UNIVERSITA' DEGLI STUDI DI VERONA

DIPARTIMENTO DI
SCIENZE DELLA VITA E DELLA RIPRODUZIONE

SCUOLA DI DOTTORATO DI SCIENZE BIOMEDICHE TRASLAZIONALI

DOTTORATO DI RICERCA IN BIOMEDICINA TRASLAZIONALE

CICLO XXV / ANNO 2010

TITOLO DELLA TESI DI DOTTORATO

WHOLE TRANSCRIPTOME ANALAYSIS BY NEXT GENERATION SEQUENCING (NGS) IN AUTISM SPECTRUM DISORDERS (ASDs)

S.S.D. BIO/13

Coordinatore: Prof. Cristiano Chiamulera

Tutor: Dott.ssa Elisabetta Trabetti

Dottoranda: Dott.ssa Chiara Zusi

INDEX

1. ABSTRACT

Autism Spectrum Disorders (ASDs) represent a group of childhood neurodevelopmental and neuropsychiatric disorders characterized by deficits in verbal communication, impairment of social interaction, and restricted and repetitive patterns of interests and behaviours. Evidences indicate that ASDs have strong genetic bases. Known chromosomal anomalies, rare genetic variants and single nucleotide polymorphisms (SNPs) have been related to ASD phenotypes in many studies. Furthermore Comparative Genomic Hybridization (CGH) studies have revealed copy number variations (CNVs) as risk factors. Recently, several studies have suggested that lymphoblastoid cells (LCLs) can discriminate between ASDs and control samples.

This study is part of a Telethon project which has been started in 2009 and involves different Italian clinical and research groups; it aims to analyze gene expression variations in ASD subjects, characterized for CNVs potentially involved in the onset of autism. Transcriptome from LCLs of 27 ASD probands and 23 health controls have been analyzed through Next Generation Sequencing technology (RNA Sequencing).

Gene set enrichment analysis (GSEA), on the total cohort and on a subgroup with a 22q13.3 deletion, revealed that autoimmune disorders and antigen processing and presentation pathways are the most enriched ones. Subgroup's GSEA highlights the involvement of axon guidance pathway, confirming that LCLs could exhibit biomarkers relevant to autism. Moreover, we demonstrate that three outlier genes cluster within a CNV on 16p13.1, suggesting that this is a potential candidate ASD region. This study provides evidence that potentially causative structural variants have a functional impact via transcriptome alterations in ASDs at a genome wide level and demonstrates the utility of integrating gene expression with mutation data.

Further analysis of differentially expressed genes and CNVs not selected in this study will help understanding the genetic bases for ASD pathophysiology and unravelling potential new pathways involved in ASDs.

1.1 Riassunto

I disordini dello spettro autistico (ASDs) sono caratterizzati dalla compromissione dell'interazione sociale e della comunicazione verbale e non verbale, e da comportamenti ripetitivi e stereotipati. L'autismo è una delle più frequenti tra le malattie complesse ereditabili, tuttavia, solo pochi geni implicati nell'eziologia sono stati identificati.

Negli ultimi anni, in diversi studi, sono stati individuati polimorfismi a singolo nucleotide (SNP), anomalie cromosomiche e rare varianti genetiche associati al fenotipo autistico. Inoltre, in studi di ibridazione genomica comparativa (CGH), come fattori di rischio sono state individuate alcune variazioni del numero di copie (CNV). Altri studi hanno evidenziato che le cellule linfoblastoidi possono discriminare tra soggetti con ASD e campioni di controllo.

Il presente studio è parte di un progetto Telethon, avviato nel 2009, che coinvolge diversi gruppi italiani di clinica e di ricerca. Questo progetto ha l'obiettivo di analizzare le variazioni di espressione genica in 27 soggetti con ASD e 23 controlli sani. I probandi selezionati presentano CNV potenzialmente coinvolte nell'insorgenza dell'autismo. Il trascrittoma di 27 probandi e 23 controlli sani è stato analizzato attraverso la tecnica di sequenziamento di nuova generazione dell'RNA (RNA sequencing).

L'analisi di arricchimento del gruppo di geni (GSEA) risultati differenzialmente espressi, compiuta sull'intera coorte e su un sottogruppo con una delezione 22q13.3, ha rilevato che i principali pathway arricchiti appartengono ai disturbi autoimmuni e al pathway di presentazione dell'antigene. La sottoanalisi compiuta sui campioni con la delezione evidenzia il coinvolgimento di geni appartenenti al pathway di orientamento degli assoni, confermando che le linee cellulari linfoblastoidi possono presentare biomarcatori rilevanti per l'autismo. Inoltre, abbiamo dimostrato che tre geni "outlier" clusterizzano all'interno di una CNV sul cromosoma 16p13.1, suggerendo che questo è un potenziale candidato locus per autismo.

Questo studio fornisce la prova che le varianti strutturali, potenzialmente causative di ASD, hanno un impatto funzionale attraverso alterazioni del trascrittoma e dimostra l'utilità di integrare i dati di espressione genica con i dati genomici.

Ulteriori analisi sui geni differenzialmente espressi e su CNVs, non selezionate in questo studio, contribuiranno a mettere in rilievo le basi genetiche e fisiopatologiche di ASD e ad evidenziare nuovi potenziali pathway coinvolti nei disordini dello spettro autistico.

2. INTRODUCTION

Autism is a neurodevelopmental disorder in the category of pervasive developmental disorders, and is characterized by severe and pervasive impairment in reciprocal socialization, qualitative impairment in communication, and repetitive or unusual behavior. The Diagnostic and Statistical Manual of Mental Disorders, 4th edition (DSM-IV-TR; American Psychiatric Association, 2000) and the International Classification of Diseases, 10th edition (ICD-10; World Health Organization, 1992), include autistic disorder, Asperger's syndrome, pervasive developmental disorder-not otherwise specified (PDD-NOS), Rett's syndrome, and childhood disintegrative disorder as pervasive developmental disorders. Clinicians and researchers use autism spectrum disorders (ASDs) to include autism, Asperger's syndrome, and PDD-NOS. For children with Rett's disorder or childhood disintegrative disorder, clinical course, patho-physiology, and the diagnostic strategies used are different.

2.1 Epidemiology

The population prevalence of ASDs is debated, estimates published in 2012 by the Centers for Disease Control and Prevention (CDC) state that the incidence of ASDs is 1 out of 88 children in the United States, as many as 1 in 500 individuals have autism, making it one of the most common neurodevelopmental disorders (Prevalence of autism spectrum disorders-Autism and developmental disabilities monitoring network, 2012). ASDs are most often diagnosed before age of four, and are at least three to four times more common in males than females (Fombonne E., 2005).

2.2 History and definitions of ASDs

In 1943 Leo Kanner, in a paper entitled Autistic Disturbances of Affective Contact, which went on to become a classic in the field of clinical psychiatry, described eleven children with "infantile autism" a distinct syndrome instead of previous depictions of such children as feeble-minded, retarded, moronic, idiotic or schizoid (Kanner L., 1943). The spectrum of clinical conditions labeled autism soon expanded beyond Kanner's first description. In 1944, one year after Kanner's paper, Hans Asperger described children that he also called "autistic", but who seemed to have high non-verbal intelligence

quotients and who used a large vocabulary appropriately. Confusion remains about the distinction between Asperger syndrome and high-functioning autism for many years.

Ever since the recognition and description of autism by Kanner, scientists and clinicians alike have been searching for the cause of this disorder or group of disorders. Initially, although Kanner in his original paper hypothesized that autism was an inborn biologic condition, misconceptions based on psychoanalytic theory led workers to view autism as psychiatric disorder. The major proponent of the purely psychiatric approach was Bruno Bettelheim (1967) who attributed the disorder to an unnurturing mother, called the "refrigerator theory". Autism remained an esoteric disorder for several decades until physicians and parents connected these symptoms with an increasing number of patients. The idea of autism as a psychiatric disorder persisted until the 1960s when Rimland (1968) hypothesized a neurologic basis.

Clinical definitions of autism continue to evolve. The current DSM-IV-TR includes autism in a broad category of pervasive developmental disorders. This spectrum blurs at the edges with disruptive behavior, communication disorders and intellectual disability at one end, and with behaviors now thought to be normal at the other. Repeated revision and expansion of the diagnostic categories has probably contributed to the gradual increase in the reported prevalence of autism spectrum disorders that has been evident since the mid-1980s.

The American Psychiatric Association (APA) has proposed new diagnostic criteria for the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) for autism. The APA's extensive development process of the fifth edition of the DSM is nearing its conclusion and final publication is planned for May 2013. The proposal by the DSM-5 Neurodevelopmental Work Group recommends a new category called autism spectrum disorder, which would incorporate several previously separate diagnoses, including autistic disorder, Asperger's disorder, childhood disintegrative disorder and PDD-NOS. Proposed DSM-5 criteria are being tested in real-life clinical settings known as field trials. Field testing of the proposed criteria for ASD does not indicate that there will be any change in the number of patients receiving care for ASDs in treatment centers, just more accurate diagnoses that can lead to more focused treatment.

Controversies over clinical definitions may only be resolved with the discovery of biomarkers (biochemical, anatomical or physiological measures) that are specific to one or more aspects of autism. Biomarkers are essential for an understanding of brain mechanisms underlying the various autisms, and for the development of useful therapeutics.

2.3 Clinical characteristics and assessment

Symptoms of ASDs affect three core domains: socialization, communication, and behavior (Figure 1). Autism is a disorder that usually begins in infancy, at the latest, in the first three years of life, but typical language development might delay identification of symptoms.

Socialisation

- Impaired use of non-verbal behaviours to regulate interactions
- Delayed peer interactions, few or no friendships, and little interaction
- Absence of seeking to share enjoyment and interests
- Delayed initiation of interactions
- Little or no social reciprocity and absence of social judgment

Communication

- Delay in verbal language without non-verbal compensation (eg, gestures)
- Impairment in expressive language and conversation, and disturbance in pragmatic language use
- Stereotyped, repetitive, or idiosyncratic language
- Delayed imaginative and social imitative play

Restricted, stereotyped, and repetitive patterns of behaviour

- Preoccupation with stereotyped or restricted interests or topics
- Adherence to routines, rigidity, and perseverative behaviour
- Stereotyped, repetitive motor mannerisms, and selfstimulatory behaviour
- Preoccupation or fascination with parts of items and unusual visual exploration

Fig 1. Core domains of autism (Levy et al, 2009).

Diagnoses show heterogeneity of clinical phenotype, severity, and type and frequency of symptoms. Indeed, autism is a heterogeneous condition; no two children or adults with autism have exactly the same profile, but difficulties fall into the three core domains that are reliably measured and usually consistent across time, even though specific behaviors may change with development. ASDs have characteristic diagnostic criteria: ages of symptom recognition, associated medical and developmental features, standard effective treatments, and usual courses of development (Figure 2).

In addition, autism affects not only social behavior and language but also many other aspects of functioning, including sensory responsiveness, play, and motor activity. Approximately 30% of autistic patients also have an associated seizure disorder. Many, but not all, individuals with autism, have mental retardation and almost all individuals with autism have a history of language delay. However, there are ASD subjects in which neither language delay nor mental retardation is present (e.g., Asperger's disorder).

	Autism	Asperger's syndrome	Pervasive developmental disorder-not otherwise specified (PDD-NOS)
Age of recognition (diagnosis*)	0–3 years (3–5 years)	>3 years (6–8 years)	Variable
Regression	About 25% (social or communication)	No	Variable
Sex ratio (male:female)	2:1	4:1	Male>female (variable)
Socialisation	Poor; >2 DSM-IV criteria	Poor	Variable
Communication	Delayed, deviant; might be non-verbal	No early delay; qualitative and pragmatic difficulties later	Variable
Behaviour	More impaired than in Asperger's syndrome or PDD-NOS (includes stereotypy)	Variable (circumscribed interests)	Variable
Intellectual disability	>60%	Mild to none	Mild to severe
Cause	More likely to establish genetic or other cause than in Asperger's syndrome or PDD-NOS	Variable	Variable
Seizures	25% over lifespan	Roughly 10%	Roughly 10%
Outcome	Poor to fair	Fair to good	Fair to good

Fig 2. Differential diagnostic features of autism spectrum disorders (Levy et al, 2009).

Diagnostic assessment of ASDs includes use of ICD or DSM-IV-TR diagnostic criteria, and standardized methods to assess core (Figure 1) and comorbid symptoms. This multidisciplinary assessment includes a review of caregiver concerns, descriptions of behavior, medical history, and questionnaires (Charman T. et al, 2002).

