University of Tartu

Faculty of Science and Technology

Institute of Molecular and Cell Biology

Chair of Biotechnology

Berit Kolk

# Characterization of the 16p11.2 600 kb BP4-BP5 CNVs in adult population cohort

Master's thesis

40 EAP

Supervisors: Katrin Männik, Ph.D.

Prof. Andres Metspalu, MD, Ph.D.

Tartu 2016

**Information sheet**

## Characterization of the 16p11.2 600 kb BP4-BP5 CNVs in adult population cohort

The 16p11.2 BP4-BP5 600 kb deletion and duplication carriers from clinical cohorts result in syndromes that affect neurodevelopment and anthropometric traits, but are also characterized by variable expressivity of associated phenotypic outcomes. The phenotype analysis showed that the 16p11.2 CNV carriers in the EGC UT adult population have characteristic features of 16p11.2 600 kb syndromes. Additionally, the adult cohort has common features, which are significantly recurrent comparing to EGC UT general population. Also, a new approach was used for finding genetic modifiers contributing to the variability of genomic disorders phenotypes. The whole-exome analysis found potential modifying substitutions for 4 adult 16p11.2 CNV carriers' specific features. According to our phenotypic and genotypic findings, it is important to conduct a detailed phenotypic assessment of individuals with particular genetic disorder and further investigate the exome or genome of the carriers to more precisely predict the severity and diverse outcomes of disease.

**Keywords:** CNV, 16p11.2, EGC UT, phenotype analysis, genetic modifiers

CERCS ERIALA: B220 Geneetika, tsütogeneetika

## Kromosoomi piirkonna 16p11.2 BP4-BP5 600 kb koopiaarvu variatsioonide iseloomustus täiskasvanute populatsioonis

Strukturaalsed muutused 16p11.2 BP4-BP5 genoomi piirkonnas väljenduvad 16p11.2 deletsiooni või duplikatsiooni sündroomina. Antud töö keskendus TÜ EGV täiskasvanute populatsiooni kohordist välja tulnud 600 kb koopiaarvu kandjatele. Läbi viidud põhjalikul fenotüübi analüüsil kinnitati 600 kb sündroomile iseloomulike sümptomite esinemise ka popuatsiooni kohordi kandjatel ning kirjeldati esmakordselt üldiste tervisehäirete sageduse tõusu (hüpotensiooni, psoriaas, funktsionaalset düspepsiat). Lisaks, viidi läbi üle-eksoomne analüüs, et leida võimalikke geneetilisi modifikaatoreid, mis vastutaks kandjate fenotüüpide varieeruvuse. Teostatud eksoomi uuringute andmetel leiti neljale 16p11.2 kandjale populatsioonist tema spetsiifilist fenotüüpi modifitseeriv geneetiline põhjendus.

**Märksõnad:** koopiaarvu variatsioonid, 16p11.2, Tartu Ülikool Eesti Geenivaramu, fenotüübi analüüs, geneetilised modifitseerijad

CERCS: B220 Genetic Engineering, Cytogenetics

**Table of contents**

**ABBREVIATIONS**

| | |
|---|---|
| 1000G | 1000 Genomes Project |
| ADHD | Attention deficit and hyperactivity disorder |
| ASD | Autism spectrum disorder |
| AVSD | Atrioventricular septal defect |
| BAF | B allele frequency |
| BMI | Body Mass Index |
| BP | Breakpoint |
| BQSR | Base Quality Score Recalibration |
| BWA | Burrows–Wheeler Aligner |
| CADD | Combined Annotation Dependent Depletion |
| CN | Copy numbers |
| CNV | Copy number variation |
| dbSNP | The Single Nucleotide Polymorphism Database |
| DEG | differentially expressed genes |
| DEL | Deletion |
| DUP | Duplication |
| ESP | Exome Sequencing Project |
| EX | Exome (ID) |
| ExAC | Exome Aggregation Consortium |
| FoSTeS | Fork Stalling and Template Switching |
| FSIQ | Full scale intelligence quotient |
| FSIQ | Full scale intelligence quotient |
| GATK | The Genome Analysis Toolkit |
| gDNA | genomic DNA |
| gVCF | genomic VCF |
| HC | Head circumference |
| HGNC | the HUGO Gene Nomenclature Committee |
| HTS | High-Throughput Screening |
| ICD-10 | International Classification of Diseases |
| ID | Intellectual disability |
| IQ | Intelligence quotient |
| LCR | low-copy repeats |
| LoF | Loss-of-function |
| LRR | Log R ratio |
| mtDNA | Mitochondrial DNA |
| NA | Not Available |
| NAHR | Non-Allelic Homologous Recombination |
| NHEJ | Non-homologous end joining |
| OR | Odd Ratio |
| PolyPhen | Polymorphism Phenotyping |
| QC | Quality Control |
| qPCR | quantitative Polymerase Chain Reaction |
| SIFT | Sorting Intolerant From Tolerant |
| SNP | Single-Nucleotide polymorphism |
| SV | Structural variation |
| Trials | Clinical Trials |
| VQSR | VariantRecalibrator |

| | |
|---|---|
| WES | Whole-exome sequencing |
| WHO | World Health Organization |

Abbreviations for genes described in this thesis are presented in the Supplementary Material Table S1.

## INTRODUCTION

The DNA copy number variations (CNVs) are major source for variation in humans and diseases. The rare intermediate-size and large non-recurrent and recurrent syndromic CNVs are frequent in the general population (10.5%). At the same time, they are also recognized to be one of the most typical causes of human disease.

This study focuses on the 600 kb deletions and duplications at the chromosome region 16p11.2, with observed frequencies 0.05% and 0.23% in population, respectively. These deletions and duplications result in mirroring phenotypes - autism spectrum disorders (ASD) vs. schizophrenia, obesity vs. underweight, macrocephaly vs. microcephaly, respectively. The 16p11.2 CNV carriers have also a broad range of features and the features vary widely in their expressivity and severity, the 600 kb CNVs are presented even in individuals with normal phenotypes. The clinical cohort's characteristics have been studied intensely, but the adult carriers phenotypes from general population have not been investigated before. At the moment, only a little is known, which secondary genetic factors influence the phenotypic variability linked to the pathogenic CNV.

In the current work, we aimed to analyze the 16p11.2 BP4-BP5 600 kb deletion and duplication carriers from EG CUT population cohort, whose phenotypes are highly variable and somewhat milder than in clinical cohort carriers. We conducted a detailed phenotypical analysis for the carriers' clinical information, to characterize the adult population features. In the second part of the study, we used exome-wide approach to find modifiers that contribute to the phenotypic variability in these disorders.

# 1 OVERVIEW OF LITERATURE

Chromosome structural variation (SV) is a common part of variation in the human genome and refers to abnormalities in chromosome structure (Weckselblatt and Rudd, 2015), that involves segments of DNA that are larger than 1 kilobase (kb) (Feuk et al., 2006) and are collectively termed as copy number variations. CNV refers to deletions or duplications; deletions are genomic losses and duplications are gains (Zarrei et al., 2015). Additionally, structural variants include inversions, alteration of the orientation of a specific genomic sequence (Alves et al., 2012), and translocations, exchanges of genomic material between two different chromosomes (Feuk et al., 2006; Scherer et al., 2007, Weckselblatt and Rudd, 2015). CNVs, inversions and translocations all result in changes in the physical arrangement of genes on chromosomes, but there are conceptual differences between the SVs; copy-neutral SV-like inversions and balanced translocations do not lead to changes in gene dosage. CNVs are the most frequent structural variations in the human genome and they contribute significantly to inter-individual genetic heterogeneity. It is identified that some CNVs have also an impact on human diseases. According to DECIPHER database (https://www.decipher.sanger.ac.uk/), as many as 70 recurrent syndromic CNVs have been linked with genomic disorders. Genomic disorders are caused by a structural alteration in the genome that might result in the complete loss or gain of a gene(s) sensitive to a dosage effect or, on the other hand, might disrupt the structural integrity of a gene (Lupski, 1998). For example, CNVs in chromosomal regions 1q21.1, 3q29, 15q11.2, 16p11.2, 16p13.1, 17q12 and 22q11.2 explain a significant proportion of risk for intellectual disability (ID; Girirajan et al., 2011), autism spectrum disorders (ASD) (Mefford et al., 2010), schizophrenia (McCarthy et al. 2009), epilepsy (Mefford et al., 2010), bipolar disease (Craddock et al., 2010; Grozeva et al., 2010) and attention deficit and hyperactivity disorder (ADHD) (Williams et al., 2010; Elia et al., 2010). Furthermore, CNV has been implicated in congenital birth defects (Mefford et al., 2007; Greenway et al., 2009) and common traits, such as cardiovascular diseases (Prakash et al., 2010; Norton et al., 2011), coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, type 1 diabetes and type 2 diabetes Xianfeng et al., 2010; The Wellcome Trust Case Control Consortium, 2010).

## 1.1 Formation mechanisms of genomic rearrangements of CNVs

Currently, four major mechanisms are known to be responsible for genomic rearrangements and CNV formation in humans i) Non-Allelic Homologous Recombination (NAHR) ii) Non-Homologous End-Joining (NHEJ) iii) Fork Stalling and Template Switching (FoSTeS) iiii) L1-mediated retrotransposition. NAHR is unequal homologous recombination between regions with high sequence resemblance but different genomic positions (Figure 1; Weckselblatt and Rudd, 2015). NAHR is a mechanism that leads to gross genome rearrangements (Parks et al., 2015). It occurs between two blocks of low-copy repeats (LCRs) (Bailey et al. 2002). LCR blocks are the typical substrates for NAHR because of their high level of sequence identity ≥97%, length ≥1 kb (Stankiewicz and Lupski, 2002). The products arising from these rearrangements can be duplication, deletion or inversion (Lupski, 1998; Stankiewicz and Lupski, 2002). In the human genome, there are many regions with high frequency of interspersed LCRs, which are possible places for NAHR-mediated rearrangements and are linked to genomic disorders (Stankiewicz and Lupski, 2002). One of the regions where NAHR-mediated CNV recurrently occurs is the 16p11.2 chromosome interval, which the current work concentrates on.

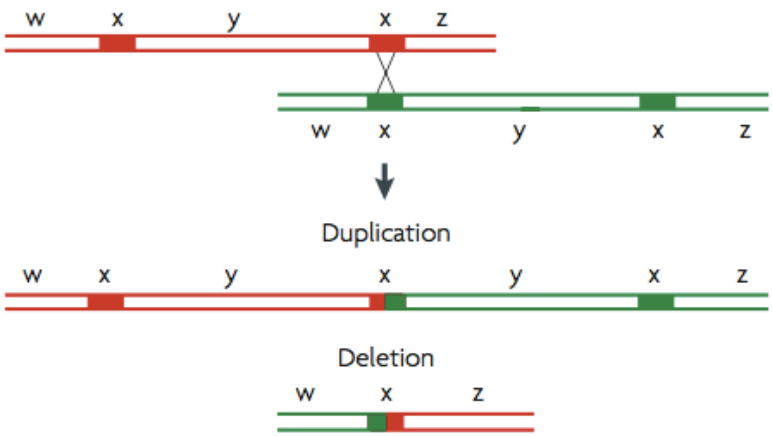


Figure 1. Mechanism of non-allelic homologous recombination (NAHR). If a direct repeat (x) is used as homology (upper panel) in a recombination repair event, NAHR will take place by uneven crossing over. In these circumstances the sequence between the repeats (y) will be reciprocally duplicated and deleted as a result of a crossover. When these products segregate from one another at the following cell division, both daughter cells have an alteration in the copy number (adapted from Hastings et al., 2009).

## 1.2 Structure of the human 16p11.2 interval

Short arm of human chromosome 16 has a complex genomic structure enriched in highly homologous and repetitive sequence blocks (LCRs). This region has been one of the most actively duplicated parts in human autosomes. The blocks act as a substrate for intra-chromosomal NAHR and make the part susceptible to recurrent structural rearrangements in several loci (Loftus et al., 1999; Martin et al., 2004; Johnson et al., 2006).

Five regions (breakpoint (BP) 1–5) on the short arm have been defined as „hotspots“ to genomic imbalances of clinical significance and are linked to neuropsychiatric phenotypes (Zufferey et al., 2012). BP4 and BP5 flank proximal 600 kb region in the 16p11.2 (Figure 2) which as a result of NAHR might be either deleted or duplicated (Zufferey et al., 2012)

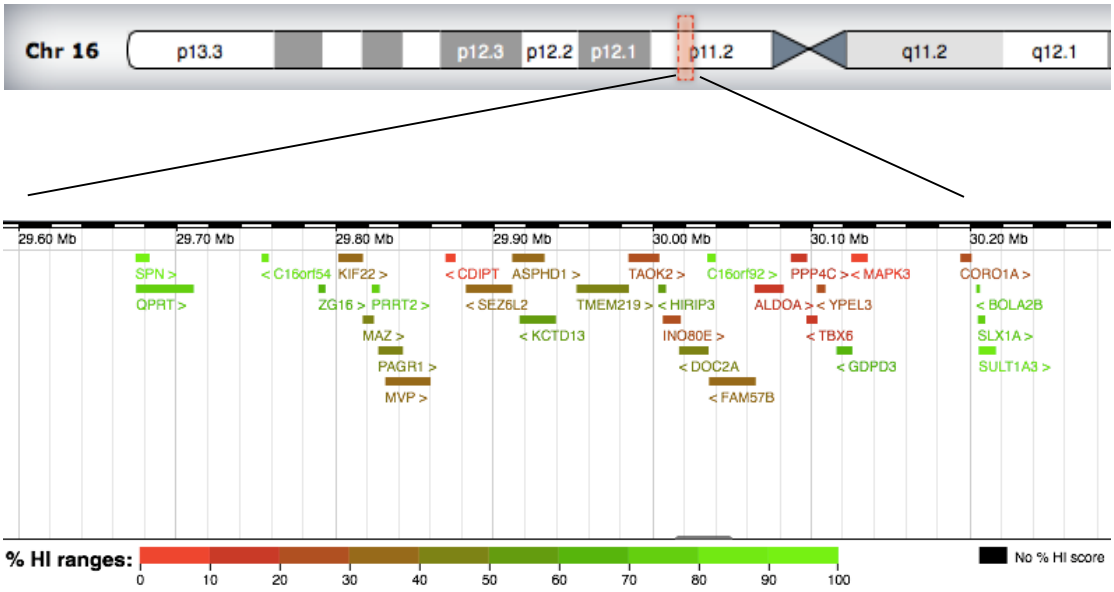


Figure 2. The human 16p11.2 BP4-BP5 600 kb chromosome region. Above is the short arm of chromosome 16. Highlighted by the red box is the 600 kb region flanked by BP4 and BP5 which is a study subject of the current work. In the middle are the unique genes encompassed by the imbalanced region. The colored scale shows the haploinsufficienies. Genomic coordinates are given according to the human genome build GRCh37/hg19. (DECIPHER database; https://www.decipher.sanger.ac.uk/)

### 1.2.1 The 16p11.2 BP4-BP5 600 kb CNV syndromes

The 16p11.2 600 kb between (29.5–30.1 Mb; GRCh37/hg19) deletion (MIM #611913) and duplication (MIM #614671) are two different conditions. Deletion and duplication in this region are associated with mirror phenotypes on body mass index (BMI) (Jacquemont et al., 2011; Zufferey et al., 2012), head circumference (Jacquemont et al., 2011) and brain volume (Qureshi et al., 2014; Maillard et al., 2014). Both 16p11.2 deletion and duplication are among the most frequent genetic causes of neurodevelopmental disorders (Cooper et al., 2011; Kaminsky et al., 2011; Hanson et al., 2015) and have been associated with ASD (Weiss et al., 2008), while only the duplication has been shown to be enriched in schizophrenia cohorts (McCarthy et al., 2009). In clinical ASD cohort, 16p11.2 600 kb CNVs account for approximately 1% of the cases (Weiss et al., 2008), and in schizophrenia cohort, 16p11.2 600 kb microduplication ranges from 0.2% to 0.6% (Giarolia et al., 2014).

16p11.2 deletion and duplication syndromes are considered to have an autosomal dominant inheritance pattern because a CNV in one copy of chromosome 16 is enough to cause the disorder. This type of CNVs occur mainly as random events during the formation of mature germ cells or in early fetal development and typically, affected people have no history of the disorder in their family (Fernandez, 2010; Rosenfeld, 2010). In approximately 80% of reported probands the 16p11.2 recurrent microdeletion is *de novo* (Fernandez, 2010). However, carriers can transmit the CNV to their children with a 50% of chance. So far, studies suggested that most of the duplications are inherited from one of the parents (Fernandez, 2010; Rosenfeld, 2010). Contrary to this, Duyzen et al. showed recently that maternal bias for *de novo* 16p11.2 600 kb deletion carriers is almost 90%. Furthermore, they observed that probands get considerably larger number of secondary deletions from mothers than from fathers. Although, there was no transmission bias observed for inherited 16p11.2 CNVs (Duyzend et al., 2016). There might be some differences in severity of the phenotype whether the CNV is inherited or is it *de novo* (newly arising in proband). Duyzend and Eichler have showed different impact on full scale intelligence quotient (FSIQ) - an inherited deletion carriers (and their family controls) have 8.33 points lower IQ than *de novo* carriers (Duyzend and Eichler, 2015). Also, families with inherited deletions have extra familial factors (environmental and genetic), which may have an impact on cognition (D'Angelo et al., 2015).

#### 1.2.1.1 16p11.2 deletion syndrome

The recurrent 16p11.2 microdeletion is rare, occurring in 1/2000 people in the general population; 1 in 100 people with autism (Weiss et al., 2008); and in around 1 out of 1000 people with a language delay or psychiatric disorder (Weiss et al., 2008; Bijlsma et al., 2009).

The most penetrant trait in deletion carriers is increased BMI (Figure 3). This feature is age-dependent – by the age of 7 years about 50% of the carries are obese (Zufferey et al., 2012). Among adult 16p11.2 deletion carriers obesity occurs with a penetrance of >70% (Zufferey et al., 2012). The deletion carriers explain 0.7% of all morbid obesity cases (BMI ≥40; p= 6.4 x $10^{-8}$; OR= 43.0) (Walters et al., 2010).

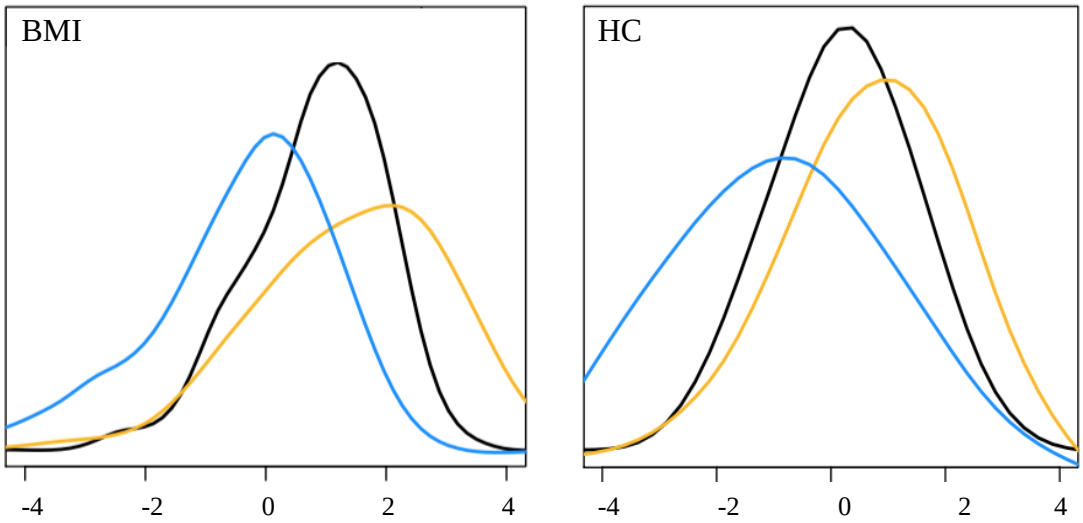


Figure 3. BMI (left) and HC (right) Z score density plots for duplication carriers, deletion carriers and for intra-familial controls. Blue line represents deletion carriers, yellow line duplication carries and black line is for intra-familial controls. BMI- body mass index; HC- head circumference. Adapted from D'Angelo et al., 2016.

The other characteristic traits of 16p11.2 microdeletion syndrome phenotype are developmental delay or learning difficulties, especially delay in speech and language development, which affects particularly expressive language, with relatively preserved receptive language (Miller et al., 2015).

Subset of 16p11.2 deletion carriers may have neurologic and motor impairments (e.g., epilepsy and paroxysmal dyskinesia syndrome). Babies may have low muscle tone,

and seizures are observed in approximately 20% of individuals with the microdeletion (Weiss et al., 2008; McCarthy et al., 2009; Shinawi et al., 2010).

Deletion carriers have tendency to increased head circumference (HC) (Figure 3; Shinawi et al., 2010; Walters et al., 2010; Maillard et al., 2015) as HC is highly correlated with brain volume (Qureshi et al., 2014; Maillard et al., 2014). Qureshi et al., suggested, that brain structure is influenced by 16p11.2 CNVs deletion is linked to unusually large brain volume (~9%; Qureshi et al., 2014). The observed association between the 16p11.2 microdeletion and increased head circumference is remarkable due to the fact that the deletion is linked also to ASD and developmental delay. Several studies have found increased head circumference in patients with autism (Butler et al., 2005; Dementieva et al., 2005; Lainhart et al., 2006; Fukumoto et al., 2008) leading to the implication that brain overgrowth in early life may be a key neurobiological mechanism in the disorder (McCarthy et al., 2009; Zufferey et al., 2012).

