zhang, L. *et al.* Network-based proteomic analysis for postmenopausal osteoporosis in Caucasian females. *Proteomics* **16**, (2016).
457. Oerter, K. E., Uriarte, M. M., Rose, S. R., Barnes, K. M. & Cutler, G. B. Gonadotropin secretory dynamics during puberty in normal girls and boys. *J. Clin. Endocrinol. Metab.* **71**, (1990).
458. Abreu, A. P. & Kaiser, U. B. Pubertal development and regulation. *The Lancet Diabetes and Endocrinology* vol. 4 at https://doi.org/10.1016/S2213-8587(15)00418-0 (2016).
459. John, G. B., Gallardo, T. D., Shirley, L. J. & Castrillon, D. H. Foxo3 is a PI3K-dependent molecular switch controlling the initiation of oocyte growth. *Dev. Biol.* **321**, (2008).
460. Moniruzzaman, M., Lee, J., Zengyo, M. & Miyano, T. Knockdown of FOXO3 induces primordial oocyte activation in pigs. *Reproduction* **139**, (2010).
461. Pelosi, E. *et al.* Constitutively active Foxo3 in oocytes preserves ovarian reserve in mice. *Nat. Commun.* **4**, (2013).
462. Mercer, M., Jang, S., Ni, C. & Buszczak, M. The Dynamic Regulation of mRNA Translation and Ribosome Biogenesis During Germ Cell Development and Reproductive Aging. *Frontiers in Cell and Developmental Biology* vol. 9 at https://doi.org/10.3389/fcell.2021.710186 (2021).
463. Jeggo, P. A., Pearl, L. H. & Carr, A. M. DNA repair, genome stability and cancer: A historical perspective. *Nature Reviews Cancer* vol. 16 at https://doi.org/10.1038/nrc.2015.4 (2016).
464. Jackson, S. P. & Bartek, J. The DNA-damage response in human biology and disease. *Nature* vol. 461 at https://doi.org/10.1038/nature08467 (2009).
465. Sharma, R., Lewis, S. & Wlodarski, M. W. DNA Repair Syndromes and Cancer: Insights Into Genetics and Phenotype Patterns. *Frontiers in Pediatrics* vol. 8 at https://doi.org/10.3389/fped.2020.570084 (2020).
466. Nelson, B. C. & Dizdaroglu, M. Implications of dna damage and dna repair on human diseases. *Mutagenesis* vol. 35 at https://doi.org/10.1093/mutage/gez048 (2020).
467. Wright, D. J. *et al.* Genetic variants associated with mosaic Y chromosome loss highlight cell cycle genes and overlap with cancer susceptibility. *Nat. Genet.* **49**, (2017).
468. Rodier, F. *et al.* Persistent DNA damage signalling triggers senescence-associated inflammatory cytokine secretion. *Nat. Cell Biol.* **11**, (2009).
469. Terzi, M. Y., Izmirli, M. & Gogebakan, B. The cell fate: senescence or quiescence. *Molecular Biology Reports* vol. 43 at https://doi.org/10.1007/s11033-016-4065-0 (2016).
470. Tavana, O. & Zhu, C. Too many breaks (brakes): Pancreatic β-cell senescence leads to diabetes. *Cell Cycle* vol. 10 at https://doi.org/10.4161/cc.10.15.16741 (2011).
471. Hernandez, A. M. *et al.* Upregulation of p21 activates the intrinsic apoptotic pathway in β-cells. *Am. J. Physiol. Endocrinol. Metab.* **304**, (2013).
472. Tay, V. S. Y. *et al.* Increased double strand breaks in diabetic β-cells with a p21 response that limits apoptosis. *Sci. Rep.* **9**, (2019).
473. Tchkonia, T., Zhu, Y., Van Deursen, J., Campisi, J. & Kirkland, J. L. Cellular senescence and the senescent secretory phenotype: Therapeutic opportunities. *Journal of Clinical Investigation* vol. 123 at https://doi.org/10.1172/JCI64098 (2013).
