



# Assessment of fertility associated variants in a Portuguese cohort of Azoospermia and Severe Oligozoospermia

Cláudia Sofia da Silva Costa

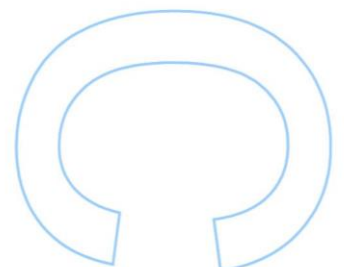
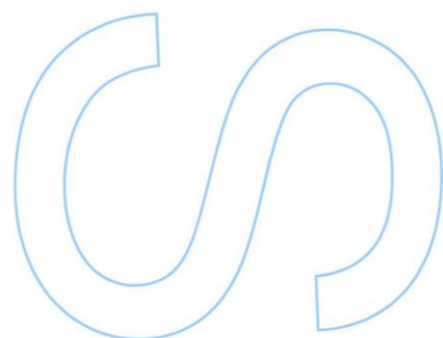
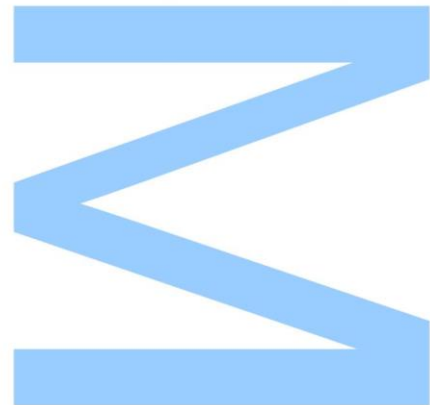
Mestrado em Genética Forense

Departamento de Biologia

2020

## Orientador

Alexandra Lopes, PhD, Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP); Instituto de Investigação e Inovação em Saúde (i3S).

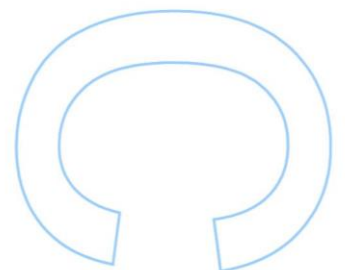
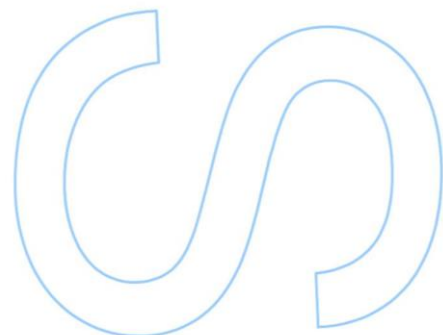
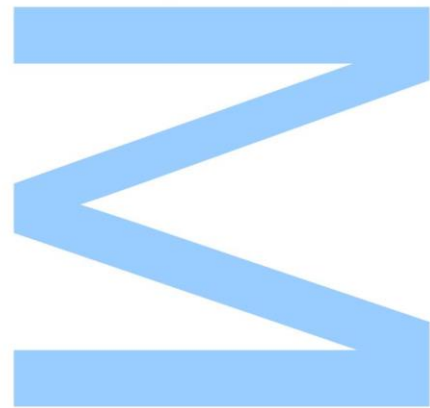




Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, \_\_\_\_ / \_\_\_\_ / \_\_\_\_



# Agradecimentos

Esta dissertação de mestrado representa o fim de uma etapa muito importante na minha vida, tanto a nível pessoal como profissional. Se hoje, a realização deste estudo foi possível, não é apenas graças ao meu esforço individual, mas também ao de muitas outras pessoas a quem gostaria de agradecer quer pela paciência, pela ajuda, pela orientação e, principalmente, pelos conhecimentos transmitidos.

Ao Professor António Amorim, pela oportunidade que me deu, ao permitir a minha passagem pelo seu grupo de investigação.

À Doutora Alexandra Lopes, um agradecimento muito especial, pois independentemente do ano atípico que vivemos, orientou-me sempre da melhor forma possível, integrando-me na sua equipa, acreditando nas minhas capacidades e transmitindo-me o seu conhecimento e aprendizagens, tanto práticos como teóricos. Ficar-lhe-ei eternamente grata, pois a profissional que sou hoje é muito graças a este ano em que tive a oportunidade de trabalhar consigo e acompanhar de perto o dia-a-dia da investigação.

A todos os investigadores, bolseiros, doutorandos e colegas do grupo de Genética Populacional e Evolução (GEPO) do IPATIMUP, pela forma como me receberam, acolheram e ajudaram ao longo deste ano. Em especial, às minhas colegas do Mestrado de Genética Forense e às minhas parceiras de almoço, pela amizade e pelos momentos de animação e descontração que terminavam em debates ideológicos bastante enriquecedores.

Por fim, mas não menos importante, quero agradecer a toda a minha família pela educação, amor e carinho que me deram e por acreditarem sempre em mim e que tudo isto era possível. Aos meus pais, um enorme bem-haja, porque sem vocês nada disto era possível. Sem vocês eu não era a mulher que sou hoje e não estava prestes a concluir esta fase tão importante da minha vida. Obrigada pela inspiração que são para mim por me apoiarem sempre em toda e qualquer decisão que tome.

E porque os amigos, são a família que escolhemos, um muito obrigada também aos meus, que são os melhores do Mundo. Em especial, à Sofia Figueiredo que, para além dos mais de 10 anos de amizade, me acompanhou diariamente ao longo deste ano de forma incansável apoiando tanto a nível profissional como pessoal.

Um enorme bem-haja a todos. A pessoa que sou hoje, devo-o a todos vós.

## Abstract

Male infertility affects about 7% of the male population. The most extreme phenotypes are azoospermia (lack of sperm in the ejaculate) and severe oligozoospermia (very low sperm counts). Although some causes are already known, about 50% of the cases remain unexplained and are catalogued as idiopathic infertility.

We selected five SNPs associated with family size in a previous GWAS study in a Hutterites population, and genotyped them in a cohort of Portuguese men with idiopathic severe oligozoospermia (n=163), azoospermia (n=219) and geographically matched normozoospermic controls (n=127), using TaqMan allelic discrimination assays. The selected SNPs were rs7174015, an intronic variant of *Ubiquitin Specific Peptidase 8 (USP8)*; rs12870438, a variant in an intron of *Epithelial Stromal Interaction Protein 1 (EPSTI1)*; rs7867029, downstream of the *Phosphoserine Aminotransferase 1 (PSAT1)* gene; rs10966811, located between *IZUMO family member 3 (IZUMO3)* and the *Tumor Suppressor Candidate 1 (TUSC1)* gene and rs10129954 in an intron of the *Double PHD Fingers 3 (DPF3)* gene.

Even though no significant associations were obtained in this study, we observed a trend towards association for three of the analyzed SNPs and a nominal association for one of them. Indeed, for USP8-rs7174015, assuming a recessive model, a nominally significant association with azoospermia was detected (OR=1.88; P=0.020) for the A allele, even though it did not pass the threshold of significance after FDR (False Discovery Rate) correction. Assuming a recessive model, there was also a borderline tendency for association as a protective factor for severe oligozoospermia (OR=0.5; nominal p=0.055) for the A allele of the intronic SNP at *EPSTI1* (rs12870438). The C allele of PSAT1-rs7867029 shows a non-significant tendency (OR=0.6; nominal p=0.078) towards a protective factor for severe oligozoospermia, assuming a dominant model. The *in silico* functional analysis showed an eQTL effect in testis for USP8-rs7174015 and some of its proxies. All the variants showing a trend towards association were located in LD blocks in genomic regions likely relevant for regulatory processes related with spermatogenesis. Overall, these results suggest that genetic susceptibility to the studied phenotypes conferred by these variants is likely conferred by deregulation of gene expression.

## Keywords

Idiopathic Infertility, spermatogenic failure, genetic association study, SNPs, *USP8*, *EPSTI1*, *PSAT1*.

## Resumo

A infertilidade masculina afeta cerca de 7% da população masculina. Os fenótipos mais extremos são azoospermia (ausência total de espermatozoides no ejaculado) e oligozoospermia grave (contagens de espermatozoides muito baixas). Embora algumas causas já sejam conhecidas, cerca de 50% dos casos permanecem inexplicados, sendo classificados como infertilidade idiopática.

Neste estudo foram selecionados cinco SNPs, associados anteriormente com número de descendentes num estudo GWAS numa população de Hutterites. Através de ensaios de discriminação alélica por TaqMan, foram genotipados homens da população portuguesa com fenótipos de oligozoospermia severa (n=163) e azoospermia (n=219). Como controlos foram também genotipados normozoospermicos (n=127) da mesma população. Os SNPs selecionados foram rs7174015, uma variante intrónica do gene *Ubiquitin Specific Peptidase 8 (USP8)*; rs12870438, uma variante localizada num intrão do gene *Epithelial Stromal Interaction Protein 1 (EPSTI1)*; rs7867029, a jusante do gene *Phosphoserine Aminotransferase 1 (PSAT1)*; rs10966811, localizado entre o gene *IZUMO family member 3 (IZUMO3)* e o gene *Tumor Suppressor Candidate 1 (TUSC1)* e rs10129954 num intrão do gene *Double PHD Fingers 3 (DPF3)*.

Apesar de não ter sido obtida nenhuma associação significativa neste estudo, observámos uma tendência para associação em 3 SNPs, sendo uma delas uma associação nominal. Então, para USP8-rs7174015, assumindo um modelo recessivo, foi detetada uma associação nominalmente significativa com a azoospermia (OR = 1,88; P = 0,020) para o alelo A, embora esta não tenha passado o limite de significância após a correção de FDR (*False discovery rate*). Assumindo um modelo recessivo, também foi notada uma tendência para associação, como fator de proteção para oligozoospermia severa, (OR = 0,5; p nominal = 0,055), para o alelo A do SNP intrónico em *EPSTI1* (rs12870438). O alelo C de PSAT1-rs7867029 mostra uma tendência não significativa (OR = 0,6; p nominal = 0,078) como fator de proteção em oligozoospermia grave, assumindo um modelo dominante. A análise funcional *in silico* mostrou um efeito eQTL em testículo para o USP8-rs7174015 e para alguns dos SNPs associados (*proxies*). Todas as variantes com tendência para associação estão localizadas em blocos de LD em regiões genómicas provavelmente relevante para processos de regulação relacionados com a espermatogénese. De uma forma global os nossos resultados sugerem que a suscetibilidade genética para os fenótipos estudados conferida por estes *loci* é provavelmente conferida pela desregulação da expressão génica.

## Palavras – Chave

Infertilidade idiopática, espermatogénese, estudo de associação genética, SNPs, gene *USP8*, gene *EPST11*, gene *PSAT1*

# Contents

<b>Agradecimentos</b> .....	<b>i</b>
<b>Abstract</b> .....	<b>ii</b>
<b>Resumo</b> .....	<b>iii</b>
<b>Index of Tables</b> .....	<b>vi</b>
<b>Index of Figures</b> .....	<b>vii</b>
<b>Abbreviations</b> .....	<b>viii</b>
<b>Introduction</b> .....	<b>1</b>
1. Infertility .....	1
1.1 Male infertility .....	1
1.1.1 Spermatogenesis .....	2
1.1.2 Causes .....	4
1.1.3 Azoospermia and Severe Oligozoospermia .....	5
2. Genomic Variation .....	6
2.1 Single nucleotide polymorphisms (SNPs) .....	7
3. GWAS overview .....	9
<b>Aim</b> .....	<b>11</b>
<b>Material and Methods</b> .....	<b>12</b>
1. Study design and study population .....	12
2. SNP selection and genotyping .....	12
2.1 Sequence to obtain individuals of reference .....	13
2.2 Genotyping .....	18
3. Statistical analysis .....	19
4. <i>In silico</i> characterization .....	20
<b>Results</b> .....	<b>21</b>
Genotyping quality control .....	21
Genotypes and association tests .....	22
Evaluation of functional annotations .....	24
<b>Discussion</b> .....	<b>31</b>
<b>Conclusion</b> .....	<b>35</b>
<b>Bibliography</b> .....	<b>36</b>
<b>Supplementary Materials</b> .....	<b>42</b>
<b>Appendix</b> .....	<b>59</b>

## Index of Tables

<b>Table 1 - Reference values, defined by WHO, used to evaluate spermograms.....</b>	<b>5</b>
<b>Table 2 - Single nucleotide polymorphisms characteristics.....</b>	<b>7</b>
<b>Table 3 - Characteristics of the five candidate SNPs. ....</b>	<b>13</b>
<b>Table 4 - PCR Master Mix protocol .....</b>	<b>14</b>
<b>Table 5 - Standard thermocycling conditions that allow the amplification of the sequences of USP8-rs7174015, PSAT1-rs7867029 and TUSC1-rs10966811.....</b>	<b>14</b>
<b>Table 6 - Thermocycling conditions that allow the amplification of the sequences of EPSTI1-rs12870438 and DPF3-rs10129954 .....</b>	<b>15</b>
<b>Table 7 - Silver Staining protocol. ....</b>	<b>16</b>
<b>Table 8 - 5ExoSap Reaction thermocycling conditions. ....</b>	<b>16</b>
<b>Table 9 - Cycle sequencing thermocycling conditions. ....</b>	<b>17</b>
<b>Table 10 - Sephadex purification protocol .....</b>	<b>17</b>
<b>Table 11 - Real Time - PCR thermocycling conditions .....</b>	<b>19</b>
<b>Table 12 - Positive controls used for each SNP. ....</b>	<b>21</b>
<b>Table 13 - Genotype and Allele Frequencies in different study groups.....</b>	<b>22</b>
<b>Table 14 - Results of the association tests derived from different comparative genetic models (additive, dominant and recessive).....</b>	<b>23</b>
<b>Table 15 - Proxies of UPS8-rs7174015 obtained by LDLink, using as a threshold <math>R^2 &gt; 0.8</math>, and all the detailed information about them .....</b>	<b>24</b>
<b>Table 16 - Proxy of PSAT1-rs7867029 obtained by LDLink, using as a threshold <math>R^2 &gt; 0.8</math>, and all the detailed information about it.....</b>	<b>25</b>
<b>Table 17 - Proxies of EPSTI1-rs12870438 obtained by LDLink, using as a threshold <math>R^2 &gt; 0.8</math>, and all the detailed information about them.....</b>	<b>25</b>
<b>Table 18- Haploreg v4.1 annotations about the transcription factor binding sites changed.....</b>	<b>29</b>
<b>Table 19 - Annotations of regulatory marks - H3K4me3 and CTCF - in the testis, mapped overlapping the USP8-rs7174015 and their proxies.....</b>	<b>30</b>
<b>Table 20 - Annotations of regulatory marks - H3K4me3 and CTCF - in the testis, overlapping EPSTI1-rs12870438 and their proxies.....</b>	<b>30</b>



## Index of Figures

<b>Figure 1 - Spermatogenesis.....</b>	<b>3</b>
<b>Figure 2 - Candidate gene vs GWAS characterization. ....</b>	<b>8</b>
<b>Figure 3 - TaqMan Genotyping Assay. ....</b>	<b>18</b>
<b>Figure 4 - Allelic Discrimination Plot for each SNP. ....</b>	<b>22</b>
<b>Figure 5 - Functional annotations of the human genome for the lead SNP variant USP8-rs7174015 and its proxies.....</b>	<b>Erro! Marcador não definido.</b>
<b>Figure 6 - Gene expression in human testicular cells.....</b>	<b>27</b>
<b>Figure 7 - Predictive functional analysis for the EPST11-rs12870438 and its proxies. .....</b>	<b>Erro! Marcador não definido.</b>

# Abbreviations

**A** – Adenine

**C** – Cytosine

**CI** – Confidence Interval

**CNV** – Copy number variation

**ddNTP** - Dideoxynucleotide triphosphate

**DNA** – Deoxyribonucleic acid

**dNTP** – deoxyribonucleotide triphosphate

**DPF3** – Double PHD Fingers 3 gene

**EPST11** – Epithelial Stromal Interaction Protein 1 gene

**eQTL** – Expression quantitative trait loci

**FDR** – False discovery rate

**G** – Guanine

**GC** – Guanine/ Cytosine

**GWAS** – Genome wide association studies

**HLA** – Human leukocyte antigen

**HWE** – Hardy – Weinberg Equilibrium

**IBS** – Iberian Population

**IZUMO3** – IZUMO family member 3 gene

**kbp** – Kilobase pair

**MAF** – Minor allele frequency

**NOA** – Non-obstructive azoospermia

**OA** – Obstructive azoospermia

**OR** – *Odds Ratio*

**PCR** – Polymerase Chain Reaction

**PSAT1** – Phosphoserine Aminotransferase 1 gene

**SNP** – Single nucleotide polymorphism

**SO** – Severe Oligozoospermia

**spz** – spermatozoa

**sQTL** – Splicing quantitative trait loci

**T** - Thymine

**TF** – Transcription factor

**TFBS** - Transcription factor binding site

**Tm** - Melting temperature

**TUSC1** – Tumor suppressor candidate 1 gene

**USP8** – Ubiquitin Specific Peptidase 8 gene

**WHO** - World Health Organization

# Introduction

## 1. Infertility

Fertility is defined as the capacity to establish a clinical pregnancy<sup>1</sup>. However, when passing 12 months or more of regular unprotected sexual intercourse and the couple cannot achieve a natural clinical pregnancy, a disease of the reproductive system can be present, leading to infertility<sup>2</sup>. The term subfertility can be used interchangeably with infertility and it is defined as any form or grade of reduced fertility in couples unsuccessfully trying to conceive<sup>1</sup>.

It is estimated that infertility affects 10-15% of couples of reproductive age worldwide<sup>3,4</sup>. However, this percentage is based on an amalgamation of numbers taken from around the world, not necessarily reflecting the rates of specific countries and regions. Considering the huge diversity of cultures and family traditions around the globe, it is very difficult to calculate the real value of the infertility rate<sup>5</sup>.

The inability to conceive affects men and women across the globe, being these concepts applied to both genders. Hereafter, the further work will be focused on male infertility.

### 1.1 Male infertility

Male infertility is a disorder that affects about 7% of the male population<sup>6,7,8</sup>. Based on the latest WHO statistics worldwide, there are approximately 50-80 million people suffering from infertility<sup>9</sup>. According to several studies, approximately 20-30% of the cases are caused only by a male factor, and 20-30% are caused by both – male and female factors<sup>9,10</sup>. However, these numbers are not consensual, since recent studies argue that a male factor was present in 20-70% of the cases under consideration. This wide range was obtained by meta-analysis excluding factors such as cultural constraints and other statistical accuracy problems<sup>5</sup>.

The reports of male infertility cases are, generally, not correct, specifically in countries where cultural differences and patriarchal societies manipulate the collection and compilation of the accurate statistics. In some regions, like Northern Africa and Middle East, the couple study is not conducted and the cause is frequently attributed to the female partner, leading to an underreporting of male infertility<sup>5</sup>. Another factor that sparse the statistics is the fact that male infertility is never defined as a disease. Thus, there is no consensus about the studies carried out. Some studies examine only males, while others examine females. However, the group sizes are not equal, being most of the male studies conducted with a small group, not representative of the larger infertile

male population<sup>5,10</sup>. The last factor - and maybe the most important - is the couples who are considered in the studies. Usually, the population in study are couples who have unprotected intercourse and the desire to have a child. This is a very specific population that does not reflect the general population<sup>5</sup>.

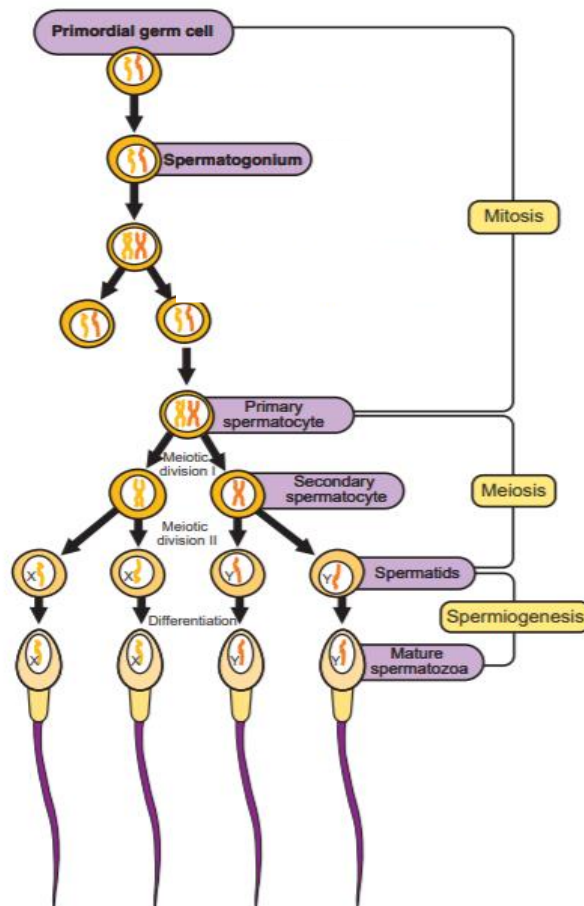
### 1.1.1 Spermatogenesis

The formation of male gametes (spermatozoa) occurs within testes through a process called spermatogenesis<sup>11</sup>. Spermatogenesis is a highly orchestrated process requiring a stem cell pool, a period of amplification of cell numbers, the completion of reduction division to haploid cells, and the morphological transformation of the haploid cells into spermatozoa<sup>12</sup> (**Figure 1**). It begins shortly before puberty and continues throughout life, with a slight decline during old age<sup>13</sup>.

The formation of sperm is a continuous process during the reproductive lifetime. It can be divided into three main steps: (1) Differentiation of spermatogonial stem cells into spermatocytes via mitotic cell division; (2) Production of haploid spermatids from tetraploid primary spermatocytes, via meiotic cell division - when the chromosomal division occurs and the haploid stage is obtained; (3) Spermiogenesis - the point at which spermatids give rise to spermatozoa<sup>14</sup>.

Spermatogonia (the first cells of spermatogenesis) are the most immature germ cells. These cells, during embryonic development, migrate to the testes and adhere to the upper surface of seminiferous tubes. However, after puberty, these cells migrate between the Sertoli cells and the central lumen of seminiferous tubes<sup>15</sup>. In the lumen of seminiferous tubes, Sertoli cells surround type A spermatogonia with cytoplasmic processes and, through to a series of mitotic divisions, they later become mature spermatogonia (type B cell). Type B spermatogonia start the prophase I of meiosis giving rise to primary spermatocytes that enter the first meiotic division leading to two secondary spermatocytes<sup>11</sup>. During the long meiotic prophase, meiotic recombination and homologous chromosome pairing occurs<sup>15</sup>. Primary spermatocytes are subclassified according to the stages of prophase 1 (i.e. preleptotene, leptotene, zygotene, and pachytene)<sup>13</sup>. In turn, the secondary meiotic division begins, leading to the division of the spermatocyte into two spermatids and each of them gets one of the homologous chromosomes (each spermatid has now 22 autosomes and one sexual chromosome - X or Y)<sup>11</sup>. During spermiogenesis, the spermatid undergoes a series of cytological events that lead to the formation of mature sperm. This final process is subdivided into four steps:

- Step 1 - occurrence of nuclear condensation and movement of nucleus to the periphery of the cell;
- Step 2 - formation of a modified lysosome, known as an acrosome and located near the cell membrane;
- Step 3 – formation of a flagellum, including the development of the axoneme;
- Step 4 – elimination of the cytoplasm by Sertoli cells<sup>11</sup>.



**Figure 1 - Spermatogenesis.** Representation of the three main steps – cellular proliferation by mitosis, two reduction divisions by meiosis and cell differentiation by spermiogenesis – until the formation of mature spermatozoa (Adapted from Rhoades, 2013).

These crucial steps are regulated by hormones and growth factors. For successful mature spermatozoa formation, there must be a correct interplay of endocrine factors within the hypothalamic-pituitary-gonadal axis and of autocrine, paracrine and juxtacrine interactions between the spermatogenic germ cells within the seminiferous tubules and the somatic cells that reside inside (Sertoli cells), between (Leydig and other interstitial cells) and within the wall (myoid cells) of the tubules, as well as of factors in the epididymis<sup>13,16</sup>. The time required to produce mature spermatozoa from the earliest stage of spermatogonia is 65 to 70 days<sup>13</sup>.

## 1.1.2 Causes

There are several possible causes underlying male infertility, such as anatomical, physiological and genetic. However, many environmental and acquired factors can also influence fertility, leading to infertility. Although some causes are already known, a substantial fraction of the cases of male infertility remain unexplained – idiopathic infertility. Male reproductive impairment may result from factors that affect sperm production, quality, function, or transport<sup>17</sup>.

The most common anatomical causes are cryptorchidism and varicocele. Cryptorchidism is the failure of one or both testes to permanently descend (also defined as undescended testis) is reported in 1% to 9% of new-borns<sup>18</sup> and adult infertility is a well-known long-term consequence<sup>19</sup>. Varicocele refers to the dilatation of testicular veins in the scrotal portion of the pampiniform plexus, causing a progressive decline of fertility<sup>20</sup>. The effect on the testicular function is variable but varicocele is identified in 35% of men with primary infertility and 81% of secondary infertility patients<sup>19,20</sup>. Among men presenting azoospermia or severe oligospermia, varicocele is detected in 4.3% to 13.3%<sup>21</sup>.

The diagnostic of male infertility is based on careful examination of medical and reproductive history. All the risk factors and behavioural patterns that affect fertility are considered. A physical examination is also performed, taking into account the secondary sex characteristics and the genitalia, and finally the spermogram is analysed<sup>22,23</sup>.

Sperm parameters are widely used to estimate the potential fertility of men. The World Health Organization (WHO) defined a standardised method for assessment of human semen that includes also the reference values that are expected in healthy men<sup>24</sup>.

When a value falls outside of the defined reference interval for spermogram parameters it only indicates the presence of an underlying pathology but it does not identify the aetiology of the condition<sup>24</sup>. In **Table 1**, it is possible to verify the reference values used by most laboratories that perform semen analysis<sup>17</sup>. The sperm count is the most important parameter in the evaluation, where the most extreme phenotypes are azoospermia (lack of sperm in the ejaculate due to non-obstructive causes) and severe oligozoospermia (very low concentration of spermatozoa in semen). These two severe manifestations of male infertility have an important genetic component<sup>17</sup> and include known genetic risk factors such as congenital genetic anomalies, Y chromosome microdeletions, karyotype abnormalities and human genome variants, single nucleotide polymorphisms (SNPs) and copy number variants (CNVs)<sup>25</sup>.

**Table 1 - Reference values, defined by WHO, used to evaluate spermograms.**

<b>Semen Parameters</b>	<b>Reference Values</b>
Semen volume	>1.5ml
Sperm concentration	>15 x10 <sup>6</sup> spz/ml
Total sperm number	>39 x10 <sup>6</sup> spz /ejaculate
Total motility	>40%
Progressive motility	>32%
Vitality	>58%
Normal morphology	>4%

### 1.1.3 Azoospermia and Severe Oligozoospermia

Azoospermia and Severe Oligozoospermia are the most extreme phenotypes of male infertility. However, inside these categories it is possible to divide clinically the cases due to obstructive or non-obstructive causes, the latter being the most severe. Severe defects in sperm production resulting in oligozoospermia are the main or contributing factor in up to a fifth of infertile couples, occurring in about 3-4% of men<sup>26</sup>. The prevalence of azoospermia is approximately 1% among all men, ranging between 10% and 15% among infertile men<sup>23</sup>. The causes of azoospermia can be divided into three main categories:

(1) Pre-testicular azoospermia, also termed secondary testicular failure, is related to endocrine abnormalities having adverse effects on spermatogenesis, where the hypothalamic – pituitary – adrenal axis is not performing properly the intrinsic stimulation of normal testes for its two gonadal functions – testosterone synthesis (Leydig cells) and spermatogenesis (within the seminiferous tubules)<sup>23,24,27</sup>. Examples of causes of pre-testicular azoospermia are hypogonadotropic hypogonadism and exogenous androgens<sup>27</sup>.

(2) Post-testicular azoospermia (or obstructive azoospermia (OA)), related to ejaculatory dysfunction or ductal obstructions, can be identified in approximately 40% of affected men<sup>27</sup>. This results of a blockage to sperm flow independent of the transport system, preventing sperm from reaching the urethral meatus<sup>23,24,27</sup>. Vasectomy, epididymal obstruction, ejaculatory duct obstruction and ejaculatory obstruction are some examples that can be included in this group<sup>27</sup>.

(3) Testicular azoospermia (non – obstructive azoospermia (NOA)), also termed primary testicular failure, encompasses spermatogenesis disorders intrinsic to the testes<sup>23,27</sup>. In these cases, the hypothalamic – pituitary – adrenal axis is functioning normally, but there is a defect in spermatogenesis reflected in the absence of sperm production<sup>24</sup>.

The variety of causes of spermatogenic failure are often classified as congenital or acquired. Among the acquired causes cryptorchidism, varicocele, chemo or radiotherapy to the testis<sup>28,19</sup>, infection or inflammation of the testis<sup>19</sup>, neoplasia of the testis<sup>19</sup>, historic of orchitis (mumps)<sup>28,19</sup>, delayed puberty<sup>28</sup>, damage to the blood supply of the testis<sup>19</sup>, testosterone replacement therapy<sup>28</sup>, anabolic steroid abuse<sup>28</sup> and exposure to environmental or occupational radiation or toxins like heavy metals and pesticides<sup>28,19</sup> are included. The genetic causes are classified as congenital and include sex chromosomal abnormalities like Klinefelter syndrome, Y chromosomal microdeletions, structural chromosomal defects<sup>19</sup> or other genetic abnormalities. Structural or numerical chromosomal abnormalities are identified in 11% to 14% of men with non-obstructive azoospermia or severe oligospermia<sup>19</sup>. Although a complete description of the testis-specific genes that direct maintenance of spermatogenic cells and adequate completion of meiosis in the adult man remains to be elucidated, it is widely accepted that several genetic abnormalities are responsible for impaired sperm production and have been associated with spermatogenic failure<sup>19</sup>.

## 2. Genomic variation

The human genome sequence is a complete DNA sequence of anatomically modern humans (*Homo sapiens sapiens*). How humans look and behave as well as their risk to develop certain diseases are determined by hundreds of complex phenotypic traits that are based on dozens to hundreds of gene variants and environmental influences<sup>29</sup>. The coding sequence covers less than 2% of the human genome, i.e. the vast majority of the genome is non-coding with a likely primary regulatory function<sup>29</sup>.

Genomic variation is a crucial biological determinant underpinning evolution and defining the heritable basis of phenotypes<sup>30</sup>, based on differences between genomes. A holy grail of genetics is to identify how genotypic differences in the 0.1% of the genome affect observable traits of individuals or their phenotypes, since the genomes of any two individuals are 99.9% identical<sup>31</sup>. Every individual carries approximately four million genetic variants that cover about 0.3% of all the sequence<sup>29</sup>. The most abundant source of genetic variation in the human genome is represented by single nucleotide polymorphisms (SNPs), which can account for heritable inter-individual differences in complex phenotypes<sup>32</sup>. However, there are other forms of variation like tandem repeats, small insertions or deletions and large copy number variants (CNVs)<sup>31,30</sup>. Genetics variants are defined as polymorphisms (or common variants) if their minor allele frequency (MAF) is at least 1% in the studied population or rare when their MAF is less than 1%<sup>29,30</sup>.



## 2.1 Single Nucleotide Polymorphisms (SNPs)

SNPs are defined as a nucleotide position along a chromosome where the DNA of different people may vary<sup>33</sup>. More often, two alternative alleles are found at a particular polymorphic site but SNPs can also be triallelic or quadriallelic<sup>34</sup>. A working definition of a SNP can be a single base change in genomic DNA where sequence alternatives exist in normal individuals and the least frequent allele has a frequency of at least 1%<sup>35</sup>.

The human genome contains nearly 11 million SNPs<sup>34</sup>, of which approximately 4 million are found at a frequency of 1-5% in human populations. The remaining 7 million present a MAF higher than 5%<sup>29</sup>. The most important characteristics of SNPs are: i) they are present across the entire human genome, within genes and in intergenic regions; ii) they represent the most common type of variants; iii) they are generally biallelic; iv) they are easily genotyped by automated technologies; v) large part of them have direct repercussions on human disease which allows the identification of individuals genetically susceptible to develop a number of multifactorial diseases, and in some cases predict the severity of the disease as well as the activity and response to medications<sup>34</sup>. SNPs are also easy and cheap to assay, making them a tool of choice for human genetics studies<sup>30</sup> (**Table 2**).

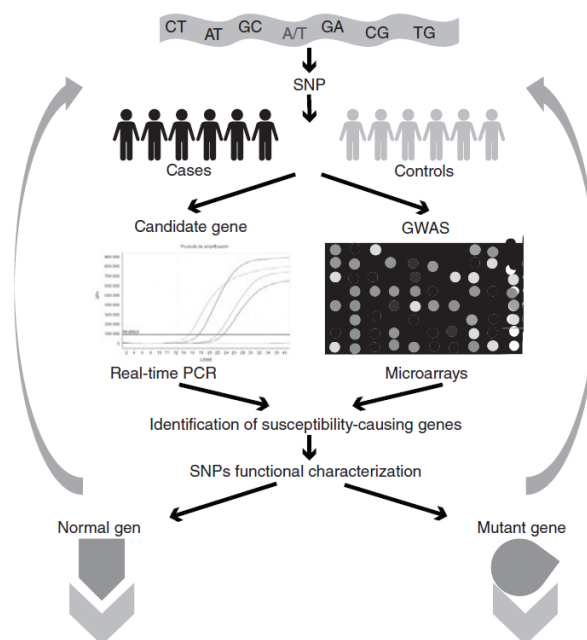
**Table 2 - Single nucleotide polymorphisms characteristics** (Adapted from Ramirez-Bello et al., 2017).

