
UNIVERSITÀ DEGLI STUDI DI TRIESTE

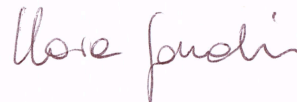
XXVIII CICLO DEL DOTTORATO DI RICERCA IN
SCIENZE DELLA RIPRODUZIONE E DELLO SVILUPPO

Research and identification of new genes and pathogenetic variants involved in Intellectual Disability

SETTORE SCIENTIFICO-DISCIPLINARE: MED03 GENETICA MEDICA

DOTTORANDA:

ILARIA GANDIN



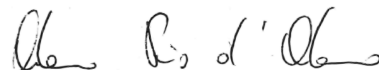
COORDINATORE:

prof. GIULIANA DECORTI



SUPERVISORE DI TESI:

prof. A. PIO D'ADAMO



ANNO ACCADEMICO 2014-2015

Abstract

Intellectual Disability (ID) is one of the most common neurodevelopmental disorders, affecting between 1.5-2% of individuals in the general population. This pathology has a serious impact on the affected individuals, their families and also on the health care system. Understanding the genetic mechanisms implicated in the disease is challenging, since ID includes a wide spectrum of possible underlying disorders and the genetic variants determining the disease are highly heterogeneous, requiring the application of different approaches and techniques, from cytogenetic analysis to the application of most recent NGS strategies. In the last decades, big steps forward have been done in the search of the genetics determinants for ID, however today less than half of the patients receive a molecular diagnosis. Moreover, the continue reporting of new ID-genes suggests that many of the genetic variants causing ID still need to be discovered and understood. The aim of the thesis is to identify new genes and genetic variants involved in ID, with the use of diverse approaches and methodologies.

The first stage of the study has been the screening of a cohort of non-syndromic ID patients with a gene-panel specific for non-syndromic ID. This approach has led to the discover of 6 novel putative mutations and, when possible, the impact of the variation has been evaluated *in silico*, with an analysis of the protein structure. The screening also led to the selection of two patients that underwent to whole-exome sequencing analysis with the aim to identify the causative variant beyond the most common genes for the disease. Notably, we have discovered 2 X-linked mutations having likely a pathogenetic role. One is a missense mutation in *CLCN4*, a gene member of the chloride channel family that only recently has been demonstrated to cause ID. Another missense mutation has been found in *ALG13* gene that has been associated with X-linked non-syndromic ID only once but without a clear explanation of its functioning.

The second part of the thesis concerns the investigation on genetic variants having a mild impact on the phenotype. We performed a Runs of Homozygosity study to investigate the role of distant inbreeding in ID. The analysis has revealed that the global amount of homozygosity and the number of homozygous stretches are associated with the degree of the impairment. Finally, we also designed an association study for the Y chromosome to investigate the presence of variants implicated in the development of cognitive function in this genomic region. We analysed the general cognitive function measured in 5 cross-sectional cohorts but did not find evidence for an association on the Y chromosome.

Sommario

La Disabilità Intellettiva (DI) è uno dei disturbi dello sviluppo più comuni, che colpisce circa il 1.5-2% della popolazione. Nella maggior parte dei casi la patologia ha un impatto importante nella vita dei pazienti e delle loro famiglie, e rappresenta anche una questione importante nella gestione della salute pubblica. Capire i meccanismi genetici coinvolti in questo disturbo è un problema complesso, poiché la DI è presente in un ampio spettro di sindromi. Inoltre, le varianti causative sono molto eterogenee e richiedono l'utilizzo di diversi approcci e tecnologie, dall'analisi citogenetica all'impiego delle più recenti strategie di NGS. Nonostante i grandi passi in avanti compiuti nello studio delle sue cause genetiche, oggi i pazienti che ricevono una diagnosi molecolare sono meno della metà. Inoltre, la continua scoperta di nuovi geni legati alla DI sta ad indicare che molte delle varianti genetiche responsabili per la patologia devono ancora essere scoperte e approfondite. Lo scopo della tesi è quello di identificare nuovi geni a varianti coinvolti nell'insorgenza della DI, tramite l'utilizzo di diversi approcci e metodologie.

La prima parte dello studio è stato lo screening di una coorte di pazienti con DI non sindromica utilizzando un pannello di geni specifico per tale disturbo. Questo approccio ha consentito di scoprire 6 nuove mutazioni candidate e, nei casi in cui è stato possibile, l'impatto delle variazioni è stato valutato *in silico* con un'analisi della struttura delle proteine. Tale screening ha anche permesso di selezionare due pazienti che sono stati analizzati con whole-exome sequencing per la ricerca delle varianti causative al di fuori dei geni più comuni per la patologia. In particolare, abbiamo scoperto la presenza di due mutazioni X-linked probabilmente patogenetiche. La prima è una mutazione missenso in *CLCN4*, un gene della famiglia dei canali del cloro che solo molto recentemente è stato dimostrato causare DI. Un'altra mutazione missenso è stata poi identificata nel gene *ALG13*, che è stato associato una sola volta alla DI non sindromica ma senza una chiara spiegazione del suo funzionamento.

La seconda parte della tesi riguarda la ricerca di varianti genetiche aventi un effetto meno evidente sul fenotipo. Abbiamo svolto un'analisi delle aree di omozigotà per indagare il ruolo dell'inbreeding nella DI. L'analisi ha rivelato che la quantità globale di omozigosi ed il numero di segmenti sono associati alla severità del deficit cognitivo. Infine, abbiamo realizzato uno studio di associazione per il cromosoma Y allo scopo di indagare la presenza di varianti coinvolte nello sviluppo della funzione cognitiva in questa regione genomica. Abbiamo analizzato la funzione cognitiva generalizzata in 5 coorti, senza però rilevare associazioni sul cromosoma Y.

Contents

1	General Introduction	1
1.1	Characteristics of Intellectual Disability	1
1.2	Genetic basis of Intellectual Disability	5
1.3	Aim of the thesis and contents	12
1.4	Patient recruitment	13
2	Target sequencing approach for non-syndromic Intellectual Disability	16
2.1	The genetic heterogeneity of Nsyn-ID	16
2.2	Targeted sequencing study	17
2.3	Protein structure	18
2.4	Novel mutations detected for NSyn-ID	20
3	Two patients with unexplained non-syndromic ID analysed with Whole-Exome Sequencing	25
3.1	Research of new genes for ID	25
3.2	Whole-Exome sequencing study	27
3.3	Discovery of novel candidate mutations	29
4	Severe cognitive impairment is associated with increased Runs of Homozygosity in Intellectual Disability	33
4.1	ROH and inbreeding	33
4.2	A Runs of Homozygosity study on ID	34
4.3	The effect of ROHs on ID severity	35
5	Y chromosome and General cognitive function	38
5.1	Possible role of Y chromosome on sexual dimorphism in diseases	38
5.2	Association study on the Y Chromosome	39
5.3	Need for higher Y chromosome's genetic variability	42
	Conclusion	48
	Appendix A	49

Appendix B	51
Bibliography	56

Chapter 1

General Introduction

This PhD thesis is an investigation on the genetic determinants in intellectual disability. This introductory chapter is aimed to give an overview on the main characteristics of the disease, the current knowledge of its genetic causes and the purposes of our study.

1.1 Characteristics of Intellectual Disability

In accordance to the Diagnostic and Statistical Manual of Mental Disorders 5 (DSM-5), Intellectual Disability (ID) is an existential condition where a cognitive ability below the average has an impact on intellectual and adaptive functioning. It involves limitations in conceptual skills, like language, reasoning and memory but also impaired social relationships and deficits in individual self-management and personal care. Because of the scarce use of standardised assessment tools in the diagnosis and the presence of differences across populations and countries, epidemiological data on ID has been always challenging to obtain. Epidemiological studies found that ID affects between 1.5 - 2% in the Western populations [1]. This makes ID one of the most common neurodevelopmental disorder and one of the most important unsolved issues in health care.

According to the DSM-5, the diagnosis of ID requires three criteria:

Deficits in intellectual functions such as reasoning, problem-solving, planning, abstract thinking, judgment, academic learning and learning from experience, practical understanding, confirmed by both clinical assessment and standardised intelligence testing. Intelligence Quotient (IQ) is the most common parameter used to assess cognitive functions. ID is characterised by a significantly lower IQ level compared to the population mean (two standard deviations), corresponding to a score of 70-75.

Deficits in adaptive functioning that result in failure to meet develop-

mental and sociocultural standards for personal independence and social responsibility. Without ongoing support, adaptive deficits limit the subject in one or more activities of daily life, such as communication, social participation, and independent living, and across multiple environments, such as home, school, work, and recreation. A common parameter for the evaluation of the adaptive functioning is the amount of assistant teaching required at school.

Paediatric onset In order to distinguish ID from dementia, intellectual and adaptive deficits must onset during the developmental period, i.e. by 18 years of age.

Intellectual functions are evaluated with an overall measure of intelligence usually referred as Intelligence Quotient (IQ). There are different type of standardised and age-dependent tests available for clinicians to estimate IQ. Wechsler Intelligence Scale for Children-fourth edition (WISC-IV) is probably the most common battery of tests being used. It offers two main measures: the Full Scale Intelligence Quotient (FSIQ), which is a four-domain test to assess the functioning of verbal comprehension, perceptual reasoning, processing speed, and working memory; and the General Abilities Index (GAI), calculated from verbal comprehension and perceptual reasoning scores, which is aimed to provide an estimate of overall reasoning ability that is less influenced by working memory and processing speed. For children aged less than four years old, scales for the assessment of Development Quotient (DV) are usually preferred instead of measures of IQ. Evaluation of DV is obtained with batteries of tests focused on motor abilities to evaluate the acquisition of developmental milestone (walking, talking, smiling, track objects with eyes) instead of cognitive or verbal skills. The most common are the Bailey scales, Griffiths scale and the Brunet-Lezine scale (focused on psychomotor development).

In the DSM-5, four severity levels are reported for ID (see Table 1.1) that were established considering not only the degree of the impairment (i.e. IQ scores) but also the amount and type of intervention needed. *Mild* category includes individuals able to learn communicative and practical life skills but slower than typical in all developmental areas. Basic self-management tasks are usually achieved and assistance is only sporadically required. People with *moderate* ID have noticeable speech and motor skills delays. Since communication is not extended on complex levels and only basic personal care tasks are carried out, they might need instruction and support in social relationships due to the poor understanding of social conventions. Individuals that fall into the *severe* category can only communicate on the most basic levels and can not perform all self-care activities independently, thus daily supervision and support is required. Frequently, people in this category have sensorial disablements and physical limitations. *Profound* ID is

Level	QI range		Proportion	Support
Mild	50-55	70-75	85%	Sporadic
Moderate	30-40	50-55	10%	Limited
Severe	20-25	35-40	3-4%	Extensive
Profound	0	20-25	1-2%	Pervasive

Table 1.1: **Degree of cognitive deficit in ID.** Cognitive impairment is classified in four levels: mild, moderate, severe and profound. For each level the corresponding IQ range, the proportion among all ID cases and the required support for the patients are reported.

the most severe form of ID. Individuals are not capable of independent living since there are significant developmental delays in all areas. Communication is usually restricted to basic form of non-verbal communication and mobility is often limited. The condition require close supervision.

In addition to the classification based on the degree of impairment, ID can also be grouped into syndromic intellectual disability (Syn-ID) and non-syndromic intellectual disability (NSyn-ID). In the syndromic form, ID is present along with one or multiple clinical features or co-morbidities. In some cases, ID could be considered one of the several symptoms in well defined syndromes (i.g. trisomy 21, fragile X syndrome, Prader-Willi syndrome, Angelman syndrome). The range of clinical features that could manifest in Syn-ID is broad (see Table 1.2): from visible anomalies as facial dysmorphisms and bone malformations, hearing or visual impairment, to more subtle malformations in the brain and other apparatuses. There could be also an extended range of neurological defects, as in the muscle tone, strength and coordination, presence of seizures.

NSyn-ID is instead delineated by the only presence of ID as clinical feature. Despite the simple definition, the diagnosis of NSyn-ID is often challenging for clinicians since the more subtle anomalies and symptoms could be extremely difficult to detect. Therefore, although this distinction is very useful for clinical purposes, it should be considered that in some cases the boundaries between Syn-ID and NSyn-IS are vague.

ID is an extremely heterogeneous condition that may result from neurological factors (*organic*) or psycho-social factors (*cultural/familiar*), or a combination of both that overall affects cognitive functioning. As regards biological determinants, there are some environmental factors well known to contribute to ID which include maternal intoxication (e.g. alcohol, drugs), prematurity, malnutrition and infectious diseases during pregnancy, premature birth, perinatal anoxia, and trauma. The proportion of ID cases with genetic base varies: for mild-moderate ID the genetic causes account for 25-50% of the cases, but they could reach the 85% of cases for more severe

Anthropometric	disproportionate short/tall stature, micro/macrocephaly, obesity
Dysmorphological	face, ears, hands, feet
Morphological	malformations in brain, bones, heart, gastrointestinal apparatus, liver, spleen, kidneys, urogenital apparatus
Sensorial	sensorineural hearing loss, visual impairment (refractive errors, strabismus, amblyopia, cortical blindness)
Neurological	seizures, hypotonia, spasticity, ataxia, dystonia
Endocrinological	hormone alterations

Table 1.2: **List of the most common areas for syndromic features in ID.**

forms. Despite the large diagnostic studies recently undertaken, the aetiology remains unknown for about 50% of the cases and this proportion is even higher for mild and borderline forms [2].

As regards ID prevalence, there is a well-known predominance of men with ID. There is indeed an excess of males over females ranging from 30% (for milder forms of ID) to 50% (for more severe forms). The male bias and the evidence for X-linked transmission in large pedigrees studies led researchers to investigate the role of X chromosome, finding a large number of genes fundamental for the development of cognitive function and thus involved in ID. One of the first syndromes identified as X-linked was the fragile X syndrome, a condition characterised by ID and very distinctive clinical features (elongated face, large ears, macroorchidism, stereotypic movements and social anxiety), which is due to defects in *FMR1* gene and represents the most common cause of X-linked ID (XL-ID) (1-2% of all ID cases [3], 25% of XL-ID cases [4]). Ever since, XL-ID has received much attention and several studies have confirmed that X-linked gene defects have important roles in the aetiology of ID and not only for men. For example, *MECP2* is another well-known XL-ID gene causing the Rett syndrome, a neurodevelopment disorder that affects almost exclusively females. XL-ID could be both syndromic and isolated, but nonsyndromic XL-ID appears to be twice as common as syndromic XL-ID: according to Piton et al. [3], of the over 100 genes identified in XL-ID, half of them are thought to cause NSyn-ID.

Despite men are ~ 1.4 fold more likely to be affected by ID, recent studies

suggest that the cumulative frequency of the detected X-linked pathogenic variants does not exceed 10%. Such discrepancy suggests that several genes or loci on the X-chromosome involved in ID are yet to be identified. However Ropers et al. [4] hypothesizes that the higher prevalence of ID and other neurodevelopmental disorders (such as autism, Tourette's syndrome) in males could be due to diverse factors, from hormonal effects to defects in gene-dosage in the two sexes (there are several genes that escape from X-inactivation having no homologue on Y chromosome). Y-chromosomal variants could be also implicated in the sex imbalance, an hypothesis arisen after the discover of the male-specific effect of *SRY* gene on the brain function.

1.2 Genetic basis of Intellectual Disability

Given the broad spectrum of possible ID phenotypes, it is not surprising that its genetics is highly complex. At present, 700 genes have now been associated to ID [5]. Encoded proteins are diverse and can be involved in one or more shared pathways or functional subclasses, taking part in complex interaction networks. The etiopathogenesis encompasses so many different genetic *entities*, varying from large chromosomal anomalies or copy number variants (CNVs), to point mutations in single genes, requiring a wide-range of diagnostic tools.

Biological processes altered in ID

Genes implicated in ID encode for diverse functional subclasses of proteins and, until recently, their function has remained unclear in most cases. According to recent reviews [2, 5], the physiopathological mechanisms of ID genes can be grouped in 4 major groups of cellular processes: neurogenesis, neuronal migration, synaptic function and regulation of transcription and translation.

Genes required for neurogenesis and the correct generation of neurons from neural progenitor in the cortical zone. The molecular mechanisms that control this neurogenesis process have not been fully understood, but it is clear that a huge number of neurons is required for the correct development of the human brain and abnormal neuron number could result in disease. In some form of microcephaly, the reduced brain size may be partially explained by the role of genes in inhibiting the proliferation of neuronal progenitors. This is the case of *ASPM*, which is a crucial gene for maintaining a symmetric, proliferative division of neuroepithelial cells during brain development.

Genes required for neuronal migration implicated in the regulation of the migration of newly born post-mitotic neurons. These genes mainly

code for proteins involved in the function of the cytoskeleton, but also cell division processes and axon/dendrite formation. Some forms of lissencephaly (brain malformation resulting in a lack of development of brain folds) are known to be caused by abnormal neuronal migration. Histological studies revealed that the affected cortex presents layers with different cell density, suggesting that many, perhaps late-born, neurons have started but not completed the migration. *PAFAH1B1* is a gene located in chromosome 17 and it is the main responsible for lissencephaly cases. Defects in *PAFAH1B1* cause delayed neuronal migration and consequently decreased brain size and distorted cellular organization in the ventricular zone.

Genes required for cellular processes involved in synaptic functions

as synaptogenesis and synaptic activities, but also synaptic plasticity, especially in postnatal stage, during learning and acquisition of intellectual performances. One of the most interesting is *FMRP*, a regulator of activity-dependent translation of mRNA-encoding proteins involved in actin/microtubule-dependent synapse growth, remodelling and function. The absence of FMRP causes fragile X syndrome: the most common monogenic cause of ID with no major apparent specific developmental brain anomaly.

Genes involved in transcription and translations

encoding transcription factors and cofactors, partners of signal transduction networks, as well as chromatin-remodeling proteins. One of the most studied is the RAS-MAPK pathway and the defects in its members like *NF1*, whose encoded protein has the capacity to regulate several intracellular processes. Mutations in this pathway impede the correct functioning of the MAPK signalling cascade, a metabolic pathway that regulates growth factors and embryological development and is associated with a particular set of intellectual disabilities, the so-called RASopathies. Given the fundamental role of the cellular processes regulated in the pathway, RASopathies are usually syndromic forms of ID.

Understanding the biological processes and the genetic networks underlying ID genes is usually very challenging, since there is little information about the gene functions and only a part of the genes involved has been discovered. A new approach has been developed in this context, which is focused on the analysis of shared molecular pathways for ID and other neurodevelopmental disorders. It is well-known that there is a high comorbidity observed between neuropsychiatric disorders such as ID, autism, schizophrenia and epilepsy, thus suggesting a possible shared genetic base. Very recently Li et al. [7] have provided new insights into this field. Authors carried out an analysis on the prevalence of *de novo* mutations, in particular deleterious mutations, among these disorders and found shared biological

process and non-coding regulatory elements of candidate genes, but also a remarkable number of candidate genes associated with more than one disorder. *SCN2A*, a sodium channel gene, was found frequently affected across all four disorders.

