THE EFFECTS OF DISEASE-CAUSING X-CHROMOSOMAL VARIANTS IN HETEROZYGOUS CARRIERS

LOTTA MIELIKÄINEN

MASTER'S THESIS

UNIVERSITY OF HELSINKI

FACULTY OF BIOLOGICAL AND ENVIRONMENTAL SCIENCES

GENETICS AND GENOMICS

FEBRUARY 2022

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta – Fakultet – Faculty The Faculty of Biological and Environmental Sciences		Koulutusohjelma – Utbildningsprogram – Degree Programme Master's Programme in Genetics and Molecular Biosciences
Tekijä – Författare – Author Lotta Mielikäinen		
Työn nimi – Arbetets titel – Title The effects of disease-causing X-chromosomal variants in heterozygous carriers		
Oppiaine/Opintosuunta – Läroämne/Studieinriktning – Subject/Study track Genetics and Genomics		
Työn laji – Arbetets art – Level Master's thesis	Aika – Datum – Month and year February 2022	Sivumäärä – Sidoantal – Number of pages 61

Tiivistelmä – Referat – Abstract

Sex determination in humans occurs via the sex chromosomes, X and Y. Females carry two X chromosomes while males are XY individuals. Due to this X chromosome distribution the expression of X-linked genes is balanced with a process called X chromosome inactivation (XCI) where one of the X chromosomes is silenced, selected either randomly or preferentially, in early female embryogenesis. X-linked disorders are more prevalent in males as, generally, in females the effects of a disease-causing variant in other of the X chromosomes can be compensated with the normal allele on the other X whereas male express the allele on their only X chromosome. However, cases of heterozygous females manifesting an assumed recessive X-linked disorder have been reported although the symptoms are usually milder in these cases than in males. One suggested reason behind this is a skewed XCI where the majority of female's cells express the mutated allele.

The main goal of this thesis was to examine how often heterozygous female carriers have symptoms of X-linked disorders. To achieve this goal, likely pathogenic and pathogenic X-chromosomal variants were retrieved from the ClinVar database and their global allele frequencies were examined from The Genome Aggregation Database (gnomAD). The genetic and phenotypic data of 500,000 individuals from the UK Biobank (UKB) were used to conduct genetic association analyses between the ClinVar variants and quantitative traits related to their reported phenotypes. The associations were tested in males and in females separately to allow for examination of sex-specific effects and inheritance models via the comparison of effect sizes.

89 (likely) pathogenic variants were detected from UKB, and the majority of these were extremely rare with minor allele frequency below 0.01% in the global population. 11 and 27 of them were selected for the association analyses for the male and female populations of UKB, respectively, after filtering out variants that did not meet requirements such as enough carriers. One to five quantitative traits were chosen for each variant resulting in 28 tests among males and 87 among females. These analyses showed few significant associations while the majority of the tested variants were observed to have no effects on the chosen trait. The most statistically significant association was observed with variant rs137852591 on the gene *AR* (androgen receptor) in males. The variant was related to lower muscle mass and shorter height that are associated partial androgen insensitivity syndrome reported in ClinVar for this variant. Nominally significant associations were seen with this variant and the same traits in heterozygous females suggesting that there might be, indeed, symptoms of the syndrome in females as well. Additionally, in both sexes variants on gene *G6PD* seemed related to traits that are characteristics of glucose 6 phosphate dehydrogenase deficiency.

The limitations of these databases must be taken into account when conducting studies utilizing them. However, this thesis demonstrated that heterozygous female carriers may have symptoms of X-linked disorders assumed to have recessive inheritance pattern. In the future, a wider set of phenotypes could be used to investigate the impacts of the X-linked variants more broadly.

Avainsanat – Nyckelord – Keywords
X chromosome, X-chromosomal variants, UK Biobank, ClinVar

Ohjaaja tai ohjaajat – Handledare – Supervisor or supervisors
Taru Tukiainen

Säilytyspaikka – Förvaringställe – Where deposited
HELDA – Digital Repository of the University of Helsinki

Muita tietoja – Övriga uppgifter – Additional information

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta – Fakultet – Faculty Bio- ja ympäristötieteellinen tiedekunta	Koulutusohjelma – Utbildningsprogram – Degree Programme Genetiikan ja molekulaaristen biotieteiden maisteriohjelma	
Tekijä – Författare – Author Lotta Mielikäinen		
Työn nimi – Arbetets titel – Title Tauteja aiheuttavien X-kromosomaalisten varianttien vaikutukset heterotsygoottisissa kantajissa		
Oppiaine/Opintosuunta – Läroämne/Studieinriktning – Subject/Study track Genetiikka ja genomiikka		
Työn laji – Arbetets art – Level Maisterintutkielma	Aika – Datum – Month and year Helmikuu 2022	Sivumäärä – Sidoantal – Number of pages 61

Tiivistelmä – Referat – Abstract

Ihmisillä sukupuoli määräytyy sukupuolikromosomien X ja Y kautta. Naiset kantavat kahta X-kromosomia, kun taas miehet ovat XY-yksilöitä. Tämän X-kromosomijakauman vuoksi X-kromosomaalisten geenien ilmentymistä tasapainotetaan prosessilla X-kromosomin inaktivaatio, jossa varhaisessa naarasalkiossa toinen, joko satunnaisesti tai ensisijaisesti valittu, X-kromosomeista hiljennetään. X-sidonnaiset taudit ovat yleisempiä miehillä, sillä yleensä naisilla tautia aiheuttavan variantin vaikutukset voidaan kompensoida toisen X-kromosomin normaalilla alleelilla, mutta miehillä ilmenee se alleeli, joka on heidän ainoassa X-kromosomissaan. On kuitenkin raportoitu tapauksia, joissa heterotsygoottinen nainen ilmentää oletettavasti resessiivisesti periytyvää X-sidonnaista tautia, vaikka oireet ovat näissä tapauksissa usein lievempiä kuin miehillä. Eräs syy tähän voisi olla vinoutunut X-kromosomin inaktivaatio, jossa suurin osa naisen soluista ilmentää mutatoitunutta alleelia.

Tämän tutkielman päätavoitteena oli tutkia, kuinka usein heterotsygoottisilla naiskantajilla on oireita X-sidonnaisista taudeista. Tämän tavoitteen saavuttamiseksi, (todennäköisesti) patogeeniset X-kromosomaaliset variantit kerättiin ClinVar -tietokannasta, ja niiden varianttien globaalit alleelifrekvenssit tutkittiin tietokannasta The Genome Aggregation Database (gnomAD). UK Biobank (UKB) -biopankin 500 000 yksilön geneettisiä ja fenotyyppisiä tietoja hyödynnettiin geneettisissä assosiaatioanalyyseissä, joissa tutkittiin ClinVar-varianttien vaikutusta niiden raportoitujen fenotyyppien kvantitatiivisiin ominaisuuksiin. Assosiaatiot testattiin erikseen miehillä ja naisilla, jotta voitaisiin huomata sukupuolispesifit vaikutukset ja periytymismallit verraten varianttien efektikokoja.

89 (todennäköisesti) patogeenistä varianttia havaittiin UKB:sta ja suurin osa näistä varianteista oli erittäin harvinaisia eli niiden alleelifrekvenssi globaalisti oli alle 0,01 %. Sen jälkeen, kun suodatettiin pois variantit, jotka eivät täyttäneet vaatimuksia, kuten tarpeeksi kantajia, niistä valittiin 11 ja 27 miesten ja naisten assosiaatioanalyyseihin. Yhdestä viiteen kvantitatiivista ominaisuutta valittiin kullekin variantille, minkä tuloksena tehtiin miehillä 28 testiä ja naisilla 87 testiä. Nämä analyysit osoittivat muutamia merkittäviä assosiaatioita, kun taas suurimmalla osalla testatuista varianteista ei havaittu olevan vaikutusta valittuun ominaisuuteen. Tilastollisesti merkittävin assosiaatio miehillä havaittiin geenin *AR* (androgeenireseptori) variantilla rs137852591. Variantti voitiin yhdistää alhaisempaan lihasmassaan sekä lyhyyteen, jotka ovat ominaisia piirteitä osittaisessa androgeeni-insensitiivisyyssydroomassa, joka on raportoitu kyseiselle variantille ClinVar:ssa. Tämä variantti yhdistettiin samoihin ominaisuuksiin myös naisten assosiaatioanalyyseissä, mikä voisi viitata siihen, että naisellakin voisi olla oireita syndroomasta. Lisäksi, molemmilla sukupuolilla variantit geenissä *G6PD* tuntuivat liittyvän ominaisuuksiin, jotka ovat tyypillisiä glukoosi-6-fosfaattidehydrogenaasin puutokselle.

Käytettyjen tietokantojen rajoitukset tulee ottaa huomioon, kun tehdään tutkimuksia niitä käyttäen. Kuitenkin, tämä tutkielma osoitti, että heterotsygoottiset kantajanaaraat voivat omata oireita X-sidonnaisista taudeista, joilla oletetaan olevan resessiivinen periytymismalli. Tulevaisuudessa, X-sidonnaisten varianttien vaikutuksia tutkiessa voitaisiin käyttää suurempaa otosta fenotyyppejä, jotta saataisiin laajempi tutkimustulos.

Avainsanat – Nyckelord – Keywords
X-kromosomi, X-kromosomaaliset variantit, UK Biobank, ClinVar

Ohjaaja tai ohjaajat – Handledare – Supervisor or supervisors
Taru Tukiainen

Säilytyspaikka – Förvaringställe – Where deposited
HELDA – Helsingin yliopiston digitaalinen arkisto

Muita tietoja – Övriga uppgifter – Additional information

Contents

ABBREVIATIONS

AF	Allele frequency
AR	Androgen receptor
BMI	Body mass index
CI	Confidence interval
gnomAD	The Genome Aggregation Database
GWAS	Genome-wide association study
ID	Intellectual disability
MAF	Minor allele frequency
OMIM	Online Mendelian Inheritance of Man
PAIS	Partial androgen insensitivity syndrome
PAR	Pseudoautosomal region
PC	Principal component
PCA	Principal component analysis
pLI	Probability of loss of function intolerance
SE	Standard error
SNV	Single nucleotide variant
UKB	UK Biobank
VCF	Variant Call Format
WGS	Whole-genome sequencing
XCI	X chromosome inactivation
XIC	X-inactivation center

INTRODUCTION

1.1 THE HUMAN X CHROMOSOME

1.1.1 Overview of the human genome and sex determination

The 3.1 giga base human nuclear genome is compressed in 22 autosomal chromosomes and two sex chromosomes, X and Y. In somatic cells chromosomes are represented as pairs, resulting in 44 autosomal chromosomes and two sex chromosomes. While every individual carries all 22 of the autosome pairs, except for certain genetic disorders (e.g., Down syndrome), the distribution of sex chromosomes is unbalanced between males (46,XY) and females (46,XX). Female gametes contain 22 autosomes and an X chromosome while male gametes have either an X or a Y chromosome in addition to the 22 autosomal chromosomes. In other words, females are homogametic and males heterogametic. In fertilization, one female gamete and one male gamete are combined to form a diploid zygote with 44 autosomes and two sex chromosomes (Figure 1). The biological sex of an individual is determined by the sex chromosome inherited from the father since all offspring receive an X chromosome from the mother. Atypical number of sex chromosomes results in aneuploidies, such as Turner syndrome (45,X0) and Klinefelter syndrome (47,XXY).

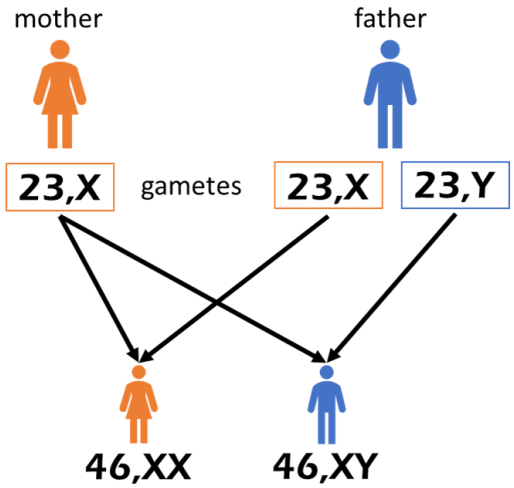

Figure 1. Sex determination in humans. Female gametes contain 22 autosomal chromosomes and an X chromosome while male gametes carry either an X or a Y chromosome. An X chromosome from mother is passed on to every offspring whereas the sex chromosome inherited from father determines the biological sex of the individual: daughters receive an X chromosome and sons inherit a Y chromosome.

The biological development into males and females is determined by the genes on sex chromosomes together with environmental factors and hormones. Primary sexual differentiation, the determination of the gonads into testis or ovaries, begins approximately in a 5-week-old embryo and it is dependent on the embryo's genotype. Gene *SRY* (sex-determining region on the chromosome Y) on chromosome Y is considered to play an important role in male development (Kashimada & Koopman, 2010). Testes are developed from gonads in males whereas the absence of *SRY* in XX embryos leads to the formation of ovaries (Kashimada & Koopman, 2010). Additionally, certain X-linked and autosomal genes, such as *DAX1* and *WNT4*, are related to the sexual differentiation by positively regulating ovarian development (Gilbert, 2000).

On the other hand, secondary sexual characteristics are mediated by the hormonal signaling rather than directly by the genotype and *SRY* although the Y-linked gene regulates male hormonal activity (Strachan & Read, 2018). For instance, genes *AMH* and *HSD17B3* are required to produce anti-Mullerian hormone and testosterone. The formation of the male urogenital system is dependent on these hormones. Additionally, without the male-determining hormones individual develops into female, meaning that female secondary sex characteristics act as default in the human development. The lack of anti-Mullerian hormone leads to the development of Fallopian tubes and uterus.

Defects in the hormonal system may cause an XY individual to express external female characteristics, such as breasts and wide hips with low amount of body hair (Strachan & Read, 2018). An example of a disorder causing feminized males is androgen insensitivity syndrome, result of a mutation in X-linked gene *AR*. In a same manner, an XX individual could have external features resembling males including increase in body hair and decreased breast size should they carry a mutation causing errors in the hormonal signaling (Strachan & Read, 2018).

1.1.2 Size and genes of the X chromosome

The length of the human X chromosome is approximately 156 Mb, which is approximately 5% of the human genome, whereas Y chromosome is only one third of its counterpart's size (57 Mb). There are an estimated 857 and 692 protein-coding and non-coding genes on the X chromosome, respectively (Howe et al., 2021). In addition, the number of pseudogenes on the X chromosome is

predicted at 888 (Howe et al., 2021). Although the X chromosome has many more genes than the Y chromosome, it is relatively gene-poor compared to autosomes that are approximately the same size (Ross et al., 2005). Small regions on the tips of X and Y chromosomes are considered as pseudoautosomal regions (PARs), containing at least 29 genes (Blaschke & Rappold, 2006). These regions are homologous between the sex chromosomes, and they are inherited in a same manner as autosomes. Generally, genes on PARs are not included when considering X-linked genes due to their inheritance pattern. The recombination of sex chromosomes in males is restricted to PARs whereas in females the two X chromosomes can recombine in a same manner as autosomes (Ross et al., 2005).

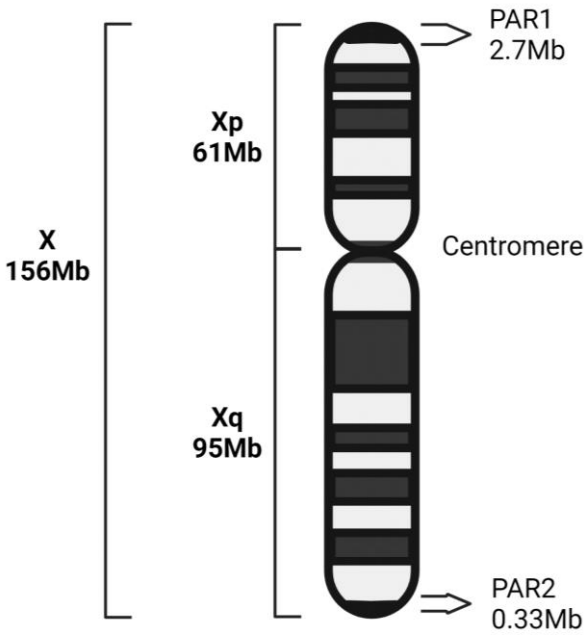

Figure 2. A simplified illustration of the human X chromosome. At the tips of the arms shown in are the pseudoautosomal regions (PARs, not in scale) that have homologues on the Y chromosome. Created with BioRender.com.

X-chromosomal genes have impact on multiple functions in humans. The target tissues of X-linked genes include, for instance, retina, blood, neural, and hepatic (Basta & Pandya, 2021). Studies have shown that X-chromosomal genes are especially involved in neurological functions (Deng et al., 2014) and naturally they have impacts on the reproductive system (Graves, 2006). The expression of X-linked genes is enriched in the human brain compared to other tissues (Nguyen &

Disteche, 2006). For example, X-chromosomal genes *DLG3*, *RPS6KA3* and *SLC25A5* associate with brain functions (Laumonnier et al., 2007). In addition to the genes expressed in the brain, an example of X-linked gene expressed in muscles is the largest gene in the human genome, *DMD* (Duchenne muscular dystrophy). Also, X-linked genes have been shown to be enriched in immune-related functions (Bianchi et al., 2012). For example, X-linked genes *CYBB* and *BTK* encode proteins that are connected to phagocyte structure and B cell development, respectively (Bianchi et al., 2012).

Although X-linked genes impact on many tissues, evidence shows that there are many X-chromosomal genes with male-specific functions (Lercher et al., 2003). These genes usually function in the testes (Lercher et al., 2003). Explanations for the enrichment of these so-called male-specific genes include sexual antagonism, the transmission time of the X chromosome and the fact that males carry only one X chromosome (Gurbich & Bachtrog, 2008). In sexual antagonism genes are positively selected in one sex whereas the other sex may have harmful consequences from those genes. As the transmission of X chromosome is female-biased and males carry only one X, antagonistic mutations are suggested to be accumulated on the X chromosome (Gurbich & Bachtrog, 2008). Usually, the male-specific genes are ampliconic, meaning that there are multiple adjacent duplications of small genomic regions, and they are expressed in the testes (Deng et al., 2014).

1.1.3 X chromosome inactivation (XCI)

In 1961, Mary Lyon suggested that to balance the expression of X-chromosomal genes in XY male and XX female mice, one of the female X chromosomes is inactivated in every cell. This X chromosome inactivation (XCI), in other words Lyonization, takes place in early mammalian female embryogenesis in the blastocyst stage (Rebuzzini et al., 2020). In each cell, only one X chromosome remains active while the majority of the genes on the other X are silenced. Additionally, individuals with sex chromosome aneuploidies have only one active X chromosome. The selection, whether the maternal or paternal X chromosome will be inactivated, is determined by each somatic cell and the choice is generally random. The inactivation is inherited through mitosis to the daughter cells meaning that the X with the same parental background is the active

one through the cell lineage. Thus, females are mosaics as approximately half of their cells express the maternal X chromosome and half have active paternal X chromosome.

XCI is initiated at the X-inactivation center (XIC, Xq13.2) that encodes *XIST* (X-inactivation-specific transcript), a long non-coding RNA that is exclusively expressed from the inactive X chromosome (Brown et al., 1991). The inactive chromosome is physically coated with the non-coding RNA (Figure 3) that recruits repressive proteins, such as Polycomb complexes. With these proteins the inactive X is epigenetically changed to contain repressive histone modifications, such as the trimethylation of lysine 9 and lysine 27 of histone 3 (H3K9me3 and H3K27me3, respectively), and it is condensed into transcriptionally inactive heterochromatin conformation. The condensed X chromosome is called Barr body (Galupa & Heard, 2018).

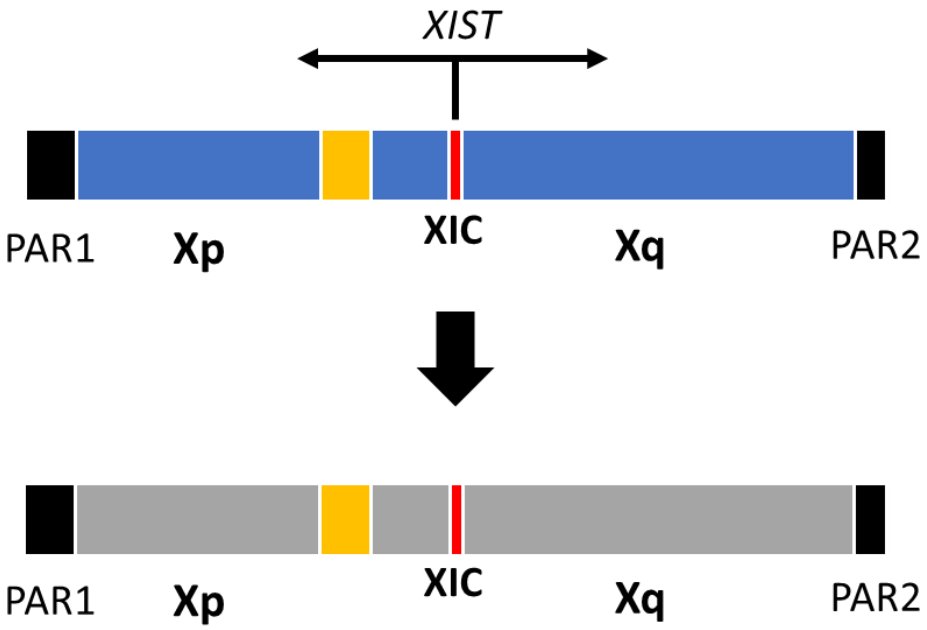

Figure 3. X chromosome inactivation is initiated from the X-inactivation centre (XCI) shown in red. *XIST* coats the X chromosome chosen to be inactivated and recruits repressive proteins to silence the genes on the X. Blue represents active X chromosome and grey is the inactivated X. Yellow indicates the centromeres. Figure not in scale.