<b>Characteristics</b>	<b>Description</b>
<b>Distribution</b>	- Approximately 11 million SNPs have been reported, which reflects, on average, one SNP per each 250 bp
<b>Location</b>	- Intergenic or intra-genic regions (both protein-coding and non-coding genes included)
<b>Allele number</b>	- Generally biallelic, but they can be triallelic, quadriallelic or tetrallelic (the last three are rare in humans)
<b>Biological effect</b>	- Neutral or functional
<b>SNPs assessment</b>	- Easy and cheap to genotype by automated technologies
<b>Uses in health</b>	- Identification of individuals genetically susceptible to develop multifactorial diseases, severity, activity and response to medications

The probability of a SNP influencing a given phenotype depends on where it occurs and the nature of the change it induces. SNPs can be categorised according to their location in the genome, whether they are found in coding regions, non-coding regions or in intergenic regions, being the majority of these variants found in non-coding regions<sup>34,29</sup>. In this context it is important to introduce another concept, linkage

disequilibrium, which is the non-random association of alleles at different *loci* in the genome. Closely spaced variants tend to be inherited together, and this has to be taken into account when using SNPs to study human disease. In fact, given any two nearby variants, one may have a direct effect on the disease under study while a proximal marker may not have any functional importance.

Association studies between genome variants and complex traits are important not only to identify genetic variation which influences disease risk but also to better understand the underlying disease mechanisms. By means of candidate gene or genome-wide association studies (GWAS) (**Figure 2**), genes involved with several complex diseases have been identified<sup>34</sup>.



**Figure 2 - Candidate gene vs. GWAS characterization.** Candidate gene or genome-wide association studies (GWAS) have enormously contributed to identify loci involved with risk for the development of different multifactorial human diseases. Molecular genetics and biochemical studies contribute to identify the functional effect of the less common alleles of the SNPs, being studies that can range from candidate gene to functional identification or vice versa<sup>34</sup>.

Both candidate gene studies and GWAS are usually based on SNP detection and the design are a case-control study. All the abnormal phenotypes were defined as cases and the normal as controls. The aim is to ascertain if there are significant differences between the two populations, allowing to define a statistical association (positive or negative) with the studied trait. The candidate gene approach begins with the selection of a putative candidate gene based on its relevance in the mechanism of the disease/trait in investigation<sup>34</sup> and real time PCR is the most common technique used nowadays in this type of study, in order to obtain the genotypes for several SNPs in the whole cohort<sup>34</sup>.

On the other hand, GWAS are based on microarrays that allow the analysis of SNPs distributed across the whole genome of an individual and are easily scalable to a large cohort. Based on the genotypes, researchers search statistical differences between the genomes of individuals carrying a particular phenotype (cases) and those that do not (controls)<sup>31</sup>. Both study designs have the aim to identify genes involved in a given disease and susceptibility-causing variants. The functional characterization of the associated alleles at each SNP is an important step to better understand the impact of these variants.

### 3. GWAS overview

Approximately half of male factor infertility cases have no known cause. However, it is likely that most idiopathic male factor infertility cases have some unidentified genetic basis<sup>8</sup>. Several SNPs have already been associated with male infertility. In 2009, Aston and Carrell performed a pilot GWAS including 80 controls and 92 azoospermic and severe oligozoospermic men, resulting in significant associations ( $p > 1 \times 10^{-5}$ ) with 21 SNPs: 4 associated with severe oligozoospermia, 16 with azoospermia and 1 with both phenotypes combined as well as with azoospermia alone<sup>36</sup>. As a follow-up, in 2010, Aston *et al.* used a targeted medium throughput approach to evaluate 172 polymorphisms in a larger cohort of infertile men and in controls – 158 controls (normozoospermic men) and 221 cases (141 SO and 80 NOA men). In this case, 9 marginally significant SNPs were associated with the studied phenotypes, being 4 associated with SO, 2 with NOA and 3 with the combination of both phenotypes<sup>22</sup>. One year later, Hu *et al.* published a study in a Chinese cohort of 1,000 individuals with NOA and 1,703 male controls that resulted in strong associations with three SNPs<sup>37</sup>. Additionally, three SNPs within the HLA (Human leukocyte antigen) region were associated with the risk of non-obstructive azoospermia in a three stage GWAS in a Chinese population<sup>38</sup>.

In 2012, Kosova *et al.* conducted a genome wide association study of two fertility traits – family size and birth rate – in 269 married men who were members of Hutterite population (founder population of European descent that proscribes contraception and has large family sizes)<sup>39</sup>. The Hutterites are among the most fertile human populations with relatively few childless couples<sup>40</sup>. They obtained 41 associated SNPs ( $p \leq 1 \times 10^{-4}$ ) and, in order to validate the most significant associations in the Hutterites and to assess the functional or clinical relevance of associated loci, a validation study in ethnically diverse men from Chicago was undertaken. In the latter study, nine significantly associated SNPs were obtained – five were associated with family size (USP8-

rs7174015, PSAT1-rs7867029, EPST11-rs12870438, DPF3-rs10129954 and TUSC1-rs10966811) and four with birth rate (rs680730, rs11236909, rs10488786 and rs724078) – but all were associated with some sperm parameter<sup>39</sup>. The selection of the candidate SNPs for this dissertation were based on the results obtained in the latter paper.

## Aim

The aim of this study was to evaluate whether five single nucleotide polymorphisms (SNPs) previously associated with family size in a cohort of Hutterite men (USP8-rs7174015, DPF3-rs10129954, EPSTI1-rs12870438, PSAT1-rs7867029 and TUSC1-rs10966811) confer risk for severe spermatogenic failure, both azoospermia and severe oligozoospermia, in the Portuguese population. To achieve this aim, an association study was carried out, with a cohort of azoospermic (n=219) and severe oligozoospermic men (n=163), having as controls normozoospermic men of matched geographic origin (n=127). Once the associations were tested, functional *in silico* analysis was carried out to ascertain how the studied variants may influence spermatogenesis.

The work developed under the scope of this dissertation was part of a larger collaborative multi-center association study in a large European cohort of infertile men (Portuguese and Spanish), led by Prof. F. David Carmona (Departamento de Genética e Instituto de Biotecnología, Universidad de Granada, Spain; Instituto de Investigación Biosanitaria IBS.Granada, Granada, Spain). The results for the Portuguese population were obtained in the course of this thesis, under the supervision of Dr. Alexandra M. Lopes (IPATIMUP/I3S). The manuscript describing this study was submitted to "Human Reproduction" journal in August 2020 and can be consulted in the Appendix, as well as the poster accepted for the scientific meeting "19<sup>th</sup> Portugaliae Genetica" (postponed to 2021 due to the COVID-19 pandemic), presenting all the preliminary results of this work.

# Material and Methods

## 1. Study Design and Study Population

According to a study in a Hutterite Population, five SNPs (USP8-rs7174015, rs10129953, EPST11-rs12870438, PSAT1-rs7867029 and TUSC1-rs10966811) are associated, positively or negatively, with family size. These SNPs were also associated with spermogram parameters in Chicago men<sup>39</sup>. To validate and better understand these associations, we conducted a case-control study in a Portuguese cohort of infertile men. This work is part of a larger collaborative association study performed in a well-powered cohort of Iberian (Portuguese and Spanish) infertile men and controls.

The total number of analysed cases was 382 (163 Severe Oligozoospermia and 219 Azoospermia cases) and the controls were 127 normozoospermic men. The samples were obtained from the National Institute of Health Doutor Ricardo Jorge, in Lisbon, and from the Serviço de Genética, Pathology Department of the Faculty of Medicine of the University of Porto. Informed written consent was signed by all participants and the approval from the local ethical committees of all participating centres was obtained in accordance with the Declaration of Helsinki.

Case classification in NOA (complete absence of sperm in the ejaculate for non-obstructive causes) or SO (< 5 million spermatozoa/ ml of semen) was performed according to the guidelines of the World Health Organization, that consists in two high-speed centrifugation processes in two different semen samples<sup>41</sup>. Only men with idiopathic infertility were included. All the participants that already had a diagnostic of abnormal karyotype, Yq deletions or testicular disorders associated with the male infertility were excluded.

The statistical analysis were performed with all genotyping results, however the genotypes of the azoospermic men and some normozoospermic men were obtained by Miriam Cérvan-Martín at Instituto de Investigación Biosanitaria de Granada (ibs.GRANADA), following the same/a similar protocol.

## 2. SNP selection and genotyping

To investigate genetic variants associated with idiopathic severe oligozoospermia and azoospermia, five candidate SNPs were selected from GWAS results in European descendent populations. The selected SNPs were rs7174015, an intronic variant of *Ubiquitin Specific Peptidase 8 (USP8)*, EPST11-rs12870438, a variant in an intron of *Epithelial Stromal Interaction Protein 1 (EPSTI1)*, rs7867029 localized downstream the *Phosphoserine Aminotransferase 1 (PSAT1)*, rs10966811 between *IZUMO family*

*member 3 (IZUMO3)* and *Tumor Suppressor Candidate 1 (TUSC1)* and DPF3-rs10129954 in an intron of *Double PHD Fingers 3 (DPF3)*. All these polymorphisms are common (MAF>1%) and have been already associated with family size in a Hutterite population and with spermogram parameters in a population of men from Chicago<sup>39</sup>, as shown in **Table 3**.

**Table 3 - Characteristics of the five candidate SNPs.**

Chr	SNP	Position	Closest Gene	Location	Distance (kbp)	Allele <sup>a</sup>	IBS MAF	Previous GWAS studies	
								Associated with family size	Associated with spermogram parameters
9	rs10966811	25,233,486	<i>TUSC1</i>	downstr.	442.903	A/G	0.41	A (risk factor)	Beat frequency
			<i>IZUMO3</i>	upstr.	687.837				
9	rs7867029	78,405,502	<i>PSAT1</i>	downstr.	75.409	C/G	0.18	C (protective factor)	Conc; % Motility
13	rs12870438	4,290,669	<i>EPST11</i>	intron		A/G	0.35	A (protective factor)	Conc; Total sperm count; Avg veloc; Mean ALH
14	rs10129954	72,683,993	<i>DPF3</i>	intron		T/C	0.48	T (risk factor)	Total motile sperm count; Linearity; Beat frequency
15	rs7174015	50,424,871	<i>USP8</i>	intron		A/G	0.47	A (risk factor)	Volume; Total sperm count; Avg veloc; Mean ALH

Associations already demonstrated in previous GWAS studies are indicated. The family size study was done in a Hutterite cohort while spermogram parameters associations were demonstrated in a population of Chicago men

Legend: Chr – chromosome; downstr – downstream nearest gene; upstr – upstream the nearest gene; IBS MAF - Minor Allele Frequency in Iberian populations in Spain of 1000 Genomes Project\_Phase3 ([www.ensembl.org](http://www.ensembl.org)); Conc – sperm concentration; Avg veloc – average velocity of sedimentation; ALH – amplitude of lateral head.

<sup>a</sup>Allele: minor/major allele (as defined in controls).

The genomic DNA was extracted from peripheral white blood cells using the QIAamp<sup>®</sup> DNA Blood Midi/Maxi (Qiagen, Hilden, Germany), following the procedures described by the manufacturers, and then diluted in the ratio of 1:20. All aliquots of severe oligozoospermia were quantified by UV/Vis spectrophotometry in Nanodrop 1000 and then selected using the concentration of 30ng/μl as a minimum threshold. All the samples that presented a concentration below the desired minimum were subjected to genomic amplification. The whole-genome amplification by illustra GenomiPhi V2 DNA Amplification Kit was carried out and after an 1:25 ratio dilution was done.

### 2.1 Sanger sequencing to obtain reference individuals

For quality control of the genotyping, individuals of reference for all the possible genotypes were obtained by Sanger sequencing of the region of interest. Several patients were sequenced to obtain three positive controls for each SNP (allele1/allele1; allele1/allele 2; allele2/allele2).

Polymerase Chain Reactions (PCRs) were undertaken with the designed primers (described in **Table S1** in the Supplementary Materials) and the PCR conditions varied according to the target amplification product.

PCR primers were designed using the Primer3web 4.1.0 Software (available on <http://primer3.ut.ee/>) and selected following the standard pre-sets: melting temperature (T<sub>m</sub>) close to 60°C, GC (Guanine/Cytosine) content of approximately 50% and hairpin and dimer primer values below 3. To avoid primers with high identity to non-target sequences were used the “BLAT” tool from UCSC Genome Browser, in order to ensure that no cross-hybridization or amplification of nonspecific PCR product occurs. For the last check, was used the “*In Silico* PCR” tool (also available on UCSC Genome Browser) to verify, with each pair of primers, if there are another possible amplification product.

The enzyme mix applied was the 2x MyTaq HS Mix (Bioline, Porto, Portugal) and, for all the candidate SNPs, the PCR mix were prepared according to following protocol:

**Table 4 - PCR Master Mix protocol**

Reagents	( $\mu$ l)
2x MyTaq HS Mix	5
Forward Primer (5mM)	0,5
Reverse Primer (5mM)	0,5
H <sub>2</sub> O	3
DNA Template ( $\geq 30$ ng/ $\mu$ l)	1

The PCR thermocycling conditions were optimized for each SNP in accordance to the designed primer pair characteristics. For 3 SNPs (USP8-rs7174015, PSAT1-rs7867029, TUSC1-rs10966811) the selected conditions were the standard to this protocol:

**Table 5 - Standard thermocycling conditions that allow the amplification of the sequences of USP8-rs7174015, PSAT1-rs7867029 and TUSC1-rs10966811.**

Temperature (°C)	Time	Cycles
94	10 min	1
94	30 s	35
60	30 s	
72	30 s	
72	10 min	1
12	$\infty$	



The annealing temperature (T<sub>m</sub>) were adapted to the characteristics of each primer pair. For the EPSTI1-rs12870438 and DPF3-rs10129954, the PCR conditions that allows the amplification of the desired product, avoiding nonspecific amplifications, were:

**Table 6 - Thermocycling conditions that allow the amplification of the sequences of EPSTI1-rs12870438 and DPF3-rs10129954.**

rs12870438			rs10129954		
Temperature (°C)	Time	Cycles	Temperature (°C)	Time	Cycles
94	10 min	1	94	15 min	1
94	30 s	5	94	30 s	10
62	30 s		64	30 s	
72	30 s		72	30 s	
94	30 s	30	94	30 s	25
60	30 s		62	30 s	
72	30 s		72	30 s	
72	10 min	1	72	10 min	1
12	∞		12	∞	

PCRs were carried out in Applied Biosystems Thermal Cycler from Applied Biosystems (Foster City, CA).

To confirm the specific amplification, PCR products were separated by electrophoresis on an acrylamide gel. A standard protocol was used for gel preparation, that included the following reagents: 3 mL Acrylamide solution + 170 µl APS + 7 µl TEMED. Acrylamide solution is prepared with 25 mL Gel Buffer 4x (1,5 M Tris HCl Buffer, pH 8,8), 20 mL Acrylamide:Bisacrylamide (19:1, 40%), 7 mL Glycerol and 43 mL H<sub>2</sub>O. After the electrophoretic race, the gel was stained following the Silver Staining protocol described in **Table 7**.

**Table 7 - Silver Staining protocol.**

Reagents	Time
1. Ethanol (10 %) (25 mL + 225 mL distilled water)	10 min
2. Nitric Acid (1%) (2.5 mL + 247.5 mL distilled water)	5 min
3. Wash with distilled water	2 x 10 s
4. Silver Nitrate (0,2 %) (0.5 g + 250 mL distilled water)	20 min
5. Wash with distilled water	2 x 10 s
6. Sodium Carbonate (0.28 M) + Formaldehyde (0.02 %) (3 g carbonate + 100 mL distilled water + 1 mL formaldehyde)	-
7. Acetic Acid (10 %)	2 min
8. Wash with water	2 x 10 s

Only the PCR products that were in concordance with the expected band size after confirmation on the acrylamide gel were sequenced. The obtained sequence was the region in which the SNP is localized.

Sanger sequencing is based on the use of dideoxynucleotides (ddNTPs) beyond the normal nucleotides (dNTPs). Each ddNTP (A, T, C, G) is fluorescently labelled. When that modified nucleotides connect to the sequence, this will prevent the addition of further nucleotides and consequently the DNA chain is terminated. The Sanger Sequencing protocol includes five main steps:

1) ExoSap Reaction.

Aim: Removal from PCR products the unincorporated primers and free nucleotides.

1 µl ExoFast + 2 µl PCR product

**Table 8 - ExoSap Reaction thermocycling conditions.** Thermal Cycler from Applied Biosystems.

Temperature (°C)	Time	Cycles
37	15 min	1
80	15 min	
12	∞	

2) Cycle Sequencing

Aim: Incorporation of ddNTPs.

1 µl Big Dye v3.1+ 1 µl Big Dye v3.1 Seq Buffer + 0.5 µl Reverse Primer +  
2.5 µl PCR product (from ExoFast reaction)

**Table 9 - Cycle sequencing thermocycling conditions.** Thermal Cycler from Applied Biosystems.

Temperature (°C)	Time	Cycles
96	1 min	1
96	15 s	35
50	5 s	
60	2 min	
12	∞	1

3) Sephadex Purification

Aim: Removal from sequencing reaction product the incorporated primers and free ddNTPs.

750 µl Sephadex + 5 µl PCR product (from Sequencing Cycle) +  
10 µl HI-DI Formamide

**Table 10 - Sephadex purification protocol.**

1. Place columns in 2 mL tubes;
2. Add 750 µl of Sephadex in each column;
3. Centrifuge for 4 minutes at 4400 rpm;
4. Place the columns in a 1.5 mL tubes, previously identified, and discard the liquid left in the 2 mL tubes;
5. Add the totally of the Cycle Sequencing product (5 µl) in the center of the Sephadex column;
6. Centrifuge for 4 minutes at 4400rpm;
7. Remove the columns. The purified products are already in the 1.5 mL tube.

4) Hi-Di Formamide

Aim: Increase the stability of the single stranded DNA molecules.

5 µl Purified Cycle Sequencing Product + 10 µl Hi-Di Formamide

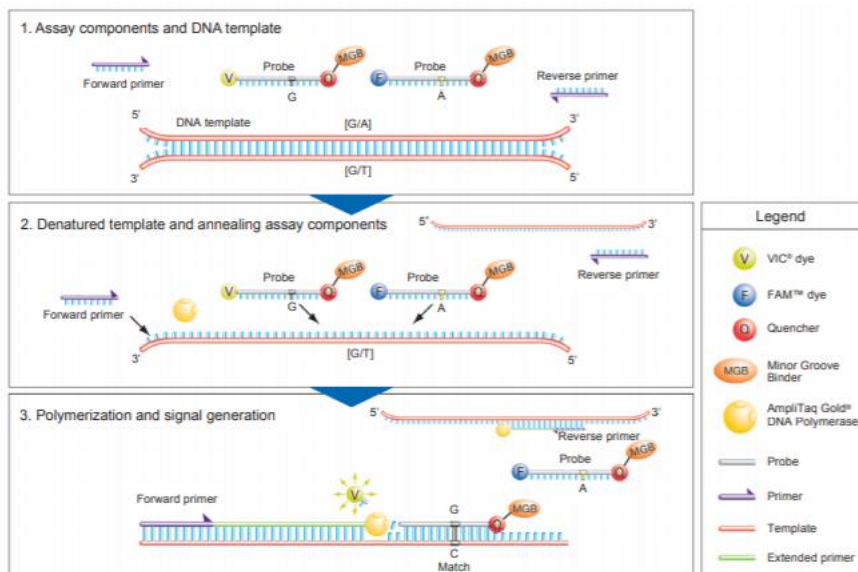
## 5) Capillary Electrophoresis

The purified sequencing products were run in the automatic sequencer ABI 3130XL (Genetic Analyser 3000<sup>®</sup> from Applied Biosystems) and analyzed using Chromas Software (Technelysium Pty, Lda).

Following the analysis of high-quality sequences, 3 individuals were selected as reference for each SNP (positive controls).

## 2.2 Genotyping

The five candidate SNPs were genotyped by TaqMan Assay on a 7500 Fast Real Time PCR System (Applied Biosystems). The TaqMan genotyping assay requires a DNA template (double-stranded), Taq polymerase enzyme, a pair of primers (forward and reverse) – which are specific to the region where the SNP is located – and a pair of probes with different fluorescent reporters. Each probe is designed in order to distinguish the alleles present in the SNP location and only the probe complementary with the DNA template emits fluorescence (**Figure 3**).



**Figure 3 - TaqMan Genotyping Assay.** Allelic discrimination is achieved by the selective annealing of TaqMan MGB probes (Biosystems, 2004)

The probes are labelled with VIC and FAM fluorescent dye according with the allele 1 or 2, respectively. This allows the differentiation of all genotypes (homozygous for most frequent and less frequent alleles and heterozygous).

The TaqMan Probe mix assay was constituted by predesigned probes (see details on the probes in **Table S2**, Supplementary Materials) and primers for the SNP target regions. The TaqMan assay used was TaqMan® Genotyping Master Mix (Applied Biosystems).

In all plates one blank was included as negative control and DNA from 3 positive controls (individuals of reference, one of each possible genotype).

((5 µl TaqMan Probe + 200 µl Master Mix + 200 µl H<sub>2</sub>O) / 96 wells)  
+ 1 µl DNA Template

**Table 11 - Real Time - PCR thermocycling conditions.**

Temperature (°C)	Time	Cycles
95	10 min	1
95	15 s	40
60	1 min	

The results obtained after real time PCR reaction were presented in an Allelic Discrimination Plot, a graphic depicting the correlation between the fluorescence of VIC and FAM dyes, and the genotype of each sample.

### 3. Statistical analysis

CaTS Power Calculator for Genetic Studies Program<sup>42</sup> was used to estimate de Statistical power of our study.

All statistical analyses were performed in Plink (v1.9) Software. We first tested for any deviation from the Hardy-Weinberg equilibrium, with the significance level set to 5%. Allelic and genotypic frequencies were calculated in both case groups (NOA and SO) and controls (Normozoospermia) and at this stage, for the cases, three possible phenotypes were defined – azoospermia, severe azoospermia and the combination of the previous two. To test for association, case-control comparisons were done by logistic regression<sup>43</sup> assuming three different models – additive (risk allele), dominant (AA+AB vs BB) and recessive (AA vs AB+BB). In order to understand the tendency of the observed association, the *Odds Ratio* (OR) was calculated using a 95% Confidence Interval. When the OR value obtained is greater than 1, the allele confers risk, while an

OR value less than 1 will be associated with protection. Only the P-values lower than 0.05 were considered statistically significant.

To correct for multiple testing, the false discovery rate (FDR) correction of the P-values were done, according to the Benjamini & Hochberg method.

#### 4. *In silico* characterization

In order to assess the putative functional impact of the SNPs showing a trend towards association, publicly available functional annotation data were explored, using different bioinformatics tools. The aim of this analysis is to find a functional variant in strong linkage disequilibrium with the variant under analysis. First of all, the proxies that present a high correlation ( $R^2 > 0.8$ ) of alleles for each studied SNP in the Iberian Population (IBS) were identified using LDLink<sup>44</sup>. Initially, all proxies were considered as candidates in the search of an underlying molecular mechanism explaining the observed association trends with male infertility traits. In order to evaluate eQTL and sQTL effects in the testis, GTEx data (<https://www.gtexportal.org.pt/>)<sup>45</sup> was consulted. Furthermore, the Single Cell Expression Atlas (<https://www.ebi.ac.uk/gxa/sc>) was then queried in order to explore single cell expression in human testis of genes modulated by the SNPs of interest.

To evaluate the presence of the different regulatory chromatin marks, such as DNase-seq hypersensitivity sites, CTCF protein binding sites, H3K4me3 and H3K27me3 histone modifications the ENCODE data the regions of interest were inspected in UCSC Genome Browser. The identifiers of the call sets used are ENCSR611DJQ (testis of 37 years old male) and ENCSR981CID (testis of 54 years old male).

Additional SNP annotation was also extracted from SNPnexus<sup>46</sup> (<https://www.snp-nexus.org/>) and Haploreg v.4.1<sup>47</sup> (<https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php>). These portals integrate the variant annotations from several databases, such as Ensembl, SIFT, Polyphen, CpG, Vista enhancers, miRbase, TarBase, TargetScan, miRNA Registry, snoRNA-LBME-DB, Road Epigenomics, Ensembl regulatory build, RegulomeDB<sup>48</sup>, and functional consequence predictions based on several algorithms like CADD, DeepSEA, EIGEN, FATHMM, fitCons, FunSeq2 GWAVA, REMM (**Tables S8 and S9** in the Supplementary Materials).

# Results

## Genotyping quality control

The aim of this study was to determine if there is an association between the selected SNPs, previously associated with fertility traits, and male infertility phenotypes. For that, a case-control study was defined to ascertain if there are significant allelic and/or genotypic differences between a cohort of Portuguese men with severe spermatogenic failure and a geographically matched control population.

The accuracy of the genotyping method (the TaqMan Genotyping Assay) was tested by including in every assay three individuals whose genotypes were previously determined by Sanger Sequencing (AA, AB or BB). In the following table (**Table 12**) a brief summary of the controls is presented, and all the detailed results are presented in **Tables S3, S4, S5 and S6** in the Supplementary Material.

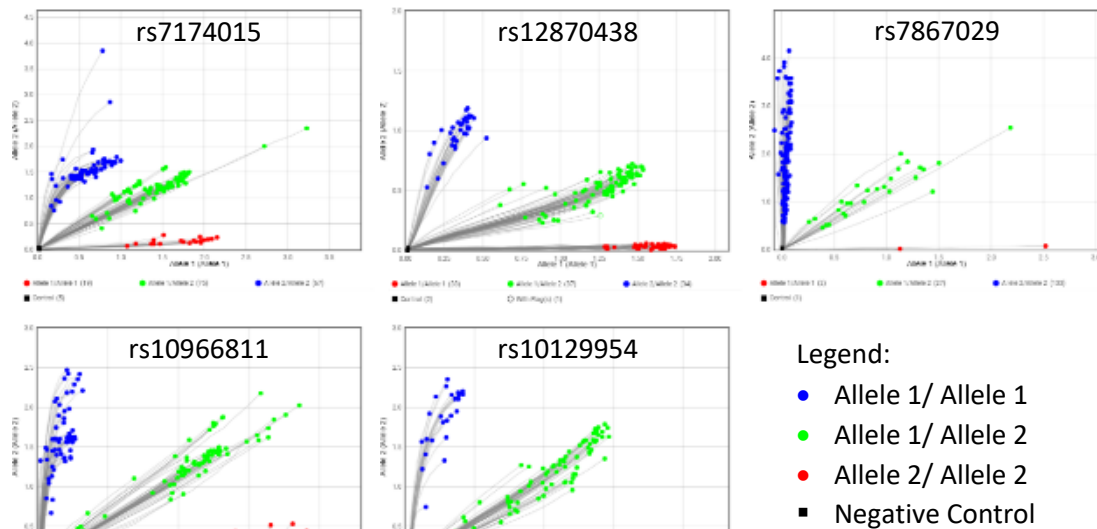
**Table 12 - Positive controls used for each SNP.**

	rs12870438		rs7867029		rs10966811		rs10129954	
	Sample	Genotype*	Sample	Genotype*	Sample	Genotype*	Sample	Genotype*
1/1	Y1627	AA	YF115	CC	Y1506	AA	Y1471	CC
1/2	Y1621	AG	YF129	CG	Y1462	AG	Y1462	CT
2/2	Y1638	GG	Y1462	GG	Y1474	GG	Y1519	TT

\*obtained by Sanger sequencing.

After several attempts varying the thermocycling conditions and the primers used, it was not possible to sequence the region of USP8-rs7174015 and, consequently, we could not obtain reference individuals for that SNP. The region flanking this SNP is rich in repetitive sequence hindering the analysis through direct sequencing. For this SNP the genotyping was performed by direct reading of fluorescence ratio for each tested individual.

The results obtained by Real Time PCR were resumed in an Allelic Discrimination Plot, for each SNP, as shown in Figure 4.



**Figure 4 - Allelic Discrimination Plot for each SNP.**

Results obtained with the TaqMan Genotyping Assay. Each point is an individual and different colours indicate different genotypes. Shown are the genotyping results obtained for the severe oligozoospermic patients. Each colour (blue, green and red) represents one genotype according to the fluorescence emitted by the dye(s) associated with the present allele(s). In a sample shown as a blue or a red dot, only fluorescence from a single dye has been detected, VIC or FAM, indicating the presence of 2 copies of Allele 1 or Allele 2, respectively. However, when a sample is represented in green, it indicates double dye fluorescence (VIC and FAM) and thus a heterozygote. Similar results were obtained for the other samples genotyped.

## Genotypes and association tests

The study in the Portuguese population was well powered for detecting large effects (EP≥0.9 for OD>1.5) but underpowered for lower ORs, as shown in **Table S7** in the Supplementary Materials. The genotyping success rate in different assays fell between 97.5% and 100%.

**Table 13 - Genotype and Allele Frequencies in different study groups.**

Chr	SNP	Position	Closest Gene	Location	Distance (kbp)	Allele <sup>a</sup>	IBS MAF	CONTROLS			CASES					
								n	Genotype counts <sup>b</sup>	AF <sup>c</sup>	n	Genotype counts <sup>b</sup>	AF <sup>c</sup>	n	Genotype counts <sup>b</sup>	AF <sup>c</sup>
9	rs10966811	25,233,486	<i>TUSC1</i>	downstr.	442.903	A/G	0.41	127	14/68/45	0.38	217	28/97/92	0.35	160	27/71/62	0.39
			<i>IZUMO3</i>	upstr.	687.837											
9	rs7867029	78,405,502	<i>PSAT1</i>	downstr.	75.409	C/G	0.18	127	2/31/94	0.14	219	2/61/156	0.15	161	2/26/133	0.09
13	rs12870438	4,290,669	<i>EPST11</i>	intron		A/G	0.35	124	25/56/43	0.43	218	33/94/91	0.37	161	19/75/67	0.35
14	rs10129954	72,683,993	<i>DPF3</i>	intron		T/C	0.48	127	17/64/46	0.39	216	42/102/72	0.43	163	25/76/62	0.39
15	rs7174015	50,424,871	<i>USP8</i>	intron		A/G	0.47	125	24/63/38	0.44	214	66/92/56	0.52	159	34/87/38	0.49

Genes: *TUSC1* - Tumor Suppressor Candidate 1; *IZUMO3* - IZUMO family member 3; *PSAT1* - Phosphoserine Aminotransferase; *EPST11* - Epithelial Stromal Interaction Protein 1; *DPF3* - Double PHD Fingers 3; *USP8* - Ubiquitin Specific Peptidase 8.

<sup>a</sup> Allele: minor/major allele (as defined in controls)

<sup>b</sup> Minor Allele Frequencies of Iberian populations in Spain in 1000 Genomes Project\_Phase3 ([www.ensembl.org](http://www.ensembl.org))

<sup>c</sup> Genotype counts: minor homozygote/heterozygote/major homozygote (as defined in controls)

<sup>d</sup> AF: allele frequency of the minor allele (as defined in controls)



The details of the variants studied as well as a summary of the genotyping results obtained for all groups are shown in **Table 13**.

According to the Principle of Hardy-Weinberg Equilibrium (HWE) the allele frequency at a given polymorphic position remains stable from one generation to the next in the absence of disturbing factors. The only significant deviation from HWE ( $p < 0.05$ ) was detected for USP8-rs7174015 in a case cohort, the group of azoospermic patients ( $p=0.041$ ) and thus it may reflect a true difference and not a genotyping error. The minor allele frequencies (MAF) of our control groups agreed with those described for the Iberian population of the 1000 Genomes Project\_Phase3.

In order to evaluate the possible effect of the candidate variants in the genetic susceptibility to spermatogenesis failure, the allele and genotype frequencies of the case groups reflecting the main clinical phenotypes (NOA, SO and NOA+SO Combined) were compared with those of the control population. These results were obtained by linear regression assuming three different models: additive, dominant and recessive (**Table 14**).