The gold standard assessment methods based on DSM-IV-TR criteria for the evaluation of suspected ASDs are the Autism Diagnostic Observation Schedule (ADOS) (Lord C. et al, 2000) and the Revised Autism Diagnostic Interview (ADI-R) (Le Couteur A. et al, 1989). These tools, requiring training and reliability testing, have improved accuracy and reliability of diagnosis. The ADI-R provides a standardized, semi-structured interview and a diagnostic algorithm for the DSM-IV and the ICD-10 definitions of autism; because of the time needed for administration (1-3h), its use is precluded in many clinical settings. The ADOS is a semi-structured, standardized assessment in which the researcher observes the social interaction, communication, play and imaginative use of materials for children suspected of having autism. Final autism case diagnosis is defined as meeting criteria on the communication, social, and repetitive behaviors domains of the

ADI-R and scoring at or above the cut off for autistic disorder on the ADOS module 1 or 2, used with children who do not consistently use phrase speech or are not verbally fluent.

A comprehensive diagnostic assessment should include medical investigation for causes and associated diagnoses (Freitag CM., 2007). An appropriate medical investigation for causes comprises a detailed history and physical examination (with careful examination for dysmorphology). Clinical genetic assessment might include laboratory studies referral to a clinical geneticist. Genetic laboratory studies should cover routine karyotype and molecular DNA testing for fragile X, or Comparative Genomic Hybridisation (CGH), or both (Schaefer GB. et al, 2008). Associated medical problems such as seizures show a need for electroencephalogram (EEG), substantial regression a need for metabolic investigation, and abnormal head size a need for neuroimaging in some.

2.4 Neurobiology

Attempts to identify unified theories explaining core and comorbid deficits have been unsuccessful, which is not surprising in view of the heterogeneous expression of ASDs. In studies (DiCicco-Bloom E. et al, 2006) of this disorder as a neurodevelopmental disorder of prenatal and postnatal brain development, researchers have attempted to elucidate these theories by examination of brain growth, functional neural networks, neuropathology, electrophysiology, and neurochemistry. Neurocognitive theories include pragmatic language impairment and difficulties in intersubjectivity (theory of mind), executive function and problem-solving mindset, weak central coherence and difficulty with integration of information into meaningful wholes (Volkmar FR. et al, 2003), and deficits in connectivity and processing demands (Minshew NJ. et al, 2007).

Neurobiological findings support different theories. Macrocephaly is noted by age 2–3 years in 20% of children with ASDs. Brain growth accelerates at 12 months (Minshew NJ. et al, 2007). These changes arise in parallel with onset of core symptoms during the first 2 years of life. Results of neuroimaging studies (Pardo CA. et al, 2007) have shown overgrowth in cortical white matter and abnormal patterns of growth in the frontal lobe, temporal lobes, and limbic structures such as the amygdala. These brain regions are

implicated in development of social, communication, and motor abilities that are impaired in ASDs.

Functional magnetic resonance imaging (MRI) has shown differences in patterns of activation and timing of synchronisation across cortical networks, with lowered functional connectivity relating to language, working memory, social cognition or perception, and problem solving. The most reliably replicated functional MRI abnormal finding is hypoactivation of the fusiform face area, associated with deficits in perception of people compared with objects (DiCicco-Bloom E. et al, 2006; Schultz RT., 2005).

Neurochemical investigations with animal models and empirical drug studies remain inconclusive. Serotonin and genetic differences in serotonin transport seem to have the most empirical evidence for a role in ASDs (Lam KSL. et al, 2006), whereas data lending support to the roles of dopaminergic and glutaminergic systems are presently less robust, but are evolving. Study of the role of the dopaminergic and cholinergic system, oxytocin, and aminoacid neurotransmitters shows promise (Lam KSL. et al, 2006). Together, results of clinical, neuroimaging, neuropathological, and neurochemical studies (Pardo CA. et al, 2007) show that ASDs are disorders of neuronal-cortical organisation that cause deficits in information processing in the nervous system, ranging from synaptic and dendritic organisation to connectivity and brain structure. These changes probably alter developmental trajectory of social communication and seem to be affected by genetic and environmental factors.

2.5 Causes

There is strong evidence for a genetic influence in ASDs and more recently there has been some association of autism with low birth weight infants and with older parents. However, the proposed etiologies that have been the most controversial, and have had the most significant effect on the health of children, concern the possible association of toxins and vaccines with autism. Mercuric compounds are nephrotoxic and neurotoxic at high doses. Thimerosal, a preservative used widely in vaccine formulations, contains ethylmercury. Thus it has been suggested that childhood vaccination with thimerosal-containing vaccine could be causally related to neurodevelopmental disorders such as autism. Results of different studies do not support a causal relationship between

childhood vaccination with thimerosal-containing vaccines and development of ASDs (Hviid A. et al, 2003; Farrington CP. et al, 2001; DeStefano F. 2007).

One of the first areas of interest in the 1960s and 1970s was the search for an infectious agent that might be involved in the etiology of ASDs (Chess S., 1971). However, after decades of research no definite role for infectious agents in autism etiology has been confirmed. On the other hand, these endeavors have led to observations that perhaps the immune system was involved in autism, and evidence continues to mount that immune abnormalities are indeed associated with ASDs.

Several sources of evidence found during last decades suggest that strong genetic components are involved in susceptibility to ASDs: there are much higher concordance rates of ASDs in monozygotic twins (92%) than dizygotic twins (10%), and recent estimate of the sibling recurrence risk ratio (λs) for autism is 22 [min 9, max 45], corresponding to an increase of 45- to 150-fold. These values make this disorder the neuropsychiatric disorder most affected by genetic factors (Bailey A. et al, 1995; Lauritsen MB. et al, 2005).

ASDs are multifactorial, with many risk factors acting together to produce the phenotype. The difference between monozygotic and dizygotic concordance rates suggests some risk factors interact (ie, gene–gene or gene–environmental interactions). These effects could be a result of toxic environmental factors or epigenetic factors that alter gene functions, in turn altering neural tissue. Epigenetic factors can be specific aspects of the physical environment (eg, biochemically active compounds) or specific types of psychological experiences (eg, stress) that alter brain chemistry, turn genes off or on at specific times during development, or regulate gene expression in other ways. The possible role of environmental and epigenetic factors is an area being studied.

2.6 Genetics of ASDs

ASD genes have been difficult to identify, despite the high heritability of the disorder. Approximately 10-15% of cases may be due to an identifiable Mendelian condition or genetic syndrome (including chromosomal anomalies). Rett syndrome, fragile X syndrome (0-20%), neurofibromatosis type I (NF1; 0.2-14%), tuberous sclerosis (TSC; 0.4-2.9%) are the most frequently cited Mendelian conditions. Other rare microdeletion or

single-gene defects have been associated with ASDs, including those found in Williams, Sotos, Cowden, Moebius, and Timothy syndromes. ASDs may also occur in some mitochondrial diseases and untreated phenylketonuria. In addition, cytogenetically visible anomalies on all chromosomes are observed in ASDs (Abrahams BS. et al, 2008; Zafeiriou DI. et al, 2007). The 22q13.3 deletion causes a neurodevelopmental syndrome, also known as Phelan-McDermid syndrome, characterized by developmental delay, severe delay or absence of expressive speech (Phelan MC. et al, 2001) and autistic-like behavior (Goizet C. et al, 2000). The genotype-phenotype study suggests that the 22q13 deletion phenotype has to be attributed to SHANK3 gene disruption. SHANK3 is strongly expressed in the cerebral cortex and cerebellum and has been proposed as the major cause for both the neurological features of the 22q13 deletion syndrome and for a monogenic form of autism (Bonaglia MC. et al, 2001).

However, the remaining 90% of ASDs, while highly familial, have unknown genetic etiology. Thus, it seems reasonable to accelerate the gene discovery process by using combinations of experimental approaches, such as the study of single gene or more simple causes, such as chromosomal copy number imbalances whose phenotypes include ASDs. It is also likely that inherited epigenetic modifications and gene by gene and gene by environmental interactions are significant sources of variation (Farber CR. et al, 2009).

Insights into underlying biological mechanisms for ASDs have been gained from study of syndromes with increased rates of this disorder. For example, functions of the genes underlying fragile X (FMR1) and Rett's syndrome (MECP2) implicate synaptic dysfunction in cause and pathogenesis (Ramocki MD. et al, 2008). Further evidence for synaptic dysfunction as a unifying cause has come from studies on rare mutations in genes coding for neural cell adhesion and synaptic molecules (Tabuchi K. et al, 2007). Convergence of genetic discoveries with implications for synaptic maturation is especially notable because findings from neuroimaging research also suggest that structural and functional brain connectivity is aberrant in ASDs (Minshew NJ. et al, 2007). Thus, genetic and neurobiological evidence point to a good causal model of this disorder genetically mediated by abnormalities of synaptic maturation and connectivity.

There are two alternative proposals, one involving numerous rare genetic mutations and the other involving fewer but more common genetic variations. Supporting the rare

mutation hypothesis are mutations in several genes and rare structural DNA variations both of which have been identified, although the pervasiveness of these effects remains controversial (Weiss LA. et al, 2008; Sebat J. et al, 2007). Data supporting the effect of common variation have been more difficult to find. Only a few common variants have been identified as possible candidate genes in linkage and association studies (Veenstra-Vanderweele J. et al, 2004), and many of these have not been replicated in subsequent independent sample studies. Moreover, more recent studies are identifying a growing number of distinct and individually rare genetic causes, suggesting that the genetic architecture of ASDs may have significant contribution from heterogeneous rare variants.

Since 2003 (Volkmar FR. et al, 2003) a number of approaches are being used to elucidate the association between specific genes and autism. Whereas genome screens (whole genome sequencing, SNP-array) search for common genetic markers in populations of multiplex families with autism, Comparative Genomic Hybridization studies search for inherited or spontaneous genetic abnormalities on an individual basis (Muhle R. et al, 2004).

Genome wide linkage scans, as genetic approach to identify susceptibility genes, is very powerful, but the heterogeneity present within autism families has led thus far to mixed success in identifying candidate genes (Benayed R. et al, 2005; Ma DQ. et al, 2005).

Genome-wide association studies (GWAS) compare the DNA of people with and without a disorder using thousands or millions of markers scattered throughout their genomes. However, GWAS can only point to large chromosome regions rather than identifying particular genes or gene mutations. The first GWAS on autism, published in 2009, implicated two regions with mild effects: one on chromosome 5, and one on chromosome 20 (Weiss LA. et al, 2009). Subsequent GWAS on autism have failed to turn up any other parts of the genome with statistical significance.

One possible interpretation of these data is that genetic risk factors may contribute combinatorially to each of the three phenotypic domains of ASDs. This line of reasoning suggests that there are genetic variants that influence independently social behavior, language development, and behavioral flexibility, and that ASDs result from a confluence of multiple genetic risk factors for each of the three domains. An alternative

hypothesis is that the core phenotypes that characterize ASDs are mediated through a common biological mechanism, and thus the three domains are not separable. In the latter hypothesis, the observed genetic complexity is due to vulnerability imparted by multiple genes encoding proteins within a common biological pathway (Campbell DB. et al, 2010).

The evaluation of linkage peaks extended beyond genes with selective brain expression to consider the complex medical conditions seen in autism patients. In addition to the well known behavioral core features, some individuals with autism exhibit gastrointestinal, immunological, or nonspecific neurological symptoms (Valicenti-McDermott M. et al, 2006; White JF. 2003). Although the degree to which individuals with autism exhibit more medical complications compared with typical individuals is debated, it is possible that autism vulnerability could include genes involved more broadly in multiple biological processes that impact the development and function of the brain and other organ systems in parallel.

2.6.1 Copy Number Variants

Another promising development in understanding the genetics of ASDs is the discovery of variations in the gene copy number as a risk factor (Sebat J. et al, 2007). Copy-number variation (CNV) is a structural variation in the genome, ranging from a few thousand to several million base pairs, in which material is either duplicated or deleted. A number of population-based studies have shown that this type of genomic variant can affect as much as 12% of the human genome (Redon R. et al, 2006). Many studies have gone on to demonstrate the importance of CNVs in determining human phenotypic variation and disease susceptibility (Perry GH. et al,2007; Craddock N. et al, 2010). CNVs can be de novo or inherited. Almost all these variations are deletions, with many fragments containing several genes (Cook EH. et al,2008). De-novo CNVs seem to be strongly associated with intellectual impairment and dysmorphology (Sebat J. et al, 2007). Most seem to be individually unique, although we do not know the full implications of them because their relation to phenotype is not established (Cook EH. et al,2008), and affected siblings do not always share specific variations (Marshall CR. et al, 2008). Furthermore, to recognize whether a given de-novo variant is abnormal is difficult because the population distribution of specific CNVs is unknown.

Comparative Genomic Hybridization (CGH) and other molecular cytogenetic techniques are methods to identify and quantify DNA copy number changes across the genome, enabling the characterization of chromosomal variations with confidence. These techniques are discovering an increasing number of copy number variations in individuals with ASDs. A catalogue of this structural variation is provided by the Database of Genomic Variants (DGV, http://projects.tcag.ca/variation/) and currently covers about 29% of the human genome.