People with the 16p11.2 microdeletion should not be regarded primarily as a malformation syndrome since most of the people with this CNV do not have major birth defects and (Miller et al., 2015). However, more careful examination of deletion patients has revealed association with congenital abnormalities, for example scoliosis and other vertebral anomalies affect ~20% of carriers (Zufferey et al., 2012; Al-Kateb et al., 2014). Also, Müllerian duct anomalies are recurrently reported in the deletion carriers (Nik-Zainal et al., 2011; Sandbacka, et al., 2013).

Additionally, it is showed that 48% deletion carriers are described to be dominance of either left-hand or mixed-hand compared with 14% of non-carrier family members (Miller, et al. 2015; Hanson et al. 2015).

#### 1.2.1.2 16p11.2 duplication syndrome

16p11.2 microduplications have been estimated to occur in general population with a frequency of 0.09% (Männik., 2015).

While 12% of duplication probands had a relative risk of being underweight (BMI<18.5; Figure 3), only 4.6% (p= 0.010) of the carriers' relatives were clinically underweight (D'Angelo et al., 2016).

Microduplication is associated with multiple psychiatric phenotypes, including schizophrenia with 14.5-fold increased risk (McCarthy et al., 2009; Crespi et al., 2010), anxiety, depression, attention deficit, hyperactivity disorder (McCarthy et al., 2009), and autism (Crespi et al., 2010). Also, compared to non-carriers duplication probands have decreased HC; z-score average 1.2 points lower ($p$<0.001; Figure 3) and about 22.3% of the carriers are present with microcephaly. This change in HC is also associated with lower non-verbal IQ (D'Angelo et al., 2016).

Epilepsy occurs in 19.4% of the clinical duplication probands and is found to 2.2% of their carrier relatives (D'Angelo et al., 2016). Reinthale et al. showed 25-fold enrichment of Rolandic and atypical Rolandic epilepsy in duplication carriers comparing to the prevalence in the general population (0.05%; Jacquemont et al., 2011; Reinthale et al., 2014)

Besides variable psychiatric phenotypes and anthropometric differences, duplication carriers may have malformations, present in 16.7% of duplication probands (D'Angelo et al., 2016). Scoliosis, genital and cardiac malformations and are the most frequent malformation conditions in duplication carriers (D'Angelo et al., 2016).

Like in case of deletion, no characteristic facial dysmorphism are occurring in duplication carriers (D'Angelo et al., 2016).

#### 1.2.1.3 Phenotypic variability

The evidence from several studies (McCarthy et al. 2009; Qureshi et al., 2014; Maillard et al., 2014; D'Angelo et al., 2016; Duyzend et al., 2016) indicates that recurrent 16p11.2 deletions and duplications are associated with variable clinical outcomes and the deletion and duplication syndromes share same features with different prevalence (Table 1).

Table 1. Frequencies of different clinical traits among the 16p11.2 deletion and duplication carriers.

| Traits | Deletion (Prevalence) | Reference | Duplication | Reference |
|---|---|---|---|---|
| Autism spectrum disorder | 15% | Zufferey et al., 2012 | 20.1% | D'Angelo et al., 2016 |
| Intellectual disability | 37.6% | Zufferey et al., 2012 | 30.5% | D'Angelo et al., 2016 |
| Epilepsy | 24% | Zufferey et al., 2012 | 21.8% | D'Angelo et al., 2016 |
| Major malformations | 21.1% | D'Angelo et al., 2016 | 16.7% | D'Angelo et al., 2016 |

Many people with the duplication are never diagnosed because some have no related health problems, the phenotype spectrum is very wide and characteristic features have many other causes. Also, the features may differ between members of the same family and some people are apparently unaffected by their 16p11.2 CNV. Phenotypic heterogeneity is associated with not only 16p11.2 syndromes, but with many genomic disorders, e.g. 15q13 deletion syndrome (MIM #612001), 1q21 recurrent deletion (MIM #612474) and recurrent duplication syndrome (MIM #612475), and 16p13 deletion (MIM #610543) and duplication (MIM# 613458) syndromes (DECIPHER database; https://www.decipher.sanger.ac.uk/). Variability of the disorders may have many explanations, for example the deletion unmasking a mutation or functional polymorphism in a recessive gene on the non-deleted homolog allele and thus cause a more severe phenotype (Pebrel-Richard et al., 2014) and additional CNVs and single nucleotide genetics modifiers of multiple functionally relevant genes (Wu et al., 2015; Duyzend et al., 2016).

16p11.2 syndromes cognitive phenotypes may vary form overall normal cognition or some speech problems to severe developmental delay and intellectual disability (Bijlsma et al., 2009). FSIQ is significantly decreased in both deletion (22.2 points) and duplication (26.3 points) carriers comparing to intra-familial controls (Figure 4; D'Angelo et al., 2016). FSIQ is a figure, which varied largely on duplication carriers and has a significantly high difference comparing to deletion carriers (Figure 4; D'Angelo et al., 2016).

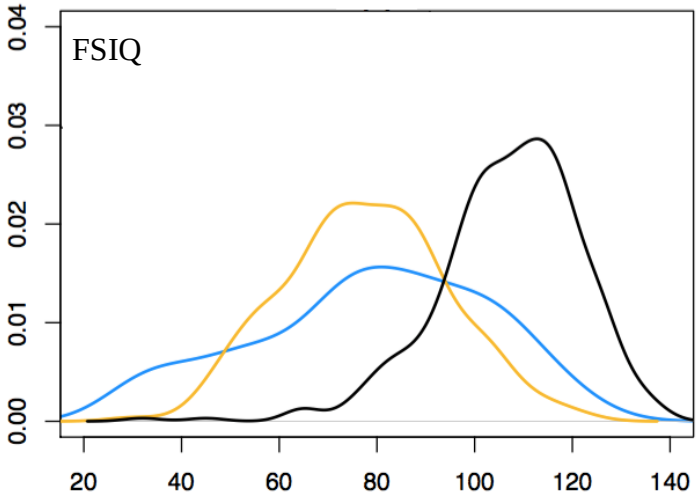

Figure 4. Distribution of FSIQ measures in clinical duplication carriers, deletion carriers and in intrafamilial non-carrier controls. Blue line represents the duplication carriers (n=154), yellow line represents deletions carriers (n=200) and black line represents familial controls (n=342). FSIQ- full scale intelligence quotient. Adapted from D'Angelo et al., 2016.

Duyzend et al. showed deletion and duplication carriers having similar percentage carrying an additional CNV (69% and 69.5%, respectively). By using FSIQ, they examined the connection between secondary-CNV burden and acuteness of phenotype. They found probands with extra CNVs tend to have (p= 0.03) lower FSIQ, comparing to probands with no other CNVs. The discovery is one explanation for the broad variation in IQ (Duyzend et al., 2016).

Also, Zufferey et al. tested the correlation between obesity and FSIQ and any behavioral trait, but no significant association was found (Zufferey et al., 2012).

The prevalence of congenital scoliosis in 16p11.2 deletion carriers is 7.5% (Wu et al., 2015). Development of scoliosis in individuals with 16p11.2 deletion is also described with additional genetic modifiers. For instance, 16p11.2 deletion, which removes one allele of the T-Box 6 gene (*TBX6;* MIM# 602427), in conjunction with a common hypomorphic allele of *TBX6* making up a risk haplotype for scoliosis. The effect of diminished *TBX6* dosage and an additional *TBX6* hypomorphic allele causes a further decrement in expression and will lead to congenital scoliosis.

In summary, the phenotypic spectrum of 16p11.2 CNVs remains to be fully characterized. Since different studies has shown that additional genetics factors play a

role in the observed severity and variability of the phenotype, further studies of additional genes and patients with a more severe phenotype than their transmitting parent are needed.

### 1.2.2 Potential candidate genes for 16p11.2 syndromes

Accodring to Decipher database, the BP4-BP5 CNV interval contains 32 genes (https://decipher.sanger.ac.uk/), but little is known about the role of most of these genes in the 16p11.2 syndrome formation. Genes, such as potassium channel tetramerization domain containing 13 *(KCTD13;* MIM# 608947), mitogen-activated protein kinase 3 (*MAPK3;* MIM# 601795), proline-rich transmembrane protein 2 (*PRRT2; MIM #*614386) and *TBX6* are well investigated and have demonstrated an association with the 16p11.2 syndromes.

*KCTD13* is a major leader of mirrored neuroanatomical phenotypes. In zebrafish embryos, the under-expression of the gene causes macrocephaly whereas the overexpression of human transcript in zebrafish embryos, the *KCTD13* influences microcephalic feature (Golzio et al., 2012). Furthermore, Golzio et al. conducted a pairwise over-expression of *KCTD13* with 16p11.2 region transcripts. They observed that *KCTD13* alone increases the expressivity of the head size 18%, but with *MVP* and *MAPK3* transcripts, the expressivity grows to 22% and 24%, respectively. The observation speculates that CNV carries have more acute phenotype, comparing to people with only heterozygous loss of function at *KCTD13* (Golzio et al., 2012).

It is considered that features, such as epilepsy, cognitive functioning, BMI, and HC are related to haploinsufficiency of distinct genes (Zufferey et al., 2012). For instance, mutations in *PRRT2* coding gene, which is mapping to the deleted interval, were identified in patients diagnosed with epilepsy and paroxysmal dyskinesia (Chen et al., 201; Crepel et al., 2011).

*TBX6*, that is located in 600 kb interval, is a candidate gene for vertebral malformations. For example, homozygous mutation in *TBX6* in mice showed rib and vertebral body anomalies (Watabe-Rudolph et al., 2002). Additionally, *TBX6* polymorphisms have been associated with congenital scoliosis in the Han population (Fei et al., 2010). Recently, Wu et al., (2015) found that 11% of congenital scoliosis

cases are explained by compound inheritance of a rare null mutation (mostly 16p11.2 deletion) and a common haplotype of T-C-A (rs3809624 (T/C), rs3809627 (C/A) and rs2289292) of *TBX6* (Wu et al., 2015).

Additionally, the pathways affected by the 600 kb BP4-BP5 16p11.2 deletions and duplications genome-wide are possibly linked to ciliary dysfunction (Migliavacca et al., 2015).

## 1.3 The Aim of the study

The purpose of this work is to characterize the 16p11.2 600kb BP4-BP5 carriers phenotype in adult general population and identify potential genetic modifiers contributing to the variable phenotypes using the Estonian population biobank samples.

# 2 METHODS

## 2.1 EGC UT population cohort characteristics

We used The Estonian Genome Center, the University of Tartu cohort (EGC UT; http://www.geenivaramu.ee/en) for the population-based study of 16p11.2 600kb CNV carriers. The current number of participants in the biobank is 51 880, which makes approximately 5% of the Estonian adult population (Leitsalu et al., 2015). From all donors, around 16 145 people have been genotyped.

For analyzing the common phenotypes, we compared deletion (n=6) and duplication (n=9) carriers to the EGC UT cohort (n=51 880). In both groups, the clinical data is assembled from the questionnaires and from national health registers. The baseline questionnaire of 16-modules covers more than 1 000 health- and lifestyle-related questions, and a uniformed report of clinical diagnoses according to the World Health Organization international classification of diseases (WHO ICD-10, http://www.who.int/classifications/icd). The data in the EGC UT database is continuously updated through follow-up interviews, as well as national electronic health databases and citizen registries.

Information modules we used from EGC UT questionnaire for common phenotype analysis included genealogy, education, diseases, health status of women (menstruation start/stop age, pregnancies, live births, miscarriages, abortions) objective data (age, height, weight, systole/diastole pressure), and disease code. Disease information extracted from the national electronic health registers and databases includes Estonian Causes of Death Registry, Estonian Cancer Registry, Database of the Tartu University Hospital, Database of the North Estonia Medical Centre in Tallinn, Estonian eHealth Foundation, as of 17$^{th}$ of September in 2015. Also, we used additional data collected at the recruitment visit of 16p11.2 CNV carriers.

Education levels were coded according to the Estonian education curriculum: 1 - less than primary; 2 - primary (currently 3 years); 3 – basic (currently 9 years, includes basic education for children with special needs); 4 – secondary (currently 12 years, includes vocational secondary); 5 – professional higher/college; 6 – university/academic degree; 7 – scientific degree (MD and PhD).

## 2.2 Phenotype analyses of the 16p11.2 CNV carriers

For conducting the phenotypic analysis, exactly the same data sources (above) were used for 16p11.2 carriers.

Two control cohorts were used in the phenotypic analyses, which differed in the number of individuals. First, the cohort of general population individuals used in analyzing the common phenotypes included EGC UT donors (n=51 880). Cohort contained 17 826 (34%) males and 34 054 (66%) females and age ranges from 18–103 years. The age, gender and education distributions of the EGC UT (n=51 880) cohort correspond to the Estonian general population. For analyzing the general characteristics (age, BMI, height, weight, education level), we used a randomly selected control cohort of fully genotyped EGC UT* donors (n=6 807). The EGC UT* cohort has the whole genotypic information and the cohort size is sufficient for analyzing the general parameters. The number of males in the cohort was 3 317 (49%) and females 3 490 (51%); the age range was 18–100 years and cohort's average educational level was 4.1 (range from 1–7; Männik et al., 2015).

The BP4-BP5 16p11.2 CNV carriers group consists of 15 individuals - deletion carriers (n=6) and duplication carriers (n=9). Carriers' cohort had 6 males and 9 females; the age range was 22–70 years; deletion carriers' mean educational level was 3.24 (range from 2–4) and the duplication carriers' 3.89 (range from 2–6). Further, we investigated if any of the 15 CNV carriers had immediate relatives (children, mother, father, siblings) who were also donors in the biobank. This resulted with 5 relatives as intra-familial controls from duplication carriers' families and 3 of deletion carriers' families. Among the family members, 4 individuals had already been on the SNP genotyping arrays and 4 had not been on one.

In the phenotypic analysis, we used R software for statistical analysis (https://www.r-project.org). From statistical tests, we used Student's t-test and Fisher's Exact Test (two-tailed) to compare the prevalence of different features in EGC UT CNV carriers to EGC UT controls.

## 2.3 Detection of the 16p11.2 CNV carriers by SNP genotyping arrays

The 16 145 individuals have been genotyped (passing QC) using three Illumina® SNP genotyping arrays i) Infinium Human 370CNV BeadChip ii) Infinium

HumanOmniExpress BeadChip (Illumina Inc., San Diego, CA; USA) and iii) HumanCore-24 BeadChip (Illumina Inc., San Diego, CA; USA). All genotyped samples were processed and the assay was performed according to a routine protocol provided by the manufacturer (Illumina Inc, http://www.illumina.com). Human 370CNV BeadChip is a genome-wide array designed especially for CNV detection studies. The total number of markers is 353 202 SNPs, which allows a median spacing of 5 kb. HumanOmniExpress BeadChip is a genome-wide platform with the total number of 713 014 markers and an advanced median spacing of 2.2 kb. HumanCore-24 BeadChip's (including PsychArray information) total number of markers is 306 670 genome-wide and the median spacing is 5.7 kb. The main difference from genome-wide platforms is that HumanCore chip contains exome-specific markers as well (240 000 markers).

### 2.4 Detection of CNVs

The generation of CNV calls from BeadChips was done by using PennCNV software (ver. June 2011). PennCNV is a method, based on Hidden Markov Model (HMM; Wang et al., 2007). The parameters for PennCNV are log R ratio (LRR) and B allele frequency (BAF). Beadstudio software (Illumina, San Diego, USA) was used for the genotype calling, signal intensity data normalization, and creating BAF and LRR at every SNP according to standard Illumina protocols. A standard SNP-based quality control was conducted for all the samples (SNP call rate >0.98). Quality control to the CNV calls removed calls with low confidence score (LOD <10, size < 20kb, or spanning < 10 probes). PennCNV inferred the copy number at every marker – CNV was called when 3 or more consecutive markers were changed. CNVs with copy number <2 were interpreted as deletions and CNVs with copy number >2 were defined as duplications. All detected 16p11.2 deletions and duplications were confirmed by quantitative PCR.

### 2.5 Confirmation of the CNV carrier status in identified 16p11.2 individuals and their immediate family members by Quantitative Polymerase Chain Reaction (qPCR)

All identified 16p11.2 BP4-BP5 CNV carriers were confirmed by quantitative polymerase chain reaction (qPCR). The quantitative PCR analysis consists of two steps: i) the empirical validation of assays and ii) the quantitative PCR analysis using

investigated samples. In empirical validation of primers amplification the efficiencies were measured based upon the generation of standard curves using genomic DNA (gDNA) dilution series. Melting curve analysis was used to check the specificity of the PCR reactions. 14 primer pairs (2 reference primers) were designed for the 16p11.2 600kb region according to the stringent parameters (primers pairs used are shown in Supplementary Material Table S2). Tools, used for primer design were Primer3 (http://www.bioinfo.ut.ee/primer3-0.4.0/) and Genometester (http://www.bioinfo.ut.ee/genometester/).

Reactions were set up in triplicate using 2.5 ng of genomic DNA. PCR reaction mix contained 2 µl qPCR mix (Solis BioDyne 5x HOT FirePol® EvaGreen®); 0.5 µl forward primer (10 pmol); 0.5 µl reverse primer (10 pmol); 4.5 µl nuclease-free water.

The experimetal test was followed by DNA copy number estimation with Biogazelle qBasePlus (https://www.biogazelle.com/qbaseplus). The analyze uses calibrator samples with known copy number and samples of interest with unknown copy number. For normal diploid copy (CN=2), there are assigned the lower boundary threshold and the upper boundary threshold. These thresholds are used for detecting deletions and duplications and are by default arranged to 1.414 (average of 1 and 2 copies) and 2.449 (average of 2 and 3 copies).

### 2.6 Whole-exome sequencing for detection of modifying variants

Whole-exome sequencing (WES) of 23 individuals from the 16p11.2 families was performed in the Center for Integrative Genomics (CIG) Genotyping Core Facility using the SureSelect Exome V5 kit (Agilent Technologies Inc., Santa Clara, CA, USA) and Illumina HiSeq2000 platform (Illumina Inc.). Genomic DNA was extracted from whole-blood according to the manufacturer's protocol (Gentra® Puregene® Handbook; Third Edition; 2011; Germany), using the Gentra Puregene Blood Kit (Gentra® Puregene®, Hilden, Germany). Evaluation of the purity of DNA extraction was assessed with The NanoDrop® ND-1000 Spectrophotometers (Thermo Scientific), Qubit 2.0 fluorometer (Invitrogen, Life technologies) and by gel electrophoresis. Sequencing libraries were constructed from 3 µg high quality genomic DNA.

### 2.6.1 Variant calling and quality control

The Genome Analysis Toolkit (GATK; https://www.broadinstitute.org/gatk/) was used following the GATK Best Practices protocol (https://www.broadinstitute.org/gatk/guide/best-practices.php) to carry out variant detection analysis and quality control (QC) on high-throughput sequencing data. Shortly, for the raw data analysis, the following pipeline was used. Sequence reads cleaning was done with Fastq-mcf (https://www.broadinstitute.org/gatk/guide/best-practices; Finseth and Harrison, 2014) 100 bp paired-end reads were aligned to the human reference genome version GRCh37/hg19 using Burrows–Wheeler Aligner (BWA) program (BWA-MEM algorithm: https://www.broadinstitute.org/gatk/guide/best-practices; Li and Durdin, 2009). Duplicate reads were removed using Picard (Picard MarkDuplicates; https://www.broadinstitute.org/gatk/guide/best-practices; Xu et al., 2012) and base quality score recalibration performed using GATK software. Single nucleotide variants (SNVs) and indels were identified using the GATK Unified Genotyper (https://www.broadinstitute.org/gatk/guide/best-practices; DePristo et al., 2011) and base quality score recalibration (BQSR; https://www.broadinstitute.org/gatk/guide/best-practices; Li et al., 2004). GATK Haplotype Caller (https://www.broadinstitute.org/gatk/guide/best-practices; Logan et al., 2015) to generate multi-sample genomic VCFs (gVCFs) files. From the gVCF cohort file, GATK Genotype gVCFs walker (https://www.broadinstitute.org/gatk/guide/best-practices) generates a group of indel calls and raw SNP. The latter will go through a variant quality score recalibration GATK VariantRecalibrator (VQSR; https://www.broadinstitute.org/gatk/guide/best-practices; Pirooznia et al., 2014).

The raw data analysis and variant calling used in this work was done with the support of Swiss Institute of Bioinformatics Vital-IT platform (http://www.vital-it.ch).

### 2.6.2 Variant annotation and interpretation

All 23 sequenced samples met established QC parameters and were used for further analysis. Average mean bait coverage of the samples was 136.6 (range from 107.88–167.95), and the percentage of all target bases with coverage greater than 30x was 97% (min: 95.15%, max: 98.03%).

Kinship analysis was done by computing correlation between samples and plotting a correlation plot using heterogenity variance function. First, variants with "PASS" and dbSNP ID were selected. Among the latter, only heterozygous variants were kept (variants are alt/ref or ref/ref for all samples in VCFs). Multi-allelic, homozygous and variants with missing calls were removed. That gave 102 000 variants for the dataset. From this set, correlation matrix was computed and clusterings were plotted. In the EGC UT cohort (carriers, intra-familial controls) all the family members grouped as the relationship expected.

For variant annotation and interpretation three different platforms were used:

i) Custom EGC UT annotation tool for diagnostic exome-sequencing. A Perl based custom script for diagnostic exome-sequencing (Kals et al., unpublished). The platform uses multi-sample vcf as an input file. The annotation information consist of variant parameters (e.g. chrom nr, posititsion, variant ID, reference, alternative, GeneID), pathogenicity scores (e.g. C-score; phyloP score; SIFT; PolyPhen), individual characteristics (e.g. genotype, count of reference/non-reference allele; read depth), population frequency parameters for identified alleles (1000G; count of population specific genotypes, based on 162 whole-exome and 96 whole-genome EST samples; ExAC data set; NHLBI ESP). The tool encompasses the GenomeTrax datasets, which show variant or gene associations with different diseases and phenotypes (e.g. HGMD inherited disease mutations, http://www.hgmd.cf.ac.uk/ac/index.php; COSMIC somatic disease mutations, http://www.cancer.sanger.ac.uk/cosmic; ClinVar, https://www.clinicalgenome.org/data-sharing/clinvar/; GWAS Catalogue https://www.ebi.ac.uk/gwas/; OMIM, http://www.omim.org).