**Table 14 – Results of the association tests derived from different comparative genetic models (additive, dominant and recessive).**

SNP	Model	Azoospermia			Severe Oligozoospermia			Combined		
		OR (95% CI)	p value	FDR B-H	OR (95% CI)	p value	FDR B-H	OR (95% CI)	p value	FDR B-H
rs10966811	Additive (risk allele, A)	0.89 (0.64-1.24)	0.495	0.619	1.06 (0.75-1.49)	0.754	0.942	0.96 (0.72-1.29)	0.792	0.792
	Dominant (AA+AG vs. GG)	0.75 (0.47-1.17)	0.204	0.588	0.87 (0.54-1.41)	0.564	0.705	0.80 (0.52-1.21)	0.281	0.468
	Recessive (AA vs. AG+GG)	1.20 (0.60-2.37)	0.608	0.608	1.64 (0.82-3.28)	0.162	0.406	1.38 (0.74-2.58)	0.314	0.392
rs7867029	Additive (risk allele, C)	1.10 (0.69-1.74)	0.691	0.691	0.64 (0.38-1.08)	0.097	0.244	0.89 (0.58-1.36)	0.590	0.737
	Dominant (CC+CG vs. GG)	1.15 (0.70-1.88)	0.578	0.587	<b>0.60</b> (0.34-1.06)	<b>0.078</b>	<b>0.389</b>	0.90 (0.57-1.42)	0.644	0.805
	Recessive (CC vs. CG+GG)	0.58 (0.08-4.14)	0.584	0.608	0.79 (0.11-5.66)	0.811	0.811	0.66 (0.12-3.67)	0.640	0.640
rs12870438	Additive (risk allele, A)	0.79 (0.58-1.08)	0.134	0.335	0.73 (0.52-1.02)	0.068	0.244	0.76 (0.57-1.02)	0.065	0.212
	Dominant (AA+AG vs. GG)	0.74 (0.47-1.17)	0.199	0.509	0.74 (0.46-1.21)	0.234	0.389	0.74 (0.49-1.13)	0.167	0.468
	Recessive (AA vs. AG+GG)	0.71 (0.40-1.25)	0.235	0.392	<b>0.53</b> (0.28-1.01)	<b>0.055</b>	<b>0.276</b>	0.63 (0.37-1.07)	0.086	0.225
rs10129954	Additive (risk allele, T)	1.20 (0.88-1.65)	0.252	0.420	1.00 (0.71-1.41)	0.987	0.987	1.11 (0.83-1.49)	0.471	0.737
	Dominant (TT+TC vs. CC)	1.14 (0.72-1.80)	0.587	0.587	0.93 (0.57-1.50)	0.751	0.751	1.04 (0.68-1.58)	0.860	0.860
	Recessive (TT vs. TC+CC)	1.56 (0.85-2.88)	0.153	0.383	1.17 (0.60-2.28)	0.640	0.811	1.39 (0.78-2.47)	0.262	0.392
rs7174015	Additive (risk allele, A)	1.34 (0.99-1.82)	0.057	0.283	1.21 (0.85-1.70)	0.289	0.481	1.29 (0.97-1.71)	0.085	0.212
	Dominant (AA+AG vs. GG)	1.23 (0.76-2.01)	0.401	0.587	1.39 (0.82-2.36)	0.220	0.389	1.30 (0.83-2.03)	0.255	0.468
	Recessive (AA vs. AG+GG)	<b>1.88</b> (1.10-3.19)	<b>0.020*</b>	<b>0.101</b>	1.15 (0.64-2.05)	0.651	0.811	1.54 (0.93-2.54)	0.090	0.225

Data are shown as odds ratio (OR), 95% confidence interval (CI), P value and FDR Benjamini–Hochberg adjustment. Numbers in bold show trend towards association and their respective OR (95% CI). Asterisk indicates  $P < 0.05$ .

Even though no significant differences were observed between case and control populations after multiple testing correction, under any of the tested models, some trends

were seen that are in concordance with Kosova *et al.*<sup>39</sup>. Assuming a dominant model, the C allele of PSAT1-rs7867029, previously associated with larger family size, shows a non-significant trend towards a protective effect for severe oligozoospermia (OR=0.6; nominal p=0.078). Another borderline trend (OR=0.5; nominal p=0.055) was detected when a recessive model is assumed for EPSTI1-rs12870438. In that case the A allele shows signs of association with SO, also as a protective factor. The opposite effect was confirmed for USP8-rs7174015, which had previously been associated with smaller family size. Assuming a recessive model, this variant showed a nominally significant association with azoospermia (OR=1.88; p=0.020) even though it did not pass the threshold of significance after FDR correction.

### ***In silico* functional analysis**

We next searched for polymorphisms in high LD ( $R^2 > 0.8$ ) with the most interesting SNPs studied – USP8-rs7174015, PSAT1-rs7867029, EPSTI1-rs12870438 - in the Iberian population of 1KGPb3, to conduct *in silico* functional analysis. In total 24 proxies were selected for USP8-rs7174015, one for PSAT1-rs7867029 and 16 for EPSTI1-rs12870438 (Tables 15, 16 and 17). From here, all analyses were performed for the lead SNPs and for all their proxies.

**Table 15 - Proxies of UPS8-rs7174015 obtained by LDLink, using as a threshold  $R^2 > 0.8$ .**

	<b>R<sup>2</sup></b>	<b>SNP</b>	<b>chr</b>	<b>Position (GRCh37)</b>	<b>Alleles</b>	<b>MAF IBS</b>
<b>rs7174015</b>	1	rs3098177	15	50783777	G/A	0.472
	1	rs2289108	15	50782335	G/A	0.472
	1	rs56398519	15	50773432	A/G	0.472
	1	rs3098171	15	50771511	C/G	0.472
	1	rs8026653	15	50736620	G/C	0.472
	1	rs11070776	15	50734588	G/A	0.472
	1	rs28582911	15	50729629	G/A	0.472
	0.9632	rs34639682	15	50791803	T/-	0.4813
	0.9632	rs36042420	15	50791585	T/-	0.4813
	0.9632	rs3098176	15	50781963	T/A	0.4813
	0.9102	rs4318151	15	50729274	A/G	0.4486
	0.8917	rs3131559	15	50780214	C/A	0.4533
	0.8917	rs3098174	15	50778853	C/T	0.4533
	0.8917	rs3131562	15	50772479	C/T	0.4533
	0.8773	rs3098169	15	50754339	G/A	0.4953
	0.8773	rs3098205	15	50742977	G/T	0.4953
	0.8773	rs3131574	15	50740939	C/A	0.4953
	0.861	rs3131568	15	50757721	C/T	0.4907
	0.861	rs10152326	15	50742248	A/G	0.4907
	0.861	rs11417752	15	50738592	_/G	0.4907
	0.861	rs12593481	15	50718276	T/C	0.4907
	0.8451	rs3131560	15	50775562	T/C	0.486

0.8451	rs3131566	15	50758662	T/C	0.486
0.8451	rs3098167	15	50750469	T/C	0.486

Table 16 - Proxy of PSAT1-rs7867029 obtained by LDLink, using as a threshold  $R^2 > 0.8$ .

	$R^2$	SNP	Chr	Position (GRCh37)	Alleles	MAF IBS
rs7867029	0.861	rs10867194	9	81018312	C/T	0.2056

Table 17 - Proxies of EPST11-rs12870438 obtained by LDLink, using as a threshold  $R^2 > 0.8$ .

	$R^2$	SNP	Chr	Position (GRCh37)	Alleles	MAF IBS
rs12870438	1	rs58357177	13	43480417	G/-	0.3458
	1	rs1535898	13	43479538	A/G	0.3458
	1	rs1535897	13	43479387	C/T	0.3458
	1	rs1830780	13	43477688	C/A	0.3458
	1	rs63351261	13	43477345	AT/-	0.3458
	0.9795	rs9594829	13	43474828	T/C	0.3411
	0.9591	rs11431398	13	43476368	_/TT	0.3458
	0.9404	rs1535900	13	43480001	G/A	0.3598
	0.9216	rs34525682	13	43482817	T/C	0.3645
	0.9191	rs10161854	13	43460000	A/G	0.3458
	0.8998	rs9590722	13	43457890	A/C	0.3505
	0.8998	rs12870885	13	43457116	T/C	0.3505
	0.8998	rs9594827	13	43455610	A/T	0.3505
	0.8998	rs9594826	13	43455550	T/C	0.3505
	0.8608	rs1044856	13	42462422	A/T	0.3224
	0.8429	rs71099806	13	43483735	A/-	0.3551

According to the data obtained from SNPnexus<sup>46</sup>, for USP8-rs7174015 and PSAT1-rs7867029, none of the lead variants nor their proxies were located in coding regions, CpG islands or miRNA target sequences. One of the proxies of EPST11-rs12870438 overlaps a miRNA (hsa-mir-221), however MIR221 does not show expression in human testis (**Figure S1** in the Supplementary Materials).

The annotations for an eQTL and sQTL effect were then accessed on the GTEx portal<sup>45</sup>. None of the SNPs presented sQTL effects in testis, however some of them presented annotations in other tissues (**Table S10** in the Supplementary Materials) highlighting the regulatory relevance of these genomic regions. The lead SNP USP8-rs7174015 and 19 of its proxies were noted as eQTLs in several tissues (**Table S11** in the Supplementary Materials), namely in testis, where they affect the expression of *USP8*, *USP50*, *AP4E1* and *RP11-562A8.5* (**Figure 6**). According to GTEx<sup>45</sup>, the

expression of these four genes in testis is considerably high (**Figures S2, S3, S4 and S5** in the Supplementary Materials), denoting that *USP50* present testis-specific expression.

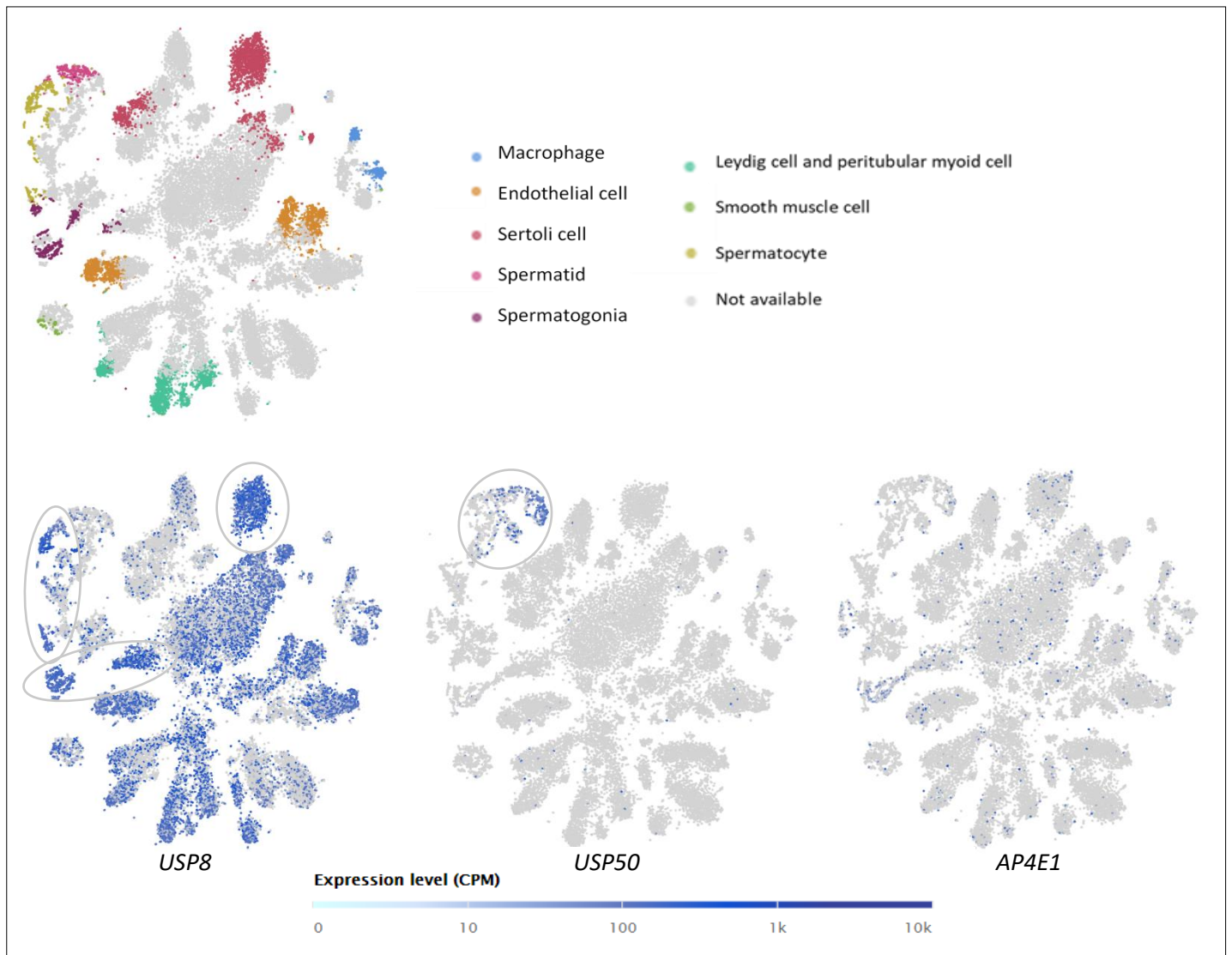
	eQTL effect in testis				Predictive functional analyses								
	<i>AP4E1</i>	<i>USP8</i>	<i>USP50</i>	<i>RP11-562A8.5</i>	RegulomeDB	CADD	fitCons	EIGEN	FATHMM	GWAVA	DeepSea	FunSeq2	ReMM
rs7174015					4								
rs10152326					6								
rs11070776					7								
rs11417752													
rs12593481					3b								
rs2289108					6								
rs28582911					7								
rs3098167					6								
rs3098169					6								
rs3098171					7								
rs3098174					6								
rs3098176													
rs3098177					7								
rs3098205					3a								
rs3131559					6								
rs3131560					6								
rs3131562					5								
rs3131566					6								
rs3131568					5								
rs3131574					5								
rs34639682													
rs36042420													
rs4318151					6								
rs56398519					5								
rs8026653													

**Figure 5 - Functional annotations of the human genome for the lead SNP variant USP8-rs7174015 and its proxies.** Overlaps are highlighted with different colors: blue for eQTL effects in testis (affected genes are coloured); orange for RegulomeDB (each tonality corresponds to each rank); green for functional prediction scores, in which the heatmap presents the probability of functionality (darker green indicates higher probability).

To ascertain the predictive functional effect of the tested variants, different scores (**Table S12** in the Supplementary Materials) were calculated with tools like CADD, deepSea, EIGEN, FATHMM, fitCONS and ReMM, comprised in the heatmap (green) of the **Figure 5**. The lead SNP USP8-rs7174015 and two of its proxies – rs12593481 and rs56398519 – showed higher probability to be causal variants in this LD block. These variants were also noted as sQTLs and eQTLs in several tissues (**Tables S10 and S11**, respectively, in the Supplementary Materials), highlighting their relevance in regulatory processes. The RegulomeDB<sup>48</sup> rank is also taken into account, in order to search SNPs within known or predictive regulatory elements such as regions of DNase hypersensitivity, transcription factor binding sites (TFBS), and promoter regions that have been biochemically characterized to regulate transcription, in intergenic regions. The scores of the studied lead variant and their most interesting proxies are shown in **Figure 5**, in which each score level corresponds to a gradient of orange (details shown in **Table S8**, in the Supplementary Materials).

Gene expression data of *USP8*, *USP50* and *AP4E1* in different testicular cells types based on single-cell RNA-seq experiments in pubertal human testes<sup>49</sup> was available in the Single Cell Expression Atlas portal. The results, shown in dimension reduction (t-

SNE) plots, were compacted in **Figure 6**. No single-cell transcriptome data was available for *RP11-562A8.5*.



**Figure 6 - Gene expression in human testicular cells.**

Single cells are represented as colored dots, and the different colors indicate the inferred cell type of testis, as demonstrated in the legend. The t-SNE plots show the specific expression patterns of *USP8*, *USP50* and *AP4E1*, being the tonality of blue correlated with expression levels (darker blue indicates higher expression).

*USP8* was mostly detected in spermatogonia, spermatocytes and Sertoli cells, while *USP50* was almost exclusively expressed in spermatids. However, a diffuse expression in multiple cell types was noted for *AP4E1*, which can be suggestive of a housekeeping role throughout the spermatogenic process.

The same approach was applied to the lead variants EPST11-rs12870438 and PSAT1-rs7867029 and their proxies. No eQTL or sQTL annotations were found in testis or in any other tissues, except for the variant rs1044856. This EPST11-rs12870438 proxy is annotated as an eQTL, affecting the expression of *EPST11* in whole blood, and thus not particularly relevant for spermatogenesis.

From the results of the predictive functional analyses (Table S13 in the Supplementary Materials), such as CADD, fitCons, EIGEN, GWAVA, FATHMM, DeepSea, FunSeq2 and ReMM, four SNPs (proxies of EPSTI1-rs12870438) – rs1535898, rs9590722, rs9594827 and rs9594829 – were highlighted as the most probable causal variants in the LD block (**Figure 7**). However, according to RegulomeDB, the variants with higher potential functional impact are EPSTI1-rs12870438 and rs9590722, the ones with a lower score (score = 4), corresponding to TF binding and DNase peak evidence (see **Table S8**, Supplementary Materials). For the lead PSAT1-rs7867029 and its proxy, no relevant score values were obtained for the predictive functional analyses (**Table S14** in the Supplementary Materials). However, a score of 5 was obtained in the RegulomeDB for the rs10867194, corresponding to TF binding or DNase peak data available.

	RegulomeDB	CADD	fitCons	EIGEN	FATHMM	GWAVA	DeepSea	FunSeq2	ReMM
<b>rs12870438</b>	4								
rs10161854	6								
rs1044856									
rs11431398	6								
rs12870885	5								
rs1535897	5								
rs1535898	5								
rs1535900	6								
rs1830780	6								
rs34525682	7								
rs58357177	-								
rs63351261	6								
rs71099806	-								
rs9590722	4								
rs9594826	5								
rs9594827	5								
rs9594829	7								

**Figure 7 - Predictive functional analysis for the rs12870438 and its proxies.**

RegulomeDB ranks are shown in orange shades (each tonality corresponds to a given rank); In shades of green are shown the results of the functional prediction analyses, in which the heatmap presents the probability of functionality (darker green indicates higher probability).

Based on Haploreg v4.1 annotations, several transcription factor binding sites are changed by the lead SNPs or their proxies (**Table 18**), standing out YY1, FOXJ1, HSF, SIX5, BCL6 and DMRT, which are relevant transcription factors for the spermatogenic process<sup>50,51,52,53,54,55</sup>. YY1 is changed by two USP8-rs7174015 proxies - rs56398519, rs2289108 – and one EPSTI1-rs12870438 proxy - rs12593481 and DMRT4 by the lead variant PSAT1-rs7867029. FOXJ1 is modified by rs56398519 and rs3098174, HSF by rs3098171, SIX5 by rs28582911, BCL6 by rs34639682, all of them proxies of USP8-rs7174015.

**Table 18 - Haploreg v4.1 annotations on the transcription factor binding sites changed by each SNP.**

SNP	Motif changed (TF)
rs3098177	MYF, PAX-5, p300
<b>rs2289108</b>	TATA, <b>YY1</b>
<b>rs56398519</b>	FAC1, <b>FOXJ1</b> , SOX
<b>rs3098171</b>	CEBPB, <b>HSF</b> , TEF
<b>rs8026653</b>	CEBPG, FOXA, <b>FOXJ1</b> , FOXJ2, FOXK1, FOXO, FOXP1, HOXA10, HOXA9, HOXB13, HOXB9, HOXD10, MEF2, NKX6-1, PAX-4
rs11070776	RAD21, SP4
<b>rs28582911</b>	<b>SIX5</b>
<b>rs34639682</b>	<b>BCL6</b> , CDC5, HNF1, IK-2, MEF2, POU5F1, STAT
rs36042420	EVI-1, HDCA2, IRF, ZFP105, p300
rs3098176	EVI-1, TATA, ZFP105
rs4318151	CDX
rs3131559	POU3F2, TATA
<b>rs3098174</b>	CEBPB, CART1, DBX1, DBX2, FOXA, <b>FOXJ1</b> , FOXJ2, FOXP1, HNF1, LHX3, NCX, PAX-6, POU2F2, POU3F4, POU6F1, TATA
rs3131562	CEBPA, CEBPB
rs3098169	FOXP1, POU1F1, POU5F1, SOX
rs3098205	BCL, BACH1, HOXD10, PLZF, RXRA, TATA
rs3131574	GCNF, GR, NR2f2, TCF12, VDR
rs3131568	IRF
rs10152326	CHD2, NRSF, RAD21
rs11417752	FOXD1, HNF4, MEF2, PITX2
<b>rs12593481</b>	CEBPA, CEBPB, CEBPD, CHOP::CEBPalpha, PAX-5, WHN, <b>YY1</b> , p300
rs3131560	GR, SOX
rs3131566	E2F, MYc, ZBTB7A
rs3098167	LUN-1, MAF
rs12870438	SEF-1
rs58357177	E2A, MYF, SEF-1, ZEB1
rs1535898	HOXA5, RXRA
rs1535897	AP-1, E2F, STAT
rs1830780	CRX, NR2F2
rs63351261	ATF3, EVI-1, HEY1, HP1-site.factor, IRX, POU2F2, TATA
rs9594829	RXRA
rs1535900	EWSR1-FLI1, FOXI1, POU2F2
<b>rs10161854</b>	AP-1, MTF1, RXRA, <b>YY1</b> , ZBTB3, ZNF143, p300
rs9594827	BARX1, BSX, DLX5, HOXB4, NKX2, NOBOX, PRRX1, SP1
rs9594826	<b>SIX5</b>
rs1044856	AP-1, HNF1, MEF2, OTX, OBOX3, OTX2, SP1
rs71099806	EVI-1, HDCA2, HOXA10, SRF, STAT, TATA, p300
<b>rs7867029</b>	CEBPA_1, <b>DMRT4</b> , E2F, NF-E2, POU3F2_1, TFE, p300
rs10867194	NRF1, OSR, SOX_16

According to the SNPnexus Annotation Tool <sup>46</sup>, based on Roadmap Epigenomics, ENCODE and Ensembl Regulatory Build databases, the studied variants overlap regulatory marks in several cell lines and tissues, supportive a putative regulatory relevance for these regions. Even though no data on regulatory marks in testis is available for these SNPs in SNPnexus, according to CHIP-seq data from ENCODE (accessed through UCSC Genome Browser) some of the variants overlap chromatin marks associated with active promoters, such as H3K4me3, and of CTCF binding sites, in adult testis (**Tables 19 and 20**).

**Table 19 - Annotations of regulatory marks (H3K4me3 and CTCF) in testis, overlapping USP8-rs7174015 and their proxies.**

SNP	H3K4me3	CTCF
	Testis	
rs7174015		
rs3098177		
rs2289108		
rs56398519		
rs3098171		
rs8026653		
rs11070776		
rs28582911		
rs34639682		
rs36042420		
rs3098176		
rs4318151		
rs3131559		
rs3098174		
rs3131562		
rs3098169		
rs3098205		
rs3131574		
rs3131568		
rs10152326		
rs11417752		
rs12593481		
rs3131560		
rs3131566		
rs3098167		

**Table 20 - Annotations of regulatory marks – (H3K4me3 and CTCF) in testis, overlapping EPSTI1-rs12870438 and their proxies.**

SNP	H3K4me3	CTCF
	Testis	
rs12870438		
rs58357177		
rs1535898		
rs1535897		
rs1830780		
rs63351261		
rs9594829		
rs11431398		
rs1535900		
rs34525682		
rs10161854		
rs9590722		
rs12870885		
rs9594827		
rs9594826		
rs1044856		
rs71099806		

USP8-rs7174015 and its proxies rs4318151, rs12593481 and rs3098167 are noted as *loci* that overlap H3K4me3 marks, while rs56398519 and rs10152236 overlap a CTCF binding site, a protein involved in the conformation of the topologically associated chromatin domains. For EPSTI1-rs12870438, no annotations were observed, however its proxies rs9590722, rs58357177, rs1535900, rs34525682 and rs12870885 are annotated as being related to regulatory marks. The former (EPSTI1-rs12870438) overlaps with H3K4me3 and the remaining three overlap CTCF binding sites. The output results from UCSC are shown in **Figures S6, S7, S8, S9, S10, S11, S12, S13, S14, S15 and S16** in the Supplementary Materials.



## Discussion

Male infertility affects approximately 7% of males of reproductive age and, in half, a cause cannot be identified, falling in the idiopathic infertility umbrella definition. However, it is predictable that many idiopathic male factor infertility cases have some unidentified genetic basis<sup>8</sup>. Understanding the genetic, genomic, and epigenetic factors affecting male fertility is important and remains an area of active investigation<sup>56</sup>, since infertility is a heterogeneous disease and successful spermatogenesis requires the proper functioning and interaction of thousands of genes. This study aimed at an evaluation of the possible implication of five SNPs previously associated with family size in a Hutterite population<sup>39</sup>, in spermatogenic failure with a case-control study design in the Portuguese population.

The results obtained showed a trend towards association for three of the SNPs analysed, with USP8-rs7174015 showing a nominal association with NOA, even though not significant following multiple testing correction. Assuming a recessive model, the A allele of USP8-rs7174015 is likely to confer risk to this phenotype, while EPST11-rs12870438 and PSAT1-rs7867029 demonstrated a trend towards association, with their minor alleles (A and C, respectively), conferring protection against SO. Therefore, major alleles at the latter loci represent risk alleles for decreased fertility and may lead to different phenotypes depending of the individual genetic background.

The selected SNP USP8-rs7174015 is located in an intron of the *Ubiquitin Specific Peptidase 8 (USP8)*, a gene that encodes a crucial enzyme for deubiquitinating proteins and sorting endosomal cargo in spermatogenic cells, being highly expressed in male germ cells<sup>57</sup>. This gene has also an important role in shaping the sperm head through direct interaction with other sorting complexes and in assembling the acrosome in differentiating sperm cells – acrosome biogenesis<sup>58,57</sup>. USP8 is also involved in several physiological functions in the cell such as cell-cycle regulation (including spermatogenesis), apoptosis and DNA repair<sup>59</sup>. Overall, *USP8* is a compelling candidate for a male fertility gene<sup>39</sup>. The results obtained for the studied variant in this gene were suggestive that, assuming a recessive model, the A allele confers risk to the most extreme male infertility phenotype – NOA – what seems to be in concordance with the *in-silico* results. The *Odds Ratio*, obtained for this nominal association, are 1.88 a large effect for which our study was well powered (Expected power E [0.984- 0.994], CaTS power calculator).

The lead SNP USP8-rs7174015 and several of its proxies are annotated as eQTLs in the testis, being involved in the expression of *USP8*, *USP50*, *AP4E1* and *RP11-562A8.5*. To the best of our knowledge there is no described evidence for the

involvement of these genes in spermatogenesis, other than *USP8* as noted earlier, but all of them are highly expressed in testis, with *USP50* showing testis-specific expression. Seen at a finer scale, the single-cell studies revealed that *USP8* is mostly expressed in spermatogonia, spermatocytes and Sertoli cells and *USP50* is almost exclusively expressed in spermatids. *AP4E1* has a diffuse expression in multiple testicular cell types, what could also suggest a role in spermatogenesis, encoding an important protein for different stages of this process. Thus, the risk allele (A) at *USP8*-rs7174015 may influence the spermatogenic process through changes in the expression level of these targets, leading to a higher predisposition to NOA. This deregulation may have several underlying mechanisms, given that this variant and some of its proxies overlap H3K4me3 marks, as well as CTCF binding sites in testis, and are annotated as modifiers of binding sites of some transcription factors related, direct or indirectly, with spermatogenesis.

Overall, the position of the lead variant and its proxies overlapping active promoters (H3K4me3) and active enhancers (H3K37ac and H3K4me1) in several tissues, including testis, suggests functional relevance in different biological processes, and a possible regulatory role in spermatogenic processes. In testis, *USP8*-rs7174015 for example, is mapped within a region enriched in H3K4me3 – an epigenetic modification of histone H3 (methylation) that is associated with open chromatin, making the DNA more accessible for the transcription factors, allowing the transcription process and thus gene expression to occur. The presence of the risk allele A may influence the binding of methyltransferases and CTCF binding, keeping the chromatin condensed and breaking the transcription process of the nearby genes.

The proxies of *USP8*-rs7174015 are also annotated within binding sites of several other transcription factors important to fertility. The variants rs2289108 and rs125593481 were found as putative modifiers of the binding site of YY1. The presence of the risk allele could decrease the binding affinity of this TF. YY1 is a transcription factor reported to play an important role in the maturation of spermatogonial stem cells, being expressed in spermatocytes, spermatogonia and spermatids but not in mature spermatozoa. This is an important transcription factor in the regulation of the spermatogenic process<sup>54,55</sup>. FOXJ1, another transcription factor, is specifically required for the formation of motile cilia, being reported as a key in the pathway of sperm motility and flagellum morphogenesis in murine models<sup>53</sup>. The binding sites for this factor are predicted to be changed by three of the proxies – rs56398519, rs3131559 and rs8026653. Finally, rs3098171, rs28582911 and rs34639682 overlap binding sites for HSF1, SIX5 and BCL6, respectively. HSF proteins are expressed during spermatogenesis, mainly in spermatocytes and spermatids<sup>60</sup>. *HSF1* is located within the azoospermia factor b (AZFb), an important region of the Y chromosome<sup>61,62</sup> that, when deleted, results in

severe male infertility. These proteins have been reported as important in the repression of sex chromatin during meiosis and the disruption of *HSF1* leads to a complete absence of mature sperm in mice, leading to sterility<sup>63</sup>. *SIX5* was reported to decrease c-kit levels in adult mice, causing an elevated spermatogenic cell apoptosis and Leydig cell hyperproliferation<sup>52</sup>. *BCL6*, annotated as a transcriptional repressor, is also related with cell apoptosis. The depletion of this TF causes testicular germ cell apoptosis in murine models<sup>50</sup>. The presence of the risk allele of the above mentioned variants, leading to a modification or abrogation of the TFBS for these TFs, may be associated with NOA, once their function is important to the spermatogenic process, justifying the nominal association obtained. *FOXJ1* could be related with astenozoospermia, characterized by alterations in sperm motility, since their function are related to the formation of cilia<sup>53</sup>, which is not the focus of our study, but may have a role in spermatogenesis. rs3098177 is located in a binding site of *PAX5*, which is another relevant transcription factor for the regulation of spermatogenesis<sup>64</sup>.

Two other trends towards association, this turn with severe oligozoospermia, were obtained for EPSTI1-rs12870438, an intronic variant of the *Epithelial Stromal Interaction Protein 1 (EPSTI1)*, and for PSAT1-rs7867029, downstream the *Phosphoserine Aminotransferase 1 (PSAT1)* gene. Assuming a recessive model for EPSTI1-rs12870438, the allele G is noted to be a risk factor, as well as the allele G of the variant PSAT1-rs7867029, but assuming a dominant model in this case. The Odds Ratio are lower in both cases, since the alleles are associated as protector factors (OR<1) rather than risk factor as in USP8-rs7174015.

Even though *EPSTI1* is highly expressed in testis, its function in this tissue is still unknown. It may be related with immune response because the transcript levels in peripheral blood correlate with lymphocyte counts<sup>65,66</sup>. The *PSAT1* variant analysed had been already associated with the risk of SO in a Japanese population<sup>67</sup> but there are no other studies on its influence in spermatogenesis.

None of these two variants or their proxies are annotated as being eQTL or sQTL in testis or located in a predictive coding effect region, CpG island or miRNAs. However, some proxies of EPSTI1-rs12870438, are located in regions with high H3K4me levels or CTCF binding sites. rs9590722 is located in an open chromatin region and rs58357177, rs1535900, rs34525682 and rs12870885 overlap CTCF binding sites. The presence of the risk alleles at those sites can block CTCF binding preventing or modifying the transcription of the genes that are regulated by this region. One proxy of this variant is also located in a binding site of an important transcription factor. rs10161854 overlaps a TFBS of *YY1*, described earlier. The presence of the A allele could prevent the ligation

of this TF, modifying the expression of some genes involved in the maturation of spermatogonial sperm cells leading in spermatogenic failure.

EPSTI1-rs12870438 showed different allele frequencies in the SO compared to the control and the azoospermia population (the latter two presented similar allele and genotype frequencies). Therefore, this SNP may be a potential diagnostic marker that allow the differentiation of SO from NOA phenotypes.

The last trend of association, obtained for PSAT1-rs7867029 with SO phenotype, could be explained by its location in a binding site of a protein of the family DMRT. DMRT are a family of testis-specific transcription factors responsible for the control of testis development and male germ cell proliferation<sup>51</sup>. However, besides one the highest DMRT4 expression was observed in adult testis, mice homozygous for a putative null mutation in DMRT4 showed normal development and fertility and thus this protein is not essential for murine viability and fertility<sup>68</sup>.

## Conclusion

Approximately half of male factor infertility cases have no known cause and are considered as idiopathic infertility. However, it is likely that most idiopathic male factor infertility cases have some unidentified genetic basis. In order to attempt to understand the genetic etiology of this condition, several studies, both genome wide association and candidate genes studies have been carried out in the last decade.

There are still few loci associated with spermatogenic failure, namely azoospermia and severe oligozoospermia, so the aim of this study was to evaluate whether 5 SNPs – USP8-rs7174015, TUSC1-rs10966811, EPSTI1-rs12870438, PSAT1-rs7867029 and DPF3-rs10129954 - previously associated with family size in an Hutterites population, are also involved in the genetic risk to azoospermia and/or severe oligozoospermia, and how it could influence, leading to spermatogenic failure. Pinpointing these risk alleles, their variants and even their genes could be used to create a multigene panel testing for male infertility. The research to validate them and identify their functions is also relevant and important to make this putative panel testing more reliable.