In particular, the application of array CGH for genome-wide screening has a major impact on clinical genetic diagnosis. First, array CGH made it possible to identify submicroscopic aberrations in at least 9% of patients with idiopathic mental retardation (Buysse K. et al, 2009a). Second, it facilitated the identification of disease genes of some well-known clinical syndromes (Vissers LE. et al, 2004; Kleefstra T. et al, 2006) and led to the characterization of many novel microdeletion and microduplication syndromes (Buysse K. et al, 2009b; Menten B. et al, 2007; Slavotinek AM. 2008). Finally, it has contributed to the elucidation of the mechanisms causing genomic disorders (Gu W. et al, 2008). Array CGH technology also facilitated the discovery of a new important source of human genetic variation.

Studies on both humans and mice have shown that CNVs can influence gene expression (Stranger BE. et al, 2007; Henrichsen CN. et al, 2009). This has led many scientists to hypothesize that these heterozygous CNVs may influence susceptibility to neurodevelopmental disorders through a gene-dosage mechanism. For instance, it is easy to imagine how mild perturbations in axon guidance molecules could affect critical stages of brain development. Others have looked for evidence of an alternative mechanism whereby deletions might unmask recessive coding changes in the opposite allele (Flipsen-ten Berg K. et al, 2007; Pagnamenta AT. et al, 2011; Vorstman JA. et al, 2011). A third possible mechanism exists for deletions or duplications with breakpoints disrupting two different genes. Where the two neighboring genes are encoded on the same chromosomal strand, such CNVs could potentially result in gene-fusion transcripts. A number of genomic rearrangements that lead to fusion transcripts have already been described in autism and schizophrenia (Pagnamenta AT. et al, 2010; Walsh T. et al, 2008; Zhou X. et al, 2010). If such transcripts are stable they may be translated into novel proteins with a possibility for deleterious gain-of-function effects.

The discovery of rare and recurrent CNVs as important pathogenic mutations in ASDs was a watershed in ASD genetics. Recurrent CNVs such as those at 16p11.2, 22q11.2, 1q21.1, 7q11.23 and 15q11-q13 show statistically significant association with ASDs. However, the functional impact of these CNVs on downstream RNA expression at both a collective and individual level remains largely unknown. Because CNVs alter copy number and must presumably act via changes in downstream gene expression, an initial study that explored the transcriptome-wide effects of CNVs reported that changes in gene copy number explained roughly 20% of detected transcriptional alterations. Although widely assumed, it remains unknown whether rare CNVs identified in autistic individuals have similar effects on transcription levels and subsequent pathophysiology. Alternative lines of evidence, such as gene-expression data, might confirm the presence of functional alterations related to a particular CNV and would thus be of significant utility.

2.6.2 Gene expression analyses

The transcriptome is the complete set of transcripts in a cell, and their quantity, for a specific developmental stage or physiological condition. Although transcriptomes are more dynamic than genomic DNA, these molecules provide direct access to gene regulation and protein information. Understanding the transcriptome is, therefore, essential for interpreting the functional elements of the genome and revealing the molecular constituents of cells and tissues, and also for understanding development and disease.

RNA-Sequencing is a recently developed approach for gene expression analysis based on deep-sequencing technologies. Next Generation Sequencing (NGS) has revolutionized the genetic landscape yielding results never before achieved. The output of this method is the number of transcripts that were produced in a cell under chosen conditions. RNA-Seq is also possible to use in studying alternative splicing, novel exons, exon-exon junctions or even novel genes. General advantages of RNA-Seq in comparison with previous techniques (i.e. microarrays, SAGE, CAGE, MPSS, Northern blot and qRT-PCR) are unlimited range of gene expression, much more higher sensitivity and ability to control regions with high similarity (single base resolution), ability to work with low amount of RNA, there is no need to know genes a priori (genes can be assembled de novo) and low-intensity background bias.

Geschwind and colleagues used RNA-Seq to profile the transcriptomes of 244 people with autism and members of their family in order to get a sense of which genetic mutations led to real changes in gene expression (Voineagu I. et al, 2011). Geschwind's study found that many of these genes with altered expression patterns were clustered in pathways known to play important roles in neuron function, development and structure (Luo R. et al, 2012). This discovery suggests common mechanisms through which seemingly different gene mutations could cause disease.

Moreover, Geschwind's team analysed gene expression patterns in the brains of 19 people with autism and without, using brain tissue collected after each person's death (Voineagu I. et al, 2011). For each brain, the scientists studied three different regions, all thought to be important in autism. They found that in people with autism, two regions of the brain that normally have distinct patterns of gene expression - the frontal and temporal lobes of the cerebral cortex - instead had almost identical patterns, with the same genes switched on or off. The altered genetics suggest a lack of specialization among some brain cells, which could lead to differences in how the brain processes information.

However, the number of postmortem brain tissues from individuals with autism is limited, which potentially erodes this avenue of analysis. An alternative that has emerged is the field of blood genomics, where peripheral lymphocytes are used to represent the transcriptome.

2.6.3 Lymphoblastoid Cell Lines

While studies of gene expression in brain tissue may lead to a better understanding of the mechanistic basis for ASD, it is not an appropriate target for diagnostic assays. Ideally, diagnostic assays should use easily obtained patient samples such as blood, considering there is evidence that gene expression and other markers exist in the peripheral blood of ASD patients.

Blood genomics is beginning to be applied to the study of central nervous system (CNS) disorders; studies have demonstrated that genes expressed in the peripheral blood can classify CNS disorders and may be able to identify and predict the genes and pathobiologies involved in these disorders (Tang Y. et al, 2003; Tang Y. et al, 2004).

In addition, several studies have suggested that B-lymphoblastoid cells lines (LCLs) transformed with Epstein-Barr virus (EBV) can be used to detect biologically plausible correlations between candidate genes and neuropsychiatric diseases, including Rett syndrome (Horike S. et al, 2005), nonspecific X-linked mental retardation (Meloni I. et al, 2002), bipolar disorder (Iwamoto K. et al, 2004), fragile X syndrome (Brown V. et al, 2001; Nishimura Y. et al, 2007) and dup(15q) (Baron CA. et al, 2006a; Baron CA. et al, 2006b).

In particular, Hu et al, 2006 provided proof-of-principle that LCLs (and possibly their precursor peripheral blood cells) exhibit biomarkers relevant to autism, and further suggest their potential usefulness as reporter cells in developing a diagnostic screen for autism (Hu VW. et al, 2006).

Using patient lymphoblastoid cells in genetic studies will ensure adequate genetic material for current and future analyses. LCLs can also provide other biomolecules for use in transcriptomics, proteomics, metabolomics, and other types of analyses. In functional analyses, LCLs have been used as surrogate or cellular models in the study of diseases. However, care must be taken to determine the suitability of using transformed B cells in certain experiments. Despite some inherent limitations, the utility of LCLs is increasingly recognized, and with appropriate infrastructure and financial support, LCLs will be an important resource for genetic and functional research of neurological disorders.

While it is unlikely that RNA-Seq studies on LCL will identify the etiology(ies) of autism, this global approach to gene expression analyses is expected to highlight molecular or pathway defects related to the pathophysiology of the condition which, in turn, can be targeted for drug therapies. Moreover, as opposed to fixed autopsy tissues in which RNA may have degraded, a live cell model can also be used to examine the functional consequences of the genomic alteration(s) and the efficacy of different pharmacological agents in ameliorating the impaired function**.**

2.7 Aim of the study

This study is part of a Telethon project which has been started in 2009 and involves different Italian clinical and research groups. It aims to analyze differentially expressed genes through the next generation sequencing technology (RNASeq) in ASD subjects, previously characterized for CNVs potentially involved in the onset of autism.

Because brain or neuronal tissue are not available, we used lymphoblastoid cells lines (LCLs). We presume that combining genome array-CGH and RNASeq there will be much more opportunities to decipher the complex nature of these diseases.

In particular, the goals of the work are: to evaluate downstream effects of CNVs on gene expression and correlate genomic imbalances and transcriptome with ASD phenotypes. This research will further help understanding the genetic bases for ASD pathophysiology and unravelling potential new pathways involved in ASDs.

3. MATERIALS AND METHODS

3.1 Sample collection

Ascertaining of autistic subjects and their affected and unaffected family member as a part of Telethon Project was done through the following neurology centers of excellence for the study of autism and neurodevelopmental disorders: Istituto IRCSS Neurologico Nazionale C. Besta (Milan), Unit of Neurology of the Oasi Institute for Research on Mental Retardation and Brain Aging (Troina-EN) , Unità di Neuropsichiatria Infantile - ASL 20 (Verona).

The enrolment of the patient will be done according the following protocol. A case individual is eligible for inclusion in this research project only if all the following criteria apply:

- Meet DSM-IV-TR criteria for PDD/ASD and reach score cut-off in Autism Diagnostic Interview-Revised (ADI-R) (Le Couteur A. et al, 1989) and/or Autism Diagnostic Observation Schedule-Generic (ADOS-G) (Lord C. et al, 2000),
- Be at least 3 years of age at the time of entering the research project
- Have at least one parent or legal guardian giving voluntary written consent for their children to take part to this research project.

The exclusion criteria are as follows:

- Presence of profound mental retardation
- History of serious head injury, encephalitis or tumors
- Presence of metabolic disease or genetic syndrome of known cause
- Age older than 18.

The following current clinical protocol will be administered to all recruited patients:

- Physical examination
- Neurological examination
- Instrumental and laboratory exams:

- genetic (Karyotype, Fragile-X, subtelomeric analysis),
- metabolic screenings (blood and urine amino acids, urinary organic acids, urinary mucopolysaccharides, blood and urine uric acid),
- neurophysiologic (sleep EEG),
- neurological imaging (investigation or acquisition of previous MRI scan of adequate quality and performed not earlier than three years).

- Battery of psychopathological and neuropsychological assessment:
  - ADI-R: a standardized, semi structured, investigator-based interview for caregiver of subjects with possible ASD. The item focus primarily on the three domains of functioning – language/communication; reciprocal social interactions; and restricted, repetitive, and stereotyped behaviors and interests- that are specified as of diagnostic importance in international diagnostic manuals.
  - ADOS-G: a semi structured assessment of communication, social interaction, and play or imaginative use of materials which is usually videotaped. It consists of standard activities that allow the examiner to observe directly behaviors that have been identified as relevant to the diagnosis of ASD at different developmental levels and chronological ages.
  - Intelligence (i.e. Wechsler Scales or Leiter-R) or developmental Scales (i.e. Griffiths Mental Developmental Scales) to evaluate cognitive abilities. In circumstances in which detailed psychometric testing has been undertaken the Vineland Adaptive Scale (Sparrow SS. et al, 1985) may be used to a screening tool to assess developmental level.

Written informed consent was obtained from all study participants after a full explanation of the study. The research protocol was in accordance with Helsinki Declaration.

Fifty nine families with an ASD proband, for a total of 217 individuals, were enrolled. Most of the families are composed of one affected child (23) or one affected child plus an healthy sibling (20). Multiplex families (9 with 2, 1 with 3 and 1 with 4 non affected children) are present and in 5 families only one parent was available. Moreover, 8 unrelated children, affected by 22q13.3 deletion syndrome, also known as Phelan-McDermid syndrome, have been collected. Subject distribution is given in figure 3.

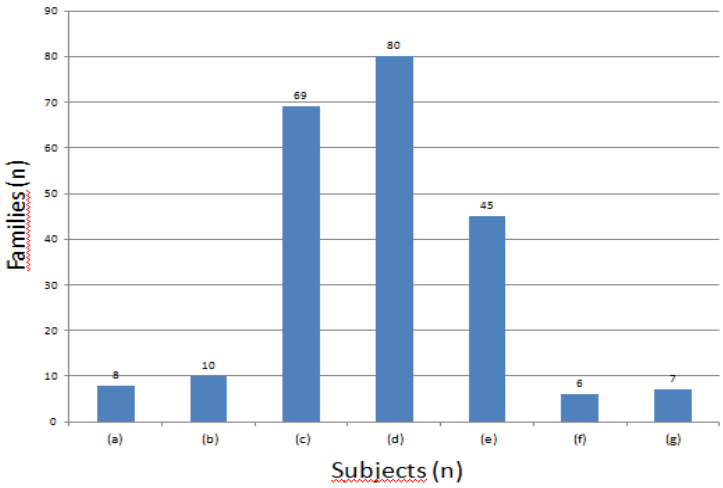

Fig 3. Families distribution according to the number of subjects. (a): 8 unfamiliar cases. (b) and (c): one affected child with one or both parents, respectively. From (d) to (g): one affected child, parents plus healthy siblings.

3.2 Lymphoblastoid cell lines (LCLs) establishment

EBV transformed lymphoblastoid cell lines were established for 106 subjects at University of Verona using standard protocol (ECACC, European Collection of Cell Coltures). LCLs are generated by EBV transformation of the B-lymphocyte component within the PBL population.