While the majority of the genes on the inactive X are silenced, more than 15%, maybe even over 20%, of them remain transcriptionally active in both the active and inactive X chromosomes (Carrel & Willard, 2005; Cotton et al., 2013; Tukiainen et al., 2017). First, it was predicted that the genes on PARs escaped the X chromosome inactivation in females (Lyon, 1962). The reasoning

behind this was that PARs are homologous between the sex chromosomes and the gene expression on these regions is therefore already balanced between males and females. Later, it was discovered that, in addition to PARs, other genes escape the inactivation as well (Carrel & Willard, 2005). However, the escape seems incomplete for the genes as they are only partially expressed from the inactive X chromosome showing heterogeneity in levels of X-linked gene expression in humans (Cotton et al., 2013). In addition to the genes on PARs, other X-linked genes have homologous genes on the Y chromosome as well which can explain their escape status (Posynick & Brown, 2019). It is also suggested that some genes may be expressed from the inactive X chromosome because they have an advantage for the individual (Posynick & Brown, 2019).

1.1.4 Skewed XCI

As the XCI is generally random the number of cells expressing maternal and paternal X chromosomes is expected to be approximately equal. Deviation from this ratio is considered as skewing of the X chromosome inactivation with the definition of either of the chromosomes being active in over 75% of the cells affecting approximately 35% of females (Amos-Landgraf et al., 2006; Minks et al., 2008). An extreme case of skewed XCI is when over 90% of the cells have the same X inactivated and 7% of females have been shown to have extremely skewed XCI in whole blood (Amos-Landgraf et al., 2006). Random and skewed XCI are illustrated in Figure 4. Individual with heterozygous genotype in addition to skewed XCI may be functionally homozygous as one of the X chromosomes is active in majority of cells.

Skewed XCI may occur by chance, by primary skewing in the time of inactivation, or due to secondary selection (Santos-Rebouças et al., 2020). In the time of the inactivation, there is a limited number of cells which is why the skewed XCI is possible via chance. Primary skewing may be caused, for example, due to mutations in genes involved in the inactivation process, such as *XIST* (Santos-Rebouças et al., 2020). Negative selection can contribute to XCI as harmful mutations in either of the X chromosomes can be compensated by the allele on the other X. Resulting in this secondary selection, the X with the non-harmful allele is preferred and is therefore chosen as the

active chromosome in more cells. Also, positive selection may occur should there be advantageous alleles on the other X chromosome (Zito et al., 2019).

The skewness varies between tissues and, in addition, evidence of age-related skewed XCI have been discovered (Busque et al., 1996; Zito et al., 2019). Increase of XCI skewness levels in blood tissues have been observed after 50-60 years of age (Busque et al., 1996; Kristiansen et al., 2005; Wong et al, 2011). Zito et al. (2019) carried out a twin-cohort study assessing the tissue-specific levels of XCI skewness. They discovered that the prevalence of skewness in XCI differs between tissues, and the study also supported previous findings of positive association between skewing levels and aging. Additionally, skewed X chromosome inactivation is possibly related to environmental factors and lifestyle, such as smoking (Zito et al. 2019). The heritability of XCI-skew has mainly been studied with blood-derived samples which have shown that skewed XCI is heritable at least in blood cells, such as granulocytes (Vickers et al., 2001; Kristiansen et al., 2005). However, other tissue types have been poorly included in these studies and no clear indication of XCI-skew being heritable is available (Zito et al., 2019).

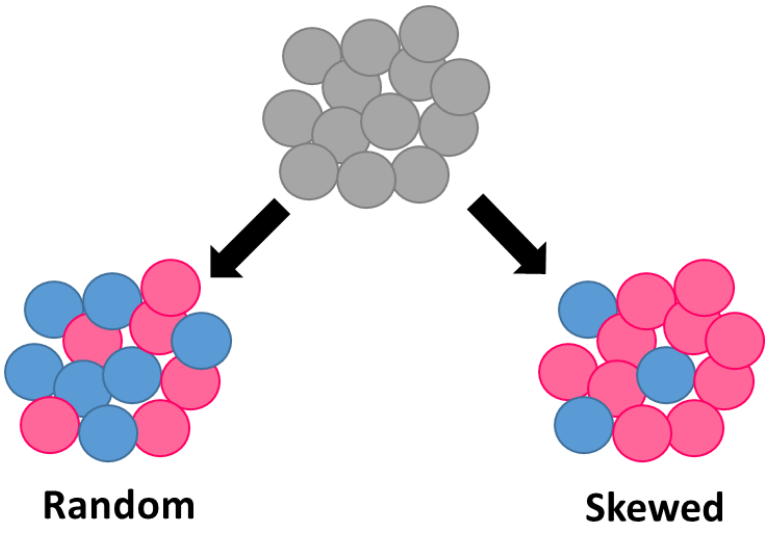

Figure 4. Random and skewed X chromosome inactivation. Blue cells illustrate cells with active paternal X chromosome and pink ones carry active maternal X. In random XCI, individual has roughly equal number of cells with inactivated paternal and maternal X chromosome. Should the individual have gone through skewed XCI, the other X chromosome is preferred over the other and is activated in more cells.

1.2 STUDY OF X-LINKED DISORDERS

1.2.1 Variants in the human genome

Human features and diseases can be roughly divided into monogenic (Mendelian) and complex traits. The presence or absence of a monogenic trait is dependent on the genotype at a single locus although it is also affected by other factors (Strachan & Read, 2018). The majority of phenotypes are, however, complex traits affected by multiple genetic variants and environmental factors (Plomin et al., 2009). The variants contributing to complex disorders usually have small effect sizes and only alter the probability of manifesting the disease unlike monogenic variants (Risch, 2000). Disease-causing and disease-associated variants are also called pathogenic variants although a variant can be classified as pathogenic without a known association. The gene's intolerance against pathogenic variants is called mutational constraint and this can be observed from large population samples where the studied gene has less variation than expected (Fuller et al., 2019).

Single nucleotide variant (SNV) is the most common type of variation in the human genome (The 1000 Genomes Project Consortium, 2015). SNV is a result of a DNA base substitution, and it is the most studied variation regarding complex traits (Goddard et al., 2016). Other types of variations include small insertions and deletions, microsatellites, and structural variants, such as inversions and copy number variants (Strachan & Read, 2018). Variants are also categorized based on their minor allele frequency (MAF) into common (MAF ≥ 5%), low-frequency (0.5% ≤ MAF < 5%) and rare (MAF < 0.5%) variants (Tam et al., 2019). Even rarer variants, denoted as extremely rare, have MAF below 0.01%. It is theorized that rarer pathogenic variants have larger effect sizes than common variations as it is not evolutionarily beneficial to have many of these large-effect disease-associated variants in the population (Strachan & Read, 2018).

1.2.2 X-linked disorders and their inheritance

Traits determined by X-chromosomal genes are considered to have an X-linked inheritance pattern. Typically, X-linked phenotypes have been further categorized into recessive and dominant

X-linked traits in a same manner as autosomal traits (Germain, 2006). Recessive alleles are manifested if individual's both copies of the gene carry the same genotype whereas dominant alleles are required to only be expressed from one copy. An exception to this is compound heterozygote who carry two different minor alleles within a gene (Miller & Piccolo, 2020). Pathogenic variants in X-chromosomal genes may lead to X-linked disorders, that are more prevalent in males than in females because the harmful effects can be compensated with the other healthy X chromosome in XX individuals (Trent, 2012). According to Online Mendelian Inheritance in Man (OMIM) database (https://www.omim.org/), which is a publicly available catalogue of human genes and phenotypes, there are nearly 220 known X-linked diseases (accessed 19.1.2022). These X-linked disorders include, for example, red-green color blindness, hemophilia A, and Duchenne muscular dystrophy. All of these are considered recessive X-linked disorders and are more common in males than in females. For example, according to MedlinePlus (https://medlineplus.gov/) 1 in 12 males and 1 in 200 females are affected by red-green color blindness.

Studies have shown that there is an enrichment of genes associated with intellectual disability (ID) on the X chromosome (Neri et al., 2018). The X-linked inheritance pattern was discovered as there is an excess of male individuals affected by ID compared to females. It is estimated that approximately 10% of ID cases in males are explained by X-chromosomal mutations (Vissers et al., 2016). Nearly 150 mutations associated with ID have been discovered on the X chromosome (Neri et al., 2018). Few examples of X-linked genes related to ID are *PQBP1*, *FLNA* and *ATRX* and associated disorders for these genes include Renpenning syndrome, Melnick-Needles syndrome, and Alpha-Thalassemia Intellectual Disability syndrome, respectively (Neri et al., 2018). The prevalence of ID-associated genes correlates with the number of genes involved in neurodevelopment on the X chromosome (Laumonnier et al., 2007).

The prevalence of X-chromosomal disorders in males can be explained by the unique inheritance of X chromosomes. Males express an X-linked disorder fully regardless of whether the inheritance mechanism is recessive or dominant as they carry a single X chromosome. Usually, under the recessive model of inheritance, females need to carry a mutated allele in both of their X chromosomes in order to express an X-linked disorder. X-linked disorders are never transmitted from father to son, but the son of a heterozygous mother is affected by recessive X-linked disorder

with a 50% possibility. Additionally, daughters of heterozygotes have a 50% chance of receiving the mutated allele without typically manifesting the disorder.

Many X-chromosomal phenotypes, however, do not seem to follow the rules of the classification into recessive and dominant traits, and the terms recessive and dominant are used less frequently when referring to X-linked disorders (Dobyns et al., 2004). The fact, that males and females carry different number of X chromosomes and females go through XCI, complicates the interpretation of X-chromosomal traits and their inheritance patterns. Regardless of the assumed recessiveness of certain X-linked disorders, their symptoms can appear in heterozygous females carrying one healthy allele and one mutant allele (Basta & Pandya, 2021). Examples of such diseases are adrenoleukodystrophy, hemophilia A, and Duchenne muscular dystrophy. However, the penetrance of X-linked disorders varies highly in females and the disorders seem less severe in females than in males (Dobyns, 2004; Germain, 2006; Migeon, 2020). Some dominant traits may be incompletely penetrant in heterozygotes while fully penetrant in hemizygous males. There are also X-linked disorders, that are only observed in females, explained by male-lethality in fetal state (Migeon, 2020). For instance, Melnick-Needles syndrome and CHILD syndrome are diseases nearly exclusively affecting females. Despite the difficulties in inheritance patterns, many of the X-chromosomal diseases have the categorization whether they are recessive or dominant in OMIM; however, there are also diseases referred only as X-linked without any further definition.

1.2.3 Contribution of XCI to X-linked disorders

The preferential inactivation of the healthy X chromosome in XX individuals can explain why heterozygous carrier females express X-linked disorders that are considered recessive. On the other hand, also heterozygous carriers unaffected by dominant alleles can be explained by XCI when the defect X chromosome is silenced in the majority of cells. As explained earlier (see section 1.1.3), females are mosaics as they have cells expressing maternal and paternal X chromosomes. Generally, in the case of random XCI, heterozygous females carry the mutated allele in approximately half of the cells and a normal allele in rest of the cells. In majority of X-linked disorder cases, this distribution is enough to counteract the effects of the mutant allele. In other words, the expression of normal allele from 50% of the heterozygous individual's cells

produces enough protein to prevent the individual manifesting the disease (Migeon, 2020). The explanation behind this is that tissues can function rather normally despite having only 50% of the gene products working properly (Juchniewicz et al., 2021). Additionally, normal cells can share their gene products to deficient cells should that be necessary (Juchniewicz et al., 2021).

Skewed XCI can have either positive or negative impacts when considering X-chromosomal disorders in heterozygous carriers. Occasionally, there is a positive selection towards the cell with the normal-functioning allele meaning that the cells with the mutated allele are lost. For example, in Wiskott-Aldrich syndrome, which is an immunodeficiency disorder, heterozygous females generally have skewed XCI favoring the cells with the healthy allele (Migeon, 2020). Conversely, studies have shown extreme skewing in favor of the mutated allele in some X-linked disorders; however, the reasons behind this remain still elusive. For instance, this pattern of skewness has been observed in heterozygous females affected by intellectual disabilities (Plenge et al., 2002; Vianna et al., 2020) and adrenoleukodystrophy (Maier et al., 2002; Engelen et al., 2014). One explanation for the skewness, in either direction, is pure chance as mentioned earlier (see section 1.1.4).

1.2.4 Current methods for studying genetic associations

Genetic association studies have been utilized to discover pathogenic variants regarding certain disorders or traits. These studies are designed to test the relationship between a variant and a phenotype, either a disease or a quantitative trait (Tam et al., 2019). In addition to commonly used SNVs, other variations, such as insertions and deletions, are utilized in association studies. Over the past 15 years or so, genome-wide association study (GWAS) has been the most implemented method in genetic association studies (Cirulli et al., 2020). In GWAS, a large number of genetic variants, nowadays usually in the order of millions, are tested for association against a certain phenotype (Visscher et al., 2017). Approximately 326,000 associations have been discovered with GWAS according to NHGRI GWAS catalog (accessed 28.1.2022; Buniello et al., 2019). These studies have given insight in genomic architecture and in different biological mechanisms (Tam et al., 2019). Typically, the variants detected with GWAS are considered

common, meaning that their MAF is over 5% in the population, and their effect size is usually small (Tam et al., 2019).

Low-frequency (0.5% ≤ MAF < 5%) and rare variants (MAF < 0.5%) are more challenging to detect with GWAS as their prevalence in the population is small. Usually, in a GWAS the power to detect these rarer variants is insufficient unless the effect sizes of the variants are very large (Strachan & Read, 2018). Next-generation sequencing, more accurate genotyping arrays, variant imputation, and large datasets enable us to study these rare variations as only few individuals carry them. In recent years, datasets with hundreds of thousands of participants have arisen giving enough power to detect rare-variant associations. For example, the UK Biobank and the FinnGen project contain genetic information from 500,000 and 350,000 individuals, respectively (Bycroft et al., 2018; https://www.finngen.fi/en 20.1.2022). In addition to large sample datasets, there are databases with information on different human variants. Databases, such as ClinVar and The HGV database, provide publicly available data of rare human variations and their relationship with phenotypes (Landrum et al. 2018; https://hgv.figshare.com/). Also, population databases, such as The Genome Aggregation Database (gnomAD) and The 1000 Genomes Project, are powerful tools in genetic studies (The 1000 Genomes Project Consortium, 2015; Karczewski et al., 2020).

A number of different software packages and pipelines have been created for running an association analysis. For instance, SNPStats, PLINK, SKAT, and REGENIE, are tools used for either single-variant association analyses, GWAS or both (Sole et al., 2006; Purcell et al., 2007; Wu et al., 2011; Mbatchou et al., 2021). Each of these tools have their own benefits and limitations regarding, for example, running time, file formats, quality control or accuracy. When designing a genetic association analysis, it is important to consider which tool works best for each purpose. For rare associations, the study design requires even more considerations. Whole-genome sequencing (WGS) provides naturally the largest number of genotype sites to be researched; however, in the case of rare variations, large sample sizes are crucial and therefore WGS can become expensive (Lee et al., 2014). Alternative options include whole-exome sequencing and genotyping arrays, but these have their own limitations. Whole-exome sequencing is restricted to the protein-coding DNA and genotyping arrays include only previously identified variations (Lee et al., 2014). Additionally, genotype imputation, where missing genotypes are predicted from a genotype or haplotype reference panel, can be conducted to increase the number of studied variations (Lee et al., 2014).

Typically, genetic association studies are performed in a case-control design, in which the allele frequencies are compared between the two groups (Tam et al., 2019). This logistic regression model is utilized when the trait of interest is binary, such as a disease. For continuous phenotypes, linear regression model is utilized (Lee et al., 2014). The computations of genetic associations are usually performed under an additive model (Equation 1) where the assumption is that the phenotype Y changes linearly by the number of copies of the minor allele (Wu et al., 2011). In the model

$$Y \sim \mu + X\beta + \varepsilon \quad \text{(Equation 1)}$$

Y represents the phenotype and μ is the mean phenotype. X determines the number of copies of the minor allele. The effect size for each copy of the alternative allele is β and ε determines the impact of errors.

There are many factors, genetic and environmental, behind complex disorders. Different confounders, such as the sex and age of the individual, need to be considered while conducting a genetic association analysis. The linear model is also adjusted for population stratification as population structure, where allele frequencies differ between populations due to ancestry, impacts the association analysis (Ma & Shi, 2020). The adjustment is usually carried out using principal components (PCs) received from principal component analysis (PCA) of genetic markers (Wu et al., 2011). The factors are included to the regression model as covariates. In this regression model

$$Y \sim \mu + Z\,\gamma + X\beta \quad \text{(Equation 2)}$$

Y and X are the phenotype and genotype, respectively, and Z represents a confounder of their association.

Association between a genotype and phenotype is typically evaluated using a p-value. A common nominal significance threshold for the p-value is 0.05. An association can be denoted as statistically significant should it reach the p-value threshold. The nominal p-value is, however, in the case of multiple tests insufficient and may produce false positive results. The threshold can be adjusted for multiple testing with Bonferroni correction, where the nominal p-values are divided by the number of individual tests (Armstrong, 2014). Other methods for correction, such as Holm-Bonferroni and Benjamini & Yekutiel, exist as well; however, Bonferroni correction is the most

commonly used in genetic association analyses. In GWAS the significance threshold for the p-value is generally accepted to be 0.05/1,000,000 = $5x10^{-8}$ as the number of independent common variants in the human genome is approximately one million (Tam et al., 2019).

1.2.5 Genetic association studies for X-chromosomal variations

The unique inheritance pattern and XCI of X-chromosomal genes in humans poses a challenge when designing a genetic association analysis for X-linked variants. This has led to X-linked variants being excluded from many association studies, such as GWAS, even though X-chromosomal genes contribute to human disease and traits (Fang et al., 2021). The exclusion is usually explained by the difficulties in implementing any analysis pipelines to X-linked variants (Wise et al., 2013). Additionally, the lack of X-chromosomal variants on genotyping chips utilized in GWAS have contributed to the absence of X-linked loci in the studies. Measures have been taken to overcome these issues, and in recent years genetic studies accounting for X-linked variations have been conducted (Tukiainen et al., 2014; Zuo et al., 2019; Martin et al., 2021).

Without proper considerations, the statistical analysis of X-linked variants results in errors and therefore it is recommended that the X-chromosomal variations are analyzed separately from autosomal variation as the number of X chromosomes differ in males and females (König et al., 2014). In the study design it is suggested that the number of female and male samples is close to equal as the excess of affected females may increase type I error rate (Özbek et al., 2018). Another consideration is the genotyping array as all of them may not include many X-chromosomal variants. Typically, in the genotype calling phase the efficient way is to use an algorithm that accounts for the sex of the sample as in the additive model the female and male genotypes are coded differently for X-linked genes (Özbek et al., 2018).

Statistical analysis of X-chromosomal variants differs from autosomal analysis if the regions of interest are not PARs which are homologous in X and Y because the coding of the genotype calls differs between the sexes (König et al., 2014). Also, if the samples contain only females, X-linked association study can be treated the same way as analysis on autosomes. However, for instance, REGENIE and PLINK have the options to account for the differences in X-chromosomal allele dosages in males and females (Purcell et al., 2007; Mbatchou et al., 2021). When interpreting the

results of association analysis of X-linked variants, it is important to consider the estimated effect sizes in males and females as the number of alleles is different between the sexes (Jons et al., 2019). The XCI status of the variant, or more precisely of the gene that the variant is on, sets expectations of the ratio of the effect sizes in females and males (Sidorenko et al., 2019). The expected female-male effect size ratio of genes going through XCI is close to 2 whereas the corresponding ratio of escape genes is close to 1 when female genotypes are coded {0,1,2} and male genotypes {0,1} (Sidorenko et al., 2019). However, male genotypes can be coded also as {0,2} which influences the expectations of the effects sizes. As not all tools are applicable to X-chromosomal studies, the selection of different tools for this purpose is limited.

2 AIMS OF THE THESIS

Phenotypic analyses of X-linked disorders are important because their inheritance patterns can be challenging to predict and the variability of the phenotypes in heterozygous females due to XCI complicates the diagnosis of these disorders, and the possibility of transferring the mutations to offspring may not be detected. In addition, X-chromosomal variants are often excluded from GWAS due to the challenges of different distribution of X-linked genes in females and males. The studies of X-chromosomal traits also give some insight into the skewing of X chromosome inactivation as some females manifesting an X-linked disorder may have skewed XCI. The main goal of this thesis is to assess how often heterozygous females express symptoms of X-linked phenotypes using publicly available and license-approved databases.

