Alma Mater Studiorum
Università di Bologna

DOTTORATO DI RICERCA IN

BIOLOGIA CELLULARE E MOLECOLARE

CICLO XXVI

Settore Concorsuale di afferenza: 05/I1
Settore Scientifico disciplinare: Bio/18

Investigating the role of Copy Number Variants in Specific Language Impairment and identification of new candidate genes

*PRESENTATA DA: **CERONI FABIOLA***

Coordinatore del corso di Dottorato	Relatore
Vincenzo Scarlato	**Elena Maestrini**
	Correlatore
	Elena Bacchelli

ESAME FINALE ANNO 2014

Table of contents

Abstract **4**

Chapter 1 Copy Number Variants.. **7**

1.1 Copy Number Variants.. 7

1.2 Mutational Mechanisms.. 7

1.2.1 Non-Allelic Homologous Recombination (NAHR)... 8

1.2.2 Non-Homologous End Joining (NHEJ) and Microhomologous-Mediated End Joining (MMEJ).. 11

1.2.3 Fork Stalling and Template Switching and MMBIR... 11

1.2.4 Retrotransposition... 13

1.3 CNV Detection Methods.. 15

1.3.1 Array-CGH.. 15

1.3.2 SNP arrays.. 16

1.3.3 Next Generation Sequencing-based approaches.. 19

1.3.4 Validation of structural variants.. 21

1.4 CNVs in Neuropsychiatric Disorders.. 22

1.5 Recurrent CNVs on chromosome 15q11-q13.. 26

1.5.1 Recurrent CNVs at 15q11.2 locus (BP1-BP2 region). 27

1.5.2 Recurrent CNVs at 15q13.2-q13.3 locus (BP4-BP5 region). 29

Chapter 2 Specific Language Impairment.. **32**

2.1 What is Specific Language Impairment?.. 32

2.2 Measurement of SLI symptoms.. 32

2.3 The genetic bases of SLI.. 34

2.4 Molecular genetic studies for SLI susceptibility.. 35

2.4.1 Linkage studies.. 35

2.4.2 Association studies.. 36

2.4.3 Nuclear families studies: evidence for linkage to chromosome 16 and 19.......... 37

2.4.4 Extended pedigree studies: evidence for linkage to chromosome 13.................. 39

2.4.5 Extended pedigrees from an isolated population: Robinson Crusoe island......... 41

2.4.6 A comparison between the different linkage studies for SLI.............................. 42

2.4.7 Targeted association studies of chromosome 16.. 44

2.5 SLI candidate genes *ATP2C2* and *CMIP*. 46

2.6 Co-morbidity of SLI with Autism Spectrum Disorders (ASD)............................... 47

2.6.1 Autism Spectrum Disorders... 47

2.6.2 Phenotypic overlap between SLI and ASD.. 48

2.6.3 Genetic overlaps between SLI and ASD... 49

2.7 The gene *CNTNAP2:* an example of functional link between neurodevelopmental disorders.. 52

Chapter 3 ***DPYD*: a candidate gene for neurodevelopmental disorders**…………… **56**

3.1 The enzyme dihydropyrimidine dehydrogenase (DPD)…………………………… 56

3.2 The dihydropyrimidine dehydrogenase deficiency.. 57

3.3 DPD: a key player in 5-FU metabolism... 58

3.4 Hypotheses for a role of *DPYD* in central nervous system...................................... 59

3.5 The structural organization of *DPYD* and genetic variation.................................... 61

3.6 The DPD protein... 63
3.7 Involvement of *DPYD* in neurodevelopmental disorders.. 65
Chapter 4 Materials and Methods.. 70
4.1 Samples: SLIC cohort.. 70
4.1.1 Subjects.. 70
4.1.2 SLIC cohort: phenotypic tests.. 71
4.2 Samples: ASD cohorts... 72
4.2.1 IMGSAC cohort: subjects.. 72
4.2.2 Italian cohort: subjects... 74
4.2.3 AGP cohort: subjects.. 74
4.2.4 ASD diagnostic instruments: ADI-R and ADOS-G... 75
4.3 Identification of CNVs in the SLIC cohort.. 75
4.3.1 Illumina Human OmniExpress array.. 75
4.3.2 QuantiSNP.. 78
4.3.3 PennCNV.. 80
4.3.4 Generation of a CNV consensus list: BEDTool.. 82
4.3.5 Overlap of the CNV consensus list with DGV.. 82
4.3.6 PLINK: CNV burden analyses... 83
4.4 Validation of CNVs by Real-Time PCR... 83
4.5 Analysis of the candidate gene *ZNF277*.. 85
4.5.1 Validation of the *ZNF277* microdeletion in the discovery pedigree G4.............. 85
4.5.2 Breakpoint characterization of the *ZNF277* microdeletion in the discovery pedigree.. 85
4.5.3 IBD and homozygosity analysis in the discovery pedigree G4........................... 87
4.5.4 Screening of Larger Cohorts: SLIC, IMGSAC and controls............................... 87
4.5.5 Gene Expression Evaluation for *ZNF277*, *DOCK4* and *IMMP2L*....................... 88
4.6 Analysis of the candidate gene *DPYD*.. 90
4.6.1 SLI probands... 90
4.6.2 Autism Cohorts: Italian, IMGSAC and AGP families... 90
4.6.3 Genotyping by endonuclease restriction analysis.. 90
4.6.4 Mutation screening of *DPYD* gene: primers and PCR conditions....................... 91
4.6.5 PCR purification... 92
4.6.6 Sequencing Reaction.. 93
4.6.7 Ethanol Precipitation of Sequencing Reactions.. 94
4.6.8 Sequence Analyses and Prediction tools... 94
4.6.9 Association Analyses with PLINK.. 95
4.6.10 Analysis of language items of ADI-R and IQ scores in the AGP cohort............. 95
Chapter 5 Results: Screening for Copy Number Variants in a SLI cohort............... 101
5.1 Detection of CNVs.. 101
5.2 CNV burden analyses... 104
5.3 Detection of rare and novel events... 105
5.4 Detection of *de novo* CNVs.. 108
5.5 CNVs in candidate genes.. 111
5.6 CNVs on chromosome 15q11-q13.. 111
5.7 Ongoing analyses.. 119

Table of contents

Chapter 6 **Results: Validation and follow-up of a homozygous exonic deletion in *ZNF277*.. 120**

6.1 Identification and validation of a homozygous microdeletion of exon 5 in *ZNF277*. 120

6.2 Homozygosity analysis in family G4……………………………………………….. 125

6.3 Screening for *ZNF277* microdeletions in the SLIC and IMGSAC cohorts……….. 126

6.4 Expression analyses of *ZNF277*, *DOCK4* and *IMMP2L*…………………………… 129

Chapter 7 **Results: Analysis of the candidate gene *DPYD*... 133**

7.1 Analysis of the candidate gene *DPYD* in SLI cases.. 133

7.1.1 *DPYD* in family G4... 133

7.1.2 Screening for the splicing variant rs3918290 in SLI probands............................ 135

7.1.3 Mutation screening of *DPYD* in SLI cases carrying the variant rs3918290......... 136

7.1.4 *DPYD* variants from exome sequences.. 141

7.2 Analysis of the candidate gene *DPYD* in ASD cases... 143

7.2.1 Screening for the splicing variant rs3918290 in ASD cases and controls............ 143

7.2.2 Statistical analysis of the rs3918290 frequency in the Italian and IMGSAC cohorts.. 145

7.2.3 Frequency and association analysis of rs3918290 in the AGP ASD cohort......... 146

7.2.4 TDT analysis for *DPYD* and *MIR137* in the AGP cohort……………………… 147

7.2.5 Mutation screening of *DPYD* in ASD probands carrying the variant rs3918290. 149

7.2.6 IQ analysis of ASD individuals carrying the variant rs3918290.......................... 151

7.2.7 Analysis of language phenotypes in ASD individuals carrying the variant rs3918290.. 152

Chapter 8 **Discussion... 155**

8.1 CNV screening in the SLIC cohort.. 155

8.2 CNVs on chromosome 15q11-q13... 158

8.3 Identification of a homozygous exonic deletion in *ZNF277*..................................... 160

8.4 Analysis of the candidate gene *DPYD* in SLI.. 163

8.5 Analysis of the candidate gene *DPYD* in ASD... 169

References **.. 174**

Abstract

Specific language impairment (SLI) is a neurodevelopmental disorder defined as an unexpected failure to develop normal language abilities despite normal intelligence and adequate educational opportunities. This condition has an estimated frequency of 5-8% in English-preschool children and it has been shown to be highly heritable. Like many behavioural traits, SLI is assumed to be a heterogeneous multifactorial disorder with a complex genetic basis. Over the years, linkage and association studies have attempted to uncover the genetic bases underlying language impairment and have identified some candidate genes possibly contributing to SLI susceptibility, such as *ATP2C2*, *CMIP* and *CNTNAP2*. However, a large portion of the genetic risk remains to be unravelled.

Since CNVs have been demonstrated to be an important source of variation in the susceptibility to other neuropsychiatric disorders, such as autism spectrum disorders (ASD) and schizophrenia, we carried out a genome-wide CNV screen on 540 individuals of the SLIC cohort in order to investigate the role of structural variants in the genetic architecture of SLI. An exploratory analysis did not find significant differences in CNV burden between 174 affected children and 40 unaffected siblings of this SLI cohort. However, these are preliminary results and further investigations are currently being performed using a larger group of independent controls.

Individual CNVs might be of interest and might reveal new candidate genes implicated in the disease susceptibility. Among the CNVs affecting genes, we focused on two particular examples of potentially contributing variants: recurrent CNVs on chromosome 15q11-q13 and a homozygous exonic microdeletion in *ZNF277*.

The locus 15q11-q13 has been implicated in a range of distinct but co-morbid neurological conditions, such as ASD, schizophrenia, epilepsy and language delay. In the SLIC cohort, we identified BP1-BP2 microdeletions at 15q11.2 (involving the genes *TUBGCP5*, *CYFIP1*, *NIPA1* and *NIPA2*) in two families, and BP4-BP5 microduplications at 15q13.2-q13.3, including the gene *CHRNA7*, in three families. Both CNVs showed an incomplete segregation with the phenotype, supporting the hypothesis that these variants are unlikely to be causative for neuropathological conditions, showing variable expressivity and incomplete penetrance. However, since these CNVs are recurrently observed in neuropsychiatric conditions, it has been proposed that they might act as modifiers and that their outcomes could depend on the interaction with other genetic and non-genetic factors. Therefore, it is plausible that microduplications at 15q11-q13 might play a role also in SLI.

We identified a microdeletion of 21,379 bp in the *ZNF277* gene, encompassing exon 5, in a single child with severe expressive and receptive language impairment. The microdeletion was inherited

from both parents, each of whom carries a heterozygous microdeletion and has a history of language problems, and was not found in the proband's affected sister or her brother who had mild language impairment. The microdeletion falls within the *AUTS1* locus, a region linked to ASD. Moreover, *ZNF277* is adjacent to the *DOCK4* and *IMMP2L* genes, which have been implicated in ASD. We screened for the presence of *ZNF277* microdeletions in cohorts of children with SLI or ASD and panels of control subjects. *ZNF277* microdeletions were at an increased allelic frequency in SLI probands (1.1%) compared to both ASD family members (0.3%) and independent controls (0.4%). We performed quantitative RT-PCR analyses of the expression of *IMMP2L*, *DOCK4* and *ZNF277* in individuals carrying either a *IMMP2L-DOCK4* microdeletion or a *ZNF277* microdeletion. While *ZNF277* microdeletions reduce the expression of *ZNF277*, they do not alter the levels of *DOCK4* or *IMMP2L* transcripts. Conversely, *IMMP2L-DOCK4* microdeletions do not affect the expression levels of *ZNF277*. We postulate that *ZNF277* microdeletions may contribute to the risk of language impairments in a complex manner, that is independent of the autism risk loci previously described in this region.

The SLI family in which the homozygous *ZNF277* microdeletion was discovered, represents an interesting example of genetic heterogeneity of the disorder within pedigrees. The affected sister, who did not carry the *ZNF277* microdeletion, had a complete dihydropyrimidine dehydrogenase (DPD) deficiency, absent in the proband with the homozygous *ZNF277* microdeletion. DPD, encoded by the gene *DPYD*, is the first rate-limiting enzyme of the pyrimidine catabolism pathway. The complete loss of the DPD activity in that child was caused by a compound heterozygosity of two deleterious variants in the gene *DPYD*: the splicing variant rs3918290, paternally inherited, and the missense change p.S201R, maternally inherited. The DPD deficiency leads to the accumulation of uracil and thymine and, like other disorders affecting the nucleotides metabolism, its clinical manifestations can include neurological problems, such as ASD features and language impairments. Therefore, we hypothesized that in this family, two risk factors (the complete loss of ZNF277 in the proband and the complete loss of DPD activity in the affected sister) could have independently contributed to the language deficits of the two affected children.

Moreover, since *DPYD* represents a good candidate for both SLI and ASD, we investigated its involvement in the susceptibility to these two neurodevelopmental disorders. Many mutations have been reported in *DPYD*, but the splicing variant rs3918290, that causes the skipping of exon 14 in the mRNA leading to a non-functional protein, is the most commonly implicated mutation in the DPD deficiency. Therefore, we decided to analyse the frequency of the splicing variant rs3918290 in a group of 166 SLI cases and three groups of ASD cases (231 Italian probands, 224 IMGSAC probands and 2681 AGP probands). We observed a higher frequency of rs3918290 in the SLI group

(1.2%) and the Italian probands (1.08%), while no difference was observed for the IMGSAC group (0.45%) and the AGP cohort (0.47%), compared to controls (~0.5% -0.6%). In the AGP collection, an association analysis of the region encompassing *DPYD* did not yield significant evidence of association. In addition, from the analysis of IQ performances and verbal abilities in the AGP cases carrying the splicing variant, we did not observe an influence of the variant on the severity of the phenotype. However, since we focused on a rare variant, it should be noted that these phenotypic analyses were based on a small number of individuals carrying rs3918290.

The mutation screening of the gene *DPYD* performed on the individuals with the splicing variant (4 SLI families including the discovery pedigree, 5 Italian families and 2 IMGSAC families) led to the identification of six known missense changes (p.C29R, p.M166V, p.S201R, p.S534N, p.I543V, p.V732I) and a novel non-coding variant in the promoter region. Several individuals were found to be compound heterozygotes for *DPYD* variants. Although the consequences of these changes are widely debated in literature, as contrasting results are frequently reported, we hypothesized that the combined effect of the mutations identified in affected individuals leads to an altered DPD activity. These hypotheses should be confirmed by enzymatic assay, however this small group of SLI and ASD cases suggests that the partial or complete loss of DPD activity might be a risk factor for language impairment or autism spectrum disorders.

In summary, these results do not support a major role of the *DPYD* gene in SLI and ASD, but suggest that rare variants in *DPYD* might contribute to a minority of cases in a complex manner, in addition to other genetic or non-genetic factors. Further investigations will be required to clarify the role of *DPYD* in SLI and ASD.

Chapter 1

Copy Number Variants

1.1 Copy Number Variants.

Analysis of the human genome has revealed an extended sequence similarity among individuals: any two humans are estimated to be approximately 99.9% identical in their DNA sequence.

The most numerous variants in the genome are individual base changes called Single-Nucleotide Polymorphisms (SNPs). Another important source of genomic variation is represented by submicroscopic variants in DNA copy number (Copy Number Variants, CNVs), which usually involve many bases. Deletions, duplications, insertions and translocations can all results in CNVs.

Two landmark studies carried out in 2004 (Iafrate et al., 2004; Sebat et al., 2004) showed that CNVs, ranging in size from kilobases (kb) to Megabases (Mb), are widespread in normal human genomes. More recently, two studies (Conrad et al., 2010; Mills et al., 2011) have estimated that CNVs account for ~4 million base pairs of genomic difference (~13%). Therefore, although the number of SNPs in the genome exceed the number of CNVs, their relative contribution to genetic heterogeneity is similar if we consider the variation in terms of nucleotides implicated.

Copy number changes of genomic segments can be tolerated, advantageous or deleterious. Certain advantageous CNVs have played a significant role in primate evolution (Bailey and Eichler, 2006; Dumas et al., 2007). CNVs can include a variable number of genes and their phenotypic effects usually depend on whether dosage-sensitive genes or regulatory regions are affected (Lupski and Stankiewicz, 2005). Disruption of regulatory regions, promoter elements or coding sequences can lead to altered gene function. Genes can be inactivated by a premature truncation or an internal deletion. CNVs can also result in fusion or abnormal gene products with a new function (Holt et al., 2012). Moreover, in some cases, deletions can have the effect of “unmasking” a recessive pathogenic allele in the heterozygous state.

1.2 Mutational mechanisms.

CNVs can be inherited or *de novo*. These structural variants can arise in both meiotic and somatic cells: in the first case, they represent constitutional genomic rearrangements, in the second case there will be mosaic populations of somatic cells carrying the CNV.

In the next paragraphs, the four major mechanisms that can lead to the formation of CNVs will be described. Speculations on the causative molecular mechanism are usually based on detailed junction analyses at base level, as the “molecular fingerprint” of the breakpoints may help to understand how the structural change has arisen.

1.2.1 Non-Allelic Homologous Recombination (NAHR).

Low copy number repeats (LCRs) or segmental duplications are DNA fragments >1 kb in size and of more than 90% of DNA sequence identity. LCRs can cause genomic instability and either mediate or stimulate CNV formation: genomic regions harbouring tandemly arranged LCRs are more prone to recurrent or non-recurrent rearrangements.

Due to their high degree of sequence identity, a misalignment between non-allelic LCRs can occur and mediate non-allelic homologous recombination (NAHR), which results in unequal crossing-over and a change in copy number (Stankiewicz and Lupski, 2002).

NAHR has been shown to require fragments of minimal length (300-500 bp) which share extremely high similarity or identity between LCRs, called *minimal efficient processing segments* (MEPS) (Rubnitz and Subramani, 1984; Waldman and Liskay, 1988). The majority of NAHR events occur between LCRs which have a sequence identity greater than ~97%, a size that ranges from about 10 to ~400 kb and are located at a distance less than ~10 Mb from each other.

The outcomes of the NAHR between LCRs depend on their orientation and their location (**Figure 1.1**). Misalignment can occur between LCRs located on homologous chromosomes (*interchromosomal*), on sister chromatids (*intrachromosomal*) or on the same chromatid (*intrachromatid*).

Interchromosomal and intrachromosomal NAHR between directly oriented LCRs results in two reciprocal products: a deletion and a reciprocal duplication of the genomic segment between them (**Figure 1.1 a and 1.1 d**). Intrachromatid NAHR between directly oriented LCRs can result in a deletion and a loop excision (**Figure 1.1 g**).

Interchromosomal and intrachromatid NAHR between LCRs in opposite orientation causes an inversion of the genomic segment flanked by them (**Figure 1.1 b and 1.1 h**). Intrachromosomal NAHR between LCRs in opposite orientation generates duplicated modules in an inverse orientation (**Figure 1.1 e**).

If the LCRs have a complex structure consisting of both direct and inverted subunits, they can predispose the region to NAHR, leading to both deletions/duplications and inversions (**Figure 1.1 c, 1.1 f, 1.1 i**).

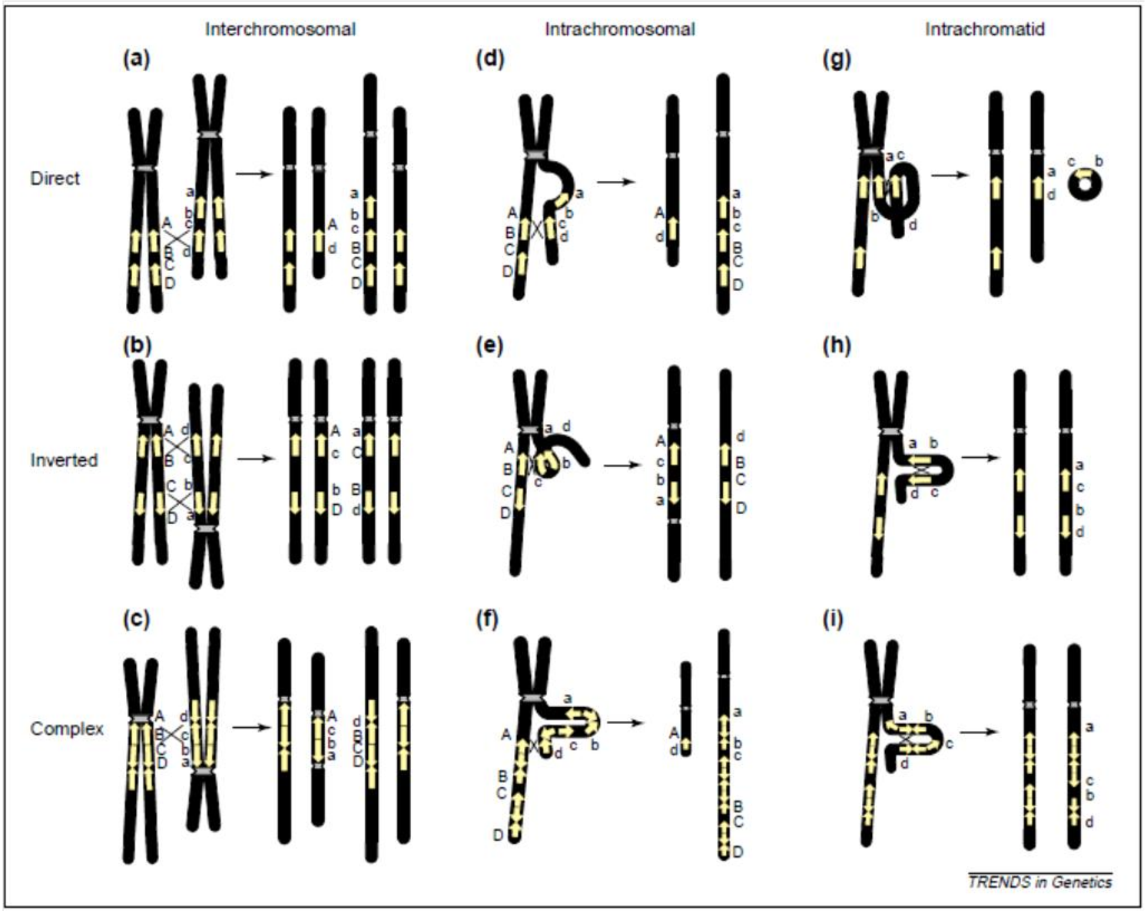


Figure 1.1 (Stankiewicz and Lupski, 2002). Schematic representation of LCRs-NAHR based mechanisms for genomic rearrangements. LCRs are depicted as yellow arrows and their orientation is indicated. The chromosome rearrangements and predicted products of recombination are listed vertically by mechanisms (interchromosomal; intrachromosomal; and intrachromatid). Interchromosomal misalignment leads to deletion/duplication (directly oriented LCRs) (a) and inversion (inverted repeats) (b). Intrachromatid loop of inverted repeats results in inversion (h). Interchromatid mispairing of direct repeats results in deletion/duplication (d). Intrachromatid misalignment of directed repeats (g) can result in deletion and an acentric fragment. Inv dup(15) and inv dup(22) chromosomes can result from interchromosomal (c) or intrachromosomal (e) unequal exchange between inverted LCRs. Also complex LCRs can be responsible for deletion/duplication (f) or inversion (i).

The distribution of the endpoints of CNVs arisen by NAHR along the LCRs seems to be not random: they have been observed to cluster in narrow “hotspots” (Lupski, 2004), where there is sufficient homology for homologous recombination.

The molecular mechanism of NAHR has been shown to account for the vast majority of the “recurrent” rearrangements, defined as those that recur in multiple individuals, share a common size and present breakpoints clustering within the same regions (**Figure 1.2 a**).

NAHR can also be mediated by highly homologous repetitive sequences, such as *Alu* sequences, L1 elements (that will be described in the paragraph 1.2.4), minisatellites and subtelomeric repeated sequences. These NAHR events explain some of the “non-recurrent” rearrangements, defined as those CNVs with different sizes and distinct breakpoints in each event (**Figure 1.2 b**). Interestingly, some non-recurrent rearrangements can map to the same genomic location, but with endpoints in many different positions: these CNVs might share a common region, indicated as the smallest region of overlap (SRO), whose change in copy number may be responsible for the common clinical features among different individuals carrying these rearrangements (Gu et al., 2008).

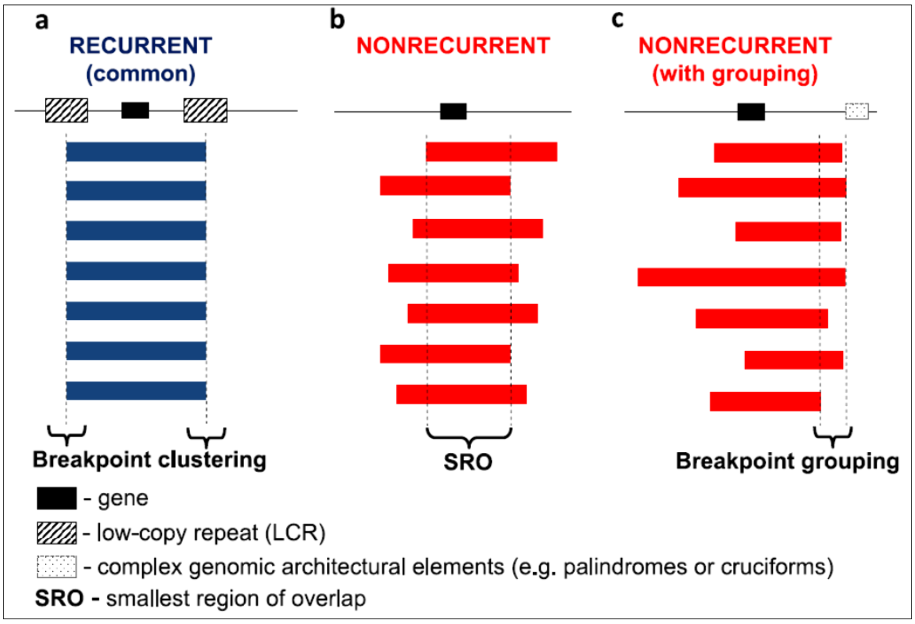

a
RECURRENT
(common)
b
NONRECURRENT
c
NONRECURRENT
(with grouping)
Breakpoint clustering
SRO
Breakpoint grouping
■ - gene
▨ - low-copy repeat (LCR)
⬚ - complex genomic architectural elements (e.g. palindromes or cruciforms)
SRO - smallest region of overlap

Figure 1.2 (from Gu et al., 2008). Recurrent and non-recurrent rearrangements. The black line indicates the genomic region hit by CNVs, blue and red bars indicate the rearrangements observed in different individuals. **a**. Recurrent CNVs have both breakpoints mapping within the same LCRs (hatching rectangles). **b**. The non-recurrent CNVs have different length and different breakpoints, but might share a common region of overlap (SRO). In this example, the SRO encompasses one gene, indicated by the black rectangle. **c**. Some non-recurrent rearrangements show one of their breakpoints within a limited genomic region: the grouping of one breakpoint may occur in proximity of an architectural element important to the rearrangements mechanism.

NAHR can occur in both meiotic and mitotic cells and certain LCRs might be involved in both types of NAHR events. Moreover, the usage of certain LCRs for mitotic NAHR events might be different among different tissues. This mechanism might play a significant role in tumorigenesis (Darai-Ramqvist et al., 2008; Fridlyand et al., 2006). Age seems to be an important risk factor for somatic structural abnormalities. Generally the rate of mutations, including CNVs, increases with age (Jacobs et al., 2012) and this accumulation, that varies with tissue type (Kennedy et al., 2012), could be due to an increased burden of somatic mutation and/or a reduced capacity of genomic maintenance.

Furthermore, post-zygotic events can explain the finding of different genomic rearrangements between monozygotic twins (Bruder et al., 2008; Forsberg et al., 2012).

1.2.2 Non-Homologous End Joining (NHEJ) and Microhomologous-Mediated End Joining (MMEJ).

Another mechanism that can lead to CNV formation is non-homologous end joining (NHEJ), one of the main pathways for repairing double-stranded DNA breaks (DSBs). When double-strand breaks are detected, NHEJ ensures that both broken DNA ends are bridged, modified, and finally ligated (Weterings and van Gent, 2004) (**Figure 1.3**). NHEJ is a homology-independent process efficient at restoring structural DNA integrity, however it can be imprecise at local sequence level and tolerates nucleotide loss or addition at the rejoining site, often leaving a “molecular scar” (Lieber, 2008).

In contrast to NAHR, NHEJ does not require LCRs, MEPS or sequence homology to mediate the recombination. However, NHEJ can frequently occur within repetitive elements, such as long terminal repeats (LTRs), short interspersed repeat elements and mammalian interspersed repeats, suggesting that it may also be stimulated and regulated by certain genome architecture.

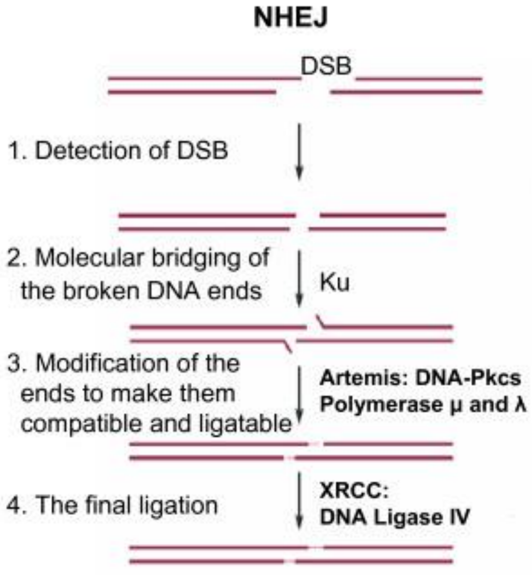

NHEJ
DSB
1. Detection of DSB
2. Molecular bridging of the broken DNA ends
Ku
3. Modification of the ends to make them compatible and ligatable
Artemis: DNA-Pkcs
Polymerase μ and λ
4. The final ligation
XRCC:
DNA Ligase IV

Figure 1.3 (from Gu et al., 2008). The NHEJ in vertebrates proceed in four steps. The two thick lines represents the double stranded DNA. When double strand breaks occur, they are detected (1) and the molecular machinery of NHEJ (Lieber et al., 2003) mediates the molecular bridging of the broken DNA ends (2). The DNA ends are modified to be compatible (3) and, finally, they are ligated (4) to restore the structural integrity.

An alternative end-joining mechanism is represented by microhomology-mediated end joining repair (MMEJ). MMEJ requires short microhomologies (5-25 bp) to anneal the ends of DSBs and leads to the deletion of the region between annealed microhomologies.

1.2.3 Fork Stalling and Template Switching and MMBIR.

Fork Stalling and Template Switching (FoSteS) is a replication-based genomic rearrangement mechanism induced by errors during DNA replication process (Lee et al., 2007). When a replication fork stalls at one position, the 3’ primer end of a DNA strand can disengage from the original template and invade another replication fork nearby (**Figure 1.4**). The invasion and annealing to the new template strand are mediated by a microhomology sequence at the 3’ end (4-15 bp) of the invading strand, which subsequently primes the DNA synthesis. Initially, the replication is characterized by low processivity, but after multiple rounds of disengaging, invasion and extension, it becomes more processive.

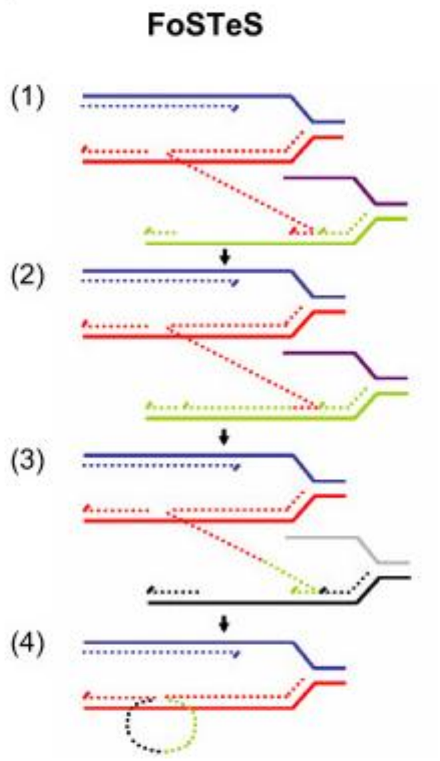


Figure 1.4 (from Gu et al., 2008). A schematic representation of the Fork Stalling and Template Switching (FoSteS) mechanism. Solid lines indicate the template DNA, dotted lines instead the newly synthesised strands. The original replication fork is represented by the red and blue lines. After the stalling, the lagging strand (red, dotted line) invades a second fork (indicated in purple and green) via microhomology (1). The 3’end of the lagging strand allows the DNA extension in the second fork (green dotted line) (2). Serial replication fork disengaging and invasion could occur several times (3) before resumption of replication on the original template (4).

Hastings and colleagues (Hastings et al., 2009) proposed a generalization of the FoSteS mechanism, known as the microhomology-mediated break-induced replication model (MMBIR), based on the mechanism of repair single double-stranded ends. The MMBIR model postulates that the 3’end of a collapsed fork could anneal to any single-stranded DNA stretch, available in physical proximity, with which it shares microhomology, such as a lagging strand of a replication fork or ssDNA exposed at excision repair tracts, at sites of transcription and at secondary DNA structures, such as cruciforms or hairpins loops caused by complex genomic architecture.

The chromosomal structural consequences of MMBIR could be summarized in this way (Hastings et al., 2009):

- template switch to sister or homologous chromosome behind the position where the fork collapsed (backward invasion), with respect to the direction of movement of the fork, generates a duplication; in the opposite case, a template switch to sister or homologous chromosome ahead the position where the fork collapsed (forward invasion) generates a deletion;
- template switch to nonhomologous sequence in another chromosome causes a translocation;
- template switch to a sequence already duplicated causes a triplication;
- template switch to the same molecule behind the position of the fork collapse could initiate a rolling-circle replication and amplification.
- whether the switch occurs in direct or opposite orientation determines if the erroneously incorporated fragment will be in direct or inverted orientation with respect to its original position.

FoSteS/MMBIR frequently generates complex non-recurrent rearrangements, e.g., deleted/duplicated regions are interrupted by normal copy regions or triplicated segments. The complexity at the joining sites is a feature of this mechanism and can help to discriminate among the different potentially causing mechanisms (NHEJ, MMEJ and MMBIR) when very small regions of microhomology are present at the boundaries: the junctions of MMBIR events are characterized by the presence of segments of DNA of variable length derived from elsewhere.

1.2.4 Retrotransposition.

Retrotransposons are mainly represented by endogenous retroviruses, Long interspersed nuclear elements 1 (LINE1 or L1) and Short interspersed nuclear elements (SINEs). L1s are the only currently active class of autonomous retrotransposons in humans.

Although ~500,000 copies can be found in the human genome (occupying ~18% of the whole genome), only 80–100 are predicted to be active full-length elements (6 Kb) and are able to transpose their own sequences or non-autonomous elements (e.g. the *Alu* sequences, the predominant SINE elements in the genome) to new genomic locations by a target primed reverse transcription (TPRT) mechanism (Goodier and Kazazian, 2008). Consistent with this model, at the L1 insertion site short “target site duplications” (TSDs), but occasionally deletions, can be generated. The insertion of retrotransposons can have variable consequences on the expression of the genes nearby, such as causing the premature termination of transcription (**Figure 1.5.6**), producing new transcription start sites (**Figure 1.5.7**) or determining the formation of new transcription modules (**Figure 1.5.8**). Sometimes, these insertions can also alter the chromatin state of the target region: structural chromatin changes mediated by methylation can initiate within transposable elements and spread to the proximal regions, causing the repression of the adjacent genes (**Figure 1.5.9**).

In addition to retrotransposition, recombination between retrotransposons can occur and lead to deletions, duplications (**Figure 1.5.5**) and rearrangements of gene sequence and this is particularly true for *Alu* elements.

Both germline and somatic L1 activity contribute significantly to structural variation in human genomes (Lupski, 2010).

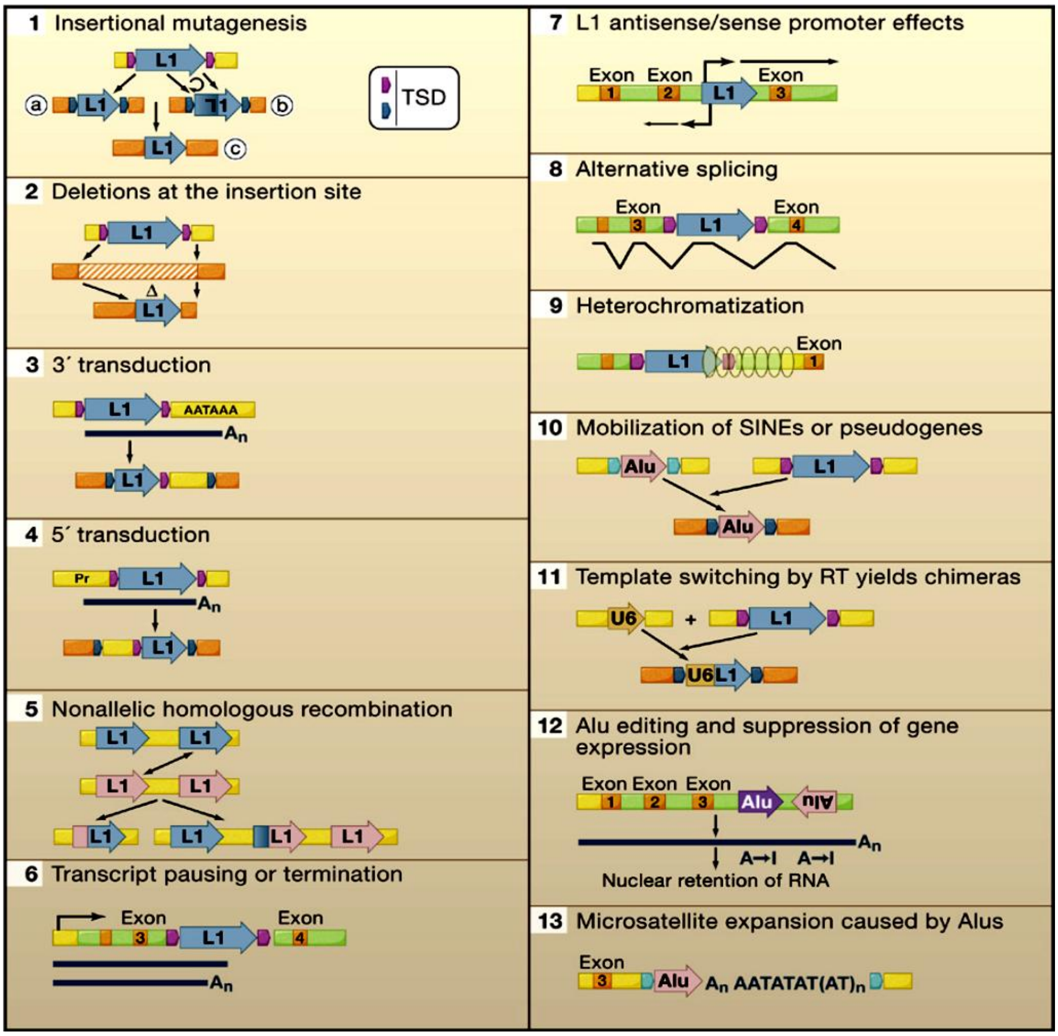

1 Insertional mutagenesis
L1
a
L1
L1
b
L1
c
TSD
2 Deletions at the insertion site
L1
Δ
L1
3 3′ transduction
L1
AATAAA
An
L1
4 5′ transduction
Pr
L1
An
L1
5 Nonallelic homologous recombination
L1
L1
L1
L1
L1
L1
L1
L1
6 Transcript pausing or termination
Exon
3
L1
Exon
4
An
7 L1 antisense/sense promoter effects
Exon
1
Exon
2
L1
Exon
3
8 Alternative splicing
Exon
3
L1
Exon
4
9 Heterochromatization
L1
Exon
1
10 Mobilization of SINEs or pseudogenes
Alu
L1
Alu
11 Template switching by RT yields chimeras
U6
+
L1
U6L1
12 Alu editing and suppression of gene expression
Exon Exon Exon
1 2 3
Alu
Alu
An
A→I A→I
Nuclear retention of RNA
13 Microsatellite expansion caused by Alus
Exon
3
Alu
An AATATAT(AT)n

Figure 1.5 (Goodier and Kazazian, 2008). Examples of genomic changes caused by retrotransposons. (1) Insertion of the L1 element to a new location. In the example a, the insertion causes the formation of TSDs. (2) The insertion of the L1 determines a deletion at the insertion site. (3 and 4) Regions flanking the retrotransposon in the original location (at 5' or 3') may be carried along with the L1 element during the retrotransposition. (5) Mispairing and crossing over between LINE or SINE elements via NAHR, leading to deletions and duplications. (6) The retrotransposon sequence can cause a pausing in the transcriptional elongation, and poly(A) signals within an L1 can lead to premature termination of transcription. (7) The antisense promoter in the L1 5' UTR can produce new transcription start sites for genes upstream of the L1 on the opposite strand. (8) Splice sites within L1s residing in introns can lead to new exons within genes, in a process called "exonization". (9) L1s can alter the chromatin state, thereby altering gene expression. (10) L1 reverse transcriptase can mobilize Alu, SVA, mRNA, and small noncoding RNAs, leading to further genome expansion. (11) Template switching of L1 reverse transcriptase from L1 RNA to other sequences, such as U6 RNA or Alu RNA, can produce chimeric insertions in the genome. (12) Editing of inverted Alus can suppress gene expression by nuclear retention of the mRNA. (13) Alu elements can promote formation and expansion of microsatellites.

1.3 CNV Detection Methods.

Traditional chromosome-banding techniques are able to detect microscopic structural variants (>3 Mb), such as reciprocal translocations, inversion, deletions and duplication, and *fluorescence in situ hybridization* (FISH) allows a more refined characterization of these aberrations. However, the resolution of this approach is low (in the range of tens of thousands kb-Mb) and these methods miss the majority of structural variants.

The development of new technologies has provided the ability to detect submicroscopic structural rearrangements and their breakpoints with a higher resolution (ranging from kb to base pair resolution) and the number of identified CNVs has dramatically increased.

Numerous genome-wide surveys of CNVs have used array-based approaches: *array comparative genomic hybridization* arrays (array-CGH) and SNP arrays. Moreover, the advent of Next Generation Sequencing (NGS) technologies has allowed sequence-based approaches for mapping CNVs at fine scale.

1.3.1 Array-CGH.

The array-CGH method is based on the competitive hybridization between two DNA samples, a reference DNA and a DNA of interest (test DNA) (Pinkel et al., 1998). The two genomes are labelled with different fluorescent dyes (e.g. Cy5 and Cy3), and they compete to hybridize to arrays that are spotted with DNA fragments (for example BACs, PCR fragments or oligonucleotides) to cover the whole genome. The fluorescence ratio between the two dyes is then determined. In array-CGH, there are no allele-specific probes, therefore the output data consists of a series of intensity measurements that reveal the copy number differences between the two DNA samples (**Figure 1.6**).

Typically, array-CGH is carried out using a 'dye-swap' method which consists of two experiments (indicated by the left and right sides of the **Figure 1.6**): after a first hybridization experiment, a second hybridization is performed reversing the labelling of the reference and test DNA samples. This allows the detection of spurious signals, which are not common to both hybridizations.

The use of BAC clones in array-CGH gives the advantage of extensive coverage of the genome, reliable mapping data and ready access to clones. The use of long oligonucleotides instead (60-100 bp) can improve the detection resolution (theoretically from 50 kb to a few kb), compared to BACs.

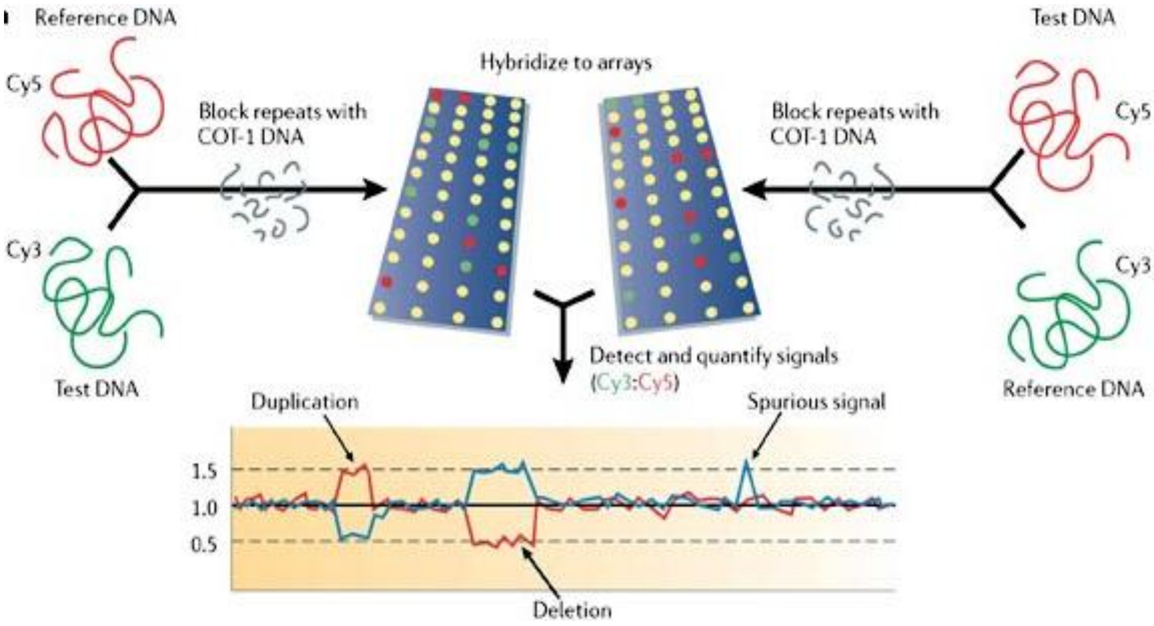


Figure 1.6 (Feuk et al., 2006a). Array-based comparative genome hybridization (array-CGH). Reference and test DNA samples are differentially labeled with fluorescent tags (Cy5 and Cy3, respectively), and are then hybridized to genomic arrays after repetitive-element binding is blocked using COT-1 DNA (which is mainly composed of repetitive sequences). After hybridization, the fluorescence ratio (Cy3:Cy5) is determined, which reveals copy-number differences between the two DNA samples. An example output for a dye-swap experiment is shown at the bottom: the red line represents the original hybridization, whereas the blue line represents the reciprocal, or dye-swapped, hybridization.

1.3.2 SNP arrays.

Whole-genome genotyping SNP arrays offer an alternative method for the identification of copy number changes. Commercial SNP-array platforms can now genotype more than two million SNPs with >99% accuracy and the genomic density of the SNPs represented in these platforms has greatly improved the resolution of CNV detection, compared to array-CGH techniques, which are limited to the detection of structural variants of ~ten to hundreds kb.

SNP arrays were originally designed to genotype simultaneously thousands of SNPs across the genome. The Affymetrix and Illumina companies developed SNP arrays based on a different chemistry (**Figure 1.7**), however, these two strategies share several aspects. Both protocols are based on a fragmentation of the genomic DNA of interest and the hybridization of ssDNA fragments to arrays containing hundreds of thousands of oligonucleotidic sequences (probes). Every SNP is interrogated by a set of probes, which are designed to be complementary to a portion of the genomic sequence containing the SNP site. Each probe is represented multiple times within an array. The number of SNPs represented on the array through these probes is proportional to the resolution of the array.

After the hybridization, a detection system measures the fluorescent signal associated with each probe. In SNP arrays, in contrast with array-CGH, the signal intensities are not compared with those of a reference genome, but with average values derived from a set of controls. A computational analysis of the raw signal data converts the intensity measures into genotype inference.

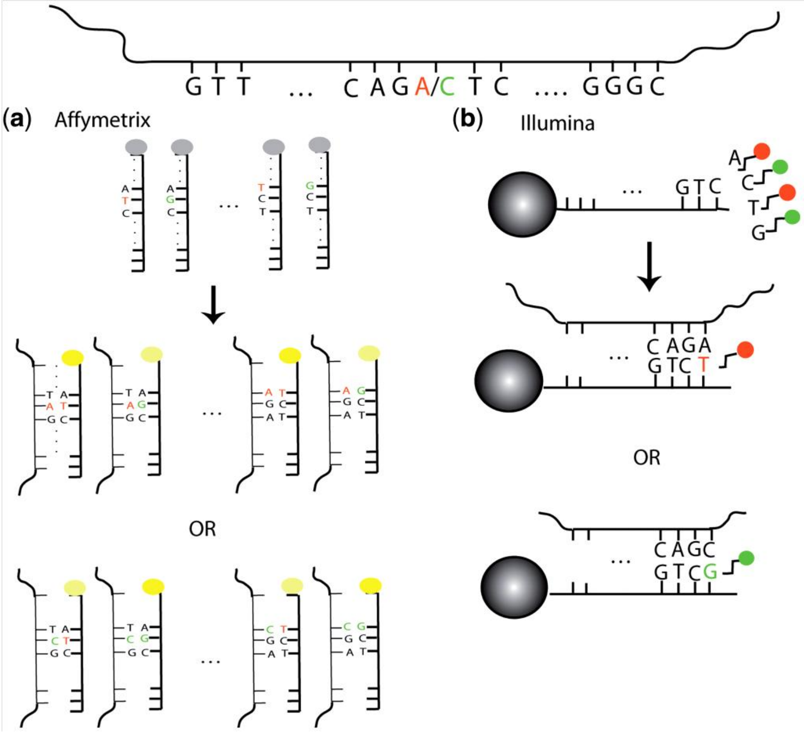


Figure 1.7 (LaFramboise, 2009). Overview of SNP array technologies. In the example, a genomic region containing the SNP A/C is shown at the top. A. Affymetrix assay: each probe (25-nt long) targets either allele A or allele B of each SNP interrogated. The DNA of interest binds to the complementary probes on the array, regardless of the allele it carries (A or C in the example). However, the efficiency is lower when there is a mismatch (indicated by a less bright yellow signal). B. Illumina BeadArray: there are 50-mer probes consisting of a sequence complementary to sequence adjacent to the SNPs interrogated. A single-base extension with labelled nucleotides results in a appropriated-colour signal.

Several algorithms have been developed for CNV discovery, such as Birdsuite (Korn et al., 2008), QuantiSNP (Colella et al., 2007) and PennCNV (Wang et al., 2007). QuantiSNP and PennCNV were originally developed for data generated by Illumina arrays. These algorithms incorporate two measures, the Log R Ratio (LRR) and the B Allele Frequency (BAF) (**Figure 1.8**), in a Hidden Markov model, that will be described in more detail in Materials and Methods.

- LRR is a normalized measure of the total signal intensity at each SNP. In autosomic regions without CNVs (copy number = 2), LRR is ~0. LRR lower than zero may indicate a deletion, LRR>0 a duplication.
- BAF represents the relative ratio of the fluorescent signals between two probes/alleles (B/A) at each SNP. BAF values range from 0 to 1: BAF close to 1 indicates that the marker is homozygous for allele B, *viceversa* BAF close to 0 indicates that the marker is homozygous for allele A. Values close to 0.5 indicate a heterozygous genotype AB. Duplicated regions

are characterized by intermediate BAF values (between 0.5 and 1 and between 0.5 and 0), correspondent to the genotypes ABB and AAB.

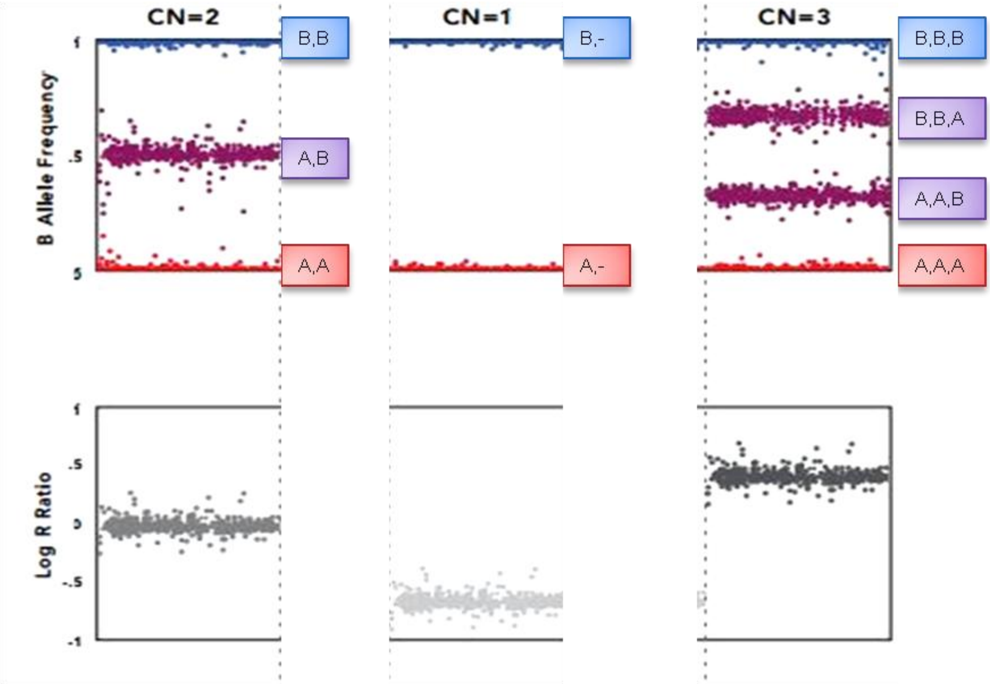


Figure 1.8. The figure shows examples of LRR and BAF plots for a deleted region (1 copy, genotype B/- or A/-), for a region with 2 normal copies (three possible genotypes for each SNP: AA, AB, BB) and a duplicated region (3 copies, four possible genotypes for each SNP: AAA, AAB, ABB, BBB).

Therefore, unlike array-CGH, SNP arrays offer the advantage of providing genotypic information, that can give support to copy number information. For example, the genotypes can reveal regions with loss of heterozygosity (LOH), that cannot be detected with array-CGH. Stretches of homozygous SNP genotypes can be a supporting evidence for the presence of a hemizygous deletion, which consists in the loss of a genomic segment on one chromosome (copy number =1). Alternatively, LOH can indicate that a genomic portion of a pair of homologous chromosomes derives from a single parent (segmental uniparental disomy, UPD). In this case LOH is defined as "copy neutral", since there is no change in copy number (copy number =2), and can be inferred only from both copy number and genotypic information. UPD can also occur in a portion of somatic cells during mitosis (referred to as acquired UPD): loss of one allele followed by reduplication determines copy neutral LOH, which is common in tumour genomes (Tuna et al., 2009). Several mechanisms have been proposed to contribute to UPD, such as mitotic non-disjunction and double-strand break repair errors.

Moreover, genotype analyses are also useful to determine the parent of origin of a *de novo* CNV and SNP data allows the ascertainment of parental consanguinity: regions of homozygosity deriving from a shared ancestry are extremely useful to discover autosomal recessive genetic causes, also in complex disorders.

1.3.3 Next Generation Sequencing-based approaches.

The characterization of human variation has been revolutionized by the recent introduction of *Next Generation Sequencing* technologies (NGS).

Massively parallel sequencing platforms, such as the HiSeq and MiSeq (*Illumina*), Ion Torrent (*Life Technologies*) and 454 Life Sciences (*Roche*), are able to perform sequencing of millions of small DNA fragments in parallel. Although these platforms use different sequencing technologies, these protocols are all based on a fragmentation of the DNA into small pieces, each of them is then sequenced and computationally aligned to the reference human genome. The sequences of each fragment ("reads") are substantially shorter than the ones that can be obtained by capillary-based sequencing technology. However, the total number of base pairs sequenced in a run is orders of magnitude higher and this has provided an extraordinary increase in DNA sequencing throughput.

New computational approaches have enabled using NGS data also to detect and map structural variants at nucleotide resolution. The main methods, represented in **Figure 1.9**, are briefly described.

a) **Read-pair analysis (RD) or paired-end mapping (PEM).** Sequences of pair of clone ends (Kidd et al., 2008; Korbel et al., 2007; Tuzun et al., 2005) or high-throughput sequencing fragments are computationally mapped to the human reference genome: abnormal mapping can reveal the presence of a CNV. For example, when the region spanned by the paired-ends in the sample genome is shorter than the correspondent region in the reference genome, this can indicate a deletion; when it is longer instead, this might indicate a simple insertion.

b) **Read depth analysis (RD)** is a method that detects structural variants by analysing read depth-of-coverage, which is measure by counting the number of reads mapping to a certain genomic window. Assuming that this number follows a Poisson distribution, the observation of an increase or a decrease of the normalized read count in a certain region may indicate a gain or a loss, respectively. A comparison between the PEM method and the Read Depth method (Yoon et al., 2009) has shown only a minority of CNVs overlap between the two call sets, indicating that these approaches have unique advantages in detecting different classes of CNVs.

c) **Split-read analysis (SR)** evaluates gapped sequence alignment for structural variants detection (Ye et al., 2009). For example, when a read spans across the breakpoint of a deletion, this sequence does not map to a single position on the reference genome, but it will be split into two fragments that map separately, indicating the position of the breakpoints.

d) **Sequence assembly (AS)** enables the fine-scale discovery of CNVs, including novel sequence insertions, which are sequences absent in the reference assembly (Hajirasouliha et al., 2010; Simpson et al., 2009). The human reference genome is itself a hybrid that contains sequences derived from different sources and individuals. Some physical gaps remain and some sequences might have been missed. Algorithms, such as EULER (Chaisson and Pevzner, 2008) and ABySS (Simpson et al., 2009), have been developed to assemble together the reads that do not map to any region of the genome ("orphan reads"): a contig formed by unmapped sequences might predict a novel inserted region.

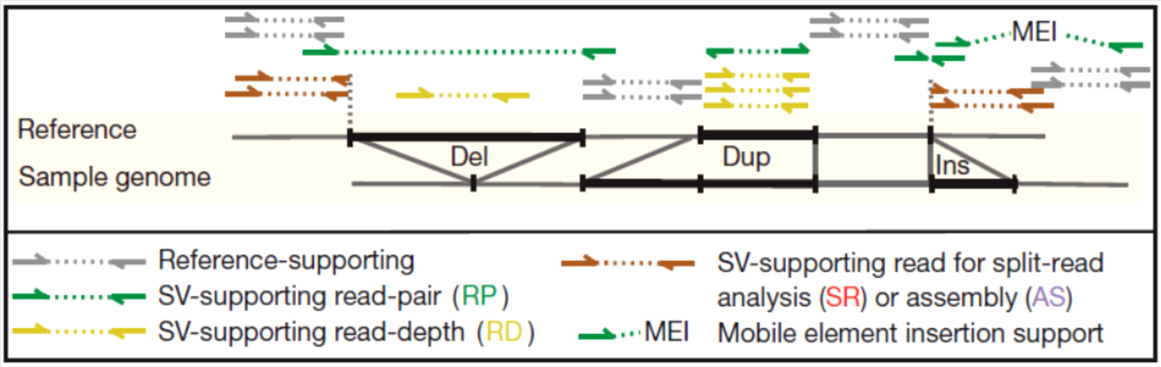

MEI
Reference
Sample genome
Del
Dup
Ins
Reference-supporting
SV-supporting read-pair (RP)
SV-supporting read-depth (RD)
SV-supporting read for split-read analysis (SR) or assembly (AS)
MEI
Mobile element insertion support

Figure 1.9 (Mills et al., 2011). Schematic representation of the sequence-based methods to detect CNVs. The arrows indicate the reads. Different sequence-based CNV-detection approaches are represented by different coloured reads.

One of the limitations of NGS technologies is represented by regions rich in GC content or with repeated architecture, that can be poorly represented in the read set or can determine erroneous mapping.

One of the main advantages of the sequence-based approaches is the base-pair resolution of the CNVs detected, which also enables the identification of the breakpoint position. Analysis of DNA region motifs surrounding the breakpoints allows to hypothesize the formation mechanism.

The Structural Variation Analysis Group of the 1000 Genomes Project (http://www.1000genomes.org), a project created with the goal of providing a deep characterization on human genetic variation in different populations, has recently applied these sequence-based approaches in order to discover CNVs larger than 50 base pairs in 185 human genomes (Mills et al., 2011). In this study, the breakpoint analysis has provided an estimation of the relative contribution of the different formation mechanisms. Nonhomology-based mechanisms (i.e., NHEJ) or MMBIR were estimated to be responsible of approximately 70.8% of the deletions, whereas 89.6% of small insertions were likely to be mediated by retrotransposition activity (*Alu* and L1 elements). Most tandem duplications showed a microhomology of 2–17 bp at the junctions and they likely arose by FoSTeS/MMBIR. Large deletions or duplications displayed extensive regions of sequence identity (>95%) at breakpoints, suggesting that they were generated by NAHR.

1.3.4 Validation of structural variants.

Genome-wide CNV screenings use computational methods to identify CNVs across the genome. These algorithms assign a measure of likelihood to each regions harbouring a copy number variant: higher scores correspond to more confident CNV calls. Generally, a cut-off for this confidence score is established to reduce the rate of false positives. Another important parameter for the selection of confident CNVs is the minimum length or the number of probes: a higher number of SNPs or longer probes (array-CGH) are stronger supporting evidence for the existence of a CNV in that region. The algorithms can also have specific parameters that can be set to reduce the number of false positives. An additional possible strategy consists in the analysis of the same data set with more than one detection algorithm: CNVs that reciprocal overlap between the call sets generated by different algorithms have higher probability to be real (Pinto et al., 2011).

However, the array-specific intensity signal variability and the DNA quality can influence the reliability of the raw data used for the CNV calling. Therefore, any structural variant detected with a prediction algorithm should be experimentally validated and several PCR-based methods are available for this purpose.

A traditional sensitive method adopted to confirm the presence of a predicted deletion or duplication is Real-Time quantitative PCR (qPCR). In a qPCR assay, fluorescent molecules (such as SYBR Green) that interact with DNA are used to monitor changes in DNA concentration. At each PCR cycle, the emitted fluorescence is measured and, during the exponential phase, it is proportional to the amount of DNA produced in the reaction. The comparison between a target region and a region with known copy number determines whether there is a gain or loss.

Multiplex PCR based-methods, alternative to qPCR, have been developed, such as Multiplex Ligation-Dependent Probe Amplification (MLPA) (Schouten et al., 2002) and Quantitative Multiplex PCR of Short Fluorescent Fragments (QMPSF) (Charbonnier et al., 2000).

If the CNV boundaries are predicted at high resolution instead, a long-range PCR could be directly performed, in order to accurately characterize the breakpoints. Long-range PCR can be an easy validation method for deletions: primers spanning the CNV breakpoints generate a PCR product shorter than the expected for the reference genomic sequence and, if resolution is sufficient, this allows the fine mapping of the boundaries. For copy number gains instead, this approach can be complicated, for example, by the two possible orientations of the duplicated region.

1.4 CNVs in Neuropsychiatric disorders.

Complex psychiatric and neurodevelopmental disorders, such as Autism Spectrum Disorders (ASD), schizophrenia (SCZ), Tourette Syndrome (TS) and bipolar disorder (BD), show a high heritability, but they have been proven to have a complex genetic architecture, in which multiple loci contribute to the overall risk.

In the past two decades, the mapping of genes underlying these diseases has been focused on two alternative hypotheses: the “common disease-common variants” model (Risch and Merikangas, 1996) and the “common disease-rare variants” model. Studies attempting to test the first hypothesis have found that common variants confer only a small or moderate level of risk (McClellan and King, 2010). These findings have suggested that the aetiology of these common diseases might be explained instead with the “common disease-rare variants” model, in which a number of different causes (SNPs or CNVs), each of them with low frequency in the population and typically highly penetrant, could collectively account for a large proportion of attributable risk (McClellan and King, 2010). These two hypotheses have been subsequently integrated in a new multifactorial model, in which common disorders could be result of a heterogeneous set of numerous rare and common variants, with different impact on the phenotype and collectively implicating a large number of different genes (Gibson, 2011).

Detection of CNVs has become an important field of genetic studies of complex disorders, as CNVs can make a substantial contribution to the genetic mechanisms underlying disease susceptibility and, in particular, rarer CNVs have been indicated as a potential source of missing heritability (Manolio et al., 2009).

Within CNV research, three study designs have been widely used.

a) Family-based approach: this approach, which examines CNVs at individuals level, enables the identification of *de novo* CNVs and allows the determination of their frequency and the association of these mutations with the disorder;

b) Case-control analysis of CNV burden: this approach analyses CNVs at population levels and examines whether the cases show a greater CNV genome-wide burden (i.e. the number of CNVs carried by an individual) compared to controls. Burden analyses can be carried out for specific categories of CNVs (such as deletions, duplications, CNVs overlapping genes or exons, etc..) and, in particular, the comparison of the collective frequency of rare variants (with frequency less than 1% in the general population) between cases and controls allows to investigate the contribution of rare CNVs to the disease.

c) Association of Target regions or Genes: this approach analyses the association of specific CNV loci with the disease phenotype.

Several interesting themes have begun to emerge from CNV studies in psychiatric and neurodevelopmental complex diseases.

Family-based studies have revealed a higher rate of *de novo* CNVs in cases compared to controls, indicating that they represent a contributing factor in 5-10% of ASD patients (Levy et al., 2011; Pinto et al., 2010; Sanders et al., 2011), 5-10% of schizophrenia cases (Kirov et al., 2012; Malhotra et al., 2011) and 4.3% of individuals affected by bipolar disorder (Malhotra et al., 2011).

De novo CNVs seem to have a stronger and more robust effect size compared to inherited CNVs. Therefore, it has been hypothesized that *de novo* mutations might be important risk factors in sporadic forms of these disorders. In autism and schizophrenia, a higher incidence of *de novo* CNVs in children from simplex families (i.e. families with one affected individual) compared to children from multiplex families (i.e. families with multiple affected individuals) was observed in some initial studies (Marshall et al., 2008; Sebat et al., 2007; Xu et al., 2008); however this difference did not emerge from more recent studies (Malhotra et al., 2011; Pinto et al., 2010).

In autism and schizophrenia, burden analyses have shown a general enrichment of potentially pathogenic duplications in a larger size range (>500 kb), and of deletions in smaller size range (30-500 kb) (International_Schizophrenia_Consortium, 2008; Pinto et al., 2010). CNVs larger than 500 kb usually contain multiple genes and can be found in a small proportion of the general population (~8%), suggesting that these large rearrangements are under purifying selection and they are likely to be potentially pathogenic. Small CNVs instead are more frequent in the general population and this complicates the identification of pathogenic CNVs; however small CNVs can also represent important risk factors and the investigation of their role needs future studies with sufficient power in sample size and resolution.

Studies including large cohorts of cases of autism (Levy et al., 2011; Sanders et al., 2011) and schizophrenia (International_Schizophrenia_Consortium, 2008; Stefansson et al., 2008) have found a significant increase in large rare/*de novo* CNV burden in cases compared to both controls and unaffected siblings. In other two studies, the burden enrichment of rare CNVs appeared more pronounced when considering only rare CNVs overlapping genes (“genic” CNVs) (Pinto et al., 2010; Walsh et al., 2008).

Regarding Tourette Syndrome, a case-control study, carried out in Caucasian individuals, did not report an increased burden of rare CNVs in cases compared to controls (Fernandez et al., 2012). Another study, carried out in two Latin-American populations, found a significant increase of large CNVs in cases (Nag et al., 2013).

Conflicting results have been reported for bipolar disorder: two studies found an enrichment of rare CNV in patients compared to controls (Priebe et al., 2012; Zhang et al., 2009), but this finding was not replicated in another two studies (Grozeva et al., 2010; McQuillin et al., 2011). However, a study focused on rare *de novo* structural variants, found a significant increase in the *de novo* rate in patients affected by BD compared to controls (Malhotra et al., 2011).

No evidence for CNV burden enrichment was observed for dyslexia (Girirajan et al., 2011).

Another interesting observation that emerged from these burden analyses is the fact that the size and the rate of rare/*de novo* CNVs seem to correlate with the severity of the phenotype. The highest burden of large rare/*de novo* CNVs has been observed in cases with intellectual disability (ID) and dysmorphic features (Girirajan et al., 2011), the lowest burden in bipolar cases, in between the extremities are schizophrenia and autism. This trend seems to support a model where neurodevelopmental disorders, based on their severity and co-morbidities of ID, are considered part of a continuum (**Figure 1.10**).

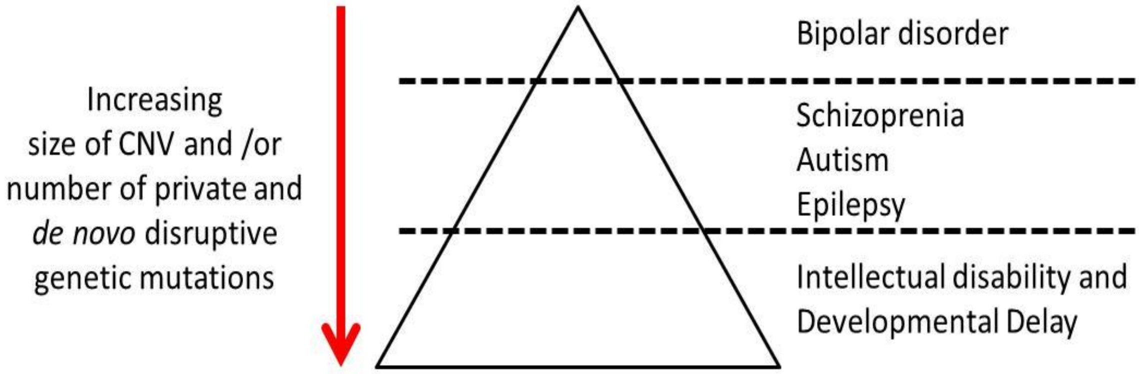


Figure 1.10 (Coe et al., 2012). An oligogenic model for neurodevelopmental disorders. In this model, a higher number of large, rare, *de novo* CNVs and, more in general, disruptive genetic mutations correlates with an increase in the severity of the clinical phenotype.

Another interesting theme emerged from CNV studies is the fact that a CNV locus can be recurrently identified in association with a variety of multiple neurological phenotypes (*pleiotropic effect*). In some cases, a CNV is necessary and sufficient to result in a specific phenotype: these CNVs are often associated with known syndromes (“syndromic CNVs”). By contrast, there are CNVs which are much more variable in their outcome. These loci exhibit a *variable expressivity*, which means that individuals carrying the same CNV show either a qualitative or quantitative phenotypic variation, and a *reduced penetrance*, as these CNVs can be identified also in asymptomatic carriers (the penetrance is the probability of expressing a trait given a certain genetic change). Therefore, in contrast with syndromic CNVs, these recurrent CNVs often show an incomplete segregation within multiplex families and are likely to be inherited from a parent, who may present any of the phenotypic manifestations associated with the CNV or have normal phenotype. Some of these CNV hotpots are 16p11.2, 15q13.3, 1q21.2, 3q29, 17q21.31 and 7q11.23

loci (Coe et al., 2012). A well-characterized example is represented by microdeletions and microduplications at 15q13.2-15q13.3, that are presented in the next paragraph.

The variable expressivity observed among the cases carrying the same CNV might be determined by secondary insults (“hits”), including additional pathogenic CNVs or damaging sequence mutations, by differences in the genetic background and epigenetic regulation and could be modulated by sex-related or age-related factors and environmental factors. Whether the CNV is maternally or paternally inherited can also have consequences on the phenotype, if the locus affected by the CNV is includes imprinted genes (Hehir-Kwa et al., 2013).

Moreover, these recurrent CNVs show that losses and gains occurring at the same locus can lead to drastically different phenotypes or, instead, to surprisingly overlapping phenotypes. An example of different outcomes associated with CNVs at the same locus is given by the 7q11.23 region. The deletions, which cause hemizygous loss of about 20 genes, are associated with the William-Beuren Syndrome, characterized by mental retardation, visuo-spatial impairments, but precocious verbal ability and highly sociable disposition. By contrast, duplications at this locus have been found in association with autism (Sanders et al., 2011), which is characterized instead by deficits in social interactions and communication, and with schizophrenia (Mulle et al., 2013).

Differences and similarities between clinical phenotypes caused by reciprocal CNVs (duplication/deletion) can be attributed in part to the variable dosage sensitivity of the genes encompassed by the CNV: for some genes, an increased or a reduced dosage can determine opposite consequences (“mirror” phenotype); for other genes, the dosage imbalance alters certain cellular functions, irrespectively of the type of copy-number change. It is worth noting that this variability is not limited to hotspot loci, but might be true also for other genomic regions.

Genic CNVs can be an important source for the discovery of the genes contributing to complex disorders and of pathways implicated in complex disorders. Rare structural variants overlapping genes or exons may be pathogenic, but, if these variants are considered individually, it might be difficult to establish their contribution to the disorder. When the rare variants in a gene or in a set of genes acting in the same pathway are taken into account collectively instead, they may highlight the processes relevant to the disorder. A number of algorithms have been developed to determine whether sets of genes hit by CNVs are significantly overrepresented, compared with all known genes.

- In autism, three studies have highlighted a CNV enrichment within or surrounding genes involved in the ubiquitination pathways (Glessner et al., 2009), gene sets related to processes such as cellular proliferation, development and maturation of synapse contacts, neuron motility, axon targeting, cell-cell adhesion and GTPase/Ras signaling (Gilman et al., 2011; Pinto et al., 2010), supporting the main role of synaptogenesis and neuronal connectivity in the disease.

- Also in schizophrenia, rare CNVs tend to target functional classes of genes related to neurodevelopment pathways and synaptic activity, including synaptic long term potentiation, glutamate receptor signaling, axon guidance signaling and ERK/MAPK signaling (Malhotra et al., 2011; Walsh et al., 2008).
- In Tourette syndrome, pathway analysis of rare genic CNVs pointed to the involvement of genes within histamine receptor signaling pathways, sphingolipid metabolism, axon guidance, cell adhesion, ubiquitination pathway, nervous system development, and synaptic structure and function processes (Fernandez et al., 2012).
- In bipolar disorder, a study (Zhang et al., 2009) reported an overrepresentation of genes involved in pathways important for learning behaviours and psychological disorders, whereas a second study that focused on *de novo* CNVs (Malhotra et al., 2011), did not identify an enrichment of genes involved in neuronal function or development, but instead genes related to cell proliferation and shape and phospholipid metabolism.

Interestingly, single case reports and network analyses indicate that the pathways implicated in different diseases may converge on common genes, such as *CNTNAP2 (contactin associated protein-like 2)* and *NRXN1 (Neurexin 1*), suggesting that there are some key genes important for several aspects of brain development and, if mutated, they can contribute to a range of disorders, depending on the genetic background. This supports the hypothesis of shared biological pathways among different neuropsychiatric conditions (Guilmatre et al., 2009).

At least for ASD, network analyses have estimated that hundreds of genes will be involved in the disorder, confirming the locus heterogeneity underlying complex neurodevelopmental diseases. So far, the rarity of the contributing variants and their heterogeneity have represented a challenge for the identification of the genes implicated in these disorders. However, in the next years, the gene discovery effort is expected to be facilitated by the increasing availability of high-throughput CNV and sequencing data. The integration of structural and sequence information will allow the capture of a larger fraction of the rare disease-causing variants and to explain a larger proportion of the risk.

1.5 Recurrent CNVs on chromosome 15q11-q13.

The proximal region of the long arm of chromosome 15 (15q11-q14) is a well-known hotspot for CNVs. The presence of complex patterns of highly homologous LCRs makes this locus one of the most unstable regions in the human genome and this can also be appreciated observing the large number of CNVs reported for this region in the DGV. The structural and sequence features of the

15q11-qq14 region can lead to a number of different rearrangements, e.g. deletions, duplications, translocations, inversions and, also, supernumerary inv-dup(15) chromosomes. Deletions and duplications are likely to be caused by NAHR between the LCRs. The breakpoints of these CNVs usually coincide with clusters of LCRs, that have been designated BP1-BP6 (**Figure 1.11**).

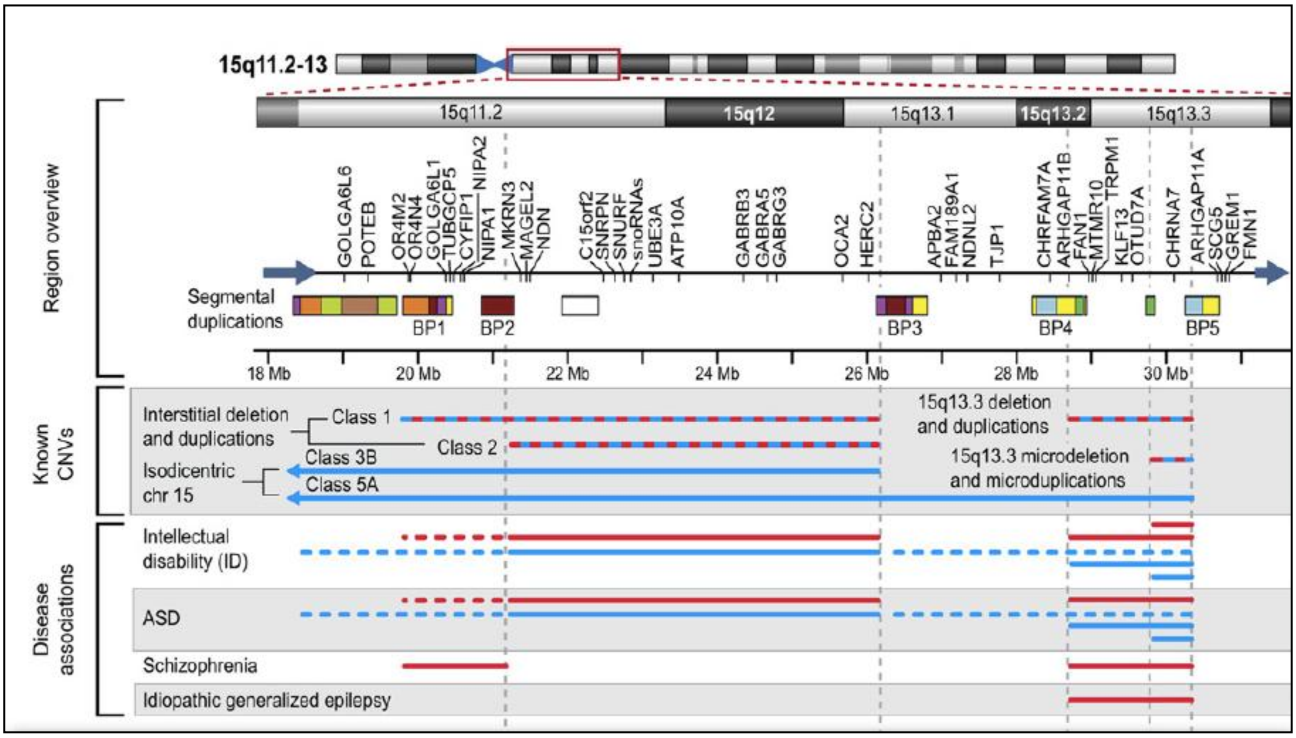


Figure 1.11 (Sanders et al., 2011). The genomic architecture of the BP1-BP5 region on chromosome 15q. The position of each BP is indicated. Class 1 indicates BP1-BP3 deletions, class 2 instead indicates the BP2-BP3 deletions.

1.5.1 Recurrent CNVs at 15q11.2 locus (BP1-BP2 region).

The BP2-BP3 region includes imprinted genes, therefore the effects of mutations in this region depend on the origin (maternal or paternal) of the chromosome in which they occur. BP1-BP3 deletions (Type I) or BP2-BP3 deletions (Type II) (**Figure 1.12**) can result in Prader-Willi Syndrome, if the deletion is inherited from the father, or Angelman Syndrome, if the deletion is inherited from the mother. Duplications instead can be associated with learning disabilities, seizures and autism (maternal duplications are present in 1-3% of individuals with ASD (Veenstra-VanderWeele and Cook, 2004)).