Genetic heterogeneity of ID and diagnostic approaches

Different criteria could be considered to classify the broad range of genetic defects causing ID. Here we chose to give an overview based on the *size* of the anomaly, from the extended chromosomal abnormalities to the single-nucleotide deleterious variations, with an eye on the diagnostic approach and the different stages of the typical genetic screening for ID.

Chromosomal anomalies

Chromosomal anomalies, for which diagnosis started under the microscope and is currently performed with cytogenetic techniques, have been extensively studied since decades and they still represent the most common known cause of ID. Despite widespread prenatal diagnosis, Down syndrome (trisomy 21) remains the most important single cause of ID (6-8% [5]). Other

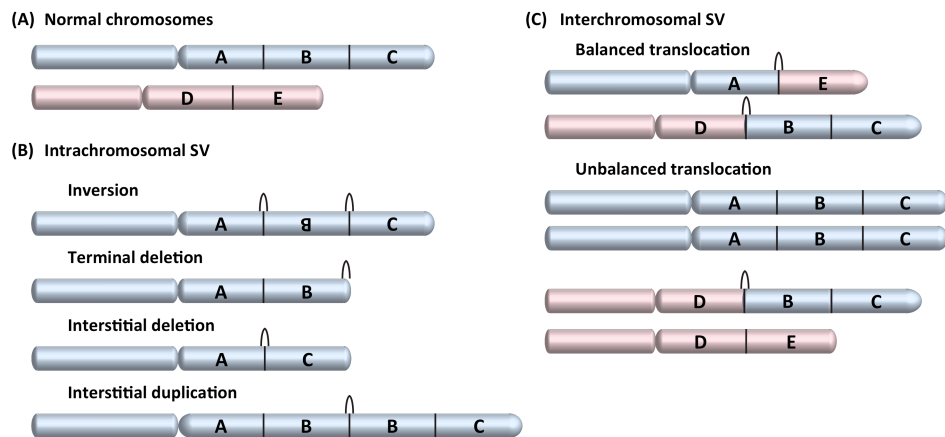


Figure 1.1: **Chromosome rearrangements** (source: Weckselbatt et al. [8]). (A) Two non homologous chromosomes. (B) Simple intrachromosomal rearrangements. Black arches indicate breakpoint junctions. (C) Interchromosomal rearrangements between two different chromosome ends. Balanced translocations are copy-neutral and do not cause a phenotype unless they disrupt developmentally important genes at breakpoints. While unbalanced translocation (represented in the figure with two copies for each chromosome), which usually occurs when one of the parents carries a balanced translocation and transmits a normal copy and a translocated copy to the offspring, results in a partial monosomy (segment E) and partial trisomy (segments B,C)

chromosomal and structural aberrations are far less common, but taken together, they account for $\sim 15\%$ of all cases [6]. Modifications could involve deletions, duplications, inversions in a single chromosome (intrachromosomal rearrangements), but they could also result in more complex structural variations involving more than one chromosome (interchromosomal translocations). Unbalanced translocations are particularly dangerous since they lead to a partial trisomy and a partial monosomy in the two chromosomes involved, as illustrated in Figure 1.1. The severity of the phenotype highly depends on the genes included in the translocated regions, as for the other structural variations after all. Uniparental disomy (UPD) occurs when the child receives two alleles from the same parent. Such anomaly can result in disease if the genomic segment involved contains recessive deleterious variants or when it encloses imprinted genes, in which either the maternal or the paternal allele copy is silenced. This is the case of Prader-Willi and Angelman that are both caused by imprint defects on chromosome 15.

One of the most common cytogenetic examinations is the banded karyogram, which produces a visible karyotype for the detection of aneuploidies and a series of lightly and darkly stained bands used as reference for the identification of chromosomal rearrangements. Subtelomeric rearrangements, i.e. deletions or duplications in the telomere-associated repeat region due to imperfect meiotic recombination, are another common cause of ID (may account for 5 - 7% of syndromic forms of ID). Such structural variants are more subtle to detect, thus usually multiprobe FISH analysis is applied, in which fluorescently-labeled telomere-specific probes are hybridised to genomic DNA.

Copy number variations

Copy Number Variations (CNVs) are genomic stretches that occur with a different number of copies compared to a reference genome assembly, arising as either deletions or duplications. With the increasing availability genome-wide microarrays, CNVs were discovered to contribute significantly to common genetic variation, covering approximately 12% of the entire genome in the normal population. Array Comparative Genomic Hybridization (aCGH) is a modification of the FISH technique in which probes have been optimised for the use in a genome-wide microarray and it is currently widely applied. Patients' DNA and a reference DNA compete to attach (hybridise) on the array and any alteration in copy number becomes visible as changes in the fluorescent signal intensity, thus aCGH has been proven to be a specific, sensitive and highthroughput technique for the detection of interstitial micro deletions and duplications. High-density SNP array is also used for CNVs detection, based on the analysis of the fluorescent intensity signals of probes. As reported in Figure 1.2, the total signal intensity (called *log R Ratio* in Illumina platforms) and allelic intensity ratio (*B Allele Frequency*) is con-

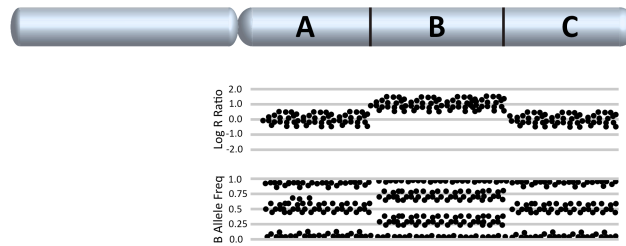


Figure 1.2: **Signal intensity analysis for CNV detection (source: Weckselbatt et al. [8]).** *Log R ratio* is the total florescent intensity signal, while *B Allele Freq* represent the ratio between the two alleles. The figure illustrates the case of an insertion. Copy-neutral segments A and C have 3 clusters in B allele frequency, corresponding to genotypes BB, AB and AA. Compared to A and C, segment B presents an increased Log R ratio and also a different pattern in the B allele frequency, corresponding to the genotypes of an duplication (BBB, BBA, BAA, AAA).

sidered at each SNP marker. Such information, together with additional parameters as inter-SNP distance and familiar relationships, is included in Bayesian framework for the prediction of deletions and duplications. The sensibility in terms of segment length of this methodology clearly depends on the resolution and the density of the markers.

CNVs have been associated with several neuro-psychiatric diseases as autism, schizophrenia, epilepsy and in particular ID. In some cases pathogenicity is clear, when the CNV disrupts well known ID genes and compromises their functioning, or when encompassing dosage-sensitive genes for which an unusual number of copies could have harmful effects, even if the encoded protein is preserved. It is estimated that CNVs, as a group, are responsible for $\sim 10\text{-}15\%$ of all ID cases. However, there are copy number variants for which a clinical interpretation is extremely difficult since considerable variation in expressivity is observed.

Recent studies have detected a CNV enrichment for ID patients compared to the general population but disease association remains unclear due to the rarity of such variations which force researchers to carry studies on very large sample sizes [9]. Moreover, deletions/insertions often contain several genes whose role need to be interpreted. Great attention has been given to *de novo* CNVs that were found to occur in the germline of 10% of patients. One of the first studies in this context was conducted by Wagenstaller et al. [10], where a cohort of 67 ID children with unexplained ID and normal karyotypes were examined with SNP-array data and several *de novo* deletions and duplications were identified most likely as causative. Interestingly, examining the size distribution of CNVs in the context of the syndromic features, Cooper et al. [9] have found larger CNVs ($\geq 400\text{Kb}$) in more severe developmental phenotypes associated with multiple congenital

abnormalities (as craniofacial and cardiovascular defects) compared to children with autism spectrum disorder. These results suggest a major effect of large CNVs in the variation of more severe phenotypes.

Therefore, while their disease risk in general is well established (with an odds ratio >20 for carriers of CNVs larger than 1.5 Mb [9]), the pathological consequences for most CNVs are not well characterised and additional studies for the genomic mapping of these effects are required.

Dysfunction of single genes

As for other Mendelian disorders, genotyping data has been used in linkage studies to investigate the cause of ID in large affected families [3]. However, in the search for the causative gene in large consanguineous families, great attention has been given to the detection of genomic areas of homozygosity, which appears as long sequences of homozygous SNPs. Stretches of homozygosity are interesting candidate regions in case of suspected recessive inheritance, since the causative genetic defect is supposed to be included in autozygosity areas. This approach, known as homozygosity mapping, led researchers to move the focus on autosomal recessive variants involved in ID. Investigation carried out in large consanguineous families led to the discovery of several autosomal ID genes [11, 12] and encouraged researchers to speculate an important part played by autosomal variants in the high rate of unexplained IDs.

Homozygosity mapping and other candidate gene studies have been conducted since decades with an increasing diagnostic yield over time, also due to the refinement provided by microarray studies, but the power in mutation detection has dramatically boosted with the introduction of next-generation sequencing (NGS), especially for sporadic ID. Direct sequencing and screening of candidate genes are now being replaced in diagnostic pipelines by more high-throughput sequencing technologies, as multiplex targeted sequencing of a few genes in large cohorts, selective targeted sequencing of up to several hundred genes, whole-exome sequencing (WES) and whole-genome sequencing (WGS).

The use of NGS strategies is particularly challenging in the case of ID, which presents great phenotype variation among patients and could involve a large range of gene and biological processes. For some clinically well-defined but very rare syndromes, exome sequencing of a small number of unrelated patients with a genotype-phenotype correlation analysis has been shown to be an efficient approach. The development of a strategy for the prioritisation of detected variants, based on overlapping clinical features in affected patients, made possible the discovery of the genetic cause of some rare Mendelian disorders [13]. The *reverse* approach is also widely used, in which variants are prioritised with computational methods based on their potential functional significance, pathway information and protein-protein

interactions [14, 15].

In case of sporadic ID, with no possibility for an analysis of overlapping mutations or clinical feature in multiple patients, trios-base sequencing studies are often performed. This approach involves the examination of the patient together with his unaffected parents. Segregation analysis is a powerful strategy to filter out detected variants when searching for the putative mutation. In 2010 Vissers et al. [16] performed family based WES approach in 10 patients with severe ID and unknown cause, and found *de novo* mutations likely to be pathogenic in six individuals. This and other similar findings led the investigators to believe that *de novo* mutations play an important role in the aetiology of ID, especially for more severe forms, providing also a possible explanation for the relatively high frequency of severe early-onset neuropsychiatric disorders in outbred populations. A *de novo* paradigm for ID has been very recently confirmed by a large-scale whole-genome sequencing study on 50 patients with severe ID, in which WGS was applied after an extensive genetic prescreening, including exome sequencing analysis [17]. Authors found a global enrichment of loss-of-function *de novo* mutations compared to previously published rates and also that *de novo* mutations occurred with higher frequency in genes previously implicated in ID-related disorders. Moreover, of the 21 patients with pathogenetic variant identified, 20 of them showed *de novo* events, including single-nucleotide variations, CNVs or structural variants, confirming that *de novo* variants are a major cause of sporadic forms of ID.

Another interesting result of the study regards the diagnostic yield of sequencing approaches, summarised in Figure 1.3. Before WGS analysis, patients underwent an extensive clinical and genetic examination, including targeted gene, SNP array and WES analysis, without obtaining a molecular diagnosis. Based on the diagnostic rates of the previous steps in the screening, the cumulative estimate for WGS to reach a conclusive genetic diagnosis was 62%. Despite the non-coding part of the genome was included in the analysis, all 78 mutations identified by WGS were found in the coding region but interestingly 65 of them were missed or detected with too low coverage in WES.

In the past 5 years NGS technologies have become more affordable in terms of costs, appearing as a very attractive strategy in the field of genetic diagnosis. However new challenges have emerged, as the interpretation of the big amount of data generated by NGS. Moreover, the recent findings of the Encyclopedia of DNA Elements (ENCODE) consortium have introduced new questions about the role of non-coding variants. The ENCODE project has identified gene-specific enhancer and repressor elements throughout the non-coding part of the genome (98.5%) which was referred to as *junk DNA* in the past. These findings suggest that non-coding CNVs may be good targets for investigation on ID since they could have disruptive effect on the regulatory landscape of the genome and result in disease [5].

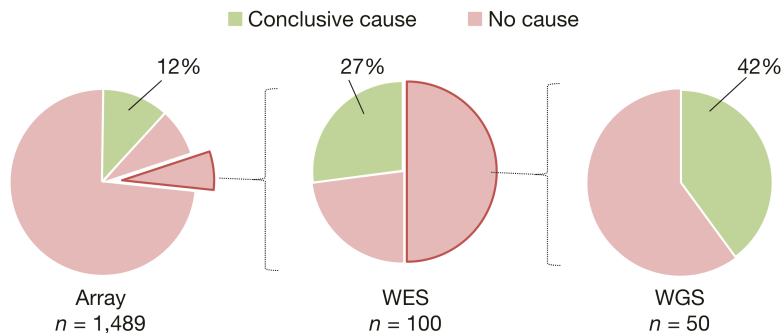


Figure 1.3: **Diagnostic yield in patients with severe ID per technology in Glissen et al. study (source: Glissen et al. [17])**. The three charts summarise the diagnostic yield of a genetic screening on a ~ 1500 cohort of severe ID patients. Brackets indicate the group of patients with no causative variant identified and subsequently analysed in the next step of the screening. The proportion of subjects diagnosed with microarrays analysis was 12%, for the others WES analysis led to a molecular diagnosis in 27% of cases. The remaining subjects were then analysed with WGS and genetic cause was identified in 62% of cases.

1.3 Aim of the thesis and contents

Despite the great steps forward in the understanding of genetic mechanisms underlying ID, the majority of the patients remain still without a molecular diagnosis, thus additional efforts are required. ID is one of the most serious neurodevelopmental disorders, which has a serious impact not only on the affected individuals and their families but also on the health care system and society. The research for the genetic determinants in ID is a complex problem, due to diverse factors. The spectrum of possible underlying disorders is extensive, requiring an ample range of clinical investigations for its determination. Then one side, the high heterogeneity of the pathology implicates that researchers are required to apply very different approaches in the analysis, from cytogenetic techniques to the application of most recent NGS strategies. On the other side, the continue reporting on new genes/genetic variants implicated in ID in the scientific community suggests that there is a lot in the genetics of ID that still need to be discovered and understood.

In this context, we set up a study aimed to identify new genes and genetic variants involved in ID and tried to approach the complexity of the pathology employing diverse methodologies. The first stage of the study has been the screening of a cohort of non-syndromic ID patients for mutations in genes already known to be associated to the disease, that is described in Chapter 1. Thanks to the design of a gene-panel specific for NSyn-ID that has been implemented in NGS technology, we were able to identify 6 novel mutations likely involved in the disease. The importance of this screening

has been not only the discover of novel variations, but also the forming of a sub-cohort of NSyn-ID patients not carrying pathogenetic mutations in the most common genes for the disease. In this way we had higher chances for finding defects in genes and genomic regions not associated with NSyn-ID yet. Two of these patients were selected and analysed with whole-exome sequencing. The analysis is described in Chapter 2. Thanks to a trio-based approach, we were able to perform an accurate selection among the $\sim 43\text{K}$ variants detected and identify two plausible candidates for having a pathogenetic role. One of those variants occurred in gene *CLCN4* that only very recently has been associated specifically to ID. The other variant is included in the *ALG13* gene that was only hypothesised to be involved in NSyn-ID in one previous investigation.

The second part of this manuscript concerns the investigation on genetic variants having a milder effect on the phenotype. In Chapter 3 we investigate the role of Runs of Homozygosity in ID. Contrary to our first hypothesis, we did not find an excess of homozygosity in syndromic ID patients compared to non-syndromic cases. However, our analysis revealed that the global amount of homozygosity and number of homozygous stretches are associated with the degree of the impairment. Indeed, despite not being the cause of the disease, the presence ROHs seems to modulate the severity of the cognitive impairment.

In the search of genetic variants having an influence on the ID phenotype, we also considered a genomic region that has been almost completely neglected in the past that is the Y chromosome. As the investigation in neurodevelopmental disorders has recently turned the focus on this region and obtained interesting results, especially for autism, we designed an association study for the Y chromosome that can potentially give new insights in the genetics of ID. In Chapter 4 we describe the implementation of a single-SNP approach for the Y chromosome and present the results of the analysis for the general cognitive function in 5 cross-sectional cohorts. No evidence for a role of Y chromosome variants in cognitive function has been found but the study could represent an important starting point for improving the knowledge of this genomic region.

1.4 Patient recruitment

A cohorts of 668 affected children have been founded specifically for the project. This was possible thanks to the collaboration of the paediatric clinics of 4 Italian hospitals which the patients were sampled from: IRCCS-Burlo Garofolo (Trieste), IRCCS-Casa Sollievo della Sofferenza (Foggia), Città della Scienza e della Salute (Torino), IRCCS-Oasi Maria SS. (Enna), which are specialised ID Health Services for patients coming from the whole Italian territory (Figure 1.4).

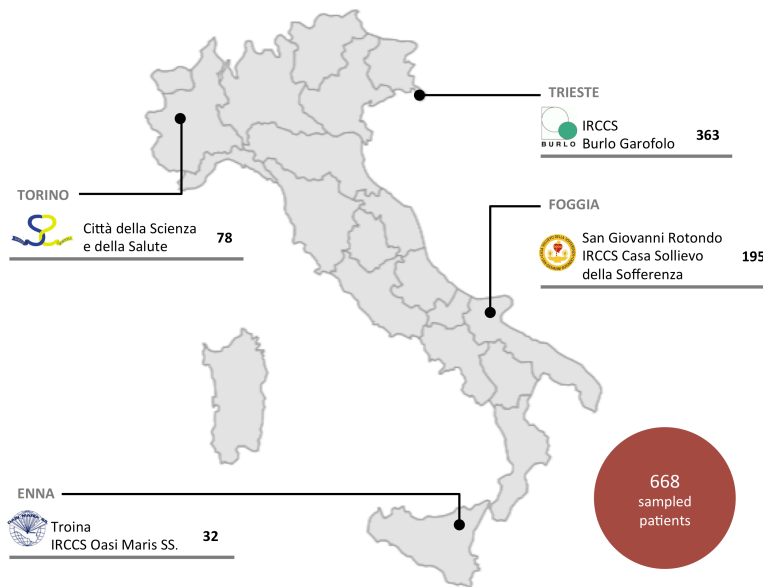


Figure 1.4: **Recruitment of ID-patients.** Affected children have been recruited in 4 Italian hospitals. For each centre, the number of subjects is reported.

In each centre, probands have been clinically evaluated. A diagnostic algorithm has been applied to exclude cases of non-genetic ID and distinguish between syndromic and non-syndromic ID, depending on the results of clinical evaluations. Since part of the study has been focused specifically on the Nsyn-ID, great deal of effort has been put for the detection of syndromic features in all areas (Table 1.3). Neurodevelopmental assessment has been collected using developmental scales (Bayley III, Brunet-Lezine, Griffiths) or intelligence scale (Weschler), according to the age of the patients.

