UNIVERSITÀ DEGLI STUDI DI TRIESTE

XXX CICLO DEL DOTTORATO DI RICERCA IN
Scienze della Riproduzione e dello Sviluppo

IDENTIFICATION OF GENETIC VARIANTS REGULATING FEMALE FERTILITY

Settore scientifico-disciplinare: MED/03

DOTTORANDA:
Caterina Maria Barbieri

COORDINATORE:
PROF. Paolo Gasparini

SUPERVISORE DI TESI:
PROF.SSA Daniela Toniolo
PROF. Paolo Gasparini

ANNO ACCADEMICO 2016/2017

UNIVERSITÀ DEGLI STUDI DI TRIESTE

XXX CICLO DEL DOTTORATO DI RICERCA IN
Scienze della Riproduzione e dello Sviluppo

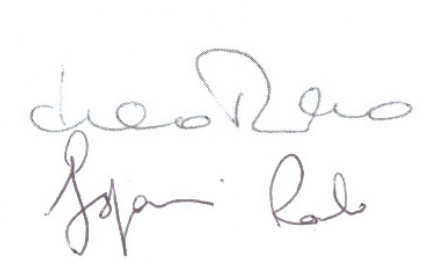
IDENTIFICATION OF GENETIC VARIANTS REGULATING FEMALE FERTILITY

Settore scientifico-disciplinare: MED/03

DOTTORANDA:
Caterina Maria Barbieri

COORDINATORE:
PROF. Paolo Gasparini

SUPERVISORE DI TESI:
PROF.SSA Daniela Toniolo
PROF. Paolo Gasparini



ANNO ACCADEMICO 2016/201

CONTENTS

LIST OF ABBREVIATION

1000GP1 1000 Genomes project phase 1

1000GP3 1000 Genomes project phase 3

AAM age at menarche

AC Alternative Allele count

AFB age at first birth

ANM age natural menopause

CADD Combined Annotation Dependent Depletion

CARL Carlantino

CNV Copy Number Variants

DP Read depth

ExAC Exome Aggregation Consortium

FVG Friuli Venezia Giulia

GIANT Genetic Investigation of ANthropometric Traits

GWAS Genome Wide Association Studies

hKO human Knockout

HSR San Raffaele Hospital

INGI Italian Network of Genetic Isolates

KO Knockout

LD linkage disequilibrium

LOF Loss Of Function

MAF minor allele frequencies

NEB number of children ever born

NRDR Non Reference Discordance Rate

OR ovarian reserve

PC principal components

QC Quality control

RPKM Reads Per Kilobase Million

UK10K the UK10K project

VBI Val Borbera

WES Whole Exome Sequencing

WGS Whole Genome Sequencing

INTRODUCTION

Women fertility

Fertility is defined as the capacity to produce offspring and it is measured with fertility rate, the average number of live births per women.
In the last decades, populations of industrialized countries have experienced a decline in total fertility far below 2.1, which is the rate considered as necessary to sustain a population size at current numbers (Skakkebaek et al., 2015) (Figure 1).

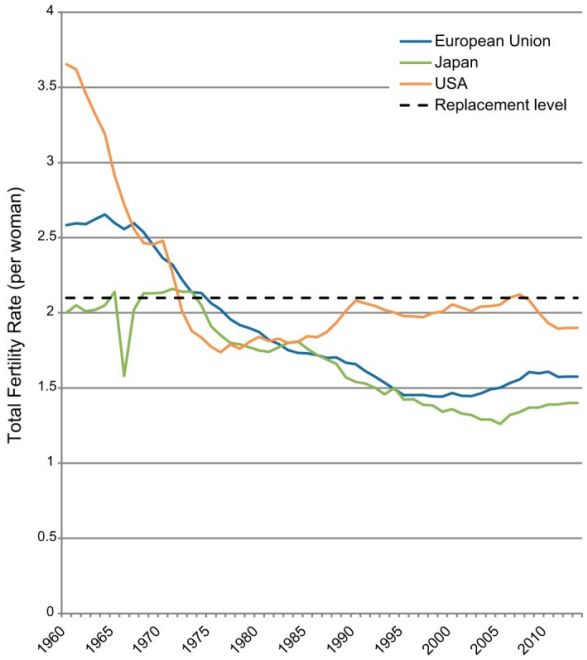


Figure 1. Total Fertility Rates (TFR) measured in European Union, Japan and United States between 1960 and 2013. Dotted line represents a fertility rate of 2.1, below which a population cannot be sustained. (Skakkebaek et al., 2015)

Fertility has become a significant political, social and economic factor.
World Health Organization defines infertility as “a disease of the reproductive system defined by the failure to achieve a clinical pregnancy after 12 months or more of regular unprotected sexual intercourse”. This definition is based on the observation that 84% of women are expected to conceive within 1 year of regular unprotected sexual intercourse; this number rises to 92% and 93% respectively after 2 or 3 years (te Velde, Eijkemans, & Habbema, 2000).

Many biological factors are known to have an impact on women fertility: some of them arise in the earlier stages of life, during foetal development, others

after birth during development. In addition, environment plays an important role in affecting fertility, including season and food intake (Sharpe & Franks, 2002).

Fertilities decline with age in both women and men but the effects of age are much more pronounced in women.

Some of common medical conditions that cause female infertility are (National Institutes of Health):

- Primary Ovarian Insufficiency (POI) or Polycystic Ovary Syndrome (PCOS) that cause failure to ovulate
- uterine fibroid, polyps or unusually shaped uterus which can affect implantation and the ability to carry a pregnancy to term
- autoimmune disorders, such as lupus, Hashimoto's and other types of thyroiditis, or rheumatoid arthritis
- silent infections as untreated gonorrhea and chlamydia that can lead to pelvic inflammatory disease, which might cause tissue scarring that blocks the fallopian tubes
- thin endometrium or endometriosis .

Diet and lifestyle may also contribute to alter women fertility: fertility rates decrease in women with extreme BMI values (obesity or extreme thinness), whereas healthy lifestyle may help to improve fertility for women with ovulatory dysfunction (Clark, Thornley, Tomlinson, Galletley, & Norman, 1998). High blood mercury concentration, due to high seafood consumption, is associated with infertility (Choy et al., 2002). Different studies show how smoking accelerates the rate of follicular depletion: on average, in smoking women menopause occurs on average 0.8-2 years earlier (Pawli & Szwed, 2007) and a large meta-analysis found that smoking women are significantly more likely to be infertile (Augood, Duckitt, & Templeton, 1998). The effect of alcohol has not been clearly established on female fertility yet, but one thing remains certain: alcohol consumption should cease during pregnancy because alcohol has detrimental effects on fetal development (Bager, Christensen, Husby, & Bjerregaard, 2017). Additionally, high levels of caffeine consumption have been associated with a delayed conception (Stanton, 1995).

Besides all these environmental causes and diseases, the ovarian reserve (OR) is fundamental for women fertility (see more details below). A small number of oocytes in initial reserve or a prematurely depletion of primordial follicles will result in infertility or in a shorter reproductive life span (Nelson, Telfer, & Anderson, 2013).

Follicle development and ovarian reserve

Germ cells separate from somatic cells during early stage of embryonic development and at this stage they are called primordial germ cells (PGCs). They start to multiply rapidly by mitotic division and in women they are transformed in oogonia. At this step, a first reserve of primordial follicles occurs: during fetal life, at around five months of gestational age, the ovary contains several million of non-growing follicles. After this initial exponential growth into oogonia (up to 7 million germ cell at 5 months of gestation), cells begin meiosis and stop in prophase-I without developing into oocytes (Fig. 2) (Monniaux et al., 2014). The ovarian reserve (OR) is established by a balance between the availability of the number of germ cells and the rate of the subsequent programmed cell death. Already at the time of birth the number of oocytes has declined to about 2 million.

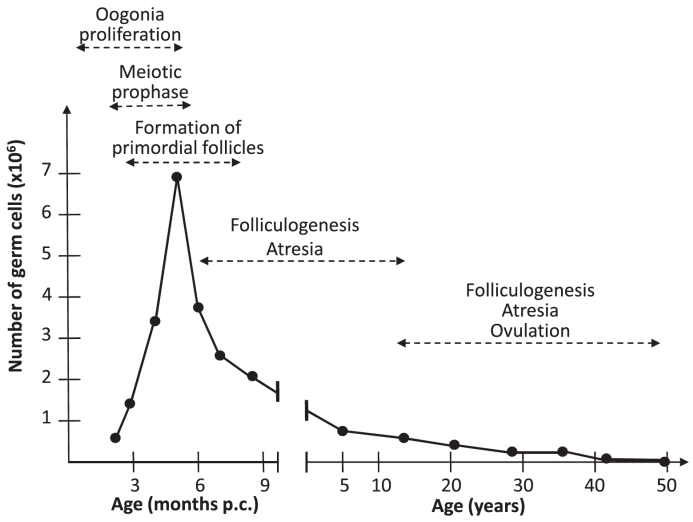


Figure 2. Numerical changes in the ovarian reserve of germ cells throughout life in humans. Data were obtained by histological analyses. Months p.c., months postconception (Monniaux et al., 2014).

Failure in mitosis/meiosis, damage in DNA, insufficient pre-granulosa cells, that surround the oocyte with major functions to produce sex steroids and myriad growth factors, may explain why so many oocytes are produced during fetal development and die before birth (Kerr, Myers, & Anderson, 2013).

Primary oocytes, surrounded by one layer of flat granulosa cells, remains in this shape until the beginning of puberty in the female (may remain at this stage for many decades). These primordial quiescent follicles support future ovulations throughout the reproductive life span. At puberty follicles start growing to full maturation and ovulation, and progressively diminish in number until menopause.

During maturation follicles increase in size, they constitute the internal and external theca and are characterized by the presence of an antrum in the granulosa (Figure 3). Granulosa and theca cells continue to undergo mitotis concomitant with an increase in antrum volume. In women, this process is long, requiring almost 1 year for a primordial follicle to grow and develop to the ovulatory stage.

Hormones and growth factors play a fundamental role in regulating follicles maturation. Gonadotropin-releasing hormone (GnRH) secreted by the hypothalamus stimulates the release of follicle-stimulating hormone (FSH) and luteinizing hormone (LH) from the anterior pituitary gland that in turn, has a stimulatory effect on antral follicles. When theca cells take form in the follicle, they show a progressive increase of LH and FSH receptors in granulosa cell that permit physiological maturation of follicles. Estrogen is fundamental in the selection of dominant follicle and a spike of LH causes ovulation.

Corpus luteum, the remaining of the ovarian follicle after the release of a mature ovum, is involved in the production of relatively high levels of progesterone, moderate levels of estradiol and inhibin A. Progesterone is fundamental for the decidualization of the endometrium, in average every 28 days if the egg is not fertilized, the corpus luteum stops secreting progesterone and degenerates in a mass of fibrous scar tissue whereas if the egg is fertilized and implanted, progesterone secretion continue maintaining a thick and bloody endometrium.

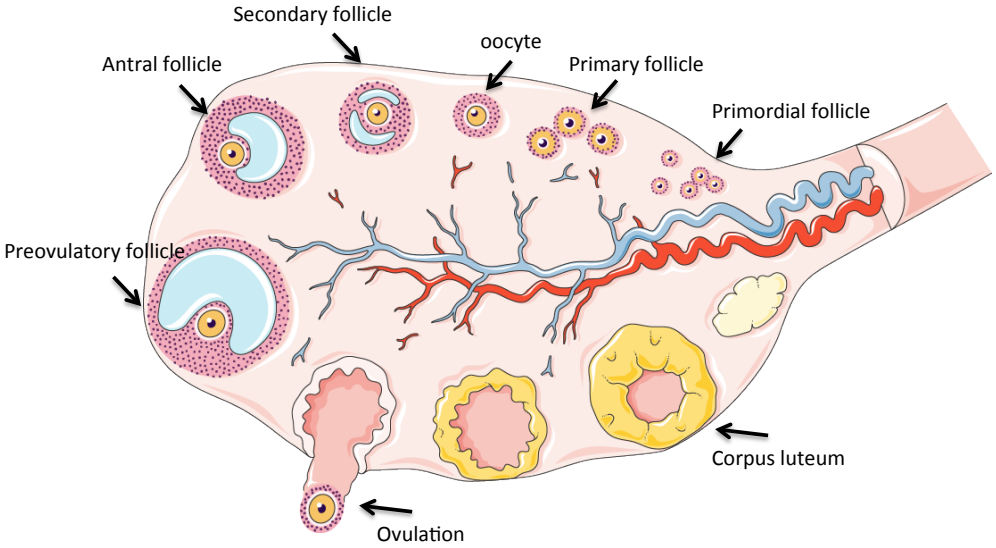


Figure 3. Overview of follicle development

Age-related decline of female fertility

Aging of females affects fertility: after the age of 30, a woman's chances of getting pregnant decrease rapidly every year because there is both a qualitative and a quantitive decline of the oocyte and follicle pool (te Velde & Pearson, 2002) (Figure 4). With age, also increases the probability that a pregnancy will terminate sooner or later after conception or implantation (e.g. miscarriage).

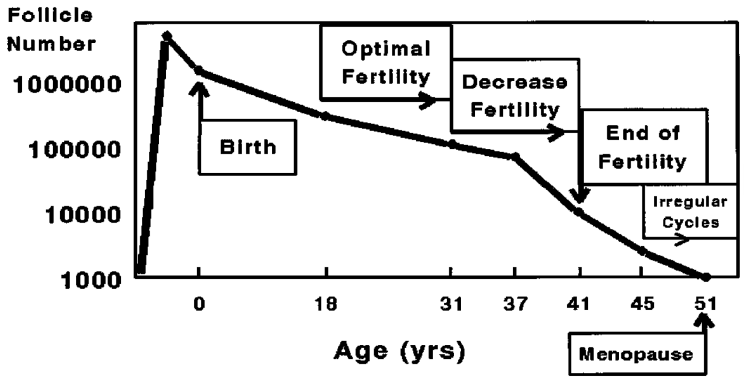


Figure 4. The declining oocyte/follicle pool during woman lifetime according to Faddy et al. 1992

Infertility is a global medico-socio-cultural problem especially in developing countries (Serour & Serour, 2017). Nowadays, reproductive behavior is deeply modified due to the access to high education, employment and career

opportunities, other life goals and to economic uncertainty (Mills, Rindfuss, McDonald, & te Velde, 2011). This has resulted in a massive delay in childbearing in Europe (Sobotka, 2004) (Fig.5) and also in certain highly educated groups in the United States (Heck, Schoendorf, Ventura, & Kiely, 1997).

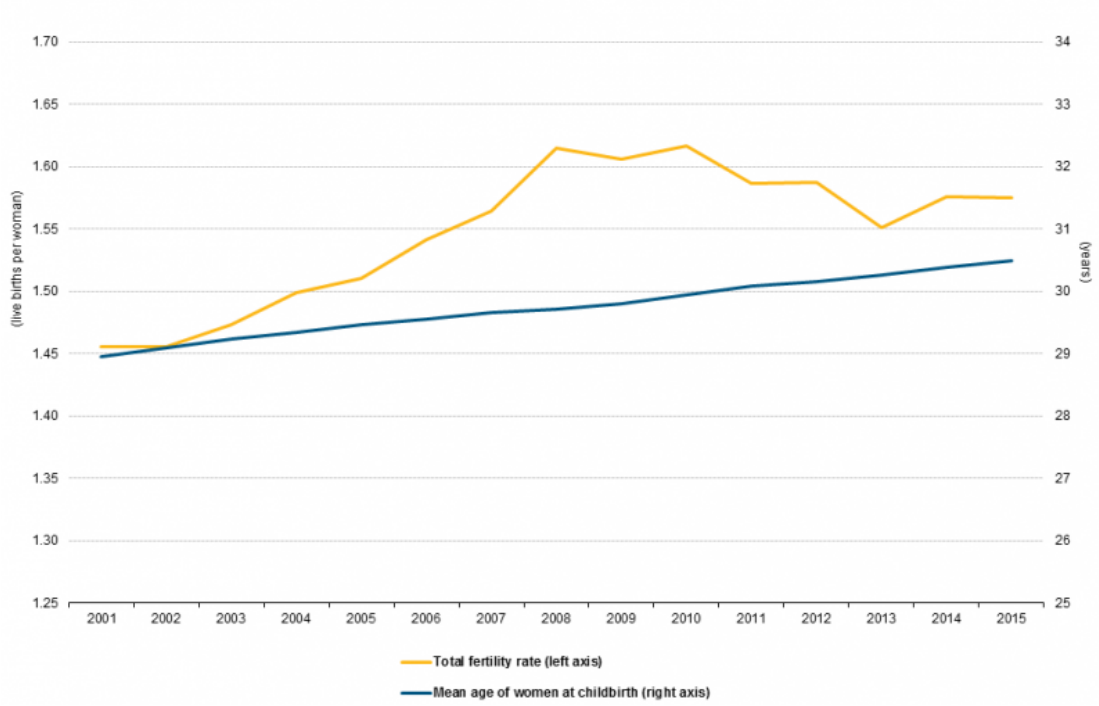


Figure 5. Increasing average age of women at childbirth between 2001 and 2015, from an average of 29.0 to 30.5 years. (Source Eurostat)

Anti Müllerian Hormone

Anti Müllerian hormone (AMH) has become the 'molecule of the moment' in the field of reproductive endocrinology (La Marca et al., 2010). The AMH evaluation can essentially lead to individualization of therapeutic strategies in women with infertile problems (Jamil et al., 2016).

AMH is a dimeric glycoprotein member of the transforming growth factor-beta (TGF-β**)** superfamily and its gene is located on chromosome 19p13.3 containing 5 exons. AMH binds to its receptor (AMHR), a single transmembrane protein with serine-threonine kinase activity specifically expressed on target organs such as granulosa cells of the ovary, Sertoli and Leyding cell of testis. *SF1*, *GATA1*, *WT1*, *DAX1* and *SOX9* genes regulate the

production of AMH that has a fundamental role in sexual differentiation (Watanabe et al., 2000). In male fetus, AMH secreted by immature Sertoli cells is involved in testes development: whereas AMH leads to regression of Müllerian duct, testosterone masculinizes the fetus, stimulating the formation of the portions of the male anatomy, such as epididymis and seminal vesicles (Matuszczak et al., 2013). In female fetus the absence of AMH permits the rise of Müllerian duct and thus the development of uterus, fallopian tubes and the upper part of the vagina. At birth AMH is higher in males respect to females where it is almost undetectable; at three months of age the hormone dosage increases in both sexes and during childhood it is stable, but the onset of testosterone synthesis in males causes a rapid decline reaching the levels observed in healthy females (Lindhardt Johansen et al., 2013) (Fig. 6). The highest levels of AMH secretion in women occur between 23-25 years old, the most fertile period in female.

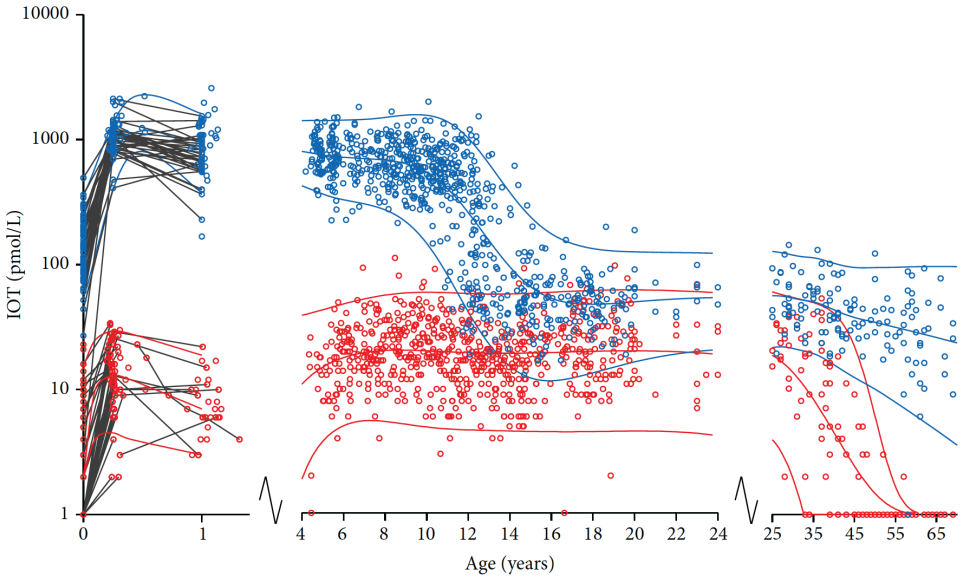


Figure 6. Serum AMH in 1953 healthy subjects (926 females and 1027 males) according to age. Females: red circles, males: blue circles. Longitudinal values during infancy are connected with grey lines. The red and blue curves represent the female and male reference ranges, respectively (median, +/-2SD). The log-scale y-axes for the DSL and Gen II assays were created using the following formulas: AMH(IOT) pmol/L = 2.0 Å~AMH(DSL) ug/L Å~ 7.14 pmol/ug and AMH(IOT) pmol/L = 0.74 Å~AMH(Gen II) ug/L Å~ 7.14 pmol/ug (Lindhardt Johansen et al., 2013)

In women AMH concentration essentially reflects the ovarian follicular pool and the reduction in the number of small growing follicles is followed by a reduction in circulating AMH. In fact AMH is produced by ovarian granulosa

cell of the preantral follicle until the stage of antral follicle when it reaches the differentiation state at which they are selected for dominance by the action of pituitary FSH (Weenen et al., 2004) (Fig. 7).

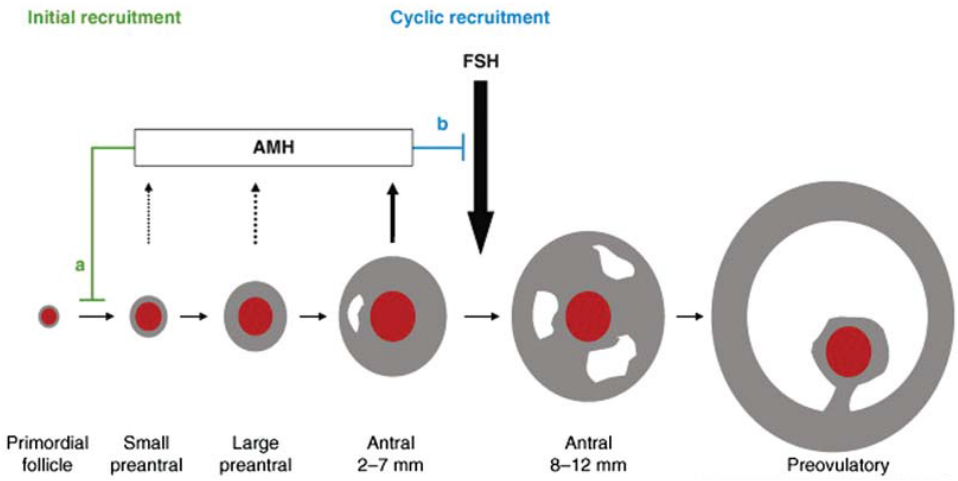


Figure 7. AMH is secreted by pre-antral and antral follicles.

AMH is secreted into the circulation and it is measurable in serum at any time throughout the cycle, in fact the majority of the studies underline that there is no fluctuation during menstrual cycle (Fig. 8) (La Marca et al., 2006; Streuli et al., 2008).

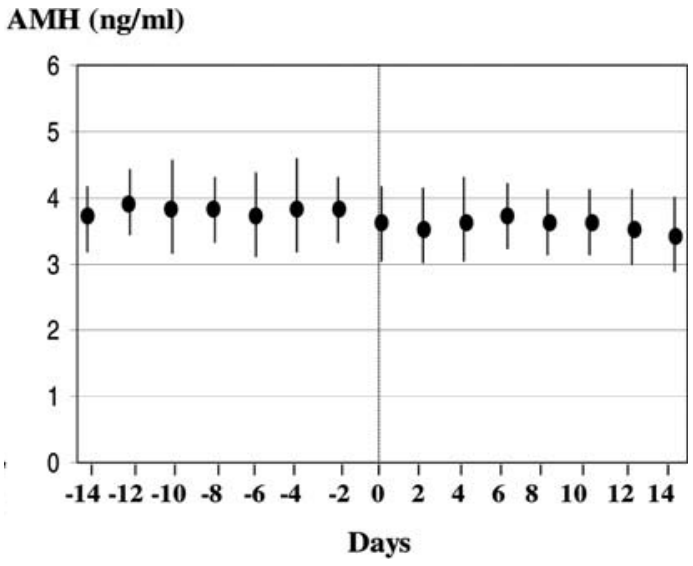


Figure 8. Circulatory pattern of AMH during the menstrual cycle of young healthy women aged 18–24 years. Day 0 = day of LH surge (La Marca et al., 2006).

In women, AMH levels show a progressive decline during reproductive life as the follicular reserve decreases, becoming undetectable after the menopause.

The correlation between AMH and the number of growing ovarian follicles is also shown by the fact that high levels of AMH are present in women with ovarian tumors and in women with polycystic ovaries (Grynnerup et al., 2012) whereas it is undetectable in patients with Turner syndrome without gonadal tissue (Lindhardt Johansen et al., 2013). In women aged 25-40 years, a value of 1.0 to 3.0 ng/mL levels of AMH is considered as “normal” for fertility (Lie Fong et al., 2012).

Younger women show a higher variability in concentration of AMH (Fig. 9) (La Marca et al., 2010), suggesting that they may underscore differences in the oocyte pool at young age.

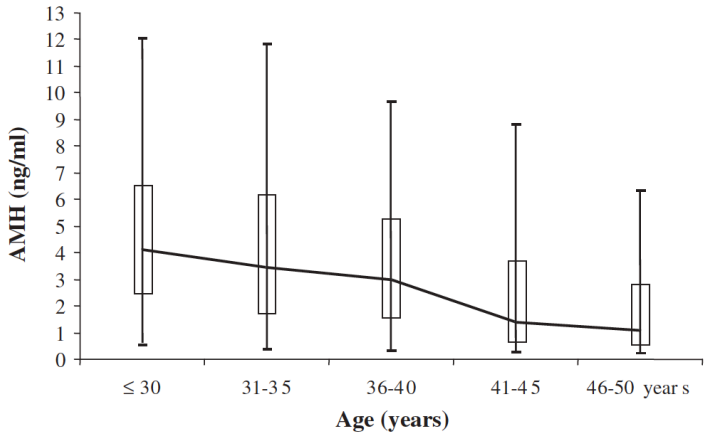


Figure 9. Serum anti-Müllerian hormone (AMH) concentrations throughout the reproductive period. Median, lower (2.5th percentile) and upper (97.5th percentile) limits and 25th and 75th percentiles are shown. (La Marca et al., 2010)

A biological role for AMH

The slow change in AMH levels across the fertile lifespan of women (Figure 6) demonstrates that AMH regulates a long-term process function.

A *in vivo* experiment with female mice AMH$^{+/+}$ and AMH$^{-/-}$ prior to puberty (Durlinger et al., 1999) showed that the number of follicles, in both groups of mice, decreases with age but the decrease is faster in AMH$^{-/-}$ mice. This is a strong evidence that the pattern of AMH expression regulates primordial follicle activation (Pankhurst, 2017) (Figure 7, line green). These observations highlight that the inhibition AMH-mediated of primordial follicle activation might

be strongest in the early stage of reproductive life of women with a decreasing effect as age increases.

A hypothesis for a potential second biological function of AMH in female takes into consideration the concentration of AMH in the hypothalamic-pituitary-gonadal axis selection of follicles with a preference for high quality oocytes (Pankhurst, 2017). When ovarian reserve is small, there is a reduced negative AMH feedback, that in turn results in a more rapid depletion of follicles. This depletion leads to a reduced probability to have a large number of high-quality follicles. Whereas, when the pool of developing small antral follicles is large, the possibility of choosing high-quality follicles is increased (Pankhurst, 2017) (Fig. 10). Indeed, this model has limitations, as it has not been confirmed that high-quality follicles are always selected.

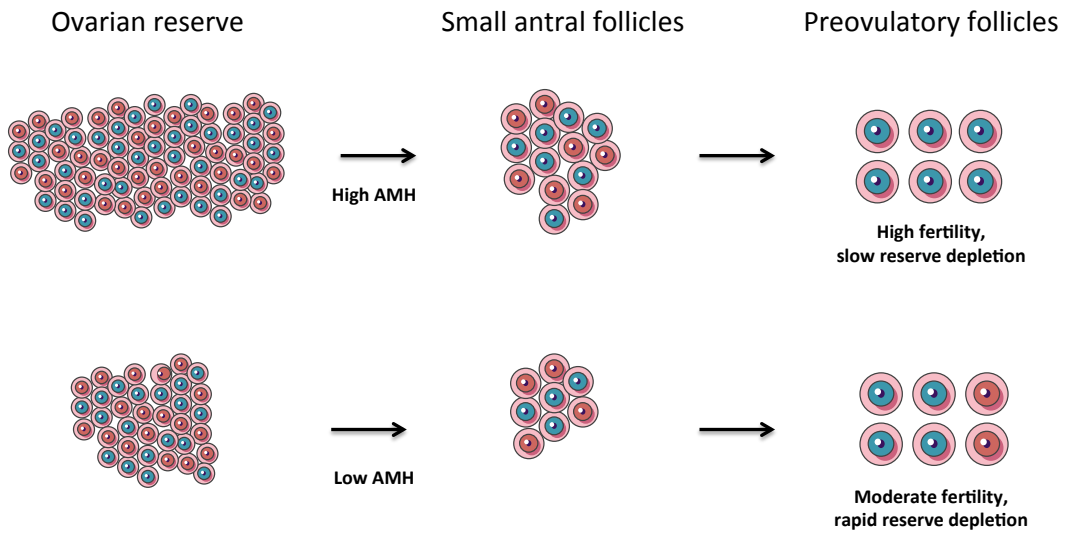


Figure 10. A simplified hypothesis of quality-based follicle selection. (Pankhurst, 2017). Blu=high quality, red=low quality.

Genetic variation in reproductive ageing

Menarche and menopause are the two events that frame the reproductive life span in women.

Different studies on families and twins, both monozygotic and dizygotic, underlined that genetic factors play a fundamental role in determining age at menarche with heritability h^2=0.63 on average (Table 1) (Towne et al., 2005).