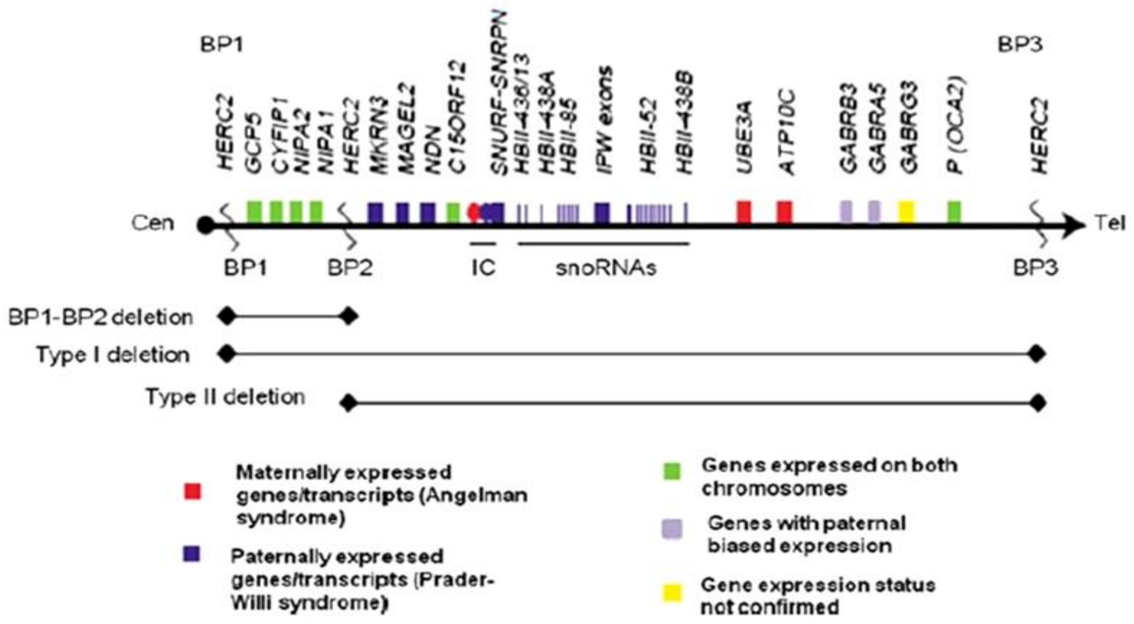


Figure 1.12 (Burnside et al., 2011). The figure shows the BP1-BP3 region on chromosome 15q: the genes are differentially coloured, depending on their imprinting.

BP1-BP2 microdeletions and microduplications have been proposed as risk factors for a range of neurological problems, in particular, language delay and developmental delay (Burnside et al., 2011; Doornbos et al., 2009), ID (Cooper et al., 2011), ASD (Doornbos et al., 2009; Sanders et al., 2011; van der Zwaag et al., 2010), schizophrenia (Kirov et al., 2009; Stefansson et al., 2008) and epilepsy (de Kovel et al., 2010), but they have been observed also in controls. A screening for large CNVs performed on 15,767 children with ID and various congenital deficits and 8,329 healthy adult controls (Cooper et al., 2011), found that BP1-BP2 microdeletions were significantly associated with neurodevelopmental disorders (p-value= 4.73 x 10^{-6}), epilepsy (p-value= 1.48 x 10^{-3}) and autism (p-value= 1.99 x 10^{-2}). Moreover, another paper (Burnside et al., 2011), describing a retrospective analysis of cases with BP1-BP2 rearrangements, showed that a high percentage of these subjects had speech delay (50% of duplications carries, 90% of deletions carriers), but also developmental delay (39% of duplications carries, 59% of deletions carriers) and ASD features (41% of duplications carries, 29% of deletions carriers). The interval between BP1 and BP2 (~500kb) contains four non-imprinted genes: *TUBGCP5, CYFIP1, NIPA1* and *NIPA2.* These genes are evolutionary conserved and they code for proteins that could be potentially implicated in neurological dysfunctions:

- *TUBGCP5* (OMIM 608147) encodes the gamma-tubulin complex component GCP5, a member of cytoskeleton tubulin complex;
- *CYFIP1* (OMIM 606322) codes for a cytoplasmic protein that interacts with FMRP, the protein implicated in fragile X syndrome;

- *NIPA1* (*non imprinted in Prader-Willi/Angelman syndrome 1*, OMIM 608145) and *NIPA2* (*non imprinted in Prader-Willi/Angelman syndrome 2*, OMIM 608146) code for magnesium transporters that can be found in a variety of neuronal and epithelial cells, suggesting that these proteins may have a role in nervous system development and maintenance.

1.5.2 Recurrent CNVs at 15q13.2-15q13.3 locus (BP4-BP5 region).

The segmental duplications BP4 and BP5 (15q13.2-15q13.3) can mediate a range of different recombination events. The typical rearrangements occurring between BP4 and BP5 are CNVs with a size of ~1.6 Mb and encompass six RefSeq genes (*MTMR15, MTMR10, TRPM1, KLF13, OTUD7A* and *CHRNA7*) and the microRNA gene *hsa-mir211*. These recurrent large BP4-BP5 deletions and duplications (**Figure 1.11**) have been found across multiple conditions and also in asymptomatic carriers. In particular, the deletion has been reported in cases of mental retardation with seizures (Sharp et al., 2008), autism (Miller et al., 2009; Pagnamenta et al., 2009), schizophrenia (International_Schizophrenia_Consortium, 2008; Stefansson et al., 2008), bipolar disorder (Ben-Shachar et al., 2009) and epilepsy (Helbig et al., 2009; Masurel-Paulet et al., 2010) and language delay (Ben-Shachar et al., 2009).

However, BP4 and BP5 have a complex organization and can lead also to smaller deletions and duplications (**Figure 1.11**). Recurrent atypical smaller microdeletions and microduplications (350-680 kb) have also been found in a similar range of neuropsychiatric phenotypes (Leblond et al., 2012; Szafranski et al., 2010).

All these CNVs include the gene *CHRNA7* (OMIM 118511), encoding the α7 subunit of the neuronal nicotinic acetylcholine receptor, and the first exon of one isoforms of *OTUD7A* (OMIM 612024), encoding a putative deubiquitinating enzyme. Between these two genes, *CHRNA7* seems to be the most likely candidate gene for cognitive and neurobehavioral deficits, as it is highly expressed in the brain and, given its functions, relates to seizures and epilepsy (Shinawi et al., 2009; Szafranski et al., 2010), therefore it has been hypothesized that abnormal dosage of *CHRNA7* could alter the neuronal homeostasis (Shinawi et al., 2009; Szafranski et al., 2010).

A recent study (Moreno-De-Luca et al., 2013) has analysed the frequency of rare CNVs recurrently found in association with neuropathological phenotypes, combining published data available for large cohorts of clinical cases and control data sets. The analysis was performed on a group of 31,516 cases (including patients with developmental delay, ID, ASD or multiple congenital abnormalities) (Cooper et al., 2011; Kaminsky et al., 2011) and 13,696 controls (Cooper et al., 2011; International_Schizophrenia_Consortium, 2008; Magri et al., 2010; Shaikh et al., 2009). As shown by **Table 1.1** and **Table 1.2**, statistical support for a pathological role was found for several

CNV loci, including the BP4-BP5 region on chromosome 15q13.2-q13.3. The BP4-BP5 deletion was detected in 88 cases, whereas the BP4-BP5 duplication was identified in 34 cases and 5 controls. Both CNV types were reported to have a statistical significant increase in cases compared to controls. However, while the deletion appears to have a complete penetrance, the reciprocal microduplication shows incomplete penetrance.

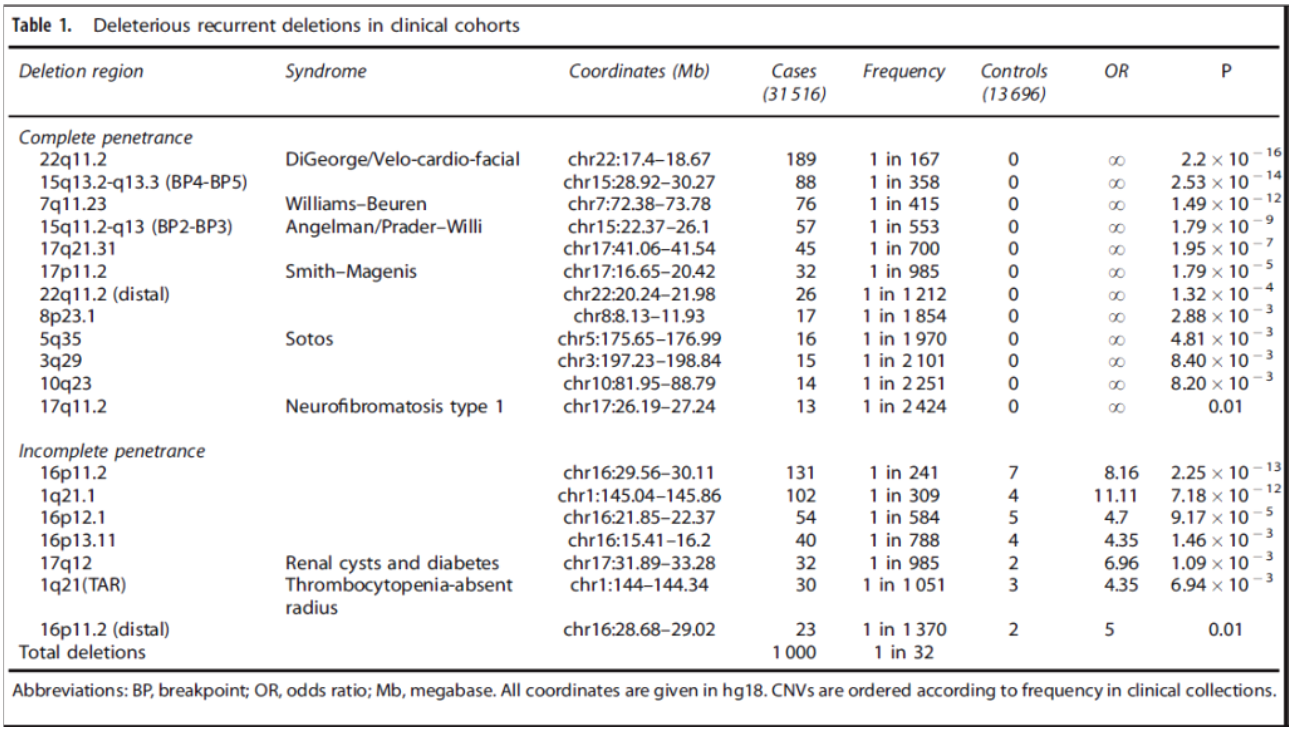


Table 1. Deleterious recurrent deletions in clinical cohorts

Deletion region	*Syndrome*	*Coordinates (Mb)*	*Cases (31 516)*	*Frequency*	*Controls (13 696)*	*OR*	P
Complete penetrance							
22q11.2	DiGeorge/Velo-cardio-facial	chr22:17.4–18.67	189	1 in 167	0	∞	2.2×10^{-16}
15q13.2-q13.3 (BP4-BP5)		chr15:28.92–30.27	88	1 in 358	0	∞	2.53×10^{-14}
7q11.23	Williams–Beuren	chr7:72.38–73.78	76	1 in 415	0	∞	1.49×10^{-12}
15q11.2-q13 (BP2-BP3)	Angelman/Prader-Willi	chr15:22.37–26.1	57	1 in 553	0	∞	1.79×10^{-9}
17q21.31		chr17:41.06–41.54	45	1 in 700	0	∞	1.95×10^{-7}
17p11.2	Smith–Magenis	chr17:16.65–20.42	32	1 in 985	0	∞	1.79×10^{-5}
22q11.2 (distal)		chr22:20.24–21.98	26	1 in 1 212	0	∞	1.32×10^{-4}
8p23.1		chr8:8.13–11.93	17	1 in 1 854	0	∞	2.88×10^{-3}
5q35	Sotos	chr5:175.65–176.99	16	1 in 1 970	0	∞	4.81×10^{-3}
3q29		chr3:197.23–198.84	15	1 in 2 101	0	∞	8.40×10^{-3}
10q23		chr10:81.95–88.79	14	1 in 2 251	0	∞	8.20×10^{-3}
17q11.2	Neurofibromatosis type 1	chr17:26.19–27.24	13	1 in 2 424	0	∞	0.01
Incomplete penetrance							
16p11.2		chr16:29.56–30.11	131	1 in 241	7	8.16	2.25×10^{-13}
1q21.1		chr1:145.04–145.86	102	1 in 309	4	11.11	7.18×10^{-12}
16p12.1		chr16:21.85–22.37	54	1 in 584	5	4.7	9.17×10^{-5}
16p13.11		chr16:15.41–16.2	40	1 in 788	4	4.35	1.46×10^{-3}
17q12	Renal cysts and diabetes	chr17:31.89–33.28	32	1 in 985	2	6.96	1.09×10^{-3}
1q21(TAR)	Thrombocytopenia-absent radius	chr1:144–144.34	30	1 in 1 051	3	4.35	6.94×10^{-3}
16p11.2 (distal)		chr16:28.68–29.02	23	1 in 1 370	2	5	0.01
Total deletions			1 000	1 in 32			

Abbreviations: BP, breakpoint; OR, odds ratio; Mb, megabase. All coordinates are given in hg18. CNVs are ordered according to frequency in clinical collections.

Table 1.1 (Moreno-De-Luca et al., 2013). Deleterious recurrent deletions in clinical cohorts.

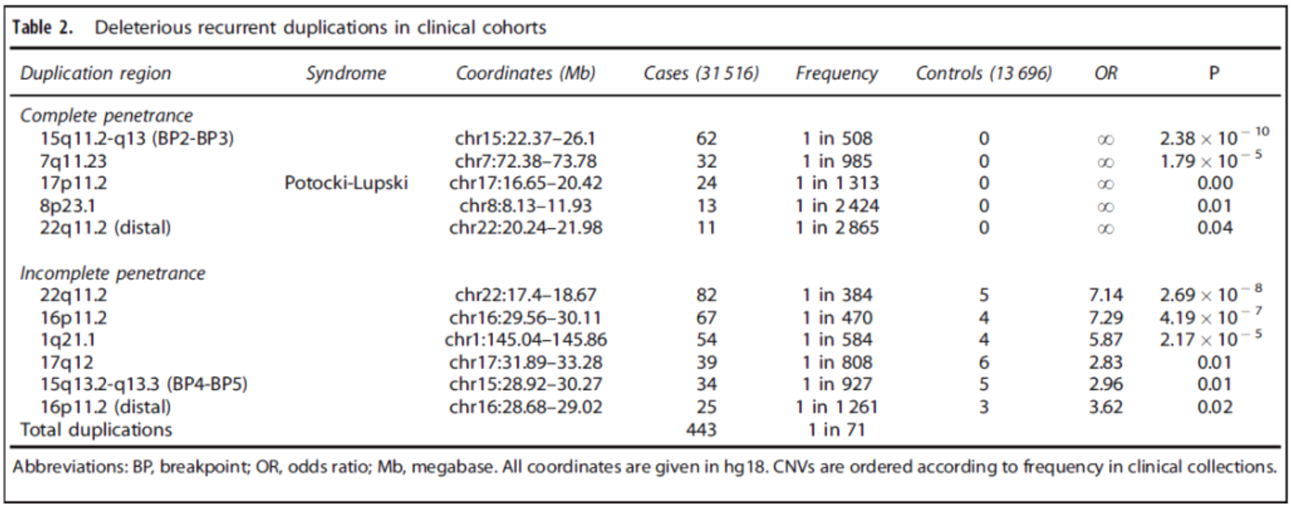


Table 2. Deleterious recurrent duplications in clinical cohorts

Duplication region	*Syndrome*	*Coordinates (Mb)*	*Cases (31 516)*	*Frequency*	*Controls (13 696)*	*OR*	P
Complete penetrance							
15q11.2-q13 (BP2-BP3)		chr15:22.37–26.1	62	1 in 508	0	∞	2.38×10^{-10}
7q11.23		chr7:72.38–73.78	32	1 in 985	0	∞	1.79×10^{-5}
17p11.2	Potocki-Lupski	chr17:16.65–20.42	24	1 in 1 313	0	∞	0.00
8p23.1		chr8:8.13–11.93	13	1 in 2 424	0	∞	0.01
22q11.2 (distal)		chr22:20.24–21.98	11	1 in 2 865	0	∞	0.04
Incomplete penetrance							
22q11.2		chr22:17.4–18.67	82	1 in 384	5	7.14	2.69×10^{-8}
16p11.2		chr16:29.56–30.11	67	1 in 470	4	7.29	4.19×10^{-7}
1q21.1		chr1:145.04–145.86	54	1 in 584	4	5.87	2.17×10^{-5}
17q12		chr17:31.89–33.28	39	1 in 808	6	2.83	0.01
15q13.2-q13.3 (BP4-BP5)		chr15:28.92–30.27	34	1 in 927	5	2.96	0.01
16p11.2 (distal)		chr16:28.68–29.02	25	1 in 1 261	3	3.62	0.02
Total duplications			443	1 in 71			

Abbreviations: BP, breakpoint; OR, odds ratio; Mb, megabase. All coordinates are given in hg18. CNVs are ordered according to frequency in clinical collections.

Table 1.2 (Moreno-De-Luca et al., 2013). Deleterious recurrent duplications in clinical cohorts.

The analysis was then restricted to 3,955 cases belonging to three of the largest ASD cohorts and included 1124 cases from the SSC, 996 from AGP and 1835 from Autism Genetic Resource Exchange (AGRE) (Geschwind et al., 2001; Itsara et al., 2010; Pinto et al., 2010; Sanders et al., 2011). In this subset, only BP4-BP5 deletions yielded a statistical significant difference between cases and controls (*p*-value =1.26 x 10^{-4}), while the BP4-BP5 duplications did not showed a significant enrichment in cases (*p*-value =0.66) (Moreno-De-Luca et al., 2013).

In conclusion, BP1-BP2 and BP4-BP5 CNVs can be observed in a wide spectrum of clinical phenotypes of variable severity and some of them, such as BP1-BP2 microdeletions and BP4-BP5 microduplications, are present also in healthy individuals, suggesting that they have variable expressivity and incomplete penetrance and are probably not sufficient to cause pathological phenotypes (Leblond et al., 2012; van Bon et al., 2009). However, since they involve interesting candidate genes and have a higher frequency in clinical cohorts, they might contribute to the susceptibility to certain neuropsychiatric disorders in specific genetic backgrounds, where secondary alterations could have an additive or epistatic effect.

Chapter 2

Specific Language Impairment

2.1 What is Specific Language Impairment?

Language is a peculiar faculty of human beings that plays a central role in social interactions. Verbal forms of communication can be found also in other organisms, but the structural complexity of human language is a unique property.

Specific Language Impairment (SLI) is a common neurodevelopmental disorder defined as an inability to develop appropriate language skills despite normal intelligence and access to adequate educational opportunities. SLI is diagnosed when expressive and receptive language abilities are severely affected, whereas non-verbal cognitive abilities are within age expectation, in the absence of any medical conditions that might underlie the language problems (e.g. hearing loss, mental retardation, autism). SLI is a heterogeneous condition that can vary in both severity and language impairment profiles. This disorder affects up to 8% of preschool children (Law et al., 2000), with a higher prevalence in males and, in some cases, it persists into adulthood (Conti-Ramsden et al., 2001).

2.2 Measurement of SLI symptoms.

The acquisition of language is one of the key milestones of childhood. Children affected by specific language impairment (SLI) struggle to acquire basic competence in one or more aspects of spoken language. Adequate language development requires the acquisition of the ability to comprehend what others say (receptive language) and to produce utterances that other people can understand (expressive language). Language abilities can be further subdivided into several domains, including the knowledge and appropriate use of:

- words and their meaning (lexicography, semantics, or vocabulary);
- the way that sequences of words combine in sentences (grammar or syntax);
- linguistic forms involved in social interactions (pragmatic language);
- the system of speech sounds that make up the language (phonology).

Children with SLI display difficulties of variable degree in one or more of these language domains (**Table 2.1**).

Characteristics of Specific Language Impairments (SLI)

Diagnostic criteria

- Language is significantly below level expected from age and IQ, usually interpreted as scoring in the lowest 10% on a standardized test of expressive and/or receptive language
- Nonverbal IQ and nonlinguistic aspects of development (self-help skills, social skills) fall within broadly normal limits
- Language difficulties cannot be accounted for by hearing loss, physical abnormalities of the speech apparatus or environmental deprivation
- Language difficulties are not caused by brain damage

*Common presenting features**

- Delay in starting to talk; first word may not appear until 2 years of age or later
- Immature or deviant production of speech sounds, especially in preschool children
- Use of simplified grammatical structures, such as omission of past tense endings or the auxiliary “is” well beyond the age when this is usually mastered
- Restricted vocabulary, in both production and comprehension
- Weak verbal short term memory, as evidenced in tasks requiring repetition of word or sentences
- Difficulties in understanding complex language, especially when the speaker talks rapidly

*SLI shows considerable heterogeneity, as well as age-related changes

Table 2.1 (Bishop, 2006). Phenotypic characteristic of SLI.

There are many tests that analyse lots of different aspects of language. These tests provide quantitative measures of different language endophenotypes, that can be also used for genetic investigations. However, an exhaustive description of them is beyond the scope of this section. Several tests have been developed for English-speaking children, such as TOLD (Test Of Language Development) and Clinical Evaluation of Language Fundamentals Revised test (CELF-R).

TOLD (Hammil et al., 1987; Hammil and Newcomer, 1988; Newcomer and Hammil, 1988) is a comprehensive test of language functioning that evaluates specific subtypes of language domains, such as comprehension, expression, grammar, syntax and phonology.

CELF-R (Semel et al., 2004; Semel et al., 1992) is widely used to identify, diagnose and follow-up language impairments in school-age children (5–17 years). CELF-R examines expressive and receptive language domains separately and then combines the results in a composite language score. Different batteries of test are available, depending on the age of the subject.

Another important language-related domain that can be evaluated is the Phonological Short Term Memory (PSTM), a working memory critical for a temporary storage and processing of incoming words or sounds. PSTM allows a representation of speech sounds (phonemes) in the brain. Based on this theory, Gathercole *et al*. (Gathercole et al., 1994) developed a test, known as “NonWord

Repetition" (NWR), to assess the capacity that phonological working memory has to store and process unfamiliar words or words with no meaning (nonsense words). In this task, the child cannot recall his stored knowledge, but he is forced to rely upon the temporary representation of the non-word in the short-term phonological store. Children with SLI poorly perform on this repetition task (Bishop et al., 1996; Gathercole et al., 1994), suggesting that the amount of memory necessary to hold novel phonological forms in their PSTM is insufficient to allow in-depth processing and transfer of this information to the long-term memory. Interestingly, this test is able to reveal deficits also in individuals reported to have language difficulties during early childhood, later resolved, indicating that this measure is a good marker for language impairments.

The Past Tense (PT) test (Marchman et al., 1999) evaluates grammatical competencies, that are not assessed by CELF and NWR tests, such as the ability to add appropriate inflectional endings to verbs, which is frequently impaired in English-speaking children affected by SLI.

Reading tests can also be informative about aspects of language, because the comprehension of a written text requires both reading abilities and language comprehension skills. Indeed, although reading impairment is distinct from specific language impairment, these deficits co-occur in ~50% of the affected individuals (Flax et al., 2003).

2.3 The genetic bases of SLI.

The evidence for an influence of genetic factors in spoken language disorders has emerged from twin and familial studies.

A review of eighteen studies relative to spoken language impairment (Stromswold, 1998) indicated that the incidence of language deficits was significantly greater in families with a SLI "proband" (the individual through whom the family was identified) than in families with an unaffected proband with normal speech and language (controls).

Studies based on the concordance rate of the disorder between co-twins allows to distinguish genetic from environmental influences, and therefore they can prove if a disorder is heritable. Monozygotic twins (MZ) can be considered 100% genetically identical, whereas dizygotic twins (DZ) are assumed to be 50% genetically similar. Three twin studies have shown an increased MZ concordance compared with DZ concordance rates (Bishop, 2002; Bishop et al., 1995; DeThorne et al., 2006; Tomblin and Buckwalter, 1998). Considering that both twin types usually share the same environment, a higher MZ concordance indicated that SLI has a strong genetic component. However, family studies have failed to detect a simple dominant or recessive pattern of inheritance, suggesting a complex genetic architecture underlying the disorder.

2.4 Molecular genetic studies for SLI susceptibility.

In order to identify the genetic risk factors conferring susceptibility to SLI, linkage and association studies have been performed. Both approaches are based on the genotyping and analysis of genetic markers, which are polymorphic variants with a known position in the genome (e.g. SNPs or microsatellites).

2.4.1 Linkage studies.

Linkage studies are based on the principle that polymorphic genetic markers located in proximity of disease-causing variants cosegregate with the affection status, across generations: the linkage between these markers and the disease-causing variants is detected when they are transmitted together to the offspring more often than expected under independent inheritance. Their probability of being separated by a crossing over during meiosis is proportional to their distance on the chromosome. These studies involve related individuals and can include large pedigrees (with extended families and/or multiple generations) or a large number of small nuclear families (consisting of a father, a mother, and their children).

Linkage methods can be divided into two main approaches:

- *"model-based" or parametric linkage analyses*, which rely on the specification of a genetic model of the disease, in particular the inheritance pattern (dominant or recessive), the penetrance levels, the expected frequencies of the disease allele in the population;
- *"model-free" or non-parametric linkage analyses* (NPL), suitable for disorders with an unknown genetic model, which is often the case of complex and heterogeneous diseases. A non-parametric linkage approach consists of the analysis of the number of *identical-by-descendent* (IBD) alleles shared by pairs of affected siblings (ASP). For any genetic locus, siblings can share both alleles (IBD=2), half their alleles (IBD=1) or none of their alleles (IBD=0), depending on the segregation pattern of parental alleles, with probabilities of 25%, 50% and 25%, respectively. This linkage sib-pair method is based on the hypothesis that chromosomal region harbouring the disease-causing genes (and thus the alleles of polymorphic markers close to these genes) are likely to be shared by ASPs. Therefore, genomic regions where the IBD alleles are shared by sib-pairs more often than expected by chance allow the identification of loci linked to the disease. These analyses can be qualitative or quantitative, depending on whether they test the correlation between genetic similarity (estimated by the IBD sharing) and binary (presence or absence) traits or trait on a continuous scale.

Results of linkage analyses are reported as Logarithm Of Odds (LOD) scores, that are a function of the recombination fraction (θ), which indicates the probability of a recombination event between two loci at meiosis. LOD scores indicate the likelihood that a marker and the disease-gene are physically linked. When the LOD scores are calculated taking into account the fact that several different genes can contribute to the susceptibility of a disorder (locus heterogeneity), they are reported as Heterogeneity LOD scores (HLOD).

Traditionally, the threshold for significant linkage is indicated as a LOD score of 3. More stringent criteria may be required, depending on the number of genotyped markers and the type of analysis used. Sometimes, lower LOD scores (~2) are reported as suggestive linkage. Linkage studies usually identify large regions of susceptibility (in the order of Mb), encompassing a large number of genes, and their success is affected by genetic and phenotypic heterogeneity of the disorder.

2.4.2 Association studies.

Association studies aim to find an allelic association between specific genetic variants and the disease phenotype in a population, assuming that the marker itself or a variant close to it confers susceptibility to the disease. When an allelic variant is associated with a trait, this result can be interpreted as:

a) a direct association, if the variant has a causal role in the phenotype susceptibility;

b) an indirect association, if the variant is in Linkage Disequilibrium with the causal variant (LD is a phenomenon arising from alleles at linked loci, that tend to cosegregate more often than expected by chance and forming “haplotypes blocks”);

c) a false positive, that may be due to chance or to problems such as population stratification or inappropriate statistical methods.

These studies are usually performed with one of two main different approaches (McCarthy et al., 2008).

- *Case-control association studies* involve large numbers of unrelated cases and controls (the sample size should be in the order of thousands of individuals) to examine if the prevalence of specific variants is significantly higher in cases than in controls. Cases and controls are required to be well-matched, in particular for ethnic background, to avoid spurious association signals that can be due to population substructures (stratification). The control group should be formed by individuals classified as “unaffected” after specific assessment for the absence of the disease (also referred to as “supernormal” controls), but often is formed by large numbers of individuals randomly collected from the population (“unscreened” controls). For common diseases, when screening for the investigated phenotype is possible, “supernormal” controls represent a better alternative as they increase the power to detect associations (Lewis and Knight, 2012).

- *Transmission disequilibrium test (TDT)* is a robust family-based strategy based on parental heterozygous markers: if an allele is associated with the disease, it will be transmitted to the affected children more often than expected by chance (50%) (Spielman et al., 1993). Although the TDT has a reduced statistical power compared to the case-control studies, it avoids issues related to the optimal selection of control samples, because the parental alleles not transmitted to the affected offspring are considered “internal controls”. TDT also allows to differentiate effects of alleles, testing whether there is a preferential paternal or maternal transmission of susceptibility alleles *(parent-of-origin*).

Results of association studies are usually expressed as P values or $–\log_{10}(P)$: very low P values provide strong evidence for association. The significance threshold depends on the number of markers: as the number of markers increases, the required number of tests increases and the significance threshold becomes more stringent. The traditionally accepted significance threshold for genome-wide association studies (GWAs) is 5×10^{-8}, which was estimated to give a probability higher than 95% of having no false positives for 1,000,000 independent tests (Risch and Merikangas, 1996). This threshold is appropriate, for example, for arrays with 1 million SNPs on them. Generally, association studies require a higher density of markers than linkage studies, but they identify candidate chromosome regions with a better resolution.

2.4.3 Nuclear families studies: evidence for *linkage* to chromosome 16 and 19.

A quantitative trait locus (QTL) genome-wide linkage analysis for SLI was undertaken by the SLI Consortium (SLIC) (SLIC, 2002). This study included 98 nuclear families, each with at least one child affected by SLI. The children’s language-related abilities were assessed using CELF-R and NWR tests, and their cognitive abilities using the Wechsler Scales of Intelligence (WISC) (as described in Materials and Methods). All probands had Expressive and Receptive Language Scores (ELS and RLS, obtained from CELF-R) >1.5 standard deviations (SD) below the normative mean for their chronological age and a Performance IQ (PIQ) >80.

The quantitative analyses were performed using two non-parametric sibling pair methods: the Haseman-Elston (HE) (Haseman and Elston, 1972) and Variance-Components (VC) methods (Pratt et al., 2000). HE is a linear regression-based method that assumes an inverse relation between the squared differences in the sib-pair trait scores and the proportion of IBD alleles shared at the loci close to genes influencing the trait (QTL). The VC linkage method separates the trait variance into three components: a major gene variance, a background polygenic variance, and a variance due to random environmental effects. This approach is different from the methods described above and estimates the relative contribution of these components and the likelihood of linkage between the phenotypic variability and shared IBD alleles at particular marker loci, under the hypothesis of a major gene (QTL) effect.

In this first SLIC study, two significant linkage regions were found: one on chromosome 16q (designated **SLI1**, OMIM 606711), linked to NWR trait (maximum LOD score of 3.55, HE analysis), and another on chromosome 19q (designated **SLI2**, OMIM 606712), linked to the ELS trait (maximum LOD score of 3.55, HE) (**Figure 2.1**).

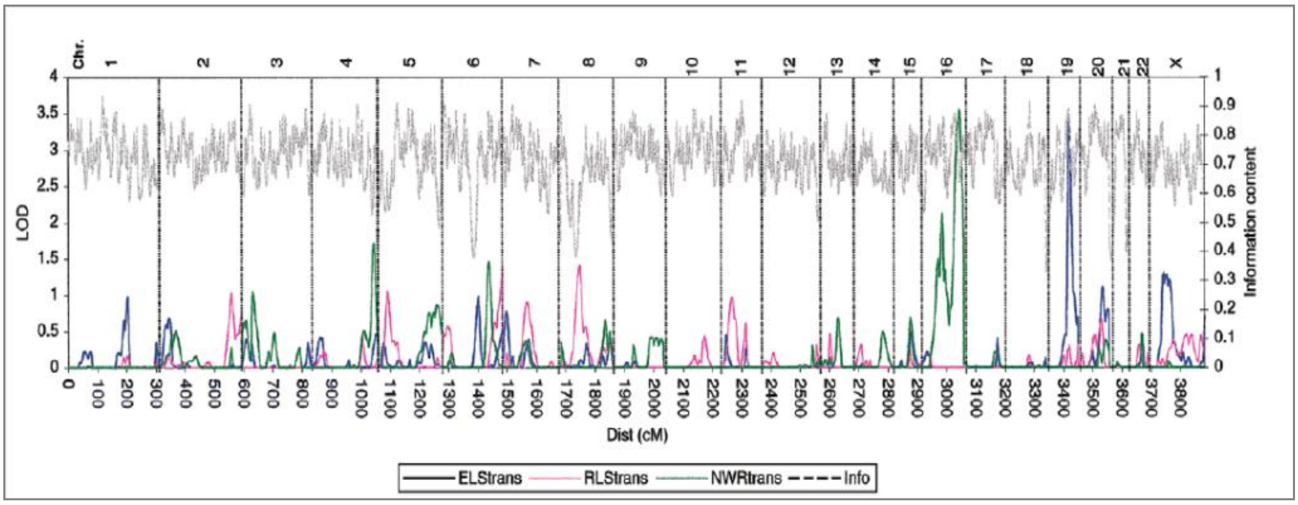


Figure 2.1 (SLIC, 2002). Genome-wide plot of HE linkage to three language-related measures (ELS, ELS; NWR) under multipoint analysis.

In 2004, after collecting 86 additional families, the SLIC performed another linkage analysis to further investigate the two regions previously identified (SLIC, 2004). In this cohort, that included also the previously collected families (n=98), giving a total number of 184 families, the linkage between chromosome 16 and NWR performance was confirmed, with a combined maximum LOD score of 7.46. The linkage on chromosome 19 also replicated, but the region was linked to different language-related measures: in first wave (the 2002 cohort) the linkage was found for ELS, in the second wave (the 2004 sample) for NWR. Taking into account both samples, the maximum LOD score on chromosome 19 with NWR was 1.4 (HE).

A study of an independent cohort of 93 nuclear families, collected through the Manchester Language Study, tried to replicate the SLIC linkage findings (Falcaro et al., 2008). The selected children had PIQ ≥80 and were mainly assessed for NWR and Past Tense marking trait (PT) task, but also ELS (CELF-R) data were available.

Linkage analyses were performed using two methods: the HE method and another method (Fulker et al., 1991), hereafter referred to as the “DF-linkage” method, which represents an extension of the classic DeFries–Fulker approach. The classical method, originally developed to estimate the heritability of traits from MZ and DZ twins and based on a regression model, was implemented to detect linkage regions in sib-pairs, using quantitative phenotypic measures and shared IBD as a measure of the genetic similarity between siblings. The DF-linkage method assumes that if a IBD value for a marker has a significant effect on the extent of regression toward the population mean, then there is evidence for linkage between the marker and a trait locus.

A weak linkage with NWR was found on chromosome 16q (maximum LOD score of 1.69, DF-linkage) while a significant linkage with ELS was identified on chromosome 19q (maximum LOD score of 5.8, DF-linkage).

Linkage to these regions was seen also to past tense phenotype. It has been suggested that, by a certain age, PT competence is either acquired or not acquired and it therefore should be analyzed as a qualitative trait (Bishop, 2005, 2014). In this study, two alternative hypotheses were tested: PT abilities were examined either on a continuous scale or as a binary trait (affected or unaffected). When PT was measured as a continuous trait, some linkage was found on both chromosome 16 (maximum LOD score of 1.8, DF-linkage) and chromosome 19 (maximum LOD score of 2.2, DF-linkage); considering PT as a binary trait instead, a linkage signal was detected only on chromosome 19 (maximum LOD score of 1.66, HE).

2.4.4 Extended pedigree studies: evidence for *linkage* to chromosome 13.

Linkage studies are based on the segregation of marker alleles with the disease phenotype, therefore extended pedigrees with multiple affected individuals are generally more informative than nuclear families, and have an increased power of identifying chromosomal regions linked with a disease. A linkage screen for SLI was carried out in five large Canadian families, two nuclear and three extended (Bartlett et al., 2002). Some of the individuals enrolled in this study were originally identified during a linkage study of schizophrenia, as they were reported to have a history of language or reading deficits (Brzustowicz et al., 2000). Although a diagnosis of SLI excludes the presence of other neurological disorders, the authors asserted that the low number of schizophrenic individuals included in the analysis (n=7) would not influence the identification of SLI susceptibility loci, considering that the largest pedigree (34 individuals) was connected to a schizophrenia family only by marriage.

From a total of 86 individuals from whom DNA was available, language phenotypes could be assessed for 73 subjects. Several tests were used, including an age-appropriate version of the language development test TOLD and reading subtests from the Woodcock Reading Mastery Test (single word and single non-word reading tests).

Each family included at least two SLI probands, defined as those with Spoken Language Quotient

Standard Score (SLQ, taken from the TOLD) of >1SD below that expected by age, PIQ of ≥80 and PIQ≥ SLQ. Three categorical diagnoses (not mutually exclusive) were derived for all family members: a) language impairment (SLQ ≤85); b) reading impairment (single non-word reading score of >1SD below PIQ); c) clinical impairment (history of language or reading difficulties).

A total of six parametric linkage analyses were performed: for each diagnostic category, two distinct models were tested, assuming either dominant or recessive inheritance. The highest LOD scores were also characterized with a Bayesian statistical approach, called Posterior Probability of Linkage (PPL) (Vieland, 1998; Vieland et al., 2001). PPL calculates a posterior probability of linkage between a marker and a trait gene, incorporating prior genomic information and prior probabilities (i.e. recombination fraction probability and linkage probability), that can be adjusted for the structure of the dataset in hand. This flexible method is particularly suited to the study of complex disorders, as it takes into account also the fact that the same susceptibility locus can show different modes of transmission among families. Moreover, it allows the accumulation of linkage evidence across many datasets. PPL values range from 0 to 1, but are commonly converted into percentages: values converging to 1 (or 100%) provide strong support to linkage peaks.

Significant evidence of linkage was found on chromosome 13q (max LOD=3.92) with a reading-based phenotype, under the recessive model of inheritance (**Figure 2.2**). The maximum PPL for this locus was 0.53, indicating a probability of 53% that a risk factor for SLI falls in this region. No linkage was detected in the two loci found by the SLIC, on chromosomes 16q (SLI1) or 19q (SLI2).

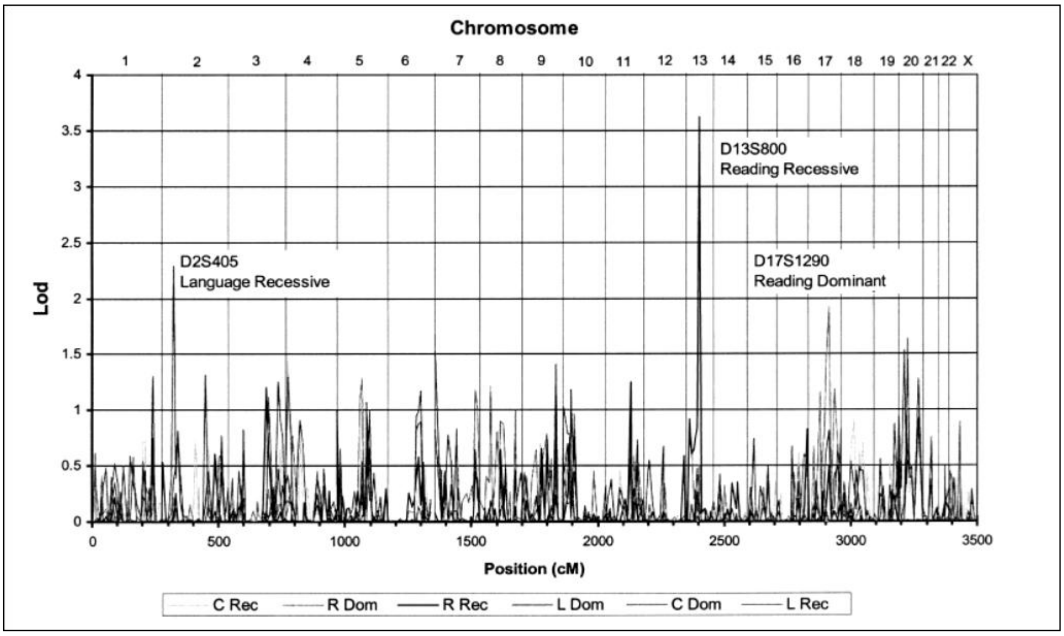


Figure 2.2 (Bartlett et al., 2002) Genome-wide plot of maximum two point heterogeneity LOD scores for all six models (R= reading discrepancy, C=clinical diagnosis, L= language impaired, Dom=dominant, Rec=recessive). The three highest peaks are labelled by marker and model tested.

In 2004, Bartlett et al. further investigated the region on chromosome 13q using a larger cohort, which included 22 additional nuclear and extended families (279 individuals from the United States), ascertained through a single proband (Bartlett et al., 2004). Assessment and proband designation were the same as described in Bartlett et al., 2002. In this second study, they also investigated a potential genetic overlap with autism, looking at two loci linked to autism, on chromosomes 2q and 7q, that will be discussed later (paragraph 2.6.3). The assessment tools, criteria for diagnosis and model parameters were the same used in their first study. The two sample sets (Canadian and US) were combined using alternative methods: the PPL method and the heterogeneity LOD score method (HLOD). Two HLOD variants were used: HLOD-P, which pools datasets to calculate one HLOD score, and HLOD-S, which calculates HLOD scores for each sample separately and then sums them across datasets.

The loci on chromosomes 2q and 7q did not provide conclusive evidence for linkage. Linkage was obtained for chromosome 13q, instead, in each separate sample set (US sample HLOD=2.616, PPL=16.8%; Canadian sample HLOD=3.565, PPL=54.2%) and when the sample were combined (HLOD-P=6.031; HLOD-S=6.181, PPL=92.3%), under the recessive reading impairment model. These analyses replicated and strengthened the findings previously reported for chromosome 13. Therefore, this region has been designated as **SLI3** (OMIM 607134).

2.4.5 Extended pedigrees from an isolated population: Robinson Crusoe island.

A linkage study involving extended pedigrees was carried out on an isolated Chilean population with an increased prevalence of SLI (known as TEL in Spanish-speaking countries) (Villanueva et al., 2011). This population inhabits the Robinson Crusoe Island, located 677 km west of Chile and belonging to the Juan Fernández archipelago. In the late 19th century, eight families repopulated the island and, as a result of its geographical isolation, the current population (633 residents, based on the 2002 Chilean census) shows a high degree of consanguinity (Villanueva et al., 2008). Given the relatively recent ancestors, a genealogical reconstruction was carried out and, interestingly, found that the vast majority of known affected individuals (84% of the individuals from whom DNA was available) descended from a single pair of founder brothers (Villanueva et al., 2011) (**Figure 2.3**).

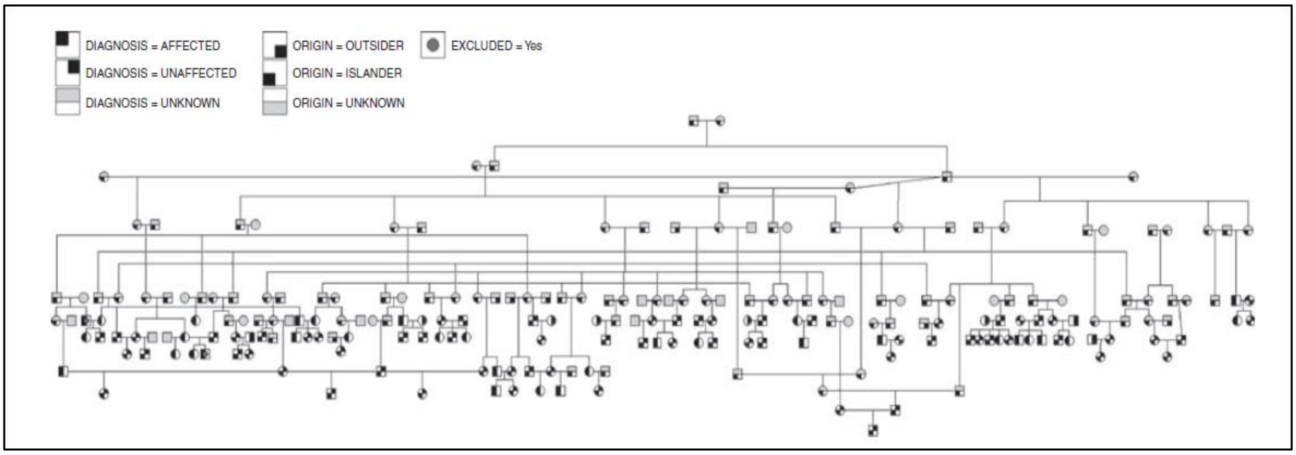


Figure 2.3 (Villanueva et al., 2011). Pedigree structure of the descendants of a single pair of founder brothers.

Island "colonising" children (i.e. related to one of the founder families) show a very high incidence of SLI (35%), compared to the frequency of SLI among the non-colonising children (3.8%), which is almost the same as that reported in mainland Chile (4%) (Villanueva et al., 2008).

Therefore, due to its derivation from small group of relatively recent founders, to its decreased genetic heterogeneity and the common environment, this Chilean founder population represents a powerful resource for the identification of genetic factors contributing to susceptibility to SLI. In particular, for complex disorders like SLI, these founder populations may be extremely useful for the identification of rare monogenic forms of the disease.

Non-parametric and parametric (assuming dominant or recessive mode of inheritance, a frequency of 35% and full penetrance) linkage analyses were performed on "colonizing" families. Five regions (on chromosomes 6, 7, 12, 13 and 17) yielded genome-wide significant linkage in non-parametric analyses. In this study, linkage was not observed for the previously implicated loci on chromosomes 16 (SLI1) or 19 (SLI2), but the region identified on chromosome 13 was close to SLI3. Across the different analyses, the most consistently linked locus (max NPL=6.73, P=4x10^{-11}) was a 48 Mb region on chromosome 7q, which overlaps with a region that has been linked to autism (*AUTS1* locus, OMIM 209850) and encompasses several interesting candidate genes, including *FOXP2* and *CNTNAP2*, both implicated in language development and discussed in more detail below.

Further studies will be required for a fine mapping of the relevant regions, in order to identify the genes contributing to the increased frequency of language impairment on the island.

2.4.6 A comparison between the different linkage studies for SLI.

None of the regions of linkage described above was replicated across all these studies. In addition to locus heterogeneity, which is an important component of complex disorders, several factors may have influenced the lack of overlap of the results across the three cohorts. The most striking differences regard the design strategy and the phenotype definition.

SLIC used non-parametric analyses and a cohort of nuclear families, the Bartlett group applied a parametric approach to a small number of large families, and the Robinson Crusoe study used both parametric and non-parametric methods with an extremely large pedigree deriving from an isolated population. Each linkage approach presents advantages and disadvantages. Non-parametric methods offer the advantage of not requiring the specification of a model and for this reason are more suitable to complex diseases with unknown mode of inheritance. Parametric analyses instead need the specification of a model, but if the parameters are correct, they provide increased power. They can be seriously affected by locus heterogeneity and mis-specification of parameters can lead to false positives. However, previous studies (Abreu et al., 1999; Greenberg et al., 1998) have shown that, if both mode of inheritance (dominant and recessive) are tested, parametric analyses can have sufficient power to detect genomic regions linked to complex disorders. Moreover, they used families with multiple affected members, that are more informative compared to small nuclear families with one affected member and are expected to have a reduced genetic heterogeneity. This is particularly true for large and multigenerational pedigrees.

Second, there are differences in the way the SLI phenotypes were measured and analysed. The absence of a standard definition for a diagnosis of SLI leads to differences in the classification of affected and unaffected individuals. The studies of extended pedigrees used binary categories of language impairment by setting an arbitrary threshold for affection. When the threshold is appropriate, this approach can increase the power and reduce the heterogeneity. SLIC instead considered language abilities along a continuous scale, using standard scores calibrated against the general population, and this allowed a quantitative analysis of three highly heritable traits (ELS, RLS and NWR).

The loci found by the three groups were linked to different phenotypes or sub-phenotypes:

- the region on chromosomes 16 was linked to NWR (Falcaro et al., 2008; SLIC, 2002);
- the region on chromosome 19 was linked to ELS (Falcaro et al., 2008; SLIC, 2002), to NWR (SLIC, 2004) and to PT (Falcaro et al., 2008);
- the region on chromosome 7 was linked to language impairment (Villanueva et al., 2011);
- the locus on chromosome 13 was linked to reading impairment in a language impaired sample (Bartlett et al., 2002). Reading impairment was not interpreted as a dyslexia phenotype, but as a language-related trait, considering that impaired language development can also lead to difficulties in reading skills (Flax et al., 2003).

Co-morbidity between reading and language abilities has been examined also by two recent GWA studies (Eicher et al., 2013; Luciano et al., 2013): association signals with both language and reading skills have been found for variants in the genes *ZNF385D* (chromosome 3p24.3), *DAZAP1* (chromosome 19p13.3), *CDC2L1*, *CDC2L2* and *RCAN3* (located on chromosome 1), while association with specific language traits has been found for variants in the genes *ABCC13* (chromosome 21q11) and *NDST4* (chromosome 4q26).

The investigation of alternative traits (“endophenotypes”) can be useful in dissecting the genetic bases underpinning a disorder like SLI, however the results of these kind of studies frequently present lack of overlap. This might indicate that the quantitative/qualitative traits chosen to evaluate language skills might be individually too restrictive to capture the complex scenario of the molecular mechanisms underlying language impairment. Complex disorders like SLI, characterized by genetic heterogeneity, are estimated to involve numerous genes and each of them might influence multiple phenotypic aspects.

2.4.7 Targeted association studies of chromosome 16.

In order to investigate the linkage of NWR trait to the SLI1 locus, a high density association screen of SLI across this region of linkage on chromosome 16q (~10 Mb) was carried out by the SLIC Consortium (Newbury et al., 2009). In this study, including 211 SLIC families, a family-based quantitative association analysis (known as QTDT) and a categorical case-control analysis were performed. The QTDT is a linkage disequilibrium test able to detect association with quantitative traits in nuclear families through a regression-based approach, which tests the correlation between a continuous trait measure and the number of alleles of a given marker carried by a child. This model takes into account variance within- and between-families. The case-control analysis instead, was carried out using the NWR trait as a binary measure: individuals with low NWR (>2 SD below the SLIC cohort mean, n=79) were defined as cases, whereas family members with above-average NWR performance (>0.5 SD above population mean, n=71) were selected as controls. To obtain a group of unrelated cases and controls, only one case or one control was selected from each family.

Strong association signals were found for two clusters of SNPs, 3Mb apart: one falling in the *CMIP* gene (c-Maf inducing protein, minP=5×10^{-7}), between exon 2 and exon 5, and the other one in the *ATP2C2* gene (ATPase, Ca^{2+} transporting, type 2C, member 2, minP=2×10^{-5}), between exon 7 and 12. Both genes are expressed in the brain and they are described later in more detail. The associations with *CMIP* and *ATP2C2* were reported to be independent, suggesting that both these genes could separately contribute to SLI susceptibility.

Both associations were followed up in a replication sample, selected from the population-based cohort “Avon Longitudinal Study of Parents and Children” (ALSPAC) (Jones et al., 2000). The children enrolled in this long-term project have been periodically checked from the age of 7 years, and a series of physical, behavioural and neuropsychological traits has been assessed, including language and reading development. From that larger group, only children with low language measures were selected (490 cases). In this second stage, the same association methods were applied, but, given the differences observed in the distributions of NWR between SLIC and ALSPAC, different case/control cut-offs were used in the case-control analyses. Significant associations were found for two markers in *ATP2C2* (minP=0.0058), that replicated the trend observed by SLIC. Regarding *CMIP*, two SNPs showed a significant association with NWR (minP=0.0182), but the genotype trends were in the opposite direction from SLIC (the genotypes associated with high NWR scores in SLIC were associated with low NWR scores in the replication cohort). Although this contrasting result may indicate a false positive, another possible explanation could be the differences in the relationship between the markers and the causal variant in the two samples (Lin et al., 2007).

The relationship between *ATP2C2* and *CMIP* markers and NWR performance was also investigated at a population level, using the entire ALSPAC cohort (n=3612), but no evidence for association emerged from this analysis, leading to the hypothesis that variants in these two genes might affect NWR only in language-impaired individuals.

CMIP and *ATP2C2* were also included among the language candidate genes in a targeted association study investigating the potential genetic overlap between SLI and dyslexia (Newbury et al., 2011). These two neurodevelopmental disorders show an extensive co-morbidity, therefore it is plausible that they might share some aetiological factors. The study focused on a set of known candidate genes for SLI and dyslexia and was performed on the SLIC cohort and two dyslexia samples. Quantitative analyses were carried out for several language and reading scores. Although the dyslexia samples did not yield significant association for any of the SLI loci, multiple SNPs in *CMIP* showed significant association with both language and reading impairments in the SLIC cohort. By contrast, the association of *ATP2C2* instead appeared to be specific to language measures (ELS, RLS, NWR).

Moreover, another study performed on the ALSPAC cohort (Scerri et al., 2011), found association between SNPs in *CMIP* and general reading skills, in particular for single word reading and single word spelling performance. The data showed that the association was not driven by reading-impaired individuals, in accord with the results obtained for dyslexia (Newbury et al., 2011). The allelic trend of the associated SNPs was consistent with the one reported for the ALSPAC language-impaired subgroup (Newbury et al., 2009). Again, association with reading abilities was not reported for *ATP2C2*.

These findings suggest that *CMIP* could contribute to normal reading variation and represent a modifier locus for language, whose effects on the phenotype might be determined by the presence of other variants, and thus it might affect both language- and reading-related processes in certain genetic backgrounds.

2.5 SLI candidate genes *ATP2C2* and *CMIP*.

CMIP (OMIM 610112) encodes a C-MAF inducing protein and it is expressed in several neuronal cells (Nagase et al., 2000). It is an adapter protein known to interact with filamin A (Grimbert et al., 2004), an actin-binding protein involved in the reorganization of the cytoskeleton during cell shape changes and migration, with RelA subunit, an anti-apoptotic factor belonging to the NF-κB family (Kamal et al., 2009), a family of transcriptional factors important for the regulation of processes associated with synaptic activity and plasticity and neurodegeneration, and with the PI3 kinase complex (Kamal et al., 2010), playing a role in the ERK signalling cascade. This evidence indicates that *CMIP* may be involved in multiple biological pathways.

ATP2C2 (OMIM 613082) encodes an ATPase (type 2C, member 2), also known as SPCA2 (*secretory pathway calcium ATPase*), that transports Ca^{2+} and Mn^{2+} into the Golgi, but is also able to interact with Ca^{2+} channels on the cell surface, eliciting entry of Ca^{2+} (Feng and Rao, 2013). Many signalling pathways use calcium as a messenger, and its homeostasis is crucial to various neuronal functions and processes, including working memory. The activity of several kinases and phosphatases required for working memory is dependent on calcium levels (Dash et al., 2007). The homeostasis of manganese ions is also tightly regulated (Tuschl et al., 2013) and a number of proteins (such as SPCA1 and SPCA2, divalent metal transporter 1, the ZIP family metal transporters and others) are suggested to be involved, however the degree of their specific contribution has still to be determined.

Although little is known about the function of the candidate genes *CMIP* and *ATP2C2* in the brain, their functions and several findings are in favour of a potential role of these genes in SLI and neurodevelopmental disorders presenting co-morbidity with SLI, such as dyslexia (as discussed before), ASD, and Attention Deficit-Hyperactivity Disorder (ADHD), as presented below.

A *de novo* deletion on chromosome 16, involving the two genes *GAN* and *CMIP*, was identified in an autistic child with severe receptive and expressive language deficits. Since mutations disrupting *GAN* cause giant axonal neuropathy, the authors hypothesized that haploinsufficiency of *CMIP* was more likely to be responsible for the ASD phenotype (Van der Aa et al., 2012).

An interesting finding regarding *ATP2C2* comes from a study of ADHD. The most recent GWA studies for ADHD (Mick et al., 2010; Neale et al., 2010a; Neale et al., 2010b; Stergiakouli et al., 2012) have failed to detect association signals reaching genome-wide significance, however signals with P values close to the threshold might indicate potentially contributing loci. In one of these studies (Lesch et al., 2008), the list of the top 30 markers located in genic regions revealed a sub-threshold signal of association within *ATP2C2* (P=8 x 10^{-7}). Interestingly, children with ADHD showing co-morbidity with SLI display reduced performance in working memory tasks, that have been shown to correlate more closely with language deficits rather than ADHD (Cohen et al., 2000; Jonsdottir et al., 2005). Considering that the SLIC studies did not exclude individuals predicted to have also reading problems, ADHD or developmental coordination disorder (representing ~1/3 of their samples) and considering that, in SLI, *ATP2C2* was found in association with NWR (Newbury et al., 2009), which is a measure of working memory, this domain has been proposed to be an ”overlapping zone” between ADHD and SLI.

Molecular mechanisms through which variants in *ATP2C2* might affect the storage and processing of verbal information are still unknown, however these findings support the idea that *ATP2C2* may be involved in neurological processes important for phonological short-term memory and may be relevant to developmental disorders characterized by working memory impairments, such as SLI and ADHD.

2.6 Co-morbidity of SLI with Autism Spectrum Disorders (ASD).

2.6.1 Autism Spectrum Disorders.

According to the fifth edition of the Diagnostic and Statistical Manual of Mental disorders (DSM-V) (APA, 2013), Autism Spectrum Disorders (ASD), also known as Pervasive Developmental Disorders (PDD), indicate an umbrella of childhood disorders that are characterized by impairments in two core domains:

1) social communication and social interaction;

2) restricted and repetitive behaviours and interests.

This diagnostic category includes:

- Autism, which presents deficits in communication, social interactions and repetitive behaviours;
- Asperger’s disorder, which is characterized by the absence of clinically significant delay in language and cognitive development;
- Childhood Disintegrative Disorder (CDD), which typically occurs later than autism and involves a more dramatic loss of skills (regression);
- Pervasive Development Disorder-not otherwise specified (PDD-NOS), which presents sub-threshold symptoms and/or later onset.

These neurodevelopmental conditions differ in the severity and the pattern of the core symptoms, developmental course, and cognitive and language abilities. The ASD have an estimated prevalence of ~60/10,000 individuals (Elsabbagh et al., 2012; Levy et al., 2009) and a male to female gender bias, with a ratio of ~4:1 for classical autism and higher ratios for ASD (Williams et al., 2008).

The importance of genetic factors in these disorders emerged from family and twin studies. Several studies have shown that ASD recur in families and siblings of affected probands have a higher prevalence of ASD compared to the general population (~25 times higher, according to the estimates of the most recent studies) (Constantino et al., 2010; Ozonoff et al., 2011). Twin studies have indicated that ASD have a high heritability (generally >80%) (Ronald and Hoekstra, 2011), except for one recent study, which reported only a modest effect for genetics (37%) (Hallmayer et al., 2011). However, concordance rates between MZ twins do not take into account the genetic factors that may differ in co-twins, such as epigenetic factors, X-inactivation and mutations *de novo* arisen after the separation of the embryos.

In support of the hypothesis of a strong genetic background, family studies have shown that siblings of autistic probands display a higher recurrence of features typical of ASD phenotypes compared with the general population. These mild forms of impairments, usually affecting only one of the core domains, are classified as “broader phenotypes”.

2.6.2 Phenotypic overlap between SLI and ASD.

In autism, which is the most severe form of ASD, verbal communication is usually abnormal, but the language profiles can be extremely varied. About 50% of autistic children do not develop any verbal language or show a marked delay in the development of spoken language (Hus et al., 2007). Autism presents also a high variation in cognitive skills, therefore low IQ scores may influence the most severe forms of language impairments.

The most frequent linguistic deficit in autistic individuals is an inappropriate use of language in social contexts (pragmatic domain). However, structural aspects (phonology, vocabulary and syntax) can also be affected, in a way that resembles SLI. Subgroups of high level functioning autistic children showed profiles similar to children with SLI on tests of phonological processing, vocabulary and higher order grammatical skills (Kjelgaard and Tager-Flusberg, 2001).

In contrast, individuals affected by SLI are mainly characterized by structural language difficulties. However, a subgroup of children with SLI have been reported to display significant difficulties also in social and communication domains (Leyfer et al., 2008). These children show more problems in socializing with their peers and processing social-affective information compared to controls. Moreover, studies of adolescents with a documented history of SLI indicated that a minority of them meet standard diagnostic criteria for autism or present with behaviours reminiscent of autism (Conti-Ramsden et al., 2006; Howlin et al., 2000; Mawhood et al., 2000).

Converging evidence of potentially shared mechanisms emerged also from family members of children affected by these disorders: first and second-degree relatives of autistic patients display a higher prevalence of language impairment than the general population, *vice versa* siblings of children with SLI present a higher risk of a diagnosis of autism compared to the general population estimates (Tomblin et al., 2003).

However, further investigations on the areas of potential phenotypic overlap between SLI and autism have provided contrasting results. Studies comparing nonword and sentence repetition tasks in children with SLI and autistic children with structural language difficulties have reported poor performances in both groups, but with different patterns of errors, that may indicate distinct underlying cognitive deficits (Riches et al., 2010; Whitehouse et al., 2008).

Therefore, the similarities and the differences in socialization and language domains have stirred a debate regarding the potential overlap between SLI and autism, leading to two alternative hypotheses: one argues that some genetic susceptibility factors might be shared by the two disorders, the other one instead argues that the similarities are superficial and different patterns of language deficits reflect alternative distinct causes.

2.6.3 Genetic overlaps between autism and SLI.

Several whole-genome linkage studies have been performed for autism and, although many loci have been implicated, replicated regions between samples are rare, reflecting the extensive heterogeneity underlying the disorder and the likely small effect size attributable to single genes. The first regions linked to autism were identified by the International Molecular Genetic Study of Autism Consortium (IMGSAC, 1998, 2001; Maestrini et al., 2010) on chromosomes 7q (designated AUTS1, 7q21-q32, OMIM 209850) and 2q (designated AUTS5, 2q24-q33, OMIM 606053) and these have been also the most consistently replicated loci (Badner and Gershon, 2002; Buxbaum et al., 2001; Schellenberg et al., 2006; Shao et al., 2002; Trikalinos et al., 2006).

A possible strategy to increase the chances of identifying contributory risk genes in a context of high heterogeneity is the study of single “endophenotypes”. The observation of broader phenotypes in family members of autistic individuals led to the hypothesis that autism could be dissected into three heritable, potentially distinct, core components (social interaction, language, and repetitive behaviour), or “endophenotypes”. Some linkage studies for autism have focused on language-related endophenotypes.

One of these linkage screens (Bradford et al., 2001) found linkage to chromosomes 7q and 13q, the latter overlapping with that identified in SLI families (Bartlett et al., 2002). Both signals appeared to be attributable to the families with ASD probands with Phrase Speech Delay (PSD) beyond 36 months of age, and parents with a history of language difficulties. In another ASD linkage study (Alarcón et al., 2002), the samples were stratified according to three endophenotypes obtained from the Autism Diagnostic Interview-Revised (ADI-R): "age at first word", "age at first phrase" and "repetitive and stereotyped behaviour". The strongest evidence of linkage was obtained for "age at first word", on a region on chromosome 7q, close to the susceptibility locus reported by IMGSAC (IMGSAC, 1998).

The other main investigated linkage locus (IMGSAC, 2001), on chromosome 2q, has also been studied in relation to language delay. Two studies (Buxbaum et al., 2001; Shao et al., 2002) found that linkage in this region was strongest when the analysis was restricted to families which included autistic children with PSD. Although the region on chromosome 2q has not been implicated in SLI studies (Bartlett et al., 2004), the evidence obtained from the ASD studies, however, seem to suggest that this locus may harbour risk factors for language development.

The SLI2 locus on chromosome 19q (SLIC, 2002) overlaps with a region of suggestive linkage with autism (Liu et al., 2001). In that study, the analyses were performed using two phenotypic categories: a strictly-defined group of autistic families, in which the probands had to meet diagnostic criteria for autism in all three core domains and an age at onset of <3 years (narrow category), and a broader category, including also individuals affected by Asperger’s disorder or other PDD. The linkage on chromosome 19q was found to be driven by the group of strict autism. One would expect that the greatest overlap with SLI would be found in the less severely affected individuals. However, in the broader group there were also patients with Asperger’s disorder, which is characterized by relatively high linguistic capabilities. Thus, it remains possible that variants that contribute to susceptibility to SLI and autism may be found on chromosome 19.

More recently, linkage studies with higher resolution and larger cohorts of ASD families (Szatmari et al., 2007; Weiss et al., 2009) have shown that the linkage regions described above failed to reach genome-wide significance. In the study realized by the Autism Genome Project (Szatmari et al., 2007), that included 1,168 families with at least two affected individuals, suggestive linkage was obtained only for a region on chromosome 11p12-p13. The samples were also stratified in categories, but, even in these subsets, the loci on 7q and 2q reached only suggestive evidence of linkage in the individuals of European ancestry, confirming the genetic heterogeneity underlying the ASD and hindering the identification of genetic cause.

A novel design strategy was used in a recent linkage study in order to specifically investigate the overlap between autism and SLI (Bartlett et al., 2013). A genome-wide analysis was performed in 70 families with at least one person with ASD and at least one person with SLI, described in a previous study (Bartlett et al., 2012). Such pedigrees were recruited in order to increase the chance of identifying loci relevant to both disorders. Moreover, the disorders were in distinct individuals and not co-morbid, allowing the examination of whether linkage signals are driven specifically by language, ASD or both. The language phenotypes, assessed in all family members, were classified in two categories: LI, which included individuals with oral language impairment or ASD, and RI, which included individuals with written language impairment (reading) or ASD. LI yielded evidence of linkage to a region of 24.2 Mb on chromosome 15q13-16.2 (maximum PPL= 0.57), whereas RI was found to be linked to a region which spans 8.9 Mb on chromosome 16p12.1-12.3 (maximum PPL= 0.36). Both loci were not linked to nonverbal IQ, suggesting these signals were not influenced by cognitive impairment. Moreover, the exclusion of the autistic or the language-impaired individuals reduced the PPL scores for both regions, indicating that both groups contributed to these linkage peaks. The genome-wide association analysis instead did not detect any strong signal, but this may be due to the regions of linkage were not adequately tagged or to the insufficient power to detect variants of small effect because of the small sample size.

Genome-wide association studies in large cohorts of autistic individuals (Anney et al., 2012; Anney et al., 2010; Wang et al., 2009; Weiss et al., 2009) have been carried out, but the common variants identified explain only a small fraction of the genetic risk. On the other hand, recent findings suggest that *de novo* and rare inherited variants of intermediate-high penetrance could collectively account for a large proportion of risk (Devlin and Scherer, 2012), according to the “common disease-rare variant” hypothesis (see paragraph 1.4). Recent high-throughput CNV screenings and exome-sequencing studies, that apply NGS technologies to sequence the coding regions of the genome, have begun to uncover a large number of individually rare sequence mutations and structural rearrangements potentially contributing to the ASD susceptibility (Devlin and Scherer, 2012). Based on these findings, hundreds of risk genes in autism are estimated to be implicated in these disorders. The next challenge will be to determine which of the potentially deleterious variants actually play a role in the disease: investigating whether they affect genes involved in interconnected pathways could narrow down the list of candidate genes and clarify the molecular mechanisms impaired in ASD.

Exome-sequencing or CNV studies for SLI have not been published yet, but these kind of studies will help to elucidate the genetic architecture of SLI, establishing the contribution of rare sequence variants and structural rearrangements and determining the genes involved in the disorder. Network analyses might provide new evidence of potential genetic overlaps between SLI and other neurodevelopmental conditions, like ASD.

2.7 The gene *CNTNAP2:* an example of a functional link between neurodevelopmental disorders.

Dysfunctions of genes involved in processes that are crucial for several aspects of brain development can have an effect on a range of neurological functions. These genes may explain shared or related genetic mechanisms present in related neurodevelopmental disorders, such as SLI, autism and others. Evidence for this hypothesis is provided by studies of biological pathways mediated by *FOXP2* (Fisher and Scharff, 2009).

The identification of *FOXP2* derives from a three generation pedigree, known as the KE family, with a rare monogenic form of a severe speech and language disorder. A point mutation that alters an invariant amino-acid residue (p.R553H, NM_014491.3) in the forkhead domain was detected in all affected members of the KE family (Lai et al., 2001). Subsequently, other cases carrying damaging mutations in *FOXP2* have been reported (Feuk et al., 2006b; Lennon et al., 2007; MacDermot et al., 2005; Shriberg et al., 2006; Zeesman et al., 2006), suggesting an important role in language development. This highly conserved gene maps on chromosome 7q31 and encodes a transcription factor protein that contains a polyglutamine tract and a forkhead DNA-binding domain. Chromatin-Immunoprecipitation (ChIP) experiments and expression analyses have found hundreds of potential FOXP2-target genes in neuronal cells, mouse models and human developing brain (Spiteri et al., 2007; Vernes et al., 2011; Vernes et al., 2007) and it seems that, in most cases, the protein acts as a repressor.

A well characterized FOXP2-target is the *CNTNAP2* gene (OMIM 604569), which is one of the largest genes in the human genome (~2.3 Mb) and is located at chromosome 7q35-36.1 (**Figure 2.4 a**). FOXP2 binds a sequence within intron 1 of *CNTNAP2* and negatively regulates this gene. Levels of *CNTNAP2* are indeed lowest at high *FOXP2* levels and *vice versa* (**Figure 2.4 b**) (Vernes et al., 2008). *CNTNAP2* codes for the *contactin-associated protein-like 2*, also known as CASPR2, a transmembrane adhesion protein that belongs to the neurexin family. The classical members of this family are on the pre-synaptic side and interact with neuroligins, on the post-synaptic side, to establish synaptic connections (Craig and Kang, 2007). The proteins of the CNTNAP family are non-classical neurexins involved in neuron-glia interactions and clustering of K^+ channels in myelinated axons (Poliak et al., 1999; Poliak et al., 2003). For the correct localization of these channels, CNTNAP2 interacts with contactin 2 (CNTN2) (Poliak et al., 2003). Interestingly, in the human fetal brain, an unusual enrichment of *CNTNAP2* levels was observed in regions important for language e.g., the perisylvian cortex (Abrahams et al., 2007), indicating that *CNTNAP2* could also represent a good candidate gene for language-related phenotypes.

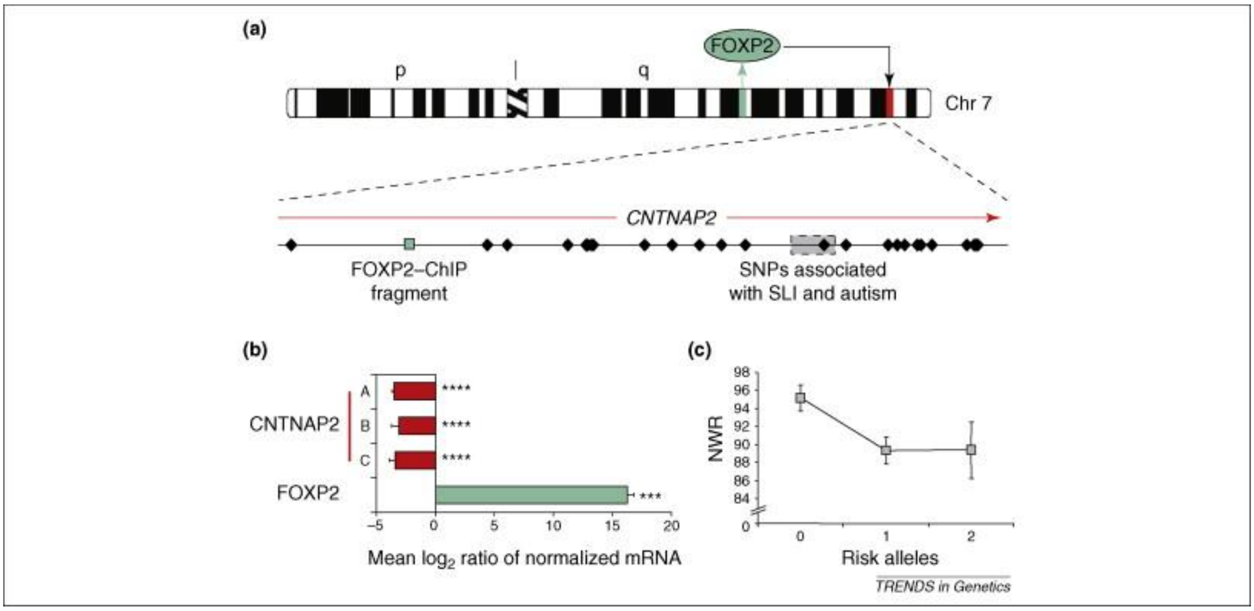


Figure 2.4 (Fisher and Scharff, 2009). Functional genetic bridges between distinct language disorders. (**a**) Locations of *FOXP2* on 7q31 and *CNTNAP2* on 7q35-36.1 are indicated. The genomic organization of *CNTNAP2* is given below. The red arrow indicates direction of transcription, diamonds represent exons and the square shows the FOXP2-bound region, mapping in intron 1. (**b**) mRNA expression in human neuron-like cells stably transfected with *FOXP2*. Levels of *CNTNAP2* mRNA (primers A-C) were inversely proportional to that of *FOXP2* (****=p<0.0001, ***=p<0.001). (**c**) The most common multimarker haplotype (ht1) for the nine SNPs (between exons 13–15) associated with deficits in NWR, negatively influenced NWR performance. When children were divided into three groups based on the numbers of carried copies of ht1, it was found that mean NWR dropped by ~6 points (~0.4SDs) as a consequence of carrying >0 risk alleles. Error bars represent standard errors.

Investigations of the potential implication of variants in *FOXP2* in complex forms of language impairment have shown that this gene is unlikely to have a direct contribution to SLI susceptibility (Meaburn et al., 2002; Newbury et al., 2002; O'Brien et al., 2003).

Common variants in *CNTNAP2* instead have shown association with language impairment. A QTDT analysis, performed on 184 SLIC families, tested the association of 38 SNPs across *CNTNAP2* with ELS, RLS and NWR measures (Vernes et al., 2008). Several markers yielded significant evidence of association, primarily to NWR (minP=$5×10^{-5}$, rs17236239) and, to a lesser extent, to RLS (minP=0.003, rs4431523) and ELS (minP=0.008, rs17236239). Interestingly, all nine SNPs associated with NWR fall in the region between exons 13 and 15. The same region (exons 13-15) has been found in association with the trait "age at first word" in an autism cohort of multiplex families (minP=0.002, rs2710102) (Alarcón et al., 2008) and the endophenotype of "early language development" (minP=0.0239, rs2710102) in a population-based cohort (the Raine sample) (Whitehouse et al., 2011). Moreover, the SNP rs2710102 was associated with NWR (p=0.0174) in a dyslexia family cohort, although the signal was driven by the opposite allele (Peter et al., 2011). In a study discussed earlier (paragraph 2.4.7) (Newbury et al., 2011), *CNTNAP2* did not show significant association in the dyslexia cohorts, whereas in the SLIC families showed a strengthened signal for NWR (minP=8 x 10^{-5}, rs17236239) and yielded association also with reading-related traits. Thus, these common variants, likely to be in linkage disequilibrium with the causal variants, suggest that *CNTNAP2* may modulate language abilities, but also may influence reading skills.

Further support for a role in language and cognitive development is provided by rare variants. A homozygous recessive frameshift mutation (3709delG) in *CNTNAP2* was identified in individuals from an isolated population (Old Order Amish), affected by a rare syndrome associated with ASD and characterized by language regression and abnormalities of neuronal migration, called Cortical Dysplasia-Focal Epilepsy syndrome (CDFE) (Strauss et al., 2006). A mutational screening for *CNTNAP2* detected 13 rare non-synonymous changes among 635 non-syndromic ASD patients (Bakkaloglu et al., 2008). Eight of them were predicted to be deleterious or altered highly conserved residues. However, these variants were inherited from an apparently unaffected parent, indicating incomplete penetrance. This suggests that some alterations of *CNTNAP2* may be required to occur in conjunction with mutations in other genes to result in neurological disorders ("multiple hit" model (Leblond et al., 2012)). Under this hypothesis, the specific outcome is determined by the nature of the mutation, the molecular pathways affected and the genetic and environmental background of subjects.

An example supporting this hypothesis is provided by a recent exome-sequencing study, in which a rare mutation (p.H275A) in *CNTNAP2* was identified in an autistic proband (O'Roak et al., 2011). This missense mutation, predicted to be deleterious, was present in the proband and an unaffected sister, inherited from the mother. In addition to the *CNTNAP2* change, the proband, who presented severe ASD, language delay and moderate intellectual disability, carried a *de novo* frameshift mutation in *FOXP1* (p.A339SfsX4), which leads to a truncated protein. The role of *FOXP1* (OMIM 605515) in neurodevelopmental disorders, including ID, ASD, language disorders and motor development delay, has recently begun to be elucidated (Bacon and Rappold, 2012). The gene is known to be closely related to *FOXP2* and the phenotypic spectra of their mutations indicate that they can operate in both different and shared pathways. Like FOXP2, FOXP1 downregulates *CNTNAP2*: in the presence of the truncated form of FOXP1, levels of *CNTNAP2* were shown to be increased (O'Roak et al., 2011).

In addition to sequence variants, complex rearrangements and CNVs involving *CNTNAP2*, in particular deletions, have also been reported across several neurodevelopmental conditions, including autism (Bakkaloglu et al., 2008; Poot et al., 2010), stuttering (Petrin et al., 2010), ADHD (Elia et al., 2010), Tourette's syndrome (Verkerk et al., 2003), schizophrenia and epilepsy (Friedman et al., 2008) and mental retardation (Zweier et al., 2009).

In conclusion, a wide set of heterogeneous mutations in *CNTNAP2* can be found in numerous and variable conditions, that present a certain degree of co-morbidity, and indicate a widespread effect of *CNTNAP2*. As demonstrated also by brain imaging studies and animal models (such as songbirds and mice) (Peñagarikano and Geschwind, 2012), this gene has a pivotal role in neurodevelopment, in particular in frontal-striatal brain circuits, and alterations of its function may affect a variety of processes, leading to distinct but overlapping phenotypes. Therefore, these findings support the idea of shared and/or intersected neurogenetic pathways converging on common genes, like *CNTNAP2.*

The falling cost of high-throughput sequencing technologies is expected to facilitate the identification of similar candidate genes.

Chapter 3

DPYD: a candidate gene for neurodevelopmental disorders

3.1. The enzyme dihydropyrimidine dehydrogenase (DPD).

The human *DPYD* gene (OMIM 612779) maps to chromosome 1p21.3 (GRCh37/hg19) and codes for the enzyme dihydropyrimidine dehydrogenase (DPD), the initial and rate-limiting factor in uracil and thymine catabolism. The pyrimidine degradation pathway consists of three consecutive steps (**Figure 3.1**):

1) Step 1: DPD catalyses the NADPH-dependent reduction of uracil to 5,6-dihydrouracil and of thymine to 5,6-dihydrothymine;
2) Step 2: dihydropyrimidinase (DHP, encoded by the gene *DPYS*) catalyses the hydrolysis of 5,6-dihydrouracil to N-carbamyl-β-alanine and of 5,6-dihydrothymine to N-carbamyl-β-aminoisobutyric acid;
3) Step 3: the third reaction is catalysed by the β-ureidopropionase (BUP-1, encoded by the gene *UPB1*), that converts N-carbamyl-β-alanine (also known as β-ureidopropionate) into β-alanine and N-carbamyl-β-aminoisobutyric acid (also known as β-ureidoisobutyrrate) into β-aminoisobutyric acid (β-AIB), producing also ammonia and CO_2.

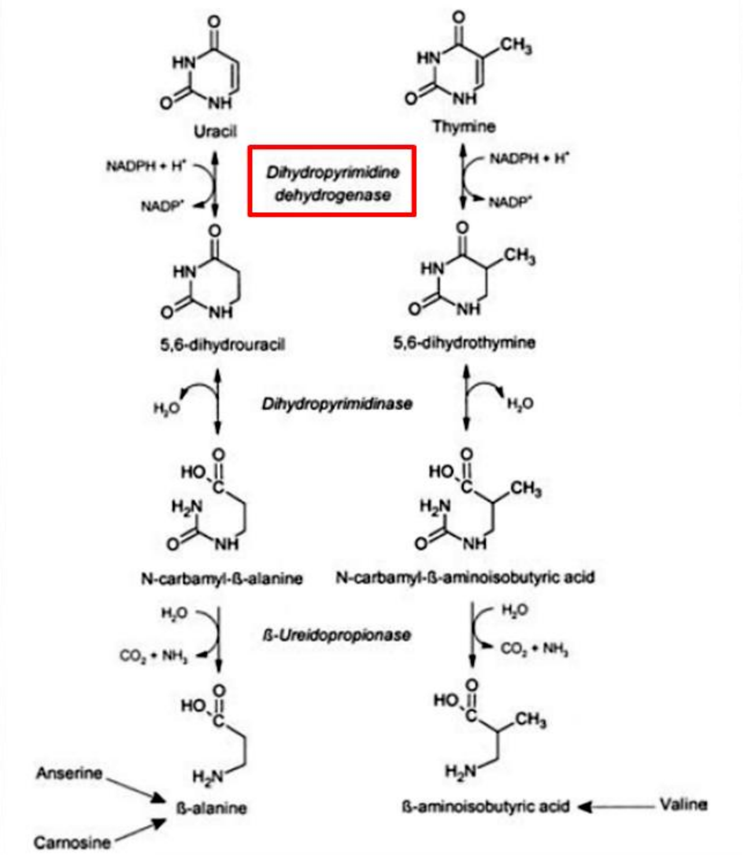


Figure 3.1 (Van Kuilenburg et al., 2004). Catabolic pathway of the pyrimidines uracil and thymine.