Thanks to the fruitful collaboration between the groups and agreed criteria for the clinical and genetic evaluations, we were able to obtain a large cohort of ID patients that have been the focus of most stages of the project. The gene-panel screening and the WES study was carried out on the sub-cohorts of NSyn-ID patients, while the homozygosity study involved all subjects.

EXAMINATIONS	EXCLUSION CRITERIA
ANAMNESIS	
Pre-/peri-/neonatal history	<i>Adverse events</i>
SPECIALIST EVALUATIONS	
Orthopaedic	<i>Bone malformations</i>
ENT	<i>Sensorineural hearing loss</i>
Ophthalmological	<i>Severe refraction disease, fundus alterations</i>
Endocrinological	<i>Hormones alterations, urogenital malformations</i>
Dysmorphological and anthropometric	<i>Major dysmorphisms, micro/macrosomia</i>
INSTRUMENTAL EVALUATION	
Brain MRI	<i>Major malformations</i>
Echocardiography	<i>Heart malformations</i>
Abdominal ultrasound	<i>Gastrointestinal, hepatic, splenic, renal malformations</i>
EEG	<i>Seizures or EEG alterations</i>

Table 1.3: **List of clinical examinations and exclusion criteria for the diagnosis of non-syndromic ID.**

Chapter 2

Target sequencing approach for non-syndromic Intellectual Disability

In this chapter we present a study on non-syndromic ID that has revealed 6 novel candidate mutations. The work has been published in Mutation Research as *Morgan A. et al "Target sequencing approach intended to discover new mutations in non-syndromic intellectual disability". Mutat Res. 2015 Nov;781:32-6.* Contributors for this work are reported at the end of the chapter.

2.1 The genetic heterogeneity of NSyn-ID

Most of the new insights on the genetics of ID emerged over the last years has been acquired thanks to studies focused on syndromic ID, which is the most common forms of ID.

Non-syndromic intellectual disability (NSyn-ID), characterised by cognitive impairment as the unique clinical feature, has been poorly investigated. Understanding the genetics behind NSyn-ID would be relevant both for patients care and for exploring the basic mechanisms underlying the human cognition and intellect. In fact genes whose alteration causes NSyn-ID could be likely involved in the learning and memory processes. Moreover, NSyn-ID shares some genetic mechanisms with the group of neurodevelopmental disorders known as Autism Spectrum Disorder [18]. In this perspective, NSyn-ID represents a relevant health care issue, which demands undeniable investigator efforts.

Targeted sequencing (TS) allows the sequencing of a selected number of genes in a multi-sample plate and offers several advantages. A coverage with high depth and a smaller portion of poorly covered regions ensure a high sensitivity and specificity for the detection of pathogenic events in the

regions of interest. Indeed, the actual coverage with the exome or whole-genome strategies is still frequently insufficient, which may result in missing mutations. Moreover, TS allows the simultaneous analysis of a high number of patients with significantly lower costs, not only in terms of the technology *per se* but especially of data analysis, storage and interpretation, when compared to WES.

Quite surprisingly, when the project was started, we could not find in literature a comprehensive list of genes involved in non-syndromic ID. Therefore a preliminary step of the study was the selection of genes related to Nsyn-ID that led to the compilation of a 71 genes list (see Table 2.1). In this context, we considered non-syndromic ID genes all genes that have been implicated at least once in Nsyn-ID. This means that genes could have been also associated to syndromic forms of ID.

With the multiple goals to (a) confirm the involvement of known genes, (b) find novel mutations in NSyn-ID-genes and (c) assess TS as NSyn-ID diagnostic tool, we analyzed 65 NSyn-ID patients with a 71 genes-panel.

2.2 Targeted sequencing study

We collected clinical information, peripheral blood gDNA samples and informed consent from 65 NSyn-ID patients (43 patients with their parents and 22 isolated ones), already found negative to the pathogenic CNVs detection via SNPs-array analysis. Our selected cohort of patients consists of 23 females and 42 males.

A specific Targeted Gene Panel has been defined by the Ion AmpliSeq™ Designer v1.2 tool (Life Technologies, CA, USA). 71 genes have been selected on the basis of the most updated literature survey (up to 2013), including all genes reported at least once as causative of NSyn-ID (Table 2.1). This selection includes 36 X-linked and 35 autosomal genes. Single-end sequencing was carried out on the Ion Personal Genome Machine System on Ion 318™ Chips (Life Technologies, CA, USA). The variants were annotated using the ANNOVAR software. CNVs analysis was performed by Conifer (COpy Number Inference From Exome Reads) tool [19]. Whenever needed, we manually investigated the raw sequence reads using the Integrative Genomics Viewer (IGV) [20] to exclude false positives calls.

The sequencing process produced a mean of 102 Mega bases of raw sequence data per individual. An average of 88% of the targeted regions was covered at least 20-fold. A mean depth of coverage of 227X and an average of 133 variants for each individual were obtained.

Candidate causative variants have been selected with a filtering procedure applied on the variants called for each proband based on two criteria:

1. variant frequency: common polymorphisms have been removed referring to 1000 Genomes Project (<http://www.1000genomes.org/>) and

the NHLBI-ESP 6500 exome project (<http://evs.gs.washington.edu/EVS/>) databases;

2. Mendelian consistency: only variants fitting one between *de novo*, X-linked or recessive inheritance patterns have been included.

All the identified candidate disease-causing variants have been analysed by direct Sanger sequencing on patients and parents.

2.3 Protein structure

In vivo functional studies are difficult to perform for mutated genes in ID patients since the affected tissue, the brain, cannot be easily accessed. Therefore the impact of missense mutations on the protein function has been assessed first using *in silico* pathogenicity predictors (Polyphen-2 [21], MutationTaster [22] and SIFT [23], PhyloP algorithm [25] for evolutionary conservation) and successively with Molecular Dynamics (MD) analysis. MD simulations provide powerful tools to estimate how the variation affects the protein structure, which is often a critical element of their function. In this study, MD analysis has been performed by the prof. Pricl's group at the Chemical Sciences Department, University of Trieste, thanks to a fruitful collaboration established for the project. Only the results of MD analysis is given here and we refer the reader to the published article for a detailed description of the protocol.

High-resolution crystallographic is used as starting geometry for both wild-type and mutant proteins, then mutation is introduced in the protein structure. Three-dimensional model structure of the full-length protein is built by a combination of homology-based techniques. To this aim, three-dimensional computational models of the proteins have been created for those candidate genes having crystal structure available. After a preliminary screening of the Protein Data Bank (PDB) repository, either partially-matching or significant homology templates have been identified as reliable starting point for 3D protein homology modeling only for three of the proteins involved: Atrx, Pqpb1 and Arid1b.

Gene	Chromosome	Reference
ACSL4	X	Nature Genet. 30: 436-440, 2002
AGTR2	X	Hum Genet. 2004 Jan;114(2):211-3
AP1S2	X	Am J Hum Genet. 2006 December; 79(6): 1119-1124
ARHGEF6	X	Nature Genet. 26: 247-250, 2000
ARID1B	6	Am J Hum Genet. 2012 March 9; 90(3): 565-572
ARX	X	J Child Neurol. 2007 Jun;22(6):744-8
ATRX	X	Am J Med Genet. 2002 Jul 1;110(3):243-7
AUTS2	7	Hum Genet (2007) 121:501-509
C10orf11	10	Europ. J. Hum. Genet. 18: 291-295, 2010
CACNG2	22	Am J Hum Genet. 2011 Mar 11;88(3):306-16
CASK	X	Nature Genet. 41: 535-543, 2009
CC2D1A	19	J. Med. Genet. 40: 729-732, 2003
CDH15	16	Am. J. Hum. Genet. 83: 703-713, 2008
CDK5R1	17	Neurogenetics. 2006 Mar;7(1):59-66
CDKL3	5	Am J Med Genet A. 2008 May 15;146A(10):1267-79
CIC	19	Nat Genet. 2010 Dec;42(12):1109-12
CRBN	3	Neurology. 2004 Nov 23;63(10):1927-31
DLG3	X	Neurogenetics. 2010 May;11(2):251-5
DOCK8	9	Genomics. 2008 Feb;91(2):195-202
DYNC1H1	14	Nat Genet. 2010 Dec;42(12):1109-12
EPB41L1	20	Am J Hum Genet. 2011 Mar 11;88(3):306-16. doi: 10.1016
FGD1	X	Clin Genet. 2002 Feb;61(2):139-45
FOXP1	3	Am J Hum Genet. 2010 Nov 12;87(5):671-8
FTSJ1	X	Am J Hum Genet. 2004 Aug;75(2):305-9
GDI1	X	Am J Med Genet A. 2011 Dec;155A(12):3067-70
GPD2	2	Hum Genet. 2009 Jan;124(6):649-58
GRIA3	X	Am J Med Genet A. 2004 Dec 1;131(2):174-8
GRIK2	6	Am J Hum Genet. 2007 Oct;81(4):792-8
GRIN1	9	Am J Hum Genet. 2011 Mar 11;88(3):306-16
GRIN2A	16	Nat Genet. 2010 Nov;42(11):1021-6
GRIN2B	12	Nat Genet. 2010 Nov;42(11):1021-6
HCFC1	X	Am J Hum Genet. 2012 Oct 5;91(4):694-702a
HUWE1	X	Am J Hum Genet. 2008 Feb;82(2):432-43
IL1RAPL1	X	Am J Med Genet A. 2011 May;155A(5):1109-14
IQSEC2	X	Nat Genet. 2010 Jun;42(6):486-8
KDM5C	X	Am J Hum Genet. 2005 Feb;76(2):227-36
KIAA2022	X	Med Genet. 2004 October; 41(10): 736-742
KIRREL3	11	Am J Hum Genet. 2008 December 12; 83(6): 703-713
MAGT1	X	Am. J. Hum. Genet. 82: 1150-1157, 2008
MBD5	2	Am. J. Hum. Genet. 81: 768-779, 2007
MECP2	X	FEBS Lett. 2000 Sep 22;481(3):285-8

Gene	Chromosome	Reference
MED23	6	Science. 2011 Aug 26;333(6046):1161-3
NLGN4	X	Am J Hum Genet. 2004 Mar;74(3):552-7
NRXN1	2	Europ. J. Hum. Genet. 20: 1240-1247, 2012
NSUN2	5	Am J Hum Genet. 2012 May 4; 90(5): 856-863
OMG	17	Neurogenetics. 2006 Mar;7(1):59-66
PAK3	X	Yi Chuan. 2007 May;29(5):523-7
PHF8	X	Mol Cell. 2010 Apr 23;38(2):165-78
PQBP1	X	Nat Genet. 2003 Dec;35(4):313-5
PRSS12	4	Am J Hum Genet. 2008 May;82(5):1158-64
PTCHD1	X	Sci Transl Med. 2010 Sep 15;2(49):49ra68
RAB39B	X	Am J Med Genet. 2000 Oct 23;94(5):376-82
RPS6KA3	X	Clin Genet. 2006 Dec;70(6):509-15
SETBP1	18	Eur J Med Genet. 2012 Mar;55(3):216-21
SHANK2	11	Nat Genet. 2010 Jun;42(6):489-91
SHANK3	22	Am J Hum Genet. 2011 Mar 11
SLC6A8	X	Am J Hum Genet. 2004 Jul;75(1):97-105
STXBP1	9	Am J Hum Genet. 2011 Mar 11;88(3):306-16
SYNGAP1	6	New Eng. J. Med. 360: 599-605, 2009
SYP	X	Nat Genet. 41: 535-543, 2009
TCF4	18	Am. J. Med. Genet. 146A: 2053-2059, 2008
TECR	19	Hum Mol Genet. 2011 Apr 1;20(7):1285-9
TRAPPC9	8	Am J Hum Genet. 2009 Dec;85(6):909-15
TSPAN7	X	J. Med. Genet. 39: 430-433, 2002
TUSC3	8	Am. J. Hum. Genet. 82: 1150-1157, 2008
UPF3B	X	Nature Genet. 39: 1127-1133, 2007
ZDHHC9	X	Am. J. Hum. Genet. 80: 982-987, 2007
ZNF41	X	Am J Hum Genet. 2003 Dec;73(6):1341-54
ZNF674	X	Am J Hum Genet. 2006 Feb;78(2):265-78
ZNF711	X	Nat Genet. 41: 535-543, 2009
ZNF81	X	J. Med. Genet. 41: 394-399, 2004

Table 2.1: **List of 71 Nsyn-ID genes.** The panel includes 35 autosomal genes and 36 X chromosome genes.

2.4 Novel mutations detected for NSyn-ID

The analysis of 65 NSyn-ID patients (10 of them affected by severe ID) led to the identification of 4 novel X-linked point mutations, located in *DLG3*, *IL1RAPL1*, *ATRX*, and *PQBP1* genes, as well as 2 novel de novo point mutations in autosomal genes *ARID1B* and *KIRREL3* (Table 2.2). *De novo* variants have been called in

Subject ID	Sex	ID Degree	Gene	Chr	Inheritance	Ref sequence	cDNA change	AC change	PhyloP	PhyloPhen2	SIFT	Mut test	Reference	MAF
117	M	Severe	DLG3	X	X-linked	NM_020730	c.G72T	p.W24C	C	B	T	D	novel	NA
4021	M	Severe	IL1RAPL1	X	X-linked	NM_014271	c.A1871G	p.Y624C	C	B	T	D	novel	NA
12-1685	M	Mild	ATRX	X	X-linked	NM_138270	c.A5342T	p.D1781V	C	D	D	D	novel	NA
13-234	M	Mild	PQBP1	X	X-linked	NM_001032383	c.C604G	p.D202H	C	D	D	D	novel	NA
663	F	Mild	ARID1B	6	<i>de novo</i>	NM_017519	c.G3293A	p.G1098E	C	D	D	D	novel	NA
12-560	M	Mild	KIRREL3	11	<i>de novo</i>	NM_032531	c.433+1G>A	-	-	-	-	-	novel	NA
11-026	M	Mild	GRIN2A	16	?	NM_001134407	c.G3200A	p.R1067Q	C	D	D	D	novel	NA

Table 2.2: List of mutations identified in our cohort of patients.

PhyloP = C conserved, N not conserved.

PolyPhen2 = D probably damaging, P possibly damaging, B benign.

SIFT = T tolerated, D deleterious.

MutationTaster = D disease causing; P polymorphism.

MAF = Minor frequency allele (data from ESP6500).

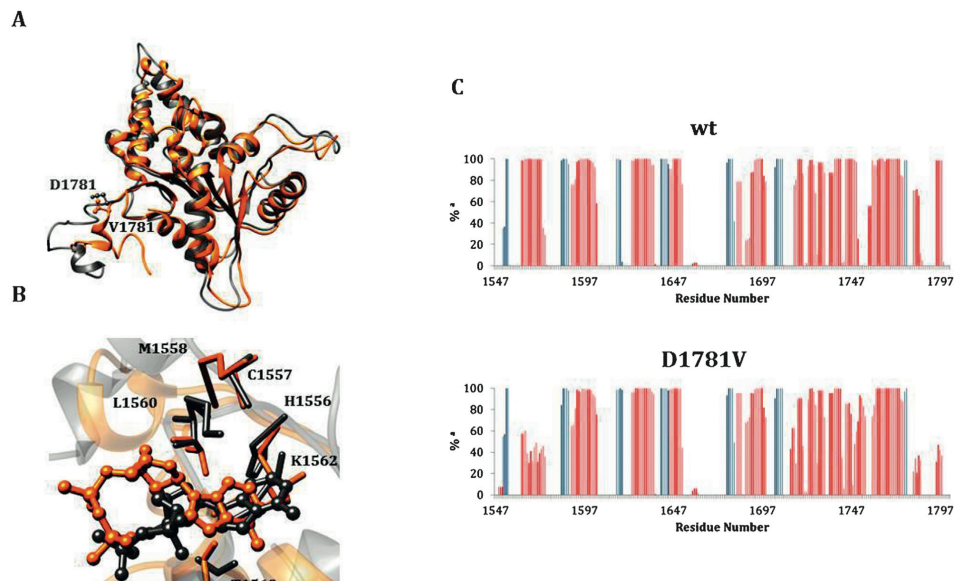


Figure 2.1: **Protein structure of *ATRX*** (source: Morgan A. et al 2015). (A) Comparison of the equilibrated MD snapshots of the Atrx wt (dim gray) with D1781V (orange) mutants and (B) the close-ups of the receptor binding site of ATP. The ligand is shown as balls and sticks colored according to the respective protein complex. Hydrogen atoms, water molecules, ions, and counterions were omitted for the sake of clarity. (C) General secondary structure description of the Atrx protein (amino acids 1547 1802) during MD simulations of the wt and D1781V. The y-axis gives the percentage of specific secondary structure for each Atrx residue.

heterozygous state and it is assumed they have a dominant effect. Both genes involved, *ARID1B* and *KIRREL3*, were previously reported associated to non-syndromic ID [26, 27]. Moreover, we found a novel missense mutation in *GRIN2A* gene predicted as highly pathogenic by all the prediction tools, although the absence of parental DNA did not permit to study the mutation segregation. No small CNV has been detected by the use of CONIFER.

We conducted an accurate patient clinical re-evaluation in order to detect any phenotypic sign but all evaluated patients showed no additional findings, confirming the previous clinical diagnosis of non-syndromic ID. Moreover, all mutations discovered in the study have been never reported before, neither as rare polymorphism in dbSNP nor as mutations in OMIM (Online Mendelian Inheritance in Man). These facts clearly support their pathogenicity and the genetic and allelic heterogeneity hypothesis behind the ID.

Thanks to the analysis of protein structure, the Atrx mutation was found to be located in the helicase/ATPase domain at the carboxyl terminus of the protein that includes a potential nucleotide-binding site for ATP constituted by residues 1594-1601. The MD simulations revealed an expansion of the ATP binding pocket from an average dynamic volume (ADV) (Figure 2.1 A,B). A different secondary structure rearrangement can be clearly observed (Figure 2.1 C). As a consequence, ATP fits less tightly in the mutant protein binding pocket, in fact the binding free

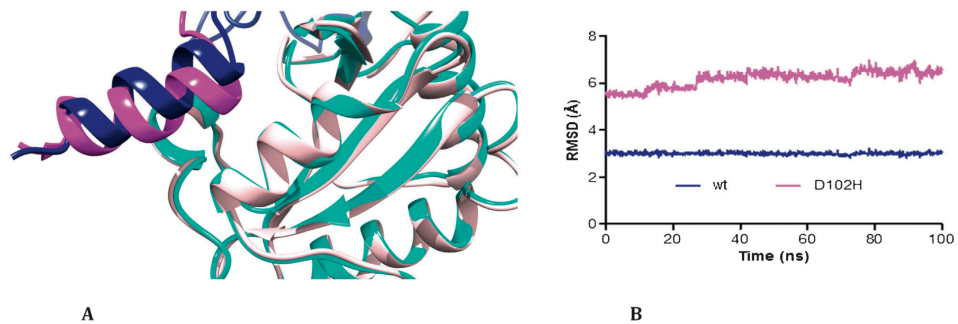


Figure 2.2: **Protein structure of *PQBP1*** (source: Morgan A. et al 2015). (A) Zoomed view of the superposition between two snapshots representing Pqbp1wt/U5, depicted pink/light gray respectively and Pqbp1 D102H/U5 complex in blue and cyan. (B) RMSD of the coordinates of the backbone atoms of Pqbp1/U5 along the MD simulation. Comparison of the equilibrated MD snapshots of the wt Pqbp1 protein (blue) with D102H (pink) mutant in complex with U5.

energy (DG_{bind}) for the wild-type/ATP Atrx complex is -17.45 ± 0.34 kcal/mol while this value drops to -14.56 ± 0.32 kcal/mol in the case of the D1781V/ATP variant assembly.