474. Roden, M. & Shulman, G. I. The integrative biology of type 2 diabetes. *Nature* vol. 576 at https://doi.org/10.1038/s41586-019-1797-8 (2019).
475. Dor, Y., Brown, J., Martinez, O. I. & Melton, D. A. *Adult pancreatic b-cells are formed*

by self-duplication rather than stem-cell differentiation. www.nature.com/nature (2004).
476. Kajimoto Y, K. H. Role of oxidative stress in pancreatic beta-cell dysfunction. *Osaka Univ. Grad. Sch. Med. Suita* (2008).
477. Lombard, D. B. *et al.* DNA repair, genome stability, and aging. *Cell* vol. 120 at https://doi.org/10.1016/j.cell.2005.01.028 (2005).
478. Blasiak, J. *et al.* DNA damage and repair in type 2 diabetes mellitus. *Mutat. Res. - Fundam. Mol. Mech. Mutagen.* **554**, (2004).
479. Tornovsky-Babeay, S. *et al.* Type 2 diabetes and congenital hyperinsulinism cause DNA double-strand breaks and p53 activity in β cells. *Cell Metab.* **19**, (2014).
480. Vujkovic, M. *et al.* Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. *Nat. Genet.* **52**, (2020).
481. Thompson, D. J. *et al.* Genetic predisposition to mosaic Y chromosome loss in blood. *Nature* **575**, (2019).
482. Tavana, O., Puebla-Osorio, N., Sang, M. & Zhu, C. Absence of p53-dependent apoptosis combined with nonhomologous end-joining deficiency leads to a severe diabetic phenotype in mice. *Diabetes* **59**, (2010).
483. Yu, S. W. *et al.* Mediation of poty(ADP-ribose) polymerase-1 - Dependent cell death by apoptosis-inducing factor. *Science (80-.).* **297**, (2002).
484. Kim, M. Y., Zhang, T. & Kraus, W. L. Poly(ADP-ribosyl)ation by PARP-1: 'PAR-laying' NAD+ into a nuclear signal. *Genes and Development* vol. 19 at https://doi.org/10.1101/gad.1331805 (2005).
485. Caldecott, K. W., Aoufouchi, S., Johnson, P. & Shall, S. *XRCC1 polypeptide interacts with DNA polymerase β and possibly poly (ADP-ribose) polymerase, and DNA ligase III is a novel molecular 'nick-sensor' in vitro*. *Nucleic Acids Research* vol. 24 (1996).
486. El-Khamisy, S. F., Masutani, M., Suzuki, H. & Caldecott, K. W. A requirement for PARP-1 for the assembly or stability of XRCC1 nuclear foci at sites of oxidative DNA damage. *Nucleic Acids Res.* **31**, (2003).
487. Gurung, R. L. *et al.* Inhibition of poly (ADP-Ribose) polymerase-1 in telomerase deficient mouse embryonic fibroblasts increases arsenite-induced genome instability. *Genome Integr.* **1**, (2010).
488. Lindahl, T., Satoh, M. S., Poirier, G. G. & Klungland, A. Post-translational modification of poly(ADP-ribose) polymerase induced by DNA strand breaks. *Trends in Biochemical Sciences* vol. 20 at https://doi.org/10.1016/S0968-0004(00)89089-1 (1995).
489. Wang, X. G., Wang, Z. Q., Tong, W. M. & Shen, Y. PARP1 Val762Ala polymorphism reduces enzymatic activity. *Biochem. Biophys. Res. Commun.* **354**, (2007).
490. Bürkle, A. PARP-1: A regulator of genomic stability linked with mammalian longevity. *ChemBioChem* **2**, (2001).
491. Masson, M. *et al.* XRCC1 Is Specifically Associated with Poly(ADP-Ribose) Polymerase and Negatively Regulates Its Activity following DNA Damage. *Mol. Cell. Biol.* **18**, (1998).
492. Hassa, P. O. *et al.* Acetylation of poly(ADP-ribose) polymerase-1 by p300/CREB-binding protein regulates coactivation of NF-κB-dependent transcription. *J. Biol. Chem.* **280**, (2005).