Aim 1: Collect pathogenic X-chromosomal variants from ClinVar, a submission-driven database with clinically relevant variants with their interpretation of clinical significance for associated conditions and examine their characteristics including population-level allele frequencies in 125,748 whole-exome sequenced individuals from the Genome Aggregation Database.

Aim 2: Investigate which of the variants found in ClinVar are present in the UK Biobank population. The Biobank has genotype and phenotype information from approximately 500,000 individuals, allowing for the analysis of rare X-chromosomal variants, should those be tolerated in a normal adult population and detected with the genotyping methods used.

Aim 3: Conduct genotype-phenotype association analyses for these X-chromosomal variants and disease-relevant quantitative phenotypes by comparing heterozygous females and reference homozygous females. In addition, replication of the reported phenotypes in ClinVar is attempted in the UK Biobank male population by conducting an association analysis only with the male-population of UKB.

3 MATERIALS AND METHODS

3.1 DATASETS

3.1.1 CLINVAR

ClinVar is a submission-driven, publicly available database comprising of clinically relevant variants maintained by National Center for Biotechnology Information, part of National Institutes of Health (Landrum et al., 2018). As of now, there are nearly 1.9 million submitted records from 2,080 submitters in ClinVar (30.11.2021). Approximately 1.18 million of the records are unique variations and they cover circa 36,000 protein-coding and non-coding genes. The database has the interpretation of a variant's effect on a certain phenotype. The possible effects include, for example, benign, pathogenic, risk factor and modifier. ClinVar also reports additional information about the submissions, such as information about the submitter and evidence for the interpretations. The submissions come from various sources, such as genetic testing laboratories, other public databases, and expert panels. For the users, ClinVar provides the reports as variant call format files (VCF) and XML files (Landrum et al., 2018).

ClinVar uses a four-star rating system to aid in assessing the reliability of the interpretations (Landrum et al., 2018). Variants without stars have not been assigned an interpretation by the submitter or they are lacking assertion criteria. To receive one star, a variant is either submitted from one source with an interpretation or it has multiple submitters with conflicting associations. Two-star rating requires multiple submitters, and their assessments for the variant must be the same. An organization can apply for the status of expert panel (three stars) or practice guideline (four stars) by contacting ClinGen (https://clinicalgenome.org/) for approval and filling in an application form. The most prevalent review status among unique variations is one star (79%),

while only 13.9%, 1% and 0.06% have two, three and four stars, respectively. The remaining variations have zero stars.

3.1.2 GNOMAD

The Genome Aggregation Database (gnomAD; https://gnomad.broadinstitute.org) has summary information from various large-scale sequencing projects publicly available (Karczewski et al., 2020). The newest version, v3.1, has 76,156 whole-genome samples and contains over 750 million short nuclear variants. The earlier version and the one used for this thesis, v2.1.1, has whole-exome sequence and whole-genome sequence information from 125,748 and 15,708 individuals, respectively. This version contains 17.2 million variants in the exome dataset and 262 million variants in the genome dataset (Karczewski et al., 2020). As majority of disease-causing variants are on the protein-coding genes, I utilized the 17.2 million variants from the exome dataset. Quality control has been conducted by gnomAD to ensure that the variants are as certain as possible. For example, samples with low sequencing quality have been removed. In addition to global allele frequencies and genotype distributions, gnomAD reports data at a subpopulation level, such as European, East Asian, and African. The majority of the samples are of European ancestry. In this project I was interested in the distribution of genotypes and the allele frequencies at the global and subpopulation level of the detected (likely) pathogenic X-chromosomal ClinVar variants.

3.1.3 UK BIOBANK

For this thesis, I utilized the UK Biobank (UKB), a database available to researchers via application, that contains genetic and phenotypic information from approximately 500,000 volunteered individuals from all over the United Kingdom (Bycroft et al., 2018). The UK Biobank, initiated in 2006, is one of the largest biomedical databases globally. The age of participants at the time of recruitment varies between 40 and 69. The phenotype data was aggregated from biological samples, questionnaires, and hospital records, and follow-ups on the health conditions of the participants are done regularly. The Biobank contains array-based genotypes that cover over 90

million imputed variants and 800,000 directly genotyped variants, as well as over 2,000 phenotypes. On the X chromosome there are roughly 3.9 million and 19,000 imputed and directly genotyped variants, respectively. For the genotyped variants, two different genotyping arrays were used, Applied Biosystems UK BiLEVE Axiom Array by Affymetrix and Applied Biosystems UK Biobank Axiom Array, which share 95% of the variants (Bycroft et al., 2018). Haplotype Reference Consortium, UK10K and 1000 Genomes phase 3 reference panels were used for the imputation (Bycroft et al., 2018). The genotype information in UKB is coded for GRCh37.

3.2 COLLECTING (LIKELY) PATHOGENIC X-CHROMOSOMAL VARIANTS

The first step of the genotype-phenotype association study is to examine how many likely pathogenic and pathogenic X-chromosomal ClinVar variants can be detected from UKB as heterozygous and hemizygous. To do this, I collected the variants and their associated phenotypes from the ClinVar database.

ClinVar provides the opportunity to download variant information in various file formats. For this thesis, I utilized the ftp pages of ClinVar (https://ftp.ncbi.nlm.nih.gov/pub/clinvar/) where it is possible to download records of the variants. I downloaded the Variant Call Format (VCF) files coded for Human Reference assembly GRCh37 (accessed 18.5.2021), and with BCFtools version 1.9 (Li, 2011) command view --regions X:2699521-154931043 I limited the variants only to the sex-specific region of X chromosome. PARs were excluded due to their different inheritance pattern. The generated VCF file was transformed into a text file and processed with R program version 4.0.0 (R Core Team, 2021) to include only the likely pathogenic and pathogenic variants. In addition, figures and tables were produced with ggplot2 (Wickham, 2016). After the combining of ClinVar and gnomAD variations (see section 3.3), the VCF file with ClinVar variants, which are not detected from gnomAD database, was transformed into a text file, and searched from UKB with the gnomAD variants.

The correlation between gene lengths collected from Ensembl (https://grch37.ensembl.org/) and the number of variants was examined with Pearson correlation coefficient using the *cor.test*

function from R (R Core Team, 2021). The significance of the correlation was evaluated by the p-value given from the function. Also, the confidence interval was retrieved from the results of the function.

3.3 COMBINING DATA FROM CLINVAR AND GNOMAD

I downloaded the exome sites for X chromosome, coded for GRCh37, as a VCF file from the gnomAD webpage (accessed 1.6.2021; https://gnomad.broadinstitute.org/downloads). The records were restricted in a same manner as the ClinVar data to select the X-chromosomal variants and to filter out the PARs of the X chromosome. The filtered VCF files from ClinVar and gnomAD were merged using BCFtools command isec which produces four files. For further examination, files with the shared variants between ClinVar and gnomAD were selected, and the records were limited to likely pathogenic and pathogenic variants. The files were converted into text files that can be further processed in R.

Variants were excluded from the analysis of global AFs and genotypes if they did not pass quality control set by gnomAD. Additionally, an arbitrary 5% minor allele frequency (MAF) threshold in any subpopulation was set to exclude overly tolerated variants from the genotype-phenotype analyses. Examination of allele frequencies, genotype distributions, and genes and phenotypes for the remaining variants was performed in R version 4.0.0 (R Core Team, 2021) using ggplot2 package (Wickham, 2016).

3.4 EXAMINING UK BIOBANK FOR THE VARIANTS

Summary statistics of imputed variants in UK Biobank was performed using QCTOOL v2 (https://www.well.ox.ac.uk/~gav/qctool_v2/). This program can be utilized for analyzing BGEN v1.2 files, that are used for the imputed variants in UKB, as QCTOOL fully supports this file format. With QCTOOL the difference in ploidy in males and females can be considered and it is therefore useful for X-chromosomal variants although other tools, such as PLINK, have the options to address the unique inheritance of X-linked genes as well. For each imputed variant an individual

has a probability for every genotype, that are homozygous for reference allele, heterozygous, or homozygous for alternative allele. The genotype counts are calculated as the sum of these probabilities within the samples meaning that the genotype counts are not completely accurate. The computation was limited to the (likely) pathogenic ClinVar variants that were indicated by their genomic location. The variants were confirmed to be shared between the datasets by ensuring their nucleotide change is the same in both sets. Variants with imputation info below 0.4, indicating unreliable imputation, were removed from further analyses.

The directly genotyped variants in UKB were provided in a binary ped file format, consisting of bed, fam and bim files. PLINK v1.9 (Purcell et al., 2007) was used to compute the genotype distributions and allele frequencies of the (likely) pathogenic ClinVar variants within the directly genotyped samples. The genomic position and nucleotide change of the variants were used to confirm the results. The mutational constraints of the genes containing the detected UKB variants were collected from gnomAD.

The summary statistics for imputed and directly genotyped variants were combined into a table in R v4.0.0 (R Core Team, 2021). Variants, which were detected from both sets, were included only as directly genotyped. Genes, phenotypes, and global allele frequencies from ClinVar and gnomAD were also included in the table. Figures were produced using R package ggplot2 (Wickham, 2016).

3.5 ASSOCIATION ANALYSIS

All association analyses in the UKB were conducted using PLINK v2.0 program with standardized multiple linear regression (Purcell et al., 2007). As a model, I used the additive model that is the most commonly used model in genetic association studies. The variants were filtered according to the number of heterozygotes and hemizygotes carrying the mutated allele, and the associated phenotypes reported in ClinVar. The number of heterozygotes and hemizygotes for the imputed variants was the number of individuals having over 90% probability of the genotype. The requirement was that the variants had an associated phenotype and a distinct quantitative trait related to it. Quantitative traits are used in order to gain enough power to detect possible differences in expression as there is more variability than with binary phenotypes. The normalization of phenotypes was carried out with PLINK as well using the option --variance-

standardize which transformed the traits linearly so that their mean is zero and standard deviation is 1. The filtered imputed and directly genotyped variants were combined to form a single bed file using PLINK. One to five quantitative traits were chosen for each variant using different databases, such as MedlinePlus (https://medlineplus.gov/) and The Genetic and Rare Diseases Information Center (https://rarediseases.info.nih.gov/). BMI, age at recruitment and PCs were used as covariates for each trait in the additive regression model. The analyses were carried out with both 10 and 20 PCs to assess the impact of population stratification.

The results were interpreted from effect size estimates, standard errors (SE), confidence intervals (CI), and p-values. Additionally, Bonferroni correction was conducted for the multiple testing to adjust the p-value thresholds for the number of individual tests. The calculation for the corrected p-value thresholds was done by dividing the nominal p-value threshold (0.05) by the number of individual analyses. The thresholds were different in male and female association analyses as there was more tests performed on females (see section 4.4.1). For the examination of the results also confidence intervals for the effect size estimates were calculated as

$$95\%\ CI = estimate \pm 1.96 * \mathrm{SE} \qquad \text{(Equation 3)}$$

3.5.1 Replication of ClinVar phenotypes in UK Biobank

The purpose of the genotype-phenotype replication was to assess the reliability of the reported associations in ClinVar. In this thesis I aimed to assess if the (likely) pathogenic X-chromosomal variants associate with quantitative traits of the reported phenotypes. The variants were filtered to include only those ones whose minor allele count in males in UKB was greater than five. The standardized multiple linear regression with additive model was fitted for one to four traits and the before-mentioned covariates. The results were examined and interpreted with R program (R Core Team, 2021), and figures produced with package ggplot2 (Wickham, 2016).

3.5.2 Phenotypes in heterozygous females in UK Biobank

To assess how often heterozygous females express the quantitative traits of the reported

phenotypes standardized multiple linear regression was used. In this test, heterozygous females were analyzed with females who are homozygous for the reference allele to assess if there are significant differences in the expression between the genotypes. Individuals with two alternative alleles were excluded from the analysis. Variants with at least ten heterozygous carriers were included in the analysis to have sufficient number of individuals with the alternative allele to reach enough power. For each variant, one to five traits were chosen and fitted to an additive multiple linear regression model with the covariates. The results were compared with the outcome of the replication analysis and processed with R version 4.0.0 (R Core Team, 2021).

4 RESULTS

4.1 (LIKELY) PATHOGENIC VARIANTS IN CLINVAR

The variant call format files of ClinVar included nearly 900,000 variants and out of them 36,889 were on the X chromosome. 36,628 X-chromosomal variants remained after exclusion of PARs which have different inheritance pattern than the sex-specific region of the X chromosome. Among the X-chromosomal variants there was a total of 12,190 likely pathogenic and pathogenic variants (33.2%) that are further examined from gnomAD and UKB (Figure 5). The clinical significance of over 30% of the X-linked variants was uncertain which was expected as it is assumed that most of the variants are extremely rare (MAF < 0.01%) and reported only from one source.

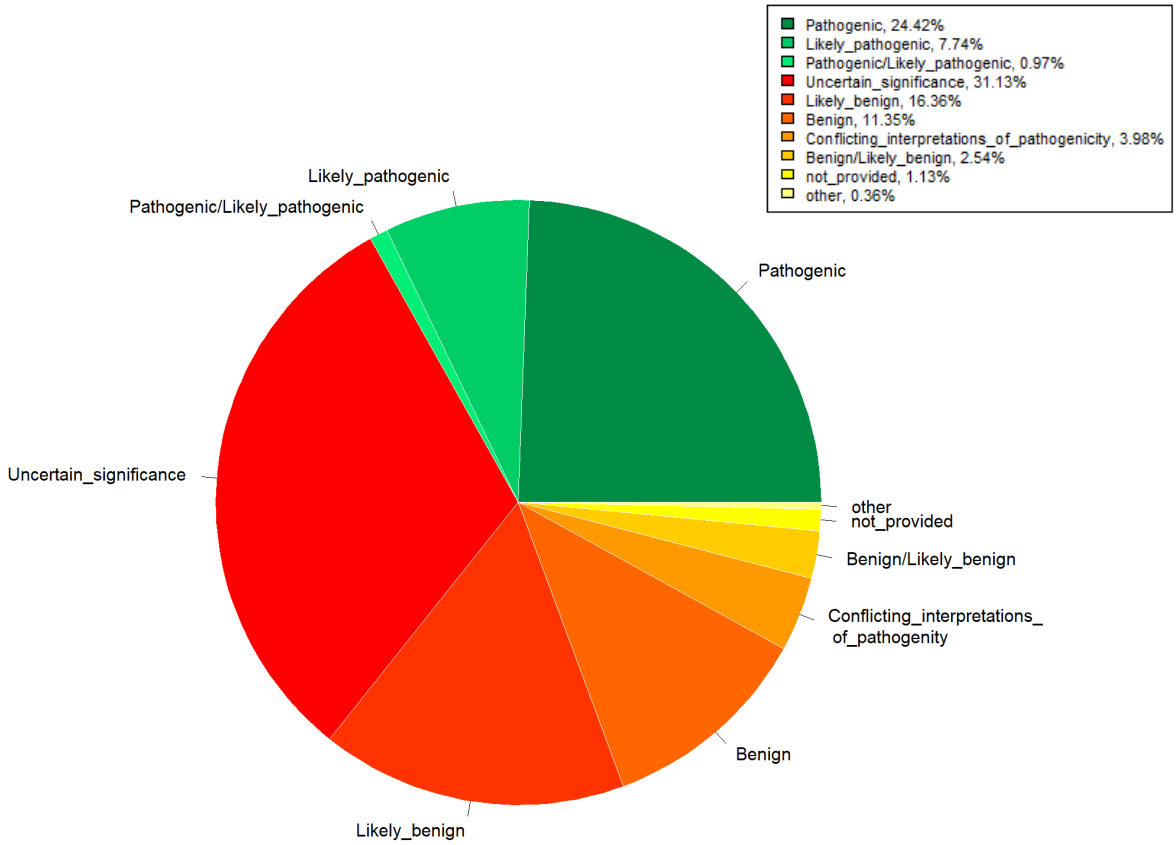

Figure 5. The clinical significance of the X-chromosomal variants in ClinVar. The variants in the green area were included in further examination. Also, few variants classified as 'other' were added to the analyses.

SNV is the most common type of variation in the human genome (The 1000 Genomes Project Consortium, 2015). Majority of the (likely) pathogenic ClinVar variants were, as expected, SNVs (66%) and approximately one fifth of the variants were deletions (Figure 6). Rest of the variants consisted of duplications, microsatellites, indels, insertions, and inversions. Missense variant was the most prevalent variation among the (likely) pathogenic X-chromosomal ClinVar variants. In addition, stop-gain and splice-region variations were observed among the ClinVar variants.

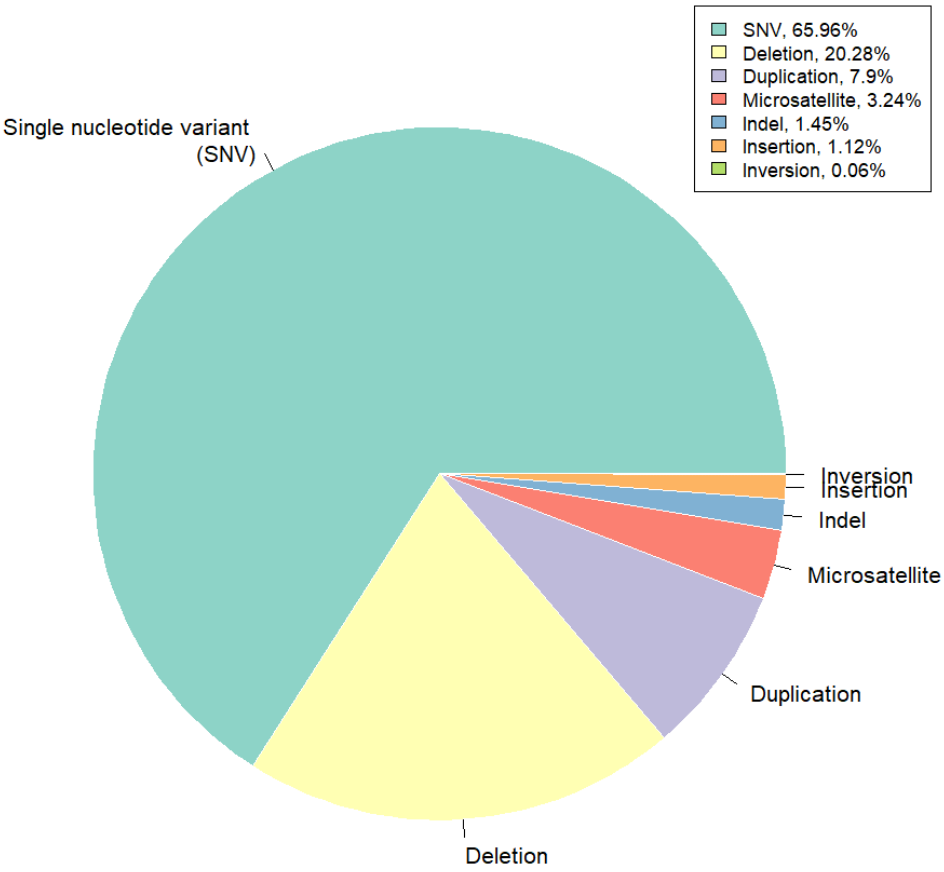

Figure 6. The different variant types among the (likely) pathogenic X-linked ClinVar variants.

The (likely) pathogenic ClinVar variants mapped together on 249 genes. The number of variants on each gene varied between one and 1,099 variants. The genes with the largest number of variants were *DMD* with 1,099 variants and *COL4A5* with 1,057 variants, associated with Duchenne muscular dystrophy and Alport syndrome, respectively. The large number of variants on *DMD* can be explained by its size as *DMD* is the longest in terms of exonic base pairs gene in the human genome. 34 genes, for instance *GLUD2*, *SOX3* and *USP51*, had only one reported (likely) pathogenic variant. 20 genes with the most variants as well as their associated phenotypes can be seen from Table 1. Among these 20 genes, variation in length can be seen; however, the Pearson correlation coefficient between the gene lengths and the number of variants was 0.43 (95% CI =

0.31 – 0.53) with p-value of $9.2x10^{-13}$ which suggests that there is a positive correlation between the variables. The correlation was computed with all the 249 genes.

Table 1. The 20 genes with the highest number of variants in ClinVar with their lengths and associated phenotypes from ClinVar.