The Pqbp1 protein has two characteristic domains that preside over protein/protein interactions, specifically with the WBP11 and U5-15kD proteins. MD simulations of the wild-type/U5 and D102H mutant pqbp1/U5 protein/protein complexes revealed that U5 interacts more loosely with the mutant protein than with the wild-type counterpart (Figure 2.2 A,B). This effect can be quantitatively appreciated by considering the root mean square deviation (RMSD) of U5 backbone atoms with time shown in Figure 2.2.

Arid1b protein is involved in chromatin remodeling. The G1098E mutant protein activity was evaluated by performing MD simulations of their interaction with an AT-rich DNA fragment. The substitution of a glycine with a charged and bulky glutamic acid results in a substantial alteration of the interactions of the mutant protein with the nucleic acid (Figure 2.3): in fact, the negatively charged side chain of E1098 points directly toward the binding region, thereby interfering, via electrostatic and steric repulsions, with the negatively charged backbone of DNA.

In this study we analysed 43 trios and additional 22 isolated patients affected by NSyn-ID, finding 6 (4 X-linked, 2 de novo) putative pathogenic mutations and a potentially causative mutation in *GRIN2A* gene. The discovery rate is therefore close to 11% in our cohort. These findings show that however a large portion of our patients is negative to the screening. We focused our study on a specific and strictly well-selected cohort of patients, affected by intellectual impairment without any other relevant signs or symptoms. Unlike syndromic ID, NSyn-ID seems therefore a truly complex phenotype, likely caused by a deep and broad interaction between gene defects and environment. We would speculate that part of the genetics behind NSyn-ID, particularly for the milder forms, could be related genes not included in the panel or to genetic variants in regulative sequences, which

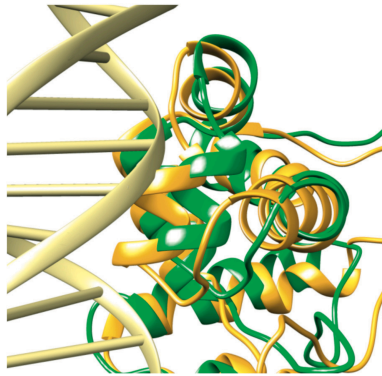


Figure 2.3: **Protein structure of ARID1B** (source: Morgan A. et al 2015). Zoomed in view of the interaction zone of DNA AT-rich sequence and Arid WT/G1098E, in green and gold respectively. Hydrogen atoms, water molecules, ions and counterions are not shown for clarity.

can not be detected with the TS approach. In addition, some environmental factors could further modulate the genetic variants magnitude, contributing therefore to arise the phenotype.

Consistently with the broad genetic heterogeneity behind the NSyn-ID, our TS analysis showed a quite low discovery rate, pointing out that several other genes remain still to be discovered; as matter of fact over the last year, several new NSyn-ID genes have been identified [21,22]. Consequently, the discovery rate could be surely improved over the time by a systematic genes list update.

In conclusion, our results represent an important contribution for future studies on NSyn-ID aimed to eventually increase evidences on the role of known genes or to discover new NSyn-ID genes. A better knowledge on NSyn-ID would be essential to the understanding of the basic abilities in human cognition and intellect, because genes involved in NSyn-ID are likely related to the learning and memory processes and the perturbation of the related pathways could affect these processes, as suggested by the *GRIN2A* involvement in neural plasticity [23].

Contributors

Anna Morgan¹, Chiara Belcaro², Pietro Palumbo³, Orazio Palumbo³, Elisa Biamino⁴, Valentina Dal Col⁵, Erik Laurini⁵, Sabrina Prici⁵, Paolo Bosco⁶, Massimo Carella³, Giovanni Battista Ferrero⁴, Corrado Romano⁶, Adamo Pio d'Adamo¹, Diego Vozzi², Flavio Faletra².

¹Department of Medical Sciences, University of Trieste, Italy

²Institute for Maternal and Child Health - IRCCS Burlo Garofolo, Trieste, Italy

³Medical Genetics Unit, IRCCS Casa Sollievo della Sofferenza, San Giovanni Rotondo, FG, Italy

⁴Department of Pediatrics, University of Torino, Torino, Italy

⁵MOSE-DEA, University of Trieste, Trieste, Italy

⁶Unit of Pediatrics and Medical Genetics, IRCCS Associazione Oasi Maria Santissima, Troina, EN, Italy

Chapter 3

Two patients with unexplained non-syndromic ID analysed with Whole-Exome Sequencing

In this chapter the Whole-Exome Sequencing analysis of two NSyn-ID patients is presented. Children were visited at the genetic clinic of Maternal-Child hospital Burlo Garofolo (Trieste, IT). Contributors for this work are reported at the end of the chapter.

3.1 Research of new genes for ID

Human brain development involves complex regulation of cellular proliferation, signaling and transcription pathways, thus many relevant genes are expected to be involved in Intellectual Disability. More than 700 ID-genes have now been identified that can be analysed for the molecular diagnosis, but a large number of new genes are discovered every year. In the *Intellectual Disability Gene Panel* used at the Dept. of Human Genetics of the Radboudumc, Nijmegen, the recognised ID genes for 2014 amount to 15 (see Table 3.1). There is evidence for a large fraction of ID genes that still remains to be discovered.

Whole-exome sequencing (WES) is the application of the next-generation technology to determine the variations of all coding regions. NGS has influenced all fields of biological research but especially the investigation of disease-causing mutations in Mendelian disorders, including those involving a large number of genes. The study of Vissers et al. [16] has demonstrated the power of this approach: they analysed ten individuals with unexplained ID and identified and validated non-synonymous de novo mutations in nine genes. Certainly the current knowledge and resources for understanding the non-coding variation is still limited, requiring extensive functional follow-up to determine the role in disease, but it seems that coding region variants plays the major role in the pathogenesis of ID.

There are no doubts about the power of approaches that allow investigators to

Gene	Inheritance	Phenotype	Reference
AHDC1	AD	Xia-Gibbs syndrome	Xia et al. 2014
FBXO31	AR	Intellectual disability	Mir et al. 2014
FMN2	AR	Intellectual disability	Law et al. 2014
HCN1	AD	Infantile epileptic encephalopathy	Nava et al. 2014
KPTN	AR	Intellectual disability	Baple et al. 2014
METTL23	AR	Intellectual disability	Bernkopf et al. 2014
NFIA	AD	Hypoplastic corpus callosum, ventriculomegaly, seizures and urinary tract defect	Rao et al. 2014
PGAP1	AR	Intellectual disability	Murakami et al. 2014
PGAP3	AR	Hypoplastic corpus callosum, ventriculomegaly, seizures and urinary tract defect	Howard et al. 2014
POGZ	AD	Autism, Intellectual disability, schizophrenia	Gilissen et al. 2014
PTDSS1	AD	Lenz-Majewski hyperostotic dwarfism	Sousa et al. 2014
SNX14	AR	Spinocerebellar ataxia, Intellectual disability	Sousa et al. 2014
SOX11	AD	Intellectual disability	Tsurusaki et al. 2014
USP9X	XL-D	Intellectual disability	Homan et al. 2014
YAP1	AD	Coloboma, ocular, with or without hearing impairment, cleft lip/palate, and/or Intellectual disability	Williamson et al. 2014

Table 3.1: **New ID genes published in 2014** Gene list obtained from the *Intellectual Disability Gene Panel* (v.DGD141114) used at the Dept. of Human Genetics of the Radboudumc, Nijmegen, The Netherlands. Inheritance model codes are: AD=Autosomal Dominant, AR=Autosomal Recessive, XL-D=X-linked Dominant. OMIM ID is referred to the phenotype.

research causative variants across all genome, but it comes also with big challenges, mainly in the interpretation of the results. Notably, it is estimated that for each individual WES data contain around 10.000 nonsynonymous variants depending on ethnicity and calling methods [28]. The discrimination of the putative variants is currently the most demanding part of the analysis and efficient strategies still need to be developed to this aim. The most common approach is to analyse the inheritance mode of the disease, and use *in silico* predictor to evaluate functional and biological significance of each variant.

The 71-gene panel screening described in Chapter 2 led to the identification of candidate causative variants in 6 NSyn-ID patients but also ensured with high probability that the remaining individuals do not carry pathogenetic mutations in the most common genes for NSyn-ID. With this precondition, we selected two affected children among the ones resulted as negative in the gene panel screening and set up a trio-based WES study aimed to reveal the genetic cause of the disease. The analysis led to the identification of 3 X-linked candidate variants: one is located in the GPR64 gene that according to the current knowledge of its function is unlikely involved in ID; while other 2 variants, carried by one patient and his affected brother, were found in CLCN4 and ALG13 genes, which can be realistically implicated in the disease. However, this study does not amounts to a final step in our project. These finding will require functional studies that, at the moment, are the only way to demonstrate the pathogenicity of candidate mutations identified.

3.2 Whole-Exome sequencing study

Two families have been selected for the WES analysis. Among patients with no molecular diagnosis after the Target Sequencing, two individuals have been chosen and analysed together with their parents. Family F1 included an affected male child 12-781, father 12-779, mother 12-780. A non-syndromic ID of moderate level was diagnosed in the boy. Family F2 included an affected male child 11-965, affected brother 11-962, father 13-1537, mother 12-1536. ID showed variability among the two brothers: in patient 11-965 a severe non-syndromic ID was diagnosed, while patient 11-962 showed a syndromic form of ID, characterised by the presence of seizures and a moderate cognitive deficit. Proband 11-965 and his parents were included in the WES analysis whereas his brother has been analysed only for the validation of candidate variants.

The WES was performed at the Europe Life Technologies Ion AmpliSeq Exome Certified Service Providers, CRIBI Sequencing Core, University of Padua, Italy. The DNA libraries were prepared employing the Ion AmpliSeq Exome Kit and were then sequenced by the Ion Proton System (Life Technologies, CA, USA), according to the standardised procedures. Reads mapping and variant calling were performed by the Ion TorrentSuite v4.0 software (Life Technologies, CA, USA) set up with standardised parameters. Single Nucleotide Variations (SNVs) and small Insertions and Deletions (INDELS) were reported into a Variant Call Format (VCF) version 4.1. SNVs and INDELS were annotated using the most updated version of ANNOVA [29].

The major issue in WES analysis is to identify the candidate variants among the several variations detected. To this aim, we set up an internal database accounting for the overall recurrent variations called in the sequenced patients, affected by diseases other than ID. In this way we were able to identify systematic errors,

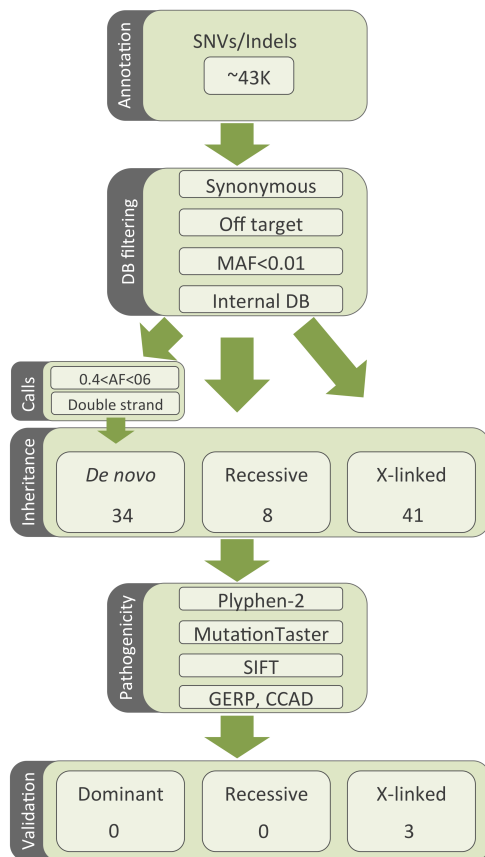


Figure 3.1: **Custom bioinformatic pipeline for variant filtering.** After annotation, variant were removed if: 1) synonymous, 2) off-target, 3) having $MAF > 0.01$, 4) systematic errors. For *de novo* variants, additional filters have been applied: frequency of the alternative allele must be between 0.4 and 0.6 and the variant must be called on the double strand. For selected variations, pathogenicity was evaluated based on *in silico* predictors and only variations showing great impact were retained. Finally, variations were validated in family members with Sanger sequencing.

i.e. base-call errors in the sequenced reads that are common in high-throughput sequencing technologies. Variations were filtered out if occurring in more than 3% of samples.

SNVs/INDELs have been filtered using a specific algorithm outlined in Figure 3.1. The first step was to remove variations leading to synonymous amino acids substitutions or if called in off-target regions. In order to exclude polymorphisms, we also removed all variations with $MAF < 0.01$ based on public databases as NCBI dbSNP build138, 1000 Genomes Project, Exome Sequencing Project (ESP) Exome Variant Server, Exome Aggregation Consortium (ExAC). After that, systematic errors were identified and discarded thanks to a frequency check in our internal database. For possible *de novo* variations, additional filters have been applied, being this category more susceptible for false positive calls. We ensured that the frequency of the alternative alleles was between 0.4 and 0.6 and that were never reported as a polymorphism. Moreover, we retained only variations called on both strands.

In order to specifically identify the putative pathogenic mutation, we performed a filter designed to select those variants segregating in the family in accordance to dominant, recessive and X-linked patterns of transmission. According to the information obtained on the family history, ID was reported as sporadic in all cases (not recurrent in the pedigree), thus dominant model was tested to identify *de novo* mutations. After that, we manually investigated the raw sequence reads for all the

candidate pathogenic variants using the Integrative Genomics Viewer (IGV) [20] to exclude likely false positive calls due to read misalignment.

The impact of selected missense mutations on the protein structure was assessed using three *in silico* predictor tools, namely, Polyphen-2 [21], MutationTaster [22], and SIFT [23]. Moreover, evolutionary conservation of nucleotides across species was evaluated using GERP [30] and CCAD algorithm [32]. Indeed, we expect pathogenetic variants to be phylogenetically conserved given the severe impact of ID phenotype. Finally, the most likely identified disease-causing SNVs/INDELS were analysed by direct Sanger sequencing in probands and their parents, to both exclude false positive SNVs calling and to perform the segregation analysis within the family.

We also investigated the presence of CNVs with the use of CoNIFER (COpy Number Inference From Exome Reads), a software for the detection of deletions and duplications on exome sequencing data [24]. Results of the CNVs analysis were then evaluated for the segregation in the families according to the three patterns of inheritance.

3.3 Discovery of novel candidate mutations

Globally, for the 6 analysed subjects, the percentage of reads mapped in target was on average 94.79% and the proportion of targeted basis covered at 20x or greater was on average 82.29% (Table 3.2).

In probands 12-781 and 11-965 at total of 45482 and 45762 SNVs/INDELS have been annotated respectively. After the application of our custom bioinformatic pipeline, variants have been selected according to three different patterns on inheritance: *de novo*, X-linked, and recessive (Figure 3.1). Under the *de novo* model 34 variations were identified but only 3 of them resulted as deleterious by the predictors. However, Sanger sequencing did not validate the mutations in the probands thus considered false positives. Considering the recessive model, 8 variants were selected but none of them were predicted as pathogenetic. The X-linked pattern showed 41 variants and 3 of them revealed high pathogenicity scores (Table 3.3). All the 3 X-linked mutations were validated with Sanger sequencing and resulted as segregating in the families.

In proband 12-781, an X-linked frameshift mutation (NM_001193466, exon8, c.2078_2079insA, p.P693fs) was identified in *GPR64* (or *ADGRG2*) gene, a member of the G protein-coupled receptor family. It was originally described as expressed in the epididymis and studied for its potential role in male fertility. More recently it was found to be up-regulated in a number of carcinomas, including breast cancer. *GPR64* has never been implicated in ID or other neurodevelopment disorders and the functional studies carried out in the mouse revealed that the gene expression was restricted to testis [31]. Although we could not exclude that this mutation is involved in the aetiology of ID, given the very few studies conducted on the function of the gene and the information available, we consider unlikely a pathogenetic role for this mutation.

Two X-linked variations were found in proband 11-965. One is a missense mutation on *CLCN4* (NM_001256944, exon10, c.G1909A, p.G637R), a member of the family of voltage-dependent chloride channel genes. The affected brother was also evaluated by Sanger sequencing and he was found to share the same hemizygous variant. *CLCN4* encodes the electrogenic chloride/proton exchanger ClC-4 and

Family	Subject ID	On Target(%)	20x Coverage(%)	SNVs	INDELs
T1	12-781	94.19	84.73	43506	1976
T1	12-779	95.03	85.53	43370	2059
T1	12-780	95.91	87.75	44848	2145
T2	11-965	94.92	86.50	43748	2014
T2	13-1536	94.32	85.54	43579	1995
T2	13-1537	94.36	87.68	43909	2051

Table 3.2: **WES data statistics.** *per*-subject summary statistics are reported (ID-patients highlighted in grey): proportion of read mapped on the target region (*On Target(%)*), proportion of target region covered at 20x or greater (*20x Coverage(%)*), number of called single nucleotide variations (*SNVs*), number of called insertions/deletions (*INDELs*).

despite little is known about the physiological role of the protein, it captured our attention since other members of CLC family are required for normal brain function. In 2013, Veeramah et al. [33] found a *de novo* non-synonymous change in *CLCN4* in a patient with refractory complex partial seizure, severe developmental delay with cognitive impairment, microcephaly, hypotonia. They performed in vitro analyses using cell-based assays revealing a great impact of *CLCN4* mutation on the ion transport by the ClC-4 exchanger. Moreover, in a very recent study conducted by Hu et al. [34] where a massive X chromosome sequencing has been applied for more than 400 ID-patients, *CLCN4* has been specifically associated to ID. Missense variants were discovered in 5 unrelated families where affected males showed variable degree of ID (ranging from mild to severe) and also intra-familial variability in the clinical features that included epilepsy, dysmorphic face, scoliosis and strabismus; in one case the ID was non-syndromic. In the study, primary neurons derived in the mouse from *Clcn4* knock-out in mice were analysed and showed a significant effect on neuronal differentiation. Moreover, in order to demonstrate the effects of mutation, authors performed analyses on *Xenopus laevis oocytes* and showed that detected mutations alter the kinetics of activation and impair the function of the ClC-4 protein.