493. Schreiber, V., Dantzer, F., Amé, J. C. & De Murcia, G. Poly(ADP-ribose): Novel functions for an old molecule. *Nature Reviews Molecular Cell Biology* vol. 7 at https://doi.org/10.1038/nrm1963 (2006).
494. Farrar, D. *et al.* Mutational Analysis of the Poly(ADP-Ribosyl)ation Sites of the Transcription Factor CTCF Provides an Insight into the Mechanism of Its Regulation by Poly(ADP-Ribosyl)ation. *Mol. Cell. Biol.* **30**, (2010).
495. Tarayrah-Ibraheim, L. *et al.* DNase II mediates a parthanatos-like developmental cell death pathway in Drosophila primordial germ cells. *Nat. Commun.* **12**, (2021).

496. Zaremba, T. *et al.* Poly(ADP-ribose) polymerase-1 (PARP-1) pharmacogenetics, activity and expression analysis in cancer patients and healthy volunteers. *Biochem. J.* **436**, (2011).
497. Cottet, F. *et al.* New polymorphisms in the human poly(ADP-ribose) polymerase-1 coding sequence: Lack of association with longevity or with increased cellular poly(ADP-ribosyl)ation capacity. *J. Mol. Med.* **78**, (2000).
498. Mohrenweiser, H. W., Xi, T., Vázquez-Matías, J. & Jones, I. M. Identification of 127 amino acid substitution variants in screening 37 DNA repair genes in humans. *Cancer Epidemiol. Biomarkers Prev.* **11**, (2002).
499. Wang, X. Bin *et al.* PARP-1 variant rs1136410 confers protection against coronary artery disease in a Chinese han population: A two-stage case-control study involving 5643 subjects. *Front. Physiol.* **8**, (2017).
500. Lockett, K. L. *et al.* The ADPRT V762A genetic variant contributes to prostate cancer susceptibility and deficient enzyme function. *Cancer Res.* **64**, (2004).
501. Li, H. *et al.* Contributions of PARP-1 rs1136410 C>T polymorphism to the development of cancer. *J. Cell. Mol. Med.* **24**, (2020).
502. Chiang, F. Y. *et al.* Association between polymorphisms in DNA base excision repair genes XRCC1, APE1, and ADPRT and differentiated thyroid carcinoma. *Clin. Cancer Res.* **14**, (2008).
503. Zhang, X. *et al.* Polymorphisms in DNA base excision repair genes ADPRT and XRCC1 and risk of lung cancer. *Cancer Res.* **65**, (2005).
504. Roszak, A., Lianeri, M., Sowińska, A. & Jagodziński, P. P. Involvement of PARP-1 Val762Ala polymorphism in the onset of cervical cancer in Caucasian women. *Mol. Diagnosis Ther.* **17**, (2013).
505. Hua, R. X. *et al.* Association between the PARP1 Val762Ala polymorphism and cancer risk: Evidence from 43 studies. *PLoS One* **9**, (2014).
506. Cao, W. H. *et al.* Analysis of genetic variants of the poly(ADP-ribose) polymerase-1 gene in breast cancer in French patients. *Mutat. Res. - Genet. Toxicol. Environ. Mutagen.* **632**, (2007).
507. Qin, Q. *et al.* PARP-1 Val762Ala polymorphism and risk of cancer: A meta-analysis based on 39 case-control studies. *PLoS One* **9**, (2014).
508. Mégnin-Chanet, F., Bollet, M. A. & Hall, J. Targeting poly(ADP-ribose) polymerase activity for cancer therapy. *Cellular and Molecular Life Sciences* vol. 67 at https://doi.org/10.1007/s00018-010-0490-8 (2010).
509. Ashworth, A. A synthetic lethal therapeutic approach: Poly(ADP) ribose polymerase inhibitors for the treatment of cancers deficient in DNA double-strand break repair. *Journal of Clinical Oncology* vol. 26 at https://doi.org/10.1200/JCO.2008.16.0812 (2008).