Gene	# of variants	Gene length	Associated disorder(s)
DMD	1,099	2,241,765	Duchenne muscular dystrophy
COL4A5	1,057	257,702	Alport syndrome
MECP2	499	76,189	Rett syndrome, encephalopathy, intellectual developmental disorder
PHEX	466	218,869	Hypophosphatemic rickets
OTC	376	68,906	Ornithine transcarbamylase deficiency
F8	344	191,153	Hemophilia A
GLA	337	10,123	Fabry disease
CDKL5	336	228,047	Developmental and epileptic encephalopathy 2
RPGR	296	58,402	Retinitis pigmentosa, macular degeneration, cone-rod dystrophy
PCDH19	235	118,630	Developmental and epileptic encephalopathy 9
ABCD1	207	19,894	Adrenoleukodystrophy
MTM1	186	104,727	Myotubular myopathy
IDS	185	56,950	Mucopolysaccharidosis II
AR	179	185,997	Androgen insensitivity, hypospadias 1, spinal and bulbar muscular atrophy of Kennedy
FLNA	176	26,115	Periventricular Heterotopia 1, Melnick-Needles syndrome, cardiac valvular dysplasia
BTK	167	36,749	Agammaglobulinemia
F9	166	32,701	Hemophilia B, thrombophilia
DDX3X	164	31,075	Intellectual developmental disorder
OFD1	159	34,649	Joubert syndrome, orofaciodigital syndrome I, Simpson-Golabi-Behmel syndrome, type 2
GJB1	146	10,323	Charcot-Marie-Tooth neuropathy

Fourth of the variants (3015, 24.7%) had not been assigned a phenotype or a disease. The most prevalent phenotype reported was Alport syndrome that is associated with gene *COL4A5*. Other commonly occurring phenotypes included Duchenne muscular dystrophy, Rett syndrome, hereditary factor VIII deficiency disease (hemophilia A) and Fabry disease. Although *DMD* had the largest number of variants, Duchenne muscular dystrophy was not assigned as the associated

phenotype for all these variants. A proportion of *DMD*-variants were associated with Becker muscular dystrophy, and some had no associated phenotype.

The ClinVar review status varies from zero to two stars among these (likely) pathogenic X-chromosomal variants. 7153 variants had one-star rating while 3627 and 1410 were rated with zero and two stars, respectively. Therefore, the majority of the variants were submitted from one source with assertion criteria while 11.6% of the variants had multiple submitters without conflicts. Nearly 30% of the variations were not provided with assertion or the required criteria to receive any stars.

4.2 GLOBAL ALLELE FREQUENCIES AND GENOTYPES

In total, there were 384,632 non-PAR X-chromosomal variants in the exome data of gnomAD v2.1.1 and only 288 of these variants were among the 12,190 (likely) pathogenic ClinVar variants. In addition, 33 variants were excluded from the examination of global AFs and genotype distributions as they failed quality control set by gnomAD. The remaining 255 variants were extremely rare (MAF < 0.01%) in the global population, except few of them being low-frequency (0.5% ≤ MAF < 5%) variants and two common variants (MAF ≥ 5%) (Figure 7).

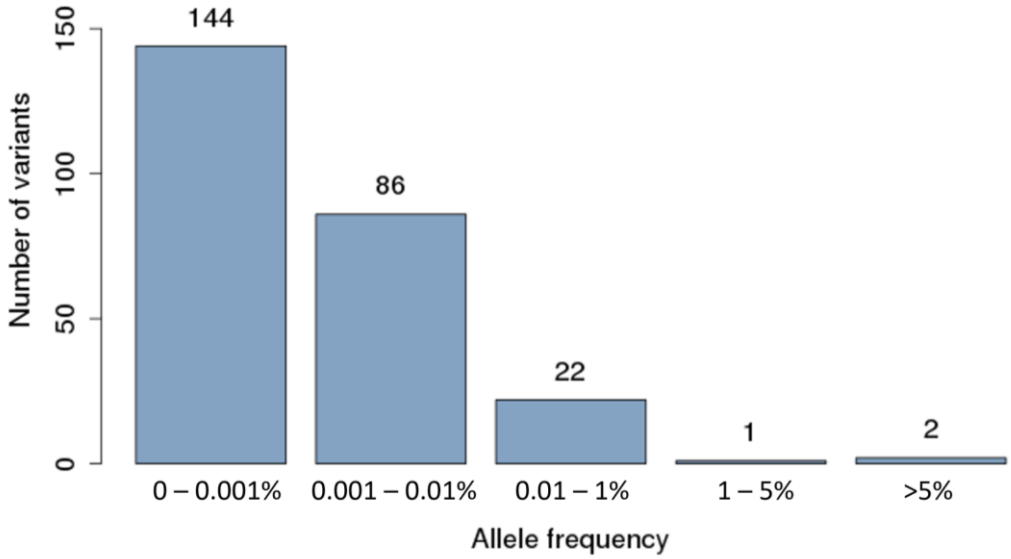

Figure 7. The distribution of global minor allele frequencies of the variants detected from gnomAD.

At the subpopulation level, three variants had alternative AF over 5% and due to this prevalence, they were filtered out of the genotype-phenotype analyses as they seemed too tolerated to be truly pathogenic. One of the excluded variants was on gene *C1GALT1C1* (rs17261572) and two on gene *SERPINA7* (rs1804495 and rs1050086), and they were associated with polyagglutinable erythrocyte syndrome and variant P of thyroxine binding globulin, respectively. In addition, one variant on gene *GLUD2* (rs9697983) had global allele frequency over 2%; however, since its AF in any subpopulation did not exceed the 5% threshold, the variant remained in the next steps. The associated phenotype in ClinVar for this variant was late-onset Parkinson disease.

The expected genotype counts are derived from allele frequencies. When the MAF is p and sample size is N, the number of hemizygotes and homozygotes for the minor allele are expected to be $p*N$ and p^2*N. The genotype distributions for the 255 variants showed that these variants do not appear very often as hemizygous or homozygous as expected for rare disease-causing mutations. 138 (~54%) and 244 (~96%) of the 255 variants had no occurrences of hemizygous and homozygous individuals in the global population, respectively, which is expected as the variants are extremely rare. As with the allele frequencies, the variants with the highest number of hemizygous and homozygous individuals were the ones with the highest global MAF.

After excluding those three variants with highest population-specific allele frequencies from the 12,190 ClinVar variants I was left with 12,187 (likely) pathogenic X-chromosomal variants to be investigated from the UK Biobank population.

4.3 VARIANTS IN UK BIOBANK POPULATION

89 of the 12,187 (likely) pathogenic variants from ClinVar were detected from UKB, and out of these 89 variants, 88 were SNVs and one was a duplication. 61 were from the directly genotyped dataset and 23 from the imputed dataset. Additionally, 5 variants were reported in both datasets and included in genotype-phenotype association examination as directly genotyped variants. 78 were asserted as pathogenic and 11 as likely pathogenic in ClinVar. Imputation info varied from ~0.11 to ~0.96 within the 21 detected imputed variants. I excluded from further investigations one

likely pathogenic and seven pathogenic variants whose imputation info was below 0.4, leaving me with 81 SNVs.

The detected variants were mainly extremely rare (MAF < 0.01%); however, 11 variants were more prevalent, and two variants exceeded even 1% MAF. The two most common variants, rs104894858 and rs9697983, are on genes *LAMP2* and *GLUD2*, respectively. The *LAMP2* variant had MAF of 3.74% and was associated with hypertrophic cardiomyopathy and Danon disease. The rs9697983 variant was the same as the one with over 2% global AF in gnomAD and its MAF in UKB population was 2.44%. The associated phenotype for the *GLUD2* variant was late-onset Parkinson disease. 23 of the detected variants were carried by at least one hemizygote individual whereas only 6 appeared as homozygous. It was expected that most of the variants would not have carriers as they were assumed to be disease-causing which was also shown from the number of hemizygotes and homozygotes.

Moderate correlation between imputation info and MAF could be seen although some variants with the lowest MAF still reached the arbitrary threshold of 0.4 while others with higher MAF were left below it (Figure 8A). Each of the five variants reported in both datasets, imputed and directly genotyped, had over 0.5 imputation info. The comparison of the MAFs from different datasets of these five variants demonstrated that all except one had very similar MAF whether it was imputed or directly genotyped (Figure 8B). Variant rs137852388 on gene *F8* had a higher MAF reported in directly genotyped dataset than in imputed summary statistics. This variant had the lowest imputation info (0.53) of the five variants which can explain the difference in MAFs.

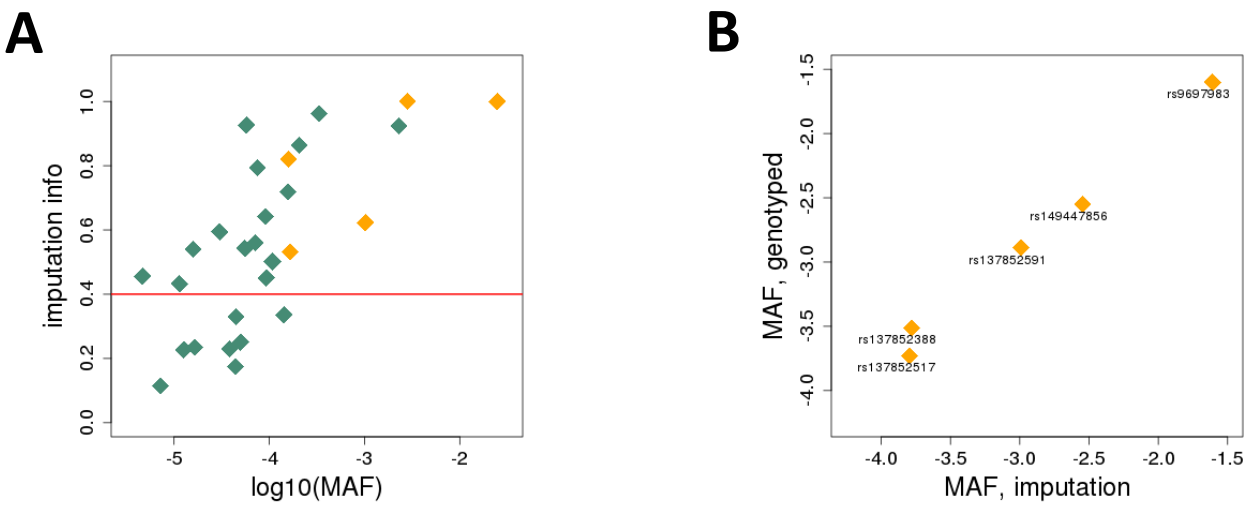

Figure 8. A) The correlation between imputation info and MAF. Variants marked in orange were included in both imputed and genotyped datasets and their MAFs for this figure was retrieved from the summary statistics of the imputation dataset. B) The comparison of MAFs for the variants detected from both datasets. The allele frequencies are given in log(10) scale in both figures.

24 of the detected variants were also reported in gnomAD after the exclusion of variants with too low imputation info. The minor allele frequency and genotype distributions in UKB were quite similar to the ones from gnomAD data (Figure 9). A Pearson correlation test indicated that the MAFs are very strongly correlated between gnomAD and UKB with correlation coefficient of 0.997 (95% CI 0.993 – 0.999, p-value = 7.25×10^{-26}). The variants not found in gnomAD were generally extremely rare within UKB population with one significant exception. The most prevalent variant in UKB, the before-mentioned *LAMP2* variant, was not present in gnomAD data which is surprising. Additionally, this variant, rs104894858, is one the detected variants with two-star rating in ClinVar increasing the expectation of it being also included in gnomAD.

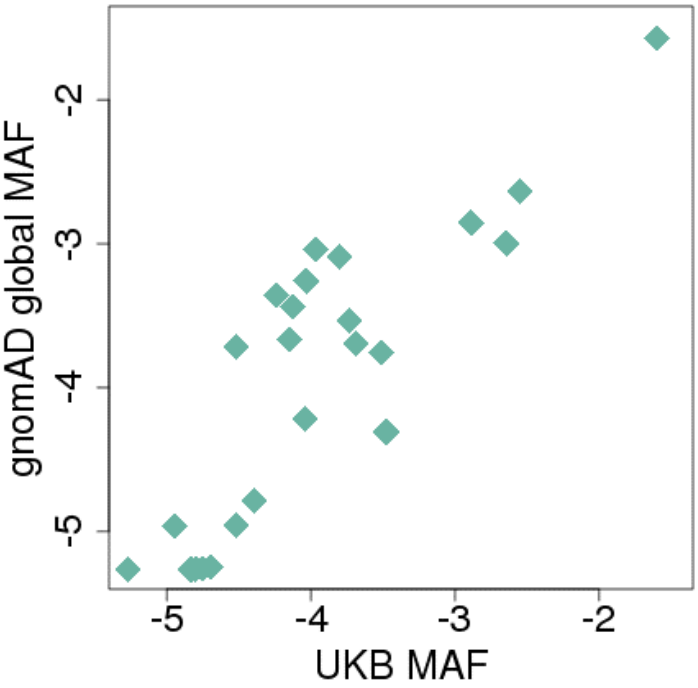

Figure 9. The comparison of MAFs of the variants detected from both gnomAD and UKB. The frequencies are given in a log(10) scale.

The variants were in the region of 36 different genes out of which 20 had only one reported variant. There were over 10 variants on genes *DKC1* (22 variants) and *AVPR2* (12 variants) which were associated with Dyskeratosis congenita and Nephrogenic diabetes insipidus, respectively. *DKC1* had 45 reported variants in ClinVar, meaning that nearly half of them were detected from UKB. Also, the number of *AVPR2* variants in ClinVar, 52, is relatively similar to the 12 in UKB. Other phenotypes associated with the detected variants included Danon disease, reducing body myopathy, Alport syndrome and type 2 3-methylglutaconic aciduria. Four variants, rs199959402 (gene *PDHA1*), rs149447856 (gene *SLC7A3*), rs782344765 (gene *RP2*) and rs766506778 (gene *CUL4B*), did not have assertion in ClinVar. The clinical significance of the latter three variants was likely pathogenic while the first one was reported as pathogenic. The Pearson correlation test showed no correlation between the gene length and the number of variants unlike among all of the X-chromosomal variants (see section 4.1).

The genes having the most variants in UKB differed from the ones in ClinVar. For example, *DMD* which had 1,099 variants reported in ClinVar had only one detected from UK Biobank. On the other hand, the gene with the most variants in UKB, *DKC1*, had only 45 variants reported in ClinVar

meaning that almost half of them were detected from UKB. Out of the 20 genes with the most variants in ClinVar, 9 had detected variants in UKB. These findings are rather interesting and will be further explored in the discussion (see section 5.2.2).

50 of the UKB variants were reviewed in ClinVar with zero stars, 20 with one star and 11 with two stars which means that most variants were unassigned with assertion criteria. This is rather unsurprising as majority of the variants are extremely rare and hence have only been detected from few individuals previously.

4.4 MUTATIONAL CONSTRAINT

The reason to inspect the mutational constraint of the genes with detected variants in UKB was to see if the variants lie on genes that are highly intolerant against pathogenic mutations which aids the interpretation whether the detected variants are truly pathogenic. The mutational constraints for the genes having detected variants in UKB were retrieved in the form of missense Z-scores and probability of loss of function intolerance (pLI) scores that indicate the intolerance against missense and loss-of-function variation, respectively. These metrics are based on the ratio of expected and observed variants in gnomAD. In both metrics, greater score indicates larger intolerance against the variation in question. However, the scales of Z-score and pLI are different as pLI is given as a probability between 0 and 1 whereas negative Z-score indicates fewer expected than observed variants and positive Z-score indicates that a gene has more expected than observed variants (Lek et al., 2016). The mutational constraints of the 36 genes with detected variants in UKB can be seen from Table 2.

Table 2. The mutational constraints of the genes with detected variants in UKB.

Gene	# of variants in UKB	Z-score (missense)	pLI
AIFM1	1	2.49	1
ANOS1	1	0.89	1
AR	3	1.23	0.99
ARSL	1	1.95	0.59
AVPR2	12	1.01	0.56
BCOR	1	1.88	1
CDKL5	1	2.74	1
CLIC2	1	1.27	0.05
COL4A5	3	2.5	1
CUL4B	1	3.77	1
DKC1	22	3.4	1
DMD	1	-2.43	1
F8	1	2.47	1
F9	1	2.18	1
FHL1	6	0.89	0.97
FLNA	1	3.78	1
G6PD	4	2	0.97
GJB1	1	2.03	0.85
GLUD2	1	0.83	0.03
GPC3	1	1.45	1
KDM5C	2	5.15	1
KDM6A	1	2.95	1
LAMP2	4	0.84	0.27
OTC	1	1.33	0.87
PDHA1	1	2.57	0.99
PLXNA3	1	3.45	1
RP2	1	0.63	0.96
RS1	1	0.97	0.96
SERPINA7	2	-0.66	0
SLC7A3	1	1.9	1
TAFAZZIN	3	2.21	0.73
TBX22	1	-0.13	0.98
TEX11	2	1.81	1
USP11	1	3.57	1
WAS	1	1.98	1
XK	2	2.12	0.95

The majority of the genes have over 0.9 pLI suggesting that they are highly intolerant against loss-of-function variants with the exception of genes *CLIC2*, *GLUD2*, and *LAMP2* that have rather low pLI scores indicating they have higher tolerance against loss-of-function variants. Additionally, the Z-scores are mainly positive meaning that there are less observed missense variants than expected to be. Few genes show, however, negative Z-scores indicating higher tolerance, and one of these genes is *DMD*. The majority of the (likely) pathogenic ClinVar *DMD* variants are nonsense and frameshift variants and only 17 of the 1,099 variants are missense variants. Hence it is surprising that the gene has a negative Z-score regarding missense variation. Additionally, as mutations on *DMD* may cause Duchenne muscular dystrophy, it is expected that the gene is intolerant against all pathogenic mutations. Interestingly, *DKC1* has the highest number of detected variants even though the mutational constraint metrics show that this gene is considerably intolerant against missense (Z-score = 3.4) and loss-of-function (pLI = 1) variants.

4.5 ASSOCIATION ANALYSES

4.5.1 Final variants and phenotypes for the analyses

Filtering of the variants resulted in 11 and 27 variants (Supplementary table 1) from the 81 UKB variants for the replication analysis in males and the genotype-phenotype association analysis in females, respectively. The four aforementioned variants (rs199959402, rs782344765, rs149447856 and rs766506778) unprovided with associated phenotypes were removed as well as two variants on gene *TEX11* due to male-specific phenotype which was spermatogenic failure. Additionally, two variants on gene *SERPINA7* were excluded given their phenotype associations with thyroxine-binding globulin deficiency generally do not cause any major health problems (Mimoto & Refetoff, 2020). From the replication analysis in males 62 variants were excluded as their alternative allele count was below five after applying the 0.9 threshold for the imputed variants (see section 3.5). The threshold of 10 heterozygous carriers for the genotype-phenotype association analysis removed 46 variants.

The tested quantitative traits were chosen according to the associated disorders of the variants. Each variant was tested against one to five traits. For instance, traits for variants on gene *AR* (androgen receptor) were testosterone levels, height, and muscle-mass whereas blood pressure

and pulse rate were used for variants on gene *LAMP2*. Other phenotypes included, for example, hand-grip strength, urine composition (creatinine, microalbumin, potassium and sodium levels in urine), blood-cell counts, and hemoglobin concentration. Variants and tested quantitative traits can be seen from Supplementary tables 2 and 3. In total, 28 and 87 multiple linear regressions were conducted in the replication and genotype-phenotype analyses, respectively.

The Bonferroni multiple testing correction was conducted by dividing the nominal p-value threshold, 0.05, by the number of independent analyses. This means that the corrected p-value threshold is 0.05/28 (~0.0018) in analyses for males and 0.05/87 (~5.7×10^{-4}) in analyses for females.

4.5.2 Replication in males

The reason behind conducting the association analysis with only the male samples from UKB was to estimate the effects of the assumed disease-causing variants and to assess the chosen quantitative traits. As men are hemizygous for X-linked variants, they should express a disease-causing variant if they carry one on their X chromosome. The expectation is that the males carrying the alternative allele would differ from the mean of the quantitative traits compared to non-carrier males. The absence of this difference would suggest that either the variant of interest is not truly pathogenic, or the quantitative trait does not represent the phenotype associated with the variant.

The replication analysis was conducted to assess the associations reported in ClinVar, although the reported phenotypes were not directly used. The analysis utilized quantitative traits that could be symptoms of the reported disorders. 11 (likely) pathogenic variants were included in the genotype-phenotype replication analysis as they met the requirements explained before. The number of PCs as covariates had only minimal effects on the results, and hence from now on I will refer to the results of the tests with 20 PCs. All results can be seen from Supplementary table 2.

Majority of the associations did not reach the statistically significant p-value even before the Bonferroni multiple correction (Supplementary table 2). Eight of the 28 tested associations were nominally significant, and four remained significant after multiple correction (Figure 10). The

results indicate that there are more associations than would occur randomly as over 28% of the associations showed statistical significance. Three of the significant Bonferroni corrected associations were with variant rs137852591 on gene *AR*. 285 males carried the alternative allele of the variant in UKB population hence the power was sufficient for the detection. The association in ClinVar for this variant was partial androgen insensitivity syndrome (PAIS) with one-star rating. The results indicate that the variant is associated with less muscle-mass and lower height as well as higher testosterone levels in serum (Figure 10) which can be associated with the reported disorder. The directions of the effects were as expected. The p-values for these associations were significantly lower than the corrected p-value threshold indicating the variant being a strong candidate to be disease-associated (Supplementary table 2). In addition to the *AR* variant, a single variant on gene *G6PD* (rs72554665) had lowering effect on glycated hemoglobin (HbA1c). However, this variant was carried only by 11 individuals lowering the power of the analysis.