Another X-linked missense mutation was found in proband 11-965 in gene *ALG13* (NM_001099922, exon23, c.C2672T, p.S891F). Again, the variation was detected also in the affected brother. *ALG13* encodes a protein that interacts with glycosylation 14 homolog to form a glycosyltransferase complex that functions during protein asparagine N-glycosylation. A *de novo* missense mutation of *ALG13* has been previously reported in one child with congenital disorders of glycosylation (CDG) who died at the age of 1 year [35]. The patient showed refractory epilepsy with intractable seizures, microcephaly, hepatomegaly, edema of the extremities, recurrent infections and increased bleeding tendency and died at 1 year. Interestingly, more recently Bissar-Tadmouri N. et al. [36] carried out a study on an Arab family with five unaffected female siblings and four male siblings with mild to moderate nonsyndromic intellectual disability as the only clinical feature. The family was analysed with WES approach and the only variation survived to filtering and validation was a novel missense variant in the *ALG13* gene. Given the previous

Subject ID	Sex	ID Degree	Gene	Chr	Inheritance	Ref sequence	cDNA change	AC change	PhyloP	PhyloPhen2	SIFT	Mut test	Reference	MAF
12-781	M	X	GPR64	X	X-linked	NM_001184835	c.C2908G	p.R970G	C	D	D	D	novel	NA
11-965	M	X	CLCN4	X	X-linked	NM_001256944	c.G1909A	p.G637R	C	P	D	D	novel	NA
11-965	M	X	ALG13	X	X-linked	NM_001099922	c.C2672T	p.S891F	C	D	-	-	novel	NA

Table 3.3: List of mutations identified in our cohort of patients.

PhyloP = C conserved, N not conserved.

PolyPhen2 = D probably damaging, P possibly damaging, B benign.

SIFT = T tolerated, D deleterious.

MutationTaster = D disease causing; P polymorphism.

MAF = Minor frequency allele (data from ESP6500).

finding for *ALG13* gene, authors conjectured a pathogenetic role for the discovered variation.

CNVs examination highlighted deletions and duplications that were already detected in the analysis of SNP array data. Variations were evaluated at the time and resulted not implicated in the disease.

Overall, this study led to the discover of 3 X-linked variations in NSynd-ID patients and two of them appeared as good candidates for having a causative role. In family F1 a single variant was detected but its implication in ID was considered improbable since the affected gene *GPR64* is involved in male fertility. Possibly a more accurate examination of low-coverage regions that could hide additional variations is required. In family F2 two potentially causative mutations were discovered. The first mutation affects *CLCN4* that very recently was demonstrated to have an effect on neuronal differentiation and cause ID in 5 unrelated families with variable ID severity and clinical features, including epilepsy and a case of NSynd-ID. Reported phenotype variability and the occurrence of epilepsy is consistent with the manifested disease in family F2. In the same family we also found a missense mutation in *ALG13* gene, that was recently associated with X-linked NSyn-ID. In the first mutation reported for *ALG13* by Timal et al. [35], the patient showed a severely reduced GlcNAc-transferase enzyme activity and thus diagnosed a congenital disorders of glycosylation type I, a recessive neurometabolic disease characterised by an enormous heterogeneity in clinical symptoms. We therefore believe that the deleterious effect of the mutation on the proteine could be demonstrated by measuring GlcNAc-transferase enzyme activity in the two affected brothers. Unfortunately, at the stage of this study DNA was the only available sample. Therefore, additional investigations will be perform to clarify the cause of the disease, primarily the analysis GlcNAc-transferase enzyme in the two affected boys and functional studies on the impact of candidate mutation in *CLCN4* channel.

Contributors

Anna Morgan¹, Chiara Belcaro², Massimo Carella³, Giovanni Battista Ferrero⁴, Corrado Romano⁵, Adamo Pio d'Adamo¹, Diego Vozzi², Flavio Faletra².

¹Department of Medical Sciences, University of Trieste, Italy

²Institute for Maternal and Child Health - IRCCS Burlo Garofolo , Trieste, Italy

³Medical Genetics Unit, IRCCS Casa Sollievo della Sofferenza, San Giovanni Rotondo, FG, Italy

⁴Department of Pediatrics, University of Torino, Torino, Italy

⁵Unit of Pediatrics and Medical Genetics, IRCCS Associazione Oasi Maria Santissima, Troina, EN, Italy

Chapter 4

Severe cognitive impairment is associated with increased Runs of Homozygosity in Intellectual Disability

This chapter describes a study on the role of Runs of Homozygosity on the cognitive deficit in ID patients, that has been published in *Genetics in Medicine* as *Gandin I. et al "Excess of runs of homozygosity is associated with severe cognitive impairment in intellectual disability." Genet Med. 2015 May;17(5):396-9.* Contributors for this work are reported at the end of the chapter.

4.1 ROH and inbreeding

The harmful effect of inbreeding is well-known by geneticists and several investigations have reported cases of neurodevelopment disorders in consanguineous families caused by recessive variants. Sons of consanguineous couples have an increased probability to be affected by a recessive pathologies due to identity-by-descent mutations, and this predisposition is inversely correlated with the distance from the common ancestor. However, only recently the attention has been directed towards the role of more ancestral inbreeding, which is difficult to be detected from the pedigree information and has a more subtle effect.

The widespread of high-density SNP-array in the last years has made possible the study of Runs of Homozygosity (ROHs), long genomic stretches of homozygosity inherited from the same ancestor, which were shown to be extremely useful to capture distant parental relatedness. In this scenario, ROHs have become a powerful tool for investigating the role of inbreeding on a large scale in outbred populations. Although first noted in fitness-related phenotypes, recently studies identified the amount and the size of Runs of Homozygosity as an important player in neurological disorders, even if the way through which homozygosity contributes in the aetiology of the disease is still unclear.

A possible interpretation is that ROHs contain specific loci harboring reces-

sive mutations that could cause the disease. Under the recessive model of inheritance, stretches of homozygosity are the first candidate regions for finding the pathogenetic mutations (homozygosity mapping). Several studies have made use of the homozygosity mapping for the identification of variants causing ID in consanguineous families, leading to the discovery of a large number of autosomal ID-genes [42, 43, 44, 12]. However, despite recessive inheritance plays an important role in ID, ancestral inbreeding has received little attention in the context of ID.

Another possibility is that, like CNVs, some ROH loci can have an effect similar to *contiguous-gene* defects. Instead of a single variant conferring disease susceptibility, there could be a cumulative or interactive effects of contiguous variants within a homozygous region. The cumulative effect of multiple recessive variants could play an important role in the aetiology of several neurological diseases, as excess of ROHs has been identified as a risk factor for Parkinson disease [37], Alzheimer disease [38], schizophrenia [39], but also speech delay and autism [40] [41].

To shed more light on the possible role of inbreeding on ID, we have set out a homozygosity study on affected individuals with two principal purposes: 1) test the extent of homozygosity as risk factor for carrying syndromic features and 2) investigate the effect of homozygosity on the degree of ID.

4.2 A Runs of Homozygosity study on ID

The study included a cohort of 668 (266 female and 402 males) affected children. We were able to collect IQ information only for 368 subjects, thus excluding the remaining individuals from all subsequent analysis. Individuals were then divided into non-severe ID cases (NSev-ID) for IQ ranging from 35 to 75, and severe ID cases (Sev-ID) for IQ below 35.

Samples were genotyped using 4 SNP-array platforms: Illumina HumanCytoSNP-12v1, HumanCNV370v1, HumanOmniExpress-12v1 and Affymetrix Genome-Wide Human SNP 6.0, depending on the Health Care Unit providing the samples. Standard genotyping quality control was performed removing low quality markers and samples.

To avoid ancestry bias we performed Principal Component Analysis (PCA) and removed outlier individuals. After this step we proceeded with the detection of ROHs using the analysis toolset PLINK [45]. We searched for continuous stretches of at least 100 homozygous autosomal markers and a minimal length of 1Mb. We chose 1Mb as minimal ROH length since sparser arrays (HumanCytoSNP-12v1 and HumanCNV370v1) are not able to identify shorter stretches. Individuals with ROHs summing up more than 6.25% of the whole genome were excluded since they are typical of descendants of close consanguineous marriages [39]. Moreover, in order to identify possible deletions that will be misclassified as ROHs, we performed CNV detection analysis using PennCNV [19].

For each subject we considered: 1) the total amount of homozygosity (KB), which is the total Kb spanned by all ROHs, and 2) the mean length of ROH segments (KBAVG). To overcome differences in ROH distribution due to the different array densities, we chose to perform a logistic regression in each dataset and then combine the results in a meta-analysis using a standard inverse variance method, in which the contribution of each dataset is weighted and correlated with its size.

4.3 The effect of ROHs on ID severity

After QC, 19 individual were excluded from genotyped samples, and other 37 were removed because identified as outliers. Moreover, we detected a total of 32 deletions and the corresponding genomic regions have been excluded from further analysis. At the end of these steps, we obtained a clean study sample of 612 individuals. As expected, the distribution of KB and KBAVG differed between the 4 arrays (Table 4.1).

As a first step, we tested the effect of ROHs on the presence of syndromic features in individuals with ID. Syndromic conditions are more common in very close inbreeding and although one mutation can be the cause of Syn-ID, different features may be due to multiple recessive homozygous mutations in different genes. We investigated the correlation between homozygosity and complexity of phenotype in ID cases. We regressed the syndromic/non-syndromic status on KB and KBAVG for 612 individuals (187 NSyn-ID and 425 Syn-ID) and no association was detected for KB (ODD=1.02, CI=[0.88,1.20], $p=7.5 \times 10^{-1}$), neither for KBAVG (ODD=0.99, CI=[0.94,1.04], $p=7.1 \times 10^{-1}$).

Based on our analysis, ROHs are not associated with the complexity of the phenotype (malformations, seizure, micro/macrocephaly, etc.). A possible explanation is that in our cohorts the majority of syndromic cases are not the result of multiple

Dataset	n SNPs	ROH Statistics					
		Min	Q1	Median	Mean	Q3	Max
TOTAL HOMOZYGOSITY (KB)							
Affymetrix	700767	7502	27860	33220	34650	38960	77480
Human OmniExpress	626132	9392	23290	28730	30950	34350	151500
Human CNV370	291981	2322	6818	10810	12450	14410	61490
Human Cyto	242657	3858	9994	13040	14810	17000	96130
MEAN LENGTH (KBAVG)							
Affymetrix	700767	1209	1209	1411	1480	1505	3099
Human OmniExpress	626132	1174	1339	1406	1489	1513	5410
Human CNV370	291981	1161	1477	1638	1845	1963	5331
Human Cyto	242657	1133	1420	1563	1703	1824	4370

Table 4.1: Detailed summary statistics for global homozygosity (KB) and mean length of ROH stretches (KBAVG) (source: Gandin et al. 2015). Total amount of homozygosity and the ROH size depend on the number of SNPs included in the array, which is reported in the second column of the table. The lower KBAVG values for denser arrays highlights that they are able to detect a greater number of short ROHs, entailing also a higher score for KB compared to the sparser arrays.

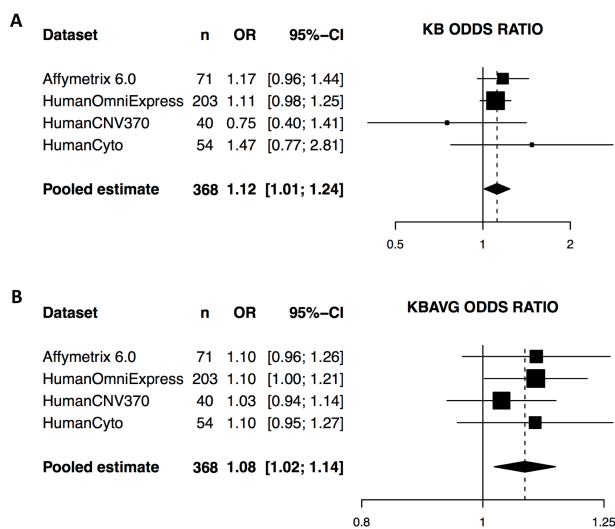


Figure 4.1: **Single dataset and combined analysis for global homozygosity (A) and mean length of ROH stretches (B)** (source: Gandin et al. 2015). In (A) we considered an increase of 10 Mb in KB, while in (B) the increase in KBAVG was set to 100 Kb. For each analysis we reported the number of individuals (n), the estimated changes in odds (OR) and the 95% confidence intervals (95%-CI). On the right side, we reported the meta-analysis forest plots: the box size in the graph is proportional to the contribution of each dataset.

gene defects but instead single-gene ID syndromes.

The second part of the study was focused on the degree of ID (294 NSev-ID and 74 Sev-ID). We found significantly larger homozygous stretches in Sev-ID compared to NSev-ID case, together with an increased amount of global homozygosity, and this despite the lower number of individuals compared to the previous analysis. In Figure 4.1 we reported the odds for severe status and the 95% confidence interval from the single logistic regressions and the meta-analysis. Notably, for every 10 Mb increase in total homozygosity the odds for a severe status are increased by 12% (ODD=1.12, CI=[1.01, 1.24], $p=3.0 \times 10^{-2}$). Moreover, there is a strong association between KBAVG and the ID degree (ODD=1.08, CI=[1.02, 1.14], $p=7.1 \times 10^{-3}$). This means that for every 100kb increase in KBAVG the odds for Sev-ID status are increased by 8%. To check the presence of confounder factors, several models were tested including sex, first 3 principal components for possible population stratification, sample origin and Syn-ID/NSyn-ID status: statistical significance for KBAVG was obtained for all the tests.

We detected a significant association with both total quantity and mean length of ROHs with the degree of ID. Not surprisingly, these results are very similar to the effect of homozygosity on ID in simplex autism [41]. The stronger effect of length with respect to total amount of ROHs can be explained with two mechanisms. On one side we have to face with an instrumental limitation: we could not analyse ROH shorter than 1Mb due to the presence of too sparse arrays. Since we are missing shorter stretches, what is driving the association are long ROHs just because they are easily detected. Moreover, it is possible that longer ROHs, which arise from closer inbreeding, could have actually a greater effect according to Szpiech [46], which reported an enrichment of deleterious variants in long ROHs in healthy individuals. Similarly, long stretches of homozygosity may contain more likely damaging variants arising from close inbreeding, which have not yet been purged

by selection. Those variants may have an important role in ID modulation.

Further investigation would require testing our findings in a larger cohort employing whole genome sequencing analysis, in order to both obtain a better estimate of ROHs extent (accounting also for shorter stretches) and try to identify which genes are mainly involved in ID degree modulation. Probably, the most important consequence of this study is that, despite the heterogeneity of ID causes in our cohorts (CNV, INDELS, single-nucleotide-variations), the distribution of homozygosity seems to have an effective impact in determining the severity of impairment. This is not surprising, since cognitive ability is a very complex trait and its total variability is unlikely determined by a single pathogenic mutation. Other environmental and genetic factors modulate the phenotype and one of the most important seems to be accounted by ROHs.

Contributors

Flavio Faletra¹, Francesca Faletra¹, Massimo Carella², Vanna Pecile¹, Giovanni Battista Ferrero³, Elisa Biamino³, Pietro Palumbo², Orazio Palumbo², Paolo Bosco⁴, Corrado Romano⁴, Chiara Belcaro⁵, Diego Vozzi¹, Adamo P. d Adamo⁵.

¹Institute for Maternal and Child Health - IRCCS Burlo Garofolo , Trieste, Italy

²Medical Genetics Unit, IRCCS Casa Sollievo della Sofferenza, San Giovanni Rotondo, FG, Italy

³Department of Pediatrics, University of Torino, Torino, Italy

⁴Unit of Pediatrics and Medical Genetics, IRCCS Associazione Oasi Maria Santissima, Troina, EN, Italy

⁵Department of Medical Sciences, University of Trieste, Italy

Chapter 5

Y chromosome and General cognitive function

In this chapter we present the design of an association study for Y chromosome to analyse the general cognitive function. Setting up and preliminary results of the analysis represent the first stage in the creation of the Ygen consortium [47]. Contributors for this work are reported at the end of the chapter.

5.1 Possible role of Y chromosome on sexual dimorphism in diseases

Intellectual disability has a significantly higher prevalence in males compared to females. X chromosome has been proved to play a fundamental role in the aetiology of ID, given the presence of a large number of gene expressed in the brain and several ID-genes identified in linkage association studies. However, the cause of male bias still needs to be clarified. The proportion of explained X-linked cases has been usually around 10-12% but this proportion is still far from the 40% excess of the disease observed in males, which is even higher if considering mild forms of ID occurring around 1.9-fold more likely in boys compared to girls [4]. Therefore, other factors are believed to increase the susceptibility in males. A new investigation field regards the effects played by hormonal mechanisms. It has been hypothesized that sex hormones in females could have a neuroprotective action. Moreover, the exposure of fetal testosterone has been proved to affect the brain development in animal models and also behaviour in cognitive development in children. Those results are currently investigated especially for male bias in autism [48].

There are many complex diseases having a large impact on public health that are sexually dimorphic and specifically more prevalent in males. Most common examples are the broad range of cardiovascular diseases and type 2 diabetes. Another important category are neurological and neuropsychiatric conditions in which the prevalence and symptomatology differ significantly between males and females. Strong male bias is observed for autism, attention-deficit hyperactivity disorder (ADHD), Tourette's syndrome, schizophrenia, Parkinson's disease. Very little is known about the underlying genetic architecture of these differences and new studies are trying to explore the role of previously overlooked genetic players, like the

Y chromosome.

The fundamental biological role of the Y chromosome is to impart male characteristics and it is carried only by men. However, it totals about 25 Mb of sequenceable DNA containing 27 genes or gene families that were entirely neglected in the search for the determinants of common complex disease. Thus their role is still largely unknown. A recent study in more than 3200 men using few Y chromosome markers discovered that haplogroup I conferred a 56% higher risk of coronary artery disease [49], demonstrating the importance of investigating this genomic region.

In this context, the laboratory of Dr. Wilson at the University of Edinburgh (UK) has recently formed a consortium aimed to investigate the role of Y chromosome on complex traits in a large number of genetic cohorts [47]. In the project more than 30 diverse traits of public health or evolutionary interest are considered, including anthropometric measures, haematological counts, cardiovascular risk factors, metabolic markers, inflammatory response and fertility. Notably, the study considers also the general cognitive function, which is a global score for fluid cognitive abilities [50]. Giving the extreme interest of this topic for neurodevelopment disorders and the innovative nature of the study, a fruitful collaboration has been established with Wilson's lab and we had the possibility to actively contribute in the design of the study and its realisation, with a primary focus on outcomes related to cognitive function trait.