3.2.1 EBV isolation

The marmoset cell line B95-8 is a continuous line permissive for Epstein-Barr virus (EBV) replication that releases high titres of transforming EBV.

The B95-8 cell line, grown in RPMI 1640 medium (Euroclone, Milano, Italy), 10% fetal calf serum (FCS; Euroclone) at 37°C in a humidified 5% CO_2 chamber, has to be expanded until the concentration of $1x10^6$/ml and a final volume of 400 ml. To induce the lytic cycle cells have to be transferred in RMPI 1640 medium, 2% FCS at 33°C with a initial concentration of $0.2x10^6$/ml for two weeks . The decrease temperature and serum concentration induce the lytic cycle of EBV. All cells have to be removed by centrifugation for 5 min at 180 x *g*, then the supernatant has to be filtered through a 0.45 µm filter. The supernatant then is aliquoted and stored at -80°C.

3.2.2 Isolation of Mononuclear Cells from Peripheral Blood

For lymphocyte separation, 6 ml of whole blood from 106 subjects is diluted to the final volume of 26 ml with medium RPMI1640 (isotonic solution) and layered carefully over Ficoll-Paque™ PLUS (without intermixing) in a centrifuge tube. After a centrifugation at room temperature (typically at 400 g for 30–40 min) mononuclear cells (lymphocytes together with monocytes and platelets) are harvested from the interface layer (Figure 4). This material is then centrifuged twice in medium RPMI1640 solution to wash the lymphocytes and to remove the platelets.

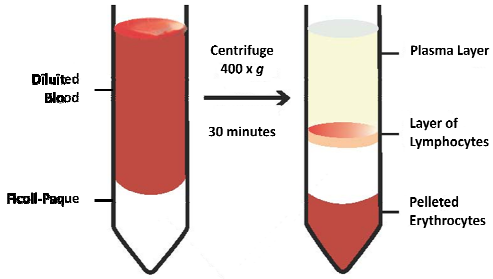

Fig 4. Lymphocyte separation using Ficoll-Paque™ PLUS. The centrifugation of diluted blood separates lymphocyte, plasma and erythrocytes, thus is possible the harvesting of lymphocytes layer.

3.2.3 B-lymphocytes infection

B lymphocytes were exposed to a standard virus dose (1 aliquot of 1ml) for an hour at 37°C, 5% CO_2. After infections, cells with EBV are centrifuged for 5 min at 400 x *g* to eliminate superrnatant, LCLs are then resuspended in complete medium (RPMI 1640 medium supplemented with 20% fetal bovine serum (FBS; Euroclone) and a mix of 2mM L-glutamine (Euroclone), fungizone 250ng/ml (Gibco, Life Technologies), and 2mM penicillin streptomycin (Life Technologies, Carlsbad, CA)supplemented with phytohemagglutinin (PHA; 500µg/ml; Sigma Aldrich) and seeded into a 24-well plate. As EBV enters B cells, the EBV genome has been reported to exist as distinct, high-copy episomal copies, as well as nonrandom integration to B cells (Gualandi G. et al, 1992; Lestou VS. et al, 1993; Leenman EE. et al, 2004; Gao J. et al, 2006). These cells are then allowed to proliferate and are subsequently maintained in growth medium before cryopreservation.

LCLs were maintained in complete medium at 37°C in a humidified 5% CO_2 chamber for about 6 weeks at which time wells with typical foci of EBV-transformed lymphoblastoid cells were scored positive. The cell lines were treated with a standardized procedure, to minimize environmental variation. Cell lines were harvested at a density of 0.8–1x10^6 cells/ml at least 85% viability for a minimum of 50 million cells per line. LCLs were spun for 5 min at 400 x g, and the resulting pellets were lysed with Trizol (Life Technologies, Carlsbad, CA) to obtain a final concentration of 10x10^6 cells/ml. Cells in Trizol were stored at -80°C. Other 25 million cells per line were spun for 5 min 400 x g, the pellet was resuspended in freezing medium (RPMI 1640, 20% FBS, 10% dimethylsulfoxide, DMSO) at the minimum concentration of 5x10^6 cells/ml. Cells were stored, first, at -80°C in Cryovials, using Nalgene® Cryo 1% freezing container, then were transferred in liquid nitrogen.

As mycoplasma infections in cultured cells tend to induce both biochemical and genetic changes, experimental results can often be misinterpreted. Therefore, contamination surveys should be periodically conducted. First, genomic DNA from each lines was extracted via the standard protocols described by Tang J. (2000). Then, each cell line was tested for the presence of mycoplasma using PCR-based method for the detection of 16S and 23S rRNA genes of 13 species using mycoplasma primers and conditions described by Sung H. (2006). Aliquots of the final PCR products were analyzed on 2% agarose gel. DNA bands were visualized with a UV transilluminator after GelRed™ staining, and then photographed.

3.3 RNA extraction

Total RNA from 50 LCLs samples was extracted using Trizol Reagent, a ready-to-use reagent for the isolation of total RNA, that maintains the integrity of the RNA, while disrupting cells and dissolving cell components.

1 mL of Trizol reagent is used per 10 x 10^6 cells that are lysed by repetitive pipetting. The homogenized samples are incubated for 5 minutes at 15 to 30°C to permit the complete dissociation of nucleoprotein complexes then is added 0.2 mL of chloroform per 1 mL of Trizol Reagent. Tubes should be capped securely and shacked vigorously by hand for 15 seconds and then let them at 15 to 30°C for 2 to 3 minutes. Samples are centrifuged at 12,000 x g for 15 minutes at 4°C. Following centrifugation, the mixture

separates into a lower red, phenol-chloroform phase, an interphase, and a colorless upper aqueous phase. RNA remains exclusively in the aqueous phase. The aqueous phase is transferred into a fresh tube.

Precipitation of the RNA from the aqueous phase is done by adding 0.5 mL of isopropyl alcohol and incubating samples at 15 to 30°C for 10 minutes. After a centrifugation at 12,000 x g for 10 minutes at 4°C the RNA precipitates and forms a gel-like pellet on the side and bottom of the tube. The supernatant is removed and the RNA pellet is washed once adding at least 1 mL of 75% ethanol. The sample is mixed by inversion and centrifuged at 7,600 x g for 5 minutes at 4°C. At the end of the procedure, briefly the RNA pellet has to be dried (air-dry for 5-10 minutes) and resuspended in 20 µl of water nuclease-free. Finally RNA is stored at -80°C.

An additional DNAse digestion step was performed to ensure that the samples were not contaminated with genomic DNA. 0.1 volume 10X TURBO DNase Buffer and 1 µL of TURBO DNase are added to a 7500ng of RNA, mixing gently. After an incubation at 37°C for 20-30 min, the resuspended DNase Inactivation Reagent is added to the solution. Finally, the solution is centrifuged at 10000 x *g* for 1.5 min and the supernatant is transferred to a fresh tube.

3.4 RNA quantification and quality analysis

RNA quantity was assessed using the ND-1000 Nanodrop (Thermo Scientific, Wilmington, DE, US) which enables highly accurate UV/Vis analyses of 1µl sample with remarkable reproducibility. Pure RNA has a 260/280 Abs ratio of 1.8/2.0 and a 260/230 Abs ratio of 1.8/2.0.

RNA was further analyzed for quality control by use of the RNA 6000 Nano LabChip (Agilent Technologies, Santa Clara, CA, US). Due to the omnipresence of RNases, and the instability of RNA, integrity checks and sample quantization are essential steps before any RNA dependent application. The RNA Nano kits (Agilent Technologies, Santa Clara, CA, US) allow the quantization and integrity analysis of total RNA as well as the visualization of rRNA impurities in mRNA samples. It also generates the RNA Integrity Number (RIN) based on a numbering system from 1 to 10, with 1 being the most degraded profile and 10 being the most intact.

150 ng of RNA were loaded into a chip system that contains up to 12 microcapillaries filled with a special matrix for the separation of RNA molecules according to their molecular weight. The chip is loaded into the Agilent 2100 Bioanalyzer, an electric field is applied and a short round of electrophoresis produces a complete separation of all the RNA molecules . Capillaries are then scanned by a laser source and two types of results are retrieved (Figure 5):

- An electropherogram with peaks corresponding to RNA molecules of different weight (28S, 18S, 5S) and whose subtended area is proportional to the abundance of a specific RNA; 28S should be twice abundant 18S area for the RNA best quality,
- An image of virtual agarose gel separation.

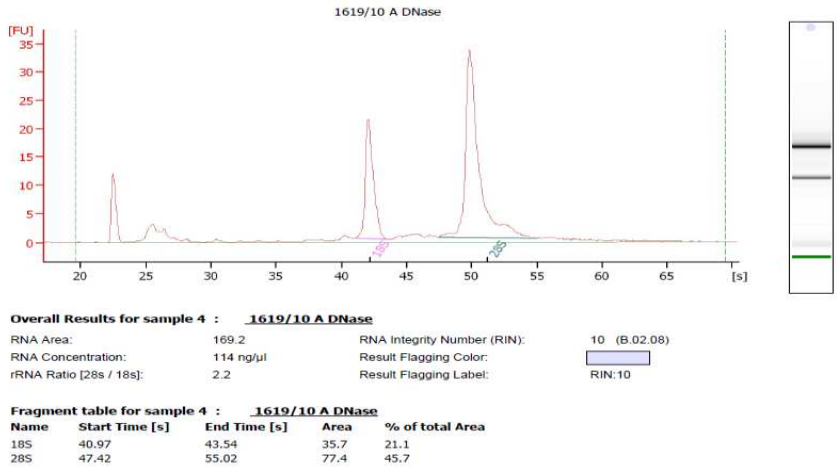

Fig 5. Agilent 2100 Bioanalyzer output of one control. This sample had a RIN of 10 and a 28S/18S of 2.2.

3.5 RNA Sequencing

RNA Sequencing experiments have been conducted in collaboration with the Functional Genomics Centre of the University of Verona (Prof. Massimo Delledonne), by the use of Illumina's next generation sequencing HiSeq 1000 instrument, TruSeq Sample Preparation Kit and cBot Automated Cluster Generation System.

Starting with 3µg of total RNA, the first step in the workflow involves purifying the poly-A containing mRNA molecules using poly-T oligo-attached magnetic beads. Following purification, the mRNA is fragmented into small pieces using divalent cations under elevated temperature. The cleaved RNA fragments are copied into first strand cDNA

using reverse transcriptase and random primers. This is followed by second strand cDNA synthesis using DNA Polymerase I and RNase H. These cDNA fragments then go through an end repair process, the addition of a single 'A' base, and then ligation of the adapters. The products are then purified and enriched with PCR to create the final cDNA library (Figure 6).

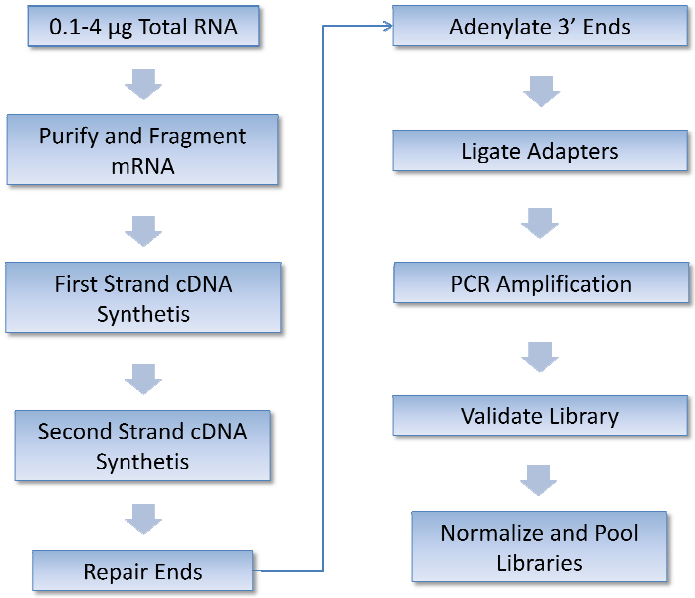

Fig 6. TruSeq RNA Sample Prep Purification Workflow.

Purify and fragment mRNA. The process purifies the poly-A containing mRNA molecules using poly-T oligo attached magnetic beads using two rounds of purification. During the second elution of the poly-A RNA, the RNA is also fragmented and primed for cDNA synthesis (Figure 7A).

Synthesize first strand cDNA. This process reverse transcribes the cleaved RNA fragments primed with random hexamers into first strand cDNA using reverse transcriptase and random primers. The optimized random hexamer priming strategy provides the most even coverage across transcripts.

Synthesize second strand cDNA. This process removes the RNA template and synthesizes a replacement strand to generate ds cDNA, AMPure beads are used to separate the ds cDNA from the second strand reaction mix(Figure 7B).