The four variants that were not significant after the correction were on genes *AR* (rs137852593), *COL4A5* (rs104886312), *GLUD2* (rs9697983), and *G6PD* (rs72554664), and their associated phenotypes were higher serum testosterone level, lower systolic blood pressure, higher mean time to correctly identify matches and higher serum glucose level, respectively (Figure 10). These results seem reasonable regarding the direction of the effects. The variants without associations included, for example, the common *LAMP2* variant (rs104894858) and it was tested against blood pressure and pulse rate. As the review status for this variant was 2 stars it would have been expected to have some effects although *LAMP2* is rather tolerant against mutations according to its mutational constraint metrics.

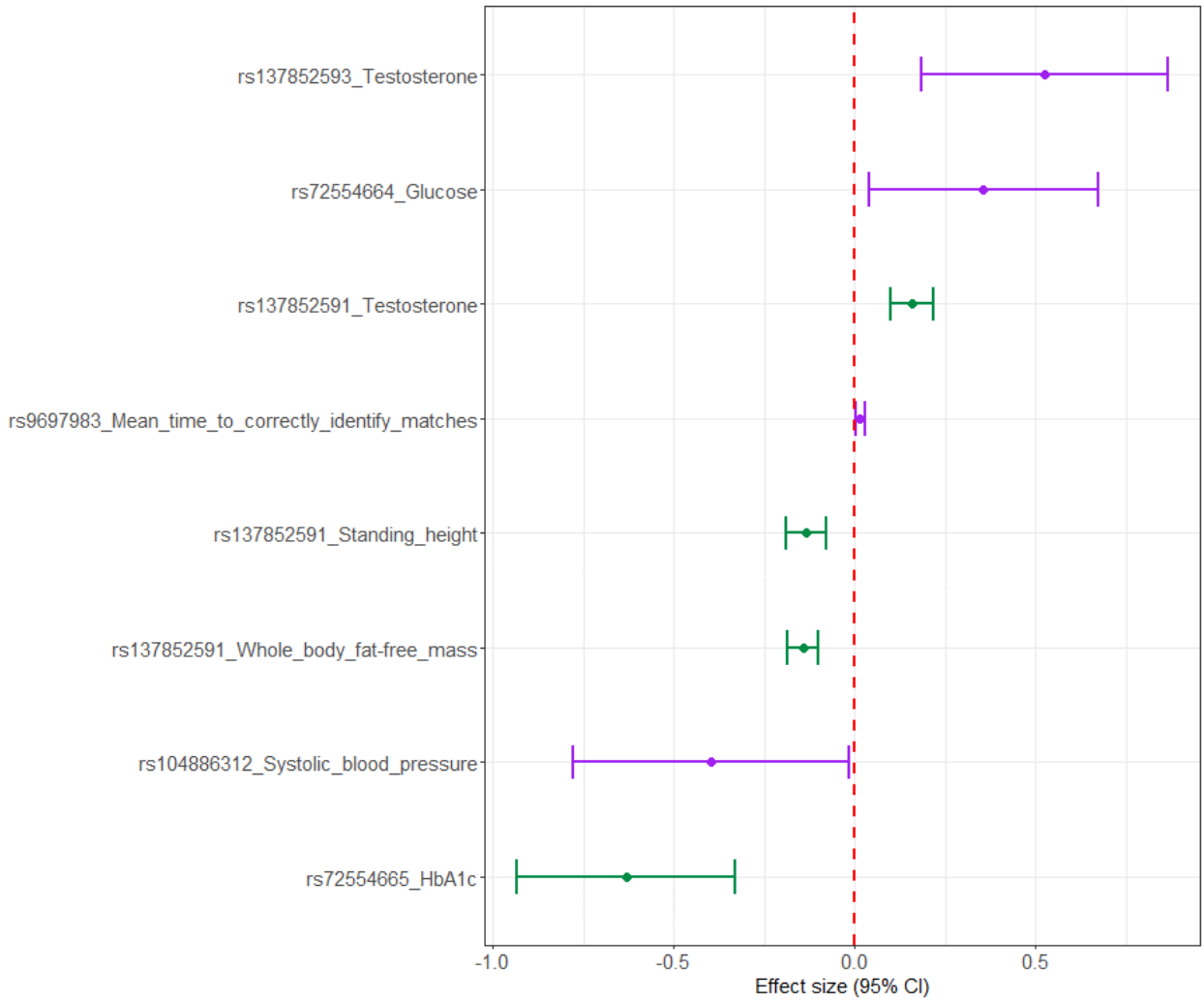

Figure 10. The tested associations with at least nominally significant p-values (<0.05) in males. Purple color indicates nominally significant associations and green associations remained significant after the Bonferroni multiple correction (p-value<0.0018). The effect sizes and their 95% confidence intervals are illustrated, and the red dashed line indicates the standardized mean 0.

4.5.3 Associations in heterozygous females

To assess how often heterozygous carrier females express the phenotypes of X-linked disorders I fitted multiple linear regression on quantitative traits associated with the disorders. 27 out of the 81 detected variants from UKB remained in the analysis after filtering. The regression analyses were done twice, using 10 or 20 PCs as covariates along with BMI and age at recruitment. With 10 PCs 12 (13.8%) and 1 (1.1%) associations of the 87 were statistically significant nominally and after Bonferroni multiple correction, respectively. Increasing the number of PCs to 20 resulted in one more nominally significant association (14.9%); however, the number of significant associations

after the multiple correction stayed the same (Supplementary table 3). As with the results in males, I will focus on the results with 20 PCs. Also in the female-analysis, more statistically significant associations can be seen than would be expected by chance.

The clearest association could be seen with variant rs5030869 on gene *G6PD*. The reported phenotype on ClinVar for this variant was glucose 6 phosphate dehydrogenase deficiency that causes the breakage of red blood cells. The quantitative traits chosen for this variant included red blood cell count, glycated hemoglobin (HbA1c) and hemoglobin concentration. The association with lower red blood cell count remained significant after multiple correction while the other two were only nominally significant (Figure 11). This variant could not be included in the replication analysis as the number of male carriers was only three. Additionally, two other variants (rs72554665 and rs72554664) on the same gene were included in the analysis; however, only variant rs72664665 had nominally significant association with HbA1c while the other variant showed no associations. The association between rs72664665 and HbA1c was also seen in males (see section 4.4.2).

Other nominally significant associations were observed with variants on genes *AR* (rs137852591), *AVPR2* (rs104894753 and rs104894748), *GLUD2* (rs9697983), and *LAMP2* (rs104894858 and rs104894857) (Figure 11). The variants rs137852591 and rs9697983 showed associations in males as well. The *AR* variant was the strongest candidate for having associations in males as the p-values were the smallest. The tested quantitative traits for these variations and the summary of the results can be seen from Supplementary table 3. The directions of the effect sizes seemed reasonable regarding the variants with significant associations.

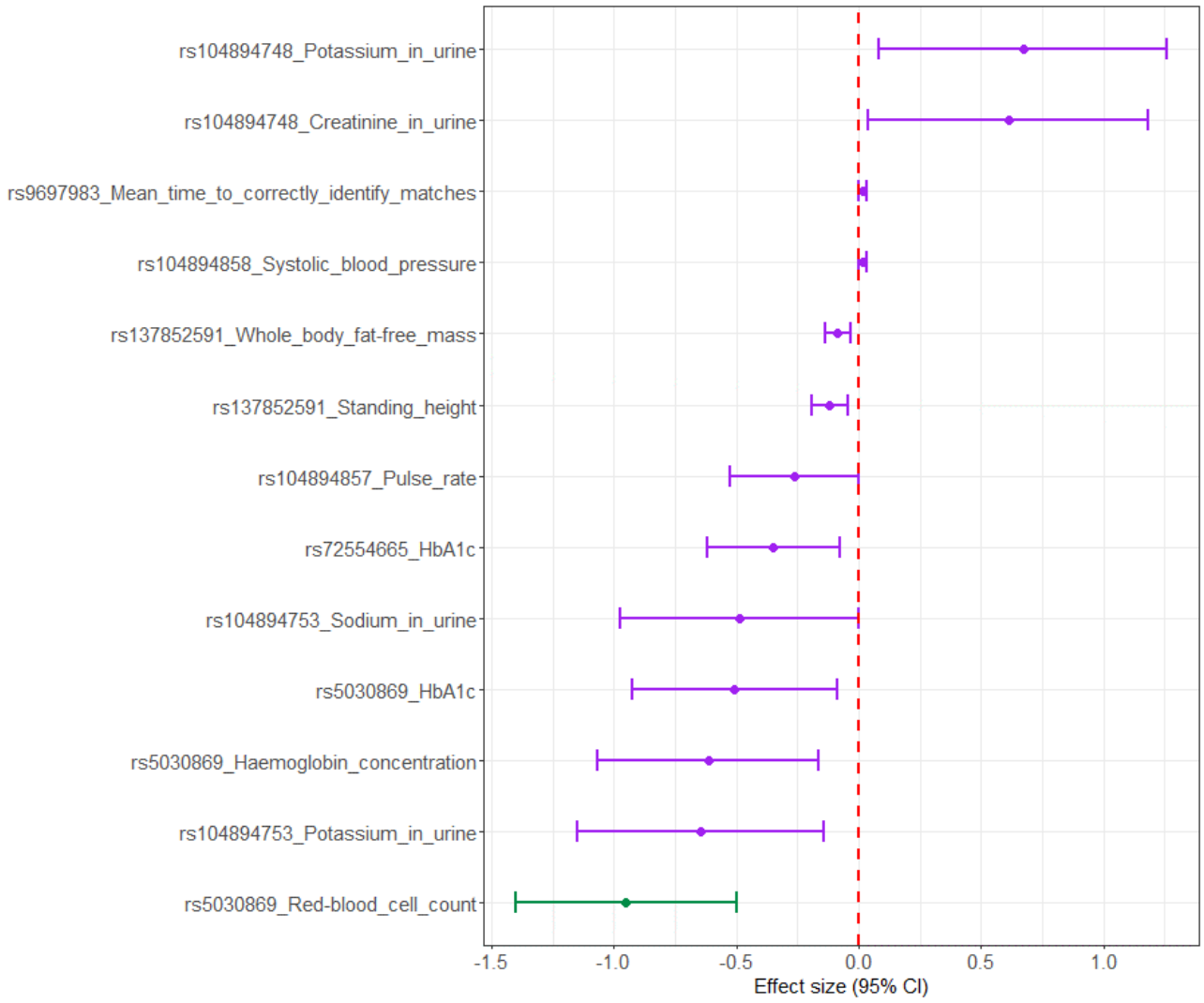

Figure 11. The tested associations with at least nominally significant p-values (<0.05) in females. Purple color indicates nominally significant associations and green associations remained significant after the Bonferroni multiple correction (p-value<$5.6x10^{-4}$). The effect sizes and their 95% confidence intervals are illustrated, and the red dashed line indicates the standardized mean 0.

4.5.4 Comparison of the results in males and females

To compare the results of the analyses regarding the associations I examined the estimated effect sizes in males and females. The effect size estimate indicates the impact of one allele of the studied variant to the tested trait. In a standardized study, as the association analyses in this thesis, 0.10 effect estimate (beta) suggests that increase of one copy of the allele changes the phenotype value 0.1 standard deviations from the mean. As females carry two X chromosomes

and males have only one, it is expected under full XCI that the effect sizes of X-linked variants are approximately twice as big in males than in females.

Few of the studied variants showed at least nominally significant associations both in males and in females. These variants included rs137852591 (*AR*), rs9697983 (*GLUD2*) and rs72554665 (*G6PD*). The *AR* variant showed association with whole body fat-free mass and standing height in both males and females. The effect sizes for the muscle mass were -0.1422 in males and -0.0861 in females indicating that the allele impact in females is a little over half of the effect in males. Also, the effect size estimates show that individuals with the mutated allele of this variant tend to have less muscle mass. Similarly, the individuals carrying this allele seem shorter than average as the effect estimates between the variant and standing height were -0.1348 and -0.1180 in males and females, respectively. These effect sizes, however, are closer to each other and the expectation of females having significantly lower effect size is not observed.

The variant rs9697983 on *GLUD2* was tested in the regression model with mean time to correctly identify matches as memory difficulties are related to Parkinson disease. The analyses showed mild increase in the trait in both males (effect size = 0.0153) and females (effect size = 0.0171). Also with this association, the effect sizes are similar between the sexes which can be explained by the chosen trait as many factors impact it. Lastly, the *G6PD* variant was associated with lower levels of glycated hemoglobin (HbA1c) with effect sizes -0.6308 in males and -0.3487 in females. This association showed difference in the estimates as the effect size in males is nearly double of the effect size in females.

5 DISCUSSION

The goal of this thesis was to examine the phenotypic effects of assumed disease-causing X-chromosomal variants in heterozygous females. In addition, genotype-phenotype associations of some of the variants were analyzed in males to see if the assigned associated phenotype for each variant could be replicated in males carrying the mutated allele. To reach these goals, I utilized three different databases, ClinVar, gnomAD and UK Biobank, to retrieve possibly pathogenic X-linked variants, examine their global allele frequencies and perform the association analyses, respectively. The association analyses were conducted using additive model in multiple linear

regression with age, BMI, and principal components as covariates. For each variant, one or multiple suitable quantitative traits were selected based on the disease-association reported in ClinVar using different sources to search for symptoms of the disorders.

The interest for this thesis arose from the unique inheritance pattern of the X chromosome and X-chromosomal phenotypes as males carry only one X in contrast to females who carry two. In the case of proposed recessive inheritance model, heterozygous carrier females should be unaffected by the mutated variants. However, many cases of assumed recessive X-linked disorders, such as hemophilia A and adrenoleukodystrophy, have been reported to manifest also in heterozygotes, although with milder symptoms, suggesting that the classification into recessive disorders should be revisited (Bryant et al., 2020; Basta & Pandya, 2021). An important factor in this unique pattern is the X chromosome inactivation in females where one of the female X chromosomes is silenced either randomly or preferentially, and it is possible that some women affected by pathogenic X-linked variants have skewed XCI.

5.1 Limitations of the used databases

The used databases, ClinVar, gnomAD and UK Biobank, have their own limitations that need to be considered when interpreting the results of this thesis. Although ClinVar is a comprehensive resource of clinically significant human variations, it has been shown that many variants included in ClinVar have been misclassified as being more pathogenic than they actually are (Shah et al., 2018; Xiang et al., 2020). The star-ratings of the variations also support this, as most variants have only one star, which means that their reported associations should be carefully considered. The low ratings are, however, unsurprising as most of the reported variants are extremely rare, possibly discovered only from one individual.

Limitations regarding gnomAD include the ancestry of the sample and the size of the sample. As the majority of the samples are from Europe the ancestral diversity in gnomAD remain rather small which restricts the studies of population-specific and rare variants as well as research of the population differences in mutational constraint (Karczewski et al., 2020). Larger sample size would increase the possibility to find the rarest variants in the human genome. Additionally, the exome

data utilized in this thesis ignores the variants outside of the protein-coding regions; however, nowadays gnomAD provides genome sequencing data containing variations from non-coding regions as well (Karczewski et al., 2020).

The UKB has its limitations as well. One significant notion is that the population of the Biobank is volunteered, mainly healthy, adults exceeding the age of 40. Assuming that the variants included in the association analyses in this thesis are pathogenic, at least the male individuals carrying them should have symptoms that might be even affecting the quality of life. Some of the associated disorders, such as Duchenne muscular dystrophy, are so severe that the life expectancy of the affected is below the 40 years and hence the patients are not included in UKB. Considering that the participants are generally healthy, it is possible that many X-chromosomal disorders are not present among the population of UKB. Also, the genetic data of UKB has been collected with genotyping arrays and computational methods and not with whole-genome sequencing meaning that many of the ClinVar variants may not even be on these arrays.

5.2 Detected variants

5.2.1 Overview of the variants in ClinVar

The examined variants were retrieved from ClinVar database that includes clinically significant human variations with evidence supporting the reported associations. The database consists of over million unique variations and out of these over 36,000 are on the X chromosome. The number of (likely) pathogenic X-linked variants was 12,190 which is over 30% of the X-chromosomal ClinVar variants. The accumulation of disease-associated variants on the X chromosome could be explained, for example, by the number of X chromosomes in the human genome. While autosomes are in homologous pairs, males do not carry homologous sex chromosomes as they are XY individuals. Pathogenic X-linked variants are therefore seen more often in males than in females, and X-chromosomal disorders seem more prevalent than autosomal disorders because of their prevalence in males. For the same reason, X-linked variants are classified as pathogenic more often than autosomal variants explaining the results.

The review status of the ClinVar variants informed that a large portion (~30%) of the (likely) pathogenic variants has zero stars and may hence have been unreliably assessed as pathogenic. The reliability of the variants with one star might also be quite low. As the majority of the retrieved (likely) pathogenic variants had low review status and only ~11.5% had a two-star rating, the clinical significance of the variants should be carefully interpreted. As mentioned, studies have shown misclassification among the variants reported in ClinVar hence it could be expected that some of the variants included in this thesis have also been misclassified as (likely) pathogenic (Shah et al., 2018; Xiang et al., 2020).

5.2.2 Gene comparison between ClinVar and UKB

The distribution of the (likely) pathogenic variants between genes showed differences between ClinVar and UKB. Genes with the highest number of variants in ClinVar had only few or even zero variants in UKB. In contrast, a number of genes with multiple detected variants in UKB were the ones with lower number of variants reported in ClinVar. For example, gene *DMD* had over a thousand variants reported in ClinVar; however, from UKB only one was detected. A few factors could explain this difference. *DMD* is the longest gene in the human genome spanning over 2 million base pairs and therefore the large number of variants in ClinVar was expected. Additionally, if these variants are truly pathogenic and contributors to Duchenne muscular dystrophy, a recessive X-linked disorder affecting males, it is unsurprising that the variations are not detected from UKB. Since the population of UKB is mainly healthy adults over the age of 40 years, the prevalence of Duchenne muscular dystrophy should be close to none as the life expectancy of individuals affected by the disease is approximately 20 years (Kliegman et al., 2020). The mutational constraint of *DMD* also supports the low number of detected variants in a normal population as the gene is extremely intolerant against pathogenic loss-of-function mutations. Additionally, it is assumed that the ClinVar variants are extremely rare which could partly explain why they are not detected from UKB. Similar explanations could also be applied to other genes that had many reported variants in ClinVar but only few in UKB.

On the contrary to the *DMD*, genes *AVPR2* and *DKC1* had many (likely) pathogenic variants in UKB compared to the number of variants detected in ClinVar. The low number of variants in ClinVar for

these genes can be explained by the fact that the genes themselves are also shorter as *DKC1* is 14934 bp long and the length of *AVPR2* is 4636 bp. It is unclear why there is such a large portion of the *DKC1* variants detected from UKB population; however, one noticeable fact is that these variants appear only as heterozygous meaning they are not present in the male population of UKB. In the case of the *AVPR2* variants, one explaining factor could be the mutational constraint of the gene which describes the tolerance of the gene against pathogenic variants (Karczewski et al., 2020). According to the reported mutational constraint metrics of *AVPR2* in gnomAD the tolerance of the gene against missense and loss-of-function variants is rather high.

5.2.3 Comparison of gnomAD and UKB

Only 288 (likely) pathogenic ClinVar variants were detected from gnomAD and out of these 24 were in UKB carried by at least one individual. The explanation behind the low number of detected variants in gnomAD could be that the variants retrieved from ClinVar are extremely rare or even reported only from one individual. The number of shared variants between gnomAD and UKB seems expected and to see if the allele frequencies are similar in these databases, I conducted a correlation test. It was challenging to expect any specific result from this test as many factors contribute to the frequency of variants. The European ancestry behind both of these databases could indicate that the MAFs would be similar. On the other hand, the difference of samples in these databases would suggest that the MAFs of disease-associated variants would be higher in gnomAD as the population in UKB consist of healthy adults whereas gnomAD has sequencing information from various disease-specific studies in addition to population genetic studies (Karczewski et al., 2020). The correlation test with high correlation coefficient and very small p-value showed that the MAFs were comparable between the databases.

One significant finding was that the most common pathogenic variant in UKB was not detected in gnomAD. The variant rs104894858 on *LAMP2* had MAF over 3% among the participants of UKB and had a two-star rating in ClinVar for being pathogenic associated with hypertrophic cardiomyopathy and Danon disease. However, the mutational constraint of *LAMP2* indicate that the variant would be, in fact, tolerated in the population. The high allele frequency in UKB and the mutational constraint suggested that the variant would have been expected to also be present in

gnomAD. However, there is no coverage in the region this variant is located indicating that the variant has not been sequenced in the projects participating in gnomAD.

5.3 Genetic analyses

5.3.1 Overview of the analyses

In this thesis, 28 and 87 genetic association analyses were conducted among the male and female populations of UKB, respectively. The analyses were done with two sets of PCs; however, the results did not differ from each other remarkably so I will only focus on the analyses with 20 PCs. 11 variants were included in the male-analyses while the female-analyses consisted of 27 variants out of the 81 variants detected from UKB. The tested phenotypes were quantitative as the power to detect differences is higher with continuous than with binary traits. Different sources were used to select the appropriate traits among the symptoms of the disorders associated with the selected variants. Also, the functions of the genes were considered when choosing the phenotypes. Nevertheless, the quantitative traits might not represent the disorders associated with the variants which must be kept in mind while interpreting the results of the association analyses. Additionally, the fact that most of these variants are extremely rare decreases the power to detect statistically significant associations.