(ii) Saphetor Web Portal (Saphetor AS, Lausanne, Switzerland; http://www.saphetor.com) is a genome-scale genetic variant analyzing tool, which uses single-sample vcf as an input file. The annotation and interpretation platform has combined over 20 leading databases (e.g. HGMD, ClinVar, ExAC, 1000 Genomes, dbSNP, Mutation Taster, SIFT, Pubmed, RefSeq, Unigene, Clinical Trials, HPO, Orphanet). Saphetor gives a possibility to make filter sets choosing from allele frequencies, predicted pathogenicity scores (SIFT score range, SIFT score assessment,

PolyPhen HVAR score assessment, PolyPhen HDIV score assessment), dbSNP database version, ClinVar class, disease associations, chromosome region, pathogenicity class, zygosity, function, call status, variant type or gene list. The Gene lists are possible to make by yourself (list of new genes) or use already existing gene list from phenotype. The candidate gene lists, used in the current work with aforementioned tools are found in the Supplementary.

(iii)Varapp Browser is a web-based tool for diagnostic WES analysis developed for the Lausanne University Hospital (CHUV) clinical research platform (Pradervand and Delafontaine, unpublished; https://www.github.com/varapp). The platform uses vcf as an input file. For data analyzing, the browser contains filter categories, such as scenario (variant is dominant; recessive; *de novo*; compound heterozygous; x linked), allele frequency, quality parameters (e.g. quality filter: PASS, quality score range), impact factors (substitution impact in the protein, such as high, medium and low) and predicted pathogenicity of a variant (PolyPhen; SIFT; CADD score predictions).

#### 2.6.2.1 Used pathogenicity prediction score levels

Pathogenicity prediction tools used in exome analysis were i) scaled the combined annotation dependent depletion (CADD) score, ii) sorting intolerant from tolerant (SIFT) score and iii) polymorphism phenotyping (PolyPhen).

i) The principle of scaled CADD score was, presented as a "meta-annotation" tool, which sorts and aggregates the information from other tools and produces a scaled score that reflects the deleteriousness of a particular variant (Kircher et al. 2014; Bandaru et al., 2015). A scaled CADD score values ≥10 are predicted to be the 10% most deleterious substitutions and score of 20 means that a variant is amongst the top 1% of deleterious variants in the human genome and so on (http://www.cadd.gs.washington.edu/info).

ii) The SIFT program gives scores that evaluate tolerance of amino-acid changes in protein function. SIFT scores range from 0 to 1. SIFT p-values below 0.05 indicate that the change is likely deleterious and scores closer to 0 show more damaging effect (Ng and Henikoffa, 2003).

iii) PolyPhen predicts the possible impact of an amino-acid substitution on the structure and function of a human protein using physical and comparative considerations, comprising interference with ligand binding sites (Adzhubei et al., 2010). It provides one of three predictions for non-synonymous variants - benign, possibly damaging, or probably damaging. PolyPhen score is from 0 (tolerated) – 1 (damaging). Scores, starting from 0.15, show possibly damaging effect (Adzhubei et al., 2010).

#### 2.6.2.2 Impact filters used in the current work

The EGC UT annotation tool we used for loss-of-function variant detection and the count of Estonian population specific genotypes, based on 162 whole-exome and 96 whole-genome Estonian samples.

The basic filters used in the current work using VarApp tool were quality filter "PASS", ExAC allele frequency (<5%), high impact and pathogenicity filters (scaled CADD score; PolyPhen; SIFT), zygosity filter (homozygous; heterozygous) When one of the observed scores had damaging prediction (scaled CADD score was ≥ 10, PolyPhen score 0.15–1.0 or SIFT scores 0–0.05), we considered the variants to start

having modifying effects on the phenotype. Used high impact factors are assumed to have a disruptive impact in the protein, probably causing protein truncation, loss of function or triggering nonsense mediated decay. This category includes variants such as frame-shift, splice acceptor, splice-donor, start-lost, stop-gained, and stop-lost. (http://www.varapp.vital-it.ch/#/)

In the Saphetor platform, we used ExAC frequency <5%, SIFT damaging assessment, PolyPhen HDIV (evaluating rare alleles at loci; http://www.annovar.openbioinformatics.org) and HVAR (used for diagnostics of Mendelian diseases; http://www.annovar.openbioinformatics.org/) database assessments, ClinVar class (likely pathogenic; pathogenic; other) prediction tool, coding impact factors (such as exon deletion; frame-shift; missense and stop-loss) and zygosity filter (homozygous; heterozygous). Analyzing the ciliopathy-associated genes and differentially expressed genes with Saphetor, we concentered on the variants without assigned rs number and no carriers in ExAC, dbSNP or 1000G databases.

# 3 RESULTS

In the genotyped EGC UT sample of 16 145 individuals, we identified 6 carriers of the 16p11.2 BP4-BP5 deletion and 9 carriers of reciprocal duplications. Showing respective prevalence of 1/2690 and 1/1794 in Estonian population. Except for one father and daughter pair with the duplication, all other analyzed family members (n=8) were non-carriers of the 16p11.2 CNVs. Altogether, we had 23 individuals from 14 families. The characteristic phenotype information of all analyzed individuals is represented in the Supplementary Materials Table S3.

## 3.1 Phenotype traits of the EGC UT 16p11.2 600kb BP4-BP5 CNV carriers

Table 2 provides summary statistics of gender, age, height, weight, BMI for the EGC UT 16p11.2 CNV carriers and their family members.

Table 2. General information of the carrier's cohort.

| General information | Gender | Mean age | Mean height (cm) | Mean weight (kg) | Mean BMI ($kg/m^2$) |
|---|---|---|---|---|---|
| Deletions (n=6) | 3M/3F | 39.3 (22–57y) | 166.2 (154.0–182.0) M:172 (163–182) F:159.7 (158–162) | M: 88.3 (80–105) F: 98.7 (87–115) | 37.1 (24.2–51.9) |
| Duplications (n=9) | 3M/6F | 44.2 (22–70y) | 170.8 (158.0–186.0) M:181.3 (176–186) F:165.5 (158–174) | M: 83.3 (80–90) F: 56.3 (49–65) | 22.8 (18.6–25.5) |
| Controls (n=8) | 4M/4F | 49.1 (22–72y) | M: 177.8 (173.0–184.0) F:165,5 (154.0–172.0) | M: 81 (73.0–110) F: 73.8 (61–92) | M: 25.8 (19.5–35.9) F: 27.4 (20.6–38.8) |

M=male; F=females; y=years; cm=centimeter; kg=kilogram; $m^2$=square meter

Intellectual disability had been diagnosed in 1 EGC UT duplication carrier (EX99) and three deletion carriers (EX107; EX110; EX119) (50%; p= 3.105e-06; OR= 184.3). In 16p11.2 clinical cohort, ID frequency has been shown to be 30.5% (D'Angelo et al., 2016) and 20.1% (Zufferey et al., 2012), respectively.

EGC UT* mean educational degree was 4.1 (18.7%, 1 275/6 807 of the carriers were

with basic or lower education level). Intra-familial controls mean education level was 3.6 (50%, 4 out of 8 carriers were with basic or lower education level). When we examined duplication (n=5) and deletion (n=3) intra-familial controls separately, we observed mean education of 4 for duplication controls and 3 for deletion controls. Conducting a statistical test for intra-familial controls and EGC UT cohort, we observed a significant difference between the cohorts (p= 0.04556; OR= 4.34). Deletion carriers' mean education was 3.24 (66.7%, 4/6 carriers were with basic or lower education level) and a significant difference occurred between deletion carriers and EGC UT* cohort educational level (p= 0.01343; OR= 8.67). Duplication carriers' mean education was 3.89 (33.3%, 3/9 carriers were with basic or lower education level). Due to small sample size, we did not compare gender differences in ID prevalence and educational levels.

EGC UT 16p11.2 CNV carriers had typical characteristics to 600 kb pathologies, like obesity and underweight. Among randomly selected population individuals (n=6 807) mean BMI was 26.7 and among intra-familial controls 26.6. Deletion carriers mean BMI was 37.1 (ranging from 24.15 to 51.9; p=0.0493) and duplication 22.8 (ranging from 18.6 to 25.5; p=0.0259). Five out of 6 deletion carriers (83%) were obese - significantly more frequent than in EGC UT* cohort (1 620 out of 6 807) (p= 0.003; OR= 16.0) but not in deletion carriers' intra-familial controls (1 out of 3; p= 0.226; OR= 7.11). All duplication carriers were under the population average BMI but none of them underweight BMI ≤18. EGC UT* controls had 81 out of 6 807 individuals (1.2%) with BMI ≤18 (Figure 5).

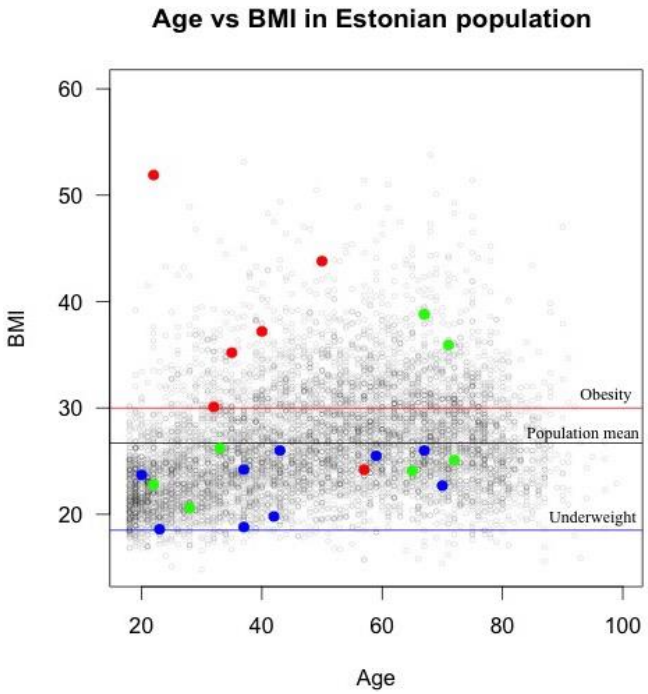


Figure 5. A scatterplot representing Estonian population BMI against age. Red dots represent the deletion, blue dots the duplication carriers and green dots are the intra-familial controls. The black line shows the EGC UT cohort average BMI (26.7); BMI ≥30 is defined as obesity (red line) and BMI ≤18 (blue line) is defined as clinical underweight according to the WHO (http://www.who.int/mediacentre/factsheets/fs311/en/).

EGC UT male population mean was 177.5cm (154–206) female population mean 164.4cm (140–194). Height of intra-familial males was average 177.8cm (173–184) and females 165.5cm (154–172), which corresponded to EGC UT* mean heights. Comparison of the height in Estonian population cohort (EGC UT*) versus 16p11.2 CNV carriers (Figure 6) showed that male duplication carriers were not taller (mean 181.3cm; range 176–186; p= 0.367) and female deletion carriers were not significantly shorter than non-carriers (mean 159.7cm; range 158–162; p=0.2184). Although the heights comparison between carriers and EGC UT* controls did not observe statistically significant difference, this is something to look in a bigger carriers' cohort. So far, no information has been published on the correlation between the 16p11.2 CNVs and height. Since the EGC UT 16p11.2 cohort was too small to make conclusion, we analyzed the adult data from the European 16p11.2 clinical cohort (deletion carriers n=78, duplication carriers n=79). The mean height of intra-familial female deletion controls were 164cm (150–176) and intra-familial male deletion controls 175.1cm (159–185.6). Mean height among deletion carriers was

159.1cm for females (149–170; p=0.002) and 171cm (155–200; p=0.129) for males. Mean height of intra-familial controls (n=57) of female duplication carriers was 165.5cm (154–167.5) and males 174.7 in males (178–182.8). The mean height in female duplication carriers was 165.3cm (150–179; p=0.923) and in male duplication 177.7cm (159–193: p= 0.304). We observe a significant difference between clinical female deletion carriers and the intra-familial controls; females in deletion carriers' clinical cohort tend to be shorter than the intra-familial deletion controls.

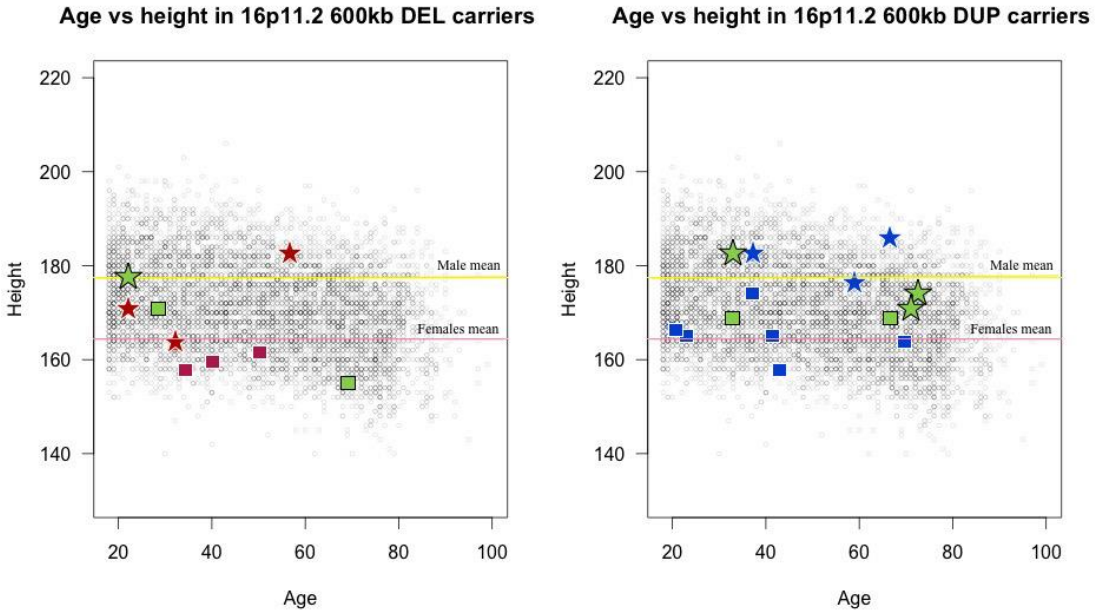


Figure 6. Comparison of the height in Estonian population cohort (EGC UT*) versus 16p11.2 CNV carriers. On the left are presented the 16p11.2 deletion carriers from the population cohort (red stars are male deletion carriers; red squares are female deletion carriers). On the right are presented the 16p11.2 duplication carriers from the population cohort (blue stars are male duplication carriers; blue squares are female duplication carriers). Green stars and squares are showing intra-familial male and female controls, respectively (on the left are deletion carriers intra-familial controls; on the right are duplication carriers' intra-familial controls). The yellow lines show Estonian male population mean height (177.5cm) and the pink lines show Estonian female mean height (164.4cm).

Among other clinical traits previously linked to the 16p11.2 CNVs, neuropsychiatric phenotypes were occurring in 7 (46.7%) EG CUT 16p11.2 CNV carriers. More specifically in duplication carriers, there were 5 (55.6%) individuals with neuropsychiatric phenotypes: schizophrenia (EX118), epilepsy (EX100), neurotic disorder (EX105), and anxiety disorder (EX99; EX113). In deletion carriers, there

were 2 individuals (2 out of 6; 33.3%): 1 with epilepsy (EX101) and 1 with neurotic disorder (EX110).

We evaluated the significances of neuropsychiatric, mental and behavioral disorders in EGC UT carriers' cohort (Table 3). Epilepsy and schizophrenia - although recurrent among 16p11.2 600kb CNV carriers - did not show higher prevalence in the EGC UT 16p11.2 CNV carriers. ID presented in 3 deletion (50%; p= 3.105e-06; OR= 184.3) and in 1 duplication carrier (p= 0.047; OR= 23.18) showed significant difference comparing to the EGC UT controls.

Table 3. Frequency of neuropsychiatric, mental and behavioral disorders in the EGC UT 16p11.2 deletion and duplication carriers.

| TRAITS (WHO ICD-10 code) | DEL (n=6) | DUP (n=9) | EGC UT cohort (n= 51 880) | P value DEL | OR (Cl95%) DEL | P value DUP | OR (Cl95%) DUP |
|---|---|---|---|---|---|---|---|
| Epilepsy (G40) | n=1 16.7% | n=1 11.1% | n=1 448 2.7% | 0.156 | 6.96 (0.15–62.32) | 0.225 | 4.35 (0.1–32.52) |
| Schizophrenia, schizotypal and delusional disorders (F20-F29) | n=0 | n=1 11.1% | n=783 0.15% | - | - | 0.128 | 8.16 (0.18–60.97) |
| Intellectual disability (F70-F79) | n=3 50% | n=1 11.1% | n=278 0.5% | 3.105e-06*** | 184.30 (24.70–1408.51) | 0.047* | 23.18 (0.52–174.18) |
| Mental and behavioral Disorders (F00-F99) | n=3 50% | n=5 55.6% | n=25 461 49% | 1 | 1.04 (0.14–7.75) | 0.749 | 1.30 (0.28–6.54) |

Fisher's Exact Test; p value: 0.05 = *, 0.01 = **, 0.001 = ***

The results on anthropometric and cognitive measures confirmed that the phenotypes of Estonian adult 16p11.2 CNV carriers were in agreement with the characteristics seen before in the 16p11.2 clinical studies (see Literature overview for more details).

The 16p11.2 rearrangements have also been associated with specific congenital abnormalities (Sampson et al., 2010; Wu et al., 2015). Similarly to observations in the 16p11.2 clinical cohorts, three deletion carriers (50%) were identified to have congenital vertebral malformations – scoliosis (EX110) or otseochondrosis (EX101; EX102). Additionally, cohort included two individuals (EX100, EX119) with

malformations of female genital organs originating from the Müllerian duct and one of them had also cardiac septa malformation (EX100). All malformations occurring in the EGC UT 16p11.2 CNV cohort corresponds to the ones described in the 16p11.2 clinical cohorts.

In addition to analysis of already known phenotype traits, the study aimed to uncover new traits in adult 16p11.2 CNV carriers that have so far remained unrecognized due to their common nature or age-dependent onset.

Common features such as hypotension, psoriasis, diseases of oesophagus, stomach and duodenum, and functional dyspepsia were significant in comparison with EGC UT controls: hypotension in duplication carriers (p= 0.008; OR= 18.52), psoriasis in deletion (p= 0.033; OR= 35.36) and duplication carriers (p= 0.047; OR=23.18) and functional dyspepsia among deletion carriers (p= 0.036; OR= 6.41). Numbers of diagnosed individuals and prevalence in the 16p11.2 CNV carriers group and the EGC UT cohort are given in the Table 4.

Table 4. Common features recurrent in EGC UT carriers' cohort.

| TRAITS (WHO ICD-10 code) | DEL (n=6) | DUP (n=9) | EGC UT cohort (n= 51 880) | P value DEL | OR (Cl95%) | P value DUP | OR (Cl95%) |
|---|---|---|---|---|---|---|---|
| Hypotension (I95) | n=0 | n=2 22.2% | n=788 1.5% | - | - | 0.008** | 18.52 (1.87–97.30) |
| Psoriasis (L40) | n=1 16.7% | n=1 11.1% | n=278 0.5% | 0.033* | 35.36 (0.75–315.03) | 0.047* | 23.18 (0.52–174.18) |
| Diseases of oesophagus, stomach and duodenum (K20-K31) | n=5 83.3% | n=5 55.6% | n=23 934 46% | 0.110 | 5.49 (0.61–259.25) | 0.745 | 1.37 (0.30–6.92) |
| Functional dyspepsia (K30) | n=3 50% | n=2 22.2% | n=7 004 13.5% | 0.036* | 6.41 (0.86–47.80) | 0.348 | 1.83 (0.19–9.62) |

Fisher's Exact Test; p value: 0.05 = *, 0.01 = **, 0.001 = ***

We also looked at the female specific phenotypes and found traits, which tend to have higher prevalence in the 16p11.2 CNV carriers compared to controls. Due to the small sample size after gender separation we analyzed female 16p11.2 CNV carriers as one group. The recurrent traits of 16p11.2 600 kb CNV female cohort were absent, scanty and rare menstruation (p= 0.037; OR= 5.24), infertility (p= 0.067; OR= 5.77), non-

inflammatory disorders of ovary, fallopian tube and broad ligament (p= 0.046; OR= 4.81), and cystitis (p= 0.039; OR= 5.0). Infections with a predominantly sexual mode of transmission had not significant trend, with an increased risk (p= 0.084; OR= 3.32). Numbers of diagnosed individuals and prevalence in the 16p11.2 CNV carriers group and the EGC UT cohort are given in the Table 5.