The hypothesis of a possible involvement of Y chromosome in ID is controversial. While there are recent studies revealing an increased expression of Y-linked genes in autism and ADHD disorders [51], up to now there is no evidence for similar results in ID. Recent reviews seem to support the intervention on Y chromosome for neurological disorders in general [52], mainly based on animal studies in which several Y-linked genes were found to be expressed in the SRY region, possibly having a regulatory role. On the other side, in literature there is no trace of ID-affected families showing a Y-linked pattern of inheritance, thus making unlikely Y chromosome variations having a pathogenetic role. However, we consider extremely interesting to investigate a possible influence of the Y chromosome on the cognitive function, which could be reflected, in case of ID, in a modulation of the cognitive impairment degree despite not being the cause of the disease.

Since this project represents the first Y chromosome association study on large scale, the first part of the work has been focused on the research of the methodology to use, which have been tested on 9 cohorts, providing preliminary results. We therefore present here the pipeline that has been developed and the preliminary association outcomes for general cognitive function, which has been measured in 5 of the cohorts.

5.2 Association study on the Y Chromosome

We considered 9 cross-sectional men-cohorts formed by healthy individuals (see Table 5.1).

General cognitive function has been calculated for 5 cohorts (see Table 5.1) as in Davies et al. [50]. Subjects were administered from 3 to 5 tests aimed to measure different cognitive domain as memory, reasoning, processing speed, spatial ability. A principal component analysis was carried out on the different cognitive scores and the measure of general cognitive function was obtained as the resulting first unrotated principal component. In this way, general cognitive function represented

Name	Label	Country	n	Reference
EGCUT Study	EGCUT	Estonia	2176	Nelis et al. 2009
AGES Reykjavik	AGES	Iceland	2216	Harris TB et al. 2007
Orkney Complex Disease Study	ORC	Scotland, UK	320	McQuillan et al. 2008
Generation Scotland	GS	Scotland, UK	3490	Smith BH et al. 2012
Lothian Birth Cohort 1921,1936	LBC	Scotland, UK	683	Deary IJ et al. 2007
INGI-Val Borbera	VB	Italy	679	Colonna V et al. 2013
INGI-Carlantino	CARL	Italy	168	Esko T et al. 2013
Korcula	KOR	Croatia	273	Polasek O et al. 2009
Silk Road	SR	Central Asia	148	Mezzavilla M et al. 2014

Table 5.1: **Cohorts included in the study.** For each cohort, name, short name, origin, number of males (n) and reference are reported. The total sample size is 10153. General cognition function was measured only in the 5 cohorts highlighted in grey (EGCUT, ORC, GS, LBC, KOR), while education attainment was obtained for all cohorts.

the variance that crosses the different fluid cognitive functions. We also considered educational attainment has been measured in all cohorts as US years of school [53].

We analysed SNP-array data genotyped with the Illumina ExomeChip platform, a custom-designed chip widely used in the last years since its affordable cost. The Y chromosome part of the chip contains markers specifically designed to represent the Y chromosome variability in world-wide populations. Based on preliminary examinations on the quality of markers, 69 SNPs of the SRY region have been selected for the analysis, which are able to uniquely identify 65 Y chromosome haplogroups. Phylogenetic tree and haplogroup definitions are reported in Figure 5.1.

For each cohort, following a standard control on marker and sample quality, selected SNPs were used to estimate the haplogroup which each individual belongs to. This step was essential for the successive inference of missing genotypes. Haplogroups were defined in terms of SNP configuration following the hierarchical structure reported in the phylogenetic tree (Figure 5.1). For each individual, his genotypes and the haplogroup definitions were compared, then the best fitting haplogroup was selected. If the best-fitting haplogroup was not unique (possibly because some missing/mismatching SNP, leading to an ambiguous assignment), then the haplogroup was considered as undetermined. If the best-fitting haplogroup did not match at least 97% of the SNPs or there were more than one mismatching SNP, the haplogroup was set undetermined as well.

After haplogroups have been assigned, we were also able to infer the genotypes of 12 SNPs that were not included in the ExomeChip but known to lie in our phylogenetic tree. Once the haplogroup was known, we could infer with certainty a SNP lying upstream in the genealogical tree, even if it was not genotyped. This technique was also used to infer missing calls among the genotyped SNPs and finally

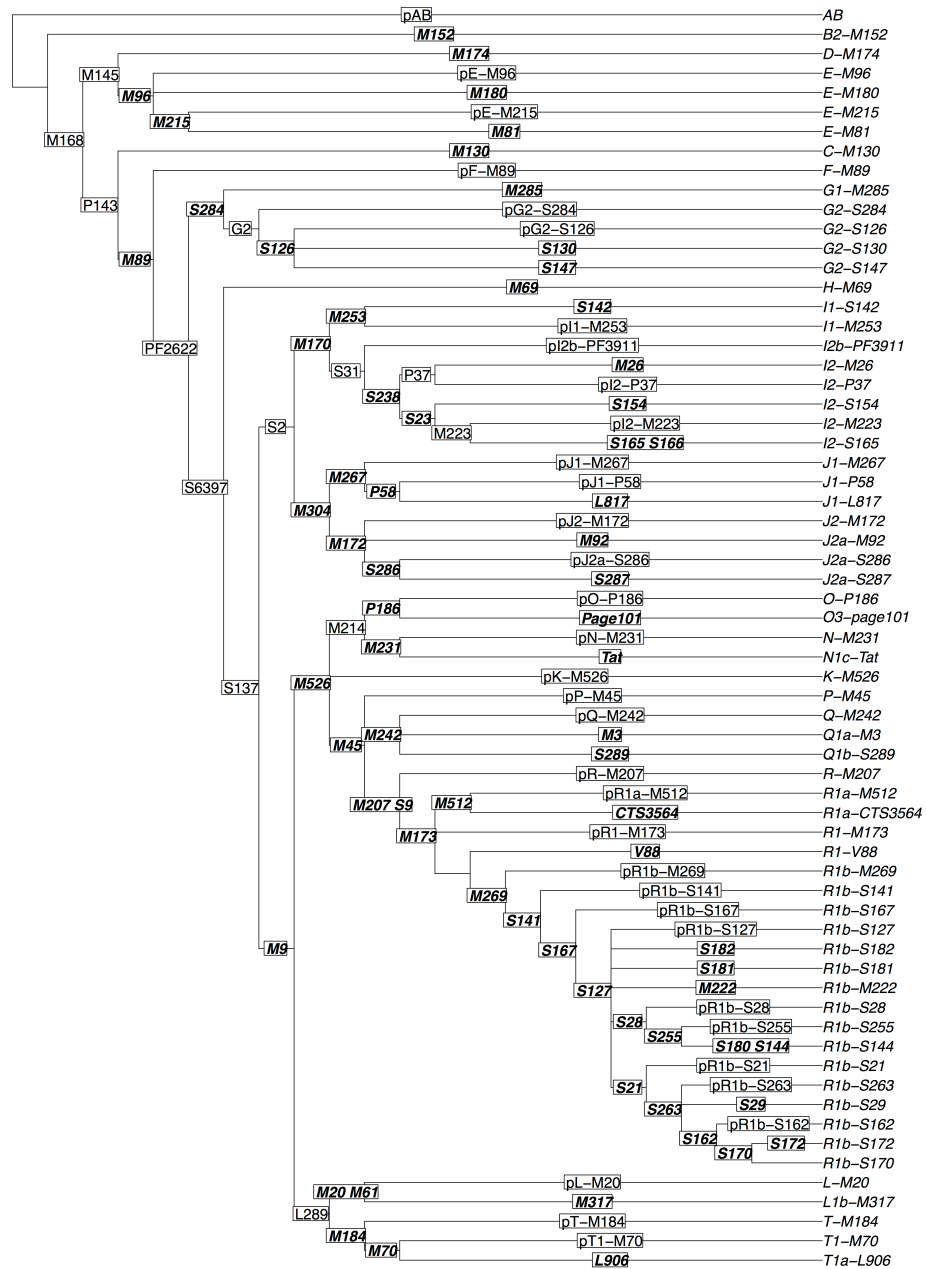


Figure 5.1: **Phylogenetic tree of the 69 SNPs for the identification of 65 haplogroups.** In the tree, labels on the branches represent the markers that are named according to the traditional nomenclature for the Y chromosome. Final leaves labels represent haplogroups. Markers in bold are genotyped SNPs, while not bolded stand for imputed SNPs that could be inferred after the haplogroup assignment for subjects. Labels with *p* prefix indicate *para-SNPs*, which are dummy tag-SNPs for those haplogroups that are not identified by a single SNP.

also to infer 32 of what we call *para-SNPs* (markers with *p* prefix in Figure 5.1) that can be thought of as tag-SNPs for paraphyletic haplogroups (see Appendix A for a detailed description). After this step, we obtained of 113 Y chromosome SNPs dataset and fully exploit the Y data available on the Illumina ExomeChip. Each of these SNPs represents many further SNPs which are phylogenetically equivalent.

Very few studies have been carried out on Y Chromosome and complex traits, and most of them used an haplogroup-based approach. This means that samples are first divided according to their haplogroups and then association is done across the categories, comparing one group versus all the others. Such approach was hardly adaptable in our study. We expect to be able to capture a large number of haplogroups thus the inter-group comparison would have led to many tests. Moreover, the high-resolution in haplogroup estimation will likely result in several sub-groups with small number of individuals that are too few to be analysed as a single category but that can be grouped together to form a more substantial group. For these reasons we opted for a single-SNP approach in the style of standard Genome-Wide Association Studies.

Y chromosome variants (and haplogroups) are characterised by high geographical specificity, which is mostly due to evolutionary mechanisms (the low effective population size means the Y chromosome is highly susceptible to genetic drift) and sociological reasons (a predominance of patrilocal societies - where men tend to stay in one place over generations, e.g. to inherit land). For this reason the first 10 Principal Components and a genomic relationship matrix (for related populations) were calculated to assist in controlling for cryptic structure across populations.

In addition to pc1-pc10, where a cohort is aware of other covariates (e.g. sampling centre) these are included. For all the analyses, we first fitted covariates in a linear regression model and then association was performed over the residuals with the PLINK function `--assoc` for quantitative traits. For studies with related populations the mixed model was also calculated. Using GenABEL's `polygenic()` function, we first calculated a mixed model including the kinship matrix as a random (polygenic) effect, then we performed association over the environment-corrected residuals with PLINK.

A standard inverse-variance meta-analysis has been applied to single-cohort results to obtain a global estimate of the association in which only SNPs having minor allele count ≥ 10 have been considered. Significance level was set to $0.05/113 = 0.0004$.

Most of the operations listed above, including haplogroup assignments, SNP imputation, association, have been implemented in a collection Unix and R scripts.

5.3 Need for higher Y chromosome's genetic variability

Haplogroup assignment was successfully performed in more the 95% of samples in the majority of the cohorts. Only in AGES and CARL the proportion of individuals with determined haplogroup was less then 95% (respectively 92% and 80%) due to the presence of few missing SNPs. Therefore, in those cohorts SNP imputation was not performed and the analysis was carried out only on the genotyped markers. The high overall scores and the correspondence between expected and ob-

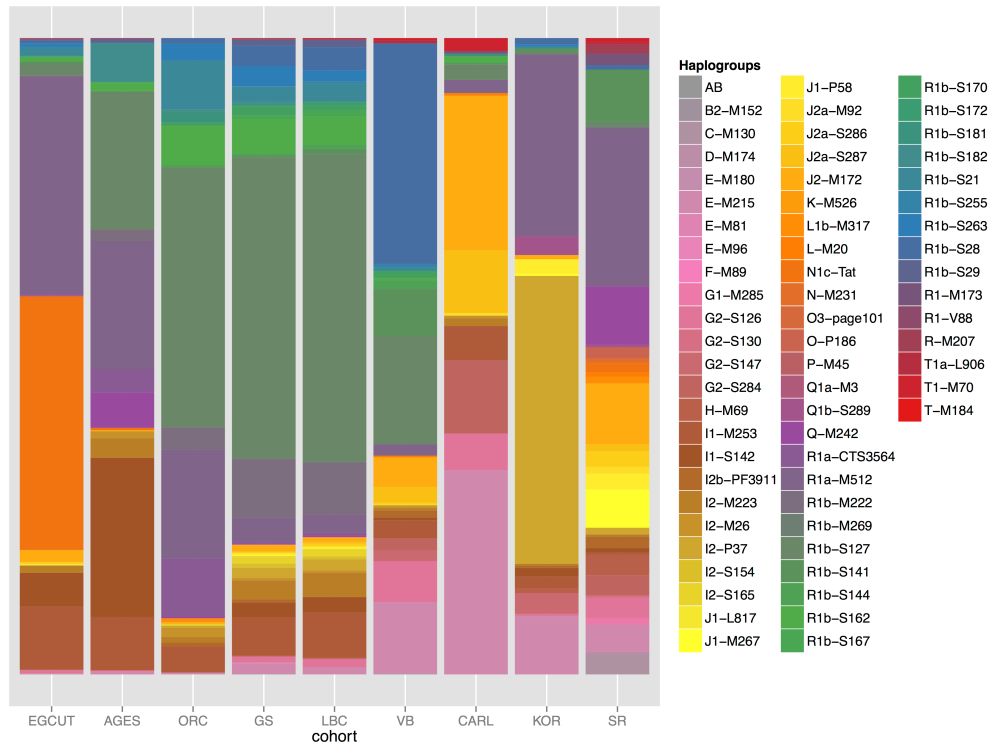


Figure 5.2: **Frequencies of estimated haplogroups.** For each cohort the haplogroups distribution is represented. Interestingly, there are some groups that represent almost half of the cohort, like N1c-Tat in EGCUT and I2-P37 in KOR, but almost completely absent in other populations. As expected, Scottish populations show a very similar frequencies pattern.

served haplogroup frequencies demonstrated our algorithm as an accurate method for haplogroup assignment. Indeed, as reported in Figure 5.2, in the Estonian cohort EGCUT the group that accounts for almost half of the cohort is N1c-Tat, which is a very common haplogroup in the Baltic region [54]. Scottish cohorts (ORC, GS, LBC) present very similar frequencies. Moreover, in Croatian cohort KOR the most frequent group is I2-P37, which is a sub-group of the very common I-M170 group in Croatian islands [55]. The most variegated is the Silk Road cohort (SR), being a mixture of several populations from the Central Asia.

In those cohorts where haplogroups have been assigned in at least 95% of the dataset, 44 SNPs were imputed, giving a 113 SNPs dataset that have been used in the single-SNP analysis. We did not find evidence for an association between Y chromosome SNPs and general cognitive function, although the underlying major problem is possibly the lack of power in the analysis.

A global representation of association result for general cognition is reported in Figure 5.3. Here, SNPs are coloured according to the estimated effect, as reported in the legend. It appears clear that none of the SNPs showed a significant effect. Only M253 and S142 have arisen with a very weak nominal significance. Details of the meta-analysis for those SNPs are reported in the forest plot in Appendix

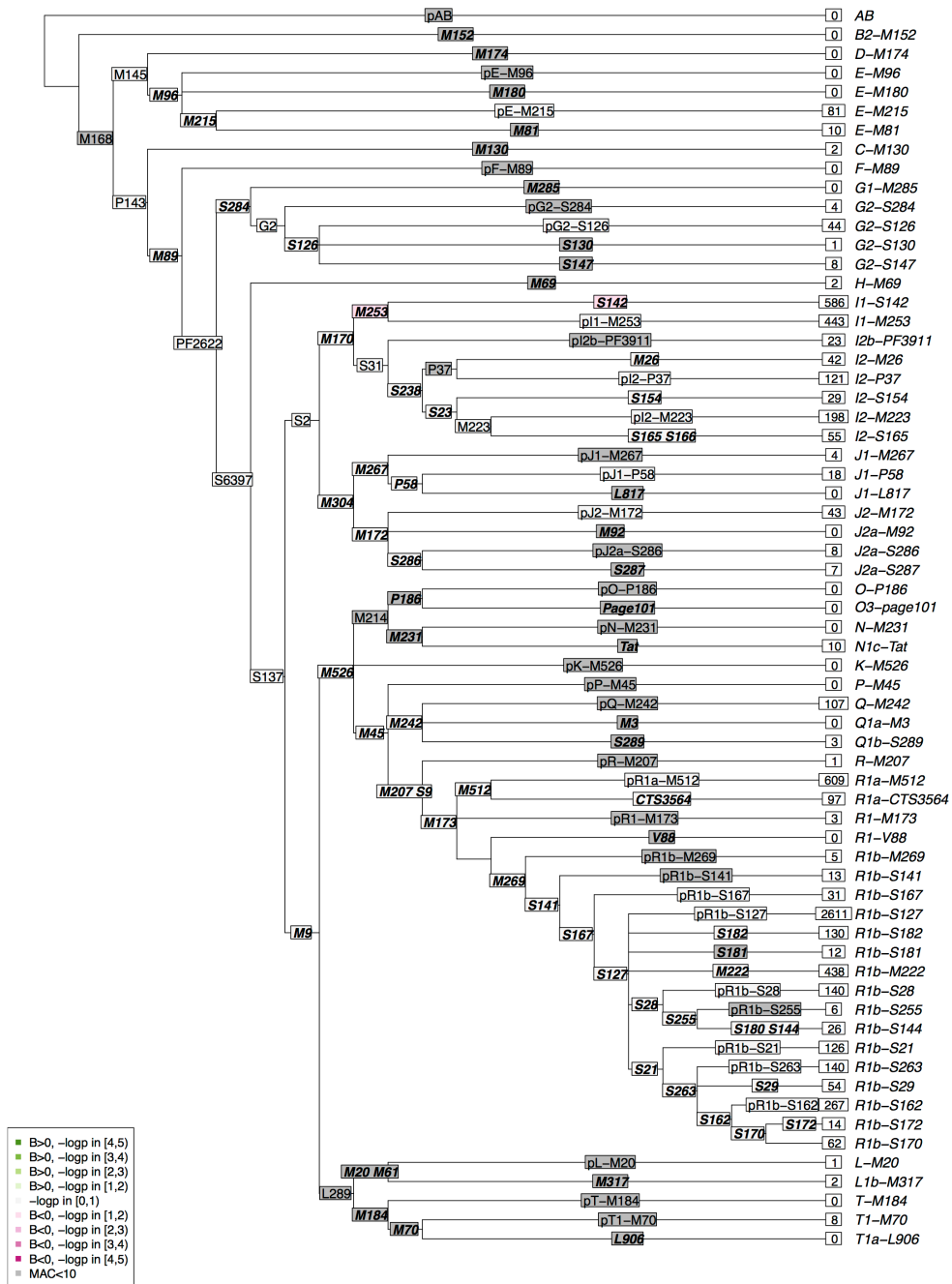


Figure 5.3: **Results of the association study for general cognition.** Results are graphically represented on the phylogenetic tree. Numbers near haplogroup labels represent the number of samples in each group. Each SNP label is coloured based on the association outcome: pink if the estimated effect is positive, green if negative. Moreover, the intensity of the colour is proportional to the significance level, as reported in the legend. SNPs in grey were not tested because too low minor allele count. No signal has been observed, only M253 and S142 show a light colouring.