Out of the 28 analyses performed in the male population resulted in eight nominally significant associations (p-value < 0.05) and four Bonferroni corrected associations (p-value < 0.05/28) whereas the female-analyses showed significance with 13 associations nominally and only one after the multiple correction (p-value < 0.05/87) out of the 87 tests. The confidence intervals of the analyses were mostly rather large, but that observation can be explained by the low number of the minor alleles. One variant (rs137852591) seemed to be a good candidate contributing to partial androgen insensitivity syndrome (PAIS) as it showed at least nominally significant associations in both males and females. I will next discuss the results of this *AR* variant and then continue to examine the other associations showing statistical significance.

5.3.2 *AR* variants

The results of this thesis showed that the variant rs137852591 on gene *AR* impacts the muscle mass and height of a carrier male as the 285 males carrying the mutated allele seem to have lower amount of muscle mass (effect size = -0.142, p-value = $5.39x10^{-11}$) and be shorter (effect size = -0.134, p-value = $2.32x10^{-06}$) than the males having the normal allele. Also, the levels of testosterone in the blood stream are higher (effect size = 0.158, p-value = $6.37x10^{-8}$) in males having this mutation. These findings indicate that the variant is indeed pathogenic and associated with PAIS as is reported in ClinVar. PAIS is characterized by abnormal sexual development in males and the features vary greatly among the individuals affected by this syndrome (Hornig & Holterhus, 2021). In PAIS the patient's response to androgens, such as testosterone, is weakened as the gene (*AR*) responsible for the production of androgen receptors is changed and unable to produce the receptors (Hornig & Holterhus, 2021). The results of these analyses are supported by the characteristics of PAIS as males manifesting it have female-like features. On average, females are shorter and have less muscle mass than males, and androgens play an important role in this. In addition, as the androgen receptors are defected, the testosterone levels in the blood stream are increased.

PAIS is assumed to have a recessive inheritance pattern meaning that heterozygous females should be unaffected by it. However, the results of this thesis showed nominally significant associations between rs137852591 and lower muscle mass (effect size = -0.0861, p-value = 0.00124) and height (effect size = -0.118, p-value = 0.00128) in the 678 carrier females. Although the associations did not remain significant after the Bonferroni correction, they could still indicate that the variant has some impacts also in females. As some phenotypes are highly correlated, using the Bonferroni correction is rather conservative and may result in losing true associations (Rice et al., 2008). Also, comparing the effect sizes in males and females would suggest that the variant, indeed, affects growth and the production of muscle mass in females as well. The effects could be result of XCI, either random or skewed, as possibly the allele must be normal in more than 50% of the cells to guarantee sufficient production of the androgen receptors. As effects can be seen in females as well, the assumed recessiveness of PAIS could be questioned. Especially the association with height showed similar sized effect in both males and females. It is possible that skewed XCI is behind the results if the carrier females have more cells with the mutated allele as

active. However, the phenotypes are highly polygenic, and the interpretation needs to be careful and requires more research. No association between elevated serum testosterone levels and the variant was shown, which could be expected as testosterone levels, production, metabolism and function are different in males and females.

Another variant (rs137852593) on the gene *AR* was tested in males with the same phenotypes as the previous one. Unfortunately, the number of females with more the 90% probability of being heterozygous was too low (7) to include this in the association analysis and the comparison of male and female effects is not possible. The very unexpected lower number of heterozygotes is probably due to the imputation as the imputation info for this variant is only 0.50. Probably the male carriers of this variant were more accurately imputed. Also, the number of males having over 90% probability of carrying the mutated allele was only 9 leaving the power to detect associations rather low as well. However, a nominally significant association with higher serum testosterone level (effect size = 0.524, p-value = 0.0025) was observed. Possibly with a larger sample size more associations with this variant could be seen.

5.3.3 Other variants

Analyses conducted in males showed additionally one statistically significant association after the Bonferroni correction with variant rs72554665 (gene *G6PD*) and lower levels of glycated hemoglobin (effect size = -0.631, p-value = 4.19×10^{-05}). In females this association was only nominally significant (effect size = -0.349, p-value = 0.0117). Another variant (rs5030869) on the same gene; however, showed significant association with lower red blood cell count with a lower p-value (effect size = -0.950, p-value = 3.56×10^{-05}) in females. Additionally, associations with hemoglobin concentration (effect size = -0.614, p-value = 0.00777) and glycated hemoglobin (effect size = -0.508, p-value = 0.0173) were observed with this variant. Unluckily, this variant was carried only by three males and the number of heterozygotes (19) was also quite low. The results with the variants on this gene could imply that they affect both males and females; however, the low number of carriers decreases the power and therefore these associations are unable to reach statistical significance after the Bonferroni correction. The results are as expected since in glucose

6 phosphate dehydrogenase deficiency red blood cells are broken down and it has also been associated with lower levels of HbA1c (Leong et al., 2020).

The analyses with rest of the variants resulted in only nominally significant associations and only one of the associations was observed in both males and females. Variant rs9697983, associated with late-onset Parkinson disease, on gene *GLUD2* was associated with longer mean time to correctly identify matches (effect sizes (F,M) = 0.0171, 0.0153; p-values (F,M) = 0.0489, 0.0204). This variant is one of the common variants with MAF of 2.4%. The observed estimated effects are very small, and the p-values are close to the nominal threshold at least in the female analysis. These results suggest that the variant might indeed have a very minor role in the development of late-onset Parkinson disease, or that there is no real association between the variant and the disease.

The majority of the variants did not show any association with the chosen phenotypes. One reason behind this observation is that the variants are, in fact, tolerated in the population and are misclassified in ClinVar. The review status of these variants supports this assumption. Also, the rarity of these variants might impact the detection of associations as the power is quite low. However, it cannot be ruled out that some of the variants might truly have a recessive inheritance pattern and might not show any phenotypic effects in females. This could be the case with those variants that were not included in the replication analysis conducted in males and had no significant associations in females. In addition, many of the selected quantitative traits, if not all of them, are highly polygenic and many factors contribute to them and therefore small genetic effects may disappear under other genetic and environmental factors. For example, blood pressure and glucose levels are highly affected by lifestyle.

6 CONCLUSIONS

This thesis demonstrates that only a small portion of the (likely) pathogenic X-chromosomal ClinVar variants can be detected from large-scale databases UKB and gnomAD with hundreds of thousands of samples. The detected variants were mainly extremely rare which is unsurprising as the majority of variants included in ClinVar have one-star rating indicating that they are submitted

only from one source. There were few exceptions; however, the common variants seemed to have very mild effects on the chosen phenotypes.

The association analyses showed few statistically significant associations in males and in females. The strongest candidates were variants on genes *AR* and *G6PD* associated with partial androgen insensitivity syndrome and glucose 6 dehydrogenase deficiency, respectively. The results could indicate that heterozygous carrier females may, indeed, have symptoms of assumed recessive X-linked disorders as has been suggested before. This supports the idea that the classification into recessively and dominantly inherited traits may not be applicable in the case of all X-chromosomal disorders; however, the low number of associations detected in this thesis is not enough to draw any major conclusions.

The detected associations in females may be due to skewed XCI where the X chromosome with the mutated allele is preferred for some reason. To confirm this prediction, further research is required. For example, the XCI status of the females could be examined. Additionally, the traits included in this thesis, such as height, blood pressure and fat-free mass, are highly polygenic meaning that many factors contribute to them. A larger number of heterozygotes could increase the power and hence aid in detecting more associations. A phenome-wide association study, that includes more phenotypes than this thesis, could be beneficial in the future regarding X-chromosomal variants.

7 ACKNOWLEDGEMENTS

I would like to thank my supervisor Taru Tukiainen for the guidance and endless support during my Master's thesis project. Especially, the support and patience through the writing process was extremely encouraging. Also, thanks to the members of the Tukiainen research group for making me feel welcome. I would also like to thank my closest ones for supporting me through this project and helping me believe in myself at the times I needed it the most. Lastly, thanks to the anonymous participants in the UK Biobank and the researchers who have worked hard to make the Biobank the way it is now.

8 REFERENCES

Amos-Landgraf, J.M., Cottle, A., Plenge, R.M., Friez, M., Schwartz, C.E., Longshore, J., and Willard, H.F. (2006). X Chromosome–Inactivation Patterns of 1,005 Phenotypically Unaffected Females. The American Journal of Human Genetics *79*, 493–499.

Armstrong, R.A. (2014). When to use the Bonferroni correction. Ophthalmic Physiol Opt *34*, 502–508.

Basta, M., and Pandya, A.M. (2021). Genetics, X-Linked Inheritance. In StatPearls, (Treasure Island (FL): StatPearls Publishing)

Bianchi, I., Lleo, A., Gershwin, M.E., and Invernizzi, P. (2012). The X chromosome and immune associated genes. Journal of Autoimmunity *38*, J187–J192.

Blaschke, R.J., and Rappold, G. (2006). The pseudoautosomal regions, SHOX and disease. Current Opinion in Genetics & Development *16*, 233–239.

Brown, C.J., Ballabio, A., Rupert, J.L., Lafreniere, R.G., Grompe, M., Tonlorenzi, R., and Willard, H.F. (1991). A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. Nature *349*, 38–44.

Bryant, P., Boukouvala, A., McDaniel, J., and Nance, D. (2020). Hemophilia A in Females: Considerations for Clinical Management. Acta Haematol *143*, 289–294.

Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Research *47*, D1005–D1012.

Busque, L., Paquette, Y., Provost, S., Roy, D.-C., Levine, R.L., Mollica, L., and Gary Gilliland, D. (2009). Skewing of X-inactivation ratios in blood cells of aging women is confirmed by independent methodologies. Blood *113*, 3472–3474.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.

Carrel, L., and Willard, H.F. (2005). X-inactivation profile reveals extensive variability in X-linked gene expression in females. Nature *434*, 400–404.

Cirulli, E.T., White, S., Read, R.W., Elhanan, G., Metcalf, W.J., Tanudjaja, F., Fath, D.M., Sandoval, E., Isaksson, M., Schlauch, K.A., et al. (2020). Genome-wide rare variant analysis for thousands of phenotypes in over 70,000 exomes from two cohorts. Nat Commun *11*, 542.

Cotton, A.M., Ge, B., Light, N., Adoue, V., Pastinen, T., and Brown, C.J. (2013). Analysis of expressed SNPs identifies variable extents of expression from the human inactive X chromosome. Genome Biol *14*, R122.

Deng, X., Berletch, J.B., Nguyen, D.K., and Disteche, C.M. (2014). X chromosome regulation: diverse patterns in development, tissues and disease. Nat Rev Genet *15*, 367–378.

Dobyns, W.B., Filauro, A., Tomson, B.N., Chan, A.S., Ho, A.W., Ting, N.T., Oosterwijk, J.C., and Ober, C. (2004). Inheritance of most X-linked traits is not dominant or recessive, just X-linked. Am. J. Med. Genet. *129A*, 136–143.

Engelen, M., Barbier, M., Dijkstra, I.M.E., Schür, R., de Bie, R.M.A., Verhamme, C., Dijkgraaf, M.G.W., Aubourg, P.A., Wanders, R.J.A., van Geel, B.M., et al. (2014). X-linked adrenoleukodystrophy in women: a cross-sectional cohort study. Brain *137*, 693–706.

Fang, H., Deng, X., and Disteche, C.M. (2021). X-factors in human disease: impact of gene content and dosage regulation. Human Molecular Genetics *30*, R285–R295.

Fuller, Z.L., Berg, J.J., Mostafavi, H., Sella, G., and Przeworski, M. (2019). Measuring intolerance to mutation in human genetics. Nat Genet *51*, 772–776.

Galupa, R., and Heard, E. (2018). X-Chromosome Inactivation: A Crossroads Between Chromosome Architecture and Gene Regulation. Annu. Rev. Genet. *52*, 535–566.

Germain, D.P. (2006). General aspects of X-linked diseases. In Fabry Disease: Perspectives from 5 Years of FOS, A. Mehta, M. Beck, and G. Sunder-Plassmann, eds. (Oxford: Oxford PharmaGenesis)

Gilbert SF. Developmental Biology. 6th edition. Sunderland (MA): Sinauer Associates; 2000. Available from: https://www.ncbi.nlm.nih.gov/books/NBK9983/,

Goddard, M.E., Kemper, K.E., MacLeod, I.M., Chamberlain, A.J., and Hayes, B.J. (2016). Genetics of complex traits: prediction of phenotype, identification of causal polymorphisms and genetic architecture. Proc. R. Soc. B. *283*, 20160569.

Graves, J.A.M. (2006). Sex Chromosome Specialization and Degeneration in Mammals. Cell *124*, 901–914.

Gurbich, T.A., and Bachtrog, D. (2008). Gene content evolution on the X chromosome. Current Opinion in Genetics & Development *18*, 493–498.

Hornig, N.C., and Holterhus, P.-M. (2021). Molecular basis of androgen insensitivity syndromes. Molecular and Cellular Endocrinology *523*, 111146.

Howe, K.L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Azov, A.G., Bennett, R., Bhai, J., et al. (2021). Ensembl 2021. Nucleic Acids Research *49*, D884–D891.

Jons, W.A., Colby, C.L., McElroy, S.L., Frye, M.A., Biernacka, J.M., and Winham, S.J. (2019). Statistical methods for testing X chromosome variant associations: application to sex-specific characteristics of bipolar disorder. Biol Sex Differ *10*, 57.

Juchniewicz, P., Piotrowska, E., Kloska, A., Podlacha, M., Mantej, J., Węgrzyn, G., Tukaj, S., and Jakóbkiewicz-Banecka, J. (2021). Dosage Compensation in Females with X-Linked Metabolic Disorders. IJMS *22*, 4514.

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443.

Kashimada, K., and Koopman, P. (2010). *Sry* : the master switch in mammalian sex determination. Development *137*, 3921–3930.

Kliegman, R., Stanton, B., St. Geme, J.W., Schor, N.F., Behrman, R.E., and Nelson, W.E. (2020). Nelson textbook of pediatrics.

König, I.R., Loley, C., Erdmann, J., and Ziegler, A. (2014). How to Include Chromosome X in Your Genome-Wide Association Study. Genet. Epidemiol. *38*, 97–103.

Kristiansen, M., Knudsen, G.P.S., Bathum, L., Naumova, A.K., Sørensen, T.I.A., Brix, T.H., Svendsen, A.J., Christensen, K., Kyvik, K.O., and Ørstavik, K.H. (2005). Twin study of genetic and aging effects on X chromosome inactivation. Eur J Hum Genet *13*, 599–606.

Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., et al. (2018). ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Research *46*, D1062–D1067.

Laumonnier, F., Cuthbert, P.C., and Grant, S.G.N. (2007). The Role of Neuronal Complexes in Human X-Linked Brain Diseases. The American Journal of Human Genetics *80*, 205–220.

Lee, S., Abecasis, G.R., Boehnke, M., and Lin, X. (2014). Rare-Variant Association Analysis: Study Designs and Statistical Tests. The American Journal of Human Genetics *95*, 5–23.

Leong, A., Lim, V.J.Y., Wang, C., Chai, J.-F., Dorajoo, R., Heng, C.-K., van Dam, R.M., Koh, W.-P., Yuan, J.-M., Jonas, J.B., et al. (2020). Association of G6PD variants with hemoglobin A1c and impact on diabetes diagnosis in East Asian individuals. BMJ Open Diab Res Care *8*, e001091.

Lercher, M.J., Urrutia, A.O., and Hurst, L.D. (2003). Evidence That the Human X Chromosome Is Enriched for Male-Specific but not Female-Specific Genes. Molecular Biology and Evolution *20*, 1113–1116.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics *27*, 2987–2993.

Lyon, M.F. (1961). Gene Action in the X-chromosome of the Mouse (Mus musculus L.). Nature *190*, 372–373.

Lyon, M.F. (1962). Sex chromatin and gene action in the mammalian X-chromosome. Am J Hum Genet *14*, 135–148.

Ma, S., and Shi, G. (2020). On rare variants in principal component analysis of population stratification. BMC Genet *21*, 34.

Maier, E.M., Kammerer, S., Muntau, A.C., Wichers, M., Braun, A., and Roscher, A.A. (2002). Symptoms in carriers of adrenoleukodystrophy relate to skewed X inactivation. Ann Neurol. *52*, 683–688.

Martin, H.C., Gardner, E.J., Samocha, K.E., Kaplanis, J., Akawi, N., Sifrim, A., Eberhardt, R.Y., Tavares, A.L.T., Neville, M.D.C., Niemi, M.E.K., et al. (2021). The contribution of X-linked coding variation to severe developmental disorders. Nat Commun *12*, 627.

Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J.A., Ziyatdinov, A., Benner, C., O'Dushlaine, C., Barber, M., Boutkov, B., et al. (2021). Computationally efficient whole-genome regression for quantitative and binary traits. Nat Genet *53*, 1097–1103.

MedlinePlus [Internet]. Bethesda (MD): National Library of Medicine (US); [updated 2021 Oct 5]. Color vision deficiency; [updated 2020 Aug 18; reviewed 2015 Jan 1; cited 2021 Oct 4]. Available from: https://medlineplus.gov/genetics/condition/color-vision-deficiency/

Migeon, B.R. (2020). X-linked diseases: susceptible females. Genetics in Medicine *22*, 1156–1174.

Miller, D.B., and Piccolo, S.R. (2020). Compound Heterozygous Variants in Pediatric Cancers: A Systematic Review. Front. Genet. *11*, 493.

Mimoto, M.S., and Refetoff, S. (2020). Clinical recognition and evaluation of patients with inherited serum thyroid hormone-binding protein mutations. J Endocrinol Invest *43*, 31–41.

Minks, J., Robinson, W.P., and Brown, C.J. (2008). A skewed view of X chromosome inactivation. J. Clin. Invest. *118*, 20–23.

Neri, G., Schwartz, C.E., Lubs, H.A., and Stevenson, R.E. (2018). X-linked intellectual disability update 2017. Am J Med Genet *176*, 1375–1388.

Nguyen, D.K., and Disteche, C.M. (2006). High expression of the mammalian X chromosome in brain. Brain Research *1126*, 46–49.

OMIM. X-linked diseases. 2022. https://www.ncbi.nlm.nih.gov/omim/?term=X-linked+diseases. (Visited 19.1.2022)

OMIM. 2021. http://omim.org/. Online Mendelian Inheritance in Man; visited 6.10.2021

Özbek, U., Lin, H.-M., Lin, Y., Weeks, D.E., Chen, W., Shaffer, J.R., Purcell, S.M., and Feingold, E. (2018). Statistics for X-chromosome associations. Genet. Epidemiol. *42*, 539–550.

Plenge, R.M., Stevenson, R.A., Lubs, H.A., Schwartz, C.E., and Willard, H.F. (2002). Skewed X-Chromosome Inactivation Is a Common Feature of X-Linked Mental Retardation Disorders. The American Journal of Human Genetics *71*, 168–173.

Plomin, R., Haworth, C.M.A., and Davis, O.S.P. (2009). Common disorders are quantitative traits. Nat Rev Genet *10*, 872–878.

Posynick, B.J., and Brown, C.J. (2019). Escape From X-Chromosome Inactivation: An Evolutionary Perspective. Front. Cell Dev. Biol. *7*, 241.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Rebuzzini, P., Zuccotti, M., and Garagna, S. (2020). X-Chromosome Inactivation during Preimplantation Development and in Pluripotent Stem Cells. Cytogenet Genome Res *160*, 283–294.

Rice, T.K., Schork, N.J., and Rao, D.C. (2008). Methods for Handling Multiple Testing. In Advances in Genetics, (Elsevier), pp. 293–308.

Risch, N.J. (2000). Searching for genetic determinants in the new millennium. Nature *405*, 847–856.

Ross, M.T., Grafham, D.V., Coffey, A.J., Scherer, S., McLay, K., Muzny, D., Platzer, M., Howell, G.R., Burrows, C., Bird, C.P., et al. (2005). The DNA sequence of the human X chromosome. Nature *434*, 325–337.

Santos-Rebouças, C.B., Boy, R., Vianna, E.Q., Gonçalves, A.P., Piergiorge, R.M., Abdala, B.B., dos Santos, J.M., Calassara, V., Machado, F.B., Medina-Acosta, E., et al. (2020). Skewed X-Chromosome Inactivation and Compensatory Upregulation of Escape Genes Precludes Major Clinical Symptoms in a Female With a Large Xq Deletion. Front. Genet. *11*, 101.

Shah, N., Hou, Y.-C.C., Yu, H.-C., Sainger, R., Caskey, C.T., Venter, J.C., and Telenti, A. (2018). Identification of Misclassified ClinVar Variants via Disease Population Prevalence. The American Journal of Human Genetics *102*, 609–619.