Table 5. Female specific features in deletion and duplication carriers and in Estonian female population cohort.

| TRAITS (WHO ICD-10 code) | DEL (n=3) | DUP (n=6) | EGC UT female cohort (n=34 054) | P value DEL | OR (Cl95%) | P value DUP | OR (Cl95%) | P value DUP& DEL | OR (Cl95%) |
|---|---|---|---|---|---|---|---|---|---|
| Absent, scanty and rare menstruation (N91) | n=1 33.3% | n=2 33.3% | n=2966 8.7% | 0.239 | 5.24 (0.09-100.73) | 0.090 | 5.24 (0.47-36.58) | 0.037* | 5.24 (0.85-24.55) |
| Infertility (N97) | n=1 33.3% | n=1 16.7% | n=1 861 5.5% | 0.155 | 8.65 (0.15-166.42) | 0.286 | 3.46 (0.07-30.94) | 0.067 | 5.77 (0.57-32.28) |
| Infections with a predominantly sexual mode of transmission (A50-A64) | n=1 33.3% | n=3 50% | n=6 763 19.9% | 0.485 | 2.02 (0.03-38.72) | 0.100 | 4.04 (0.54-30.14) | 0.084 | 3.23 (0.64-15.0) |
| Non-inflammatory disorders of ovary, fallopian tube and broad ligament (N83) | n=1 33.3% | n=2 33.3% | n=3 207 9.4% | 0.257 | 4.81 (0.08-92.56) | 0.103 | 4.81 (0.43-33.56) | 0.046* | 4.81 (0.78-22.53 |
| Cystitis (N30) | n=1 33.3% | n=3 50% | n=14 077 41.3% | 1 | 0.71 (0.01-13.63) | 0.696 | 1.42 (0.19-10.59) | 0.039* | 5.0 (0.95-49.0) |

Fisher's Exact Test; p value: 0.05 = *, 0.01 = **, 0.001 = ***

## 3.2 Whole-exome sequencing in a cohort of 16p11.2 families

### 3.2.1 Quality control and variant calling

In exome sequencing, the observed average transition-transversion ratio (Ti/Tv) in SNPs was 2.37 (range 2.33–2.4), sample contamination values were average 0.0369 (range 0.018–0.154). The mean number of novel variants was 1621.4 per individual (range 1504–1881). Ti/Tv shows the number of transition mutations (pyrimidine-pyrimidine or purine-purine) divided by transversion mutations (pyrimidine to a purine or purine to a pyrimidine) and transitions are present at a higher rate. The expected Ti/Tv value for SNPs in exome sequencing is expected to situate between 2.0 and 3.0 (Guo et al., 2014). Sample contamination values show the level of accidentally mixed DNA from two or more individuals and moderate contamination levels stay within 5%–20% (Flickinge et al., 2015). Our observed scores stay within the expected values.

Considering the mean bait coverage and the percentage of all targeted bases, the sequencing quality achieved in this work is very high for all tested samples (see Methods and Materials for more details).

### 3.2.2 Exome-wide variant burden in the 600kb BP4-BP5 16p11.2 CNV carriers

The approach of identifying genetic modifiers is different from search for causal mutations in case of Mendelian diseases, especially in population cohort, where full trios are not available. This approach is emerging and no established standards for this type of analysis are available.

The total number of different exome-wide variants in the cohort of EGC UT 16p11.2 CNV carriers and their intra-familial members was 226 782 with average 87 266.5 per person. Deletion carriers had altogether 170 699 different alterations with 86 626–88 778 variants per person and duplication carriers had 188 026 with 85 971–88 411 variants per person, while intra-familial controls had 198 319 variants with 86 125–87 819 variants per person. When we limited from all the variants to protein disrupting variants, deletion carriers had 1 255, duplications 1 412 and controls 1 747 different variants. We observed that deletion carriers tend to have less protein disrupting variants than controls. Deletion carriers' genetic background contains less severe variants, otherwise the phenotype maybe lethal

#### 3.2.2.1 Variant burden in the 16p11.2 600kb BP4-BP5 genomic interval

16p11.2 600kb interval consists of 32 genes (*SPN, QPRT, C16orf54, ZG16, KIF22, MAZ, PRRT2, PAGR1* (*C16orf53*), *MVP, CDIPT, CDIPT-AS, SEZ6L2,* ASPHD1, *KCTD13, TMEM219, TAOK2, HIRIP3, INO80E, DOC2A, C16orf92, FAM57B, ALDOA, PP4C, TBX6, YPEL3, GDPD3, MAPK3, CORO1A, PPP4C, RN7SKP127, SLC7A5P1, MIR3680-2)* (https://www.decipher.sanger.ac.uk/). In total, we found different 126 variants in this interval. None of them were deleterious loss-of-function variants.

Deletion carriers had an average of 32 (ranging from 23 to 45; p=0.01) variants per individual; duplication carriers had 51 (range 39–59; p=0.09) and controls 46 (range 40–56) variants per person. After filtering for coding impact (exon deletion, frame-shift, missense or stop-loss substitutions) an average of 5.8 variants (range 5–7; p=0.02) per deletion carrier, 7.3 (range 6–9; p=0.94) per duplication carrier was identified and an average of 7.4 variants (range 6–10) intra-familial controls. We observe that deletion carriers have significantly less variants and coding impact variants in the 600kb region.

Genes *C16orf54, CDIPT-AS1, MIR3680-2, RN7SKP127, SLC7A5P1* and *YPEL3* (MIM #609724) did not have any protein truncating variants present in carriers or intra-familial members. *HIRIP3* (MIM #603365*)* had one heterozygous variant in control. Additionally, genes *FAM57B* (MIM #615175) and *ALDOA* (MIM #103850) had heterozygous variants, but none of them were homozygous among the 16p11.2 CNV carriers. Next we focused on particular genes in the 16p11.2 that have been associated with the phenotypes - *MAPK3, MVP, KCTD13* (Golzio et al., 2012; Migliavacca et al., 2015), *PRRT2* (Barge-Schaapveld et al., 2011; Crepel et al., 2011) and *TBX6* (Wu et al., 2015). *MAPK3, MVP, KCTD13* nor *PRRT2* genes contained substitutions which would have causative effect on the EGC UT carriers' phenotypes.

Search for variants in the *TBX6* gene resulted a total of 6 different variants in 23 carriers (deletion n=5 variants, duplication n=3 variants, controls n=3 variants). From these, 2 were intronic variants (rs3833842 and c.1098-19G>T), one was missense (rs56098093), one in splice region (rs112565029; and two were synonymous (c.864C>A and rs2289292) (Table 6). The synonymous variant rs2289292 in the last

exon of *TBX6*, along with SNPs rs3809624 (T/C) and rs3809627 (C/A) in the 5′ noncoding region, is part of the risk haplotype for congenital scoliosis identified by Wu et al. The three SNPs are tagging together a haploblock. Unfortunately the variants in non-coding region were not captured, and we do not know are these present in our carriers of the rs2289292. However, using this SNP we can at least tag a part of risk haplotype for vertebral malformations and we analyzed segregation of this variant together with spinal phenotypes in the 16p11.2 deletion carriers. Out of six deletion carriers one was with scoliosis (EX110) and two with osteochondrosis (EX101; EX102). Two deletion carriers (EX110, EX102) with congenital spinal defects indeed had the same rs2289292 tag-SNP. Third individual with congenital spinal defect (EX101) had a hemizygous variant c.864C>A (CADD= 9.6, SIFT= NA, ExAC= 2.6*e-04, ClinVar= NA).

Table 6. The variants in *TBX6* gene, which were occurring in EGC UT 16p11.2 CNV carriers' cohort.

| Variants in *TBX6* gene | rs2289292 | rs3833842 | c.1098-19G>T | c.864C>A | rs112565029 | rs56098093 c.484G>A |
|---|---|---|---|---|---|---|
| Position start | 30097630 | 30098189 | 30097778 | 30098148 | 30102391 | 30100401 |
| Impact | Synonymous | Intron | Intron | Synonymous | Splice region | Missense |
| ExAC | 0.322 | 0.352 | 3.7*e-05 | 2.6*e-04 | 0.06 | 0.00454 |
| CADD (scaled) | 2.0 | NA | 0.03 | 9.6 | 7.23 | 36 |
| Samples | DEL (n=2) DUP (n=6) | DEL (n=2) DUP (n=9) | DEL (n=1) | DEL (n=1) | DUP (n=2) | DEL (n=1) |

ExAC allele frequencies (n= 60706 unrelated individuals); EST- allele frequencies based on 162 whole-exomes.

The latter observation suggested that combination of *TBX6* null allele and suggestive hypomorphic risk haplotype might explained the scoliosis phenotype in the EGC UT 16p11.2 deletion carrier with scoliosis and be linked also to other vertebrae malformations, for example, osteochondrosis.

#### 3.2.2.2 Variants associated with cardiac septa defect in the 16p11.2 CNV carrier with congenital malformations of heart and Müllerian duct anomalies (EX100)

We filtered the list of 59 genes associated with atrioventricular septal defect (AVSD; in Supplementary Material Table S4) (D'Alessandro et al., 2015) against the duplication carrier EX100, with cardiac developmental disorder and congenital Müllerian duct anomaly. The total number of different variants in these candidate genes in the carriers (EX100) was 510. Using coding impact factors, number of variants was reduced to 87, of which 40 were homozygous. None of the homozygous variants showed significant causative effect on the phenotypes.

Analyzing the 87 variants further, we identified two missense mutations in EX100 (rs667782; p.Ile56Thr), which evaluation was either probably damaging or damaging. The rs667782 (CADD= 18.54, SIFT= 0.14, ExAC= 0.362, PolyPhen= 0.869) in gene hydrolethalus syndrome 1 (*HYLS1*; MIM #610693) was recurrent. Heterozygous mutation p.Ile56Thr (CADD= 23.9, SIFT= 0, ExAC= 0.00431, PolyPhen=0.968) in protein tyrosine phosphatase, non-receptor type 11 (*PTPN11*; MIM #176876) had predicted damaging effect and the variant is not present in ExAC, dbSNP or 1000G databases. Heterozygous dominant missense mutations in *PTPN11* gene result in Noonan syndrome (MIM #163950) and other rasopathies (http://www.omim.org). 16p11.2 duplication carrier (EX100) severe phenotypic features (heart defect and urogenital malformations) are probably caused by the p.Ile56Thr in *PTPN11* + 16p11.2 600 kb duplication.

#### 3.2.2.3 Variant burden in the 16p11.2 BP2-BP3 220kb genomic interval

Besides 16p11.2 BP4-BP5 600kb CNV, the proximal short arm of chromosome 16 includes a distal BP2-BP3 220kb region (Itsara et al., 2009), which microdeletion is associated with obesity and developmental delay (Bijlsma et al., 2009, Bachmann-Gagescu et al., 2010; Barge-Schaapveld et al., 2011). 220kb region 2 encompasses 9 genes: *ATXN2L* (MIM #607931), *TUFM* (MIM #602389), *SH2B1* (MIM #608937), *ATP2A1* (MIM #108730), *RABEP2* MIM #611869), *CD19* (MIM #107265), *NFATC2IP* (MIM #614525), *SPNS1* (MIM #612583), *LAT* (MIM #602354). We filtered the data against the genes and further used scaled CADD score and ExAC

frequency to identify potentially causative alterations. A deletion carrier EX102 in the EGC UT cohort, with the highest BMI 51.9 in the sample set, had 5 different alterations in the SH2B adaptor protein 1 gene (*SH2B1*; MIM #608937) – 5' UTR variant rs7198606; two intronic variants rs28433345 and rs386493443; splice region variant rs117918991; missense rs7498665. Different studies have presented *SH2B1* important role in obesity (Willer et al., 2009; Thorleifsson et al., 2009; Bochukova et al., 2010; Beckers et al., 2011; Tang et al., 2014). EX102 had a homozygous missense variant rs7498665 (c.1450A>G; CADD= 4.1, SIFT= 1, ExAC= 0.34, PolyPhen= 0) that has previously been described to be associated with obesity (Tang et al., 2014). In addition, EX102 had a heterozygous splice region variant rs117918991 (CADD= 10.56, SIFT= NA, ExAC= 0.01, PolyPhen= NA). The splice region variant rs117918991 and missense variant rs7498665 might have an impact on the phenotype of EX102. In the EGC UT carriers' cohort, this SNP is only present in EX102.

#### 3.2.2.4 Variant burden in the ciliopathy-associated genes

Ciliary dysfunction has previously shown to be affected by the 16p11.2 CNVs and some of the ciliopathy-associated genes (e.g. *CEP290*, *BBS6*) were shown to rescue the 16p11.2 triggered phenotype in zebrafish (Migliavacca et al., 2015). We analysed 173 ciliary genes (Migliavacca et al., 2015; gene list in Supplementary Material Table S5). In the latter, 16p11.2 deletion probands carried 1089.5 substitutions per individual, of which average 26 per individual were predicted to be damaging by SIFT score and approximately 5 of the damaging were in homozygote state. EX119 had a homozygous missense variant rs61734902 (CADD= 28.7, SIFT= 0, ExAC= NA, PolyPhen= 0.874) in katanin-interacting protein *(KIAA0556;* MIM #616650) and a homozygous missense variant rs4762 (CADD= 11.57, SIFT= 0.01, ExAC= 0.12, PolyPhen= 1) in angiotensinogen gene (*AGT;* MIM #106150). In addition, the EX119 had a heterozygous missense variant rs137853921 (CADD= 28.7, SIFT= 0, ExAC= 0.0043, PolyPhen= 0.954) in Bardet-Biedl syndrome gene (*BBS5;* MIM #615983). The mentioned variants are possible modifiers of the severe phenotype of EX119 (phenotypic features: mild mental retardation, cardiomyopathy, congenital malformations of uterus and cervix, female infertility). Variants in *BBS5* have been associated with Bardet-Biedl syndrome (Yadav et al., 2013) and EX119 have characteristic features to the disease.

Duplication carriers had total of 1097.8 variants per person (range from 1033–1170). The number of variants predicted to be damaging by SIFT was an average of 27.1 (range from 22–30) and homozygous damaging variants 5.8 (range from 3–9) per individual. Intra-familial controls had average 1105 substitutions per individual, of which 27.6 were with damaging assessment and 6.6 were homozygous damaging variants. The number of variants between carriers and intra-familial controls were similar.

From the aforementioned probably deleterious substitutions, we looked for variants not presented in dbSNP database and not in ExAC (the variants we analyzed here were in heterozygote state). We did not observe causative variants in the genes, which effect could affirm the carriers' specific features.

#### 3.2.2.5 Variant burden in the differentially expressed genes (DEG) associated with the 16p11.2 CNVs

Using the gene dosage model, 1 188 differentially expressed genes (DEGs) were previously identified in association with the 16p11.2 deletions and duplications (Migliavacca et al., 2015). From 1 188 DEG, we analyzed 1088 genes with Saphetor platform (Saphetor did not identify 100 genes from the list). The full list of 16p11.2 DEGs is provided in the Supplementary Material Table S6.

EGC UT 16p11.2 deletion individuals carried 4 860 (range from 4 739–4 978) substitutions per individual in the DEGs, from which homozygous were 1 927 variants (range from 1 815–1 927) and of these 13 variants (range from 10–18) per individual had damaging prediction. Heterozygous variants with damaging prediction in deletion carriers were 66.5 (range from 61–75) per individual. Duplication carriers had total of 4 816 variants per person (range from 4 719–4 898) and homozygous were 1 892 (range from 1 753–2 011) variants, from which 13 (range from 9–16) had damaging prediction. In case of duplication carriers, 65 (range from 57–75) heterozygous variants were predicated to be damaging. Intra-familial controls had average 4 751 variants (range from 4 663–4 803) in DEG set; homozygous were 1 974 (range from 1 813–2 087) variants and 16 (range from 11–22) had damaging prediction and heterozygous damaging variants occurred average 64 (range from 58–68). All the aforementioned variant quantities were similar between duplication carriers, deletion carriers and intra-familial controls. We did not observe any

significant variants, except the one, p.Ile56Thr in *PTPN11*, which was already described before in the current work (see above). This observation gives meaningful significance to the mutation (p.Ile56Thr) in *PTPN11* gene in the current analyze.

# 4 DISCUSSION

The 16p11.2 600 kb BP4-BP5 CNVs are among frequent genetic contributors to obesity and underweight (Jacquemont et al., 2011), and neuropsychiatric and neurodevelopmental disorders, such as intellectual disability (D'Angelo et al., 2016), language disorder (Shinawi, et al., 2010), and autism, (Weiss et al., 2008; Kumar et al., 2008). The reciprocal 600 kb deletion and duplication prevalence in general population is ~0.04% (Stefansson et al., 2014; Kirov et al., 2014). The current study focused on the carriers of the 16p11.2 600 kb CNVs in an adult population cohort from Estonia.

Adult CNV carriers had typical characteristics to the 16p11.2 pathology, like obesity and underweight. In the EGC UT cohort, 83% of the deletion carriers were obese and only one del carrier had BMI 24.15, which fits into normal range. Similarly the obesity penetrance among 600 kb deletion carriers was reported as 70% in a large 16p11.2 clinical cohort with a few individuals with normal BMI (Zufferey et al., 2012). The 16p11.2 duplication carriers were all between the population mean BMI (BMI 26.7) and underweight level (BMI <18.50). None of the duplication carriers were underweight, but 2 were close to the threshold (BMI 18.6 and 18.8). The frequency of underweight in the 16p11.2 clinical duplication cohort is around 12% (D'Angelo et al., 2016) hence the difference in population cohort is presumable, also showing the milder end of the syndrome's phenotype. Conclusively, these outcomes assured the earlier studies (Walters, et al., 2010; D'Angelo et al., 2016). Unlike BMI, 16p11.2 CNVs do not have conclusive implications on the carriers' height. Comparison of the height in Estonian population cohort versus 16p11.2 CNV carriers did not show significant difference between duplication males and controls (duplication male carriers being taller than non-carriers), and deletions females and controls (deletion female carriers being shorter than non-carriers). This could be explained by the small cohort size of the carriers. We analyzed the heights from the European 16p11.2 clinical adult cohort and we observed that female deletion carriers were significantly shorter than controls (female carriers' mean height was 159.1cm vs. deletion intra-familial controls 164cm). Analyzing the clinical male duplication carriers, we did not observe significant difference (male duplication carriers 177.7cm vs. male duplication controls 174.7cm).

In the current study's cohort schizophrenia, different neurodevelopmental disorders, and epilepsy were present, but had insignificant relationship with the 16p11.2 CNV carrier status. In clinical cohorts the frequencies are relevant. For example, frequency of epilepsy is 24% in deletion cohort (Zufferey et al., 2012) and 21.8% in duplication cohort (D'Angelo et al., 2016) and ASD frequencies are 15% for deletion (Zufferey et al., 2012) and 20.1% for duplication carriers (D'Angelo et al., 2016). Thus, likely reason might be the size of the current cohort. The second explanation would be that it is a general population cohort the individuals are with milder neuropsychiatric features whereas 16p11.2 clinical cohort contains the severe cases of the characteristics.

The average educational level in the EGC UT cohort has shown to be 4.1, which shows that most individuals in this group have completed at least the secondary school (Männik et al., 2015). EGC UT 16p11.2 CNV carrier cohort mean education was under the population average - deletion carriers 3.24 and duplication carriers 3.89. None of the deletion carriers were with higher education than secondary, whereas two out of 9 duplication carriers were with college/university degree. In clinical cohort, the mean effect of the reciprocal deletion and duplication on cognition is similarly decreasing, and the variation among duplication carriers is significantly greater with distinctive severe and mild subgroups (D'Angelo et al., 2016). A significant difference between intra-familial and EGC UT control groups were observed, but this observation is probably random. Although the 16p11.2 are known for their association with ASD, we did not find any CNV carriers with autism diagnosis in the EGC UT cohort. The explanation might be that people with autism have not become gene donors and they belong to the clinical cohort. For example, duplication carriers with ASD have more severe phenotype (FISQ decrease is impaired significantly; D'Angelo et al., 2016) and it has been previously shown that 15% of children carrying the reciprocal deletion had ASD and none of the fully assessed adult deletion carriers met the criteria for ASD (Zufferey et al. 2012). Current study's cohort had deletion and duplication carriers with different malformations of spine, heart, uterus and cervix. Previous studies in clinical 16p11.2 cohorts showed that the same set of malformations, vertebral (3.8% for deletion and 0.7% for duplication) genital (3.1% and 5.9%) and cardiac (4.6% and 4.4%), described in he current study, are the most frequent congenital defects in clinical carriers' cohort (Zufferey et al., 2012; D'Angelo et al. 2016). In the population

carriers' cohort, both cardiac malformation and genitalia deformity frequencies were 16.7% in deletion and 11.1% in duplication carriers, and vertebral malformation occurred in 50% of the deletion carriers (duplication carriers were without spinal malformations). In spite of the fact, that clinical carriers cohort had generally more severe phenotype than EG CUT population cohort, these malformations were presented more recurrently.

The novel part of the current study was the analysis of common phenotypes. No previously published study has investigated the genomic disorders relationship to common features. We had the access to accurate clinical data, to conduct this kind of analysis (detailed clinical information on the CNV carriers and EGC UT population cohort). This kind of study is important for clinical practice, because common characteristics influence the course and severity of the disease and this gives opportunity to prepare a proper treatment plan. With the phenotypic analysis, we found three significantly recurrent disease categories compared to controls - hypotension in the duplication carriers, psoriasis in both CNV carriers and functional dyspepsia in deletion carriers. Diseases of oesophagus, stomach and duodenum showed non-significant trend with increased risk.

Previous information from literature shows that women with the 600kb deletion have problems with reproductive system (Sandbacka et al., 2013; Tewes et al., 2015). We found absent, scanty and rare menstruation, non- inflammatory disorders of ovary, fallopian tube, broad ligament, and cystitis occurring significantly more often in female 16p11.2 CNV carriers than with controls. Due to small sample size and rarity, infertility and infections with a predominantly sexual mode of transmission showed non-significant trend with increased risk.

From the results of this study, we have evidence that 16p11.2 600kb carriers have problems with digestive system and psoriasis. Moreover, there is a clear significance of difficulties in female health features in the 16p11.2 CNV cohort. We researched the 16p11.2 CNV phenotypes describing articles by Shinaw et al., Zufferey et al., and D'Angelo et al., whether they had reported aforementioned common traits occurring in the carriers. These studies have so far excluded common features from the analysis and their association has not been investigated in the 16p11.2 CNV carriers (Shinaw et al. 2010; Zufferey et al., 2012; D'Angelo et al., 2016). However, gastrointestinal disorders are known significant comorbid conditions of ASD (reviewed in Hsiao et al., 2013) and patients with ID have also higher rates of gastrointestinal disorders

(Krahn and Fox, 2013). Studies conducted to investigate the link between obesity and functional gastrointestinal disorders, such as functional dyspepsia, did not show a consistent connection (Ho and Spiegel, 2008).
In summary, the phenotype analysis of this study confirmed the characteristic features of 16p11.2 pathologies and identified new potential associations with height and common disease features.