In this study we observed a trend towards association for three of the analyzed SNPs and a nominal association for one of them. The A allele (risk factor) of USP8-rs7174015 located in *USP8*, was nominally associated with azoospermia. There is a borderline tendency for association of the allele A (protective factor) of the intronic SNP at *EPSTI1*, EPSTI1-rs12870438, with severe oligozoospermia. The C allele (protective factor) of PSAT1-rs7867029 shows also a non-significant tendency towards severe oligozoospermia. All the variants are localized in LD blocks with a high probability to be genomic regions relevant for regulatory processes related with the spermatogenic process, being located in regions with histone modifications, binding sites of CTCF or transcription factors, important for spermatogenesis, binding sites. The *in silico* analysis showed also the eQTL effect of some variants associated with USP8-rs7174015, regulating genes with high expression in testis. Overall deregulation of gene expression seems to be an important mechanism underlying the genetic susceptibility to the studied phenotypes, what makes them good putative markers to add to the panel testing.

The larger study in which this thesis is included in has been submitted for publication and replicated and strengthened the results presented in the Portuguese population, confirming the trends observed for the described markers.

## Bibliography

1. Vander Borcht, M. & Wyns, C. Fertility and infertility: Definition and epidemiology. *Clin. Biochem.* **62**, 2–10 (2018).
2. Zegers-Hochschild, F. *et al.* International Committee for Monitoring Assisted Reproductive Technology (ICMART) and the World Health Organization (WHO) revised glossary of ART terminology, 2009\*. *Fertil. Steril.* **92**, 1520–1524 (2009).
3. Babakhanzadeh, E., Nazari, M., Ghasemifar, S. & Khodadadian, A. Some of the factors involved in male infertility: A prospective review. *Int. J. Gen. Med.* **13**, 29–41 (2020).
4. Sharlip, I. D. *et al.* Best practice policies for male infertility. *Fertil. Steril.* **77**, 873–882 (2002).
5. Agarwal, A., Mulgund, A., Hamada, A. & Chyatte, M. R. A unique view on male infertility around the globe. *Reprod. Biol. Endocrinol.* **13**, 1–9 (2015).
6. Ibtisham, F. *et al.* Progress and future prospect of in vitro spermatogenesis. *Oncotarget* **8**, 66709–66727 (2017).
7. Thoma, M. E. *et al.* Prevalence of infertility in the United States as estimated by the current duration approach and a traditional constructed approach. *Fertil. Steril.* **99**, 1324–1331 (2013).
8. Aston, K. I. Genetic susceptibility to male infertility: News from genome-wide association studies. *Andrology* **2**, 315–321 (2014).
9. Kumar, N. & Singh, A. Trends of male factor infertility, an important cause of infertility: A review of literature. *J. Hum. Reprod. Sci.* **8**, 191–196 (2015).
10. Masoumi, S. Z. *et al.* An Epidemologic Survey On The Causes Of Infertility In Patien Referred To Infertility Center In Fatimieh Hospital In Hamadan. Iranian Journal Reproductive Medicine. *Iran J Reprod Med* **13**, 513–516 (2015).
11. Vanputte, C. *et al.* *Seeley's Anatomy & Physiology.* (2019).
12. Griswold, M. D. Spermatogenesis: The commitment to Meiosis. *Physiol. Rev.* **96**, 1–17 (2016).
13. Rhoades, R. A. . & Bell, D. R. *Medical Physiology. Principles for Clinical Medicine.* *Lippincott Williams & Wilkins* **53**, (2013).

14. Nishimura, H. & L'Hernault, S. W. Spermatogenesis. *Curr. Biol. Mag.* **27**, R988–R994 (2017).
15. Donnell, L. O., Stanton, P. & Kretser, D. M. De. Endocrinology of the Male Reproductive System and Spermatogenesis. *Endotex [Internet]* (2017).
16. Matzuk, M. M. & Lamb, D. J. The biology of infertility: research advances and clinical challenges. *Nat. Med.* **14**, 1197–1213 (2008).
17. Tournaye, H., Krausz, C. & Oates, R. D. Novel concepts in the aetiology of male reproductive impairment. *Lancet Diabetes Endocrinol.* **5**, 544–553 (2017).
18. Gurney, J. K. *et al.* Risk factors for cryptorchidism. *Nature Reviews Urology* **176**, 139–148 (2017).
19. Berookhim, B. M. & Schlegel, P. N. Azoospermia due to spermatogenic failure. *Urol. Clin. North Am.* **41**, 97–113 (2014).
20. Santana, V. P., Miranda-Furtado, C. L., de Oliveira-Gennaro, F. G. & dos Reis, R. M. Genetics and epigenetics of varicocele pathophysiology: an overview. *J. Assist. Reprod. Genet.* **34**, 839–847 (2017).
21. Czaplicki, M., Bablok, L. & Janczewski, Z. Varicocelectomy in patients with azoospermia. *J. Reprod. Syst.* **3**, 51–55 (1979).
22. Aston, K. I., Krausz, C., Laface, I., Ruiz-Castané, E. & Carrell, D. T. Evaluation of 172 candidate polymorphisms for association with oligozoospermia or azoospermia in a large cohort of men of European descent. *Hum. Reprod.* **25**, 1383–1397 (2010).
23. Hwang, K. *et al.* Evaluation of the azoospermic male: a committee opinion. *Fertil. Steril.* **109**, 777–782 (2018).
24. Oates, R. Evaluation of the azoospermic male. *Asian J. Androl.* **14**, 82–87 (2012).
25. Krausz, C. & Riera-Escamilla, A. Genetics of male infertility. *Nat. Rev. Urol.* **15**, 369–384 (2018).
26. Reijo, R., Alagappan, R. K., Patrizio, P. & Page, D. C. Severe oligozoospermia resulting from deletions of azoospermia factor gene on Y chromosome. *Lancet* **347**, 1290–1293 (1996).
27. Committee, P. & Society, A. Evaluation of the azoospermic male. *Fertil. Steril.* **90**, 74–77 (2008).

28. Tiseo, B. C., Hayden, R. P. & Tanrikut, C. Surgical management of nonobstructive azoospermia. *Asian J. Urol.* **2**, 85–91 (2015).
29. Carlberg, C., Ulven, S. M. & Molnár, F. *Human Genomic Variation. Nutrigenomics* (2016). doi:10.1007/978-3-319-30415-1
30. Barnes, M. R. & Breen, G. Genetic variation analysis for biomedical researchers: a primer. *Methods Mol. Biol.* **628**, 1–20 (2010).
31. Halldórsson, B. V. & Sharan, R. Network-based interpretation of genomic variation data. *J. Mol. Biol.* **425**, 3964–3969 (2013).
32. Suh, Y. & Vijg, J. SNP discovery in associating genetic variation with human disease phenotypes. *Mutat. Res. - Fundam. Mol. Mech. Mutagen.* **573**, 41–53 (2005).
33. Communications, L. H. N. C. for B., Medicine, U. S. N. L. of, Health, N. I. of & Services, D. of H. & H. *Genomic Research.* (2020).
34. Ramírez-Bello, J. & Jiménez-Morales, M. [Functional implications of single nucleotide polymorphisms (SNPs) in protein-coding and non-coding RNA genes in multifactorial diseases]. *Gac. Med. Mex.* **153**, 238–250 (2017).
35. Brookes, A. J. The essence of SNPs. *Gene* **234**, 177–186 (1999).
36. Aston, K. I. & Carrell, D. T. Genome-wide study of single-nucleotide polymorphisms associated with azoospermia and severe oligozoospermia. *J. Androl.* **30**, 711–725 (2009).
37. Hu, Z. *et al.* A genome-wide association study in Chinese men identifies three risk loci for non-obstructive azoospermia. *Nat. Genet.* **44**, 183–186 (2012).
38. Zhao, H. *et al.* A genome-wide association study reveals that variants within the HLA region are associated with risk for nonobstructive azoospermia. *Am. J. Hum. Genet.* **90**, 900–906 (2012).
39. Kosova, G., Scott, N. M., Niederberger, C., Prins, G. S. & Ober, C. Genome-wide association study identifies candidate genes for male fertility traits in humans. *Am. J. Hum. Genet.* **90**, 950–961 (2012).
40. Ober, C., Hyslop, T. & Hauck, W. W. Inbreeding effects on fertility in humans: Evidence for reproductive compensation. *Am. J. Hum. Genet.* **64**, 225–231 (1999).



41. Cooper, T. G. *et al.* World Health Organization reference values for human semen characteristics. *Hum. Reprod. Update* **16**, 231–245 (2009).
42. Skol, A. D., Scott, L. J., Abecasis, G. R. & Boehnke, M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat. Genet.* **38**, 209–213 (2006).
43. Hill, A. *et al.* Stepwise distributed open innovation contests for software development: Acceleration of genome-wide association analysis. *Gigascience* **6**, 1–10 (2017).
44. Machiela, M. J. & Chanock, S. J. LDlink: A web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* **31**, 3555–3557 (2015).
45. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
46. Oscanoa, J. *et al.* SNPnexus: a web server for functional annotation of human genome sequence variation (2020 update). *Nucleic Acids Res.* **48**, W185–W192 (2020).
47. Ward, L. D. & Kellis, M. HaploReg: A resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40**, 930–934 (2012).
48. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22**, 1790–1797 (2012).
49. Guo, J. *et al.* The Dynamic Transcriptional Cell Atlas of Testis Development during Human Puberty. *Cell Stem Cell* **26**, 262-276.e4 (2020).
50. Kojima, S. *et al.* Testicular germ cell apoptosis in Bcl6-deficient mice. *Development* **128**, 57–65 (2001).
51. Zhang, T., Zarkower, D. & Biology, C. DMRT and mammalian spermatogenesis. 195–202 (2015). doi:10.1016/j.scr.2017.07.026.DMRT
52. Sarkar, P. S., Paul, S., Han, J. & Reddy, S. Six5 is required for spermatogenic cell survival and spermiogenesis. *Hum. Mol. Genet.* **13**, 1421–1431 (2004).
53. Beckers, A. *et al.* The FOXJ1 target Cfap206 is required for sperm motility, mucociliary clearance of the airways and brain development. *Development* **147**, (2020).

54. Kim, J. S., Chae, J. H., Cheon, Y. P. & Kim, C. G. Reciprocal localization of transcription factors YY1 and CP2c in spermatogonial stem cells and their putative roles during spermatogenesis. *Acta Histochem.* **118**, 685–692 (2016).
55. Bajusz, I., Henry, S., Sutus, E., Kovács, G. & Pirity, M. K. Evolving role of RING1 and YY1 binding protein in the regulation of germ-cell-specific transcription. *Genes (Basel)*. **10**, 1–33 (2019).
56. Thirumavalavan, N., Gabrielsen, J. S. & Lamb, D. J. Where are we going with gene screening for male infertility? *Fertil. Steril.* **111**, 842–850 (2019).
57. Berruti, G., Ripolone, M. & Ceriani, M. USP8, a regulator of endosomal sorting, is involved in mouse acrosome biogenesis through interaction with the spermatid ESCRT-0 complex and microtubules. *Biol. Reprod.* **82**, 930–939 (2010).
58. Nakamura, N. Ubiquitination Regulates the Morphogenesis and Function of Sperm Organelles. *Cells* **2**, 732–750 (2013).
59. Kimura, Y. & Tanaka, K. Regulatory mechanisms involved in the control of ubiquitin homeostasis. *J. Biochem.* **147**, 793–798 (2010).
60. Steven S. Witkin, Tomi T. Kanninen, and G. S. The Role of Heat Shock Proteins in Reproductive System Development and Function. *Adv. Anatomy, Embryol. Cell Biol.* **222**, 1–27 (2017).
61. Shinkay, T. *et al.* Molecular characterization of heat shock-like factor encoded on the human Y chromosome, and implications for male infertility. *Biol. Reprod.* **71**, 297–306 (2004).
62. Tessari, A. *et al.* Characterization of HSFY, a novel AZFb gene on the Y chromosome with a possible role in human spermatogenesis. *Mol. Hum. Reprod.* **10**, 253–258 (2004).
63. Åkerfelt, M. *et al.* Heat shock transcription factor 1 localizes to sex chromatin during meiotic repression. *J. Biol. Chem.* **285**, 34469–34476 (2010).
64. Adams, B. *et al.* Pax-5 encodes the transcription factor BSAP and is expressed in B lymphocytes, the developing CNS, and adult testis. *Genes Dev.* **6**, 1589–1607 (1992).
65. Nielsen, H. L., Ronnov-Jessen, L., Villadsen, R. & Petersen, O. W. Identification of EPST11, a novel gene induced by epithelial-stromal interaction in human breast cancer. *Genomics* **79**, 703–710 (2002).

66. Buess, M. *et al.* Characterization of heterotypic interaction effects in vitro to deconvolute global gene expression profiles in cancer. *Genome Biol.* **8**, (2007).
67. Sato, Y. *et al.* An association study of four candidate loci for human male fertility traits with male infertility. *Hum. Reprod.* **30**, 1510–1514 (2015).
68. Balciuniene, J., Bardwell, V. J. & Zarkower, D. Mice Mutant in the DM Domain Gene *Dmrt4* Are Viable and Fertile but Have Polyovular Follicles. *Mol. Cell. Biol.* **26**, 8984–8991 (2006).

## Supplementary Materials

**Table S1 - Primers used in this study.**

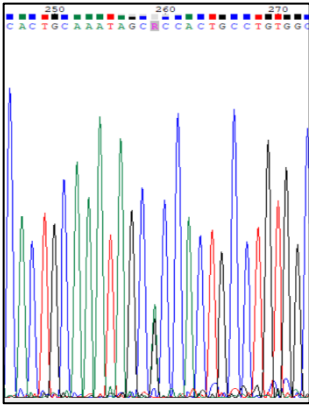
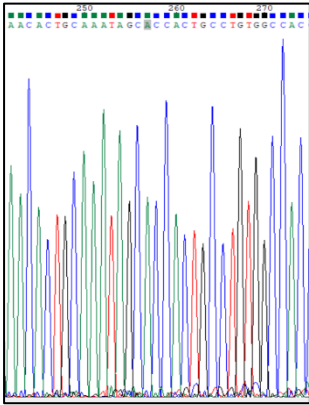
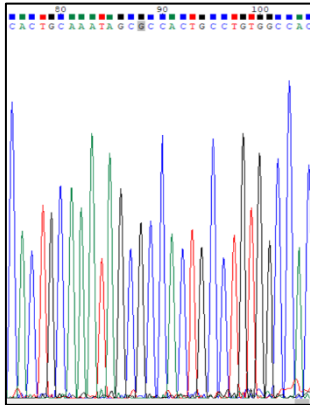
SNP	Primer	Length (bp)	GC%	Tm (°C)	Sequence	Fragment length
rs10966811	F	22	40,91	58,93	CATTCACAGCACTGAGAAATCA	225 bp
	R	20	50,00	59,48	GCAAAGTGAAAGGAGGCAGT	
rs7867029	F	21	47,62	60,18	ACCCAAGTCATTCTCCATTC	177 bp
	R	26	30,77	58,81	ACTTAGGATATAAACTGCGATGAAA	
rs12870438	F	25	40,00	59,88	CCATCCTAGAGATTGAAAGTATGGA	206 bp
	R	23	39,13	60,20	TGGCAACCCCTCTTCTTATTATT	
rs10129954	F	20	55,00	62,43	TGGACAAGCTGTCACCAAGC	228 bp
	R	21	47,62	62,15	TGCAGTAAGCCATGATTGTGC	
rs7174015	F1	23	43,48	58,16	ACACTTATACCCGACTTTGACCT	180 bp
	R1	25	40,00	57,54	CTCTGACAAGCTTATAGGGTCTTTA	
	F2	20	60,00	60,01	GCCCTAGCACCTGTTCTCTG	916 bp
	R2	20	40,00	60,21	CGGCTGAAATGCAAAAATCT	

**Table S2 - TaqMan probes used in this study.**

SNP	Context Sequence [VIC/FAM]	Design Strand	Allele 1	Dye	Allele 2	Dye
rs10966811	CCGTATTTATTGTCAATTACTCTCC [A/G] TTTTTAACTACTTCACAGGCAA	F	C_26249696_10_V	VIC	C_26249696_10_M	FAM
rs7867029	CAAGTAATAGTTCATATTGCCACAT [C/G] ATTTGAAATATATCATCTATTAGTT	F	C_31364474_20_V	VIC	C_31364474_20_M	FAM
rs7174015	TCTGCCTACAATCCCAGGCCTTACT [G/A] TAGCTCCTAAAAGTGTTCAGTT	F	C_32072246_20_V	VIC	C_32072246_20_M	FAM
rs10129954	ATGAGGAGTTTGGGTTGTATTCAA [C/T] GTGATGAAAGTGTGAGAGCATGAG	R	C_30534824_10_V	VIC	C_30534824_10_M	FAM
rs12870438	AAATGAGGACAACACTGCAAATAGC [A/G] CCTACTGCCTGTGGCCACCCGATGCA	F	C_3123309_10_V	VIC	C_3123309_10_M	FAM

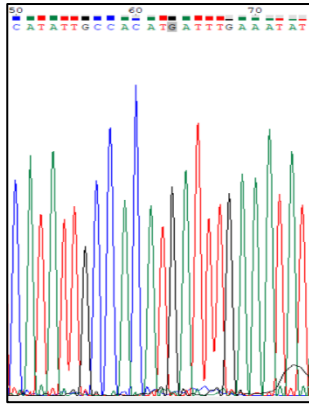
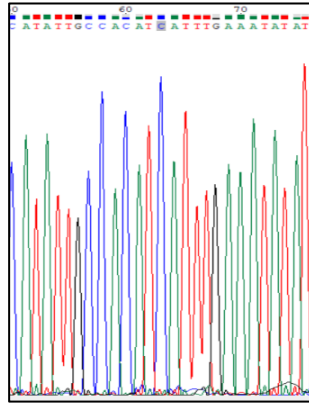
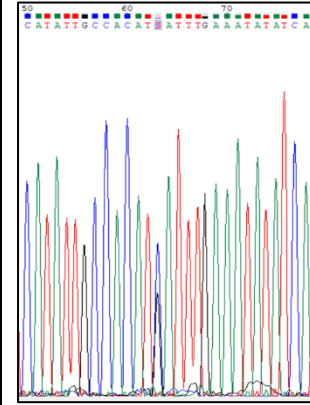
For each SNP, two different probes were necessary in order to identify the two alternative alleles. The specific sequence of the probes was not known since they are designed by the manufacturers.

**Table S3 - Sanger Sequencing of reference individuals for EPST11-rs12870438**

rs12870438			
Sample	Y1621	Y1627	Y1638
Genotype obtained by Sanger Sequencing	AG	AA	GG
			
Expected Genotyping Result	Allele 1/ Allele 2 (VIC/ FAM)	Allele 1/ Allele 1 (VIC/ VIC)	Allele 2/ Allele 2 (FAM/ FAM)

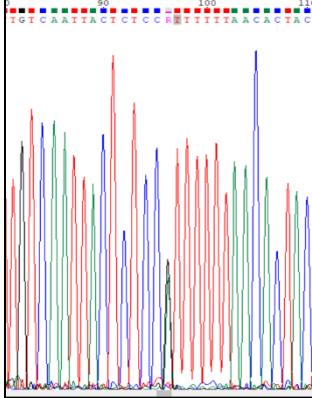
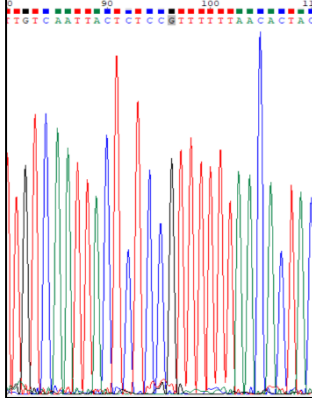
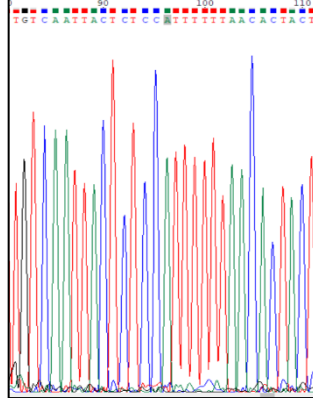
The electropherogram and expected genotyping results for EPST11-rs12870438 for three individuals of reference with the possible genotypes

**Table S4 - Sanger Sequencing of reference individuals for PSAT1-rs7867029**

rs7867029			
Sample	Y1462	YF115	YF129
Genotype obtained by Sanger Sequencing	GG	CC	CG
			
Expected Genotyping Result	Allele 2/ Allele 2 (FAM/ FAM)	Allele 1/ Allele 1 (VIC/ VIC)	Allele 1/ Allele 2 (VIC/ FAM)

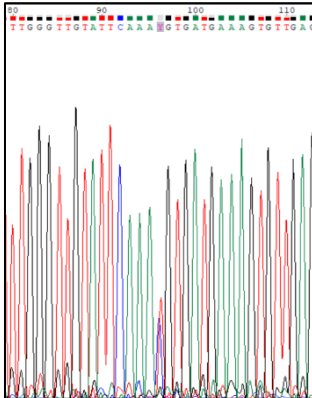
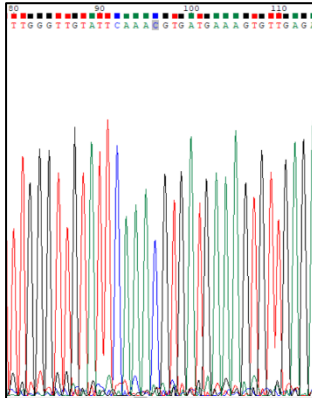
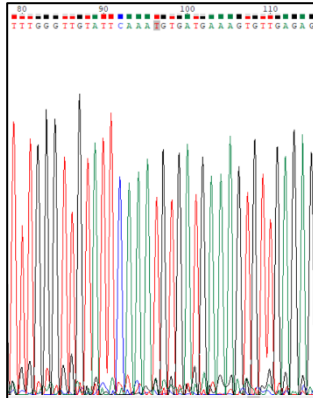
The electropherogram and expected genotyping results for PSAT1-rs7867029 for three individuals of reference with the possible genotypes

**Table S5 - Sanger Sequencing of reference individuals for rs10966811**

rs10966811			
Sample	Y1462	Y1474	Y1506
Genotype obtained by Sanger Sequencing	GA	GG	AA
			
Expected Genotyping Result	Allele 1/ Allele 2 (VIC/ FAM)	Allele 1/ Allele 1 (VIC/ VIC)	Allele 2/ Allele 2 (FAM/ FAM)

The electropherogram and expected genotyping results for rs10966811 for three individuals of reference with the possible genotypes

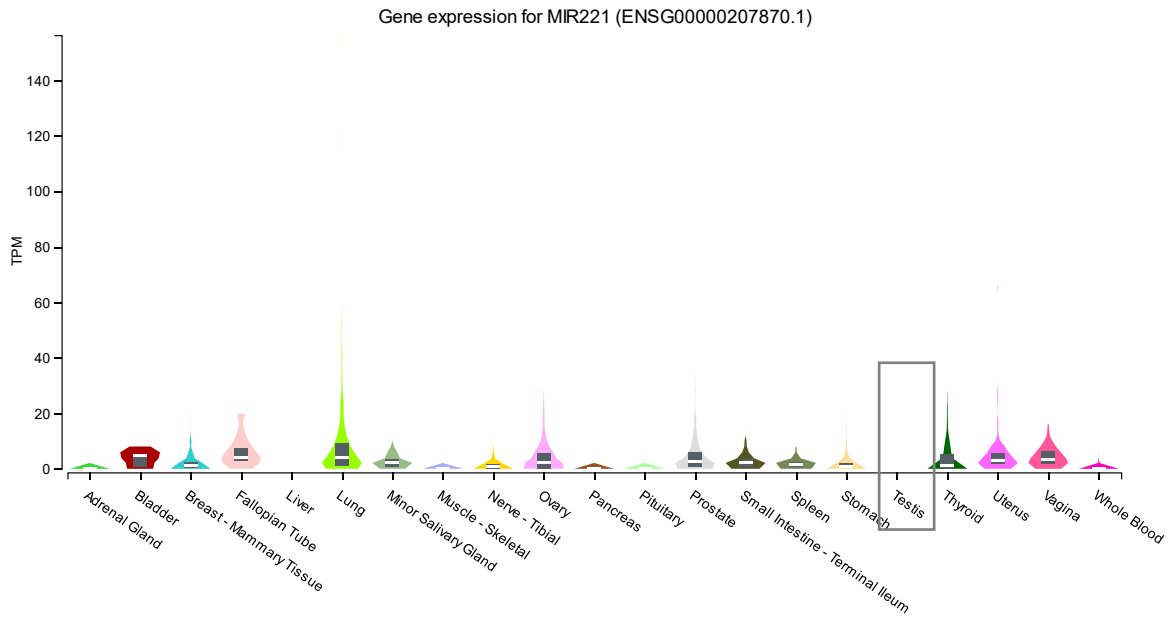
**Table S6 - Sanger Sequencing of reference individuals for DPF3-rs10129954**

rs10129954			
Sample	Y1462	Y1471	Y1519
Genotype obtained by Sanger Sequencing	CT	CC	TT
			
Expected Genotyping Result	Allele 1/ Allele 2 (VIC/ FAM)	Allele 1/ Allele 1 (VIC/ VIC)	Allele 2/ Allele 2 (FAM/ FAM)

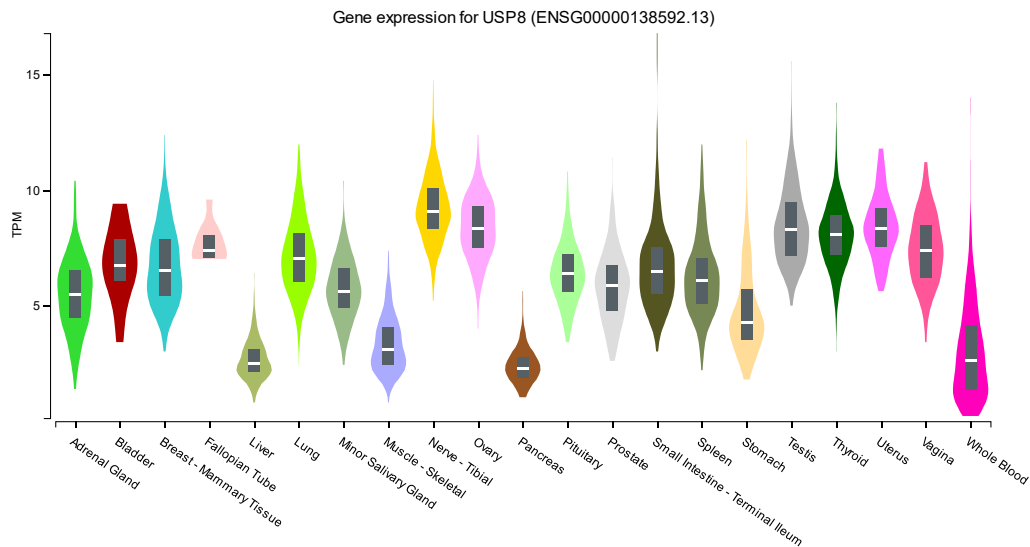
The electropherogram and expected genotyping results for DPF3-RS for three individuals of reference with the possible genotypes

**Table S7 - Estimation of the statistical power of our study for 380 patients and 130 controls.**

SNP	Ref allele	IBS MAF	Expected Odds Ratio (Genotype Relative Risk)									Expected Power
			1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	
rs10966811	A	0.481	0.103	0.249	0.454	0.656	0.811	0.908	0.959	0.983	0.994	
rs7867029	C	0.182	0.083	0.178	0.327	0.504	0.674	0.808	0.899	0.952	0.979	
rs12870438	A	0.346	0.099	0.238	0.438	0.643	0.805	0.907	0.961	0.985	0.995	
rs10129954	T	0.481	0.103	0.249	0.454	0.656	0.811	0.908	0.959	0.983	0.994	
rs7174015	A	0.472	0.103	0.250	0.454	0.657	0.812	0.909	0.960	0.984	0.994	



**Figure S1 - Gene Expression of MIR221 in human different tissues of the GTEx database.**



**Figure S2 - USP8 gene expression in several human tissues of the GTEx database.**

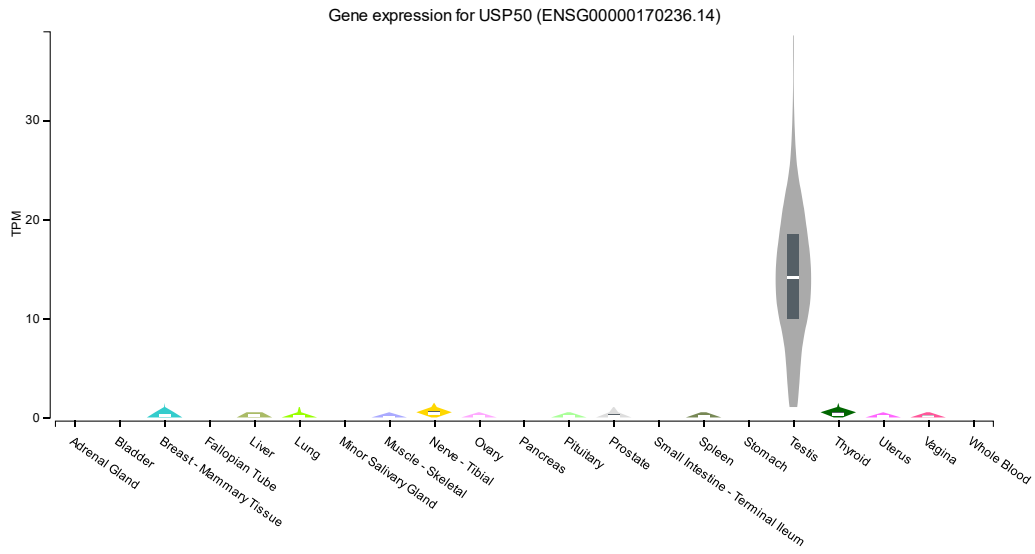


Figure S3 - Human testis - specific expression of USP50 of the GTEx database.

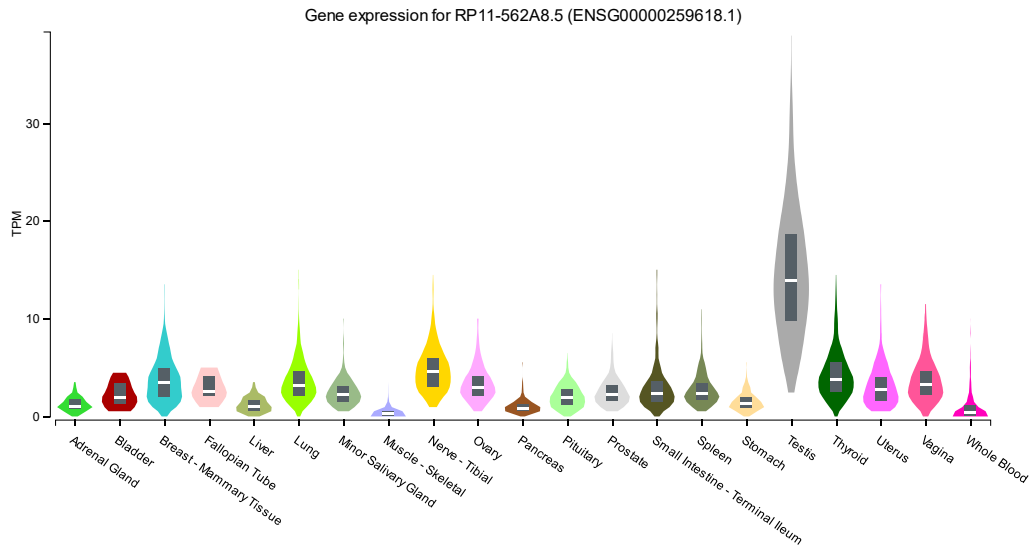
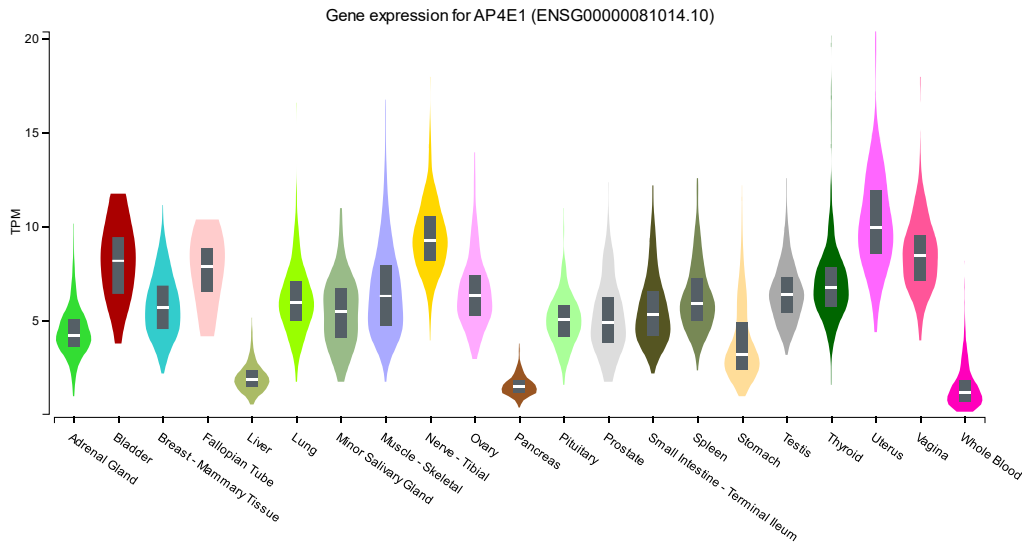


Figure S4 - RP11-582A8.5 gene expression in human different tissues of the GTEx database. Higher expression in testis.