Study	Heritability (±se)	Type of relative pairs in study	No. of pairs
Van den Ackker et al., 1987 (London)	0.72	Unspecified twins	364
Van den Ackker et al., 1987 (Birmingham)	0.54	Unspecified twins	98
Treloar and Martin, 1990	0.61–0.68	MZ and DZ twins	1,177 MZ; 711 DZ
Meyer et al., 1991	0.71	MZ and DZ twins	1,233 MZ; 751 DZ
Kaprio et al., 1995	0.74	MZ and DZ twins	234 MZ; 189 DZ
Loesch et al., 1995	0.95	MZ and DZ twins	44 MZ; 42 DZ
Snieder et al., 1998	0.45	MZ and DZ twins	275 MZ; 353 DZ
Doughty and Rodgers, 2000	0.54±0.08	Various relative pairs	1,178
Rowe, 2000	0.44± 0.15	Various sib pairs	505
Kirk et al., 2001	0.5	MZ and DZ twins	1,373 MZ; 1,310 DZ

Table 1. A comparison between heritability value for age of menarche estimated by various studies (Towne et al., 2005), (Van den Akker, Stein, Neale, & Murray, 1987), (Treloar & Martin, 1990), (Meyer, Eaves, Heath, & Martin, 1991), (Kaprio et al., 1995), (Loesch, Hopper, Rogucka, & Huggins, 1995), (Snieder, Macgregor, & Spector, 1998), (Doughty & Rodgers, 2000), (Rowe, 2000), (Kirk et al., 2001)

The difference in each study might reflect measurement bias or the effect of environmental factors but all confirm that individual differences in age at menarche are highly heritable.

Age of natural menopause is also largely under genetic control. The irreversible end of a women's reproductive life varies widely between 50-60 years and is modulated by both genetic and environmental factors. Only a small proportion of this large variation in age of menopause can be explained by environmental factors (van Noord, Peeters, Grobbee, Dubas, & te Velde, 1999) with the highest effect due to smoke: smokers women have menopause on average 0.8-2 years earlier than non-smokers (Pawli & Szwed, 2007) (Figure 11).

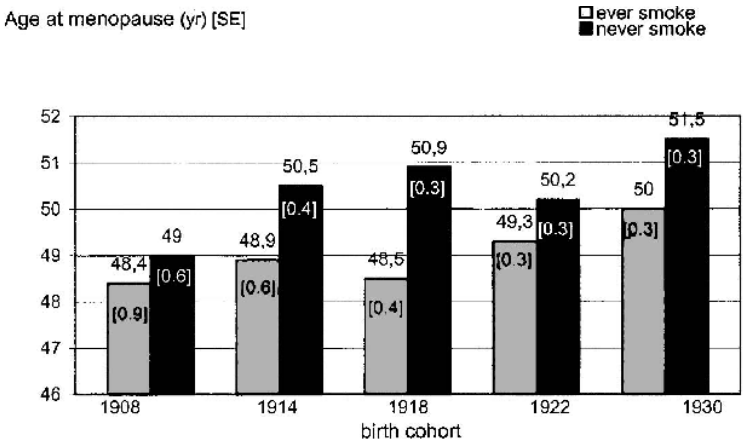


Figure 11. Age at menopause (y) grouped according to birth cohort and smoking status. (Rödström et al., 2003)

A study of twins in United Kingdom with a model specifying additive genetic and unique environmental variance component estimates a heritability of 63% [95% confidence interval (CI), 0.53-0.71] (Snieder et al., 1998).

This high estimate of heritability was confirmed in the multi-generational Framingham Heart Study, a community-based epidemiological study (Murabito et al. 2005). Women from the original cohort collected starting in 1948 and offspring collected later (Dawber, Meadors, & Moore, 1951) (Swan, 2000) were analyzed: heritability was calculated using the variance-components methods implemented in SOLAR (Almasy & Blangero, 1998) that takes into account the entire set of familial relationships. The separate estimations for (1) the original cohort, (2) the offspring and (3) for the entire set of women provide convincing evidence of the importance of genetic factors in determining both age of menarche and age of menopause (Table 2).

	n	H2, mean	95% CI	P value
Original cohort				
Crude	1500	0.65	0.42, 0.89	<0.0001
Multivariable-adjusted[a]	984	0.74	0.31, 1.00	<0.002
Offspring cohort				
Crude	932	0.59	0.37, 0.81	<0.0002
Multivariable-adjusted[b]	680	0.48	0.15, 0.81	0.003
Pooled original and offspring cohorts				
Crude, generation adjusted	2432	0.49	0.37, 0.61	<0.0001
Multivariable-adjusted[c]	1672	0.52	0.35, 0.69	<0.0001

Table 2. Estimated heritability of age at natural menopause using variance-components methodology in SOLAR: 1296 Framingham Heart Study families (Murabito et al., 2005)

a. Adjusted for mean body mass index, cigarette smoking, and parity.

b. Adjusted for mean body mass index, cigarette smoking, parity, mean alcohol intake, oral contraception use ever, and age at menarche.

c Adjusted for generation (original cohort/offspring), mean body mass index, cigarette smoking, and parity.

Although both menarche and menopause are under genetic control, it seems that there is no correlation between the onset and cessation of menstruation, suggesting different genetic mechanisms (Snieder et al., 1998). Many aspects of human women health are related to menarche and menopause (respectively figure 12 A and 12 B) and deep insights into the genetic basis may have clinical value for understanding, preventing or treating disorders of puberty and/or menopause.

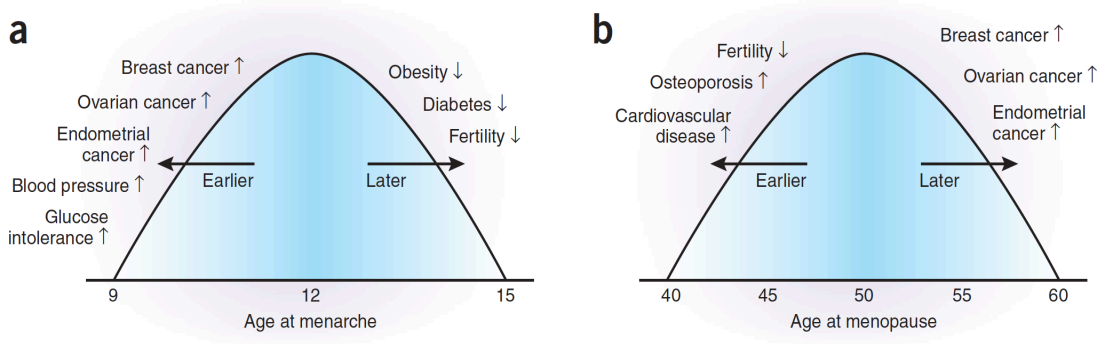


Figure 12. Potential health impacts of ovarian aging. The timing of both age at menarche (a) and age at natural menopause (b) has a wide range of effects on human health. Shown are a range of conditions, in which age at menarche and age at natural menopause are known or suspected to either increase (↑) or reduce (↓) risks of these disorders. (Hartge, 2009)

An earlier age of menarche and a later age of menopause are associated with an increased risk to develop breast, ovarian and endometrial cancer (Velie, Nechuta, & Osuch, 2005).

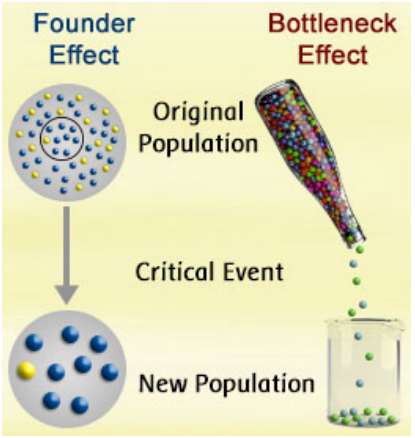
Girls with earlier age at menarche tend to have higher body mass index (BMI) and adiposity in comparison to girls with a later age at menarche (Anderson, Dallal, & Must, 2003). An earlier menopause increases risk of osteoporosis (Khosla & Monroe, 2017).

A set of genes involved in menopause and menarche have been found in recent large genetic studies when assessing the contribution of common variation to these physiological conditions (Stolk et al., 2012 and J. R. Perry et al., 2014). However, the total effect of the identified variants explains only a small percentage of the total hidden heritability and they are not sufficient to predict a potential personalized range of fertility in women. A larger contribution of rare variants has not been assessed yet and it might help in developing predictive tools and in understanding the genetic variation architecture that confers risk for each of the menarche/menopause-related disease conditions.

The resource of isolated cohorts

A particular resource for identifying rare variants are the genetically isolated populations (Varilo & Peltonen, 2004), that derive from a small number of founders, introducing their pool of genetic variations, and persist in a state of

geographical or cultural isolation for a large number of generations during which they do not receive any genetic contribution from the outside. They are also defined as “isolates”.



Ernst Mayr (Provine, 2004) defines the founding effect, which intervenes at the initial stage of isolation, as "the formation of a new population from a small number of individuals (founders) carrying only a small fraction of the genetic variability from their ancestral populations". A population may descend from a small number of founders mainly for two reasons: either a small number of individuals colonize a previously uninhabited place, or a drastic reduction in population size due to sudden environmental changes or extreme circumstances such as famines or wars (bottle neck effect) (Arcos-Burgos & Muenke, 2002). The few surviving individuals then expand in favorable conditions.

Isolation leads to the formation of a different genetic pool compared to the genome of parental or surrounding populations. A further differentiation of the genetic pool occurs mainly to selective adaptation and genetic drift (in the absence of migration). Genetic drifts act much more on small populations, while in large or rapidly expanding populations it is partially offset by the high number of progeny per generation (Cavalli-Sforza “storia e geografia dei geni umani” 2009). During this events some rare allele are lost due to the lack in founders or some may be lost in the next generation. Instead, rare alleles that survive during first generation could be found enriched in the isolated population.

Characteristics of an isolated population are (1) a high degree of inbreeding (that might cause an increase incidence of recessive disease), (2) an increase in regions of homozigosity and (3) extended region of Linkage Disequilibrium (LD). An additional advantage of isolated populations is the environmental and cultural homogeneity: individuals are exposed to the same factors and in this way the power to identify genetic effects increased.

In Italy, thanks to the morphological and geographical position, there are several isolated populations. In 2009 it was founded the Italian Network of Genetic Isolates (INGI) with the aim of a large collaboration between different independent projects: Val Borbera, Carlantino and Friuli Venezia Giulia genetic park (Figure 13).

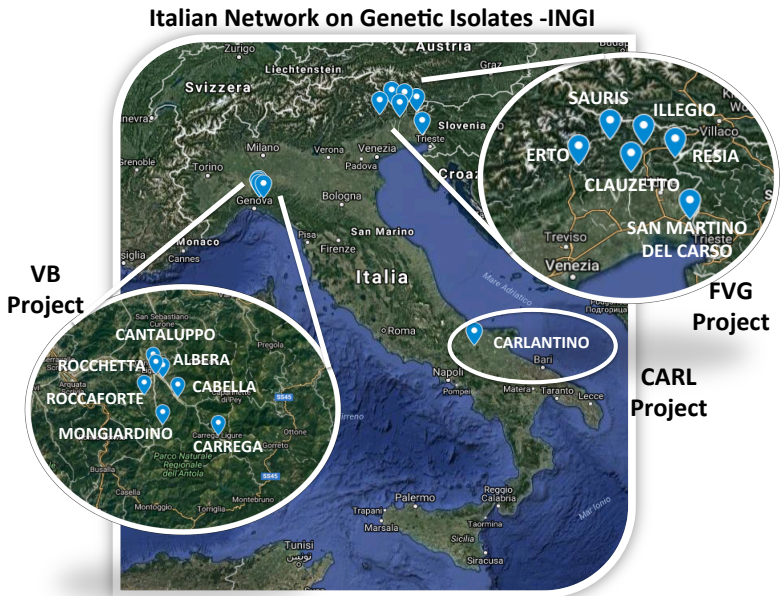


Figure 13. Italian Network of Genetic Isolates:
Val Borbera project, Friuli Venezia Giulia project and Carlantino project

Val Borbera project

The Val Borbera project started in 2005 with the aim to investigate the genetic factors responsible for human complex disease. Population from Val Borbera (VB), a large valley in Northwest Italy, isolated from the surrounding areas by mountains and by a deep canyon on its western side, lives in seven villages. According to the characteristics of the isolated population described above, VB population shows extended region of homozigosity (ROH) and a highest levels of LD compared with other populations as reported in the following figures (Colonna et al., 2012).

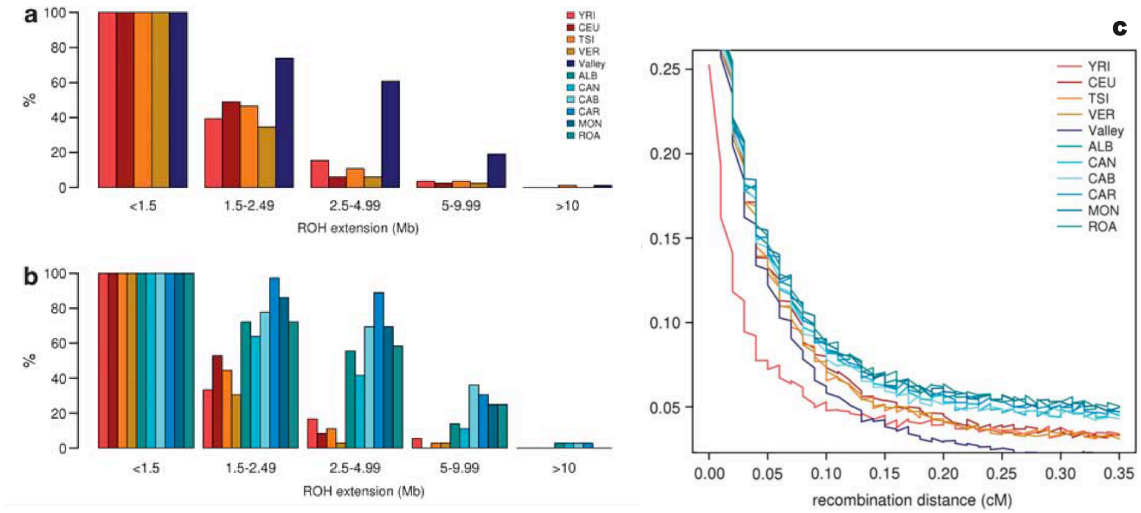


Figure 14. Statistics on the extension of the ROHs in the valley (a) and villages (b) with respect to other populations. (c) Decay of LD with increasing recombination distance measured as average r^2 within recombination distance bins. (Colonna et al., 2012)

An analysis of the first two principal components (PC) estimated between randomly selected VB samples of equal size and several populations of European ancestry, shows the significance of the first PC (Fig. 15). The first significant PC separates African from non-African populations, and the second PC, although not significant, distinguishes the valley from the other European populations, suggesting a poor recent genetic exchanges between the VB population and the other Italian and European populations.

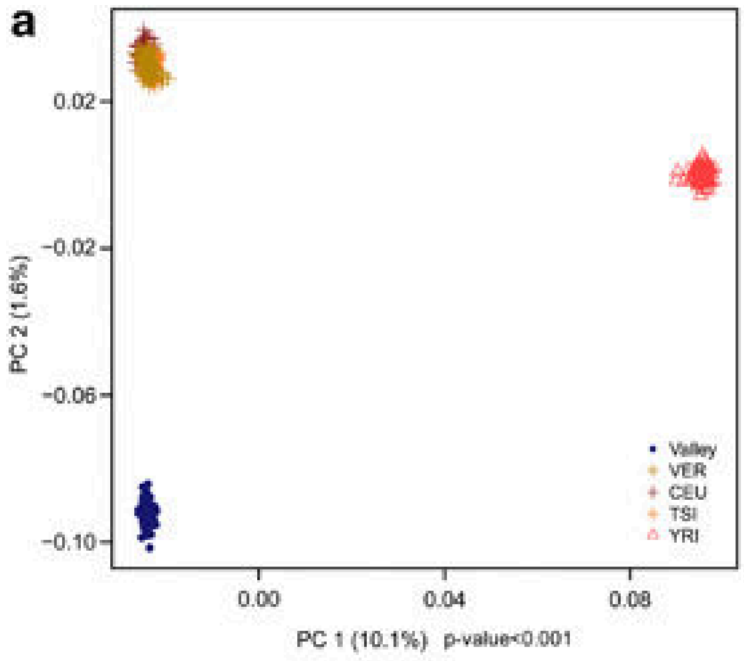


Figure 15. Population multi dimensional plot Valley: Val Borbera, VER:Veneto region, North-East Italy, TSI:Tuscany, Central Italy, CEU: Europe, YRI: Africa (Colonna et al., 2012).

Thanks to the collection of birth, marriage and death records extracted from church and city archives, a genealogical pedigree of about 50,000 people going back up to 16 generation was created.
For this project 1,803 participants were enrolled: the sample was enriched in females (56%) and in older people (mean age 55 years, range 10-102 years) (Traglia et al., 2009) (Figure 16).

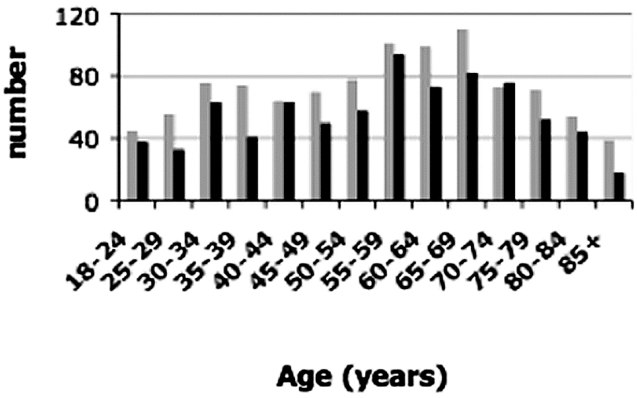


Figure 16. Age and sex distribution of the participants to the study: 5 years periods were considered.In black are the males, in grey are the females (Traglia et al., 2009).

Different phenotypic traits were collected including hematological parameters, anthropometric measures and cardiovascular values. 1785 out of 1,803 samples were genotyped with three different Illumina chips (see materials and methods) and 433 out of 1,785 samples, randomly selected, were whole genome sequenced with a mean coverage of 6X.
The Ethical committee of the San Raffaele Hospital in Milan and the Piemonte Region approved the project; all participants signed an informed consent.

Friuli Venezia Giulia genetic park

The Friuli Venezia Giulia (FVG) project collected 1,840 people (3-92 years of age) from 6 different isolated villages in the Italian FVG region (Clauzetto, Erto, Illegio, Resia, Sauris, San Martino del Carso). The high level of genomic homozygosity and the elevated linkage disequilibrium (Esko et al., 2013) confirmed that it is a genetic isolate. For 1,610 out of 1,840 samples genotype and phenotype data are available; 1,470 are adult (>=18 years). 386 samples, selected to better represent each of the different villages, were whole genome

sequenced, 200 with a mean of 4X as coverage and 186 with a mean of 10X. Ethical committee of the IRCCS Burlo-Garofolo approved the project and a written informed consensus was signed from each participant.

Carlantino project

Carlantino (CARL) is a small village in southern Italy in the province of Foggia. 1,563 individuals took part to the project and for 630 individuals genotype and phenotype data are available. 133 samples were whole genome sequenced with a mean of 10X as coverage. The project was approved by the local administration of Carlantino, the Health Service of Foggia Province and ethical committee of the IRCCS Burlo-Garofolo of Trieste. Subjects gave their written informed consent for participating in these studies.

Cohort	villages	participants	genotyped	exome chip	WGS
VBI	7	1803	1785	1803	**433**
FVG	6	1840	1610	1581	**380**
CARL	1	1563	630	820	**133**
tot	**14**	**5206**	**4025**	**4204**	**946**

Table 3. Data available from INGI cohort

GWAS

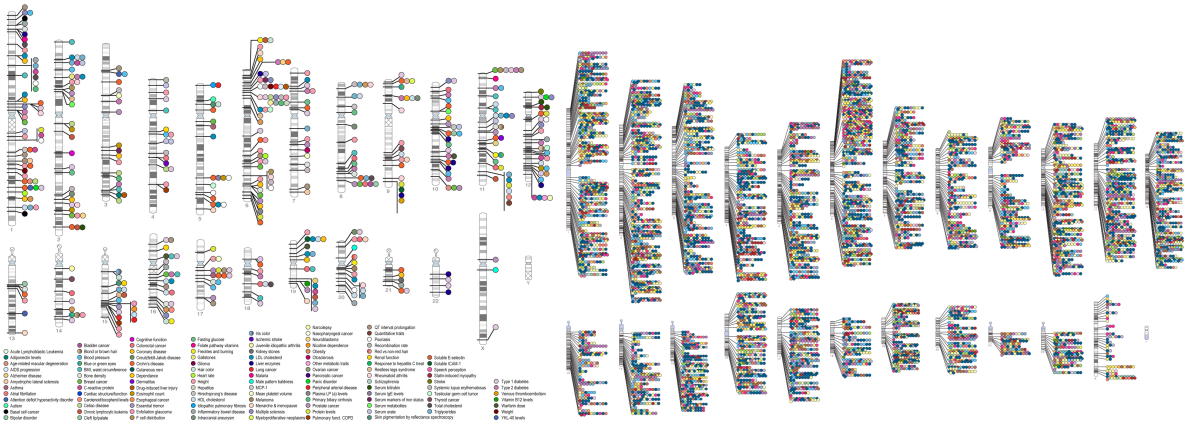
Genome-wide association studies (GWAS) are used to identify DNA variants and genes involved in common human diseases or with a quantitative trait or phenotype. The goal is to identify small variations, single nucleotide polymorphisms (SNPs) or small insertion or deletion (INDELs), which occur with different frequency in people. The study design and analysis may be based on case-control comparison (affected individuals versus controls) or on the genetic variation affecting the quantitative phenotypic variation in the entire population.

The past 10 years have seen an exponential increase of associated genetic loci predisposing to complex diseases: thousands loci have been identified (Hindorff et al., 2009) and in many instances the genetic findings have been a

breakthrough to identify causal mechanism of disease (Voight et al., 2012) and biological targets for new drugs (Chatenoud, Warncke, & Ziegler, 2012).

The GWAS catalog (source GWAS catalog), provided jointly by NHGRI and EMBL-EBI, is the largest collection of published GWAS including at least 100.000 SNPs and association with p-value < 1.0×10^{-5}.

As show in figure 17 A and B great striders have been made in order to identify more loci associated with traits in less than 10 years. The diagram shows all SNP-trait associations with p-values <= 5.0×10^{-8}, mapped onto the human genome by chromosomal locations and displayed on the human karyotype published by the EMBL-EBI and NHGRI GWAS catalogue. On the left loci identified up to June 2009, on the right loci identified up June 2017.



A B

Figure 17. A. Loci identified up to June 2009; B. Loci identified up to June 2017.

The completion of the Human Genome Project (Lander et al., 2001) and the improvements of technologies have allowed this spectacular result. Genotypes could be determined with different commercialized chip (whole genome or exomes) and the number of variants could be increased in genotyped sample using imputation, a cost-effective strategy for expanding the number of variants. Imputation is based on Linkeage Disequilibrium (LD) and permit, thanks to the availability of reference panels of sequenced samples of different ancestry, to infer the presence of variants and to fill 'holes' in study genotypes from SNP arrays (Marchini & Howie, 2008).

Several reference panels have been generated since 2005. The first was generated by the International HapMap consortium and included 269 samples and one million SNPs (phase I) (International, Consortium, & The International HapMap, 2005). Following releases were phase II with 3.1 million SNPs

(Frazer et al., 2007) and HapMap phase III (The International HapMap 3 Consortium et al., 2010) with 1.6 million SNPs in 1184 individuals from 11 populations. Finally, the last release begins to gain in imputation accuracy for low-frequency and rare variants.

1000 Genome Project has permitted a further improvement in imputation panels. The first phase of the project combined low read depth WGS (2-4x) and target deep exome sequencing (50-100x) in 1092 individuals (McVean et al., 2012) while in phase III it was expanded to 2504 individuals from 26 population (1000 Genomes Project Consortium et al., 2015). The advantage to use different populations is to obtain a profile of rare and common variants with considerable geographic differentiation.

1000 Genomes Release	Variants	Individuals	Populations
Phase 3	84.4 million	2504	26
Phase 1	37.9 million	1092	14
Pilot	14.8 million	179	4

Table 4. 1000 Genome Project data available

Others panels based on WGS have been generated in different cohorts: a great example is the UK10K Project that collects WGS (~7x) of 3781 samples of British ancestry from two population-based cohorts. The UK10K reference panel is enriched in rare variants and increases the accuracy of imputation in European population (Huang et al., 2015).

Also the Haplotype Reference consortium has combined low-read-depth WGS data (4-8x) from 20 studies of mainly European ancestry (McCarthy et al., 2016) including 64.976 haplotypes. This consortium is currently continuing to incorporate samples from worldwide population: this effort is fundamental since rare variants are, on average, younger than common variants, showing more geographical clusters and are more difficult to impute (Bomba, Walter, & Soranzo, 2017).

GWAS have allowed the identification of thousands of robust associations with complex diseases and traits but it is now required a valuable and

compelling annotation of the clinical and phenotypic consequences of genome sequence variation, through experiments for investigating the transcription, translation and the epigenetic regulation in order to understand the biological relationship between genotype and phenotype (Bomba et al., 2017).

Statistical power limitation and the contribution of large consortia

A large sample size is required to achieve an adequate statistical power. Since GWAS test a large number of SNP markers, this lead to a large number of multiple comparisons leading to an increase of false positive rates. Bonferroni correction (the significant threshold is set to 0.05 divided by the total number of SNPs analyzed) is usually applied to decrease false positives (Hong & Park, 2012). The genome-wide significance P-value threshold of 5×10^{-8} has become a standard for common-variant GWAS (Fadista, Manning, Florez, & Groop, 2016).

Genetic Power calculator developed by Purcell (Purcell, Cherny, & Sham, 2003) under various assumption about genetic models (i.e. additive, dominant, recessive), allele frequencies, disease prevalence and number of SNPs markers permits to evaluate a correct statistical power and sample size.

For example, a smaller sample size is required with a strong effect size and common SNPs or under a dominant model. Nowadays with the genotype chips available and the possibility to infer with dense reference panel, enriched also in rare variants, a large sample size is required (Hong & Park, 2012).

The attendance of international consortia is essential to reach statistical power. The aim is to group different cohorts with a common interest to study in detail a phenotypic trait or disease.

Human knockout

The study of SNPs and INDELs effects is very useful to identify human knockouts. The loss of function mutations (LoF) in homozygous state might generate a so-called 'human gene knockout' and which might be used as model to identify the function of many genes. LoF are deleterious variants in protein coding genes defined as frameshift, splice acceptor variant, splice donor variant, stop gained, stop lost, start lost, transcript ablation, transcript amplification.

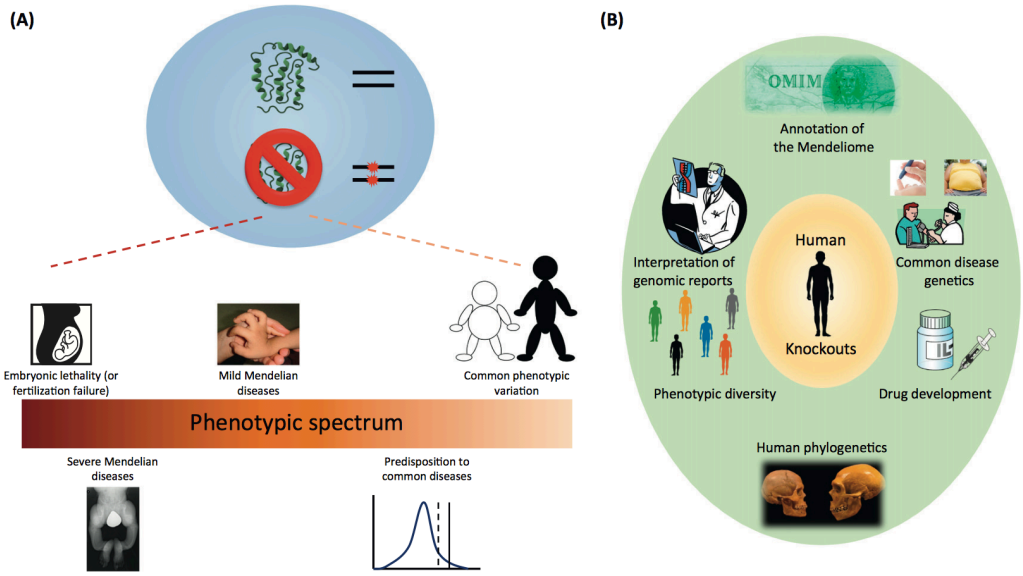


Figure 18. (A) Phenotypes associated with human knockout events can span the entire phenotypic spectrum and are not limited to diseases. (B) Research in human knockouts has a wide range of basic and translational applications. (Alkuraya, 2015)

Phenotypes associated with human knockout events potentially span the entire phenotypic spectrum and are not limited to diseases. A gene with LoF might cause embryonic lethality or severe Mendelian diseases if the gene is important for life, whereas some genes might be inactivated without essential clinical impact, referred as LoF-tolerate genes.

In literature, it is reported that on average every person carries ~200 variants predicted to cause loss of function of protein-coding gene and ~20 genes carrying homozygous LoF alleles and hence are likely completely inactivated (V. M. Narasimhan et al., 2016) (Callaway, 2014).

The human knockout studies represent a valuable resource in discovering not only genes correlated with Mendelian diseases, but also with common genetic diseases. This might shed light in identification of target genes for the development of new drug and also for studying the human phylogenetics (Figure 18 B).

ANALYSIS AND RESULTS

During these 3 years I focused my work on genetics of fertile window in women life, contributing to the identification of genetic variants that regulate the onset of first menstruation in females, i.e “age at menarche”, and the end of fertile life in women, i.e. “age of menopause”.
GWAS analyses, developed and applied in the last 10 years, provide a powerful approach for the discovery of variants and genes contributing to risk of complex diseases. In general the effect of common variants associated with complex traits highlighted in these previous GWAS identified only a small proportion of the predicted genetic variation.
My project aimed to define the genetic architecture of age at menarche, age at menopause and some reproduction behaviors, testing the genetic contribution of low-frequency and rare variants with several approaches in order to improve the understanding of women infertility.
When I started my work, different GWAS had already been performed on “**age at menarche**”. Perry et al., 2014 reported a first set of 106 loci associated with age at menarche: collectively these signals explained only 3% of the trait genetic variance. Many of these signals were already associated with other pubertal traits in both sexes, rare disorders of puberty and nuclear hormone receptors. It was also shown that some of the loci were enriched in imprinted regions, demonstrating parent-of-origin-specific associations.