3.2 Dihydropyrimidine dehydrogenase deficiency.

The deficiency of dihydropyrimidine dehydrogenase (OMIM#274270) causes an autosomic recessive disease caused by homozygous or heterozygous-compound mutations in the *DPYD* gene. A total absence of DPD activity causes a large accumulation of uracil and thymine (*thymine-uraciluria*) in blood, urine and cerebrospinal liquid. The phenotypic outcomes of the disorder are extremely variable, ranging from asymptomatic conditions to neurological abnormalities (van Kuilenburg et al., 2002; Van Kuilenburg et al., 1999).

The first case of deficiency of dihydropyrimidine dehydrogenase was described in a 4 year old boy with transient seizures, speech retardation and behavioural problems (van Gennip et al., 1981). Since then, several cases of thymine-uraciluria associated with similar neurological problems have been reported (Berger et al., 1984; Brockstedt et al., 1990). An example showing how the clinical presentation of the disorder can be heterogeneous, also within the same family, is provided by two male siblings, born from two first-cousin Asian parents, both with a diagnosis of thymine-uraciluria (Henderson et al., 1995). Interestingly, although levels of pyrimidines in urine and of enzymatic activity indicated a total absence of DPD in the two children, they had different phenotypes: the proband presented facial dysmorphism, absent in the older brother, who was reported to have instead problems in phonology, for which he received speech therapy.

Further evidences for the high variability associated with the disorder were provided by a study of 22 children with complete deficiency of DPD and onset of the clinical phenotype during childhood (Van Kuilenburg et al., 1999). Convulsive disorders, motor retardation and mental retardation were observed in the majority of cases, whereas growth retardation, microcephaly, autism and dysmorphism were less frequently observed. A minority of cases did not present any of the previously mentioned abnormalities, but they had other neurological problems, such as lethargy, dizziness, monoplegia and, interestingly, minor difficulties in learning speech and language. A mutation screening of *DPYD* in this group of patients identified 7 mutations: 2 microdeletions causing a frameshift with the introduction of a premature stop codon (*DPYD*7* [295-298delTCAT] e *DPYD*3* [1897delC]), 4 missense changes (p.C29R, p.R235W, p.R886H, p.V995F) and a splice-site mutation [IVS14 +1 G>A, also known as allele *DPYD*2A*], that resulted to be the most common variant (observed in 52% of analysed cases). Individuals carrying the same mutation were reported to have different clinical features, making the establishment of a correlation of genotype-phenotype difficult and complex.

However, an important point of this and subsequent studies is the frequent observation of neurological abnormalities, although of different entities, in patients with complete deficiency of DPD, suggesting that this gene might be implicated in neurodevelopment.

3.3 DPD: a key player in 5-FU metabolism.

Besides its crucial role in the catabolic pathway of pyrimidines, DPD is involved also in the catabolism of the widely used anti-neoplastic agents 5-Fluorouracil (5-FU) and its orally active prodrug capecitabine (Thorn et al., 2011).

The activation mechanism is based on the conversion of 5-FU to 5-fluoro-2'-deoxyuridine monophosphate (FdUMP), which competes with the natural substrates of thymidylate synthase (TS), preventing the pyrimidine synthesis (**Figure 3.2**). However, more than 85% of the administered 5-FU is rapidly degraded by DPD (catabolic pathway).

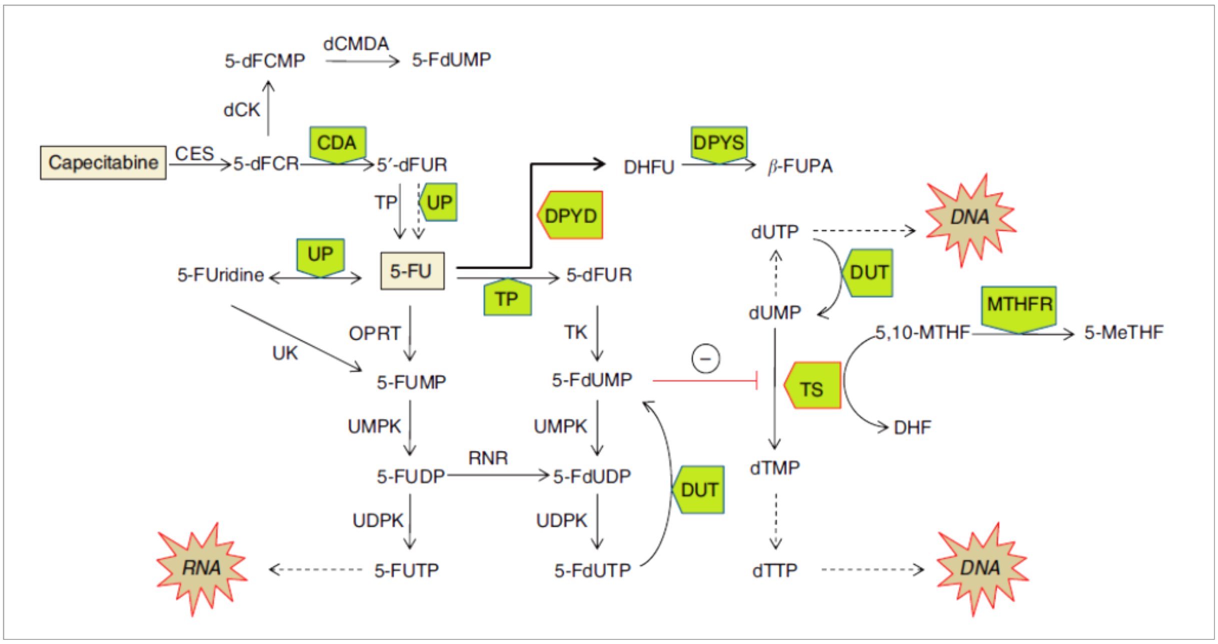


Figure 3.2 (Loganayagam et al., 2013). Anabolic and catabolic pathways of the 5-FU. In the degradation pathway, 5-FU is converted to dihydrofluorouracil (DHFU) by DPD (indicated with *DPYD*). DHFU is subsequently converted to fluoro-β-ureidopropionate (β-FUPA) by dihydropyrimidinase (indicated with *DPYS*).

The efficacy of the 5-FU anti-tumoral treatment depends on a narrow therapeutic window: as for other anti-neoplastic agents, there is a delicate equilibrium between toxic and therapeutic effects. Therefore, the breakdown of 5-FU is an important step of its metabolic regulation. Mutations in the *DPYD* gene causing partial or complete deficiency of DPD activity are associated with mild and severe toxicity in cancer patients receiving 5-fluorouracil chemotherapy, which can be lethal in the most extreme cases (Ezzeldin et al., 2003; van Kuilenburg et al., 2001). The typical toxic reactions to 5-FU when DPD is partially/completely deficient are diarrhoea, fever, neutropenia and mucositis. Given the availability of quantitative measures of these symptoms, toxicity reactions are classified in different categories: grades 0-2 indicate a mild-moderate toxicity, grades 3-4 a severe toxicity. In some cases, myelosuppression and CNS alterations, such as acute cerebellar ataxia, mental deterioration and myelopathy can occur (van Kuilenburg et al., 2003). Investigation on 5-FU-dependent neuronal alteration and neurological abnormalities reported in children with complete DPD deficiency might contribute to the understanding of the role of endogenous pyrimidines in neuronal activity.

Since the late 1980s, numerous cases with severe toxic reactions to 5-FU have been reported and have contributed to the identification of several variants in the *DPYD* gene. In addition to patients with homozygous mutations (Van Kuilenburg et al., 1999; Vreken et al., 1996), patients with multiple heterozygous mutations have been described (Gross et al., 2003), suggesting that the DPD deficiency might be determined by complex patterns of variants. Moreover, in accordance with the studies of thymine-uraciluria, individuals with the same genotypic profiles can display variable responses to 5-FU administration.

Considering the wide use of 5-FU in cancer chemotherapy and the absence of a clear genotype-phenotype correlation, it would be important to assay the DPD activity before the 5-FU treatment, in order to exclude an adverse reaction in the patient. Levels of DPD activity can be determined with a radio-enzymatic assay (Johnson et al., 1997). This method measures DPD activity using radio-labelled substrates, such as uracil, thymine or 5-FU, and its sensitivity allows the discrimination between partial and complete deficiency. Various tissues can be examined: peripheral blood mononuclear cells are particularly suitable to the measurement of the DPD activity, because the subsequent enzymes of the pyrimidine catabolic pathway (DHP and UP) are absent. Valid alternative methods have been also introduced, such as the determination of plasmatic uracil/dihydrouracil ratios (Ciccolini et al., 2006; Zhou et al., 2007) and the uracil breath test (UraBT) (Mattison et al., 2004), in which 2-^{13}C-uracil is orally administered to the patient: as CO_2 is one of the final products of the pyrimidine catabolism, estimates of ^{13}C levels in exhaled CO_2 allow to rapidly identify partial and profound DPD deficiencies.

3.4 Hypotheses for a role of *DPYD* in the central nervous system.

The frequent observation of neurological abnormalities, often with a childhood onset, in patients with DPD deficiency led researchers to hypothesize a role for *DPYD* in neurodevelopment.

It has been suggested that an altered homeostasis of β-alanine, a structural analog of gamma-aminobutyric acid (GABA) and glycine (**Figure 3.3**), which are major inhibitory neurotransmitters in the central nervous system, may account for some of the clinical abnormalities described in patients with DPD deficiency (Van Kuilenburg et al., 1999).

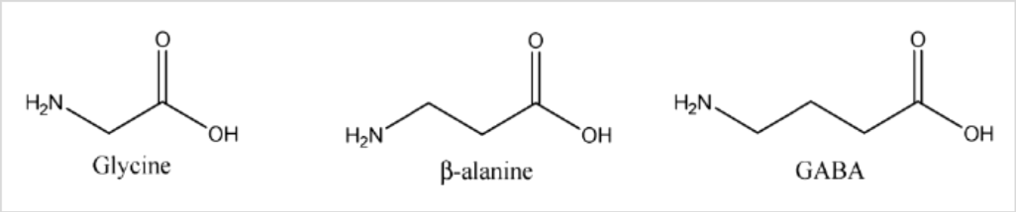


Figure 3.3 (Tiedje et al., 2010). Glycine, β-alanine, gamma-aminobutyric acid (GABA).

β-alanine biosynthesis can occur via three main ways:

1) Deamination and carboxylation of uracil (pyrimidine catabolism), mainly occurring in liver;
2) Interchangeable conversion of L-alanine and pyruvate;
3) L-aspartate decarboxylation by gut microbes.

Molecules of β-alanine can reach the central nervous system (CNS) crossing the blood brain barrier. Moreover, this aminoacid can be converted to malonate semialdehyde and, within the brain, the reverse reaction (transamination of malonate semialdehyde, catalysed by the enzyme GABA-T) can represent an additional source of β-alanine. β-alanine is present throughout the CNS and high-affinity uptake systems in glial and neuronal cells contribute to the regulation of its concentration, supporting a role in the modulation of the neuronal response (Tiedje et al., 2010). β-alanine can be recognized by multiple receptors, such as $GABA_A$ and glycine receptors, indicating that it can behave as an agonist of these inhibitory neurotransmitters. It is also a potent blocker of GABA re-uptake in glial cells.

Alterations of the balance between inhibitory and excitatory neurotransmissions can generate seizures. To contrast this hyperexcitability, GABAergic inhibition can be potentiated and the reuptake of GABA temporarily blocked (Pfeiffer et al., 1996), with an anticonvulsive effect. Since β-alanine is an agonist of GABA, it is possible that this neurotransmitter could be also involved in response to convulsions, a frequently observed symptom in DPD-deficient patients.

Although reduced levels of β-alanine would be expected in patients with DPD deficiency, they were reported to be only slightly lower in urine and plasma, and normal in cerebrospinal liquid, compared with controls (Van Kuilenburg et al., 2004). This finding suggested that alternative pathways for the production of β-alanine might compensate the effect of the decreased production of β-alanine from the pyrimidine pathway.

On the other side, DPD has a crucial role in the catabolism of pyrimidines. Pyrimidines are essential precursors for DNA and RNA synthesis, but they have also many more roles and are important for the activity of the central nervous system. The availability of pyrimidines in the cell is determined by the correct balance between synthesis (*de novo* and salvage pathways) and degradation. Therefore, alterations of the homeostasis of pyrimidines in brain, in particular in early developmental stages, may lead to neurological abnormalities. In addition to the DPD deficiency, there are other disorders caused by inborn errors in the nucleotide metabolism, such as dihydropyrimidinase (DHP) deficiency and β-ureidopropionase (BUP-1) deficiency (DHP and BUP-1 are the second and the third enzymes of the pyrimidine catabolic pathway, see paragraph 3.1) and they also present neurological dysfunctions, such as myelination delay, epileptic attacks, speech and developmental delay (Micheli et al., 2011). Therefore, it is possible to hypothesize that metabolic changes influencing pyrimidine homeostasis, and their downstream products, may account for some of the clinical manifestations observed in patients with DPD deficiency. However, the actual molecular link of purines and pyrimidines altered metabolism with the development and functionality of the central nervous system remains to be unravelled.

3.5 The structural organization of *DPYD* and genetic variation.

Two transcript variants encoding different isoforms have been found for the gene *DPYD* (**Figure 3.4**):

-isoform 1 (NM_000110.3, chr1:97543300-98386615, hg19, strand -), which consists of 23 coding exons and codes for a protein of 1025 aminoacids (NP_000101) with a molecular weight of 111 kDa;

-isoform 2 (NM_001160301.1, chr1:98185314-98386615, hg19, strand -), which includes only 6 coding exons and codes for a protein of 173 aminoacids (NP_001153773), of which the first 161 aa are common to isoform 1 protein.

Recently, two antisense non-coding genes, *DPYD-AS1* (NR_046590, chr1:97561479-97788511, strand +) and *DPYD-AS2* (NR_046591.1, chr1:98262477-98263607, strand +), have been mapped at the 3' end and the 5' end of the isoform 1, respectively, but their function is still unknown.

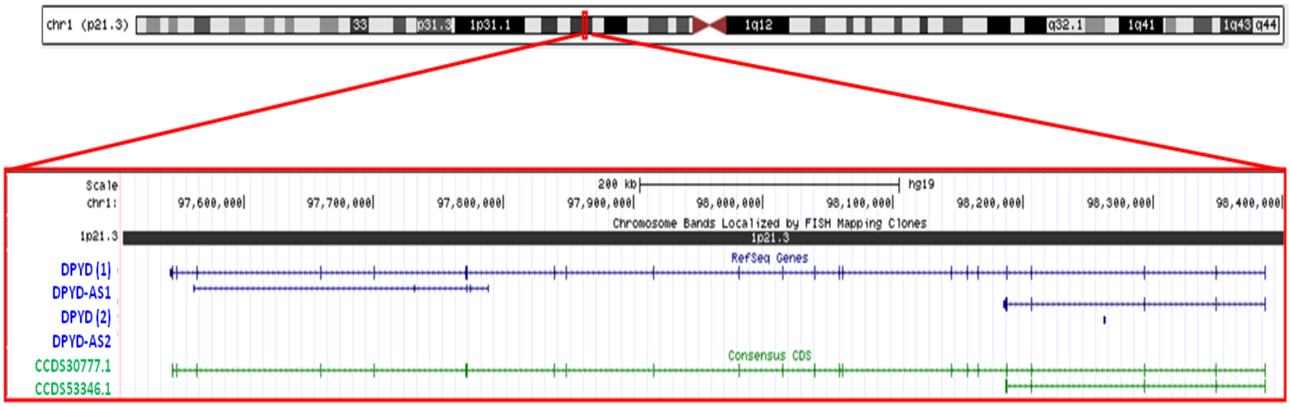


Figure 3.4. Schematic representation of the isoforms of *DPYD*, *DPYD-AS1* and *DPYD-AS2* from the Genome Browser UCSC (Feb. 2009, GRCh37/hg19).

The promoter for *DPYD* has also been characterized, with the identification of two essential regulatory elements in the region flanking the 5' of the *DPYD* gene: the regulatory element I (between -23 and -42), and the regulatory element II (between -72 and -51) (Shestopal et al., 2000) (**Figure 3.5**). A subsequent study (Zhang et al., 2006b) showed that the Sp1 and Sp3 transcription factors bind to *DPYD* promoter. Three Sp-target sites were identified: SpA (from -148 to -140), SpB (from -68 to -60) and SpC (from -37 to -19). The major promoter activity was detected for SpB, suggesting that it may function as an upstream enhancer, while SpC may represent an element of the basal promoter.

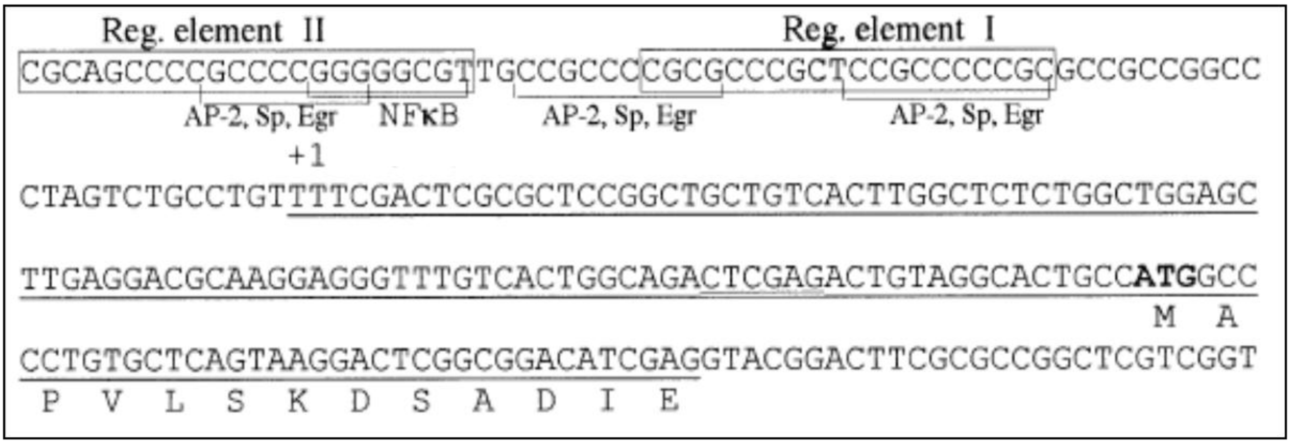

Reg. element II
Reg. element I
CGCAGCCCCGCCCCGGGGGCGTTGCCGCCCCGCGCCCGCTCCGCCCCCGCGCCGCCGGCC
AP-2, Sp, Egr
NFκB
AP-2, Sp, Egr
AP-2, Sp, Egr
+1
CTAGTCTGCCTGTTTTCGACTCGCGCTCCGGCTGCTGTCACTTGGCTCTCTGGCTGGAGC
TTGAGGACGCAAGGAGGGTTTGTCACTGGCAGACTCGAGACTGTAGGCACTGCCATGGCC
M A
CCTGTGCTCAGTAAGGACTCGGCGGACATCGAGGTACGGACTTCGCGCCGGCTCGTCGGT
P V L S K D S A D I E

Figure 3.5 (Shestopal et al., 2000). Localization of the two regulatory regions described in the *DPYD* promoter. Base positions are indicated respective to the transcription start site (+1). Exon 1 is underlined and the aminoacids are indicated under the correspondent triplettes. Putative binding sites for transcription factors AP-2, Sp, Egr, NF-κB are also indicated.

A large number of polymorphisms have been found in the coding, intronic and untraslated regions of the gene *DPYD*, and these findings have been mainly driven by studies regarding 5-FU toxicity. Among the missense amino-acid changes known to have a deleterious effect on the DPD activity, the most common variant in DPD-deficient individuals is the point mutation in the donor splice site of intron 14 (IVS14 +1 G> A, *DPYD**2A, rs3918290). The change of the first base of the donor splice site (GT) prevents the recognition of the site by the splicing machinery, resulting in a mRNA lacking exon 14, which is translated into a non-functional protein missing the correspondent 55 amino-acids (Vreken et al., 1996).

There are two other mutations reported to be consistently associated with a decreased DPD activity:

- rs55886062 (c.1679T> G, *DPYD**13), which determines the substitution of isoleucine with a serine (p.I560S),
- rs67376798 (c.2846A>T), which substitutes aspartic acid 949 with valine (p.D949V).

Several other variants have been identified in DPD deficient patients or individuals with adverse responses to 5-FU treatment, however their role in DPD deficiency has not been completely elucidated, because different studies often report contrasting results. A highly variable effect on DPD activity has been observed for heterozygous variants in *DPYD* (Amstutz et al., 2011).

Two main explanations have been hypothesized for the lack of correlation between mutations and DPD activity and/or 5-FU toxicity. One suggestion is that there could be an allelic regulation mechanism, leading to increased expression of the *wild-type* allele in the presence of a mutated allele on the other chromosome (Amstutz et al., 2011). The alternative hypothesis suggests that some sequence variants might represent "protective" alleles, which lead to an above-average enzymatic activity. The effect of a damaging mutation on one allele could be then compensated by a protective variant on the other copy of the gene. This hypothesis is supported by the observation of a broad range of DPD activity levels in the general population and by the recent findings of an *in vitro* assay (Offer et al., 2013), described in the next paragraph, where the variants C29R and S534N were shown to determine an increased DPD activity.

Moreover, even though the vast majority of the intronic variants are expected to be non-functional and, for this reason, are usually not investigated in mutational screenings, part of this genetic variation may contribute to the large variability in DPD activity, also in the general population. Intronic regions can harbour variants with a great functional impact, such as the mutations affecting the splicing. A cryptic splice donor site has been identified in intron 10 (c.1129-5923, rs75017182, chr1:98045449): the change C>G creates a splice donor site, causing a frameshift and a premature stop codon (van Kuilenburg et al., 2010). Interestingly, this variant is in linkage with a haplotype hapB3, previously found to be associated with severe adverse reactions to 5-FU (Amstutz et al., 2008).

Epigenetic factors could also influence the regulation of DPD activity. The promoter methylation has been investigated as a potential contributing factor. In a study including only five cancer patients (Ezzeldin et al., 2005), a partial methylation of the DPYD promoter was reported to be correlated to the reduced activity of the enzyme. However, these findings have not been subsequently replicated in larger independent samples of cancer patients (Amstutz et al., 2008; Savva-Bordalo et al., 2010; Schwab et al., 2008), indicating that the hypermethylation of the promoter is unlikely to be an important contributing mechanism.

3.6 The DPD protein.

A deep knowledge of the protein sequence and structure offers the possibility to predict and test the effect of point variants. The human DPD protein (Q12882-1, 1025 aa) has been purified from liver, one of the tissues where *DPYD* expression is higher, and that allowed the characterization of its enzymatic activity and the localization of functional domains. This cytosolic enzyme works as a homodimer and several important regions have been identified in the protein, including the uracil-binding site, the FMN-interaction site, the NADPH-binding region, iron/sulphur domains and the FAD-binding site. Moreover, these domains are conserved across evolution and present a high amino-acid identity between mammalians and *Drosophila melanogaster* (>85% identity between human and *Drosophila* for the first four domains, 54% for FAD-binding site) (Mattison et al., 2002).

Rat, pig and bovine DPD proteins have also been purified. Studies of the 3-dimensional structure of the pig protein (Dobritzsch et al., 2002; Dobritzsch et al., 2001) have been carried out: these models allow the prediction of potential conformational changes and altered interactions caused by mutations of amino-acids (**Figure 3.6**).

In the past, bacterial expression systems have been used to characterize the most interesting variants found in DPD-deficient patients. Recently, a new cellular system has been set up to test the effect of *DPYD* variants *in vitro* (Offer et al., 2013). In the human cell line HEK293T, endogenous DPD activity is not detectable, therefore, transfecting these cells with *DPYD* expression constructs containing a certain variant allows the determination of its effect on the activity of the protein. Moreover, the authors suggest that this system is suitable to test the effect of variants in the heterozygous state, by co-expression of the *DPYD* allele carrying the mutation with the *wild-type DPYD* allele.

All these approaches (comparative sequence analysis, expression systems and protein studies) are expected to provide new insights into the structural basis of DPD deficiencies caused by naturally occurring point mutations in the human *DPYD* gene.

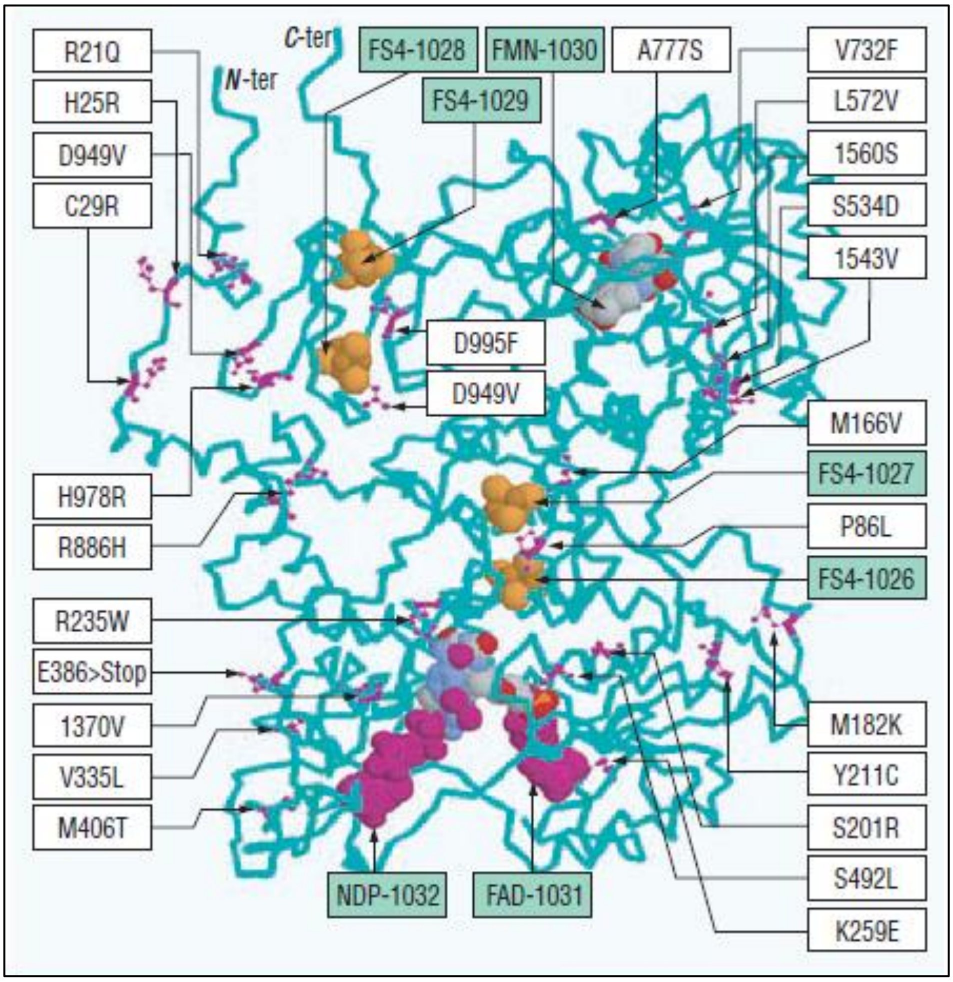


Figure 3.6 (Ezzeldin and Diasio, 2004). Stereo view of the DPD structure, obtained from the pig liver protein. Functional domains are indicated in green. The position of known sequence variants frequently reported in literature is also indicated.

3.7 Involvement of *DPYD* in neurodevelopmental disorders.

The identification of rare sequence and structural variants in *DPYD* in neurological disorders has provided new evidence for the hypothesis of a potential involvement of this gene in processes important for brain development.

In a recent study (Carter et al., 2011), hemizygous deletions involving the gene *DPYD* have been described in four individuals with ASD and severe speech delay, belonging to three unrelated families (**Figure 3.7 b**).

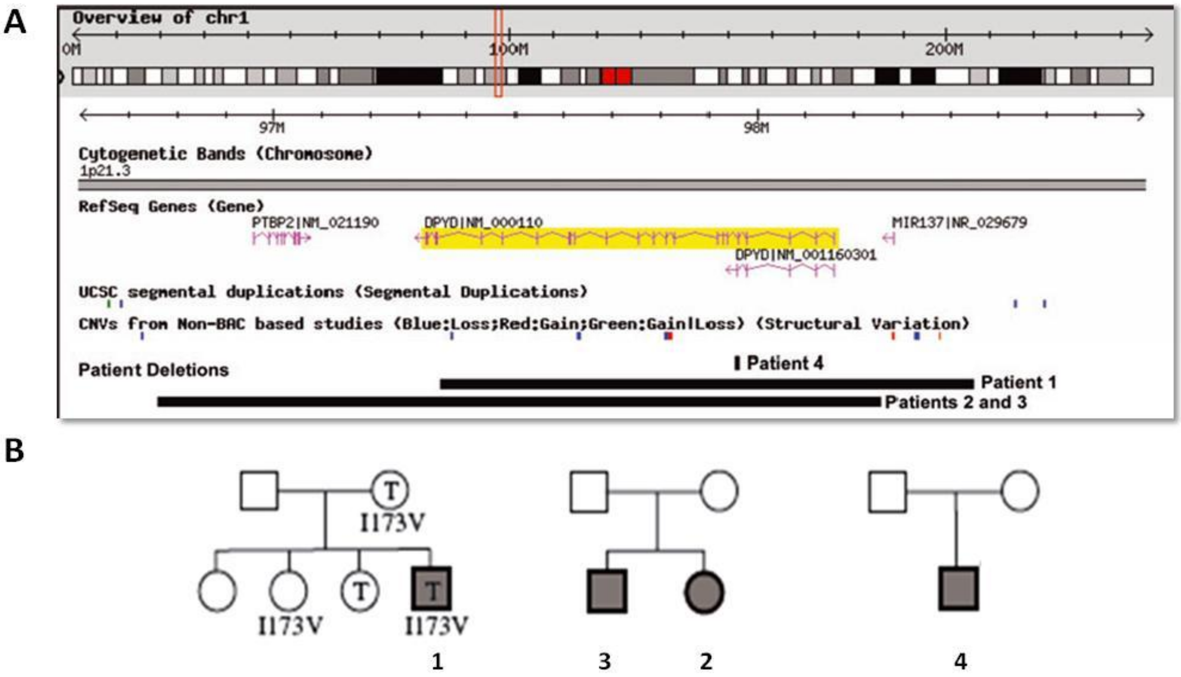


Figure 3.7 (Carter et al., 2011)**. A.** The figure shows the position of the three deletions (black bars) detected in the study of Carter et al. **B.** Pedigrees of the three families described in this study. The four affected individuals, each carrying a *DPYD* deletion, are indicated in grey. In family 1, I173V indicates the missense mutation in *PTCHD1* and T indicates a translocation t(19;21)(p13.3;q22.1).

- Patient 1 carried a *de novo* deletion of ~1Mb on chromosome 1p21.3, involving *DPYD* and the adjacent gene *MIR137* (**Figure 3.7a**). In addition, he had a missense mutation (p.I173V) in *PTCHD1* and translocation t(19;21)(p13.3;q22.1), both inherited from the mother (Noor et al., 2010). These two variants were also present, separately, in unaffected sisters (**Figure 3.7 b**).
- Patients 2 and 3 presented a translocation of a 1.5 Mb region on chromosome 1, including *DPYD* and the adjacent gene *PTBP2* (**Figure 3.7 a**), to the short arm of chromosome 10, a rearrangement inherited from the mother.
- Patient 4 was shown to have a 10Kb intragenic deletion, involving only exon 6 of *DPYD* (**Figure 3.7 a**), inherited from the mother. This CNV is predicted to cause a frameshift, with a premature truncation of the protein.

In order to check whether the non-deleted copy of *DPYD* carried a damaging sequence mutation, leading to a complete loss of the functional protein DPD in compound heterozygosity, a mutational screening of the coding regions of the gene was performed in the four probands. However, no coding variants were reported in these individuals. From the screening of 300 unrelated autism probands instead, four missense mutations (p.P3C, p.C622Y, p.T793I, p.P1023T) and the splicing variant IVS14+1 G>A (rs3918290) were identified, each of them present in 1/300 probands. All these variants were inherited from an unaffected parent, but absent in 48 controls. The changes p.P3C and p.T793I are not reported in dbSNP nor in the ESP database (http://evs.gs.washington.edu/EVS/), that collects data from exome sequences of a large number of European American (EA) and AfricanAmerican (AA) individuals. p.C622Y (rs201433243) is also rare (minor allelic frequency=0.023% in EA, not reported in AA subjects), and p.P1023T (rs114096998) has not been identified in EA individuals, but it is reported in African American individuals with a minor allelic frequency of 3.8%.

Then, this study suggested that *de novo* and inherited variants in *DPYD* could contribute to the ASD susceptibility, probably in conjunction with mutations in other genes. An example is given by the family 1, in which the proband, in addition to the *de novo* deletion in *DPYD*, had also inherited a missense change in *PTCHD1*, which is a X-linked autism susceptibility gene. This family was originally included in a study (Noor et al., 2010) that reported mutations in *PTCHD1* in 1% of males with ASD and with ID. Therefore, in some individuals, the presence of mutations in *PTCHD1* could modulate the phenotype associated with the partial loss of DPD.

A CNV screening carried out in ~700 ASD unrelated cases (Prasad et al., 2012) also detected two rare inherited deletions in *DPYD* and an inherited exonic copy number loss in *UPB1*, the gene coding for the enzyme β-ureidopropionase.

Deletions encompassing *DPYD* have been identified also in individuals with Intellectual Disability (ID) (Willemsen et al., 2011). This study described five individuals, belonging to three different families, with moderate ID and language deficits. The deletions reported in this study are represented in **Figure 3.8**.

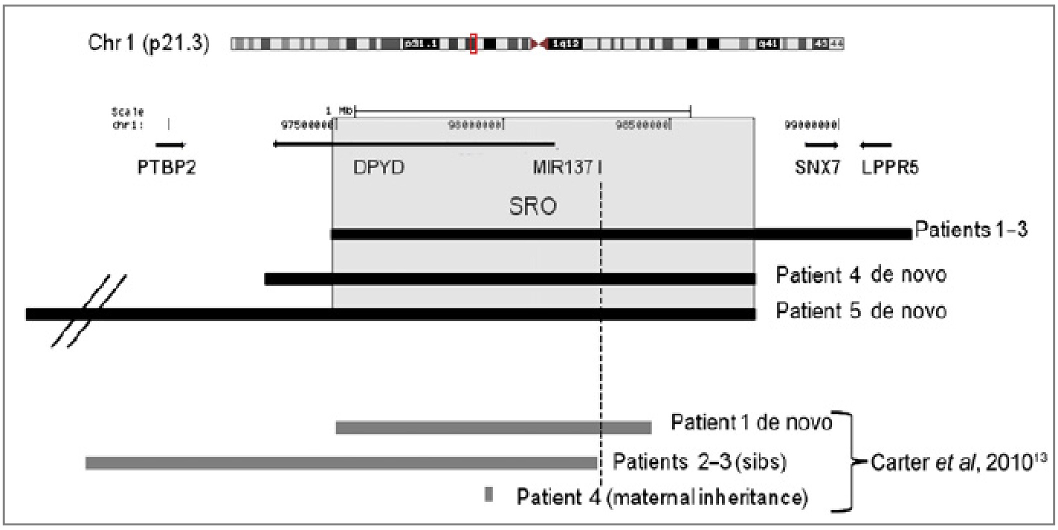


Figure 3.8 (Willemsen et al., 2011). Schematic representation of the deletions on chromosome 1p21.3 identified by Willemsen et al. The shortest overlapping region among the deletions is shown in grey. The overlap with the deletions reported in Carter *et al*. is also shown.

Patients 1, 2, and 3 were siblings and presented a 1.75 Mb deletion on chromosome 1p21.3, involving the genes *DPYD*, *SNX7* and *LPPR5* and the microRNA *MIR137*. The parents were not available, therefore the heritability could not be checked. Patient 4 had a 1.41 Mb *de novo* deletion that included the two adjacent genes *DPYD* and *MIR137*. Patient 5 carried a larger *de novo* deletion (2.45 Mb), encompassing the three genes *PTBP2*, *DPYD* and *MIR137*. The shortest overlapping region (SRO) across the different CNVs was a region of 1.22 Mb that includes only *DPYD* and *MIR137*. Metabolic tests in individuals 1, 2 and 4 did not show the typical thymine-uraciluria profile that would be expected from the partial loss of DPD. Expression analyses in lymphoblastoid cell lines showed instead decreased levels of the precursor and mature *MIR137* and, in addition, a significant increase of the expression of three genes (*MITF*, *EZH2 e KLF4*), negatively regulated by *MIR137*. Since also the microRNA is a good candidate gene for neurological abnormalities, as it is expressed in neuronal tissues, particularly in the hippocampus, the authors suggested that the ID phenotype was more likely to be associated with the haploinsufficiency of *MIR137*, although they did not exclude a possible involvement of *DPYD* in neurological disorders.

Another study has reported an individual with mild developmental delay/ID and dysmorphic features carrying a *de novo* copy number loss (570 kb) involving both *DPYD* and *MIR137* (Battaglia et al., 2013).

De novo mutations in *DPYD* have been detected also in individuals with schizophrenia (Xu et al., 2012). The analysis of 795 exomes, deriving from 231 simplex families (146 Afrikaner and 85 US trios), highlighted four genes affected by more than one *de novo* event and, thus, indicated as potential susceptibility genes: *DPYD, LAMA2, TRRAP* and *VPS39.* In particular, the variants identified in *DPYD* were a nonsense mutation (c.1863G>A, p.W621*) and a missense change (c.1615G>A, p.G539R). The patient carrying the p.G539R variant had increased levels of thymine and uracil in urines, suggesting a deleterious effect of the variant.

A previous GWA study for schizophrenia (Schizophrenia_Psychiatric_GWAS_Consortium, 2011) had identified a strong association signal for the SNP rs1625579 ($P= 1.6 \times 10^{-11}$), mapping to the intron 3 of the miR137-Host Gene *MIR137HG*. The association of this region has been subsequently replicated ($P= 1.72 \times 10^{-12}$, rs1198588, falling within an intergenic region ~38 kb upstream the gene *MIR137HG*) in a larger cohort of over 21,000 cases and 38,000 controls (Ripke et al., 2013). The SNP rs1625579 falls in a linkage disequilibrium block (D >0.9) extending to the 5’ of *DPYD* (**Figure 3.9**) and rs1198588 represents an expression QTL for *DPYD*.

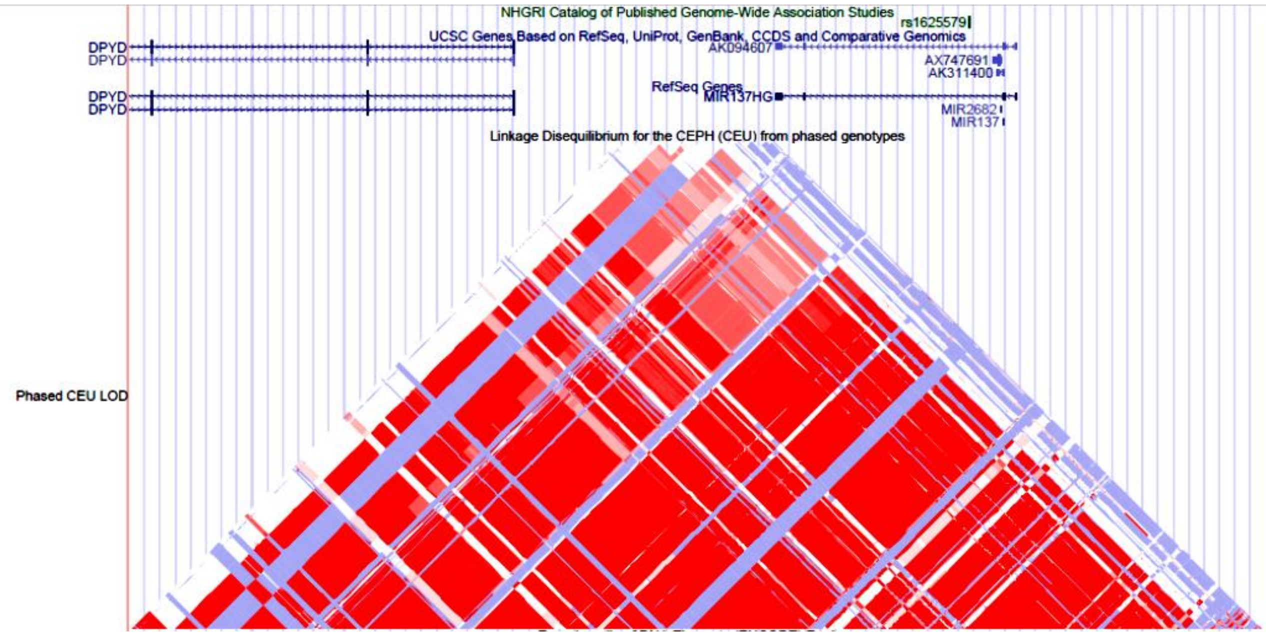


Figure 3.9. Linkage Disequilibrium block including *DPYD*, *MIR137HG* and *MIR137*. At the top, the position of the SNP rs1625579, found in association with schizophrenia, is indicated. The miR-137 Host Gene (*MIR137HG)* contains 5 exons: mature microRNA-137 is encoded in exon 3 of *MIR137HG*.

In Xu *et al.*, sequence variants in *MIR137* have not been detected, therefore, taking into account the linkage disequilibrium between *MIR137* and *DPYD,* the association signal might reflect also the involvement of *DPYD* in schizophrenia susceptibility.

In addition, a putative miR-137-target site has been identified within *DPYD* (Ripke et al., 2013), indicating that this gene might be regulated by miR-137 and supporting the hypothesis that both genes could be implicated in a range of neuropsychiatric disorders. The phenotypic outcomes may depend on the nature of the mutations occurring in these genes and whether they affect one of them (this would be the case of point mutations) or both (like in the case of large deletions).

Moreover, a recent screening has found that miR-137 is one of the direct FOXP2-targets in embryonic mouse brain (Vernes et al., 2011). These findings establish intriguing connections between these three candidate genes *FOXP2*, *MIR137* and *DPYD*, suggesting that they might be all involved in networks important for brain development.

Chapter 4

Materials and Methods

4.1 Samples: SLIC cohort.

4.1.1 Subjects.

The individuals belonging to the SLIC cohort have been recruited from 4 centres in the United Kingdom: the Cambridge Language and Speech Project (CLASP), the Newcomen Centre at Guy's Hospital (London), the Child Life and Health Department at the University of Edinburgh and the University of Manchester (Falcaro et al., 2008; Newbury et al., 2011; SLIC, 2002, 2004).

- CLASP consists of an epidemiological investigation of speech and language problems (Burden et al., 1996). The recruitment of children into the study has followed a multi-stage ascertainment procedure. First, at age 36 months, the population was defined by means of a questionnaire; then, at age 39 months, language abilities were assessed in more detail and, finally, at age 45 months, screen–positive cases were examined in depth. At 8 years of age, the children and their siblings were tested by the CELF-R and Wechsler Scales of Intelligence–Third UK Edition (WISC-III) (Wechsler, 1992).
- The cases recruited at Guy's Hospital were selected through three special schools for language disorders and through "Afasic", a support organization for people with developmental and language impairments (http://www.afasic.org.uk/). Therefore, these individuals were considered a self-referred sample as they attended special schools because of their persistent language problems.
- The cases referred by the University of Edinburgh were selected originally to participate in a study of children from eastern and central Scotland with severe receptive-language impairments (Clark et al., 2007). All probands needed specialist educational support and were selected to have a historical language comprehension score >2 SD below that expected for their age, although the majority were confirmed to also have expressive problems. A detailed family history was taken from the families that decided to take part in the study.
- The Manchester Language Study (http://www.manchesterlanguagestudy.co.uk) has been running nationwide since 1995 (Conti-Ramsden and Botting, 1999; Falcaro et al., 2008), following the progress of 242 children who were attending language units at 7 years of age. The children have been assessed at several stages, including at the age of 17 years. The study involved their families and schools. A group of age-matched peers were also recruited into the study at the age of 16 years for comparison purposes.

Whole-blood (Guy's and Edinburgh) or buccal swab (Cambridge and Manchester) samples were collected from all available family members, regardless of their language ability. DNA was extracted using standard protocols, and all buccal swab DNAs were pre-amplified using a rolling circle whole-genome amplification protocol (HY-Genomiphi, GE Healthcare). Only genomic samples were included in the CNV analyses.

4.1.2 SLIC Cohort: phenotypic tests.

Language skills of all available children belonging to the SLIC collection were assessed by one of the four centres across the UK using two baseline tests: Clinical Evaluation of Language Fundamentals Revised test (CELF-R) and NonWord Repetition (NWR), as described in the first paper of the SLI Consortium (SLIC, 2002).

CELF-R evaluates expressive and receptive language domains independently and then combines the results in a composite language score. It consists of different batteries of tests, depending on the age of the subject. Raw scores were transformed to obtain an age-normalised standardized receptive language score (RLS) and an expressive language score (ELS), each with mean 100 and Standard Deviation of 15 in the general-population calibration sample (Semel et al., 1992).

NWR was used as a marker of phonological short-term memory (Gathercole et al., 1994). In this test, children are required to hear and repeat a sequence of tape-recorded single nonsense words of increasing length and complexity (such as "woogalammic" or "perplisteronk") and each repetition is scored as correct or incorrect. All available children of age 7.5–18 years were assessed with the 28-item NWR test. Standard scores for a British population were then obtained by use of norms extended, for older children and in many cases for adults.

In order to exclude mental retardation as a possible cause for language impairment, cognitive abilities were assessed with the WISC-III test (Wechsler, 1992). This consists of a multi-test battery that evaluates both verbal IQ (VIQ) and performance IQ (PIQ). The verbal abilities are examined by subtests of vocabulary, similarities, comprehension and abstract reasoning, whereas the performance tasks are primarily based on visual and constructional clues (e.g., mazes, block design, picture concepts and abstract visual problem solving). WISC Perceptual Organisation Index (POI) is a composite score of the non-verbal subtests Picture Completion, Picture Arrangement, Block Design and Object Assembly. Verbal IQ and PIQ can be combined in a composite scores to give a full-scale IQ. The WISC-III does not include reading or writing tests.

In some subgroups (Guy's, Manchester and Cambridge), reading abilities were assessed using single-word reading (Read), single-word spelling (Spell) and reading comprehension (Comp) tests from the Wechsler Objective Reading Dimensions (WORD) (Rust et al., 1993).

In each family, the proband was defined as the individual through whom the family was identified and, either currently or in the past, had language skills (ELS and RLS) ≥1.5 SD below the normative mean for their chronological age, on the receptive and/or expressive scales of the CELF-R battery. When considering family segregation patterns, children were classified as affected if they were probands or siblings with ELS or RLS scores ≥ 1.5 SD below the normative mean. Children with both ELS and RLS scores ≤ 0.5 SD below the mean were classified as unaffected. For the other family members that did not meet these thresholds, the affection status was considered unknown. For many adults, data was unknown, but self-reported problems were taken into account.

Any proband or sibling found to have WISC POI composite score of <77.5 (1.5 SD below that expected for their age) or PIQ <80 was excluded from the CNV genome screen. Additional exclusion criteria included: MZ twinning, chronic illness requiring multiple hospital visits or admissions, deafness, an ICD-10/DSM-IV diagnosis of childhood autism, English being a second language, care provision by local authorities, and known neurological disorders. In the Guy’s Hospital sample, those families with chromosome abnormalities, including fragile X, were excluded by cytogenetic testing.

4.2 Samples: ASD cohorts.

4.2.1 IMGSAC cohort: subjects.

The International Molecular Genetic Study of Autism Consortium (IMGSAC) (IMGSAC, 1998, 2001) collected multiplex ASD families from different countries (UK, Netherlands, France, USA, Germany, Denmark, Greece). The individuals had predominantly a Caucasian origin (>90%). Clinical diagnosis of ASD was made using the Autism Diagnostic Interview-Revised (ADI-R) (Lord et al., 1994) and Autism Diagnostic Observation Schedule (ADOS) (Lord et al., 1989) or ADOS-Generic (ADOS-G) (Lord et al., 2000), that will be described later.

A clinical evaluation was undertaken in order to exclude known medical disorders etiologically associated with autism. The multiplex families had fragile X testing and were karyotyped whenever possible. The following items were used to exclude individuals so as to keep the sample set more homogeneous:

- Any medical condition likely to be aetiological (e.g. tuberous sclerosis, fragile X, focal epilepsy, infantile spasms, single gene disorders involving the central nervous system).
- Any neurological disorder involving pathology above the brain stem, other than uncomplicated non-focal epilepsy.
- Contemporaneous evidence, or unequivocal retrospective evidence, of probable neonatal brain damage.

- Clinically significant visual or auditory impairment after correction.
- Rearing in adoptive or foster homes.
- Institutional rearing during the first 4 years when there is any possibility that this led to an autistic-like picture.
- Any circumstances that might possibly account for the picture of autism (e.g. very severe nutritional or psychological deprivation).
- Birth in a place making it difficult to obtain satisfactory obstetric data (this would ordinarily exclude those born in a developing country).
- Autism secondary to some other psychiatric disorder (e.g., schizophrenia), but not psychiatric co-morbidity.
- Observational data that cast doubt on the diagnosis.
- *In vitro* fertilisation as a means of conception.
- Cases arising from consanguineous marriage.

A number of features were not used as exclusion criteria these are: epilepsy (unless focal); psychiatric co-morbidity; head circumference over 97%; mental illness in a parent and belonging to an ethnic minority.

In a previous study (IMGSAC, 2001), the affection status was classified in a hierarchical fashion:

- "case type 1" if they had a clinical diagnosis of autism, met ADI-R and ADOS or ADOS-G algorithm criteria for autism, had a history of language delay and a PIQ ≥35.
- "case type 2" if they had a clinical diagnosis of autism, atypical autism, Asperger syndrome, or PDD-NOS and met at least ADOS-G criteria for PDD; there was no requirement for a history of language delay, and individuals were allowed to fall one point below threshold on one behavioral domain of the ADI-R.
- "case type 3" if they had a clinical diagnosis of autism or another PDD and either met ADI-R criteria for autism or fell one point below threshold in one behavioral domain, but failed to meet ADOS-G criteria for PDD.

In this project, all three classes of individuals were included as ASD cases.

Genomic DNA was extracted from blood using Nucleon® kit (IMGSAC, 1998). In a minority of cases in which a blood sample could not be obtained, DNA was extracted from buccal swabs (IMGSAC, 1998). In addition, whenever possible, lymphoblastoid cell lines (LCLs) have been generated from peripheral blood leukocytes.

4.2.2 Italian cohort: subjects.

The Italian ASD cohort was formed by Italian simplex families, each with one individual affected by ASD, collected by the clinical team of Professor Agatino Battaglia and Dr Raffaella Tancredi at the Institute Stella Maris (Pisa, Italy) and the group of Professor Luigi Mazzone from the University of Catania (Catania, Italy). Phenotypes were assessed using the two diagnostic tools: ADI-R (Lord et al., 1994) and ADOS-G (Lord et al., 2000). A clinical evaluation was undertaken in order to exclude known medical disorders etiologically associated with autism. Standard karyotyping and fragile-X testing were obtained for probands whenever possible.

DNA was extracted from blood with the QIAGEN DNA Blood extraction kit.

4.2.3 AGP cohort: subjects.

The third group of cases was part of the cohort of families collected by the Autism Genome Project (AGP), an ongoing international project which gathers more than 50 research groups from different countries in North America and Europe (Pinto et al., 2010), including the group headed by Elena Maestrini. The affection status was determined using the diagnostic tools ADI-R and ADOS. Three categories of affected individuals were established (strict, broad and spectrum ASD), based on proband diagnostic measures:

- the strict class included affected individuals who met criteria for autism on both ADI-R and ADOS instruments;
- the broad class included individuals who met full autism criteria on one diagnostic instrument and ASD criteria on the other one;
- the spectrum class included all individuals who were classified as ASD on both the ADI-R and ADOS or who were not evaluated on one of the instruments but were diagnosed with autism on the other instrument.

Given the international and multi-site nature of the project and the range of chronological and mental ages of the probands, a range of cognitive tests were used, and standard scores were combined across tests to provide consolidated IQ estimates. Subjects from all classes (strict, broad, and spectrum) were included in the association analyses performed in this thesis.

DNA, extracted from blood, buccal-swabs or cell-lines, has been previously genotyped with Illumina Human 1M-single and Illumina Human 1M-Duo BeadChip arrays (Anney et al., 2010; Pinto et al., 2010).

298 affected individuals and 592 parents from IMGSAC multiplex families, 27 affected individuals and 54 parents from IMGSAC simplex families and 69 Italian trios were included in the AGP sample.

4.2.4 ASD diagnostic instruments: ADI-R and ADOS-G.

The ADI-R (Lord et al., 1994) is a standardized interview conducted with caregivers of autistic individuals. The questionnaire explores the areas of communication, social interactions, restricted and repetitive behaviours and developmental history. An algorithm was generated to make a standard diagnosis on the basis of the ADI-R scores obtained in each area and it includes only the items that more closely depicted the phenotypic abnormalities described in the Diagnostic and Statistical Manual of mental disorders-4th edition (DSM-IV) ([APA], 1994). The algorithm specifies a cut-off score for each of three core domains of autism, therefore, only the individuals who meet all the cut-offs, meet diagnostic criteria for autism, the most severe form of ASD.

The ADOS-G (Lord et al., 2000) is an interactive test that aims to assess social interactions, communication, play and spontaneous behaviours in a standardized context. ADOS-G is an implementation of ADOS (Lord et al., 1989), which was proposed as a complementary instrument to ADI. ADOS-G has four possible modules that provide different structured and unstructured situations in order to evaluate social-communicative skills of the referred subject. As for ADI-R, subsets of items in each module of ADOS-G were selected to generate the diagnostic algorithm. Classification is made on the basis of exceeding cut-offs in social behaviour, communication and social-communication totals.

4.3 Identification of CNVs in the SLIC cohort.

A genome-wide CNV screen of 542 individuals from 170 2-generation families from the SLIC cohort (119 individuals from Cambridge collection, 165 from the Edinburgh collection, 210 from the Guy’s collection and 48 from the Manchester collection) was performed to investigate copy number variation burden in individuals with SLI.

This study utilized genome-wide SNP data generated on the Illumina Human OmniExpress (v12.1) beadchip. CNVs were identified using QuantiSNP (Colella et al., 2007) and PennCNV (Wang et al., 2007) algorithms. All the data were generated in Build GRCh37/hg19.

4.3.1 Illumina Human OmniExpress array.

The Illumina Human OmniExpress beadchip generates genotype calls for more than 700,000 markers using the Infinium HD technology (**Figure 4.1**).

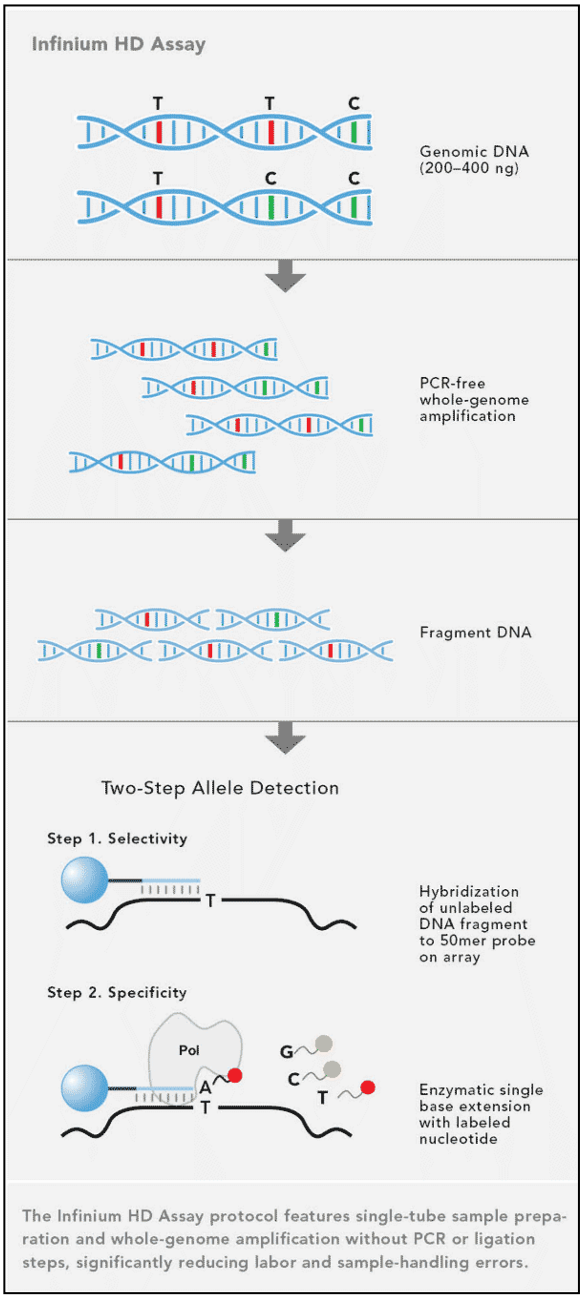

Infinium HD Assay
T
T
C
T
C
C
Genomic DNA (200–400 ng)
PCR-free whole-genome amplification
Fragment DNA
Two-Step Allele Detection
Step 1. Selectivity
T
Hybridization of unlabeled DNA fragment to 50mer probe on array
Step 2. Specificity
Pol
A
T
G
C
T
Enzymatic single base extension with labeled nucleotide
The Infinium HD Assay protocol features single-tube sample preparation and whole-genome amplification without PCR or ligation steps, significantly reducing labor and sample-handling errors.

Figure 4.1. Schematic representation of the Illumina Infinium HD Assay protocol.

This assay is based on a two-step detection process:

a) fragments of the DNA of interest selectively hybridize to specific probes (50-mer oligonucleotides), designed to be complementary to the loci of interest, but stopping one base before the interrogated SNP;

b) an enzymatic single-base extension incorporates a nucleotide labelled with a fluorescent dye, complementary to the base present at the SNP site.

For each SNP, Cy3 and Cy5 fluorescence signals specify the two alleles (generally referred to as allele A and allele B). Dual-colour staining of the labelled nucleotides is followed by an image scanning performed by Illumina’s iScan imaging system, which detects both colour and signal intensity. Homozygous genotypes are specified by red/red or green/green signals, heterozygous genotypes are indicated by red/green (yellow) signals.

The data generated by a SNP genotyping array can be analyzed and visualized with GenomeStudio Data Analysis software (Illumina). This program can covert the raw image scan into quantitative values and calculate the signal intensities for alleles A and B at each SNP, indicated as X and Y, respectively. The allele specific intensities are normalized using a proprietary algorithm of the Illumina GenomeStudio software: this procedure adjusts for background and makes markers more comparable to each other. Normalized allelic intensities are used to calculate the total fluorescent intensity signal (R) and the allelic intensity ratio (θ). R is calculated as a combined SNP intensity: $R_{observed}$ = X+Y, while θ is calculated as arctan(Y/X)/(π/2). R and θ values are calibrated to canonical genotype clusters generated from a large panel of normal samples, used to determine the R and theta values expected for each genotype (AA, AB and BB). R and θ are then converted to two important measures: Log R Ratio (LRR) and B Allele Frequency (BAF).

LRR, which represents a normalized measure of the total signal intensity at each SNP, derives from the comparison of the $R_{observed}$ with the R obtained from a reference sample population ($R_{expected}$) and it is calculated as $\log_2(R_{observed}/R_{expected})$. In autosomic regions without CNVs (copy number =2 for autosomes), LRR is ~0: LRR lower than zero may indicate a deletion, LRR higher than zero a duplication.

BAF, which derives from the normalized θ, represents the proportion contributed by allele B to the total copy number. BAF represents an estimate of $N_B/(N_A+N_B)$, where N_A and N_B are the number of A and B alleles, respectively, therefore its value range from 0 to 1. BAF close to 1 indicates that all alleles for that marker are B alleles (e.g. BB, BBB or B/-), *viceversa* BAF close to 0 indicates that all alleles for that SNP are A alleles (e.g. AA, AAA or A/-), values close to 0.5 indicate a heterozygous genotype AB.

These two transformed parameters, LRR and BAF, are plotted along each chromosome for all SNPs on the array and can be then visually inspected (**Figure 4.2**). The exported values of LRR and BAF for each SNP in each individual can be used for the identification of changes in copy number by QuantiSNP and PennCNV.

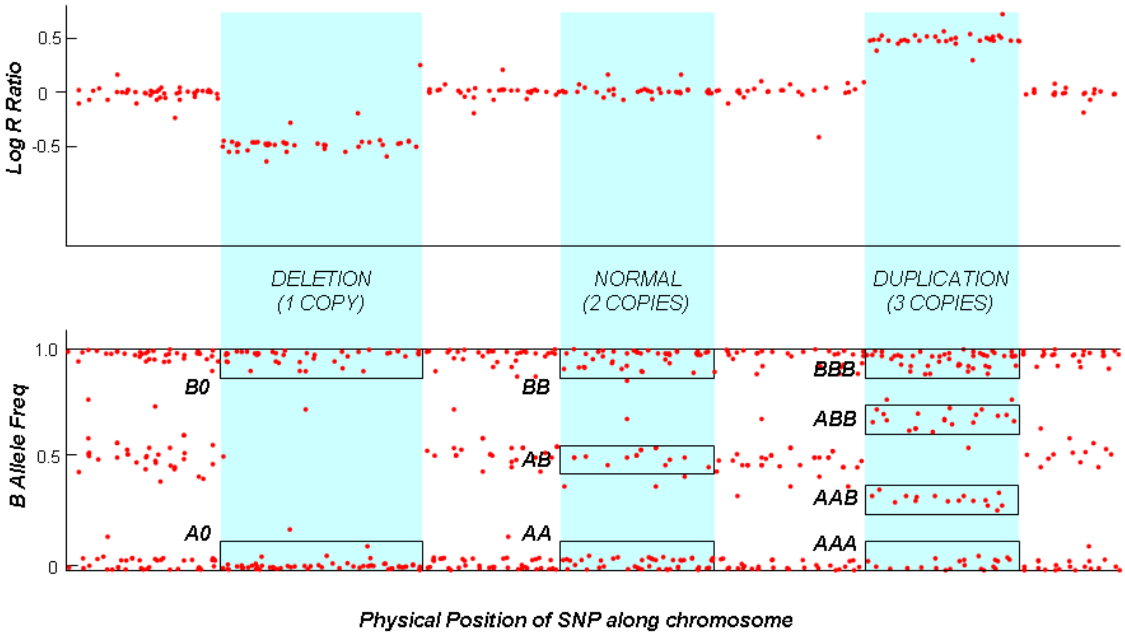


Figure 4.2. Examples of LRR and BAF plots for a deleted region (1 copy, genotype B/- or A/-), region with 2 normal copies (three possible genotypes for each SNP: AA, AB, BB) and duplicated region (3 copies, four possible genotypes for each SNP: AAA, AAB, ABB, BBB).

4.3.2 QuantiSNP.

The algorithm QuantiSNP (Colella et al., 2007) was originally developed for Illumina Infinium SNP genotyping data. This program incorporates the LRR and the BAF values simultaneously in an Objective Bayes Hidden Markov Model (OB-HMM). The HMM provides a statistically powerful framework suitable for CNV detection from signal intensity data. This model sets *a priori* probability of observing copy number changes between SNP loci at a certain distance. The hidden state represents the unknown copy number at each SNP site. Instead of using only three possible states (loss, normal and gain), QuantiSNP adopts a six-state definition for modelling CNV events (**Table 4.1**) and the states are inferred using the LRR and the BAF, and assumed to be independent.

Copy no. state	Copy no.	Description	CNV genotypes
1	0	Deletion of 2 copies	Null
2	1	Deletion of 1 copy	A, B
3	2	Normal state	AA, AB, BB
4	2	Copy neutral with LOH	AA, BB
5	3	Single copy duplication	AAA, AAB, ABB, BBB
6	4	Doubles copy duplication	AAAA, AAAB, AABB, ABBB, BBBB

Table 4.1 (Wang et al., 2007). Hidden states and correspondent copy numbers and genotypes. This table is valid for autosomes.

When a copy number variation is detected, QuantiSNP assigns a Bayes Factor to all of the possible copy number states, a measure of likelihood that the region harbours that copy number. Among all the possible copy number states, the most probable is indicated with the Maximum Log Bayes Factor. Therefore, higher values of Log Bayes Factor indicate a stronger support for the presence of CNV in the specified position given by the available SNP data.

We downloaded QuantiSNP package (version 2, April 2010) from the website https://sites.google.com/site/quantisnp/downloads.

CNV detection performed with QuantiSNP requires:

- Configuration files (provided with the package), that contain default model parameters. These files are levels-hd.dat (list of the copy number states and associated mean levels for LRR, suitable for Illumina-Omni arrays and other Illumina Infinium users) and params.dat (list of hyperparameter settings involved in the statistical model underlying QuantiSNP).
- MATLAB Run-Time Libraries;
- Input signal file, which is a text file with 5 columns for each marker (SNP Name, Chromosome, Position, Log R Ratio, B Allele Frequency), as exported from Genome Studio;
- Emitters (optional), which represents the number of iterations used for the EM algorithm during learning. We used the default number 10.
- lsetting (optional), which indicates the length used to calculate transition probabilities. We set the default value (2,000,000).
- GC file (optional): SNP data can present wave-like artefacts in the Log R Ratio, called “genomic waves” (Diskin et al., 2008), that are not platform-specific, but depend on local genomic features, in particular the GC content. In order to reduce false CNV calls, QuantiSNP can incorporate local GC content information to remove these artefacts. We used the local GC content-based correction for Build 37.

Chromosome X needs a special processing: using --doXcorrect in the command line, the program adjusts the LRR for the X chromosome at zero for females or the deletion level for males.

After CNV calling, QuantiSNP generates two output files for each sample: a list of putative copy number alterations and a summary of quality parameters.

The quality control (QC) output file reports three measures for each chromosome:

- an estimated probability of outliers (outlier rate);
- a measure of the standard deviation of LRR values (SD_LRR);
- a measure of the spread of BAF distribution for heterozygous genotypes (SD_BAF).

A high outlier rate and high standard deviations provide a measure of the noise in the data, which can also be visually deduced from scatter plots of the two metrics within Genome Viewer (Illumina), that indicate bad quality samples.

For each individual, we calculated the average of LRR_SD and BAF_SD values, taking into account all chromosomes, and we filtered out samples with LRR_SD > 0.3 and BAF_SD > 0.15 to exclude low quality samples. These thresholds were established based on the quality parameters values specified by previous CNV studies (Marenne et al., 2011; Pankratz et al., 2011; Sanders et al., 2011).

The output file with the list of putative CNVs reports a Maximum Bayes Factor for every copy number variant identified. Since Log Bayes Factors of less than 10 are frequently associated to false positives (>10%), only CNV calls including three or more probes and with a log Bayes Factor ≥10 were kept in the analysis.

4.3.3 PennCNV.

PennCNV (Wang et al., 2007) represents an integrated HMM algorithm, originally developed for the Illumina Infinium assay, but then extended to other SNP genotyping platforms. This algorithm is based on more realistic models for state transition between different copy number states. Indeed, in addition to LRR and BAF, PennCNV also incorporates the distance between adjacent SNPs and the population frequency of the B allele for each SNP (PFB). The list of PFB for all SNPs is compiled using a large set of control individuals with no clinical phenotypes and, if possible, with mixed ethnic backgrounds, and contributes to a more accurate modelling of the likelihood of copy number genotypes. All these values (LRR, BAF, PFB and distance) determine the probability of having a copy number change at a specific position.

Chromosome X needs a special treatment: using the --chrx argument, LRR levels are adjusted, so that the average LRR is either 0 for females or the values expected for a single copy deletion for males. After this procedure, the CNV calling is performed in a similar way as for autosomes.

The PennCNV package (version of June 2011) was freely downloaded at http://www.openbioinformatics.org/penncnv/penncnv_download.html.

PennCNV individual-based calling.

PennCNV requires several files for CNV detection:

- a signal input file for each sample, which is a text file containing the SNP name, LRR and BAF values for each marker;

- the HMM file (“hhall.hmm” file in our analysis, supplied with the package), which specifies the HMM model, that indicates the expected signal intensity values and the expected transition probability for different copy number states;
- the PFB file (Build 37), which is a four column text file containing SNP name, chromosome, position, and population of the B allele for each marker;
- the GC file (Build 37), which specifies the content in G and C base pairs (ranging from 0% to 100%) of a 1 Mb genomic region surrounding each marker (500 kb each side) and allows an adjustment of LRR to reduce GC-content-caused fluctuation. The file was downloaded from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database (gc5Base.txt.gz).

For each sample, a quality summary is provided, containing mean, median and standard deviation for LRR and BAF, BAF_DRIFT, waviness Factor (WF) and GC-waviness factor (GCWF). WF is a measure of the total signal fluctuation across the genome in the given individual, while GCWF value indicates the signal intensity fluctuation correlating with the GC-content. Since the WF and the GCWF are based on the median absolute deviation of signal intensities, they are less sensitive to extreme values (such as those within CNVs) than a standard deviation measure.

After the CNV calling, samples were discarded if the quality control (QC) summary reported LRR_SD > 0.35, BAF drift >0.002 and waviness factors (WF value) less than -0.04 or higher than 0.04. These thresholds were established taking into account the quality parameters values specified by previous CNV studies (Marenne et al., 2011; Pankratz et al., 2011; Sanders et al., 2011).

Only CNV calls including three or more markers and with a confidence score ≥10 were taken forward. This confidence score is a log Bayes factor and, as for QuantiSNP, it has been suggested that a value of ~10 or larger can be a reliable threshold.

For recently developed SNP arrays with high-density markers, PennCNV tends to split large CNVs (such as those >500kb) in smaller fragments. Therefore, the raw calls were visually inspected and adjacent boundaries were joined when appeared close enough to belong to the same rearrangement.

Trio and quartet base-calling.

PennCNV offers a Father-Mother-Offspring trio calling algorithm which incorporates information from related individuals in a Bayesian approach for *a posteriori* validation. The principle is that pedigree information could be integrated to determine the most likely configuration, improving the number of detected CNVs and the accuracy for boundary mapping.

The -trio or -quartet argument in the command line specifies to use family-based CNV detection algorithm to jointly update CNV status for a father-mother-offspring trio (parents-child) or quartet (parents and two children).

After the individual-based calling step, 92 SLI families with both parents genotyped on the array and passing the quality control (including 167 individuals - 87 affected individuals, 28 unaffected individuals and 52 of unknown affection status), were also analysed as trios or quartets. PennCNV cannot use this approach on families with 3 or more children; therefore, we divided the families into trios and quartets, and then combined the CNV calls together. This method was used to identify *de novo* CNVs.

4.3.4 Generation of a CNV consensus list: BEDTool.

Only CNVs detected by both algorithms QuantiSNP and PennCNV and meeting the quality criteria that we set, were included in subsequent analyses as “high-confidence CNVs”.

A CNV was considered detected by both algorithms if the two independent calls were found to reciprocally overlap for at least 50% of their length. The innermost boundaries of the overlap were then used to define the CNV. This was achieved using the BEDTool “intersectBed”, which returns overlapping “features” between two BED files. BED (Browser Extensible Data) format, as described on the UCSC Genome Browser website, is a tab-limited file which requires at least three columns (chromosome, start position, end position), but can have additional columns. It represents a simple way to annotate genomic features, such as SNPs or structural variants. As QuantiSNP and PennCNV, BEDTools works in a “command line” environment. In cases where one programme identified a large CNV and the other one smaller overlapping CNVs, if the total overlapping segments intersected 50% or more of the larger CNV then the CNV was called as one CNV using the innermost boundaries of the overlapping segments.

4.3.5 Overlap of the CNV consensus list with DGV.

To identify rare and novel CNVs, the calls were compared to CNV loci published in the Database of Genomic Variants (DGV; http://dgv.tcag.ca/) (Macdonald et al., 2013). We used the version of DGV available in January 2012. The list of structural variants present in this catalogue was downloaded from the UCSC Genome Browser (hg19, Jan2012).

Using BEDTool, CNVs were classified as “rare” if they overlapped less than 50% and 5 or less times with CNVs in the DGV, as “novel” if no CNVs were reported in the same genomic region, otherwise they were classified as “common”.

4.3.6 PLINK: CNV burden analyses.

PLINK is a free whole-genome association analysis tool (http://pngu.mgh.harvard.edu/purcell/plink/), running in a command line environment (Purcell et al., 2007). A recent application of PLINK allows CNV burden analyses. The program does not identify CNVs, but offers functions for downstream analysis of CNV data.

In PLINK, CNVs are considered as segments and the analyses require:

-a MAP file, with dummy entries corresponding to the start and stop sites of all "segments";

-a FAM file, which contains 6 fields: FID (Family ID), IID (Individual ID), Paternal ID, Maternal ID, Sex (1=male; 2=female; other=unknown) and phenotype (1= unaffected, 2 =affected, 0 or -9 missing);

-a CNV list, a text file with 8 columns: FID, IID, CHR (Chromosome), BP1 (Start position, bp), BP2 (End position, bp), TYPE (number of copies), SCORE (confidence score associated with the variant), SITES (number of probes included by the variant).

By default, a summary of the PLINK analysis is reported in a file named "plink.cnv.summary", that represents a count of CNVs, in cases and controls, assuming that the individuals are unrelated.

The arguments *--cnv-indiv-perm* and *--mperm 10000* allow a set of global tests of CNV burden in cases versus controls to be performed. In this test (1-sided), the number of CNVs (RATE), the proportion of samples with one or more segments (PROP), the total kb length spanned per person (TOTKB) and the average segment size per person (AVGKB) are the metrics compared between cases and controls and evaluated by permutation.

We also supplied a "gene.list" file (Build 37) to add an extra test, that takes into account the number of genes spanned by CNVs (GRATE), the number of CNVs with at least one gene (GPROP) and the number of genes per total CNV kb (GRICH).

4.4 Validation of CNVs by Real-Time PCR.

The presence of selected microdeletions and microduplications was validated by quantitative PCR (qPCR) using iQ SYBR Green Supermix (Bio-Rad). Briefly, SYBR Green is a non-specific dye that emits fluorescence when it intercalates with double stranded DNA. In a qPCR experiment, the fluorescence is measured after each extension step and this allows the monitoring of the increasing amount of DNA produced during the PCR reaction. The analysis of qPCR data is based on the threshold Cycle (Ct), which represents the cycle at which the fluorescence passes the threshold level, within the exponential phase, and it is a relative measure of the amount of target in the reaction.

All the primers used for the CNV validation were designed with Primer3 (http://bioinfo.ut.ee/primer3-0.4.0/primer3/) and are listed in **Table 4.2**. Oligonucleotide primers were designed taking into account an optimal primer length of 20-22 nucleotides, a GC content of 40–60% and an optimal PCR product size of 90-140 bp. In order to evaluate the PCR efficiency, a standard curve was set up for each primer pair, using three replicates of a control DNA and 5 template concentrations (five DNA template amounts deriving from a 1:4 serial dilution). The primers used for the CNV validation were selected to have a PCR efficiency in the range of 90%-110%. We made sure the concentration of the DNA we would be using occurred within the serial dilution range.

Each reaction was set up in triplicate using the following conditions:

qPCR Reaction MIX	**Amount (µl)**
DNA template	5
iQ SYBR Green Supermix	12.5
Primer F (5µM)	1
Primer R (5µM)	1
H_2O	5.5
Final volume	25

For the relative quantification, the DNA template was used with a concentration of 5 ng/µl.

The qPCR program included three main stages:

- Initial denaturation: incubation at 95°C for 4’ 30”;
- Amplification: 40 cycles of incubation at 95°C for 30”, at 60°C for 30”, at 72°C for 30”;
- Melting curve: incubation at 95°C for 1’, followed by an incubation at 55°C for 30”, repeated for 81 times with an increase of 0.5°C each cycle.

The melting curve analysis, also called dissociation curve, consists of a step in which the temperature is gradually increased while the fluorescence is constantly monitored. When the temperature is high enough, the double strands of DNA fragments are denatured and the SYBR Green dye dissociates from the double stranded DNA, causing a decrease in fluorescence. Analysis of melting curves allows the detection of primer-dimers or other non-specific PCR products that could reduce the PCR efficiency and give spurious fluorescence signals. Primer-dimers usually have lower Tm compared to the PCR products, since they have a smaller size. The melting temperature (Tm) depends on primer features (such as sequence complementarity and G-C composition) and reaction conditions.

For each sample, the qPCR data were compared against a control gene (*ZNF423*) and a control subject, genotyped on the same SNP array and predicted to have two normal copies of the tested region. The number of copies of each amplified fragment was calculated using the $2^{-\Delta\Delta Ct}$ method (Livak and Schmittgen, 2001):

$$\Delta Ct_{\text{sampleDNA}} = Ct_{\text{fragment of interest}} - Ct_{\textit{ZNF423}}$$

$$\Delta Ct_{\text{controlDNA}} = Ct_{\text{fragment of interest}} - Ct_{\textit{ZNF423}}$$

$$\Delta\Delta Ct = \Delta Ct_{\text{sampleDNA}} - \Delta Ct_{\text{controlDNA}}$$

$$\text{Copy number} = 2 \text{ x } 2^{-\Delta\Delta Ct}$$

4.5 Analysis of the candidate gene *ZNF277.*

4.5.1 Validation of the *ZNF277* microdeletion in the discovery pedigree G4.

All the members of the discovery family G4 were analyzed for the presence of the *ZNF277* microdeletion (minimum predicted size: 4,153 bp, chr7:111,955,948-111,960,100, hg19) by qPCR, except the sister G4_5, for whom insufficient DNA was available. Four primer pairs were used (**Table 4.2**): chr7_fg1, mapping to intron 1 of *DOCK4*, and chr7_fg4, mapping to intron 8 of *ZNF277*, were used as control probes for the regions predicted to be not deleted, whereas chr7_fg2 and chr7_fg3, mapping to intron 4 and exon 5 of *ZNF277* respectively, were used to confirm the predicted deletion. The sister G4_5 was instead tested using a PCR assay, described in the next paragraph.

4.5.2 Breakpoint characterization of the *ZNF277* microdeletion in the discovery pedigree.

To further confirm the qPCR results, two PCR assays were performed in all the members of family G4. A primer pair (ZNF_break) spanning the *ZNF277* microdeletion breakpoints was used to show the presence of the microdeletion, in the heterozygous or homozygous form. A primer pair that amplifies exon 5 of *ZNF277* (*ZNF277*_x5) was used instead to detect the presence of at least one non-deleted allele.

These PCR assays were performed using the following conditions:

PCR mix	**Amount (µl)**	**Program**	
NH_4 Buffer (10X)	2.5	Initial Denaturation	95°C x 4 minutes
$MgCl_2$ (50 mM)	1.25	Amplification (30 cyles)	95°C x 30 seconds
dNTPs (10 µM)	0.5		55°C x 30 seconds
Primer F (10 µM)	0.5		72°C x 30 seconds
Primer R (10µM)	0.5	Final extension	72°C x 7 minutes
BioTaq:Pfu (9:1)*	0.1		
H_2O	14.65		
Template DNA (5 ng/µl)	5		
Final volume	25		

* BioTaq DNA Polymerase 5 u/µl (*Bioline*), Pfu DNA polymerase 2.5 u/µl (*Thermo Scientific*).

The PCR products were visualized using a UV transilluminator after electrophoresis on a 2% agarose gel and SYBR safe staining (*Invitrogen*). The 1kb Plus DNA ladder (*Invitrogen*) product size was loaded to check the PCR fragment size.

PCR purification.