Phenotypic variability in 16p11.2 deletion and duplication carriers is wide and studies have been performed on the importance of secondary mutational hits at other loci, among 16p11.2 CNV carriers, affecting phenotype severity (Girirajan and Eichler, 2010; Girirajan et al., 2010; Girirajan et al., 2012). In the clinical cases (more severe phenotypes compared to carriers from population cohort), approximately 70% of the deletion and duplication carriers had a secondary CNV and 35% had two or more CNVs (Duyzend et al., 2016). Several of the secondary CNVs have already been implicated as risk factors for autism and developmental delay, for example 2x40 kb deletion of the *TOP3B* locus on chromosome 22q11.22 (Stoll et al., 2013). Duyzend et al. (2016) reported that 16p11.2 CNV carriers who carry other CNVs have lower IQs compared to those without extra CNVs (Duyzend et al., 2016). In general, deletion carriers have significantly diminished rate of second pathogenic CNVs compared to the duplication carriers (D'Angelo et al., 2016). However, these examples might explain only a small part of the variable expressivity in 16p11.2 syndromes. Therefore, we conducted whole-exome sequencing to 15 CNV carriers and their 8 intra-familial members to find genetic factors modulating the variability and explain the co-morbidities.
This type of research is the first to investigate the genetic background of the genomic disorder. We were not expecting to identify rare loss-of-function variants, instead slightly deleterious mutations. We were analyzing the adult carriers set from population cohort with "milder" phenotype. If they contain rare causative loss-of-function variants additionally to the deleterious CNV, they probably would not be in the population cohort. Also, we searched for variants in genes, which were associated with the 600 kb interval. In conjunction with 16p11.2 CNVs, the modifying variants could be common and do not have to be rare deleterious mutation.

We compared the amount of the protein disrupting variants between deletion carriers,

duplication carriers and controls, the deletion carriers tend to have slightly decreased rate of deleterious variants in contrast to controls. This observation was expected, because otherwise the high amount of mutations in carriers would result in critical condition and not in population cohort.

We also looked at genes in the 16p11.2 600 kb locus. Among the 32 genes, the *TBX6* (transcription factor that functions in early embryogenesiss) was the most significant candidate to explain specific phenotypes. *TBX6* belongs into phylogenetically well-conserved T-box gene family (Yi et al., 1999). It consists of 8 exons, the transcript is 1806 bp long and the protein size is 436 amino acids. An abnormal function or expression of various T-box genes has been associated with congenital malformations in humans (Naiche et al., 2005; Plageman and Yutzey, 2005). *TBX6* expression has shown to be essential for normal vertebral formation (Watabe-Rudolph et al., 2002; White et al., 2003). Although *TBX6* was the main candidate for congenital vertebral defects seen in 16p11.2 deletion carriers, the exact mechanism why scoliosis and hemivertebra are present only in a subset of 16p11.2 deletion carriers was not explained until recently. The effect of diminished *TBX6* dosage and an additional *TBX6* hypomorphic allele causes a further decrement in expression and will lead to congenital scoliosis.

In our 16p11.2 CNV carriers' data set, we had three deletion carriers with developmental spine disorders. All the latter deletion carriers had the tag-SNP rs3809624 that describes the risk haplotype for scoliosis. The capture kit used in the exome analysis catches only variants from protein coding regions, therefore only this tag-SNP from described risk haplotype was described. Although the other two SNPs were located in the 5′ noncoding region we were assuming that the two deletion carriers have the described risk haplotype. Additional deletion carrier with osteochondrosis and without this tag-SNP was the only individual who had a homozygous variant c.864C>A in the *TBX6* gene, which possibly contributes to the phenotype. Based on that we might have explained the scoliosis phenotype in our cohort's deletion carrier and, furthermore, could suggest that this genetic combination might be responsible also for other spinal malformations, for example osteochondrosis. Hence, it would be useful for clinicians to conduct more profound

assessment of vertebral conditions of the 16p11.2 600kb CNV carriers and carry out targeted genetic testing for risk variants in the *TBX6*.

In addition to vertebral disorders, *TBX6* has also been associated with disorder of the Müllerian duct (Sandbacka et al., 2013; Tewes et al., 2015), and cardiac defects (Gavrilov et al., 2013) as *TBX6* is a general developmental regulator (Callery et al., 2010). We analyzed whether variants in this gene might be associated with mentioned congenital defects. In this cohort, 1 female duplication (EX100) and 1 deletion carrier (EX119) had Müllerian duct malformation. Duplication carrier (EX100) had also cardiac developmental disorder. None of the two were carriers of a significant finding in the *TBX6*. We analyzed other candidate genes for the malformations. *PTPN11* was associated with AVSD and also belonged to differentially expressed genes set in 16p11.2 CNV carriers. The duplication carrier with cardiac and urogenital malformations was the only carrier to have the missense mutation p.Ile56Thr in *PTPN11*. ExAC database shows that mutations in *PTPN11* gene are rare and LoF variants are not tolerated (observed number of LoF variants 0). Gene *PTPN11* has been associated with Noonan syndrome and congenital heart defects. Carrier's (EX100) features - the atrial septal defect and urogenital malformations – are also the characteristic features of Noonan syndrome. According to the described case, EX100 severe phenotype is probably explained by pathogenic missense mutation p.Ile56Thr in *PTPN11* + 16p11.2 duplication. While analyzing possibly causative variants in the other individual with urogenital malformation (EX119), we found a heterozygous variant rs137853921 in *BBS5* gene related to Bardet-Biedl syndrome. *BBS5* also belongs into ciliopathy-associated genes. The same variant (rs137853921) has also been mentioned before as being possibly pathogenic (Filges et al., 2013). Bardet-Biedl syndrome characteristic features are cardiomyopathy (Yadav et al., 2013), genital abnormalities (Forsythe and Beales, 2013), different neuropsychiatric impairments and mental retardation (Baker et al., 2011), which are also represented in our 16p11.2 duplication carrier (EX119). The other two homozygous missense variants (rs61734902 in *KIAA0556* and rs4762 in *AGT*) in ciliary associated genes might also have a modifying effect on the EX119 phenotype.

From differentially expressed genes associated with the 16p11.2 deletions and duplications, we filtered out probably damaging variants without assigned rs number and no carriers in ExAC, dbSNP or 1000G databases. The filtered variants were compared to specific features of the carriers. With this analyze, we re-identified the

variant p.Ile56Thr in *PTPN11*, which was described above. This observation confirms the effect of the missense mutation on the duplication carrier (EX100). According to used filter criteria, no other substitutions were associated with carriers' characteristic phenotypes, but further detailed analyze is necessary for the DE genes.

In addition to the 600 kb CNV, the 16p11.2 chromosome interval harbors also a distal BP2-BP3 220kb CNV, which has been shown to interact with the 600 kb CNVs on the chromatin level (Loviglio et al., Mol Psych in press). The 220 kb deletion (MIM #613444) has been associated with developmental delay, behavioral problems, and extreme childhood obesity (Bochukova et al., 2010; Barge-Schaapveld et al., 2011). SH2B Adaptor Protein 1 (*SH2B1*) encoded protein is involved in leptin (Duan et al., 2004; Li et al., 2007) and insulin signaling in lipid metabolism (Morris et al., 2009). *SH2B1* is one of the most significant genes in 220kb interval. GWAS studies have showed *SH2B1* gene as susceptibility locus for obesity (Willer et al., 2009; Thorleifsson et al., 2009) and both GWAS studies identified SNP rs7498665 in association with obesity. In addition a meta-analysis conducted by Tang et al. also identified a significant association with overweight/obesity (Tang et al., 2014). In EG CUT carriers' cohort, the deletion carrier EX102, who had the highest BMI of 51.9, had 5 alterations in *SH2B1* gene, of which one is the same described missense variant rs7498665. The SNP was homozygous in only EX102 among all CNV carriers and might be a modifier in the EX102's extremely high BMI. Also, the heterozygous splice region variant rs117918991 in *SH2B1* gene in EX102 may have an impact on the phenotype.

Knowing a little about the mechanisms involved in a disease helps to orientate exome analysis to corresponding pathways. 16p11.2 CNV syndromes, for instance, are known to be caused by the 600 kb deletion or duplication, but exact molecular mechanisms behind the disease warrant further investigation. The 16p11.2 600kb interval comprises 32 genes and the CNVs perturb the localized genes and genes on its flanks (Migliavacca et al., 2015). In the current work, exome analysis was directed to the aforementioned genes, differentially expressed genes associated with 16p11.2 CNVs (Migliavacca et al., 2015), interaction partners of genes involved in the 600kb interval (MetaCore, GeneCards, STRING) and genes, which were linked to the features of the 16p11.2 syndromes (Sampson et al., 2010; Nik-Zainal et al., 2011; Sandbacka, et al. 2013; Wu et al., 2015; D'Alessandro et al., 2016).

In conclusion, our results have described 5 of the population 16p11.2 600 kb carrier's specific phenotypes and have shown the importance of analyzing the genomic background, despite knowing the genetic cause of a disease.

## CONCLUSION

The 16p11.2 BP4-BP5 region is predisposed to rearrangements and diseases. The structural variations in this genomic position result in 16p11.2 deletion or duplication syndrome. In the current work, we aimed to characterize the genotypic and phenotypic features in 16p11.2 600kb BP4-BP5 CNV carriers in EGC UT adult population cohort.

We confirmed the frequent characteristic features of 16p11.2 600 kb syndromes, such as high vs. low BMI, neuropsychiatric disorders (schizophrenia, epilepsy, neurotic disorder, anxiety disorder), intellectual disability, and malformations of vertebrae, heart and genitalia. The current work is first to describe common features significantly recurrent in EGC UT 16p11.2 CNV carriers' cohort. Hypotension frequency was higher among duplication carriers, psoriasis in deletion and duplication carriers, and functional dyspepsia in deletion cohort. Additionally, several female health specific traits were significantly more presented in the female CNV carriers' cohort. We used an exome-wide approach to find genetic modifiers resulting in the variability of genomic disorders' phenotypes. We might have explained 4 16p11.2 CNV carriers' specific features with genetic substitutions found by exome analysis. First, a common haplotype of *TBX6* in case of scoliosis and osteochondrosis. Secondly a missense mutation in *PTPN11* resulting in severe duplication phenotype with atrial septal defect and urogenital malformation. Thirdly, a missense variant in gene *BBS5*, explaining Bardet-Biedl syndrome-like characteristics in 16p11.2 deletion carrier and finally, variants in *SH2B1* modifying the extreme obesity in a deletion carrier. The observed alterations identified in the study need further functional studies and confirmation.

According to our phenotypic and genotypic findings, we conclude that it is important to conduct a detailed phenotypic assessment of individuals with particular genetic alterations (e.g. 16p11.2 600 kb deletion and duplication) and further investigate the exome (or genome) of the carriers to more precisely predict the severity and diverse outcomes of disease. The secondary hits may be necessary for a clinically ascertainable phenotype and therefore are important to take into account in genetic counseling.

**Kromosoomi piirkonna 16p11.2 BP4-BP5 600 kb koopiaarvu variatsioonide iseloomustus täiskasvanute populatsioonis**

Berit Kolk

## Kokkuvõte

Kromosoomi piirkond 16p11.2 BP4-BP5 on disponeeritud ümberkorraldustele ja muutustest tingitud haigustele. Strukturaalsed muutused antud genoomi piirkonnas väljenduvad 16p11.2 deletsiooni või duplikatsiooni sündroomina. Käesoleva uurimuse eesmärgiks on välja tuua 16p11.2 600kb BP4-BP5 koopiaarvu kandjate genotüübi ja fenotüübi iseloomustus TÜ EGV täiskasvanute populatsiooni andmete alusel. Meie tulemused tõendavad, et 16p11.2 600 kb sündroomile on iseloomulikud ja sageli esinevad järgmised sümptoomid: kõrge või madal KMI, psühhiaatrilised häired (skisofreenia, neuroos, ärevushäired), epilepsia, vaimse arengu mahajäämus, samuti lülisamba, südame ja genitaalide anomaaliad. Antud töö kirjeldab esmakordselt olulisi sagedasi haiguslikke tunnuseid EGV koopiaarvu kandjate kohordis. Nii esines hüpotensiooni sagedamini duplikatsiooni kandjatel, psoriaasi esines nii deletsiooni kui duplikatsiooni kandjatel ja funktsionaalset düspepsiat enam deletsiooni kohordis. Lisaks olid naiste tervisehäiretega seotud kõrvalekalded oluliselt enam väljendunud koopia arvu variatsioonidega naistel.

Kasutasime üle-eksoomset analüüsi, et leida võimalikke geneetilisi modifikaatoreid, mis vastutaksid kohordis esinevate fenotüüpide mitmekesisuse eest. Teostatud eksoomi uuringute andmetel on 16p11.2 kandjatel võimalik kirjeldada 4 erinevat fenotüüpi. Esiteks levinud *TBX6* haplotüüp, millega seostub skolioos ja osteokondroos. Teiseks *missense* mutatsioon *PTPN11* geenis, mille korral lisandub duplikatsiooni raskele fenotüübile kodade vaheseina defekt ja urogenitaaltrakti anomaalia. Kolmandaks *BBS5* geenis *missense* variant, mis võib põhjustada 16p11.2 deletsiooni kandjatel Bardet-Biedl sündroomiga sarnaseid tunnuseid. Neljandaks *SH2B1* geeni variant, mis mõjutab äärmusliku rasvumise kujunemist deletsiooni kandjatel. Kinnitamaks töös tuvastatud variantide olulisust fenotüüpide modifitseerimisel, vajavad need variandid jätkuvat uurimist ning nende kahjuliku mõju kinnitamist.

Toetudes fenotüübi ja genotüübi vaheliste seoste uurimistulemustele saame väita, et detailne fenotüübi kirjeldamine isikutel, kellel esineb geneetiline muutus (16p11.2 600 kb deletsioon või duplikatsioon), on väga oluline. Edasine täpne geneetilise tausta (eksoomi või genoomi) uurimine kandjatel annab võimaluse prognoosida haiguse raskust ja kliiniliste sümptomite varieeruvust. Teiseks annab täpsuselt kirjeldatud fenotüüp olulise aluse kliinilises praktikas ja on kasutatav geneetilise konsultatsiooni läbiviimisel.

**ACKNOWLEDGMENTS**

I would like to thank my supervisor Katrin Männik from the bottom of our hearts for her help and guidance during my master's study. I thank her for being not only a brilliant supervisor, but also a good friend and a role model. Also, I would like to thank my supervisor prof. Andres Metspalu, whose advice I appreciate so much and whose support has always been very important to me.

Special thanks go to Margit Nõukas, who helped me with my questions and difficulties during my study.

I would like to thank Mart Kals, Mari-Liis Tammesoo, Reedik Mägi, Steven Smith, Kairit Mikkel, Helene Alavere, Dr. Anneli Kolk, Dr. Anu Reigo, Heidi Saulep, Viljo Soo, Pille Kool and Tiit Nikopensius for their help and advice. I would like to thank the whole biotechnology collative in Estonian Genome Center.

I am very grateful to Alexandre Reymond for giving me the opportunity to do a part of my master's thesis in his lab in the University of Lausanne, for his good supervision and for useful discussions on science.

Many thanks go to Hannes Richter, Anne Maillard, Sylvain Pradervand, Julien Delafontaine, Andreas Massouras, Jacqueline Chrast.

Furthermore, I would like to thank my precious family and good friends for all the support and encouragement during my master studies.

Especially big thanks go to my lovely mother and father, whose advice, support and help have always been priceless.

# 5 REFERENCES


Abe, K., Takamatsu, N., Ishikawa, K., Tsurumi, T., Tanimoto, S., Sakurai, Y., Lisse, T., Imai, K., Serikawa, T., and Mashimo, T. (2015). "Novel ENU-Induced Mutation in Tbx6 Causes Dominant Spondylocostal Dysostosis-Like Vertebral Malformations in the Rat." Plos One. **10**(6): e0130231.

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., and Sunyaev, S. R. (2010). "A method and server for predicting damaging missense mutations." Nat Methods. **7**: 248–249.

Al-Kateb, H., Khanna, G., Filges, I., et al. (2014). "Scoliosis and vertebral anomalies: additional abnormal phenotypes associated with chromosome 16p11.2 rearrangement." Am J Med Genet A. **164**: 1118–1126.

Alves J. M., L. A. M., Chikhi, L., and Amorim, A. (2012). "On the structural plasticity of the human genome: chromosomal inversions revisited." Current Genomics. **13**(8): 623–632.

Atanda, A., Shah, S. A. and O'Brien, K. (2011). "Osteochondrosis: Common Causes of Pain in Growing Bones." Am Fam Physician. **83**(3): 285–291.

Bachmann–Gagescu, R., Mefford, H. C., Cowan, C., Glew, G. M., Hing, A. V., Wallace, S., Bader, P. I., Hamati, A., Reitnauer, P. J., Smith, R., Stockton, D. W., Muhle, H., Helbig, I., Eichler, E. E., Ballif, B. C., Rosenfeld, J., and Tsuchiya, K. D. (2010). "Recurrent 200–kb deletions of 16p11.2 that include the SH2B1 gene are associated with developmental delay and obesity." Genet Med. **12**: 641–647.

Bailey, J. A., Gu, Z., Clark, R. A., Reinert, K., Samonte, R. V., Schwartz, S., Adams, M. D., Myers, E. W., Li, P. W., and Eichler, E. E. (2002). "Recent segmental duplications in the human genome." Science. **297**(5583): 1003–1007.

Baker, K., Northam, G. B., Chong, W. K., Banks, T., Beales, P., and Baldeweg, T. (2011). "Neocortical and hippocampal volume loss in a human ciliopathy: a quantitative MRI study in Bardet-Biedl syndrome." American Journal of Medical Genetics Part A. **155**: 1–8.

Barge-Schaapveld, D. Q., Maas, S. M., Polstra, A., Knegt, L. C., and Hennekam, R. C. (2011). "The atypical 16p11.2 deletion: a not so atypical microdeletion syndrome?" Am J Med Genet A. **155A**(5): 1066–1072.

Becker-Heck, A., Zohn, I., Okabe, N., Pollock, A., Niswander, L., Omran, H., Burdine, R. D., et al. (2011). "The coiled-coil domain containing protein CCDC40 is essential for motile cilia function and left-right axis formation." Nat Genet. **43**(1): 79–84.

Beckers, S., Zegers, D., Van Gaal, L. F., and Hul, W. V. (2011). "Replication of the SH2B1 rs7498665 Association with Obesity in a Belgian Study Population." The European Journal of Obesity. **4**: 473–477.

Bijlsma, E. K., Gijsbers, A. C., Schuurs-Hoeijmakers J. H., van Haeringen, A., Fransen van de Putte, D. E., Anderlid, B. M., Lundin, J., van Essen, A. J., Hordijk, R., Sikkema-Raddatz,

B., Verbruggen, K. T., et al. (2009). "Extending the phenotype of recurrent rearrangements of 16p11.2: deletions in mentally retarded patients without autism and in normal individuals." Eur J Med Genet. **52**: 77–87.

Bochukova, E. G., Huang, N., Keogh, J., Henning, E., Purmann, C., Blaszczyk, K., Saeed, S., Hamilton-Shield, J., Clayton-Smith, J., O'Rahilly, S., Hurles, M. E., and Farooqi, I. S. (2010). "Large, rare chromosomal deletions associated with severe early-onset obesity." Nature. **463**(7281): 666–670.

Budisteanu, M., Arghir, A., Cardos, G., Chirieac, S., Macureanu, S., Blancho, D., Dupont, J.-M. and Lungeanu, A. (2008). "Phenotypic variability in Angelman syndrome — report of two cases." Maedica - a Journal of Clinical Medicine. **3**(3): 192–197.

Butler, M. G., et al. (2005). "Subset of individuals with autism spectrum disorders and extreme macrocephaly associated with germline PTEN tumour suppressor gene mutations." J Med Genet. **42**: 318–321.

Callery, E. M., Thomsen, G. H., and Smith, J. C. (2010). "A divergent Tbx6-related gene and Tbx6 are both required for neural crest and intermediate mesoderm development in Xenopus." Dev Biol. **340**(1): 75–87.

Carvalho, C. M. B., Zhang, F., and Lupski, J. R. (2010). "Genomic disorders: A window into human gene and genome evolution." PNAS. **107**(1): 1765–1771.

Cerrato, F., Crescenzo, A. D. and Riccio, A. (2014). "Looking for CDKN1C enhancers." European Journal of Human Genetics. **22**: 442–443.

Chen, X., Li, X., Wang, P., Liu, Y., Zhang, Z., Zhao, G., Xu, H., Zhu, J., Qin, X., Chen, S., Hu, S. and Kong, X. (2010). "Novel Association Strategy with Copy Number Variation for Identifying New Risk Loci of Human Diseases." PLoS One. **5**(8): e12185.

Clancy, S. (2008). "Copy number variations (CNVs) have been linked to dozens of human diseases, but can they also represent the genetic variation that was so essential to our evolution?" Nature Education. **1**(1): 95.

The Wellcome Trust Case Control Consortium. (2010). "Genome-wide association study of copy number variation in 16,000 cases of eight common diseases and 3,000 shared controls." Nature. **464**(7289): 713–720.