**Figure S5 - Gene expression of AP4E1 in human different tissues of the GTEx database.**

**Table S8 - RegulomeDB scoring scheme.**

Score	Supporting data
1a	eQTL + TF binding + matched TF motif + matched DNase Footprint + DNase peak.
1b	eQTL + TF binding + any motif + DNase Footprint + DNase peak.
1c	eQTL + TF binding + matched TF motif + DNase peak.
1d	eQTL + TF binding + any motif + DNase peak.
1e	eQTL + TF binding + matched TF motif.
1f	eQTL + TF binding / DNase peak.
2a	TF binding + matched TF motif + matched DNase Footprint + DNase peak.
2b	TF binding + any motif + DNase Footprint + DNase peak.
2c	TF binding + matched TF motif + DNase peak.
3a	TF binding + any motif + DNase peak.
3b	TF binding + matched TF motif.
4	TF binding + DNase peak.
5	TF binding or DNase peak.
6	Motif hit
7	Other

**Table S9 - Tools for generating functional prediction scores.**

Method	Predicted effect	Score range	Note
<b>CADD</b>	Benign to Pathogenic	[1, 99]	Score above 10 is considered for potentially pathogenic variants.
<b>fitCons</b>	Non-functional to Functional	[0, 1]	Higher scores indicating more potential for interesting genomic function.
<b>EIGEN</b>	Non-functional to Functional	[-5, 40]	With median score of around 0, higher scores indicating more likely to be functional.
<b>EIGEN-PC</b>	Non-functional to Functional	[-5, 100]	With median score of around 0, higher scores indicating more likely to be functional.
<b>FATHMM</b>	Deleterious to Neutral/Benign	[0, 1]	Scores above 0.5 are predicted to be deleterious. Scores close to the extremes (0 or 1) yield the highest accuracy.
<b>GWAVA</b>	Non-functional to Functional	[0, 1]	Higher scores indicating more likely to be functional.
<b>DeepSEA</b>	Functional to Non-functional	[0, 1]	Lower scores indicating higher likelihood of functional significance.
<b>FunSeq2</b>	Non-functional to Functional	[0, 6]	Higher scores indicating more likely to be functional.
<b>ReMM</b>	Non-deleterious to Deleterious	[0, 1]	Higher scores indicating higher prediction of deleteriousness.

**Table S10 - GTEx data for sQTL for the lead variant USP8-rs7174015 and its proxies, in several tissues.**

GTEx data for sQTL		
SNP	Tissue	P-value
rs10152326	Cells_Cultured_fibroblasts	2.31E-07
	Muscle_Skeletal	2.06E-08
	Muscle_Skeletal	9.77E-08
	Nerve_Tibial	1.87E-07
	Skin_Sun_Exposed_Lower_leg	2.26E-06
	Thyroid	1.45E-06
rs11070776	Artery_Tibial	4.15E-06
	Cells_Cultured_fibroblasts	2.61E-08
	Esophagus_Muscularis	2.99E-06
	Muscle_Skeletal	2.86E-10
	Muscle_Skeletal	6.24E-07
	Nerve_Tibial	6.99E-07
	Nerve_Tibial	5.58E-06
	Skin_Not_Sun_Exposed_Suprapubic	4.92E-07
	Skin_Sun_Exposed_Lower_leg	2.73E-07
	Thyroid	4.49E-06
rs12593481	Cells_Cultured_fibroblasts	1.82E-07
	Muscle_Skeletal	5.50E-09
	Muscle_Skeletal	9.54E-08
	Nerve_Tibial	2.72E-06
	Nerve_Tibial	1.44E-07
	Skin_Sun_Exposed_Lower_leg	8.87E-07
	Thyroid	1.01E-06
rs2289108	Cells_Cultured_fibroblasts	2.07E-07
	Muscle_Skeletal	1.97E-08
	Muscle_Skeletal	2.70E-07
	Nerve_Tibial	2.88E-06
	Nerve_Tibial	2.10E-06
	Skin_Not_Sun_Exposed_Suprapubic	9.18E-07
rs28582911	Artery_Tibial	3.78E-06
	Cells_Cultured_fibroblasts	1.98E-08
	Muscle_Skeletal	1.56E-10
	Muscle_Skeletal	6.28E-07
	Nerve_Tibial	8.30E-07
	Skin_Not_Sun_Exposed_Suprapubic	6.86E-07
	Skin_Sun_Exposed_Lower_leg	2.75E-07
	Thyroid	5.93E-06
rs3098167	Cells_Cultured_fibroblasts	9.00E-08
	Cells_EBV-transformed_lymphocytes	7.38E-07
	Muscle_Skeletal	2.25E-09
	Muscle_Skeletal	5.87E-08
	Nerve_Tibial	1.37E-06
	Nerve_Tibial	1.58E-07
	Skin_Sun_Exposed_Lower_leg	1.01E-06
	Thyroid	2.34E-06
rs3098169	Cells_Cultured_fibroblasts	2.19E-07
	Cells_EBV-transformed_lymphocytes	1.47E-06
	Muscle_Skeletal	8.85E-09
	Muscle_Skeletal	1.22E-07
	Nerve_Tibial	3.61E-06
	Nerve_Tibial	1.25E-07
	Skin_Sun_Exposed_Lower_leg	1.83E-06
	Thyroid	1.97E-06
rs3098171	Artery_Tibial	1.89E-06
	Cells_Cultured_fibroblasts	1.56E-08
	Esophagus_Muscularis	5.48E-06
	Lung	4.13E-06
	Muscle_Skeletal	2.21E-10
	Muscle_Skeletal	1.11E-06
	Nerve_Tibial	6.59E-07
	Nerve_Tibial	3.99E-06
	Skin_Not_Sun_Exposed_Suprapubic	3.19E-07
	Skin_Sun_Exposed_Lower_leg	3.10E-07
Thyroid	5.40E-06	
rs3098174	Artery_Tibial	1.36E-07
	Breast_Mammary_Tissue	1.83E-08
	Cells_Cultured_fibroblasts	4.55E-10
	Colon_Transverse	7.35E-07

	Esophagus_Mucosa	5.34E-06
	Esophagus_Muscularis	3.16E-07
	Lung	2.93E-08
	Muscle_Skeletal	2.08E-12
	Muscle_Skeletal	9.02E-06
	Nerve_Tibial	3.41E-08
	Nerve_Tibial	4.18E-06
	Skin_Not_Sun_Exposed_Suprapubic	6.09E-08
	Skin_Sun_Exposed_Lower_leg	3.33E-09
	Stomach	1.35E-06
	Thyroid	7.17E-07
	Thyroid	1.44E-06
	rs3098177	Artery_Tibial
Cells_Cultured_fibroblasts		1.28E-08
Esophagus_Muscularis		5.48E-06
Lung		4.13E-06
Muscle_Skeletal		1.86E-10
Muscle_Skeletal		1.02E-06
Nerve_Tibial		4.27E-07
Nerve_Tibial		3.94E-06
Skin_Not_Sun_Exposed_Suprapubic		3.19E-07
Skin_Sun_Exposed_Lower_leg		1.95E-07
rs3098205	Thyroid	5.40E-06
	Cells_Cultured_fibroblasts	9.79E-07
	Muscle_Skeletal	4.97E-07
	Muscle_Skeletal	7.07E-08
	Nerve_Tibial	4.06E-08
rs3131559	Thyroid	1.04E-06
	Artery_Tibial	1.36E-07
	Breast_Mammary_Tissue	1.83E-08
	Cells_Cultured_fibroblasts	4.55E-10
	Colon_Transverse	7.35E-07
	Esophagus_Mucosa	5.34E-06
	Esophagus_Muscularis	3.16E-07
	Lung	2.93E-08
	Muscle_Skeletal	2.08E-12
	Muscle_Skeletal	9.02E-06
	Nerve_Tibial	3.41E-08
	Nerve_Tibial	4.18E-06
	Skin_Not_Sun_Exposed_Suprapubic	6.09E-08
	Skin_Sun_Exposed_Lower_leg	3.33E-09
	Stomach	1.35E-06
rs3131560	Thyroid	7.17E-07
	Thyroid	1.44E-06
	Cells_Cultured_fibroblasts	9.00E-08
	Cells_EBV-transformed_lymphocytes	7.38E-07
	Muscle_Skeletal	2.25E-09
	Muscle_Skeletal	5.87E-08
	Nerve_Tibial	1.37E-06
	Nerve_Tibial	1.58E-07
rs3131562	Skin_Sun_Exposed_Lower_leg	1.01E-06
	Thyroid	2.34E-06
	Artery_Tibial	2.00E-07
	Breast_Mammary_Tissue	1.83E-08
	Cells_Cultured_fibroblasts	5.55E-10
	Colon_Transverse	7.35E-07
	Esophagus_Mucosa	5.34E-06
	Esophagus_Muscularis	3.16E-07
	Lung	2.93E-08
	Muscle_Skeletal	2.08E-12
	Muscle_Skeletal	9.02E-06
	Nerve_Tibial	3.41E-08
	Nerve_Tibial	4.18E-06
	Skin_Not_Sun_Exposed_Suprapubic	6.09E-08
	Skin_Sun_Exposed_Lower_leg	3.33E-09
rs3131566	Stomach	1.35E-06
	Thyroid	7.17E-07
	Thyroid	1.44E-06
	Cells_Cultured_fibroblasts	9.00E-08
rs3131566	Cells_EBV-transformed_lymphocytes	7.38E-07
	Muscle_Skeletal	2.25E-09
	Muscle_Skeletal	5.87E-08
	Muscle_Skeletal	5.87E-08

	Nerve_Tibial	1.37E-06
	Nerve_Tibial	1.58E-07
	Skin_Sun_Exposed_Lower_leg	1.01E-06
	Thyroid	2.34E-06
rs3131568	Cells_Cultured_fibroblasts	7.51E-08
	Cells_EBV-transformed_lymphocytes	1.47E-06
	Muscle_Skeletal	1.11E-09
	Muscle_Skeletal	4.46E-08
	Nerve_Tibial	8.27E-07
	Nerve_Tibial	1.90E-07
	Skin_Sun_Exposed_Lower_leg	6.18E-07
	Thyroid	3.62E-06
rs3131574	Cells_Cultured_fibroblasts	1.44E-06
	Muscle_Skeletal	3.90E-07
	Muscle_Skeletal	9.43E-08
	Nerve_Tibial	3.39E-08
	Thyroid	1.04E-06
rs4318151	Artery_Tibial	4.04E-07
	Breast_Mammary_Tissue	8.91E-08
	Cells_Cultured_fibroblasts	1.37E-09
	Esophagus_Muscularis	5.85E-07
	Lung	4.18E-08
	Muscle_Skeletal	7.76E-12
	Nerve_Tibial	3.35E-07
	Skin_Not_Sun_Exposed_Suprapubic	2.99E-07
	Skin_Sun_Exposed_Lower_leg	5.55E-09
	Stomach	1.17E-06
	Thyroid	1.90E-06
Thyroid	6.84E-07	
rs56398519	Cells_Cultured_fibroblasts	3.00E-07
	Muscle_Skeletal	6.98E-08
	Muscle_Skeletal	3.29E-07
	Nerve_Tibial	7.12E-06
rs7174015	Artery_Aorta	4.78E-06
	Artery_Tibial	2.44E-06
	Cells_Cultured_fibroblasts	1.86E-08
	Esophagus_Muscularis	3.36E-07
	Lung	3.37E-06
	Muscle_Skeletal	2.10E-10
	Muscle_Skeletal	2.82E-06
	Nerve_Tibial	1.67E-07
	Nerve_Tibial	3.22E-06
	Skin_Not_Sun_Exposed_Suprapubic	3.05E-07
	Skin_Sun_Exposed_Lower_leg	1.01E-07
	Thyroid	2.29E-06

**Table S11 - GTEx data of eQTL for the lead SNP USP8-rs7174015 and its proxies in several tissues, including testis.**

GTEx for eQTL			
SNP	Tissue	P-value	Gene
rs10152326	Artery_Tibial	5.44444E-05	
	Brain_Cerebellar_Hemisphere	0.000045581	
	Brain_Nucleus_accumbens_basal_ganglia	1.36543E-06	
	Cells_Cultured_fibroblasts	3.7709E-06	
	Cells_Cultured_fibroblasts	0.000234978	
	Esophagus_Mucosa	0.000138102	
	Esophagus_Muscularis	0.000092706	
	Muscle_Skeletal	2.58634E-05	
	Muscle_Skeletal	1.37386E-06	
	Skin_Not_Sun_Exposed_Suprapubic	0.000260063	
	Testis	1.71658E-05	AP4E1
	Testis	6.83245E-08	USP8
	Testis	1.18869E-05	USP50
	Testis	0.000174309	RP11-562A8.1
Whole_Blood	0.000057503		
rs11070776	Artery_Tibial	0.000173456	
	Brain_Cerebellar_Hemisphere	5.29961E-06	
	Brain_Cerebellum	0.000144612	
	Brain_Nucleus_accumbens_basal_ganglia	9.31896E-06	

	Cells_Cultured_fibroblasts	9.62283E-06	
	Esophagus_Mucosa	0.000191805	
	Esophagus_Muscularis	0.000127847	
	Muscle_Skeletal	4.74849E-05	
	Muscle_Skeletal	2.34823E-05	
	Prostate	9.52952E-06	
	Skin_Not_Sun_Exposed_Suprapubic	2.13244E-05	
	Testis	6.14497E-05	APE41
	Testis	5.7852E-10	USP8
	Testis	1.67704E-05	USP50
	Testis	0.000099852	RP11-562A8.1
	Whole_Blood	1.60573E-06	
rs12593481	Artery_Tibial	3.82589E-05	
	Brain_Cerebellar_Hemisphere	4.23929E-05	
	Brain_Nucleus_accumbens_basal_ganglia	1.11514E-06	
	Cells_Cultured_fibroblasts	3.89103E-06	
	Cells_Cultured_fibroblasts	9.29803E-05	
	Esophagus_Mucosa	0.000164062	
	Esophagus_Muscularis	0.000133552	
	Muscle_Skeletal	1.38579E-05	
	Muscle_Skeletal	9.61096E-07	
	Muscle_Skeletal	0.000204074	
	Skin_Not_Sun_Exposed_Suprapubic	0.000207451	
	Testis	1.01226E-05	AP4E1
	Testis	5.01266E-08	USP8
	Testis	1.90068E-05	USP50
	Testis	0.000160938	RP11-562A8.1
	Whole_Blood	4.39318E-05	
	rs2289108	Artery_Tibial	0.000084984
Brain_Cerebellar_Hemisphere		7.11687E-06	
Brain_Cerebellum		0.000123545	
Brain_Nucleus_accumbens_basal_ganglia		9.08476E-06	
Cells_Cultured_fibroblasts		5.32235E-07	
Esophagus_Mucosa		4.37724E-05	
Esophagus_Muscularis		0.000175841	
Muscle_Skeletal		0.000146417	
Muscle_Skeletal		8.93207E-06	
Prostate		6.52818E-05	
Skin_Not_Sun_Exposed_Suprapubic		1.13291E-05	
Testis		0.000129438	AP4E1
Testis		1.05666E-08	USP8
Testis		3.09802E-05	USP50
Thyroid		0.000519009	
Whole_Blood		1.10359E-06	
rs28582911		Artery_Tibial	0.000236294
	Brain_Cerebellar_Hemisphere	6.70094E-06	
	Brain_Cerebellum	0.000106084	
	Brain_Nucleus_accumbens_basal_ganglia	9.31896E-06	
	Cells_Cultured_fibroblasts	9.98435E-06	
	Esophagus_Mucosa	0.000210156	
	Esophagus_Muscularis	0.000117675	
	Muscle_Skeletal	6.82995E-05	
	Muscle_Skeletal	3.63363E-05	
	Prostate	9.59053E-06	
	Skin_Not_Sun_Exposed_Suprapubic	8.79114E-06	
	Testis	0.000123341	AP4E1
	Testis	6.22339E-10	USP8
	Testis	2.73893E-05	USP50
	Testis	0.000160282	RP11-562A8.1
	Whole_Blood	1.14064E-06	
	rs3098167	Artery_Tibial	8.23965E-05
Brain_Cerebellar_Hemisphere		4.23929E-05	
Brain_Nucleus_accumbens_basal_ganglia		7.82342E-07	
Brain_Putamen_basal_ganglia		9.89281E-05	
Cells_Cultured_fibroblasts		4.54732E-06	
Esophagus_Mucosa		0.000129549	
Esophagus_Muscularis		8.76778E-05	
Muscle_Skeletal		7.01891E-05	
Muscle_Skeletal		1.30798E-06	
Muscle_Skeletal		0.000318032	
Skin_Not_Sun_Exposed_Suprapubic		0.000115011	
Testis		2.78853E-05	AP4E1

	Testis	1.17625E-07	USP8
	Testis	9.95515E-06	USP50
	Whole_Blood	4.01132E-05	
rs3098169	Artery_Tibial	9.47298E-05	
	Brain_Cerebellar_Hemisphere	0.000045581	
	Brain_Nucleus_accumbens_basal_ganglia	9.67631E-07	
	Brain_Putamen_basal_ganglia	8.88044E-05	
	Cells_Cultured_fibroblasts	0.00000392	
	Esophagus_Mucosa	0.000106488	
	Esophagus_Muscularis	0.000084905	
	Muscle_Skeletal	5.31118E-05	
	Muscle_Skeletal	2.1368E-06	
	Skin_Not_Sun_Exposed_Suprapubic	0.00019545	
	Testis	2.78853E-05	AP4E1
	Testis	1.17625E-07	USP8
	Testis	9.95515E-06	USP50
	Whole_Blood	8.50657E-05	
rs3098171	Artery_Tibial	0.00018862	
	Brain_Cerebellar_Hemisphere	5.29961E-06	
	Brain_Cerebellum	0.000144612	
	Brain_Nucleus_accumbens_basal_ganglia	9.31896E-06	
	Cells_Cultured_fibroblasts	9.92845E-06	
	Esophagus_Mucosa	0.000181072	
	Esophagus_Muscularis	0.000104604	
	Muscle_Skeletal	9.27758E-05	
	Muscle_Skeletal	5.49018E-05	
	Prostate	1.28636E-05	
	Skin_Not_Sun_Exposed_Suprapubic	1.12827E-05	
	Testis	9.95543E-05	AP4E1
	Testis	1.17835E-09	USP8
	Testis	1.42732E-05	USP50
Testis	0.00018444	RP11-562A8.1	
Whole_Blood	9.29553E-07		
rs3098174	Adipose_Subcutaneous	0.00001048	
	Adipose_Visceral_Omentum	0.00003203	
	Artery_Tibial	0.000342721	
	Brain_Cerebellar_Hemisphere	5.1426E-06	
	Brain_Cerebellum	0.000126959	
	Brain_Nucleus_accumbens_basal_ganglia	2.69514E-05	
	Cells_Cultured_fibroblasts	2.46125E-08	
	Cells_Cultured_fibroblasts	5.93132E-09	
	Esophagus_Muscularis	9.16301E-05	
	Esophagus_Muscularis	0.000304306	
	Lung	3.36318E-05	
	Muscle_Skeletal	2.30259E-07	
	Skin_Not_Sun_Exposed_Suprapubic	6.4791E-06	
	Skin_Not_Sun_Exposed_Suprapubic	0.00029693	
	Testis	0.000231568	AP4E1
	Testis	2.44877E-10	USP8
	Testis	5.34409E-09	USP50
	Testis	6.04553E-06	RP11-562A8.1
Whole_Blood	2.21805E-06		
rs3098177	Artery_Tibial	0.000266476	
	Brain_Cerebellar_Hemisphere	5.29961E-06	
	Brain_Cerebellum	0.000144612	
	Brain_Nucleus_accumbens_basal_ganglia	9.31896E-06	
	Cells_Cultured_fibroblasts	7.3169E-06	
	Esophagus_Mucosa	0.000181072	
	Esophagus_Muscularis	0.000104604	
	Muscle_Skeletal	9.33158E-05	
	Muscle_Skeletal	6.58318E-05	
	Prostate	1.28636E-05	
	Skin_Not_Sun_Exposed_Suprapubic	1.12827E-05	
	Testis	9.95543E-05	AP4E1
	Testis	1.17835E-09	USP8
	Testis	1.42732E-05	USP50
Testis	0.00018444	RP11-562A8.1	
Whole_Blood	1.31413E-06		
rs3098205	Artery_Tibial	0.000244245	
	Brain_Cerebellar_Hemisphere	6.23203E-05	
	Brain_Nucleus_accumbens_basal_ganglia	1.38703E-06	
	Cells_Cultured_fibroblasts	4.31876E-07	

	Cells_Cultured_fibroblasts	3.80683E-05	
	Esophagus_Mucosa	0.000014412	
	Esophagus_Muscularis	0.000118364	
	Muscle_Skeletal	2.40601E-05	
	Muscle_Skeletal	1.94459E-07	
	Skin_Not_Sun_Exposed_Suprapubic	0.000216196	
	Testis	0.000025545	AP4E1
	Testis	3.95772E-07	USP8
	Testis	2.22885E-05	USP50
Whole_Blood	4.39813E-05		
rs3131559	Adipose_Subcutaneous	0.00001048	
	Adipose_Visceral_Omentum	0.00003203	
	Artery_Tibial	0.000342721	
	Brain_Cerebellar_Hemisphere	5.1426E-06	
	Brain_Cerebellum	0.000126959	
	Brain_Nucleus_accumbens_basal_ganglia	2.69514E-05	
	Cells_Cultured_fibroblasts	2.46125E-08	
	Cells_Cultured_fibroblasts	5.93132E-09	
	Esophagus_Muscularis	9.16301E-05	
	Esophagus_Muscularis	0.000304306	
	Lung	3.36318E-05	
	Muscle_Skeletal	2.30259E-07	
	Skin_Not_Sun_Exposed_Suprapubic	6.4791E-06	
	Skin_Not_Sun_Exposed_Suprapubic	0.00029693	
	Testis	0.000231568	AP4E1
	Testis	2.44877E-10	USP8
	Testis	5.34409E-09	USP50
Testis	6.04553E-06	RP11-562A8.1	
Whole_Blood	2.21805E-06		
rs3131560	Artery_Tibial	8.23965E-05	
	Brain_Cerebellar_Hemisphere	4.23929E-05	
	Brain_Nucleus_accumbens_basal_ganglia	7.82342E-07	
	Brain_Putamen_basal_ganglia	9.89281E-05	
	Cells_Cultured_fibroblasts	4.54732E-06	
	Esophagus_Mucosa	0.000129549	
	Esophagus_Muscularis	8.76778E-05	
	Muscle_Skeletal	7.01891E-05	
	Muscle_Skeletal	1.30798E-06	
	Muscle_Skeletal	0.000318032	
	Skin_Not_Sun_Exposed_Suprapubic	0.000115011	
	Testis	2.78853E-05	AP4E1
	Testis	1.17625E-07	USP8
	Testis	9.95515E-06	USP50
Whole_Blood	4.01132E-05		
rs3131562	Adipose_Subcutaneous	0.00001048	
	Adipose_Visceral_Omentum	0.00003203	
	Artery_Tibial	0.000325146	
	Brain_Cerebellar_Hemisphere	5.1426E-06	
	Brain_Cerebellum	0.000126959	
	Brain_Nucleus_accumbens_basal_ganglia	2.69514E-05	
	Cells_Cultured_fibroblasts	4.50379E-08	
	Cells_Cultured_fibroblasts	5.06119E-09	
	Esophagus_Muscularis	9.16301E-05	
	Esophagus_Muscularis	0.000304306	
	Lung	3.36318E-05	
	Muscle_Skeletal	2.30259E-07	
	Skin_Not_Sun_Exposed_Suprapubic	6.4791E-06	
	Skin_Not_Sun_Exposed_Suprapubic	0.00029693	
	Testis	0.000231568	AP4E1
	Testis	2.44877E-10	USP8
	Testis	5.34409E-09	USP50
Testis	6.04553E-06	RP11-562A8.1	
Whole_Blood	2.03027E-06		
rs3131566	Artery_Tibial	8.23965E-05	
	Brain_Cerebellar_Hemisphere	4.23929E-05	
	Brain_Nucleus_accumbens_basal_ganglia	7.82342E-07	
	Brain_Putamen_basal_ganglia	9.89281E-05	
	Cells_Cultured_fibroblasts	4.54732E-06	
	Esophagus_Mucosa	0.000129549	
	Esophagus_Muscularis	8.76778E-05	
	Muscle_Skeletal	7.01891E-05	
	Muscle_Skeletal	1.30798E-06	

	Muscle_Skeletal	0.000318032	
	Skin_Not_Sun_Exposed_Suprapubic	0.000115011	
	Testis	2.78853E-05	AP4E1
	Testis	1.17625E-07	USP8
	Testis	9.95515E-06	USP50
rs3131568	Whole_Blood	4.01132E-05	
	Artery_Tibial	9.47298E-05	
	Brain_Cerebellar_Hemisphere	5.91472E-05	
	Brain_Nucleus_accumbens_basal_ganglia	4.62788E-07	
	Cells_Cultured_fibroblasts	3.3196E-06	
	Esophagus_Mucosa	0.000182714	
	Esophagus_Muscularis	9.91767E-05	
	Muscle_Skeletal	8.92265E-05	
	Muscle_Skeletal	1.27124E-06	
	Skin_Not_Sun_Exposed_Suprapubic	0.000187054	
	Testis	2.78853E-05	AP4E1
	Testis	1.17625E-07	USP8
	Testis	9.95515E-06	USP50
	Whole_Blood	5.34864E-05	
rs3131574	Brain_Cerebellar_Hemisphere	6.23203E-05	
	Brain_Nucleus_accumbens_basal_ganglia	9.90315E-07	
	Brain_Putamen_basal_ganglia	8.88044E-05	
	Cells_Cultured_fibroblasts	8.17233E-07	
	Cells_Cultured_fibroblasts	8.53139E-05	
	Esophagus_Mucosa	0.000013624	
	Esophagus_Muscularis	0.000118364	
	Muscle_Skeletal	2.73548E-05	
	Muscle_Skeletal	2.20258E-07	
	Skin_Not_Sun_Exposed_Suprapubic	0.000276033	
	Testis	0.000025545	AP4E1
	Testis	3.95772E-07	USP8
	Testis	2.22885E-05	USP50
	Whole_Blood	6.68245E-05	
rs4318151	Adipose_Subcutaneous	5.19873E-06	
	Adipose_Visceral_Omentum	8.20931E-06	
	Artery_Tibial	0.000274011	
	Brain_Cerebellar_Hemisphere	3.71672E-06	
	Brain_Nucleus_accumbens_basal_ganglia	4.66297E-05	
	Cells_Cultured_fibroblasts	5.29182E-08	
	Cells_Cultured_fibroblasts	1.17104E-08	
	Esophagus_Muscularis	6.93392E-05	
	Lung	3.37049E-05	
	Muscle_Skeletal	1.06858E-07	
	Nerve_Tibial	0.000314605	
	Skin_Not_Sun_Exposed_Suprapubic	3.56149E-06	
	Testis	0.000239875	AP4E1
	Testis	1.80148E-10	USP8
	Testis	3.32586E-09	USP50
	Testis	5.03769E-06	RP11-562A8.1
	Whole_Blood	5.9828E-06	
	rs56398519	Artery_Tibial	0.000050816
Brain_Cerebellar_Hemisphere		5.83071E-06	
Brain_Cerebellum		0.000127401	
Brain_Nucleus_accumbens_basal_ganglia		8.21114E-06	
Brain_Putamen_basal_ganglia		7.98972E-05	
Cells_Cultured_fibroblasts		8.52695E-07	
Cells_Cultured_fibroblasts		0.000302409	
Esophagus_Mucosa		3.79734E-05	
Muscle_Skeletal		0.000064559	
Muscle_Skeletal		2.10788E-06	
Skin_Not_Sun_Exposed_Suprapubic		1.08385E-05	
Skin_Sun_Exposed_Lower_leg		0.000143232	
Testis		5.60779E-05	AP4E1
Testis		2.56803E-09	USP8
Testis		1.04344E-05	USP50
Thyroid		0.000322179	
Thyroid		0.000367115	
Whole_Blood		2.53365E-07	
rs7174015	Artery_Tibial	0.000216121	
	Brain_Cerebellar_Hemisphere	5.29961E-06	
	Brain_Cerebellum	0.000138165	
	Brain_Nucleus_accumbens_basal_ganglia	9.31896E-06	



	Cells_Cultured_fibroblasts	7.13283E-06	
	Esophagus_Mucosa	0.000283085	
	Esophagus_Muscularis	7.16878E-05	
	Muscle_Skeletal	5.99913E-05	
	Muscle_Skeletal	0.00016635	
	Prostate	2.29885E-05	
	Skin_Not_Sun_Exposed_Suprapubic	6.95486E-06	
	Testis	5.05685E-05	AP4E1
	Testis	1.14029E-09	USP8
	Testis	3.21206E-05	USP50
	Testis	0.0002328	RP11-562A8.1
	Whole_Blood	3.46872E-06	

**Table S12 - Scores obtained for the functional predictive tests for USP8-rs7174015 and its proxies according to the SNPnexus database.**

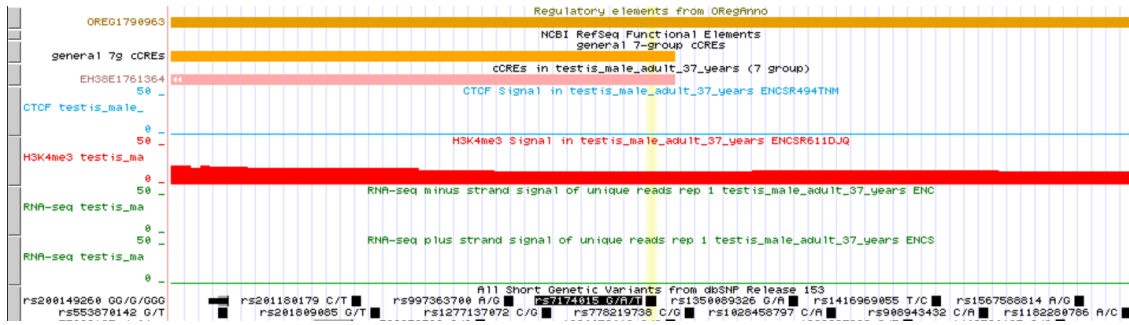
SNP	Function predictive effect							
	CADD	fitCons	EIGEN	FATHMM	GWAVA	DeepSea	FunSeq2	ReMM
rs7174015	7,888; 7,305	0,085055	1,263628	0,24567; 0,38127	0,66	0,015285; 0,023881	1,378127	0,174
rs3098177	1,618; 1,107	0,106106	-0,22873	0,05900; 0,15004	0,19	0,232330; 0,205510	0,580878	0,011
rs2289108	2,538	0,088506	0,058305	0,12883	0,13	0,083479	0,580878	0,011
rs56398519	0,549	0,099367	-0,240353	0,0742	0,12	0,12811	0,580878	0,482
rs3098171	1,279	0,099367	-0,259828	0,05063	0,03	0,27279	0,580878	0,017
rs11070776	0,737; 0,504; 0,535	0,099367	-0,726639	0,04123; 0,04853; 0,04724	0,01	0,345190; 0,42250; 0,442880	0,580878	0,01
rs28582911	1,849	0,106106	-0,360355	0,06679	0,09	0,10657	0,580878	0,012
rs4318151	0,081	0,088506	-0,733786	0,04902	0,1	0,48584	0,580878	0,007
rs3131559	0,285	0,106106	-0,726949	0,02239	0,15	0,31491	0,580878	0,008
rs3098174	0,629	0,099367	-0,644072	0,06085	0,12	0,138300; 0,200540; 0,193260	0,580878	0,009
rs3131562	6,805	0,088506	0,022646	0,04833	0,09	0,053188	1,203713	0,029
rs3098169	0,091; 0,064	0,088506	-0,454919	0,05321; 0,09565	0,21	0,106380; 0,136360	0,580878	0,007
rs3098205	0,833	0,048788	-0,474181	0,0962	0,18	0,1105	0,580878	0,029
rs3131574	1,788; 1,684; 2,387	0,088506	-0,505159	0,03510; 0,06122; 0,06013	0,12	0,20445	0,736264	0,022
rs3131568	0	0,099367	-0,415523	0,09994	0,18	0,20601	0,736264	0,01
rs10152326	2,307; 2,090	0,062308	-0,81643	0,02424; 0,03819	0,04	0,399340; 0,378010	0,580878	0,018
rs12593481	4,956	0,093295	-0,406658	0,03547	0,67	0,040078	0,736264	0,071
rs3131560	5,769	0,088506	-0,416689	0,06598	0,11	0,16677	0,580878	0,027
rs3131566	0,055	0,062308	-1,260296	0,01822	0,37	0,3493	0,736264	0,011
rs3098167	5,003	0,099367	-0,37746	0,03706	0,17	0,079779	0,580878	0,032