Thanks to 1000 Genome project and the possibility to impute our genotyping chip data to 1000G reference panel phase 3 we obtained a new resource that allowed to include also low frequency (1%< Minor allele frequency <5%) and rare variants in GWAS (Minor allele frequency <1%). In fact, from 142,722 variants with frequency <=5% in HapMap Phase 3 Release 2 we moved to 10,123,788 low frequency and rare variants in 1000 Genomes Project Phase 3. In addition it was possible to increase the number of samples analyzed and to replicate significant results in independent group of samples (deCoDE study).
In the first two chapters, I report our approaches for the GWAS analysis of “age at menarche” by two separate GWAS. In one we tested a dataset of 192,974 women for low-frequency protein-coding variants genotyped by exome array and X-linked loci, two resources not yet investigated. A second

GWAS was done with a double sample size and genotype chip data imputed to the 1000G reference panel.

Similar was the background of the trait "**age of menopause**": 18 common genetic loci were already associated with this trait in two previous meta-analysis (Stolk et al., 2012 and J. R. Perry et al., 2014) but altogether these variants explained less than 5% of the genetic variance in age at menopause. We performed a meta-analysis with a larger sample (nearly 70,000 women) and for the first time, low frequency coding variants were analyzed for this trait.

In addition, I focused my attention on some reproduction behaviors: **age at first birth** (AFB) and **number of children ever born** (NEB). These could be best standards to measure lifetime reproductive success and to indicate biological fitness. In a previous published work (Tropf et al., 2015), a subset of 6,758 samples from UK and Netherlands were analyzed and results showed significant additive genetic effects on both traits explaining 15% of NEB and 10% of AFB but the underlying mechanism of AFB and NEB are poorly understood. We contributed to the GWAS of both traits and a meta analysis was performed reaching 251,151 samples for AFB and 343,072 for NEB and an additional a genome-wide gene-based analysis using VEGAS (Liu et al., 2010) was performed to increase power to find statistically significant association.

Thanks to the next generation technologies we **whole genome sequenced** (WGS) 947 samples from Italian Network Genetic Isolate (INGI): an extremely useful resource in order to build a custom reference panel to infer our genotypes. In comparison with 1000G reference panel, a custom panel using a denser scaffold of known haplotypes building with data of the same population, improves the imputation quality and specifically the low and rare frequency spectrum of variants (Cocca et al., in preparation). Very rare variants typical of our population could be lost with only 1000G reference panel due to the lack of some specific block of haplotypes.

The analysis of these sequences also leads to the characterization of deleterious variants and specifically to the identification of knockout human genes in our population.

Finally, since **anti mullerian hormone** (AMH) is a marker of ovarian reserve, I have also examined in depth this trait. We have performed a GWAS on the quantitative trait in INGI fertile women with genotype data imputed to Italian Reference Panel enriched in Italian rare variants and we have conducted an analysis in endometriosis patients because 50% of women affected by endometriosis, an inflammatory disease characterized by growth of endometrial tissue at ectopic location, are infertile and cryopreservation could be considered before the accelerated follicular depletion during disease progression.

In summary, the first 4 chapters report our GWAS results that were already published on menarche (Lunetta et al., 2015) (Day et al., 2017), on menopause (Day et al., 2015) and on human behaviors correlated to fertility, such as AFB and NEB (Barban et al., 2016). My contribution to all these papers that I coauthored has been to analyze the INGI-VB cohort and to participate to the global quality control for metanalyses especially in relation to analysis on age at menarche.
In chapter 5, I describe how Whole Genome Sequenced INGI data was obtained and the analyses on human knockouts. Finally, in the last two chapters the results on AMH in healthy and endometriosis affected women are reported.

1. **Rare coding variants and X-linked loci associated with age at menarche.** (Lunetta et al., 2015)

We have participated with our cohorts in GWAS organized from Reprogen (Lunetta et al., 2015) consortium (Reprogen) in order to study two overlooked sources of variants that in previous works (J. R. B. Perry et al., 2014) were not analyzed: low-frequency protein coding variants genotyped with exome chip and X-linked variants. 19 studies with a total of 76,657 European ancestry women were collected for the analysis with exome array data available (61,734 low-frequency variants – MAF <5% - passed QC). Meta-analysis for age at menarche (see Materials and methods - GWAS age at menarche on exome array), restricted to 9-17 years of age, was performed and different significant signals were identified.

The most significant signal was identified in ***ALMS1* gene** (**rs45501594**, MAF 1%, P=4.6x10^{-10}) (Fig. 18). Deleterious mutations in this gene are associated with Alstrom's syndrome (OMIN 203800), a rare autosomal recessive disease characterized by multiorgan dysfunction and in particular sensorineural hearing loss, childhood obesity, diabetes mellitus, menstrual irregularities and hypogonadotropic hypogonadism in males (Joy et al., 2007).

Our signal was confirmed in 116,317 independent women from deCODE (source deCODE) and 23andMe (source 23andMe) that were selected for follow-up. A combined meta-analysis of discovery phase and follow-up data permitted to identify 5 variants genome-wide significant (Fig. 19).

Table 1 | Association statistics for the novel low-frequency and X-chromosome variants.

Gene	SNP	Location	Alleles*	Discovery			Follow-up			Combined		
				Effect (s.e.)	P	N	Effect (s.e.) | VE†	P	N	Effect (s.e.)	P	N
Exome array												
ALMS1	rs45501594	2p13.1	G/C/1.1%	0.26 (0.04)	4.6E−10	57,867	0.23 (0.03) | 0.12%	2.2E-11	116,317	0.24 (0.03)	6.8E-20	174,184
LAMB2	rs35713889	3p21.31	T/C/4.4%	0.11 (0.02)	5.0E−07	58,695	0.08 (0.02) | 0.04%	2.2E-06	116,317	0.09 (0.01)	1.0E-11	175,012
TNRC6A	rs113388806	16p12.1	T/A/4.7%	0.09 (0.02)	1.7E−05	76,657	0.08 (0.02) | 0.04%	1.4E-07	116,317	0.08 (0.01)	1.1E-11	192,974
TACR3	rs144292455	4q24	T/C/0.08%	0.71 (0.15)	1.3E−06	68,487	1.25 (0.25) | 0.20%	8.0E-07	116,317	0.84 (0.13)	2.8E-11	184,804
PRKAG1	rs1126930	12q13.12	C/G/3.4%	−0.11 (0.02)	4.4E−07	76,657	−0.08 (0.02) | 0.02%	3.6E-05	116,317	−0.09 (0.01)	9.6E-11	192,974
1000G X-chromosome												
IGSF1	rs762080	Xq26.2	A/C/24%	−0.07 (0.01)	4.1E−12	76,831	−0.04 (0.01) | 0.04%	6.7E-03	39,486	−0.06 (0.008)	9.4E-13	116,317
FAAH2	rs5914101	Xp11.21	A/G/24%	−0.07 (0.01)	1.1E−09	76,831	−0.03 (0.01) | 0.03%	2.0E-02	39,486	−0.05 (0.009)	4.9E-10	116,317

deCODE, Diabetes Epidemiology: Collaborative analysis of Diagnostic criteria in Europe; SNP, single-nucleotide polymorphism; VE, variance explained.
*Refers to effect allele/other allele/effect allele frequency.
†Beta (standard error) from the combined replication samples | VE in the deCODE study. Units are on a 1-year scale.

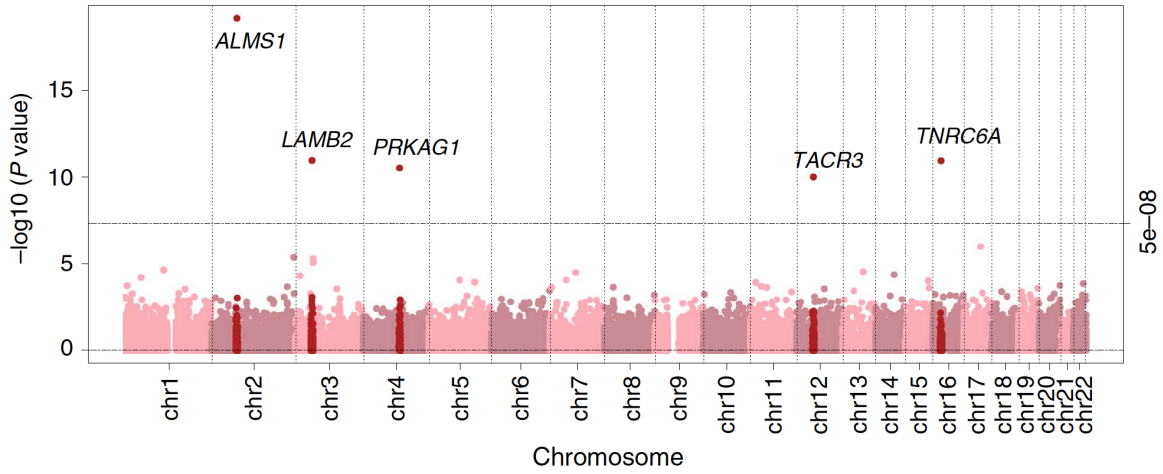


Figure 19. A 'Manhattan plot' of menarche association statistics for the genotyped low-frequency exome array variants. Test statistics are shown from the exome-chip discovery-phase samples, with the exception of the five labelled loci that indicate results from the combined discovery and replication set (Lunetta et al., 2015).

Missense variant in *ALMS1* remained in the combined meta-analysis the strongest signal with an effect size more than double (0.23 years later age in puberty) than any genetic variant previously reported for this trait.

A rare stop-gain (r**s144292455**, MAF=0.08%) in ***TACR3*** (tachykinin receptors 3) is associated with 1.25 years-later age at menarche with combined p-value = 2.8 x 10^{-11}. This premature stop codon in *TACR3* (p.W275X) is associated with hypogonadropic hygonadism (IHH), a rare reproductive disorder (Phenotype MIM number: 614840). This variant both in homozygous and heterozygous state have been already reported in male with 'early androgen deficiency'; however heterozygous cases in IHH showed evidence of spontaneous neuroendocrine recovery.

rs35713889 (MAF = 4%; P=1.1 x 10^{-11}) is associated with 0.08 years-later age at menarche and it is a missense variant in ***LAMB2*** gene that encodes a subunit of Laminin, a glycoprotein in extracellular matrix fundamental for organization of the cell, for migration and attachment. Mutations in *LAMB2* cause Pierson's syndrome (OMIM 609049) characterized by congenital nephrotic syndrome and ocular anomalies. Common variants in or near other Laminin genes have already been reported as associated with different complex traits including colorectal cancer, coffee consumption, type 2

diabetes.

TNRC6A is a member of the trinucleotide repeat containing 6 protein family with a role in post-transcriptional gene silencing through the RNA interference (RNAi) and microRNA pathways. **rs113388806** in this gene is associated with later age at menarche ($p=1.1 \times 10^{-11}$, MAF= 4.7%) and this finding increases the number of epigenetic mechanisms implicated in the regulation of puberty (Lomniczi, Wright, & Ojeda, 2015).

The last significant signal in combined meta-analysis is a low frequency missense variant (**rs1126930**, $p=9.6 \times 10^{-11}$, MAF=3.4%) in ***PRKAG1*** gene which encodes a regulatory subunit of the AMP-activated protein kinase (*AMPK*). *AMPK* is an important energy-sensing enzyme that monitors cellular energy status and it is activated in response to cellular metabolic stresses. *PRKAG1* is overexpressed in ovarian carcinomas (Li, Liu, Chiu, Chan, & Ngan, 2012).

GWAS on X-chromosome

A second GWAS was performed with data on X-chromosome imputed with 1000 Genome reference panel in up to 76,831 women available thanks to 23andMe database (see Materials and methods - GWAS age at menarche on X chromosome). Our results identified two significant signals in/near *IGSF1* and *FAAH2*, both association were confirmed in 39,486 women from deCODE project.

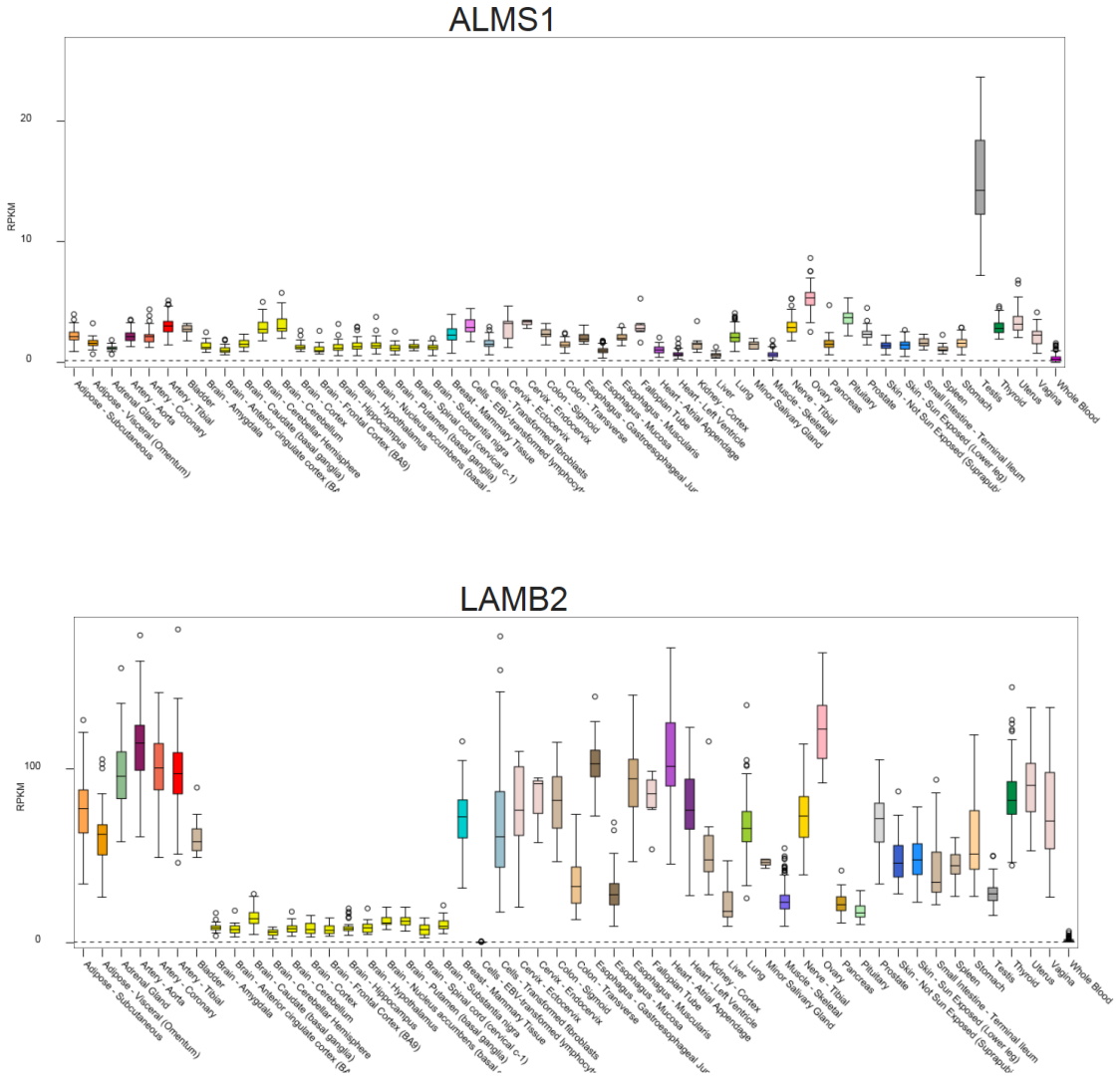
IGSF1 encodes the immunoglobulin superfamily member 1 high expressed in the pituitary gland and testis and common variants (lead SNP: **rs762080**, MAF=24%**,** $p=9.4 \times 10^{-13}$) in this gene are robustly associated with age at menarche in our analysis. In literature rare mutations in *IGSF1* are associated with hypotiroidism, delayed puberty and macro-orchidism in males (OMIM 300888).

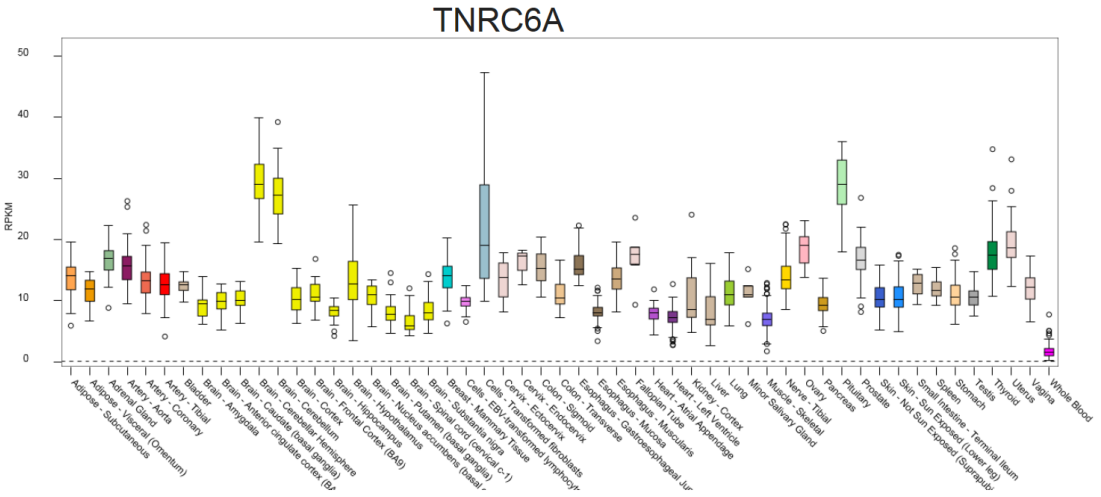
The second significant signal on X-chromosome is in the intronic region of ***FAAH2*** (**rs5914101**, MAF=24%, $p=1.9 \times 10^{-10}$): this is a critical region for Turner's syndrome, a genetic condition that affects development in females

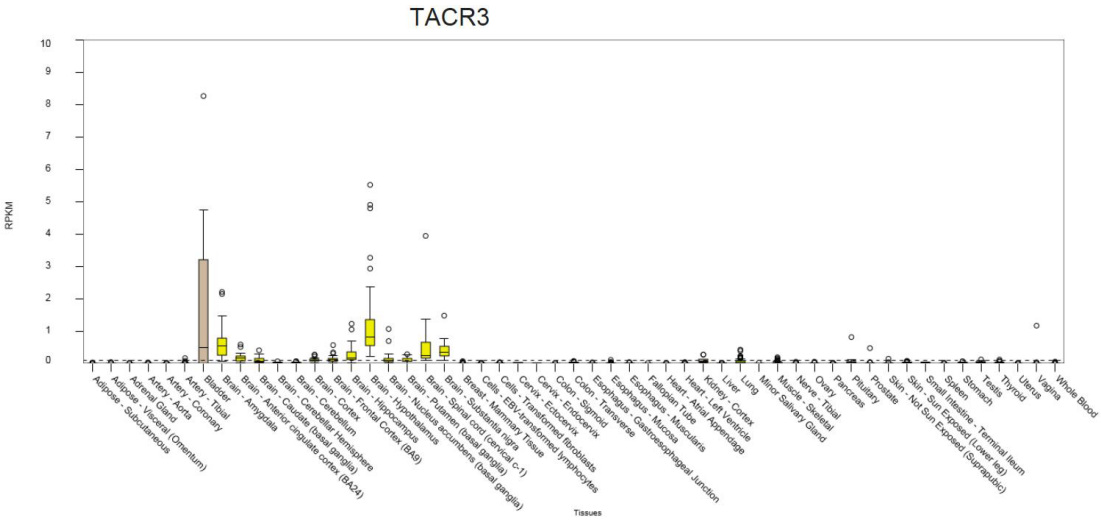
and cause primary ovarian insufficiency.

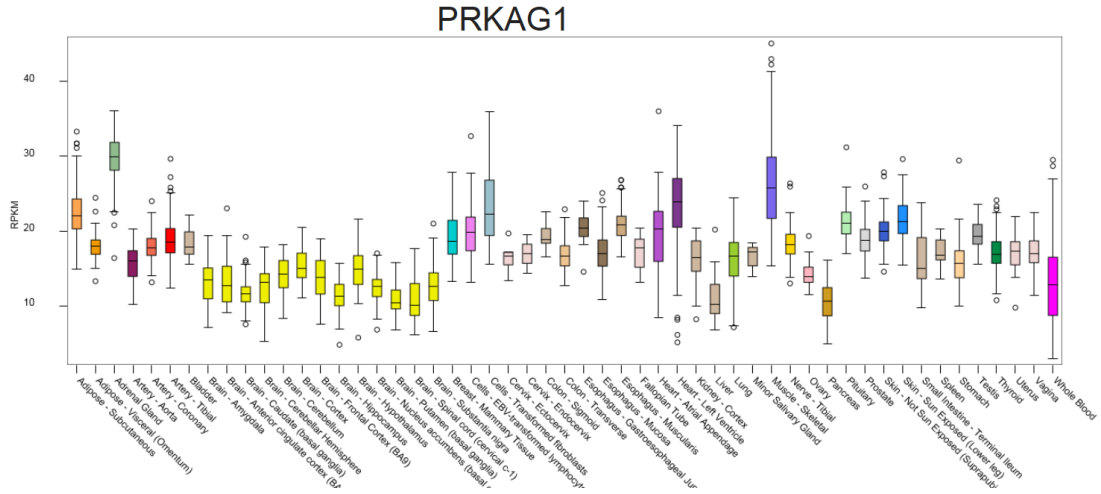
Gene expression data

In order to further characterize the function of significant signals, an additional analysis was performed on the 7 significant genes using expression data on 53 tissue types from Genotype-Tissue Expression consortium. Genes, as showed in the following figures, have a high relative tissue expression in ovary and hypothalamus (Figure 20) but any significant associations were found with mRNA transcript abundance.




TNRC6A
RPKM
0
10
20
30
40
50
Adipose - Subcutaneous
Adipose - Visceral (Omentum)
Adrenal Gland
Artery - Aorta
Artery - Coronary
Artery - Tibial
Bladder
Brain - Amygdala
Brain - Anterior cingulate cortex (Br
Brain - Caudate (basal ganglia)
Brain - Cerebellar Hemisphere
Brain - Cerebellum
Brain - Cortex
Brain - Frontal Cortex (BA9)
Brain - Hippocampus
Brain - Hypothalamus
Brain - Nucleus accumbens (basal ganglia)
Brain - Putamen (basal ganglia)
Brain - Spinal cord (cervical c-1)
Brain - Substantia nigra
Breast - Mammary Tissue
Cells - EBV-transformed lymphocytes
Cervix - Ectocervix
Cervix - Endocervix
Colon - Sigmoid
Colon - Transverse
Esophagus - Gastroesophageal Jur
Esophagus - Mucosa
Esophagus - Muscularis
Fallopian Tube
Heart - Atrial Appendage
Heart - Left Ventricle
Kidney - Cortex
Liver
Lung
Minor Salivary Gland
Muscle - Skeletal
Nerve - Tibial
Ovary
Pancreas
Pituitary
Prostate
Skin - Not Sun Exposed (Suprapubi
Skin - Sun Exposed (Lower leg)
Small Intestine - Terminal Ileum
Spleen
Stomach
Testis
Thyroid
Uterus
Vagina
Whole Blood


TACR3
RPKM
0
1
2
3
4
5
6
7
8
9
10
Adipose - Subcutaneous
Adipose - Visceral (Omentum)
Adrenal Gland
Artery - Aorta
Artery - Coronary
Artery - Tibial
Bladder
Brain - Amygdala
Brain - Anterior cingulate cortex (BA24)
Brain - Caudate (basal ganglia)
Brain - Cerebellar Hemisphere
Brain - Cerebellum
Brain - Cortex
Brain - Frontal Cortex (BA9)
Brain - Hippocampus
Brain - Hypothalamus
Brain - Nucleus accumbens (basal ganglia)
Brain - Putamen (basal ganglia)
Brain - Spinal cord (cervical c-1)
Brain - Substantia nigra
Breast - Mammary Tissue
Cells - EBV-transformed lymphocytes
Cervix - Ectocervix
Cervix - Endocervix
Colon - Sigmoid
Colon - Transverse
Esophagus - Gastroesophageal Junction
Esophagus - Mucosa
Esophagus - Muscularis
Fallopian Tube
Heart - Atrial Appendage
Heart - Left Ventricle
Kidney - Cortex
Liver
Lung
Minor Salivary Gland
Muscle - Skeletal
Nerve - Tibial
Ovary
Pancreas
Pituitary
Prostate
Skin - Not Sun Exposed (Suprapubic)
Skin - Sun Exposed (Lower leg)
Small Intestine - Terminal Ileum
Spleen
Stomach
Testis
Thyroid
Uterus
Vagina
Whole Blood
Tissues


PRKAG1
RPKM
10
20
30
40
Adipose - Subcutaneous
Adipose - Visceral (Omentum)
Adrenal Gland
Artery - Aorta
Artery - Coronary
Artery - Tibial
Bladder
Brain - Amygdala
Brain - Anterior cingulate cortex (Br
Brain - Caudate (basal ganglia)
Brain - Cerebellar Hemisphere
Brain - Cerebellum
Brain - Cortex
Brain - Frontal Cortex (BA9)
Brain - Hippocampus
Brain - Hypothalamus
Brain - Nucleus accumbens (basal ganglia)
Brain - Putamen (basal ganglia)
Brain - Spinal cord (cervical c-1)
Brain - Substantia nigra
Breast - Mammary Tissue
Cells - EBV-transformed lymphocyte
Cervix - Ectocervix
Cervix - Endocervix
Colon - Sigmoid
Colon - Transverse
Esophagus - Gastroesophageal Jur
Esophagus - Mucosa
Esophagus - Muscularis
Fallopian Tube
Heart - Atrial Appendage
Heart - Left Ventricle
Kidney - Cortex
Liver
Lung
Minor Salivary Gland
Muscle - Skeletal
Nerve - Tibial
Ovary
Pancreas
Pituitary
Prostate
Skin - Not Sun Exposed (Suprapubi
Skin - Sun Exposed (Lower leg)
Small Intestine - Terminal Ileum
Spleen
Stomach
Testis
Thyroid
Uterus
Vagina
Whole Blood

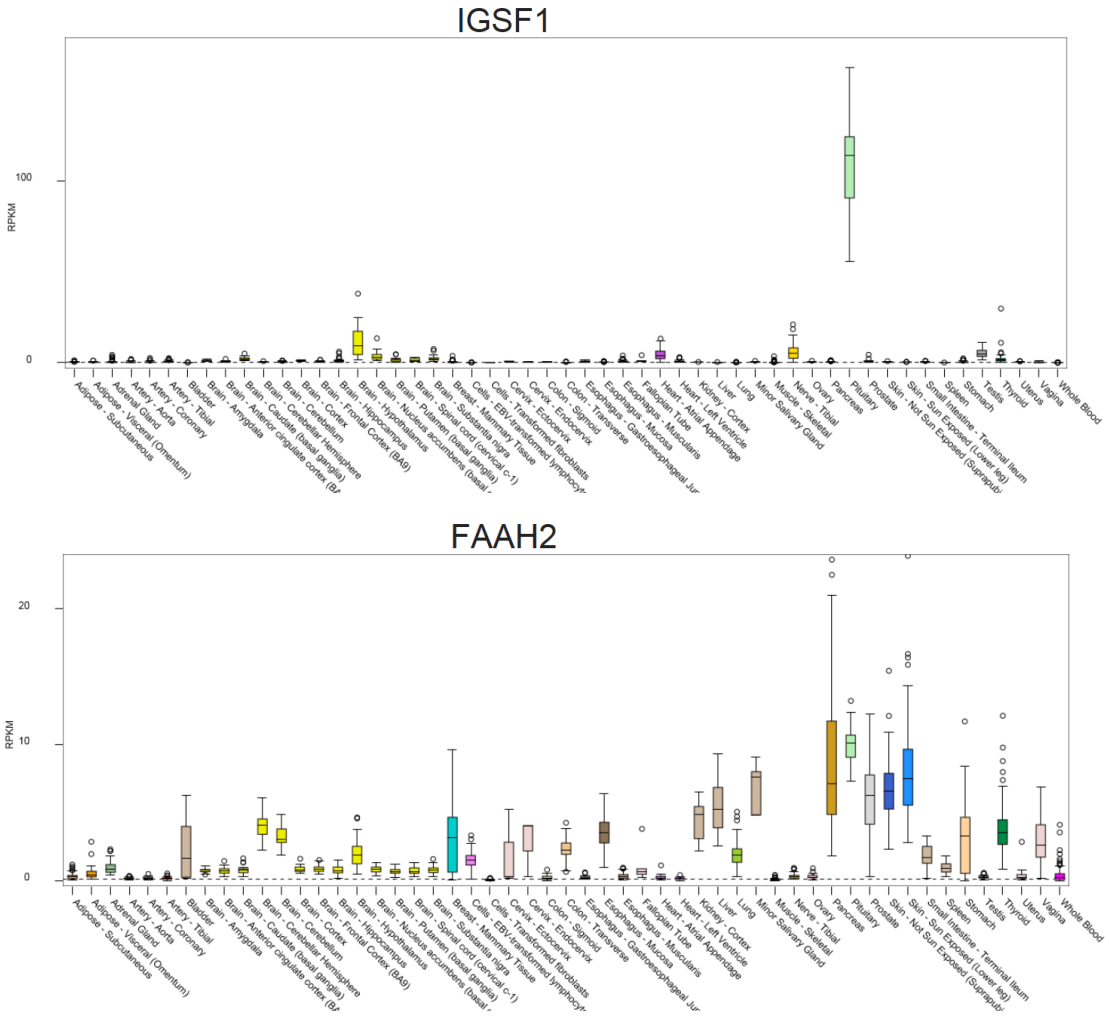


Figure 20. Relative gene expression profiles of highlighted genes using GTeX
(Lunetta et al., 2015)

Based on this evidence, we might conclude that genetic architecture of puberty timing is a complex set of hundred or even thousands of variants, similar to other complex traits. In fact several low-frequency exonic variants and two common signals on X chromosome that reach the significance threshold explained only 0.5% of the variance of age at menarche in the deCODE study. This suggests that also these genetic sources are not enough to find the missing heritability of menarche.

2. Genomic analyses identify hundreds of variants associated with age at menarche and support a role for puberty timing in cancer risk. (Day et al., 2017)

Thanks to the coordination of ReproGen consortium a second meta-analysis on age at menarche (AAM) was performed.