Cooper, G. M., Coe, B. P., Girirajan, S., Rosenfeld, J. A., Alexander, N., Gorski, J., Kussmann, J., Shashi, V., Johnson, K., Ballif, B. C., Shaffer, L. G., Eichler, E. E., et al. (2011). "A copy number variation morbidity map of developmental delay." Nat Genet. **43**: 838–846.

Crepel, A., Steyaert, J., De la Marche, W., De Wolf, V., Fryns, J. P., Noens, I., Devriendt, K. and Peeters, H. (2011). "Narrowing the critical deletion region for autism spectrum disorders on 16p11.2." Am J Med Genet B Neuropsychiatr Genet. **156**: 243–245.

Crespi, B., Stead, P. and Elliot, M. (2010). "Evolution in health and medicine Sackler colloquium: comparative genomics of autism and schizophrenia." Natl Acad. Sci.: 1736–

1741.

D'Angelo, D., Lebon, S., Chen, Q., Martin-Brevet, S., Männik, K., Metspalu, A., Mignot, C., Mukherjee, P., Owen, M. J., Passeggeri, M., Rooryck, C., Rosenfeld, J. A., Spence, S. J., Steinman, K. J., Tjernagel, J., van Haelst, M., Shen, Y., Draganski, B., Spiro, J. E., Reymond, A., Jacquemont, S., Chung, W. K., et al.; for the ECHO study, the 16p11.2 European Consortium and the Simons VIP Consortium. (2016). "Defining the effect of the 16p11.2 duplication on cognition, behavior, and medical comorbidities." JAMA Psychiatry. **73**(1): 20–30.

Dementieva, Y. A., et al. (2005). "Accelerated head growth in early development of individuals with autism." Pediatr. Neurol. **32**: 102–108.

DePristo, M. A., Banks, E., Poplin, R. E., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). "A framework for variation discovery and genotyping using next-generation DNA sequencing data." Nat Genet. **43**: 491–498.

Duan, C., Li, M., and Rui, L. (2004). "SH2-B promotes insulin re- ceptor substrate 1 (IRS1)- and IRS2-mediated activation of the phosphatidylinositol 3-kinase pathway in response to leptin." J Biol Chem. **279**: 43684–43691.

Duyzend, M. H., and Eichler, E. E. (2015). "Genotype-First Analysis of the 16p11.2 Deletion Defines a New Type of "Autism"." Biological Psychiatry. **77**: 769–771.

Duyzend, M. H., Nuttle, X., Coe, B. P., Baker, C., Nickerson, D. A., Bernier, R., and Eichler, E. E. (2016). "Maternal Modifiers and Parent-of-Origin Bias of the Autism-Associated 16p11.2 CNV." The American Journal of Human Genetics **98**: 45–57.

Elia, J., Gai, X., Xie, H. M., Perin, J. C., Geiger, E., Shaikh, T. H., Hakonarson, H., White, P. S., et al. (2010). "Rare structural variants found in attention-deficit hyperactivity disorder are preferentially associated with neurodevelopmental genes." Mol Psychiatry. **15**(6): 637–646.

Fernandez, B. A., Roberts, W., Chung, B., et al. (2010). "Phenotypic spectrum associated with de novo and inherited deletions and duplications at 16p11.2 in individuals ascertained for diagnosis of autism spectrum disorder." J. Med. Genet. **47**: 195–203.

Feuk, L., Carson, A. R. and Scherer, S. W. (2006). "Structural variation in the human genome." Nature Reviews Genetics. **7**: 85-97.

Finseth, F. R. and Harrison, R. G. (2014). "A Comparison of Next Generation Sequencing Technologies for Transcriptome Assembly and Utility for RNA-Seq in a Non-Model Bird." PLoS One. **9**(10): e108550.

Flickinger, M., Jun, G., Abecasis, G. R., Boehnke, M., and Kang, H. M. (2015). "Correcting for Sample Contamination in Genotype Calling of DNA Sequence Data." The American Journal of Human Genetics. **97**(2): 284–290.

Forsythe, E., and Beales, F. L. (2011). "Bardet–Biedl syndrome." European Journal of Human Genetics. **21**: 8–13.

Fromer, M., Pocklington, A.J., Kavanagh, D.H., Williams, H.J., Dwyer, S., Gormley, P., Georgieva, L., Rees, E., Palta, P., Ruderfer, D.M., et al. (2014). "De novo mutations in schizophrenia implicate synaptic networks." Nature. **506**: 179–184.

Fukumoto, A., et al. (2008). "Growth of head circumference in autistic infants during the first year of life." J. Autism Dev. Disord. **38**: 411–418.

Gavrilov, S., Harvey, R. P. and Papaioannou, V. E. (2013). "Lack of Genetic Interaction between Tbx20 and Tbx3 in Early Mouse Heart Development." PLoS One. **8**: e70149.

Giarolia, G., Bassa, N., Strydoma, A., Rantella, K. and McQuillina, A. (2014). "Does rare matter? Copy number variants at 16p11.2 and the risk of psychosis: A systematic review of literature and meta-analysis." Schizophr Res. **159**(2-3): 340-346.

Girirajan, S., and Eichler, E. E. (2010). "Phenotypic variability and genetic susceptibility to genomic disorders." Hum Mol Genet. **19**(R2): R176–R187.

Girirajan, S., Campbell, C. D., Eichler, E. E. (2011). "Human copy number variation and complex genetic disease." Annu Rev Genet. **45**: 203–226.

Golzio, C., Willer, J., Talkowski, M.E., Oh, E.C., Taniguchi, Y., Jacquemont, S., Reymond, A., Sun, M., Sawa, A., Gusella, J.F., et al. (2012). "KCTD13 is a major driver of mirrored neuroana- tomical phenotypes of the 16p11.2 copy number variant." Nature. **485**: 363–367.

Greenway, S. C., Pereira, A. C., Lin, J. C., DePalma, S. R., Israel, S. J., Mesquita, S. M., Ergul, E., Conta, J. H., Korn, J. M., McCarroll S. A., Gorham, J. M., Gabriel, S., Altshuler, D. M., Quintanilla-Dieck Mde, L., Artunduaga, M. A., Eavey, R. D., Plenge, R. M., Shadick, N. A., Weinblatt, M. E., De, Jager, P. L., Hafler, D. A., Breitbart, R. E., Seidman, J. G. and Seidman, C. E. (2009). "De novo copy number variants identify new genes and loci in isolated sporadic tetralogy of Fallot." Nat Genet. **41**(8): 931–935.

Grozeva, D., Kirov, G., Ivanov, D., Jones, I. R., Jones, L., Green, E. K., McGuffin, P., Holmans, P. A., Owen, M. J., O' Donovan, M. C., Craddock, N., et al.; Wellcome Trust Case Control Consortium. (2010). "Rare copy number variants: a point of rarity in genetic risk for bipolar disorder and schizophrenia." Arch Gen Psychiatry. **67**(4): 318–327.

Guo, Y., Zhao, S., Samuels, D. C., Shyr, Y., et al. (2014). "Multi-perspective quality control of Illumina exome sequencing data using QC3." Genomics. **103**(5–6): 323–328.

Hanson, E., Bernier, R., Porche, K., Jackson, F. I., Goin-Kochel, R. P., Wallace, A. S., Ledbetter, D. H., Spiro, J. E. and Chung, W. K. (2015). "The Cognitive and Behavioral Phenotype of the 16p11.2 Deletion in a Clinically Ascertained Population." Biological Psychiatry. **77**: 785–793.

Hastings, P. J., Lupski, J. R., Rosenberg, S. M. and Ira, G. (2009). "Mechanisms of change in gene copy number." Natrue Reviews | Genetics. **10**: 551–564.

Henrichsen, C. N., Vinckenbosch, N., Zöllner, S., Chaignat, E., Pradervand, S., Schütz, F.,Ruedi , M., Kaessmann, H.,and Reymond, A. (2009). "Segmental copy number variation shapes tissue transcriptomes." Nat Genet. **41**: 424–429.

Hotta, K., Kitamoto, T., Kitamoto, A., Mizusawa, S., Nakao, K., Sekine, A., et al. (2011). "Computed tomography analysis of the association between the SH2B1 rs7498665 single-nucleotide polymorphism and visceral fat area." Journal of Human Genetics. **56**: 716–719.

Itsara, A., Cooper, G. M., Baker, C., Girirajan, S., Nickerson, D. A., Eichler, E. E., et al. (2009). "Population analysis of large copy number variants and hotspots of human genetic disease." Am J Hum Genet. **84**(2): 148–161.

Jacquemont, S., Reymond, A., Zufferey, F., Harewood, L., Walters, R. G., Kutalik, Z., et al. (2011). "Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus." Nature. **478**: 97–102.

Johnson, M. E., Cheng, Z., Morrison, V. A., Scherer, S., Ventura, M., Gibbs, R. A., Green, E. D. and Eichler, E. E. (2006). "Recurrent duplication-driven transposition of DNA during hominoid evolution." Proc Natl Acad Sci U S A. **103**(47): 17626–17631.

Kaminsky, E. B., Kaul, V., Paschall, J., Church, D. M., Bunke, B., Kunig, D., Moreno-de-Luca, D., Aston, E., Whitby, H., Shetty, S., Rossi, M. R., Rudd, M. K., South, S. T., Brothman, A. R., Sanger, W. G., Iyer, R. K., Crolla, J. A., Thorland, E. C., Aradhya, S., Ledbetter, D. H., Martin, C. L., et al. (2011). "An evidence-based approach to establish the functional and clinical significance of copy number variants in intellectual and developmental disabilities." Genet Med. **13**: 777–784.

Kim, S. J., Miller, J. L., Kuipers, P. J., German, J. R., Beaudet, A. L., Sahoo, T. and Driscoll, D. J. (2011). "Unique and atypical deletions in Prader-Willi syndrome reveal distinct phenotypes." European Journal of Human Genetics. **20**(3): 283–290.

Kircher, M., et al. (2014). "A general framework for estimating the relative pathogenicity of human genetic variants." Nat Genet. **46**(3): 310–315.

Kirov, G., Rees, E., Walters, J. T., Escott-Price, V., Georgieva, L., Richards, A. L., et al. (2014). "The penetrance of copy number variations for schizophrenia and developmental delay." Biol Psychiatry. **75**: 378–385.

Lainhart, J. E., et al. (2006). "Head circumference and height in autism: a study by the Collaborative Program of Excellence in Autism." Am J Med Genet. **140**: 2257–2274.

Leitsalu, L., Haller, T., Esko, T., et al. (2014). "Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu." Int J Epidemiol. **44**(4): 1137–1147.

Li, H., and Durbin, R. (2009). "Fast and accurate short read alignment with Burrows-Wheeler transform." Bioinformatics. **25**: 1754–1760.

Li, M., Nordborg, M. and Li, L. M. (2004). "Adjust quality scores from alignment and improve sequencing accuracy." Nucleic Acids Res. **32**: 5183–5191.

Li, Y., Dai ,Q., Jackson, J. C. and Zhang, J. (2008). "Overweightisassociatedwithdecreased cognitive functioning among school-age children and adolescents." Obesity (Silver Spring). **16**: 1809–1815.

Li, Z., Zhou, Y., Carter-Su, C., Myers, M. G., and Rui, L. (2007). "SH2B1 enhances leptin signaling by both Janus kinase 2 Tyr813 phosphorylation-dependent and -independent mechanisms." Mol Endocrinol. **21**: 2270–2281.

Loftus, B. J., Kim, U. J., Sneddon, V. P., Kalush, F., Brandon, R., Fuhrmann, J., Mason, T., Crosby, M. L., Barnstead, M., Cronin, L., et al. (1999). "Genome duplications and other features in 12 Mb of DNA sequence from human chromosome 16p and 16q." Genomics **60**(3): 295–308.

Logan, C. V., Cossins, J., Cruz, P. M. R., Sheridan, E., Jayawant, S., Palace, J., Johnson, C. A., Beeson, D., et al. (2015). "Congenital Myasthenic Syndrome Type 19 Is Caused by Mutations in COL13A1, Encoding the Atypical Non-fibrillar Collagen Type XIII α1 Chain." The American Journal of Human Genetics. **97**(6): 878–885.

Lupski, J. R. (1998). "Genomic disorders: Structural features of the genome can lead to DNA rearrangements and human disease traits." Trends Genet. **14**: 417–422.

MacDonald, J. R., Ziman, R., Yuen, R. K., Feuk, L., and Scherer, S. W. (2014). "The Database of Genomic Variants: a curated collection of structural variation in the human genome." Nucleic Acids Res. **42**: D986-D992.

Maillard, A. M., Ruef, A., Pizzagalli, F., Migliavacca, E., Hippolyte, L., Adaszewski, S., et al. (2014). "The 16p11.2 locus modulates brain structures common to autism, schizophrenia and obesity." Mol Psychiatry. **20**: 140-147.

Männik, K., Mägi, R., Macé, A., Cole, B., Guyatt, A. L., Timpson, N., Metspalu, A., Reymond, A., et al. (2015). "Copy Number Variations and Cognitive Phenotypes in Unselected Populations." JAMA. **313**(20): 2044–2054.

Martin, J., Han, C., Gordon, L. A., Terry, A., Prabhakar, S., She, X., Xie. G., Hellsten, U., Chan, Y. M., Altherr, M., et al. (2004). "The sequence and analysis of duplication-rich human chromosome 16." Nature. **432**(7020): 988–994.

McCarthy, S. E., Makarov, V., Kirov, G., et al.; Wellcome Trust Case Control Consortium. (2009). "Microduplications of 16p11.2 are associated with schizophrenia." Nat Genet. **41**(11): 1223–1227.

Mefford, H. C., Clauin, S., Sharp, A. J., Moller, R. S., Ullmann, R., Kapur, R., Pinkel, D., Cooper, G. M., Ventura, M., Ropers, H. H., Tommerup, N., Eichler, E. E. and Bellanne-Chantelot, C. (2007). "Recurrent reciprocal genomic rearrangements of 17q12 are associated with renal disease, diabetes, and epilepsy." Am J Hum Genet. **81**(5): 1057–1067.

Mehta, D., Iwamoto, K., Ueda, J., Bundo, M., Adati, N., Kojima, T and Kato, T. (2014). "Comprehensive survey of CNVs influencing gene expression in the human brain and its implications for pathophysiology." Neurosci Res. **79**: 22–23.

Merla, G., Howald, C., Henrichsen, C.N., Lyle, R., Wyss, C., Zabot, M.T., Antonarakis, S.E., and Reymond, A. (2006). "Submicroscopic deletion in patients with Williams-Beuren syndrome influences expression levels of the nonhemizygous flanking genes." Am J Hum

Genet**. 79**: 332–341.

Migliavacca, E., Golzio, C., Mannik, K., Katsanis, N., Reymondet, A., et al. (2015). "A Potential Contributory Role for Ciliary Dysfunction in the 16p11.2 600 kb BP4-BP5 Pathology." American journal of human genetics. **96**(5): 784–796.

Miller, M., Young G. S., Hutman, T., Johnson, S., Schwichtenberg, A. J. and Ozonoff, S. (2015). "Early pragmatic language difficulties in siblings of children with autism: implications for DSM-5 social communication disorder?" J Child Psychol Psychiatry. **56**(7): 774–781.

Morris, D. L., Cho, K. W., Zhou, Y., and Rui, L. (2009). "Sh2b1 enhances insulin sensitivity by both stimulating the insulin receptor and inhibiting tyrosine dephosphorylation of insulin receptor substrate proteins." Diabetes. **58**: 2039–2047.

Nair, K. G., Shalia, K. K., Ashavaid, T. F. and Dalal, J. J. (2003). "Coronary heart disease, hypertension, and angiotensinogen gene variants in Indian population." J Clin Lab Anal **17**(5): 141–146.

Nik-Zainal, S., Strick, R., Storer, M., Huang, N., Rad, R., Willatt, L., Fitzgerald, T., Martin, V., Sandford, R. and Carter, N.P. (2011). "High incidence of recurrent copy number variants in patients with isolated and syndromic Müllerian aplasia." J Med Genet. **48**: 197–204.

Norton, N., Li, D., Rieder, M. J., Siegfried, J. D., Rampersaud, E., Züchner, S., Mangos, S., Gonzalez-Quintana, J., Wang, L., McGee, S., Reiser, J., Martin, E., Nickerson, D. A. and Hershberger, R. E. (2011). "Genome-wide studies of copy number variation and exome sequencing identify rare variants in BAG3 as a cause of dilated cardiomyopathy." Am J Hum. Genet. **88**(2): 273–282.

Parks, M. M., Lawrence, C. E. and Raphael, B. J. (2015). "Detecting non-allelic homologous recombination from high-throughput sequencing data." Genome Biology. **16**(72).

Pebrel-Richard, C., Debost-Legrand, A., Eymard-Pierre, E., Tchirkov, A., Vago, F., Goumy, C., Francannet., et al. (2014). "An unusual clinical severity of 16p11.2 deletion syndrome caused by unmasked recessive mutation of CLN3." Eur J Hum Genet. **22**(3): 369–373.

Pinhoa, E., Gomesa, A. A., Silvaa, M. J., Torresa, T. P., Coelhoa, A., Almeidab, P. B., Lourencoa, P. and Bettencourta, P. (2013). "Atrial Septal Defect in a Very Old Woman." Cardiology Research. **4**(1): 41–44.

Pirooznia, M., Kramer, M., Parla, J., Goes, F. D., Potash, J. B., McCombie, W. R. and Zandi, P. P (2014). "Validation and assessment of variant calling pipelines for next-generation sequencing." Hum Genomics. **8**(1): 14.

Prakash, S. K., LeMaire, S. A., Guo, D. C., Russell, L., Regalado, E. S., Golabbakhsh, H., Johnson, R. J., Safi, H. J., Estrera, A. L., Coselli, J. S., Bray, M. S., Leal, S. M., Milewicz, D. M., Belmont, J. W. (2010). "Rare copy number variants disrupt genes regulating vascular smooth muscle cell adhesion and contractility in sporadic thoracic aortic aneurysms and dissections." Am J Hum Genet. **87**(6): 743–756.

Qureshi, A. J., Mueller, S., Snyder, A. Z., Mukherjee, P., Berman, J. I., and T. P. L. Roberts, Nagarajan, S. S., Spiro, J. E., Chung, W. K., Sherr, E. H. and Buckner R. L. (2014). "Opposing Brain Differences in 16p11.2 Deletion and Duplication Carriers." Neurobiology of Disease. **34**(34): 11199–11211.

Reinthaler, E. M., Lal, D., Lebon, S., Jacquemont, S., Reymond, A., Zimprich, F., Neubauer, B. A., et al. (2014). "16p11.2 600 kb Duplications confer risk for typical and atypical Rolandic epilepsy." Human Molecular Genetics. **23**(22): 6069–6080.

Reymond, A., Henrichsen, C.N., Harewood, L., and Merla, G. (2007). "Side effects of genome structural changes." Curr Opin Genet Dev. **17**: 381–386.

Ricard, G., Molina, J., Chrast, J., Wenli, G., Gheldof, N., Pradervand, S., Schütz, F., Young, J. I., Lupski, J. R., Reymond, A. and Walz, K. (2010). "Phenotypic Consequences of Copy Number Variation: Insights from Smith-Magenis and Potocki-Lupski Syndrome Mouse Models." PLoS Biology. **8**(11): 1-12.

Rosenfeld, M. J. (2010). "Nontraditional families and childhood progress through school." Demography. **47**(3): 755–775.

Sampson, M. G., Coughlin, C. R., 2nd, Kaplan, P., Conlin, L. K., Meyers, K.E., Zackai, E. H., Spinner, N. B. and Copelovitch, L. (2010). " Evidence for a recurrent microdeletion at chromo- some 16p11.2 associated with congenital anomalies of the kidney and urinary tract (CAKUT) and Hirschsprung disease." Am J Med Genet A. **152A**: 2618–2622.

Sandbacka, M., Laivuori, H., Freitas, É., Halttunen, M., Jokimaa, V., Morin-Papunen, L., Rosenberg, C., Aittomäki, K. (2013). "TBX6, LHX1 and copy number variations in the complex genetics of Müllerian aplasia." Orphanet J Rare Dis. **8**: 125.

Santoro, S. D., Giacheti, C. M., Rossi, N. F., Campos, L. M. G. and Pinato, L. (2016). "Correlations between behavior, memory, sleep-wake and melatonin in Williams-Beuren syndrome." Physiology & Behavior. **159**: 14–19.

Scherer, S. W., Lee, C., Birney E., Altshuler, D. M., Eichler, E. E., Carter, N. P., Hurles, M. E., and Feuk, L.Scherer, S. W., Lee, C., Birney E., Altshuler, D. M., Eichler, E. E., Carter, N. P., Hurles, M. E., and Feuk, L. (2007). "Challenges and standards in integrating surveys of structural variation." Nature Genetics Supplement. **39**.

Shinawi, M., Liu, P., Kang, S. H., Shen, J., Belmont, J. W., Scott, D. A., Probst, F. J., Craigen, W. J., Graham, B. H., Pursley, A., et al. (2010). "Recurrent reciprocal 16p11.2 rearrangements associ- ated with global developmental delay, behavioural problems, dysmorphism, epilepsy, and abnormal head size." J. Med. Genet. **47**: 332–341.

Stankiewicz, P., and Lupski, J. R. (2002). "Genome architecture, rearrangements and genomic disorders." Trends Genet. **18**: 74–82.

Stefansson, H., Meyer-Lindenberg, A., Steinberg, S., Magnusdottir, B., Morgen, K., Arnarsdottir, S., et al. (2014). "CNVs conferring risk of autism or schizophrenia affect cognition in controls." Nature. **505**: 361–366.

Stoll, G., Pietiläinen, O. P. H., Linder, B., Suvisaari, J., Brosi, C., Salomaa, V., Hurles, M., Stefansson, H., et al. (2013). "Deletion of TOP3β, a component of FMRP-containing mRNPs, contributes to neurodevelopmental disorders." Nature Neuroscience. **16**: 1228–1123.