Sidorenko, J., Kassam, I., Kemper, K.E., Zeng, J., Lloyd-Jones, L.R., Montgomery, G.W., Gibson, G., Metspalu, A., Esko, T., Yang, J., et al. (2019). The effect of X-linked dosage compensation on complex trait variation. Nat Commun *10*, 3009.

Silva, T.H. da, Anequini, I.P., Fávero, F.M., Voos, M.C., Oliveira, A.S.B., Telles, J.A.R., and Caromano, F.A. (2020). Functional performance and muscular strength in symptomatic female carriers of Duchenne muscular dystrophy. Arq. Neuro-Psiquiatr. *78*, 143–148.

Sole, X., Guino, E., Valls, J., Iniesta, R., and Moreno, V. (2006). SNPStats: a web tool for the analysis of association studies. Bioinformatics *22*, 1928–1929.

Strachan, T. & Read, A. P. (2019) Human molecular genetics. Fifth edition. Boca Raton, Florida ;: CRC Press.

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. Nat Rev Genet *20*, 467–484.

The 1000 Genomes Project Consortium, Corresponding authors, Auton, A., Abecasis, G.R., Steering committee, Altshuler, D.M., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Trent, R. (2012) Molecular Medicine: Genomics to Personalized Healthcare. 4th edition. San Diego: Elsevier Science & Technology.

Tukiainen, T., Pirinen, M., Sarin, A.-P., Ladenvall, C., Kettunen, J., Lehtimäki, T., Lokki, M.-L., Perola, M., Sinisalo, J., Vlachopoulou, E., et al. (2014). Chromosome X-Wide Association Study Identifies Loci

for Fasting Insulin and Height and Evidence for Incomplete Dosage Compensation. PLoS Genet *10*, e1004127.

Tukiainen, T., Villani, A.-C., Yen, A., Rivas, M.A., Marshall, J.L., Satija, R., Aguirre, M., Gauthier, L., Fleharty, M., Kirby, A., et al. (2017). Landscape of X chromosome inactivation across human tissues. Nature *550*, 244–248.

Vianna, E.Q., Piergiorge, R.M., Gonçalves, A.P., dos Santos, J.M., Calassara, V., Rosenberg, C., Krepischi, A.C.V., Boy da Silva, R.T., dos Santos, S.R., Ribeiro, M.G., et al. (2020). Understanding the Landscape of X-linked Variants Causing Intellectual Disability in Females Through Extreme X Chromosome Inactivation Skewing. Mol Neurobiol *57*, 3671–3684.

Vickers, M.A., McLeod, E., Spector, T.D., and Wilson, I.J. (2001). Assessment of mechanism of acquired skewed X inactivation by analysis of twins. Blood *97*, 1274–1281.

Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. The American Journal of Human Genetics *101*, 5–22.

Vissers, L.E.L.M., Gilissen, C., and Veltman, J.A. (2016). Genetic studies in intellectual disability and related disorders. Nat Rev Genet *17*, 9–18.

Wickham, H. (2016). ggplot2: elegant graphics for data analysis (Cham: Springer International Publishing).

Wise, A.L., Gyi, L., and Manolio, T.A. (2013). eXclusion: Toward Integrating the X Chromosome in Genome-wide Association Analyses. The American Journal of Human Genetics *92*, 643–647.

Wong, C.C.Y., Caspi, A., Williams, B., Houts, R., Craig, I.W., and Mill, J. (2011). A Longitudinal Twin Study of Skewed X Chromosome-Inactivation. PLoS ONE *6*, e17873.

Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. The American Journal of Human Genetics *89*, 82–93.

Xiang, J., Yang, J., Chen, L., Chen, Q., Yang, H., Sun, C., Zhou, Q., and Peng, Z. (2020). Reinterpretation of common pathogenic variants in ClinVar revealed a high proportion of downgrades. Sci Rep *10*, 331.

Zito, A., Davies, M.N., Tsai, P.-C., Roberts, S., Andres-Ejarque, R., Nardone, S., Bell, J.T., Wong, C.C.Y., and Small, K.S. (2019). Heritability of skewed X-inactivation in female twins is tissue-specific and associated with age. Nat Commun *10*, 5339.

Zuo, X.-Y., Feng, Q.-S., Sun, J., Wei, P.-P., Chin, Y.-M., Guo, Y.-M., Xia, Y.-F., Li, B., Xia, X.-J., Jia, W.-H., et al. (2019). X-chromosome association study reveals genetic susceptibility loci of nasopharyngeal carcinoma. Biol Sex Differ *10*, 13.

Supplementary table 1. Detected variants in UKB. Variants coded as green were included in both association analyses, variants coded as red were exclusive to female-analysis and the variant coded as blue was exclusive for male-analysis. (*the number of individuals reaching the 0.9 threshold of the genotype)

Variant	Gene	Imputation info	MAF	Hetero count	Hemi count	Phenotype
rs80338710	ARSL	0.114	$7.16x10^{-6}$	4 (*0)	1(*0)	X-linked chondrodysplasia punctata 1
rs137852517	ANOS1	0.820	$1.86x10^{-4}$	79	42	Hypogonadotropic hypogonadism 1 with or without anosmia (Kallmann syndrome 1)
rs62653623	CDKL5	NA	$9.40x10^{-6}$	7	0	Early infantile epileptic encephalopathy 2
rs104894934	RS1	NA	$4.01x10^{-5}$	3	0	Juvenile retinoschisis
rs199959402	PDHA1	0.456	$4.65x10^{-6}$	3(*1)	1(*0)	not provided
rs128627256	DMD	NA	$1.63x10^{-5}$	11	0	Becker muscular dystrophy
rs104894954	XK	NA	$1.58x10^{-6}$	1	0	McLeod neuroacanthocytosis syndrome
rs104894953	XK	NA	$4.43x10^{-6}$	3	0	McLeod neuroacanthocytosis syndrome
rs67486158	OTC	0.962	$3.33x10^{-4}$	177(*169)	73(*69)	Ornithine carbamoyltransferase deficiency
rs121434619	BCOR	NA	$1.45x10^{-6}$	1	0	Oculofaciocardiodental syndrome
rs189751539	KDM6A	0.250	$5.01x10^{-5}$	26(*21)	11(*8)	Kabuki syndrome 2
rs782344765	RP2	0.642	$9.15x10^{-5}$	50(*21)	19(*8)	not provided
rs777516785	USP11	0.433	$1.14x10^{-5}$	7(*1)	2(*1)	Bronchiectasis
rs782290433	WAS	0.541	$1.59x10^{-5}$	9(*4)	2(*1)	X-linked severe congenital neutropenia
rs782367133	KDM5C	0.335	$1.43x10^{-4}$	75(*8)	32(*7)	Smith-Magenis Syndrome-like
rs782600511	KDM5C	0.331	$4.46x10^{-5}$	23(*2)	11(*0)	Mental retardation, syndromic, Claes-Jensen type, X-linked
rs137852593	AR	0.502	$1.08x10^{-4}$	48(*7)	33(*9)	Prostate cancer susceptibility
rs137852571	AR	0.230	$3.82x10^{-5}$	22(*5)	6(*0)	Prostate cancer, somatic
rs137852591	AR	0.623	$1.29x10^{-3}$	678	285	Partial androgen insensitivity syndrome
rs140984555	TEX11	0.794	$7.48x10^{-5}$	46(*33)	11(*7)	Spermatogenic failure, X-linked, 2
rs143246552	TEX11	0.927	$5.73x10^{-5}$	46(*30)	10(*7)	Spermatogenic failure, X-linked, 2
rs149447856	SLC7A3	1	$2.82x10^{-3}$	1469	647	not provided
rs104894811	GJB1	NA	$2.01x10^{-5}$	15	0	Charcot-Marie-Tooth Neuropathy X Type 1
rs200060292	TBX22	0.864	$2.07x10^{-4}$	106(*85)	49(*42)	Abruzzo-Erickson syndrome
rs61754490	SERPINA7	0.924	$2.28x10^{-3}$	1195(*989)	506(*423)	Thyroxine-binding globulin, Chicago
rs28933689	SERPINA7	NA	$2.39x10^{-5}$	18	0	Thyroxine-binding globulin, Chicago
rs104886312	COL4A5	NA	$6.64x10^{-5}$	42	7	Alport syndrome 1, X-linked recessive
rs104886150	COL4A5	NA	$1.46x10^{-5}$	11	0	Alport syndrome 1, X-linked recessive
rs104886303	COL4A5	NA	$1.33x10^{-6}$	1	0	Alport syndrome 1, X-linked recessive
rs104894859	LAMP2	NA	$2.95x10^{-6}$	2	0	Danon disease
rs104894858	LAMP2	NA	$3.89x10^{-2}$	17083	7502	Hypertrophic cardiomyopathy
rs104894857	LAMP2	NA	$9.20x10^{-5}$	59	0	Danon disease
rs137852527	LAMP2	NA	$3.10x10^{-5}$	20	1	Danon disease
rs766506778	CUL4B	0.543	$5.56x10^{-5}$	28(*12)	14(*8)	not provided
rs9697983	GLUD2	1	$2.51x10^{-2}$	12653	5431	Parkinson disease, late-onset
rs724160014	AIFM1	0.227	$1.27x10^{-5}$	7(*1)	3(*0)	Deafness, X-linked 5
rs104894854	GPC3	NA	$1.18x10^{-5}$	8	0	Simpson-Golabi-Behmel syndrome type 1
rs122458140	FHL1	NA	$9.30x10^{-6}$	7	0	Scapuloperoneal myopathy, X-linked dominant
rs122458143	FHL1	NA	$6.64x10^{-6}$	5	0	Myopathy, reducing body, X-linked, early-onset, severe
rs122459146	FHL1	NA	$2.00x10^{-5}$	15	0	Myopathy, reducing body, X-linked, early-onset, severe
rs122458144	FHL1	NA	$1.40x10^{-6}$	1	0	Myopathy, reducing body, X-linked, childhood-onset
rs122458145	FHL1	NA	$1.06x10^{-5}$	8	0	Myopathy, reducing body, X-linked, childhood-onset
rs122459149	FHL1	NA	$1.33x10^{-6}$	1	0	Emery-Dreifuss muscular dystrophy 6
rs137852237	F9	NA	$5.37x10^{-6}$	4	0	Thrombophilia, X-linked, due to factor IX defect
rs193922112	AVPR2	NA	$1.34x10^{-6}$	1	0	Nephrogenic diabetes insipidus

rs104894751	AVPR2	NA	$6.64x10^{-6}$	5	0	Nephrogenic diabetes insipidus, X-linked
rs104894760	AVPR2	NA	$2.39x10^{-5}$	18	0	Nephrogenic diabetes insipidus
rs104894758	AVPR2	NA	$3.99x10^{-6}$	3	0	Nephrogenic diabetes insipidus, X-linked
rs104894747	AVPR2	NA	$2.66x10^{-5}$	20	0	Nephrogenic diabetes insipidus, X-linked
rs104894757	AVPR2	NA	$1.77x10^{-5}$	12	0	Nephrogenic diabetes insipidus, X-linked
rs104894748	AVPR2	NA	$1.46x10^{-5}$	11	0	Nephrogenic diabetes insipidus, X-linked
rs104894755	AVPR2	NA	$4.02x10^{-6}$	3	0	Nephrogenic diabetes insipidus, X-linked
rs104894750	AVPR2	NA	$2.68x10^{-6}$	2	0	Nephrogenic diabetes insipidus, X-linked
rs193922117	AVPR2	NA	$1.33x10^{-6}$	1	0	Nephrogenic diabetes insipidus
rs193922122	AVPR2	NA	$3.99x10^{-6}$	3	0	Nephrogenic diabetes insipidus
rs104894753	AVPR2	NA	$1.99x10^{-5}$	15	0	Nephrogenic diabetes insipidus, X-linked
rs137853311	FLNA	0.594	$3.01x10^{-5}$	15(*6)	8(*2)	Periventricular nodular heterotopia 1
rs104894941	TAFAZZIN	NA	$1.49x10^{-5}$	11	0	3-Methylglutaconic aciduria type 2
rs104894937	TAFAZZIN	NA	$1.20x10^{-5}$	9	0	3-Methylglutaconic aciduria type 2
rs132630277	TAFAZZIN	NA	$5.36x10^{-6}$	4	0	3-Methylglutaconic aciduria type 2
rs202070666	PLXNA3	0.560	$7.15x10^{-5}$	39(*2)	14(*0)	Short stature
rs72554664	G6PD	0.450	$9.33x10^{-5}$	48(*14)	22(*10)	Glucose 6 phosphate dehydrogenase deficiency
rs72554665	G6PD	0.719	$1.57x10^{-4}$	86(*47)	27(*12)	Glucose 6 phosphate dehydrogenase deficiency
rs398123546	G6PD	0.235	$1.65x10^{-5}$	9(*0)	3(*1)	Glucose 6 phosphate dehydrogenase deficiency
rs5030869	G6PD	NA	$3.04x10^{-5}$	19	3	Glucose 6 phosphate dehydrogenase deficiency
rs121912303	DKC1	NA	$2.95x10^{-6}$	2	0	Dyskeratosis congenita, X-linked
rs199422242	DKC1	NA	$2.99x10^{-6}$	2	0	Dyskeratosis congenita, X-linked
rs28936072	DKC1	NA	$2.95x10^{-6}$	2	0	Dyskeratosis congenita, X-linked
rs121912296	DKC1	NA	$1.60x10^{-6}$	1	0	Dyskeratosis congenita, X-linked
rs121912292	DKC1	NA	$2.21x10^{-5}$	15	0	Dyskeratosis congenita, X-linked
rs121912302	DKC1	NA	$7.38x10^{-6}$	5	0	Dyskeratosis congenita, X-linked
rs199422243	DKC1	NA	$2.98x10^{-6}$	2	0	Dyskeratosis congenita, X-linked
rs121912304	DKC1	NA	$5.90x10^{-6}$	4	0	Dyskeratosis congenita, X-linked
rs121912301	DKC1	NA	$2.95x10^{-6}$	2	0	Dyskeratosis congenita, X-linked
rs121912297	DKC1	NA	$1.52x10^{-6}$	1	0	Dyskeratosis congenita, X-linked
rs199422244	DKC1	NA	$4.52x10^{-6}$	3	0	Dyskeratosis congenita, X-linked
rs199422245	DKC1	NA	$1.03x10^{-5}$	7	0	Dyskeratosis congenita, X-linked
rs199422247	DKC1	NA	$2.07x10^{-5}$	14	0	Dyskeratosis congenita, X-linked
rs199422248	DKC1	NA	$1.49x10^{-6}$	1	0	Dyskeratosis congenita, X-linked
rs121912300	DKC1	NA	$2.96x10^{-6}$	2	0	Dyskeratosis congenita, X-linked
rs137854492	DKC1	NA	$4.43x10^{-6}$	3	0	Dyskeratosis congenita, X-linked
rs199422249	DKC1	NA	$1.48x10^{-5}$	10	0	Dyskeratosis congenita, X-linked
rs199422251	DKC1	NA	$4.48x10^{-6}$	3	0	Dyskeratosis congenita, X-linked
rs199422252	DKC1	NA	$2.95x10^{-6}$	2	0	Dyskeratosis congenita, X-linked
rs121912299	DKC1	NA	$8.90x10^{-6}$	6	0	Dyskeratosis congenita, X-linked
rs121912295	DKC1	NA	$1.77x10^{-5}$	12	0	Dyskeratosis congenita, X-linked
rs121912289	DKC1	NA	$8.85x10^{-6}$	6	0	Dyskeratosis congenita, X-linked
rs137852388	F8	0.533	$3.08x10^{-4}$	164	50	Hereditary factor VIII deficiency disease
rs398122917	CLIC2	0.176	$4.41x10^{-5}$	23(*1)	10(*1)	Mental retardation, X-linked, syndromic 32

Supplementary table 2. Results of the association analyses conducted in males. Hemi_count indicates the number of males carrying the mutated allele. (*nominally significant, **significant after multiple correction)

Variant	Gene (length)	MAF (%)	Hemi_count	Tested trait	ClinVar phenotype	Beta_10PC (Beta_20_PC)	SE_10PC (SE_20PC)	P-value_10PC (P-value_20PC)
rs137852517	*ANOS1* (203,313bp)	0.0186	30	Testosterone	Hypogonadotropic hypogonadism 1 with or without anosmia (Kallmann syndrome 1)	0.1443 (0.1439)	0.0917 (0.0917)	0.1156 (0.1167)
rs67486158	*OTC* (68,906bp)	0.0333	69	Heel bone mineral density	Ornithine carbamoyltransferase deficiency	-0.1065 (-0.1074)	0.0779 (0.0779)	0.1716 (0.1681)
rs67486158	*OTC* (68,906bp)	0.0333	69	Microalbumin in urine	Ornithine carbamoyltransferase deficiency	-0.0539 (-0.0520)	0.1019 (0.1019)	0.5966 (0.6101)
rs137852593	*AR* (185,997bp)	0.0108	9	Standing height	Prostate cancer susceptibility	-0.0972 (-0.0902)	0.1541 (0.1539)	0.5285 (0.5576)
rs137852593	*AR* (185,997bp)	0.0108	9	Testosterone	Prostate cancer susceptibility	0.5253 (0.5243)	0.1737 (0.1736)	0.0025* (0.0025*)
rs137852593	*AR* (185,997bp)	0.0108	9	Whole body fat-free mass	Prostate cancer susceptibility	-0.1649 (-0.1637)	0.1297 (0.1294)	0.2038 (0.2060)
rs137852591	*AR* (185,997bp)	0.1289	285	Standing height	Partial androgen insensitivity syndrome	-0.1356 (-0.1348)	0.0286 (0.0285)	$2.13x10^{-6}$** ($2.32x10^{-6}$**)
rs137852591	*AR* (185,997bp)	0.1289	285	Testosterone	Partial androgen insensitivity syndrome	0.1590 (0.1587)	0.0293 (0.0293)	$6.04x10^{-8}$** ($6.37x10^{-8}$**)
rs137852591	*AR* (185,997bp)	0.1289	285	Whole body fat-free mass	Partial androgen insensitivity syndrome	-0.1432 (-0.1422)	0.0217 (0.0217)	$4.38x10^{-11}$** ($5.39x10^{-11}$**)
rs200060292	*TBX22* (17,014bp)	0.0207	42	Standing height	Abruzzo-Erickson syndrome	0.0776 (0.0771)	0.0739 (0.0737)	0.2936 (0.2955)
rs104886312	*COL4A5* (257,702bp)	0.0066	7	Diastolic blood pressure	Alport syndrome 1, X-linked recessive	-0.2358 (-0.2342)	0.1988 (0.1987)	0.2355 (0.2385)
rs104886312	*COL4A5* (257,702bp)	0.0066	7	Systolic blood pressure	Alport syndrome 1, X-linked recessive	-0.3974 (-0.3959)	0.1945 (0.1945)	0.0411* (0.0418*)
rs104894858	*LAMP2* (41,539bp)	3.8895	7502	Diastolic blood pressure	Hypertrophic cardiomyopathy	0.0009 (0.0009)	0.0059 (0.0059)	0.8860 (0.8804)
rs104894858	*LAMP2* (41,539bp)	3.8895	7502	Pulse rate	Hypertrophic cardiomyopathy	0.0012 (0.0013)	0.0060 (0.0060)	0.8467 (0.8246)
rs104894858	*LAMP2* (41,539bp)	3.8895	7502	Systolic blood pressure	Hypertrophic cardiomyopathy	-0.0067 (-0.0066)	0.0058 (0.0058)	0.2493 (0.2537)
rs9697983	*GLUD2* (2,333bp)	2.5052	5431	Hand grip strength, left	Parkinson disease, late-onset	0.0060 (0.0059)	0.0066 (0.0066)	0.3575 (0.3644)
rs9697983	*GLUD2* (2,333bp)	2.5052	5431	Hand grip strength, right	Parkinson disease, late-onset	-0.0052 (-0.0051)	0.0066 (0.0066)	0.4269 (0.4401)
rs9697983	*GLUD2* (2,333bp)	2.5052	5431	Mean time to correctly identify matches	Parkinson disease, late-onset	0.0156 (0.0153)	0.0066 (0.0066)	0.0181* (0.0204*)
rs72554664	*G6PD* (16,182bp)	0.0093	10	Glucose	Glucose 6 phosphate dehydrogenase deficiency	0.3545 (0.3560)	0.1613 (0.1613)	0.0280* (0.0274*)
rs72554664	*G6PD* (16,182bp)	0.0093	10	Glycated hemoglobin (HbA1c)	Glucose 6 phosphate dehydrogenase deficiency	-0.1538 (-0.1566)	0.1563 (0.1563)	0.3252 (0.3164)
rs72554664	*G6PD* (16,182bp)	0.0093	10	Hemoglobin concentration	Glucose 6 phosphate dehydrogenase deficiency	-0.0743 (-0.0771)	0.1725 (0.1725)	0.6666 (0.6551)
rs72554664	*G6PD* (16,182bp)	0.0093	10	Red blood cell count	Glucose 6 phosphate dehydrogenase deficiency	-0.2125 (-0.2236)	0.1698 (0.1697)	0.2107 (0.1877)
rs72554665	*G6PD* (16,182bp)	0.0157	12	Glucose	Glucose 6 phosphate dehydrogenase deficiency	-0.1601 (-0.1579)	0.1494 (0.1494)	0.2840 (0.2907)
rs72554665	*G6PD* (16,182bp)	0.0157	12	Glycated hemoglobin (HbA1c)	Glucose 6 phosphate dehydrogenase deficiency	-0.6291 (-0.6308)	0.1540 (0.1540)	$4.39x10^{-5}$** ($4.19x10^{-5}$**)
rs72554665	*G6PD* (16,182bp)	0.0157	12	Hemoglobin concentration	Glucose 6 phosphate dehydrogenase deficiency	-0.0150 (-0.0159)	0.1446 (0.1446)	0.9171 (0.9123)
rs72554665	*G6PD* (16,182bp)	0.0157	12	Red blood cell count	Glucose 6 phosphate dehydrogenase deficiency	-0.0977 (-0.1096)	0.1423 (0.1422)	0.4923 (0.4408)
rs137852388	*F8* (191,153bp)	0.0308	50	Mean corpuscular volume	Hereditary factor VIII deficiency disease	0.0390 (0.0346)	0.0695 (0.0694)	0.5746 (0.6184)
rs137852388	*F8* (191,153bp)	0.0308	50	Red blood cell count	Hereditary factor VIII deficiency disease	-0.0198 (-0.0152)	0.0701 (0.0701)	0.7774 (0.8278)