In comparison to previous study (Lunetta et al., 2015) in this analysis the sample size doubled and genotype chip data was imputed to the 1000G reference panel, a more dense panel (see Materials and methods - GWAS age at menarche imputed to 1000G). 329,345 women of European ancestry participant were included in the analysis, 179,117 from ReproGen consortium (40 studies), 76,831 from 23andMe (source 23andMe) and 73,397 from UK Biobank study (Collins, 2012). In total 389 indipendent signals (37,925 variants) were associated with AAM with $P < 5 \times 10^{-8}$ across all 23 chromosome (Fig. 21).

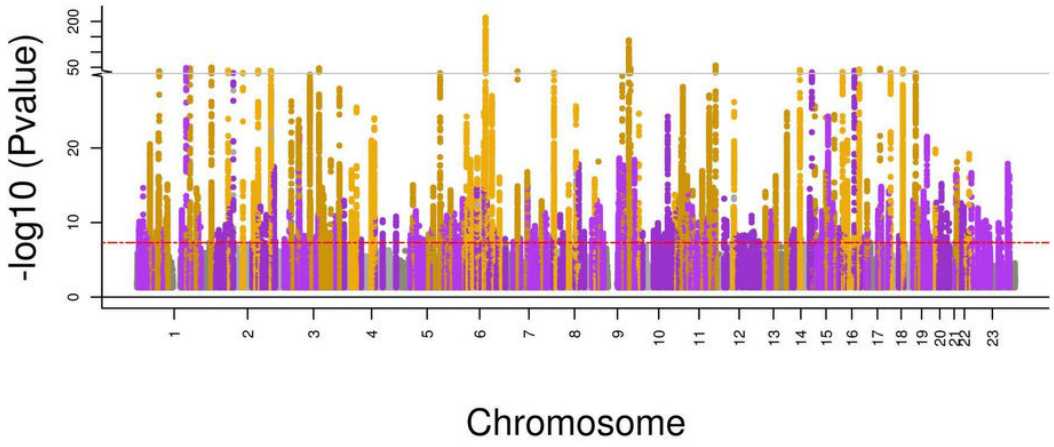


Figure 21. Manhattan plot displaying the genomic locations of the 389 genome-wide significant loci. Previously identified genome-wide significant loci are shown in gold, and new loci are shown in purple. SNPs within 300 kb of the lead SNP at each locus are highlighted (Day et al., 2017).

A subset of 42 index variants were selected because they have a per-allele effect range from ~1 week to 5 month. Among these, 16 variants have low frequency (MAF <5%) and 26 are INDELs.

A replication in deCODE study showed concordant direction of the effect for 367 signals (94.3%) and a combined meta-analysis revealed 368 significant signals. The 389 index SNP in aggregate explained 7.4% of the variance in deCODE and 7.2% in UK Biobank, corresponding to ~25% of the estimated heritability of trait. An analysis on age at voice braking in 55,871 men in

23andMe shows a strong sharing with genetic architectures of AAM in women. 327 of 389 AAM signal have directionally consistent trends.

Implicated genes and tissue

Different analyses were performed to study the implicated genes in AAM: mapping of non-synonymous SNPs, expression quantitative trait locus analysis (eQTL) and integration of Hi-C chromatin-interaction data.

In total 32 gene have non-synonymous variants (8 lead SNP and 24 high correlated – r^2>0.8) (Tab. 5) and are located in genes disrupted in rare disorders of puberty: *GNRH1*, *FUT2*, *KISS1* and *CYP19A1*.

Chr.	Position	r^2	rsID	Ref. Seq.Name	dbSNP functional annotation
1	204190659	0.96	rs4889	KISS1	NSM
3	88140191	0.98	rs7653652	ZNF654	NSM;INT
3	184321878	0.88	rs2178403	EIF4G1	NSM
4	94249688	0.93	rs11722476	SMARCAD1	NSM
5	64724489	1	rs80170948	SREK1IP1	NSM
5	168506153	1	rs2305734	RARS	NSM
6	29792252	0.95	rs45479291	HCG4	NSM
6	41936044	0.82	rs1051130	CCND3	NSM
6	75634715	0.99	rs17414086	SENP6	NSM
6	75715572	0.92	rs9250	SENP6	NSM
7	130023656	1	rs11556924	ZC3HC1	NSM
8	25423284	1	rs6185	GNRH1	NSM
8	52940311	0.94	rs33977775	NPBWR1	NSM
9	7174673	1	rs913588	KDM4C	NSM
9	108965390	0.81	rs7021366	CTNNAL1	NSM
9	109119576	0.86	rs1051474	TMEM245	NSM
11	237087	0.87	rs1045288	PSMD13	NSM
11	244106	0.91	rs10902112	PSMD13	NSM;INT
11	244108	0.91	rs7107362	PSMD13	NSM;INT
11	244115	0.91	rs7128044	PSMD13	NSM;INT
11	244129	0.88	rs7116130	PSMD13	NSM;INT
11	244141	0.88	rs7128029	PSMD13	NSM;INT
11	27658369	0.86	rs6265	BDNF	NSM
11	119172947	0.81	rs643423	NLRX1	NSM
11	119182117	0.83	rs4245191	NLRX1	NSM
11	119188695	1	rs1815811	PDZD3	NSM
12	108224853	1	rs3764002	WSCD2	NSM
14	60437039	0.9	rs1254319	C14orf39	NSM
15	51222375	1	rs28757184	CYP19A1	NSM
15	74044292	1	rs5742915	PML	NSM
16	19872042	0.88	rs61742688	GPRC5B	NSM;SYN
16	20359488	0.9	rs9652588	PDILT	NSM
16	20359494	0.9	rs9652589	PDILT	NSM
16	24793633	1	rs113388806	TNRC6A	NSM
17	81246424	0.93	rs2725405	SLC38A10	NSM
19	1819126	0.91	rs2396359	REXO1	NSM
19	35733804	0.88	rs231591	KMT2B	NSM
19	48703417	0.85	rs601338	FUT2	NSN
19	48703728	0.96	rs602662	FUT2	NSM
19	58478128	0.94	rs893185	ZNF446	NSM
20	56248973	0.98	rs3827103	MC3R	NSM

Table 5. Non-synonymous variants (Day et al., 2017)

In order to understand if significant variants influence gene expression, public gene expression dataset were examined (see Materials and methods – Gene expression data integration). First, a comparison was done with available whole-blood eQTL data set published by Westra (Westra et al., 2013): Summary Mendelian Randomization approach (Zhu et al., 2016) prioritized 113 transcripts, 60 of which had evidence for causal or pleiotropic effects (see Materials and methods - Mendelian Randomization analyses). Secondly, an analysis on 46 GTEx tissues have shown that 5 tissues, involved in central nervous system, are positively enriched for AAM (Figure 22). Later AAM was associated with higher transcript levels of *LIN28B* in the pituitary, *NCOA6* in the cerebellum and *HSD17B12* in various tissues.

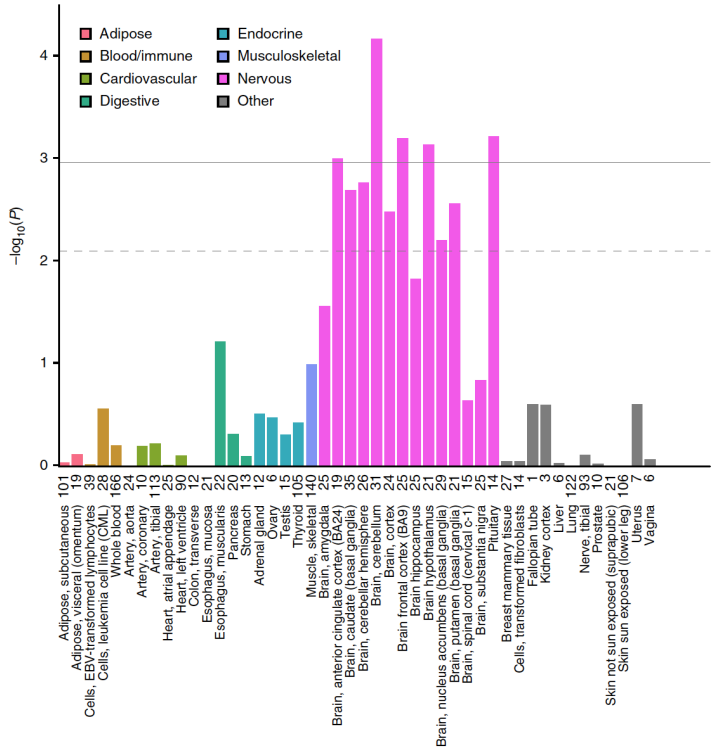


Figure 22. GTEx tissue enrichment using LD score regression. Numbers on the x axis correspond to sample number for each tissue. The dashed line represents significance at FDR < 5%, and the solid horizontal line represents Bonferroni-corrected significance for the number of tissues tested (Day et al., 2017).

335 of the 389 loci were located in regions of chromatin looping containing chromatin contact points (TADs): data of significant Hi-C interactions and contact domains were obtained from Rao at al (Rao et al., 2014) in order to identify possible distal causal genes. 66 signals have a direct physical connection; in fact they are located in a specific contact point. 22 meta-

analysis significant signal, that were in gene desert regions, were in TADs contained notable distal candidate genes as *INHBA*, *BDNF*, *JARID2* and several other AAM signals resided within one TAD containing the same single gene for example TACR3 with a signal in 5'UTR, 2 signal upstream and 1 signal downstream.

Transcription factor binding enrichment

2382 transcription factors were tested, in order to identify network involved in regulation of AAM, combining DNase-1 hypersensitive sites and chromatin sites in 111 cell types and tissue.

16 transcription factor binding motifs were enriched for co-occurrence with AAM-associated variants (Table 6).

Transcription factor	Genes	Start	Stop	FDR
ELF1_1	13	41506055	41556418	0.000685671
SPIC_1	12	101871335	101880775	0.000840042
SPIB_2	19	50922195	50934309	0.003093769
TEAD1_4	11	12695969	12966284	0.00600102
MYF6_2	12	81101408	81103256	0.007099359
GATA_3	-			0.008257109
CTCF_1	16	67596310	67673088	0.010031085
FOXC2_1	16	86600857	86602537	0.015555478
HESX1_1	3	57231944	57234280	0.015559573
RXRB_1	6	33161362	33168630	0.017855659
NR2F1_2	5	92919043	92930315	0.018061308
TAL1_5	1	47681962	47698007	0.0190937
NR2C2_1	3	14989091	15090786	0.019105021
SMAD3_3	15	67358195	67487533	0.027643611
PITX1_1	5	134363424	134369964	0.036859677
MSX1_5	4	4861392	4865660	0.039625118

Table 6. transcription factor enriched for co-occurrence with AAM-associated variants. False Discovery Rate < 0.05 (Day et al., 2017).

Pathway analyses

Software MAGENTA (Ayellet et al., 2010) was used to identify pathway associated with AAM: ten pathways reached the significance (see Materials and methods – Pathway analyses). Five out of 10 pathways were related to transcription factor binding and the other five pathways were related

respectively to peptide hormone binding, PI3-kinase binding, angiotensin-stimulated signaling, neuron development and γ-aminobutyric acid (GABA)-type B receptor signaling.

The strongest AAM signal is in LIN28B gene, which is linked to the repression of let-7 family of microRNAs. From literature we know that let-7 miRNA targets are enriched for variants associated with type 2 diabetes and transgenic Lin28a/b mice demonstrate that both alter pubertal grown and glycaemic control. This suggest that there could be a link between puberty timing and type 2 diabetes in humans.

Imprinted genes and parent-of-origin effects.

This findings confirmed enrichment in imprinted genes associated with age at menarche (Table 7), in particular we identified a rare 5' UTR variant rs530324840 in MKRN3 and a rare intergenic variant at the DLK1 locus both associated with AAM under the paternal model but not under the maternal one (see Materials and methods – Parent-of-origin-specific associations and variance).

		Allele		Freq.		Additive		Maternal		Paternal		
Marker	Position (hg38)	A1	A2	A1 (%)	Locus	*P*	β^a	*P*	β^a	*P*	β^a	$P_{mat\ vs.\ pat}$[b]
rs530324840[c]	15:23,565,461	A	C	0.80	*MKRN3*	4.4×10^{-4}	−0.206	2.0×10^{-1}	0.098	6.4×10^{-11}	−0.523	1.3×10^{-7}
rs184950120[c]	15:23,565,696	T	C	0.26	*MKRN3*	1.0×10^{-2}	−0.265	9.8×10^{-1}	0.003	1.5×10^{-4}	−0.502	4.9×10^{-2}
rs12148769[c]	15:23,906,947	A	G	10.1	*MKRN3*	5.8×10^{-6}	−0.078	3.4×10^{-1}	−0.022	9.2×10^{-8}	−0.120	2.3×10^{-3}
rs138827001[d]	14:100,771,634	T	C	0.36	*DLK1*	6.8×10^{-6}	−0.387	8.8×10^{-1}	−0.018	4.7×10^{-10}	−0.704	1.4×10^{-4}
rs10144321[d]	14:100,416,068	G	A	23.0	*DLK1*	5.6×10^{-6}	−0.056	4.0×10^{-1}	−0.014	1.9×10^{-7}	−0.084	9.7×10^{-3}
rs7141210[d]	14:100,716,133	T	C	38.2	*DLK1*	4.5×10^{-2}	0.021	1.5×10^{-1}	−0.021	2.3×10^{-5}	0.059	4.0×10^{-4}
rs61992671[e]	14:101,065,517	A	G	48.5	*MEG9*	4.7×10^{-3}	−0.029	6.0×10^{-8}	−0.077	2.7×10^{-1}	0.015	1.9×10^{-5}

Table 7. Parent-of-origin-specific associations between sequence variants at *MKRN3*, *DLK1* and *MEG9* and AAM in Iceland (*N* = 39,543) (Day et al., 2017)

[a]The effect of allele A1 in years per allele. [b]*P* value for heterogeneity between paternal and maternal allele associations. [c]rs530324840 is a new variant identified by the parent-of-origin-specific analysis. rs184950120 is the rare variant identified by the meta-analysis. rs12148769 is the previously reported intergenic common signal (ref. 3). [d]rs138827001 is a new variant identified by the parent-of-origin-specific analysis. rs10144321 and rs7141210 are previously reported common variants (ref. 3). [e]rs61992671 is a suggestive new parent-of-origin-specific association signal.

Disproportionate genetic effects on early or late puberty timing.

The impacts of genetic and environmental factors show age-related difference: heritability and effect are higher for early AAM than for late AAM (h^2_{snp}= 28.8% for early and h^2_{snp}= 21.5% for late AAM). 57.7% of autosomal index SNPs had larger effect estimates on early than on late AAM.

Effects of puberty timing on cancer risk

Increasing AAM, using a model adjusted for BMI, was associated with lower risk for breast and ovarian cancer in women; this could be mediated by a shorter duration of exposure to sex steroids.

All this findings suggest the genetic complexity of the puberty regulation and the large set of new associated genes discovered.

3. Large-scale genomic analyses link reproductive aging to hypothalamic signaling, breast cancer susceptibility and BRCA1-mediated DNA repair. (Day et al., 2015)

Natural age of menopause (ANM), as already reported, has a substantial impact on infertility: it is estimated that fertility ceases 10 years before menopause and this could be relevant for women that decide to delay childbearing. Early menopause is associated with many aspect of human health: a lower risk of breast cancer but higher risks of type II diabetes, cardiovascular disease and osteoporosis.

The aim of this study is to analyze both common and low-frequency coding variants that influenced natural age of menopause (ANM): a meta-analysis of up to 69,360 women of European ancestry with genotype date imputed to HapMap2 and a meta-analysis of up to 39,026 women genotyped on exome arrays were performed with ANM as trait (see Materials and methods - GWAS age at menopause on all genome and GWAS age at menopause on exome chip data).

54 independent signals across the genome were significant associated (1,208 SNPs with $P< 5X10^{-}8$) and their minor allele frequency ranged from 7 to 49% and effect size from 0.07 to 0.88 years for allele with no significant heterogeneity between studies. The top 54 SNPs explain 6% of the variance in ANM, if we consider the top 29,958 independent variants with P< 0.05 the variance explained increase to 21% (Table 8) (see Materials and methods - Estimating variance).

Cut Off for SNP set	Number of SNPs	Variance Estimate	SE	p
All QD'd (unclumped)	2797986	0.3407	0.1789	0.02676
0.05	29958	0.2099	0.0974	0.01261
0.005	5602	0.1717	0.0466	6.47E-05
5.00E-04	1070	0.0848	0.0242	4.16E-05
5.00E-05	378	0.0825	0.0186	1.33E-09
5.00E-06	208	0.0785	0.0177	2.50E-12
5.00E-07	128	0.0604	0.0161	2.06E-11
GWAS sigificant	54	0.0567	0.0161	6.26E-12

Table 8. Estimation of variance (Day et al., 2015)

Using variants from exome array, we identified only one significant signal: two low-frequency highly correlated (r2=0.73, D'=1) missense variants in *HELB* gene, a DNA helicase that unwinds DNA during replication, transcription, repair. A combined analysis with our women and other 10,157 sample from deCODE study confirmed the signal on *HELB* and found a new one on *SLCO4A1* gene, a solute carrier that transport organic anions such as thyroid hormones and estrone-3-sulfate. Both this signals, in *HELB* and in *SLCO4A1*, were identified also by our HapMap 2-based GWAS meta done in parallel (in purple in figure 23).

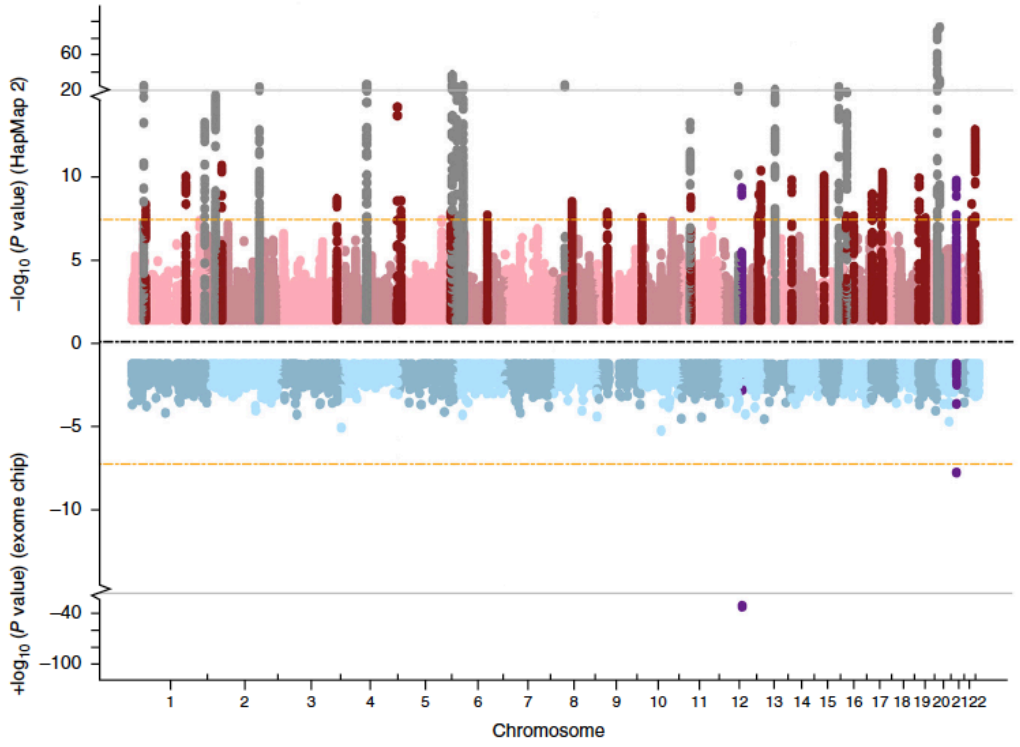


Figure 23. Miami plot of HapMap and exome SNP associations. Log-transformed *P* values are shown for association with ANM for SNPs from HapMap 2 (top; pink) and SNPs from the meta-analysis of exome chip data (bottom; blue). Previously known signals are shown in gray, and newly discovered signals are shown in red (HapMap 2) or purple (exome chip and HapMap 2). The yellow lines correspond to genome-wide significant levels in each direction; the gray lines indicate where the *y* axis has been truncated (Day et al., 2015).

The common rs12371165 in *HELB* is fully explained by the two rare exome chip SNPs which are in high LD each other (Figure 24). The substitution of aspartate by a non polar residue at amino acid 506 of DNA helicase B affect the binding of the helicase to replication protein A as demonstrated with functional studies (Guler et al., 2012). The three significant variants on *SLCO4A1* are non-redundant signals, the association with each of these

variants are unaffected by the presence of the others two.

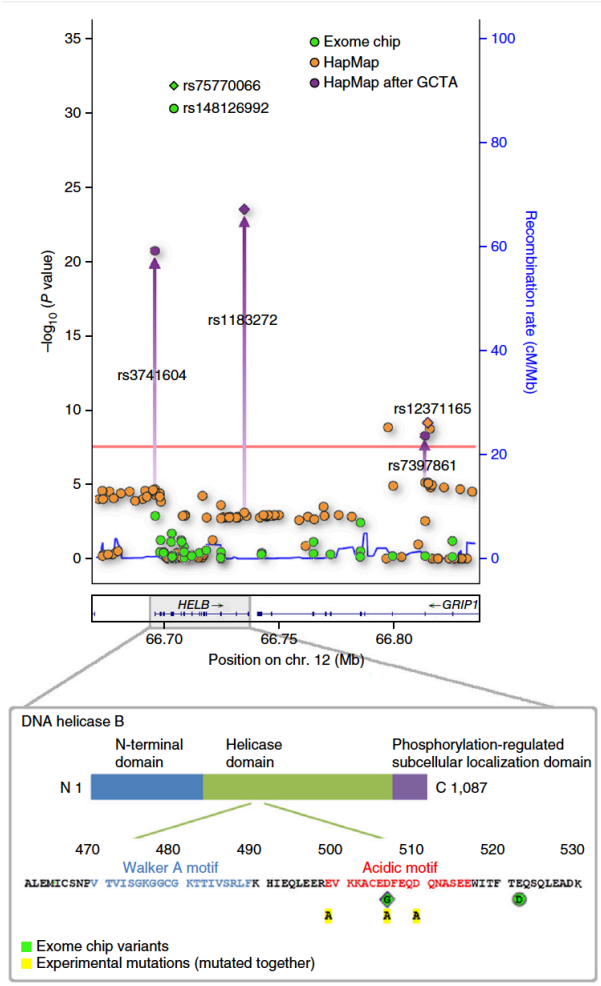


Figure 24. Multiple signals at HELB and relationship to DNA helicase B protein sequence. Positions are given in Build 37 coordinates of the reference genome. The top signal from the exome chip analysis maps to an acidic motif of DNA helicase B and results in the replacement of an acidic aspartate residue by a nonpolar glycine residue (Day et al., 2015).

ANM SNPs strongly enriched in DNA damage response pathways

Enrichment in pathway involved in DNA damage response (DDR) among the significant signal resulted from an analysis using MAGENTA (Ayellet et al., 2010): 29 regions GWAS-significant contain one or more DDR genes within 500kb (see Materials and methods – Pathway identification) (Figure 25). rs1799949 on chr7 with p-value 8.4×10^{-11} is highly correlated with 4 common non-synonymous variants in *BRCA1*; this signal is an eQTL for *BRCA1* in multiple tissues, including blood, skin, brain and adipose. The data show that ANM-lowering allele reduce *BRCA1* expression in blood; *BRCA1* directly inhibits the transcriptional activation function of ERα and the altered estrogen

signaling could in turn affect ANM. A STRING (Szklarczyk et al., 2015) analysis identified 15 ANM signal direct linked to *BRCA1* and in particular 7 genes encoded binding partner of *BRCA1* (*BRE*, *MSH6*, *POLR2H*, *FAM175A*, *UIMC1*, *RAD51* and *CHEK2*).

In addition to genes involved in homologous recombination for the repair of double-strand breaks as BRCA1, other DDR mechanism are presented from ANM signals: *MSH5* e *MSH6* genes as an example of mismatch repair, *APEX1* and *PARP2* involved in base-excision repair, *CHEK2* and *BRSK1* as damage checkpoint. This show how DNA damage during fetal development could results in apoptosis and thus in a reduction of oocytes pool both during mitosis and meiosis phases of oocyte development.

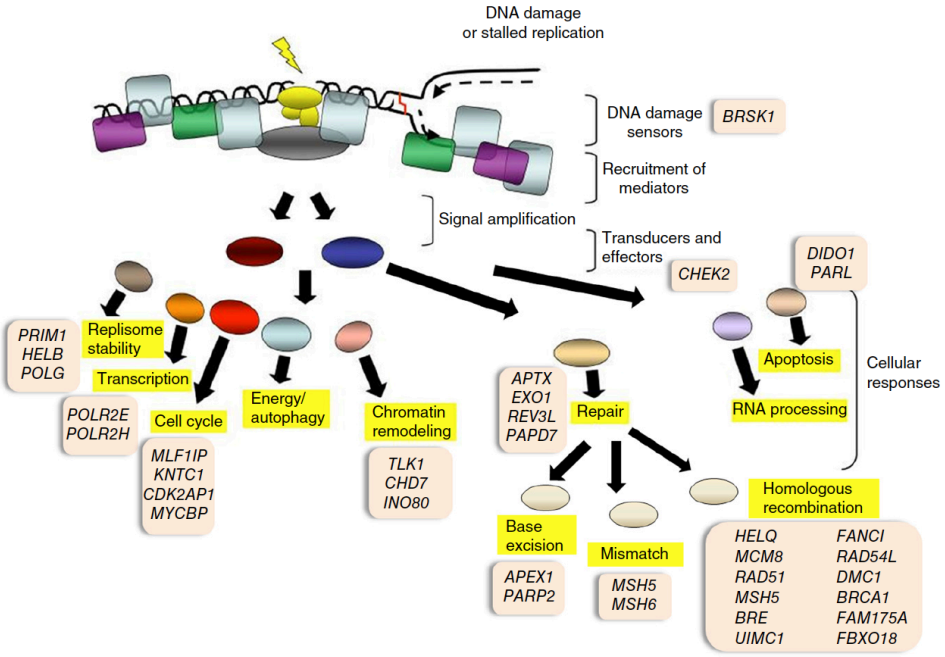


Figure 25. Classification of the genes identified as being involved in DDR pathways at genetic loci associated with ANM (Day et al., 2015).

ANM SNPs enriched in known POI genes and correlation with other traits and disease

Four GWAS significant regions are in or near genes already reported associated with POI: *MCM8* (Desai et al., 2017), *EIF2B4* (Fogli et al., 2003), *POLG* (Trifunovic et al., 2004) and *MSH5* (Mandon-Pépin et al., 2008). In particular recessive mutations in *MCM8* identified in a recent study (AlAsiri et al., 2015) show as deficiency in double-strand break repair cause DNA instability with a great effect on the oocyte pool causing depletion of oocytes

and thus ANM.

We identified an overlap between the significant signals and GWAS catalog for other traits as liver enzymes, lipids, urate, height and fasting glucose but no one overlap with any autoimmune traits.

Significant signals for GWAS ANM were found in or near 5 genes reportedly causal for hypogonadotropic hypogonadism: CHD7, FGFR1 and SOX10 already involved in Kallmann syndrome (anosmic hypogonadotropic hypogonadism due to failure of embryonic migration of gonadotropin-releasing hormone (GnRH)-secreting neurons from the olfactory bulb to the hypothalamus (Silveira & Latronico, 2013)), KISS1R encoded the receptor for kisspeptin (a key hypothalamic activator of the reproductive hormone axis) and TAC3 encoded a member of the tachykinin family of secreted neuropeptides. Recent studies have identified expression of TAC3, KISS1R and kisspeptin in ovarian granulosa cells: these neuropeptides and their receptors could have a peripheral action (Garcìa-Ortega et al., 2014).

A Mendelian randomization approach was used to test the causal relationship between breast cancer and ANM: a positive correlation between the risk of breast cancer and increasing in ANM was identified due to the probability of prolonged estrogen and progesterone exposure.

In conclusion, we underlined the importance of repair DNA mechanisms: in fact during fetal development 7 million of oogonia are produced by mitosis and DNA damage in this phase and during recombination at meiosis could cause apoptosis and a reduction of the initial oocyte pool. Aberrant repair throughout life could affect the rate of atresia and thus ANM. Also this work underlines the usefulness of genetics in identifying biological pathways in the studied traits.

4. Genome-wide analysis identifies 12 loci influencing human reproductive behavior. (Barban et al., 2016)

We examined in depth two traits strictly connected to fertility: age at first birth (AFB) and number of children ever born (NEB), in collaboration with Nicola Barban from University of Oxford. Human reproductive behavior may influence the fitness, infertility and also risk of neuropsychiatric disorders. In industrialized societies prenatal, infant and child mortality is reduced thanks to improvement in care and hygiene and so NEB could be an optimal standard to measure lifetime reproductive success indicating biological fitness.

A large meta-analysis of 62 cohorts of European ancestry (251,151 and 343,072 individuals respectively for AFB and NEB) was performed separately for men, women and both sex combined for the two traits. For AFB analysis only people with offspring were considered, while for NEB trait only men with age >= 55 years and women >=45 years (end reproductive period) were included in the analysis. AFB and NEB analysis showed respectively 10 loci significantly associated (9 in both sex combined and one in women only) and 3 loci (2 in both sex combined and 1 only in men) (Fig. 26, Tab. 9).