Stranger, B. E., Forrest, M.S., Dunning, M., Ingle, C.E., Beazley, C., Thorne, N., Redon, R., Bird, C.P., de Grassi, A., Lee, C., et al. (2007). "Relative impact of nucleotide and copy number variation on gene expression phenotypes." Science. **315**: 848–853.

Tang, L., Ye, H., Hong, Q., Chen, F., Mai, Y., Duan, S., et al. (2014). "Meta-analyses between 18 candidate genetic markers and overweight/obesity." Diagnostic Pathology. **9**: 56.

Tewes, A. C., Rall, K. K., Romer, T., Hucke, J., Wieacker, P., Ledig, S., et al. (2015). "Variations in RBM8A and TBX6 are associated with disorders of the mullerian ducts." American Society for Reproductive Medicine. **103**(5).

van Dam, T. J., Wheway, G., Slaats, G. G., Huynen, M. A., and Giles, R. H.; SYSCILIA Study Group (2013). "The SYSCILIA gold standard (SCGSv1) of known ciliary components and its applications within a systems biology consortium." Cilia. **2**: 7.

Walters, R. G., Jacquemont, S., Valsesia, A., Walley, A. J., Coin, L. J. M., Blakemore, A. I. F., Froguel, P., Beckmann, J. B., et al. (2010). "A novel highly-penetrant form of obesity due to microdeletions on chromosome 16p11.2." Nature. **463**: 671–675.

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F. A., Hakonarson, H. and Bucan, M. (2007). "PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data." Genome Res. **17**(111): 1665–1674.

Watabe-Rudolph, M., Schlautmann, N., Papaioannou, V. E., and Gossler, A. (2002). "The mouse rib-vertebrae mutation is a hypomorphic Tbx6 allele." Mech Dev. **119**: 251–256.

Weckselblatt, B., Rudd, M. K. (2015). "Human Structural Variation: Mechanisms of Chromosome Rearrangements." Trends in Genetics. **31**(10): 587–599.

Weger, W. (2010). "Current status and new developments in the treatment of psoriasis and psoriatic arthritis with biological agents." Br J Pharmacol. **160**(4): 810–820.

Weiss, L. A., Shen, Y., Korn, J.M., Arking, D.E., Miller, D.T., Fossdal, R., Saemundsen, E., Stefansson, H., Ferreira, M.A.R., Green, T., et al.; Autism Consortium. (2008). "Association between microdeletion and microduplication at 16p11.2 and autism." N Engl J Med. **358**: 667–675.

Weksberg, R., Shuman, C. and Beckwith, J. B. (2010). "Beckwith-Wiedemann syndrome." Eur J Hum Genet. **18**: 8–14.

White, P. H., Farkas, D. R., McFadden, E. E., and Chapman, D. L. (2003). "Defective somite patterning in mouse embryos with reduced levels of Tbx6." Development. **130**: 1681–1690.

Williams, N. M., Zaharieva, I., Martin, A., Owen, M. J., O'Donovan, M., Thapar, A., et al. (2010). "Rare chromosomal deletions and duplications in attention-deficit hyperactivity disorder: a genome-wide analysis." Lancet. **376**: 1401–1408.

Wu, N., Ming, X., Xiao, J., Wu, Z., Chen, X., Shinawi, M., Shen, Y., et al. (2015). "TBX6 Null Variants and a Common. Hypomorphic Allele in Congenital Scoliosis." N Engl J Med. **372**: 341–350.

Xu, H., Luo, X., Chen, S., et al. (2012). "FastUniq: A Fast De Novo Duplicates Removal Tool for Paired Short Reads." PLoS One. **7**(12): e52249.

Yadav, D. K., Beniwal, M. K., and Jain, A. (2013). "Bardet-Biedl Syndrome - A Rare Cause of Cardiomyopathy." Indian Pediatrics. **50**(6): 599–601.

Zarrei, M., MacDonald, J. R., Merico, D., and Scherer, S. W. (2015). "A copy number variation map of the human genome." NATURE REVIEWS | GENETICS. **16**: 172–183.

Zufferey, F., Sherr, E. H., Beckmann, N. D., Hanson, E., Maillard, A. M., Hippolyte, L. et al. (2012). "A 600 kb deletion syndrome at 16p11.2 leads to energy imbalance and neuropsychiatric disorders." J Med Genet. **49**: 660–668.

**Web-based references**

http://www.cancer.sanger.ac.uk/cosmic
http://www.bioinfo.ut.ee/genometester/
http://www.bioinfo.ut.ee/primer3-0.4.0/
http://www.cadd.gs.washington.edu/info
http://www.geenivaramu.ee/en)
http://www.hgmd.cf.ac.uk/ac/index.php
http://www.illumina.com
http://www.ncbi.nlm.nih.gov/SNP/
http://www.omim.org
http://www.saphetor.com
http://www.vital-it.ch
http://www.who.int/classifications/icd
http://www.who.int/mediacentre/factsheets/fs311/en/
https://www.decipher.sanger.ac.uk/
https://www.github.com/varapp
https://www.broadinstitute.org/gatk/
https://www.broadinstitute.org/gatk/guide/best-practces.php
https://www.clinicalgenome.org/data-sharing/clinvar/
https://www.decipher.sanger.ac.uk/
https://www.ebi.ac.uk/gwas/
http://www.exac.broadinstitute.org
https://www.biogazelle.com/qbaseplus
https://www.r-project.org

# 6 SUPPLEMENTARY MATERIALS

Table S1: Abbreviations for genes described in this thesis

| GENE symbol | Gene Name | Entrez Gene (GeneID) | Chromosomal location |
|---|---|---|---|
| *ACVR2B* | Activin A Receptor, Type IIB | 93 | 3p22 |
| *AGT* | Angiotensinogen | 183 | 1q42.2 |
| *ALDOA* | Aldolase, Fructose-Bisphosphate Ap | 226 | 16p11.2 |
| *AMH* | Anti-Mullerian Hormone | 268 | 19p13.3 |
| *AMHR2* | Anti-Mullerian Hormone Type 2 Receptor | 110542 | 15 F3 |
| *ASPHD1* | Aspartate Beta-Hydroxylase Domain | 253982 | 16p11.2 |
| *ATP2A1* | ATPase sarcoplasmic/endoplasmic reticulum Ca2+ transporting 1 | 487 | 16p12.1 |
| *ATXN2L* | ataxin 2 like | 11273 | 16p11 |
| *BBS5* | Bardet-Biedl syndrome 5 | 129880 | 2q31.1 |
| *BBS9* | Bardet-Biedl syndrome 9 | 27241 | 7p14 |
| *BMPR1a* | bone morphogenetic protein receptor type 1A | 657 | 10q22.3 |
| *BOLA2/2B* | bolA family member 2 | 552900 | 16p11.2 |
| *C16orf54* | chromosome 16 open reading frame 54 | 283897 | 16p11.2 |
| *C16orf92* | chromosome 16 open reading frame 92 | 146378 | 16p11.2 |
| *C16orf53* | chromosome 16 open reading frame 92 | 79447 | 16p11.2 |
| *CC2D2A* | coiled-coil and C2 domain containing 2A | 57545 | 4p15.32 |
| *CCDC40* | coiled-coil domain containing 40 | 55036 | 17q25.3 |
| *CD19* | CD19 molecule | 930 | 16p11.2 |
| *CDH5* | cadherin 5 | 1003 | 16p11.2 |
| *CDIPT* | CDP-diacylglycerol--inositol 3-phosphatidyltransferase | 10423 | 16p11.2 |
| *CEP120* | centrosomal protein 120 | 153241 | 5q23.2 |
| *CEP152* | centrosomal protein 152 | 22995 | 15q21.1 |
| *CFC1* | cripto, FRL-1, cryptic family 1 | 55997 | 2q21.1 |
| *CHD7* | chromodomain helicase DNA binding protein 7 | 55636 | 8q12.2 |
| *CLNK* | cytokine dependent hematopoietic cell linker | 116449 | 4p16.1 |
| *CORO1A* | coronin 1A | 11151 | 16p11.2 |
| *CRELD1* | cysteine rich with EGF like domains 1 | 78987 | 3p25.3 |
| *DOC2A* | double C2 domain alpha | 8448 | 16p11.2 |
| *ECI1* | enoyl-CoA delta isomerase 1 | 1632 | 16p13.3 |
| *ESR1* | estrogen receptor 1 | 2099 | 6q25.1 |
| *F5* | coagulation factor V | 2153 | 1q23 |
| *FAM57B* | family with sequence similarity 57 member B | 83723 | 16p11.2 |
| *FOXH1* | forkhead box H1 | 8928 | 8q24.3 |
| *GATA4* | GATA binding protein 4 | 2626 | 8p23.1-p22 |
| *GDF1* | growth differentiation factor 1 | 2657 | 19p12 |
| *GDPD3* | glycerophosphodiester phosphodiesterase domain containing 3 | 79153 | 16p11.2 |
| *HIRIP3* | HIRA interacting protein 3 | 8479 | 16p11.2 |
| *HNF1B* | HNF1 homeobox B | 6928 | 17q12 |
| *HOXA5* | homeobox A5 | 3202 | 7p15.2 |
| *HOXA9* | homeobox A9 | 3205 | 7p15.2 |
| *IGF2* | insulin like growth factor 2 | 3481 | 11p15.5 |
| *INO80E* | INO80 complex subunit E | 283899 | 16p11.2 |
| *IRS1* | insulin receptor substrate 1 | 3667 | 2q36 |
| *IRS2* | insulin receptor substrate 2 | 8660 | 13q34 |
| *KAT6B* | lysine acetyltransferase 6B | 23522 | 10q22.2 |
| *KCTD13* | potassium channel tetramerization domain containing 13 | 253980 | 16p11.2 |
| *KIF22* | kinesin family member 22 | 3835 | 16p11.2 |
| *LAT* | linker for activation of T-cells | 27040 | 16p11.2 |
| *KIAA0556* | KIAA0556 | 23247 | 16p12.1 |

| | | | |
|---|---|---|---|
| *LHX1* | LIM homeobox 1 | 3975 | 17q12 |
| *MAPK3* | mitogen-activated protein kinase 3 | 5595 | 16p11.2 |
| *MAPT* | microtubule associated protein tau | 4137 | 17q21.1 |
| *MAZ* | MYC associated zinc finger protein | 4150 | 16p11.2 |
| *MDM4* | MDM4, p53 regulator | 4194 | 1q32 |
| *MVP* | major vault protein | 9961 | |
| *NFATC2IP* | nuclear factor of activated T-cells 2 interacting protein | 84901 | 16p11.2 |
| *NIPBL* | NIPBL, cohesin loading factor | 25836 | 5p13.2 |
| *NKX2-5* | NK2 homeobox 5 | 1482 | 5q34 |
| *NODAL* | nodal growth differentiation factor | 4838 | 10q22.1 |
| *NPIP* | nuclear pore complex interacting protein family member A1 pseudogen | 730153 | 16p11.2 |
| *NR2F2* | nuclear receptor subfamily 2 group F member 2 | 7026 | 15q26 |
| *PAGR1* | PAXIP1 associated glutamate rich protein 1 | 79447 | 16p11.2 |
| *PGR* | progesterone receptor | 5241 | 11q22-q23 |
| *PPP4C* | protein phosphatase 4 catalytic subuni | 5531 | 16p11.2 |
| *PRRT2* | proline rich transmembrane protein 2 | 112476 | 16p11.2 |
| *PTPN11* | protein tyrosine phosphatase, non-receptor type 11 | 5781 | 12q24 |
| *QPRT* | quinolinate phosphoribosyltransferase | 23475 | 16p11.2 |
| *RABEP2* | rabaptin, RAB GTPase binding effector protein 2 | 79874 | 16p11.2 |
| *SEZ6L2* | seizure related 6 homolog like 2 | 26470 | 16p11.2 |
| *SH2B1* | SH2B adaptor protein 1 | 25970 | 16p11.2 |
| *SLX1A* | SLX1 homolog A, structure-specific endonuclease subunit | 548593 | 16p11.2 |
| *SLX1B* | SLX1 homolog B, structure-specific endonuclease subunit | 79008 | 16p11.2 |
| *SPN* | sialophorin | 6693 | 16p11.2 |
| *SPNS1* | SPNS sphingolipid transporter 1 (putative) | 83985 | 16p11.2 |
| *SULT1A3/4* | sulfotransferase family 1A member 1?? | | |
| *TAOK2* | TAO kinase 2 | 9344 | 16p11.2 |
| *TBX20* | T-box 20 | 57057 | 7p14.3 |
| *TBX5* | T-box 5 | 6910 | 12q24.1 |
| *TBX6* | T-box 6 | 6911 | 16p11.2 |
| *TLL1* | tolloid like 1 | 7092 | 4q32.3 |
| *TMEM219* | transmembrane protein 219 | 124446 | 16p11.2 |
| *TUFM* | Tu translation elongation factor, mitochondrial | 7284 | 16p11.2 |
| *WISP2* | WNT1 inducible signaling pathway protein 2 | 8839 | 20q13.12 |
| *WT1* | Wilms tumor 1 | 7490 | 11p13 |
| *YPEL3* | yippee like 3 | 83719 | 16p11.2 |
| *ZFPM2* | zinc finger protein, FOG family member 2 | 23414 | 8q23 |
| *ZG16* | zymogen granule protein 16 | 653808 | 16p11.2 |
| *ZIC3* | Zic family member 3 | 7547 | Xq26.2 |

Table S2. Primer pairs designed for copy number profiling using real-time quantitative PCR.

| Oligo name | Sequence 5'-3' | Product lenght | Position on genome |
|---|---|---|---|
| Primer pair I | CCATGGACAAATATTGCCAGCC | 141bp | chr16:29337336 |
| | TTTCCGACGTGGAATGCAGA | | chr16:29337476 |
| Primer pair II | AAGCCATTGTTGCGCTCATTC | 121bp | chr16:29646308 |
| | TTGCTTGTAATCGCACACCCA | | chr16:29646428 |
| Primer pair III | TGAGTGCCCCATCAGTTGTTCA | 110bp | chr16:29653173 |
| | TAGCAGCCATCACAGGTTTCCA | | chr16:29653282 |
| Primer pair IV | TTCCATAAGAACTCAGCCCGCA | 60bp | chr16:29855950 |
| | TTCCGAGGTCTCAAAGCCAAAGA | | chr16:29856009 |
| Primer pair V | TGCTGTGCCATGGACTGTGATT | 95bp | chr16:29991079 |
| | TGCCTGCCTGGTTTCCCTAAAT | | chr16:29991173 |
| Primer pair VI | TTGGCCAGCTTTGAATCCCA | 61bp | chr16:30014875 |
| | AAATGACAACGCTGAGGTCGCA | | chr16:30014935 |
| Primer pair VII | TTCATGTCGGCAGCCTGAAACT | 126bp | chr16:30117626 |
| | AGGGTGGTGCTGTTAAGTGGTCAA | | chr16:30117751 |
| Primer pair VIII | CGTCGTCGCATATTCGAAAGGA | 98bp | chr16:30177171 |
| | TCCTGGTCATCAAATCAGGCAA | | chr16:30177268 |
| Primer pair IX | TGAGCCAACTGCAGGATGAAGA | 56bp | chr16:30179542 |
| | CTGGGATGCTGACACTTTCCAA | | chr16:30179597 |
| Primer pair X | TTTCACAGCAGCTCCAAGGTGA | 95bp | chr16:30191225 |
| | TGTGCTTCCCAACCAAAGCA | | chr16:30191319 |
| Primer pair XI | TTTTCCTCAAACGCCCACTTCC | 110bp | chr16:30593784 |
| | TGATGGCGTTCACACAAACTGG | | chr16:30593883 |
| Primer pair XII | ACACAGCGCAACAAACAGCCTT | 74bp | chr16:30722099 |
| | TGACCTTCCTCTTCGTTGGCAGTA | | chr16:30722172 |
| Reference primer pair I | TGCGAAACTGCGTGGACATT | 70bp | chr1:28661159 |
| | ATGCGGAAGCCCATTTCCAT | | chr16:28661228 |
| Reference primer pair II | CTGTGACCTGCAGCTCATCCT | 120bp | chr3:113954875 |
| | TAAGTTCTCTGACGTTGACTGATGTG | | chr16:113954994 |

| Family | V code | Exom ID | 16p11.2 CNV | Sex | Age | BMI | Characteristic features to the 16p11.2 del/dup syndrome | Other features |
|---|---|---|---|---|---|---|---|---|
| 1 | V09503 | EX99 | DUP | F | 24 | 18.6 | Mild mental retardation, significant impairment of behavior requiring attention or treatment; Recurrent depressive disorder, current episode moderate; Dysthymia; Anxiety disorder, unspecified; Obsessive-compulsive disorder; Adjustment disorders; Intentional self-harm | Cerebral palsy; Hypotension; Cystitis; Non-inflammatory disorders of ovary, fallopian tube and broad ligament; Absent, scanty and rare menstruation; Excessive, frequent and irregular menstruation; Congenital malformations of ovaries, fallopian tubes and broad ligaments; Congenital obstructive defects of renal pelvis and congenital malformations of ureter; Functional dyspepsia |
| 2 | V09186 | EX100 | DUP | F | 20 | 23.7 | Epilepsy: Localization-related (focal)(partial) symptomatic epilepsy and epileptic syndromes with simple partial seizures; Petit mal status epilepticus; Congenital malformations of cardiac septa: Atrial septal defect; Congenital malformations of uterus and cervix: Doubling of uterus with doubling of cervix and vagina | Disorders of thyroid gland; Angina pectoris; Atrioventricular and left bundle-branch block; Other cardiac arrhythmias; Asthma; Seropositive rheumatoid arthritis; Dorsalgia; Cystitis; Inflammatory diseases of female pelvic organs; Absent, scanty and rare menstruation; Acute bronchitis; Hypertensive heart disease; Tension-type headache |
| 3 | V21252 | EX101 | DEL | F | 50 | 43.8 | Obesity | Gastritis and duodenitis; Psoriasis; Seropositive rheumatoid arthritis; Deforming dorsopathies; Dorsalgia;Spinal osteochondrosis; Cystitis; Acute bronhities; Type 2 diabetes mellitus; Essential (primary) hypertension; Hypercholesterolemia |
| 4 | V02715 | EX102 | DEL | M | 67 | 51.9 | Depressive episode; Specified forms of tremor; Unspecified epilepsy; Obesity: due to excess calories | Malignant neoplasm of breast; Malignant neoplasm of corpus uteri; Other extrapyramidal and movement disorders; Oesophagitis; Gastro-oesophageal reflux disease; Dermatitis and eczema; Deforming dorsopathies; Dorsalgia; Spinal osteochondrosis; Congenital leukonychia; Acute bronhities; Essential (primary) hypertension; Chronic ischaemic heart disease |
| 5 | V22648 | EX103 (father) | DUP | M | 67 | 26.0 | | Conductive and sensorineural hearing loss; Other rheumatic heart diseases; Other cardiac arrhythmias; Vasomotor and allergic rhinitis; Asthma; Duodenal ulcer; Psoriasis; Dorsalgia; Acute bronhities; Dyslipidemi; Mixed hyperlipidaemia; Essential (primary) hypertension; Chronic ischaemic heart disease |
| | V22673 | EX104 (child) | Control | NA | NA | 26.2 | Mild depressive episode | Allergic contact dermatitis; Dorsalgia |
| 6 | V28938 | EX105 (mother) | DUP | F | 42 | 19.8 | Post-traumatic stress disorder; Neurotic disorder, unspecified | Leiomyoma of uterus; Hypotension; Vasomotor and allergic rhinitis; Oesophagitis; Gastritis and duodenitis; Duodenal ulcer; Gastro- |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | | oesophageal reflux disease; Dorsalgia; Other inflammation of vagina and vulva; Non-inflammatory disorders of ovary, fallopian tube and broad ligament; Syncope and collapse; Acute bronchitis; Transient cerebral ischaemic attacks and related syndromes |
| | V44037 | EX106 (child) | Control | M | 22 | 22.8 | | Dorsalgia |
| 7 | V33975 | EX107 | DEL | M | | 24.15 | Mild mental retardation; Significant impairment of behavior requiring attention or treatment; Persistent delusional disorders | Functional dyspepsia; Duodenal ulcer; Gout; Hypertensive heart disease; Hypertensive heart and renal disease; Angina pectoris |
| 8 | V35427 | EX108 | DUP | M | 37 | 24.2 | | Conductive and sensorineural hearing loss; Functional dyspepsia; Essential (primary) hypertension |
| | V35633 | EX109 (father) | Control | M | 72 | 25.1 | | Dorsalgia; Conductive and sensorineural hearing loss; Transient cerebral ischaemic attacks and related syndromes: Vertebro-basilar artery syndroms; Hypertensive heart disease; Gastro-oesophageal reflux disease; Gout; Syncope and collapse |
| 9 | V43909 | EX110 | DEL | M | 32 | 30.1 | Neurotic disorder; Moderate mental retardation; Mild mental retardation, Significant impairment of behavior requiring attention or treatment; Scoliosis | Functional dyspepsia; Acute bronhities |
| 10 | V13801 | EX111 (father) | DUP | M | 59 | 25.5 | | Melanocytic naevi; Vasomotor and allergic rhinitis; Gastritis and duodenitis; Other dermatitis; Dorsalgia; Acute bronchitis; Essential (primary) hypertension; Angina pectoris |
| | V26590 | EX112 (mother) | Control | NA | NA | 24.1 | Other mood (affective) disorders; Other anxiety disorders | Dorsalgia; Nonorganic sleep disorders; Other headache syndromes; Transient cerebral ischaemic attacks and related syndromes; Angina pectoris; Other cardiac arrhythmias; Complications and ill-defined descriptions of heart disease; Gastro-oesophageal reflux disease; Gastritis and duodenitis; Psoriasis; Other dermatitis; Other arthritis; Cystitis |
| | V13124 | EX113 (proband) | DUP | F | 37 | 18.8 | Moderate depressive episode; Other anxiety disorders; Ovarian dysfunction | Dorsalgia; Cystitis; Inflammatory diseases of female pelvic organs; Endometriosis of uterus; Female infertility; Excessive, frequent and irregular menstruation; Acute bronhities; Somatoform autonomic dysfunction; Essential (primary) hypertension |
| | V11819 | EX114 (sibling) | Control | M | 35 | 19.5 | Depressive episode | Dorsalgia; Allergic contact dermatitis; Other arthritis; Other inflammation of vagina and vulva |
| 11 | V11257 | EX115 (proband) | DEL | F | 40 | 37.2 | | Allergic contact dermatitis; Dorsalgia; Inflammatory disease of cervix uteri; Acute bronchitis; Other headache syndromes |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | V12736 | EX116 (mother) | Control | F | 67 | 38.8 | | Dorsalgia; Nonorganic sleep disorders; Hypertensive heart disease |
| | V40443 | EX117 (mat uncle) | Control | NA | NA | 35.9 | | Mononeuropathies of upper limb; Dorsalgia; Hypertensive heart disease with (congestive) heart failure; Essential (primary) hypertension |
| 12 | V14358 | EX118 | DUP | F | 43 | 26.0 | Undifferentiated schizophrenia: Paranoid schizophrenia, Hebephrenic schizophrenia; Delusional disorder; Acute and transient psychotic disorders: Acute polymorphic psychotic disorder with symptoms of schizophrenia, Acute and transient psychotic disorder, unspecified; Moderate depressive episode; Recurrent depressive disorder | |
| 13 | V12867 | EX119 (proband) | DEL | F | 35 | 35.2 | Reaction to severe stress, and adjustment disorders; Adjustment disorders; Mild mental retardation -> With the statement of no, or minimal, impairment of behavior; Congenital malformations of uterus and cervix-> Other doubling of uterus, Bicornate uterus; Obesity; Ovarian dysfunction | Leiomyoma of uterus; Benign neoplasm of ovary; Ovarian dysfunction; Other rheumatic heart diseases; Cardiomyopathy; Paroxysmal tachycardia; Functional dyspepsia; Seborrhoeic dermatitis; Vulvovaginal ulceration and inflammation in diseases classified elsewhere; Other inflammation of vagina and vulva; Non-inflammatory disorders of ovary, fallopian tube and broad ligament; Absent, scanty and rare menstruation; Female infertility; Excessive, frequent and irregular menstruation; Essential (primary) hypertension |
| | V38423 | EX120 (sibling) | Control | F | | 20.6 | Recurrent depressive disorder, current episode moderate, current episode severe without psychotic symptoms; Neurotic disorder, unspecified; Scoliosis | Trichomoniasis;<br>Dorsalgia; Cystitis; Inflammatory diseases of female pelvic organs |
| 14 | V03575 | EX121 | DUP | F | 70 | 22.7 | Other specified congenital malformations of integument; Chronic ischaemic heart disease | Malignant neoplasm of breast; Malignant neoplasm of corpus uteri; Oesophagitis;<br>Dermatitis and eczema; Acute bronhities;<br>Infections with a predominantly sexual mode of transmission |