Fig 7. Optimized TruSeq RNA sample preparation. Starting with total RNA, mRNA is polyA-selected and fragmented (A). It then undergoes first- and second-strand synthesis to produce products ready for library construction (B).

The TruSeq DNA Sample Preparation Kits are used to prepare DNA libraries with insert sizes from 300-500bp for, single, paired-end, and multiplex sequencing. Library construction begins with either double-stranded cDNA synthesized from RNA. Blunt-end DNA fragments are generated using a combination of fill-in reactions and exonuclease activity.

Adenylate 3' ends. A single 'A' nucleotide is added to the 3' ends of the blunt fragments to prevent them from ligating to one another during the adapter ligation reaction. A corresponding single 'T' nucleotide on the 3' end of the adapter provides a complementary overhang for ligating the adapter to the fragment. This strategy ensures a low rate of chimera (concatenated template) formation.

Ligate adapters. This process ligates multiple indexing adapters to the ends of the ds cDNA, preparing them for hybridization onto a flow cell. Adapters contain unique index sequences that are ligated to sample fragments allowing the samples to be pooled and then individually identified during downstream analysis. Ligated fragments of the range of 75bp are isolated via gel extraction and amplified using limited cycles of PCR.

Enrich DNA fragments. This process uses PCR to selectively enrich those DNA fragments that have adapter molecules on both ends and to amplify the amount of DNA in the library. The PCR is performed with a PCR primer cocktail that anneals to the ends of the adapters. The number of PCR cycles should be minimized to avoid skewing the representation of the library. PCR enriches for fragments that have adapters ligated on both ends. Fragments with only one or no adapters on their ends are by-products of inefficiencies in the ligation reaction. Neither species can be used to make clusters, as fragments without any adapters cannot hybridize to surface bound primers in the flow

cell, and fragments with an adapter on only one end can hybridize to surface bound primers but cannot form clusters.

Validate library. Illumina recommends performing the following procedures for quality control analysis on sample library and quantification of the DNA library templates.

- *Quantify library*. In order to achieve the highest quality of data on Illumina sequencing platforms, it is important to create optimum cluster densities across every lane of every flow cell. This requires accurate quantization of DNA library templates. Quantify your libraries using qPCR according to the Illumina Sequencing Library qPCR Quantification Guide.
- *Quality control*. First, load 1 µl of the resuspended construct on an Agilent Technologies 2100 Bioanalyzer using a DNA specific chip such as the Agilent DNA-1000. Then, check the size and purity of the sample. The final product should be a band at approximately 100 bp (for paired-end read libraries) (Figure 8).

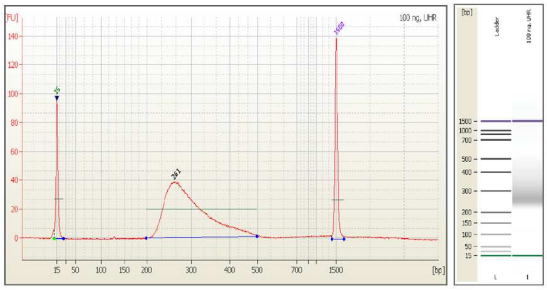

Fig 8. Example of TruSeq RNA sample prep v2 library size distribution and 260bp PCR product.

Normalize and pool libraries. Following the denaturation and amplification steps libraries can be pooled with up to 8 samples per lane (40 sample per flow cell) for cluster generation on cBot. This process describes how to prepare DNA templates that will be applied to cluster generation.

Processed samples can be amplified on cBot Automated Cluster Generation System and used with Illumina's next generation sequencing HiSeq 1000 instrument.

The Illumina flow cell accommodate multiplexed pool of five samples. During the cluster generation process, library fragments are hybridized to oligonucleotides, that correspond to the sequence of the adapters ligated during the sample preparation

stage, on the surface of the flow cell. Single-stranded, adapter-ligated fragments are bound to the surface of the flow cell exposed to reagents for polymerase-base extension. The hybridized fragments are copied by 3' extension and the original fragments are denatured leaving the immobilized copies on the flow cell surface. During isothermal amplification, the fragments loop over to hybridize to neighboring oligonucleotides to form a DNA bridge. The DNA bridge is copied and forms a double-stranded "DNA bridge", which is then denatured to form two ssDNA strands. Repeated denaturation and extension (35 times) results in localized amplification of single molecules in millions of unique locations across the flow cell surface. After the amplification process is completed, the reverse strands are removed by specific base cleavage, leaving the forward strands. After this process the flow cell contains >200,000,000 clusters with ~1,000 molecules/cluster, and is ready to be sequencing on the Illumina HiSeq (Figure 9).

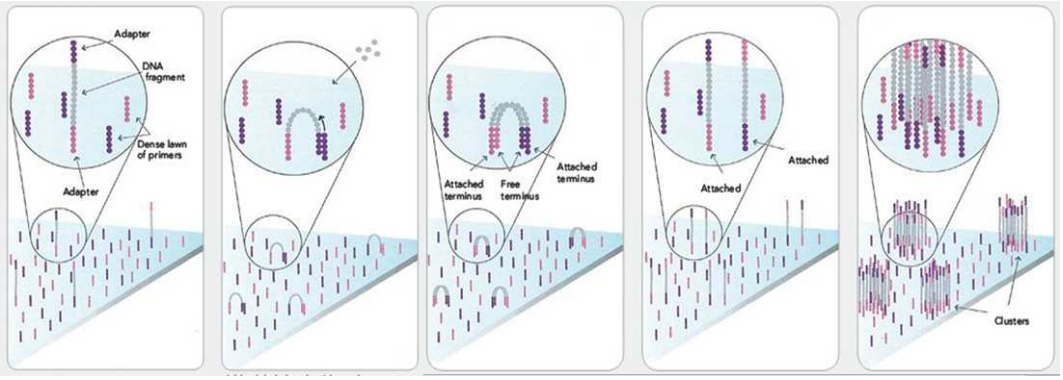

Fig 9. Cluster generation on a flowcell using the illumina cBot Cluster Generation System. Figure 8A shows hybridization of library fragments to the oligonucleotide on the surface of flow cell (A), the looping of a single-stranded copied fragment (B), subsequent amplification forming a double-stranded "DNA bridge"(C), and denaturation (D). The final illustration in Figure (E) shows a clonal cluster after the isothermal amplification and denaturation process has been repeated several times. Image from http://www.illumina.com/.

The Illumina sequencing run generates raw data in the form of a series of images, which are analyzed in three different steps: image analysis, base calling, and sequence analysis. Illumina Sequencing Control Software (SCS) runs on the instrument, to perform real-time image analysis and base calling. Any signal above background identifies the physical location of a cluster, and the fluorescent emission identifies which of the four bases was incorporated at that position. This image represent the data collected for the first base, cycles are repeated, one base at time, generating a series of images each representing a single base extension at a specific cluster. Base calls are derived with an algorithm that

identifies the emission color over time. At this time reports of useful Illumina reads range from 100 bases (Figure 10).

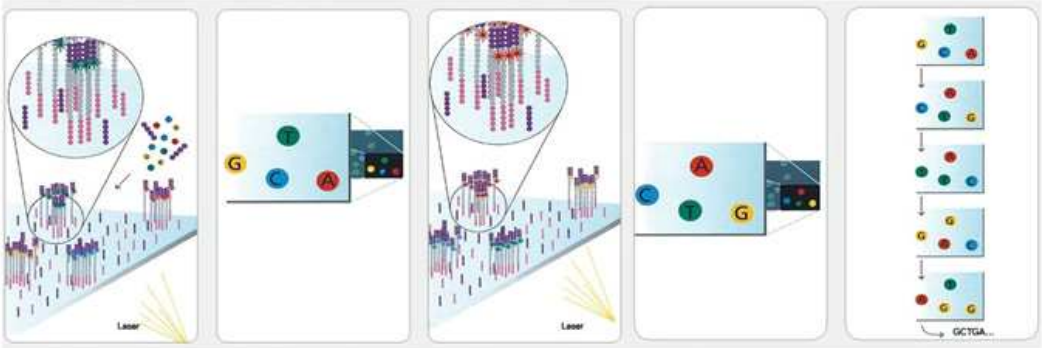

Fig 10. Illumina HiSeq 1000 sequencing chemistry. Figure 10A shows the nucleotide incorporation and laser excitation during the first cycle for four clusters. Figure 10B illustrates the first image base calls of the clusters. Figure 410C shows the nucleotide incorporation and laser excitation during a second cycle. Figure 10D is the second cycle base calls of the clusters. Figure 10E shows the base calls of the four clusters for the first five cycles of the sequencing run. Image from http://www.illumina.com/.

Because RNA-Seq reads are short, the first task is challenging. Current mapping strategies (Marioni JC. et al, 2008; Mortazavi A. et al, 2008; Sultan M. et al, 2008) include alignment procedures designed to localize reads to known exons in the genome.

3.6 Reads alignment

Sequence alignment is the first step in the analysis of a new RNA-Seq data. Read alignment was conducted uging the human genome sequence as reference. The purpose of alignment is to determine the location of the reference genome sequence that matches with the read sequence . Once each read has been aligned downstream bioinformatic tools estimate the sequence of all transcripts targeted by the reads (e.g. for digital gene expression).

Alignment was conducted using the computer program Bowtie. Bowtie (Langmead B. et al, 2009) enables ultrafast and memory-efficient alignment of large sets of sequencing reads to a reference genome without relying on existing annotation. Bowtie indexes the reference genome using a technique borrowed from data-compression, the Burrows-Wheeler transform (Burrows M. et al, 1994; Ferragina P. et al, 2001).

3.7 Transcript quantification

After aligning reads to the reference genome the gene expression levels of each transcript can be estimated.

Transcript quantification is the estimation of relative abundances at gene levels. RSEM (RNA-Seq by Expectation Maximization) is a software package for performing gene level quantification from RNA-Seq raw data (Li B. et al, 2011). The primary output of RSEM consists of one file for gene (or isoform)-level estimates. Abundance estimates are given in terms of two measures. The first is an estimate of the number of fragments that are derived from a given gene. The second measure of abundance is the estimated fraction of transcripts made up by a given gene.

3.8 Differential Expression (DE) analysis

Reads counts were used for statistical analysis using DESeq package of R/Bioconductor environment, an open source statistical package (http://www.r-project.org/), to determine differential expression at the gene level. Since RNA-seq technology results in discrete measurement of gene expression and statistical methods developed based on normal distributions are not directly applicable, researchers have developed several methods based on discrete distributions (binomial, Poisson, and negative binomial) to model the gene counts directly. For the DESeq analysis, differential expression testing was performed using the negative binomial test on variance estimated and size factor normalized data (Anders S. et al, 2010). Benjamini-Hochberg correction is performed to filter for a set of differentially expressed genes that have a certain false discovery rate (FDR) due to multiple hypothesis testing.

3.9 Gene Set Enrichment Analysis (GSEA)

Lists of differentially expressed genes identified as significant by statistical analysis of gene expression data and differentially expressed of almost 2 folds in ASDs respect to controls have been tested for pathways analysis.

In an attempt to uncover common fractions among the DE genes, we classified genes into gene ontology groups using DAVID (Dennis G. et al, 2003), bioinformatics resource, which is able to extract biological features/meaning associated with large gene lists.

Fisher's exact test was used to calculate a p-value determining the probability that the biological function assigned to the analysis is explained by chance alone. GSEA, or pathway enrichment analysis, provides a means of integrating the results of DE genes in a known molecular/biological pathway. The common gene set/pathway databases include KEGG (Kanehisa M. et al, 2000), Biocarta and the gene ontology (GO) databases (Gene Ontology Consortium, 2006).

3.10 Outlier-Gene Analysis

Outlier-gene analysis have been performed on cases' group (27 subjects) calculating the Z score for each gene by using the "scale" function in R. First was calculated the mean and SD for each expressed gene in cases. Then was selected a cutoff of 3.5 standard deviations to define whether a gene was an outlier in probands.

4. RESULTS

This study is part of a Telethon project which has been started in 2009 and involves different Italian clinical and research groups; it aims to analyze gene expression variations in ASD subjects, characterized for CNVs potentially involved in the onset of autism. LCLs transcriptome of 27 ASD probands and 23 health controls have been analyzed through Next Generation Sequencing technology (RNA Sequencing).

4.1 Genomic characterization

A total of 27 ASD subjects, carrying genomic imbalances, from the whole sample have been selected for gene expression analysis: eight subjects were selected for the presence of the 22q13.3 deletion syndrome, involving the last 5 Mb, and 19 subjects for the presence of potentially causative CNVs.

CNVs potentially involved in the onset of ASDs were selected for the presence of at least one of the following four criteria: genomic imbalances already associated with ASDs in literature or containing genes previously correlated to ASD; CNVs within genes involved in brain development, neuronal functioning, synaptic plasticity or, in general, genes expressed in central nervous system; regions involving unknown transcripts and gene-desert regions; rare CNVs (frequency < 1%).