Prior to sequencing, the PCR products were purified with Exonuclease I (ExoI) and Shrimp Alcaline Phospatase (SAP). The enzyme ExoI removes single-strand DNA (primers or intermediate products), the enzyme SAP removes the unconsumed dNTPs remaining in the PCR mixture. The PCR cleanup was performed adding the Exo-SAP reaction mix to the PCR products, followed by a two-step incubation:

EXOSAP mix	**Amount (µl)**	**Program**	
SAP Buffer (10X)	1	Treatment	37°C x 30 minutes
SAP enzyme	1	Enzymatic inactivation	80°C x 20 minutes
ExoI enzyme	0.1		
H_2O	0.9		
PCR product	7		
Final volume	10		

Sanger Sequencing reaction.

Sanger sequencing was performed to verify the predicted breakpoints of the *ZNF277* microdeletion in the G4 family, using primers flanking the predicted deletion boundaries (ZNF_break). The reaction was performed using the ABI PRISM BigDye v3.1 Terminator Cycle Sequencing (*Life Technologies*).

Sequencing mix	**Amount (µl)**	**Program**	
Big Dye terminator buffer (5X)	1.75	Initial Denaturation	95°C x 1 minute
Big Dye terminator	0.5	Amplification (35 cycles)	95°C x 10 seconds
Primer (F or R, 10µM)	0.5		50°C x 10 seconds
H_2O	5.25		60°C x 4 minutes
PCR product	2		
Final volume	10		

Ethanol-EDTA Precipitation of Sequencing Reactions.

The products of sequence reactions were precipitated using the following procedure.

1) A “precipitation mix” was prepared and added to each sample (volume of the sequence reaction: 10 µl). This solution was composed by 50 µl of 100% EtOH, 2 µl of NaAc (sodium acetate) 3M pH 5.2, 2 µl of EDTA (ethylenediaminetetraacetic acid) 125 mM.
2) The samples were spun in a centrifuge for 30 minutes at 3000 *g* (4°C).

3) The surnatant was discarded and the microplate was spun inverted for 10 seconds at 190 *g*.
4) The pellet was rinsed with 70 µl of 70% EtOH.
5) The samples were spun in a centrifuge for 15 minutes at 1650 *g* (4°C).
6) The surnatant was discarded and the microplate was spun inverted for 10 seconds at 190 *g*.
7) The pellet was dried at room temperature and stored away from direct light.

The purified sequencing products were sent to the DNA sequencing facility at the Zoology Department (Oxford, UK).

4.5.3 IBD and homozygosity analysis in the discovery pedigree G4.

In order to test whether G4 was a consanguineous family, we used the Pairwise IBD Estimation tool in PLINK (http://pngu.mgh.harvard.edu/~purcell/plink/ibdibs.shtml) to estimate the proportion of IBD in each pair of individuals within the same family of the SLIC cohort. The analysis was performed using genome-wide SNPs genotyped on the on Illumina Human OmniExpress beadchip. Given IBS information, this statistic evaluates the IBD alleles in order to estimate whether two individuals look more similar than expected by chance in a random sample. The proportion of IBD is calculated as PI_HAT =P(IBD=2)+0.5*P(IBD=1). The expected values for a pair of unrelated individuals is zero, for a parent-child pair or a sibling pair it is 0.5 and for a pair of half-siblings it is 0.25.

SNP data were used to analyse the haplotypes of a 10 Mb region encompassing the gene *ZNF277* (chr7:106790042-11673037, hg19), in family G4. This analysis was performed using the Haplotyping tool of Merlin (http://www.sph.umich.edu/csg/abecasis/merlin/tour/haplotyping.html). The command requires a *.ped* file, that provides pedigree information (a family identifier, an individual identifier, a link to each parent and finally an indicator of each individual's sex) and various types of genetic data, including genotypes and phenotypes; a *.map* file, which provides the chromosomal locations of the markers used in the analysis; and a *.dat* file, which describes the *.ped* file structure. The output file lists the two haplotypes for each individual and indicates the location of recombination events.

4.5.4 Screening of Larger Cohorts: SLIC, IMGSAC and controls.

The same PCR assays used to confirm the presence of the *ZNF277* microdeletion in family G4 were used in a PCR-based screening for *ZNF277* microdeletions of three separate cohorts: a cohort of families containing individuals with SLI, a cohort of families containing individuals with ASD and a control cohort. Amplification with the primer pair spanning the deletion breakpoints identified *ZNF277* microdeletion carriers, while the primer pair amplifying exon 5 of *ZNF277* was subsequently used to test whether the identified microdeletions were in the heterozygous or homozygous form.

The SLI screening cohort consisted of DNA from 1234 individuals from 322 SLIC families (545 parents, 318 SLI probands, 371 sibs). This cohort included the 512 individuals who comprised the CNV study but included many additional SLI subjects (144 additional probands) and their family members (550 individuals).

The ASD cohort consisted of DNA from 1021 individuals from 252 IMGSAC multiplex families (454 parents, 412 affected children, 155 affected and unaffected sibs).

The control cohort consisted of DNA from 224 non-related UK Caucasian blood donors from the ECACC Human Random Control (HRC) panel [http://www.hpacultures.org.uk/products/dna/hrcdna/hrcdna.jsp]. In addition, we had access to sequence data from 130 unrelated Caucasian samples through an in-house project at the Wellcome Trust Centre for Human Genetics–the 500 Whole-Genome Sequences Project [WGS500 Consortium] (Palles et al., 2013).

We used the two-tailed Fisher's exact test (1 degree of freedom) to test whether the allelic frequency of *ZNF277* microdeletions was significantly different between SLI probands and control individuals.

4.5.5 Gene Expression Evaluation for *ZNF277, DOCK4* and *IMMP2L.*

RNA samples were not available for the discovery individuals of family G4. We therefore chose to examine expression levels of *ZNF277*, *DOCK4* and *IMMP2L* by qPCR in cDNA derived from lymphoblastoid cell lines (LCLs) from ten individuals belonging to 4 ASD families, five of whom carried a heterozygous *ZNF277* microdeletion, and in cDNA derived from blood from the parents of a single Dutch multiplex ASD family (15-0084) previously described (Pagnamenta et al., 2010) with a *IMMP2L-DOCK4* microdeletion (chr7:110876742-111470446, hg19).

<u>RNA extraction from LCLs and cDNA synthesis.</u>

EBV-transformed peripheral LCLs were grown in RPMI 1640 media (Sigma) supplemented with 10% fetal bovine serum (PAA), L-glutamine (final concentration 2 mM) and penicillin (Sigma) (500 U/ml) and streptomycin (Sigma) (5 μg/ml). When the cells reached an amount of approximately 1 x 10^7, RNA was extracted using the RNeasy Mini kit (QIAGEN, Crawley, UK), according to the manufacturer's suggested protocol.

1. After determining the number of cells, the amount correspondent to 1 x 10^7 cells was spun for 5 min at 150 x g, then the supernatant was removed by aspiration.

2. 600 µl of lysing Buffer RLT was added to the pelleted cells mixed by vortexing or pipetting, ensuring that no cell clumps were visible before proceeding to the next step. Incomplete homogenization leads to significantly reduced RNA yields and can cause clogging of the RNeasy spin column.
3. We added 1 volume (600 µl) of 70% Ethanol to the homogenized lysate, and mixed well by pipetting.
4. 700 µl of the sample was transferred to an RNeasy spin column placed in a 2 ml collection tube. Each column was spun for 15 s at 8000 x g (10,000 rpm), the flow-through was discarded.
5. 700 µl Buffer RW1 was added to the RNeasy spin column, that was spun for 15 s at 8000 x g (10,000 rpm), to wash the spin column membrane. The flow-through was discarded.
6. 500 µl Buffer RPE was added to the RNeasy spin column that was spun for 15 s at 8000 x g (10,000 rpm), to wash the spin column membrane. The flow-through was discarded.
7. 500 µl Buffer RPE was added to the RNeasy spin column that was spun for 2 minutes at 8000 x g (10,000 rpm), to wash the spin column membrane. The long centrifugation dries the spin column membrane, ensuring that no ethanol is carried over during RNA elution. Residual ethanol may interfere with downstream reactions.
9. After placing the RNeasy spin column in a new collection tube, the column was spun for 1 minute at 8000 x g, to dry the membrane. This step was performed to eliminate any possible residue of Buffer RPE, or residual flow-through remained on the outside of the RNeasy spin column.

10 The RNeasy spin column was then transferred to a new collection tube, then 30 µl of RNAse-free water was added to the column and spun for 1 minute at 8000 x g. The collected RNA was conserved at -80°C.

cDNA was synthesized using the QuantiTect Reverse Transcription kit (QIAGEN, Crawley, UK), using approximately 1 µg of RNA as template, according to the manufacturer's protocol.

qPCR analyses.

All the primers were designed with Primer3 (http://bioinfo.ut.ee/primer3-0.4.0/primer3/), as described before, and are listed in **Table 4.2**. Moreover, to ensure the specific amplification of cDNA, the forward and the reverse primers of each pair were designed in different adjacent exons or across exon-exon boundaries. The qPCR reaction was performed as stated in **section 4.4**. Two primer pairs were designed for each gene, *ZNF277* (NM_021994.2), *DOCK4* (NM_014705.3), *IMMP2L* (NM_032549.3). Four housekeeping were tested (*ACTB, B2M, TFRC* and *GUSB*). Among these, the housekeeping gene *GUSB* (NM_000181.3) had expression levels more similar to our genes of interest and was used as a normalizer. Expression levels were also normalized against a control individual. The $2^{-\Delta\Delta Ct}$ method was applied to estimate the difference in the expression of the three genes between samples.

Statistical significance was calculated with the Student T-test, assuming unequal variance between the two independent sample groups for the expression analysis.

4.6 Analysis of the candidate gene *DPYD.*

4.6.1 SLI probands.

A screening for the splicing variant rs3918290 was performed in a group of 166 independent SLI cases. The screening was performed by PCR amplification of the fragment containing exon 14 and exon-intron boundaries (primers and conditions are described in **Table 4.2**), followed by Sanger sequencing, as previously described (**section 4.5.2**).

4.6.2 Autism Cohorts: Italian, IMGSAC and AGP families.

A screening for the splicing variant rs3918290 was performed in two groups of cases and a group of control individuals.

The first group of cases was formed by 231 Italian simplex families, each with one individual affected by ASD (413 parents, 231 probands, 48 unaffected sibling). The second group of cases was constituted of 224 multiplex families belonging to the IMGSAC Consortium (365 parents, 224 probands, 161 affected siblings, 115 unaffected siblings). The third group of cases was part of the cohort of families collected by the AGP Consortium. Subjects from all phenotypic classes (strict, broad, and spectrum) were included in the analyses for *DPYD*. These individuals had been previously genotyped on the Illumina Human 1M-single and Illumina Human 1M-Duo BeadChip arrays, which also include the SNP rs3918290. The association analyses were performed on the individuals who have been successfully genotyped for rs3918290 and for another 99 independent SNPs of the surrounding region (chr1:97499599-98599144), including *DPYD* and *MIR137*.

The group of controls genotyped for rs3918290 was formed by 449 unrelated and healthy Italian individuals. DNA was obtained from blood or buccal-swabs.

4.6.3 Genotyping by endonuclease restriction analysis.

We used the bioinformatic program INSIZER (http://zeon.well.ox.ac.uk/git-bin/insizer, not available anymore) to find a restriction endonuclease that could specifically discriminate between the two allelic variants of the splice donor site at intron 14 of *DPYD*. We interrogated the sequence of a PCR fragment including exon 14 (413 bp) and we found that the enzyme HpyCH4IV (NEB, *New England Biolabs*) recognizes the site:

5'. . . A▼C G T . . . 3'
3'. . . T G C▲A . . . 5'

In our PCR fragment, these 4 bp can be found only at the junction between exon 14 and intron 14, therefore, the target site is unique in our sequence. When the splice donor site is not mutated, the enzyme HpyCH4IV cuts the PCR fragment, producing two restriction fragments of different lengths (278 bp and 135 bp) that can be easily distinguished on an agarose gel. When the base G is substituted by the base A (rs3918290), the enzyme does not recognize the target site and does not cut the sequence. Therefore, the distinct gel band patterns produced by the digestion with HpyCH4IV allowed the genotyping of the variant rs3918290 in the Italian and IMGSAC ASD families and the control samples.

The PCR fragment containing exon 14 was amplified with the primers *DPYD*_x14F and *DPYD*_x14R and the conditions are reported in the **Table 4.2**. Subsequently, the digestion reaction was assembled using with the following conditions.

Digestion mix	**Amount (µl)**	**Program**	
NEB buffer 1 (10X)	1.5	Digestion	37°C for 3 hours
HpyCH4IV enzyme (10 U/µl)	0.1	Heat inactivation	65°C for 2 minutes
H_2O	10.4		
PCR product	3		
Final volume	15		

The restriction fragments were separated by electrophoresis on a 2% agarose gel, using GelRed staining (*Biotium*). The digestion profiles were visualized with a UV transilluminator. The 100 bp DNA ladder (*NEB*) product size was loaded to check the restriction fragment size.

4.6.4 Mutation screening of *DPYD* gene: primers and PCR conditions.

The 23 coding regions and the flanking intronic regions of the isoform 1 of *DPYD* (NM_000110.3), the coding region of exon 6 of isoform 2 (NM_001160301.1), and regulatory elements at 5' of the gene (chr1:98386616-98386699, hg19) were analysed during a mutation screening. These regions were amplified separately using the primers and the conditions listed in **Table 4.2.** All the primers were designed using Primer3 (http://bioinfo.ut.ee/primer3-0.4.0/primer3/).

The PCR assays were performed using the kit provided with AmpliTaq Gold DNA Polymerase (*Applied Biosystem*). AmpliTaq Gold is a *Hot Start* polymerase, that is inactive at room temperature and is activated during the initial denaturing step at 95°C. The reactions were set up in a final volume of 15 µl, with ~30 ng of template DNA (3 µl of 10 ng/µl).

The PCR conditions were optimized for each primer pair used in the screening. For some of them, we used a traditional PCR program, with one annealing temperature (Ta):

Program	
Initial Denaturation	95°C for 15 minutes
Amplification (30 cyles)	95°C for 30 seconds
	Ta°C for 30 seconds
	72°C for 30 seconds
Final extension	72°C for 7 minutes

For other fragments, in order to increase the specificity and the yield of the reaction, we used a Touch-Down PCR program (TD) (Korbie and Mattick, 2008). In a TD program, the initial Ta (1) is higher to ensure a specific annealing of the primers to the template, then it is progressively decreased until it reaches a second, lower Ta (2), which is maintained constant for remaining amplification cycles.

TD Program	
Initial Denaturation	95°C for 15 minutes
Touch-Down (Ta decreases 0.5°C at each cycle)	95°C for 30 seconds
	Ta_1°C for 30 seconds
	72°C for 30 seconds
Amplification (30 cyles)	95°C for 30 seconds
	Ta_2°C for 30 seconds
	72°C for 30 seconds
Final extension	72°C for 7 minutes

The amplification of the fragment containing exon 1 was particularly problematic, since the region has a high GC content. The GC pair has a higher number of hydrogen bonds compared to the AT pair, therefore, GC-rich stretches are more stable and require a higher melting temperature. Since Dimethyl sulfoxide (DMSO) improves the denaturation of the template DNA, it was added to the PCR mix to help the reaction and overcome the problem of non-specific annealing.

The PCR products were visualized with a UV transilluminator after electrophoresis on a 2% agarose gel and GelRed staining (*Biotium*). The 100 bp DNA ladder (*NEB*) product size was loaded to check the PCR fragment size.

4.6.5 PCR purification.

The PCR products were purified with Exonuclease I (ExoI) and Shrimp Alcaline Phosphatase (SAP), as described in section 4.5.2.

For exon 1, the PCR product with the correct size was purified with the GEL/PCR Extraction & Purification Kit (*Fisher Molecular Biology*), using the manufacturer's instructions.

1. PCR products were loaded on a 2% agarose gel and separated by electrophoresis.
2. The band of the correct size was excised from the gel with a clean scalpel and transferred into a microcentrifuge tube (the kit requires a maximum volume of the gel slice of 300 mg).
3. 500 µl of FSDF Buffer were added to the sample and mixed by vortexing.
4. The sample was incubated at 55°C for 10-15 minutes and vortexed every 2-3 min until the gel slice was completely dissolved.
5. After cooling down at room temperature, 800 µl of the sample mixture were transferred to a FSDF Column, inserted in a collection tube.
6. The column was spun at full speed in a microcentrifuge for 30 seconds, then the flow-through was discarded.
7. 750 µl of Wash Buffer (after addition of ethanol) were added to the FSDF Column.
8. The column was spun at full speed in a microcentrifuge for 30 seconds, then the flow-through was discarded.
9. The FSDF Column was placed to a new microcentrifuge tube.
10. 40 µl of Elution Buffer were added to the membrane of the FSDF Column and left for 2 minutes.
11. DNA was eluted spinning the tube at full speed in a microcentrifuge for 2 minutes.

4.6.6 Sequencing Reaction.

Sanger sequencing was performed with the ABI PRISM BigDye Terminator v3.1 Cycle Sequencing kit (*Life Technologies*), using the following conditions.

Sequencing mix	**Amount (µl)**	**Program**	
Big Dye terminator buffer (5X)	1.75	Initial Denaturation	96°C for 1 minute
Big Dye terminator	0.5	Amplification (25 cycles)	96°C for 10 seconds
Primer (F or R, 10µM)	0.16		50°C for 5 seconds
H_2O	6.59		60°C for 4 minutes
PCR product	2		
Final volume	10		

The primers used for the sequencing reactions were the same used for PCR amplification, except for the fragment amplifying exon 1, that needed an internal primer (5’-attaaaggccagtccccaga-3’) and the addition of DMSO 5% to the sequencing reaction mix (final reaction volume=10 µl).

4.6.7 Ethanol Precipitation of Sequencing Reactions.

The products of sequence reactions were precipitated using the following procedure.

1. A "precipitation mix" was prepared and added to each sample (volume of the sequence reaction: 10 µl). This solution was composed by 10 ul of H_2O, 2.5 volumes of 100% EtOH (55 µl), 1/10 of the volume of sodium acetate (NaAc) 3M pH 5.2 (2 µl).
2. The samples were spun in a centrifuge for 30 minutes at 3000 rpm (4°C).
3. The surnatant was discarded.
4. The pellet was rinsed with 70 µl of 70% EtOH.
5. The samples were spun in a centrifuge for 10 minutes at 3000 rpm (4°C).
6. The surnatant was discarded.
7. The microplates were spun upside down for 1 minute at 300 rpm, to remove traces of EtOH or NaAc.
8. The pellet was dried at room temperature and stored away from direct light.

The purified sequencing products were then resuspended in 15µl of Injection Solution (DNA Sequencing Reaction Cleanup kit, *Millipore*) by pipetting up and down several times and/or using a microplate shaker. The sequences products were run on the ABI PRISM 3730 DNA analyser (*Applied Biosystem*).

4.6.8 Sequence Analyses and Prediction tools.

The sequences were analysed and compared to the reference sequence using the software Sequencher 5.0 (*Gene Code Corporation*).

We used two online bioinformatic tools to predict the possible impact of non-synonymous coding SNPs on the structure and function of the human protein DPD: PolyPhen-2 (Polymorphism Phenotyping v2, http://genetics.bwh.harvard.edu/pph2) (Adzhubei et al., 2010) and SIFT (Sorting Tolerant From Intolerant, http://sift.jcvi.org) (Kumar et al., 2009). SIFT is a sequence-based algorithm that uses sequence homology to predict the effect of amino-acid replacements, assuming that important positions in a protein sequence have been conserved throughout evolution. Polyphen-2 instead is prediction algorithm, that incorporates sequence conservation information and protein structure annotations to predict the impact of the non-synonymous change. For SIFT, the score ranges from 0 (damaging) to 1 (neutral); for Polyphen-2, the score ranges from 0 (neutral) to 1 (damaging).

4.6.9 Association Analyses with PLINK.

In order to test whether there is a preferential transmission of the allele A of the splicing variant rs3918290 to individuals with ASD compared to the reference allele G, we performed a Transmission Disequilibrium Test (TDT) on 8,363 individuals of the AGP cohort (5200 males and 3163 females). In this large cohort, rs3918290 was successfully genotyped in 2,681 affected individuals and their parents, who were all included in this family-based association analysis. The test was performed using PLINK v1.07 (http://pngu.mgh.harvard.edu/~purcell/plink), that required a *.map* file with the SNP rs3918290, a *.ped* file, which specified the relationships among the individuals and a phenotype file, that specified their affection status. This analysis returned a *p-value* calculated with a χ^2 test with one degree of freedom.

A TDT test was also performed for the SNPs included in the genomic region chr1:97500000-98598500 (hg19), that covers the genes *DPYD* and *MIR137*. We selected a subset of 99 independent SNPs (not in LD) out of 900 SNPs present on the Illumina 1M array in this region, using the “indep” command from the PLINK analysis suite (http://pngu.mgh.harvard.edu/~purcell/plink).

SNPs that had missing genotypes in more that 10% of the samples and SNPs that were not in Hardy-Weinberg equilibrium were also excluded.

4.6.10 Analysis of language items of ADI-R and IQ scores in the AGP cohort.

Since we hypothesized the involvement of *DPYD* in SLI, we investigated a possible effect of the variant rs3918290 on language development of ASD individuals of the AGP cohort. The ASD phenotype of the individuals included in this cohort was assessed using the diagnostic tools ADOS-G and ADI-R. As previously described, ADI-R is structured interview that includes some questions regarding speech development and language abilities. The analysis was performed using available information for age of first single words (item M9_12_8), age of first single phrases (item M10_13_9) and the overall level of current language (item M30_19_14). Age of first single words and age of single phrases were expressed in months. Assuming that the ability of saying and using few meaningful words is a milestone that should be achieved within 24 months, we classified the children in two categories, depending on the age at which they acquired this skill (if the age of first words was <24 months, the child had a “normal” speech onset; if it was ≥ 24 months, then was considered “delayed”). We also classified the children in two classes depending on age of first phrases. The ability to formulate simple phrases should be acquired within 36 months, therefore we classified subjects with age of first phrases ≥ 36 months as “severe language-delayed” and subjects with age of first phrases < 36 months as “normal”. A similar item is represented by the overall level of current language, which evaluates the spontaneous use of social language. Individuals with the ability of formulating phrases with more than 5 words were reported to have “normal” language development, whereas individuals unable of formulating phrases with more than 5 words were considered to have a “severe delay” in language development.

For each sub-phenotype, we compared the number of ASD cases carrying the splicing variant rs3918290 with the ASD cases *wild-type* for that SNP, using a χ^2 test.

We also compared verbal, performance and full scale IQ scores of the AGP cases carrying the splicing variant rs3918290 with those of the affected individuals *wild-type* for the SNP. The distributions of these scores were analysed with a Wilcoxon Whitney Mann test.

Table 4.2 PRIMERS

control gene **validation by RealtimePCR**

primer name	sequence (5'-3')	primer size	product size (bp)	fragment position (Hg19)
ZNF423_RT_F	ccaacacagtcacaaacaggaa	22	97	chr16:49529702-49529798
ZNF423_RT_R	gcctaagcaacagagaatggag	22		

BP1-BP2 deletion **validation by RealtimePCR**

primer name	sequence (5'-3')	primer size	product size (bp)	fragment position (Hg19)	in the figures
chr15_del_2F	ACAGTACAGGGCTGGAGTTTCA	22	97	chr15:22868848-22868944	TUGCP5_ex20
chr15_del_2R	TCATGGATGGTTGACAGATACC	22			
chr15_del_3F	CTGAACAAAGAGCTGCGAAGTA	22	95	chr15:23045509-23045603	NIPA1_ex5
chr15_del_3R	GGAGGAATTTGGTATTCTGGTG	22			
chr15_del_4F	ggaaacatccttgccagtgt	20	128	chr15:23,661,907-23,662,034	intergenic-fg
chr15_del_4R	aaagcagcaatacattctctcc	22			

BP4-BP5 duplication **validation by RealtimePCR**

primer name	sequence (5'-3')	primer size	product size (bp)	fragment position (Hg19)	in the figures
chr15_dupl_1F	aatcaggcagaccttgctctt	21	104	chr15:31,908,496-31,908,599	fg1
chr15_dupl_1R	tccagttaattcctcagcatgt	22			
chr15_dupl_2F	gctgattctggcctggtagtaa	22	94	chr15:32,085,461-32,085,554	fg2
chr15_dupl_2R	cttgtgtgatgctcttggacac	22			
chr15_dupl_3F	agtagccaaggaagtgaagtgc	22	119	chr15:32,403,945-32,404,063	fg3
chr15_dupl_3R	CATCTGGGAAACGAACAGTC	20			
chr15_dupl_4F	TAAGGTTCCCTTGGATGATCTG	22	119	chr15:32,929,848-32,929,966	fg4
chr15_dupl_4R	TTCCTCTTTCTGATGGTCTCCT	22			

***ZNF277* deletion** **validation by RealtimePCR**

primer name	sequence (5'-3')	primer size	product size (bp)	fragment position (Hg19)
chr7_Fg1_F	cagtgtcagggtatgctttcct	22	95	chr7:111,845,306-111,845,400
chr7_Fg1_R	tacagggctgctgagtgataaa	22		
chr7_Fg2_F	cagaaatggttgcacgatagaa	22	136	chr7:111957488-111957623
chr7_Fg2_R	ctgctctaatcaaggtggttcc	22		
chr7_Fg3_F	agtttcctcccatccagacttt	22	127	chr7:111958153-111958279
chr7_Fg3_R	TGGTATCATTTCGTTCTTGCTG	22		
chr7_Fg4_F	cctggtttgggaatgataagaa	22	121	chr7:111977177-111977297
chr7_Fg4_R	gttaaatctggcggtagacagg	22		

***ZNF277* screening** **PCR**

primer name	sequence (5'-3')	primer size	product size (bp)	fragment position (Hg19)	PCR conditions
ZNF277_break_F	aattgcccagcatccaatta	20	466 (deleted)	chr7:111941453-111963297	**2.5 Mg^{2+}** 95°C-55°C-72°C (30",30", 30") for 30 cycles
ZNF277_break_R	tgacattttcctggggatct	20			
ZNF277_x5_F	tgcaaaaacaggaatggaga	20	434	chr7:111958084-111958517	
ZNF277_x5_R	tacccctctgctgcttagag	20			

expression analyses **RealtimePCR**

primer name	sequence (5'-3')	primer size	product size (bp)
GUSB_x8F	cacctagaatctgctggctact	22	93
GUSB_x9R	agagttgctcacaaaggtcaca	22	
IMMP2L_x2F	ctgtgtggcaagagtagaagga	22	100
IMMP2L_x3R	cctcactttccagtggttcaaa	22	
IMMP2L_x3F	ccactggaaagtgaggaattt	21	136
IMMP2L_x4-5R	tttgtgtcctatggttctgaca	22	
RT_DOCK4_ex47F	AGCTGTCAACCGATATTCTTCC	22	128
RT_DOCK4_ex48R	AGCTTGAGGTAGATGGACTTGG	22	

primer name	sequence (5'-3')	primer size	product size (bp)
RT_DOCK4_ex17F	CATGAAGGCCACAAAGGAGT	20	130
RT_DOCK4_ex18R	AGAGAGACAGCCAGTGATCTTG	22	
ZNF277_x1F	GAATGCAGGAAGACCGTGAT	20	102
ZNF277_x2R	GACTTTCTGGCAGGGAAAGC	20	
ZNF277_x3F	AATTAATTCCACTGCTCCATTTG	23	129
ZNF277_x5R	TGCTGTTGTTCCAGAATTTCTCT	23	

DPYD screening

					PCR conditions		
primer name	sequence (5'-3')	primer size (bp)	product size (bp)	exon	[Mg^{2+}]	PCR Program	primer for sequencing
DPYD_x1_F	gcggactgcttttacctttg	20	679	1	2 mM +	TD=65°C-60°C (30",30", 30") 10+30 cycles	attaaaggccagtccccaga
DPYD_x1_R	aggcttcctgaaatctcttcc	21			DMSO 5%		DMSO 5%
DPYD_x2_F	ggggcctgtaagaggaaatc	20	641	2	2.5 mM	95°C-55°C-72°C (30",30", 30") for 30 cycles	2F
DPYD_x2_R	aatcacggctgtactttaatacct	24					
DPYD_x3_F	ttgataacgaaactccactttga	23	426	3	2.5 mM	95°C-55°C-72°C (30",30", 30") for 30 cycles	3R
DPYD_x3_R	tgaatggtggcaatgaactc	20					
DPYD_x4_F	aggagtgccaaagatgaaaca	21	363	4	2.5 mM	95°C-55°C-72°C (30",30", 30") for 30 cycles	4F
DPYD_x4_R	tggatttgctaagacaagctg	21					
DPYD_x5_F	tgtttgtcgtaatttggctgt	21	327	5	2.5 mM	95°C-55°C-72°C (30",30", 30") for 30 cycles	5R
DPYD_x5_R	tgggtatcaacagagcacca	20					
DPYD_x6_iso1_F	gccataactcctcatctacttgac	24	477	6_isoform 1	2.5 mM	95°C-55°C-72°C (30",30", 30") for 30 cycles	6R
DPYD_x6_iso1_R	ccatctgtgagcctgaagtt	20					
DPYD_x7_F	aagattggtcaaagattggtca	22	272	7	2.5 mM	95°C-55°C-72°C (30",30", 30") for 30 cycles	7R
DPYD_x7_R	tgcttctgcctgatgtagctt	21					
DPYD_x8_F	cactggcttttcttctgcatt	21	392	8	2.5 mM	TD=62°C-55°C (30",30", 30") 14+30 cycles	8F
DPYD_x8_R	ggcagtcattcttctggatattg	23					
DPYD_x9_F	ttgatttgcttacagatgttttcc	24	351	9	2.5 mM	95°C-55°C-72°C (30",30", 30") for 30 cycles	9F
DPYD_x9_R	aaggttgggtgtgagagctg	20					
DPYD_x10_F	tggaaaactgcaagatgcaa	20	377	10	2.5 mM	95°C-55°C-72°C (30",30", 30") for 30 cycles	10F
DPYD_x10_R	gacaatttcaacattctagcgatt	24					
DPYD_x11_F	tggtgaaagaaaaagctgcat	21	427	11	2.5 mM	95°C-55°C-72°C (30",30", 30") for 30 cycles	11F
DPYD_x11_R	tgaaaaacaattccctgaaagc	22					
DPYD_x12_F	tgtgaggtgtaaagttaagtcagtg	25	589	12	2.5 mM	95°C-55°C-72°C (30",30", 30") for 30 cycles	12F
DPYD_x12_R	tggcccaatttttaatcaact	21					

DPYD screening

					PCR conditions		
primer name	sequence (5'-3')	primer size (bp)	product size (bp)	exon	[Mg^{2+}]	PCR Program	primer for sequencing
DPYD_x13_F	actcttcatactgcctttgaaatta	25	612	13	2.5 mM	95°C-55°C-72°C (30",30", 30") for 30 cycles	13F
DPYD_x13_R	tatatgcctgccccttcttc	20					
DPYD_x14_F	aaaaatgtgagaagggacctca	22	413	14	2.5 mM	95°C-55°C-72°C (30",30", 30") for 30 cycles	14F
DPYD_x14_R	tgcatcagcaaagcaactg	19					
DPYD_x15_F	taattccaaagccccaaatg	20	498	15	2.5 mM	95°C-55°C-72°C (30",30", 30") for 30 cycles	15F
DPYD_x15_R	aggtagtgtgtgaaatccaagg	22					
DPYD_x16_F	gctgtgatgcagaaaacagaa	21	342	16	2.5 mM	95°C-55°C-72°C (30",30", 30") for 30 cycles	16F
DPYD_x16_R	aaacaatgcagacctggaagt	21					
DPYD_x17_F	tcttgcacgtctccagcttt	20	369	17	2.5 mM	TD=62°C-55°C (30",30", 30") 14+30 cycles	17R
DPYD_x17_R	tcctgtgtttgtgggatcaa	20					
DPYD_x18_F	aagagctgcatgaaaatgttg	21	250	18	2.5 mM	TD=65°C-60°C (30",30", 30") 10+30 cycles	18R
DPYD_x18_R	gggatcataaagggcacaaa	20					
DPYD_x19_F	aacatccattaacaaattaacatgc	25	394	19	2.5 mM	TD=62°C-55°C (30",30", 30") 14+30 cycles	19F
DPYD_x19_R	cattgcatttgtgagatggag	21					
DPYD_x20_F	atcatgcctcaaacagtgc	19	461	20	2.5 mM	TD=62°C-55°C (30",30", 30") 14+30 cycles	20R
DPYD_x20_R	tggctgtaatcaagtctccttc	22					
DPYD_x21_F	catgaaacaatccctagacaca	22	479	21	2.5 mM	TD=62°C-55°C (30",30", 30") 14+30 cycles	21R
DPYD_x21_R	catgcttgccagtgttctaaa	21					
DPYD_x22_F	aaaaacaggaaaatgctgagtg	22	378	22	2.5 mM	TD=62°C-55°C (30",30", 30") 14+30 cycles	22R
DPYD_x22_R	gggtgacaggacagaaagatg	21					
DPYD_x23_F	tcatagtgtggctcctctgc	20	366	23 (coding)	2.5 mM	TD=62°C-55°C (30",30", 30") 14+30 cycles	23F
DPYD_x23_R	TGGAAAGAGCTGAACACAAGG	21					
DPYD_x6_iso2_F	tttcaaaaaccctgaaactaagtaa	25	444	6_isoform2 (coding)	2.5 mM	95°C-55°C-72°C (30",30", 30") for 40 cycles	6_iso2_F
DPYD_x6_iso2_R	GGAAGGGTCCCAAAATGAAA	20					

Chapter 5

Results

Screening for Copy Number Variants in a SLI cohort.

5.1. Detection of CNVs.

In order to investigate the role of CNVs in the genetic risk for SLI, 542 individuals, belonging to 170 families recruited by the SLI Consortium, were genotyped using the Illumina Human OmniExpress (v12.1) beadchip, which includes more than 700,000 SNPs. This cohort was formed by probands, parents, affected and unaffected siblings.

SNP data were analysed using the CNV detection algorithms PennCNV and QuantiSNP, which utilise the Log R ratio and B allele frequency to identify (“call”) CNVs across the genome.

The quality of the CNVs predicted by each algorithm was examined as described in Materials and Methods. Following quality control, the predicted CNVs were taken forward if they were found to have been called by both algorithms and the algorithm calls overlapped by 50% or more each way. This generated a list of “high-confidence” calls, that included 6,229 CNVs for the entire cohort. As a further control, we checked the distribution of CNVs across the cohort and found that the number of CNVs per individual, including chromosome X, ranged from 0 to 324 (**Figure 5.1**), with a mean of 11.41 and a median of 11. The majority of the individuals (97%) carried 5-20 CNVs.

A CNV screen performed on a control population of 1,000 US Caucasian subjects, genotyped with the Affymetrix GeneChip Mapping 500K array (Li et al., 2009), found an average of ~9 autosomic CNVs/individual, ranging from 1 to 32.

Since the majority of the SLIC individuals had an average number of CNVs that matched with these estimates, while two samples had more than 50 predicted CNVs, we decided to exclude these individuals from further analyses, that were performed then on a group of 540 subjects (total number of “high-confidence” CNVs =5835). In this sample set, the number of CNVs per person ranged from zero to 43, with a mean of 10.72 and a median of 10.5.

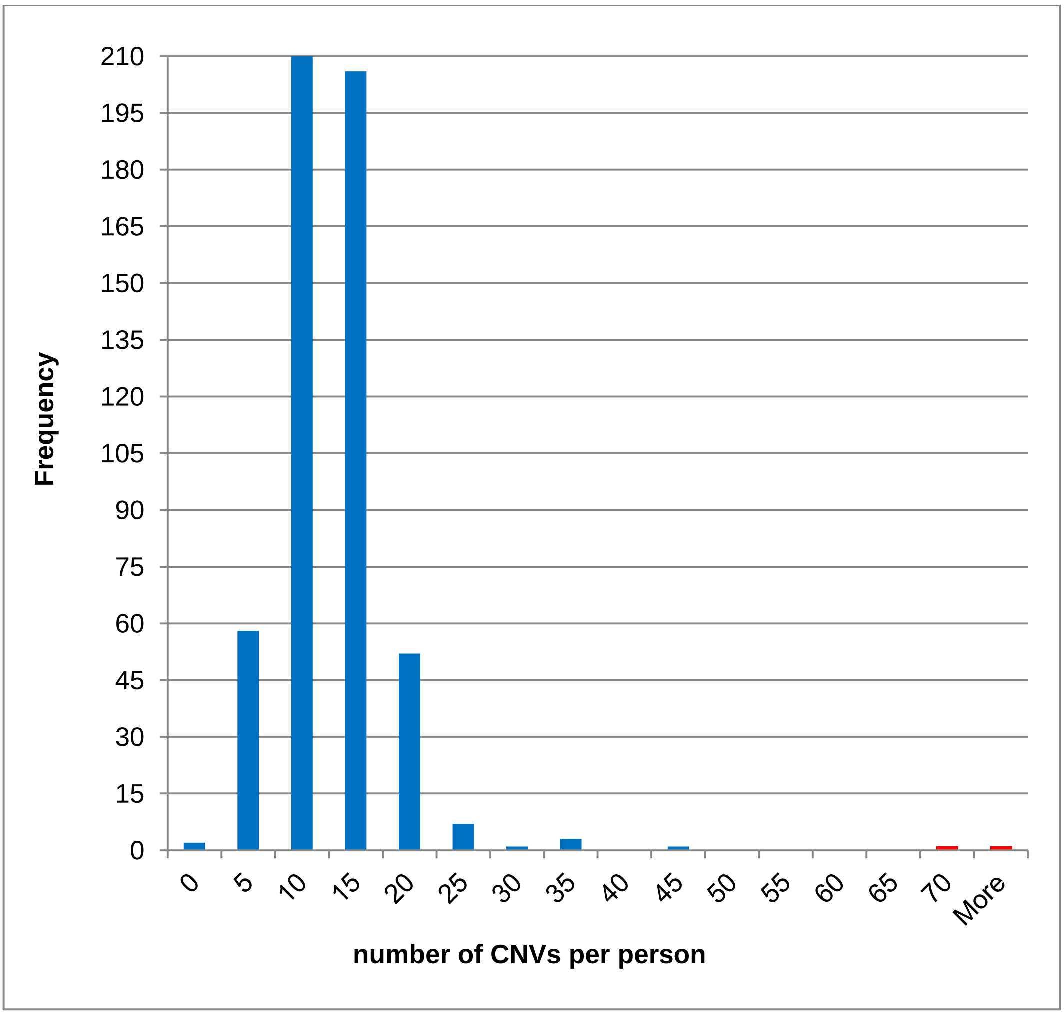


Figure 5.1. Distribution of the average number of CNVs per each person of the SLIC cohort included in the CNV screening (n=542).

Moreover, we looked at the amount of kilobases globally spanned by CNVs in each individual. We found that the mean of kb included by CNVs in each person was 731.6 kb and the median of kb included by CNVs was 567.2, ranging from 0 to 14793.6 kb. For the majority of the subjects (96%), the burden of CNVs involved a portion of the genome that ranges from 150 kb to 2,500 kb (**Figure 5.2**). Nine individuals had a CNV burden of more than 2,500 kb. At the extreme, one sample was predicted to have a CNV burden of ~15,060 kb, given by 30 CNVs: 3 of them, falling in intergenic regions, had a size >3.5 Mb kb each.

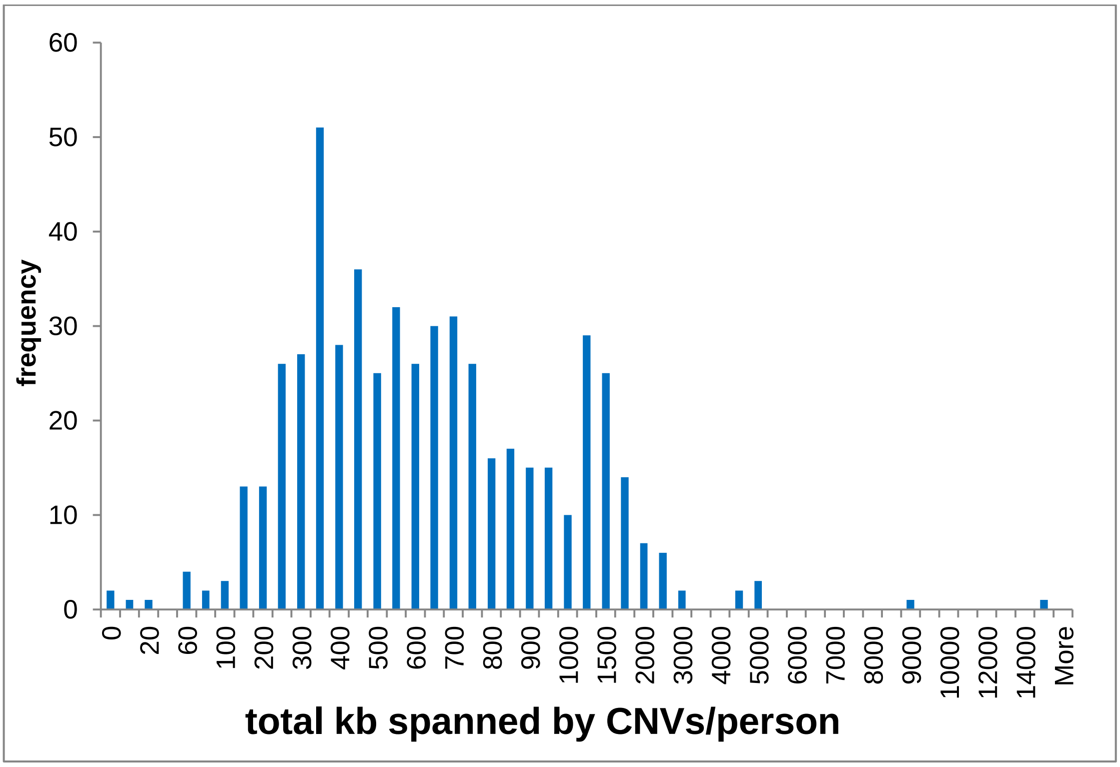


Figure 5.2. Distribution of the total number of kilobases (kb) spanned by the CNVs per each person of the SLIC cohort included in the CNV screening (n=540).

73% (n=4247) of the identified CNVs were deletions (**Figure 5.3**), a number significantly higher compared to the duplications (T-test *p*-value <0.0001). The duplications resulted to be significantly larger (142.1 kb vs 45.9 kb, T-test *p*-value <0.0001). A higher number of deletions and larger duplications reflect the trend which is usually observed in the general population (Shaikh et al., 2009).

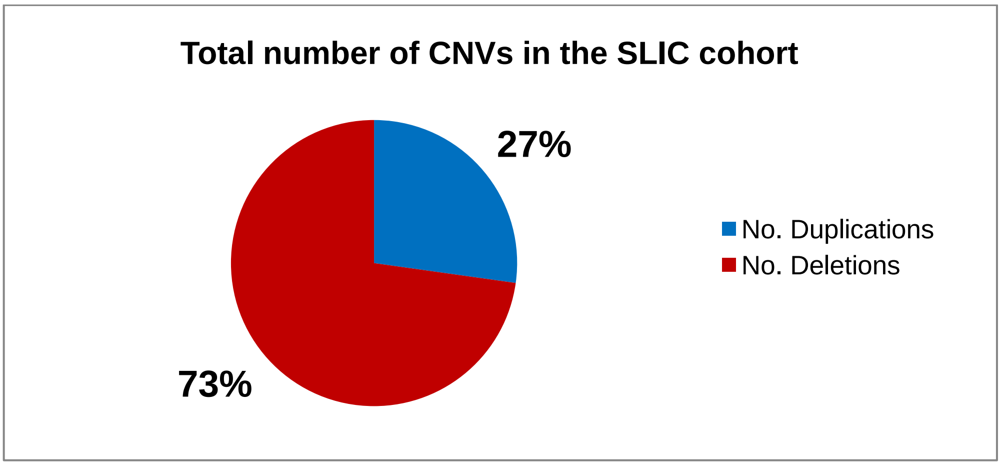


Figure 5.3. The graph shows the enrichment of the deletions (n=4247) compared to the duplications (n=1588) identified in the CNV screening performed in the SLIC cohort.

5.2 CNV burden analyses.

From the entire cohort, 214 individuals were selected to investigate the burden between cases and controls. This analysis included 174 individuals (98 independent individuals and 38 sibling pairs, total number of independent individuals =136) classified as language-impaired cases and 40 siblings (36 independent subjects and two sibling pairs) as unaffected and selected as “super-controls”. 25 super-controls were related to individuals included in the group of cases. Taking into account the overlapping families between the two groups, the total number of cases and super-controls represented 148 independent families. The remaining children had intermediate language ability and were defined to have “unknown” language status. These individuals were removed from the following analyses.

Although the size of this control sample set was very small and included related individuals, we decided to perform an exploratory CNV burden analysis using these internal controls, since a larger group of unaffected individuals genotyped on the same SNP array (Illumina Human OmniExpress v12.1), and therefore with CNVs called with the same genomic coverage and resolution, was not available at the time of the analysis.

When the analysis was restricted to the language-impaired cases and the super-controls, the list of CNVs included 2393 calls (**Table 5.1**). The burden analysis was performed using PLINK to test whether the language-impaired cases presented a higher global burden of CNVs compared to unaffected siblings. This analysis consists of several tests that examine the number of CNVs per person, average size of CNV, total length of CNVs, and the number of genes spanned by CNVs. No significant differences in CNV burden were observed between cases and controls, as shown by **Table 5.1**.

Burden analysis Tests (cases versus controls)	**TEST**	**AFF**	**UNAFF**	***P*-value**
No. of segments (CNVs)	N	1963	430	
Average no. of segments (CNVs)	RATE	11.28	10.49	0.18
Proportion of sample with one or more segment	PROP	1	1	1
Total kb length spanned	TOTKB	798.5	622.5	0.06
Average segment size	AVGKB	69.81	63.06	0.21
No. regions/genes spanned by CNVs	GRATE	14.17	13.46	0.45
No. CNVs with at least one gene	GPROP	0.95	0.98	0.88
No. regions/genes per total CNV kb	GRICH	0.02	0.02	0.86

Table 5.1. Results of the PLINK burden analysis performed on the list of CNVs detected in cases and controls.

5.3 Detection of rare and novel events.

Then we focused on the rare and novel CNVs. Common CNVs were classified as those that overlapped more than 50% or five or more times with the CNVs reported in the Database of Genomic Variants (DGV) database (version of January 2012). After filtering out the common CNVs, the remaining CNVs were classified as:

- Rare CNVs, if they overlapped with a CNV in the DGV by <50% and 5 or less times;
- Novel CNVs, if they did not appear in the DGV.

484 rare and novel CNVs were identified within 294 of the initial individuals (**Figure 5.4**). The number of deletions significantly exceeded the number of duplications (T-test *p*-value<0.0001) in the category of novel CNVs and when considering novel and rare CNVs together. Again, in the total group of rare and novel CNVs, the average size of duplications resulted to be significantly larger than for deletions (74.3 kb vs 38.2 kb, T-test *p*-value<0.0001).

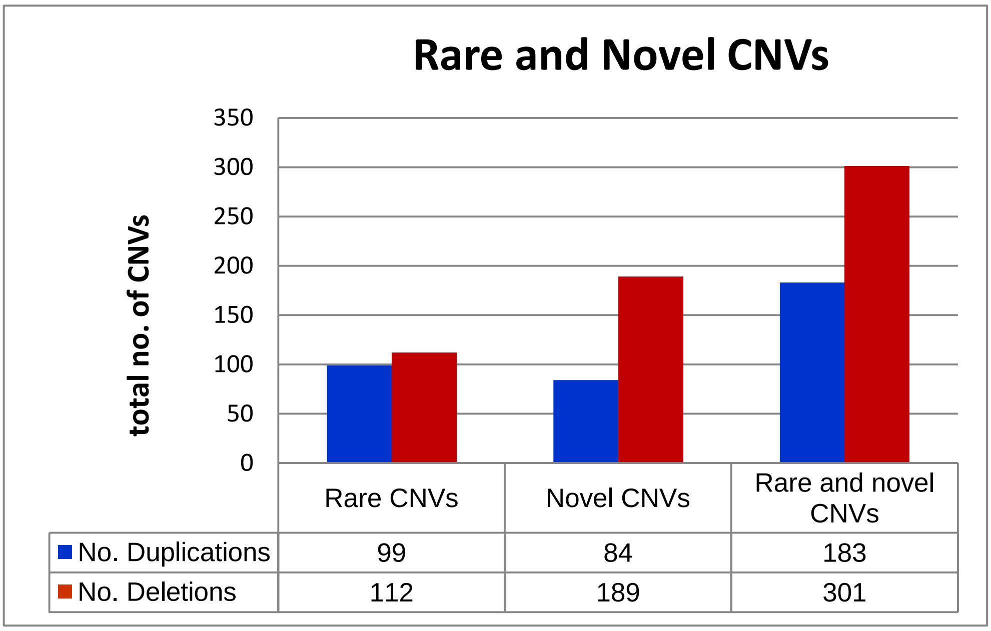


Figure 5.4. 484 rare and novel CNVs: comparison of the total number of deletions and duplications for each category of CNVs (rare, novel or rare and novel CNVs).

Within the class of rare and novel CNVs, we found 284 CNVs overlapping RefSeq genes. All predicted CNVs require a validation (for example by qPCR) before performing in depth analyses for CNV burden or pathway enrichment analyses. To prioritize the validation of CNVs, we can focus on events that are more likely to have a deleterious effect. For this reason, we compiled a list of rare and novel CNVs overlapping exons. The list of rare and novel exonic deletions (**Table 5.2**) and duplications (**Table 5.3**) detected in the individuals classified as affected is reported. Rare and novel exonic events identified in the group of “super-controls” are also reported, in order to evaluate which genes are affected by CNVs that are specific of language impaired cases and are absent in unaffected siblings.

RARE AND NOVEL EXONIC DELETIONS

Indiv. ID	affection	Coordinates (hg19)	cn	score	RefSeq Genes
C94_3	affected	chr1:93639101-93651547	1	13.432	CCDC18 TMED5
E10_4	unaffected	chr1:206317334-206329582	1	148.244	CTSE
E18_3	affected	chr1:206317334-206329582	1	134.387	CTSE
E22_3	affected	chr1:206317334-206329582	1	13.991	CTSE
G69_5	affected	chr1:206317334-206329582	1	126.542	CTSE
C61_4	affected	chr1:226018027-226020988	1	181.415	EPHX1
GCC8_3	unaffected	chr2:37241050-37285840	1	19.951	HEATR5B
G85_3	affected	chr2:183244353-183307061	1	514.243	PDE1A
G25_3	affected	chr3:7348129-7391647	1	79.048	GRM7
E18_5	affected	chr6:151865531-151869219	1	128.395	C6orf97
G65_4	**affected**	**chr7:1022728-1052353**	**1**	**16.345**	**CYP2W1 C7orf50**
C26_4	affected	chr7:16567628-16576251	1	197.496	LRRC72
E38_3	affected	chr7:81788703-81926808	1	182.705	CACNA2D1
GCC8_5	unaffected	chr7:91588036-91610111	1	15.908	AKAP9
C4_4	affected	chr10:74268031-74362355	1	142.423	MICU1 MIR1256
G25_3	affected	chr11:5255221-5261374	1	236.651	HBD
G25_4	affected	chr11:5255221-5261374	1	29.177	HBD
G20_5	affected	chr11:74771478-74806005	1	17.978	OR2AT4
G20_6	affected	chr11:114163716-114173835	1	184.421	NNMT
E13_3	affected	chr11:118055801-118076069	1	170.047	AMICA1
E24_4	affected	chr12:3859449-3872140	1	154.762	EFCAB4B
C42_4	affected	chr12:54568590-54593953	1	25.449	SMUG1
C42_4	affected	chr12:111946837-111979060	1	14.519	ATXN2
C88_4	affected	chr12:123792848-123825559	1	12.845	SBNO1
E39_3	affected	chr13:92408505-92524032	1	699.837	GPC5
E39_4	affected	chr13:92408505-92524032	1	105.795	GPC5
E39_6	affected	chr13:92408505-92524032	1	92.675	GPC5
E6_3	affected	chr16:55618451-55623080	1	103.719	LPCAT2
G77_3	unaffected	chr17:2227855-2247982	1	16.101	TSR1 SRR SNORD91B SNORD91A SGSM2
G20_5	affected	chr17:19998377-20103560	1	141.521	SPECC1
GCC8_5	unaffected	chr18:29193385-29207155	1	102.765	B4GALT6
M19_3	**affected**	**chr18:77914538-77916015**	**1**	**10.565**	**PARD6G-AS1 PARD6G**
G69_3	unaffected	chr22:25315753-25403765	1	34.917	SGSM1 TMEM211

Table 5.2. The table lists rare and novel exonic deletions found in affected individuals and super-controls. CNVs indicated in bold were predicted to be *de novo* (see **Table 5.4**). For each CNV, copy number (cn) and confidence score of the algorithm prediction (score) are reported.

One of the interesting deletions that emerge from this list, detected in one affected individual, E38_3, is an intragenic copy loss of exon 4 in *CACNA2D1*, a gene coding for the α_2-δ1 subunit of Voltage-dependent calcium (Ca_v) channels. This gene is expressed in several tissues, including the CNS (Dolphin, 2013). In particular, in rat brains, its mRNA has been found in areas important for learning and memory, such as neocortex, hippocampus cerebellum and baso-lateral amigdala (Cole et al., 2005). Moreover, *CACNA1C*, the gene coding for the α1-C subunit of L-type voltage-gated calcium channel, is a widely recognized susceptibility gene for schizophrenia and bipolar disorder (Bhat et al., 2012) and it is involved in learning, memory and brain plasticity. Since memory deficits seem to be an important contributing factor for SLI, *CACNA2D1* is an interesting functional candidate gene for language impairment.

RARE AND NOVEL EXONIC DUPLICATIONS

Indiv. ID	affection	Coordinates (hg19)	cn	score	RefSeq Genes
G20_5	affected	chr1:27930898-27973748	3	14.088	FGR
G14_5	unaffected	chr1:111415841-111422817	3	11.084	CD53
G62_4	unaffected	chr1:153023650-153042554	3	13.979	SPRR2A
M68_4	affected	chr1:185110367-185128250	3	141.907	SWT1 TRMT1L
G12_3	affected	chr1:198574902-198609305	3	10.097	PTPRC
G14_3	affected	chr1:198599301-198660724	3	24.11	PTPRC
G20_5	affected	chr1:198601197-198656407	3	17.784	PTPRC
C55_3	affected	chr2:28846035-28920335	3	768.149	PLB1
E5_3	affected	chr2:37247717-37285840	3	115.403	HEATR5B
G42_3	affected	chr2:98620765-98814054	3	95.129	VWA3B
G14_5	unaffected	chr2:197045077-197123318	3	15.24	HECW2
E9_3	affected	chr2:201766236-201943431	3	25.018	FAM126B NDUFB3 NIF3L1 ORC2
G12_3	affected	chr2:201823460-201943431	3	12.955	FAM126B NDUFB3 ORC2
C20_5	affected	chr4:15791464-15840839	3	142.645	CD38
E38_4	affected	chr4:39426809-39572921	3	100.138	C4orf34 KLB LIAS LOC401127 RPL9 UGDH
M68_4	affected	chr4:76712173-76824078	3	568.956	PPEF2 USO1
G14_5	unaffected	chr5:39172131-39222427	3	16.134	FYB
E15_3	affected	chr5:151273630-151357064	3	568.113	GLRA1
E45_4	unaffected	chr6:74315347-74382213	3	291.513	SLC17A5
G20_5	affected	chr6:109758678-109775436	3	11.941	MICAL1 PPIL6 SMPD2
M39_3	affected	chr7:5133936-5240541	3	36.187	WIPI2 ZNF890P
E5_3	**affected**	**chr7:129709081-129738451**	**3**	**107.987**	**KLHDC10**
E30_4	affected	chr9:2718654-2735415	3	269.613	KCNV2
C17_3	affected	chr9:138518114-138555161	3	45.777	GLT6D1
G16_3	affected	chr10:17080633-17211383	3	134.74	CUBN TRDMT1
G16_4	unaffected	chr10:17080633-17211383	3	123.9	CUBN TRDMT1
G20_5	affected	chr10:73500599-73521371	3	151.475	C10orf54 CDH23
G20_5	affected	chr11:63971083-64055905	3	15.331	BAD DNAJC4 FERMT3 FKBP2 GPR137 NUDT22 PLCB3 PPP1R14B STIP1 TRPT1 VEGFB
G60_4	affected	chr11:122430752-122684597	3	33.445	UBASH3B
C95_3	affected	chr11:122455520-122675454	3	26.886	UBASH3B
G14_5	unaffected	chr12:15089416-15119104	3	12.789	ERP27 ARHGDIB
M69_4	affected	chr12:112219696-112298257	3	407.208	ALDH2 MAPKAP5
G45_3	affected	chr13:46168409-46274638	3	431.298	FAM194B
G34_4	affected	chr14:68026119-68036445	3	18.683	PLEKHH1
E39_3	affected	chr15:71432350-71466241	3	215.547	THSD4
E39_4	affected	chr15:71432350-71466241	3	132.718	THSD4
M18_3	affected	chr15:78907656-78974545	3	15.855	CHRNA3 CHRNB4
G50_3	**unaffected**	**chr16:9288453-9356311**	**3**	**177.691**	**MIR548X**
G14_5	unaffected	chr18:2634848-2705700	3	21.106	CBX3P2 SMCHD1
E37_5	unaffected	chr18:9942075-10056733	3	553.227	VAPA
G87_3	affected	chr19:15710360-15732070	3	79.078	CYP4F8
E1_4	unaffected	chr19:58397475-58417426	3	107.728	ZNF814 ZNF417
M19_3	affected	chrX:35827927-36025401	3	52.77	CXorf22 LOC101928564
E45_5	affected	chrX:64346959-64764336	3	21.928	ZC3H12B LAS1L
GCC13_0	affected	chrX:73422412-73564051	3	160.484	FTX MIR421 MIR374B MIR374C ZCCHC13 MIR545 MIR374A

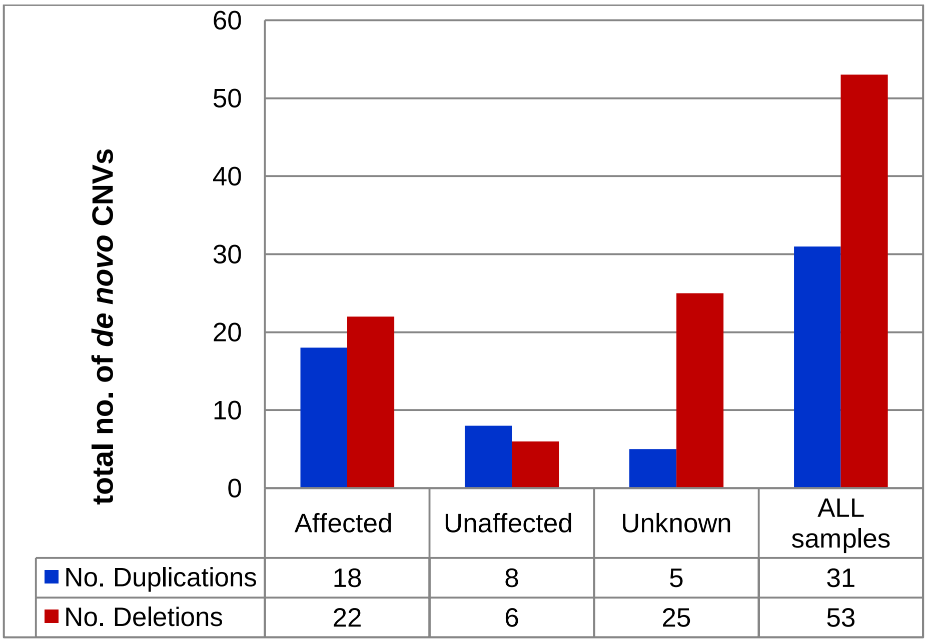
Table 5.3. The table lists rare and novel exonic duplications found in affected individuals and super-controls. CNVs indicated in bold were predicted to be *de novo* (see **Table 5.4**). For each CNV, copy number (cn) and confidence score of the algorithm prediction (score) are reported.

A duplication predicted to involve the first two exons of the gene *KLHDC10* (kelch domain containing 10, OMIM 615152) might be interesting. This gene has been found mainly to be involved in the activation of apoptosis signal-regulating kinase-1 via its kelch repeat domain in response to oxidative stress (Sekine et al., 2012). However, a disrupting deletion in a gene of the same family, *KLHL23*, has been found in a ASD family (Holt et al., 2012). Moreover, 3 kelch proteins (*KLHL17, KLHL22* and *LZTR-1*) emerged from a functional category enrichment analysis of genes hit by *de novo* CNVs in schizophrenia (Malhotra et al., 2011). *De novo* mutations with large effect are rare, because of purifying selection, therefore the occurrence of *de novo* mutations in the same gene or genes with related functions in multiple unrelated patients might indicate that these genes are implicated in the disorder. Therefore, it is possible that rare variants affecting kelch-proteins might contribute to neuropsychiatric disorders. Further characterization of this CNV will be required.

5.4 Detection of *De Novo* CNVs.

167 individuals from 92 families (87 affected individuals, 28 unaffected individuals and 52 of unknown affection status) had both parents available for analysis, allowing the identification of putative *de novo* CNVs within the PennCNV trio and quartet-based CNV calling algorithms.

A total number of 84 *de novo* CNVs was identified within 60 individuals (28 affected individuals, 11 unaffected individuals and 21 with affection status unknown) (**Figure 5.5**). In this total group of subjects, we identified more deletions (63.1%, n=53), than duplications (36.9%, n=31), in line with that observed for all high-confidence events. However, this difference is not driven by the subset of the individuals with a full diagnosis of SLI (affected), as they carry a similar number of putative *de novo* deletions and duplications.



	Affected	Unaffected	Unknown	ALL samples
🟦 No. Duplications	18	8	5	31
🟥 No. Deletions	22	6	25	53

Figure 5.5. Distribution of putative *de novo* deletions and duplications among affected individuals (n=28), unaffected individuals (n=11) and subjects with unknown affection status (n=21).

Generally, *de novo* CNVs are particularly interesting, as deleterious mutations not inherited from the parents may account for sporadic cases of a disorder (i.e. affected individuals with no family history of the disease) and might be extremely useful for the identification of new candidate genes. As for the class of rare and novel CNVs, also for putative *de novo* CNVs, we filtered the CNVs for gene content to prioritize the validation of this group of variants. We identified 43 CNVs overlapping RefSeq Genes, 25 of which were deletions (**Table 5.4**) and 18 were duplications (**Table 5.5**).

***De novo* genic deletions**						
IID	**affection**	**Coordinates (hg19)**	**cn**	**score**	**Ref Seq Genes**	**exonic/intronic**
G38_3	unaffected	chr3:189738195-189739056	0	231.736	LEPREL1	intronic
G72_4	unknown	chr3:189738195-189739056	0	24.046	LEPREL1	intronic
G6_4	affected	chr8:51031221-51033517	0	23.929	SNTG1	intronic
G33_5	affected	chr10:54016099-54016782	0	16.498	PRKG1	intronic
E45_5	affected	chr1:25598276-25642596	1	160.905	RHD	exonic
G45_5	**unknown**	**chr1:211447625-211466761**	**1**	**137.472**	**RCOR3**	**exonic**
E12_3	affected	chr2:65486928-66364645	1	690.822	ACTR2 SPRED2	exonic
E29_3	affected	chr4:120289042-120381341	1	17.087	LINC01061 LOC645513	exonic
G45_5	**unknown**	**chr5:142770135-142776684**	**1**	**126.311**	**NR3C1**	**intronic**
M28_3	affected	chr5:148883634-148903068	1	125.178	CSNK1A1	exonic
G65_3	**unknown**	**chr5:152968158-153023165**	**1**	**46.367**	**GRIA1**	**intronic**
G65_4	**affected**	**chr7:1022728-1052353**	**1**	**16.345**	**CYP2W1 C7orf50**	**exonic**
G45_5	unknown	chr8:79622363-79643639	1	17.64	ZC2HC1A LOC101241902	exonic
G33_5	affected	chr11:18949220-18956690	1	166.263	MRGPRX1	exonic
E20_3	affected	chr11:55365761-55427700	1	34.539	OR4C11 OR4P4 OR4S2	exonic
G16_4	unaffected	chr11:55365761-55427700	1	16.6	OR4C11 OR4P4 OR4S2	exonic
G45_5	unknown	chr11:72778457-72800970	1	129.526	FCHSD2	exonic
E45_4	unaffected	chr11:84841573-84918098	1	44.775	DLG2	exonic
C47_4	**unknown**	**chr11:114170465-114173835**	**1**	**130.554**	**NNMT**	**intronic**
M41_4	unknown	chr14:50098031-50129548	1	146.715	DNAAF2 POLE2	exonic
E44_4	unknown	chr14:104164522-104169017	1	134.257	KLC1 XRCC3	exonic
G45_5	**unknown**	**chr15:62316035-62340126**	**1**	**134.882**	**VPS13C**	**exonic**
E13_4	unknown	chr16:32137965-32392598	1	11.283	LOC390705 TP53TG3D HERC2P4	exonic
M19_3	**affected**	**chr18:77914538-77916015**	**1**	**10.565**	**PARD6G-AS1 PARD6G**	**exonic**
E6_4	affected	chr19:53932295-54004939	1	43.974	TPM3P9 ZNF761 ZNF813	exonic

Table 5.4. List of predicted *de novo* genic deletions detected in children of the SLIC cohort. CNVs were found in SLI cases, unaffected individuals and individuals whose affection status was classified as "unknown". The overlap with exonic or intronic regions of RefSeq Genes (UCSC Genome Browser, hg19) is also indicated. For each CNV, the copy number (cn) and the minimum confidence score from QuantiSNP or PennCNV (score) are also reported. Rare/novel *de novo* CNVs are highlighted in bold.

Among the rare genic *de novo* deletions, we found an intronic copy loss in the gene *GRIA1* (OMIM 138248), which maps on chromosome 5q31.1 and codes for the subunit GluR1 of the ionotropic glutamate receptor AMPA1. Glutamate receptors are the predominant receptors mediating excitatory neurotransmission in the mammalian brain. In particular, *GRIA1* has been shown to play an important role in learning processes and memory (Lee and Kirkwood, 2011; Mead and Stephens, 2003) and it is one of candidate genes for schizophrenia (Ayalew et al., 2012). Therefore, this gene is an intriguing functional candidate gene for a neurodevelopmental disorder like SLI. Although the deletion is predicted to be intronic, in **Table 5.4** we report here the minimal predicted coordinates by PennCNV and QuantiSNP, therefore it might be possible that the deletion could be larger than estimated and might encompass the adjacent exon. Validation will be required to define the CNV breakpoints more accurately.

***De novo* genic duplications**

IID	affection	Coordinates (hg19)	cn	score	Ref Seq Genes	exonic/intronic
G12_3	affected	chr2:61107207-61171045	3	109.997	LINC01185 REL PUS10	exonic
G57_5	unaffected	chr4:58053743-58098554	3	361.955	IGFBP7-AS1	exonic
E25_4	**unknown**	**chr6:35356143-35356640**	**3**	**15.722**	**PPARD**	**intronic**
E5_3	**affected**	**chr7:129709081-129738451**	**3**	**107.987**	**KLHDC10**	**exonic**
C47_3	**unknown**	**chr8:73602555-73635954**	**3**	**389.609**	**KCNB2**	**intronic**
E49_4	unaffected	chr8:144611729-144697653	3	45.583	ZC3H3 GSDMD MROH6 NAPRT1 EEF1D TIGD5 TSTA3 PYCRL	exonic
E5_3	affected	chr8:144974963-145018354	3	17.911	PLEC	exonic
E49_4	unaffected	chr9:139693596-139947473	3	50.851	CCDC183 CCDC183-AS1 RABL6 MIR4292 C9orf172 PHPT1 MAMDC4 EDF1 TRAF2 MIR4479 BC034456 FBXW5 C8G LCN12 PTGDS LCNL1 C9orf142 CLIC3 ABCA2 C9orf139 FUT7 NPDC1 ENTPD2 CCDC183 CCDC183-AS1 C9orf139 C9orf142 C9orf172	exonic
E33_3	affected	chr11:54794237-55035985	3	20.579	TRIM48	exonic
E49_4	unaffected	chr11:55367889-55427700	3	30.34	OR4C11 OR4P4 OR4S2	exonic
E12_4	unaffected	chr11:64570925-64606177	3	14.197	MEN1 CDC42BPG	exonic
G68_4	unknown	chr15:24409977-24714849	3	53.014	PWRN2	exonic
G35_3	unaffected	chr16:1807723-1842209	3	17.301	MAPK8IP3 NME3 MRPS34 EME2 SPSB3 NUBP2 IGFALS	exonic
G50_3	**unaffected**	**chr16:9288453-9356311**	**3**	**177.691**	**MIR548X**	**exonic**
G16_3	affected	chr16:34546530-34687052	3	145.404	RP11-488I20.3	exonic
G33_5	affected	chr17:44165803-44350090	3	34.099	KANSL1 KANSL1-AS1	exonic
G16_3	affected	chr17:44238126-44350090	3	15.542	KANSL1 KANSL1-AS1	exonic
M28_3	affected	chr22:21105255-21463730	3	185.271	PI4KA SERPIND1 SNAP29 CRKL AIFM3 LZTR1 THAP7 THAP7-AS1 TUBA3FP P2RX6 SLC7A4 Mir_649 P2RX6P LOC400891 BCRP2 P2RX6P	exonic

Table 5.5. List of predicted *de novo* genic deletions detected in children of the SLIC cohort. CNVs were found in SLI cases, unaffected individuals and individuals whose affection status was classified as "unknown". The overlap with exonic or intronic regions of RefSeq Genes (UCSC Genome Browser, hg19) is also indicated. For each CNV, the copy number (cn) and the minimum confidence score from QuantiSNP or PennCNV (score) are also reported. Rare/novel *de novo* CNVs are highlighted in bold.

The 17q21.31 deletion syndrome, a multisystem disorder characterized by ID, hypotonia and distinctive facial features, is caused by haploinsufficiency of *KANSL1*, which encodes an evolutionarily conserved regulator of the chromatin modifier KAT8 (Koolen et al., 2012). Reciprocal microduplications (~ 500-650 kb) in this region have been also described, generally, in association with variable, but milder phenotypes (Grisart et al., 2009; Kitsiou-Tzeli et al., 2012). We identified two smaller duplications on chromosome 17q21.31 (**Table 5.5**), encompassing the 5' end of the gene *KANSL1*, in two unrelated affected individuals. The region presents frequent microduplication in the general population. Looking at the number of gains overlapping the interval chr17:44165803-44350090, we estimated that a duplication in this region (of different length) can be found in ~5.4% of healthy individuals. However, the validation and the characterization of these *de novo* events might be interesting and these duplications might reveal other contributing factors to SLI.

5.5 CNVs in candidate genes.

In addition to genic rare/novel or *de novo* CNVs, rare homozygous deletions predicted to have a deleterious effect on genes can be another interesting source for the identification of new candidate genes. As an example, we identified an exonic homozygous copy loss in the *ZNF277* gene, on chromosome 7q31.1, in the proband G4_4. The minimum predicted size of the deletion included only exon 5. The lack of this exon (92 bp) causes a frameshift in the transcript, introducing a premature stop codon in the mRNA. Therefore, the complete absence of exon 5, predicted to alter the *ZNF2777* transcript and possibly causing nonsense mediated decay (NMD), appeared to be an interesting finding to follow-up. We decided then to validate and investigate its possible role in SLI genetic risk (**Chapter 6**).

5.6 CNVs on chromosome 15q11-q13.

In addition to the previously mentioned classes of rare, novel and *de novo* CNVs, we identified rare CNVs in the region 15q11-q13. CNVs in this locus are rare in the general population, but they are recurrently found in a range of neurological conditions, including ID, ASD, schizophrenia, epilepsy and language delay (Burnside et al., 2011; Cooper et al., 2011; Moreno-De-Luca et al., 2013; Shinawi et al., 2009; Stefansson et al., 2008), as described in chapter 1 (paragraph 1.5), suggesting that they might be variants with variable expressivity implicated in different neuropsychiatric phenotypes. In the SLIC cohort two interesting types of CNVs on chromosome 15q11-q13 were identified and validated: microdeletions between the breakpoints BP1 and BP2 and microduplications between BP4 and BP5.

Microdeletions in the BP1-BP2 region were predicted in two families - family G46, with a size of 475,950 bp (chr15:22750305-23226254), and family E21, with a size of 522,429 bp (chr15:22750305-23272733), as shown by the **Figure 5.6**. These deletions include four non-imprinted genes (*TUBGCP5*, *CYFIP1, NIPA1* and *NIPA2*) and were confirmed by qPCR, using two primer pairs, one designed in exon 20 of *TUBGCP5* and another one in exon 5 of *NIPA1*.

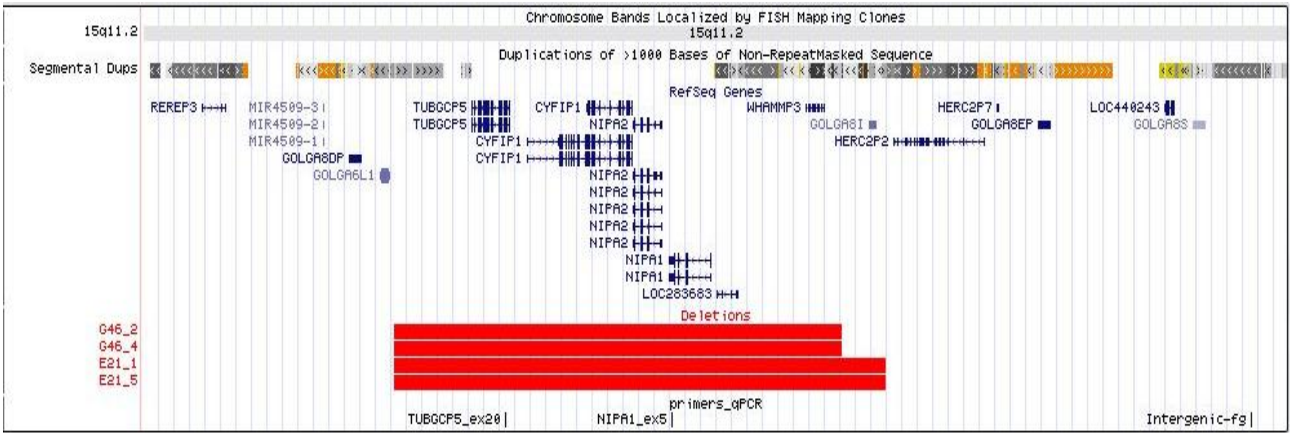


Figure 5.6. BP1-BP2 region on chromosome 15q and the Refseq genes present in this region (UCSC Genome Browser, hg19). The deletions identified in this region are indicated as red bars. The three PCR fragments used for the Realtime validation are indicated.

qPCR confirmed that the deletion in family G46 was present in the proband G46_4, inherited from the mother G46_2 (**Figure 5.7 a**). The event was absent in the sibling G46_3, who also had SLI, and the sibling G46_5, who was not severely enough affected to be classified as SLI but had expressive language score 1.3 SD below that expected for his age. DNA was not available for the father G46_1.

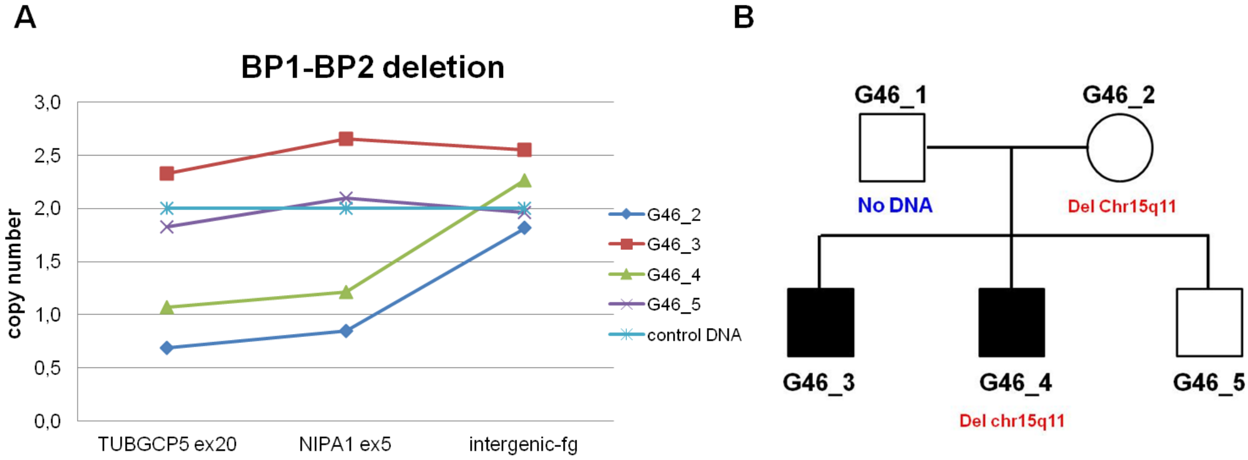


Figure 5.7. Validation of the BP1-BP2 deletion in family G46 by qPCR. **A**. The graph shows a heterozygous deletion in G46_2 and G46_4, that includes the fragments in TUBGCP5_ex20 and in NIPA1_ex5. The third fragment was localized in an intergenic region outside the predicted CNV. **B.** Pedigree of family G46. The children G46_3 and G46_4 were both affected, while the affection status of G46_5 was classified as unknown as his expressive language difficulties were not extreme enough to warrant a diagnosis of SLI.

The two affected children both had low expressive language scores and receptive language scores, but the proband G46_4 had lower performance scores for the NWR test compared to the rest of the family (**Table 5.6**).

individual		ELS	RLS	NWR
G46_1	father			
G46_2	mother			81
G46_3	Affected brother	59	72	76
G46_4	proband	54	76	55
G46_5	Brother (borderline affected)	80	97	96

Table 5.6. Phenotypic test scores for family G46 (CELF-R Expressive language score ELS; CELF-R Receptive language score RLS, Nonword repetition scores NWR) All scores are age-normalised and have a population mean of 100 and a SD of 15. SLI is diagnosed in our data set as expressive or receptive language abilities at least 1.5 SD below that expected for chronological age.

In family E21, a BP1-BP2 deletion was present in the father E21_1 and inherited by E21_5, whose affection status is unknown as CELF data were not available, but who has NWR performance in the normal range for her age. The CNV was absent in the proband E21_3 and the sibling E21_4, who also has missing CELF data but NWR performance in the normal range (**Figure 5.8**).

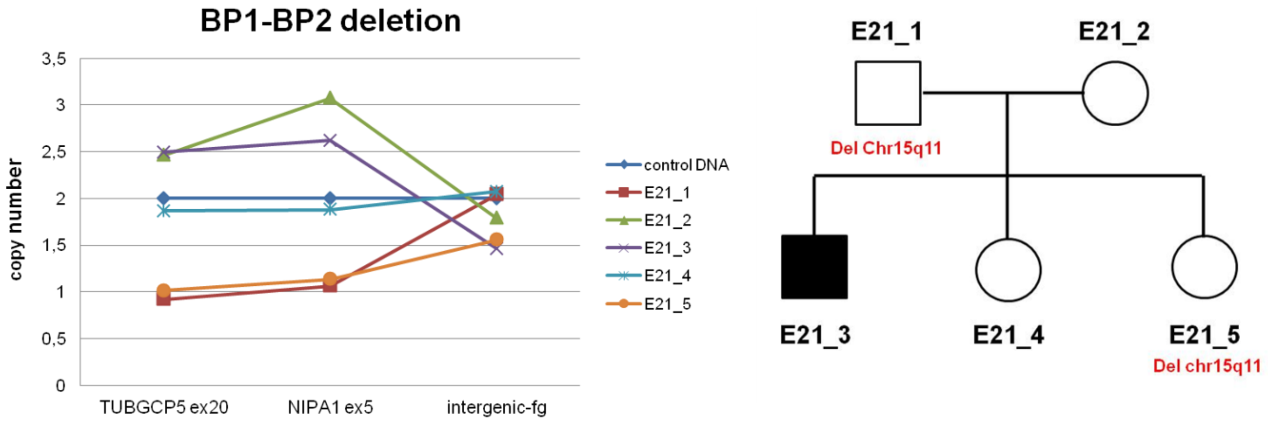


Figure 5.8. A. Validation of the BP1-BP2 deletion in family E21 by qPCR. The graph shows a heterozygous deletion in E21_1 and E21_5 (fragments TUBGCP5_ex20 and NIPA1_ex5). The third fragment was localized in an intergenic region outside the predicted CNV. **B.** Pedigree of family E21. Affection status was unknown for the children E21_4 and E21_5.