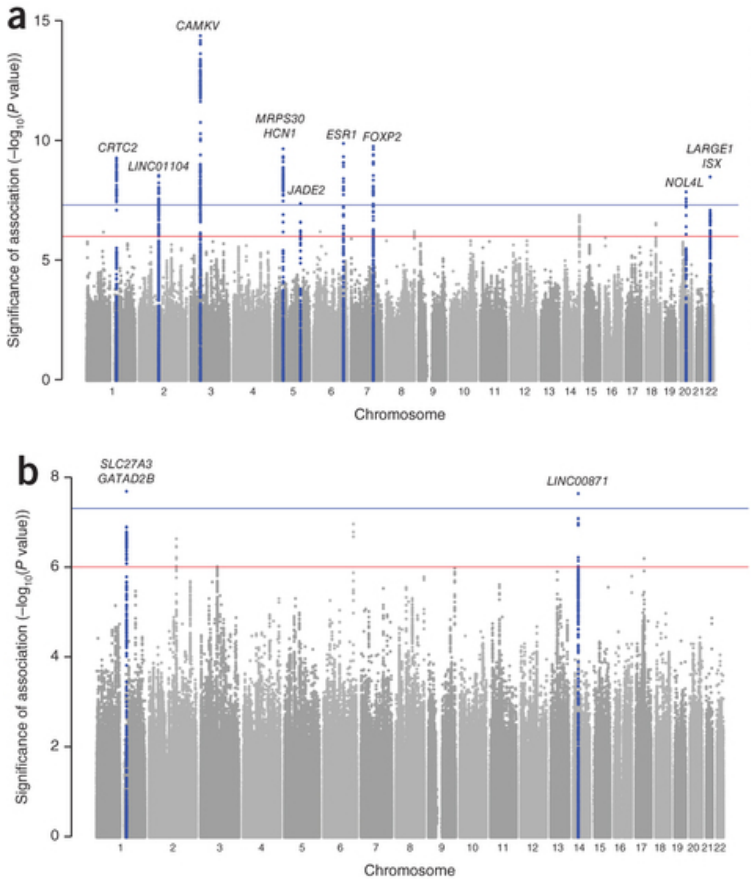


Figure 26 (a,b). SNPs are plotted on the *x* axis according to their position on each chromosome against association with AFB (a) and NEB (b). The solid blue line indicates the threshold for genome-wide significance ($P < 5 \times 10^{-8}$), and the red line represents the threshold for suggestive hits ($P < 5 \times 10^{-6}$). Blue points represent SNPs in a 100- kb region centered on genome-wide significant hits. Loci are annotated with the names of the genes closest to the significant SNPs. (Barban et al., 2016)

SNP	Chr.	Position (bp)	Nearest gene	Annotation	Effect allele/ other allele	EAF	β	P value	n (pooled)	β (men)	P value (men)	β (women)	P value (women)
Age at first birth													
rs10908557	1	153,927,052	*CRTC2*	N, R, ctQ, ctM	C/G	0.695	0.091	**5.59×10^{-10}**	249,025	0.185	2.98×10^{-7}	0.076	5.38×10^{-6}
rs1160544	2	100,832,218	*LINC01104*	R, cQ, cM	A/C	0.395	−0.082	**2.90×10^{-9}**	250,330	−0.042	2.12×10^{-1}	−0.092	**5.00×10^{-9}**
rs2777888	3	49,898,000	*CAMKV*	N, R, ctQ, ctM	A/G	0.507	0.106	**4.58×10^{-15}**	250,941	0.155	2.40×10^{-6}	0.095	**6.07×10^{-10}**
rs6885307	5	45,094,503	*MRPS30, HCN1*	R, ctQ, cM	A/C	0.799	−0.107	**2.32×10^{-10}**	248,999	−0.131	2.07×10^{-3}	−0.104	**3.90×10^{-8}**
rs10056247	5	133,898,136	*JADE2*	cQ, cM	T/C	0.289	0.082	**4.37×10^{-8}**	249,429	0.050	1.68×10^{-1}	0.089	1.28×10^{-7}
rs2347867	6	152,229,850	*ESR1*	cM	A/G	0.649	0.091	**1.38×10^{-10}**	248,039	0.098	4.69×10^{-3}	0.097	**1.80×10^{-9}**
rs10953766	7	114,313,218	*FOXP2*	cM	A/G	0.429	0.087	**1.82×10^{-10}**	248,039	0.106	1.31×10^{-3}	0.089	**8.41×10^{-9}**
rs2721195	8	145,677,011	*CYHR1*	R, cQ, ctM	T/C	0.469	−0.073	6.25×10^{-7}	250,493	−0.014	6.85×10^{-1}	−0.099	**6.13×10^{-9}**
rs293566	20	31,097,877	*NOL4L*	cQ, cM	T/C	0.650	0.081	**1.41×10^{-8}**	245,995	0.110	1.47×10^{-3}	0.079	1.31×10^{-6}
rs242997	22	34,503,059	*LARGE1, ISX*		A/G	0.613	−0.084	**3.38×10^{-9}**	238,002	−0.139	8.51×10^{-5}	−0.076	1.82×10^{-6}
Number of children ever born													
rs10908474	1	153,753,725	*SLC27A3, GATAD2B*		A/C	0.384	0.020	**2.08×10^{-8}**	342,340	0.021	8.10×10^{-4}	0.020	7.89×10^{-6}
rs13161115	5	107,050,002	*EFNA5, FBXL17*	cM	C/G	0.234	−0.041	1.34×10^{-2}	341,737	−0.041	**1.37×10^{-8}**	0.005	3.29×10^{-1}
rs2415984	14	46,873,776	*LINC00871*	cM	A/G	0.470	−0.020	**2.34×10^{-8}**	315,167	−0.029	2.41×10^{-6}	−0.016	3.71×10^{-4}

Table 9. The rows in bold correspond to the independent signals reaching $P < 5 \times 10^{-8}$ in the meta-analysis. Annotation shows for each of the 12 independent lead SNPs (excluding rs10908474 on chromosome 1) whether it is (i) in strong LD ($r^2 > 0.8$) with a nonsynonymous variant (N) or one or more variants prioritized by RegulomeDB (R) with evidence of having functional consequences (defined by a score <4); (ii) associated with an eQTL in *cis* and/or *trans* (ctQ); and (iii) associated with an meQTL in *cis* and/or *trans* (ctM). EAF, effect allele frequency of the pooled meta-analysis; *β*, effect size in the AFB and NEB analyses. All *P* values are from the sample-size-weighted fixed-effects meta-analysis. (Barban et al., 2016)

A gene-based approach using VEGAS (Liu et al., 2010) confirmed 7 loci from SNP-based GWAS for trait AFB and identified 3 new loci: *SLF2*, *ENO4* and *TRAF3*-AMN. For NEB trait was confirmed the signal on gene *GATAD2B* and one new on chromosome 17.

Causal variants

In order to identify potentially causal variants in both analyses performed, the functional annotation showed non-synonymous SNPs in high LD with our significant variants on chr1 in gene CREB-regulated transcription co-activator 2 (*CRTC2*) (rs11264680, r^2=0.98) and CREB protein 3 like 4 (*CREB3L4*) (rs11264743, r^2=0.94). This latter SNP is considered probably damaging and deleterious respect to PolyPhen and SIFT. *CRTC2* gene is a mediator of FSH (follicle-stimulating hormone) and TGF-β1-stimulated steroidogenesis in ovarian granulosa cell (Fang et al., 2012); *CREB3L4* is highly expressed in the ovaries, uterus, placenta, prostate, testis and has a role in male germ development (Adham et al., 2005). Our significant variant on chr3 is in LD with two nonsynonymous SNPs in *MST1R* gene (rs2230590, r^2=0.95 and rs1062633, r^2=0.95).

An analysis with RegulomeDB (Boyle et al., 2012) identified, among the significant variants, 50 SNPs that might have functional consequences influencing downstream gene expression (supplementary Table 6 of Barban et al., 2016). In particular stand out two sets of SNPs, one on chromosome 1 (18 SNPs) where the most promising variants, rs6680140, is located in a site of acetylation of histone H3 at lysine 27 near active regulation elements and another on chromosome 3 (25 SNPs) in a transcription factor binding site. Genes that bind one or more of 18 variants on chromosome 1 are *CREBBP*, *HNF4A*, *CDX2* and *ERG* and they may act upstream in the causal pathway and influence the expression of causal genes at this locus. On chromosome 3, 7 were eQTL for *HYAL3* and 10 for *RBM6* in monocytes.

Other 2 variants on chromosome 3, rs2247510 and rs1800688, are in H3K27ac sites and DNA I hypersensitivity cluster.

eQTL and meQTL analyses

For the 12 independent lead SNPs, local (cis, exons or methylation sites <1

Mb from SNP) and genome-wide (trans, exons or methylation sites >5 Mb from SNP) effects were identified between SNP and exon/methylation sites. Whole-blood BIOS eQTL and methylation quantitative trait locus database were used for this analysis (Zhernakova, Deelen, Vermaat, Iterson, & Van, 2015) (Bonder, Luijk, Zhernakova, & Moed, 2016). Seven SNPs were associated in *cis* with the expression of 54 genes and 5 are in high LD with the strongest eQTL for at least one gene in the corresponding locus: this means that the SNPs associated with NEB and AFB and the SNPs associated with expression tag the same functional site (supplementary Table 7 of Barban et al., 2016). 3 variants are associated with expression of 8 genes in *trans* (supplementary Table 8 of Barban et al., 2016). Also meQTL analysis that establish quantitative differences in DNA methylation between the two alleles show significant association: 11 of 12 lead SNPs are associated with DNA methylation of 131 genes in *cis* and 3 SNPs with 10 genes in *trans*.

Functional network and polygenic prediction

No enrichment in biological functions, neither genes and tissue sets were highlighted: this results reflect the need for a larger sample size but also because phenotypes are influenced by a mixture of biological, psychological and socio-environmental factors.

The polygenic scores for AFB and NEB were calculated using PRSice (Euesden, Lewis, & O'Reilly, 2015) and sets of significant SNPs in the respectively meta-analysis. Then, a regression model was used to predict the same phenotypes in 4 independent cohorts: HRS, LifeLines, STR and TwuinsUK. The mean predictive power of the polygenic scores is 0.9% for AFB and 0.2% for NEB. These results, also if the power of the polygenic score is low, showed that an increase of 1 s.d. in the NEB polygenic score is associated with 9% decrease in the probability of women remaining childless.

Signal associated with related traits and disease

Signal on chromosome 2 and on chromosome 3 were already associated with educational attainment (Okbay et al., 2016), the locus on chromosome 5 with age at menarche (J. R. B. Perry et al., 2014), locus on chromosome 6 with age at first intercourse (Day et al., 2016). The SNPs in chromosome 3 are in

LD with rs2013208 (r^2=0.81) already associated with HDL cholesterol (Global Lipids Genetics Consortium et al., 2013) and with rs7613875 (r^2=0.81) associated with BMI (Locke et al., 2015).

None of the associated variants are associated with age of menopause whereas rs9589, rs6803222 and rs9858889 on chromosome 3 are associated with age of menarche.

NEB showed significant and negative genetic correlation with years of education and age at first sexual intercourse (Figure 23 blue bar). AFB showed significant e positive genetic correlation with age at menarche and menopause, with voice breaking, birth weight, age at first intercourse and also with years of education. Smoking, triglyceride levels, risk of diabetes and BMI were associated with a lower genetic risk having an increase of AFB (Figure 27 red bar).

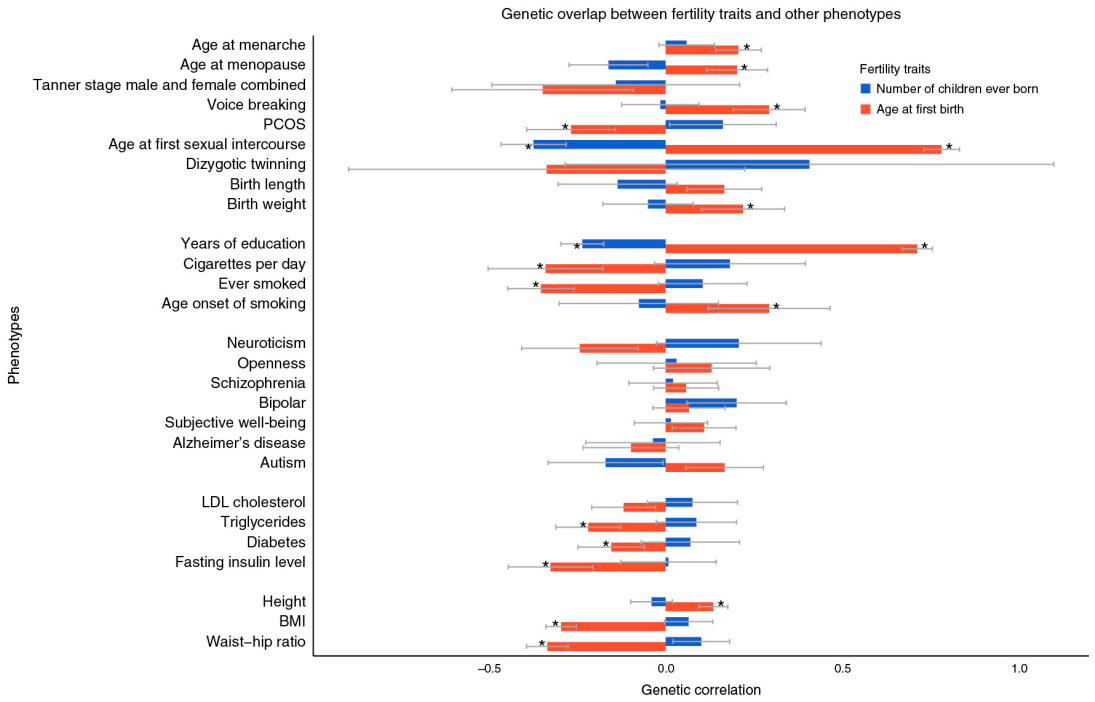


Figure 27. Genetic overlap between AFB or NEB and other related traits. Results from LD Score regressions show estimates of genetic correlation with developmental, reproductive, behavioral, neuropsychiatric and anthropometric phenotypes for which GWAS summary statistics were available in the public domain. The lengths of the bars correspond to estimates of genetic correlation. Gray error bars represent 95% confidence intervals. An asterisk indicates that the estimate of genetic correlation is statistically significant after controlling for multiple testing (P < 0.05/27 = 1.85E10^{-3}). (Barban et al., 2016)

In conclusion, the two most interesting loci in our analysis are located on chromosome 1 and 3: *CRTC2* and *CREB3L4* genes, that have non-

synonymous variants in high LD with the lead SNPs, are critical signal mediators in follicle stimulating hormone and transforming growth factor β1 stimulated steroidogenesis in ovarian the first whereas the second gene is high expressed in ovaries, uterus, placenta, prostate, testis and has a role in male germ cell development. Lead SNPs are associated with methylation status of these two genes and expression of CRTC2 in lymphocytes and whole blood. The association with other variants in locus on chromosome 1 may be mediated by alterations in cAMP responsive element binding men and women.

On chromosome 3, the lead SNPs rs2777888 is associated with the expression of RNF123 involved in cellular transition from quiescent to a proliferative state, it influences the expression of RNA-binding motif proteins (RBM5 and RBM6) that have role in cell cycle arrest, and it also affects the expression of LAMP2 on chromosome X. LAMP2 has a role in the acrosome reaction, allowing sperm to penetrate and fertilize ova, in fact encodes a lysosomal membrane protein (Tsukamoto et al., 2013).

This work has permitted to identify coding and regulatory variants that are potentially causal and even if AFB and NEB are also subject to cultural and social environment, the analysis underlines the genetic factor that partly drive reproductive behaviors.

5. WGS INGI data

As previously reported, whole genome sequencing was performed for a subset of 946 INGI individuals.

After all filtering steps (see materials and methods – WGS INGI data), 926 samples were retained. We identified approximately 27M sites (i.e. 24,557,366 SNVs and 2,061,725 INDELs).

Overall, 7.1 M sites (26%) were common (MAF>5%), 3.1M (12%) were low frequency (MAF between 1% and 5%) and 16.6M (62%) were rare (MAF <1%) with a partition similar in all cohorts. Singletons were >6M (24%) (6,193,486 SNPs and 273,679 INDELs) (Table 10).

INGI All samples				
	CARL	FVG	VBI	**INGI**
Samples	124	378	424	**926**
Average coverage	6.31	7.23	6.12	**6.55**
Sites	13,370,262	17,002,010	19,361,094	**26,619,091**
SNPs	12,208,629	15,521,313	17,830,208	**24,557,366**
INDELs	1,161,633	1,480,697	1,530,886	**2,061,725**
Sites MAF<=1%	3,627,622	7,283,720	9,416,028	**16,685,951**
Sites 1%<MAF<=5%	3,007,162	3,069,534	3,121,545	**3,125,971**
Sites MAF>5%	6,735,478	6,648,756	6,823,521	**7,123,064**
Singletons SNPs	2,061,824	2,784,746	3,554,744	**6,193,486**
Singletons INDELs	92,372	131,275	133,156	**273,679**
Average singleton per sample	17,285	7,671	8,646	**6,925**

Table 10. Number of variants shared or private in INGI population

For each individual we identified on average ~3.5M variant sites including ~0.56M indels and ~6.000 singletons.

A comparison between INGI cohorts showed that about 50% of the called variants were shared among at least two populations while another 50% were private (Table 11, Figure 28).

	Shared variants				**Private variants**		
	COMMON	**VBI-CARL**	**VBI-FVG**	**FVG-CARL**	**CARL only**	**VBI only**	**FVG only**
SNPs	9,045,896	721,758	1,834,514	354,720	2,086,255	6,228,040	4,286,183
INDELs	886,471	69,417	213,110	53,191	151,822	359,316	325,779
TOT	9,932,367	791,175	2,047,624	407,911	2,238,077	6,587,356	4,611,962

Table 11. Number of variants shared or private in INGI population

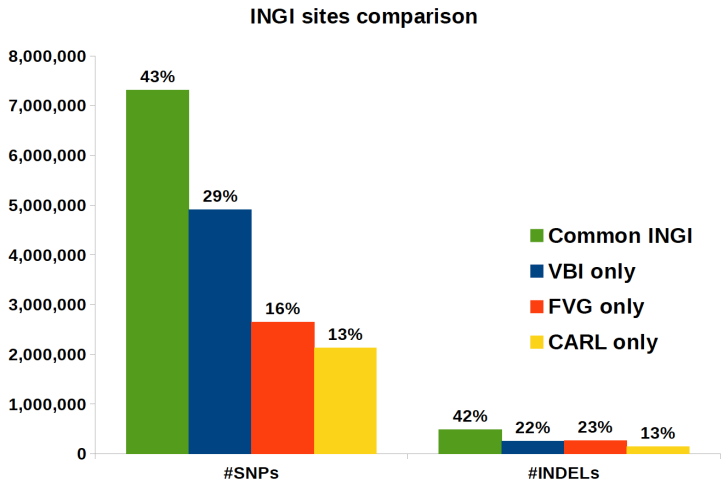


Figure 28. INGI sites comparison

Interestingly, most of the variants shared within INGI cohorts were found in the outbred references (~ 14 M in 1000 Genome project and ~15 M sites in UK10K) while most of the private ones remained private (>10 M with 1KG and >8M with UK10K). About 10% of the sites shared between two or all three populations, was not found in the EUR population or in UK10K dataset suggesting that they may be characteristics of the general Italian population. The majority of the private variants are within the range of low (MAF 1-5%) and rare frequencies (MAF < 1%).

After the QC step an Italian reference panel with WGS data was built: a `core' INGI+TSI panel was created merging data from the different INGI cohorts with those from the TSI cohort from the 1KGP3 (INGI+TSI), using the method implemented by the IMPUTE2 software.

We assessed the quality of INGI panel using two different parameters: the r^2 metric, which measures the correlation between the true genotype and the imputed genotype for the subset genotyped with commercial chips and the IMPUTE info score parameter, which provide a measure of the observed statistical information associated with the allele frequency estimate for each variant (Marchini & Howie, 2010).

The usage of “ad hoc” reference panels permits to identify specific and rare variants for each study.

Loss of function and human knockouts

In the final clean dataset of WGS INGI data, we focused the attention on Loss of Function (LoF) variants that might have different effects on genes: LoF may be maintained in a population because they have a mild effect or are in heterozygosity otherwise they could have a stronger effect on phenotype. Interestingly, some may be also found in homozygosity and can be defined as human Knock-out (hKO). Among the deleterious variants in the analyzed cohorts (See materials and methods - Human knockout), 506 presenting with a CADD score >20 (Kircher et al., 2014) were found in homozygous state in at least one individual, in one population. Stop gain variants have in average a higher CADD score (Figure 29).

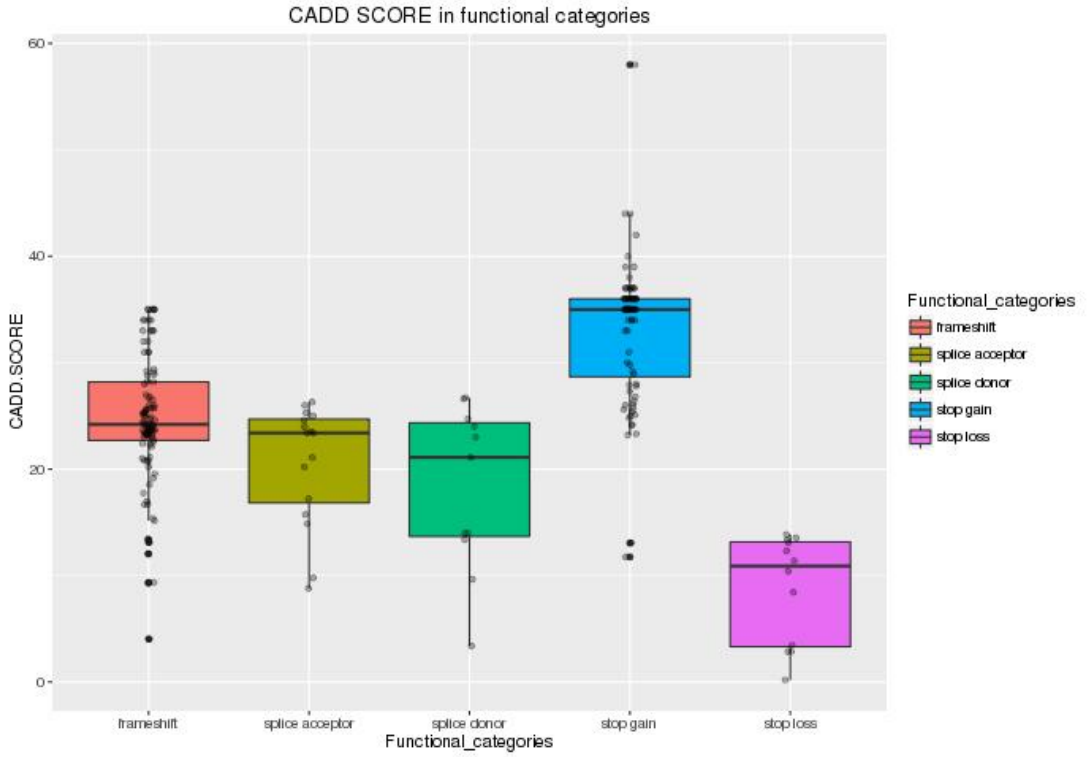


Figure 29. CADD score divided in functional categories

Transcript analysis revealed that only 205 affected all transcripts in 195 different genes**.** Among the 205 variants, the majority (150, 73%) were shared among all 3 populations (Figure 30). They span the entire spectrum of frequency but more than half (~60%) had frequency >=0.05**.**

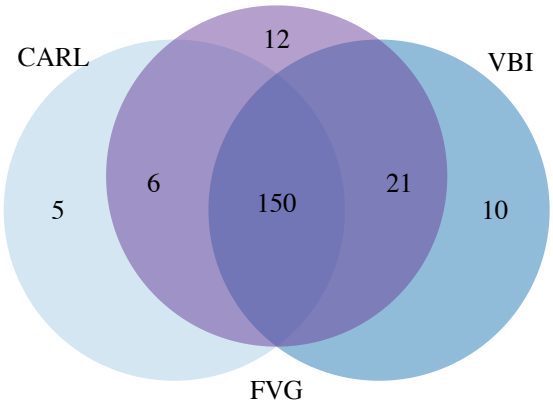


Figure 30. Number of hKO shared among INGI population.

Gene ontology analysis was useful to perform enrichment analysis on gene sets. The aim is also to identify if in our INGI population we have some KO gene involved in fertility.

The analysis revealed an excess of olfactory receptors hKO, several of the hKOs are located in genes involved in hair/skin/epithelium or eye phenotypes, and many members of gene families. In these categories KO may not cause a phenotypic effects because their possible function may be replaced by another gene of the same family and especially for OR could become a pseudogene (humans have approximately 400 functional genes coding for olfactory receptors, and other 600 candidates are pseudogenes).

Common human knockout could have been maintained because they may have beneficial effects, such as a stop gain (rs497116 Arg/X) in *CASP12* associated with resistance to infection and sepsis (Saleh et al., 2004).

Indeed, rare hKO could have a phenotypic effect: we identified a frameshift that cause KO of GnRH-II gene (Gonadotropin-Releasing Hormone 2) correlated with fertility.

GnRH2 knockout

In mammals gonadotrophin-releasing hormone I and the type I GnRH receptor drive the reproductive hormonal cascade in mammals: stimulate synthesis and secretion of luteinizing hormone (LH) and follicle stimulating

hormone (FSH). Mammals have a second hormone and receptors called GnRH-II system but in many species, one or both genes of GnRH-II are disrupted. Humans have only a gene encoding GnRH-II but no the functional receptors (Stewart, Katz, Millar, & Morgan, 2009). Human GnRH-II contains four exons, the first of which is non coding. From literature we known that different disruptive mutation appears within mammalian in GnRH-II genes (Table 12) and they cause gene knockout.

Species	DNA sequence feature		
	coding exon 1	coding exon 2	coding exon 3
Chimpanzee *(Pan troglodytes)*	codon 31 = stop codon (UGA) intact splice donor (GTGGGT)	intact splice acceptor (TCTAAG) codons intact intact splice donor (GTGAGT)	intact splice acceptor (CCGCAG) codons intact 1st AATAAA lies within MRPS26 gene
Orangutan *(Pongo pygmaeus)*	codon 30 = stop codon (UAG) intact splice donor (GTGGGT)	intact splice acceptor (CTGCAG) codons intact intact splice donor (GTGAGC)	uncharacterised
Mouse-lemur *(Microcebus murinus)*	codon 14 = frameshift (+CC) codon 31 = amide donor residue (Gly) deleted codon 33 = frameshift (+C) intact splice donor (GTGAGT)	intact splice acceptor (TCTGAA) codon 43 = frameshift (+5 bp) intact splice donor (GTAAGT)	intact splice acceptor (CTGCAG) codon 2 = frameshift (+C) codon 3 = frameshift (+GA) codon 9 = frameshift (+A)
Rabbit *(Oryctolagus cuniculus)*	start of coding region missing (45–60 bp) codon 18 = mutated di-basic cleavage motif (Arg to Thr) codon 19 = stop codon (UGA) intact splice donor (GTGGGT)	no splice acceptor codon 2 = frameshift (+A) intact splice donor (GTGAGT)	uncharacterised
Pika *(Ochotona princeps)*	codon 6 = frameshift (–G) codon 21 = frameshift (–AC) codon 23 = insert UUC (Phe) di-basic cleavage site mutated no splice donor	no splice acceptor codon 11 = stop codon (UAA) codon 19 = frameshift (+CG) codon 23 = frameshift (+CC) codon 24 = frameshift (+4 bp) codon 30 = frameshift (+GG)	uncharacterised
Cat *(Felis catus)*	His to Arg at GnRH-II position 5 Gly to Arg at GnRH-II position 10 intact splice donor (GTAGGT)	intact splice acceptor (TCTGAG) codon 27 = frameshift intact splice donor (GTAAGT)	intact splice acceptor (GCGCAG) codons intact
Dog *(Canis familiaris)*	Start codon mutated (AUG to UGA) codon 37 = mutated di-basic cleavage motif (Arg to Gly) intact splice donor (GTGGGT)	uncharacterised	uncharacterised

Table 12. Disruptive mutations within mammalian GnRH2 genes (Stewart et al., 2009).

We identified a frameshift (rs16996832: insertion of GCCC - E/EPX) in GnRH-II in homozygous state in 38 INGI samples (respectively 22 in VB, 14 in FVG and 2 samples in CARL) with a frequency of 0.18 from INGI cohort and 0.10 in European population of 1000 Genome. Individuals with this mutation in homozygosity have GnRH-II KO rendering their product non-functional but also do not show any particular phenotypes since GnRH-I system is driving the reproductive hormonal cascade. This example shows no effect cause from KO due to the redundancy of function from another gene of the same family.

6. Quantitative AMH GWAS with WGS INGI samples

A GWAS on quantitative trait Anti Müllerian hormone (AMH), a marker of ovarian reserve, was performed on INGI women. 475 women with less than 41 years were included in analysis (244 from VB cohort, 152 from FVG and 79 from CARL).

We performed the analysis with data imputed with 2 different reference panels: the first analysis with 1000G phase1 (1000GP1) panel and the second with Italian reference panel (see WGS INGI data).

Due to the small number of women that respect all the criteria of inclusion (see material and methods), analysis has only suggestive signal that no reach the threshold of significance.

Interesting, in the comparison of both analysis, is evident the enrichment of variants with lowest p-value in signals with data imputed with Italian reference panel (Figure 31).

The signal on **chromosome 5** is in an intergenic region and is enriched in rare variants as reported in table 13. In this locus rare variants are better imputed with Italian reference panel in comparison to 1000GP1 imputation and they reach a suggestive p-value.

RS id	chr	pos	A1	A2	Effect	P-value	Dir	freq vb	freq fvg	freq carl	info Italian	info 1000G
rs12655444	5	81853493	C	T	1.38	4.38E-06	+++	0.01	0.02	0.01	0.923	0.879
rs74354356	5	81855591	A	AG	-1.35	2.46E-06	---	0.01	0.02	0.01	0.943	-
rs78788723	5	81856711	A	C	-1.38	1.30E-06	---	0.01	0.02	0.01	0.927	0.912
rs79107879	5	81856874	T	C	-1.40	2.16E-06	---	0.01	0.02	0.01	0.929	0.945
rs75757001	5	81861862	G	A	1.37	3.07E-06	+++	0.01	0.02	0.01	0.935	0.851
rs114906912	5	81864971	C	T	1.36	3.75E-06	+++	0.01	0.02	0.01	0.924	0.836
rs10514241	5	81868187	T	G	-1.36	3.44E-06	---	0.01	0.02	0.01	0.932	0.851
rs75983158	5	81869241	A	C	-1.35	3.63E-06	---	0.01	0.02	0.01	0.920	0.843
rs12655100	5	81870579	A	G	-1.38	2.75E-06	---	0.01	0.02	0.01	0.926	0.822
rs16899734	5	81872254	C	G	-1.43	1.61E-06	---	0.01	0.02	0.01	0.915	0.811
rs78460703	5	81873459	C	G	-1.42	1.65E-06	---	0.01	0.02	0.01	0.915	0.809
rs57253569	5	81876751	T	A	1.44	1.46E-06	+++	0.01	0.02	0.01	0.917	0.783
rs78135984	5	81881180	C	T	1.49	7.14E-07	+++	0.01	0.02	0.01	0.867	0.753
rs74291681	5	81881336	T	C	-1.47	1.10E-06	---	0.01	0.02	0.01	0.903	0.745
rs2385889	5	81882497	A	G	-1.51	5.92E-07	---	0.01	0.02	0.01	0.884	0.721

Table 13. Suggestive signals in AMH GWAS with data imputed with Italian reference panel, chr 5.