Table S3. EGC UT cohort carriers and intra-familial controls detailed clinical information.

Table S4. A list of genes underlying syndromes associated with AVSD.

| Genes associated with AVSD<br><br>(D'Alessandro et al., 2016) |
| --- |
| *ANKRD11 ARL6 ATR BBS1 BBS10 BBS12 BBS2 BBS4 BBS5 BBS7 BBS9 BRAF CDKN1C CENPJ CEP152 CEP290 CEP63 CHD7 COL1A1 COL1A2 COL2A1 COL5A1 COL5A2 DHCR7 DOK7 ELN EVC EVC2 GDF6 GLI3 HYLS1 IRF6 KRAS MAP2K1 MDM4 MKKS MKS1 NIPBL NPHP4 NRAS OFD1 PTPN11 RAF1 RAPSN RBBP8 RECQL4 SETBP1 SMC1A SMC3 SOS1 TBX1 TBX5 TFAP2A TRIM32 TTC8 WNK1 WNK4 ZFPM2 ZIC3* |

AVSD: an atrioventricular septal defect

Table S5. A list of ciliopathy-associated genes used in the exome analysis in Saphetor platform.

| Ciliopathy-associated genes<br><br>(van Dam et al., 2013) |
| --- |
| *ACE AGT AGTR1 AHI1 AIPL1 ALMS1 ANKS6 APC2 ARL13B ARL6 ARMC4 ARVCF B9D1 B9D2 BAZ1B BBIP1 BBS1 BBS10 BBS12 BBS2 BBS4 BBS5 BBS7 BBS9 BMPER C21orf59 C2CD3 C5orf42 CC2D2A CCDC103 CCDC114 CCDC151 CCDC28B CCDC39 CCDC40 CCDC65 CCNO CDC73 CDKN1C CENPF CEP120 CEP164 CEP290 CEP41 CLIP2 COMT CPT2 CRB1 CRX CSPP1 DNAAF1 DNAAF2 DNAAF3 DNAAF5 DNAH11 DNAH5 DNAI1 DNAI2 DNAL1 DRC1 DYNC2H1 DYX1C1 ELN ESCO2 ETFA ETFB ETFDH EYA1 FLNB GDF6 GLIS2 GP1BB GTF2I GTF2IRD1 GUCY2D H19 HIRA HYDIN IFT140 IFT172 IFT27 IFT80 IMPDH1 INPP5E INVS IQCB1 KAT6B KCNJ13 KCNQ1OT1 KIAA0556 KIAA0586 KIF7 LCA5 LIMK1 LRAT LRP5 LRRC6 LZTFL1 MKKS MKS1 NEK1 NEK8 NME8 NMNAT1 NOTCH2 NPHP1 NPHP3 NPHP4 NSD1 OCRL OFD1 PEX1 PEX10 PEX11B PEX12 PEX13 PEX14 PEX16 PEX19 PEX2 PEX26 PEX3 PEX5 PEX6 PKD1 PKD2 PKHD1 PRKCSH RD3 RDH12 REN RFC2 RPE65 RPGR RPGRIP1 RPGRIP1L RSPH1 RSPH3 RSPH4A RSPH9 SDCCAG8 SEC63 SETD2 SHANK3 SIX1 SIX5 SPAG1 SPATA7 TBL2 TBX1 TCTN1 TCTN2 TCTN3 TMEM138 TMEM216 TMEM231 TMEM237 TMEM67 TRIM32 TSC1 TSC2 TTC21B TTC8 TULP1 UFD1L VHL WDPCP WDR19 WDR34 WDR35 WDR60 ZMYND10 ZNF423* |

Table S6. 16p11.2 dosage-dependently differently expressed genes in human transcriptome used in the exome-analysis in Saphetor platform.

| 16p11.2 dosage-dependently differently expressed genes in human transcriptome<br><br>(Migliavacca et al., 2015) |
| --- |
| *AAAS AAMP ABHD13 ABHD2 ABHD3 ABI1 ACADSB ACBD5 ACER3 ACSL4 ACTL6A ADAM10 ADAM9 ADCY7 ADK AEBP2 AFF4 AGFG1 AGGF1 AGL AGTPBP1 AHNAK AHR AKIRIN1 ALAD ALDOA ALG10B ALKBH8 AMBRA1 AMY1A AMY1B AMY1C AMY2A ANAPC10 ANGEL2 ANKIB1 ANKLE2 ANKRD13C ANKRD39 ANKRD52 ANXA11 AP1AR AP1S3 AP2S1 AP3D1 APC API5 APOBEC3C APOOL APPL1 AQR ARF1 ARFIP2 ARID1B ARID2 ARID4B ARL13B ARL17A ARL2 ARL5B ARL6IP6 ARMC8 ARMCX5 ARPC5 ARPC5L ARRB2 ASAP1 ASCC3 ASPM ASXL2 ATAD2 ATF1 ATF2 ATG4B ATL2 ATL3 ATP11B ATP13A1 ATP13A3 ATP5S ATP6V0A2 ATP6V0E1 ATP6V1G2 ATRN ATXN3 ATXN7 AURKAIP1 AURKB AZI2 AZIN1 B3GNT2 B4GALT4 BAZ1B BBS10 BBS4 BBS7 BCAP31 BCL2 BCLAF1 BDP1 BHLHE22 BICD2 BIRC6 BIVM BLVRA BLZF1 BMP7 BMPR2 BOLA2 BOLA2B BRAP BRD9 BRIP1 BRIX1 BSG BTAF1 C17orf75 C17orf80 C19orf12 C19orf43 C1orf109 C1orf162 C2orf69 C3orf17 C3orf38 C4orf46 C5orf24 CALCOCO1 CALM3 CAPN12 CAPNS1 CASC4 CASP6 CASZ1 CBL CBLL1 CBWD2 CBX3 CCAR1 CCDC18 CCDC24 CCDC28B CCDC50 CCDC82 CCDC84 CCDC88A CCDC91 CCDC93 CCNI CCNJ CCNT2 CCT6B CD2AP CD44 CD58 CD63 CD79A CD84 CDC14A CDC73 CDIPT CDK12 CDK16 CDK8 CDV3 CENPK CEP120 CEP170 CEP290 CEP350 CEP57 CGRRF1 CHCHD7 CHD4 CHD6 CHD7 CHD9 CHMP5 CHRNB2 CHST11 CHUK CIB1 CIR1 CISD2 CLIP1 CLN5 CLN6 CMAS CMC1 CMIP CNNM3 CNOT7 CNOT8 CNST COG5 COMMD2 COPE CORO1A CORO7 CPEB4 CPSF2 CPSF3L CPSF6 CR1 CRCP CREB1 CREB5 CREBBP CREBL2 CSGALNACT2 CTH CTSS CTTNBP2NL CWF19L1 CXorf56 CYBRD1 CYFIP1 CYP1A2 CYP51A1 DAPP1 DARS DBR1 DBT DCAF10 DCAF17 DCBLD2 DCTN4 DCUN1D1 DDHD1 DDX18 DDX20 DDX23 DDX27 DDX41 DDX59 DEAF1 DENND4A DENND4B DENND4C DHFRL1 DHPS DHX32 DICER1 DLAT DLG1 DMXL1 DNAJA3 DNAJB6 DNAJC1 DNAJC10 DNAJC16 DNAJC21 DNAJC27 DNAJC4 DPP8 DR1 DTX3 DYNC1LI2 DYRK2 EBF1 EEF1A1 EEF1D EFCAB7 EFR3B EHBP1L1 EHHADH EID1 EIF1 EIF1AX EIF2AK1 EIF2AK2 EIF2S3 EIF3J EIF4E EIF4G1 EIF5 EIF5A ELF1 ELF2 EMILIN2 ENOX2 ENPP5 ENTPD6 EP300 EP400 EPC2 EPRS EPS15 ERCC4 ERGIC1 ERGIC2 ERGIC3 ERI1 ERLIN1 ESF1 ETNK1 EWSR1 EXOC5 EXOC8 F11R FAHD1 FAM73A FAM76A FAM76B FAM92A1 FAN1 FBXO21 FBXO22 FBXO28 FBXO45 FBXW11 FBXW4 FCF1 FEM1B FGFR1OP2 FGFR2 FKBP1A FKBP4 FLI1 FLVCR2 FLYWCH1 FLYWCH2 FNBP4 FNDC3A FRS2 FSCN1 FUBP3 FXR1 FXYD5 FYN FYTTD1 G2E3 G3BP1 GABPA GALNT1 GALNT6 GANC GBA2 GCA GCNT2 GCOM1 GDF11 GDPD3 GET4 GFM1 GGA1 GGA3 GGACT GGPS1 GIGYF1 GINS2 GKAP1 GLG1 GLS GLYR1 GNAI2 GNAS GNB1 GOLGA2 GON4L GOPC GPAA1 GPATCH8 GPR156 GPRASP2 GPX7 GSPT1 GSS GTDC1 GTF2F2 GTSE1 GTSF1 H2AFY HACD3 HAUS2 HAUS6 HCFC2 HDGF HEATR3 HELLS HELZ HEMK1 HERC1 HERC4 HEXA HIPK1 HIRIP3 HOXA7 HSF2 HSP90B1 HTATIP2 ICK IER5 IFIH1 IFRD1 IFT74 IKBIP IL10 IL10RB IL21R ILDR1 ILF3 ILK ILVBL INO80E INPPL1 INSIG1 INTS4 IPMK IPP IRAK1BP1 IREB2 IRF3 IRF8 ITCH ITFG2 ITPR1 IWS1 JDP2 JMJD1C KBTBD2 KBTBD7 KCNQ1OT1 KCTD13 KCTD2 KCTD3 KDM1B KDM2A KDM4A KDM4B KDM5A KDM6A KIF21B KIF22 KLF7 KLHDC1 KLHDC4 KLHL20 KLHL23 KLHL5 KLHL9 KPNB1 KRAS KRIT1 LATS2 LGALS8 LGMN LIG4 LILRB3 LIMS1 LMF2 LRCH3 LRCH4 LRRC43 LRRC57 LRRFIP1 LUZP1 LY75 LYN LYRM5 M6PR MALAT1 MAML2 MAP1LC3A MAP2K3 MAP3K11 MAP3K13 MAP3K2 MAP3K8 MAPK1IP1L MAPK3 MARCKS MARK4 MASTL MATR3 MAZ MBIP MBNL1 MCFD2 MCL1 MDM4 MED1 MED23 MED30 MED4 MED7 MEX3C MGA MGEA5 MIER3 MIR155HG MLEC MOCS2 MORC2 MORC3 MPHOSPH9 MRPL4 MRPL44 MRPL50 MRPS5 MTF1 MTF2 MTM1 MTMR10 MVP MYO3B MYO9A N4BP1 NAA35 NAB1 NANOS1 NAP1L1 NAV1 NBPF1 NCEH1 NCF4 NCOA2 NCOA3 NDUFA10 NDUFC2 NDUFV1 NEDD1 NEK1 NFATC3 NFIC NFKBIB NFXL1 NHLRC3 NIPBL NKIRAS1 NKTR NOL9 NPEPPS NPFF NR1D1 NR2C2 NR3C1 NRBF2 NSF NUCKS1 NUDT19 NUDT21 NUP160 NUP50 NUS1 NUSAP1 ODF2L OGFOD1 OS9 OTUD6B PACRGL PACS2 PAIP1 PAK1 PAK2 PANK2 PAPD4 PAPOLA PARD3 PARD6B PARG PARP11 PBRM1 PCBP2 PCGF5 PDE12 PDE7A PDLIM7 PDS5B PEX1 PEX13 PEX7 PGGT1B PGK1 PGM1 PGM2 PGPEP1 PGRMC2 PHAX PHF3 PHF6 PHYKPL PIAS1 PIAS2 PICALM PIF1 PIGN PIKFYVE PJA1 PKP4 PLA2G12A PLAA PLCL2 PLEKHG4 PLOD3 PLXNC1 PMPCA* |

*PNPLA6 POC1B POLH POLK POLR2E POLR2K POLR3F POLRMT POM121 POM121C POMP POTEKP PPARD PPFIBP1 PPIG PPIL2 PPIL4 PPM1A PPM1B PPP1CB PPP1R10 PPP1R15B PPP1R3D PPP2R5E PPP4C PPP4R2 PPP4R3A PPP6C PREPL PRIM2 PRKAA1 PRKCI PRKD2 PRKY PRMT6 PROSC PRPF39 PRPF40A PRRC2B PSMA2 PSMC2 PSMG4 PSPH PTGR1 PTK2 PTP4A1 PTPN11 PTPN2 PTPN22 PURA PURB PWP1 PXDN PYHIN1 QARS QPRT RAB14 RAB18 RAB1B RAB33B RAB39B RAB4A RAB4B RAB8B RABL2A RABL2B RAD18 RAD21 RAD23A RANBP9 RAP2A RAP2C RASA2 RASGRP3 RASSF6 RB1CC1 RBBP4 RBL2 RBM12B RBM15 RBM17 RBM25 RBM27 RBMS2 RCHY1 RDH11 REEP3 REST RFK RFXAP RHEB RICTOR RIOK3 RLIM RNF10 RNF11 RNF115 RNF123 RNF13 RNF138 RNF213 RNFT1 RNMT ROBO1 RORA RP2 RPAIN RPL13P5 RPL14 RPL15 RPL18A RPL18AP3 RPL31 RPL38 RPL5 RPL7A RPLP2 RPS10 RPS10P7 RPS17 RPS4X RPS4XP6 RRBP1 RREB1 RSF1 RSPH10B2 RSPRY1 RSRC1 RSRC2 RSU1 RUFY3 RWDD3 SAMD4A SAMD9L SAP25 SART3 SBF1 SBF2 SBNO1 SCAMP2 SCAMP3 SCAP SCAPER SCD SCLT1 SCRN3 SDCCAG3 SDHAF3 SEC22B SEC23A SEC23B SEC24D SECISBP2L SEH1L SENP6 SEPSECS SEPW1 SERINC1 SERP1 SERPINB1 SFPQ SFXN1 SGSM2 SH3BGRL SH3RF1 SIAE SIKE1 SIN3B SLAIN1 SLAIN2 SLC16A1 SLC16A7 SLC25A1 SLC30A5 SLC35D1 SLC37A3 SLC38A10 SLC38A9 SLC39A6 SLC9A7 SLX1A SLX1B SMAD2 SMAD4 SMARCA2 SMARCC1 SMARCE1 SMC6 SMCHD1 SMCR8 SMG1 SMG7 SNAP23 SNAPC3 SNRNP70 SNRPE SNRPN SNX19 SNX30 SOCS4 SOCS5 SOD2 SOX15 SP100 SPA17 SPAG1 SPECC1 SPHK2 SPICE1 SPN SPTAN1 SQLE SREBF2 SRPK2 SRRM2 SRSF1 SRSF10 SRSF3 SRSF4 SSB SSBP2 ST6GAL1 STAC2 STAM2 STOML1 STRN STT3B STUB1 STX17 STXBP3 STYX SULT1A1 SULT1A2 SULT1A3 SULT1A4 SUMO1 SUN1 SUPT3H SVIP SYPL1 TAB1 TAB2 TAF2 TANK TAOK2 TATDN3 TAX1BP1 TBCCD1 TBCE TBL1XR1 TCEB1 TCEB3 TET2 TET3 TFCP2 TFIP11 TGFBRAP1 THAP5 THAP6 THOC5 THOC7 THRA THRAP3 THUMPD3 TIA1 TIAL1 TLK1 TLR10 TM2D1 TMCC1 TMED3 TMED7 TMEFF2 TMEM117 TMEM128 TMEM135 TMEM161B TMEM165 TMEM170A TMEM182 TMEM209 TMEM219 TMEM222 TMEM231 TMEM5 TMEM55A TMEM64 TMEM69 TMEM87B TMEM97 TMOD3 TMX4 TNFAIP8 TNFSF11 TNRC6B TOMM22 TOP1 TPR TRAPPC2 TRAPPC9 TRIM11 TRIM14 TRIM32 TRIM34 TRIM5 TRIM6-TRIM34 TRMT5 TSC22D2 TSPYL2 TSR1 TTC3 TTC33 TTC37 TTC39C TTPAL TTYH3 TUBGCP2 TUBGCP6 TULP3 TWF1 TXLNG TXN TXN2 UBA1 UBAP2 UBB UBE2D1 UBE2D3 UBE2Z UBE3A UBL3 UBN1 UBN2 UFM1 UGCG UPF1 UQCRC1 USF2 USP11 USP14 USP28 USP32 USP38 USP42 USP46 UTP23 VAMP4 VCPIP1 VDAC3 VEZT VGLL4 VPS13A VPS13B VPS28 VPS35 WASF2 WBP4 WBSCR16 WDR11 WDR43 WDR53 WDR6 WDR76 WIPF1 WTAP WWP1 XPOT XRCC4 YEATS4 YIPF3 YIPF4 YPEL3 YTHDC1 YTHDC2 YY1 ZBED4 ZBTB1 ZBTB10 ZBTB17 ZBTB21 ZBTB34 ZBTB6 ZDHHC21 ZDHHC3 ZEB1 ZFP1 ZFP30 ZFP36L1 ZFP62 ZFP69 ZFX ZFY ZIK1 ZMYM2 ZMYM5 ZMYM6 ZNF10 ZNF117 ZNF131 ZNF138 ZNF140 ZNF141 ZNF148 ZNF160 ZNF180 ZNF19 ZNF224 ZNF225 ZNF227 ZNF230 ZNF234 ZNF253 ZNF254 ZNF256 ZNF257 ZNF260 ZNF264 ZNF267 ZNF268 ZNF273 ZNF274 ZNF284 ZNF302 ZNF304 ZNF347 ZNF350 ZNF383 ZNF397 ZNF398 ZNF41 ZNF419 ZNF430 ZNF432 ZNF439 ZNF440 ZNF442 ZNF45 ZNF496 ZNF510 ZNF542P ZNF544 ZNF546 ZNF548 ZNF555 ZNF565 ZNF569 ZNF57 ZNF571 ZNF576 ZNF585A ZNF594 ZNF600 ZNF607 ZNF614 ZNF616 ZNF623 ZNF624 ZNF625 ZNF644 ZNF649 ZNF654 ZNF655 ZNF658 ZNF658B ZNF664 ZNF665 ZNF670 ZNF675 ZNF680 ZNF7 ZNF700 ZNF701 ZNF708 ZNF709 ZNF763 ZNF765 ZNF770 ZNF776 ZNF780A ZNF780B ZNF782 ZNF788 ZNF813 ZNF814 ZNF829 ZNF84 ZNF846 ZNF850 ZNF93 ZSCAN12 ZSCAN30*