Phenotypic data for all siblings of family E21 was available only for the NWR test and the scores were in the normal range for all the children. Expressive and receptive language scores instead, necessary to determine the affection status, were available only for the proband, who did not carry the deletion (**Table 5.7**).

individual		ELS	RLS	NWR
E21_1	father			88
E21_2	mother			94
E21_3	proband	86	107	99
E21_4	sister			93
E21_5	sister			92

Table 5.7. Phenotypic test scores for family E21 (Expressive language score ELS; Receptive language score RLS, Nonword repetition scores NWR).

BP4-BP5 microduplications involving *CHRNA7* were detected in three families: G68, G79 and E38 (**Figure 5.9**). In the three families, this duplication was predicted to occur together with a deletion of the 5’ of *LOC100288637* and the 3’ end of an isoform of *ARHGAP11B* (UCSC isoform uc001zeu.3) (chr15: 30936285-30968006, minimum length= 31,721 bp), overlapping with the segmental duplications of BP4 and reported to accompany the majority of the microduplications involving *CHRNA7* (Szafranski et al., 2010).

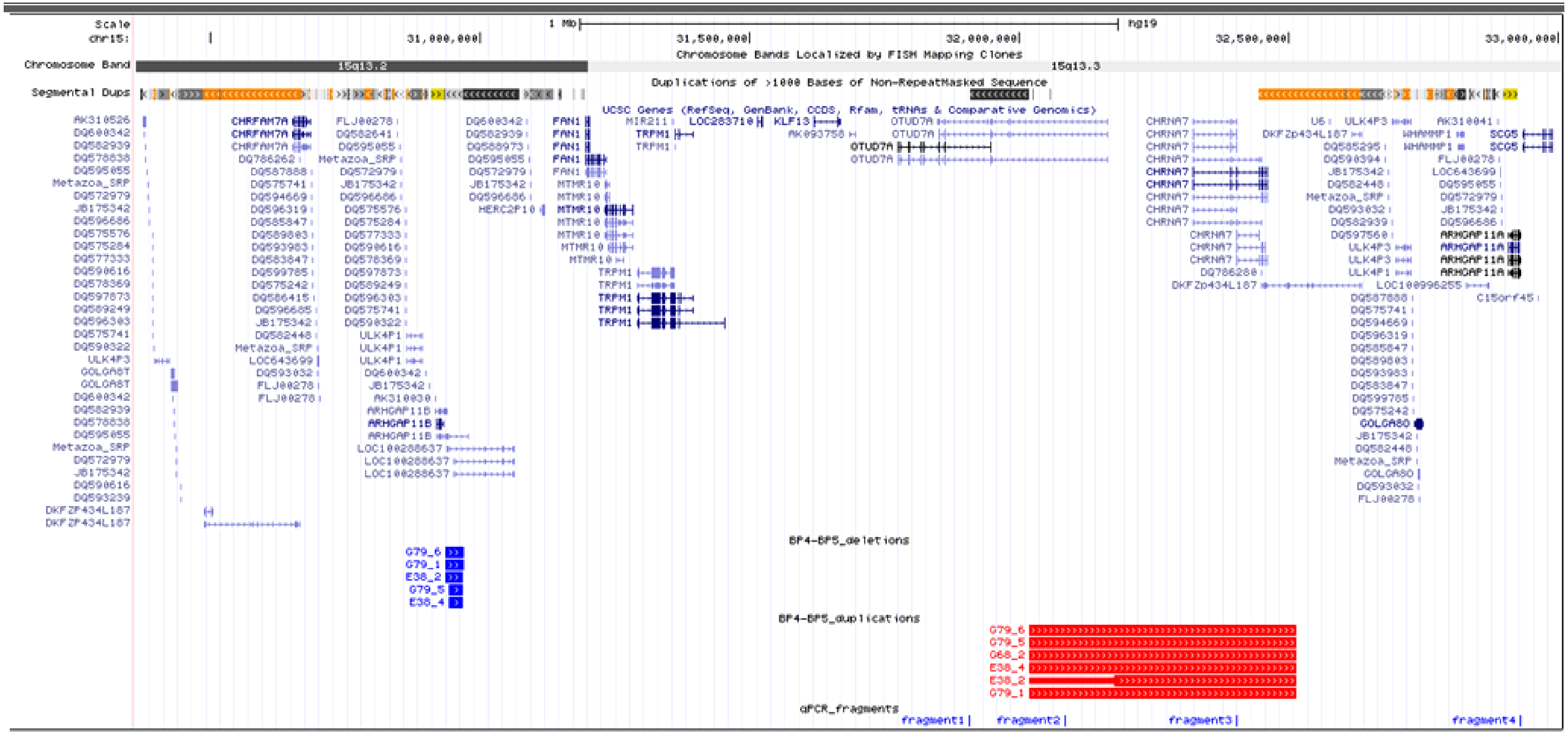


Figure 5.9. BP4-BP5 region on chromosome 15q and the UCSC genes present in this region (UCSC Genome Browser, hg19). The duplications identified in this region are indicated as red bars, the deletions as blue bars. The four PCR fragments used for the qPCR validation are indicated as fragments 1-4.

Again, all predicted BP4-BP5 events were verified by qPCR across each of the family units. In family G68, the CNV (chr15:32018731-32514341, predicted size: 495,611 bp) was predicted to be present in the mother G68_2, but not transmitted to the children, two of them with full diagnosis of SLI (G68_3 and G68_5) and one with affection status unknown (G68_4), but with expressive and receptive language problems (**Table 5.8**). The presence and the segregation of this duplication was checked by qPCR in the entire family, except for the proband G68_5, for whom DNA was not sufficient for the experiment (**Figure 5.10**). The CNV was demonstrated to be absent in the other two children, G68_3 and G68_4. Using the available SNP data for all the member of the family, we performed a haplotype analysis of this genomic region with the program Merlin and we found that all the children share the same maternal haplotype for the BP4-BP5 locus, therefore, we excluded the presence of the duplication also in G68_5.

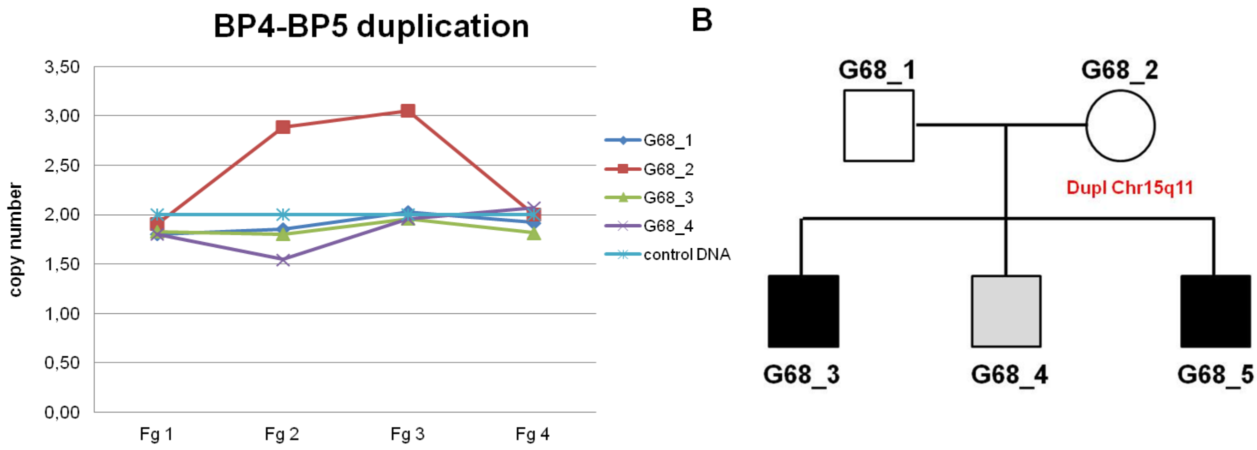


Figure 5.10. A. Validation of the BP4-BP5 deletion in family G68 by qPCR. Fragments 2 and 3 were designed inside the region predicted to be duplicated, the fragments 1 and 4 instead in the region outside, as indicated in the previous **figure 5.9**. The graph shows that the duplication was present only in G68_2. G68_5 was not tested. **B.** Pedigree of family G68. G68_3 and G68_5 are affected (black filling), the affection status of G68_4 was classified as unknown, but he displayed some expressive and receptive language deficits (grey filling).

individual		ELS	RLS	NWR
G68_1	father			104
G68_2	mother			64
G68_3	affected brother	73	105	92
G68_4	brother	78	80	105
G68_5	proband	76	80	73

Table 5.8. Phenotypic test scores for family G68 (Expressive language score ELS; Receptive language score RLS, Nonword repetition scores NWR). G68_3 and G68_5 are classified as affected, G68_4 was classified to have unknown affection status.

In family G79, the BP4-BP5 duplication (chr15:32018731-32514341, predicted size: 495,611 bp) was transmitted from the father to three children G79_4, G79_5, G79_6 and the segregation of the CNV was confirmed by qPCR (**Figure 5.11**).

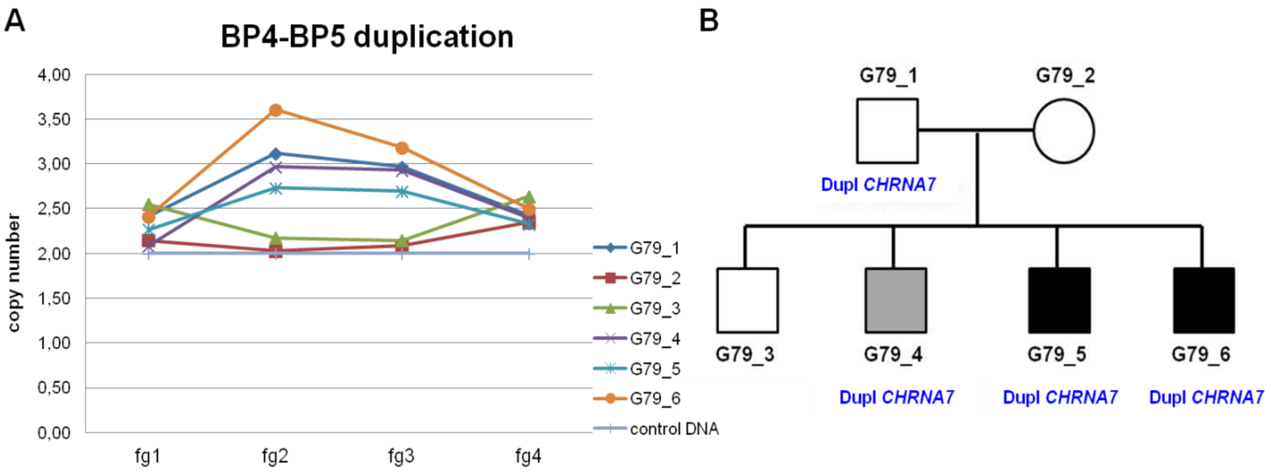


Figure 5.11. A. Validation of the BP4-BP5 deletion in family G79 by qPCR. Fragments 2 and 3 were designed inside the region predicted to be duplicated, the fragments 1 and 4 instead in the region outside, as indicated in the previous **figure 5.9**. The graph shows that the duplication was present in G79_1, G79_4, G79_5 and G79_6. **B.** Pedigree of family G79. G79_5and G79_6 are affected (black filling), the affection status of G79_4 was classified as unknown, but he displayed some language problems (grey filling).

The proband G79_5 obtained low scores in all the phenotypic tests. His younger sibling, who also inherited the microduplication, appeared to have an expressive language impairment but performed in the normal range on the tests of receptive language ability and non-word repetition. The eldest brother (G79_3) did not inherit the duplication and performed in the normal range in all tests. The remaining child (G79_4) inherited the duplication and showed lower scores than expected in the tests of expressive language and nonword repetition but not bad enough to be labelled as "affected" (**Table 5.9**).

individual		ELS	RLS	NWR
G79_1	father			
G79_2	mother			
G79_3	unaffected brother	108	112	98
G79_4	brother	84	105	76
G79_5	proband	64	70	72
G79_6	affected brother	76	101	99

Table 5.9. Phenotypic test scores for family G79 (Expressive language score ELS; Receptive language score RLS, Nonword repetition scores NWR).

In family E38, the BP4-BP5 duplication occurred in the mother E38_2 (chr15:32176304-32514341, hg19, predicted size of 338,038 bp) and was transmitted to the affected sibling, E38_4 (chr15:32018731-32514341, hg19, predicted size of 495,611 bp) (**Figure 5.12**). The proband, E38_3 and his sib, E38_4 are half-siblings, on the maternal side, and both with a diagnosis of SLI, and similar profiles of severe and widespread language impairment (**Table 5.10**). The pedigree structure of this family may lead to the hypothesis that the children are likely to share strong genetic risk factors deriving from the mother. However, the segregation pattern of the chromosome 15q CNV indicates that, if the *CHRNA7* duplication contributes to SLI susceptibility in the sibling, it is separate from other risk factors that may be inherited by the proband.

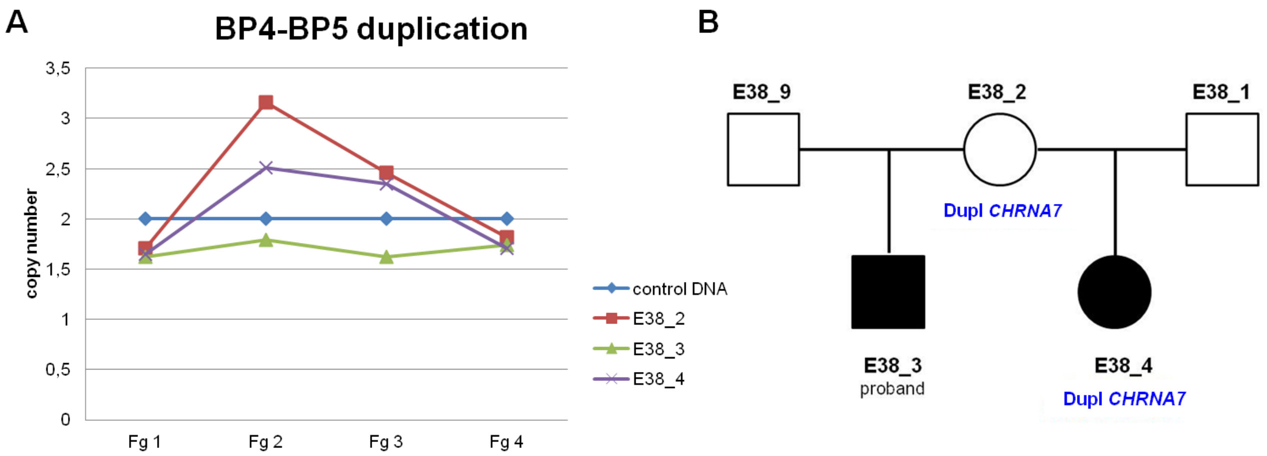


Figure 5.12. A. Validation of the BP4-BP5 duplication in family E38 by qPCR. The fragments 2 and 3 were designed inside the region predicted to be duplicated, the fragments 1 and 4 instead in the region outside. The graph shows that duplication was present in E38_2 and E38_4. **B.** Pedigree of family E38. The two children E38_3 and E38_4 are half-sibling and are both affected.

individual		ELS	RLS	NWR
E38_9	father of E38_3			
E38_1	father of E38_4			
E38_2	mother			88
E38_3	proband	73	74	69
E38_4	Affected half-sib	72	72	79

Table 5.10. Phenotypic test scores for family E38 (Expressive language score ELS; Receptive language score RLS, Nonword repetition scores NWR).

In conclusion, we identified and validated two types of recurrent CNVs on chromosome 15q in the SLI families. We observed an incomplete segregation pattern for both the BP1-BP2 deletion and the BP4-BP5 duplication, as they were identified in affected and unaffected individuals and were not present in all the affected members of these families. The BP1-BP2 microdeletion, was inherited by the proband only in one of the two families in which it was detected. In the other one, it was inherited by a sibling with unknown affection status. The BP4-BP5 microduplication was not transmitted to the children in family G68, it was inherited by one affected child in family E38 and was inherited by all children with language deficits in family G79.

5.7 Ongoing analyses.

Due to the relatedness between the cases and the unaffected siblings used as ‘super-controls’ within the CNV burden analysis reported above, further analysis is currently being conducted using unrelated samples. A list of high confidence CNVs for 127 independent cases affected with SLI and a control population of 269 unrelated individuals, who were unselected in terms of language performance, were compared to assess differences in the burden of CNVs for a number of tests. In summary, the results so far indicate a general trend that the independent cases have on average more CNVs which are of a larger size and span more genes than the control samples. When the analysis was extended to include a further 385 individuals from the SLI cohort, who were a mixture of parents, affected siblings and unaffected siblings, the trends remained when compared to controls. Rare and novel CNVs were interrogated in the same way but there was no overall difference between the independent cases and controls. These data suggest that common CNVs play a role in the pathogenicity of SLI and that compared to other neurodevelopmental disorders, such as autism and ADHD, the increased burden is not driven by rare events. The evidence that the extended SLI family sample set remains significant for the trends observed in the independent cases indicates shared genetic factors that are inherited and could influence the phenotypic outcomes in the wider family, who often present with other language and/or reading difficulties, along with other genetic and environmental factors.

Chapter 6

Results

Validation and follow-up of a homozygous exonic deletion in *ZNF277*.

During the CNV screening, an intragenic homozygous deletion involving the gene *ZNF277* (NM_021994, OMIM 605465) was identified.

ZNF277 maps to chromosome 7q31.1, in a region previously found to be in linkage with ASD (*AUTS1,* OMIM 209850). A fine mapping study of the *AUTS1* locus found association of ASD with SNPs in two genes that are proximal to *ZNF277*, *DOCK4* (dedicator of cytokinesis 4, OMIM 607679) and *IMMP2L* (IMP2 inner mitochondrial membrane protease-like, OMIM 605977) (Maestrini et al., 2010). An additional investigation (Pagnamenta et al., 2010) described a rare microdeletion involving the 3’ end of *DOCK4* (exons 27-52) and the 5’ end of *IMMP2L* genes (exons 1-3), that cosegregated with the presence of dyslexia in an extended family. Therefore, given the interesting genomic location of the gene *ZNF277* and the potential phenotypic and genetic overlap between ASD and SLI, we decided to characterize this homozygous deletion.

6.1. Identification and validation of a homozygous microdeletion of exon 5 in *ZNF277*.

The *ZNF277* homozygous microdeletion was identified during the CNV screening of the SLI cohort in a single child (G4_4) of a Caucasian family (**Figure 6.1 b**).

The child G4_4 met full criteria for a clinical diagnosis of SLI and did not develop language skills until the age of 4-5 years. She was dependent on being shown what to do with toys and her thinking was slightly rigid, but overall she appeared sociable and no other obvious autistic behaviors were reported or observed. Her non-verbal intelligence was below average (Performance IQ=75). She attended a special unit for speech and language impaired children.

The proband had two siblings, an older brother (G4_3) and a younger sister (G4_5). All three children presented with a similar pattern of speech and language impairment, which primarily affected the expressive domain (**Table 6.1**). They all presented with delayed word and phrase speech, unintelligible speech with poor articulation and impaired word retrieval. However, the three children differed in terms of the severity of their impairment and their non-verbal attainment. The younger sister had a higher non-verbal IQ than the proband (PIQ=94) and also had a diagnosis of SLI, although she appeared less severely affected than the proband. The brother had a particularly high non-verbal IQ (PIQ=127) and, although he was reported to have had an early speech and language delay, he did not have a diagnosis of SLI. He did attend a special educational unit and at age 10 years had a significant verbal performance discrepancy and impaired sentence recall. Both parents reported a family history of speech or language problems: the father speech impairment and the mother dyslexia.

individual		ELS	RLS	NWR
G4_1	father			64
G4_2	mother			88
G4_3	brother	99	110	79
G4_4	proband	50	70	72
G4_5	sister	54	83	85

Table 6.1. Phenotypic test scores for family G4 (Expressive language score ELS; Receptive language score RLS, Nonword repetition scores NWR).

The predicted microdeletion included only three SNP probes (rs11769219, rs4727766 and rs7802828), had a minimum predicted size of 4,153 bp and overlapped exon 5. The gene is formed by 12 coding exons (NM_021994.2), but the absence of exon 5 causes a frame-shift mutation and introduces a premature stop codon in exon 7. Given the likelihood of nonsense-mediated mRNA decay, this homozygous microdeletion would thus be predicted to result in a complete lack of functional protein in the affected individual (Khajavi et al., 2006).

A qPCR experiment of available family members in this pedigree demonstrated that one copy of the microdeletion was transmitted to the proband from each parent, who were both heterozygotes (**Figure 6.1 a**). However, the microdeletion was not transmitted to the proband's brother, who presented with an early expressive speech and language impairment but did not have a diagnosis of SLI. Insufficient DNA was available for the qPCR assay in the proband's sister, who was also included in the CNV screening and was predicted to have two normal copies of this genomic region.

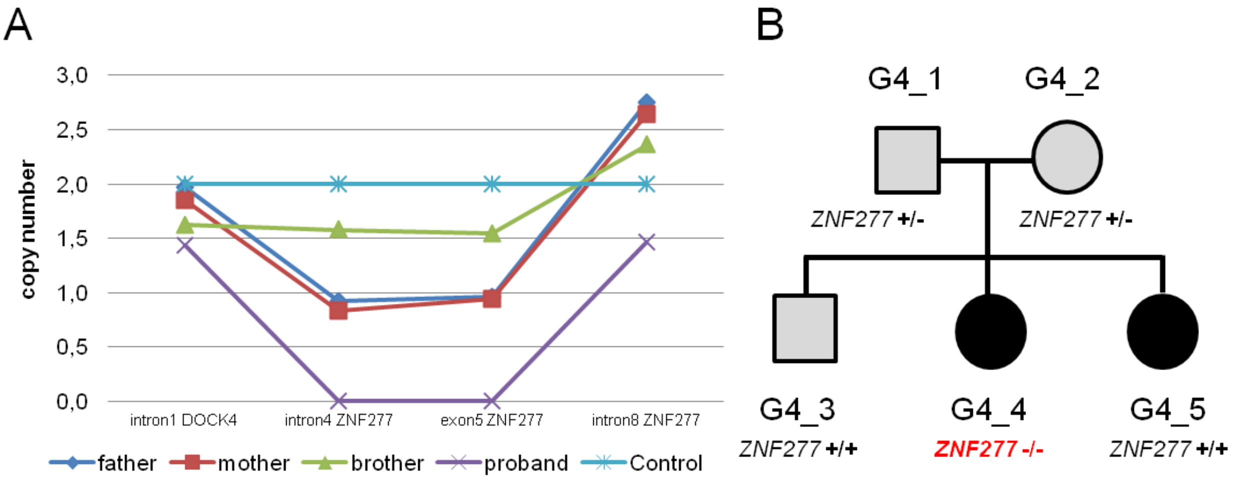


Figure 6.1. A. Results of the qPCR validation of the microdeletion overlapping exon 5 of *ZNF277* in family G4. The qPCR fragments, designed within the predicted deletion region and used in the qPCR experiment, map in intron 4 and exon 5 of *ZNF277*. The graph shows that the proband G4_4 has copy number of zero in the region encompassing exon 5, while the father G4_1 and the mother G4_2 carry the microdeletion in the heterozygous state. The CNV is absent in the brother G4_3. **B**. Pedigree of family G4, where black indicates diagnosis of SLI and grey indicates language problems.

A similar microdeletion was observed in an in-house sequencing database at the Wellcome Trust Centre for Human Genetics in a heterozygous form (1/130 samples of the 500 Whole-Genome Sequences Project [WGS500 Consortium]) (Palles et al., 2013). This deletion, encompassing exon 5, had a size of 21,379 bp, with breakpoints located within intron 4 and intron 5 of *ZNF277*. NGS technologies offer the advantage of identifying CNVs with a bp resolution, allowing a fine mapping of the breakpoints. SNP array instead have a lower resolution and do not detect the precise boundaries of a CNV event. Since the size of this deletion was compatible with the maximum and minimum predicted size of the microdeletion detected in family G4, we tested whether this CNV had arisen from recombination events between the same breakpoints. Primers spanning the microdeletion breakpoints were designed: this PCR fragment allowed the specific amplification of the allele carrying the deletion. Indeed, this PCR product could be detected only in the proband G4_4 (*ZNF277* -/-) and the parents (*ZNF277* +/-), but not in the siblings G4_3 and G4_5 (*ZNF277* +/+) (**Figure 6.2 b**). Conversely, the amplification of a fragment including exon 5 of *ZNF277* gave a PCR product for the parents and the siblings, but not for the proband, who does not have any copy of this genomic region (**Figure 6.2 a**).

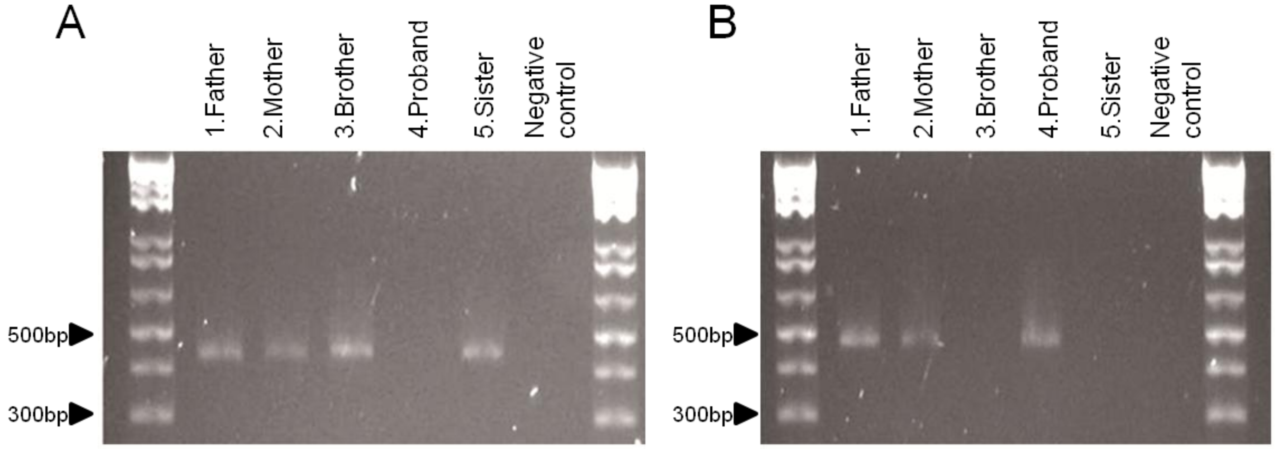


Figure 6.2. A. Amplification of exon 5 of *ZNF277* indicates the presence of at least one allele without the microdeletion in the parents and the siblings G4_3 and G4_5, while the PCR product is absent in G4_4, confirming the homozygous copy loss of this region. **B.** The figure shows the results of a PCR amplification across the microdeletion breakpoints: only the allele with the microdeletion can be amplified and visualized as a band of 466 bp in the parents and in the proband. In both gels, 1kb Plus DNA ladder was loaded at the extremities.

Sanger sequencing validated the boundaries in the discovery individual G4_4 (chr7:111941769-111963147, hg19, 21,379 bp) and further confirmed the presence of the microdeletion also in the parents (**Figure 6.3**). This allowed the accurate detection of the breakpoint boundaries, which lie in two LINE elements, L2c and L1M4. L2c (chr7:111941666-111941883, hg19, strand +) belongs to the L2 LINE family and L1M4 (chr7:111961275-111963848, hg19, strand -) to the L1 LINE family, which promote structural variation through NAHR. BLAST alignment of the entire sequence of these two elements did not reveal extended homology between them. However, sequencing of the breakpoints revealed 2 bp microhomology at the junctions, suggesting that this deletion may be generated through a microhomology-mediated repair mechanism (Vissers et al., 2009).

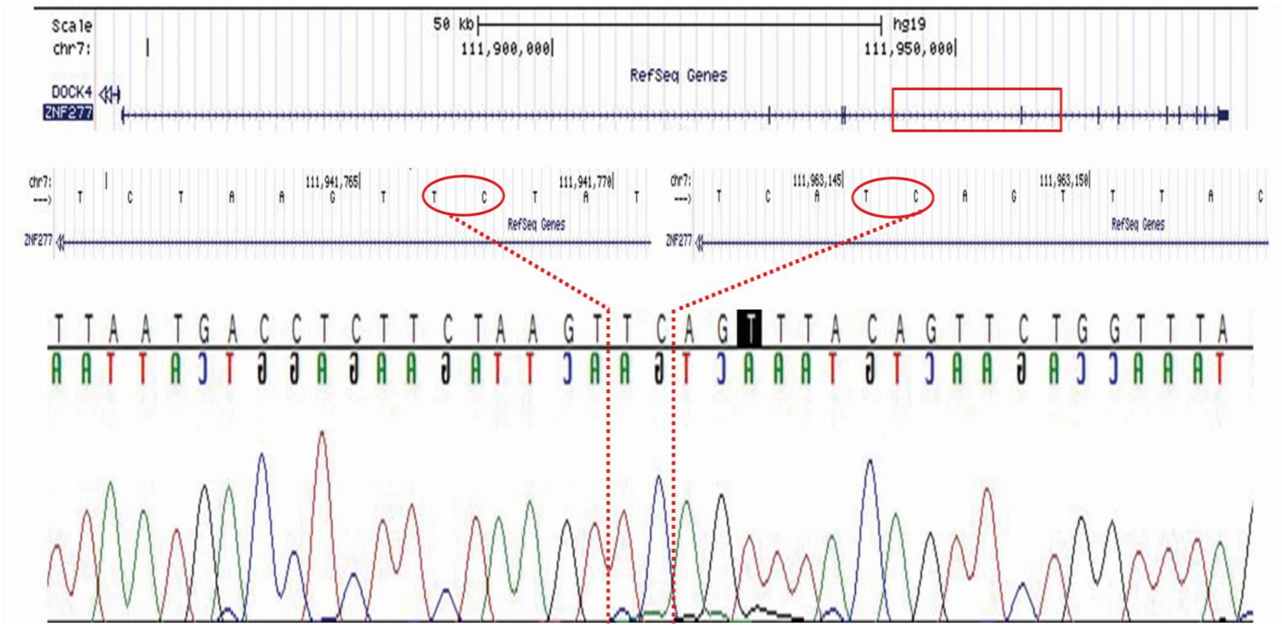


Figure 6.3. Molecular characterization of the *ZNF277* microdeletion in the discovery pedigree. Sequence electropherograms from the PCR products spanning the microdeletion in *ZNF277*. The rectangle indicates the genomic position of the microdeletion in *ZNF277*. The 2 bp (TC), common to both ends, are delimited by dotted lines and circled in the Reference sequence.

At the time of detection (January 2012), there were no overlapping deletions described in the DGV (Iafrate et al., 2004; Macdonald et al., 2013; Zhang et al., 2006a). The latest version of the DGV (January 2014) does report 5 CNVs in *ZNF277* **(Figure 6.4):**

- Two large duplications, a 1,054,909 bp duplication and a 1,152,320 bp duplication, both encompassing the genes *C7orf53, C7orf60, DOCK4, IFRD1, TMEM168* and *ZNF277* (Itsara et al., 2009; Simon-Sanchez et al., 2007).
- A small insertion (267 bp) within intron 1, identified by the 1000 Genomes Consortium Pilot Project;
- Two deletions: a copy number loss of ~45 kb involving the last 8 exons (5-12) of the gene (chr7:111952128-111997265) (Kidd et al., 2008) and a 21 kb copy number loss (chr7:111941766-111963145), involving exon 5 and described by the 1000 Genomes Consortium Pilot Project (Abecasis et al., 2012).

The recently reported microdeletion encompassing exon 5 (indicated in **Figure 6.4** as esv2656841) corresponds to the microdeletion that we detected in family G4 and was identified in 8 samples out of 1151. Assuming that these samples carry the deletion in the heterozygous state, the allelic frequency would be of 0.35%.

In the SLI sample set, no duplications were identified for *ZNF277*.

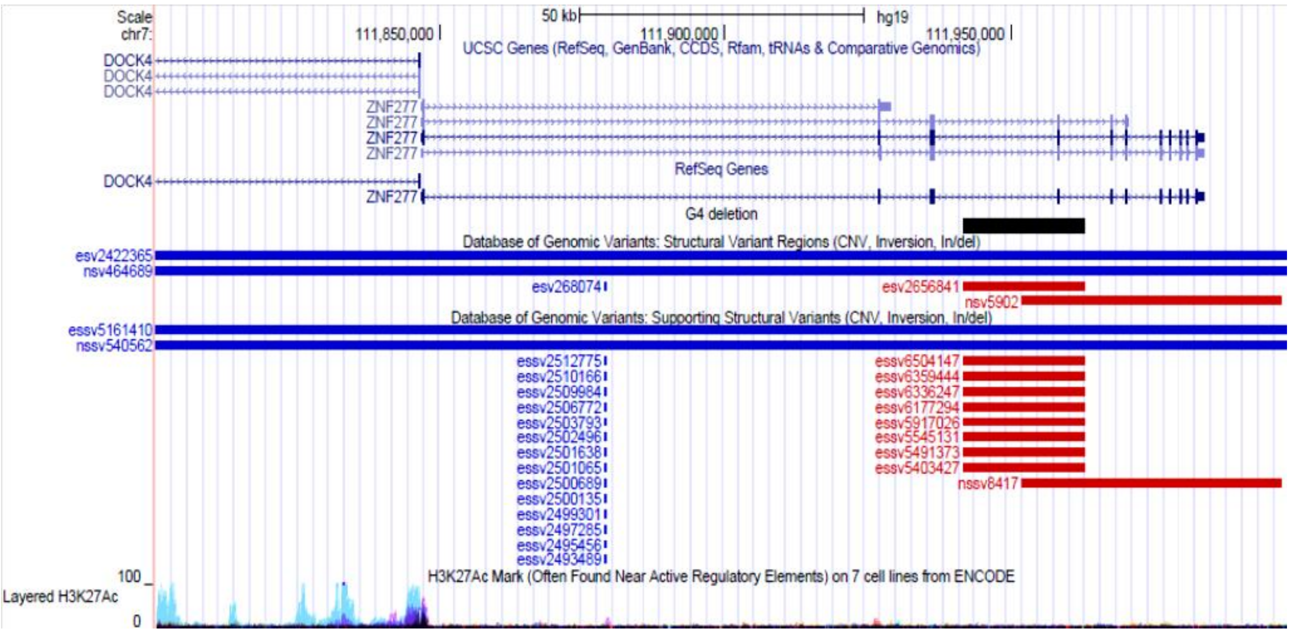


Figure 6.4. Structural variants reported in the last version of DGV (January 2014) within the gene *ZNF277*. The deletion identified in G4_4 is indicated in black. DGV shows also single supporting deletions (red) and duplications (blue) for each CNV reported in this region.

6.2 Homozygosity analysis in family G4.

The presence of a rare homozygous deletion in proband G4_4 led us to hypothesize that the parents might be distantly related. The clinical reports for this family did not contain any information suggesting a possible consanguinity between the parents. Therefore, in order to test this hypothesis, we estimated the proportion of pairwise IBD using genome-wide SNPs in each pair of individuals within the family unit, using PLINK. The proportion of IBD is calculated as PI_HAT =P(IBD=2)+0.5*P(IBD=1). This statistic evaluates the IBD alleles in order to estimate whether two individuals look more similar than expected by chance in a random sample. The expected values for a pair of unrelated individuals is zero, for a parent-child pair or a sibling pair it is 0.5 and for a pair of half-siblings it is 0.25. As shown in **Figure 6.5**, the pairwise IBDs in family G4 were in agreement with the expected values, indicating that there is no evidence of consanguinity in this family. Accordingly, the inbreeding coefficient estimated from genome-wide SNP data for family G4 was ~ 0.

Screening for runs of homozygous genotypes in proband G4_4 did not reveal the presence of homozygous segments longer than 1 Mb throughout the genome, suggesting lack of consanguinity.

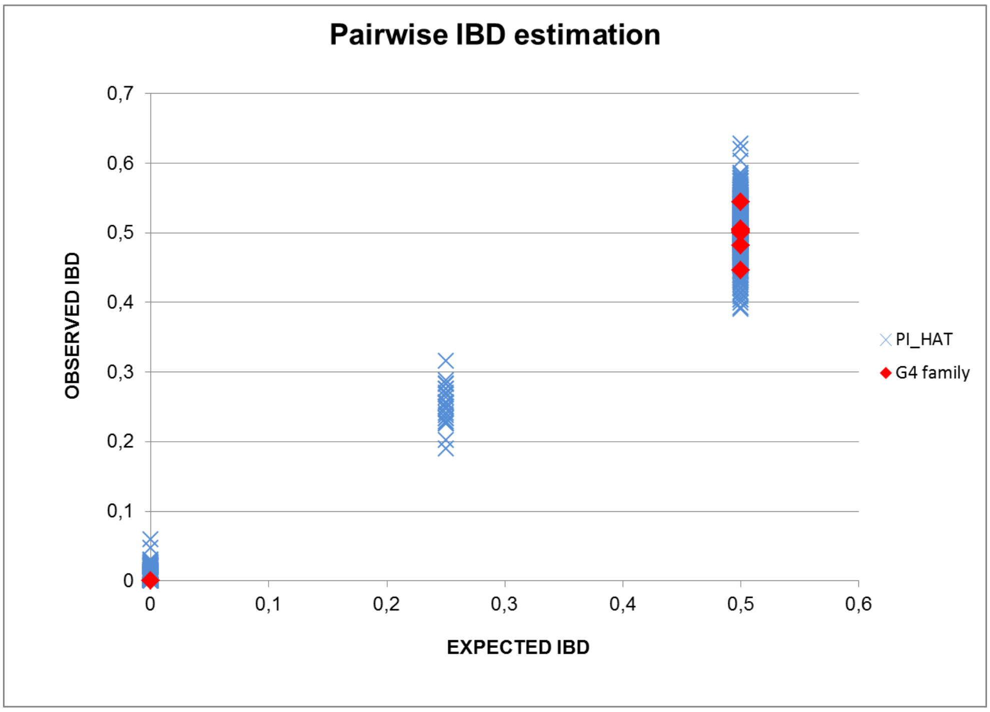


Figure 6.5. Genome-wide pairwise IBD analysis performed on the SLIC cohort, using PLINK. The PI_HAT values obtained for family G4 are indicated in red. The parents pair had an estimated IBD of zero, the other pairs (parent-child or sibling pairs) had an estimated IBD of ~0.5.

Inspection of the haplotypes of the G4 family in a 10 Mb region encompassing the gene *ZNF277*, on chromosome 7q31.1, demonstrated that the proband G4_4 carries a region of homozygosity which spans about 800 kb (chr7:111616692-112460775, hg19) and includes the 5' of *DOCK4*, *ZNF277*, *IFRD1* (OMIM 603502), *LSMEM1*, *TMEM168* and three uncharacterized transcripts (*LOC100996249*, *AC002463.3* and *C7orf60*). For this region, both siblings G4_3 and G4_5 inherited the other parental haplotypes, confirming the experimental results obtained during the validation of the *ZNF277* microdeletion.

6.3. Screening for *ZNF277* microdeletions in the SLIC and IMGSAC cohorts.

In addition to the discovery family G4, another 1229 individuals from the SLI cohort were screened, giving a total cohort size of 322 families (1234 individuals - 545 parents, 318 probands and 371 siblings). The screening led to the identification of 16 additional individuals with the *ZNF277* microdeletion. All individuals carried the deletion in a heterozygous form, five of whom were probands (allelic frequency 0.8%), 6 parents (allelic frequency 0.6%) and 5 siblings (allelic frequency 0.7%) (**Figure 6.6**), giving an allelic frequency of 0.8% in the entire cohort (20/2468 chromosomes). Across all SLI probands (i.e. independent cases including the discovery proband), the allelic frequency of microdeletions was therefore 1.1% (7/636 chromosomes).

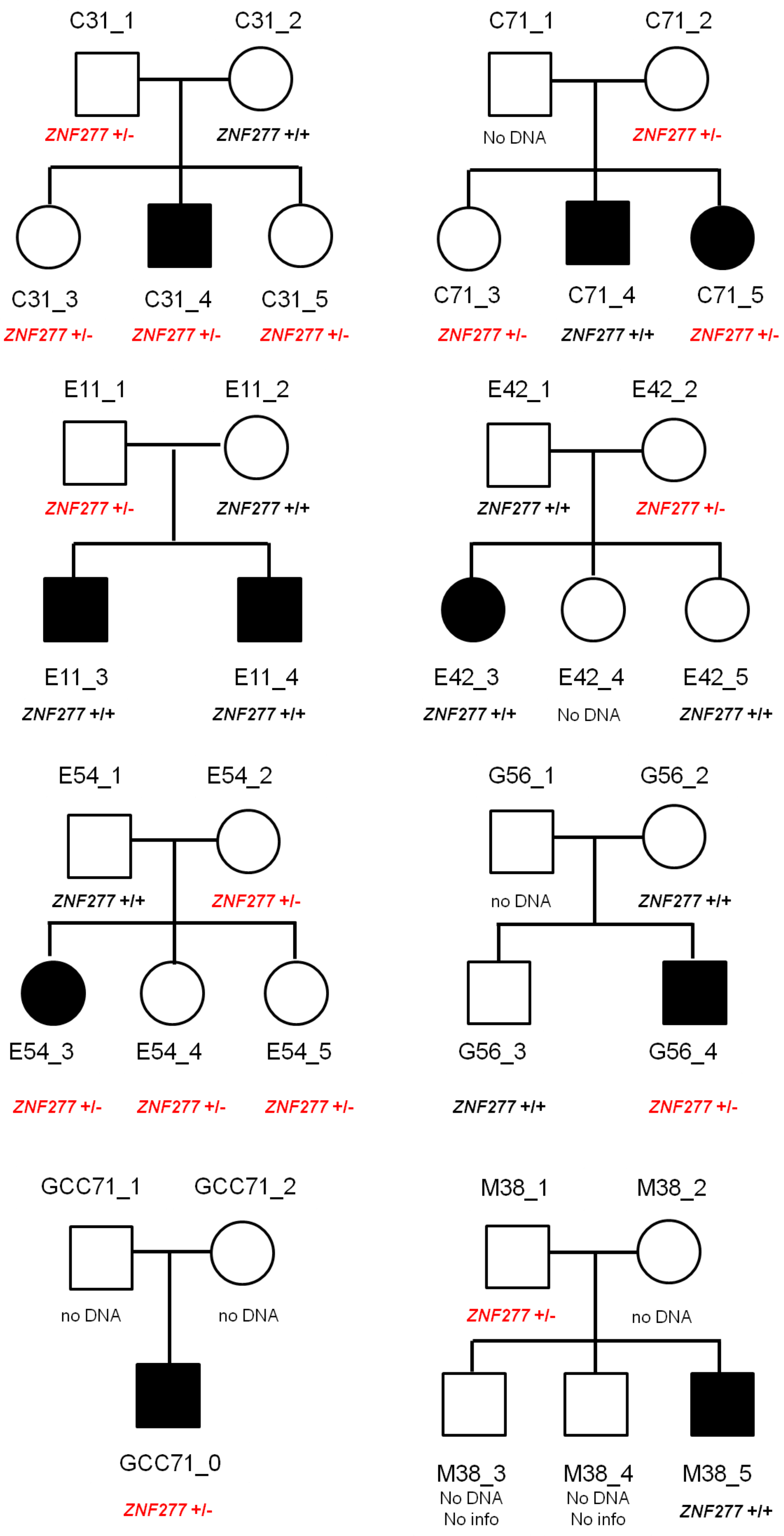


Figure 6.6. Pedigree of the SLIC families carrying the *ZNF277* microdeletion. Black filling means full diagnosis of SLI.

In contrast, the microdeletion was observed in the heterozygous form in 1 of 130 unrelated samples in our in-house sequencing cohort (allelic frequency 0.4%) and 2 of 224 ECACC control individuals (allelic frequency 0.4%) giving a control population allelic frequency of 0.4% (3/708 chromosomes). The frequency in our control group is in agreement with the estimated frequency of the “esv2656841” variant in DGV (0.35% assuming that all the individuals carry the deletion in the heterozygous state), identified by the 1000 Genome Consortium.

Moreover, since the gene *ZNF277* maps to *AUTS1*, a locus previously implicated in autism, we investigated whether the *ZNF277* microdeletion could represent a risk factor also for ASDs. Screening of 252 multiplex ASD families (1021 individuals- 454 parents, 412 affected children, 155 sibs) from the IMGSAC cohort identified heterozygous *ZNF277* microdeletions in 4 ASD families (**Figure 6.7**). Four mothers carried the microdeletion (allelic frequency 0.4%) and it was inherited by 3 affected children (allelic frequency 0.4%) giving a frequency of 0.3% (7/2042 chromosomes) across the entire cohort. All of the ASD families were ascertained as multiplex pedigrees and thus included more than one affected child. Unlike the SLI families, in many cases, there was no single designated proband within the family units. All three ASD cases who inherited the microdeletion had affected siblings who did not inherit the microdeletion rendering the derivation of an objective proband frequency problematic.

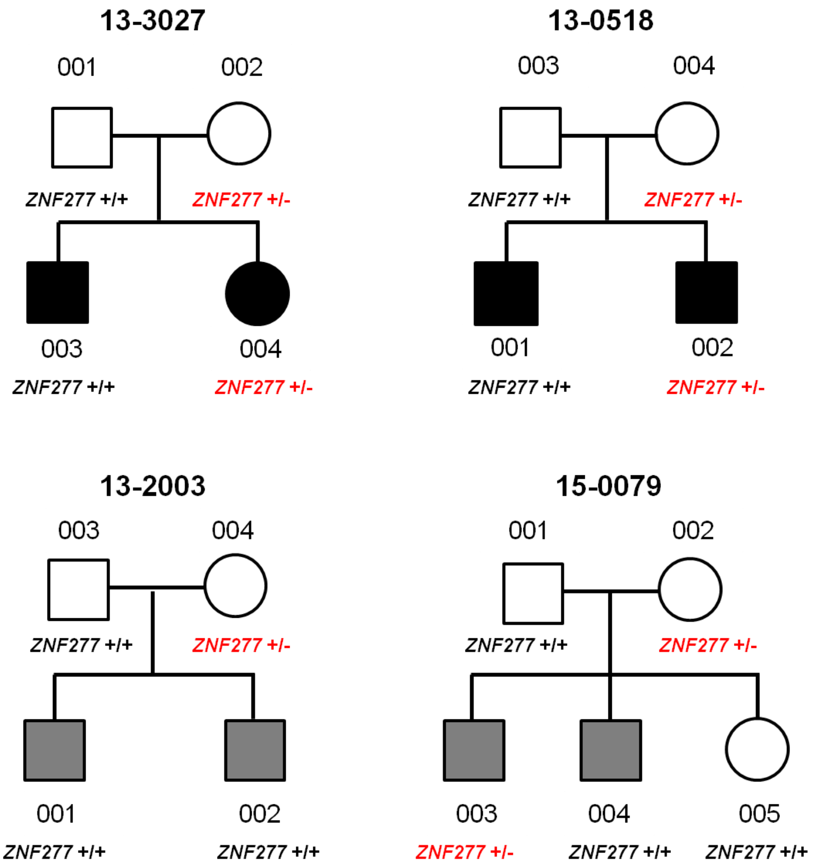


Figure 6.7. Pedigree of the IMGSAC families carrying the *ZNF277* microdeletion. Black filling means full diagnosis of autism, grey filling indicates a diagnosis of PDD.

6.4 Expression analyses of *ZNF277*, *DOCK4* and *IMMP2L*.

In a previous study (Pagnamenta et al., 2010), a microdeletion involving the 5’ end of *IMMP2L* (exons 1-3) and the 3’ end of *DOCK4* (exons 27-52) was identified and characterized in a Dutch family (15-0084) with ASD and dyslexia. The microdeletion (chr7:110876742-111470446, hg19), leading to a fusion transcript between *IMMP2L* and *DOCK4*, was maternally inherited by all three children: two of them presented with ASD and one presented with reading impairment. Interestingly, the simultaneous presence of the *IMMP2L-DOCK4* microdeletion and of a *CNTNAP5* microdeletion of paternal origin, on chromosome 2q14.3, was found to segregate with autism, whereas the *IMMP2L-DOCK4* deletion segregated with dyslexia in the maternal extended family (**Figure 6.8**).

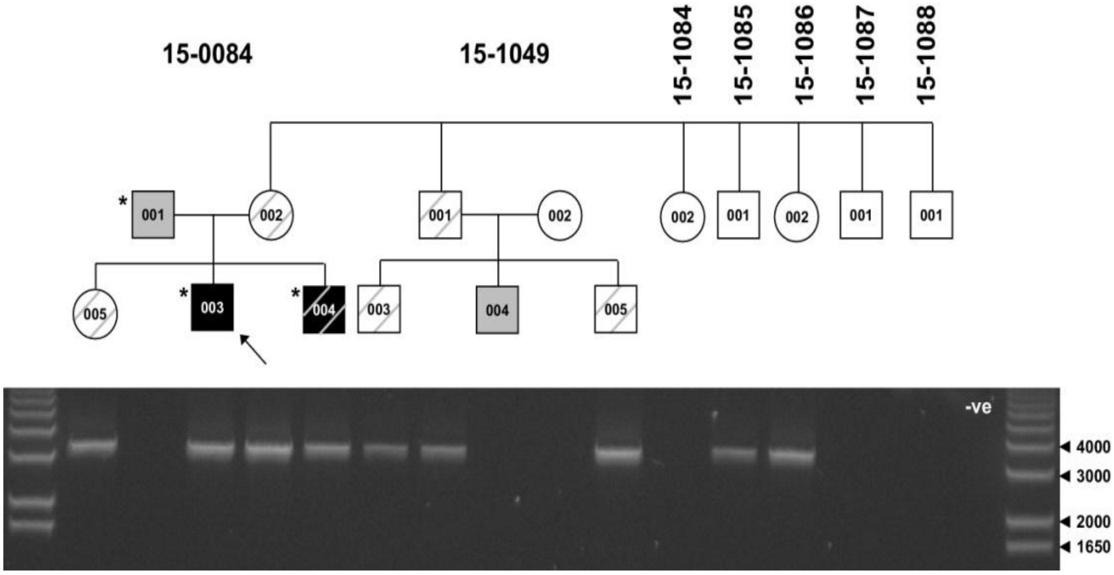


Figure 6.8 (Pagnamenta et al., 2010). Inheritance pattern of the *IMMP2L-DOCK4* deletion within the extended family. Long-range PCR products of 3087 bp are visible only where this deletion is present. Gel lanes are aligned with the pedigree, with the proband indicated by an arrow. Dark shading indicates ADI-defined autism, lighter shading indicates Asperger syndrome or autistic features, and diagonal stripes indicate dyslexic diagnosis or reading impaired. Asterisk indicates presence of *CNTNAP5* deletion.

Since numerous deletions in *IMMP2L* are reported in the DGV, while copy number losses in *DOCK4* are rare and generally small, the haploinsufficiency of *DOCK4* was proposed to be a risk factor for dyslexia susceptibility, and in addition with other variants, for autism susceptibility. Moreover, *DOCK4* encodes for a guanine nucleotide exchange factor (GEF) for Rac1 and Rap1 and positively regulates the dendritic spine formation (Ueda et al., 2013), and thus represents an interesting functional candidate for neurodevelopmental disorders.

Interestingly, *ZNF277* and its adjacent gene *DOCK4* are separated only by 180 bp and are transcribed in opposite directions (**Figure 6.9**). Since the transcription start sites of *DOCK4* and

ZNF277 are so close, it might be possible that regulatory elements for *ZNF277* might lie within *DOCK4* and *vice versa*.

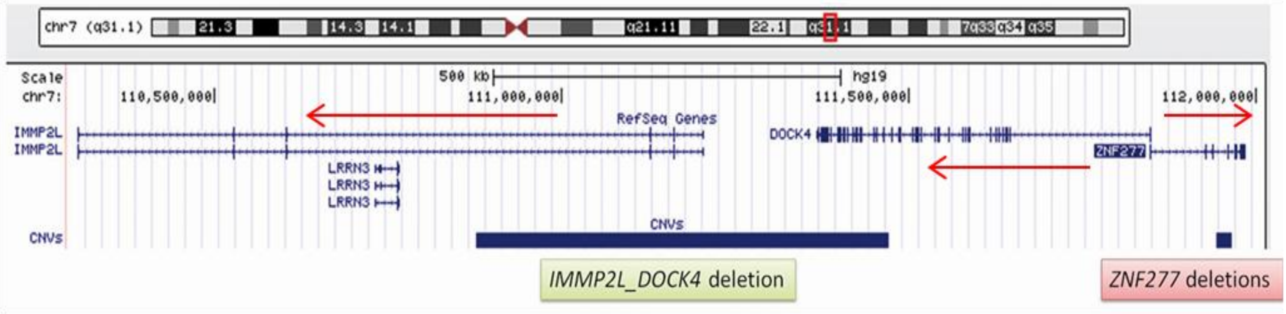


Figure 6.9. Schematic representation showing *ZNF277, DOCK4, IMMP2L* loci with respect to chromosome 7. The blue bars underneath show the chromosomal position of the two types of deletions that were analyzed in this study: the *IMMP2L-DOCK4* deletion includes exons 27-52 of *DOCK4* and exons 1-3 of *IMMP2L*. *ZNF277* deletion includes exon 5. The red arrows indicate the direction of transcription of each gene.

We assessed the effects of the *ZNF277* microdeletion on the expression of *ZNF277*, *DOCK4* and *IMMP2L* by qPCR. Conversely, we also assessed the effect of the *IMMP2L-DOCK4* deletion on the expression of *ZNF277*.

In the SLIC cohort we did not identify any CNV overlapping *DOCK4*, whereas 5 deletions and 1 duplication, all common, were predicted for *IMMP2L* in 6 families. RNA from lymphoblastoid cell lines or blood was not available for these SLI families or for the pedigree G4 carrying the *ZNF277* microdeletion. However, lymphoblastoid cell lines were available for four ASD families in which the *ZNF277* microdeletion was detected in the heterozygous state and RNA from blood was available for the parents of the ASD/dyslexia family 15-0084, in which the mother carried the *IMMP2L-DOCK4* deletion.

Two qPCR fragments were tested to analyse the *ZNF277* transcript levels: one within the region encompassed by exons 1-2 (which lies outside of the microdeletion) and another one within the region encompassed by exons 3-5 (included by the microdeletion). The expression pattern for the fragments was decreased in both cases when compared to individuals without the CNV (**Figure 6.10**) whilst that in exon 5 was significantly lower (p=0.035), indicating that the microdeletion causes a decreased expression of the entire *ZNF277* transcript and supporting the hypothesis of nonsense mediated decay. However, in its heterozygous form, the *ZNF277* microdeletion did not significantly alter the expression of the genes *DOCK4* or *IMMP2L* (**Figure 6.10**).

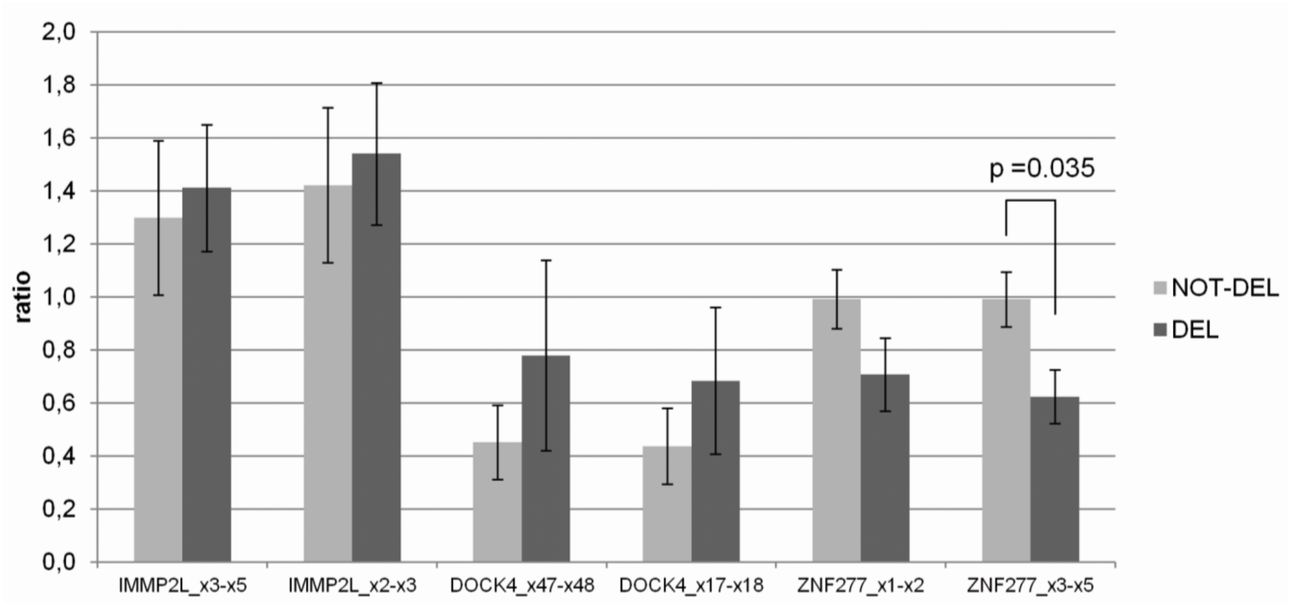


Figure 6.10. The graph shows the ratio of *IMMP2L*, *DOCK4* and *ZNF277* transcript levels, normalized using *GUSB* as a reference. The ratio has been calculated as an average of 5 samples for each group of individuals, belonging to 4 ASD families: “not-del” indicates the group of individuals with two *wild-type* copies of *ZNF277*, “del” the group of individuals with the heterozygous *ZNF277* microdeletion. Bars indicate the standard errors.

Similarly, when we tested the effect of the deletion within *DOCK4* and *IMMP2L*, we did not observe a difference in *ZNF277* levels between the individual carrying the deletion (mother) and the individual carrying normal copies of *IMMP2L* and *DOCK4* (father) (**Figure 6.11**). Note that this deletion has previously been shown to decrease the expression level of *DOCK4* (Pagnamenta et al., 2010).

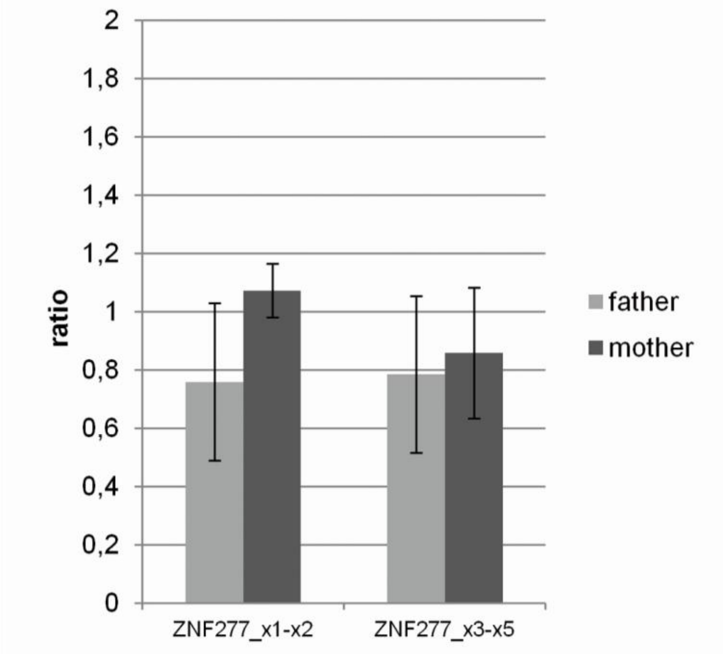


Figure 6.11. Ratio of *ZNF277* transcription levels in an individual with *IMMP2L-DOCK4* microdeletion (the mother 15-0084-002), compared to an individual with two normal copies of *IMMP2L* and *DOCK4* (the father 15-0084-001). The ratio has been normalized on *GUSB* expression levels. Standard error bars are indicated.

In conclusion, we validated an interesting homozygous exonic microdeletion in the gene *ZNF277* in a child with severe language impairment. This microdeletion was found in the heterozygous state in other individuals of the SLIC cohort, in which the probands showed an increased allelic frequency (1.1% in independent SLI probands) compared to both ASD family members (0.3%) and independent controls (0.4%). Moreover, although *ZNF277* falls within the ASD linkage locus *AUTS1* and its neighbouring genes are *IMMP2L* and *DOCK4*, previously implicated in dyslexia and autism, we observed that microdeletions encompassing exon 5 of *ZNF277* reduce the expression of *ZNF277*, but do not alter the levels of *DOCK4* or *IMMP2L* transcripts. Conversely, the 594 kb *IMMP2L-DOCK4* deletion described in the ASD/dyslexia family 15-0084 does not affect the expression levels of *ZNF277*. Taken together, these results suggest that *ZNF277* microdeletions may contribute to the risk of language impairments in a complex manner that is independent of the autism risk loci previously described in this region.

Chapter 7

Results

Analysis of the candidate gene *DPYD*

7.1 Analysis of the candidate gene *DPYD* in SLI cases.

7.1.1 *DPYD* in family G4.

The interesting findings in family G4 led us to re-analyse the clinical reports of these individuals in detail. We found that, after a hospitalization, the younger sister G4_5, who was also diagnosed with SLI, had received a diagnosis of dihydropyrimidine dehydrogenase (DPD) deficiency. This autosomic recessive disorder, caused by homozygous or heterozygous compound mutations in the gene *DPYD*, can be recognized by increased levels of thymine and uracil in blood and urine. As discussed in the introduction, *DPYD* also represents a candidate gene for neurological dysfunctions, since rare CNVs and sequence variants in this gene have been described in individuals with ASD, ID, schizophrenia and other neurological abnormalities. Therefore, we hypothesized that *DPYD* mutations in the sister G4_5 could have contributed to her language deficits.

First, in order to identify the causal mutations of the DPD deficiency in G4_5, we carried out a mutational screening of all the coding parts of the two isoforms of *DPYD*. We detected two potentially causing mutations: a splice site mutation in intron 14 (rs3918290, also known as IVS14+1G>A or *DPYD*2A*) and a missense change in exon 6 of the isoform 1 (rs72549308, **A**GT⇒**C**GT, pS201R).

The splice site mutation rs3918290 is well-known to be implicated in DPD deficiency: the disruption of the donor splice site in intron 14 determines the absence of exon 14 (165 bp) in the mRNA, resulting in a shortened protein which lacks the corresponding 55 amino acids and has no residual activity.

The missense variant rs72549308 is not reported in the database of the NHLBI Exome Sequencing Project (ESP, http://evs.gs.washington.edu/EVS/), that collects the exome sequence data of more than 200,000 individuals from the Unites States, and no allelic frequency is reported in dbSNP138, suggesting that rs72549308 is rare in the population. We used the bioinformatics tools Polyphen2 and SIFT to predict the impact of the substitution of the amino acid serine 201 with arginine and both predicted a deleterious effect for this change (PolyPhen2 score of 1.00, SIFT score 0). This is in agreement with a reported association of this missense change with decreased DPD activity (Ezzeldin and Diasio, 2004).

Analysis of the segregation pattern of the two coding variants detected in G4_5 revealed that she had inherited the missense change from the mother and the splice site mutation from the father (**Figure 7.1**). The proband in this family, G4_4 did not inherit either of the *DPYD* mutations, whereas the older brother G4_3 inherited just the splice site variant. Therefore, the DPD deficiency reported in G4_5 resulted from a compound heterozygous of two damaging mutations, whereas the parents and the brother carried only one deleterious variant in the heterozygous state and were also unlikely to be affected by DPD deficiency.

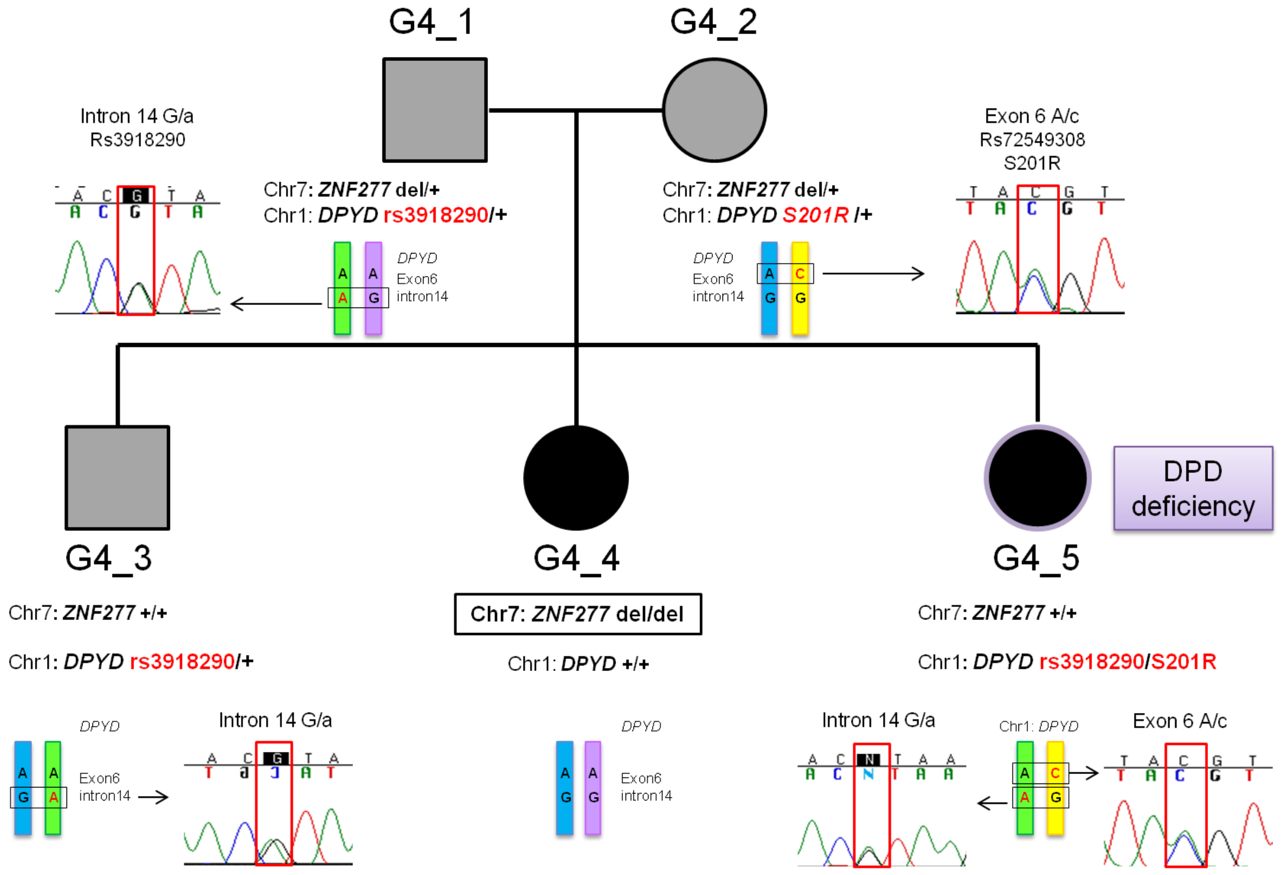


Figure 7.1. Pedigree structure of family G4 with the variants detected in this study: the deletion of exon 5 in *ZNF277* (on chromosome 7) and the sequence variants identified during the mutational screening of the gene *DPYD* (on chromosome 1). The two copies of chromosome 1 are indicated with different colours for each parental allele: in the children, the maternal and paternal copies of chromosome 1 are deduced from the segregation pattern of the two mutations detected in exon 6 and intron 14.

In this family, then, the complete loss of DPD activity was found only in one out of two affected children. However, since SLI is a complex disorder in which multiple risk factors are likely to be implicated, it remains possible that this heterogeneity could be observed also within families and that damaging mutations in *DPYD* might still contribute to SLI in a complex manner.

7.1.2 Screening for the splicing variant rs3918290 in SLI probands.

Considering that *DPYD* represents a good candidate gene for a range of neurological deficits, including language and speech delay, and that we identified two deleterious mutations in this gene in a child with language impairment, we decided to investigate whether *DPYD* could contribute to SLI susceptibility. Over the years, several variants have been reported in this gene, however their functional impact on DPD activity is often uncertain. Although the SNP rs3918290 is rare in the general population, it is the most frequent mutation found in individuals with DPD deficiency and it is widely recognised as a damaging mutation. Therefore, we decided to carry out a preliminary analysis for a possible role of the gene *DPYD* in SLI focusing on this splice site variant (rs3918290).

A group of 166 language-impaired independent cases from the SLIC cohort was screened for the presence of the variant rs3918290. Sequencing of a PCR fragment including rs3918290 was performed and the variant was found in 3 cases, giving an allelic frequency of 0.9% (A=3/G=329). Including also G4_5 in this analysis, the allelic frequency raises to 1.2% (A=4/G=330).

We compared the observed frequency of the rs3918290 minor allele in the SLI samples with that reported by several databases: the 1000 Genome Project (http://www.1000genomes.org/), The HapMap Project (http://hapmap.ncbi.nlm.nih.gov/) and the ESP database (http://evs.gs.washington.edu/EVS/). In **Table 7.1** the allelic frequencies of rs3918290 are reported for different populations.

Controls	**Allele G**	**Allele A**	**Genotype G|G**	**Genotype G|A**	**Allele count**	**Genotype count**
1000Genomes_ALL	0.997	**0.003**	0.995	0.005	2178 (G)/6(A)	1086 (G|G)/6 (G|A)
1000Genomes_EUR	0.993	**0.007**	0.987	0.013	753 (G)/5(A)	374 (G|G)/5 (G|A)
1000Genomes_CEU	0.994	**0.006**	0.988	0.012	169 (G)/1(A)	84 (G|G)/1 (G|A)
1000Genomes_TSI	0.995	**0.005**	0.990	0.010	195 (G)/1(A)	97 (G|G)/1 (G|A)
1000Genomes_GBR	1.000		1.000		178 (G)	89 (G|G)
HAPMAP-CEU	0.996	**0.004**	0.991	0.009	225 (G)/1(A)	112 (G|G)/1(G|A)
HAPMAP-TSI	0.994	**0.006**	0.989	0.011	175 (G)/1(A)	87 (G|G)/1 (G|A)
ESP6500: European-American	0.994	**0.006**	0.988	0.012	8550 (G)/50(A)	4250(G|G)/50(G|A)
SLI cases	0.988	**0.012**	0.976	0.024	330 (G)/4 (A)	163 (G|G)/ 4 (G|A)

Table 7.1. Allelic and genotypic frequencies for rs3918290 reported in three databases for populations of European origin: 1000 Genome Project, HapMap, ESP database, compared with those found in the SLI cases. In the 1000 Genome Project, 26 populations have been analysed (ALL), but then they have been divided in five super-populations, one of those is represented by the Europeans (EUR). EUR includes CEU (Utah Residents (CEPH) with Northern and Western European ancestry), TSI (Tuscans in Italy), and GBR (British in England and Scotland).

For individuals with European ancestry, the frequency of the minor allele is lower (~0.6%) than that observed in our SLI cases (1.2%). However, the difference between these controls and the SLI cases is not statistically significant (one-tailed Fisher's exact test, *p-value*= 0.1423 against the ESP-EA controls, *p-value*= 0.4524 against CEU controls of the 1000 Genome Project, *p-value*= 0.4319 against EUR controls-1000 Genome Project).

7.1.3 Mutation screening of *DPYD* in SLI cases carrying the variant rs3918290.

In order to check whether the three additional SLI cases carrying the splice variant rs3918290, also had a second functional sequence mutation in *DPYD*, we screened the coding regions and the two known regulatory regions in the 5' flanking region of both isoforms in these probands and their family members. We identified several known variants, all common in the population, except for a novel variant in the promoter region found in one family and detailed below (**Figure 7.2 c, Table 7.3**).

In family E6 (**Figure 7.2 a)**, we identified four changes:

1. the splice variant rs3918290, inherited by the proband E6_3 from his mother E6_2;
2. a missense mutation in exon 2 (rs1801265, also referred to as *DPYD*9A*, **T**GT⇒**C**GT, pC29R), inherited from the father by the affected sibling E6_4;
3. a missense change in exon 6 of isoform 1 (rs2297595, **A**TG⇒**G**TG, p.M166V), carried by all family members;
4. a missense mutation in exon 13 (rs1801158, A**G**T⇒A**A**T, p.S534N), inherited from the mother by the affected sibling E6_4.

In family M18 (**Figure 7.2 b)**, we identified two changes:

1. the splice mutation rs3918290, inherited by the proband M18_3 and the unaffected child M18_4, from the mother M18_2.
2. a common missense mutation in exon 2 (rs1801265, **T**GT⇒**C**GT, pC29R), carried by both parents and the unaffected sibling M18_4.

In family E30 (**Figure 7.2 c)**, we identified five changes:

1. the splice variant rs3918290, inherited from the father E30_1 by the proband E30_3 and her affected sibling E30_4;
2. a novel C/T change in the promoter region (chr1:98,386,652, hg19), inherited by the affected sibling E30_4, from the mother E30_2;
3. a common missense mutation in exon 2 (rs1801265, **T**GT⇒**C**GT, pC29R), identified only in the father E30_1;

4. a common missense mutation in exon 13 (rs1801159, **A**TA ⇒ **G**TA, p.I543V), found only in the father E30_1;
5. a common missense change in exon 18 (rs1801160, **G**TT⇒**A**TT, pV732I) inherited by the proband E30_3 from the mother E30_2.

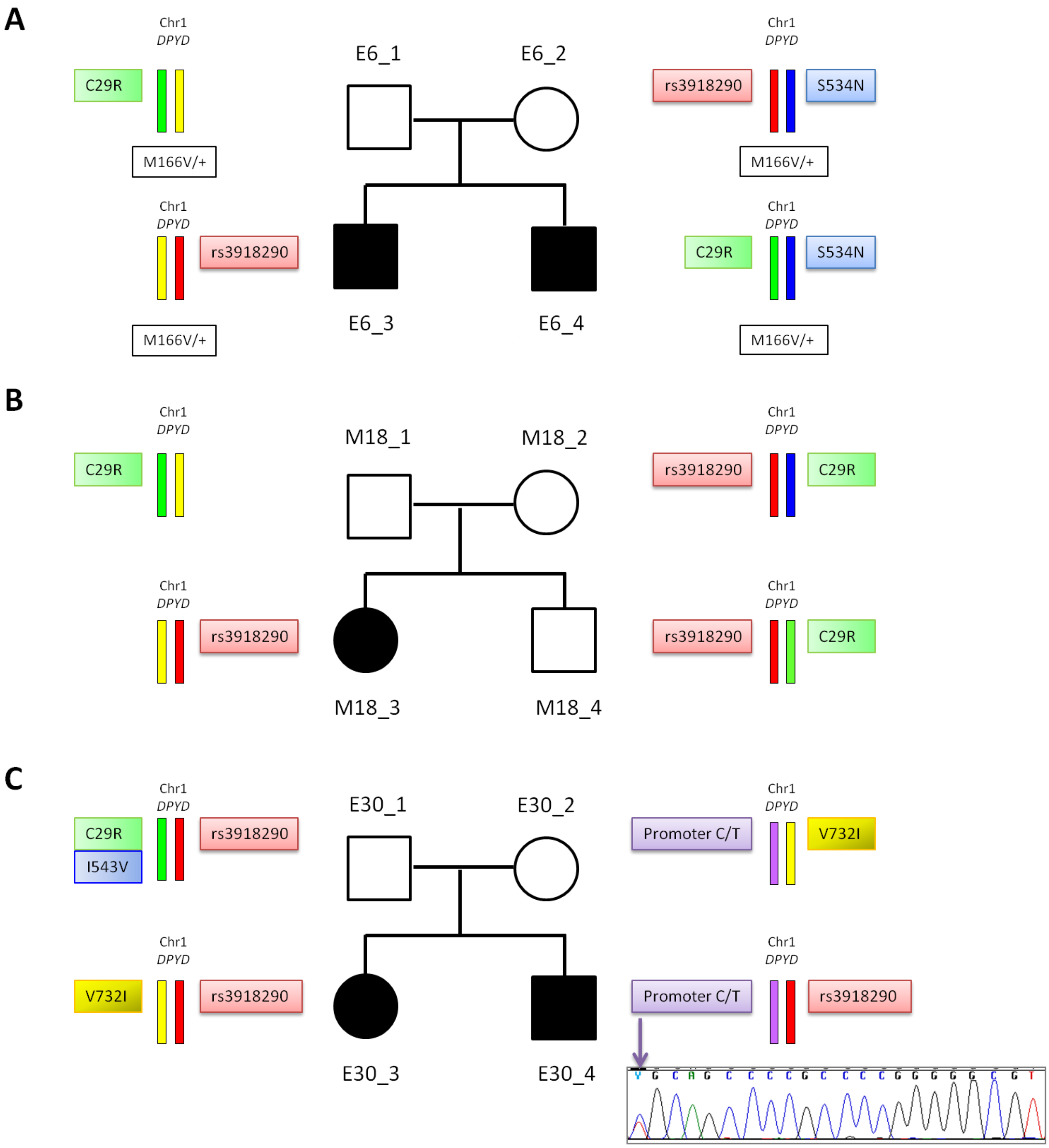


Figure 7.2. Pedigree of the three SLI families in which the mutation screening of *DPYD* has been performed. Paternal and maternal haplotypes for *DPYD* are represented with rectangles of different colours. The haplotypes of the children have been deduced from the inheritance pattern of all the variants identified during the screening. Electropherogram is shown for the variant C/T (strand -) identified in the regulatory region flanking the 5’ of the gene.

All the variants identified are summarized in **Table 7.3**. For each change, the expected frequencies observed in the general population are reported in **Table 7.4,** in which the functional effect for the protein DPD predicted by Polyphen and SIFT is also indicated.

Phenotypic scores of ELS, RLS and NWR language tests are reported in **Table 7.2**.

Indiv.		**ELS**	**RLS**	**NWR**	**AFF**
E6_1	father			94	**0**
E6_2	mother				**0**
E6_3	proband	76	87	92	**2**
E6_4	brother	67	89	67	**2**
M18_1	father			111	**0**
M18_2	mother			104	**0**
M18_3	proband	95	131	115	**2**
M18_4	brother	112	117	113	**1**
E30_1	father				**0**
E30_2	mother			81	**0**
E30_3	proband	70	59	55	**2**
E30_4	brother	72	85	91	**2**

Table 7.2. Phenotypic scores for ELS, RLS and NWR tests available for the three families carrying the splicing variant in intron 14. Individuals were classified as affected (AFF=2) if they were probands or siblings with ELS or RLS scores ≥ 1.5 SD below the normative mean. Individuals with both ELS and RLS scores ≤ 0.5 SD below the mean were classified as unaffected (AFF=1). For the other family members that did not meet these threshold, the affection status was considered unknown (AFF=0).