B. An interesting fact emerging from the examination of the tree in Figure 5.3 is the large number of grey labels. They represent SNPs that were discarded in the meta-analysis because having minor allele count less than 10 in all cohorts. Notably, 76 out of the 144 SNPs appear to be very rare in our cohorts, despite the considerable global number of samples ($n=6799$). This fact can be considered a direct consequence of the presence of very common haplogroups (see Figure 5.2) in some of the cohorts that indicate small Y chromosome's variability with respect to our selection of SNPs. Very large groups (like N1c-Tat in EGCUT, R1b-127 in Scottish cohorts, I2-P37 in KOR) are actually subdivided in several sub-groups defined by some specific SNPs that were not included in the ExomeChip and thus impossible to detect by our algorithm. For the same reason, it should be noted that the global effect of SNPs calculated in the meta-analysis is obtained with the contribution of only a subset of the cohorts. Indeed, most of the analysed SNPs are polymorphic in some cohorts and monomorphic in others. This fact decreases even further the power of the study.

A recent genome-wide association study focused on the general cognition found an overlap between SNPs associated with the phenotype of interest and the educational attainment [50]. We therefore analysed also educational attainment coded as US years of school attended, that was measured in all cohorts. The outcomes of the two traits have been compared. Results of the association for education attainment are represented in Figure 5.4. The strongest signal has been found for SNP M145 ($\beta=0.37$, $SE=0.12$, $P=0.0025$), together with pE-M215 and P143 (see Appendix B) that are highly correlated with M145 due to the strong linkage disequilibrium. Despite macro-group E was represented in the general cognition analysis by 299 samples, it did not show any signal. Moreover, while I1 branch seems to have a negative effect on general cognition, although with nominal significance, in the case of educational attainment the estimated beta is positive, again with nominal significance ($\beta=0.12$, $SE=0.06$, $P=0.0315$).

The findings of this study led to the conclusion that there is no evidence for a role of the selected Y chromosome SNPs in determining general cognitive function in our cohorts. Education attainment has been also examined given the correlation between the two traits, but no SNPs were found to be associated and outcomes were not overlapping, even when considering results at the nominal significance level.

We also observed a great proportion of SNPs too rare to be tested in the association study. The 69 selected SNPs were designed to represent the worldwide variation of Y chromosome, however for some cohorts they are not able to provide a haplogroup subdivision with high resolution. It is interesting to note that for educational attainment the number of analysed cohorts has been increased (9 cohorts for a total of 10153 subjects) and this has resulted in a lower number of discarded SNPs (69 SNPs compared to 75 SNPs excluded for general cognition).

A problematic aspect of this study seems to be the reduce Y chromosome variability provided by our cohorts. However, findings here reported represent only the preliminary results of an international consortium that is expected to receive the contribution of a large number of groups, providing cohorts from almost all world populations [47]. The Illumine ExomeChip has been largely used in the last years since its affordable costs and the number of genetic cohorts that have been genotyped using this platform is now remarkable. Moreover, the collaboration will be facilitated by providing a collection of scripts for the execution of the analyses in the

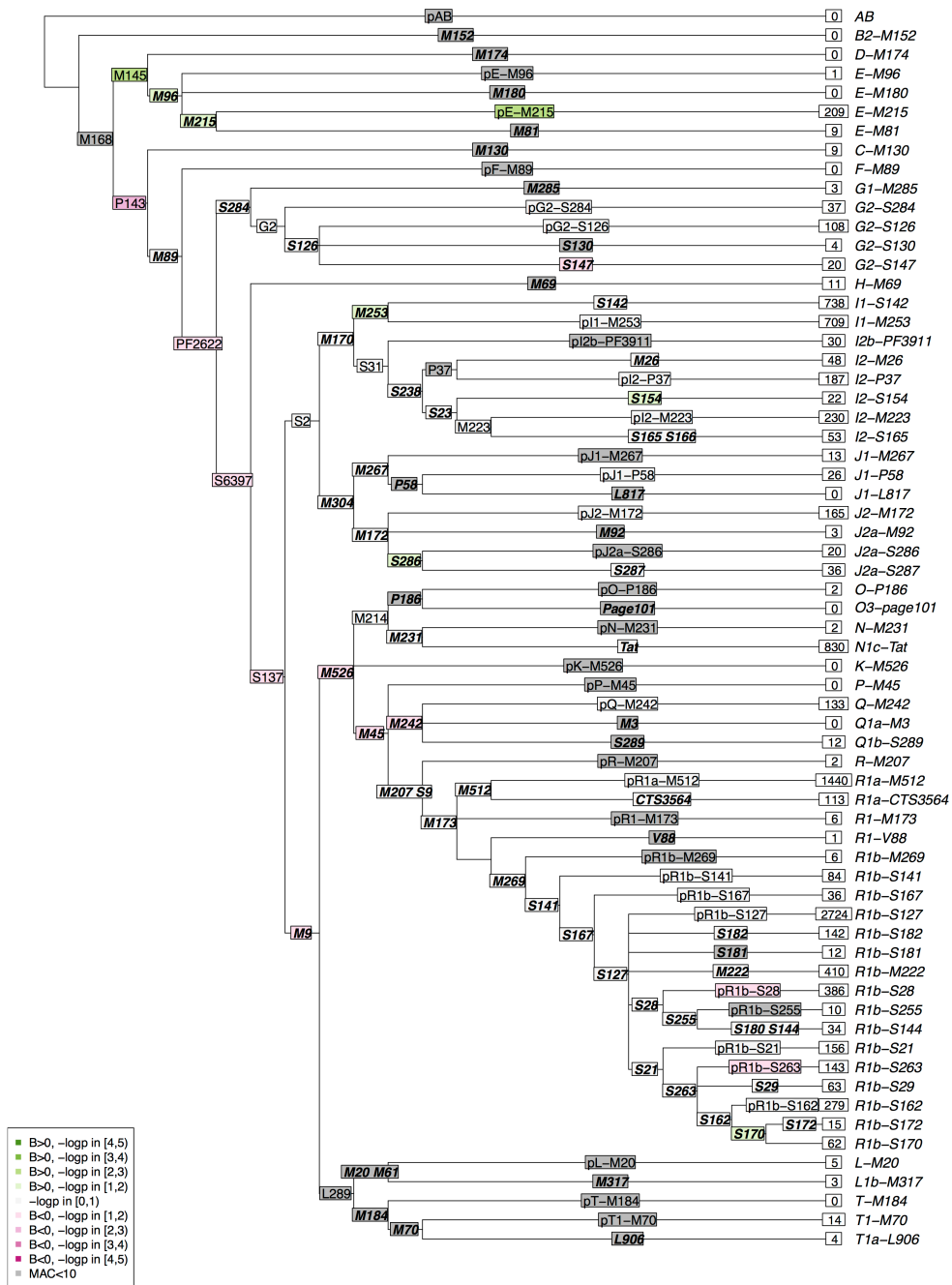


Figure 5.4: **Results of the association study for educational attainment.** Given the correlation usually observed with general cognition, also education attainment has been analysed. The strongest signal has been found in the E branch that, although represented in general cognition cohorts, did not show association in the previous analysis. SNP M253 has opposite effect compared to the general cognition analysis.

single cohorts, making the contribution of the groups simpler and faster. Thus the achievement of the required power in terms of sample size and Y chromosome variability in the next future appears plausible. The inclusion of cohorts with different ancestries as Africans, South Americans, East Asians, South Asians, will increase the chances of testing a large proportion of the SNPs in our selection and the effect of markers will be easily interpreted in terms of haplogroups, as reported in the graphical tree-representation. Any finding regards Y chromosome determinants for the cognitive function could clearly lead to fundamental new insights for the study of neurological and neuropsychiatric disorders.

Contributors

Central team: Peter K. Joshi¹, Tonu Esko^{2 3 4 5}, James F. Wilson^{1 6}.

Beta testers: Claudia Schurmann^{7 8}, Hannele Mattsson^{9 10}, Bram Prims¹¹, Asa Johansson¹².

Cohorts PIs: David Porteous^{13 14}, Ian Deary^{14 15}, Albert V. Smith^{16 17}, Eleftheria Zeggini¹¹, Markus Perola^{2 17}, Harry Campbell¹, Ruth Loos^{7 8}, Andres Metspalu^{2 18}, Caroline Hayward⁶, Daniela Toniolo¹⁹, Paolo Gasparini^{20 21 22}.

¹Centre for Global Health Research, Usher Institute for Population Health Sciences and Informatics, University of Edinburgh, Edinburgh, Scotland

²Estonian Genome Center, University of Tartu, Tartu, Estonia

³Division of Endocrinology and Center for Basic and Translational Obesity Research, Boston Children's Hospital, Cambridge, MA, USA

⁴Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, USA

⁵Department of Genetics, Harvard Medical School, Boston, MA, USA

⁶MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Edinburgh, Scotland

⁷The Charles Bronfman Institute for Personalized Medicine, 103 Icahn School of Medicine at Mount Sinai, New York, USA

⁸The Genetics of Obesity and Related Metabolic Traits Program, Icahn School of 106 Medicine at Mount Sinai, New York, USA.

⁹Unit of Public Health Genomics, National Institute for Health and Welfare, Helsinki, Finland

¹⁰Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland.

¹¹Human Genetics, Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK

¹²Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden

¹³Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden

¹⁴Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh, UK

¹⁵Psychology, University of Edinburgh, Edinburgh, UK

¹⁶Icelandic Heart Association, Kopavogur, Iceland

¹⁷Faculty of Medicine, University of Iceland, Reykjavik, Iceland

¹⁸Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia.

¹⁹Division of Genetics and Cell Biology, San Raffaele Scientific Institute, Milan, Italy

²⁰Institute for Maternal and Child Health - IRCCS Burlo Garofolo, Trieste, Italy.

²¹Sidra Medical and Research Center, Doha, Qatar

²²Department of Medical Sciences, University of Trieste, Italy

Conclusion

Intellectual Disability is a complex disease and this thesis represent an attempt to shed more light on its the genetic bases using diverse approaches.

In the first part we have focused on the search of causing mutations in a Mendelian setting, where the disorder is transmitted according to dominant, recessive or X-linked pattern. Thanks to a pre-screening strategy on the most common genes for the non-syndromic ID, we have been able to identify novel candidate mutations for the disease. Some of these variations are located in genes already associated to the disease, but others have been found in genes only recently implicated on ID or only hypothesised to be involved.

In the second part of the work, ID have been approached as a more complex disorder, considering that its phenotype is unlikely determined by a single variant. We have focused our attention on other genetic factors that, despite not being the main cause of the disease, can influence the phenotype in patients. We have shown that excess of homozygosity has an effective impact in determining the severity of impairment. Such result has led to speculations about a possible cumulative effect of multiple recessive variants in the modulation of the degree of the impairment. Based on recent findings for other neurodevelopment disorders, we have also analysed the possible role of Y chromosome variants in the development of the cognitive function. Our preliminary results do not show associations but the study will gain more power adding more cohorts in the next future.

We can conclude that the research on the causing mutations is a fundamental part in the investigation on ID, since there is a large number of genetic variants with pathogenetic role that still need to be discovered. At the same time, it is important to approach ID as polygenic disorder, since cognitive function is the result of the effect of many genetic factors.

Appendix A

We describe here the inference of missing SNPs in the association study for the Y chromosome.

After the estimation of haplogroups for individuals, we were also able to infer 12 SNPs and 33 *para*-SNPs. In Figure A.1 a graphical representation of the mechanism is reported for zoomed part of the tree. Haplogroup assignment was performed on the genotyped SNPs (Fig. A.1 (a)).

Since there were 12 SNPs that have not been included in the ExomeChip but known to lie in our phylogenetic tree (Fig. A.1 (b)), it was possible to infer with certainty the genotypes for our subjects.

Some of the haplogroup are uniquely identified by single SNPs, like D-M174 which is determined by SNP M174. But there are others, like E-M521, which are paraphyletic groups and thus not represented by a unique marker. In order to include in the analysis the possible effects of paraphyletic groups, we defined 33 *para*-SNPs, that are dummy SNPs which capture a haplotype effect in one binary variable (Fig. A.1 (c)). These SNPs do not really exist, but they are an efficient way to capture all the information in the genotypes along with the known genealogical relationships. These SNPs are assigned dummy positions and alleles (I/O, I=*in* O=*out*), and can be thought as tag-SNPs for the paraphyletic groups.

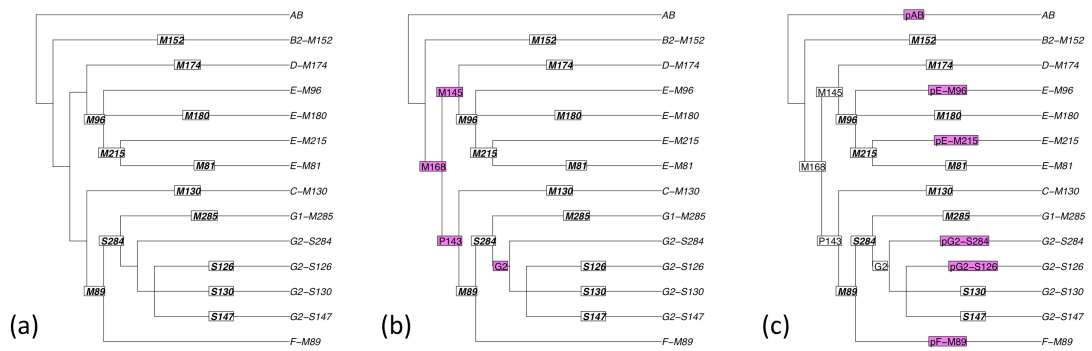
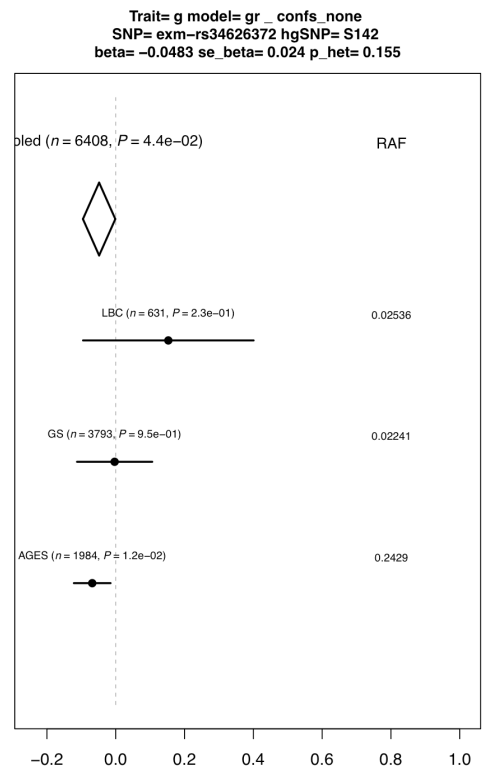
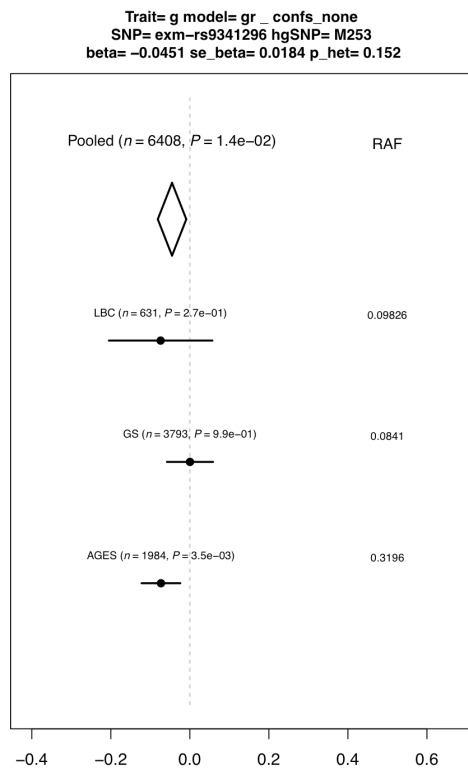
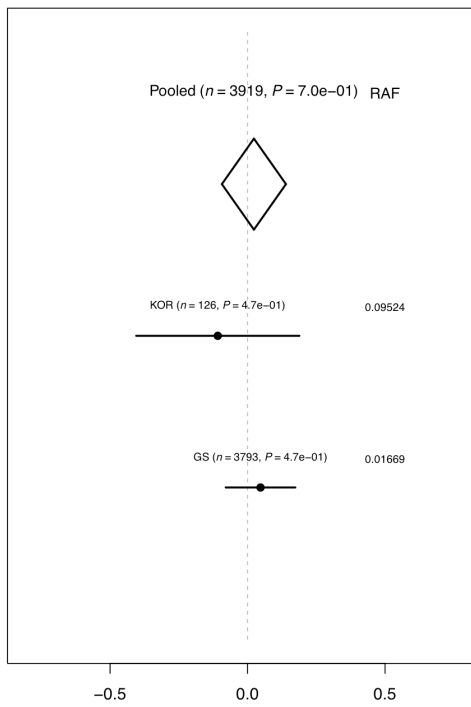


Figure A.1: **Imputation of Y chromosome SNPs.** (a) Genotyped SNPs. (b) Imputed SNPs. (g) *para*-SNPs.