On **chromosome 12** we have a very similar scenario: the signal is enriched in variants better imputed with Italian reference panel (Table 14), 15 variants in this locus reach the suggestive limit.

RS id	chr	pos	A1	A2	Effect	P-value	Dir	freq vb	freq fvg	freq carl	info Italian	info 1000G
rs61930911	12	86288138	A	G	-0.37	2.72E-06	---	0.22	0.22	0.22	0.993	0.990
rs7301600	12	86289984	G	C	0.41	2.05E-06	+++	0.16	0.16	0.19	0.994	0.987
rs73177142	12	86290735	C	T	0.37	2.43E-06	+++	0.21	0.22	0.22	0.992	0.988
rs61930913	12	86291121	G	A	0.37	2.07E-06	+++	0.22	0.22	0.22	0.994	0.987
rs11117088	12	86291542	G	T	0.41	1.88E-06	+++	0.16	0.16	0.19	0.994	0.987
rs4143239	12	86291728	T	A	0.37	2.07E-06	+++	0.22	0.22	0.22	0.994	0.987
rs10863094	12	86292351	C	T	0.37	2.07E-06	+++	0.22	0.22	0.22	0.994	0.987
rs79932392	12	86292574	GCTA	G	-0.37	2.07E-06	---	0.22	0.22	0.22	0.994	-
rs2897186	12	86293716	G	A	0.37	2.06E-06	+++	0.22	0.22	0.22	0.994	0.987
rs11117089	12	86294163	T	C	-0.37	2.06E-06	---	0.22	0.22	0.22	0.994	0.987
rs58780912	12	86296085	G	A	0.42	1.78E-06	+++	0.16	0.16	0.19	0.994	0.986
rs11609918	12	86297970	A	G	-0.42	1.44E-06	---	0.16	0.16	0.19	0.988	0.987
rs11117093	12	86302072	A	T	-0.41	2.50E-06	---	0.16	0.15	0.19	0.995	0.987
rs11117097	12	86303286	A	G	-0.41	2.48E-06	---	0.16	0.15	0.19	0.996	0.976
rs12313362	12	86304321	A	G	-0.41	2.46E-06	---	0.16	0.15	0.19	0.996	0.987

Table 14. Suggestive signal in AMH GWAS with data imputed with Italian reference panel on chr 12.

Signal on chromosome 19 is on gene ***PPP5C*** and for each variant with a suggestive p-value the functional consequences is reported in table 15. *PPP5C* gene encodes a serine/threonine phosphatase, which is a member of the protein phosphatase catalytic subunit family. In literature is reported that the product of this gene participate in signaling pathways in response to hormones or cellular stress, and elevated levels of this protein may be associated with breast cancer development [provided by RefSeq, Feb 2011].

The last suggestive signal is on chromosome 22, on gene ***POLDIP3***; 8 variants (5 are intron variants, 1 is an upstream gene variant and 1 is a missense variant) have suggestive p-value in Italian reference panel and these variants are better imputed in comparison with 1000GP1 imputation (Table 15). *POLDIP3* encodes an RRM (RNA recognition motif)-containing protein that participates in the regulation of translation by recruiting ribosomal protein S6 kinase beta-1 to mRNAs. RPKM (Reads Per Kilobase Million) analysis, quantifying gene expression from RNA sequencing data, shows a ubiquitous expression of *POLDIP3* in ovary (RPKM 49.2) (Figure 32).

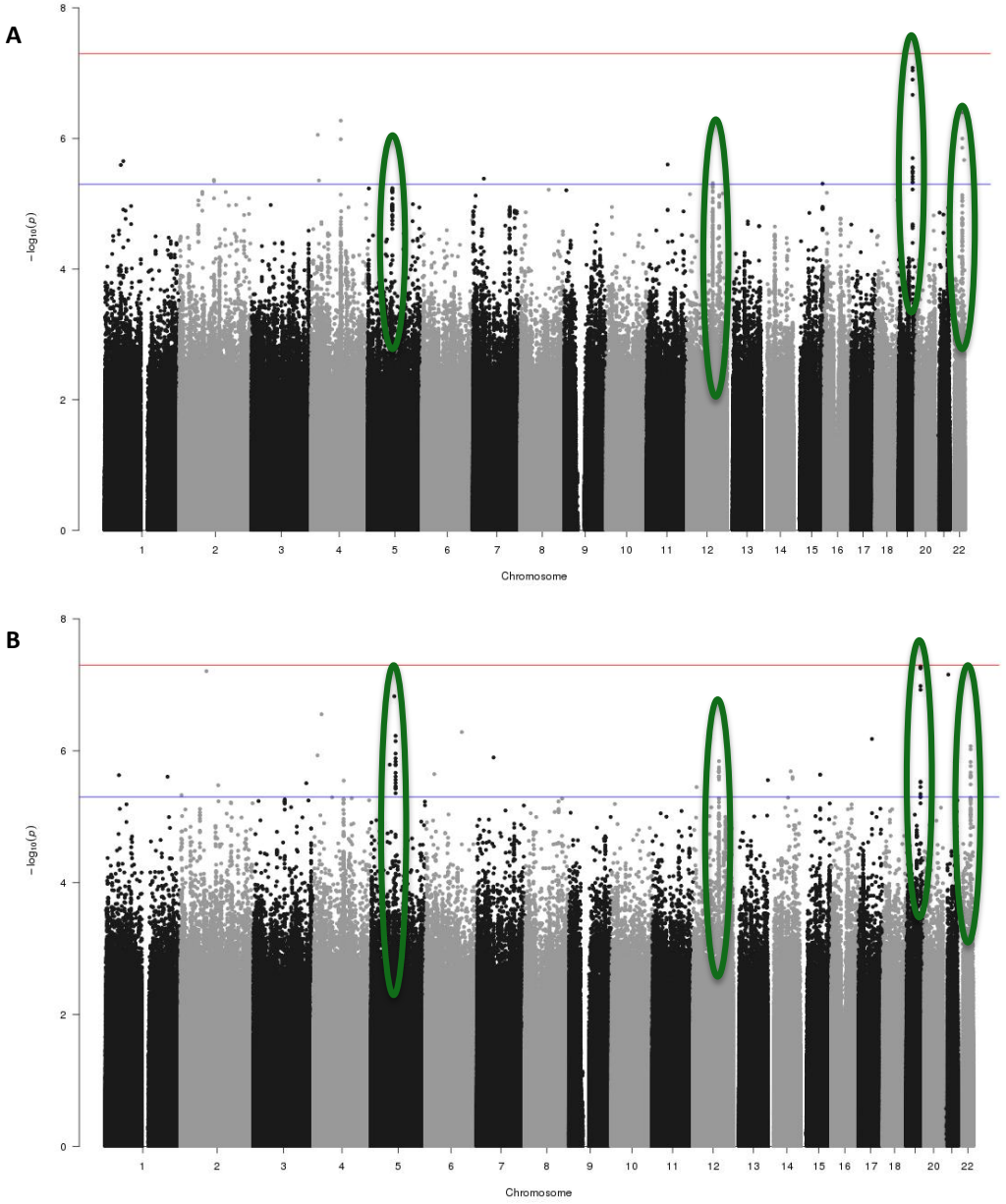


Figure 31. SNPs are plotted on the *x* axis according to their position on each chromosome against association with AMH (A) with data imputed to 1000G phase I and (B) with data imputed with Italian reference panel. The solid blue line indicates the threshold for genome-wide significance ($P < 5 \times 10^{-8}$), and the red line represents the threshold for suggestive hits ($P < 5 \times 10^{-6}$). Signal encircled in green improve in data imputed with Italian reference panel.

RS id	chr	pos	A1	A2	Effect	P-value	Dir	freq vb	freq fvg	freq carl	info Italian	info 1000G	cons
rs16980461	19	46856133	G	A	0.56	3.53E-06	+++	0.08	0.07	0.08	0.991	0.973	intron
rs917948	19	46857015	C	T	0.56	4.50E-06	+++	0.08	0.07	0.08	0.987	0.968	synonymous
rs917946	19	46860058	A	G	-0.57	3.00E-06	---	0.08	0.07	0.08	0.991	0.975	intron
rs55842357	19	46863135	C	T	0.57	2.93E-06	+++	0.08	0.07	0.08	0.991	0.957	intron
rs2239380	19	46870302	T	C	-0.61	5.61E-08	---	0.09	0.08	0.09	0.991	0.975	intron
rs56203900	19	46888566	C	T	0.59	1.19E-07	+++	0.09	0.08	0.09	0.990	0.980	intron
rs73043492	19	46888894	G	A	0.55	4.93E-06	+++	0.08	0.07	0.08	0.994	0.979	intron
rs58701793	19	46890247	G	C	0.60	1.04E-07	+++	0.09	0.08	0.09	0.995	0.989	intron
rs8811	19	46893810	G	A	0.61	5.48E-08	+++	0.09	0.08	0.09	0.997	0.996	3_prime_UTR
rs741231	19	46894163	C	A	0.61	5.25E-08	+++	0.09	0.08	0.09	0.997	-	downstream

RS id	chr	pos	A1	A2	Effect	P-value	Dir	freq vb	freq fvg	freq carl	info Italian	info 1000G	cons
rs66748203	22	42999394	D	A	-0.43	3.23E-06	---	0.12	0.14	0.14	0.995	-	intron
rs137108	22	43002362	A	G	-0.46	9.48E-07	---	0.12	0.14	0.13	0.982	0.986	intron
rs146980368	22	43003683	G	A	0.45	1.47E-06	+++	0.12	0.14	0.13	0.985	0.970	intron
rs148867399	22	43003771	A	G	-0.44	2.50E-06	---	0.12	0.14	0.14	0.996	0.974	intron
rs137110	22	43007597	A	C	-0.44	2.45E-06	---	0.12	0.14	0.14	0.996	0.986	intron
rs137113	22	43009817	G	C	0.44	2.32E-06	+++	0.12	0.14	0.14	0.997	0.986	intron
rs28627172	22	43010817	G	A	0.45	8.51E-07	+++	0.13	0.14	0.14	0.999	0.974	missense
rs137114	22	43011246	C	T	0.44	2.13E-06	+++	0.12	0.14	0.14	0.997	0.979	upstream

Table 15. Suggestive signals in AMH GWAS with data imputed with Italian reference panel. Signal on chr 19 and on chr 22

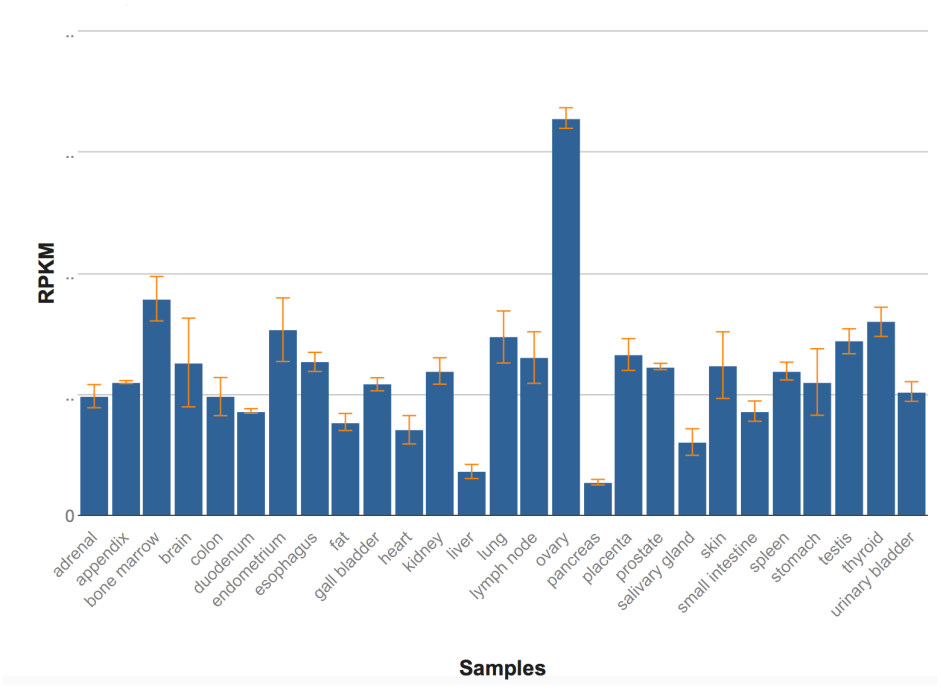


Figure 32. Analysis RPKM (Reads Per Kilobase per Million mapped reads), a method of quantifying gene expression from RNA sequencing data by normalizing for total read length and the number of sequencing reads, for gene *POLDIP3*.

An analysis with Variant Effect Predictor (VEP) (McLaren et al., 2016) for variants with suggestive p-value in AMH GWAS imputed with Italian reference panel showed that the majority are intron variants, no one have high effect, there is only 1 missense with moderate effect in *POLDIP3* (Figure 33).

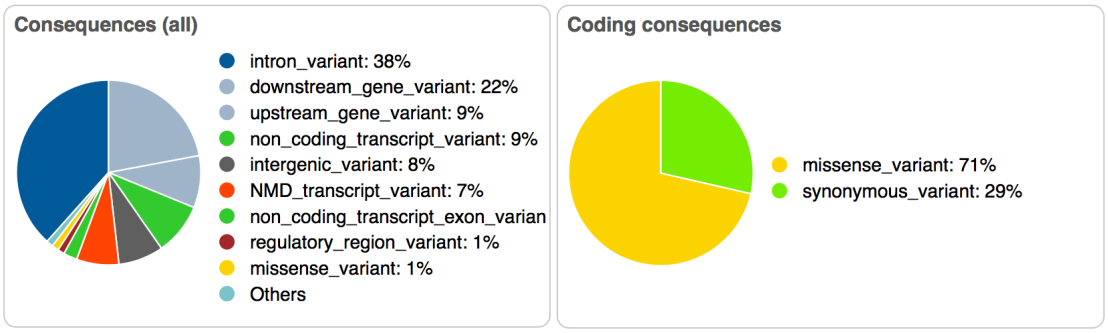


Figure 33. Functional consequences of suggestive variants of AMH GWAS according to VEP.

PANTHER pathway analysis showed a significant result emphasize in yellow in figure 34 and table 16. DNA replication pathway has a fold enrichment of 46.97 with p-value of $2.18E^{-2}$. Panther "DNA replication" pathway includes a series of integrated protein-protein and protein-DNA interactions and enzymatic reactions to ensure high accuracy of DNA replication.

	Homo sapiens (REF)	upload_1				
PANTHER Pathways	**n° genes**	**n° genes**	**expected**	**Fold Enrichment**	**+/-**	**raw P value**
DNA replication	28	1	0.02	46.97	+	2.18E-02
EGF receptor signaling pathway	135	1	0.1	9.74	+	9.85E-02
Cadherin signaling pathway	148	1	0.11	8.89	+	1.07E-01
Gonadotropin-releasing hormone receptor pathway	227	1	0.17	5.79	+	1.60E-01

Table 16. PANTHER Overrepresentation Test (Released 20171205). Annotation Version and Release Date: PANTHER version 13.0 Released 2017-11-12. Reference List: Homo sapiens all genes in database. Analyzed List: upload_1 (suggestive gene in AMH GWAS with Italian reference panel)

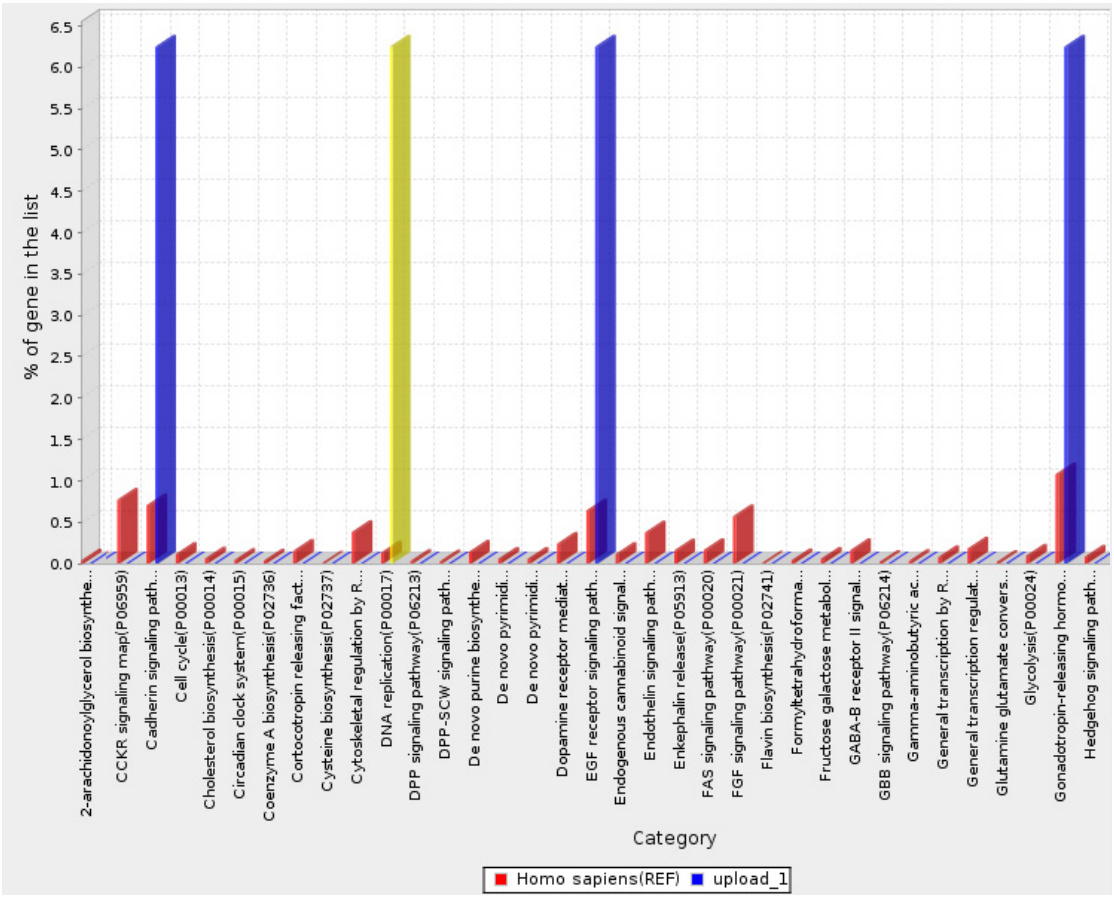


Figure 34. PANTHER Overrepresentation Test (Released 20171205). Annotation Version and Release Date: PANTHER version 13.0 Released 2017-11-12. Reference List: Homo sapiens all genes in database. Analyzed List: upload_1 (suggestive gene in AMH GWAS with Italian reference panel). Only a part of results are showed.

In conclusion, even if the sample size is very small in our AMH GWAS, we found “DNA replication” pathway enriched in our suggestive signal and a signal on gene *POLDIP3*, involved in regulation of translation, overexpressed in the ovary.

We demonstrate the improvement of imputation quality with Italian reference panel in comparison with 1000G phase I panel, especially for rare variants.

7. **Fertility preservation in endometriosis patients: is AMH a reliable marker of the ovarian follicle density?** (Garavaglia et al., 2017)

In our study we have evaluated pre-surgical serum AMH levels in a cohort of endometriosis patients undergoing ovarian cortex resections for fertility preservation at San Raffaele Hospital in Milan in the years 2011–2014, compared to a large control group, and then we have tested the hypothesis that the preoperative AMH values may correlate with the individual follicular density.

A statistical analysis on AMH values of 201 untreated endometriosis patients and 387 normal women shows that the median of AMH residual levels is significantly lower in cases compared to controls ($P=1.0x10^{-3}$) and the significance increase when the threshold of >36 years was applied (in women with >36 years there was a depletion of ovarian reserve) with a p-value of $2.7x10^{-4}$. In summary, AMH decreased as expected in all women, but the decrease was faster in endometriosis cases (Fig. 35)

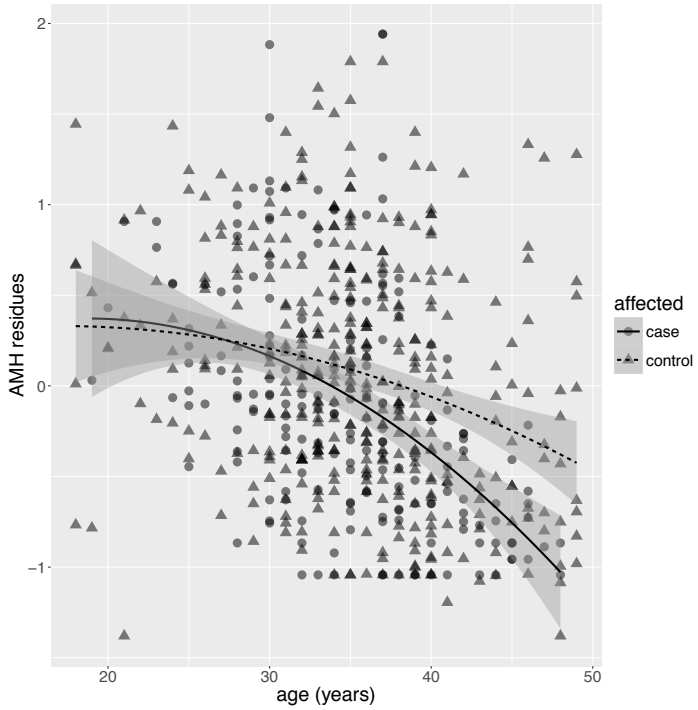


Figure 35. Multiple regression analysis of AMH and age. Scatter plot of the correlation between AMH residues and age in cases (solid circles) and controls (solid triangles). Linear interpolation for cases (solid lines) and controls (dashed lines) together with confidence intervals (gray shadowed area) are reported. (Garavaglia et al., 2017)

Ovarian biopsies were collected in a subset of 24 endometriosis cases and 33 controls. In this samples the number of primordial follicles decreased significantly with age as expected, but this decrease did not show any significant differences between case and controls in this study, even if we consider only older cases (>36 years).

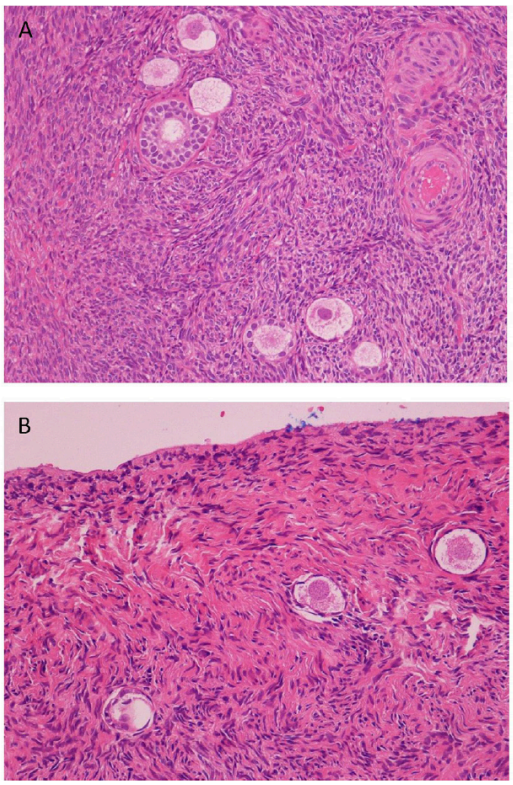


Figure 36: Histological analysis of cortical ovarian strips. (a) Hematoxylin/ eosin staining shows primordial, early, and mature primary follicles of one representative ovarian cortical strip of a control (125×). (B) Hematoxylin/eosin staining shows a few primordial follicles (severe ipotrophia for age of the patient) in a fibrosclerotic tonaca albuginea of one representative cortical strip of an endometriosis patient (200×). (Garavaglia et al., 2017)

We confirmed a statistically significant correlation between AMH serum levels and the number of primordial follicles with a regression analysis adjusted for age, kit and laboratory test kit. The estimated effect (beta value) of the regression curve was higher in endometriosis patients, suggesting a faster depletion of primordial follicles in cases.

We underlines the importance to measure AMH levels in women with endometriosis in order to consider a cryopreservation before the accelerated follicular depletion during disease progression.

MATERIALS AND METHODS

INGI genotyping and imputation

Val Borbera samples

1664 samples were genotyped with Illumina 370K Quad v3 chip and additional 121 samples with Illumina 700K Omni Express array.

Imputation was performed with software IMPUTE2 (Howie, Marchini, & Stephens, 2011) after SNP filtering; SNPs with minor allele frequency < 1%, Hardy-Weinberg p-value < 10^{-6}, call rate < 0.9 were removed. The panel for imputation used was the 1000G Phase I Integrated Release Version 3 [MARCH 2012].

1786 VB samples were genotyped with Human Exome 12v1-2_A chip. After variants calling with software Zcall (Goldstein et al., 2012), different filters were applied for QC (HWE>=10^{-6} (applied only to common SNPs with MAF>5%), CALLRATE >= 95 and MIND >= 95%) and a subset of 1779 samples with 243,961 markers genotyped was obtained.

Friuli Venezia Giulia samples

1696 samples were genotyped: 1330 with Ilumina 370K Quad v3 chip and additional 366 samples with Illumina 700K Omni Express array. 1913 samples were also genotyped with HumanExome and with Illumina Omni Express Exome array. For the analysis people with age < 18 were excluded. Different filter of quality were applied: sample call rate 95%, HWE>=10^{-6}, SNP call rate 99%.

Carlantino samples

630 samples were genotyped with Illumina Infinium Duo and Infinium Quad chip and 820 also with exome chip Illumina Omni Express Exome array.

People with age <18 were excluded for the analysis and quality filter were applied before analysis: sample call rate 95%, HWE>=10^{-6}, SNP call rate 99%.

1. Rare coding variants and X-linked loci associated with age at menarche. (Lunetta et al., 2015)

GWAS age at menarche on exome array

A linear regression model on age at menarche between age 9 and 17, adjusted for birth year and principal component was used in each study performed, using the skatMeta/seqMeta package in R. Variants with MAF>5% in the meta-analysis were excluded. SKAT test assume that rare variant effects are random and can contain risk, protective or null rare alleles.

A burden test was performed using seqMeta, a gene-based test, for low-frequency variants, which assume that the entire set of rare variants have the same effect direction. Only a first subset of variants with MAF<1% and separately other subsets for all non-synonymous annotated as 'damaging' and for loss of function, were used in both tests. Study significance threshold was calculated including 2 tests x 3 filters x number of genes ($P<1.14 \times 10^{-6}$).

Meta-analysis was performed with software METAL (significance threshold $P<5 \times 10^{-8}$). Follow-up was run on 2 different groups of women: 76,831 from 23andME with exome array available and 39,486 from deCODE study with imputation data available. Variants of both studies had passed genotyping QC or imputation quality score > 0.4.

GWAS age at menarche on X chromosome

76,831 women from 23andME study of European ancestry provided informed consent to participate in the analysis. SNPs with HWE < 10^{-20}, call rare < 95% or with large difference in frequency compared to 1000G were removed. Genotype data were imputed against the 1000G reference panel march 2012 'v3' release. A linear regression model assuming additive allelic effects was performed and women from deCODE were used as replication.

2. Genomic analyses identify hundreds of variants associated with age at menarche and support a role for puberty timing in cancer risk. (Day et al., 2017)

GWAS age at menarche imputed to 1000G

Each study performed an additive linear regression model for association with AAM, using as covariates age at visit and other specific for each population for example PCA. Data were imputed to 1000 Genome Project reference panel and a standardized quality control pipeline was applied at study-level. Software METAL was used for meta-analysis: first ReproGen consortium studies were combined and only SNPs present in over half of studies were taken forward and secondly ReproGen data were combined with UK Biobank and 23andMe studies. Variants with combined MAF>0.1% were analyzed and using a distance-based metric (any SNPs passing the significance threshold within 1 MB of another significant SNP were considered to be located in the same locus) a list of index variants was defined.

39,486 women from deCODE study was used as replica.

Parent-of-origin-specific associations and variance

Women from deCODE study with parental origins of alleles determined by a combination of genealogy and long-range phasing was used to study parent-of-origin-specific allelic associations.

The variance explained by each associated variant was calculated with the following formulas where *f is MAF,* f_h is homozygous frequency of the variant and β is the effect of specific model:

additive model $2 f (1-f) \beta_a^2$

recessive model $f_h (1-f_h) \beta_r^2$

maternal model $f (1-f) \beta_m^2$

paternal model $f (1-f) \beta_p^2$

Variance explained across multiple SNPs was calculated summing the variance of the single variants.

Mendelian randomization analyses

Mendelian randomization analyses was performed in order to understand if there is a causal effect of puberty timing on the risks of cancer sex-steroid-

sensitive. In Mendelian randomization, the basic principle utilized is that genetic variants that alter the level or mirror the biological effects of a modifiable environmental exposure that itself alters disease risk should be related to disease risk to the extent predicted by their influence on exposure to the environmental risk factor (George Davey Smith and Shah Ebrahim). Genotyped data of individuals with cancer were available thanks to Breast Cancer Association Consortium (BCAC), Endometrial Cancer Association Consortium (ECAC) and Ovarian Cancer Association Consortium (OCAC). Genetic variants predicted AAM was tested in a logistic regression model for association with each cancer.

A BMI-adjusted analysis was performed including AAM genetic risk score as a covariate.

Pathway analyses

A gene set enrichment analysis (GSEA) implemented in MAGENTA (Meta-Analysis Gene-set Enrichment of variaNT Associations) software was performed (Ayellet et al., 2010). The aim is to test for enrichment of genetic associations in predefined biological processes or sets of functionally related genes. For each gene, a single index SNP with the lowest p-value, within 110kb upstream and 4kb downstream of gene mapped in the genome, was considered and analysis give a gene score corrected for confounding factors such gene size and SNP density. The number of scores in a given pathway were ranked and then compared to 1,000,000 randomly permuted pathway of the same size. An empirical GSEA p-value was created for each pathway.