510. Bryant, H. E. *et al.* Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase. *Nature* **434**, (2005).
511. Farmer, H. *et al.* Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature* **434**, (2005).
512. Nakamura, K. *et al.* Poly (ADP-ribose) polymerase inhibitor exposure reduces ovarian reserve followed by dysfunction in granulosa cells. *Sci. Rep.* **10**, (2020).
513. Ooi, S. K. T. & Bestor, T. H. The Colorful History of Active DNA Demethylation. *Cell* vol. 133 at https://doi.org/10.1016/j.cell.2008.06.009 (2008).
514. Seisenberger, S. *et al.* The Dynamics of Genome-wide DNA Methylation Reprogramming in Mouse Primordial Germ Cells. *Mol. Cell* **48**, (2012).
515. Seki, Y. *et al.* Extensive and orderly reprogramming of genome-wide chromatin modifications associated with specification and early development of germ cells in mice. *Dev. Biol.* **278**, (2005).
516. Reik, W., Dean, W. & Walter, J. Epigenetic reprogramming in mammalian development. *Science* vol. 293 at https://doi.org/10.1126/science.1063443 (2001).

517. Surani, M. A., Hayashi, K. & Hajkova, P. Genetic and Epigenetic Regulators of Pluripotency. *Cell* vol. 128 at https://doi.org/10.1016/j.cell.2007.02.010 (2007).
518. Sasaki, H. & Matsui, Y. Epigenetic events in mammalian germ-cell development: Reprogramming and beyond. *Nature Reviews Genetics* vol. 9 at https://doi.org/10.1038/nrg2295 (2008).
519. Smith, Z. D. *et al.* A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature* **484**, (2012).
520. Kurimoto, K. & Saitou, M. Epigenome regulation during germ cell specification and development from pluripotent stem cells. *Current Opinion in Genetics and Development* vol. 52 at https://doi.org/10.1016/j.gde.2018.06.004 (2018).
521. Hikabe, O. *et al.* Reconstitution in vitro of the entire cycle of the mouse female germ line. *Nature* **539**, (2016).
522. Saffman, E. E. & Lasko, P. Germline development in vertebrates and invertebrates. *Cellular and Molecular Life Sciences* vol. 55 at https://doi.org/10.1007/s000180050363 (1999).
523. Hayashi, K. & Saitou, M. Perspectives of germ cell development in vitro in mammals. *Animal Science Journal* vol. 85 at https://doi.org/10.1111/asj.12199 (2014).
524. Yao, C., Yao, R., Luo, H. & Shuai, L. Germline specification from pluripotent stem cells. *Stem Cell Research and Therapy* vol. 13 at https://doi.org/10.1186/s13287-022-02750-1 (2022).
525. Hayashi, K., Ohta, H., Kurimoto, K., Aramaki, S. & Saitou, M. Reconstitution of the mouse germ cell specification pathway in culture by pluripotent stem cells. *Cell* **146**, (2011).
526. Ohinata, Y. *et al.* Blimp1 is a critical determinant of the germ cell lineage in mice. *Nature* **436**, (2005).
527. McLaren, A. & Southee, D. Entry of mouse embryonic germ cells into meiosis. *Dev. Biol.* **187**, (1997).
528. Eriksson, N. *et al.* Web-based, participant-driven studies yield novel genetic associations for common traits. *PLoS Genet.* **6**, (2010).
529. Hayashi, K. *et al.* Offspring from oocytes derived from in vitro primordial germ cell-like cells in mice. *Science (80-.).* **338**, (2012).
530. Hayashi, K. & Saitou, M. Generation of eggs from mouse embryonic stem cells and induced pluripotent stem cells. *Nat. Protoc.* **8**, (2013).
531. IDTDNA. Perform gene knockout in your research. https://sfvideo.blob.core.windows.net/sitefinity/docs/default-source/application-guide/perform-gene-knockout-with-the-alt-r-crispr-cas-system-reference-guide.pdf?promo_name=Reference Protocol&promo_id=1d&promo_creative=Gene Knockout with Alt-R&promo_position=Get Started Box (2022).