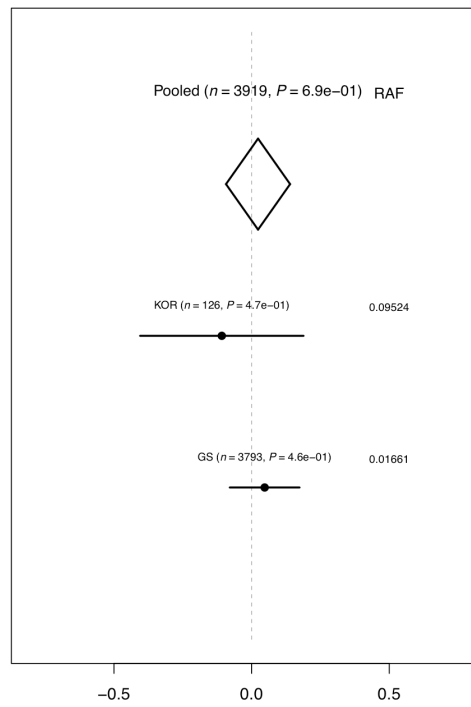
Appendix B



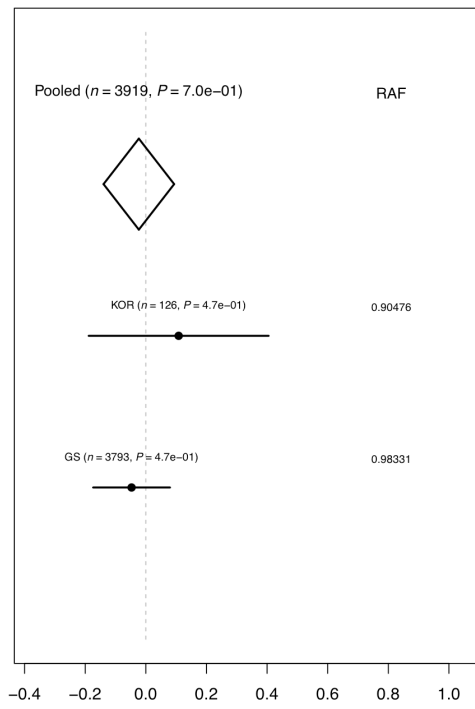
Trait= g model= gr _ confs_none
SNP= rs3848982 hgSNP= M145
beta= 0.0231 se_beta= 0.0594 p_het= 0.345



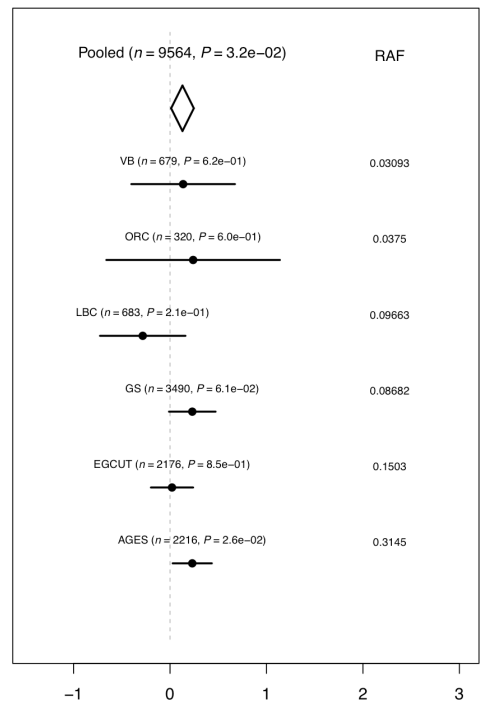
Trait= g model= gr _ confs_none
SNP= exm-rs2032654 hgSNP= M215
beta= 0.0234 se_beta= 0.0594 p_het= 0.344



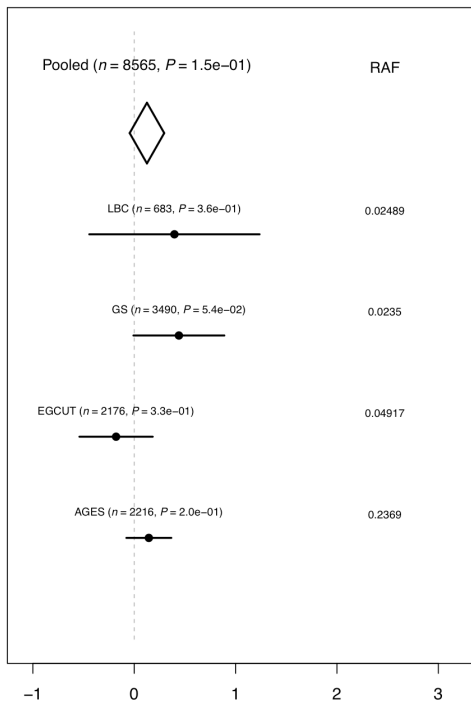
Trait= g model= gr _ confs_none
 SNP= rs4141886 hgSNP= P143
 beta= -0.0231 se_beta= 0.0594 p_het= 0.345



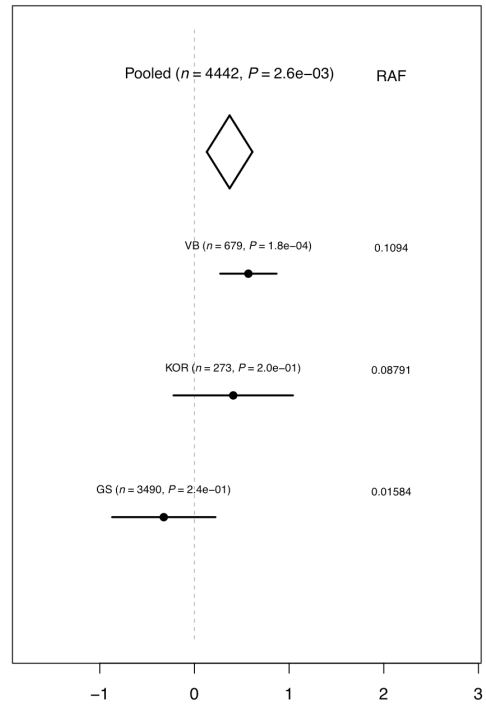
Trait= edu model= gr _ confs_none
 SNP= exm-rs9341296 hgSNP= M253
 beta= 0.129 se_beta= 0.0601 p_het= 0.309



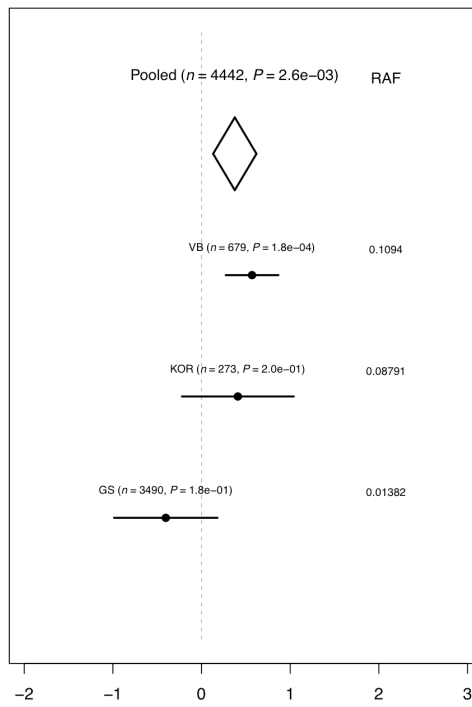
Trait= edu model= gr _ confs_none
 SNP= exm-rs34626372 hgSNP= S142
 beta= 0.125 se_beta= 0.087 p_het= 0.168



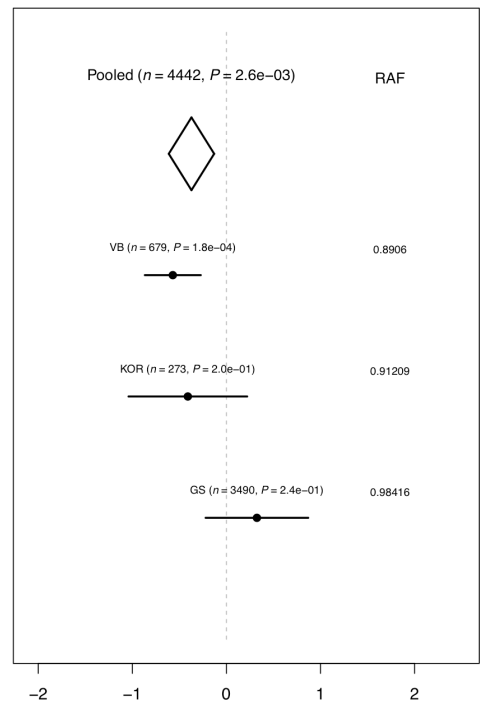
Trait= edu model= gr _ confs_none
 SNP= rs3848982 hgSNP= M145
 beta= 0.371 se_beta= 0.123 p_het= 0.0187



Trait= edu model= gr _ confs _ none
 SNP= pE-M215 hgSNP= pE-M215
 beta= 0.375 se_beta= 0.125 p_het= 0.0143



Trait= edu model= gr _ confs _ none
 SNP= rs4141886 hgSNP= P143
 beta= -0.371 se_beta= 0.123 p_het= 0.0187



Bibliography

- [1] Leonard H. et al. *The epidemiology of mental retardation: challenges and opportunities in the new millennium*. Ment Retard Dev Disabil Res Rev. 2002;8(3):117-34.
- [2] Chelly J et al. *Genetics and pathophysiology of mental retardation*. Eur J Hum Genet. 2006 Jun;14(6):701-13. Review.
- [3] Piton A. et al. *XLID-Causing Mutations and Associated Genes Challenged in Light of Data From Large-Scale Human Exome Sequencing*. Am J Hum Genet. 2013 Aug 8;93(2):368-83. doi: 10.1016/j.ajhg.2013.06.013.
- [4] Ropers H. et al. *X-linked mental retardation: many genes for a complex disorder*. Curr Opin Genet Dev. 2006 Jun;16(3):260-9.
- [5] Vissers LE *Genetic studies in intellectual disability and related disorders*. Nature. 2014 Jul 17;511(7509):344-7. doi: 10.1038/nature13394.
- [6] *Diagnostic Yield of Various Genetic Approaches in Patients With Unexplained Developmental Delay or Mental Retardation*. Am J Med Genet A. 2006 Oct 1;140(19):2063-74.
- [7] Li J et al. *Genes with de novo mutations are shared by four neuropsychiatric disorders discovered from NPdenovo database*. Mol Psychiatry. 2016 Feb;21(2):290-7. doi: 10.1038/mp.2015.40.
- [8] Weckselbatt et al. *Human Structural Variation: Mechanisms of Chromosome Rearrangements*. Trends Genet. 2015 Oct;31(10):587-99. doi: 10.1016/j.tig.2015.05.010.
- [9] Cooper GM et al. *A Copy Number Variation Morbidity Map of Developmental Delay*. Nat Genet. 2011 Aug 14;43(9):838-46. doi: 10.1038/ng.909.
- [10] Wagenstaller J et al. *Copy-Number Variations Measured by Single-Nucleotide Polymorphism Oligonucleotide Arrays in Patients with Mental Retardation*. Am J Hum Genet. 2007 Oct;81(4):768-79.
- [11] Kiss AW et al. *Autosomal recessive mental retardation: homozygosity mapping identifies 27 single linkage intervals, at least 14 novel loci and several mutation hotspots*. Hum Genet. 2011 Feb;129(2):141-8. doi: 10.1007/s00439-010-0907-3.
- [12] Najmabadi H et al. *Deep sequencing reveals 50 novel genes for recessive cognitive disorders*. Nature. 2011 Sep 21;478(7367):57-63. doi: 10.1038/nature10423.

- [13] Ng et al. *Exome sequencing identifies the cause of a mendelian disorder*. Nat Genet. 2010 Jan;42(1):30-5. doi: 10.1038/ng.499.
- [14] Kong W et al. *SCN8A mutations in Chinese children with early onset epilepsy and intellectual disability*. Epilepsia. 2015 Mar;56(3):431-8. doi: 10.1111/epi.12925.
- [15] Kou Y et al. *Network- and attribute-based classifiers can prioritize genes and pathways for autism spectrum disorders and intellectual disability*. Am J Med Genet C Semin Med Genet. 2012 May 15;160C(2):130-42. doi: 10.1002/ajmg.c.31330.
- [16] Vissers LE et al. *A de novo paradigm for mental retardation*. Nat Genet. 2010;42(12):1109-1112. doi:10.1038/ng.712.
- [17] Glissen C et al. *Genome sequencing identifies major causes of severe intellectual disability* Nature. 2014 Jul 17;511(7509):344-7. doi: 10.1038/nature13394.
- [18] Kaufman L et al. *The genetic basis of non-syndromic intellectual disability: a review*. J. Neurodev. Disord. 2 (2010) 182-209, <http://dx.doi.org/10.1007/s11689-010-9055-2>.
- [19] Wang K et al. *PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data*. Genome Res. 2007;17(11):1665-74. doi:10.1101/gr.6861907.
- [20] Robinson JT et al. *Integrative genomics viewer*. Nat Biotechnol. 2011 Jan;29(1):24-6. doi: 10.1038/nbt.1754.
- [21] Adzhubei I et al. *Predicting functional effect of human missense mutations using PolyPhen-2*. Curr. Protoc. Hum. Genet. 2013, <http://dx.doi.org/10.1002/0471142905.hg0720s76>.
- [22] Schwarz JM et al. *MutationTaster evaluates disease-causing potential of sequence alterations*. Nat. Methods 7 2010 575-576, <http://dx.doi.org/10.1038/nmeth0810-575>.
- [23] Ng PC et al. *SIFT: predicting amino acid changes that affect protein function*. Nucleic Acids Res. 2003 Jul 1;31(13):3812-4.
- [24] Krumm N et al. *Copy number variation detection and genotyping from exome sequence data*. Genome Res. 2012 Aug;22(8):1525-32. doi: 10.1101/gr.138115.112.
- [25] Pollard KS et al. *Detection of nonneutral substitution rates on mammalian phylogenies*. Genome Res. 2010 Jan;20(1):110-21. doi: 10.1101/gr.097857.109.
- [26] Hoyer J et al. *Haploinsufficiency of ARID1B, a member of the SWI/SNF-a chromatin-remodeling complex, is a frequent cause of intellectual disability*. Am. J. Hum. Genet. 90 (2012) 565-572, <http://dx.doi.org/10.1016/j.ajhg.2012.02.007>.

- [27] Bhalla K et al. *Alterations in CDH15 and KIRREL3 in patients with mild to severe intellectual disability*. Am. J. Hum. Genet. 83 (2008) 703-713, <http://dx.doi.org/10.1016/j.ajhg.2008.10.020>.
- [28] Wu J et al. *Integrating Multiple Genomic Data to Predict Disease-Causing Nonsynonymous Single Nucleotide Variants in Exome Sequencing Studies*. PLoS Genet. 2014 Mar 20;10(3):e1004237. doi: 10.1371/journal.pgen.1004237.
- [29] Wang K et al. *ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data*. Nucleic Acids Res. 2010 Sep;38(16):e164. doi: 10.1093/nar/gkq603.
- [30] Cooper GM et al., 2005. *Distribution and Intensity of Constraint in Mammalian Genomic Sequence*. Genome Res. 2005 Jul;15(7):901-13.
- [31] Obermann H. et al. *HE6, a two-subunit heptahelical receptor associated with apical membranes of efferent and epididymal duct epithelia*. Mol Reprod Dev. 2003 Jan;64(1):13-26.
- [32] Kircher M et al. *A general framework for estimating the relative pathogenicity of human genetic variants*. Nat Genet. 2014 Mar;46(3):310-5. doi: 10.1038/ng.2892.
- [33] Veeramah KR et al. *Exome sequencing reveals new causal mutations in children with epileptic encephalopathies*. Epilepsia. 2013 Jul;54(7):1270-81. doi: 10.1111/epi.12201.
- [34] Hu H et al. *X-exome sequencing of 405 unresolved families identifies seven novel intellectual disability genes*. Mol Psychiatry. 2016 Jan;21(1):133-48. doi: 10.1038/mp.2014.193. Epub 2015 Feb 3.
- [35] Timal S. et al. *Gene identification in the congenital disorders of glycosylation type I by whole-exome sequencing*. Hum Mol Genet. 2012 Oct 1;21(19):4151-61. doi: 10.1093/hmg/dds123.
- [36] Bissar-Tadmouri N. et al. *X Chromosome Exome Sequencing Reveals a Novel ALG13 Mutation in a Nonsyndromic Intellectual Disability Family With Multiple Affected Male Siblings* Am J Med Genet A. 2014 Jan;164A(1):164-9.
- [37] Sim n-S nchez J et al. *Cooperative genome-wide analysis shows increased homozygosity in early onset Parkinson's disease*. PLoS One. 2012;7(3):e28787. doi:10.1371/journal.pone.0028787.
- [38] Ghani M et al. *Evidence of recessive Alzheimer disease loci in a Caribbean Hispanic data set: genome-wide survey of runs of homozygosity*. JAMA Neurol. 2013;70(10):1261-7. doi:10.1001/jamaneurol.2013.3545.
- [39] Keller MC et al. *Runs of homozygosity implicate autozygosity as a schizophrenia risk factor*. PLoS Genet. 2012;8(4):e1002656. doi:10.1371/journal.pgen.1002656.
- [40] Lin PI et al. *Runs of homozygosity associated with speech delay in autism in a taiwanese han population: evidence for the recessive model*. PLoS One. 2013;8(8):e72056. doi:10.1371/journal.pone.0072056.

- [41] Gamsiz ED et al. *Intellectual disability is associated with increased runs of homozygosity in simplex autism*. Am J Hum Genet. 2013;93(1):103-109. doi:10.1016/j.ajhg.2013.06.004.
- [42] Abou JR et al. *Homozygosity mapping in 64 Syrian consanguineous families with non-specific intellectual disability reveals 11 novel loci and high heterogeneity*. Eur J Hum Genet. 2011;19(11):1161-6. doi:10.1038/ejhg.2011.98.
- [43] Caliskan Met al. *Exome sequencing reveals a novel mutation for autosomal recessive non-syndromic mental retardation in the TECR gene on chromosome 19p13*. Hum Mol Genet. 2011;20(7):1285-9. doi:10.1093/hmg/ddq569.
- [44] Najmabadi H et al. *Homozygosity mapping in consanguineous families reveals extreme heterogeneity of non-syndromic autosomal recessive mental retardation and identifies 8 novel gene loci*. Hum Genet. 2007 Mar;121(1):43-8. Epub 2006 Nov 21.
- [45] Purcell S et al. *PLINK: a tool set for whole-genome association and population-based linkage analyses*. Am J Hum Genet. 2007;81(3):559-75. doi:10.1086/519795.
- [46] Szpiech ZA et al. *Long runs of homozygosity are enriched for deleterious variation*. Am J Hum Genet. 2013;93(1):90-102. doi:10.1016/j.ajhg.2013.05.003.
- [47] Gandin I. et al. *Y chromosome variation and complex traits: the Ygen consortium*. (Abstract) Presented at the 65th Annual Meeting of The American Society of Human Genetics, October 7, 2015 in Baltimore, MD.
- [48] Baron-Cohen S et al. *Why Are Autism Spectrum Conditions More Prevalent in Males?* PLoS Biol. 2011 Jun;9(6):e1001081. doi:10.1371/journal.pbio.1001081.
- [49] Charchar F. et al. *Inheritance of coronary artery disease in men: an analysis of the role of the Y chromosome*. Lancet. 2012 Mar 10;379(9819):915-22. doi:10.1016/S0140-6736(11)61453-0.
- [50] Davies G et al. *Genetic contributions to variation in general cognitive function: a meta-analysis of genome-wide association studies in the CHARGE consortium (N = 53 949)*. Mol Psychiatry. 2015 Feb;20(2):183-92. doi: 10.1038/mp.2014.188.
- [51] Ross J. et al. *Behavioral phenotypes in males with XYY and possible role of increased NLGN4Y expression in autism features*. Genes Brain Behav. 2015 Feb;14(2):137-44. doi: 10.1111/gbb.12200.
- [52] Like H et al. *Biological factors underlying sex differences in neurological disorders*. Int J Biochem Cell Biol. 2015 Aug;65:139-50. doi: 10.1016/j.biocel.2015.05.024.
- [53] Rietveld C et al. *GWAS of 126,559 individuals identifies genetic variants associated with educational attainment*. Science. 2013 Jun 21;340(6139):1467-71. doi: 10.1126/science.1235488.

- [54] Pliss L et al. *Y-Chromosomal Lineages of Latvians in the Context of the Genetic Variation of the Eastern-Baltic Region*. Ann Hum Genet. 2015 Nov;79(6):418-30. doi: 10.1111/ahg.12130.
- [55] Barack L et al. *Y chromosomal heritage of Croatian population and its island isolates*. Eur J Hum Genet. 2003 Jul;11(7):535-42.