Gene expression data integration

A LD score regression to specifically expressed gene (LDSC-SEG) was performed to identify tissue and cell types most relevant to gene associated with AAM. This approach uses stratified LD score regression to test whether disease heritability is enriched in regions surrounding genes with the highest specific expression in a given tissue (Finucane et al., 2017). Genes were ranked by a *t*-statistic for differential expression for each tissue and the contribution of the annotation of the top 10% of genes by this ranking were estimated.

To identify specific eQTL-linked gene both summary Mendelian randomization (SMR) and MetaXcan was performed. The first approach was run against whole-blood eQTL available data set by Westra to map potentially functional genes to trait-associated SNPs. The second approach is a meta-analysis extension of the PrediXcan that incorporates information from gene expression and GWAS data to translate evidence of association with a phenotype from SNP level to the gene. Gene expression information was taken from the GTEx project: DNA and RNA from multiple tissues were sequenced from almost 1,000 European, African and Asian dead individuals. MetaXcan in this analysis were targeted to tissue with previous evidence of association with AAM.

3. Large-scale genomic analyses link reproductive aging to hypothalamic signaling, breast cancer susceptibility and BRCA1-mediated DNA repair. (Day et al., 2015)

GWAS age at menopause on all genome

A total sample size of 69,360 individuals of European descent (33 studies) was collected for GWAS on age at natural menopause. Data were imputed to HapMap2 panel and for the cohort with data imputed to 1000 Genome project were take into consideration only SNPs included in HapMap2. GWAS was performed for each cohort separately and an additive model including top principal components and study-specific covariates was run. Quality controls were applied: SNPs with MAF < 1% and imputation quality <0.4 were removed. Meta-analysis was performed with METAL using an inverse variance-weighted. Only SNPs that had data from more than 50% of the studies were considered for results and the cut off of significant pvalue was applied according to Bonferroni correction ($P < 5x10^{-8}$).

GWAS age at menopause on exome chip data

Twenty-two studies of European ancestry performed R package skatMeta or seqMeta analysis with exome genotyping data available on autosomal and on X-chromosome. A single-variant meta-analysis was performed in METAL with a total sample size of 39,026; the cut off of significance was set to $P< 5x10^{-8}$

according to Bonferroni correction. Replication was performed in the deCODE study (n=10,157).

Conditional analysis

GCTA software package was used to identify independent signals. LD between variants was estimated in genotype of 3 different studies: Rotterdam study I (n=5,974) and 2 EPIC-InterAct data set (n=7,397 and 9,294). We assumed zero correlation between SNPs more than 10Mb apart.

The independence of the exome array and HapMap2 signals was confirmed with a conditional analysis in the Women's Genome Health study (n=11,664). Regression analysis including all significant index SNPs in additive model was performed.

Causative gene identification

The likely causative genes were annotated selecting genes identified by GRAIL (Raychaudhuri et al., 2009) or STRING programs (Szklarczyk et al., 2015), genes in which the top SNP was a coding variants or in an eQTL. GRAIL program suggests the most likely causal gene at each locus and a $P< 0.05$ was taken to indicate a suggested causal gene.

Pathway identification

Several softwares were used to test for signal enrichment: Panther, Ingenuity, GO, KEGG, Reactome and Biocarta using MAGENTA (Ayellet et al., 2010). We also tested 4 custom pathways for gene involving in age at menarche (n=154), ovarian function (n=130), POI (n=31) and monogenic disorders of puberty (n=21).

Estimating variance

Restricted maximum likelihood (REML) implemented in GCTA was used to estimate the total variance explained by ANM-associated SNPs. Variance was calculated at varying significance thresholds ($5x10^{-7}$, $5x10^{-6}$, $5x10^{-5}$, $5x10^{-4}$, 0.005, 0.05 and all SNP passing the quality control).

4. Genome-wide analysis identifies 12 loci influencing human reproductive behavior. (Barban et al., 2016)

GWAS reproductive behavior

Sixty-two cohorts participated to the analysis of two traits regarding human reproductive behavior: age at first birth (AFB) and number of children ever born (NEB). GWAS was performed on data imputed to HapMap 2 CEU. In the analysis only people with European ancestry and with all covariates data were included: sex, age of birth and cohort-specific covariates. An additive linear model with the entire set of covariates was performed for each cohort and a QC protocol, similar of the GIANT consortium's QC used in the recent study of human height (Wood et al., 2014), was utilized to verified the quality of analysis. Sample-size weighted meta-analysis was performed with software METAL in which only variants in at least 50% of the participants for a given phenotypes were included. Sample size of meta-analysis reaches 251,151 individuals for AFB pooled and 343,072 for NEB pooled. Meta-analyses were performed also divided by sex: 189,656 women and 48,408 men for AFB and 225,230 women and 103,909 men for NEB.

The lead SNPs for each significant locus was identified with the PLINK clumping function (Chang et al., 2014).

A gene-based analysis was performed with VEGAS (Mishra & Macgregor, 2015): a 50kb extra window surrounding the genes were considered in the analysis.

Functional variant analysis using RegulomeDB

RegulomeDB (Boyle et al., 2012) integrates results from ENCODE project (E. N. C. O. D. E. P. Consortium et al., 2012) and Roadmap Epigenomics project (R. E. Consortium et al., 2015). 322 significantly associated SNPs ($P< 5x10^{-8}$) with AFB and/or NEB in meta-analysis were tested and those with RegulomeDB score < 4 were considered been functional and analyzed in details regards eQTL and their protein-binding capacity.

eQTL and meQTL analyses.

Spearman's rank correlations between SNPs and local or global exons/methylation sites were computed for lead SNPs in GWAS to identified the possible local (cis, exons/methylation sites < 1 MB from the SNP) and genome-wide (trans, exons/methylation sites > 5 MB from the SNP) effects. LD between strongest eQTLs identified for these exons/methylation sites and the corresponding SNP identified in the GWAS were computed using BIOS genotypes. The BIOS consortium used samples from five Dutch cohorts: The Leiden Longevity Study (Deelen et al., 2014), The Rotterdam Study (Ho fman et al., 2009), The LifeLines-DEEP cohort (Tigchelaar et al., 2015), The Cohort on Diabetes and Atherosclerosis Maastricht (CODAM) (Van Dam, Boer, Feskens, & Seidell, 2001) and The Netherlands Twin Register (Willemsen et al., 2013). Bonferroni multiple testing correction was performed for the 12 SNPs tested.

Gene prioritization

All genes located within 40 kb of the lead SNPs were used as input for four bioinformatics tools in order to identify potentially causal genes: ToppGene (Chen et al., 2009), Endeavor (Tranchevent et al., 2008), MetaRanker (Pers et al., 2013) and DECIPT (Pers et al., 2015). For the first three softwares a list of genes with a role in fertility was used as training (*BRCA1, EGFR, ERBB2, ERBB3, ERBB4, HSD17B1, RBM5, ESR1, ESR2 and FSHB*). MetaRanker use SNPs that reached $P<5x10^{-4}$ and their chromosomal position as input. We performed a gene prioritization procedure using also software DECIPT giving as input all meta-analysis SNPs that reached $P<5x10^{-4}$.

Functional network enrichment

To identified gene set, tissue and cell type enrichment among the variants with p-value less than 5×10^{-4} in meta analysis, DECIPT software was used (Pers et al., 2015). DEPICT was only able to perform analyses for AFB and NEB pooled, and AFB women due to the number of samples. 5 prioritized candidate gene sets were combined for the analysis: closest genes to the lead SNPs, closest genes to the non-synonymous SNPs in high LD ($r^2>0.50$) with the corresponding lead SNP, closest genes to other types of SNPs in

very high LD (r^2>0.80) with the corresponding lead SNP, and expression probe gene names of cis, and trans eQTLs.

Polygenic score prediction

Polygenic scores were calculated for AFB and NEB, based on GWA meta-analysis results and using regression models to predict the same phenotypes in four independent cohorts: HRS, Lifelines, STR and TwinsUK. Polygenic scores for NEB were tested to predict childlessness at the end of the reproductive period (using age 45 for women and 55 for men). Finally, we also tested the predictive value of our polygenic scores for AFB for age at menarche (using TwinsUK) and age at menopause (using Lifelines).

5. WGS INGI DATA

Genotype calls for autosomal chromosomes were separately produced for each population, data were annotated using bcftools (v.1.2) to add information about Ancestral Allele and allele frequencies from 1000GP phase 3 (1KGP3) populations and rsIDs from dbSNP v.141. Furthermore functional annotations, Polyphen and Sift information's and CADD score were added.

Very stringent filters were applied for quality check both by samples and by site. Individuals with excess of singletons (> than 60000 singletons: one sample from the FVG cohort) and excess of heterozygosity (> than 3 SD of the mean: one from the CAR cohort, one from the FVG cohort and 4 samples from the VBI cohort) were removed from the analysis. Individuals with a Non-Reference Discordance Rate (NRDR) greater than 5% were also removed (i.e. 8 samples from CAR, 1 from FVG and 5 from VBI cohort).

Sites with Hardy-Weinberg equilibrium exact test p-value below the threshold of $1x10^{-8}$ and heterozygosity rate greater than 3 standard deviations of the mean were removed. For variants overlapping SNP array data available for each cohort, the Concordance and the Non-Reference Discordance Rate (NRDR) were calculated. Sites falling outside the boundaries of 3 standard deviation of the mean (5552, 2577 and 2502 sites from CAR, FVG and VBI respectively) were removed.

Human knockout

To identify hKO we considered only deleterious variants in protein coding genes: we first selected variants with high impact as defined by VEP (McLaren et al., 2016) (i.e. frameshift, splice acceptor variant, splice donor variant, stop gained, stop lost, start lost, transcript ablation, transcript amplification) and among those we further selected for CADD score >=20. A total of 12,231 variants (8832 SNV and 3399 indels) were selected and 5,916 had a CAD score >=20. Among the variants those presenting at least one homozygous individual in one population were defined putative hKO. After filter application, the average number of hKO per individual was 20 (12-31), in agreement with previous determinations (Vagheesh M Narasimhan et al., 2016).

6. GWAS ON AMH QUANTITATIVE TRAIT

Women with less than 41 years, without any kind of ovaries surgeries were included in GWAS analysis performed with software GEMMA (Zhou & Stephens, 2014). AMH trait was normalized with square root and age, BMI and smoke behaviours were used as covariates. Analysis was performed both with data imputed to 1000G phase I and with Italian reference panel. Data were filtered to 0.4 for info quality of imputation and for MAF < 1%. Manhattan plot was performed in R language with qqman package (source qqman R packages). Variants were analysed with VEP (McLaren et al., 2016) for functional categories and a gene ontology analysis was performed with Panther (source PANTHER).

7. Fertility preservation in endometriosis patients: is AMH a reliable marker of the ovarian follicle density? (Garavaglia et al., 2017)

Human subject

A totoal of 202 women in premenopausal period with endometriosis undergoing surgery for the first time were included in the study. The mean

age of women was 34.7 ± 5.9 years (19-48 age range). The day before surgery AMH dosage was performed.

From a subgroup of 25 patients, ovarian biopsies were collected with micro-scissors with cold blade from healthy ovarian cortex, far from the cyst. Tissue samples of about 4x4 mm were transferred to the pathology laboratory in formaldehyde. Biopsies from 33 patients undergoing surgery for different cause from endometriosis were used as control.

A set of 200 women randomly selected from Val Borbera isolate population and other 200 from a cohort undergoing ART for male infertility at the Obstetrics and Gynecology Unit at San Raffaele Scientific Institute were used as controls for serum AMH dosage.

Hormone assay

Different immunoenzymatic assay kits were used for AMH serum levels: the GenII ELISA (Beckman Coulter) and the EIA AMH/MHS kit (Immunotech, Beckman Coulter). The limit of detection (LOD) of both assays was 0.14 ng/ml.

Tissue Preparation and Follicle counts

All sections of biobsies were stained with hematoxylin and eosin and analyzed using an Olympus microscope at 200× magnification. According to Gougeon's protocol (1986) we defined the developmental stages of follicles: the primordial follicle is a structure containing the primary oocyte surrounded by a single flat cells stratum, while the primitive follicle is an element assembled by the primary oocyte surrounded by the *zona pellucida* and by 1–2 cubical cells of the granulosa tissue. The follicle density was determined by dividing the number of follicles counted in 10 non-adjacent fields of 1 mm^2 each by the volume of tissue analyzed (0.5 mm^3).

Statistical analysis

Statistics analyses were performed in R 3.1.1 (source R Project for Statistical Computing) for analyzing follicles counts and the levels of AMH in the total sample (N = 602). Outlier values were excluded using a 3 SD threshold based on Shapiro–Wilk normality tests assessing the normality of the distributions.

We analyzed 588 AMH serum levels in 201 endometriosis cases and 387 controls and 57 follicle counts in 24 cases and 33 controls after outliers exclusion. Normality Shapiro-Wilk test failed, so follicle counts were square root transformed whereas AMH dosage was transformed with a rank-based inverse normalization using the quantile R function qnorm (p, μ, σ). The continuous distributions of each transformed trait were represented by density() R function. The R wilcox.test() function and/or t.test() function were used to estimate significant difference in the median and/or mean of case/control population. The glm() R function was used to fit generalized linear models in regression analyses. ggplot2 library was used for exploratory data analysis and plotting results.

CONCLUSION AND DISCUSSION

The work described in my thesis was aimed at identifying new genes and pathway involved in women fertility and indeed we succeeded in the task.

GWAS design in general, over the last decade, has led to a remarkably diverse set of discoveries in human genetics especially thanks to a continuously increase of samples number and to the possibility to study common and rare variants. For all complex traits, which have been studied with this kind of approach, it was shown that many genetic loci could contribute to the genetic variation of the trait in population. On average the proportion of genetic variance explained by each variant is very small and the genetic architecture of most traits studied is dominated by the additive effect of hundreds of variants with small effect. My results reveal a similar trend for fertility: for age at menarche, the 389 independent signals explained 7.4% of the population variance, for age at menopause the top 54 SNPs explain 6% of the variance that increase to 21% if we consider the top 29,958 independent SNPs with association p-value<0.05.

Loci identification may lead to new clues to the understanding of the mechanisms controlling each trait. Also in our case large efforts have been made in order to understand some of the biological mechanisms.

Regarding age at menarche the results of three GWAS showed that the effect of the variants and their heritability were higher for early versus late puberty timing. We discovered a significant role of imprinted genes in the regulation of age at menarche, namely we found that rare coding mutations in MKRN3 and DLK1 genes, when paternally inherited and expressed, confer a substantial decrease of age at menarche.

The relationship between puberty timing and BMI was already known (Mohamad et al., 2013) but our observations confirmed a strong inverse genetic correlation between age at menarche and **BMI**: earlier age of menarche is associated with higher BMI.

In addition the results show a complex association among age at menarche, BMI and adult cancer risk: it was reported (Hamajima et al., 2012) that lower adolescent BMI and earlier menarche were associated with a higher breast cancer risk. Adjusted for BMI, in order to see the effect of menarche, results of our analysis show evidence for association of an increased age at menarche with a lower risk of breast cancer estrogen receptor positive and ovarian

cancer. We speculate that this could be mediated by a shorter duration of exposure to sex steroids.

Also age of menopause is associated with cancer risk for the same reason, the prolonged exposure of estrogen and progesterone. Individuals carrying a higher number of variants that show an increase of age at menopause are more susceptible to breast cancer.

Some of the loci for menopause confirmed the involvement of **hypothalamic or pituitary activity in controlling age of menopause**: 5 loci, significantly associated with AAM, contain genes reported to be responsible for hypogonadotropic hypogonadism. This is a condition characterized by a low level of circulating estrogen and progesterone in females, caused by an impaired secretion of follicle-stimulating hormone (FSH) and luteinizing hormone (LH) by the pituitary gland in the brain.

Interestingly, two-thirds of the significant loci associated with age at menopause have a substantial role in **DNA damage response**, a remarkable excess of DDR genes in comparison to the number originally estimated.

BRCA1 is one of the genes in the list. It encodes a protein involved in repairing damaged DNA. Rare loss of function mutations in this gene were associated with breast cancer predisposition but variants in this gene had not been mapped to any complex trait. Our data showed that BRCA1 directly inhibits the transcriptional activation of ERα and thus variants that reduce BRCA1 expression altered estrogen signaling.

We identified 13 genes involved in homologous recombination-mediator repair. Two of the genes are specific for meiotic repair (MSH5 and DMC1): aberrant meiotic recombination can cause meiotic arrest and affect the viability of oocytes. We know that menopause occurs when the number of oocytes falls below approximately 1,000 (Gold, 2011) and thus a process that affects the size of the oocytes pool may affect age at menopause.

Also other repair mechanisms, mismatch repair, base-excision repair, mitotic repair and repair checkpoints are involved and due to the characteristics of oocyte, the prolonged state of cell cycle arrest until 50-60 years, oocyte may be particularly sensitive to DNA damage and the aberrant repair through life could affect the rate of oocyte loss as during female life the majority of oocyte

are lost by atresia

Only 2 genes, *MSH6* and *SPPL3*, are significantly associated with both menarche and menopause. The first gene, *MSH6*, is one of the genes involved in DDR pathway whereas the second common gene, a Signal Peptide Peptidase, has no known function. All together, the results suggest that the two traits are regulated quite differently.

Even though traits as age at first birth and number of children ever born are subject to individual choice, to cultural, social and economic environment, we demonstrated the importance of steroidogenesis pathway, in fact a significant signal in both AFB and NEB analysis is on *CRTC2* gene. The protein encoded by this gene has an important mediator role in follicole-stimulating hormone (FSH) and transforming growth factor (TGF)-beta1 stimulated steroidogenesis in ovarian granulosa cell, the biological process by which steroids are generated from cholesterol.

A common scenario in these results is also the evidence of involvement of genes with a role in cell cycle progression due to a control in methylation and expression of other genes.

Two genes, *FOXP2* and *FBXL17*, are significant associated respectively in AFB and NEB analysis and are in common with significant signal in menarche. The *FOXP2* encode for a transcription factor that may regulate hundreds of genes, it is active in several tissues, including the brain, both before and after birth. On gene *FBXL17* very few details are known, it forms a complex with SKP1 and cullin protein that acts as protein-ubiquitin ligases.

AMH GWAS on INGI women showed suggestive signal on gene *POLDIP3*, involved in regulation of translation, overexpressed in the ovary and “DNA replication” pathway enrichment.

We demonstrate the improvement of imputation quality with Italian reference panel in comparison with 1000G phase I panel, especially for rare variants.

Analysis on AMH in endometriosis patients allowed to emphasize the importance of ovarian cortical tissue cryopreservation for women that will have surgery for endometriosis. In fact, our results indicate that ovarian reserve in women with endometriosis and with more than 36 years decrease more rapidly than in healthy women.

In conclusion, many new genes and biological pathways were discovered that shed light on a number of reproduction related traits: they may be relevant for predicting early infertility before entry into menopause. The identification of novel genes and pathways will certainly provide new targets for biomarkers discovery and to study new drugs for personalized medicine.

Our approach emphasizes the ways in which disease risks are unique and different in patients with same disease but with different genome variants. Those disease risks are based on the predispositions written in genome at birth, combined with lifestyle and environment.
Nowadays the possibility to look and to integrate “omics” data of different types to advance hypotheses and answer to biological questions is increased.
Genomics data, analyzed together with trascriptomics, epigenomics, proteomics and metabolomics could be an enormous and extraordinary source that, analyzed with appropriate bioinformatics softwares, could bring a great advantage in personalized medicine.

BIBLIOGRAPHY

1000 Genomes Project Consortium, {fname}, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., … Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Adham, I. M., Eck, T. J., Mierau, K., Müller, N., Sallam, M. A., Paprotta, I., … Engel, W. (2005). Reduction of spermatogenesis but not fertility in Creb3l4-deficient mice. *Molecular and Cellular Biology*, *25*(17), 7657–64. https://doi.org/10.1128/MCB.25.17.7657-7664.2005

AlAsiri, S., Basit, S., Wood-Trageser, M. A., Yatsenko, S. A., Jeffries, E. P., Surti, U., … Rajkovic, A. (2015). Exome sequencing reveals MCM8 mutation underlies ovarian failure and chromosomal instability. *Journal of Clinical Investigation*, *125*(1), 258–262. https://doi.org/10.1172/JCI78473

Alkuraya, F. S. (2015). Human knockout research: New horizons and opportunities. *Trends in Genetics*, *31*(2), 108–115. https://doi.org/10.1016/j.tig.2014.11.003

Almasy, L., & Blangero, J. (1998). Multipoint Quantitative-Trait Linkage Analysis in General Pedigrees. *American Journal of Human Genetics*, *62*(5), 1198–211. https://doi.org/10.1086/301844

Anderson, S. E., Dallal, G. E., & Must, A. (2003). Relative weight and race influence average age at menarche: results from two nationally representative surveys of US girls studied 25 years apart. *Pediatrics*, *111*(4 Pt 1), 844–850. https://doi.org/10.1542/peds.111.4.844

Arcos-Burgos, M., & Muenke, M. (2002). Genetics of population isolates. *Clinical Genetics*, *61*(4), 233–247. https://doi.org/10.1034/j.1399-0004.2002.610401.x

Augood, C., Duckitt, K., & Templeton, a a. (1998). Smoking and female infertility: a systematic review and meta-analysis. *Human Reproduction (Oxford, England)*, *13*(6), 1532–1539. https://doi.org/10.1093/humrep/13.6.1532

Ayellet, V. S., Groop, L., Mootha, V. K., Daly, M. J., & Altshuler, D. (2010). Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genetics*, *6*(8). https://doi.org/10.1371/journal.pgen.1001058

Bager, H., Christensen, L. P., Husby, S., & Bjerregaard, L. (2017). Biomarkers for the Detection of Prenatal Alcohol Exposure: A Review. *Alcoholism, Clinical and Experimental Research*, *41*(2), 251–261. https://doi.org/10.1111/acer.13309

Barban, N., Jansen, R., de Vlaming, R., Vaez, A., Mandemakers, J. J., Tropf, F. C., … Mills, M. C. (2016). Genome-wide analysis identifies 12 loci influencing human reproductive behavior. *Nature Genetics*, *48*(12), 1462–1472. https://doi.org/10.1038/ng.3698

Bomba, L., Walter, K., & Soranzo, N. (2017). The impact of rare and low-frequency genetic variants in common disease. *Genome Biology*, *18*(1), 77. https://doi.org/10.1186/s13059-017-1212-4

Bonder, M. J., Luijk, R., Zhernakova, D. V, & Moed, M. (2016). Disease variants alter transcription factor levels and methylation of their binding sites. *Nature Genetics*, *49*(1), 131–138. https://doi.org/10.1038/ng.3721

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., … Snyder, M. (2012). Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research*, *22*(9), 1790–1797. https://doi.org/10.1101/gr.137323.112

Callaway, E. M. (2014). Geneticists tap human knockouts. *Nature*, 548. https://doi.org/10.1038/514548a

Chang, C. C., Chow, C. C., Tellier, L. C. A. M., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2014). Second-generation PLINK: rising to the challenge of larger and richer datasets, 1–16. https://doi.org/10.1186/s13742-015-0047-8

Chatenoud, L., Warncke, K., & Ziegler, A. (2012). Clinical Immunologic Interventions for the Treatment of Type 1 Diabetes, 1–18.

Chen, J., Bardes, E. E., Aronow, B. J., & Jegga, A. G. (2009). ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Research*, *37*(SUPPL. 2), 305–311. https://doi.org/10.1093/nar/gkp427

Choy, C. M. Y., Lam, C. W. K., Cheung, L. T. F., Briton-Jones, C. M., Cheung, L. P., & Haines, C. J. (2002). Infertility, blood mercury concentrations and dietary seafood consumption: A case-control study. *BJOG: An International Journal of Obstetrics and Gynaecology*, *109*(10), 1121–1125. https://doi.org/10.1016/S1470-0328(02)02984-1

Clark, a. M., Thornley, B., Tomlinson, L., Galletley, C., & Norman, R. J. (1998). Weight loss in obese infertile women results in improvement in reproductive outcome for all forms of fertility treatment. *Human Reproduction (Oxford, England)*, *13*(6), 1502–1505. https://doi.org/10.1093/humrep/13.6.1502

Collins, R. (2012). What makes UK Biobank special? *The Lancet*, *379*(9822), 1173–1174. https://doi.org/10.1016/S0140-6736(12)60404-8

Colonna, V., Pistis, G., Bomba, L., Mona, S., Matullo, G., Boano, R., … Toniolo, D. (2012). Small effective population size and genetic homogeneity in the Val Borbera isolate. *European Journal of Human Genetics*, (June 2012), 89–94. https://doi.org/10.1038/ejhg.2012.113

Consortium, E. N. C. O. D. E. P., Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C., & Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. https://doi.org/10.1038/nature11247

Consortium, R. E., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., …

Ziegler, S. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, *518*(7539), 317–330. https://doi.org/10.1038/nature14248

Dawber, T. R., Meadors, G. F., & Moore, F. E. (1951). Epidemiological Approaches to Heart Disease: The Framingham Study. *American Journal of Public Health and the Nations Health*, *41*(3), 279–286. https://doi.org/10.2105/AJPH.41.3.279

Day, F. R., Helgason, H., Chasman, D. I., Rose, L. M., Loh, P.-R., Scott, R. A., … Perry, J. R. B. (2016). Physical and neurobehavioral determinants of reproductive onset and success. *Nature Genetics*, *advance on*(6), 617–623. https://doi.org/10.1038/ng.3551

Day, F. R., Ruth, K. S., Thompson, D. J., Lunetta, K. L., Pervjakova, N., Chasman, D. I., … Murray, A. (2015). Large-scale genomic analyses link reproductive aging to hypothalamic signaling, breast cancer susceptibility and BRCA1-mediated DNA repair. *Nature Genetics*, *47*(11), 1294–1303. https://doi.org/10.1038/ng.3412

Day, F. R., Thompson, D. J., Helgason, H., Chasman, D. I., Finucane, H., Sulem, P., … Perry, J. R. B. (2017). Genomic analyses identify hundreds of variants associated with age at menarche and support a role for puberty timing in cancer risk. *Nature Genetics*, *49*(6), 834–841. https://doi.org/10.1038/ng.3841

Deelen, J., Beekman, M., Uh, H. W., Broer, L., Ayers, K. L., Tan, Q., … Slagboom, P. E. (2014). Genome-wide association meta-analysis of human longevity identifies a novel locus conferring survival beyond 90 years of age. *Human Molecular Genetics*, *23*(16), 4420–4432. https://doi.org/10.1093/hmg/ddu139

Desai, S., Wood-Trageser, M., Matic, J., Chipkin, J., Jiang, H., Bachelot, A., … Rajkovic, A. (2017). MCM8 and MCM9 nucleotide variants in women with primary ovarian insufficiency. *Journal of Clinical Endocrinology and Metabolism*, *102*(2), 576–582. https://doi.org/10.1210/jc.2016-2565

Doughty, D., & Rodgers, J. L. (2000). Behavior Genetic Modeling of Menarche in U.S. Females. *Genetic Influences on Human Fertility and Sexuality*, 169–181.