532. IDTDNA. Alt-R CRISPR-Cas9 System. https://sfvideo.blob.core.windows.net/sitefinity/docs/default-source/protocol/alt-r-crispr-cas9-user-guide-ribonucleoprotein-electroporation-neon-transfection-system0601611532796e2eaa53ff00001c1b3c.pdf?sfvrsn=6c43407_28 (2022).
533. McKinnon, K. M. Flow cytometry: An overview. *Curr. Protoc. Immunol.* **2018**, (2018).
534. Lai, Y. S. *et al.* SRY (sex determining region Y)-box2 (Sox2)/poly ADP-ribose polymerase 1 (Parp1) complexes regulate pluripotency. *Proc. Natl. Acad. Sci. U. S. A.* **109**, (2012).
535. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2-ΔΔCT method. *Methods* **25**, (2001).
536. Saitou, M. & Hayashi, K. Mammalian in vitro gametogenesis. *Science* vol. 374 at https://doi.org/10.1126/science.aaz6830 (2021).
537. Saitou, M., Barton, S. C. & Surani, M. A. A molecular programme for the specification of germ cell fate in mice. *Nature* vol. 418 at https://doi.org/10.1038/nature00927 (2002).
538. Yamaji, M. *et al.* Critical function of Prdm14 for the establishment of the germ cell lineage in mice. *Nat. Genet.* **40**, (2008).
539. Günesdogan, U., Magnúsdóttir, E. & Surani, M. A. Primoridal germ cell specification: A

context-dependent cellular differentiation event. *Philosophical Transactions of the Royal Society B: Biological Sciences* vol. 369 at https://doi.org/10.1098/rstb.2013.0543 (2014).
540. Saitou, M. & Miyauchi, H. Gametogenesis from Pluripotent Stem Cells. *Cell Stem Cell* vol. 18 at https://doi.org/10.1016/j.stem.2016.05.001 (2016).
541. Yokobayashi, S. *et al.* Clonal variation of human induced pluripotent stem cells for induction into the germ cell fate. *Biol. Reprod.* **96**, (2017).
542. Hayashi, K., Hikabe, O., Obata, Y. & Hirao, Y. Reconstitution of mouse oogenesis in a dish from pluripotent stem cells. *Nat. Protoc.* **12**, (2017).
543. Miyauchi, H. *et al.* Bone morphogenetic protein and retinoic acid synergistically specify female germ-cell fate in mice. *EMBO J.* **36**, (2017).
544. Saitou, M., Payer, B., O'Carroll, D., Ohinata, Y. & Surani, M. A. Blimp1 and the emergence of the germ line during development in the mouse. *Cell Cycle* vol. 4 at https://doi.org/10.4161/cc.4.12.2209 (2005).
545. Iwase, A. *et al.* Anti-Müllerian hormone as a marker of ovarian reserve: What have we learned, and what should we know? *Reprod. Med. Biol.* **15**, (2016).
546. Iwase, A. *et al.* Clinical application of serum anti-Müllerian hormone as an ovarian reserve marker: A review of recent studies. *J. Obstet. Gynaecol. Res.* **44**, (2018).
547. Tsai, C. C., Su, P. F., Huang, Y. F., Yew, T. L. & Hung, S. C. Oct4 and Nanog Directly Regulate Dnmt1 to Maintain Self-Renewal and Undifferentiated State in Mesenchymal Stem Cells. *Mol. Cell* **47**, (2012).
548. Nussinov, R. The spatial structure of cell signaling systems. *Phys. Biol.* **10**, (2013).
549. Kholodenko, B. N. Spatially distributed cell signalling. *FEBS Letters* vol. 583 at https://doi.org/10.1016/j.febslet.2009.09.045 (2009).
550. Jacob, S. & Moley, K. H. Gametes and embryo epigenetic reprogramming affect developmental outcome: Implication for assisted reproductive technologies. *Pediatric Research* vol. 58 at https://doi.org/10.1203/01.PDR.0000179401.17161.D3 (2005).