Array-CGH analysis revealed 24 interesting copy number variants in 19 patients. In table 1 the patients and their relative CNVs are given.

Sample ID	CHR	Size (kb)	Type
867/10	chr20:14105407-14439297	333.89	del
1380/10	chr16:81640000-81672514	32.51	del
1414/10	chr9:1495137-1582306	87.17	del
1449/10	chr10:34925193-34961448	36.26	del
1494/10	chr9:7003793-8207768	1203.98	del
1531/10	chr2:160120557-160360995	240.44	del
1553/10	chr4:187179173-187355125	175.95	dup
	chr6:24390126-24491606	101.48	del
1593/10	chr17:30711469-30787396	75.93	del

	chr16:14876356-16199682	1323.33	dup
1613/10	chr6:71018624-71095644	77.02	dup
1649/10	chr5:155189273-155433297	244.02	del
	chr17:58403458-58441144	37.69	dup
	chrX:15238518-15604496	365.98	dup
1654/10	chr8:47062622-47857974	795.35	del
1675/10	chrX:140145377-140591773	446.4	dup
1692/10	chr3:61640711-61686826	46.12	del
1730/10	chr13:42312014-42721603	409.59	dup
1733/10	chr6:45107837-45138076	30.24	del
	chr22:24170487-24281368	110.88	del
1773/09	chrX:112662992-112854590	191.6	dup
1853/09	chr2:43574302-43749060	174.76	dup
3676	4p16.2	146	dup
3843	20p12.1	25	dup

Table 1. List of subjects with CNVs selected for gene expression analysis. CNV the nucleotide position in reference genome (GRCh37) or cytogenetic band position for each CNV, size (kb) and type (duplication, dup or a deletion, del) are given for each subject (sample ID).

Five CNVs are rare variants, 6 CNVs contains gene desert region, 3 CNVs contain genes and/or genomic regions already associated with autism, the last 10 CNVs involve genes with a neuronal function.

Thirteen CNVs are deletions with a medium size of 256.79 kb (min. 30.24 kb - max. 1203.98 kb) whereas 11 CNVs are duplications with an average size of 306.66 kb (min. 25 kb – max. 1323.33 kb).

All the 50 subjects selected for gene expression analysis, listed in table 2, included 27 patients (case group) and 23 non affected subjects (control group). These latter healthy siblings of probands or other health siblings enrolled in the study and matched for sex and age.

Autism patients with 22 deletion syndrome and those with autism, belonging to family collection, are listed in Table2 together with controls. Sex and age are also indicated.

Sample ID	Affection status	Sex	Age
ME1	Autism (22q13.3 deletion syndrome)	F	6
ME2	Autism (22q13.3 deletion syndrome)	F	5
ME3	Autism (22q13.3 deletion syndrome)	F	11

ME4	Autism (22q13.3 deletion syndrome)	M	11
ME5	Autism (22q13.3 deletion syndrome)	M	21
ME7	Autism (22q13.3 deletion syndrome)	M	5
ME8	Autism (22q13.3 deletion syndrome)	F	5
ME9	Autism (22q13.3 deletion syndrome)	F	15
867/10	Autism	M	7
1380/10	Autism	F	5
1414/10	Autism	F	3
1449/10	Autism	F	9
1494/10	Autism	M	18
1531/10	Autism	M	10
1553/10	Autism	M	3
1593/10	Autism	F	6
1613/10	Autism	M	8
1649/10	Autism	F	11
1654/10	Autism	M	3
1675/10	Autism	M	4
1692/10	Autism	M	6
1730/10	Autism	M	9
1733/10	Autism	M	5
1773/09	Autism	M	9
1853/09	Autism	M	21
3676	Autism	F	7
3843	Autism	M	6
40/11	Control	M	7
106/10	Control	M	7
130/11	Control	F	5
259/10	Control	M	8
1375/10	Control	F	14
1412/10	Control	F	9
1415/10	Control	M	8
1421/10	Control	M	9
1422/10	Control	F	18
1448/10	Control	M	6
1471/10	Control	F	7
1493/10	Control	F	4
1538/10	Control	M	3
1551/10	Control	F	5
1596/10	Control	M	2
1619/10	Control	F	13
1622/10	Control	F	10
1652/10	Control	M	16
1695/10	Control	F	6

1736/10	Control	F	12
1762/09	Control	M	13
1774/09	Control	F	7
1857/09	Control	F	22

Table 2. Expression analysis cohort. 50 subjects with affection status, sex and age.

The case group included 16 males and 11 females, with a mean age of 8 ± 5.06 years. The control group included 10 males and 13 females, with a mean age of 9 ± 4.96 years.

4.2 LCL achievement

LCL establishment and array CGH experiments were carried out at the same time. Thus, LCLs were established in our laboratory for 106 subjects, consisting of 67 ASD probands, and 39 healthy siblings. Starting from peripheral blood samples, leukocytes were separated and transformed by Epstein-Barr virus (EBV). Each LCL was maintained in culture for about six weeks.

For each LCL, at least 25 millions of cells, with a mean viability of 87%, have been stocked in liquid nitrogen, while 100 millions cells for line (88% mean viability) were pelleted and resuspended in 10 ml TRIzol® for gene expression experiments.

4.3 RNA obtainment, quantification and quality analysis

RNA from the selected 50 LCLs has been extracted in triplicate showing a mean yield of 75.85 µg (min 36.24 µg and max 105.72 µg) concentration of 1914 ng/µL (min 896 ng/µL and max 2925 ng/µL), an absorbance A260/280 ratio of 1.97 and an absorbance A260/230 ratio of 2.1.

An additional DNAse digestion step was performed to ensure that the samples were not contaminated with genomic DNA. After DNase treatment, the average yield of RNA was 6,45 µg (min 4.84 µg and max 8.21 µg)174.2 ± 13 ng/µl (min 130.82 ng/µl and max 222.02 ng/µl) and each RNA sample had an A260:A280 ratio above 1.95 and A260:A230 ratio above 1.70 (Table3).

RNA is a thermodynamically stable molecule, which is, however, rapidly digested in the presence of the nearly ubiquitous RNase enzymes. The average RIN value for the 50 samples, after DNase treatment, is 9.9 (min 9.3 – max 10), based on a scale from 1 to 10

(Table 3). The RIN value recommended for obtaining meaningful gene expression data is greater than or equal to 8. As an example, the following figure shows the RNA Bioanalyzer trace of the 1774/09 sample (Figure 11).

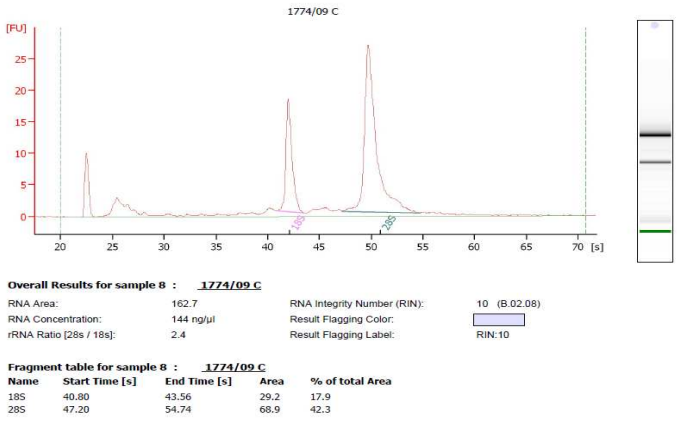

Fig 11. Agilent BioAnalyzer output of 1774/09 sample. The electropherogram and associated image of virtual agarose separation are given. In the output RIN value, RNA concentration and 28S/18S ratio are indicated.

Sample ID	Yield (µg)	ng/ul	260/280	260/230	RIN
40/11	6.64	179.36	2.01	1.93	9.7
106/10	5.33	144	1.98	1.77	9.8
130/11	6.72	181.75	2.01	1.87	9.6
259/10	6.49	175.49	2.01	1.98	10
867/10	6.54	176.83	2.00	1.85	9.9
1375/10	6.22	168.21	2.01	1.93	10
1380/10	6.90	186.47	1.96	1.90	9.8
1412/10	6.24	168.65	1.98	1.77	9.9
1414/10	6.56	177.31	1.99	1.96	10
1415/10	6.35	171.6	2.02	1.95	10
1421/10	6.59	178.23	2.00	1.97	9.7
1422/10	6.56	177.27	2.00	1.87	9.4
1448/10	6.49	175.29	1.99	1.98	9.9
1449/10	6.36	171.78	1.99	1.90	9.4
1471/10	5.86	158.31	1.99	2.00	9.6
1493/10	6.90	186.46	2.00	1.95	9.9
1494/10	6.47	174.95	2.00	1.95	10
1531/10	6.41	173.17	2.00	1.93	9.8
1538/10	4.84	130.82	1.99	1.86	9.8
1551/10	6.45	174.36	1.99	1.99	10
1553/10	6.47	174.96	1.97	1.89	9.3
1593/10	6.91	186.89	1.96	1.92	9.6
1596/10	6.29	169.92	1.98	1.87	9.8
1613/10	6.56	177.18	2.00	2.00	10
1619/10	6.30	170.19	1.96	1.82	10
1622/10	6.50	175.56	2.00	1.96	10
1649/10	5.89	159.3	1.99	1.76	9.8
1652/10	6.60	178.46	1.99	1.94	9.5

1654/10	6.35	171.51	1.94	1.95	9.3
1675/10	6.79	183.4	2.01	2.01	10
1692/10	7.14	192.93	1.99	1.99	9.9
1695/10	6.24	168.75	1.97	1.85	9.4
1730/10	7.16	193.55	2.00	1.94	9.7
1733/10	8.21	222.02	2.01	2.04	9.9
1736/10	6.50	175.81	1.96	1.95	9.3
1762/09	6.53	176.61	1.99	1.99	10
1773/09	6.29	169.89	2.00	1.88	9.7
1774/09	6.48	175.18	2.01	1.98	10
1853/09	6.46	174.67	1.96	1.88	9.8
1857/09	6.76	182.58	2.00	1.89	10
ME1	5.79	156.44	1.99	1.80	9.7
ME2	6.33	171.14	1.98	1.94	9.9
ME3	6.65	179.77	2.00	1.70	9.8
ME4	5.72	154.55	1.99	1.67	10
ME5	6.57	177.5	1.99	1.89	10
ME7	5.98	161.5	2.01	1.84	10
ME8	6.22	168.11	2.01	1.93	9.8
ME9	6.22	168.2	2.02	1.94	9.5
3676	6.85	185.26	2.00	1.86	10
3843	6.58	177.92	1.97	1.95	9.6

Table 3. RNA quantity and quality after DNase treatment for each sample. Total amount of RNA (µg), RNA concentration (ng/µl), 260/280 and 260/230 ratio and RIN value are given.

4.4 RNA Sequencing

We sequenced 50 RNA specimens, obtaining an average of 20 million reads per sample, with a range from 10.2 million to 57.4 million reads. 22000, out of 37500 genes present in human db, resulted to be expressed in at least one sample (by at least one read). The top 10 most expressed genes are EEF1A1, ACTB, GAPDH, ACRG1, EEF2, RPL3, CD74, TMSB4X, ENO1 and RPLP0 (table 4), accounting for the 7.11% of the total reads per samples. The most expressed gene is EEF1A1, eukaryotic translation elongation factor 1 alpha 1, which is covered on average by 41000 reads for each sample.

Gene ID	baseMean	foldChange	log2FoldChange	pval	gene symbol
ENSG00000156508	410944.42	1.00	0.01	0.93	EEF1A1
ENSG00000075624	186797.85	1.06	0.08	0.21	ACTB
ENSG00000111640	170848.71	1.06	0.08	0.37	GAPDH
ENSG00000184009	110510.42	0.99	-0.02	0.86	ACTG1
ENSG00000167658	98452.05	0.98	-0.03	0.81	EEF2
ENSG00000100316	91541.19	1.02	0.03	0.69	RPL3
ENSG00000019582	91303.86	1.02	0.02	0.54	CD74
ENSG00000205542	91254.01	1.06	0.08	0.23	TMSB4X
ENSG00000074800	88252.13	1.01	0.02	0.82	ENO1
ENSG00000089157	82516.02	1.01	0.02	0.74	RPLP0

Table 4. List of the most expressed genes in LCLs. Ensemble gene name, base mean reads, fold change (number describing how much a expression changes in cases respect controls), log2fold change, p-value and gene symbol are given. Gene description is given in Appendix 1.

4.5 Differentially expressed genes (DEG) and gene set enrichment analysis (GSEA)

Analysis of differential expression genes has been performed with the case-control design including 27 ASD and 23 control samples.