Supplementary table 3. Results of the association analyses conducted in females. Hetero_count indicates the number of heterozygotes carrying the mutated allele. (*nominally significant, **significant after multiple correction)

Variant	Gene (length)	MAF (%)	Hetero_count	Tested trait	ClinVar phenotype	Beta_10PC (Beta_20_PC)	SE_10PC (SE_20PC)	P-value_10PC (P-value_20PC)
rs137852517	*ANOS1* (203,313bp)	0.0186	94	Testosterone	Hypogonadotropic hypogonadism 1 with or without anosmia (Kallmann syndrome 1)	0.2114 (0.2135)	0.1162 (0.1162)	0.0690 (0.0662)
rs128627256	*DMD* (2,241,765bp)	0.0016	11	Hand grip strength, left	Becker muscular dystrophy	-0.1924 (-0.1926)	0.2827 (0.2822)	0.4961 (0.4950)
rs128627256	*DMD* (2,241,765bp)	0.0016	11	Hand grip strength, right	Becker muscular dystrophy	0.0727 (0.0718)	0.2831 (0.2827)	0.7974 (0.7996)
rs128627256	*DMD* (2,241,765bp)	0.0016	11	Whole body fat-free mass	Becker muscular dystrophy	-0.2057 (-0.1993)	0.2081 (0.2074)	0.3228 (0.3365)
rs67486158	*OTC* (68,906bp)	0.0333	169	Heel bone mineral density	Ornithine carbamoyltransferase deficiency	0.0229 (0.0184)	0.1013 (0.1013)	0.8213 (0.8561)
rs67486158	*OTC* (68,906bp)	0.0333	169	Microalbumin in urine	Ornithine carbamoyltransferase deficiency	-0.0985 (-0.0892)	0.1316 (0.1317)	0.4540 (0.4981)
rs137852591	*AR* (185,997bp)	0.1289	678	Standing height	Partial androgen insensitivity syndrome	-0.1201 (-0.1180)	0.0367 (0.0366)	0.0011* (0.0013*)
rs137852591	*AR* (185,997bp)	0.1289	678	Testosterone	Partial androgen insensitivity syndrome	0.0174 (0.0171)	0.0425 (0.0425)	0.6820 (0.6868)
rs137852591	*AR* (185,997bp)	0.1289	678	Whole body fat-free mass	Partial androgen insensitivity syndrome	-0.0902 (-0.0861)	0.0268 (0.0267)	0.0008* (0.0012*)
rs104894811	*GJB1* (10,323bp)	0.0020	15	Hand grip strength, left	Charcot-Marie-Tooth Neuropathy X Type 1	-0.0681 (-0.0539)	0.2426 (0.2421)	0.7790 (0.8238)
rs104894811	*GJB1* (10,323bp)	0.0020	15	Hand grip strength, right	Charcot-Marie-Tooth Neuropathy X Type 1	-0.0859 (-0.0709)	0.2426 (0.2422)	0.7234 (0.7697)
rs104894811	*GJB1* (10,323bp)	0.0020	15	Whole body fat-free mass	Charcot-Marie-Tooth Neuropathy X Type 1	-0.2642 (-0.2787)	0.1850 (0.1844)	0.1532 (0.1307)
rs200060292	*TBX22* (17,014bp)	0.0207	85	Standing height	Abruzzo-Erickson syndrome	0.1124 (0.1010)	0.1038 (0.1036)	0.2791 (0.3296)
rs104886312	*COL4A5* (257,702bp)	0.0066	42	Diastolic blood pressure	Alport syndrome 1, X-linked recessive	0.0664 (0.0645)	0.1521 (0.1520)	0.6625 (0.6712)
rs104886312	*COL4A5* (257,702bp)	0.0066	42	Systolic blood pressure	Alport syndrome 1, X-linked recessive	0.1174 (0.1164)	0.1456 (0.1455)	0.4200 (0.4240)
rs104886150	*COL4A5* (257,702bp)	0.0015	11	Diastolic blood pressure	Alport syndrome 1, X-linked recessive	-0.3374 (-0.3437)	0.2863 (0.2862)	0.2387 (0.2299)
rs104886150	*COL4A5* (257,702bp)	0.0015	11	Systolic blood pressure	Alport syndrome 1, X-linked recessive	-0.0989 (-0.1022)	0.2740 (0.2740)	0.7182 (0.7093)
rs104894858	*LAMP2* (41,539bp)	3.8895	17083	Diastolic blood pressure	Hypertrophic cardiomyopathy	0.0087 (0.0083)	0.0078 (0.0078)	0.2659 (0.2900)
rs104894858	*LAMP2* (41,539bp)	3.8895	17083	Pulse rate	Hypertrophic cardiomyopathy	0.0026 (0.0026)	0.0081 (0.0081)	0.7492 (0.7492)
rs104894858	*LAMP2* (41,539bp)	3.8895	17083	Systolic blood pressure	Hypertrophic cardiomyopathy	0.0169 (0.0167)	0.0075 (0.0075)	0.0244* (0.0258*)
rs104894857	*LAMP2* (41,539bp)	0.0092	59	Diastolic blood pressure	Danon disease	-0.0738 (-0.0738)	0.1292 (0.1291)	0.5678 (0.5678)
rs104894857	*LAMP2* (41,539bp)	0.0092	59	Pulse rate	Danon disease	-0.2621 (-0.2626)	0.1337 (0.1337)	0.0500 (0.0495*)
rs104894857	*LAMP2* (41,539bp)	0.0092	59	Systolic blood pressure	Danon disease	-0.0665 (-0.0654)	0.1235 (0.1235)	0.5901 (0.5961)
rs137852527	*LAMP2* (41,539bp)	0.0031	20	Diastolic blood pressure	Danon disease	-0.0065 (-0.0129)	0.2122 (0.2121)	0.9756 (0.9516)
rs137852527	*LAMP2* (41,539bp)	0.0031	20	Pulse rate	Danon disease	-0.1391 (-0.1395)	0.2197 (0.2197)	0.5269 (0.5256)
rs137852527	*LAMP2* (41,539bp)	0.0031	20	Systolic blood pressure	Danon disease	-0.0641 (-0.0666)	0.2029 (0.2029)	0.7520 (0.7426)
rs9697983	*GLUD2* (2,333bp)	2.5052	12653	Hand grip strength, left	Parkinson disease, late-onset	-0.0050 (-0.0049)	0.0086 (0.0086)	0.5578 (0.5660)
rs9697983	*GLUD2* (2,333bp)	2.5052	12653	Hand grip strength, right	Parkinson disease, late-onset	-0.0166 (-0.0166)	0.0086 (0.0086)	0.0535 (0.0540)
rs9697983	*GLUD2* (2,333bp)	2.5052	12653	Mean time to correctly identify matches	Parkinson disease, late-onset	0.0179 (0.0171)	0.0087 (0.0087)	0.0386* (0.0489*)
rs122459146	*FHL1* (22,159bp)	0.0020	15	Hand grip strength, left	Myopathy, reducing body, X-linked, early-onset, severe	-0.0293 (-0.0064)	0.2426 (0.2422)	0.9039 (0.9790)
rs122459146	*FHL1* (22,159bp)	0.0020	15	Hand grip strength, right	Myopathy, reducing body, X-linked, early-onset, severe	-0.0807 (-0.0639)	0.2427 (0.2423)	0.7396 (0.7920)
rs122459146	*FHL1* (22,159bp)	0.0020	15	Whole body fat-free mass	Myopathy, reducing body, X-linked, early-onset, severe	-0.2101 (-0.1943)	0.1849 (0.1844)	0.2559 (0.2920)
rs104894760	*AVPR2* (4,636bp)	0.0024	18	Creatinine in urine	Nephrogenic diabetes insipidus	0.0479 (0.0451)	0.2279 (0.2279)	0.8334 (0.8432)
rs104894760	*AVPR2* (4,636bp)	0.0024	18	Cystatin C	Nephrogenic diabetes insipidus	0.0275 (0.0189)	0.2093 (0.2092)	0.8954 (0.9279)
rs104894760	*AVPR2* (4,636bp)	0.0024	18	Microalbumin in urine	Nephrogenic diabetes insipidus	0.0096 (0.0154)	0.4970 (0.4970)	0.9845 (0.9753)

rs104894760	*AVPR2* (4,636bp)	0.0024	18	Potassium in urine	Nephrogenic diabetes insipidus	-0.0441 (-0.0420)	0.2340 (0.2339)	0.8506 (0.8574)
rs104894760	*AVPR2* (4,636bp)	0.0024	18	Sodium in urine	Nephrogenic diabetes insipidus	0.0287 (0.0205)	0.2270 (0.2268)	0.8993 (0.9280)
rs104894747	*AVPR2* (4,636bp)	0.0027	20	Creatinine in urine	Nephrogenic diabetes insipidus	0.0690 (0.0676)	0.2279 (0.2279)	0.7620 (0.7667)
rs104894747	*AVPR2* (4,636bp)	0.0027	20	Cystatin C	Nephrogenic diabetes insipidus	-0.0823 (-0.0850)	0.2036 (0.2036)	0.6860 (0.6764)
rs104894747	*AVPR2* (4,636bp)	0.0027	20	Microalbumin in urine	Nephrogenic diabetes insipidus	-0.1108 (-0.1121)	0.3758 (0.3758)	0.7680 (0.7655)
rs104894747	*AVPR2* (4,636bp)	0.0027	20	Potassium in urine	Nephrogenic diabetes insipidus	0.2347 (0.2382)	0.2340 (0.2339)	0.3158 (0.3085)
rs104894747	*AVPR2* (4,636bp)	0.0027	20	Sodium in urine	Nephrogenic diabetes insipidus	0.2220 (0.2165)	0.2270 (0.2268)	0.3281 (0.3398)
rs104894757	*AVPR2* (4,636bp)	0.0018	12	Creatinine in urine	Nephrogenic diabetes insipidus	0.0220 (0.0083)	0.2784 (0.2783)	0.9371 (0.9762)
rs104894757	*AVPR2* (4,636bp)	0.0018	12	Cystatin C	Nephrogenic diabetes insipidus	0.1739 (0.1902)	0.2793 (0.2792)	0.5335 (0.4958)
rs104894757	*AVPR2* (4,636bp)	0.0018	12	Microalbumin in urine	Nephrogenic diabetes insipidus	-0.0161 (-0.0065)	0.6985 (0.6986)	0.9816 (0.9926)
rs104894757	*AVPR2* (4,636bp)	0.0018	12	Potassium in urine	Nephrogenic diabetes insipidus	-0.2096 (-0.2231)	0.2860 (0.2860)	0.4636 (0.4352)
rs104894757	*AVPR2* (4,636bp)	0.0018	12	Sodium in urine	Nephrogenic diabetes insipidus	0.3028 (0.2803)	0.2772 (0.2770)	0.2747 (0.3115)
rs104894748	*AVPR2* (4,636bp)	0.0015	11	Creatinine in urine	Nephrogenic diabetes insipidus	0.6176 (0.6107)	0.2916 (0.2915)	0.0341* (0.0361*)
rs104894748	*AVPR2* (4,636bp)	0.0015	11	Cystatin C	Nephrogenic diabetes insipidus	-0.2539 (-0.2500)	0.2808 (0.2807)	0.3659 (0.3731)
rs104894748	*AVPR2* (4,636bp)	0.0015	11	Microalbumin in urine	Nephrogenic diabetes insipidus	-0.0675 (-0.0668)	0.4448 (0.4448)	0.8793 (0.8807)
rs104894748	*AVPR2* (4,636bp)	0.0015	11	Potassium in urine	Nephrogenic diabetes insipidus	0.6753 (0.6707)	0.2993 (0.2992)	0.0241* (0.0250*)
rs104894748	*AVPR2* (4,636bp)	0.0015	11	Sodium in urine	Nephrogenic diabetes insipidus	0.1838 (0.1702)	0.3046 (0.3043)	0.5462 (0.5760)
rs104894753	*AVPR2* (4,636bp)	0.0020	15	Creatinine in urine	Nephrogenic diabetes insipidus	-0.4497 (-0.4398)	0.2497 (0.2496)	0.0717 (0.0781)
rs104894753	*AVPR2* (4,636bp)	0.0020	15	Cystatin C	Nephrogenic diabetes insipidus	-0.0457 (-0.0463)	0.2373 (0.2372)	0.8474 (0.8454)
rs104894753	*AVPR2* (4,636bp)	0.0020	15	Microalbumin in urine	Nephrogenic diabetes insipidus	-0.0249 (-0.0273)	0.4973 (0.4973)	0.9600 (0.9562)
rs104894753	*AVPR2* (4,636bp)	0.0020	15	Potassium in urine	Nephrogenic diabetes insipidus	-0.6541 (-0.6460)	0.2563 (0.2563)	0.0107* (0.0117*)
rs104894753	*AVPR2* (4,636bp)	0.0020	15	Sodium in urine	Nephrogenic diabetes insipidus	-0.5006 (-0.4872)	0.2487 (0.2484)	0.0441* (0.0499*)
rs104894941	*TAFAZZIN* (10,212bp)	0.0015	11	Diastolic blood pressure	3-Methylglutaconic aciduria type 2	0.2617 (0.2634)	0.2862 (0.2861)	0.3605 (0.3572)
rs104894941	*TAFAZZIN* (10,212bp)	0.0015	11	Pulse rate	3-Methylglutaconic aciduria type 2	-0.0906 (-0.0945)	0.2967 (0.2966)	0.7600 (0.7499)
rs104894941	*TAFAZZIN* (10,212bp)	0.0015	11	Systolic blood pressure	3-Methylglutaconic aciduria type 2	0.1454 (0.1481)	0.2739 (0.2738)	0.5955 (0.5886)
rs104894941	*TAFAZZIN* (10,212bp)	0.0015	11	White blood cell count	3-Methylglutaconic aciduria type 2	-0.3458 (-0.3371)	0.2947 (0.2946)	0.2405 (0.2526)
rs72554664	*G6PD* (16,182bp)	0.0093	14	Glucose	Glucose 6 phosphate dehydrogenase deficiency	0.1544 (0.1592)	0.3093 (0.3093)	0.6176 (0.6068)
rs72554664	*G6PD* (16,182bp)	0.0093	14	Glycated hemoglobin (HbA1c)	Glucose 6 phosphate dehydrogenase deficiency	-0.1476 (-0.1338)	0.2806 (0.2806)	0.5990 (0.6336)
rs72554664	*G6PD* (16,182bp)	0.0093	14	Hemoglobin concentration	Glucose 6 phosphate dehydrogenase deficiency	-0.0633 (-0.0580)	0.2715 (0.2714)	0.8156 (0.8309)
rs72554664	*G6PD* (16,182bp)	0.0093	14	Red blood cell count	Glucose 6 phosphate dehydrogenase deficiency	0.3926 (0.3988)	0.2704 (0.2703)	0.1465 (0.1402)
rs72554665	*G6PD* (16,182bp)	0.0157	47	Glucose	Glucose 6 phosphate dehydrogenase deficiency	-0.1455 (-0.1446)	0.1503 (0.1503)	0.3329 (0.3360)
rs72554665	*G6PD* (16,182bp)	0.0157	47	Glycated hemoglobin (HbA1c)	Glucose 6 phosphate dehydrogenase deficiency	-0.3522 (-0.3487)	0.1384 (0.1384)	0.0109* (0.0117*)
rs72554665	*G6PD* (16,182bp)	0.0157	47	Hemoglobin concentration	Glucose 6 phosphate dehydrogenase deficiency	0.0091 (0.0142)	0.1503 (0.1503)	0.9519 (0.9243)
rs72554665	*G6PD* (16,182bp)	0.0157	47	Red blood cell count	Glucose 6 phosphate dehydrogenase deficiency	-0.1123 (-0.1053)	0.1497 (0.1497)	0.4530 (0.4818)
rs5030869	*G6PD* (16,182bp)	0.0030	19	Glucose	Glucose 6 phosphate dehydrogenase deficiency	-0.1142 (-0.1152)	0.2518 (0.2519)	0.6501 (0.6475)

rs5030869	*G6PD* (16,182bp)	0.0030	19	Glycated hemoglobin (HbA1c)	Glucose 6 phosphate dehydrogenase deficiency	-0.4844 (-0.5076)	0.2132 (0.2133)	0.0230* (0.0173*)
rs5030869	*G6PD* (16,182bp)	0.0030	19	Hemoglobin concentration	Glucose 6 phosphate dehydrogenase deficiency	-0.5845 (-0.6136)	0.2304 (0.2305)	0.0111* (0.0078*)
rs5030869	*G6PD* (16,182bp)	0.0030	19	Red blood cell count	Glucose 6 phosphate dehydrogenase deficiency	-0.8769 (-0.9495)	0.2296 (0.2297)	0.0001** ($3.56x10^{-5}$**)
rs121912292	*DKC1* (14,934bp)	0.0022	15	Hemoglobin concentration	Dyskeratosis congenita, X-linked	0.1340 (0.1330)	0.2511 (0.2511)	0.5936 (0.5964)
rs121912292	*DKC1* (14,934bp)	0.0022	15	Platelet count	Dyskeratosis congenita, X-linked	0.0684 (0.0662)	0.2558 (0.2557)	0.7891 (0.7959)
rs121912292	*DKC1* (14,934bp)	0.0022	15	White blood cell count	Dyskeratosis congenita, X-linked	-0.1024 (-0.1021)	0.2511 (0.2511)	0.6834 (0.6841)
rs199422247	*DKC1* (14,934bp)	0.0021	14	Hemoglobin concentration	Dyskeratosis congenita, X-linked	0.1404 (0.1370)	0.2599 (0.2599)	0.5891 (0.5981)
rs199422247	*DKC1* (14,934bp)	0.0021	14	Platelet count	Dyskeratosis congenita, X-linked	0.0474 (0.0361)	0.2648 (0.2648)	0.8580 (0.8914)
rs199422247	*DKC1* (14,934bp)	0.0021	14	White blood cell count	Dyskeratosis congenita, X-linked	-0.2620 (-0.2557)	0.2600 (0.2599)	0.3136 (0.3252)
rs199422249	*DKC1* (14,934bp)	0.0015	10	Hemoglobin concentration	Dyskeratosis congenita, X-linked	-0.1556 (-0.1585)	0.3242 (0.3242)	0.6313 (0.6249)
rs199422249	*DKC1* (14,934bp)	0.0015	10	Platelet count	Dyskeratosis congenita, X-linked	-0.0295 (-0.0393)	0.3302 (0.3302)	0.9287 (0.9053)
rs199422249	*DKC1* (14,934bp)	0.0015	10	White blood cell count	Dyskeratosis congenita, X-linked	-0.2165 (-0.2055)	0.3242 (0.3241)	0.5042 (0.5260)
rs121912295	*DKC1* (14,934bp)	0.0018	12	Hemoglobin concentration	Dyskeratosis congenita, X-linked	0.1743 (0.1762)	0.2807 (0.2807)	0.5348 (0.5302)
rs121912295	*DKC1* (14,934bp)	0.0018	12	Platelet count	Dyskeratosis congenita, X-linked	-0.5308 (-0.5418)	0.2860 (0.2860)	0.0635 (0.0581)
rs121912295	*DKC1* (14,934bp)	0.0018	12	White blood cell count	Dyskeratosis congenita, X-linked	-0.1980 (-0.1917)	0.2808 (0.2807)	0.4807 (0.4946)
rs137852388	*F8* (191,153bp)	0.0308	164	Mean corpuscular volume	Hereditary factor VIII deficiency disease	0.04282 (0.0358)	0.0768 (0.0768)	0.5772 (0.6412)
rs137852388	*F8* (191,153bp)	0.0308	164	Red blood cell count	Hereditary factor VIII deficiency disease	-0.1039 (-0.0937)	0.0774 (0.0774)	0.1795 (0.2261)