Durlinger, A. L. L., Kramer, P., Karels, B. a S., Jong, F. H. D. E., Uilenbroek, J. a N. T. H. J., Grootegoed, J. A., … Vigier, B. (1999). ¨ llerian Hormone in the Mouse Ovary *. *Endocrinology*, *140*(12), 5789–5796. https://doi.org/10.1210/en.140.12.5789

Esko, T., Mezzavilla, M., Nelis, M., Borel, C., Debniak, T., Jakkula, E., … D'Adamo, P. (2013). Genetic characterization of northeastern Italian population isolates in the context of broader European genetic diversity. *European Journal of Human Genetics : EJHG*, *21*(6), 659–65. https://doi.org/10.1038/ejhg.2012.229

Euesden, J., Lewis, C. M., & O'Reilly, P. F. (2015). PRSice: Polygenic Risk Score software. *Bioinformatics*, *31*(9), 1466–1468. https://doi.org/10.1093/bioinformatics/btu848

Fadista, J., Manning, A. K., Florez, J. C., & Groop, L. (2016). The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. *European Journal of Human Genetics*, *24*(8), 1202–1205.

https://doi.org/10.1038/ejhg.2015.269

Fang, W. L., Lee, M. T., Wu, L. S., Chen, Y. J., Mason, J., Ke, F. C., & Hwang, J. J. (2012). CREB coactivator CRTC2/TORC2 and its regulator calcineurin crucially mediate follicle-stimulating hormone and transforming growth factor β1 upregulation of steroidogenesis. *Journal of Cellular Physiology*, *227*(6), 2430–2440. https://doi.org/10.1002/jcp.22978

Finucane, H., Reshef, Y., Anttila, V., Slowikowski, K., Gusev, A., Byrnes, A., … Price, A. (2017). Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *bioRxiv*, 103069. https://doi.org/10.1101/103069

Fogli, A., Rodriguez, D., Eymard-Pierre, E., Bouhour, F., Labauge, P., Meaney, B. F., … Boespflug-Tanguy, O. (2003). Ovarian failure related to eukaryotic initiation factor 2B mutations. *American Journal of Human Genetics*, *72*(6), 1544–1550. https://doi.org/10.1086/375404

Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., … Stewart, J. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, *449*(7164), 851–61. https://doi.org/10.1038/nature06258

Garavaglia, E., Sala, C., Taccagni, G., Traglia, M., Barbieri, C., Ferrari, S., … Toniolo, D. (2017). Fertility Preservation in Endometriosis Patients: Anti-Müllerian Hormone Is a Reliable Marker of the Ovarian Follicle Density. *Frontiers in Surgery*, *4*(July), 1–6. https://doi.org/10.3389/fsurg.2017.00040

Garcìa-Ortega, J., Pinto, F. M., Fernandez-Sanchez, M., Prados, N., Cejudo-Rom??n, A., Almeida, T. A., … Candenas, L. (2014). Expression of neurokinin B/NK3 receptor and kisspeptin/KISS1 receptor in human granulosa cells. *Human Reproduction*, *29*(12), 2736–2746. https://doi.org/10.1093/humrep/deu247

Global Lipids Genetics Consortium, Willer, C. J., Schmidt, E. M., Sengupta, S., Peloso, G. M., Gustafsson, S., … Abecasis, G. R. (2013). Discovery and refinement of loci associated with lipid levels. *Nature Genetics*, *45*(11), 1274–83. https://doi.org/10.1038/ng.2797

Gold, E. B. (2011). The Timing of the Age at Which Natural Menopause Occurs. *Obstetrics and Gynecology Clinics of North America*, *38*(3), 425–440. https://doi.org/10.1016/j.ogc.2011.05.002

Goldstein, J. I., Crenshaw, A., Carey, J., Grant, G. B., Maguire, J., Fromer, M., … Neale, B. M. (2012). Zcall: A rare variant caller for array-based genotyping. *Bioinformatics*, *28*(19), 2543–2545. https://doi.org/10.1093/bioinformatics/bts479

Grynnerup, A. G. A., Lindhard, A., & Sørensen, S. (2012). The role of anti-Müllerian hormone in female fertility and infertility - An overview. *Acta Obstetricia et Gynecologica Scandinavica*, *91*(11), 1252–1260. https://doi.org/10.1111/j.1600-0412.2012.01471.x

Guler, G. D., Liu, H., Vaithiyalingam, S., Arnett, D. R., Kremmer, E., Chazin, W. J., & Fanning, E. (2012). Human DNA helicase B (HDHB) binds to replication protein

A and facilitates cellular recovery from replication stress. *Journal of Biological Chemistry*, *287*(9), 6469–6481. https://doi.org/10.1074/jbc.M111.324582

Hamajima, N., Hirose, K., Tajima, K., Rohan, T., Friedenreich, C. M., Calle, E. E., … Fukao, A. (2012). Menarche, menopause, and breast cancer risk: Individual participant meta-analysis, including 118 964 women with breast cancer from 117 epidemiological studies. *The Lancet Oncology*, *13*(11), 1141–1151. https://doi.org/10.1016/S1470-2045(12)70425-4

Hartge, P. (2009). Genetics of reproductive lifespan. *Nature Genetics*, *41*(6), 637–638. https://doi.org/10.1038/ng0609-637

Heck, K. E., Schoendorf, K. C., Ventura, S. J., & Kiely, J. L. (1997). Delayed Childbearing by Education Level in the United States, 1969–1994. *Maternal and Child Health Journal*, *1*(2), 81–88. https://doi.org/10.1023/a:1026218322723

Hindorff, L. a, Sethupathy, P., Junkins, H. a, Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. a. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(23), 9362–7. https://doi.org/10.1073/pnas.0903103106

Hofman, A., Breteler, M. M. B., Van Duijn, C. M., Janssen, H. L. A., Krestin, G. P., Kuipers, E. J., … Witteman, J. C. M. (2009). The rotterdam study: 2010 objectives and design update. *European Journal of Epidemiology*, *24*(9), 553–572. https://doi.org/10.1007/s10654-009-9386-z

Hong, E. P., & Park, J. W. (2012). Sample size and statistical power calculation in genetic association studies. *Genomics & Informatics*, *10*(2), 117–22. https://doi.org/10.5808/GI.2012.10.2.117

Howie, B., Marchini, J., & Stephens, M. (2011). Genotype Imputation with Thousands of Genomes. *G3*, *1*(6), 457–470. https://doi.org/10.1534/g3.111.001198

Huang, J., Howie, B., McCarthy, S., Memari, Y., Walter, K., Min, J. L., … Soranzo, N. (2015). Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nature Communications*, *6*, 8111. https://doi.org/10.1038/ncomms9111

International, T., Consortium, H., & The International HapMap, C. (2005). A haplotype map of the human genome. *Nature*, *437*(7063), 1299–1320. https://doi.org/10.1038/nature04226

Jamil, Z., Fatima, S. S., Ahmed, K., & Malik, R. (2016). Anti-Mullerian Hormone: Above and beyond Conventional Ovarian Reserve Markers. *Disease Markers*, *2016*. https://doi.org/10.1155/2016/5246217

Joy, T., Cao, H., Black, G., Malik, R., Charlton-Menys, V., Hegele, R. a, & Durrington, P. N. (2007). Alstrom syndrome (OMIM 203800): a case report and literature review. *Orphanet Journal of Rare Diseases*, *2*(Omim 203800), 49. https://doi.org/10.1186/1750-1172-2-49

Kaprio, J., Rimpelä, A., Winter, T., Viken, R. J., Rimpelä, M., & Rose, R. J. (1995). Common genetic influences on BMI and age at menarche. *Hum Biol*, *67*(5), 739–753.

Kerr, J. B., Myers, M., & Anderson, R. A. (2013). The dynamics of the primordial follicle reserve. *Reproduction*, *146*(6). https://doi.org/10.1530/REP-13-0181

Khosla, S., & Monroe, D. G. (2017). Regulation of Bone Metabolism by Sex Steroids. *Cold Spring Harbor Perspectives in Medicine*, a031211. https://doi.org/10.1101/cshperspect.a031211

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, *46*(3), 310–315. https://doi.org/10.1038/ng.2892

Kirk, K. M., Blomberg, S. P., Duffy, D. L., Heath, A. C., Owens, I. P. F., & Martin, N. G. (2001). Natural Selection and Quantitative Genetics of Life-History Traits in Western Women: a Twin Study. *Evolution*, *55*(2), 423–435. https://doi.org/10.1111/j.0014-3820.2001.tb01304.x

La Marca, A., Sighinolfi, G., Giulini, S., Traglia, M., Argento, C., Sala, C., … Toniolo, D. (2010). Normal serum concentrations of anti-Mullerian hormone in women with regular menstrual cycles. *Reproductive BioMedicine Online*, *21*(4), 463–469. https://doi.org/10.1016/j.rbmo.2010.05.009

La Marca, A., Stabile, G., Carducci Artenisio, A., & Volpe, A. (2006). Serum anti-Mullerian hormone throughout the human menstrual cycle. *Human Reproduction*, *21*(12), 3103–3107. https://doi.org/10.1093/humrep/del291

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., … International Human Genome Sequencing, C. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

Li, C., Liu, V. W., Chiu, P. M., Chan, D. W., & Ngan, H. Y. (2012). Over-expressions of AMPK subunits in ovarian carcinomas with significant clinical implications. *BMC Cancer*, *12*(1), 357. https://doi.org/10.1186/1471-2407-12-357

Lie Fong, S., Visser, J. A., Welt, C. K., De Rijke, Y. B., Eijkemans, M. J. C., Broekmans, F. J., … Laven, J. S. E. (2012). Serum anti-m??llerian hormone levels in healthy females: A nomogram ranging from infancy to adulthood. *Journal of Clinical Endocrinology and Metabolism*, *97*(12), 4650–4655. https://doi.org/10.1210/jc.2012-1440

Lindhardt Johansen, M., Hagen, C. P., Johannsen, T. H., Main, K. M., Picard, J.-Y., Jørgensen, A., … Johansen, M. L. (2013). Anti-Müllerian Hormone and Its Clinical Use in Pediatrics with Special Emphasis on Disorders of Sex Development. *International Journal of …*, *2013*, 198698. https://doi.org/10.1155/2013/198698

Liu, J. Z., McRae, A. F., Nyholt, D. R., Medland, S. E., Wray, N. R., Brown, K. M., … MacGregor, S. (2010). A versatile gene-based test for genome-wide association

studies. *American Journal of Human Genetics*, *87*(1), 139–145. https://doi.org/10.1016/j.ajhg.2010.06.009

Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., … Speliotes, E. K. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature*, *518*(7538), 197–206. https://doi.org/10.1038/nature14177

Loesch, D. Z., Hopper, J. L., Rogucka, E., & Huggins, R. M. (1995). Timing and genetic rapport between growth in skeletal maturity and height around puberty: similarities and differences between girls and boys. *American Journal of Human Genetics*, *56*(3), 753–9. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1801156&tool=pmcentrez&rendertype=abstract

Lomniczi, A., Wright, H., & Ojeda, S. R. (2015). Epigenetic regulation of female puberty. *Frontiers in Neuroendocrinology*, *36*, 90–107. https://doi.org/10.1016/j.yfrne.2014.08.003

Lunetta, K. L., Day, F. R., Sulem, P., Ruth, K. S., Tung, J. Y., Hinds, D. A., … Perry, J. R. B. (2015). Rare coding variants and X-linked loci associated with age at menarche. *Nature Communications*, *6*, 7756. https://doi.org/10.1038/ncomms8756

Mandon-Pépin, B., Touraine, P., Kuttenn, F., Derbois, C., Rouxel, A., Matsuda, F., … Fellous, M. (2008). Genetic investigation of four meiotic genes in women with premature ovarian failure. *European Journal of Endocrinology*, *158*(1), 107–115. https://doi.org/10.1530/EJE-07-0400

Marchini, J., & Howie, B. (2008). Comparing Algorithms for Genotype Imputation. *American Journal of Human Genetics*, *83*(4), 535–539. https://doi.org/10.1016/j.ajhg.2008.09.007

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, *11*(7), 499–511. https://doi.org/10.1038/nrg2796

Matuszczak, E., Hermanowicz, A., Komarowska, M., & Debek, W. (2013). Serum AMH in physiology and pathology of male gonads. *International Journal of Endocrinology*, *2013*. https://doi.org/10.1155/2013/128907

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., … Haplotype Reference Consortium. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, *48*(10), 1279–1283. https://doi.org/10.1038/ng.3643

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., … Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*(1), 122. https://doi.org/10.1186/s13059-016-0974-4

McVean, G. A., Altshuler (Co-Chair), D. M., Durbin (Co-Chair), R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., … McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491*(7422), 56–65.

https://doi.org/10.1038/nature11632

Meyer, J. M., Eaves, L. J., Heath, a C., & Martin, N. G. (1991). Estimating genetic influences on the age-at-menarche: a survival analysis approach. *American Journal of Medical Genetics*, *39*(2), 148–54. https://doi.org/10.1002/ajmg.1320390207

Mills, M., Rindfuss, R. R., McDonald, P., & te Velde, E. (2011). Why do people postpone parenthood? Reasons and social policy incentives. *Human Reproduction Update*, *17*(6), 848–860. https://doi.org/10.1093/humupd/dmr026

Mishra, A., & Macgregor, S. (2015). VEGAS2: Software for More Flexible Gene-Based Testing. *Twin Research and Human Genetics*, *18*(1), 86–91. https://doi.org/10.1017/thg.2014.79

Mohamad, K., Jamshidi, L., & Nouri Jelyani, K. (2013). Is age of menarche related with body mass index? *Iranian Journal of Public Health*, *42*(9), 1043–1048.

Monniaux, D., Clément, F., Dalbiès-Tran, R., Estienne, A., Fabre, S., Mansanet, C., & Monget, P. (2014). The ovarian reserve of primordial follicles and the dynamic reserve of antral growing follicles: what is the link? *Biology of Reproduction*, *90*(4), 85. https://doi.org/10.1095/biolreprod.113.117077

Murabito, J. M., Yang, Q., Fox, C., Wilson, P. W. F., & Cupples, L. A. (2005). Heritability of age at natural menopause in the framingham heart study. *Journal of Clinical Endocrinology and Metabolism*, *90*(6), 3427–3430. https://doi.org/10.1210/jc.2005-0181

Narasimhan, V. M., Hunt, K. A., Mason, D., Baker, C. L., Karczewski, K. J., Barnes, M. R., … Maher, E. R. (2016). Health and population effects of rare gene knockouts in adult humans with related parents, *352*(6284), 474–477. https://doi.org/10.1126/science.aac8624.Health

Narasimhan, V. M., Hunt, K. A., Mason, D., Baker, C. L., Karczewski, K. J., Barnes, M. R., … van Heel, D. A. (2016). Health and population effects of rare gene knockouts in adult humans with related parents. *Science*, *352*(6284), 474–477. https://doi.org/10.1126/science.aac8624

Nelson, S. M., Telfer, E. E., & Anderson, R. A. (2013). The ageing ovary and uterus: New biological insights. *Human Reproduction Update*, *19*(1), 67–83. https://doi.org/10.1093/humupd/dms043

Okbay, A., Beauchamp, J. P., Fontana, M. A., Lee, J. J., Pers, T. H., Rietveld, C. A., … Benjamin, D. J. (2016). Genome-wide association study identifies 74 loci associated with educational attainment. *Nature*, *533*(7604), 539–542. https://doi.org/10.1038/nature17671

Pankhurst, M. W. (2017). A putative role for anti-Müllerian hormone (AMH) in optimising ovarian reserve expenditure. *Journal of Endocrinology*, *233*(1), R1–R13. https://doi.org/10.1530/JOE-16-0522

Pawli, R., & Szwed, A. (2007). Cigarette smoking and age at natural menopause of

women in Poland, *2*(1), 1–5.

Perry, J. R. B., Day, F., Elks, C. E., Sulem, P., Thompson, D. J., Ferreira, T., … Ong, K. K. (2014). Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature*, *514*(7520), 92–7. https://doi.org/10.1038/nature13545

Perry, J. R., Hsu, Y. H., Chasman, D. I., Johnson, A. D., Elks, C., Albrecht, E., … Murray, A. (2014). DNA mismatch repair gene MSH6 implicated in determining age at natural menopause. *Hum Mol Genet*, *23*(9), 2490–2497. https://doi.org/10.1093/hmg/ddt620

Pers, T. H., Dworzynski, P., Thomas, C. E., Lage, K., & Brunak, S. (2013). MetaRanker 2.0: a web server for prioritization of genetic variation data. *Nucleic Acids Research*, *41*(Web Server issue), 104–108. https://doi.org/10.1093/nar/gkt387

Pers, T. H., Karjalainen, J. M., Chan, Y., Westra, H.-J., Wood, A. R., Yang, J., … Franke, L. (2015). Biological interpretation of genome-wide association studies using predicted gene functions. *Nature Communications*, *6*, 5890. https://doi.org/10.1038/ncomms6890

Provine, W. B. (2004). Ernst Mayr: Genetics and speciation. *Genetics*, *167*(3), 1041–1046. https://doi.org/10.1016/j.tree.2005.10.001

Purcell, S., Cherny, S. S., & Sham, P. C. (2003). Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits S. *Bioinformatics Applications Note*, *17*(2), 192–193.

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., … Aiden, E. L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, *159*(7), 1665–1680. https://doi.org/10.1016/j.cell.2014.11.021

Raychaudhuri, S., Plenge, R. M., Rossin, E. J., Ng, A. C. Y., Purcell, S. M., Sklar, P., … Williams, N. M. (2009). Identifying relationships among genomic disease regions: Predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genetics*, *5*(6). https://doi.org/10.1371/journal.pgen.1000534

Rödström, K., Bengtsson, C., Milsom, I., Lissner, L., Sundh, V., & Bjoürkelund, C. (2003). Evidence for a secular trend in menopausal age: a population study of women in Gothenburg. *Menopause (New York, N.Y.)*, *10*(6), 538–543. https://doi.org/10.1097/01.GME.0000094395.59028.0F

Rowe, D. C. (2000). Environmental and Genetic Influences on Pubertal Development: Evolutionary Life History Traits? In J. L. Rogers, R. B. Miller, & D. C. Rowe (Eds.), *Genetic Influences on Human Fertility and Sexuality* (pp. 147–168). Boston: Kluwer Academic Publishers.

Saleh, M., Vaillancourt, J. P., Graham, R. K., Huyck, M., Srinivasula, S. M., Alnemri, E. S., … Nicholson, D. W. (2004). Differential modulation of endotoxin responsiveness by human caspase-12 polymorphisms. *Nature*, *429*(6987), 75–79. https://doi.org/10.1038/nature02451

Serour, G. I., & Serour, A. G. (2017). Ethical issues in infertility. *Best Practice and Research: Clinical Obstetrics and Gynaecology*, *43*, 21–31. https://doi.org/10.1016/j.bpobgyn.2017.02.008

Sharpe, R. M., & Franks, S. (2002). Environment, lifestyle and infertility -- an inter-generational issue: Serviço de Descoberta da Universidade de Coimbra, 33–40. Retrieved from http://eds.a.ebscohost.com/eds/detail/detail?sid=cad67123-d407-4586-9a9e-7c13a58d27f1%40sessionmgr4006&vid=0&hid=4113&bdata=Jmxhbmc9cHQtYnImc2l0ZT1lZHMtbGl2ZQ%3D%3D#AN=47028713&db=a9h

Silveira, L. F. G., & Latronico, A. C. (2013). Approach to the patient with hypogonadotropic hypogonadism. *The Journal of Clinical Endocrinology and Metabolism*, *98*(5), 1781–8. https://doi.org/10.1210/jc.2012-3550

Skakkebaek, N. E., Rajpert-De Meyts, E., Buck Louis, G. M., Toppari, J., Andersson, A.-M., Eisenberg, M. L., … Juul, A. (2015). Male Reproductive Disorders and Fertility Trends: Influences of Environment and Genetic Susceptibility. *Physiological Reviews*, *96*(1), 55–97. https://doi.org/10.1152/physrev.00017.2015

Snieder, H., Macgregor, A. J., & Spector, T. D. (1998). Genes control the cessation of a woman's reproductive life: A twin study of hysterectomy and age at menopause. *Journal of Clinical Endocrinology and Metabolism*, *83*(6), 1875–1880. https://doi.org/10.1210/jc.83.6.1875

Sobotka, T. (2004). Is lowest-low fertility in Europe explained by the postponement of childbearing? *Population and Development Review*, *30*(2), 195–220. https://doi.org/10.1111/j.1728-4457.2004.010_1.x

Stewart, A. J., Katz, A. A., Millar, R. P., & Morgan, K. (2009). Retention and Silencing of Prepro-GnRH-II and Type II GnRH Receptor Genes in Mammals. *Neuroendocrinology*.

Stolk, L., Perry, J. R. B., Chasman, D. I., He, C., Mangino, M., Sulem, P., … Lunetta, K. L. (2012). Meta-analyses identify 13 loci associated with age at menopause and highlight DNA repair and immune pathways. *Nature Genetics*, *44*(3), 260–8. https://doi.org/10.1038/ng.1051

Streuli, I., Fraisse, T., Pillet, C., Ibecheole, V., Bischof, P., & de Ziegler, D. (2008). Serum antimüllerian hormone levels remain stable throughout the menstrual cycle and after oral or vaginal administration of synthetic sex steroids. *Fertility and Sterility*, *90*(2), 395–400. https://doi.org/10.1016/j.fertnstert.2007.06.023

Swan, H. J. C. (2000). The framingham offspring study: A commentary. *Journal of the American College of Cardiology*, *35*(5), 13B–17B. https://doi.org/10.1016/S0735-1097(00)80043-1

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., … Von Mering, C. (2015). STRING v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, *43*(D1), D447–D452. https://doi.org/10.1093/nar/gku1003

te Velde, E. R., Eijkemans, M. J. C., & Habbema, J. D. F. (2000). Variation in couple fecundity and time to pregnancy, an essential concept in human reproduction. *The Lancet*, *355*, 1928–1929. https://doi.org/10.1016/S0140-6736(00)03202-5

te Velde, E. R., & Pearson, P. L. (2002). The variability of female reproductive aging. *Human Reproduction Update*, *8*(2), 141–154. https://doi.org/10.1093/humupd/8.2.141

The International HapMap 3 Consortium, Altshuler, D. M., Gibbs, R. A., Peltonen, L., Dermitzakis, E., Schaffner, S. F., … McEwen, J. E. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, *467*(7311), 52–8. https://doi.org/10.1038/nature09298

Tigchelaar, E. F., Zhernakova, A., Dekens, J. A. M., Hermes, G., Baranska, A., Mujagic, Z., … Feskens, E. J. M. (2015). Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ Open*, *5*(8), e006772. https://doi.org/10.1136/bmjopen-2014-006772

Towne, B., Czerwinski, S. A., Demerath, E. W., Blangero, J., Roche, A. F., & Siervogel, R. M. (2005). Heritability of age at menarche in girls from the Fels Longitudinal Study. *American Journal of Physical Anthropology*, *128*(1), 210–219. https://doi.org/10.1002/ajpa.20106

Traglia, M., Sala, C., Masciullo, C., Cverhova, V., Lori, F., Pistis, G., … Toniolo, D. (2009). Heritability and demographic analyses in the large isolated population of val borbera suggest advantages in mapping complex traits genes. *PLoS ONE*, *4*(10), 1–10. https://doi.org/10.1371/journal.pone.0007554

Tranchevent, L. C., Barriot, R., Yu, S., Van Vooren, S., Van Loo, P., Coessens, B., … Moreau, Y. (2008). ENDEAVOUR update: a web resource for gene prioritization in multiple species. *Nucleic Acids Research*, *36*(Web Server issue), 377–384. https://doi.org/10.1093/nar/gkn325

Treloar, S. A., & Martin, N. G. (1990). Age at menarche as a fitness trait: nonadditive genetic variance detected in a large twin sample. *American Journal of Human Genetics*, *47*(1), 137–48. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2349942%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1683767

Trifunovic, A., Wredenberg, A., Falkenberg, M., Spelbrink, J. N., Rovio, A. T., Bruder, C. E., … Larsson, N.-G. (2004). Letters To Nature. *Nature*, *429*(May), 417–423. https://doi.org/10.1038/nature02544.1.

Tropf, F. C., Stulp, G., Barban, N., Visscher, P. M., Yang, J., Snieder, H., & Mills, M. C. (2015). Human fertility, molecular genetics, and natural selection in modern societies. *PLoS ONE*, *10*(6), 1–14. https://doi.org/10.1371/journal.pone.0126821

Tsukamoto, S., Hara, T., Yamamoto, A., Ohta, Y., Wada, A., Ishida, Y., … Kokubo, T. (2013). Functional Analysis of Lysosomes During Mouse Preimplantation Embryo Development, *59*(1).

Van Dam, R. M., Boer, J. M. A., Feskens, E. J. M., & Seidell, J. C. (2001). Parental

history off diabetes modifies the association between abdominal adiposity and hyperglycemia. *Diabetes Care*, *24*(8), 1454–1459. https://doi.org/10.2337/diacare.24.8.1454

Van den Akker, O., Stein, G., Neale, M., & Murray, R. (1987). Genetic and Environmental Variation in Menstrual Cycle: Histories of Two British Twin Samples., (4), 541–548. https://doi.org/https://doi.org/10.1017/S0001566000006929

van Noord, P. A., Peeters, P. H., Grobbee, D. E., Dubas, J. S., & te Velde, E. (1999). Onset of natural menopause. *J Clin Epidemiol*, *52*(12), 1290–1292. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=10580794

Varilo, T., & Peltonen, L. (2004). Isolates and their potential use in complex gene mapping efforts. *Current Opinion in Genetics and Development*, *14*(3), 316–323. https://doi.org/10.1016/j.gde.2004.04.008

Velie, E. M., Nechuta, S., & Osuch, J. R. (2005). Lifetime reproductive and anthropometric risk factors for breast cancer in postmenopausal women. *Breast Disease*, *24*(1), 17–35. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-33747338458&partnerID=40&md5=92b228ac23350b4a1590f9cd220a413c

Voight, B. F., Peloso, G. M., Orho-Melander, M., Frikke-Schmidt, R., Barbalic, M., Jensen, M. K., … Kathiresan, S. (2012). Plasma HDL cholesterol and risk of myocardial infarction: A mendelian randomisation study. *The Lancet*, *380*(9841), 572–580. https://doi.org/10.1016/S0140-6736(12)60312-2

Watanabe, K., Clarke, T. R., Lane, A. H., Wang, X., & Donahoe, P. K. (2000). Endogenous expression of Müllerian inhibiting substance in early postnatal rat sertoli cells requires multiple steroidogenic factor-1 and GATA-4-binding sites. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(4), 1624–9. https://doi.org/97/4/1624 [pii]

Weenen, C., Laven, J. S. E., von Bergh, A. R. M., Cranfield, M., Groome, N. P., Visser, J. A., … Themmen, A. P. N. (2004). Anti-M??llerian hormone expression pattern in the human ovary: Potential implications for initial and cyclic follicle recruitment. *Molecular Human Reproduction*, *10*(2), 77–83. https://doi.org/10.1093/molehr/gah015

Westra, H.-J., Peters, M. J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., … Franke, L. (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature Genetics*, *45*(10), 1238–43. https://doi.org/10.1038/ng.2756

Willemsen, G., Vink, J. M., Abdellaoui, A., Den Braber, A., Van Beek, J. H. D. A., Draisma, H. H. M., … Boomsma, D. I. (2013). The Adult Netherlands Twin Register: Twenty-Five Years of Survey and Biological Data Collection. *Twin Research and Human Genetics*, *16*(1), 271–281. https://doi.org/10.1017/thg.2012.140

Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., … Frayling, T. M. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet*, *46*(11), 1173–1186. https://doi.org/10.1038/ng.3097\rhttp://www.nature.com/ng/journal/v46/n11/abs/ng.3097.html#supplementary-information

Zhernakova, D. V, Deelen, P., Vermaat, M., Iterson, M. Van, & Van, M. (2015). Hypothesis-free identification of modulators of genetic risk factors, 1–25. https://doi.org/10.1101/033217

Zhou, X., & Stephens, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods*, *11*(4), 407–409. https://doi.org/10.1038/nmeth.2848

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., … Yang, J. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics*, *48*(5), 481–7. https://doi.org/10.1038/ng.3538

WEBLIOGRAPHY

23andMe, https://www.23andme.com/en-int/research/

deCODE, https://www.decode.com/

Eurostat, http://ec.europa.eu/eurostat/statistics-explained/index.php/Fertility_statistics

Gene Ontology Consortium, http://geneontology.org/

Genetic Power Calculator, http://zzz.bwh.harvard.edu/gpc/

GWAS Catalog, https://www.ebi.ac.uk/gwas/

PANTHER, http://pantherdb.org/

Qqman R packages, http://cran.r-project.org/web/packages/qqman/

R Project for Statistical Computing, https://www.r-project.org/

VEP, http://www.ensembl.org/info/docs/tools/vep/index.html

LIST OF PUBBLICATIONS

1. Dr. Layal Chaker , Dr. Stefan Groeneweg , Dr. Yong Li , Ms. Celia Di Munno , Ms. Caterina Barbieri , Prof. Henry Völzke , Dr. Serenna Sanna , Dr. Anna Kottgen , Dr. Theo Visser , Dr. Marco Medici at al. **"Genome-wide analyses identify novel players in thyroid hormone regulation."** Submitted to Nature Genetics, 2nd Feb 18.

2. Dr. Evangelos Evangelou at al (including C. Barbieri). **"Genetic analysis of over one million people identifies 535 novel loci for blood pressure."** Submitted to Nature Genetics, 31st Jan 18.

3. Garavaglia E. at al (including C. Barbieri). **"Fertility preservation in endometriosis patients: is AMH a reliable marker of the ovarian follicle density?"** – Front Surg. 2017 Jul 25;4:40. doi: 10.3389/fsurg.2017.00040. eCollection 2017.

4. Wain L. at al (including C. Barbieri). **"Novel blood pressure locus and gene discovery using GWAS and expression datasets from blood and the kidney"** – Hypertension. 2017 Jul 24. doi: 10.1161/HYPERTENSIONAHA.117.09438.

5. Day FR at al (including C. Barbieri). **Genomic analyses identify hundreds of variants associated with age at menarche and support a role for puberty timing in cancer risk.** Nat Genet. 2017 Jun;49(6):834-841. doi: 10.1038/ng.3841. Epub 2017 Apr 24.

6. Gunter Schumann at al (including C. Barbieri). **'*KLB* is associated with alcohol drinking, and its gene product β-Klotho is necessary for FGF21 regulation of alcohol preference'** – *Proc Natl Acad Sci U S A. 2016 Dec 13, DOI: 10.1073/pnas.1611243113*

7. Desai S at al (including C. Barbieri). November 2016 '**MCM8 and MCM9 Nucleotide Variants in Women with Primary Ovarian Insufficiency.'** – J Clin Endocrinol Metab *(2016)* PMID: 27802094

8. Iotchkova V at al (including C. Barbieri). November 2016 '**Discovery and refinement of genetic loci associated with cardiometabolic risk using dense imputation maps'** – *Nature Genetics (2016)* doi:10.1038/ng.3668

9. Galesloot TE at al (including C. Barbieri). **Meta-GWAS and Meta-Analysis of Exome Array Studies Do Not Reveal Genetic Determinants of Serum Hepcidin**. PLoS One. 2016 Nov 15;11(11):e0166628. doi: 10.1371/journal.pone.0166628. eCollection 2016.

10. Galesloot TE at al (including C. Barbieri). **'Meta-GWAS and Meta-Analysis of Exome Array Studies Do Not Reveal Genetic Determinants of Serum**

Hepcidin.' - PLoS One. 2016 Nov 15, doi: 10.1371/journal.pone.0166628.

11. Barban N at al (including C. Barbieri). October 2016 '**Genome-wide analysis identifies 12 loci influencing human reproductive behavior.'** – *Nature Genetics (2016)* doi: 10.1038/ng.3698

12. Seung-Hoan Choi at al (including C. Barbieri). February 2016 '**Six novel loci associated with circulating VEGF levels identified by a meta-analysis of genome-wide association studies'** – *PLoS Genet. 2016 Feb 24;12(2):e1005874. doi: 10.1371/journal.pgen.1005874. eCollection 2016.*

13. Kathryn L. Lunetta at al (including C. Barbieri). August 2015 **'Rare coding variants and X-linked loci associated with age at menarche'** – *Nature Communications* **6**, Article number: 7756 doi:10.1038/ncomms8756

14. Day FR at al (including C. Barbieri). February 2015 '**Large-scale genomic analyses link reproductive aging to hypothalamic signaling, breast cancer susceptibility and BRCA1-mediated DNA repair.'** – Nature Genetics 2015 Nov;47(11):1294-303. doi: 10.1038/ng.3412. Epub 2015 Sep 28.

15. Jennifer Wessel at al (including C. Barbieri). January 2015 '**Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility'** - *Nature Communications* **6**, Article number: 5897 doi:10.1038/ncomms6897

ACKNOWLEDGMENTS

Un ringraziamento alla prof.ssa Daniela Toniolo che mi ha guidato durante questo percorso di dottorato con saggi consigli e mi ha dato la straordinaria opportunità di entrare a contatto con la ricerca a livello internazionale.

Un ringraziamento particolare va anche al Prof. Paolo Gasparini e a tutto il suo gruppo di Trieste: la continua collaborazione nel progetto INGI, le lunghe discussioni e critici scambi di idee hanno reso possibile parte del lavoro.

Un sentito grazie a Cinzia, Michela e Max che sono sempre stati disponibili a confronti scientifici.

Grazie alla mia famiglia che mi ha sostenuta e sempre incoraggiata durante tutto il percorso. Grazie a Alan e a tutti i miei amici con cui ho condiviso fatiche, ansie, gioie e soddisfazioni.