This analysis identified 295 differentially expressed genes (DEG) with a statistically significant nominal p-value ($p< 0.05$). The top 10 most differentially expressed genes between cases and controls are RP11-632C17_A.1, ZNRD1, SFRP1, RNF183, TRIM26, HLA-DRA, VPS37B, FTL, TRIM27 and PPP1R18 (Table 5). However, after correction of multiple tests no gene was significantly DE.

Gene ID	baseMean	foldChange	log2FoldChange	pval	gene symbol
ENSG00000230202	205.97	172.44	7.43	3.24E-05	RP11-632C17__A.1
ENSG00000224859	20.21	0.01	-6.75	0.0004	ZNRD1
ENSG00000104332	61.02	0.08	-3.72	0.0005	SFRP1
ENSG00000165188	31.91	0.24	-2.08	0.0017	RNF183
ENSG00000230230	122.09	0.10	-3.26	0.0033	TRIM26
ENSG00000206308	10936.83	1.76	0.82	0.0034	HLA-DRA
ENSG00000139722	655.03	1.41	0.50	0.0037	VPS37B
ENSG00000087086	29840.91	1.19	0.25	0.0046	FTL
ENSG00000237071	100.69	0.12	-3.11	0.0047	TRIM27
ENSG00000231247	186.46	0.07	-3.87	0.0049	PPP1R18

Table 5. The top ten differentially expressed genes between cases and controls. Gene description is given in Appendix 1.

The 266 genes, which are differentially expressed of almost 2 folds in cases respect to controls, have been selected for DAVID pathway analysis. 138 out of 266 genes are recognized by the tool for the following analysis. KEGG functional classification of genes demonstrated that 10 different pathways are over represented (Table 6). One of the enriched pathways is reported in Figure 12.

Term	Count	%	pvalue	padj
hsa05320:Autoimmune thyroid disease	7	5.07	1.38E-06	8.15E-05
hsa05310:Asthma	6	4.35	1.61E-06	9.52E-05
hsa04612:Antigen processing and presentation	8	5.80	1.66E-06	9.78E-05
hsa05330:Allograft rejection	6	4.35	4.94E-06	2.91E-04
hsa05322:Systemic lupus erythematosus	8	5.80	5.47E-06	3.22E-04
hsa05332:Graft-versus-host disease	6	4.35	7.42E-06	4.38E-04
hsa04940:Type I diabetes mellitus	6	4.35	1.08E-05	6.37E-04
hsa04672:Intestinal immune network for IgA production	6	4.35	2.33E-05	0.0013748
hsa05416:Viral myocarditis	6	4.35	1.42E-04	0.0083314
hsa04514:Cell adhesion molecules (CAMs)	7	5.07	3.28E-04	0.0191427

Table 6. DAVID annotation chart of DEG, showing the enriched functional annotation pathways associated to gene list, listed according to their enrichment p-value. Number of genes (count) involved in given term, percentage respect to input gene list (%) and padj (adjusted p-value) are shown.

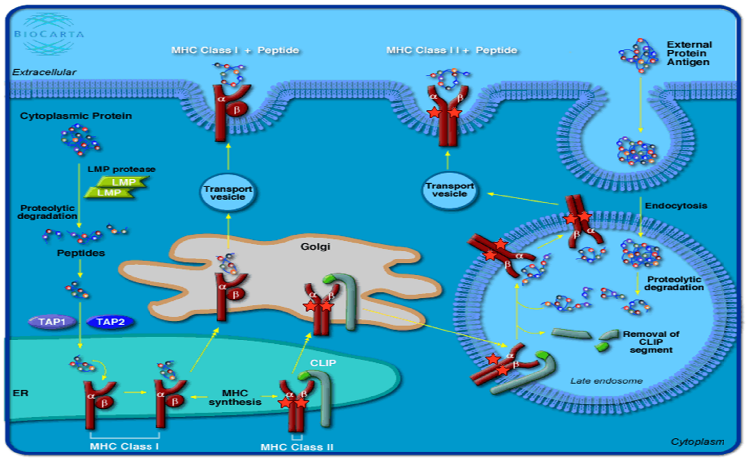

Fig 12. Antigen processing and presentation pathway. Major histocompatibility complex, class II, genes are indicated by a red star. (Source: BIOCARTA databases).

Six genes are shared in the four most significant enriched pathways and all of them belong to HLA gene region (chromosome 6p21.3). Thus, focus on all HLA genes resulted differentially expressed in DEG analysis and through the Ensembl gene name it was possible to attribute a haplotype to seven out of 10 HLA genes (Table 7).

Gene ID	foldChange	pval	Gene symbol	Haplotype
ENSG00000241106	0.55	0.036	HLA-DOB	6p21.3
ENSG00000232062	28.78	0.042	HLA-DQA1	Ssto
ENSG00000206301	0.3	0.013	HLA-DQA2	Mcf
ENSG00000237541	0.08	0.041	HLA-DQA2	6p21.3
ENSG00000231286	1.95	0.017	HLA-DQB1	Mcf
ENSG00000232351	1.74	0.037	HLA-DQB1-AS1	Mcf
ENSG00000206308	1.76	0.003	HLA-DRA	mcf/ssto
ENSG00000228080	3.03	0.029	HLA-DRB1	ssto
ENSG00000227669	1.7	0.044	HLA-H	dbb
ENSG00000206341	1.4	0.047	HLA-H	6p21.3

Table 7. List of HLA-DEG. Haplotype name or map position are given. Gene description is given in Appendix 1.

4.6 DEG analysis and GSEA on 22q13.3qter

A differential expression analysis was performed on a subgroup of patients characterized for the deletions at the chromosome 22q13.3qter. Eight ASDs cases (3 males and 5 females; mean age 10 ± 5.97) and eight controls (2 males and 6 females; mean age 10 ± 5.96).

The top 10 most differentially expressed genes between cases and controls were FRMD4B, BRD2, PNRC2, HMGA2, RP11-290F5.2, HLA-DMA, HIST1H1C, RP11-366L20.2, SHROOM3 and ZBED4 (Table 8). In this analysis three genes resulted differentially expressed in ASDs respect to controls with a significant p-adjusted value (p-adj < 0.05).

gene ID	baseMean	foldChange	log2FoldChange	pval	padj	gene symbol
ENSG00000114541	120.59	0.17	-2.52	4.22E-08	0.001	FRMD4B
ENSG00000234704	792.65	23.75	4.57	1.08E-07	0.002	BRD2
ENSG00000189266	1607.62	0.31	-1.68	7.97E-07	0.009	PNRC2
ENSG00000149948	57.81	0.05	-4.45	1.39E-05	0.113	HMGA2
ENSG00000180712	515.47	0.4	-1.31	2.99E-05	0.194	RP11-290F5.2
ENSG00000241394	405.66	0.02	-5.72	3.70E-05	0.2	HLA-DMA
ENSG00000187837	843.54	2.27	1.18	4.83E-05	0.224	HIST1H1C
ENSG00000197301	7.82	0.08	-3.59	6.71E-05	0.272	RP11-366L20.2
ENSG00000138771	165.7	0.2	-2.32	0.0001	0.364	SHROOM3
ENSG00000100426	928.64	0.5	-1.01	0.0003	1	ZBED4

Table 8. The most DEG in 22q13.3qter analysis. Ensemble gene name, fold change value, log2 Fold Change, p-value, p-value adjusted and gene name. Gene description is given in Appendix 1.

Four hundred and forty eight genes with a significant nominal p-value and differentially expressed of almost 2 times in ASDs respect to controls were submitted to DAVID functional annotation chart. The tool recognized 341 genes. Results of the pathway enrichment analysis are listed in table 9. Analysis a total of 10 pathways were identified using KEGG database.

Term	Count	%	PValue	padj
hsa04612:Antigen processing and presentation	9	2.65	2.95E-04	0.036
hsa05322:Systemic lupus erythematosus	9	2.65	9.74E-04	0.115
hsa04940:Type I diabetes mellitus	6	1.77	0.0016	0.186
hsa04360:Axon guidance	9	2.65	0.0052	0.478
hsa05330:Allograft rejection	5	1.47	0.0063	0.546
hsa05332:Graft-versus-host disease	5	1.47	0.0084	0.651
hsa05416:Viral myocarditis	6	1.77	0.0157	0.862
hsa04672:Intestinal immune network for IgA production	5	1.47	0.0184	0.902
hsa05320:Autoimmune thyroid disease	5	1.47	0.0211	0.93
hsa04514:Cell adhesion molecules (CAMs)	7	2.06	0.0565	0.999

Table 9. DAVID annotation chart of DEG analysis on 8 ASD subjects with 22qter deletion and 8 controls.

4.7 Outlier genes

The term “outlier genes” is used to reflect that the dysregulated gene is contained within a structural variation and shows significant alteration in gene expression. The identification of outlier genes has been performed in all the 27 ASD cases. We have identified 68 dysregulated genes in 17 ASD subjects (Table 10).

ID sample	gene ID	CHR	POS	gene symbol
867/10	ENSG00000114473	3	197615946	IQCG
867/10	ENSG00000112511	6	33378176	PHF1
867/10	ENSG00000204531	6	31130253	POU5F1
867/10	ENSG00000211934	14	106452671	IGHV1-2
1414/10	ENSG00000081985	1	67773047	IL12RB2
1414/10	ENSG00000118689	6	108977549	FOXO3
1414/10	ENSG00000179165	6	36358778	PXT1
1414/10	ENSG00000167123	9	131174030	CERCAM
1414/10	ENSG00000171840	12	673462	NINJ2
1414/10	ENSG00000088387	13	99445741	DOCK9
1414/10	ENSG00000088387	13	99445741	DOCK9
1414/10	ENSG00000170468	14	73957644	AC005280.1

1414/10	ENSG00000101335	20	35169887	MYL9
1449/10	ENSG00000214776	12	9620148	RP11-726G1.1
1494/10	ENSG00000129474	14	23447821	JUB
1494/10	ENSG00000205609	16	28390900	EIF3CL
1531/10	ENSG00000162738	1	160389283	VANGL2
1531/10	ENSG00000038427	5	82849181	VCAN
1531/10	ENSG00000106268	7	2281857	NUDT1
1531/10	ENSG00000171914	15	62682725	TLN2
1553/10	ENSG00000196284	6	44777054	SUPT3H
1553/10	ENSG00000182534	17	74680745	MXRA7
1593/10	ENSG00000157873	1	2487078	TNFRSF14
1593/10	ENSG00000145495	5	10353815	MARCH6
1593/10	ENSG00000107372	9	74966341	ZFAND5
1593/10	ENSG00000230224	10	102008028	PHBP9
1593/10	ENSG00000187066	11	64852451	AP003068.6
1593/10	ENSG00000227825	12	98847613	SLC9A7P1
1593/10	ENSG00000119688	14	74766602	ABCD4
1593/10	ENSG00000166783	16	15729851	KIAA0430
1593/10	ENSG00000250251	16	15219099	PKD1P6
1593/10	ENSG00000179889	16	15068599	PDXDC1
1593/10	ENSG00000182149	16	71958311	IST1
1593/10	ENSG00000255185	16	70096958	RP11-106J23.2
1593/10	ENSG00000237296	16	22448329	RP11-368J21.1
1593/10	ENSG00000247061	17	2574352	AC005696.2
1593/10	ENSG00000171791	18	60790579	BCL2
1593/10	ENSG00000101346	20	30804468	POFUT1
1593/10	ENSG00000197976	X	1710518	AKAP17A
1613/10	ENSG00000131016	6	151561134	AKAP12
1654/10	ENSG00000185090	1	38259474	MANEAL
1654/10	ENSG00000186603	1	45792545	HPDL
1654/10	ENSG00000137198	6	16290176	GMPR
1654/10	ENSG00000104267	8	86376243	CA2
1654/10	ENSG00000127564	16	3018103	PKMYT1
1654/10	ENSG00000174292	17	7284365	TNK1
1654/10	ENSG00000005206	19	2328629	AC004410.1
1654/10	ENSG00000188130	22	50683879	MAPK12
1675/10	ENSG00000198794	15	75287939	SCAMP5
1692/10	ENSG00000197262	17	34639793	CCL4L2
1730/10	ENSG00000185668	1	38509523	POU3F1
1730/10	ENSG00000116741	1	192778273	RGS2

1730/10	ENSG00000082781	3	124567206	ITGB5
1730/10	ENSG00000137731	11	117672445	FXYD2
1730/10	ENSG00000182253	15	99645286	SYNM
1730/10	ENSG00000100979	20	44527399	PLTP
1733/10	ENSG00000117115	1	17393256	PADI2
1733/10	ENSG00000236756	10	75006946	C10orf103
1733/10	ENSG00000211895	14	106173457	IGHA1
1733/10	ENSG00000188636	22	44888452	LDOC1L
1733/10	ENSG00000130822	X	152952080	PNCK
1773/09	ENSG00000182095	7	5346421	TNRC18
1773/09	ENSG00000156299	21	32638612	TIAM1
1853/09	ENSG00000015568	2	110550335	RGPD5
1853/09	ENSG00000134762	18	28570052	DSC3
3676	ENSG00000204469	6	31588450	PRRC2A
3676	ENSG00000234745	6	31321649	HLA-B
3843	ENSG00000147394	X	152082986	ZNF185

Table 10. List of the 68 dysregulated genes. For each sample, the Ensembl gene name, the position on chromosome and gene symbol are indicated. Gene description is given in Appendix 1.