DPYD	**E30_1**	**E30_2**	**E30_3**	**E30_4**	**E6_1**	**E6_2**	**E6_3**	**E6_4**	**M18_1**	**M18_2**	**M18_3**	**M18_4**
promoter	-	SNP (C/T)	-	SNP (C/T)	-	-	-	-	-	-	-	-
exon 2	rs1801265 (C/T, C29R)	-	-	-	rs1801265 (C/T, C29R)	-	-	rs1801265 (C/T, C29R)	rs1801265 (C/T, C29R)	rs1801265 (C/T, C29R)	-	rs1801265 (C/T, C29R)
exon 6 isoform1	-	-	-	-	rs2297595 (A/G, M166V)	rs2297595 (A/G, M166V)	rs2297595 (A/G, M166V)	rs2297595 (A/G, M166V)	-	-	-	-
exon 13	rs1801159 (A/G, I543V)	-	-	-	-	rs1801158 (G/A, S534N)	-	rs1801158 (G/A, S534N)	-	-	-	-
ex/intr 14	rs3918290 (G/A, splicing)	-	rs3918290 (G/A, splicing)	rs3918290 (G/A, splicing)	-	rs3918290 (G/A, splicing)	rs3918290 (G/A, splicing)	-	-	rs3918290 (G/A, splicing)	rs3918290 (G/A, splicing)	rs3918290 (G/A, splicing)
exon 18	-	rs1801160 (G/A, V732I)	rs1801160 (G/A, V732I)	-	-	-	-	-	-	-	-	-

Table 7.3 . Variants identified in the mutation screening of *DPYD* (Gene accession number: NM_000110.3), carried out in the 3 SLI families with the variant rs3918290.

chr position (hg19)	**rsID**	**Position**	**AA Change**	**cDNA**	**Ref Base** NCBI 37	**dbSNP**	**ESP (EA)**		**ConservationScore**			**Prediction**	
				Pos.	**Strand -**	**MAF (%)**	**Allele Count**	**MAF %**	**Phast Cons**	**GERP**	**Grantham Score**	**Polyphen (score)**	**SIFT (score)**
1:98386652	unknown	promoter	**-**		C	NA	NA	NA				NA	NA
1:98348885	**rs1801265**	Exon 2	**C29R**	85	C	C= 26.30 T= 73.69	T=6666 C=1934	22.488	0.997	5.84	180	benign (0.00)	Tolerared (0.18)
1:98165091	**rs2297595**	Exon 6	**M166V**	496	A	G= 6.562 A= 93.438	G=852 A=7748	9.907	0.999	5.26	21	probably-damaging (1.00)	damaging (0.05)
1:97981421	**rs1801158**	Exon 13	**S534N**	1601	G	A= 1.458 G= 98.542	A=174 G=8422	2.0242	0.95	5.2	46	probably-damaging (0.996)	damaging (0.00)
1:97981395	**rs1801159**	Exon 13	**I543V**	1627	A	A= 80.182 G= 19.818	G=1704 A=6896	19.814	0.026	-3.13	29	benign (0.00)	tolerated (1.00)
1:97915614	**rs3918290**	Splice donor site intron14	**-**		G	A= 0.358 G= 99.642	A=50 G=8550	0.5814	0.997	5.31	NA	NA	NA
1:97770920	**rs1801160**	Exon 18	**V732I**	2194	G	A= 4.543 G= 95.457	A=402 G=8198	4.6744	0.918	5.55	29	probably-damaging (0.998)	damaging (0.00)

Table 7.4. Variants in *DPYD* (Gene accession number: NM_000110.3) identified in the mutational screening. For each variant, position and allelic frequencies reported in dbSNP and ESP (only for European-American individuals) databases are indicated. Functional effects of aminoacid substitutions were predicted using Polyphen2 and SIFT. Residue conservation was estimated by three tools. **PhastCons** describes the degree of sequence conservation among 17 vertebrate species, expressed with values ranging from a minimum of 0 to a maximum of 1 (Siepel et al., 2005). The Genomic Evolutionary Rate Profiling (**GERP**) score ranges from -12.3 to 6.17, with 6.17 being the most conserved (Cooper et al., 2005). **Grantham Scale** ranks amino acid substitutions in classes of increasing chemical dissimilarity (Grantham, 1974), based on chemical properties, including polarity and molecular volume. The changes can be considered conservative (0-50), moderately conservative (51-100), moderately radical (101-150), or radical (≥151).

Interestingly, the variant C/T (chr1:98,386,652, hg19) in the 5’ flanking region of *DPYD* (identified in family E30) is not described in dbSNP 138, in the ESP database or the 1000 Genome Project. This nucleotide falls within a CpG island and represents the first nucleotide of a positive regulatory region (element II, that includes the nucleotides from -72 to -51, considering the transcription start site as the position +1), previously described by Shestopal et al. (2000). A subsequent study (Zhang et al., 2006b) demonstrated that the ubiquitously expressed transcription factors Sp1 and Sp3 bind to the *DPYD* promoter. They found 3 Sp-binding sites, designated as SpA (from -148 to -140), SpB (from -68 to -60) and SpC (from -37 to -19), with the major promoter activity detected for SpB, suggesting that, when bound by Sp1, it may function as an upstream enhancer.

Moreover, we consulted the UCSC Genome Browser (hg19) to find which transcription factor target sites are reported for the region flanking the 5’ of *DPYD* by the ENCODE Project. Using the track that shows the sequences bound by Transcription Factors identified by ChIP-seq, we saw that the region overlapping element II has several predicted target sites (**Figure 7.3**). In particular, the C/T variant falls within 11 putative target sites, with the highest cluster scores (in black) for the components of RNA Polymerase II and Egr-1 (*early growth response,* known also as Zif268). Zif268/Egr-1 belongs to the family of Egr C2H2-type zinc-finger proteins and binds to the DNA sequence 5'-CGCCCCCGC-3' (EGR-site), activating the transcription of target genes required for mitogenesis and differentiation. In brain, Zif268/Egr-1 has a key role in different types of synaptic plasticity, memory consolidation and reconsolidation processes (Veyrac et al., 2014). Other potential binding sites have been predicted for E2F, ZNF263, GATA-1, AP-2gamma.

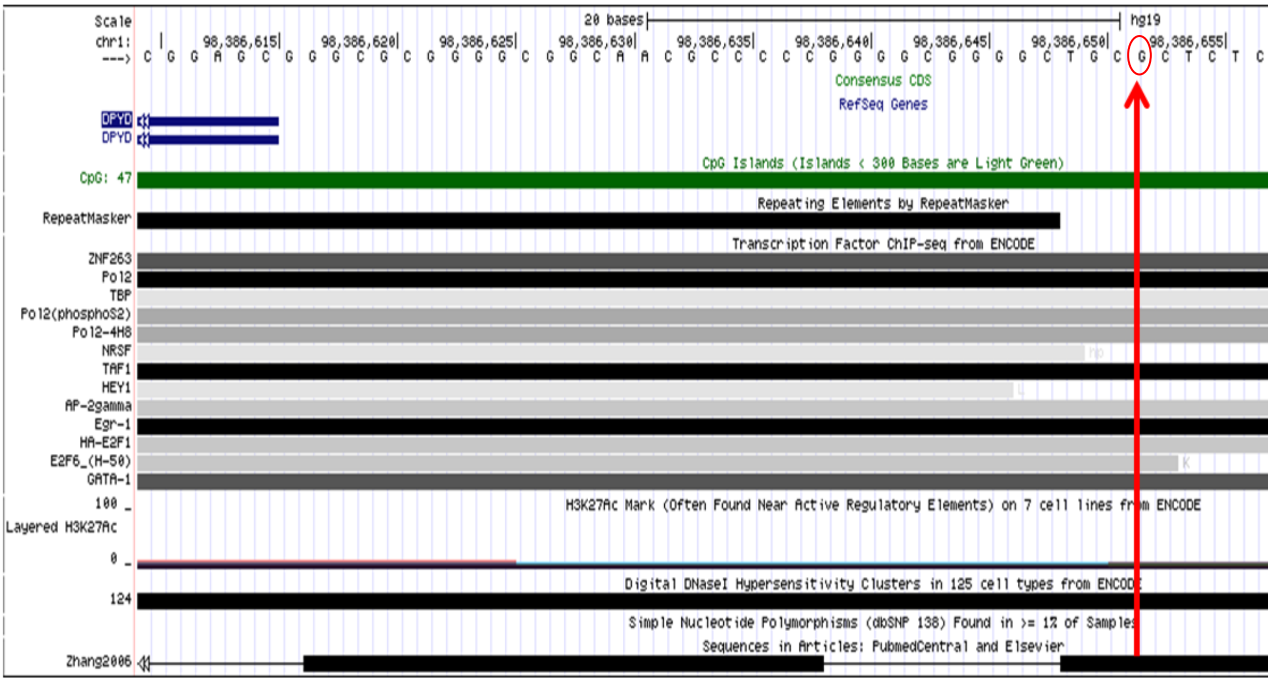

Scale
chr1:
--->
20 bases
hg19
98,386,615
98,386,620
98,386,625
98,386,630
98,386,635
98,386,640
98,386,645
98,386,650
98,386,655
C G G A G C G G G C G C G G G G C G G C A A C G C C C C C G G G G C G G G C T G C G C T C T C
Consensus CDS
RefSeq Genes
DPYD
DPYD
CpG Islands (Islands < 300 Bases are Light Green)
CpG: 47
Repeating Elements by RepeatMasker
RepeatMasker
Transcription Factor ChIP-seq from ENCODE
ZNF263
Pol2
TBP
Pol2(phosphoS2)
Pol2-4H8
NRSF
TAF1
HEY1
AP-2gamma
Egr-1
HA-E2F1
E2F6_(H-50)
GATA-1
100 _
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
0 _
Digital DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE
124
Simple Nucleotide Polymorphisms (dbSNP 138) Found in >= 1% of Samples
Sequences in Articles: PubmedCentral and Elsevier
Zhang2006

Figure 7.3. The figure shows the transcription factors reported in UCSC Genome Browser by the track provided by the ENCODE Project. These transcription factor binding sites have been identified by ChIP-seq (chromatin immunoprecipitation with antibodies specific to the transcription factor followed by DNA sequencing of the precipitated fragments), and the grey scale indicates the cluster scores, with black being the highest scores.

7.1.4 *DPYD* variants from exome sequences.

Exome data for 45 independent probands of the SLIC cohort were also available. In these individuals, the splicing variant in intron 14 of *DPYD* was not identified, but 4 missense coding variants were detected in 24 individuals and confirmed by Sanger sequencing: p.M166V (11/45 individuals, allelic frequency 12.2%), p.S534N (3/45 individuals, allelic frequency 3.3%), p.I543V (13/45 individuals, one of whom was homozygous for the variant, allelic frequency 14.4%), p.V732I (6/45 individuals, allelic frequency 6.7%). Some of the probands (7 individuals) carried more than one of these changes.

Since p.M166V (rs2297595), p.I543V (rs1801159) and p.V732I (rs1801160) do not affect the enzymatic activity of DPD (Ezzeldin and Diasio, 2004; Offer et al., 2013), whereas p.S534N (rs1801158, exon 13) seems to have an effect on DPD activity, we decided to check the segregation of this variant in the three families in which it was identified: E35, E53, G67 (**Figure 7.4**).

In family E53 (**Figure 7.4 a**), the missense variant p.S534N was maternally inherited from the proband E53_3 and and the sibling E53_4, who also had full diagnosis of SLI (**Table 7.5**).

In family G67 (**Figure 7.4 b**), p.S534N was present in the proband G67_4 and absent in the sister G67_3, whose affection status was classified as unknown (**Table 7.5**). Both children carried another missense change in the same exon, p.I543V. We could not assess the inheritance of these mutations, because DNA for the parents was not available.

In family E35 (**Figure 7.4 c**), the variant p.S534N was maternally inherited by the proband E35_3 and his sister E35_5, while it was absent in the brother E35_4. For both siblings E35_4 and E35_5, the affection status was classified as unknown (**Table 7.5**). The father instead carried the missense change p.I543V, on exon 13, inherited by all children (**Figure 7.4**).

From the segregation pattern of p.S534N and p.I543V then, we deduced that, in both families, these variants were carried by different parental alleles.

The probands of these three families did not carry the splicing mutation rs3918290 and the exome sequencing did not detect other variants in the coding region of *DPYD*.

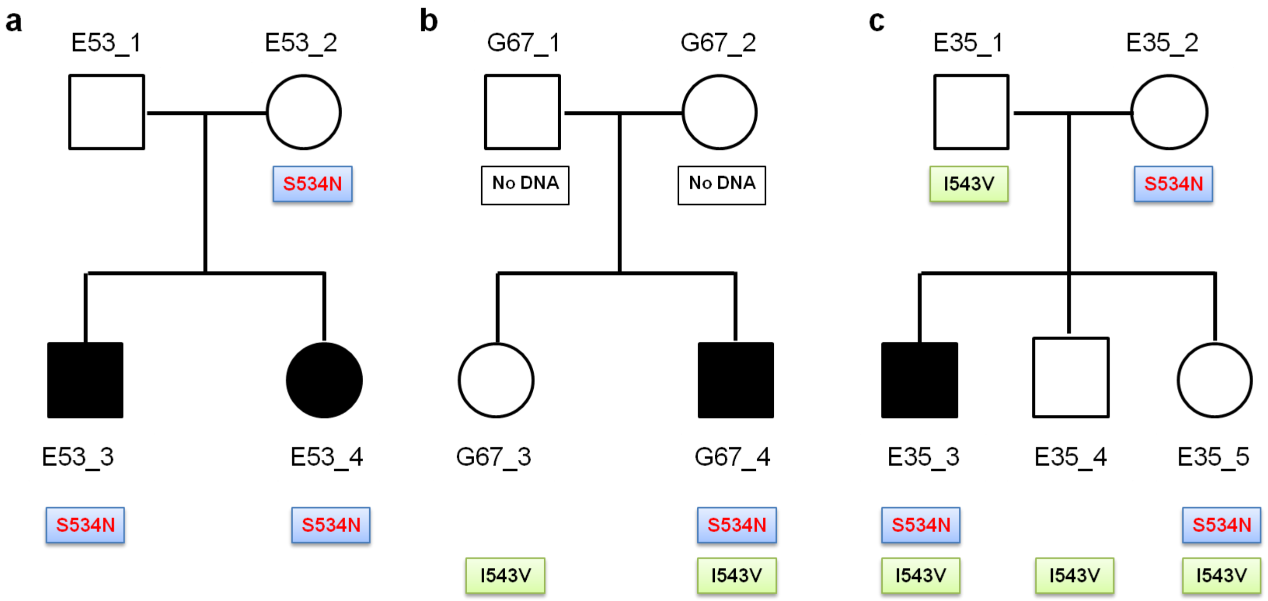


Figure 7.4. Pedigree of the three families carrying the p.S534N variant.

individuals	gender	proband	ELS	RLS	NWR	AFF	S534N (minor allele: A)	I543V (minor allele: G)
E53_1	male	Father			91	**0**	G/G	A/A
E53_2	female	Mother			55	**0**	**G/A**	A/A
E53_3	male	Proband	64	65		**2**	**G/A**	A/A
E53_4	female	Sister	73	85	77	**2**	**G/A**	A/A
G67_1	male	Father				**0**	no DNA	no DNA
G67_2	female	Mother			108	**0**	no DNA	no DNA
G67_3	female	Sister			106	**0**	G/G	**A/G**
G67_4	male	Proband	62	74	66	**2**	**G/A**	**A/G**
E35_1	male	Father			104	**0**	G/G	**A/G**
E35_2	female	Mother			91	**0**	**G/A**	A/A
E35_3	male	Proband	72	72	92	**2**	**G/A**	**A/G**
E35_4	male	Brother	86	93	103	**0**	G/G	**A/G**
E35_5	female	Sister			116	**0**	**G/A**	**A/G**

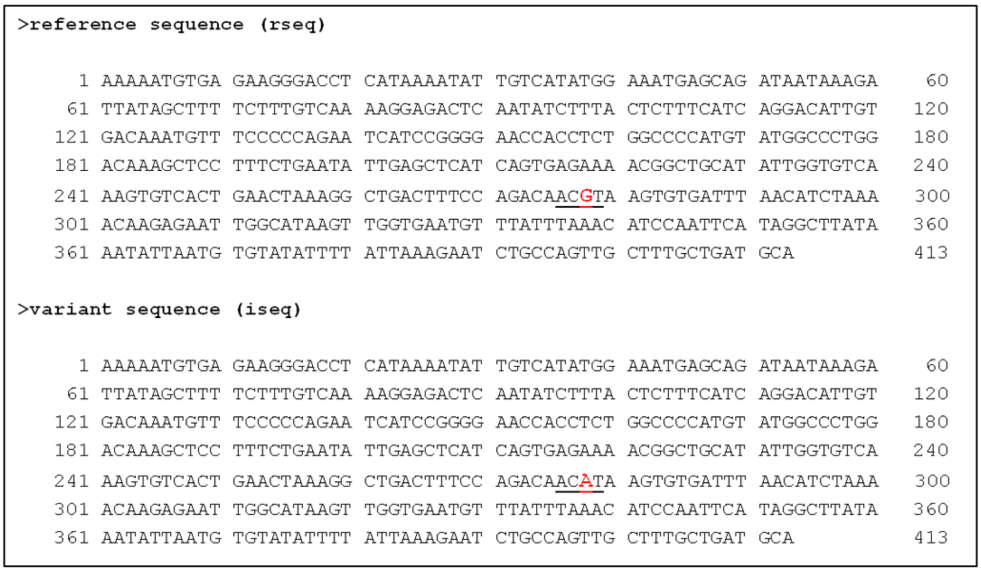
Table 7.5. Phenotypic scores for ELS, RLS and NWR tests available for the three families carrying the p.S534N variant. AFF=2 means a full diagnosis of SLI, AFF=0 means affection status unknown.

7.2 Analysis of the candidate gene *DPYD* in ASD cases.

7.2.1 Screening for the splicing variant rs3918290 in ASD cases and controls.

Since rare variants in *DPYD* have been reported for autism and autistic features have been observed in patients with DPD deficiency, we decided to investigate also the frequency of the splice variant rs3918290 in ASD cohorts. The analysis was carried out in two stages. In the first one, the SNP rs3918290 was genotyped in 231 Italian simplex families, 224 IMGSAC multiplex families and 449 Italian unrelated controls, with a restriction endonuclease analysis. Subsequently, we extended the analysis to the entire cohort of families of the AGP Consortium, which have been genotyped with the Illumina Infinium 1M-single and 1M-duo SNP array, that include the variant of interest rs3918290.

Given the size of the ASD cohorts in the first stage, we developed a restriction enzyme assay to allow the genotyping of the SNP rs3918290. The assay used the restriction endonuclease HpyCH4IV which specifically recognizes the site 5'...A|C**G**T...3', which also corresponds to the sequence that includes the splice donor site of intron 14. Therefore, HpyCH4IV is able to cut a PCR fragment containing the splice site with the reference allele G. In our case, the PCR fragment used to amplify the exon 14 and the surrounding intronic regions had a size of 413 bp and a unique target site for HpyCH4IV (**Figure 7.5**). In the presence of the *wild-type* donor splice site, the PCR fragment was cut in two fragments of 278 bp and 135 bp. In the presence of the variant rs3918290 or another mutation altering the target sequence, the restriction site was not recognized by the enzyme HpyCH4IV and the PCR fragment was not cut.



```
>reference sequence (rseq)

    1 AAAAATGTGA GAAGGGACCT CATAAAATAT TGTCATATGG AAATGAGCAG ATAATAAAGA    60
   61 TTATAGCTTT TCTTTGTCAA AAGGAGACTC AATATCTTTA CTCTTTCATC AGGACATTGT   120
  121 GACAAATGTT TCCCCCAGAA TCATCCGGGG AACCACCTCT GGCCCCATGT ATGGCCCTGG   180
  181 ACAAAGCTCC TTTCTGAATA TTGAGCTCAT CAGTGAGAAA ACGGCTGCAT ATTGGTGTCA   240
  241 AAGTGTCACT GAACTAAAGG CTGACTTTCC AGACAACGTA AGTGTGATTT AACATCTAAA   300
  301 ACAAGAGAAT TGGCATAAGT TGGTGAATGT TTATTTAAAC ATCCAATTCA TAGGCTTATA   360
  361 AATATTAATG TGTATATTTT ATTAAAGAAT CTGCCAGTTG CTTTGCTGAT GCA          413

>variant sequence (iseq)

    1 AAAAATGTGA GAAGGGACCT CATAAAATAT TGTCATATGG AAATGAGCAG ATAATAAAGA    60
   61 TTATAGCTTT TCTTTGTCAA AAGGAGACTC AATATCTTTA CTCTTTCATC AGGACATTGT   120
  121 GACAAATGTT TCCCCCAGAA TCATCCGGGG AACCACCTCT GGCCCCATGT ATGGCCCTGG   180
  181 ACAAAGCTCC TTTCTGAATA TTGAGCTCAT CAGTGAGAAA ACGGCTGCAT ATTGGTGTCA   240
  241 AAGTGTCACT GAACTAAAGG CTGACTTTCC AGACAACATA AGTGTGATTT AACATCTAAA   300
  301 ACAAGAGAAT TGGCATAAGT TGGTGAATGT TTATTTAAAC ATCCAATTCA TAGGCTTATA   360
  361 AATATTAATG TGTATATTTT ATTAAAGAAT CTGCCAGTTG CTTTGCTGAT GCA          413
```

Figure 7.5. The figure shows the PCR fragment used to amplify exon 14 and the surrounding intronic regions. This fragment contains only one restriction site for the enzyme HpyCH4IV and the position of the SNP rs3918290 is indicated in red: it is represented by the base G in the reference sequence, by the base A in the variant sequence.

Visualization of the restriction products on a 2% agarose gel allowed the discrimination of homozygous individuals of the reference allele at the splice donor site (G/G), who presented two bands (278 bp and 135 bp), and heterozygotes individuals (G/A), who displayed three bands (the 278 bp and 135 bp fragments derived from the reference allele G and the 413 bp fragment from the mutated allele A at the splice donor site) – **Figure 7.6**. Homozygous individuals for the mutated allele (A/A), would have shown only a band of 413 bp, but were not identified in this screening. Since a mutation of any of the four bases of the target site could prevent the recognition by the enzyme, the presence of the splicing variant in all identified heterozygotes individuals was then confirmed by Sanger sequencing.

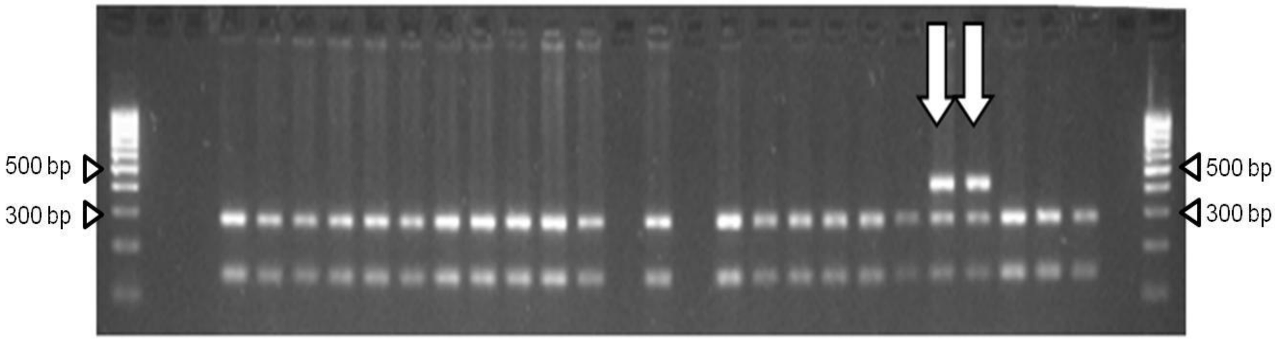


Figure 7.6. Example of genotyping of the SNP rs3918290 using the restriction endonuclease HpyCH4IV. The arrows indicate two of the individuals identified during the screening, carrying the variant rs3918290 (G/A) in the heterozygous state. The ladder 100 bp was loaded in the first and the last lane.

In the IMGSAC cohort, which is formed by multiplex families, the screening was performed on one affected case per family. In cases where mutations were identified, all family members were screened in order to check the inheritance of the variant, the segregation pattern and detect possible *de novo* mutations.

From restriction endonuclease analysis in the sample of 224 IMGSAC cases, rs3918290 was identified in two families: 203 and 171. In family 203, the variant was transmitted from the father 203.1 to the affected son 203.3, but not to the other affected child 203.4. In family 171, the variant was transmitted from the mother 171.5 to the affected son 171.6 and to the unaffected child 171.7. The presence of the variant and its segregation were confirmed in both families with Sanger sequencing (**Figure 7.7**).

In the Italian ASD cohort, including 231 cases, restriction endonuclease analysis was performed on all available family members and identified 9 families in which one or more individuals showed a 3 bands digestion profile. Sanger sequencing confirmed the presence of the SNP rs3918290 in 8 out of 9 families (**Figure 7.7**). In 5 of them (SM2, SM38, SM161, C19 e C22), the variant was present in the proband, inherited from one of the parents, whereas in three families (SM23, SM77 e C38) it was carried by one parent but not transmitted to the children. In the ninth family, the mother

SM81.2 and the unaffected child SM81.4 showed the presence of a different variant that altered the target site of the enzyme HpyCH4IV (5'…A**C**GT…3' →5'…A**T**GT…3'), corresponding to a rare synonymous SNP (rs3918289, aminoacid N635) of the last base of the exon 14 (**Figure 7.7**).

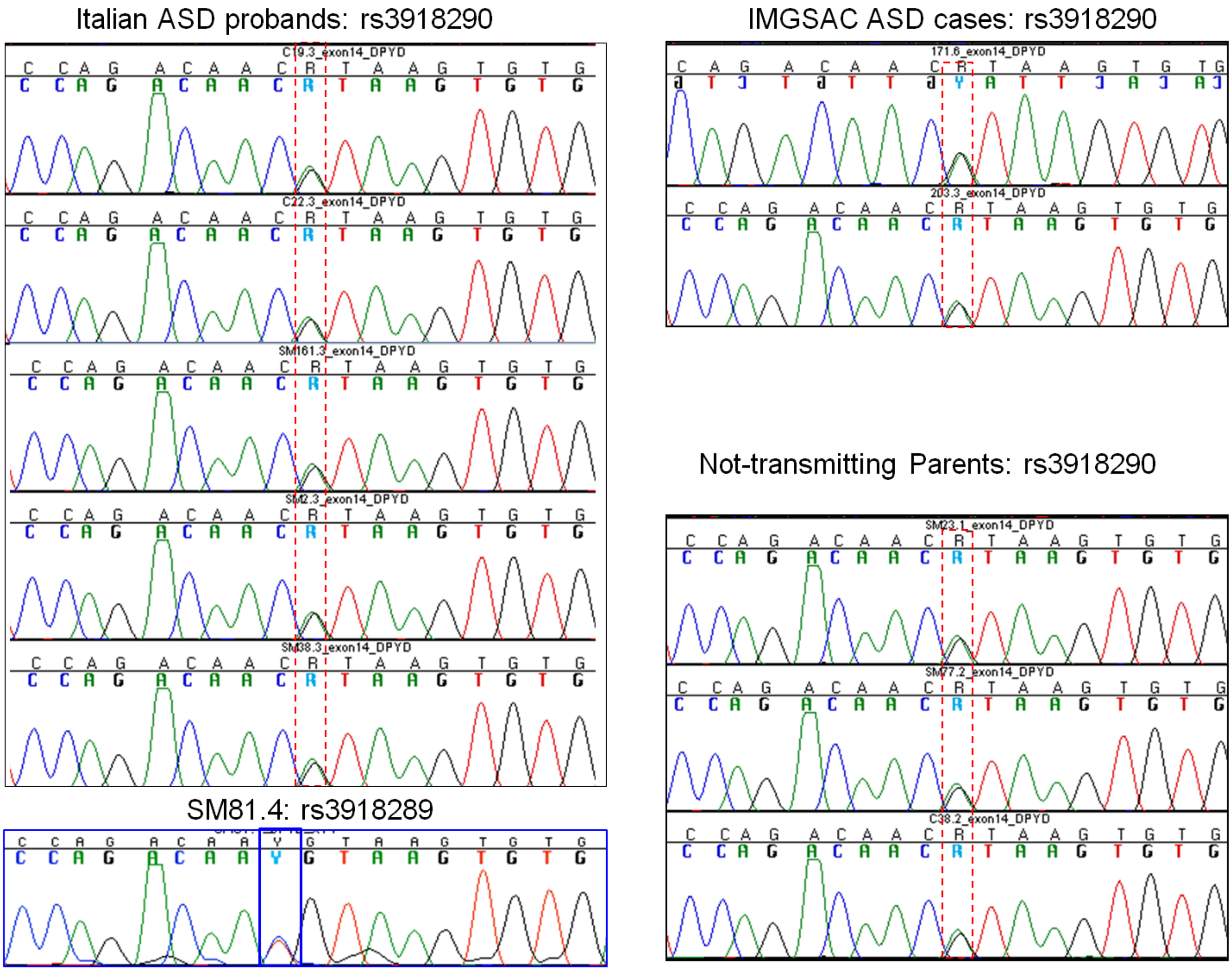


Figure 7.7. The figure shows the electropherograms of the individuals in which the Sanger sequencing confirmed the presence of the variant rs3918290 (G/A, indicated by the red rectangle): the Italian probands SM38.3, C22.3, SM161.3, SM2.3 and C19.3; the IMGSAC cases 203.3 and 171.6; and the parents SM23.1 (father), SM77.2 (mother) e C38.2 (mother). In family SM81 instead, the Sanger sequencing revealed the presence of the SNP rs3918289 (C/T, indicated by the blue rectangle), that corresponds to the last base of exon 14 and to the second base of the HpyCH4IV target site (5'…A**C**GT…3' →5'…A**T**GT…3').

The analysis of 449 unrelated controls identified 4 subjects carrying the splice site mutation rs3918290 in the heterozygous state, all confirmed by Sanger sequencing.

7.2.2 Statistical analysis of the rs3918290 frequency in the Italian and IMGSAC cohorts.

The SNP rs3918290 was identified in 7 individuals with ASD (171.6, 203.3, SM2.3, SM38.3, SM161.3, C22.3, C19.3) and 4 unaffected controls. Thus, the splicing variant is present in 5/231 Italian independent probands, with an allelic frequency of 1.08% versus an allelic frequency in Italian controls of 0.45% (4/898). Although the frequency in affected individuals is higher than in controls, as observed in the SLI cohort, the difference is not statistically significant (*p-value* = 0.1541, one-tailed Fisher's exact test).

The allelic frequency observed in our control group is similar to the one observed in the sample of Tuscan controls in the database of 1000 Genome Project, with a Minor Allele Frequency (MAF) for rs3918290 in these individuals of 0.5% (A=1/G=195) (**Table 7.6**).

In the IMGSAC sample, in which the families were all Caucasian but recruited from different Countries (UK, Netherlands, France, USA, Germany, Denmark, Greece), the splicing variant was identified in 2 out of 224 families. Since these families included multiple affected individuals, in order to calculate the allelic frequency of the variant in independent individuals we selected *a priori* one affected member in each family and we obtained an estimated allelic frequency of 0.45% (2/448). We compared this frequency with the data reported in public databases for samples of European origin, already shown earlier in **Table 7.1** The frequency of the minor allele A in the CEU samples of the HapMap project is 0.4% (A=1/G=225), in the CEU samples of the 1000 Genome Project is 0.6% (A=1/G=169) and in European-American (EA) individuals of the ESP database is 0.6% (A=50/G=8550) (**Table 7.6**). Therefore, the allelic frequency in the IMGSAC cases was similar to the MAF reported in ethnicity-matched controls, showing no difference between the cases and controls. Therefore, the findings obtained from the multiplex families of the IMGSAC cohort did not replicate the trend observed for the Italian ASD simplex families and the SLIC families.

7.2.3 Frequency and association analysis of rs3918290 in the AGP ASD cohort.

To further investigate this discrepancy and analyse this variant in a larger number of ASD individuals, we extended the analysis to the entire AGP cohort. The genotyping of more than 1M SNPs distributed across the whole genome has been recently performed in 2,705 families, simplex and multiplex, collected by the AGP Consortium (Anney et al., 2012). Since the SNP rs3918290 was included among the markers interrogated by the SNP arrays used for the genotyping (Infinium 1M and 1M-Duo), we calculated the frequency of the splicing variant in this larger cohort of ASD subjects. The SNP rs3918290 was successfully genotyped in 2,681 independent cases and the allele A was detected in the heterozygous state in 25 of them, giving an allelic frequency of 0.47% (25/2681 probands), again not significantly different from the frequency reported in European populations (**Table 7.6**) and supporting that observed in the IMGSAC samples. When we compared this frequency with the one reported in the ESP database (EA: 0.58%), we did not find a significant difference (one-tailed χ^2 test, *p-value*= 0.1826).

	heterozygotes	tot individuals	tot alleles	Allelic freq. %
Italian ASD probands	5	231	462	1.08
Italian Controls	4	449	898	0.45
1000 Genomes Tuscans	1	98	196	0.51
IMGSAC cases	2	224	448	0.45
AGP cases	25	2681	5362	0.47
HapMap CEU controls	1	113	226	0.44
1000 Genomes CEU controls	1	85	170	0.59
ESP EA controls	50	4300	8600	0.58

Table 7.6. Summary of the allelic frequencies of rs3918290 used in the statistical analyses.

Given the availability of genotype information for the parents of the affected individuals of the AGP cohort, we also performed a *Transmission Disequilibrium Test* (TDT) on these AGP samples, using the program PLINK. However, as shown by the **Table 7.7**, we did not observe a preferential transmission of allele A to individuals with ASD, therefore, the splicing variant did not result to be associated with the ASD phenotype in the AGP families.

SNP	allele	% in cases	T	NT	O.R.	CHISQ	P
rs3918290	A	0.5%	25	25	1	0	1
	G	99.5%	25	25			

Table 7.7. Output of the TDT analysis in the AGP cohort: this table shows that there is not a preferential transmission of allele A to the affected children, compared to allele G (T= parents transmitting allele A to the affected child; NT: parents NOT transmitting allele A to the affected child). We found 25 parents transmitting allele A versus 25 parent who did not transmitted allele A to the affected children, *p-value* =1.

7.2.4 TDT analysis for *DPYD* and *MIR137* in the AGP cohort.

Rare structural variants involving *DPYD* and its adjacent gene *MIR137* have been found in ASD cases (Carter et al., 2011; Prasad et al., 2012) and in individuals with ID (Willemsen et al., 2011). In schizophrenia, rare sequence mutations have been identified in *DPYD* (Xu et al., 2012), while common variant association has been detected for the region surrounding *MIR137* (Ripke et al., 2013; Schizophrenia_Psychiatric_GWAS_Consortium, 2011) (**Figure 7.8**).

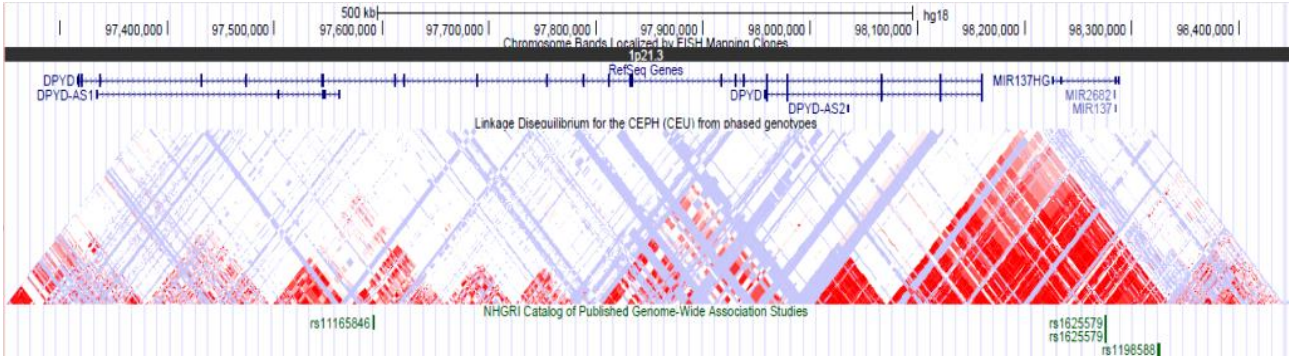


Figure 7.8. The figure shows the genomic location of *DPYD* and *MIR137* (UCSC Genome Browser, hg18). LD plots SNPs genotyped in the HapMap CEPH (CEU) population are shown. In addition, the position of SNPs found in association with schizophrenia, rs1625579 (Schizophrenia_Psychiatric_GWAS_Consortium, 2011) and rs1198588 (Bergen et al., 2012), is indicated.

Since *MIR137* and *DPYD* are both interesting candidates for neurological phenotypes, it has been suggested that both genes could be implicated in neurodevelopmental disorders. Therefore, in order to investigate whether variants in the two genes were in association with the ASD phenotype in our sample, we decided to extend the TDT analysis to the SNPs of the genomic region including *DPYD* and *MIR137* (chr1:97499599-98599144, hg19), that have been genotyped in the AGP cohort. The analysis was performed on ninety-nine independent markers (i.e. not in LD), including the SNP rs1625579 previously found in association with schizophrenia. Few SNPs reached a *p-value* <0.05 (minimum *p*-value= 0.01494, **Table 7.8**), all within *DPYD*. However, the signals for the SNPs that showed a preferential transmission of the minor allele to the affected individuals did not survive multiple testing correction.

coordinates	position	SNP	A1	A2	T	U	OR	CHISQ	*P*-value
chr1:98336940	intron 2	rs17117281	A	G	0	4	0	4	0.0455
chr1:98296425	intron 2	rs11802430	G	A	15	6	2.5	3.857	0.04953
chr1:98226950	intron 3	rs6604874	A	C	5	0	NA	5	0.02535
chr1:98198206	intron 4	rs4554755	A	G	1097	1214	0.9036	5.923	0.01494
chr1:98079228	intron 8	rs7533902	A	G	959	856	1.12	5.845	0.01562
chr1:98075118	intron 8	rs12068454	A	G	21	10	2.1	3.903	0.04819
chr1:98073410	intron 8	rs12144462	A	G	392	331	1.184	5.147	0.02329
chr1:98036219	intron 11	rs2811216	A	G	6	16	0.375	4.545	0.03301
chr1:98014940	intron 12	rs2786525	A	C	6	16	0.375	4.545	0.03301
chr1:98003908	intron 12	rs4950041	G	A	1006	911	1.104	4.708	0.03002
chr1:98001389	intron 12	rs12063030	A	C	1001	911	1.099	4.236	0.03957
chr1:98000861	intron 12	rs2786519	G	A	1016	918	1.107	4.966	0.02585
chr1:97578236	intron 20	rs12076846	G	A	3	11	0.2727	4.571	0.03251

Table 7.8. The SNPs that obtained a *p*-value< 0.05 in the TDT analysis for the region chr1:97499599-98599144, performed on the AGP cohort, are listed in this table. For each SNP, the position in the gene *DPYD* (NM_000110.3) is indicated. A1 indicates the minor allele, A2 the major allele. The numbers of transmitted minor allele count (T) and untrasmitted allele count (U) are reported. Odds ratio (OR) and *p*-value calculated for each SNP are also reported.

Among the SNPs examined in this TDT analysis, three of the missense variants identified during the mutation screening of *DPYD* carried out in the SLI families (p.C29R, p.M166V and p.I543V) were included. However, we did not detect a statistically different transmission of the minor allele to the affected individuals for any of these coding variants (**Table 7.9**).

coordinates	position	SNP	A1	A2	T	U	OR	CHISQ	*P*-value	aa change
chr1:98348885	exon 2	rs1801265	G	A	881	872	1.01	0.046	0.8298	**C29R**
chr1:98165091	exon 6	rs2297595	G	A	449	421	1.067	0.901	0.3425	**M166V**
chr1:97981395	exon 13	rs1801159	G	A	842	819	1.028	0.319	0.5725	**I543V**

Table 7.9. Results of the TDT association analysis for 3 common missense changes identified during the mutation screening of the gene *DPYD*.

7.2.5 Mutation screening of *DPYD* in ASD probands carrying the variant rs3918290.

In order to identify other possible variants in *DPYD* in the 7 ASD probands from the Italian and IMGSAC cohorts carrying the variant rs3918290, we completed a mutation screen of the gene in these individuals. The coding regions of both isoforms, the splice sites and the two regulatory regions at 5’ flanking region of the gene (Shestopal et al., 2000) were examined by Sanger sequencing. This screening identified 2 missense variants in four families: one in exon 2 (rs1801265) and another one in exon 13 (rs1801158), identified also in the SLI samples (**Table 7.3**).

The SNP rs1801265 (p.C29R), a common variant predicted to have neutral effect for the protein, was identified in 3 families: 171, SM2 and C22.

In the IMGSAC family 171 (**Figure 7.9 a**):

- the splicing variant rs3918290 was inherited from the mother 171.5 by the proband 171.6, and by the unaffected brother 171.7;
- the missense change p.C29R was present in the heterozygous state in the proband 171.6, the mother 171.5 and the unaffected brother 171.7, whereas the father 171.5 carried the minor allele C in the homozygous state. This segregation pattern indicated that the children had the two variants on different chromosomes, the rs1801265 on the paternal chromosome and the rs3918290 on the maternal one.

In family SM2 (**Figure 7.9 b**):

- the splicing variant rs3918290 was inherited from the father by the proband SM2.3, and not by the unaffected brother SM2.4;

- the missense change p.C29R was inherited by the proband SM2.3 from the father and was absent in the unaffected brother SM2.4, therefore we deduced that this variant lies on the same paternal haplotype that carries the splicing variant.

In family C22 (**Figure 7.9 c**):

- the splicing variant rs3918290 was inherited from the mother C22.2 by the proband C22.3;
- both parents carried the missense variant p.C29R in the heterozygous state and the proband C22.3 was found to be heterozygote as well, therefore it was not possible to establish whether the two variants identified in the proband were on the same parental haplotype.

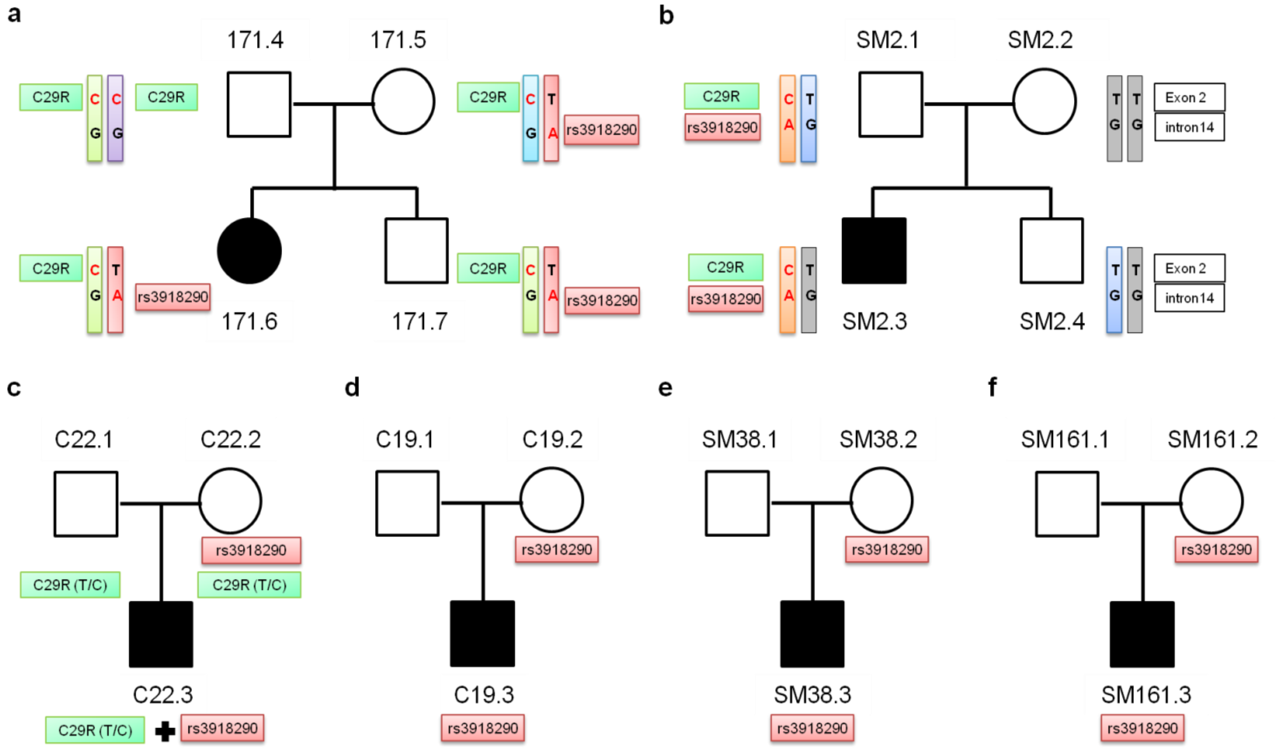


Figure 7.9. Pedigree of six families included in the mutational screening of *DPYD*. The SNP G/A in the splice donor site of intron 14 (rs3918290) is indicated in red, the SNP rs1801265 (T/C, p.C29R) in exon 2 is indicated in green. For the pedigrees 171 (a) and SM2 (b), the haplotypes for *DPYD* were deduced from the segregation of these two variants. Families C19 (d), SM38 (e) and SM161 (f) carried only the splicing variant rs3918290.

In the IMGSAC multiplex family 203, we identified the splicing variant and the missense change p.S534N (rs1801158) (**Figure 7.10**):

- the splicing variant rs3918290 was inherited from the father by the affected child 203.3, and not by the affected sibling 203.4;
- the missense change p.S534N was identified again in the proband 203.3, inherited from the father, but not in the affected brother 203.4. From this segregation pattern, we deduced that the two mutations are likely to lie on the same paternal haplotype, which is not shared by the affected siblings.

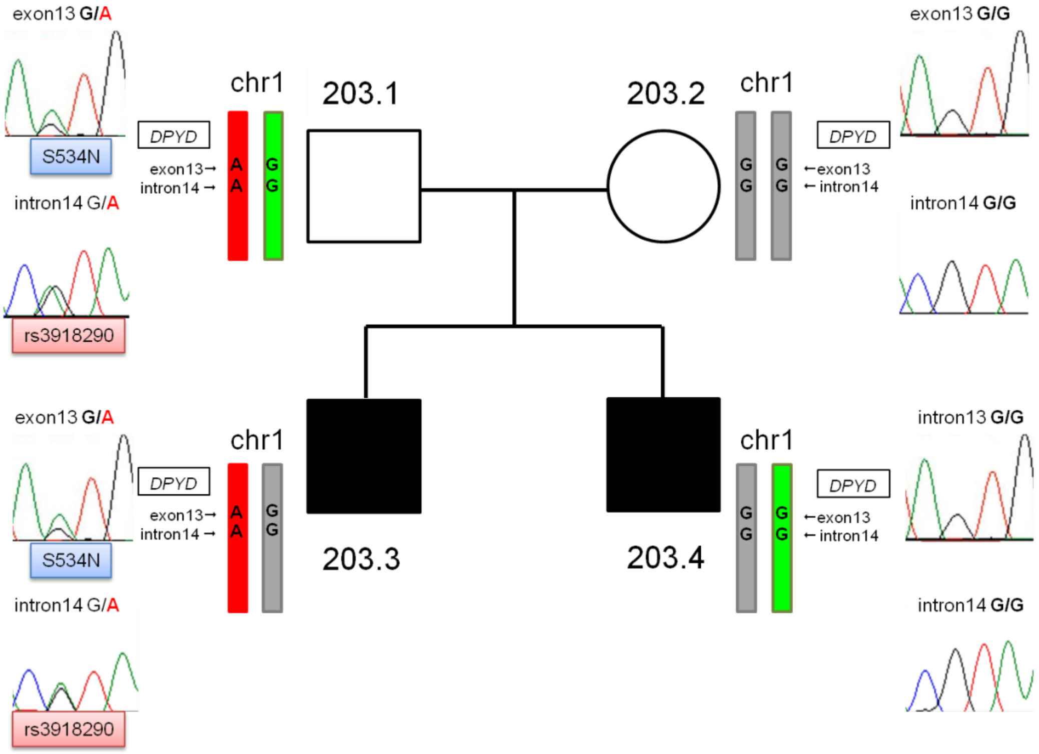


Figure 7.10. Pedigree of family 203. The electropherograms of the two SNPs identified in *DPYD* are shown for each family member. Based on these variants we could discriminate between the paternal haplotypes, but not between the maternal haplotypes (and for this reason they are indicated in grey). As represented in the figure, the siblings inherited different paternal haplotypes for *DPYD*.

7.2.6 IQ analysis of ASD individuals carrying the variant rs3918290.

In order to evaluate whether the splicing variant rs3918290, and therefore the partial loss of DPD, could have an effect on cognitive functioning in ASD individuals carrying this variant, we decided to analyse the IQ scores of these subjects in comparison with the entire AGP cohort.

From this large group, Full scale IQ (FIQ) measures were available for 1246 affected individuals. 37% of the subjects had normal cognitive functioning (FIQ≥86). The mean FIQ was 76 with a standard deviation of 25. Verbal IQ (VIQ) and performance IQ (PIQ) scores were also available for the majority of ASD cases genotyped for the SNP rs3918290. The VIQ, PIQ and FIQ scores were analysed separately and compared between ASD patients carrying the splicing variant in *DPYD* and those who were *wild-type* for the SNP, using a non parametric test, the Wilcoxon-Mann-Whitney test. **Figure 7.11** shows that the distribution of the FIQ scores in ASD cases with the variant r3918290 did not significantly differ from the distribution of values measured in individuals without the variant. In contrast, non-verbal (PIQ) scores were significantly higher in individuals with the splicing mutation (*p-value*=0.02962). Although the median VIQ value for individuals carrying the splicing variant (median VIQ=59) was lower than that of *wild-type* individuals, the distribution of VIQ scores in the two groups did not show significant differences (*p-value*=0.5628).

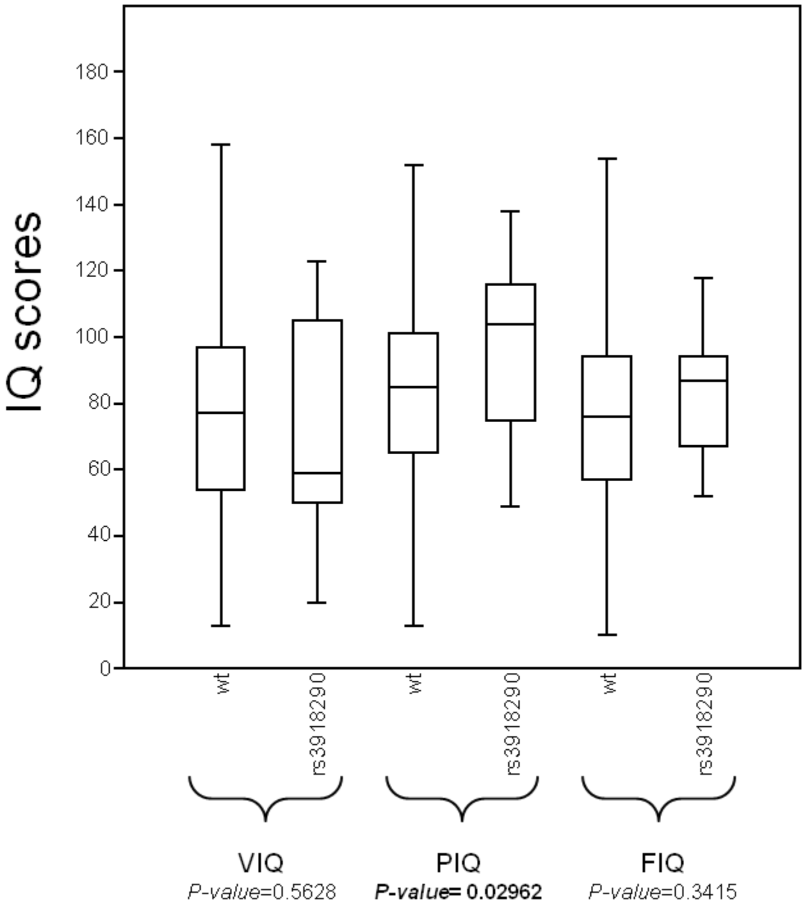


Figure 7.11. The graph shows verbal, performance and full-scale IQ score analyses in ASD cases of the AGP cohort, genotyped for the SNP rs3918290. For each IQ measure, the distribution of the scores in individuals with the splicing variant (rs3918290) and individuals *wild-type* for the SNP (wt) is represented with a box plot. *P-values* were obtained from the Wilcoxon-Mann-Whitney test.

7.2.7 Analysis of language phenotypes in ASD individuals carrying the variant rs3918290.

The autism affection status of the individuals of the AGP cohort was derived from two diagnostic tools: Autism Diagnostic Interview-revised (ADI-R) (Lord et al., 1994) and Autism Diagnostic Observation Schedule-Generic (ADOS-G) (Lord et al., 2000). The ADOS-G schedule consists of a series of interacting tasks that aim to examine communication and behaviours in a standardized context. ADI-R instead is a structured interview conducted with the parents of individuals referred for a possible ASD condition. The questionnaire spans the three main domains that are impaired in autism: language and communication skills, social interactions and patterns of behaviour. Since we hypothesized the involvement of *DPYD* in SLI, we investigated a possible influence of the variant rs3918290 on language development of ASD individuals of the AGP cohort. We examined ADI-R items relative to language development and abilities (the overall level of current language, the age of first single words and the age of first phrases) in individuals for whom the rs3918290 has been successfully genotyped.

The ADI-R item regarding the overall level of current language evaluates the spontaneous use of social language, based on the formulation of sentences of at least three words. We performed a qualitative analysis dividing the children into two categories: “normal” language development (ability of formulating phrases with more than 5 words) and “delayed” language development (inability of formulating phrases with more than 5 words). For each category, the number of affected children carrying the splice variant rs3918290 was compared to affected children who are *wild-type* for rs3918290, using a χ^2 test. However, no statistical difference between the two categories was detected (**Table 7.10**).

Overall level of language	**normal**	**delay**	total	**χ2**	***P-value***
with rs3918290 (G/A)	21	5	26	1.4011	0.2365
without rs3918290 (G/G)	1831	781	2612		
total	1852	786			

Table 7.10. Overall level of language abilities in ASD cases of the AGP cohort, genotyped for rs3918290. For each category (“normal” or “severe delay”), the number of individuals (with or without the variant rs3918290) is reported. The *p-value*, calculated with a χ^2 test, does not indicate a statistical significant difference between the two groups.

The same qualitative analysis was performed for “age of first words” and “age of first phrase” items. Generally, most babies say meaningful first words by age of 18 months. When this milestone is not achieved within 24 months, language abilities are considered abnormal. We performed a qualitative analysis subdividing the children in two categories: “normal” (age of first words < 24 months) and “delayed” onset of speech (age of first words ≥ 24 months). The results of the χ^2 test are shown in **Table 7.11**. No statistical difference between the two categories was detected.

Age of first words	**Normal**	**Delayed**	total	**χ2**	***P-value***
with rs3918290 (G/A)	12	12	24	0.563617261	0.452806
without rs3918290 (G/G)	1075	1461	2536		
total	1087	1473			

Table 7.11. Comparison of ASD cases of the AGP cohort, for whom age of first words and rs3918290 genotypes were available. The *p-value*, calculated with a χ^2 test, does not indicate a statistical significant difference between the two groups.

Within 36 months, children should be able to formulate simple phrases, formed by at least two or three words. Children who do not develop this ability within 3 years are referred for language delay. The results of the qualitative analysis performed comparing the number of affected children carrying splice variant rs3918290 with the number of ASD children without the variant, taking into account “normal” (age of first phrases < 36 months) and phrase speech “delay” (age of first phrases ≥ 36 months), are shown in **Table 7.12**. As for the two previous tests, no statistical difference between the two categories was identified.

Age of first phrases	Normal	Delayed	total	χ2	P-value
with rs3918290 (G/A)	5	19	24	0.1698095	0.68028
without rs3918290 (G/G)	617	1905	2522		
total	622	1924			

Table 7.12. Comparison of ASD cases of the AGP cohort, for whom age of first phrases and rs3918290 genotypes were available. The *p-value*, calculated with a χ^2 test, does not indicate a statistical significant difference between the two groups.

In conclusion, we investigated the role of *DPYD* in SLI and ASD, focusing on the most commonly implicated variant in DPD deficiency, rs3918290. In the SLI cohort and the Italian ASD cohort, we observed a slightly increased frequency of this SNP, compared to controls. However, this trend was not replicated in the IMGSAC cohort of multiplex families or the large number of cases belonging to the AGP cohort. Furthermore, in the AGP cohort, we did not observe a preferential allelic transmission of the minor allele of rs3918290 or other missense changes identified during mutation screening of the gene (p.C29R, p.M166V, p.I543V).

In order to test whether the presence of the damaging splicing variant could have an effect on phenotypic features rather than ASD diagnosis *per se*, we analysed IQ scores of ASD cases carrying the rs3918290, but we found that they did not significantly diverge from the cognitive abilities of the rest of the cases, except for PIQ, that was reported to be slightly higher in the rs3918290-carriers. When we stratified the AGP samples for language endophenotypes (age at first words, age at first phrases, overall level of language), again, we did not observe a significant difference in language delay between individuals with the splicing variant and individuals without it. However, it is worth noting that we examined a variant which is rare in the population, therefore our analyses included only a very small number of cases carrying the variant rs3918290, even in the large AGP cohort. The mutation screening of *DPYD* in the ASD probands with the splicing mutation identified two common missense changes, p.C29R and p.S534N, and their putative effect on DPD activity will be later discussed.

Chapter 8

Discussion

8.1 CNV screening in the SLIC cohort.

Specific language impairment (SLI) is a common neurodevelopmental disorder diagnosed in children with an unexpected failure in the acquisition of language abilities, given adequate educational opportunities and in the absence of other medical conditions that could have an influence on language, such as hearing deficits or intellectual disability (Tomblin et al., 1996).

Twin and family studies have provided evidence for the role of a strong genetic background in SLI, but the inheritance pattern suggests that several loci and environmental factors contribute to the overall risk in a complex manner (Stromswold, 1998).

So far, the genetic bases of SLI have been investigated by genome-wide linkage studies, which yielded borderline significant *p-values* for few susceptibility loci: SLI1 on chromosome 16q, SLI2 on chromosome 19q (SLIC, 2002, 2004), SLI3 on chromosome 13q (Bartlett et al., 2004; Bartlett et al., 2002) and a region on chromosome 7q (Villanueva et al., 2011). Further investigations on the SLI1 region with a targeted association study led to the identification of two candidate genes, *ATP2C2* and *CMIP* (Newbury et al., 2009). Association with SLI has been detected also for variants in *CNTNAP2*, a gene on chromosome 7q35-q36.1 implicated in a broad spectrum of neurodevelopmental disorders. Moreover, two recent GWAs (Eicher et al., 2013; Luciano et al., 2013) have reported association of language skills and related traits with variants in the genes *ABCC13*, *DAZAP1*, *ZNF385D*, *COL4A2* and *NDST4*.

A large portion of the genetic risk factors contributing to SLI, however, remains to be unravelled. Recent studies have shown that CNVs can be an important source of susceptibility to complex psychiatric disorders, such as ASD, ADHD and schizophrenia (Coe et al., 2012). Therefore, in order to investigate whether CNVs play a role also in the genetic architecture of SLI, we performed a genome-wide CNV screen in 540 individuals of the SLIC cohort, formed by families with one or multiple individuals with SLI. To our knowledge, this is the first study of CNVs within a SLI cohort.

A high-confidence list of CNVs was generated using QuantiSNP (Colella et al., 2007) and PennCNV (Wang et al., 2007) algorithms. Moreover, using the annotations in DGV, a database that collects structural variants observed in the general population, we compiled a list of rare and novel CNVs.

A role for rare *de novo* CNVs has been proposed for neuropsychiatric disorders such as ASD (Levy et al., 2011; Sebat et al., 2007), schizophrenia (Xu et al., 2008) and bipolar disorder (Malhotra et al., 2011). Rare *de novo* events are an extremely important source for the discovery of risk variants with a strong penetrance on the phenotype. Therefore, we also generated a list of predicted *de novo* CNVs for the families in which SNP data from both parents were available for the analysis.

In order to investigate whether there are differences in CNV burden between language impaired individuals and subjects with normal language abilities, we carried out preliminary CNV burden analyses comparing the children with a formal diagnosis of SLI (174 individuals from 136 independent families) with unaffected siblings (40 subjects from 38 independent families). These individuals could be considered ideal “super-controls”, as they share the same ethnic background of the affected siblings and their language abilities were assessed with the same tests used for the cases, that excluded the presence of language impairment. Moreover, cases and controls were genotyped on the same SNP array and this allows a comparison between two homogeneous sets of CNV calls. The use of different array types could introduce a bias in the analysis, due to different genomic coverage and resolution and possible array-specific artefacts (Pinto et al., 2011). On the other hand, a major disadvantage of using unaffected siblings as controls is that they are not independent. Hence, further analysis is currently being carried out by the SLI research group in Oxford, using a set of unrelated controls which have been genotyped using the same array (Illumina Human OmniExpress beadchip).

The comparison of several classes of CNVs (the global list of CNVs, only deletions, only duplications, rare and novel CNVs and *de novo* CNVs) did not detect any significant difference between cases and controls. This could be due to the small sample size of the controls and to the fact that the burden analysis did not take into account their non-independent nature. For this reason, these results should be only considered as preliminary, and warrant further investigations using a larger set of independent controls (and this stage is being carried out by the SLI research group in Oxford).

In contrast to CNV screenings for autism, schizophrenia and ID, that found an increased burden of rare/*de novo* CNVs in cases, the exploratory analyses that were performed for SLI yielded a result more similar to that obtained for dyslexia, bipolar disorder and Tourette Syndrome (Coe et al., 2012). As previously discussed (chapter 1, paragraph 1.4), the comparison of CNV findings in different neurodevelopmental disorders with a certain degree of co-morbidity (ID, schizophrenia, autism, bipolar disorder and dyslexia) has shown that the difference in burden of large rare/*de novo* CNV between cases and controls increases with the severity of the phenotype (Coe et al., 2012). In particular, since the highest CNV burden has been observed in Developmental Delay and ID and the lowest in bipolar disorder and dyslexia, it has been suggested that large rare/*de novo* CNVs might correlate with the level of cognitive impairment. This model then postulates that this range of disorders might be considered a part of a continuum and the overall level of disruptive mutations increases with the phenotypic severity (Coe et al., 2012). SLI is a neurodevelopmental condition in which the language domain is specifically impaired, while non-verbal IQ performances are in the normal range. Therefore, according to this model, it is conceivable that a high difference in rare/*de novo* CNV burden might not be observed in cases compared to controls.

Although the class of rare/*de novo* CNVs did not show an overall increased burden, some CNVs might be of interest. The validation of the predicted CNVs will prioritize the events occurring in interesting functional candidates. For example, among the genic CNVs including exonic regions, we identified a rare *de novo* duplication predicted to involve the first two exons of the gene *KLHDC10* (kelch domain containing 10, OMIM 615152). It has been implicated in response to oxidative stress, but the presence of a kelch domain, which is generally involved in protein-protein interactions, suggests that this protein might have various functions. A disrupting inherited deletion in a gene of the same family, *KLHL23*, has been previously identified in an ASD family (Holt et al., 2012). Proteins containing the kelch domain have been also detected as one of the functional categories enriched among *de novo* genic CNVs in schizophrenia (Malhotra et al., 2011). The occurrence of *de novo* mutations in the same gene or genes with related functions in multiple unrelated patients with the same neuropsychiatric disorder or co-morbid conditions might indicate that these genes are implicated in important neurological processes. Therefore, validation of this CNV will be required in order to investigate whether rare variants in *KLHDC10* might be possible contributing factors also for SLI.

Another example of an interesting candidate is given by *CACNA2D1*. Widespread expression of this gene, encoding the α_2-δ1 subunit of Voltage-dependent calcium (Ca_v) channels, has been detected in rat brain with an enrichment in regions involved in learning and memory (Cole et al., 2005). Another gene, coding for the subunit α1-C of voltage calcium channel, *CACNA1C*, has been implicated in schizophrenia and other neuropsychiatric conditions (Bhat et al., 2012) and its inactivation in mouse hippocampus and neocortex has been shown to impair memory formation (White et al., 2008). Weak verbal short term memory, as evidenced by tasks requiring repetition of words or sentences (such as the NWR test), is an important feature of SLI and association with the NWR trait has been found for *ATP2C2*, which codes for a calcium ATPase (Newbury et al., 2009). Calcium is an important ion in the regulation of many neuronal cell functions (Zheng and Poo, 2007), therefore it is plausible that calcium homeostasis might be important also for neuronal processes required for verbal memory and language acquisition.

We also identified a homozygous deletion involving the gene *ZNF277*, predicted to cause the complete lack of exon 5. The absence of this exon is predicted to cause a frameshift that would introduce a premature stop codon in exon 7. Little is known about this gene, but it is conserved throughout evolution and it is expressed in brain (Liang et al., 2000), in particular in neocortex and hippocampus. We therefore validated this exonic homozygous deletion. The characterization and the follow-up of this CNV will be discussed in the paragraph 8.3.

CNVs at loci implicated in microdeletion/microduplication syndromes can be found also in multiple neuropsychiatric disorders, such as ASD, ADHD, epilepsy and ID. For example, during our CNV screening, we identified common duplications in the region 17q21.31. The 17q21.31 locus is associated with *KANSL1*-related ID syndrome, which can be caused by a 500-650 kb heterozygous deletion at chromosome 17q21.31 that includes the gene *KANSL1* (Koolen et al., 2012; Zollino et al., 2012). The KANSL1 protein plays a role in chromatin modification regulating KAT8, which influences gene expression through histone H4 lysine 16 (H4K16) acetylation (Koolen et al., 2012). Reciprocal microduplications have been described, generally, in association with variable, but milder phenotypes (Grisart et al., 2009; Kitsiou-Tzeli et al., 2012). In the SLIC cohort, two smaller *de novo* duplications (one of ~184 kb and the other one of ~112 kb) on chromosome 17q21.31, encompassing the 5' end of the gene, were predicted in two unrelated affected individuals. Although duplications overlapping this region are commonly reported in the DGV, the validation and characterization of the breakpoint positions will be needed to clarify the effect of these CNVs and the potential implication of *KANSL1* gene in SLI.

8.2 CNVs on chromosome 15q11-q13.

A chromosomal region in which CNVs are recurrently observed across multiple neuropsychiatric conditions is the 15q11-q13 locus (Coe et al., 2012; Cooper et al., 2011; Moreno-De-Luca et al., 2013). The presence of complex patterns of highly homologous LCRs makes this locus one of the most unstable regions in the human genome and can lead to a number of different rearrangements (Szafranski et al., 2010). Deletions and duplications are likely to be caused by NAHR between the LCRs, which are organized in clusters, that have been designated BP1-BP6.

Microduplications and microdeletions between BP1 and BP2 include four non-imprinted genes, *TUBGCP5*, *CYFIP1*, *NIPA1* and *NIPA2*, all possible candidates for neuronal alterations (Burnside et al., 2011), while microduplications and microdeletions between the breakpoints BP4 and BP5 include the candidate gene *CHRNA7*, coding for the α7 subunit of the neuronal nicotinic acetylcholine receptor (Shinawi et al., 2009). Both types of CNVs have been associated with a wide spectrum of phenotypes, primarily represented by neurological and developmental problems. CNVs at the BP1-BP2 region have been observed in individuals with language delay and developmental delay (Burnside et al., 2011; Doornbos et al., 2009), ID (Cooper et al., 2011), ASD (Doornbos et al., 2009; Sanders et al., 2011; van der Zwaag et al., 2010), schizophrenia (Kirov et al., 2009; Stefansson et al., 2008) and epilepsy (de Kovel et al., 2010). Interestingly, a retrospective analysis of individuals carrying BP1-BP2 CNVs found that speech delay was one of the most frequent features of microdeletion carriers (Burnside et al., 2011).

CNVs at the BP4-BP5 region have been described in individuals with intellectual disability with seizures (Sharp et al., 2008), autism (Miller et al., 2009; Pagnamenta et al., 2009), schizophrenia (International_Schizophrenia_Consortium, 2008; Stefansson et al., 2008), bipolar disorder (Ben-Shachar et al., 2009), epilepsy (Helbig et al., 2009; Masurel-Paulet et al., 2010) and language delay (Ben-Shachar et al., 2009).

In the CNV screening performed on the SLIC cohort, we identified and validated microdeletions (~500 kb) between the breakpoints BP1 and BP2 in two families, and microduplications (~495 kb) between the breakpoints BP4 and BP5 in another three families.

In family G46, the BP1-BP2 microdeletion was inherited by the proband, but not by the affected sibling. In family E21 instead, the paternal BP1-BP2 deletion was not inherited by the proband, but was inherited by his sister E21_5, whose language status was classified as unknown. The frequency of this deletion in SLI probands was then 0.76% (1 out of 132 independent probands).

In family G68, the BP4-BP5 microduplication, identified in the mother, was not transmitted to the affected children. In family E38, we found that the microduplication in the 15q11-q13 region was not inherited by the proband E38_3, but was inherited by the affected half-sibling E38_4. In family G79 instead, the paternal BP4-BP5 microduplication was inherited by all children with language deficits and was absent in the unaffected child G79_3. As for the BP1-BP2 microdeletion, the frequency of the BP4-BP5 microduplication in SLI probands was 0.76% (1 out of 132 independent probands).

Therefore, both the BP1-BP2 microdeletion and the BP4-BP5 microduplication were identified in affected and unaffected individuals and were not present in all the affected members of these families. This incomplete segregation suggests that these CNVs, if implicated in SLI, might be contributing risk variants with an incomplete penetrance and variable expressivity, as has previously been suggested.

These findings are not surprising, as these types of CNVs can be found also in healthy controls (Cooper et al., 2011; Leblond et al., 2012; Moreno-De-Luca et al., 2013; Shaikh et al., 2009). Although the study carried out by Leblond *et al.* (2012) did not find a different frequency for BP1-BP2 microdeletions between ASD cases (0.32%) and controls (0.25%), the study conducted by

Cooper *et al*. on a large cohort of 15,767 cases with ID, developmental delay and/or other abnormalities and 8,329 healthy controls, found that microdeletions at 15q11.2 were significantly enriched in cases (*p-value*=2.5 x 10^{-5}).

A similar situation has been described for BP4-BP5 microduplications. Although the phenotypes observed in microduplication carriers are similar to those of microdeletion carriers (developmental delay, intellectual disability, ADHD, ASD) (Szafranski et al., 2010), the microduplications do not show an increased frequency in cases of ASD or epilepsy compared to controls (Helbig et al., 2009; Leblond et al., 2012). Therefore, it has been hypothesized that haploinsufficiency of *CHRNA7* might have a stronger effect on the phenotype compared to the microduplication, but this does not exclude that the microduplication could also predispose to neurodevelopmental or neuropsychiatric disorders.

In conclusion, our findings support the hypothesis that these CNVs are not sufficient to cause a pathological phenotype, but may represent modifiers in a range of neurodevelopmental conditions, including SLI, acting in concert with other factors (genetic, epigenetic or environmental), that determine then a variable phenotypic outcome (Leblond et al., 2012; van Bon et al., 2009).

8.3 Identification of a homozygous exonic deletion in *ZNF277.*

During the CNV screening in the SLIC cohort, we identified a novel homozygous microdeletion of exon 5 of the *ZNF277* gene in a proband with severe receptive and expressive language impairment. The characterization of this microdeletion, predicted to result in a frameshift of the transcript and in the introduction of a premature stop codon, allowed the fine mapping of the deletion breakpoints. This 21,379 bp microdeletion was not found in the proband's affected sister or her brother, who had mild language impairment. However, it was inherited from both parents, each of whom carries a heterozygous microdeletion and have a history of language problems. IBD analyses demonstrated that the parents were not consanguineous.

Screening of an additional 321 SLI families indicated that the allelic frequency of *ZNF277* microdeletions was more than twice that observed in control cohorts (1.1% vs 0.4%), although the rarity of the microdeletion meant that this difference did not reach significance when examined with a two-tailed Fisher’s exact test (*p-value* =0.206).

Although a diagnosis of SLI excludes the presence of other clinical conditions that affect language, SLI presents co-morbidity with other neurodevelopmental disorders, such as ASD (as discussed in **paragraph 2.6.2**). Recent studies investigated whether the phenotypic similarities of the language deficits may reflect potentially shared causes. An example of a functional link between SLI and ASD, in addition to other neurodevelopmental disorders, is given by *CNTNAP2* (*contactin* *associated protein-like 2*): variants in and disruptions of this gene are reported to be associated with language endophenotypes in both SLI and ASD (Peñagarikano and Geschwind, 2012), suggesting that *CNTNAP2* might harbour susceptibility risk factors that could impair language skills in distinct language-related disorders.

The *ZNF277* microdeletion falls in the *AUTS1* region of linkage to ASD (7q21-q32, OMIM 209850). In this region, association with ASD was found for two genes that are proximal to *ZNF277*, *DOCK4* (dedicator of cytokinesis 4) and *IMMP2L* (IMP2 inner mitochondrial membrane protease-like) (Maestrini et al., 2010). A rare microdeletion involving these two genes was described in a family with autism and dyslexia: deletions of *DOCK4* were suggested to be a risk factor for reading impairment, and, in conjunction with other variants, for ASD (Pagnamenta et al., 2010).

Therefore, given the genomic position of the *ZNF277* gene within a known ASD risk locus, as well as the prior observations of phenotypic and genetic overlaps between SLI and autism, we postulated that disruption of this gene may be relevant for both disorders. However, screening of a cohort of ASD multiplex families found that the frequency of *ZNF277* microdeletions in individuals with autism was similar to that observed in controls. At the time of the discovery, the *ZNF277* microdeletion that we describe was not documented in the Database of Genomic Variants, possibly because of the small number of standard array SNPs contained within the deleted segment. We specifically searched supplementary data from CNV studies of ASD and found that the microdeletion had been previously characterized during a CNV screen of the Simons Simplex Collection (Sanders et al., 2011) and occurred at a frequency of 0.3% (6/2248 chromosomes). This figure matches the one observed in our ASD cohort, reinforcing our conclusion that the *ZNF277* microdeletion does not contribute to ASD susceptibility.

Since *DOCK4* lies head-to-head with the *ZNF277* gene, we investigated whether there is a reciprocal position effect for deletions. We found that the heterozygous microdeletion of exon 5 in *ZNF277* does not affect the expression of autism candidate genes *IMMP2L* and *DOCK4*. Similarly, a microdeletion involving the 3' end of *DOCK4* (exons 27-52) and the first three exons of *IMMP2L*, decreases the expression level of *DOCK4* but not *ZNF277* (Pagnamenta et al., 2010). Taken together, these data suggest that *ZNF277* microdeletions may play a role in SLI susceptibility that is distinct from the autism risk loci described in the *AUTS1* region.

In the SLI families where DNA of both parents was available, we observed that none of the observed heterozygous *ZNF277* microdeletions was *de novo* and that, in many cases, the segregation with language impairment was incomplete. In three families, the microdeletion was not inherited by the proband, while in another three families, the microdeletion was inherited by unaffected siblings. These data suggest that the heterozygous copy loss of exon 5 of *ZNF277* may represent a low penetrant risk factor rather than a highly penetrant variant, while the homozygous copy loss may be expected to have a greater impact on the SLI susceptibility.

In neurodevelopmental disorders, it is hypothesised that multiple common and rare variants act in concert to determine the phenotype in a complex manner. Under this hypothesis, it is perhaps not surprising to find risk variants in both affected and unaffected family members, or to observe transmission to only a subset of affected individuals of the family. Such findings are consistent with a multigenic threshold model (Cook and Scherer, 2008). Some variants may be highly penetrant, while others may be individually insufficient to cause the disorder, but they may combine with other risk loci and/or environmental factors to cross the risk threshold. Even well-established risk loci for autism, ID, schizophrenia and other neurodevelopmental syndromes provide examples of imperfect segregation, as shown by exonic CNVs in *NRXN1* (Bucan et al., 2009), missense mutations in *SHANK2* (Berkel et al., 2010), rare sequence and structural variants in *CNTNAP2* (Bakkaloglu et al., 2008; Gregor et al., 2011) and microdeletions and microduplications at 16p11.2 (McCarthy et al., 2009; Weiss et al., 2008). Interestingly, single case reports and network analyses indicate that the pathways implicated in different diseases may converge on common genes, such as *CNTNAP2* and *NRXN1*, suggesting that there are some key genes important for several aspects of brain development and, if mutated, they can contribute to a range of disorders, depending on the genetic background. This supports the hypothesis of shared biological pathways among different neuropsychiatric conditions.

Given these data, some researchers proposed a “dual-hit” model (Girirajan et al., 2010). Under such a hypothesis, the phenotypic effects of copy number events, even those of high penetrance, may be modulated by a second independent genetic “insult” which may take the form of an additional CNV or a rare coding mutation. Support for this model comes from studies of single language-impaired cases (Newbury et al., 2013) and larger cohorts of individuals with ASD (Girirajan et al., 2013; O'Roak et al., 2011) or particular microdeletion/duplication syndromes (Girirajan et al., 2012; Girirajan et al., 2010; Leblond et al., 2012). In the current discovery family with *ZNF277* microdeletion, there were no obviously co-segregating second hits. A rare duplication of 72 Kb was observed to occur only in the proband (chr2:41263841-41336618, hg19) and a novel 9 Kb deletion was observed in the mother and sister (chr2:125099924-125109738, hg19). However, neither of these events affect any coding sequence. Larger or more in-depth studies would therefore be required to further investigate possible genetic modulators of *ZNF277* microdeletions. It is conceivable that the phenotypic variability associated with heterozygous *ZNF277* microdeletions might be modulated by a combination of pathogenic single nucleotide variants in other genomic locations, small changes in regulatory regions or other factors.

ZNF277 is an evolutionary conserved zinc finger gene with 12 exons (Liang et al., 2000). It is expressed in several tissues, including the brain, particularly in the neocortex and hippocampus in early mid-fetal development. Although the function of *ZNF277* has not been studied in humans, the mouse *Zfp277* gene, which shows more than 80% homology to the human gene at the amino acid level, has been implicated in the epigenetic regulation of cellular memory (Negishi et al., 2010). *Zfp277*$^{-/-}$ mice were born healthy and fertile, indicating that the knockout is not lethal (Negishi et al., 2010), consistent with our finding of viability for humans with no functional *ZNF277*. Interestingly, the Zfp277 protein directly interacts with Bmi-1, a key component of the Polycomb Repressor Complex (PRC1). This complex has an important role in the maintenance of adult stem cells from numerous tissues, including the central nervous system (Molofsky et al., 2005; Molofsky et al., 2003).

In conclusion, we propose that the disruption of *ZNF277* may contribute to SLI in a complex genetic model. We further hypothesize that this risk is distinct from the autism risk loci previously described in this region. Further studies will be required to replicate these findings and characterize the function of the human protein ZNF277, clarifying its potential implication in language development.

8.4 Analysis of the candidate gene *DPYD* in SLI.

After the identification of the *ZNF277* microdeletion in the proband G4_4, we consulted the clinicians who recruited the family, in order to obtain additional phenotypic information. From the clinical reports, we found that the affected sister G4_5, who did not carry the *ZNF277* deletion but was affected with SLI, had a metabolic disorder known as dihydropyrimidine dehydrogenase (DPD) deficiency. This autosomal recessive disease is caused by mutations in the gene encoding the DPD enzyme, *DPYD*. Partial or complete loss of the activity of DPD, which is the initial and rate-limiting enzyme of the catabolism of pyrimidines, causes an accumulation of uracil and thymine and reduces the amount of the neurotransmitter β-alanine produced by this pathway (**Figure 8.1**).

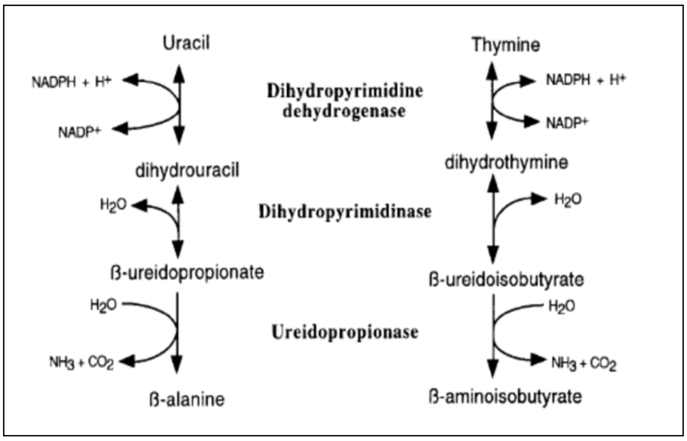


Figure 8.1 (Van Kuilenburg et al., 1999). Catabolic pathway of pyrimidines.

The clinical presentation of children affected by DPD deficiency can be highly heterogeneous, however neurological dysfunctions have been frequently observed and, interestingly, in some cases language impairment and speech delay have been reported (Henderson et al., 1995; van Gennip et al., 1981; Van Kuilenburg et al., 1999). Other disorders in which pyrimidine or purine homeostasis is altered, also present a spectrum of neurological abnormalities, indicating that the fine-regulation of nucleotide metabolism is important for CNS development (Micheli et al., 2011). Moreover, rare variants in *DPYD* have been found in a range of neurodevelopmental conditions, such as ASD, schizophrenia and ID (Carter et al., 2011; Willemsen et al., 2011; Xu et al., 2012). Therefore, several lines of evidence indicate that *DPYD* is an interesting functional candidate gene for neurological phenotypes. Thus, we decided to investigate whether *DPYD* could have a contributing role in SLI, both in this family and across the larger SLIC cohort.

In the discovery pedigree G4, we found that the affected child G4_5 was heterozygous compound for two deleterious mutations in *DPYD*: a paternally inherited splice site mutation in intron 14 (rs3918290, known also as *DPYD*2A*), and a maternally inherited missense change in exon 6 (rs72549308, p.S201R). G4_4 did not show any variant in *DPYD*, whereas the older brother G4_3 carried only the paternal splice site variant.