**Table S13 - Scores obtained for the functional predictive tests for rs12870438 and its proxies, according to SNPnexus database.**

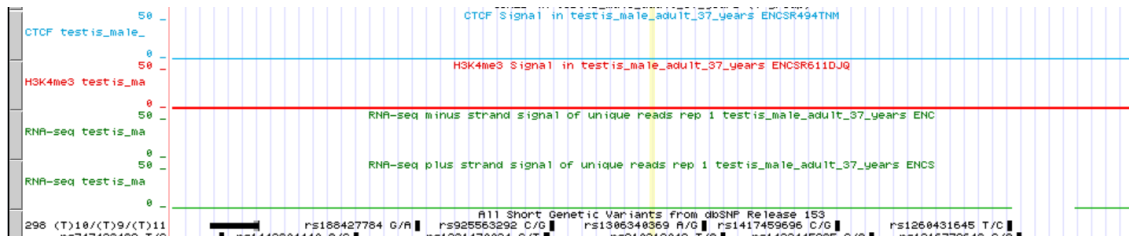
SNP	Functional predictive effect							
	CADD	fitCons	EIGEN	FATHMM	GWAVA	DeepSea	FunSeq2	ReMM
rs12870438	2,423	0,062308	0,056621	0,15532	0,13	0,035054	0,011479	0,078
rs58357177	0	0,164513	0	0	0	0,034408	0	0,194
rs1535898	14,72	0,062308	0,982871	0,67038	0,18	0,027322	0,634313	0,941
rs1535897	2,322; 3,183	0,106106	0,127086	0,17854; 0,15422	0,05	0,049245; 0,032204	0,166865	0,518
rs1830780	0,054	0,099367	-0,777478	0,0378	0,01	0,3964	0,011479	0,015
rs63351261	0	0,088506	0	0	0	0,0984	0	0,418
rs9594829	10,78	0,106106	0,450494	0,22564	0,1	0,021424	0,011479	0,634
rs11431398	0	0,106106	0	0	0	0	0	0,52
rs1535900	3,575	0,164513	0,035426	0,05512	0,04	0,13695	0,011479	0,052
rs34525682	1,442	0,106106	-0,477165	0,0549	0	0,24516	0,011479	0,019
rs10161854	0,139	0,053691	-0,676741	0,0428	0,06	0,20912	0	0,024
rs9590722	4,157	0,119485	0,127844	0,11759	0,16	0,10992	0,155386	0,718
rs12870885	2,004	0,074388	-0,18402	0,02896	0,1	0,20756	0,155386	0,065
rs9594827	7,045	0,053691	0,047557	0,14829	0,06	0,092822	0	0,736
rs9594826	3,007; 3,284; 2,896	0,053691	-0,075663	0,13183; 0,13640; 0,12420	0,05	0,17377; 0,090951; 0,122700	0	0,048
rs71099806	0	0,106106	0	0	0	0,29545	0	0,019

**Table S14 - Scores obtained for the functional predictive tests for rs12870348 and its proxy, according to SNPnexus database.**

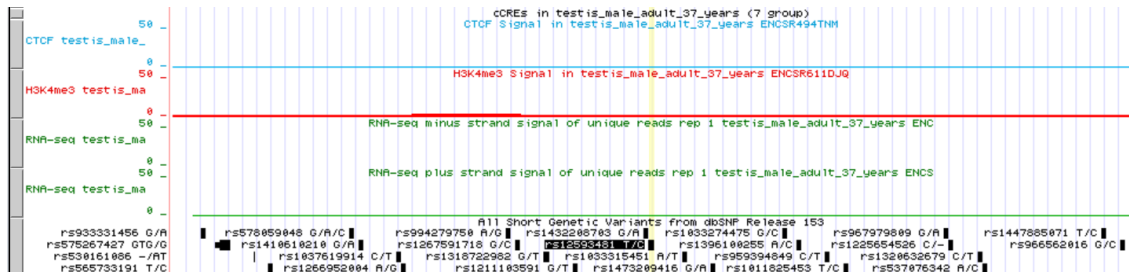
SNP	Functional predictive effect							
	CADD	fitCons	EIGEN	FATHMM	GWAVA	DeepSea	FunSeq2	ReMM
rs7867029	2,035	0,074388	-0,190303	0,05038	0,11	0,061662	0,002295	0,013
rs10867194	2,071	0,078448	-0,156187	0,06271	0,11	0,15092	0,155386	0,02



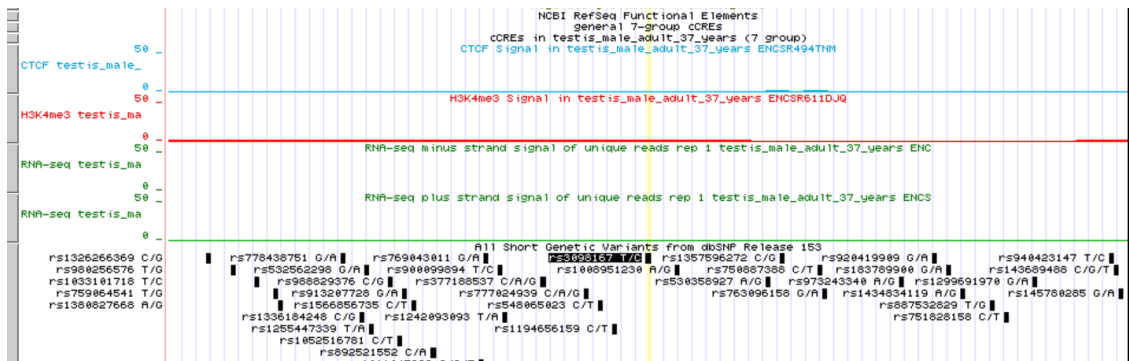
**Figure S6 – USP8-rs7174015 overlapping a binding site of H3K4me3 in adult testis** (output from UCSC Genome Browser based on the information of ENCODE database).



**Figure S7 - rs4318151 overlapping a binding site of H3K4me3 in adult testis** (output from UCSC Genome Browser based on the information of ENCODE database).



**Figure S8 - rs12593481 overlapping a binding site of H3K4me3 in adult testis** (output from UCSC Genome Browser based on the information of ENCODE database).



**Figure S9 - rs3098167 overlapping a binding site of H3K4me3 in adult testis** (output from UCSC Genome Browser based on the information of ENCODE database).

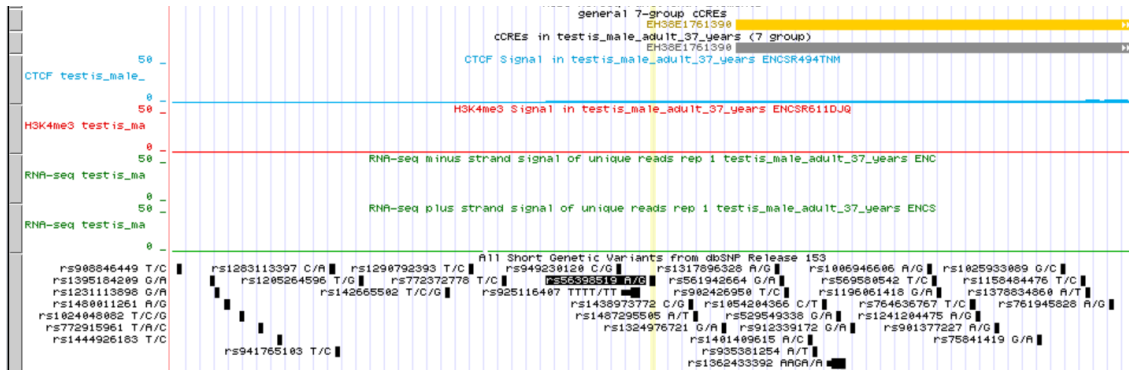


Figure S10 - rs56398519 overlapping a binding site of CTFP in adult testis (output from UCSC Genome Browser based on the information of ENCODE database).

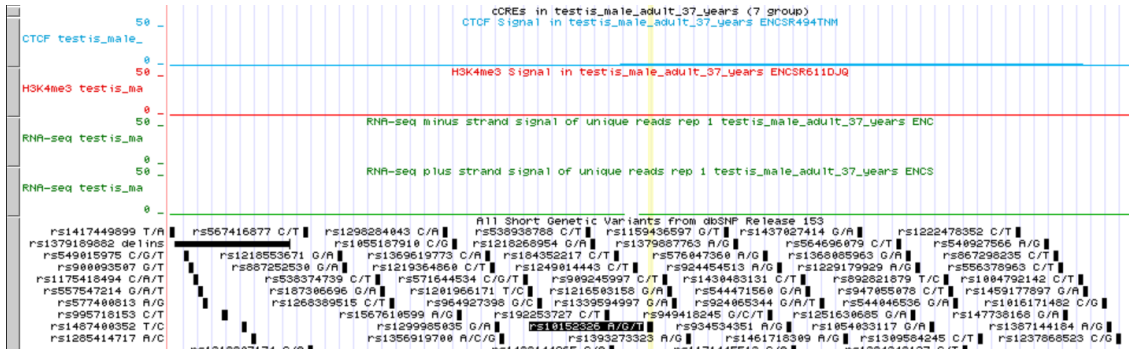


Figure S11 - rs10152326 overlapping a binding site of CTFP in adult testis (output from UCSC Genome Browser based on the information of ENCODE database).

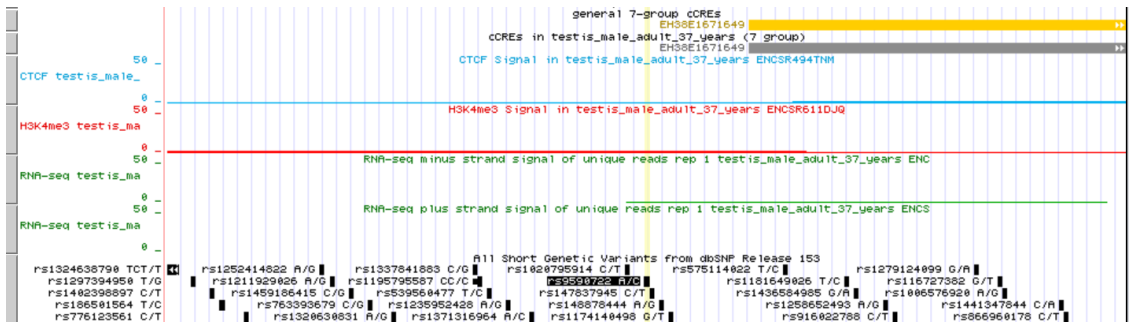


Figure S12 - rs9590722 overlapping a binding site of H3K4me3 in adult testis (output from UCSC Genome Browser based on the information of ENCODE database).

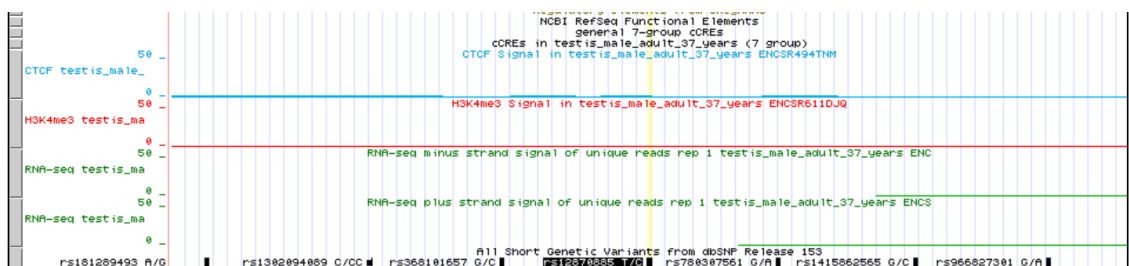
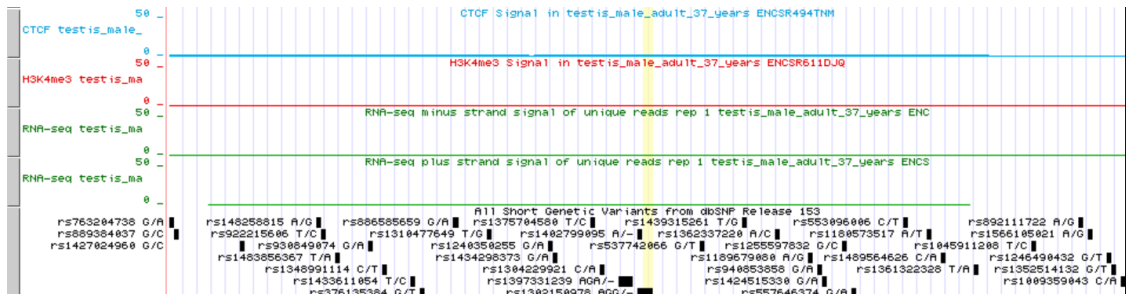
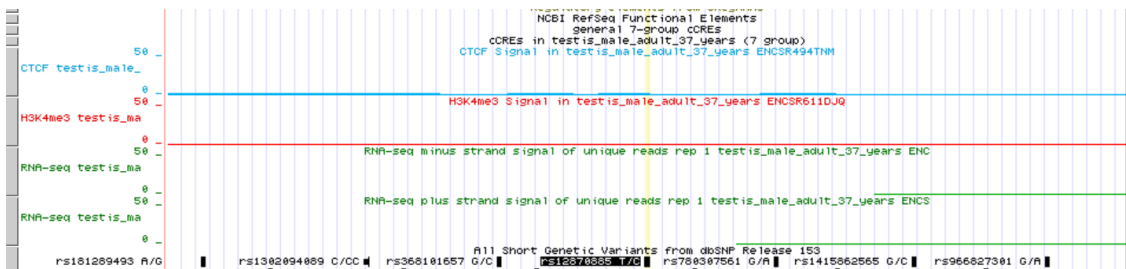


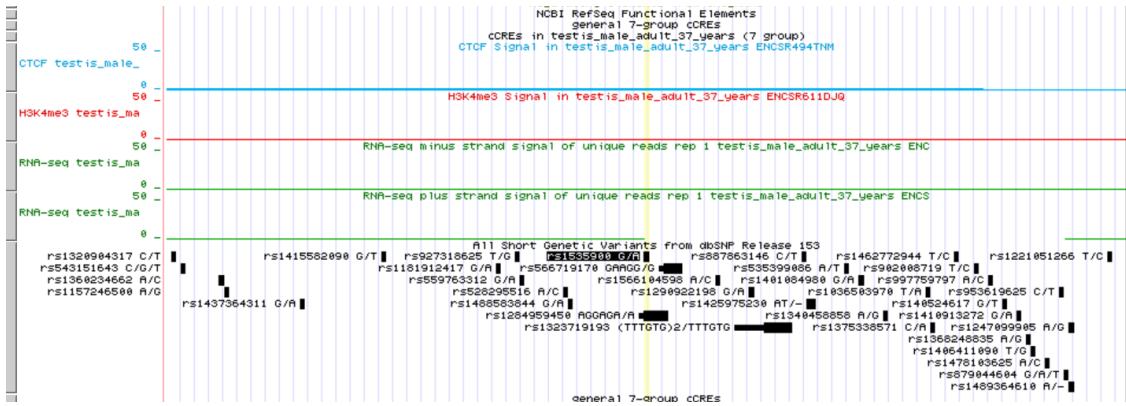
Figure S13 - rs12870885 overlapping a binding site of CTFP in adult testis (output from UCSC Genome Browser based on the information of ENCODE database).



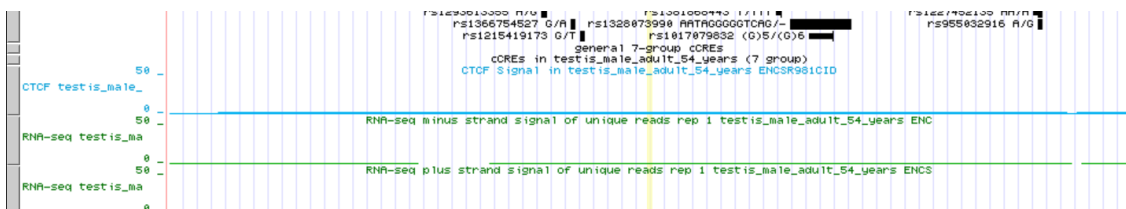
**Figure S14 - rs58357177 overlapping a binding site of CTCF in adult testis** (output from UCSC Genome Browser based on the information of ENCODE database).



**Figure S15 - rs12870885 overlapping a binding site of CTCF in adult testis** (output from UCSC Genome Browser based on the information of ENCODE database).



**Figure S16 - rs1535900 overlapping a binding site of CTCF in adult testis** (output from UCSC Genome Browser based on the information of ENCODE database).



**Figure S17 - rs34525682 overlapping a binding site of CTCF in adult testis** (output from UCSC Genome Browser based on the information of ENCODE database).

# Appendix

## **Poster 1**

Assessment of fertility associated variants in a Portuguese cohort of Azoospermia and Severe Oligospermia

# Assessment of fertility associated variants in a Portuguese cohort of Azoospermia and Severe Oligozoospermia

Cláudia Costa<sup>1,2,3</sup>, Miriam Cerván-Martín<sup>4,5</sup>, Chiranan Khantham<sup>6</sup>, Inés Linares-Burguet<sup>4</sup>, Patrícia I. Marques<sup>1,2,7</sup>, Filipa Carvalho<sup>1,8</sup>, Alberto Barros<sup>8</sup>, Susana Seixas<sup>1,2,8</sup>, Iris Caetano<sup>9</sup>, João Gonçalves<sup>9,10</sup>, Rogelio J. Palomino-Morales<sup>11,5</sup>, F. David Carmona<sup>4,5</sup>, Alexandra M. Lopes<sup>1,2</sup>

<sup>1</sup>Instituto de Investigação e Inovação em Saúde, Universidade do Porto (I3S), Porto, Portugal. <sup>2</sup>Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP), Porto, Portugal. <sup>3</sup>Faculdade de Ciências da Universidade do Porto (FCUP). <sup>4</sup>Departamento de Genética e Instituto de Biotecnología, Universidad de Granada, Spain. <sup>5</sup>Instituto de Investigación Biosanitaria de Granada (ibs.GRANADA). <sup>6</sup>Department of Pharmaceutical Sciences, Faculty of Pharmacy, Chiang Mai University, Chiang Mai, Thailand. <sup>7</sup>Genetic Diversity Group, Instituto de Investigação e Inovação em Saúde (I3S), Universidade do Porto, Porto, Portugal. <sup>8</sup>Serviço de Genética, Departamento de Patologia, Faculdade de Medicina da Universidade do Porto, Porto, Portugal. <sup>9</sup>Departamento de Genética Humana, Inst. Nac. Saúde Dr Ricardo Jorge, Lisbon, Portugal; <sup>10</sup>ToxOmics - Centro de Toxicogenómica e Saúde Humana, Nova Medical School, Lisbon, Portugal. <sup>11</sup>Departamento de Bioquímica y Biología Molecular I, Universidad de Granada, Granada, Spain.

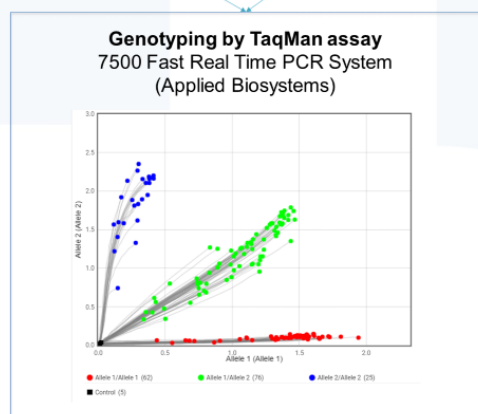
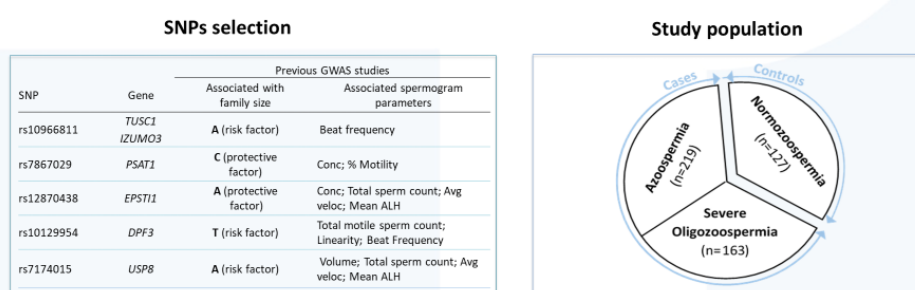


## Background

Male infertility affects about 7% of men worldwide. The complexity of this condition is reflected in a very heterogeneous phenotypic expression, from lowered sperm counts (oligozoospermia) with or without other alterations in its quality, to a complete absence of sperm in the ejaculate (azoospermia)<sup>2</sup>. Although in some patients with severe spermatogenic failure the underlying cause can be determined (i.e., azoospermia factor deletions – AZF), in most cases it remains unexplained – idiopathic infertility.

Spermatogenesis is a complex process that involves thousands of genes and carefully orchestrated interactions between somatic and germ cells. In this work we aimed to replicate the association of 5 non-coding SNPs with fertility traits reported in previous GWAS studies<sup>1,3,4</sup>, in our cohort of Portuguese men with severe spermatogenic failure.

## Patients and Methods



Statistical analysis performed using **PLINK (v1.9)** Software:

- Allele and genotype frequencies;
- Hardy-Weinberg test;
- Logistic regression assuming different models – additive, dominant, recessive – with the FDR (False Discovery Rate) Benjamini–Hochberg adjustment.

## Results & Discussion

### Genotyping & QC

The genotyping success rate was **over 97.5%**. The genotype frequencies of all SNPs show **no significant divergence from HWE** in controls ( $P > 0.05$ ; Table 1).

**Table 1.** Genotype and allele frequencies in the study population

Chr	SNP	Position	Closest Gene	Location	Distance (kb)	Allele	IBS	MAF <sup>a</sup>	CONTROLS			CASES				
									n	Genotypes <sup>c</sup>	AF <sup>d</sup>	n	Genotypes <sup>c</sup>	AF <sup>d</sup>	n	Genotypes <sup>c</sup>
9	rs10966811	25 233 486	TUSC1 IZUMO3	downstr. upstr.	442.903 687.837	A/G	0.41	127	14/68/45	0.38	217	28/97/92	0.35	160	27/71/62	0.39
9	rs7867029	78 405 502	PSAT1	downstr.	75.409	C/G	0.18	127	2/31/94	0.14	219	2/61/156	0.15	161	2/26/133	0.09
13	rs12870438	4 290 669	EPST11	intron		A/G	0.35	124	25/56/43	0.43	218	33/94/91	0.37	161	19/75/67	0.35
14	rs10129954	72 683 993	DPF3	intron		T/C	0.48	127	17/64/46	0.39	216	42/102/72	0.43	163	25/76/62	0.39
15	rs7174015	50 424 871	USP8	intron		A/G	0.47	125	24/63/38	0.44	214	66/92/56	0.52	159	34/87/38	0.49

Genes: TUSC1 Tumor Suppressor Candidate 1 gene; IZUMO3 IZUMO family member 3 gene; PSAT1 Phosphoserine Aminotransferase 1 gene; EPST11 Epithelial Stromal Interaction Protein 1 gene; DPF3 Double PHD Fingers 3 gene; USP8 Ubiquitin Specific Peptidase 8 gene.

<sup>a</sup>Allele: minor/major allele (as defined in controls)  
<sup>b</sup>Minor Allele Frequencies of Iberian populations in Spain in 1000 Genomes Project\_Phase3 (www.ensembl.org)  
<sup>c</sup>Genotypes: minor homozygote/heterozygote/major homozygote (as defined in controls)  
<sup>d</sup>AF: allele frequency of the minor allele (as defined in controls)

## Association tests (results in Table 2)

**Table 2.** The association derived from different comparative genetic models (additive, dominant and recessive)

SNP	Model	Azoospermia			Severe Oligozoospermia			Combined		
		OR (95% CI)	p value	FDR	OR (95% CI)	p value	FDR	OR (95% CI)	p value	FDR
rs10966811	Additive (risk allele, A)	0.89 (0.64-1.24)	0.495	0.619	1.06 (0.75-1.49)	0.754	0.942	0.96 (0.72-1.29)	0.792	0.792
	Dominant (AA+AG vs. GG)	0.75 (0.47-1.17)	0.204	0.588	0.87 (0.54-1.41)	0.564	0.705	0.80 (0.52-1.21)	0.281	0.468
	Recessive (AA vs. AG+GG)	1.20 (0.60-2.37)	0.608	0.608	1.64 (0.82-3.28)	0.162	0.406	1.38 (0.74-2.58)	0.314	0.392
rs7867029	Additive (risk allele, C)	1.10 (0.69-1.74)	0.691	0.691	0.64 (0.38-1.08)	0.097	0.244	0.89 (0.58-1.36)	0.590	0.737
	Dominant (CC+CG vs. GG)	1.15 (0.70-1.88)	0.578	0.587	0.60 (0.34-1.06)	<b>0.078</b>	0.389	0.90 (0.57-1.42)	0.644	0.805
	Recessive (CC vs. CG+GG)	0.58 (0.08-4.14)	0.584	0.608	0.79 (0.11-5.66)	0.811	0.811	0.66 (0.12-3.67)	0.640	0.640
rs12870438	Additive (risk allele, A)	0.79 (0.58-1.08)	0.134	0.335	0.73 (0.52-1.02)	0.068	0.244	0.76 (0.57-1.02)	0.065	0.212
	Dominant (AA+AG vs. GG)	0.74 (0.47-1.17)	0.199	0.509	0.74 (0.46-1.21)	0.234	0.389	0.74 (0.49-1.13)	0.167	0.468
	Recessive (AA vs. AG+GG)	0.71 (0.40-1.25)	0.235	0.392	0.53 (0.28-1.01)	<b>0.055</b>	0.276	0.63 (0.37-1.07)	0.086	0.225
rs10129954	Additive (risk allele, T)	1.20 (0.88-1.65)	0.252	0.420	1.00 (0.71-1.41)	0.987	0.987	1.11 (0.83-1.49)	0.471	0.737
	Dominant (TT+TC vs. CC)	1.14 (0.72-1.80)	0.587	0.587	0.93 (0.57-1.50)	0.751	0.751	1.04 (0.68-1.58)	0.860	0.860
	Recessive (TT vs. TC+CC)	1.56 (0.85-2.88)	0.153	0.383	1.17 (0.60-2.28)	0.640	0.811	1.39 (0.78-2.47)	0.262	0.392
rs7174015	Additive (risk allele, A)	1.34 (0.99-1.82)	0.057	0.283	1.21 (0.85-1.70)	0.289	0.481	1.29 (0.97-1.71)	0.085	0.212
	Dominant (AA+AG vs. GG)	1.23 (0.76-2.01)	0.401	0.587	1.39 (0.82-2.36)	0.220	0.389	1.30 (0.83-2.03)	0.255	0.468
	Recessive (AA vs. AG+GG)	<b>1.88 (1.10-3.19)</b>	<b>0.020</b>	<b>0.101</b>	1.15 (0.64-2.05)	0.651	0.811	1.54 (0.93-2.54)	0.090	0.225

Data are shown as odds ratio (OR), 95% confidence interval (CI), P value and FDR Benjamini–Hochberg adjustment. Boxes show trend toward association. Bold numbers indicate P value < 0.05

- rs10966811 (TUSC1/IZUMO3) - No association with male infertility was detected in our study however, the A allele had been described as a risk factor for reduced family size in a Hutterite population<sup>1</sup>. Even though the functional impact of this SNP is unknown, TUSC1 transcripts are mostly expressed in testis and IZUMO3 is an essential gene for gamete fusion<sup>4</sup>.
- rs7867029 (PSAT1) – The C allele shows a non-significant trend (OR=0.6; nominal p=0.078) towards a protection from severe oligozoospermia, assuming a dominant model. This allele has been associated with larger family size<sup>1</sup>, although there is still no functional evidence for a role of PSAT1 in spermatogenesis.
- rs12870438 (EPST11) - Assuming a recessive model, there is also a borderline trend for association of the A allele with severe oligozoospermia, as a protective factor (OR=0.5; nominal p=0.055). This gene seems to be related with immune response but its function in testis is still unknown in spite of its high level of expression in the male gonad<sup>4</sup>.
- rs10129954 (DPF3) - Although the T allele has been already associated with smaller family size and as a risk factor for lower total motile sperm count<sup>4</sup>, we did not detect association with either severe oligozoospermia or azoospermia in our cohort. It is known that DPF3 plays a key role in preparing the DNA in mature sperm for early embryogenesis<sup>3</sup>. This polymorphism may affect histone modification by altering the splicing of DPF3<sup>3</sup>.
- rs7174015 (USP8) - Assuming a recessive model, a nominally significant association with azoospermia was detected for the A allele (OR=1.88; P=0.020), even though it did not pass the threshold of significance after FDR correction. The A allele of this SNP was already associated with smaller family size<sup>4</sup>. USP8 encodes a crucial enzyme for protein deubiquitinating and sorting endosomal cargo in spermatogenic cells. This gene has also an important role in shaping the sperm head and in acrosome assembly in differentiating sperm cells<sup>4</sup>.

## Conclusion

Even though all of these SNPs had been already associated with fertility traits in other populations (Hutterites and Japanese), in our study we only observed a trend towards association that did not stand after FDR correction, for 3 of these SNPs (rs12870438, rs7867029, rs7174015). For these SNPs our results show a similar trend of that already described in the literature in populations of European ancestry. This work is part of a larger ongoing association study in a well-powered cohort of Portuguese and Spanish infertile men and controls. Thus, we expect that with a larger sample size we will better evaluate the impact of these SNPs in male infertility phenotypes in Europeans.

## Acknowledgments

Portuguese funds supported this work through FCT-Fundação para a Ciência e a Tecnologia, Ministério da Ciência, Tecnologia e Inovação in the framework of the project "Institute for Research and Innovation in Health Sciences" (POCI-01-0145-FEDER-007274). AML and PIM are funded by FCT: IF/01262/2014 and SFRH/BPD/120777/2016, respectively.

This work was also supported by the Spanish Ministry of Economy and Competitiveness through the Spanish National Plan for Scientific and Technical Research and Innovation (ref. SAF2016-78722-R) and the "Ramón y Cajal" program to D. Carmona ref. RYC-2014-16458), which include FEDER funds.

## References

- Kosova, G., Scott, N. M., Niederberger, C., Prins, G. S., & Ober, C. (2012). Genome-wide association study identifies candidate genes for male fertility traits in humans. *American Journal of Human Genetics*, 90(6), 950–961. <https://doi.org/10.1016/j.ajhg.2012.04.016>
- Krausz, C., & Riera-Escamilla, A. (2018). Genetics of male infertility. *Nature Reviews Urology*, 15(6), 369–384. <https://doi.org/10.1038/s41585-018-0003-3>
- Liu, S. Y., Zhang, C. J., Peng, H. Y., Sun, H., Lin, K. Q., Huang, X. Q., ... Yang, Z. Q. (2016). Strong association of SLC1A1 and DPF3 gene variants with idiopathic male infertility in Han Chinese. *Asian Journal of Andrology*, 18(October 2015), 486–492. <https://doi.org/10.4103/1008-682X.178850>
- Sato, Y., Hasegawa, C., Tajima, A., Nozawa, S., Yoshiike, M., Koh, E., ... Iwamoto, T. (2018). Association of TUSC1 and DPF3 gene polymorphisms with male infertility. *Journal of Assisted Reproduction and Genetics*, 35(2), 257–263. <https://doi.org/10.1007/s10815-017-1052-x>

**INSTITUTO DE INVESTIGAÇÃO E INOVAÇÃO EM SAÚDE**  
UNIVERSIDADE DO PORTO

Rua Alfredo Allen, 208  
4200-135 Porto  
Portugal  
+351 220 408 800

[www.i3s.up.pt](http://www.i3s.up.pt)

## **Article 1**

Evaluation of male fertility-associated loci in a European population of patients with severe spermatogenic impairment



**Running title:** Genetic determinants of severe spermatogenic impairment

**Title:** Evaluation of male fertility-associated *loci* in a European population of patients with severe spermatogenic impairment

Miriam Cerván-Martín<sup>1,2,\*</sup>, Bossini-Castillo L<sup>1,2,\*</sup>, Rocío Rivera-Egea<sup>3,4</sup>, Nicolás Garrido<sup>4,5</sup>, Saturnino Lujan<sup>5</sup>, Gema Romeu<sup>5</sup>, Samuel Santos-Ribeiro<sup>6,7</sup>, *IVIRMA Group*<sup>8</sup>, *Lisbon Clinical Group*<sup>8</sup>, José A. Castilla<sup>2,9,10</sup>, M. Carmen Gonzalvo<sup>2,9</sup>, F. Javier Vicente<sup>2,11</sup>, **Cláudia Costa**<sup>12,13</sup>, Inés Llinares-Burguet<sup>1</sup>, Chiranan Khantham<sup>14</sup>, Miguel Burgos<sup>1</sup>, Francisco J. Barrionuevo<sup>1</sup>, Rafael Jiménez<sup>1</sup>, Josvany Sánchez-Curbelo<sup>15</sup>, Olga López-Rodrigo<sup>15</sup>, M. Fernanda Peraza<sup>15</sup>, Iris Pereira-Caetano<sup>16</sup>, Patricia I. Marques<sup>12,13</sup>, Filipa Carvalho<sup>17</sup>, Alberto Barros<sup>17</sup>, Lluís Bassas<sup>15</sup>, Susana Seixas<sup>12,13</sup>, João Gonçalves<sup>16,18</sup>, Sara Larriba<sup>19</sup>, Alexandra M. Lopes<sup>12,13</sup>, Rogelio J. Palomino-Morales<sup>20,2,¶</sup>, F. David Carmona<sup>1,2,¶</sup>.