551. Yamashiro, C., Sasaki, K., Yokobayashi, S., Kojima, Y. & Saitou, M. Generation of human oogonia from induced pluripotent stem cells in culture. *Nat. Protoc.* **15**, (2020).
552. Lin, X. *et al.* PARP inhibitors trap PARP2 and alter the mode of recruitment of PARP2 at DNA damage sites. *Nucleic Acids Res.* **50**, (2022).
553. Chen, H., Pu, X. Y., Zhang, R. P. & A, Z. C. Association of common SNP rs1136410 in PARP1 gene with the susceptibility to male infertility with oligospermia. *J. Assist. Reprod. Genet.* **31**, (2014).
554. Laniel, M. A., Bergeron, M. J., Poirier, G. G. & Guérin, S. L. A nuclear factor other than Sp1 binds the GC-rich promoter of the gene encoding rat poly(ADP-ribose) polymerase in vitro. *Biochem. Cell Biol.* **75**, (1997).
555. Maymon, B. B. S. *et al.* Role of poly(ADP-ribosyl)ation during human spermatogenesis. *Fertil. Steril.* **86**, (2006).
556. Hayashi, Y., Saitou, M. & Yamanaka, S. Germline development from human pluripotent stem cells toward disease modeling of infertility. *Fertility and Sterility* vol. 97 at https://doi.org/10.1016/j.fertnstert.2012.04.037 (2012).
557. Shi, J. *et al.* Age at menarche and age at natural menopause in East Asian women: a genome-wide association study. *Age (Omaha).* **38**, (2016).
558. Wadhwa, P. D., Buss, C., Entringer, S. & Swanson, J. M. Developmental origins of health and disease: Brief history of the approach and current focus on epigenetic mechanisms. *Seminars in Reproductive Medicine* vol. 27 at https://doi.org/10.1055/s-0029-1237424 (2009).
559. Barker, D. J. P. The origins of the developmental origins theory. in *Journal of Internal Medicine* vol. 261 (2007).
560. Aiken, C. E., Tarry-Adkins, J. L., Penfold, N. C., Dearden, L. & Ozanne, S. E. Decreased ovarian reserve, dysregulation of mitochondrial biogenesis, and increased lipid peroxidation in female mouse offspring exposed to an obesogenic maternal diet. *FASEB J.* **30**, (2016).

561. O'connor, J. M., Sedghi, T., Dhodapkar, M., Kane, M. J. & Gross, C. P. Factors Associated With Cancer Disparities Among Low-, Medium-, and High-Income US Counties. doi:10.1001/jamanetworkopen.2018.3146.
562. Gottschalk, M. S., Eskild, A., Hofvind, S., Gran, J. M. & Bjelland, E. K. Temporal trends in age at menarche and age at menopause: a population study of 312 656 women in Norway. *Hum. Reprod.* **35**, 464–471 (2020).
563. Dratva, J. *et al.* Is age at menopause increasing across Europe? Results on age at menopause and determinants from two population-based studies. *Menopause* **16**, (2009).
564. Pakarinen, M., Raitanen, J., Kaaja, R. & Luoto, R. Secular trend in the menopausal age in Finland 1997-2007 and correlation with socioeconomic, reproductive and lifestyle factors. *Maturitas* **66**, (2010).
565. Ettorre, V. M. & Bachmann, G. A. Childhood predictors of age at natural menopause. *Case Reports in Women's Health* vol. 24 at https://doi.org/10.1016/j.crwh.2019.e00148 (2019).
566. Fudvoye, J. & Parent, A.-S. Secular trends in growth. *Ann. Endocrinol. (Paris).* **78**, (2017).
567. Altman, D. G. & Bland, J. M. Statistics notes: How to obtain the P value from a confidence interval. *BMJ* **343**, (2011).
568. Forman, M. R., Mangini, L. D., Thelus-Jean, R. & Hayward, M. D. Life course origins of the ages at menarche and menopause. *Adolesc. Health. Med. Ther.* (2013) doi:10.2147/AHMT.S15946.