The splicing variant rs3918290 is one of the few mutations with a well-established role in DPD deficiency: the disruption of the donor splice site in intron 14 causes the skipping of exon 14 in the mRNA, that would result in a protein lacking the correspondent 55 amino acids with no residual activity.

The missense variant p.S201R is also predicted to have a deleterious effect on DPD activity and it is rare in the general population. Analysis of the crystal structure of the pig DPD protein indicates that the substitution of serine at position 201 with an arginine, which is an amino-acid with large side chain, could interfere with the electron flow, as this residue participates in a domain interface close to FAD and cluster Fe-S (van Kuilenburg et al., 2002). A compound heterozygosity for rs3918290 and S201R has been previously reported in a Scottish patient with motor impairment, intellectual disability and history of regression (van Kuilenburg et al., 2002). In this individual, no DPD activity was detected in peripheral blood mononuclear cells (PBM). Therefore, the compound heterozygosity rs3918290/S201R genetically confirmed the diagnosis of DPD deficiency in G4_5. In this family, the complete loss of DPD activity was found only in one out of two affected children, since we did not identify any variant in *DPYD* in the proband G4_4, who carries the homozygous copy loss in *ZNF277*. Therefore these findings suggest that, if the complete losses of ZNF277 and DPD activity represent risk factor for SLI, they contribute independently to language impairment in the affected members of family G4. SLI is a complex disorder in which multiple risk factors are likely to be implicated, therefore it remains possible that this heterogeneity could be observed also within families and that damaging mutations in *DPYD* might contribute to SLI in a complex manner.

Since rs3918290 is one of the most common mutations in DPD deficiency, we analysed the frequency of this splicing variant in a group of 166 SLIC cases. We identified the variant rs3918290 in 3 SLI cases in addition to G4_5, thus resulting in a minor allelic frequency (MAF) of 1.2% (A=4; G=330). This frequency was twice the reported frequency in controls with European ancestry (~0.6% in 1000 Genomes CEU samples and ESP EA controls). Although the frequency is double the expected, it does not reach statistical significance, perhaps due to the rarity of the splice site mutation.

To check whether the three affected children, carrying the splice variant rs3918290, also had a compound heterozygosity that could determine the complete loss of DPD activity, a mutation screening of the coding regions and the regulatory regions at 5' of *DPYD* was performed in the three SLI cases and their family members. Across the three families (**Figure 8.2**), we identified a novel variant in the region flanking the 5' of the gene (chr1:98386652, hg19), not reported in public databases, and five known coding variants (p.C29R (rs1801265), p.M166V (rs2297595), p.S534N (rs1801158), p.I543V (rs1801159) and p.V732I (rs1801160)), all common in the population (allelic frequency >1%).

The non-coding change maps at -72 nucleotides from the transcription start site and was found in a single family (E30) from the three studied. In this 5' flanking region, two GC boxes, essential for the promoter activity, have been found to be target sites for the ubiquitous transcription factors Sp1 and Sp3: SpB (from -68 to -60) and SpC (from -37 to -19) (Shestopal et al., 2000; Zhang et al., 2006b). This novel change might interfere with the transcription of the gene, therefore further investigations would be required to establish the consequences of this variant on *DPYD* expression.

The impact the five known common missense changes on DPD activity is controversially debated in the literature.

- The p.C29R substitution is predicted to have a benign effect on the DPD protein by bioinformatic tools. Although DPD-deficient patients heterozygous compound rs3918290/pC29R have been described (Vreken et al., 1997b) and the recombinant expressed DPD protein carrying the C29R substitution in *E.coli* was reported to have no residual activity (Vreken et al., 1997a), homozygosity for this mutation has been reported in individuals with normal DPD activity (Collie-Duguid et al., 2000; Seck et al., 2005). Moreover, in a recent study (Offer et al., 2013) *DPYD* variants were expressed in a mammalian cellular model, and the enzymatic activity of the expressed DPD protein was determined. In this study, the change p.C29R showed higher enzymatic activity compared with *wild-type* DPD, leading the researcher to hypothesize that it might be a "protective" allele.
- The p.S534N is predicted to have a damaging effect on the protein by bioinformatic tools, and the residue Ser534 is highly conserved throughout evolution. This variant was previously associated with toxicity reactions to 5-FU (Loganayagam et al., 2013) and decreased DPD activity in a control population (Seck et al., 2005); however in the study performed by Offer *et al*. (2013) the expression of the protein carrying this variant in a cellular model was associated with a significantly increased DPD activity (Offer et al., 2013) and this is supported by studies showing a lack of correlation between p.S534N and 5-FU sensitivity (Amstutz et al., 2009; Schwab et al., 2008). Offer *et al*. suggested that the S534N substitution may alter the protein structure of the conserved loop structure that covers the active site and that this might increase the substrate turnover.
- The change p.V732I is predicted deleterious for the protein, but it has been associated with normal DPD activity (Ezzeldin and Diasio, 2004; Offer et al., 2013; Seck et al., 2005) and it has been shown not to significantly alter the enzymatic activity in the human cellular model (Offer et al., 2013).
- p.M166V is also predicted damaging for the protein activity and has been found in association with 5-FU toxicity (Gross et al., 2008), however the majority of the studies agree on the fact that p.M166V is a common polymorphism with no effect on the protein activity (Ezzeldin and Diasio, 2004; Seck et al., 2005).
- The change p.I543V is also reported as a common change with neutral effect on the protein.

Four of the common changes (p.M166V, p.S534N, p.I543V, p.V732I) identified in these three SLI families were found also in the exome sequences of a group of 45 independent probands of the

SLIC cohort. Their allelic frequencies were not significantly different from those found in the general population. As for the previously investigated SLI families, 7 individuals carried multiple variants in *DPYD*, while 17 individuals were found to have only one of these missense mutations.

Since none of these variants was consistently reported to have a damaging effect on the protein activity, we do not predict that any of the individuals of the SLI families that we analysed have a complete loss of DPD activity. The only exception is for E30_4, (**Figure 8.2 c**), who is compound heterozygote for the splicing variant rs3918290, paternally inherited, and for a novel promoter variant, maternally inherited, which could possibly affect the gene expression level, thereby reducing or abolishing DPD residual activity.

For the other individuals, we can tentatively formulate a hypothesis for the combined effect of *DPYD* variants, given the well-established damaging effect of rs3918290, the possible higher activity for p.C29R and p.S534N (Offer et al., 2013), and a neutral effect for p.M166V, p.I543V and p.V732I (Ezzeldin and Diasio, 2004) (**Figure 8.2**):

- the three probands E6_3 (rs3918290/+), M18_3 (rs3918290/+), E30_3 (rs3918290/p.V732I) are likely to have a partial loss of DPD activity;
- compound heterozygosity of the two protective alleles p.C29R/p.S534N may determine an increased DPD activity in the affected sibling E6_4.
- The compound heterozygosity of a damaging and a protective allele, such as rs3918290/p.C29R (in E30_1, M18_2 and M18_4) or rs3918290/p.S534N (in E6_2), might lead to a balanced effect on DPD activity.

Interestingly, all probands in these 3 families would manifest an altered level of DPD activity, meaning that any level of dysregulation might be harmful.

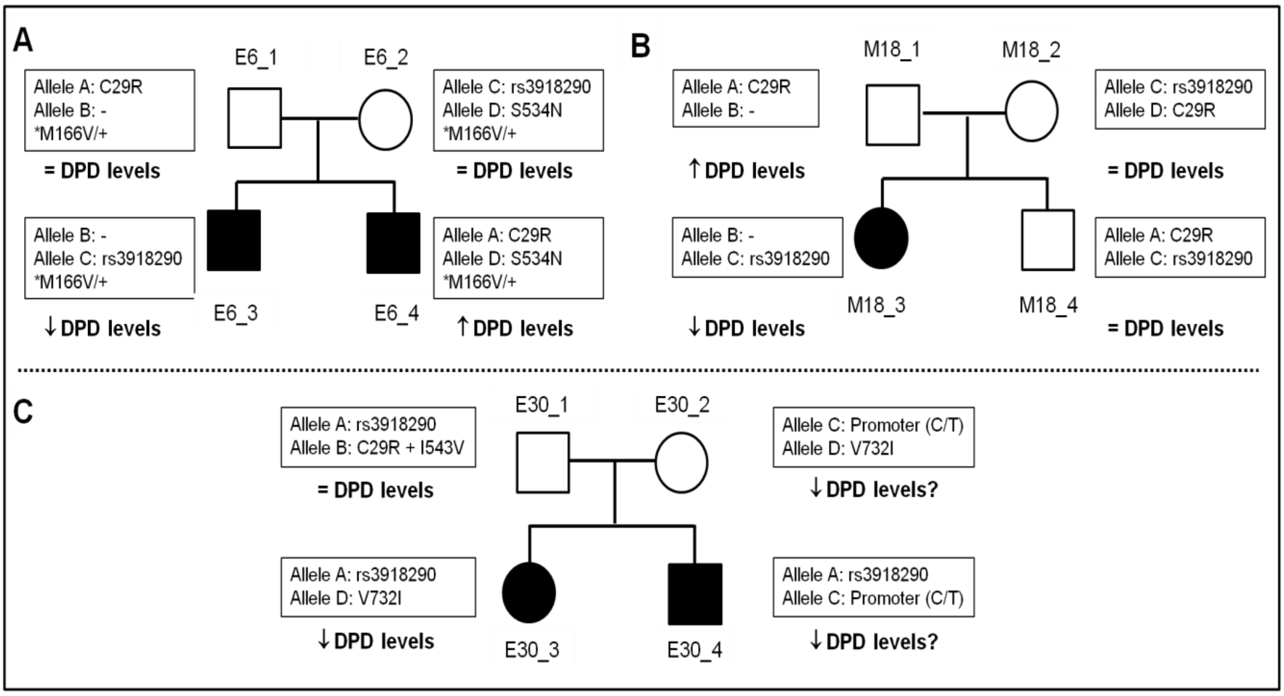


Figure 8.2. Hypotheses of DPD activity levels in three SLI families carrying the splicing mutation rs3918290, based on the mutation screening results. Alleles A and B indicate the paternal alleles for *DPYD*, alleles C and D the maternal alleles. *M166V was identified in all family members of E6, therefore it was not possible to determine the haplotype segregation for this change. Increased (↑), decreased (↓) and balanced effects (=) on DPD activity are indicated.

Taken together, the genetic profile of *DPYD* in the families of rs3918290 carriers revealed complex patterns of mutations and the simultaneous presence of multiple coding variants complicate the prediction of their effect on DPD activity, since the effect of damaging mutations might be compensated by potentially “protective” alleles. Therefore, the genotype-phenotype correlation is usually not straightforward and our hypotheses should be verified measuring the effective levels of DPD activity. However, it is possible that damaging mutations in *DPYD*, like rs3918290, or an accumulation of multiple mutations could lead to functional effects and that they could be contributing factors for neurological abnormalities in at least some language-impaired individuals. Moreover, variants in intronic and non-coding regions of the gene, not examined in this study, could also influence the transcription levels of this enzyme.

In conclusion, these results indicate that *DPYD* is an interesting candidate gene for SLI. In addition to a case of complete DPD deficiency, we observed an increased allelic frequency of the splicing variant rs3918290 in SLI probands compared to controls, although this difference was not significant. The rarity of the variant makes it difficult to provide statistical evidence for its involvement in SLI, as very large sample sizes would be needed even if there was a real effect, and it cannot be excluded that the reported frequency difference is a chance finding. To further investigate the influence of *DPYD* in language development, we are planning to extend the analysis of rs3918290 to a population-based sample, the “Avon Longitudinal Study of Parents and Children”

(ALSPAC) cohort (previously described in **paragraph 2.4.7**), with the aim to test whether this mutation in *DPYD* is associated with low language performance.

8.5 Analysis of the candidate gene *DPYD* in ASD.

Autism Spectrum Disorders (ASD) are a heterogeneous group of neurodevelopmental conditions characterized by impairment in social interactions and repetitive and restricted patterns of behaviours. Autism, which is the most severe form of ASD, presents deficits also in the language and communication domain. The clinical manifestations of autistic features are variable, ranging from mild to severe abnormalities. In the past thirty years, epidemiological studies have shown that ASD are highly heritable, highlighting the importance of genetic factors in ASD. However, the genetic causes underlying this group of disorders have been determined only for a minority of cases (Devlin and Scherer, 2012; Schaaf and Zoghbi, 2011). In about 10% of cases, classified as "syndromic", ASD features can be observed together with genetic syndromes or known Mendelian conditions, such as Fragile X syndrome, Rett syndrome, *PTEN* macrocephaly syndrome or Neurofibromatosis. Rare chromosomal abnormalities (e.g. 45 X0 Turner syndrome and trisomy 21) and CNVs at loci associated with known syndromes (e.g. the maternal duplication in the Prader-Willi/Angelman syndrome region at chromosome 15q11-q13) can account for 7-20% of ASD cases. Known metabolic disorders, such as phenylketonuria and creatine deficiency, have been estimated to contribute to ~5% of cases of ASD. The genetic causes of the rest of non-syndromic or idiopathic autism have yet to be identified.

Rare, but penetrant mutations have been identified in genes, such as *NLGN3, NLGN4, SHANK1, SHANK2, SHANK3, NRXN1, NRXN3, CNTNAP2* (Betancur, 2011; Devlin and Scherer, 2012), important for the formation, maturation and regulation of neuronal circuits (Peñagarikano and Geschwind, 2012; Südhof, 2008). Moreover, the NGS technologies, which allow the examination of the whole exomes without selecting candidate genes *a priori*, have identified rare sequence variants in genes, such as *AMT* and *PEX7,* involved in neurometabolic disorders (Yu et al., 2013), but not directly implicated in synaptic activity. Highly penetrant mutations in *AMT* (coding for an enzyme essential for glycine degradation) classically cause nonketotic hyperglycinemia and mutations in *PEX7* (coding for a receptor required for import of proteins into the peroxisomes) classically cause rhizomelic chondrodysplasia punctata: interestingly, Yu et al. (2013) demonstrated that hypomorphic mutations in these genes, in homozygous state or in compound heterozygosity, could result in atypical milder forms of these diseases, associated with a ASD phenotype. This study then suggested that unexpected pathways may be involved in ASD aetiology.

Since neurological abnormalities and autistic features have been observed in patients with DPD deficiency (Van Kuilenburg et al., 1999) and rare CNVs and sequence variants in *DPYD* have been reported for autism (Carter et al., 2011), *DPYD* represents an interesting non-synaptic candidate gene for ASD. Therefore, we decided to investigate its role also in this neurodevelopmental disorder. As for the SLIC cohort, our initial analysis evaluated the frequency of the splice variant rs3918290 in affected individuals. The analysis was carried out in two stages. In the first one, the SNP rs3918290 was genotyped in 231 Italian simplex families, 224 IMGSAC multiplex families and 449 Italian unrelated controls, by restriction endonuclease analysis. Subsequently, we extended the analysis to the entire cohort of ASD families of the AGP Consortium, which have been genotyped with the Illumina Infinium 1M-single and 1M-Duo SNP array (Anney et al., 2012), that include also the variant of interest rs3918290.

In the first stage, we identified the polymorphism rs3918290 in 2/224 IMGSAC multiplex families (203 and 171), 8/231 Italian simplex ASD families (5/231 probands) and 4/449 unrelated Italian controls. Segregation analyses showed that the variant did not completely segregate with the phenotype. The Minor Allele Frequency (MAF) for rs3918290 in the Italian trio sample was 1.08% compared to 0.45% in Italian controls (4/898), which is similar to the MAF reported for Tuscan controls in the database of 1000 Genomes Project (0.5%). Although the frequency in affected individuals is higher than in controls, this difference is not statistically significant (*p-value*= 0.1541).

In the IMGSAC sample, we obtained an estimated allelic frequency of 0.45% (2/448), selecting *a priori* one affected member in each family. In this case, the MAF for rs3918290 did not deviate from the frequencies reported in public databases for samples of European origin (0.4% - 0.6%). Therefore, the findings obtained from the multiplex families of the IMGSAC cohort did not replicate the trend observed for the Italian ASD simplex families.

To further investigate this variant in a larger number of affected individuals, we extended the analysis to the entire AGP cohort. The SNP rs3918290 was successfully genotyped in 2,681 independent cases and the allele A was detected in 25 of them, giving an allelic frequency of 0.47% (25/2681 probands), again not significantly different from the MAF reported in European populations (0.4%-0.6%).

Given the availability of genotype information for rs3918290 for the parents of the affected individuals, we also performed a *Transmission Disequilibrium Test* (TDT) on these AGP samples, to test whether the minor allele A was preferentially transmitted to individuals with ASD. However, we did not observe a preferential transmission, suggesting that the splicing variant is not associated with the ASD phenotype in the whole group of AGP families. To test whether other variants in

DPYD and the region encompassing its adjacent gene *MIR137*, another good candidate gene for neurodevelopmental disorders (Willemsen et al., 2011), were associated with ASD in the AGP group, we extended the TDT analysis to a set of 99 independent SNPs in the genomic region including *DPYD* and *MIR137* (chr1:97499599-98599144). After Bonferroni correction for multiple testing, no SNPs reached a *p-value* <0.05, indicating that common variants do not contribute to the ASD phenotype in these families.

Although rs3918290 showed no effect on the ASD categorical phenotype, it could have an effect on specific phenotypic features of ASD, such as cognitive functioning or language abilities. For this reason, we decided to investigate if the presence of the mutation rs3918290 could determine lower performances on Verbal IQ (VIQ), non-verbal IQ (PIQ) and Full scale-IQ tests (FIQ). We analysed the AGP IQ scores for the cases genotyped for the SNP rs3918290, comparing the individuals carrying the variants with those without it. We found that the distribution of the VIQ and FIQ scores did not significantly differ between the two groups and PIQ scores appeared instead slightly higher in individuals with the splicing variant (*P*-value= 0.03). Therefore, we did not observe an effect of the variant rs3918290 on cognitive abilities in the ASD phenotype.

Qualitative analyses were also performed for language endophenotypes. Information regarding language development in the autistic individuals of the AGP cohort were provided by three Autism Diagnostic Interview-revised (ADI-R) (Lord et al., 1994) items: the overall level of current language, the age of first single words and the age of first phrases. For each test, we subdivided the individuals into two categories: subjects with “normal” development and subjects with “delayed” language development, according to commonly used thresholds. We did not find a significant difference for any of these endophenotypes. It should be noted that given the rarity of the splicing variant, these endophenotype analysis were based on a very small number of cases carrying the variant, compared to the cohort size.

In order to check whether the ASD probands carried other variants in the coding or 5’ regulatory regions of *DPYD*, that could potentially cause a complete loss of DPD activity in conjunction with rs3928190, we performed a mutation screening of the gene in the IMGSAC and Italian probands carrying the splicing variant, as we did for the SLI families. This screening identified 2 common missense variants, identified also in the SLI samples and described above: one in exon 2 (rs1801265, p.C29R) and another one in exon 13 (rs1801158, p.S534N).

As previously discussed, both p.C29R and p.S534N have been suggested to be protective alleles that determine an above-average activity of the protein DPD (Offer et al., 2013). However, this effect could be “neutralized” in the presence of the disrupting mutation rs3928190 on the same allele. Based on this assumption, a partial loss of DPD activity could be hypothesized at least for the probands 203.3 (with p.S534N and rs3918290 on the paternal allele) (**Figure 8.3 d**) and SM2.3 (with p.C29R and rs3918290 on the paternal allele) (**Figure 8.3 e**), and for C19.3, SM38.3 and SM161.3 (all rs3918290/+) (**Figure 8.3 a, b, c**). In family 171 (**Figure 8.3 f**), the compound heterozygosity p.C29R/rs3918290 in the proband 171.6, the unaffected sibling and the mother, may instead determine a balance between the effects of the two variants. For family C22 (**Figure 8.3 g**), the outcome is more uncertain, because the haplotype pattern of the two mutations p.C29R and rs3918290 in the proband is unknown.

Therefore, we did not detect any complete loss of DPD activity in any of the ASD probands carrying the splicing mutation, but in at least 5 of them we can hypothesize a partial DPD deficiency. Again, as we observed also for SLI families, the interpretation of the genotypic profile of *DPYD* variants in terms of DPD activity is often complex and should be confirmed by an enzymatic assay.

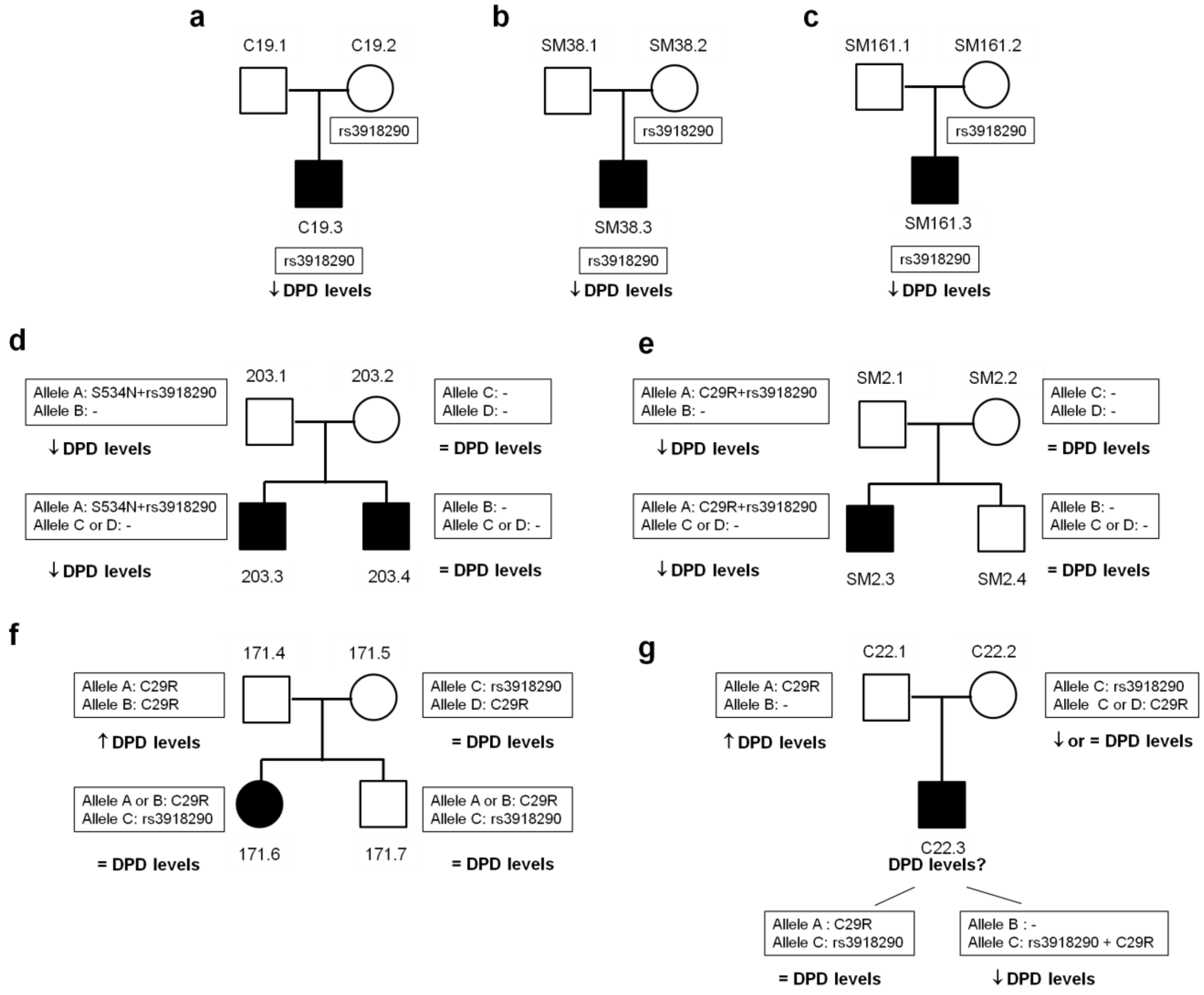


Figure 8.3. Hypotheses of DPD activity levels in ASD families carrying the splicing mutation rs3918290, based on the mutation screening results. Increased (↑), decreased (↓) and balanced effects (=) on DPD activity are indicated.

From the result of the mutation screening then, it is not excluded that the splice site variant could have a role, at least in a subgroup of individuals with ASD, with incomplete penetrance and variable expressivity. In addition, its effect may depend also on the interactions with other mutations in the same gene or in other genomic locations. We cannot exclude the hypothesis that the ASD probands carrying the splicing site mutation in *DPYD* could have other variants (CNVs or sequence mutations), that could contribute to the ASD susceptibility, together with *DPYD*. According to a "multiple hit hypothesis", some rare variants could manifest a pathogenic role only in the simultaneous presence of other variants, rare or common, located in different genomic loci.

An exome study for schizophrenia (Xu et al., 2012) identified two rare *de novo* mutations in *DPYD* in two independent cases, suggesting that rare and penetrant point mutations in this gene could contribute to this disorder. To date, exome studies of ASD have not identified rare *de novo* mutations in *DPYD* (Neale et al., 2012; O'Roak et al., 2011; Sanders et al., 2012). However, it is not excluded that future studies, not focused only on *de novo* events and including larger number of individuals, will identify new variants in *DPYD* and/or that they will clarify its contribution to ASD aetiology.

In conclusion, although rare mutations in *DPYD* have implicated this interesting candidate gene in several neurodevelopmental conditions, including ASD, in this study we did not find striking evidence for a possible involvement of *DPYD* in ASD. Our analysis was mainly focused on the variant which is the most commonly reported in individuals with DPD deficiency, but further investigations would be required, for example, to characterize the genetic background of rs3918290 carriers. Therefore, it remains possible that different mutations in *DPYD*, with variable penetrance, could act as risk factors for ASD susceptibility and that their contribution may depend also on the genetic background of the individual or other non-genetic factors.

References.

Abecasis, G. R., A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin, R. E. Handsaker, H. M. Kang, G. T. Marth, G. A. McVean, and G. P. Consortium, 2012, An integrated map of genetic variation from 1,092 human genomes: Nature, v. 491, p. 56-65.

Abrahams, B. S., D. Tentler, J. V. Perederiy, M. C. Oldham, G. Coppola, and D. H. Geschwind, 2007, Genome-wide analyses of human perisylvian cerebral cortical patterning: Proc Natl Acad Sci U S A, v. 104, p. 17849-54.

Abreu, P. C., D. A. Greenberg, and S. E. Hodge, 1999, Direct power comparisons between simple LOD scores and NPL scores for linkage analysis in complex diseases: Am J Hum Genet, v. 65, p. 847-57.

Adzhubei, I. A., S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev, 2010, A method and server for predicting damaging missense mutations: Nat Methods, v. 7, p. 248-9.

Alarcón, M., B. S. Abrahams, J. L. Stone, J. A. Duvall, J. V. Perederiy, J. M. Bomar, J. Sebat, M. Wigler, C. L. Martin, D. H. Ledbetter, S. F. Nelson, R. M. Cantor, and D. H. Geschwind, 2008, Linkage, association, and gene-expression analyses identify CNTNAP2 as an autism-susceptibility gene: Am J Hum Genet, v. 82, p. 150-9.

Alarcón, M., R. M. Cantor, J. Liu, T. C. Gilliam, D. H. Geschwind, and A. G. R. E. Consortium, 2002, Evidence for a language quantitative trait locus on chromosome 7q in multiplex autism families: Am J Hum Genet, v. 70, p. 60-71.

Amstutz, U., S. Farese, S. Aebi, and C. R. Largiadèr, 2008, Hypermethylation of the DPYD promoter region is not a major predictor of severe toxicity in 5-fluorouracil based chemotherapy: J Exp Clin Cancer Res, v. 27, p. 54.

Amstutz, U., S. Farese, S. Aebi, and C. R. Largiadèr, 2009, Dihydropyrimidine dehydrogenase gene variation and severe 5-fluorouracil toxicity: a haplotype assessment: Pharmacogenomics, v. 10, p. 931-44.

Amstutz, U., T. K. Froehlich, and C. R. Largiadèr, 2011, Dihydropyrimidine dehydrogenase gene as a major predictor of severe 5-fluorouracil toxicity: Pharmacogenomics, v. 12, p. 1321-36.

Anney, R., L. Klei, D. Pinto, J. Almeida, E. Bacchelli, G. Baird, N. Bolshakova, S. Bölte, P. F. Bolton, T. Bourgeron, S. Brennan, J. Brian, J. Casey, J. Conroy, C. Correia, C. Corsello, E. L. Crawford, M. de Jonge, R. Delorme, E. Duketis, F. Duque, A. Estes, P. Farrar, B. A. Fernandez, S. E. Folstein, E. Fombonne, J. Gilbert, C. Gillberg, J. T. Glessner, A. Green, J. Green, S. J. Guter, E. A. Heron, R. Holt, J. L. Howe, G. Hughes, V. Hus, R. Igliozzi, S. Jacob, G. P. Kenny, C. Kim, A. Kolevzon, V. Kustanovich, C. M. Lajonchere, J. A. Lamb, M. Law-Smith, M. Leboyer, A. Le Couteur, B. L. Leventhal, X. Q. Liu, F. Lombard, C. Lord, L. Lotspeich, S. C. Lund, T. R. Magalhaes, C. Mantoulan, C. J. McDougle, N. M. Melhem, A. Merikangas, N. J. Minshew, G. K. Mirza, J. Munson, C. Noakes, G. Nygren, K. Papanikolaou, A. T. Pagnamenta, B. Parrini, T. Paton, A. Pickles, D. J. Posey, F. Poustka, J. Ragoussis, R. Regan, W. Roberts, K. Roeder, B. Roge, M. L. Rutter, S. Schlitt, N. Shah, V. C. Sheffield, L. Soorya, I. Sousa, V. Stoppioni, N. Sykes, R. Tancredi, A. P. Thompson, S. Thomson, A. Tryfon, J. Tsiantis, H. Van Engeland, J. B. Vincent, F. Volkmar, J. A. Vorstman, S. Wallace, K. Wing, K. Wittemeyer, S. Wood, D. Zurawiecki, L. Zwaigenbaum, A. J. Bailey, et al., 2012, Individual common variants exert weak effects on the risk for autism spectrum disorders: Hum Mol Genet, v. 21, p. 4781-92.

Anney, R., L. Klei, D. Pinto, R. Regan, J. Conroy, T. R. Magalhaes, C. Correia, B. S. Abrahams, N. Sykes, A. T. Pagnamenta, J. Almeida, E. Bacchelli, A. J. Bailey, G. Baird, A. Battaglia, T. Berney, N. Bolshakova, S. Bolte, P. F. Bolton, T. Bourgeron, S. Brennan, J. Brian, A. R. Carson, G. Casallo, J. Casey, S. H. Chu, L. Cochrane, C. Corsello, E. L. Crawford, A. Crossett, G. Dawson, M. de Jonge, R. Delorme, I. Drmic, E. Duketis, F. Duque, A. Estes, P. Farrar, B. A. Fernandez, S. E. Folstein, E. Fombonne, C. M. Freitag, J. Gilbert, C. Gillberg, J. T. Glessner, J. Goldberg, J. Green, S. J. Guter, H. Hakonarson, E. A. Heron, M. Hill, R. Holt, J. L. Howe, G. Hughes, V. Hus, R. Igliozzi, C. Kim, S. M. Klauck, A. Kolevzon, O. Korvatska, V. Kustanovich, C. M. Lajonchere, J. A. Lamb, M. Laskawiec, M. Leboyer, A. Le Couteur, B. L. Leventhal, A. C. Lionel, X. Q. Liu, C. Lord, L. Lotspeich, S. C. Lund, E. Maestrini, W. Mahoney, C. Mantoulan, C. R. Marshall, H. McConachie, C. J. McDougle, J. McGrath, W. M. McMahon, N. M. Melhem, A. Merikangas, O. Migita, N. J. Minshew, G. K. Mirza, J. Munson, S. F. Nelson, C. Noakes, A. Noor, G. Nygren, G. Oliveira, K. Papanikolaou, J. R. Parr, B. Parrini, T. Paton, A. Pickles, J. Piven, D. J. Osey, A. Poustka, F.

Poustka, et al., 2010, A genome-wide scan for common alleles affecting risk for autism: Human Molecular Genetics, v. 19, p. 4072-4082.

APA, American Psychiatric Association, 1994, Diagnostic and statistical manual of mental disorders: DSM-IV.

APA, American Psychiatric Association, 2013, Diagnostic and statistical manual of mental disorders (5th ed.), Arlington, VA: American Psychiatric Publishing.

Ayalew, M., H. Le-Niculescu, D. F. Levey, N. Jain, B. Changala, S. D. Patel, E. Winiger, A. Breier, A. Shekhar, R. Amdur, D. Koller, J. I. Nurnberger, A. Corvin, M. Geyer, M. T. Tsuang, D. Salomon, N. J. Schork, A. H. Fanous, M. C. O'Donovan, and A. B. Niculescu, 2012, Convergent functional genomics of schizophrenia: from comprehensive understanding to genetic risk prediction: Mol Psychiatry, v. 17, p. 887-905.

Bacon, C., and G. A. Rappold, 2012, The distinct and overlapping phenotypic spectra of FOXP1 and FOXP2 in cognitive disorders: Hum Genet, v. 131, p. 1687-98.

Badner, J. A., and E. S. Gershon, 2002, Regional meta-analysis of published data supports linkage of autism with markers on chromosome 7: Mol Psychiatry, v. 7, p. 56-66.

Bailey, J. A., and E. E. Eichler, 2006, Primate segmental duplications: crucibles of evolution, diversity and disease: Nat Rev Genet, v. 7, p. 552-64.

Bakkaloglu, B., B. J. O'Roak, A. Louvi, A. R. Gupta, J. F. Abelson, T. M. Morgan, K. Chawarska, A. Klin, A. G. Ercan-Sencicek, A. A. Stillman, G. Tanriover, B. S. Abrahams, J. A. Duvall, E. M. Robbins, D. H. Geschwind, T. Biederer, M. Gunel, R. P. Lifton, and M. W. State, 2008, Molecular cytogenetic analysis and resequencing of contactin associated protein-like 2 in autism spectrum disorders: Am J Hum Genet, v. 82, p. 165-73.

Bartlett, C. W., J. F. Flax, Z. Fermano, A. Hare, L. Hou, S. A. Petrill, S. Buyske, and L. M. Brzustowicz, 2012, Gene × gene interaction in shared etiology of autism and specific language impairment: Biol Psychiatry, v. 72, p. 692-9.

Bartlett, C. W., J. F. Flax, M. W. Logue, B. J. Smith, V. J. Vieland, P. Tallal, and L. M. Brzustowicz, 2004, Examination of potential overlap in autism and language loci on chromosomes 2, 7, and 13 in two independent samples ascertained for specific language impairment: Hum Hered, v. 57, p. 10-20.

Bartlett, C. W., J. F. Flax, M. W. Logue, V. J. Vieland, A. S. Bassett, P. Tallal, and L. M. Brzustowicz, 2002, A major susceptibility locus for specific language impairment is located on 13q21: Am J Hum Genet, v. 71, p. 45-55.

Bartlett, C. W., L. Hou, J. F. Flax, A. Hare, S. Y. Cheong, Z. Fermano, B. Zimmerman-Bier, C. Cartwright, M. A. Azaro, S. Buyske, and L. M. Brzustowicz, 2013, A Genome Scan for Loci Shared by Autism Spectrum Disorder and Language Impairment: Am J Psychiatry.

Battaglia, A., V. Doccini, L. Bernardini, A. Novelli, S. Loddo, A. Capalbo, T. Filippi, and J. C. Carey, 2013, Confirmation of chromosomal microarray as a first-tier clinical diagnostic test for individuals with developmental delay, intellectual disability, autism spectrum disorders and dysmorphic features: Eur J Paediatr Neurol, v. 17, p. 589-99.

Ben-Shachar, S., B. Lanpher, J. R. German, M. Qasaymeh, L. Potocki, S. C. Nagamani, L. M. Franco, A. Malphrus, G. W. Bottenfield, J. E. Spence, S. Amato, J. A. Rousseau, B. Moghaddam, C. Skinner, S. A. Skinner, S. Bernes, N. Armstrong, M. Shinawi, P. Stankiewicz, A. Patel, S. W. Cheung, J. R. Lupski, A. L. Beaudet, and T. Sahoo, 2009, Microdeletion 15q13.3: a locus with incomplete penetrance for autism, mental retardation, and psychiatric disorders: J Med Genet, v. 46, p. 382-8.

Bergen, S. E., C. T. O'Dushlaine, S. Ripke, P. H. Lee, D. M. Ruderfer, S. Akterin, J. L. Moran, K. D. Chambert, R. E. Handsaker, L. Backlund, U. Ösby, S. McCarroll, M. Landen, E. M. Scolnick, P. K. Magnusson, P. Lichtenstein, C. M. Hultman, S. M. Purcell, P. Sklar, and P. F. Sullivan, 2012, Genome-wide association study in a Swedish population yields support for greater CNV and MHC involvement in schizophrenia compared with bipolar disorder: Mol Psychiatry, v. 17, p. 880-6.

Berger, R., S. A. Stoker-de Vries, S. K. Wadman, M. Duran, F. A. Beemer, P. K. de Bree, J. J. Weits-Binnerts, T. J. Penders, and J. K. van der Woude, 1984, Dihydropyrimidine dehydrogenase deficiency leading to thymine-uraciluria. An inborn error of pyrimidine metabolism: Clin Chim Acta, v. 141, p. 227-34.

Berkel, S., C. R. Marshall, B. Weiss, J. Howe, R. Roeth, U. Moog, V. Endris, W. Roberts, P. Szatmari, D. Pinto, M. Bonin, A. Riess, H. Engels, R. Sprengel, S. W. Scherer, and G. A. Rappold, 2010, Mutations in the SHANK2 synaptic scaffolding gene in autism spectrum disorder and mental retardation: Nat Genet, v. 42, p. 489-91.

Betancur, C., 2011, Etiological heterogeneity in autism spectrum disorders: more than 100 genetic and genomic disorders and still counting: Brain Res, v. 1380, p. 42-77.

Bhat, S., D. T. Dao, C. E. Terrillion, M. Arad, R. J. Smith, N. M. Soldatov, and T. D. Gould, 2012, CACNA1C (Cav1.2) in the pathophysiology of psychiatric disease: Prog Neurobiol, v. 99, p. 1-14.

Bishop, D. V., 2002, The role of genes in the etiology of specific language impairment: J Commun Disord, v. 35, p. 311-28.

Bishop, D. V., 2005, DeFries-Fulker analysis of twin data with skewed distributions: cautions and recommendations from a study of children's use of verb inflections: Behav Genet, v. 35, p. 479-90.

Bishop, D. V., 2006, What Causes Specific Language Impairment in Children?: Curr Dir Psychol Sci, v. 15, p. 217-221.

Bishop, D. V., 2014, Problems with tense marking in children with specific language impairment: not how but when: Philos Trans R Soc Lond B Biol Sci, v. 369, p. 20120401.

Bishop, D. V., T. North, and C. Donlan, 1995, Genetic basis of specific language impairment: evidence from a twin study: Dev Med Child Neurol, v. 37, p. 56-71.

Bishop, D. V., T. North, and C. Donlan, 1996, Nonword repetition as a behavioural marker for inherited language impairment: evidence from a twin study: J Child Psychol Psychiatry, v. 37, p. 391-403.

Bradford, Y., J. Haines, H. Hutcheson, M. Gardiner, T. Braun, V. Sheffield, T. Cassavant, W. Huang, K. Wang, V. Vieland, S. Folstein, S. Santangelo, and J. Piven, 2001, Incorporating language phenotypes strengthens evidence of linkage to autism: Am J Med Genet, v. 105, p. 539-47.

Brockstedt, M., C. Jakobs, L. M. Smit, A. H. van Gennip, and R. Berger, 1990, A new case of dihydropyrimidine dehydrogenase deficiency: J Inherit Metab Dis, v. 13, p. 121-4.

Bruder, C. E., A. Piotrowski, A. A. Gijsbers, R. Andersson, S. Erickson, T. Diaz de Ståhl, U. Menzel, J. Sandgren, D. von Tell, A. Poplawski, M. Crowley, C. Crasto, E. C. Partridge, H. Tiwari, D. B. Allison, J. Komorowski, G. J. van Ommen, D. I. Boomsma, N. L. Pedersen, J. T. den Dunnen, K. Wirdefeldt, and J. P. Dumanski, 2008, Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles: Am J Hum Genet, v. 82, p. 763-71.

Brzustowicz, L. M., K. A. Hodgkinson, E. W. Chow, W. G. Honer, and A. S. Bassett, 2000, Location of a major susceptibility locus for familial schizophrenia on chromosome 1q21-q22: Science, v. 288, p. 678-82.

Bucan, M., B. S. Abrahams, K. Wang, J. T. Glessner, E. I. Herman, L. I. Sonnenblick, A. I. Alvarez Retuerto, M. Imielinski, D. Hadley, J. P. Bradfield, C. Kim, N. B. Gidaya, I. Lindquist, T. Hutman, M. Sigman, V. Kustanovich, C. M. Lajonchere, A. Singleton, J. Kim, T. H. Wassink, W. M. McMahon, T. Owley, J. A. Sweeney, H. Coon, J. I. Nurnberger, M. Li, R. M. Cantor, N. J. Minshew, J. S. Sutcliffe, E. H. Cook, G. Dawson, J. D. Buxbaum, S. F. Grant, G. D. Schellenberg, D. H. Geschwind, and H. Hakonarson, 2009, Genome-wide analyses of exonic copy number variants in a family-based study point to novel autism susceptibility genes: PLoS Genet, v. 5, p. e1000536.

Burden, V., C. M. Stott, J. Forge, and I. Goodyer, 1996, The Cambridge Language and Speech Project (CLASP). I . Detection of language difficulties at 36 to 39 months: Dev Med Child Neurol, v. 38, p. 613-31.

Burnside, R. D., R. Pasion, F. M. Mikhail, A. J. Carroll, N. H. Robin, E. L. Youngs, I. K. Gadi, E. Keitges, V. L. Jaswaney, P. R. Papenhausen, V. R. Potluri, H. Risheg, B. Rush, J. L. Smith, S. Schwartz, J. H. Tepperberg, and M. G. Butler, 2011, Microdeletion/microduplication of proximal 15q11.2 between BP1 and BP2: a susceptibility region for neurological dysfunction including developmental and language delay: Hum Genet, v. 130, p. 517-28.

Buxbaum, J. D., J. M. Silverman, C. J. Smith, M. Kilifarski, J. Reichert, E. Hollander, B. A. Lawlor, M. Fitzgerald, D. A. Greenberg, and K. L. Davis, 2001, Evidence for a susceptibility gene for autism on chromosome 2 and for genetic heterogeneity: Am J Hum Genet, v. 68, p. 1514-20.

Carter, M. T., S. M. Nikkel, B. A. Fernandez, C. R. Marshall, A. Noor, A. C. Lionel, A. Prasad, D. Pinto, A. M. Joseph-George, C. Noakes, C. Fairbrother-Davies, W. Roberts, J. Vincent, R. Weksberg, and S. W. Scherer, 2011, Hemizygous deletions on chromosome 1p21.3 involving the DPYD gene in individuals with autism spectrum disorder: Clin Genet, v. 80, p. 435-43.

Chaisson, M. J., and P. A. Pevzner, 2008, Short read fragment assembly of bacterial genomes: Genome Res, v. 18, p. 324-30.

Charbonnier, F., G. Raux, Q. Wang, N. Drouot, F. Cordier, J. M. Limacher, J. C. Saurin, A. Puisieux, S. Olschwang, and T. Frebourg, 2000, Detection of exon deletions and duplications of the mismatch repair genes in hereditary nonpolyposis colorectal cancer families using multiplex polymerase chain reaction of short fluorescent fragments: Cancer Res, v. 60, p. 2760-3.

Ciccolini, J., C. Mercier, A. Evrard, L. Dahan, J. C. Boyer, F. Duffaud, K. Richard, C. Blanquicett, G. Milano, A. Blesius, A. Durand, J. F. Seitz, R. Favre, and B. Lacarelle, 2006, A rapid and inexpensive method for anticipating severe toxicity to fluorouracil and fluorouracil-based chemotherapy: Ther Drug Monit, v. 28, p. 678-85.

Clark, A., A. O'Hare, J. Watson, W. Cohen, H. Cowie, R. Elton, J. Nasir, and J. Seckl, 2007, Severe receptive language disorder in childhood--familial aspects and long-term outcomes: results from a Scottish study: Arch Dis Child, v. 92, p. 614-9.

Coe, B. P., S. Girirajan, and E. E. Eichler, 2012, The genetic variability and commonality of neurodevelopmental disease: Am J Med Genet C Semin Med Genet, v. 160C, p. 118-29.

Cohen, N. J., D. D. Vallance, M. Barwick, N. Im, R. Menna, N. B. Horodezky, and L. Isaacson, 2000, The interface between ADHD and language impairment: an examination of language, achievement, and cognitive processing: J Child Psychol Psychiatry, v. 41, p. 353-62.

Cole, R. L., S. M. Lechner, M. E. Williams, P. Prodanovich, L. Bleicher, M. A. Varney, and G. Gu, 2005, Differential distribution of voltage-gated calcium channel alpha-2 delta (alpha2delta) subunit mRNA-containing cells in the rat central nervous system and the dorsal root ganglia: J Comp Neurol, v. 491, p. 246-69.

Colella, S., C. Yau, J. M. Taylor, G. Mirza, H. Butler, P. Clouston, A. S. Bassett, A. Seller, C. C. Holmes, and J. Ragoussis, 2007, QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data: Nucleic Acids Res, v. 35, p. 2013-25.

Collie-Duguid, E. S., M. C. Etienne, G. Milano, and H. L. McLeod, 2000, Known variant DPYD alleles do not explain DPD deficiency in cancer patients: Pharmacogenetics, v. 10, p. 217-23.

Conrad, D. F., D. Pinto, R. Redon, L. Feuk, O. Gokcumen, Y. Zhang, J. Aerts, T. D. Andrews, C. Barnes, P. Campbell, T. Fitzgerald, M. Hu, C. H. Ihm, K. Kristiansson, D. G. Macarthur, J. R. Macdonald, I. Onyiah, A. W. Pang, S. Robson, K. Stirrups, A. Valsesia, K. Walter, J. Wei, C. Tyler-Smith, N. P. Carter, C. Lee, S. W. Scherer, M. E. Hurles, and W. T. C. C. Consortium, 2010, Origins and functional impact of copy number variation in the human genome: Nature, v. 464, p. 704-12.

Constantino, J. N., Y. Zhang, T. Frazier, A. M. Abbacchi, and P. Law, 2010, Sibling recurrence and the genetic epidemiology of autism: Am J Psychiatry, v. 167, p. 1349-56.

Conti-Ramsden, G., and N. Botting, 1999, Characteristics of children attending language units in England: a national study of 7-year-olds: Int J Lang Commun Disord, v. 34, p. 359-66.

Conti-Ramsden, G., N. Botting, Z. Simkin, and E. Knox, 2001, Follow-up of children attending infant language units: outcomes at 11 years of age: Int J Lang Commun Disord, v. 36, p. 207-19.

Conti-Ramsden, G., Z. Simkin, and N. Botting, 2006, The prevalence of autistic spectrum disorders in adolescents with a history of specific language impairment (SLI): J Child Psychol Psychiatry, v. 47, p. 621-8.

Cook, E. H., and S. W. Scherer, 2008, Copy-number variations associated with neuropsychiatric conditions: Nature, v. 455, p. 919-23.

Cooper, G. M., B. P. Coe, S. Girirajan, J. A. Rosenfeld, T. H. Vu, C. Baker, C. Williams, H. Stalker, R. Hamid, V. Hannig, H. Abdel-Hamid, P. Bader, E. McCracken, D. Niyazov, K. Leppig, H. Thiese, M. Hummel, N. Alexander, J. Gorski, J. Kussmann, V. Shashi, K. Johnson, C. Rehder, B. C. Ballif, L. G. Shaffer, and E. E. Eichler, 2011, A copy number variation morbidity map of developmental delay: Nat Genet, v. 43, p. 838-46.

Cooper, G. M., E. A. Stone, G. Asimenos, E. D. Green, S. Batzoglou, A. Sidow, and N. C. S. Program, 2005, Distribution and intensity of constraint in mammalian genomic sequence: Genome Res, v. 15, p. 901-13.

Craig, A. M., and Y. Kang, 2007, Neurexin-neuroligin signaling in synapse development: Curr Opin Neurobiol, v. 17, p. 43-52.

Darai-Ramqvist, E., A. Sandlund, S. Müller, G. Klein, S. Imreh, and M. Kost-Alimova, 2008, Segmental duplications and evolutionary plasticity at tumor chromosome break-prone regions: Genome Res, v. 18, p. 370-9.

Dash, P. K., A. N. Moore, N. Kobori, and J. D. Runyan, 2007, Molecular activity underlying working memory: Learn Mem, v. 14, p. 554-63.

de Kovel, C. G., H. Trucks, I. Helbig, H. C. Mefford, C. Baker, C. Leu, C. Kluck, H. Muhle, S. von Spiczak, P. Ostertag, T. Obermeier, A. A. Kleefuss-Lie, K. Hallmann, M. Steffens, V. Gaus, K. M. Klein, H. M. Hamer, F. Rosenow, E. H. Brilstra, D. K. Trenité, M. E. Swinkels, Y. G. Weber, I. Unterberger, F. Zimprich, L. Urak, M. Feucht, K. Fuchs, R. S. Møller, H. Hjalgrim, P. De Jonghe, A. Suls, I. M. Rückert, H. E. Wichmann, A. Franke, S. Schreiber, P. Nürnberg, C. E. Elger, H. Lerche, U. Stephani, B. P. Koeleman, D. Lindhout, E. E. Eichler, and T. Sander, 2010, Recurrent microdeletions at 15q11.2 and 16p13.11 predispose to idiopathic generalized epilepsies: Brain, v. 133, p. 23-32.

DeThorne, L. S., S. A. Hart, S. A. Petrill, K. Deater-Deckard, L. A. Thompson, C. Schatschneider, and M. D. Davison, 2006, Children's history of speech-language difficulties: genetic influences and associations with reading-related measures: J Speech Lang Hear Res, v. 49, p. 1280-93.

Devlin, B., and S. W. Scherer, 2012, Genetic architecture in autism spectrum disorder: Curr Opin Genet Dev, v. 22, p. 229-37.

Diskin, S. J., M. Li, C. Hou, S. Yang, J. Glessner, H. Hakonarson, M. Bucan, J. M. Maris, and K. Wang, 2008, Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms: Nucleic Acids Res, v. 36, p. e126.

Dobritzsch, D., S. Ricagno, G. Schneider, K. D. Schnackerz, and Y. Lindqvist, 2002, Crystal structure of the productive ternary complex of dihydropyrimidine dehydrogenase with NADPH and 5-iodouracil. Implications for mechanism of inhibition and electron transfer: J Biol Chem, v. 277, p. 13155-66.

Dobritzsch, D., G. Schneider, K. D. Schnackerz, and Y. Lindqvist, 2001, Crystal structure of dihydropyrimidine dehydrogenase, a major determinant of the pharmacokinetics of the anti-cancer drug 5-fluorouracil: EMBO J, v. 20, p. 650-60.

Dolphin, A. C., 2013, The α2δ subunits of voltage-gated calcium channels: Biochim Biophys Acta, v. 1828, p. 1541-9.

Doornbos, M., B. Sikkema-Raddatz, C. A. Ruijvenkamp, T. Dijkhuizen, E. K. Bijlsma, A. C. Gijsbers, Y. Hilhorst-Hofstee, R. Hordijk, K. T. Verbruggen, W. S. Kerstjens-Frederikse, T. van Essen, K. Kok, A. T. van Silfhout, M. Breuning, and C. M. van Ravenswaaij-Arts, 2009, Nine patients with a microdeletion 15q11.2 between breakpoints 1 and 2 of the Prader-Willi critical region, possibly associated with behavioural disturbances: Eur J Med Genet, v. 52, p. 108-15.

Dumas, L., Y. H. Kim, A. Karimpour-Fard, M. Cox, J. Hopkins, J. R. Pollack, and J. M. Sikela, 2007, Gene copy number variation spanning 60 million years of human and primate evolution: Genome Res, v. 17, p. 1266-77.

Eicher, J. D., N. R. Powers, L. L. Miller, N. Akshoomoff, D. G. Amaral, C. S. Bloss, O. Libiger, N. J. Schork, B. F. Darst, B. J. Casey, L. Chang, T. Ernst, J. Frazier, W. E. Kaufmann, B. Keating, T. Kenet, D. Kennedy, S. Mostofsky, S. S. Murray, E. R. Sowell, H. Bartsch, J. M. Kuperman, T. T. Brown, D. J. Hagler, A. M. Dale, T. L. Jernigan, B. St Pourcain, G. Davey Smith, S. M. Ring, J. R. Gruen, and N. u. Pediatric Imaging, and Genetics Study, 2013, Genome-wide association study of shared components of reading disability and language impairment: Genes Brain Behav, v. 12, p. 792-801.

Elia, J., X. Gai, H. M. Xie, J. C. Perin, E. Geiger, J. T. Glessner, M. D'arcy, R. deBerardinis, E. Frackelton, C. Kim, F. Lantieri, B. M. Muganga, L. Wang, T. Takeda, E. F. Rappaport, S. F. Grant, W. Berrettini, M. Devoto, T. H. Shaikh, H. Hakonarson, and P. S. White, 2010, Rare structural variants found in attention-deficit hyperactivity disorder are preferentially associated with neurodevelopmental genes: Mol Psychiatry, v. 15, p. 637-46.

Elsabbagh, M., G. Divan, Y. J. Koh, Y. S. Kim, S. Kauchali, C. Marcín, C. Montiel-Nava, V. Patel, C. S. Paula, C. Wang, M. T. Yasamy, and E. Fombonne, 2012, Global prevalence of autism and other pervasive developmental disorders: Autism Res, v. 5, p. 160-79.

Ezzeldin, H., and R. Diasio, 2004, Dihydropyrimidine dehydrogenase deficiency, a pharmacogenetic syndrome associated with potentially life-threatening toxicity following 5-fluorouracil administration: Clin Colorectal Cancer, v. 4, p. 181-9.

Ezzeldin, H., M. R. Johnson, Y. Okamoto, and R. Diasio, 2003, Denaturing high performance liquid chromatography analysis of the DPYD gene in patients with lethal 5-fluorouracil toxicity: Clin Cancer Res, v. 9, p. 3021-8.

Ezzeldin, H. H., A. M. Lee, L. K. Mattison, and R. B. Diasio, 2005, Methylation of the DPYD promoter: an alternative mechanism for dihydropyrimidine dehydrogenase deficiency in cancer patients: Clin Cancer Res, v. 11, p. 8699-705.

Falcaro, M., A. Pickles, D. F. Newbury, L. Addis, E. Banfield, S. E. Fisher, A. P. Monaco, Z. Simkin, G. Conti-Ramsden, and S. Consortium, 2008, Genetic and phenotypic effects of phonological short-term memory and grammatical morphology in specific language impairment: Genes Brain Behav, v. 7, p. 393-402.

Feng, M. Y., and R. Rao, 2013, New insights into store-independent Ca(2+) entry: secretory pathway calcium ATPase 2 in normal physiology and cancer: Int J Oral Sci, v. 5, p. 71-4.

Fernandez, T. V., S. J. Sanders, I. R. Yurkiewicz, A. G. Ercan-Sencicek, Y. S. Kim, D. O. Fishman, M. J. Raubeson, Y. Song, K. Yasuno, W. S. Ho, K. Bilguvar, J. Glessner, S. H. Chu, J. F. Leckman, R. A. King, D. L. Gilbert, G. A. Heiman, J. A. Tischfield, P. J. Hoekstra, B. Devlin, H. Hakonarson, S. M. Mane, M. Günel, and M. W. State, 2012, Rare copy number variants in tourette syndrome disrupt genes in histaminergic pathways and overlap with autism: Biol Psychiatry, v. 71, p. 392-402.

Feuk, L., A. R. Carson, and S. W. Scherer, 2006a, Structural variation in the human genome: Nat Rev Genet, v. 7, p. 85-97.

Feuk, L., A. Kalervo, M. Lipsanen-Nyman, J. Skaug, K. Nakabayashi, B. Finucane, D. Hartung, M. Innes, B. Kerem, M. J. Nowaczyk, J. Rivlin, W. Roberts, L. Senman, A. Summers, P. Szatmari, V. Wong, J. B. Vincent, S. Zeesman, L. R. Osborne, J. O. Cardy, J. Kere, S. W. Scherer, and K. Hannula-Jouppi, 2006b, Absence of a paternally inherited FOXP2 gene in developmental verbal dyspraxia: Am J Hum Genet, v. 79, p. 965-72.

Fisher, S. E., and C. Scharff, 2009, FOXP2 as a molecular window into speech and language: Trends Genet, v. 25, p. 166-77.

Flax, J. F., T. Realpe-Bonilla, L. S. Hirsch, L. M. Brzustowicz, C. W. Bartlett, and P. Tallal, 2003, Specific language impairment in families: evidence for co-occurrence with reading impairments: J Speech Lang Hear Res, v. 46, p. 530-43.

Forsberg, L. A., C. Rasi, H. R. Razzaghian, G. Pakalapati, L. Waite, K. S. Thilbeault, A. Ronowicz, N. E. Wineinger, H. K. Tiwari, D. Boomsma, M. P. Westerman, J. R. Harris, R. Lyle, M. Essand, F. Eriksson, T. L. Assimes, C. Iribarren, E. Strachan, T. P. O'Hanlon, L. G. Rider, F. W. Miller, V. Giedraitis, L. Lannfelt, M. Ingelsson, A. Piotrowski, N. L. Pedersen, D. Absher, and J. P. Dumanski, 2012, Age-related somatic structural changes in the nuclear genome of human blood cells: Am J Hum Genet, v. 90, p. 217-28.

Fridlyand, J., A. M. Snijders, B. Ylstra, H. Li, A. Olshen, R. Segraves, S. Dairkee, T. Tokuyasu, B. M. Ljung, A. N. Jain, J. McLennan, J. Ziegler, K. Chin, S. Devries, H. Feiler, J. W. Gray, F. Waldman, D. Pinkel, and D. G. Albertson, 2006, Breast tumor copy number aberration phenotypes and genomic instability: BMC Cancer, v. 6, p. 96.

Friedman, J. I., T. Vrijenhoek, S. Markx, I. M. Janssen, W. A. van der Vliet, B. H. Faas, N. V. Knoers, W. Cahn, R. S. Kahn, L. Edelmann, K. L. Davis, J. M. Silverman, H. G. Brunner, A. G. van Kessel, C. Wijmenga, R. A. Ophoff, and J. A. Veltman, 2008, CNTNAP2 gene dosage variation is associated with schizophrenia and epilepsy: Mol Psychiatry, v. 13, p. 261-6.

Fulker, D. W., L. R. Cardon, J. C. DeFries, W. J. Kimberling, B. F. Pennington, and S. D. Smith, 1991, Multiple regression analysis of sib-pair data on reading to detect quantitative trait loci: Reading Writing Interdiscip J, v. 3, p. 299–313

Gathercole, S. E., C. S. Willis, A. D. Baddeley, and H. Emslie, 1994, The Children's Test of Nonword Repetition: a test of phonological working memory: Memory, v. 2, p. 103-27.

Geschwind, D. H., J. Sowinski, C. Lord, P. Iversen, J. Shestack, P. Jones, L. Ducat, S. J. Spence, and A. S. Committee, 2001, The autism genetic resource exchange: a resource for the study of autism and related neuropsychiatric conditions: Am J Hum Genet, v. 69, p. 463-6.

Gibson, G., 2011, Rare and common variants: twenty arguments: Nat Rev Genet, v. 13, p. 135-45.

Gilman, S. R., I. Iossifov, D. Levy, M. Ronemus, M. Wigler, and D. Vitkup, 2011, Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses: Neuron, v. 70, p. 898-907.

Girirajan, S., Z. Brkanac, B. P. Coe, C. Baker, L. Vives, T. H. Vu, N. Shafer, R. Bernier, G. B. Ferrero, M. Silengo, S. T. Warren, C. S. Moreno, M. Fichera, C. Romano, W. H. Raskind, and E. E. Eichler, 2011, Relative burden of large CNVs on a range of neurodevelopmental phenotypes: PLoS Genet, v. 7, p. e1002334.

Girirajan, S., M. Y. Dennis, C. Baker, M. Malig, B. P. Coe, C. D. Campbell, K. Mark, T. H. Vu, C. Alkan, Z. Cheng, L. G. Biesecker, R. Bernier, and E. E. Eichler, 2013, Refinement and discovery of new hotspots of copy-number variation associated with autism spectrum disorder: Am J Hum Genet, v. 92, p. 221-37.

Girirajan, S., J. A. Rosenfeld, B. P. Coe, S. Parikh, N. Friedman, A. Goldstein, R. A. Filipink, J. S. McConnell, B. Angle, W. S. Meschino, M. M. Nezarati, A. Asamoah, K. E. Jackson, G. C. Gowans, J. A. Martin, E. P. Carmany, D. W. Stockton, R. E. Schnur, L. S. Penney, D. M. Martin, S. Raskin, K. Leppig, H. Thiese, R. Smith, E. Aberg, D. M. Niyazov, L. F. Escobar, D. El-Khechen, K. D. Johnson, R. R. Lebel, K. Siefkas, S. Ball, N. Shur, M. McGuire, C. K. Brasington, J. E. Spence, L. S. Martin, C. Clericuzio, B. C. Ballif, L. G. Shaffer, and E. E. Eichler, 2012, Phenotypic heterogeneity of genomic disorders and rare copy-number variants: N Engl J Med, v. 367, p. 1321-31.

Girirajan, S., J. A. Rosenfeld, G. M. Cooper, F. Antonacci, P. Siswara, A. Itsara, L. Vives, T. Walsh, S. E. McCarthy, C. Baker, H. C. Mefford, J. M. Kidd, S. R. Browning, B. L. Browning, D. E. Dickel, D. L. Levy, B. C. Ballif, K. Platky, D. M. Farber, G. C. Gowans, J. J. Wetherbee, A. Asamoah, D. D. Weaver, P. R. Mark, J. Dickerson, B. P. Garg, S. A. Ellingwood, R. Smith, V. C. Banks, W. Smith, M. T. McDonald, J. J. Hoo, B. N. French, C. Hudson, J. P. Johnson, J. R. Ozmore, J. B. Moeschler, U. Surti, L. F. Escobar, D. El-Khechen, J. L. Gorski, J. Kussmann, B. Salbert, Y. Lacassie, A. Biser, D. M. McDonald-McGinn, E. H. Zackai, M. A. Deardorff, T. H. Shaikh, E. Haan, K. L. Friend, M. Fichera, C. Romano, J. Gécz, L. E. DeLisi, J. Sebat, M. C. King, L. G. Shaffer, and E. E. Eichler, 2010, A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay: Nat Genet, v. 42, p. 203-9.

Glessner, J. T., K. Wang, G. Cai, O. Korvatska, C. E. Kim, S. Wood, H. Zhang, A. Estes, C. W. Brune, J. P. Bradfield, M. Imielinski, E. C. Frackelton, J. Reichert, E. L. Crawford, J. Munson, P. M. Sleiman, R. Chiavacci, K. Annaiah, K. Thomas, C. Hou, W. Glaberson, J. Flory, F. Otieno, M. Garris, L. Soorya, L. Klei, J. Piven, K. J. Meyer, E. Anagnostou, T. Sakurai, R. M. Game, D. S. Rudd, D. Zurawiecki, C. J. McDougle, L. K. Davis, J. Miller, D. J. Posey, S. Michaels, A. Kolevzon, J. M. Silverman, R. Bernier, S. E. Levy, R. T. Schultz, G. Dawson, T. Owley, W. M. McMahon, T. H. Wassink, J. A. Sweeney, J. I. Nurnberger, H. Coon, J. S. Sutcliffe, N. J. Minshew, S. F. Grant, M. Bucan, E. H. Cook, J. D. Buxbaum, B. Devlin, G. D. Schellenberg, and H. Hakonarson, 2009, Autism genome-wide copy number variation reveals ubiquitin and neuronal genes: Nature, v. 459, p. 569-73.

Goodier, J. L., and H. H. Kazazian, 2008, Retrotransposons revisited: the restraint and rehabilitation of parasites: Cell, v. 135, p. 23-35.

Grantham, R., 1974, Amino acid difference formula to help explain protein evolution: Science, v. 185, p. 862-4.

Greenberg, D. A., P. Abreu, and S. E. Hodge, 1998, The power to detect linkage in complex disease by means of simple LOD-score analyses: Am J Hum Genet, v. 63, p. 870-9.

Gregor, A., B. Albrecht, I. Bader, E. K. Bijlsma, A. B. Ekici, H. Engels, K. Hackmann, D. Horn, J. Hoyer, J. Klapecki, J. Kohlhase, I. Maystadt, S. Nagl, E. Prott, S. Tinschert, R. Ullmann, E. Wohlleber, G. Woods, A. Reis, A. Rauch, and C. Zweier, 2011, Expanding the clinical spectrum associated with defects in CNTNAP2 and NRXN1: BMC Med Genet, v. 12, p. 106.

Grimbert, P., A. Valanciute, V. Audard, P. Lang, G. Guellaën, and D. Sahali, 2004, The Filamin-A is a partner of Tc-mip, a new adapter protein involved in c-maf-dependent Th2 signaling pathway: Mol Immunol, v. 40, p. 1257-61.

Grisart, B., L. Willatt, A. Destrée, J. P. Fryns, K. Rack, T. de Ravel, J. Rosenfeld, J. R. Vermeesch, C. Verellen-Dumoulin, and R. Sandford, 2009, 17q21.31 microduplication patients are characterised by behavioural problems and poor social interaction: J Med Genet, v. 46, p. 524-30.

Gross, E., B. Busse, M. Riemenschneider, S. Neubauer, K. Seck, H. G. Klein, M. Kiechle, F. Lordick, and A. Meindl, 2008, Strong association of a common dihydropyrimidine dehydrogenase gene polymorphism with fluoropyrimidine-related toxicity in cancer patients: PLoS One, v. 3, p. e4003.

Gross, E., T. Ullrich, K. Seck, V. Mueller, M. de Wit, C. von Schilling, A. Meindl, M. Schmitt, and M. Kiechle, 2003, Detailed analysis of five mutations in dihydropyrimidine dehydrogenase detected in cancer patients with 5-fluorouracil-related side effects: Hum Mutat, v. 22, p. 498.

Grozeva, D., G. Kirov, D. Ivanov, I. R. Jones, L. Jones, E. K. Green, D. M. St Clair, A. H. Young, N. Ferrier, A. E. Farmer, P. McGuffin, P. A. Holmans, M. J. Owen, M. C. O'Donovan, N. Craddock, and W. T. C. C. Consortium, 2010, Rare copy number variants: a point of rarity in genetic risk for bipolar disorder and schizophrenia: Arch Gen Psychiatry, v. 67, p. 318-27.

Gu, W., F. Zhang, and J. R. Lupski, 2008, Mechanisms for human genomic rearrangements: Pathogenetics, v. 1, p. 4.

Guilmatre, A., C. Dubourg, A. L. Mosca, S. Legallic, A. Goldenberg, V. Drouin-Garraud, V. Layet, A. Rosier, S. Briault, F. Bonnet-Brilhault, F. Laumonnier, S. Odent, G. Le Vacon, G. Joly-Helas, V. David, C. Bendavid, J. M. Pinoit, C. Henry, C. Impallomeni, E. Germano, G. Tortorella, G. Di Rosa, C. Barthelemy, C. Andres, L. Faivre, T. Frébourg, P. Saugier Veber, and D. Campion, 2009, Recurrent rearrangements in synaptic and neurodevelopmental genes and shared biologic pathways in schizophrenia, autism, and mental retardation: Arch Gen Psychiatry, v. 66, p. 947-56.

Hajirasouliha, I., F. Hormozdiari, C. Alkan, J. M. Kidd, I. Birol, E. E. Eichler, and S. C. Sahinalp, 2010, Detection and characterization of novel sequence insertions using paired-end next-generation sequencing: Bioinformatics, v. 26, p. 1277-83.

Hallmayer, J., S. Cleveland, A. Torres, J. Phillips, B. Cohen, T. Torigoe, J. Miller, A. Fedele, J. Collins, K. Smith, L. Lotspeich, L. A. Croen, S. Ozonoff, C. Lajonchere, J. K. Grether, and N. Risch, 2011, Genetic heritability and shared environmental factors among twin pairs with autism: Arch Gen Psychiatry, v. 68, p. 1095-102.

Hammil, D. D., V. L. Brown, S. C. Larsen, and J. L. Wiederhold, 1987, Test of Adolescent Language-2, Austin, TX.

Hammil, D. D., and P. L. Newcomer, 1988, Test of Language Development-2, Intermediate, Austin, TX.

Haseman, J. K., and R. C. Elston, 1972, The investigation of linkage between a quantitative trait and a marker locus: Behav Genet, v. 2, p. 3-19.

Hastings, P. J., G. Ira, and J. R. Lupski, 2009, A microhomology-mediated break-induced replication model for the origin of human copy number variation: PLoS Genet, v. 5, p. e1000327.

Hehir-Kwa, J. Y., R. Pfundt, J. A. Veltman, and N. de Leeuw, 2013, Pathogenic or not? Assessing the clinical relevance of copy number variants: Clin Genet, v. 84, p. 415-21.

Helbig, I., H. C. Mefford, A. J. Sharp, M. Guipponi, M. Fichera, A. Franke, H. Muhle, C. de Kovel, C. Baker, S. von Spiczak, K. L. Kron, I. Steinich, A. A. Kleefuss-Lie, C. Leu, V. Gaus, B. Schmitz, K. M. Klein, P. S. Reif, F. Rosenow, Y. Weber, H. Lerche, F. Zimprich, L. Urak, K. Fuchs, M. Feucht, P. Genton, P. Thomas, F. Visscher, G. J. de Haan, R. S. Møller, H. Hjalgrim, D. Luciano, M. Wittig, M. Nothnagel, C. E. Elger, P. Nürnberg, C. Romano, A. Malafosse, B. P. Koeleman, D. Lindhout, U. Stephani, S. Schreiber, E. E. Eichler, and T. Sander, 2009, 15q13.3 microdeletions increase risk of idiopathic generalized epilepsy: Nat Genet, v. 41, p. 160-2.

Henderson, M., S. Jones, P. Walker, J. Duley, and H. Simmonds, 1995, Heterogeneity of symptomatology in two male siblings with thymine uraciluria: Journal of Inherited Metabolic Disease, v. 18, p. 85-86.

Holt, R., N. H. Sykes, I. C. Conceição, J. B. Cazier, R. J. Anney, G. Oliveira, L. Gallagher, A. Vicente, A. P. Monaco, and A. T. Pagnamenta, 2012, CNVs leading to fusion transcripts in individuals with autism spectrum disorder: Eur J Hum Genet, v. 20, p. 1141-7.

Howlin, P., L. Mawhood, and M. Rutter, 2000, Autism and developmental receptive language disorder--a follow-up comparison in early adult life. II: Social, behavioural, and psychiatric outcomes: J Child Psychol Psychiatry, v. 41, p. 561-78.

Hus, V., A. Pickles, E. H. Cook, S. Risi, and C. Lord, 2007, Using the autism diagnostic interview--revised to increase phenotypic homogeneity in genetic studies of autism: Biol Psychiatry, v. 61, p. 438-48.

Iafrate, A. J., L. Feuk, M. N. Rivera, M. L. Listewnik, P. K. Donahoe, Y. Qi, S. W. Scherer, and C. Lee, 2004, Detection of large-scale variation in the human genome: Nat Genet, v. 36, p. 949-51.

IMGSAC, 1998, A full genome screen for autism with evidence for linkage to a region on chromosome 7q. International Molecular Genetic Study of Autism Consortium: Hum Mol Genet, v. 7, p. 571-8.

IMGSAC, 2001, A genomewide screen for autism: strong evidence for linkage to chromosomes 2q, 7q, and 16p: Am J Hum Genet, v. 69, p. 570-81.

International_Schizophrenia_Consortium, 2008, Rare chromosomal deletions and duplications increase risk of schizophrenia: Nature, v. 455, p. 237-41.

Itsara, A., G. M. Cooper, C. Baker, S. Girirajan, J. Li, D. Absher, R. M. Krauss, R. M. Myers, P. M. Ridker, D. I. Chasman, H. Mefford, P. Ying, D. A. Nickerson, and E. E. Eichler, 2009, Population analysis of large copy number variants and hotspots of human genetic disease: Am J Hum Genet, v. 84, p. 148-61.

Itsara, A., H. Wu, J. D. Smith, D. A. Nickerson, I. Romieu, S. J. London, and E. E. Eichler, 2010, De novo rates and selection of large copy number variation: Genome Res, v. 20, p. 1469-81.

Jacobs, K. B., M. Yeager, W. Zhou, S. Wacholder, Z. Wang, B. Rodriguez-Santiago, A. Hutchinson, X. Deng, C. Liu, M. J. Horner, M. Cullen, C. G. Epstein, L. Burdett, M. C. Dean, N. Chatterjee, J. Sampson, C. C. Chung, J. Kovaks, S. M. Gapstur, V. L. Stevens, L. T. Teras, M. M. Gaudet, D.

Albanes, S. J. Weinstein, J. Virtamo, P. R. Taylor, N. D. Freedman, C. C. Abnet, A. M. Goldstein, N. Hu, K. Yu, J. M. Yuan, L. Liao, T. Ding, Y. L. Qiao, Y. T. Gao, W. P. Koh, Y. B. Xiang, Z. Z. Tang, J. H. Fan, M. C. Aldrich, C. Amos, W. J. Blot, C. H. Bock, E. M. Gillanders, C. C. Harris, C. A. Haiman, B. E. Henderson, L. N. Kolonel, L. Le Marchand, L. H. McNeill, B. A. Rybicki, A. G. Schwartz, L. B. Signorello, M. R. Spitz, J. K. Wiencke, M. Wrensch, X. Wu, K. A. Zanetti, R. G. Ziegler, J. D. Figueroa, M. Garcia-Closas, N. Malats, G. Marenne, L. Prokunina-Olsson, D. Baris, M. Schwenn, A. Johnson, M. T. Landi, L. Goldin, D. Consonni, P. A. Bertazzi, M. Rotunno, P. Rajaraman, U. Andersson, L. E. Beane Freeman, C. D. Berg, J. E. Buring, M. A. Butler, T. Carreon, M. Feychting, A. Ahlbom, J. M. Gaziano, G. G. Giles, G. Hallmans, S. E. Hankinson, P. Hartge, R. Henriksson, P. D. Inskip, C. Johansen, A. Landgren, R. McKean-Cowdin, D. S. Michaud, B. S. Melin, U. Peters, A. M. Ruder, H. D. Sesso, G. Severi, X. O. Shu, K. Visvanathan, et al., 2012, Detectable clonal mosaicism and its relationship to aging and cancer: Nat Genet, v. 44, p. 651-8.

Johnson, M. R., J. Yan, L. Shao, N. Albin, and R. B. Diasio, 1997, Semi-automated radioassay for determination of dihydropyrimidine dehydrogenase (DPD) activity. Screening cancer patients for DPD deficiency, a condition associated with 5-fluorouracil toxicity: J Chromatogr B Biomed Sci Appl, v. 696, p. 183-91.

Jones, R. W., S. Ring, L. Tyfield, R. Hamvas, H. Simmons, M. Pembrey, J. Golding, and A. S. Team, 2000, A new human genetic resource: a DNA bank established as part of the Avon longitudinal study of pregnancy and childhood (ALSPAC): Eur J Hum Genet, v. 8, p. 653-60.

Jonsdottir, S., A. Bouma, J. A. Sergeant, and E. J. Scherder, 2005, The impact of specific language impairment on working memory in children with ADHD combined subtype: Arch Clin Neuropsychol, v. 20, p. 443-56.

Kamal, M., A. Pawlak, F. BenMohamed, A. Valanciuté, K. Dahan, M. Candelier, P. Lang, G. Guellaën, and D. Sahali, 2010, C-mip interacts with the p85 subunit of PI3 kinase and exerts a dual effect on ERK signaling via the recruitment of Dip1 and DAP kinase: FEBS Lett, v. 584, p. 500-6.

Kamal, M., A. Valanciute, K. Dahan, V. Ory, A. Pawlak, P. Lang, G. Guellaen, and D. Sahali, 2009, C-mip interacts physically with RelA and inhibits nuclear factor kappa B activity: Mol Immunol, v. 46, p. 991-8.

Kaminsky, E. B., V. Kaul, J. Paschall, D. M. Church, B. Bunke, D. Kunig, D. Moreno-De-Luca, A. Moreno-De-Luca, J. G. Mulle, S. T. Warren, G. Richard, J. G. Compton, A. E. Fuller, T. J. Gliem, S. Huang, M. N. Collinson, S. J. Beal, T. Ackley, D. L. Pickering, D. M. Golden, E. Aston, H. Whitby, S. Shetty, M. R. Rossi, M. K. Rudd, S. T. South, A. R. Brothman, W. G. Sanger, R. K. Iyer, J. A. Crolla, E. C. Thorland, S. Aradhya, D. H. Ledbetter, and C. L. Martin, 2011, An evidence-based approach to establish the functional and clinical significance of copy number variants in intellectual and developmental disabilities: Genet Med, v. 13, p. 777-84.

Kennedy, S. R., L. A. Loeb, and A. J. Herr, 2012, Somatic mutations in aging, cancer and neurodegeneration: Mech Ageing Dev, v. 133, p. 118-26.

Khajavi, M., K. Inoue, and J. R. Lupski, 2006, Nonsense-mediated mRNA decay modulates clinical outcome of genetic disease: Eur J Hum Genet, v. 14, p. 1074-81.

Kidd, J. M., G. M. Cooper, W. F. Donahue, H. S. Hayden, N. Sampas, T. Graves, N. Hansen, B. Teague, C. Alkan, F. Antonacci, E. Haugen, T. Zerr, N. A. Yamada, P. Tsang, T. L. Newman, E. Tüzün, Z. Cheng, H. M. Ebling, N. Tusneem, R. David, W. Gillett, K. A. Phelps, M. Weaver, D. Saranga, A. Brand, W. Tao, E. Gustafson, K. McKernan, L. Chen, M. Malig, J. D. Smith, J. M. Korn, S. A. McCarroll, D. A. Altshuler, D. A. Peiffer, M. Dorschner, J. Stamatoyannopoulos, D. Schwartz, D. A. Nickerson, J. C. Mullikin, R. K. Wilson, L. Bruhn, M. V. Olson, R. Kaul, D. R. Smith, and E. E. Eichler, 2008, Mapping and sequencing of structural variation from eight human genomes: Nature, v. 453, p. 56-64.

Kirov, G., D. Grozeva, N. Norton, D. Ivanov, K. K. Mantripragada, P. Holmans, N. Craddock, M. J. Owen, M. C. O'Donovan, I. S. Consortium, and W. T. C. C. Consortium, 2009, Support for the involvement of large copy number variants in the pathogenesis of schizophrenia: Hum Mol Genet, v. 18, p. 1497-503.

Kirov, G., A. J. Pocklington, P. Holmans, D. Ivanov, M. Ikeda, D. Ruderfer, J. Moran, K. Chambert, D. Toncheva, L. Georgieva, D. Grozeva, M. Fjodorova, R. Wollerton, E. Rees, I. Nikolov, L. N. van de Lagemaat, A. Bayés, E. Fernandez, P. I. Olason, Y. Böttcher, N. H. Komiyama, M. O. Collins, J. Choudhary, K. Stefansson, H. Stefansson, S. G. Grant, S. Purcell, P. Sklar, M. C. O'Donovan, and

M. J. Owen, 2012, De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia: Mol Psychiatry, v. 17, p. 142-53.

Kitsiou-Tzeli, S., H. Frysira, K. Giannikou, A. Syrmou, K. Kosma, G. Kakourou, E. Leze, C. Sofocleous, E. Kanavakis, and M. Tzetis, 2012, Microdeletion and microduplication 17q21.31 plus an additional CNV, in patients with intellectual disability, identified by array-CGH: Gene, v. 492, p. 319-24.

Kjelgaard, M. M., and H. Tager-Flusberg, 2001, An Investigation of Language Impairment in Autism: Implications for Genetic Subgroups: Lang Cogn Process, v. 16, p. 287-308.