Affiliations:

<sup>1</sup>Departamento de Genética e Instituto de Biotecnología, Universidad de Granada, Spain. <sup>2</sup>Instituto de Investigación Biosanitaria ibs.GRANADA, Granada, Spain. <sup>3</sup>Andrology Laboratory and Sperm Bank, IVIRMA Valencia, Valencia, Spain. <sup>4</sup>IVI Foundation, Health Research Institute La Fe, Valencia, Spain. <sup>5</sup>Servicio de Urología. Hospital Universitari i Politecnic La Fe e Instituto de Investigación Sanitaria La Fe (IIS La Fe), Valencia, Spain. <sup>6</sup>IVI-RMA Lisbon, Lisbon, Portugal. <sup>7</sup>Department of Obstetrics and Gynecology, Faculty of Medicine, University of Lisbon, Lisbon, Portugal. <sup>8</sup>See supplemental note. <sup>9</sup>Unidad de Reproducción, UGC Obstetricia y Ginecología, HU Virgen de las Nieves, Granada, Spain. <sup>10</sup>CEIFER Biobanco - NextClinics, Granada, Spain. <sup>11</sup>UGC de Urología, HU Virgen de las Nieves, Granada, Spain. <sup>12</sup>Instituto de Investigação e Inovação em Saúde, Universidade do Porto (I3S), Porto, Portugal. <sup>13</sup>Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP), Porto, Portugal. <sup>14</sup>Department of Pharmaceutical Sciences, Faculty of Pharmacy, Chiang Mai University, Chiang Mai, Thailand. <sup>15</sup>Laboratory of Seminology and Embryology, -Andrology Service-Fundació Puigvert, Barcelona, Spain. <sup>16</sup>Departamento de Genética Humana, Instituto Nacional de Saúde Dr. Ricardo Jorge, Lisbon, Portugal. <sup>17</sup>Serviço de Genética, Departamento de Patologia, Faculdade de Medicina da Universidade do Porto, Porto, Portugal. <sup>18</sup>ToxOmics - Centro de Toxicogenómica e Saúde Humana, Nova Medical School, Lisbon, Portugal. <sup>19</sup>Human Molecular Genetics Group, Bellvitge Biomedical Research Institute (IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain. <sup>20</sup>Departamento de Bioquímica y Biología

Molecular I, Universidad de Granada, Granada, Spain. \*Contributed equally. <sup>†</sup>Share senior authorship.

*Correspondence to:*

Rogelio J. Palomino-Morales, PhD. Centro de Investigación Biomédica (CIBM), Universidad de Granada. Parque Tecnológico Ciencias de la Salud. Avda. del Conocimiento S/N. 18016-Armilla (Granada), Spain.

E-mail: [rpm@ugr.es](mailto:rpm@ugr.es) / Tel: (+34) 958 243088

**CAPSULE:** This study provides insight into the genetic component of severe spermatogenic impairment of idiopathic origin, and the putative pathogenic mechanisms leading to this condition.

## ABSTRACT

**Objective:** To evaluate whether 5 single-nucleotide polymorphisms (SNP) previously associated with reduced family size in Hutterites are also involved in the genetic risk to severe spermatogenic failure (SpF).

**Design:** Genetic association study.

**Setting:** Fertility clinics, hospitals, research Institutes and University laboratories.

**Patient(s):** Seven hundred and twenty five Iberian SpF patients, including 495 men diagnosed with non-obstructive azoospermia (NOA) and 230 with severe oligospermia (SO), and 1,058 fertile male controls.

**Intervention(s):** None.

**Main Outcome Measure(s):** All subjects were genotyped for *USP8*-rs7174015, *DPF3*-rs10129954, *EPSTI1*-rs12870438, *PSAT1*-rs7867029, and *TUSC1*-rs10966811 using the TaqMan technology. Case-control association analyses by logistic regression on the genotypes assuming different models and in silico functional characterization of risk variants were conducted.

**Result(s):** A significant difference in the allele frequencies of *USP8*-rs7174015 was observed between the NOA group and both the control group ( $P_{\text{ADD}}=0.0402$ ,  $\text{OR}=1.18$ ) and the SO group ( $P_{\text{ADD}}=0.0323$ ,  $\text{OR}=0.77$ ). Differences in the genotype distributions were also suggested in those comparisons ( $P_{\text{GENO}}=0.0709$  and  $P_{\text{GENO}}=0.0178$ , respectively). The association signal was stronger under a recessive model ( $P_{\text{REC}}=0.0226$ ,  $\text{OR}=1.33$ , and  $P_{\text{REC}}=0.0048$ ,  $\text{OR}=0.56$ , respectively, being the latter significant after controlling for multiple testing:  $P_{\text{BONF}}=0.0242$ ). Other genetic associations for *EPSTI1*-rs12870438 and *PSAT1*-rs7867029 with SO and between *TUSC1*-rs10966811 and TESE success in the context of NOA were suggestive. The analysis of functional annotations evidenced cis-eQTL effects of such SNPs likely due to modification of binding motif sites for relevant transcription factors.

**Conclusion(s):** We identified putative markers of NOA, SO, and TESE success in a population of European descent. Genetic susceptibility is likely conferred by deregulation of gene expression during spermatogenesis.

**Key Words:** polymorphisms, spermatogenesis, non-obstructive azoospermia, oligospermia, infertility.

## 1. INTRODUCTION

Male infertility is considered one of the major health concerns in developed societies, affecting 10-15% of couples in childbearing age worldwide. The clinical manifestations of this condition are highly heterogeneous due to the large variety of possible causes that can lead to it. These include, for example, physical, environmental, or genetic cues, being the latter responsible for a large proportion of male infertility cases (1). Indeed, it has been reported that the two most extreme phenotypes of male infertility, *i.e.* severe oligospermia (SO, very low concentration of spermatozoa in semen) and non-obstructive azoospermia (NOA, complete lack of sperm in the ejaculate due to non-obstructive causes), have an important genetic component (2). Known genetic risk factors for these two severe manifestations of male infertility include congenital genetic anomalies, such as point mutations on genes with key roles in the spermatogenic process, Y chromosome microdeletions, and karyotype abnormalities, as well as common variation of the human genome, mostly single-nucleotide polymorphisms (SNPs) and copy-number variants (CNVs) (3).

One of the most successful strategies to investigate the possible influence of common genetic variation in the development of complex traits is the genome-wide association study (GWAS) approach, in which millions of genetic polymorphisms are interrogated in a hypothesis-free fashion across the whole genome (4). In a previous study, Kosova and colleagues (5) performed a GWAS to determine the possible causes of reduced male fertility in a study cohort composed of Hutterite men with reported fatherhood. Hutterites are a North American ethno-religious population of European descent in which contraception is proscribed, resulting in large family sizes. The authors described different genes associated with family size and several semen parameters, including *TUSC1* (MIM\*610529; encoding the tumor suppressor candidate 1, which is downregulated in non-small-cell lung cancer and small-cell lung cancer cell lines), *PSAT1* (MIM\*610936; encoding a phosphoserine aminotransferase expressed in the testis), *EPSTI1* (MIM\*607441; encoding the epithelial stromal interaction protein 1 highly expressed in the testis), *USP8* (MIM\*603158; encoding a ubiquitin specific protein that regulates endosome morphology and it is also highly expressed in the testis), and *DPF3* (MIM\*601672; encoding a transcription regulator involved in chromatin remodeling) (5).

Taking all the above into consideration, we decided to analyze for the first time whether the genetic markers of male fertility identified in the Hutterite population conferred risk also to severe spermatogenic failure (SpF), in a large cohort of Iberian men diagnosed with SO and NOA. Specific clinical entities of NOA, as well as probability

of success in sperm retrieval with testicular sperm extraction (TESE) techniques, were also tested for association.

## **2. MATERIAL AND METHODS**

### **Study design and study population**

An Iberian population of 725 infertile men due to SpF (comprising 495 NOA patients and 230 SO patients) and 1,058 unaffected male controls of European descent were enrolled in this study. SpF cases were recruited in different clinics of the 'IVI-RMA Global' group as well as public centers and Hospitals from Spain and Portugal. The control population included 700 population-representative healthy subjects with self-reported fatherhood (provided by the National DNA Bank Carlos III, University of Salamanca, Spain) as well as 350 men with tested fertility by semen analysis (spermatozoa number and motility), as previously described (6). Case and control populations were matched by age, ethnicity and geographical origin.

Informed written consents were signed by all participants, and the procedures followed in this study were approved by the local ethical committees of every participating center, according to the tenets of the Declaration of Helsinki.

The selection criteria used to include the infertile men were based on a thorough exam of individuals showing total absence of sperm in ejaculated (NOA) or <5 million spermatozoa/mL semen (SO) confirmed by two high-speed centrifugation processes in two different semen samples, consistent with the guidelines of the World Health Organization (7). The medical history records were revisited to extract information related to physical examination, karyotype analysis, endocrine analysis of follicle stimulating hormone (FSH), luteinizing hormone (LH), and testosterone, as well as Y chromosome microdeletions screening, and patients with known causes of infertility were excluded from the study. In this regard, only individuals with normal karyotype, absence of Yq deletions, and a normal history of testicular development were included. In addition, those patients with a testicular biopsy performed, were classified into different subgroups according to clinical and histological data, including hypospermatogenesis (HS, extremely low numbers of mature motile sperm cells in few testicular locations), maturation arrest of germ cells (MA, >90% of maturation arrest of the germ line either at spermatogonia or at primary spermatocyte levels), and Sertoli-cell only syndrome (SCO, total absence of germ cells). Two additional subgroups were also established accordingly with the outcome in the TESE techniques (including both gross and micro-TESE), named TESE- (including those NOA individuals in which no mature sperm cell could be retrieved from the biopsy) and TESE+ (patients with a successful sperm extraction from the biopsy). All the available information about clinical features of the patients is shown in **Supplemental Table 1**.

## **SNP selection and genotyping**

Tree intronic variants of *USP8* (rs7174015), *DPF3* (rs10129954), and *EPST11* (rs12870438) as well as two intergenic variants of the regions harboring *PSAT1* (rs7867029) and *TUSC1* (rs10966811) were selected to determine their possible association with male infertility traits in our study population. The SNPs selection was based on the findings by Kosova *et al.* (5), in which they were reported to correlate with family size in a Hutterite population and with semen parameters in an independent cohort of Chicago men.

Genomic DNA was extracted from peripheral white blood cells using the QIAamp® DNA Blood Midi/Maxi (Qiagen, Hilden, Germany) or MagNA Pure LC – DNA LV Isolation kit I (Roche, Basel, Switzerland), following the procedures described by the manufacturers. The genotyping was carried out using the TaqMan™ SNP genotyping technology (Applied Biosystems, Foster City, California, USA). The real-time quantitative PCRs and the post-PCR allelic discriminations were performed with predesigned TaqMan™ probes (assay IDs: C\_\_26249696\_10, C\_\_31364474\_20, C\_\_32072246\_20, C\_\_30534824\_10 and C\_\_3123309\_10) on a 7900HT Fast Real-Time PCR System (Applied Biosystems, Foster City, California, USA), as described elsewhere (6).

## **Statistical analysis**

CaTS Power Calculator for Genetic Studies program (8) was used to estimate the statistical power of our study. All the statistical analyses were performed with the software Plink v1.9 (9). Possible deviance from Hardy-Weinberg equilibrium (HWE) was evaluated in both cases and controls at the 5% significance level. To test for association between the candidate SNPs and male infertility traits, different case-control comparisons were conducted. In a first step, the whole group of SpF cases was compared against the control one. Afterwards, SpF men were divided into two different subgroups (SO and NOA) and, finally, the NOA set was further subdivided into four additional subgroups (SCO, MA, HS and TESE-). All the established case subgroups were tested against both controls and the remaining cases not showing the specific clinical phenotype for every subgroup (in order to eliminate having NOA or SO as possible confounding variable). Allele and genotype frequencies of every tested group were compared by means of logistic regression with geographical origin (Spain or Portugal) as covariate, and assuming additive, recessive, dominant, and 2 degree of freedom (genotypic) models. P-values, odds ratios (ORs), and their 95% confidence intervals (CI) were then calculated, and P-values lower than 0.05 were considered

statistically significant. Possible multiple testing effects were evaluated with the Bonferroni method.

### ***In silico characterization of associated variants***

Publicly available functional annotation data were explored to evaluate the possible functional implications of the observed associations using different bioinformatic tools. In a first step, we identified all the proxies ( $r^2 > 0.8$ ) of the associated lead SNPs in the European population using LDLink (10). All proxies were considered equally as candidates for prioritizing causality or hypothesising possible underlying molecular mechanisms for the observed associations with male infertility traits. The GTEx Portal (<https://www.gtexportal.org/>) (11) was used to prioritise eQTL and sQTL effect in the testis. Single-cell expression in the human testis of genes influenced by the studied SNPs was queried in the Single Cell Expression Atlas portal (<https://www.ebi.ac.uk/gxa/sc>) (12). Furthermore, we downloaded the call sets from the ENCODE portal (13) (<https://www.encodeproject.org/>) with the following identifiers: ENCFF323BCL, ENCFF608KRZ; ENCFF300WML, ENCFF559LDF, ENCFF644JKD, ENCFF767LMP, ENCFF788RFY, ENCFF855EVV, ENCFF286DAB, ENCFF509DBT, ENCFF316MJM, ENCFF610XSK, ENCFF819NRA, ENCFF711LHL and ENCFF881OHS, to evaluate different regulatory chromatin marks, such as DNase-seq hypersensitivity sites, CTCF protein ChIP-seqs, H3K4me3, H3K4me1, H3K27ac, H3K9me3, and H3K27me3 histone modification ChIP-seqs. SNP-based information was also extracted from Haploreg v.4.1. (14) (<https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php>) and SNPnexus (15) (<https://www.snp-nexus.org/>) to further assess the potential significance of the candidate sequence variants. These portals integrate the variant annotations from different databases, such as Ensembl, SIFT, Polyphen, CpG, Vista enhancers, miRbase, TarBase, TargetScan, miRNA Registry, snoRNA-LBME-DB, Roadmap Epigenomics, Ensembl regulatory build, RegulomeDB (16), and functional consequence predictions based on several algorithms such as: CADD, DeepSEA, EIGEN, FATHMM, fitCons, FunSeq2 GWAVA, REMM (**Supplemental Tables 2 and 3**).

In addition, to provide an illustrative picture of the putative functional role of the tested variants, we conducted enrichment analyses of both gene ontology (GO) terms and protein-protein interactions, considering all predicted transcription factors whose binding sites (TFBS) were altered by the lead SNPs and their proxies according to position weight matrices (PWM), using the tools for that purpose of the Retrieval of Interacting Genes/Proteins (STRING) portal (17).



### 3. RESULTS

This study was conducted with an appropriate overall statistical power, as shown in **Supplemental Table 4**. No significant deviation from HWE either in cases or controls was observed ( $p < 0.05$ ). The genotyping success rate for every analyzed SNP was over 98%, and the minor allele frequencies (MAF) of the control groups were consistent with those of both the Iberian and the European populations of the 1KGPh3 (18). The above evidences reinforce the reliability of the generated data and the proper implementation of the methodology used.

#### ***Susceptibility to non-obstructive azoospermia and specific histological manifestations***

In a first approach, we compared the allele and genotype frequencies of the five analyzed SNPs between the SpF group (which comprises all the infertile individuals of our study cohort) with those of the control population. No significant differences between them were observed under any of the tested models (**Table 1**).

Subsequently, we compared the NOA group and the different NOA subgroups against the control one. Significant P-values were observed in the analysis of the *USP8*-rs7174015 SNP frequencies of NOA cases against controls under both the additive and recessive models ( $P_{ADD}=0.0402$ ,  $OR=1.18$ ,  $P_{REC}=0.0226$ ,  $OR=1.33$ ), and a suggestive P-value was obtained in the genotypic model ( $P_{GENO}=0.0709$ ) (**Table 1**). Consistent with this, similar results were obtained when the NOA group was compared against the SO one ( $P_{ADD}=0.0323$ ,  $OR=0.77$ ;  $P_{REC}=0.0048$ ,  $OR=0.56$ ;  $P_{GENO}=0.0178$ ) (**Table 2**). The association under the recessive model remained significant after adjustment for multiple testing ( $P_{REC-BONF}=0.0242$ ).

In addition, a trend towards association was evident for this SNP when the allele frequencies between the subgroup of NOA patients with a negative TESE outcome were compared against both the control one ( $P_{ADD}=0.0594$ ,  $OR=1.28$ ,  $P_{REC}=0.0977$ ,  $OR=1.38$ ) and the subgroup of NOA patients with a positive TESE outcome ( $P_{ADD}=0.0865$ ,  $OR=1.4$ ) (**Tables 1, 2**). Finally, suggestive P-values were also yielded in the HS+ vs HS- comparison under both the additive ( $P_{ADD}=0.0727$ ,  $OR=0.64$ ) and recessive ( $P_{REC}=0.0824$ ,  $OR=0.48$ ) models (**Table 2**).

The subphenotype analysis between NOA cases with and without specific histological patterns/TESE success also reached statistical significance in the analysis of the *TUSC1*-rs10966811 polymorphism. The minor allele of such SNP showed a significant recessive risk for the HS subphenotype ( $P_{REC}=0.0205$ ,  $OR=2.88$ ). Consistent

with this observation, the *TUSC1*-rs10966811 genotype frequencies were also significantly different between the NOA subgroup of patients with HS and that without this specific spermatogenic failure ( $P_{\text{GENO}}=0.0295$ ). Similarly, the comparison between TESE+ vs TESE- NOA patients also evidenced that this same minor allele conferred risk for an unsuccessful TESE in a recessive manner ( $P_{\text{REC}}=0.0407$ ,  $\text{OR}=0.44$ ) (**Table 2**). The remaining analyzed SNPs (*DPF3*-rs10129954, *EPST11*-rs12870438 and *PSAT1*-rs7867029) showed no evidence of association with any of the histological patterns considered (either when the NOA subgroups were compared against the control population or in the intra-disease comparisons).

### ***Susceptibility to severe oligospermia***

A protective effect for SO predisposition was evidenced for the minor allele of *EPST11*-rs12870438 in the case-control comparison under both the additive and dominant models ( $P_{\text{ADD}}=0.0229$ ,  $\text{OR}=0.75$ ,  $P_{\text{DOM}}=0.0388$ ,  $\text{OR}=0.70$ ). The genotype distribution of this SNP was considerably different (albeit not significant) between the SO group and the control one ( $P_{\text{GENO}}=0.0745$ ) (**Table 1**). Suggestive P-values were also found for *PSAT1*-rs7867029 in the SO vs controls analysis under both the additive and dominant models ( $P_{\text{ADD}}=0.0728$ ,  $\text{OR}=0.71$ ;  $P_{\text{DOM}}=0.0548$ ,  $\text{OR}=0.67$ ) (**Table 1**).

On the other hand, when the SO group was compared against the NOA one (in order to detect SO-specific associations), significant differences in the allele frequencies were found for *PSAT1*-rs7867029 considering both additive and dominant effects ( $P_{\text{ADD}}=0.0351$ ,  $\text{OR}=0.66$ ;  $P_{\text{DOM}}=0.0187$ ,  $\text{OR}=0.61$ ). The genotype distributions between SO and NOA groups for this SNP also differed significantly ( $P_{\text{GENO}}=0.0487$ ) (**Table 2**).

No evidence of association was observed in any of the tests performed between SO versus both NOA and control groups for *DPF3*-rs10129954 or *TUSC1*-rs10966811 (**Tables 1, 2**).

### ***Evaluation of functional annotations***

We further searched for functional annotations of the 5 polymorphisms included in this study and their proxies ( $r^2>0.8$ ) in the European population of the 1KGPh3 (**Supplemental Tables 5-7**). None of the lead or proxy variants were located in coding regions, CpG Islands, or miRNA target sequences according to SNPnexus (15). Because of that, we decided to focus on other possible regulatory effects that may alter the normal gene expression levels in the testis, exploring first the transcriptome data in the GTEx project (analysis release V8) (11).

As indicated in **Figure 1**, the lead SNP variant *USP8*-rs7174015 and 19 of its proxies displayed evidences of functionality in the testicular tissue as expression quantitative trait *loci* (eQTL), with 11 of them affecting the expression levels of *USP8*, *USP50*, and *AP4E1*, and the remaining ones influencing also the *RP11-562A8.5* transcription levels (**Figure 1**). Interestingly, these four genes showed a considerable high expression in the testis according to both the Human Protein Atlas (19) (<http://www.proteinatlas.org>) and the GTEx database (11) (**Supplemental Figures 1-4**). Indeed, a testis-specific expression was evident for *USP50* and *RP11-562A8* (**Supplemental Figures 2 and 4**). Besides, the SNPs in this LD block were also annotated as eQTLs and sQTLs in multiples tissues, including ovary (**Supplemental Table 5 and 6**).

At the cellular level, recently published data from single-cell RNA-seq experiments on puberty human testes (**Figure 2A**) (20) showed that: 1) *USP8* was mostly expressed in spermatogonia, spermatocytes, spermatids, and Sertoli cells (**Figure 2B**), 2) *USP50* was detected almost exclusively in spermatocytes and spermatids (**Figure 2C**), and 3) *AP4E1* had a diffuse expression in multiple cell types (**Figure 2D**), thus suggesting a possible role of their encoded proteins in the spermatogenic process. No single-cell transcriptome data was available for *RP11-562A8*.

Moreover, six of the above mentioned linked SNPs (including *USP8*-rs7174015) overlapped with chromatin marks related to active enhancers (H3K37ac and H3K4me1), active promoters (H3K4me3), and with a TFBS of CTCF (which is involved in the conformation of the topologically associated domains) in the adult testis, according to ChIP-seq ENCODE data (13) (**Figure 1 and Supplemental Table 5**). These variants also mapped into *loci* with a number of different overlapping regulatory marks in multiple tissues (including ovary) and cell lines according to Roadmap Epigenomics, ENCODE, and Ensembl Regulatory Build databases (13, 21, 22), thus supporting the putative regulatory relevance for this region. The output data obtained from Haploreg (14) for the *USP8*-rs7174015 LD block highlighted a large number of TFBS that were predicted to be altered by such linked SNPs based on PWM data (**Supplemental Tables 5 and 7**). We decided to prioritize then according to overlap with putative testis-specific TFBS by querying the GeneCards Suite (23) and by performing a comprehensive bibliographic search. Notably, 8 out of all the tested SNPs were predicted to change the binding motif site of transcription factors potentially involved in the testicular function (**Figure 2, Supplemental Tables 5 and 8**). For instance, rs3098174 and rs56398519 were predicted to change the TFBS of FOXJ1, a transcription factor specifically required for the formation of motile cilia and which have been reported as an important member of a pathway involved in sperm motility and flagellum morphogenesis in murine models (24).

Similarly, the rs3098171 SNP modified the TFBS of HSF1, a stress-inducible and DNA-binding transcription factor that plays a central role in the activation of the heat shock response (HSR), and which has been proposed essential for spermatogenesis (25). Both rs12593481 and rs3131574 SNPs were annotated to alter the TFBS of PAX5 and NR6A1, respectively. These transcription factors have a known key role in spermatogenesis and are highly related to sperm formation and male infertility (26) (**Figure 2 and Supplemental Table 5**). Different scores indicative of a possible functional effect of the tested variants were also calculated with tools like RegulomeDB, CADD, deppSEA, EIGEN, FATHMM, fitCons and ReMM (**Figure 2 and Supplemental Table 5**). Overall, both *USP8*-rs7174015 and rs12593481 showed higher scores, thus suggesting that they are the most likely causal variants of this LD block. The *USP8*-rs7174015 SNP and its proxies were also annotated as eQTLs and sQTLs in multiples tissues (**Supplemental Table 5 and 6**), which highlights the high relevance of this genomic region in regulatory processes.

On the other hand, *TUSC1*-rs10966811, *EPSTI1*-rs12870438, *PSAT1*-rs7867029 and their corresponding proxies showed no significant effects on gene expression in the testis according to GTEx (11). However, rs10812205 (a *TUSC1*-rs10966811 proxy) as well as rs58357177, rs9590722, rs9594826, and rs9594827 (all of them *EPSTI1*-rs12870438 proxies) overlapped with an open chromatin state in the testis according to ChIP-seq data from ENCODE (13), and other regulatory marks in multiple tissues. Furthermore, the SNPs rs10966813 and rs11789162 (proxies of *TUSC1*-rs10966811) were located in predicted target sequences of DMRT2 (rs10966813), DMRT7 and DMRT1 (rs11789162) according to Haploreg (14), a family of transcription factors with a key role in male sex determination and spermatogenesis (27). The RegulomeDB score and the other functional prediction scores also suggested that the SNPs rs10812205, rs62534083, rs1535898, rs9590722, rs9594827, and rs9594829 were more likely to exert the functional effect (**Supplemental Table 7**).

Finally, to provide a global overview of the possible pathways involved in male infertility associated to the putative causal variants, we accomplished a protein-protein interaction (PPI) and biological pathway enrichment analysis with 199 transcription factors that had target sequences altered by such SNPs (**Supplemental Table 5 and 7**). The molecular network of the selected proteins had significantly more interactions than expected (number of nodes, 98; number of edges, 459; average node degree, 9,37; clustering coefficient, 0.372; expected number of edges, 89; PPI enrichment,  $P < 1 \times 10^{-16}$ , **Supplemental Figure 5**). Regarding the functional enrichment of the network, biological processes with the highest significant p-values were those related with gene expression

regulation processes (**Supplemental Table 9**), consistent with the evidences described above. Interestingly, spermatogenesis (GO:0007283) was one of the GO terms significantly enriched in the transcription factor set ( $P=0.0004$ ). Indeed, some members of this biological process, such as YY1, BCL6, HOXA10, ZBTB16 (PLZF), and PAX5 (highlighted in red in **Supplemental Figure 5**) represented relevant nodes in the PPI network.

#### 4. DISCUSSION

Idiopathic male infertility is supposed to have a complex etiology likely influenced by genetic, epigenetic, and environmental factors (3). Regarding its genetic basis, it has been estimated that the most severe expressions of this condition (NOA and SO) have a high heritability with a polygenic inheritance, in which many *loci* may exert an additive effect on the pathogenic phenotype (1). In the present study, we aimed to perform the first attempt to evaluate the possible implication in the development of SpF of 5 SNPs, previously associated with a reduced fertility in men (5), in the largest European case-control cohort included in a genetic study to date.

Our results suggest that both *EPSTI1*-rs12870438 and *PSAT1*-rs7867029 are involved in the pathological mechanisms underlying SO, whereas the intergenic SNP *USP8*-rs7174015 may contribute to the genetic susceptibility to overall NOA. Additionally, the minor allele of *TUSC1*-rs10966811 (A) was associated with a higher predisposition to HS in a context of NOA and, consequently, a better probability of TESE success.

Consistent with our observations, Kosova and colleagues (5) described that the risk alleles of the associated variants correlated with a decreased fertility in their study cohort. It could be speculated that the presence of such genetic variants may lead to different phenotypes related to male fertility depending on the specific genetic background of the individual, ranging from mild outcomes (such as reduced sperm counts or low birth rates) to more severe conditions like SO or NOA, which supports the notion of idiopathic male infertility as complex disease (1). In addition, *PSAT1*-rs7867029 and *USP8*-rs7174015 were significantly associated with SO predisposition and *EPSTI1*-rs12870438 with NOA risk in a low-powered Japanese population comprising 76 NOA patients, 50 SO patients, and 791 fertile men (28). However, the authors did not observe a correlation of such SNPs with semen parameters in an independent study cohort of Japanese males composed of 791 fertile men and 1224 young men from the general population (29). In a subsequent study, the same group also reported significant associations of *TUSC1*-rs10966811 (associated with HS under a NOA microenvironment in our study cohort) and *DPF3*-rs10129954 (which did not yield significant P-values in our analyses) with SO and SpF, respectively (30). The discrepancy of the results could be due to different genetic architectures of the regions encompassing those SNPs between Japanese and Iberian populations, or to a possible type I error affecting their results as a consequence of a reduced power (the case population included only 83 NOA patients and 62 SO patients). Indeed, for *DFP3*-rs10129954 the authors obtained significant P-values under opposite models (recessive and dominant) (30).

Despite the above, our results clearly suggest that *TUSC1*-rs10966811 may represent a potential marker of disease outcome after NOA development. The *TUSC1*-rs10966811\*G allele is associated with the most severe manifestation of this pathology (complete lack of sperm cells in the testis biopsy), whereas the presence of the *TUSC1*-rs10966811\*A allele is associated with the HS phenotype, the milder histological pattern of NOA. The functional annotations of this SNP are consistent with this idea. *TUSC1*-rs10966811 is located in a target sequence for YY1, a transcription factor that has been reported to play a major role in spermatogonial stem cell (SSC) maturation, being expressed in spermatocytes, spermatogonia, and spermatids, but not in mature spermatozoa (31, 32). The *TUSC1*-rs10966811 polymorphism represents a crucial position in the consensus sequence recognized by YY1, and the presence of the G allele correlates with a drastic decrease of the binding affinity (**Supplemental Table 8 and Supplemental Figure 6**). Other important transcription factors for the spermatogenic process have also predicted target sequences in the flanking regions of different *TUSC1*-rs10966811 proxies, such as BCL6, a repressor whose depletion causes testicular germ cell apoptosis in murine models (33). This protein is predicted to bind the genomic region containing rs10966813, showing a lower affinity in the presence of the rs10966813\*G allele, which is highly linked to *TUSC1*-rs10966811\*G (the risk allele for unsuccessful TESE). In addition, DMRT proteins are a family of testis-specific transcription factors that play a pivotal role in male sex determination and differentiation by controlling testis development and male germ cell proliferation (27). In this regard, the *TUSC1*-rs10966811 proxies rs10966813 and rs11789162 overlap with binding motifs of some members of this family, including DMRT1, DMRT2, and DMRT7. The gene encoding DMRT1 is a confirmed NOA-susceptibility *locus* (34-37), and the screening of its sequence to detect point mutations has been recently incorporated by some physicians in the routine clinical practice of idiopathic NOA to increase the diagnostic efficiency (38). Moreover, it has been reported that open chromatin in SSCs is considerably enriched in TFBS for DMRT1 (39). Besides, additional transcription factors involved in spermatogenesis have also predicted binding motifs within the *TUSC1*-rs10966811 haplotype block (**Supplemental Table 8**), suggesting that such block could have a potential interest for the development of prognostic markers of NOA.

On the other hand, our data suggest that the intergenic variant *USP8*-rs7174015\*A confers risk to NOA development acting as recessive allele. This result seems consistent, as the allele frequencies were significantly different between the NOA group and both the control population and the SO group. The results of our *in silico* analyses were also concordant with this association. Interestingly, *USP8*-rs7174015 is annotated

as an eQTL in the testis, affecting the expression of *USP8*, *USP50*, *AP4E1*, and *RP11-562A8.5*. The first of them has been reported to be highly expressed in male germ cells, in which it is involved in acrosome biogenesis (40, 41). Regarding *USP50*, *AP4E1*, and *RP11-562A8.5*, although their possible involvement in spermatogenesis has not been previously described, all three genes show a high expression in the testis (11). Indeed, *USP50* has a testis-specific expression, mostly in spermatocytes (**Figure 2**). Therefore, our data suggest that *USP8*-rs7174015\*A could exert its pathogenic influence in NOA predisposition by deregulating the baseline gene expression of *USP8*, *USP50*, *AP4E1*, and *RP11-562A8.5*. Such deregulation could be a consequence of an alteration of a binding protein motif by *USP8*-rs7174015\*A or any of its proxies (**Supplemental Table 8 and Supplemental Figure 7**). In this context, a proxy of this SNP, rs12593481, is located within a consensus sequence for PAX5 and YY1, which are relevant transcription factors in the regulation of the spermatogenic process (31, 32, 42).

Another highly linked SNP to *USP8*-rs7174015 is rs3098171, which maps into a putative TFBS for the stress-inducible protein HSF1. The encoding gene of this transcription factor is located within the azoospermia factor b (AZFb) region of the Y chromosome, and deletion of this region results in severe male infertility (43, 44). HSF proteins are expressed during mammalian spermatogenesis, mainly in spermatocytes and round spermatids (25). Disruption of different HSF members, such as HSF1 and HSF2, leads to male sterility and complete lack of mature sperm in mice, as these proteins have been reported to play an essential role in the repression of sex chromatin during meiosis (45). In this regard, the rs3098171\*G risk allele, which significantly reduces the expression of *USP8*, decreases drastically the affinity of HSF1 for the TFBS in which this SNP is located. Finally, it should be also noted that, at least, 5 proxies of *USP8*-rs7174015 are annotated to map active enhancers, active promoters, and/or TFBS in the testis through ChIP-seq studies according to ENCODE (13) (**Figure 2**), which strongly support a putative functional implication related to their position in the genome.