Koolen, D. A., J. M. Kramer, K. Neveling, W. M. Nillesen, H. L. Moore-Barton, F. V. Elmslie, A. Toutain, J. Amiel, V. Malan, A. C. Tsai, S. W. Cheung, C. Gilissen, E. T. Verwiel, S. Martens, T. Feuth, E. M. Bongers, P. de Vries, H. Scheffer, L. E. Vissers, A. P. de Brouwer, H. G. Brunner, J. A. Veltman, A. Schenck, H. G. Yntema, and B. B. de Vries, 2012, Mutations in the chromatin modifier gene KANSL1 cause the 17q21.31 microdeletion syndrome: Nat Genet, v. 44, p. 639-41.

Korbel, J. O., A. E. Urban, J. P. Affourtit, B. Godwin, F. Grubert, J. F. Simons, P. M. Kim, D. Palejev, N. J. Carriero, L. Du, B. E. Taillon, Z. Chen, A. Tanzer, A. C. Saunders, J. Chi, F. Yang, N. P. Carter, M. E. Hurles, S. M. Weissman, T. T. Harkins, M. B. Gerstein, M. Egholm, and M. Snyder, 2007, Paired-end mapping reveals extensive structural variation in the human genome: Science, v. 318, p. 420-6.

Korbie, D. J., and J. S. Mattick, 2008, Touchdown PCR for increased specificity and sensitivity in PCR amplification: Nat Protoc, v. 3, p. 1452-6.

Korn, J. M., F. G. Kuruvilla, S. A. McCarroll, A. Wysoker, J. Nemesh, S. Cawley, E. Hubbell, J. Veitch, P. J. Collins, K. Darvishi, C. Lee, M. M. Nizzari, S. B. Gabriel, S. Purcell, M. J. Daly, and D. Altshuler, 2008, Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs: Nat Genet, v. 40, p. 1253-60.

Kumar, P., S. Henikoff, and P. C. Ng, 2009, Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm: Nat Protoc, v. 4, p. 1073-81.

LaFramboise, T., 2009, Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances: Nucleic Acids Res, v. 37, p. 4181-93.

Lai, C. S., S. E. Fisher, J. A. Hurst, F. Vargha-Khadem, and A. P. Monaco, 2001, A forkhead-domain gene is mutated in a severe speech and language disorder: Nature, v. 413, p. 519-23.

Law, J., J. Boyle, F. Harris, A. Harkness, and C. Nye, 2000, Prevalence and natural history of primary speech and language delay: findings from a systematic review of the literature: Int J Lang Commun Disord, v. 35, p. 165-88.

Leblond, C. S., J. Heinrich, R. Delorme, C. Proepper, C. Betancur, G. Huguet, M. Konyukh, P. Chaste, E. Ey, M. Rastam, H. Anckarsäter, G. Nygren, I. C. Gillberg, J. Melke, R. Toro, B. Regnault, F. Fauchereau, O. Mercati, N. Lemière, D. Skuse, M. Poot, R. Holt, A. P. Monaco, I. Järvelä, K. Kantojärvi, R. Vanhala, S. Curran, D. A. Collier, P. Bolton, A. Chiocchetti, S. M. Klauck, F. Poustka, C. M. Freitag, R. Waltes, M. Kopp, E. Duketis, E. Bacchelli, F. Minopoli, L. Ruta, A. Battaglia, L. Mazzone, E. Maestrini, A. F. Sequeira, B. Oliveira, A. Vicente, G. Oliveira, D. Pinto, S. W. Scherer, D. Zelenika, M. Delepine, M. Lathrop, D. Bonneau, V. Guinchat, F. Devillard, B. Assouline, M. C. Mouren, M. Leboyer, C. Gillberg, T. M. Boeckers, and T. Bourgeron, 2012, Genetic and functional analyses of SHANK2 mutations suggest a multiple hit model of autism spectrum disorders: PLoS Genet, v. 8, p. e1002521.

Lee, H. K., and A. Kirkwood, 2011, AMPA receptor regulation during synaptic plasticity in hippocampus and neocortex: Semin Cell Dev Biol, v. 22, p. 514-20.

Lee, J. A., C. M. Carvalho, and J. R. Lupski, 2007, A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders: Cell, v. 131, p. 1235-47.

Lennon, P. A., M. L. Cooper, D. A. Peiffer, K. L. Gunderson, A. Patel, S. Peters, S. W. Cheung, and C. A. Bacino, 2007, Deletion of 7q31.1 supports involvement of FOXP2 in language impairment: clinical report and review: Am J Med Genet A, v. 143A, p. 791-8.

Lesch, K. P., N. Timmesfeld, T. J. Renner, R. Halperin, C. Röser, T. T. Nguyen, D. W. Craig, J. Romanos, M. Heine, J. Meyer, C. Freitag, A. Warnke, M. Romanos, H. Schäfer, S. Walitza, A. Reif, D. A. Stephan, and C. Jacob, 2008, Molecular genetics of adult ADHD: converging evidence from genome-wide association and extended pedigree linkage studies: J Neural Transm, v. 115, p. 1573-85.

Levy, D., M. Ronemus, B. Yamrom, Y. H. Lee, A. Leotta, J. Kendall, S. Marks, B. Lakshmi, D. Pai, K. Ye, A. Buja, A. Krieger, S. Yoon, J. Troge, L. Rodgers, I. Iossifov, and M. Wigler, 2011, Rare de novo and transmitted copy-number variation in autistic spectrum disorders: Neuron, v. 70, p. 886-97.

Levy, S. E., D. S. Mandell, and R. T. Schultz, 2009, Autism: Lancet, v. 374, p. 1627-38.

Lewis, C. M., and J. Knight, 2012, Introduction to genetic association studies: Cold Spring Harb Protoc, v. 2012, p. 297-306.

Leyfer, O. T., H. Tager-Flusberg, M. Dowd, J. B. Tomblin, and S. E. Folstein, 2008, Overlap between autism and specific language impairment: comparison of Autism Diagnostic Interview and Autism Diagnostic Observation Schedule scores: Autism Res, v. 1, p. 284-96.

Li, J., T. Yang, L. Wang, H. Yan, Y. Zhang, Y. Guo, F. Pan, Z. Zhang, Y. Peng, Q. Zhou, L. He, X. Zhu, H. Deng, S. Levy, C. J. Papasian, B. M. Drees, J. J. Hamilton, R. R. Recker, J. Cheng, and H. W. Deng, 2009, Whole genome distribution and ethnic differentiation of copy number variation in Caucasian and Asian populations: PLoS One, v. 4, p. e7958.

Liang, H., W. Guo, and L. Nagarajan, 2000, Chromosomal mapping and genomic organization of an evolutionarily conserved zinc finger gene ZNF277: Genomics, v. 66, p. 226-8.

Lieber, M. R., 2008, The mechanism of human nonhomologous DNA end joining: J Biol Chem, v. 283, p. 1-5.

Lieber, M. R., Y. Ma, U. Pannicke, and K. Schwarz, 2003, Mechanism and regulation of human non-homologous DNA end-joining: Nat Rev Mol Cell Biol, v. 4, p. 712-20.

Lin, P. I., J. M. Vance, M. A. Pericak-Vance, and E. R. Martin, 2007, No gene is an island: the flip-flop phenomenon: Am J Hum Genet, v. 80, p. 531-8.

Liu, J., D. R. Nyholt, P. Magnussen, E. Parano, P. Pavone, D. Geschwind, C. Lord, P. Iversen, J. Hoh, J. Ott, T. C. Gilliam, and A. G. R. E. Consortium, 2001, A genomewide screen for autism susceptibility loci: Am J Hum Genet, v. 69, p. 327-40.

Livak, K. J., and T. D. Schmittgen, 2001, Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method: Methods, v. 25, p. 402-8.

Loganayagam, A., M. Arenas Hernandez, A. Corrigan, L. Fairbanks, C. M. Lewis, P. Harper, N. Maisey, P. Ross, J. D. Sanderson, and A. M. Marinaki, 2013, Pharmacogenetic variants in the DPYD, TYMS, CDA and MTHFR genes are clinically significant predictors of fluoropyrimidine toxicity: Br J Cancer, v. 108, p. 2505-15.

Lord, C., S. Risi, L. Lambrecht, E. H. Cook, B. L. Leventhal, P. C. DiLavore, A. Pickles, and M. Rutter, 2000, The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism: J Autism Dev Disord, v. 30, p. 205-23.

Lord, C., M. Rutter, S. Goode, J. Heemsbergen, H. Jordan, L. Mawhood, and E. Schopler, 1989, Autism diagnostic observation schedule: a standardized observation of communicative and social behavior: J Autism Dev Disord, v. 19, p. 185-212.

Lord, C., M. Rutter, and A. Le Couteur, 1994, Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders: J Autism Dev Disord, v. 24, p. 659-85.

Luciano, M., D. M. Evans, N. K. Hansell, S. E. Medland, G. W. Montgomery, N. G. Martin, M. J. Wright, and T. C. Bates, 2013, A genome-wide association study for reading and language abilities in two population cohorts: Genes Brain Behav, v. 12, p. 645-52.

Lupski, J. R., 2004, Hotspots of homologous recombination in the human genome: not all homologous sequences are equal: Genome Biol, v. 5, p. 242.

Lupski, J. R., 2010, Retrotransposition and structural variation in the human genome: Cell, v. 141, p. 1110-2.

Lupski, J. R., and P. Stankiewicz, 2005, Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes: PLoS Genet, v. 1, p. e49.

MacDermot, K. D., E. Bonora, N. Sykes, A. M. Coupe, C. S. Lai, S. C. Vernes, F. Vargha-Khadem, F. McKenzie, R. L. Smith, A. P. Monaco, and S. E. Fisher, 2005, Identification of FOXP2 truncation as a novel cause of developmental speech and language deficits: Am J Hum Genet, v. 76, p. 1074-80.

Macdonald, J. R., R. Ziman, R. K. Yuen, L. Feuk, and S. W. Scherer, 2013, The database of genomic variants: a curated collection of structural variation in the human genome: Nucleic Acids Res.

Maestrini, E., A. T. Pagnamenta, J. A. Lamb, E. Bacchelli, N. H. Sykes, I. Sousa, C. Toma, G. Barnby, H. Butler, L. Winchester, T. S. Scerri, F. Minopoli, J. Reichert, G. Cai, J. D. Buxbaum, O. Korvatska, G. D. Schellenberg, G. Dawson, A. de Bildt, R. B. Minderaa, E. J. Mulder, A. P. Morris, A. J.

Bailey, A. P. Monaco, and Imgsac, 2010, High-density SNP association study and copy number variation analysis of the AUTS1 and AUTS5 loci implicate the IMMP2L-DOCK4 gene region in autism susceptibility: Molecular Psychiatry, v. 15, p. 954-968.

Magri, C., E. Sacchetti, M. Traversa, P. Valsecchi, R. Gardella, C. Bonvicini, A. Minelli, M. Gennarelli, and S. Barlati, 2010, New copy number variations in schizophrenia: PLoS One, v. 5, p. e13422.

Malhotra, D., S. McCarthy, J. J. Michaelson, V. Vacic, K. E. Burdick, S. Yoon, S. Cichon, A. Corvin, S. Gary, E. S. Gershon, M. Gill, M. Karayiorgou, J. R. Kelsoe, O. Krastoshevsky, V. Krause, E. Leibenluft, D. L. Levy, V. Makarov, A. Bhandari, A. K. Malhotra, F. J. McMahon, M. M. Nöthen, J. B. Potash, M. Rietschel, T. G. Schulze, and J. Sebat, 2011, High frequencies of de novo CNVs in bipolar disorder and schizophrenia: Neuron, v. 72, p. 951-63.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. Mackay, S. A. McCarroll, and P. M. Visscher, 2009, Finding the missing heritability of complex diseases: Nature, v. 461, p. 747-53.

Marchman, V. A., B. Wulfeck, and S. Ellis Weismer, 1999, Morphological productivity in children with normal language and SLI: a study of the English past tense: J Speech Lang Hear Res, v. 42, p. 206-19.

Marenne, G., B. Rodríguez-Santiago, M. G. Closas, L. Pérez-Jurado, N. Rothman, D. Rico, G. Pita, D. G. Pisano, M. Kogevinas, D. T. Silverman, A. Valencia, F. X. Real, S. J. Chanock, E. Génin, and N. Malats, 2011, Assessment of copy number variation using the Illumina Infinium 1M SNP-array: a comparison of methodological approaches in the Spanish Bladder Cancer/EPICURO study: Hum Mutat, v. 32, p. 240-8.

Marshall, C. R., A. Noor, J. B. Vincent, A. C. Lionel, L. Feuk, J. Skaug, M. Shago, R. Moessner, D. Pinto, Y. Ren, B. Thiruvahindrapduram, A. Fiebig, S. Schreiber, J. Friedman, C. E. Ketelaars, Y. J. Vos, C. Ficicioglu, S. Kirkpatrick, R. Nicolson, L. Sloman, A. Summers, C. A. Gibbons, A. Teebi, D. Chitayat, R. Weksberg, A. Thompson, C. Vardy, V. Crosbie, S. Luscombe, R. Baatjes, L. Zwaigenbaum, W. Roberts, B. Fernandez, P. Szatmari, and S. W. Scherer, 2008, Structural variation of chromosomes in autism spectrum disorder: Am J Hum Genet, v. 82, p. 477-88.

Masurel-Paulet, A., J. Andrieux, P. Callier, J. M. Cuisset, C. Le Caignec, M. Holder, C. Thauvin-Robinet, B. Doray, E. Flori, M. P. Alex-Cordier, M. Beri, O. Boute, B. Delobel, A. Dieux, L. Vallee, S. Jaillard, S. Odent, B. Isidor, C. Beneteau, J. Vigneron, F. Bilan, B. Gilbert-Dussardier, C. Dubourg, A. Labalme, C. Bidon, A. Gautier, P. Pernes, J. M. Pinoit, F. Huet, F. Mugneret, B. Aral, P. Jonveaux, D. Sanlaville, and L. Faivre, 2010, Delineation of 15q13.3 microdeletions: Clin Genet, v. 78, p. 149-61.

Mattison, L. K., H. Ezzeldin, M. Carpenter, A. Modak, M. R. Johnson, and R. B. Diasio, 2004, Rapid identification of dihydropyrimidine dehydrogenase deficiency by using a novel 2-13C-uracil breath test: Clin Cancer Res, v. 10, p. 2652-8.

Mattison, L. K., M. R. Johnson, and R. B. Diasio, 2002, A comparative analysis of translated dihydropyrimidine dehydrogenase cDNA; conservation of functional domains and relevance to genetic polymorphisms: Pharmacogenetics, v. 12, p. 133-44.

Mawhood, L., P. Howlin, and M. Rutter, 2000, Autism and developmental receptive language disorder--a comparative follow-up in early adult life. I: Cognitive and language outcomes: J Child Psychol Psychiatry, v. 41, p. 547-59.

McCarthy, M. I., G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. Ioannidis, and J. N. Hirschhorn, 2008, Genome-wide association studies for complex traits: consensus, uncertainty and challenges: Nat Rev Genet, v. 9, p. 356-69.

McCarthy, S. E., V. Makarov, G. Kirov, A. M. Addington, J. McClellan, S. Yoon, D. O. Perkins, D. E. Dickel, M. Kusenda, O. Krastoshevsky, V. Krause, R. A. Kumar, D. Grozeva, D. Malhotra, T. Walsh, E. H. Zackai, P. Kaplan, J. Ganesh, I. D. Krantz, N. B. Spinner, P. Roccanova, A. Bhandari, K. Pavon, B. Lakshmi, A. Leotta, J. Kendall, Y. H. Lee, V. Vacic, S. Gary, L. M. Iakoucheva, T. J. Crow, S. L. Christian, J. A. Lieberman, T. S. Stroup, T. Lehtimäki, K. Puura, C. Haldeman-Englert, J. Pearl, M. Goodell, V. L. Willour, P. Derosse, J. Steele, L. Kassem, J. Wolff, N. Chitkara, F. J. McMahon, A. K. Malhotra, J. B. Potash, T. G. Schulze, M. M. Nöthen, S. Cichon, M. Rietschel, E. Leibenluft, V. Kustanovich, C. M. Lajonchere, J. S. Sutcliffe, D. Skuse, M. Gill, L. Gallagher, N. R. Mendell, N. Craddock, M. J. Owen, M. C. O'Donovan, T. H. Shaikh, E. Susser, L. E. Delisi, P. F.

Sullivan, C. K. Deutsch, J. Rapoport, D. L. Levy, M. C. King, J. Sebat, and W. T. C. C. Consortium, 2009, Microduplications of 16p11.2 are associated with schizophrenia: Nat Genet, v. 41, p. 1223-7.

McClellan, J., and M. C. King, 2010, Genetic heterogeneity in human disease: Cell, v. 141, p. 210-7.

McQuillin, A., N. Bass, A. Anjorin, J. Lawrence, R. Kandaswamy, G. Lydall, J. Moran, P. Sklar, S. Purcell, and H. Gurling, 2011, Analysis of genetic deletions and duplications in the University College London bipolar disorder case control sample: Eur J Hum Genet, v. 19, p. 588-92.

Meaburn, E., P. S. Dale, I. W. Craig, and R. Plomin, 2002, Language-impaired children: No sign of the FOXP2 mutation: Neuroreport, v. 13, p. 1075-7.

Mead, A. N., and D. N. Stephens, 2003, Selective disruption of stimulus-reward learning in glutamate receptor gria1 knock-out mice: J Neurosci, v. 23, p. 1041-8.

Micheli, V., M. Camici, M. G. Tozzi, P. L. Ipata, S. Sestini, M. Bertelli, and G. Pompucci, 2011, Neurological disorders of purine and pyrimidine metabolism: Curr Top Med Chem, v. 11, p. 923-47.

Mick, E., A. Todorov, S. Smalley, X. Hu, S. Loo, R. D. Todd, J. Biederman, D. Byrne, B. Dechairo, A. Guiney, J. McCracken, J. McGough, S. F. Nelson, A. M. Reiersen, T. E. Wilens, J. Wozniak, B. M. Neale, and S. V. Faraone, 2010, Family-based genome-wide association scan of attention-deficit/hyperactivity disorder: J Am Acad Child Adolesc Psychiatry, v. 49, p. 898-905.e3.

Miller, D. T., Y. Shen, L. A. Weiss, J. Korn, I. Anselm, C. Bridgemohan, G. F. Cox, H. Dickinson, J. Gentile, D. J. Harris, V. Hegde, R. Hundley, O. Khwaja, S. Kothare, C. Luedke, R. Nasir, A. Poduri, K. Prasad, P. Raffalli, A. Reinhard, S. E. Smith, M. M. Sobeih, J. S. Soul, J. Stoler, M. Takeoka, W. H. Tan, J. Thakuria, R. Wolff, R. Yusupov, J. F. Gusella, M. J. Daly, and B. L. Wu, 2009, Microdeletion/duplication at 15q13.2q13.3 among individuals with features of autism and other neuropsychiatric disorders: J Med Genet, v. 46, p. 242-8.

Mills, R. E., K. Walter, C. Stewart, R. E. Handsaker, K. Chen, C. Alkan, A. Abyzov, S. C. Yoon, K. Ye, R. K. Cheetham, A. Chinwalla, D. F. Conrad, Y. Fu, F. Grubert, I. Hajirasouliha, F. Hormozdiari, L. M. Iakoucheva, Z. Iqbal, S. Kang, J. M. Kidd, M. K. Konkel, J. Korn, E. Khurana, D. Kural, H. Y. Lam, J. Leng, R. Li, Y. Li, C. Y. Lin, R. Luo, X. J. Mu, J. Nemesh, H. E. Peckham, T. Rausch, A. Scally, X. Shi, M. P. Stromberg, A. M. Stütz, A. E. Urban, J. A. Walker, J. Wu, Y. Zhang, Z. D. Zhang, M. A. Batzer, L. Ding, G. T. Marth, G. McVean, J. Sebat, M. Snyder, J. Wang, E. E. Eichler, M. B. Gerstein, M. E. Hurles, C. Lee, S. A. McCarroll, J. O. Korbel, and G. Project, 2011, Mapping copy number variation by population-scale genome sequencing: Nature, v. 470, p. 59-65.

Molofsky, A. V., S. He, M. Bydon, S. J. Morrison, and R. Pardal, 2005, Bmi-1 promotes neural stem cell self-renewal and neural development but not mouse growth and survival by repressing the p16Ink4a and p19Arf senescence pathways: Genes Dev, v. 19, p. 1432-7.

Molofsky, A. V., R. Pardal, T. Iwashita, I. K. Park, M. F. Clarke, and S. J. Morrison, 2003, Bmi-1 dependence distinguishes neural stem cell self-renewal from progenitor proliferation: Nature, v. 425, p. 962-7.

Moreno-De-Luca, D., S. J. Sanders, A. J. Willsey, J. G. Mulle, J. K. Lowe, D. H. Geschwind, M. W. State, C. L. Martin, and D. H. Ledbetter, 2013, Using large clinical data sets to infer pathogenicity for rare copy number variants in autism cohorts: Mol Psychiatry, v. 18, p. 1090-5.

Mulle, J. G., A. E. Pulver, J. A. McGrath, P. S. Wolyniec, A. F. Dodd, D. J. Cutler, J. Sebat, D. Malhotra, G. Nestadt, D. F. Conrad, M. Hurles, C. P. Barnes, M. Ikeda, N. Iwata, D. F. Levinson, P. V. Gejman, A. R. Sanders, J. Duan, A. A. Mitchell, I. Peter, P. Sklar, C. T. O'Dushlaine, D. Grozeva, M. C. O'Donovan, M. J. Owen, C. M. Hultman, A. K. Kähler, P. F. Sullivan, G. Kirov, S. T. Warren, and T. M. G. o. S. Consortium, 2013, Reciprocal Duplication of the Williams-Beuren Syndrome Deletion on Chromosome 7q11.23 Is Associated with Schizophrenia: Biol Psychiatry.

Nag, A., E. G. Bochukova, B. Kremeyer, D. D. Campbell, H. Muller, A. V. Valencia-Duarte, J. Cardona, I. C. Rivas, S. C. Mesa, M. Cuartas, J. Garcia, G. Bedoya, W. Cornejo, L. D. Herrera, R. Romero, E. Fournier, V. I. Reus, T. L. Lowe, I. S. Farooqi, C. A. Mathews, L. M. McGrath, D. Yu, E. Cook, K. Wang, J. M. Scharf, D. L. Pauls, N. B. Freimer, V. Plagnol, A. Ruiz-Linares, and T. S. A. I. C. f. Genetics, 2013, CNV analysis in Tourette syndrome implicates large genomic rearrangements in COL8A1 and NRXN1: PLoS One, v. 8, p. e59061.

Nagase, T., R. Kikuno, A. Hattori, Y. Kondo, K. Okumura, and O. Ohara, 2000, Prediction of the coding sequences of unidentified human genes. XIX. The complete sequences of 100 new cDNA clones from brain which code for large proteins in vitro: DNA Res, v. 7, p. 347-55.

Neale, B. M., Y. Kou, L. Liu, A. Ma'ayan, K. E. Samocha, A. Sabo, C. F. Lin, C. Stevens, L. S. Wang, V. Makarov, P. Polak, S. Yoon, J. Maguire, E. L. Crawford, N. G. Campbell, E. T. Geller, O.

Valladares, C. Schafer, H. Liu, T. Zhao, G. Cai, J. Lihm, R. Dannenfelser, O. Jabado, Z. Peralta, U. Nagaswamy, D. Muzny, J. G. Reid, I. Newsham, Y. Wu, L. Lewis, Y. Han, B. F. Voight, E. Lim, E. Rossin, A. Kirby, J. Flannick, M. Fromer, K. Shakir, T. Fennell, K. Garimella, E. Banks, R. Poplin, S. Gabriel, M. DePristo, J. R. Wimbish, B. E. Boone, S. E. Levy, C. Betancur, S. Sunyaev, E. Boerwinkle, J. D. Buxbaum, E. H. Cook, B. Devlin, R. A. Gibbs, K. Roeder, G. D. Schellenberg, J. S. Sutcliffe, and M. J. Daly, 2012, Patterns and rates of exonic de novo mutations in autism spectrum disorders: Nature, v. 485, p. 242-5.

Neale, B. M., S. Medland, S. Ripke, R. J. Anney, P. Asherson, J. Buitelaar, B. Franke, M. Gill, L. Kent, P. Holmans, F. Middleton, A. Thapar, K. P. Lesch, S. V. Faraone, M. Daly, T. T. Nguyen, H. Schäfer, H. C. Steinhausen, A. Reif, T. J. Renner, M. Romanos, J. Romanos, A. Warnke, S. Walitza, C. Freitag, J. Meyer, H. Palmason, A. Rothenberger, Z. Hawi, J. Sergeant, H. Roeyers, E. Mick, J. Biederman, and I. I. C. Group, 2010a, Case-control genome-wide association study of attention-deficit/hyperactivity disorder: J Am Acad Child Adolesc Psychiatry, v. 49, p. 906-20.

Neale, B. M., S. E. Medland, S. Ripke, P. Asherson, B. Franke, K. P. Lesch, S. V. Faraone, T. T. Nguyen, H. Schäfer, P. Holmans, M. Daly, H. C. Steinhausen, C. Freitag, A. Reif, T. J. Renner, M. Romanos, J. Romanos, S. Walitza, A. Warnke, J. Meyer, H. Palmason, J. Buitelaar, A. A. Vasquez, N. Lambregts-Rommelse, M. Gill, R. J. Anney, K. Langely, M. O'Donovan, N. Williams, M. Owen, A. Thapar, L. Kent, J. Sergeant, H. Roeyers, E. Mick, J. Biederman, A. Doyle, S. Smalley, S. Loo, H. Hakonarson, J. Elia, A. Todorov, A. Miranda, F. Mulas, R. P. Ebstein, A. Rothenberger, T. Banaschewski, R. D. Oades, E. Sonuga-Barke, J. McGough, L. Nisenbaum, F. Middleton, X. Hu, S. Nelson, and P. G. C. A. Subgroup, 2010b, Meta-analysis of genome-wide association studies of attention-deficit/hyperactivity disorder: J Am Acad Child Adolesc Psychiatry, v. 49, p. 884-97.

Negishi, M., A. Saraya, S. Mochizuki, K. Helin, H. Koseki, and A. Iwama, 2010, A novel zinc finger protein Zfp277 mediates transcriptional repression of the Ink4a/arf locus through polycomb repressive complex 1: PLoS One, v. 5, p. e12373.

Newbury, D. F., E. Bonora, J. A. Lamb, S. E. Fisher, C. S. Lai, G. Baird, L. Jannoun, V. Slonims, C. M. Stott, M. J. Merricks, P. F. Bolton, A. J. Bailey, A. P. Monaco, and I. M. G. S. o. A. Consortium, 2002, FOXP2 is not a major susceptibility gene for autism or specific language impairment: Am J Hum Genet, v. 70, p. 1318-27.

Newbury, D. F., F. Mari, E. Sadighi Akha, K. D. Macdermot, R. Canitano, A. P. Monaco, J. C. Taylor, A. Renieri, S. E. Fisher, and S. J. Knight, 2013, Dual copy number variants involving 16p11 and 6q22 in a case of childhood apraxia of speech and pervasive developmental disorder: Eur J Hum Genet, v. 21, p. 361-5.

Newbury, D. F., S. Paracchini, T. S. Scerri, L. Winchester, L. Addis, A. J. Richardson, J. Walter, J. F. Stein, J. B. Talcott, and A. P. Monaco, 2011, Investigation of dyslexia and SLI risk variants in reading- and language-impaired subjects: Behav Genet, v. 41, p. 90-104.

Newbury, D. F., L. Winchester, L. Addis, S. Paracchini, L. L. Buckingham, A. Clark, W. Cohen, H. Cowie, K. Dworzynski, A. Everitt, I. M. Goodyer, E. Hennessy, A. D. Kindley, L. L. Miller, J. Nasir, A. O'Hare, D. Shaw, Z. Simkin, E. Simonoff, V. Slonims, J. Watson, J. Ragoussis, S. E. Fisher, J. R. Seckl, P. J. Helms, P. F. Bolton, A. Pickles, G. Conti-Ramsden, G. Baird, D. V. Bishop, and A. P. Monaco, 2009, CMIP and ATP2C2 modulate phonological short-term memory in language impairment: Am J Hum Genet, v. 85, p. 264-72.

Newcomer, P. L., and D. D. Hammil, 1988, Test of Language Development-2, Primary, Austin, TX.

Noor, A., A. Whibley, C. R. Marshall, P. J. Gianakopoulos, A. Piton, A. R. Carson, M. Orlic-Milacic, A. C. Lionel, D. Sato, D. Pinto, I. Drmic, C. Noakes, L. Senman, X. Zhang, R. Mo, J. Gauthier, J. Crosbie, A. T. Pagnamenta, J. Munson, A. M. Estes, A. Fiebig, A. Franke, S. Schreiber, A. F. Stewart, R. Roberts, R. McPherson, S. J. Guter, E. H. Cook, G. Dawson, G. D. Schellenberg, A. Battaglia, E. Maestrini, L. Jeng, T. Hutchison, E. Rajcan-Separovic, A. E. Chudley, S. M. Lewis, X. Liu, J. J. Holden, B. Fernandez, L. Zwaigenbaum, S. E. Bryson, W. Roberts, P. Szatmari, L. Gallagher, M. R. Stratton, J. Gecz, A. F. Brady, C. E. Schwartz, R. J. Schachar, A. P. Monaco, G. A. Rouleau, C. C. Hui, F. Lucy Raymond, S. W. Scherer, J. B. Vincent, and A. G. P. Consortium, 2010, Disruption at the PTCHD1 Locus on Xp22.11 in Autism spectrum disorder and intellectual disability: Sci Transl Med, v. 2, p. 49ra68.

O'Brien, E. K., X. Zhang, C. Nishimura, J. B. Tomblin, and J. C. Murray, 2003, Association of specific language impairment (SLI) to the region of 7q31: Am J Hum Genet, v. 72, p. 1536-43.

O'Roak, B. J., P. Deriziotis, C. Lee, L. Vives, J. J. Schwartz, S. Girirajan, E. Karakoc, A. P. Mackenzie, S. B. Ng, C. Baker, M. J. Rieder, D. A. Nickerson, R. Bernier, S. E. Fisher, J. Shendure, and E. E. Eichler, 2011, Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations: Nat Genet, v. 43, p. 585-9.

Offer, S. M., N. J. Wegner, C. Fossum, K. Wang, and R. B. Diasio, 2013, Phenotypic profiling of DPYD variations relevant to 5-fluorouracil sensitivity using real-time cellular analysis and in vitro measurement of enzyme activity: Cancer Res, v. 73, p. 1958-68.

Ozonoff, S., G. S. Young, A. Carter, D. Messinger, N. Yirmiya, L. Zwaigenbaum, S. Bryson, L. J. Carver, J. N. Constantino, K. Dobkins, T. Hutman, J. M. Iverson, R. Landa, S. J. Rogers, M. Sigman, and W. L. Stone, 2011, Recurrence risk for autism spectrum disorders: a Baby Siblings Research Consortium study: Pediatrics, v. 128, p. e488-95.

Pagnamenta, A. T., E. Bacchelli, M. V. de Jonge, G. Mirza, T. S. Scerri, F. Minopoli, A. Chiocchetti, K. U. Ludwig, P. Hoffmann, S. Paracchini, E. Lowy, D. H. Harold, J. A. Chapman, S. M. Klauck, F. Poustka, R. H. Houben, W. G. Staal, R. A. Ophoff, M. C. O'Donovan, J. Williams, M. M. Nöthen, G. Schulte-Körne, P. Deloukas, J. Ragoussis, A. J. Bailey, E. Maestrini, A. P. Monaco, and I. M. G. S. O. A. Consortium, 2010, Characterization of a family with rare deletions in CNTNAP5 and DOCK4 suggests novel risk loci for autism and dyslexia: Biol Psychiatry, v. 68, p. 320-8.

Pagnamenta, A. T., K. Wing, E. Sadighi Akha, S. J. Knight, S. Bölte, G. Schmötzer, E. Duketis, F. Poustka, S. M. Klauck, A. Poustka, J. Ragoussis, A. J. Bailey, A. P. Monaco, and I. M. G. S. o. A. Consortium, 2009, A 15q13.3 microdeletion segregating with autism: Eur J Hum Genet, v. 17, p. 687-92.

Palles, C., J. B. Cazier, K. M. Howarth, E. Domingo, A. M. Jones, P. Broderick, Z. Kemp, S. L. Spain, E. Guarino, E. Guarino Almeida, I. Salguero, A. Sherborne, D. Chubb, L. G. Carvajal-Carmona, Y. Ma, K. Kaur, S. Dobbins, E. Barclay, M. Gorman, L. Martin, M. B. Kovac, S. Humphray, A. Lucassen, C. C. Holmes, D. Bentley, P. Donnelly, J. Taylor, C. Petridis, R. Roylance, E. J. Sawyer, D. J. Kerr, S. Clark, J. Grimes, S. E. Kearsey, H. J. Thomas, G. McVean, R. S. Houlston, I. Tomlinson, C. Consortium, and W. Consortium, 2013, Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas: Nat Genet, v. 45, p. 136-44.

Pankratz, N., A. Dumitriu, K. N. Hetrick, M. Sun, J. C. Latourelle, J. B. Wilk, C. Halter, K. F. Doheny, J. F. Gusella, W. C. Nichols, R. H. Myers, T. Foroud, A. L. DeStefano, and C. o. a. M. G. L. PSG-PROGENI and GenePD Investigators, 2011, Copy number variation in familial Parkinson disease: PLoS One, v. 6, p. e20988.

Peter, B., W. H. Raskind, M. Matsushita, M. Lisowski, T. Vu, V. W. Berninger, E. M. Wijsman, and Z. Brkanac, 2011, Replication of CNTNAP2 association with nonword repetition and support for FOXP2 association with timed reading and motor activities in a dyslexia family sample: J Neurodev Disord, v. 3, p. 39-49.

Petrin, A. L., C. M. Giacheti, L. P. Maximino, D. V. Abramides, S. Zanchetta, N. F. Rossi, A. Richieri-Costa, and J. C. Murray, 2010, Identification of a microdeletion at the 7q33-q35 disrupting the CNTNAP2 gene in a Brazilian stuttering case: Am J Med Genet A, v. 152A, p. 3164-72.

Peñagarikano, O., and D. H. Geschwind, 2012, What does CNTNAP2 reveal about autism spectrum disorder?: Trends Mol Med, v. 18, p. 156-63.

Pfeiffer, M., A. Draguhn, H. Meierkord, and U. Heinemann, 1996, Effects of gamma-aminobutyric acid (GABA) agonists and GABA uptake inhibitors on pharmacosensitive and pharmacoresistant epileptiform activity in vitro: Br J Pharmacol, v. 119, p. 569-77.

Pinkel, D., R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W. L. Kuo, C. Chen, Y. Zhai, S. H. Dairkee, B. M. Ljung, J. W. Gray, and D. G. Albertson, 1998, High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays: Nat Genet, v. 20, p. 207-11.

Pinto, D., K. Darvishi, X. Shi, D. Rajan, D. Rigler, T. Fitzgerald, A. C. Lionel, B. Thiruvahindrapuram, J. R. Macdonald, R. Mills, A. Prasad, K. Noonan, S. Gribble, E. Prigmore, P. K. Donahoe, R. S. Smith, J. H. Park, M. E. Hurles, N. P. Carter, C. Lee, S. W. Scherer, and L. Feuk, 2011, Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants: Nat Biotechnol, v. 29, p. 512-20.

Pinto, D., A. T. Pagnamenta, L. Klei, R. Anney, D. Merico, R. Regan, J. Conroy, T. R. Magalhaes, C. Correia, B. S. Abrahams, J. Almeida, E. Bacchelli, G. D. Bader, A. J. Bailey, G. Baird, A. Battaglia, T. Berney, N. Bolshakova, S. Bölte, P. F. Bolton, T. Bourgeron, S. Brennan, J. Brian, S. E. Bryson,

A. R. Carson, G. Casallo, J. Casey, B. H. Chung, L. Cochrane, C. Corsello, E. L. Crawford, A. Crossett, C. Cytrynbaum, G. Dawson, M. de Jonge, R. Delorme, I. Drmic, E. Duketis, F. Duque, A. Estes, P. Farrar, B. A. Fernandez, S. E. Folstein, E. Fombonne, C. M. Freitag, J. Gilbert, C. Gillberg, J. T. Glessner, J. Goldberg, A. Green, J. Green, S. J. Guter, H. Hakonarson, E. A. Heron, M. Hill, R. Holt, J. L. Howe, G. Hughes, V. Hus, R. Igliozzi, C. Kim, S. M. Klauck, A. Kolevzon, O. Korvatska, V. Kustanovich, C. M. Lajonchere, J. A. Lamb, M. Laskawiec, M. Leboyer, A. Le Couteur, B. L. Leventhal, A. C. Lionel, X. Q. Liu, C. Lord, L. Lotspeich, S. C. Lund, E. Maestrini, W. Mahoney, C. Mantoulan, C. R. Marshall, H. McConachie, C. J. McDougle, J. McGrath, W. M. McMahon, A. Merikangas, O. Migita, N. J. Minshew, G. K. Mirza, J. Munson, S. F. Nelson, C. Noakes, A. Noor, G. Nygren, G. Oliveira, K. Papanikolaou, J. R. Parr, B. Parrini, T. Paton, A. Pickles, M. Pilorge, et al., 2010, Functional impact of global rare copy number variation in autism spectrum disorders: Nature, v. 466, p. 368-72.

Poliak, S., L. Gollan, R. Martinez, A. Custer, S. Einheber, J. L. Salzer, J. S. Trimmer, P. Shrager, and E. Peles, 1999, Caspr2, a new member of the neurexin superfamily, is localized at the juxtaparanodes of myelinated axons and associates with K+ channels: Neuron, v. 24, p. 1037-47.

Poliak, S., D. Salomon, H. Elhanany, H. Sabanay, B. Kiernan, L. Pevny, C. L. Stewart, X. Xu, S. Y. Chiu, P. Shrager, A. J. Furley, and E. Peles, 2003, Juxtaparanodal clustering of Shaker-like K+ channels in myelinated axons depends on Caspr2 and TAG-1: J Cell Biol, v. 162, p. 1149-60.

Poot, M., V. Beyer, I. Schwaab, N. Damatova, R. Van't Slot, J. Prothero, S. E. Holder, and T. Haaf, 2010, Disruption of CNTNAP2 and additional structural genome changes in a boy with speech delay and autism spectrum disorder: Neurogenetics, v. 11, p. 81-9.

Prasad, A., D. Merico, B. Thiruvahindrapuram, J. Wei, A. C. Lionel, D. Sato, J. Rickaby, C. Lu, P. Szatmari, W. Roberts, B. A. Fernandez, C. R. Marshall, E. Hatchwell, P. S. Eis, and S. W. Scherer, 2012, A discovery resource of rare copy number variations in individuals with autism spectrum disorder: G3 (Bethesda), v. 2, p. 1665-85.

Pratt, S. C., M. J. Daly, and L. Kruglyak, 2000, Exact multipoint quantitative-trait linkage analysis in pedigrees by variance components: Am J Hum Genet, v. 66, p. 1153-7.

Priebe, L., F. A. Degenhardt, S. Herms, B. Haenisch, M. Mattheisen, V. Nieratschker, M. Weingarten, S. Witt, R. Breuer, T. Paul, M. Alblas, S. Moebus, M. Lathrop, M. Leboyer, S. Schreiber, M. Grigoroiu-Serbanescu, W. Maier, P. Propping, M. Rietschel, M. M. Nöthen, S. Cichon, and T. W. Mühleisen, 2012, Genome-wide survey implicates the influence of copy number variants (CNVs) in the development of early-onset bipolar disorder: Mol Psychiatry, v. 17, p. 421-32.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham, 2007, PLINK: a tool set for whole-genome association and population-based linkage analyses: Am J Hum Genet, v. 81, p. 559-75.

Riches, N. G., T. Loucas, G. Baird, T. Charman, and E. Simonoff, 2010, Sentence repetition in adolescents with specific language impairments and autism: an investigation of complex syntax: Int J Lang Commun Disord, v. 45, p. 47-60.

Ripke, S., C. O'Dushlaine, K. Chambert, J. L. Moran, A. K. Kähler, S. Akterin, S. E. Bergen, A. L. Collins, J. J. Crowley, M. Fromer, Y. Kim, S. H. Lee, P. K. Magnusson, N. Sanchez, E. A. Stahl, S. Williams, N. R. Wray, K. Xia, F. Bettella, A. D. Borglum, B. K. Bulik-Sullivan, P. Cormican, N. Craddock, C. de Leeuw, N. Durmishi, M. Gill, V. Golimbet, M. L. Hamshere, P. Holmans, D. M. Hougaard, K. S. Kendler, K. Lin, D. W. Morris, O. Mors, P. B. Mortensen, B. M. Neale, F. A. O'Neill, M. J. Owen, M. P. Milovancevic, D. Posthuma, J. Powell, A. L. Richards, B. P. Riley, D. Ruderfer, D. Rujescu, E. Sigurdsson, T. Silagadze, A. B. Smit, H. Stefansson, S. Steinberg, J. Suvisaari, S. Tosato, M. Verhage, J. T. Walters, D. F. Levinson, P. V. Gejman, C. Laurent, B. J. Mowry, M. C. O'Donovan, A. E. Pulver, S. G. Schwab, D. B. Wildenauer, F. Dudbridge, J. Shi, M. Albus, M. Alexander, D. Campion, D. Cohen, D. Dikeos, J. Duan, P. Eichhammer, S. Godard, M. Hansen, F. B. Lerer, K. Y. Liang, W. Maier, J. Mallet, D. A. Nertney, G. Nestadt, N. Norton, G. N. Papadimitriou, R. Ribble, A. R. Sanders, J. M. Silverman, D. Walsh, N. M. Williams, B. Wormley, M. J. Arranz, S. Bakker, S. Bender, E. Bramon, D. Collier, B. Crespo-Facorro, J. Hall, C. Iyegbe, A. Jablensky, R. S. Kahn, L. Kalaydjieva, S. Lawrie, C. M. Lewis, et al., 2013, Genome-wide association analysis identifies 13 new risk loci for schizophrenia: Nat Genet, v. 45, p. 1150-9.

Risch, N., and K. Merikangas, 1996, The future of genetic studies of complex human diseases: Science, v. 273, p. 1516-7.

Ronald, A., and R. A. Hoekstra, 2011, Autism spectrum disorders and autistic traits: a decade of new twin studies: Am J Med Genet B Neuropsychiatr Genet, v. 156B, p. 255-74.

Rubnitz, J., and S. Subramani, 1984, The minimum amount of homology required for homologous recombination in mammalian cells: Mol Cell Biol, v. 4, p. 2253-8.

Rust, J., S. Golombok, and G. Trickey, 1993, Wechsler objective reading dimensions, Psychological Corporation, Sidcup.

Sanders, S. J., A. G. Ercan-Sencicek, V. Hus, R. Luo, M. T. Murtha, D. Moreno-De-Luca, S. H. Chu, M. P. Moreau, A. R. Gupta, S. A. Thomson, C. E. Mason, K. Bilguvar, P. B. Celestino-Soper, M. Choi, E. L. Crawford, L. Davis, N. R. Wright, R. M. Dhodapkar, M. DiCola, N. M. DiLullo, T. V. Fernandez, V. Fielding-Singh, D. O. Fishman, S. Frahm, R. Garagaloyan, G. S. Goh, S. Kammela, L. Klei, J. K. Lowe, S. C. Lund, A. D. McGrew, K. A. Meyer, W. J. Moffat, J. D. Murdoch, B. J. O'Roak, G. T. Ober, R. S. Pottenger, M. J. Raubeson, Y. Song, Q. Wang, B. L. Yaspan, T. W. Yu, I. R. Yurkiewicz, A. L. Beaudet, R. M. Cantor, M. Curland, D. E. Grice, M. Günel, R. P. Lifton, S. M. Mane, D. M. Martin, C. A. Shaw, M. Sheldon, J. A. Tischfield, C. A. Walsh, E. M. Morrow, D. H. Ledbetter, E. Fombonne, C. Lord, C. L. Martin, A. I. Brooks, J. S. Sutcliffe, E. H. Cook, D. Geschwind, K. Roeder, B. Devlin, and M. W. State, 2011, Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism: Neuron, v. 70, p. 863-85.

Sanders, S. J., M. T. Murtha, A. R. Gupta, J. D. Murdoch, M. J. Raubeson, A. J. Willsey, A. G. Ercan-Sencicek, N. M. DiLullo, N. N. Parikshak, J. L. Stein, M. F. Walker, G. T. Ober, N. A. Teran, Y. Song, P. El-Fishawy, R. C. Murtha, M. Choi, J. D. Overton, R. D. Bjornson, N. J. Carriero, K. A. Meyer, K. Bilguvar, S. M. Mane, N. Sestan, R. P. Lifton, M. Günel, K. Roeder, D. H. Geschwind, B. Devlin, and M. W. State, 2012, De novo mutations revealed by whole-exome sequencing are strongly associated with autism: Nature, v. 485, p. 237-41.

Savva-Bordalo, J., J. Ramalho-Carvalho, M. Pinheiro, V. L. Costa, A. Rodrigues, P. C. Dias, I. Veiga, M. Machado, M. R. Teixeira, R. Henrique, and C. Jerónimo, 2010, Promoter methylation and large intragenic rearrangements of DPYD are not implicated in severe toxicity to 5-fluorouracil-based chemotherapy in gastrointestinal cancer patients: BMC Cancer, v. 10, p. 470.

Scerri, T. S., A. P. Morris, L. L. Buckingham, D. F. Newbury, L. L. Miller, A. P. Monaco, D. V. Bishop, and S. Paracchini, 2011, DCDC2, KIAA0319 and CMIP are associated with reading-related traits: Biol Psychiatry, v. 70, p. 237-45.

Schaaf, C. P., and H. Y. Zoghbi, 2011, Solving the autism puzzle a few pieces at a time: Neuron, v. 70, p. 806-8.

Schellenberg, G. D., G. Dawson, Y. J. Sung, A. Estes, J. Munson, E. Rosenthal, J. Rothstein, P. Flodman, M. Smith, H. Coon, L. Leong, C. E. Yu, C. Stodgell, P. M. Rodier, M. A. Spence, N. Minshew, W. M. McMahon, and E. M. Wijsman, 2006, Evidence for multiple loci from a genome scan of autism kindreds: Mol Psychiatry, v. 11, p. 1049-60, 979.

Schizophrenia_Psychiatric_GWAS_Consortium, 2011, Genome-wide association study identifies five new schizophrenia loci: Nat Genet, v. 43, p. 969-76.

Schouten, J. P., C. J. McElgunn, R. Waaijer, D. Zwijnenburg, F. Diepvens, and G. Pals, 2002, Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification: Nucleic Acids Res, v. 30, p. e57.

Schwab, M., U. M. Zanger, C. Marx, E. Schaeffeler, K. Klein, J. Dippon, R. Kerb, J. Blievernicht, J. Fischer, U. Hofmann, C. Bokemeyer, M. Eichelbaum, and G.-F. T. S. Group, 2008, Role of genetic and nongenetic factors for fluorouracil treatment-related severe toxicity: a prospective clinical trial by the German 5-FU Toxicity Study Group: J Clin Oncol, v. 26, p. 2131-8.

Sebat, J., B. Lakshmi, D. Malhotra, J. Troge, C. Lese-Martin, T. Walsh, B. Yamrom, S. Yoon, A. Krasnitz, J. Kendall, A. Leotta, D. Pai, R. Zhang, Y. H. Lee, J. Hicks, S. J. Spence, A. T. Lee, K. Puura, T. Lehtimäki, D. Ledbetter, P. K. Gregersen, J. Bregman, J. S. Sutcliffe, V. Jobanputra, W. Chung, D. Warburton, M. C. King, D. Skuse, D. H. Geschwind, T. C. Gilliam, K. Ye, and M. Wigler, 2007, Strong association of de novo copy number mutations with autism: Science, v. 316, p. 445-9.

Sebat, J., B. Lakshmi, J. Troge, J. Alexander, J. Young, P. Lundin, S. Månér, H. Massa, M. Walker, M. Chi, N. Navin, R. Lucito, J. Healy, J. Hicks, K. Ye, A. Reiner, T. C. Gilliam, B. Trask, N. Patterson, A. Zetterberg, and M. Wigler, 2004, Large-scale copy number polymorphism in the human genome: Science, v. 305, p. 525-8.

Seck, K., S. Riemer, R. Kates, T. Ullrich, V. Lutz, N. Harbeck, M. Schmitt, M. Kiechle, R. Diasio, and E. Gross, 2005, Analysis of the DPYD gene implicated in 5-fluorouracil catabolism in a cohort of Caucasian individuals: Clin Cancer Res, v. 11, p. 5886-92.

Sekine, Y., R. Hatanaka, T. Watanabe, N. Sono, S. Iemura, T. Natsume, E. Kuranaga, M. Miura, K. Takeda, and H. Ichijo, 2012, The Kelch repeat protein KLHDC10 regulates oxidative stress-induced ASK1 activation by suppressing PP5: Mol Cell, v. 48, p. 692-704.

Semel, E., E. Wiig, and W. Secord, 2004, Clinical evaluation of language fundamentals, fourth edition-Screening test (CELF-4screening test), Toronto, Canada, The Psychological Corporation/A Harcourt Assessment Company.

Semel, E. M., E. H. Wiig, and W. Secord, 1992, Clinical Evaluation of Language Fundamentals–Revised., Psychological Corporation,San Antonio.

Shaikh, T. H., X. Gai, J. C. Perin, J. T. Glessner, H. Xie, K. Murphy, R. O'Hara, T. Casalunovo, L. K. Conlin, M. D'Arcy, E. C. Frackelton, E. A. Geiger, C. Haldeman-Englert, M. Imielinski, C. E. Kim, L. Medne, K. Annaiah, J. P. Bradfield, E. Dabaghyan, A. Eckert, C. C. Onyiah, S. Ostapenko, F. G. Otieno, E. Santa, J. L. Shaner, R. Skraban, R. M. Smith, J. Elia, E. Goldmuntz, N. B. Spinner, E. H. Zackai, R. M. Chiavacci, R. Grundmeier, E. F. Rappaport, S. F. Grant, P. S. White, and H. Hakonarson, 2009, High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications: Genome Res, v. 19, p. 1682-90.

Shao, Y., K. L. Raiford, C. M. Wolpert, H. A. Cope, S. A. Ravan, A. A. Ashley-Koch, R. K. Abramson, H. H. Wright, R. G. DeLong, J. R. Gilbert, M. L. Cuccaro, and M. A. Pericak-Vance, 2002, Phenotypic homogeneity provides increased support for linkage on chromosome 2 in autistic disorder: Am J Hum Genet, v. 70, p. 1058-61.

Sharp, A. J., H. C. Mefford, K. Li, C. Baker, C. Skinner, R. E. Stevenson, R. J. Schroer, F. Novara, M. De Gregori, R. Ciccone, A. Broomer, I. Casuga, Y. Wang, C. Xiao, C. Barbacioru, G. Gimelli, B. D. Bernardina, C. Torniero, R. Giorda, R. Regan, V. Murday, S. Mansour, M. Fichera, L. Castiglia, P. Failla, M. Ventura, Z. Jiang, G. M. Cooper, S. J. Knight, C. Romano, O. Zuffardi, C. Chen, C. E. Schwartz, and E. E. Eichler, 2008, A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures: Nat Genet, v. 40, p. 322-8.

Shestopal, S. A., M. R. Johnson, and R. B. Diasio, 2000, Molecular cloning and characterization of the human dihydropyrimidine dehydrogenase promoter: Biochim Biophys Acta, v. 1494, p. 162-9.

Shinawi, M., C. P. Schaaf, S. S. Bhatt, Z. Xia, A. Patel, S. W. Cheung, B. Lanpher, S. Nagl, H. S. Herding, C. Nevinny-Stickel, L. L. Immken, G. S. Patel, J. R. German, A. L. Beaudet, and P. Stankiewicz, 2009, A small recurrent deletion within 15q13.3 is associated with a range of neurodevelopmental phenotypes: Nat Genet, v. 41, p. 1269-71.

Shriberg, L. D., K. J. Ballard, J. B. Tomblin, J. R. Duffy, K. H. Odell, and C. A. Williams, 2006, Speech, prosody, and voice characteristics of a mother and daughter with a 7;13 translocation affecting FOXP2: J Speech Lang Hear Res, v. 49, p. 500-25.

Siepel, A., G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, and D. Haussler, 2005, Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes: Genome Res, v. 15, p. 1034-50.

Simon-Sanchez, J., S. Scholz, H. C. Fung, M. Matarin, D. Hernandez, J. R. Gibbs, A. Britton, F. W. de Vrieze, E. Peckham, K. Gwinn-Hardy, A. Crawley, J. C. Keen, J. Nash, D. Borgaonkar, J. Hardy, and A. Singleton, 2007, Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals: Hum Mol Genet, v. 16, p. 1-14.

Simpson, J. T., K. Wong, S. D. Jackman, J. E. Schein, S. J. Jones, and I. Birol, 2009, ABySS: a parallel assembler for short read sequence data: Genome Res, v. 19, p. 1117-23.

SLIC, 2002, A genomewide scan identifies two novel loci involved in specific language impairment: Am J Hum Genet, v. 70, p. 384-98.

SLIC, 2004, Highly significant linkage to the SLI1 locus in an expanded sample of individuals affected by specific language impairment: Am J Hum Genet, v. 74, p. 1225-38.

Spielman, R. S., R. E. McGinnis, and W. J. Ewens, 1993, Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM): Am J Hum Genet, v. 52, p. 506-16.

Spiteri, E., G. Konopka, G. Coppola, J. Bomar, M. Oldham, J. Ou, S. C. Vernes, S. E. Fisher, B. Ren, and D. H. Geschwind, 2007, Identification of the transcriptional targets of FOXP2, a gene linked to speech and language, in developing human brain: Am J Hum Genet, v. 81, p. 1144-57.

Stankiewicz, P., and J. R. Lupski, 2002, Genome architecture, rearrangements and genomic disorders: Trends Genet, v. 18, p. 74-82.

Stefansson, H., D. Rujescu, S. Cichon, O. P. Pietiläinen, A. Ingason, S. Steinberg, R. Fossdal, E. Sigurdsson, T. Sigmundsson, J. E. Buizer-Voskamp, T. Hansen, K. D. Jakobsen, P. Muglia, C. Francks, P. M. Matthews, A. Gylfason, B. V. Halldorsson, D. Gudbjartsson, T. E. Thorgeirsson, A. Sigurdsson, A. Jonasdottir, A. Bjornsson, S. Mattiasdottir, T. Blondal, M. Haraldsson, B. B. Magnusdottir, I. Giegling, H. J. Möller, A. Hartmann, K. V. Shianna, D. Ge, A. C. Need, C. Crombie, G. Fraser, N. Walker, J. Lonnqvist, J. Suvisaari, A. Tuulio-Henriksson, T. Paunio, T. Toulopoulou, E. Bramon, M. Di Forti, R. Murray, M. Ruggeri, E. Vassos, S. Tosato, M. Walshe, T. Li, C. Vasilescu, T. W. Mühleisen, A. G. Wang, H. Ullum, S. Djurovic, I. Melle, J. Olesen, L. A. Kiemeney, B. Franke, C. Sabatti, N. B. Freimer, J. R. Gulcher, U. Thorsteinsdottir, A. Kong, O. A. Andreassen, R. A. Ophoff, A. Georgi, M. Rietschel, T. Werge, H. Petursson, D. B. Goldstein, M. M. Nöthen, L. Peltonen, D. A. Collier, D. St Clair, K. Stefansson, and GROUP, 2008, Large recurrent microdeletions associated with schizophrenia: Nature, v. 455, p. 232-6.

Stergiakouli, E., M. Hamshere, P. Holmans, K. Langley, I. Zaharieva, Z. Hawi, L. Kent, M. Gill, N. Williams, M. J. Owen, M. O'Donovan, A. Thapar, d. Genetics, and P. G. Consortium, 2012, Investigating the contribution of common genetic variants to the risk and pathogenesis of ADHD: Am J Psychiatry, v. 169, p. 186-94.

Strauss, K. A., E. G. Puffenberger, M. J. Huentelman, S. Gottlieb, S. E. Dobrin, J. M. Parod, D. A. Stephan, and D. H. Morton, 2006, Recessive symptomatic focal epilepsy and mutant contactin-associated protein-like 2: N Engl J Med, v. 354, p. 1370-7.

Stromswold, K., 1998, Genetics of spoken language disorders: Hum Biol, v. 70, p. 297-324.

Szafranski, P., C. P. Schaaf, R. E. Person, I. B. Gibson, Z. Xia, S. Mahadevan, J. Wiszniewska, C. A. Bacino, S. Lalani, L. Potocki, S. H. Kang, A. Patel, S. W. Cheung, F. J. Probst, B. H. Graham, M. Shinawi, A. L. Beaudet, and P. Stankiewicz, 2010, Structures and molecular mechanisms for common 15q13.3 microduplications involving CHRNA7: benign or pathological?: Hum Mutat, v. 31, p. 840-50.

Szatmari, P., A. D. Paterson, L. Zwaigenbaum, W. Roberts, J. Brian, X. Q. Liu, J. B. Vincent, J. L. Skaug, A. P. Thompson, L. Senman, L. Feuk, C. Qian, S. E. Bryson, M. B. Jones, C. R. Marshall, S. W. Scherer, V. J. Vieland, C. Bartlett, L. V. Mangin, R. Goedken, A. Segre, M. A. Pericak-Vance, M. L. Cuccaro, J. R. Gilbert, H. H. Wright, R. K. Abramson, C. Betancur, T. Bourgeron, C. Gillberg, M. Leboyer, J. D. Buxbaum, K. L. Davis, E. Hollander, J. M. Silverman, J. Hallmayer, L. Lotspeich, J. S. Sutcliffe, J. L. Haines, S. E. Folstein, J. Piven, T. H. Wassink, V. Sheffield, D. H. Geschwind, M. Bucan, W. T. Brown, R. M. Cantor, J. N. Constantino, T. C. Gilliam, M. Herbert, C. Lajonchere, D. H. Ledbetter, C. Lese-Martin, J. Miller, S. Nelson, C. A. Samango-Sprouse, S. Spence, M. State, R. E. Tanzi, H. Coon, G. Dawson, B. Devlin, A. Estes, P. Flodman, L. Klei, W. M. McMahon, N. Minshew, J. Munson, E. Korvatska, P. M. Rodier, G. D. Schellenberg, M. Smith, M. A. Spence, C. Stodgell, P. G. Tepper, E. M. Wijsman, C. E. Yu, B. Rogé, C. Mantoulan, K. Wittemeyer, A. Poustka, B. Felder, S. M. Klauck, C. Schuster, F. Poustka, S. Bölte, S. Feineis-Matthews, E. Herbrecht, G. Schmötzer, J. Tsiantis, K. Papanikolaou, E. Maestrini, E. Bacchelli, F. Blasi, S. Carone, C. Toma, H. Van Engeland, M. de Jonge, C. Kemner, F. Koop, M. Langemeijer, et al., 2007, Mapping autism risk loci using genetic linkage and chromosomal rearrangements: Nat Genet, v. 39, p. 319-28.

Südhof, T. C., 2008, Neuroligins and neurexins link synaptic function to cognitive disease: Nature, v. 455, p. 903-11.

Thorn, C. F., S. Marsh, M. W. Carrillo, H. L. McLeod, T. E. Klein, and R. B. Altman, 2011, PharmGKB summary: fluoropyrimidine pathways: Pharmacogenet Genomics, v. 21, p. 237-42.

Tiedje, K. E., K. Stevens, S. Barnes, and D. F. Weaver, 2010, Beta-alanine as a small molecule neurotransmitter: Neurochem Int, v. 57, p. 177-88.

Tomblin, J. B., and P. R. Buckwalter, 1998, Heritability of poor language achievement among twins: J Speech Lang Hear Res, v. 41, p. 188-99.

Tomblin, J. B., L. L. Hafeman, and M. O'Brien, 2003, Autism and autism risk in siblings of children with specific language impairment: Int J Lang Commun Disord, v. 38, p. 235-50.

Tomblin, J. B., N. L. Records, and X. Zhang, 1996, A system for the diagnosis of specific language impairment in kindergarten children: J Speech Hear Res, v. 39, p. 1284-94.

Trikalinos, T. A., A. Karvouni, E. Zintzaras, T. Ylisaukko-oja, L. Peltonen, I. Järvelä, and J. P. Ioannidis, 2006, A heterogeneity-based genome search meta-analysis for autism-spectrum disorders: Mol Psychiatry, v. 11, p. 29-36.

Tuna, M., S. Knuutila, and G. B. Mills, 2009, Uniparental disomy in cancer: Trends Mol Med, v. 15, p. 120-8.

Tuschl, K., P. B. Mills, and P. T. Clayton, 2013, Manganese and the brain: Int Rev Neurobiol, v. 110, p. 277-312.

Tuzun, E., A. J. Sharp, J. A. Bailey, R. Kaul, V. A. Morrison, L. M. Pertz, E. Haugen, H. Hayden, D. Albertson, D. Pinkel, M. V. Olson, and E. E. Eichler, 2005, Fine-scale structural variation of the human genome: Nat Genet, v. 37, p. 727-32.

Ueda, S., M. Negishi, and H. Katoh, 2013, Rac GEF Dock4 interacts with cortactin to regulate dendritic spine formation: Mol Biol Cell, v. 24, p. 1602-13.

van Bon, B. W., H. C. Mefford, B. Menten, D. A. Koolen, A. J. Sharp, W. M. Nillesen, J. W. Innis, T. J. de Ravel, C. L. Mercer, M. Fichera, H. Stewart, L. E. Connell, K. Ounap, K. Lachlan, B. Castle, N. Van der Aa, C. van Ravenswaaij, M. A. Nobrega, C. Serra-Juhé, I. Simonic, N. de Leeuw, R. Pfundt, E. M. Bongers, C. Baker, P. Finnemore, S. Huang, V. K. Maloney, J. A. Crolla, M. van Kalmthout, M. Elia, G. Vandeweyer, J. P. Fryns, S. Janssens, N. Foulds, S. Reitano, K. Smith, S. Parkel, B. Loeys, C. G. Woods, A. Oostra, F. Speleman, A. C. Pereira, A. Kurg, L. Willatt, S. J. Knight, J. R. Vermeesch, C. Romano, J. C. Barber, G. Mortier, L. A. Pérez-Jurado, F. Kooy, H. G. Brunner, E. E. Eichler, T. Kleefstra, and B. B. de Vries, 2009, Further delineation of the 15q13 microdeletion and duplication syndromes: a clinical spectrum varying from non-pathogenic to a severe outcome: J Med Genet, v. 46, p. 511-23.

Van der Aa, N., G. Vandeweyer, E. Reyniers, S. Kenis, L. Dom, G. Mortier, L. Rooms, and R. F. Kooy, 2012, Haploinsufficiency of CMIP in a girl with autism spectrum disorder and developmental delay due to a de novo deletion on chromosome 16q23.2: Autism Res, v. 5, p. 277-81.

van der Zwaag, B., W. G. Staal, R. Hochstenbach, M. Poot, H. A. Spierenburg, M. V. de Jonge, N. E. Verbeek, R. van 't Slot, M. A. van Es, F. J. Staal, C. M. Freitag, J. E. Buizer-Voskamp, M. R. Nelen, L. H. van den Berg, H. K. van Amstel, H. van Engeland, and J. P. Burbach, 2010, A co-segregating microduplication of chromosome 15q11.2 pinpoints two risk genes for autism spectrum disorder: Am J Med Genet B Neuropsychiatr Genet, v. 153B, p. 960-6.

van Gennip, A., V. B.-B. E, and W. S, 1981, Liquid chromatography of urinary pyrimidines for the evaluation of primary and secondary abnormalities of pyrmidine metabolism, New York, Marcel Dekker, p. 285-296.

van Kuilenburg, A. B., R. A. De Abreu, and A. H. van Gennip, 2003, Pharmacogenetic and clinical aspects of dihydropyrimidine dehydrogenase deficiency: Ann Clin Biochem, v. 40, p. 41-5.

van Kuilenburg, A. B., D. Dobritzsch, R. Meinsma, J. Haasjes, H. R. Waterham, M. J. Nowaczyk, G. D. Maropoulos, G. Hein, H. Kalhoff, J. M. Kirk, H. Baaske, A. Aukett, J. A. Duley, K. P. Ward, Y. Lindqvist, and A. H. van Gennip, 2002, Novel disease-causing mutations in the dihydropyrimidine dehydrogenase gene interpreted by analysis of the three-dimensional protein structure: Biochem J, v. 364, p. 157-63.

van Kuilenburg, A. B., J. Meijer, A. N. Mul, R. Meinsma, V. Schmid, D. Dobritzsch, R. C. Hennekam, M. M. Mannens, M. Kiechle, M. C. Etienne-Grimaldi, H. J. Klümpen, J. G. Maring, V. A. Derleyn, E. Maartense, G. Milano, R. Vijzelaar, and E. Gross, 2010, Intragenic deletions and a deep intronic mutation affecting pre-mRNA splicing in the dihydropyrimidine dehydrogenase gene as novel mechanisms causing 5-fluorouracil toxicity: Hum Genet, v. 128, p. 529-38.

van Kuilenburg, A. B., E. W. Muller, J. Haasjes, R. Meinsma, L. Zoetekouw, H. R. Waterham, F. Baas, D. J. Richel, and A. H. van Gennip, 2001, Lethal outcome of a patient with a complete dihydropyrimidine dehydrogenase (DPD) deficiency after administration of 5-fluorouracil: frequency of the common IVS14+1G>A mutation causing DPD deficiency: Clin Cancer Res, v. 7, p. 1149-53.

Van Kuilenburg, A. B., A. E. Stroomer, H. Van Lenthe, N. G. Abeling, and A. H. Van Gennip, 2004, New insights in dihydropyrimidine dehydrogenase deficiency: a pivotal role for beta-aminoisobutyric acid?: Biochem J, v. 379, p. 119-24.

Van Kuilenburg, A. B., P. Vreken, N. G. Abeling, H. D. Bakker, R. Meinsma, H. Van Lenthe, R. A. De Abreu, J. A. Smeitink, H. Kayserili, M. Y. Apak, E. Christensen, I. Holopainen, K. Pulkki, D. Riva,

G. Botteon, E. Holme, M. Tulinius, W. J. Kleijer, F. A. Beemer, M. Duran, K. E. Niezen-Koning, G. P. Smit, C. Jakobs, L. M. Smit, and A. H. Van Gennip, 1999, Genotype and phenotype in patients with dihydropyrimidine dehydrogenase deficiency: Hum Genet, v. 104, p. 1-9.

Veenstra-VanderWeele, J., and E. H. Cook, 2004, Molecular genetics of autism spectrum disorder: Mol Psychiatry, v. 9, p. 819-32.

Verkerk, A. J., C. A. Mathews, M. Joosse, B. H. Eussen, P. Heutink, B. A. Oostra, and T. S. A. I. C. f. Genetics, 2003, CNTNAP2 is disrupted in a family with Gilles de la Tourette syndrome and obsessive compulsive disorder: Genomics, v. 82, p. 1-9.

Vernes, S. C., D. F. Newbury, B. S. Abrahams, L. Winchester, J. Nicod, M. Groszer, M. Alarcón, P. L. Oliver, K. E. Davies, D. H. Geschwind, A. P. Monaco, and S. E. Fisher, 2008, A functional genetic link between distinct developmental language disorders: N Engl J Med, v. 359, p. 2337-45.

Vernes, S. C., P. L. Oliver, E. Spiteri, H. E. Lockstone, R. Puliyadi, J. M. Taylor, J. Ho, C. Mombereau, A. Brewer, E. Lowy, J. Nicod, M. Groszer, D. Baban, N. Sahgal, J. B. Cazier, J. Ragoussis, K. E. Davies, D. H. Geschwind, and S. E. Fisher, 2011, Foxp2 regulates gene networks implicated in neurite outgrowth in the developing brain: PLoS Genet, v. 7, p. e1002145.

Vernes, S. C., E. Spiteri, J. Nicod, M. Groszer, J. M. Taylor, K. E. Davies, D. H. Geschwind, and S. E. Fisher, 2007, High-throughput analysis of promoter occupancy reveals direct neural targets of FOXP2, a gene mutated in speech and language disorders: Am J Hum Genet, v. 81, p. 1232-50.

Veyrac, A., A. Besnard, J. Caboche, S. Davis, and S. Laroche, 2014, The transcription factor zif268/egr1, brain plasticity, and memory: Prog Mol Biol Transl Sci, v. 122, p. 89-129.

Vieland, V. J., 1998, Bayesian linkage analysis, or: how I learned to stop worrying and love the posterior probability of linkage: Am J Hum Genet, v. 63, p. 947-54.

Vieland, V. J., K. Wang, and J. Huang, 2001, Power to detect linkage based on multiple sets of data in the presence of locus heterogeneity: comparative evaluation of model-based linkage methods for affected sib pair data: Hum Hered, v. 51, p. 199-208.

Villanueva, P., Z. de Barbieri, H. M. Palomino, and H. Palomino, 2008, [High prevalence of specific language impairment in Robinson Crusoe Island. A possible founder effect]: Rev Med Chil, v. 136, p. 186-92.

Villanueva, P., D. F. Newbury, L. Jara, Z. De Barbieri, G. Mirza, H. M. Palomino, M. A. Fernández, J. B. Cazier, A. P. Monaco, and H. Palomino, 2011, Genome-wide analysis of genetic susceptibility to language impairment in an isolated Chilean population: Eur J Hum Genet, v. 19, p. 687-95.

Vissers, L. E., S. S. Bhatt, I. M. Janssen, Z. Xia, S. R. Lalani, R. Pfundt, K. Derwinska, B. B. de Vries, C. Gilissen, A. Hoischen, M. Nesteruk, B. Wisniowiecka-Kowalnik, M. Smyk, H. G. Brunner, S. W. Cheung, A. G. van Kessel, J. A. Veltman, and P. Stankiewicz, 2009, Rare pathogenic microdeletions and tandem duplications are microhomology-mediated and stimulated by local genomic architecture: Hum Mol Genet, v. 18, p. 3579-93.

Vreken, P., A. B. Van Kuilenburg, R. Meinsma, G. P. Smit, H. D. Bakker, R. A. De Abreu, and A. H. van Gennip, 1996, A point mutation in an invariant splice donor site leads to exon skipping in two unrelated Dutch patients with dihydropyrimidine dehydrogenase deficiency: J Inherit Metab Dis, v. 19, p. 645-54.

Vreken, P., A. B. Van Kuilenburg, R. Meinsma, and A. H. van Gennip, 1997a, Dihydropyrimidine dehydrogenase (DPD) deficiency: identification and expression of missense mutations C29R, R886H and R235W: Hum Genet, v. 101, p. 333-8.

Vreken, P., A. B. Van Kuilenburg, R. Meinsma, and A. H. van Gennip, 1997b, Identification of novel point mutations in the dihydropyrimidine dehydrogenase gene: J Inherit Metab Dis, v. 20, p. 335-8.

Waldman, A. S., and R. M. Liskay, 1988, Dependence of intrachromosomal recombination in mammalian cells on uninterrupted homology: Mol Cell Biol, v. 8, p. 5350-7.

Walsh, T., J. M. McClellan, S. E. McCarthy, A. M. Addington, S. B. Pierce, G. M. Cooper, A. S. Nord, M. Kusenda, D. Malhotra, A. Bhandari, S. M. Stray, C. F. Rippey, P. Roccanova, V. Makarov, B. Lakshmi, R. L. Findling, L. Sikich, T. Stromberg, B. Merriman, N. Gogtay, P. Butler, K. Eckstrand, L. Noory, P. Gochman, R. Long, Z. Chen, S. Davis, C. Baker, E. E. Eichler, P. S. Meltzer, S. F. Nelson, A. B. Singleton, M. K. Lee, J. L. Rapoport, M. C. King, and J. Sebat, 2008, Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia: Science, v. 320, p. 539-43.

Wang, K., M. Li, D. Hadley, R. Liu, J. Glessner, S. F. Grant, H. Hakonarson, and M. Bucan, 2007, PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data: Genome Res, v. 17, p. 1665-74.

Wang, K., H. Zhang, D. Ma, M. Bucan, J. T. Glessner, B. S. Abrahams, D. Salyakina, M. Imielinski, J. P. Bradfield, P. M. Sleiman, C. E. Kim, C. Hou, E. Frackelton, R. Chiavacci, N. Takahashi, T. Sakurai, E. Rappaport, C. M. Lajonchere, J. Munson, A. Estes, O. Korvatska, J. Piven, L. I. Sonnenblick, A. I. Alvarez Retuerto, E. I. Herman, H. Dong, T. Hutman, M. Sigman, S. Ozonoff, A. Klin, T. Owley, J. A. Sweeney, C. W. Brune, R. M. Cantor, R. Bernier, J. R. Gilbert, M. L. Cuccaro, W. M. McMahon, J. Miller, M. W. State, T. H. Wassink, H. Coon, S. E. Levy, R. T. Schultz, J. I. Nurnberger, J. L. Haines, J. S. Sutcliffe, E. H. Cook, N. J. Minshew, J. D. Buxbaum, G. Dawson, S. F. Grant, D. H. Geschwind, M. A. Pericak-Vance, G. D. Schellenberg, and H. Hakonarson, 2009, Common genetic variants on 5p14.1 associate with autism spectrum disorders: Nature, v. 459, p. 528-33.

Wechsler, D., 1992, Wechsler Intelligence Scale for Children–Third UK Edition. , Psychological Corporation, London.

Weiss, L. A., D. E. Arking, M. J. Daly, A. Chakravarti, and G. D. P. o. J. H. a. t. A. Consortium, 2009, A genome-wide linkage and association scan reveals novel loci for autism: Nature, v. 461, p. 802-8.

Weiss, L. A., Y. Shen, J. M. Korn, D. E. Arking, D. T. Miller, R. Fossdal, E. Saemundsen, H. Stefansson, M. A. Ferreira, T. Green, O. S. Platt, D. M. Ruderfer, C. A. Walsh, D. Altshuler, A. Chakravarti, R. E. Tanzi, K. Stefansson, S. L. Santangelo, J. F. Gusella, P. Sklar, B. L. Wu, M. J. Daly, and A. Consortium, 2008, Association between microdeletion and microduplication at 16p11.2 and autism: N Engl J Med, v. 358, p. 667-75.

Weterings, E., and D. C. van Gent, 2004, The mechanism of non-homologous end-joining: a synopsis of synapsis: DNA Repair (Amst), v. 3, p. 1425-35.

White, J. A., B. C. McKinney, M. C. John, P. A. Powers, T. J. Kamp, and G. G. Murphy, 2008, Conditional forebrain deletion of the L-type calcium channel Ca V 1.2 disrupts remote spatial memories in mice: Learn Mem, v. 15, p. 1-5.

Whitehouse, A. J., J. G. Barry, and D. V. Bishop, 2008, Further defining the language impairment of autism: is there a specific language impairment subtype?: J Commun Disord, v. 41, p. 319-36.

Willemsen, M. H., A. Vallès, L. A. Kirkels, M. Mastebroek, N. Olde Loohuis, A. Kos, W. M. Wissink-Lindhout, A. P. de Brouwer, W. M. Nillesen, R. Pfundt, M. Holder-Espinasse, L. Vallée, J. Andrieux, M. C. Coppens-Hofman, H. Rensen, B. C. Hamel, H. van Bokhoven, A. Aschrafi, and T. Kleefstra, 2011, Chromosome 1p21.3 microdeletions comprising DPYD and MIR137 are associated with intellectual disability: J Med Genet, v. 48, p. 810-8.

Williams, E., K. Thomas, H. Sidebotham, and A. Emond, 2008, Prevalence and characteristics of autistic spectrum disorders in the ALSPAC cohort: Dev Med Child Neurol, v. 50, p. 672-7.

Xu, B., I. Ionita-Laza, J. L. Roos, B. Boone, S. Woodrick, Y. Sun, S. Levy, J. A. Gogos, and M. Karayiorgou, 2012, De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia: Nat Genet, v. 44, p. 1365-9.

Xu, B., J. L. Roos, S. Levy, E. J. van Rensburg, J. A. Gogos, and M. Karayiorgou, 2008, Strong association of de novo copy number mutations with sporadic schizophrenia: Nat Genet, v. 40, p. 880-5.

Ye, K., M. H. Schulz, Q. Long, R. Apweiler, and Z. Ning, 2009, Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads: Bioinformatics, v. 25, p. 2865-71.

Yoon, S., Z. Xuan, V. Makarov, K. Ye, and J. Sebat, 2009, Sensitive and accurate detection of copy number variants using read depth of coverage: Genome Res, v. 19, p. 1586-92.

Yu, T. W., M. H. Chahrour, M. E. Coulter, S. Jiralerspong, K. Okamura-Ikeda, B. Ataman, K. Schmitz-Abe, D. A. Harmin, M. Adli, A. N. Malik, A. M. D'Gama, E. T. Lim, S. J. Sanders, G. H. Mochida, J. N. Partlow, C. M. Sunu, J. M. Felie, J. Rodriguez, R. H. Nasir, J. Ware, R. M. Joseph, R. S. Hill, B. Y. Kwan, M. Al-Saffar, N. M. Mukaddes, A. Hashmi, S. Balkhy, G. G. Gascon, F. M. Hisama, E. LeClair, A. Poduri, O. Oner, S. Al-Saad, S. A. Al-Awadi, L. Bastaki, T. Ben-Omran, A. S. Teebi, L. Al-Gazali, V. Eapen, C. R. Stevens, L. Rappaport, S. B. Gabriel, K. Markianos, M. W. State, M. E. Greenberg, H. Taniguchi, N. E. Braverman, E. M. Morrow, and C. A. Walsh, 2013, Using whole-exome sequencing to identify inherited causes of autism: Neuron, v. 77, p. 259-73.

Zeesman, S., M. J. Nowaczyk, I. Teshima, W. Roberts, J. O. Cardy, J. Brian, L. Senman, L. Feuk, L. R. Osborne, and S. W. Scherer, 2006, Speech and language impairment and oromotor dyspraxia due to deletion of 7q31 that involves FOXP2: Am J Med Genet A, v. 140, p. 509-14.

Zhang, D., L. Cheng, Y. Qian, N. Alliey-Rodriguez, J. R. Kelsoe, T. Greenwood, C. Nievergelt, T. B. Barrett, R. McKinney, N. Schork, E. N. Smith, C. Bloss, J. Nurnberger, H. J. Edenberg, T. Foroud, W. Sheftner, W. B. Lawson, E. A. Nwulia, M. Hipolito, W. Coryell, J. Rice, W. Byerley, F. McMahon, T. G. Schulze, W. Berrettini, J. B. Potash, P. L. Belmonte, P. P. Zandi, M. G. McInnis, S. Zöllner, D. Craig, S. Szelinger, D. Koller, S. L. Christian, C. Liu, and E. S. Gershon, 2009, Singleton deletions throughout the genome increase risk of bipolar disorder: Mol Psychiatry, v. 14, p. 376-80.

Zhang, J., L. Feuk, G. E. Duggan, R. Khaja, and S. W. Scherer, 2006a, Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome: Cytogenet Genome Res, v. 115, p. 205-14.

Zhang, X., L. Li, J. Fourie, J. R. Davie, V. Guarcello, and R. B. Diasio, 2006b, The role of Sp1 and Sp3 in the constitutive DPYD gene expression: Biochim Biophys Acta, v. 1759, p. 247-56.

Zheng, J. Q., and M. M. Poo, 2007, Calcium signaling in neuronal motility: Annu Rev Cell Dev Biol, v. 23, p. 375-404.

Zhou, Z. W., G. Q. Wang, d. S. Wan, Z. H. Lu, Y. B. Chen, S. Li, G. Chen, and Z. Z. Pan, 2007, The dihydrouracil/uracil ratios in plasma and toxicities of 5-fluorouracil-based adjuvant chemotherapy in colorectal cancer patients: Chemotherapy, v. 53, p. 127-31.

Zollino, M., D. Orteschi, M. Murdolo, S. Lattante, D. Battaglia, C. Stefanini, E. Mercuri, P. Chiurazzi, G. Neri, and G. Marangi, 2012, Mutations in KANSL1 cause the 17q21.31 microdeletion syndrome phenotype: Nat Genet, v. 44, p. 636-8.

Zweier, C., E. K. de Jong, M. Zweier, A. Orrico, L. B. Ousager, A. L. Collins, E. K. Bijlsma, M. A. Oortveld, A. B. Ekici, A. Reis, A. Schenck, and A. Rauch, 2009, CNTNAP2 and NRXN1 are mutated in autosomal-recessive Pitt-Hopkins-like mental retardation and determine the level of a common synaptic protein in Drosophila: Am J Hum Genet, v. 85, p. 655-66.