In relation to the SO-associated polymorphisms *EPST11*-rs12870438 and *PSAT1*-rs7867029, their allele frequencies in the SO group differed from those in both the control population and the NOA group (with the two latter cohorts showing similar allele and genotype frequencies), which could be indicative of a potential value of these two SNPs as diagnostic markers for SO compared to NOA. Interestingly, the rs9594826 variant, highly linked to *EPST11*-rs12870438, overlaps a target sequence of the transcription factor SIX5, which has been reported to decrease c-kit levels in adult mice, causing an elevated spermatogenic cell apoptosis and Leydig cell hyperproliferation (46). In this



case, a significant decrease in the binding affinity of SIX5 was also evident when the SO risk allele was present in the motif sequence (**Supplemental Table 8 and Supplemental Figure 8**).

In summary, our study may represent an important contribution to the current knowledge about the molecular mechanisms underlying SpF. We have evaluated the association of previously reported genetic factors associated with fertility in a well characterized cohort of SpF men of European ancestry. Our findings may shed light into their putative role in the development of specific male infertility histological patterns, and their possible use as prognostic markers of TESE success. Therefore, this study may represent a solid basis for future approaches aimed at developing more effective diagnostic and prognostic markers for severe cases of male infertility. In this sense, we identified 2 promising molecular markers of SO (*EPST11*-rs12870438 and *PSAT1*-rs7867029), one of NOA (*USP8*-rs7174015), and a potential marker of TESE success (*TUSC1*-rs10966811). The latter could be a good candidate to be included in a panel of SNPs that could anticipate the probability of unsuccessful surgeries for retrieving viable sperm cells from the testis, which represent around half of the total surgeries currently performed in NOA patients (47).

## **5. ACKNOWLEDGEMENTS**

We thank the National DNA Bank Carlos III (University of Salamanca, Spain) for supplying part of the control DNA samples from Spain, as well as all patients and controls for kindly agreeing to their essential collaboration. This work was supported by the Spanish Ministry of Economy and Competitiveness through the Spanish National Plan for Scientific and Technical Research and Innovation (ref. SAF2016-78722-R) the “Ramón y Cajal” program (ref. RYC-2014-16458), and the “Juan de la Cierva Incorporación” program (ref. IJC2018-038026-I), which include FEDER funds. SL received support from the Spanish Ministry of Science and Innovation (grants FIS-ISCIII DTS18/00101, co-funded by FEDER funds/European Regional Development Fund (ERDF)-a way to build Europe-), and from Generalitat de Catalunya (grant 2017SGR191). SL is sponsored by the “Researchers Consolidation Program” from the SNS-Dpt. Salut Generalitat de Catalunya (Exp. CES09/020). JG was partially funded by FCT/MCTES, through national funds attributed to Center for Toxicogenomics and Human Health - ToxOmics (UIDB/00009/2020). PIM is supported by the FCT post-doctoral fellowship (SFRH/BPD/120777/2016), financed from the Portuguese State Budget of the Ministry for Science, Technology and High Education and from the European Social Fund, available through the Programa Operacional do Capital Humano. This article is related to the Ph.D. Doctoral Thesis of Miriam Cerván-Martín.

## Figure Legends.

**Figure 1.** Enrichment of functional annotations of the human genome for the *USP8*-rs7174015 variant and its proxies. Overlaps are highlighted with different colors: blue for expression quantitative trait locus (eQTL) effects in the testis (affected genes are shown); green for active enhancers, active promoters, and transcription factor binding sites (TFBS) from chromatin immunoprecipitation followed by sequencing (ChIP) experiments in the testis (using ENCODE data); orange for other epigenetic marks of the ENCODE and Roadmap Epigenomics projects (such as histone methylation and DNAase hypersensitivity); violet for TFBS modifications related to transcription factors involved in spermatogenesis based on protein weight matrix (PWM) data; and pink for functional prediction scores, in which the heatmap displays the probability of functionality for each tested variant (dark pink indicates higher probability), according to the different calculation methods described in Supplemental Tables 2 and 3.

**Figure 2.** Gene expression in testicular cells of human adolescence subjects. **(A)** Dimension reduction (t-SNE) plots of single-cell transcriptome data in puberty human testes (n=31,671) based on RNA-seq dataset from Guo et al. (20). Single cells are represented as colored dots and the different colors indicate cluster identities. Specific expression patterns of *USP8* **(B)**, *UPS50* **(C)**, and *AP4A1* **(D)** projected on the t-SNE plot are shown. Tonality of blue correlates with expression levels (with dark blue indicating high expression) and gray indicates low or no expression.

## REFERENCES

1. Cervan-Martin M, Castilla JA, Palomino-Morales RJ, Carmona FD. Genetic Landscape of Nonobstructive Azoospermia and New Perspectives for the Clinic. *Journal of clinical medicine* 2020;9.
2. Tournaye H, Krausz C, Oates RD. Novel concepts in the aetiology of male reproductive impairment. *The lancet Diabetes & endocrinology* 2017;5:544-53.
3. Krausz C, Riera-Escamilla A. Genetics of male infertility. *Nature reviews Urology* 2018;15:369-84.
4. Hofker MH, Fu J, Wijmenga C. The genome revolution and its role in understanding complex diseases. *Biochimica et biophysica acta* 2014;1842:1889-95.
5. Kosova G, Scott NM, Niederberger C, Prins GS, Ober C. Genome-wide association study identifies candidate genes for male fertility traits in humans. *American journal of human genetics* 2012;90:950-61.
6. Cerván-Martín M, Suazo-Sánchez I, Rivera-Egea R, Garrido N, Lujan S, Romeu G *et al.* Intronic variation of the SOHLH2 gene confers risk to male reproductive impairment. *Fertility and sterility* 2020;*In press*.
7. Cooper TG, Noonan E, von Eckardstein S, Auger J, Baker HW, Behre HM *et al.* World Health Organization reference values for human semen characteristics. *Human reproduction update* 2010;16:231-45.
8. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nature genetics* 2006;38:209-13.
9. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* 2015;4:7.
10. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* 2015;31:3555-7.
11. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S *et al.* The Genotype-Tissue Expression (GTEx) project. *Nature genetics* 2013;45:580-5.
12. Papatheodorou I, Moreno P, Manning J, Fuentes AM, George N, Fexova S *et al.* Expression Atlas update: from tissues to single cells. *Nucleic acids research* 2020;48:D77-D83.

13. Sloan CA, Chan ET, Davidson JM, Malladi VS, Strattan JS, Hitz BC *et al.* ENCODE data at the ENCODE portal. *Nucleic acids research* 2016;44:D726-32.
14. Ward LD, Kellis M. HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic acids research* 2016;44:D877-81.
15. Oscanoa J, Sivapalan L, Gadaleta E, Dayem Ullah AZ, Lemoine NR, Chelala C. SNPnexus: a web server for functional annotation of human genome sequence variation (2020 update). *Nucleic acids research* 2020.
16. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome research* 2012;22:1790-7.
17. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J *et al.* STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic acids research* 2015;43:D447-52.
18. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO *et al.* A global reference for human genetic variation. *Nature* 2015;526:68-74.
19. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A *et al.* Proteomics. Tissue-based map of the human proteome. *Science* 2015;347:1260419.
20. Guo J, Nie X, Giebler M, Mlcochova H, Wang Y, Grow EJ *et al.* The Dynamic Transcriptional Cell Atlas of Testis Development during Human Puberty. *Cell stem cell* 2020;26:262-76 e4.
21. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* 2015;518:317-30.
22. Zerbino DR, Wilder SP, Johnson N, Juettemann T, Flicek PR. The ensembl regulatory build. *Genome biology* 2015;16:56.
23. Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S *et al.* The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Current protocols in bioinformatics* 2016;54:1 30 1-1 3.
24. Beckers A, Adis C, Schuster-Gossler K, Tveriakhina L, Ott T, Fuhl F *et al.* The FOXJ1 target Cfap206 is required for sperm motility, mucociliary clearance of the airways and brain development. *Development* 2020;*In press*.

25. Widlak W, Vydra N. The Role of Heat Shock Factors in Mammalian Spermatogenesis. *Advances in anatomy, embryology, and cell biology* 2017;222:45-65.
26. Fang F, Angulo B, Xia N, Sukhwani M, Wang Z, Carey CC *et al.* A PAX5-OCT4-PRDM1 developmental switch specifies human primordial germ cells. *Nature cell biology* 2018;20:655-65.
27. Zhang T, Zarkower D. DMRT proteins and coordination of mammalian spermatogenesis. *Stem cell research* 2017;24:195-202.
28. Sato Y, Tajima A, Tsunematsu K, Nozawa S, Yoshiike M, Koh E *et al.* An association study of four candidate loci for human male fertility traits with male infertility. *Hum Reprod* 2015;30:1510-4.
29. Sato Y, Tajima A, Tsunematsu K, Nozawa S, Yoshiike M, Koh E *et al.* Lack of replication of four candidate SNPs implicated in human male fertility traits: a large-scale population-based study. *Hum Reprod* 2015;30:1505-9.
30. Sato Y, Hasegawa C, Tajima A, Nozawa S, Yoshiike M, Koh E *et al.* Association of TUSC1 and DPF3 gene polymorphisms with male infertility. *Journal of assisted reproduction and genetics* 2018;35:257-63.
31. Kim JS, Chae JH, Cheon YP, Kim CG. Reciprocal localization of transcription factors YY1 and CP2c in spermatogonial stem cells and their putative roles during spermatogenesis. *Acta histochemica* 2016;118:685-92.
32. Bajusz I, Henry S, Sutus E, Kovacs G, Purity MK. Evolving Role of RING1 and YY1 Binding Protein in the Regulation of Germ-Cell-Specific Transcription. *Genes* 2019;10.
33. Kojima S, Hatano M, Okada S, Fukuda T, Toyama Y, Yuasa S *et al.* Testicular germ cell apoptosis in Bcl6-deficient mice. *Development* 2001;128:57-65.
34. Lopes AM, Aston KI, Thompson E, Carvalho F, Goncalves J, Huang N *et al.* Human spermatogenic failure purges deleterious mutation load from the autosomes and both sex chromosomes, including the gene DMRT1. *PLoS genetics* 2013;9:e1003349.
35. Tewes AC, Ledig S, Tuttelmann F, Kliesch S, Wieacker P. DMRT1 mutations are rarely associated with male infertility. *Fertility and sterility* 2014;102:816-20 e3.

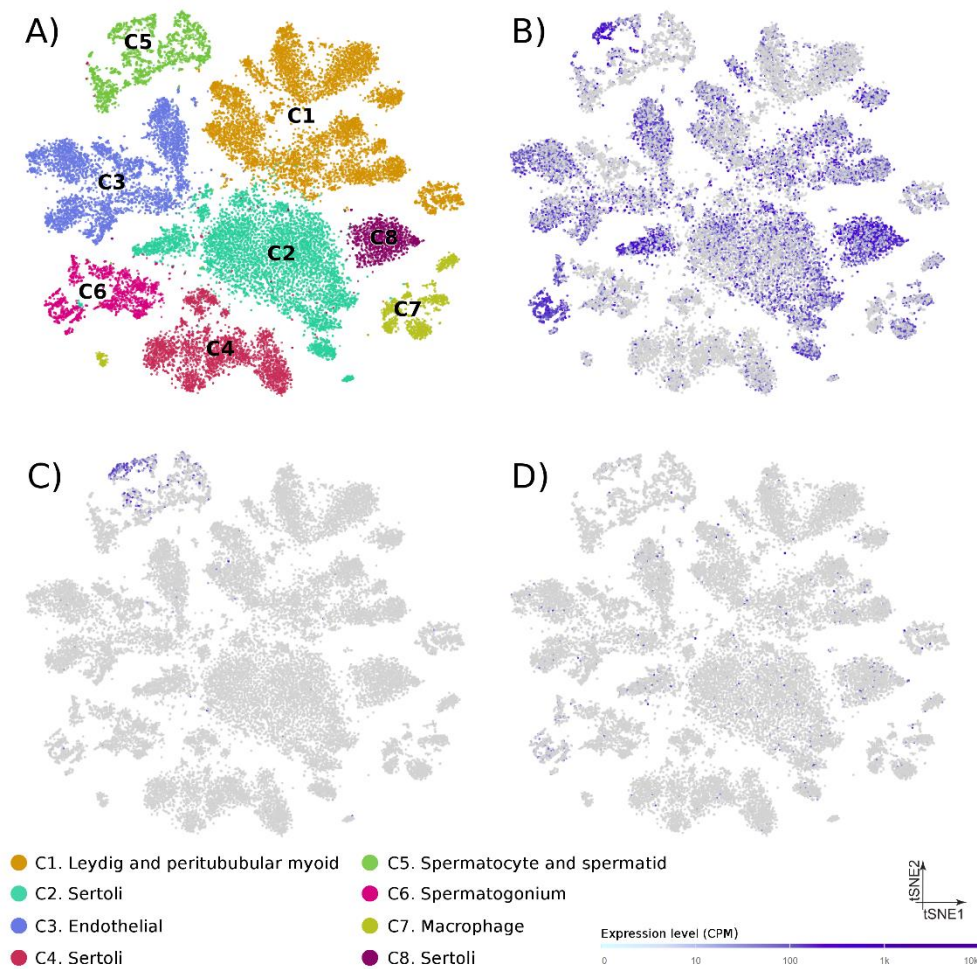
36. Araujo TF, Friedrich C, Grangeiro CHP, Martelli LR, Grzesiuk JD, Emich J *et al.* Sequence analysis of 37 candidate genes for male infertility: challenges in variant assessment and validating genes. *Andrology* 2020;8:434-41.
37. Lima AC, Carvalho F, Goncalves J, Fernandes S, Marques PI, Sousa M *et al.* Rare double sex and mab-3-related transcription factor 1 regulatory variants in severe spermatogenic failure. *Andrology* 2015;3:825-33.
38. Tuttelmann F, Ruckert C, Ropke A. Disorders of spermatogenesis: Perspectives for novel genetic diagnostics after 20 years of unchanged routine. *Medizinische Genetik : Mitteilungsblatt des Berufsverbandes Medizinische Genetik eV* 2018;30:12-20.
39. Guo J, Grow EJ, Yi C, Mlcochova H, Maher GJ, Lindskog C *et al.* Chromatin and Single-Cell RNA-Seq Profiling Reveal Dynamic Signaling and Metabolic Transitions during Human Spermatogonial Stem Cell Development. *Cell stem cell* 2017;21:533-46 e6.
40. Berruti G, Ripolone M, Ceriani M. USP8, a regulator of endosomal sorting, is involved in mouse acrosome biogenesis through interaction with the spermatid ESCRT-0 complex and microtubules. *Biology of reproduction* 2010;82:930-9.
41. Berruti G, Paiardi C. USP8/UBPy-regulated sorting and the development of sperm acrosome: the recruitment of MET. *Reproduction* 2015;149:633-44.
42. Adams B, Dorfler P, Aguzzi A, Kozmik Z, Urbanek P, Maurer-Fogy I *et al.* Pax-5 encodes the transcription factor BSAP and is expressed in B lymphocytes, the developing CNS, and adult testis. *Genes & development* 1992;6:1589-607.
43. Tessari A, Salata E, Ferlin A, Bartoloni L, Slongo ML, Foresta C. Characterization of HSFY, a novel AZFb gene on the Y chromosome with a possible role in human spermatogenesis. *Molecular human reproduction* 2004;10:253-8.
44. Shinka T, Sato Y, Chen G, Naroda T, Kinoshita K, Unemi Y *et al.* Molecular characterization of heat shock-like factor encoded on the human Y chromosome, and implications for male infertility. *Biology of reproduction* 2004;71:297-306.
45. Akerfelt M, Vihervaara A, Laiho A, Conter A, Christians ES, Sistonen L *et al.* Heat shock transcription factor 1 localizes to sex chromatin during meiotic repression. *The Journal of biological chemistry* 2010;285:34469-76.
46. Sarkar PS, Paul S, Han J, Reddy S. Six5 is required for spermatogenic cell survival and spermiogenesis. *Human molecular genetics* 2004;13:1421-31.

47. Vloeberghs V, Verheyen G, Haentjens P, Goossens A, Polyzos NP, Tournaye H.  
How successful is TESE-ICSI in couples with non-obstructive azoospermia? Hum  
Reprod 2015;30:1790-6.



SNP_ID	eQTL effects in the testis				ChIP-seq annotations in the testis			Other epigenetic marks	TFBS changed related to spermatogenesis	Functional prediction scores							
	AP4E1	USP8	USP50	RP11-562A8.5	Enhancers	Promoters	TFBS			RegulomeDB	CADD	DEEPSEA	EIGEN	FATHMM	FITCONS	FUNSEQ2	GWAVA
rs10152326																	
rs11070776																	
rs12593481																	
rs2289108																	
rs28582911																	
rs3098167																	
rs3098169																	
rs3098171																	
rs3098174																	
rs3098177																	
rs3098205																	
rs3131559																	
rs3131560																	
rs3131562																	
rs3131566																	
rs3131568																	
rs3131574																	
rs4318151																	
rs56398519																	
rs7174015																	

**Figure 1** - Enrichment of functional annotations of the human genome for the USP8-rs7174015 variant and its proxies.



**Figure 2** - Gene expression in testicular cells of human adolescence subjects. (A) Dimension reduction (t-SNE) plots of single-cell transcriptome data in puberty human testes (n=31,671) based on RNA-seq dataset from Guo et al. (20). Single cells are represented as as colored dots and the different colors indicate cluster identities. Specific expression patterns of *USP8* (B), *UPS50* (C), and *AP4A1* (D) projected on the t-SNE plot are shown. Tonality of blue correlates with expression levels (with dark blue indicating high expression) and gray indicates low or no expression

**Table 21** - Analysis of the genotype and allele frequencies of the tested genetic variants comparing subgroups of clinical phenotypes of male infertility against fertile controls.

SNP	1/2	Subgroup (N)	Genotype, N (%)				MAF (%)	Additive			Recessive			Dominant		Genotypic
			1/1	1/2	2/2	P-value		OR [CI 95%]*	P-value	OR [CI 95%]*	P-value	OR [CI 95%]*	P-value			
rs10129954	T/C	Controls (n = 1049)	220	501	328	0,4485	NA	NA	NA	NA	NA	NA	NA	NA	NA	
		SpF (n = 709)	139	344	226	0,4386	0,9563	1.00 [0.87-1.16]	0,7	0.95 [0.74-1.23]	0,6761	1.05 [0.84-1.30]	0,7821			
		SO (n = 222)	47	96	79	0,4279	0,9992	1.00 [0.80-1.25]	0,4821	1.15 [0.77-1.72]	0,5508	0.90 [0.64-1.27]	0,5193			
		NOA (n = 487)	92	248	147	0,4435	0,8727	1.01 [0.87-1.19]	0,588	0.93 [0.70-1.22]	0,4761	1.09 [0.86-1.39]	0,5502			
		SCO (n = 101)	23	51	27	0,4802	0,3117	1.16 [0.87-1.55]	0,5376	1.17 [0.71-1.91]	0,3108	1.27 [0.80-2.01]	0,5741			
		MA (n = 51)	11	28	12	0,4902	0,2648	1.26 [0.84-1.89]	0,6254	1.19 [0.59-2.38]	0,207	1.54 [0.79-3.00]	0,4497			
		HS (n = 48)	7	24	17	0,3958	0,4602	0.85 [0.56-1.30]	0,4439	0.72 [0.32-1.65]	0,6403	0.86 [0.47-1.60]	0,7274			
		TESE- (n = 140)	28	77	35	0,475	0,4636	1.10 [0.86-1.40]	0,6977	0.92 [0.59-1.43]	0,1402	1.36 [0.90-2.03]	0,2148			
		rs10966811	A/G	Controls (n = 1047)	136	520	391	0,3782	NA	NA	NA	NA	NA	NA	NA	NA
SpF (n = 707)	97			319	291	0,3628	0,2533	0.92 [0.79-1.06]	0,8327	1.03 [0.77-1.39]	0,0835	0.83 [0.68-1.02]	0,1635			
SO (n = 220)	34			100	86	0,3818	0,8223	0.97 [0.76-1.24]	0,5376	1.16 [0.73-1.83]	0,4481	0.88 [0.63-1.23]	0,5016			
NOA (n = 487)	63			219	205	0,3542	0,191	0.90 [0.76-1.06]	0,9548	0.99 [0.71-1.38]	0,0779	0.82 [0.65-1.02]	0,1854			
SCO (n = 100)	10			50	40	0,35	0,4008	0.87 [0.64-1.20]	0,3903	0.74 [0.38-1.47]	0,5761	0.89 [0.58-1.35]	0,6573			
MA (n = 51)	5			27	19	0,3627	0,72	0.92 [0.60-1.42]	0,4907	0.72 [0.28-1.85]	0,9897	1.00 [0.55-1.80]	0,7728			
HS (n = 48)	10			17	21	0,3854	0,9298	1.02 [0.66-1.58]	0,1317	1.76 [0.84-3.66]	0,3402	0.75 [0.41-1.36]	0,1099			
TESE- (n = 140)	13			66	61	0,3286	0,1011	0.80 [0.61-1.05]	0,2201	0.69 [0.38-1.25]	0,1605	0.77 [0.54-1.11]	0,2617			
rs12870438	A/G			Controls (n = 1048)	155	502	391	0,3874	NA	NA	NA	NA	NA	NA	NA	
		SpF (n = 711)	101	324	286	0,3699	0,3529	0.93 [0.80-1.08]	0,7861	0.96 [0.72-1.28]	0,2642	0.89 [0.72-1.09]	0,5336			
		SO (n = 220)	24	100	96	0,3364	<b>2,29E-02</b>	0.75 [0.59-0.96]	0,1162	0.67 [0.40-1.10]	<b>3,88E-02</b>	0.70 [0.50-0.98]	<b>7,45E-02</b>			
		NOA (n = 491)	77	224	190	0,3849	0,9243	0.99 [0.85-1.16]	0,6533	1.07 [0.79-1.46]	0,6405	0.95 [0.75-1.19]	0,7316			

		SCO (n = 102)	16	47	39	0,3873	0,9635	0.99 [0.74-1.34]	0,8307	1.06 [0.61-1.87]	0,8237	0.95 [0.63-1.45]	0,932
		MA (n = 51)	7	23	21	0,3627	0,5218	0.87 [0.57-1.33]	0,7797	0.89 [0.39-2.03]	0,4816	0.81 [0.45-1.45]	0,7794
		HS (n = 48)	7	26	15	0,4167	0,6148	1.12 [0.73-1.71]	0,9391	0.97 [0.42-2.22]	0,441	1.28 [0.68-2.41]	0,7019
		TESE- (n = 141)	19	64	58	0,3617	0,4134	0.90 [0.69-1.16]	0,6876	0.90 [0.54-1.50]	0,3877	0.85 [0.60-1.22]	0,6828
rs7174015	A/G	Controls (n = 1048)	257	541	250	0,5033	NA	NA	NA	NA	NA	NA	NA
		SpF (n = 706)	189	351	166	0,5163	0,2097	1.10 [0.95-1.27]	0,1911	1.17 [0.93-1.47]	0,466	1.09 [0.86-1.39]	0,4042
		SO (n = 221)	44	119	58	0,4683	0,3802	0.90 [0.71-1.14]	0,3204	0.82 [0.55-1.22]	0,6622	0.92 [0.63-1.34]	0,6048
		NOA (n = 485)	145	232	108	0,5381	4,02E-02	1.18 [1.01-1.38]	2,26E-02	1.33 [1.04-1.71]	0,2963	1.15 [0.88-1.50]	7,09E-02
		SCO (n = 102)	29	53	20	0,5441	0,213	1.21 [0.90-1.62]	0,3443	1.25 [0.79-1.96]	0,2819	1.32 [0.79-2.21]	0,4586
		MA (n = 51)	16	27	8	0,5784	0,1132	1.40 [0.92-2.13]	0,2259	1.46 [0.79-2.71]	0,1774	1.70 [0.79-3.70]	0,2878
		HS (n = 47)	8	26	13	0,4468	0,3802	0.82 [0.54-1.27]	0,3204	0.67 [0.31-1.47]	0,665	0.86 [0.44-1.68]	0,6057
		TESE- (n = 141)	44	71	26	0,5638	0,0594	1.28 [0.99-1.65]	0,0977	1.38 [0.94-2.03]	0,1611	1.38 [0.88-2.16]	0,1671
rs7867029	C/G	Controls (n = 1050)	15	251	784	0,1338	NA	NA	NA	NA	NA	NA	NA
		SpF (n = 711)	10	155	546	0,1231	0,3597	0.90 [0.73-1.12]	0,943	1.03 [0.44-2.43]	0,3081	0.88 [0.70-1.12]	0,57
		SO (n = 221)	3	37	181	0,0973	0,0728	0.71 [0.49-1.03]	0,8494	0.87 [0.22-3.50]	0,0548	0.67 [0.45-1.01]	0,1534
		NOA (n = 490)	7	118	365	0,1347	0,9019	0.99 [0.78-1.24]	0,9668	1.02 [0.40-2.58]	0,8844	0.98 [0.76-1.26]	0,9868
		SCO (n = 103)	2	27	74	0,1505	0,5421	1.14 [0.75-1.71]	0,727	1.31 [0.29-5.84]	0,5694	1.14 [0.72-1.79]	0,8276
		MA (n = 50)	1	10	39	0,12	0,6734	0.87 [0.46-1.64]	0,7665	1.37 [0.17-10.97]	0,5862	0.83 [0.41-1.65]	0,789
		HS (n = 48)	1	15	32	0,1771	0,2394	1.40 [0.80-2.45]	0,7368	1.43 [0.18-11.52]	0,2336	1.46 [0.78-2.74]	0,49
		TESE- (n = 141)	4	29	108	0,1312	0,9095	0.98 [0.67-1.42]	0,2036	2.07 [0.67-6.35]	0,6195	0.90 [0.59-1.36]	0,3245

\*Odds ratio for the minor allele. SpF, spermatogenic failure; NOA, non-obstructive azoospermia; SCO, Sertoli cell-only; MA, meiotic arrest; HS, hypospermatogenesis; TESE, testicular sperm extraction; SO, severe oligospermia.

**Table 22** - Analysis of the allele and genotype frequencies of the tested genetic variants in Iberian infertile men accordingly to the presence ("with manifestation") and absence ("without manifestation") of specific clinical phenotypes

SNP	1/2	Subgroup (N)	With Manifestation				Without Manifestation				Additive		Recessive		Dominant		Genotypic
			1/1	½	2/2	MAF (%)	1/1	1/2	2/2	MAF (%)	P-value	OR [CI 95%]**	P-value	OR [CI 95%]**	P-value	OR [CI 95%]**	P-value
rs10129954	T/C	SO/NOA (n = 222/487)	47	96	79	0,4279	92	248	147	0,4435	0,7559	0.96 [0.76-1.22]	0,7171	1.08 [0.71-1.63]	0,4325	0.87 [0.61-1.24]	0,5803
		SCO/noSCO (n = 101/130)	23	51	27	0,4802	25	66	39	0,4462	0,5245	1.13 [0.78-1.64]	0,5187	1.24 [0.65-2.34]	0,6871	1.13 [0.63-2.02]	0,7952
		MA/noMA (n = 51/180)	11	28	12	0,4902	37	89	54	0,4528	0,3862	1.22 [0.78-1.91]	0,8525	1.08 [0.50-2.32]	0,2418	1.55 [0.74-3.24]	0,4924
		HS/noHS (n = 48/183)	7	24	17	0,3958	41	93	49	0,4781	0,2133	0.74 [0.46-1.19]	0,2413	0.59 [0.24-1.43]	0,3809	0.73 [0.37-1.47]	0,4421
		TESE-/TESE+ (n = 140/92)	28	77	35	0,475	16	46	30	0,4239	0,2538	1.26 [0.85-1.86]	0,6221	1.19 [0.60-2.35]	0,1948	1.47 [0.82-2.63]	0,4291
		SO/NOA (n = 220/487)	34	100	86	0,3818	63	219	205	0,3542	0,4268	1.10 [0.87-1.40]	0,4662	1.19 [0.74-1.92]	0,5477	1.11 [0.79-1.56]	0,7137
rs10966811	A/G	SCO/noSCO (n = 100/130)	10	50	40	0,35	15	62	53	0,3538	0,9608	0.99 [0.66-1.48]	0,7534	0.87 [0.37-2.04]	0,8936	1.04 [0.61-1.77]	0,9262
		MA/noMA (n = 51/179)	5	27	19	0,3627	20	85	74	0,3492	0,8442	1.05 [0.65-1.70]	0,7049	0.82 [0.29-2.33]	0,6135	1.18 [0.62-2.26]	0,7571
		HS/noHS (n = 48/182)	10	17	21	0,3854	15	95	72	0,3434	0,47	1.20 [0.73-1.96]	<b>2,05E-02</b>	2.88 [1.18-7.07]	0,5707	0.83 [0.43-1.59]	<b>0,0295</b>
		TESE-/TESE+ (n = 140/92)	13	66	61	0,3286	17	37	38	0,3859	0,1983	0.78 [0.53-1.14]	<b>4,07E-02</b>	0.44 [0.20-0.97]	0,7109	0.90 [0.53-1.54]	0,1164
		SO/NOA (n = 220/491)	24	100	96	0,3364	77	224	190	0,3849	0,1257	0.83 [0.65-1.05]	0,1024	0.65 [0.39-1.09]	0,3205	0.84 [0.60-1.18]	0,2359
		SCO/noSCO (n = 102/130)	16	47	39	0,3873	20	64	46	0,4	0,7351	0.94 [0.64-1.37]	0,917	1.04 [0.51-2.13]	0,5725	0.86 [0.50-1.47]	0,8137
rs12870438	A/G	MA/noMA (n = 51/181)	7	23	21	0,3627	29	88	64	0,4033	0,5193	0.86 [0.54-1.36]	0,6358	0.80 [0.33-1.98]	0,5673	0.83 [0.43-1.58]	0,8124
		HS/noHS (n = 48/184)	7	26	15	0,4167	29	85	70	0,3886	0,5325	1.16 [0.73-1.84]	0,7741	0.88 [0.35-2.17]	0,2649	1.49 [0.74-2.98]	0,4186
		TESE-/TESE+ (n = 141/93)	19	64	58	0,3617	20	40	33	0,4301	0,1689	0.77 [0.53-1.12]	0,11	0.57 [0.28-1.14]	0,4356	0.81 [0.47-1.39]	0,2722
		SO/NOA (n = 221/485)	44	119	58	0,4683	145	232	108	0,5381	<b>3,23E-02</b>	0.77 [0.61-0.98]	<b>4,84E-03</b>	0.56 [0.38-0.84]	0,5185	0.88 [0.60-1.30]	<b>1,78E-02</b>
		SCO/noSCO (n = 102/128)	29	53	20	0,5441	33	69	26	0,5273	0,7789	1.06 [0.72-1.55]	0,6459	1.15 [0.64-2.06]	0,9733	0.99 [0.51-1.91]	0,885
		MA/noMA (n = 51/179)	16	27	8	0,5784	46	95	38	0,5223	0,2298	1.33 [0.84-2.11]	0,4234	1.32 [0.66-2.64]	0,2446	1.66 [0.71-3.90]	0,459
rs7174015	A/G	HS/noHS (n = 47/183)	8	26	13	0,4468	54	96	33	0,5574	0,0727	0.64 [0.40-1.04]	0,0824	0.48 [0.21-1.10]	0,2532	0.64 [0.30-1.37]	0,184

		TESE-/TESE+ (n = 141/91)	44	71	26	0,5638	21	46	24	0,4835	0,0865	1.40 [0.95-2.04]	0,1738	1.52 [0.83-2.79]	0,15	1.59 [0.85-3.00]	0,2286
rs7867029	C/G	SO/NOA (n = 221/490)	3	37	181	0,0973	7	118	365	0,1347	3,51E-02	0.66 [0.45-0.97]	0,8202	1.18 [0.28-4.97]	1,87E-02	0.61 [0.40-0.92]	0,0487
		SCO/noSCO (n = 103/129)	2	27	74	0,1505	2	32	95	0,1395	0,6068	1.15 [0.67-1.96]	0,8594	1.20 [0.16-8.71]	0,6053	1.17 [0.65-2.11]	0,8734
		MA/noMA ( n = 50/182)	1	10	39	0,12	3	49	130	0,1511	0,2618	0.67 [0.33-1.35]	0,7842	1.38 [0.14-14.05]	0,1921	0.60 [0.28-1.29]	0,366
		HS/noHS (n = 48/184)	1	15	32	0,1771	3	44	137	0,1359	0,4862	1.25 [0.66-2.37]	0,7374	1.49 [0.15-15.29]	0,5046	1.27 [0.63-2.57]	0,7846
		TESE-/TESE+ (n = 141/91)	4	29	108	0,1312	0	22	69	0,1209	0,7643	1.09 [0.62-1.91]	0,9986	1.036e+09 [0.00-Inf]	0,8794	0.95 [0.51-1.77]	0,8637

\*Odds ratios (OR) and 95% confidence intervals (CI) considering NOA as cases and SO as controls: additive=1.29 (1.02-1.64), recessive=1.78 (1.19-2.65), dominant=1.14 (0.77-1.67). \*\*Odds ratio for the minor allele. NOA: non-obstructive azoospermia; SCO, Sertoli cell-only; MA, meiotic arrest; HS, hypospermatogenesis; TESE, testicular sperm extraction; SO, severe oligospermia.