Among them, the three outlier genes KIAA0430 (chromosome 16: 15,688,243-15,737,023), PDXDC1 (chromosome 16:14,876,356-16,199,683) and PKD1P6 (chromosome 16: 15,219,099-15,248,421) are located in a region that was found duplicated in one autistic subject (1593/10) and reported as an ASD potentially involved CNV on chromosome 16 (already found in ASDs) (Table 9). The position of the duplication is chr16:14876356-16199682 and has a size of 1323.326 KB.

ENGname	CHR	POS	gene name	ID sample	CNV
ENSG00000166783	16	15688243-15737023	KIAA0430	1593/10	chr16:14876356-16199682
ENSG00000179889	16	15068833-15131552	PDXDC1	1593/10	chr16:14876356-16199682
ENSG00000250251	16	15219099-15248421	PKD1P6	1593/10	chr16:14876356-16199682

Table 9. The three outlier genes located within a duplicated CNV. The Ensembl gene name, the chromosome, the nucleotide position, the gene symbol, Id sample and the position of the CNV.

5. DISCUSSION

In this study, 19 ASD individuals, carrying genomic imbalances, potentially involved in the disease, 8 individuals with 22qter deletion syndrome and 23 controls, have been analyzed for whole gene expression profile by next generation sequencing (RNA-Seq), to identify altered pathway between cases and controls and better understand the pathological role of the CNVs in ASDs. Expression experiments were carried out using mRNA extracted from LCLs.

5.1 Technical controls

From the methodological point of view, both quality and quantity of RNA were found to be optimal. The mean RIN value observed in the 50 samples is 9.9, which is much higher compared to the threshold of 8, usually considered as standard quality cut-off for RNA-Seq analyses. These data confirmed that starting materials were good enough for subsequent experiments. Although it is well recognized that biases and sequence errors can be introduced during RNA fragmentation and cDNA synthesis (Kassahn KS. et al, 2011), technical quality controls associated with library generation and deep sequencing were exceeded.

5.2 Copy Number Variants

The discovery of rare and recurrent CNVs as important pathogenic mutations in ASDs was a watershed in ASD genetics. Recurrent CNVs such as those at 16p11.2, 22q11.2, 1q21.1, 7q11.23 and 15q11-q13 show statistically significant association with the disease (Sebat J . et al, 2007; Marshall CR. et al, 2008; Ramalingam A. et al, 2011). Our incomplete understanding of structural polymorphism, such as CNVs, and the appreciation that ASDs show a high degree of clinical variance and incomplete penetrance is blurring clear-cut genotype-phenotype correlations. The advent of commercially available microarrays has facilitated the implementation of molecular karyotyping in laboratories. The resolution offered by these arrays allows the detection of larger aberrations with high confidence, but at the same time leads to the detection of a high number of smaller CNVs for which causality often remains to be elucidated.

Since each of the genomic variants were private (observed sporadically and not common to two or more subjects) in our study, it was not possible to state a correlation between CNV and disease status. Lack of recurrence may in fact reflect an underlying reality that autistic behavior can result from many different genomic defects. However, 3 of the observed CNV (located in 20p12.1, 4p16.2 and 16p13.11 regions) were already associated with ASD in other studies (Sebat J . et al, 2007; Marshall CR. et al, 2008; Ramalingam A. et al, 2011)

5.3 Differentially expressed genes (DEG) and gene set enrichment analysis (GSEA)

DEG and GSEA were performed on the whole cohort (27 cases and 23 controls). Gene expression information was integrated within our study to check whether a specific biological pathway is associated with the disorder. The aim is to attempt to identify those pathways that are statistically over-represented or under-represented in DEG list.

Even if 295 genes were DE, no significant adjusted p-value was observed after correction for multiple test. Some explanations of this result can be attributed to the fact that there is no real difference between cases and controls or the method of correction used was too severe or that the study is underpowered to find significant differences in the two groups. Thus, we submitted to GSEA a set of 266 DEG that showed a FC of 2 and a p-value < 0.05. This analysis identified a total of 10 enriched pathways (Table 4), that are predominately related to autoimmune disease and antigen processing and presentation pathways.

It is no surprise to see immune gene associations in ASDs, as numerous researchers have reported immune abnormalities in autism for over 20 years. It has become increasingly clear that inflammatory processes are related with autism and recent genetic research has also associated numerous immune function genes with autism (Gregg JP. et al, 2008; Enstrom AM. et al, 2009; Morgan JT. et al, 2010; Voineagu I. et al, 2011).

Moreover, Sweeten and colleagues (2003a) observed that monocyte counts were increased in autistic children suggesting that the immune system was over activated in ASDs. Two studies found that the prevalence of autoimmune disorders in the families with autistic children was higher than in control subjects (Comi AM. et al, 1999; Sweeten TL. et al, 2003b).

Interestingly, a metanalysis study found that SNPs of human leukocyte antigen (HLA) region were associated with schizophrenia (Stefansson H. et al, 2009). Current research is increasingly demonstrating a role for HLA proteins in neural cell interactions, synaptic function, cerebral hemispheric specialization, central nervous system (CNS) development (Xiao BG. et al, 1998; Huh GS. et al, 2000; Boulanger LM. et al, 2004; Cullheim S. et al, 2007; Ohtsuka M. et al, 2008) and even neurological disorders (Bailey SL. et al, 2006).

It has been reported that autism subjects often show associations with HLA genes/haplotypes, suggesting an underlying dysregulation of the immune system mediated by HLA genes (Torres AR. et al, 2002; Chien YL. et al, 2012; Mostafa GA. et al, 2013).

The HLA region on chromosome 6p21 (about $4x10^6$ bp) is of major interest in basic research as well as medicine as genes/proteins in this region are involved in many biological processes such as histocompatibility, inflammation, ligands for immune cell receptors, and the complement cascade. The HLA region has 20 typical HLA genes and 112 nontypical HLA genes that are inherited together as frozen blocks of DNA called ancestral or extended haplotypes.

We focus on 7 HLA genes and 1 HLA pseudogene that resulted differentially expressed (DE) (Table 5), confirming the implication of HLA loci in our ASD dataset. Interestingly, we found that MHC class I (i.e. HLA-H pseudogene) and MHC class II (i.e. HLA-DOB, HLA-DQA1, HLA-DQA2, HLA-DQB1, HLA-DQB1-AS1, HLA-DRA and HLA-DRB1) genes are differentially expressed in cases respect to controls. Moreover, since complete DNA sequences have been published for 8 of the most common haplotypes in the European population (Horton R. et al, 2008), it was possible to attribute reference haplotypes for some of aforementioned HLA genes. For example, the highest expression of HLA-SSTO haplotype in cases compared to controls (FC: 28.78; p-value: 0.042) suggests that it is the most frequent haplotype in our samples.

HLA-DM and –DQ, class II proteins have structures like classical antigen-binding HLA proteins and work in the cytoplasm of antigen presenting cells. It is now known that DQ2 is a poor substrate for DM and it has been proposed that antigen presentation in the thymus and periphery can be affected by impaired DQ-DM interactions so as to

promote autoimmune disease (Hou T. et al, 2011). The altered expression of HLA-DQ in our ASD subjects is a proof in support of this theory. In ASDs the over-expression of HLA-DQA1 and HLA-DQB1 and HLA-DQB1-AS could be a compensative response to the under-expression of HLA-DQA2 aiming to form the heterodimer .

5.4 DEG and GSEA on 22q13.3qter

Classification of autistic patients on the basis of genotypic and phenotypic information is one effective way to identify more homogeneous subgroups and hasten the identification of genes underlying autism (Folstein SE. et al, 2001; Belmonte MK et al, 2004; Veenstra-Vanderweele J. et al, 2004; Muhle R. et al, 2004). So we made a differential expression analysis on the subgroup of the 8 ASD patients with the deletion 22q13.3qter, against 8 sex and age matched controls.

Affected individuals with 22q13.3 deletions were enrolled because SHANK3 gene, a strong candidate gene for the neurobehavioral symptoms, is located in that deleted region. SHANK3 gene has became the best candidate gene for the neurological deficits (developmental delay and absent speech) in the last years (Bonaglia MC. et al, 2001).

In our dataset SHANK3 gene results under-expressed with a FC of 0.8, a p-value of 0.74 and a baseMean of 5.83 reads. Probably, we were not able to detect any significant difference due to the fact that the gene is less expressed in LCLs than in brain and because lymphoblastoid cell lines are not neuronal cells and some of our findings might not reflect the pathophysiology in ASD brains.

However, we also explore the functional impact or 22q13.3 deletion. Previously, it was unclear which genes are dysregulated nearby the 22qter region or whether there is a common expression signature shared by 22q13.3 cases. We found that 6 genes, BRD1 (p-value 0.002), PIM3 (p-value 0.032), TRABD (p-value 0.016), SBF1 (p-value 0.007), ZBED4 (p-value 0.0003) and CHKB (p-value 0.024), surrounding SHANK3 gene region in the nearest 500 kb, are all under-expressed with a significant p-value. Our analysis, then, shows a positive correlation between expression level and deletion.

The most differentially expressed genes are: FRMD4B (FERM domain containing 4B), BRD2 (bromodomain containing 2) and PNRC2 (proline-rich nuclear receptor coactivator 2). FRMD4B and PNRC2 genes are located respectively to 3p14.1 and 1p36.11 loci. BRD2

gene encodes a transcriptional regulator that belongs to the BET (bromodomains and extra terminal domain) family of proteins. This protein associates with transcription complexes and with acetylated chromatin during mitosis and has been implicated in juvenile myoclonic epilepsy. Notably, BRD2 gene maps to the major histocompatability complex (MHC) class II region on chromosome 6p21.3, the same region identified in the DE study of the whole sample. However, sequence comparison suggests that the protein is not involved in the immune response.

GSEA demonstrates that there is specific enrichment of autoimmune disorders pathway (systemic lupus erythematosus, type I diabetes mellitus) in probands, supporting the hypothesis that ASDs are associated with immune genes.

Moreover, this analysis highlights the axon guidance pathway. Axon guidance (also called axon pathfinding) is a subfield of neural development concerning the process by which neurons send out axons to reach the correct targets. Axons often follow very precise paths in the nervous system. Genetics and biochemistry have identified a large set of molecules that affect axon guidance. How all of these pieces fit together is less understood.

Previous studies found decreased expression of axon-guidance proteins in the brains of subjects with autism, and suggest that dysfunctional axon-guidance protein expression may play an important role in the pathophysiology of autism (Suda S. et al, 2011). Our differential expression analysis revealed that significant difference concerning neuronal growth is detectable in LCLs.

5.5 Outlier genes

We next asked whether CNVs could cause transcriptional changes and, conversely, whether dysregulated genes can aid in characterizing structural chromosomal variation. To assess which dysregulated genes could be associated to higher risk of disease, we investigated expression variance in each of the ASD probands (27 cases) and we looked for genes (outliers) with an expression deviation of at least 3.5 SD. To explore the functional impact of CNVs in ASDs at a genome-wide scale, our query searched for the overlap between location of structural-variation data and transcriptional data as well.

We found 68 outlier genes in 17 of the 27 ASD subjects. In particular, we observed an hyper-expressed region corresponding to a CNV presenting a duplication. In fact, three outlier genes (KIAA0430, PDXDC1 and PKD1P6), belonging to a unique subject (1593/10), cluster within a potentially involved CNV of several KB on chromosome 16, already found associated in autism (Ramalingam A. et al, 2011).

PKD1P6 (polycystic kidney disease 1 - autosomal dominant - pseudogene 6) is a known pseudogene, that has 6 alternative processed transcripts that don't contain an open reading frame (ORF) and cannot be translated. KIAA0430 (limkain-b1) encodes a putative peroxisomal protein that appears to be an essential regulator of oogenesis required for female meiotic progression. It acts by down-regulating RNA transcripts, either at transcriptional of post-transcriptional level. In humans, it may be autoantigenic. PDXDC1 encodes a pyridoxal-dependent decarboxylase domain containing 1. Even if RNA-Seq expression data from Illumina's Human BodyMap 2.0 project detected expression of these three genes in brain tissue (http://www.ensembl.info/blog/2011/05/24/human-bodymap-2-0-data-from-illumina/), the three genes we found within the duplication are no obvious candidates for autism.

The 16p13.11 aberration is located in areas with high locus control regions (LCR) content. There are several paralogous repeats at the proximal and distal breakpoint (She X. et al, 2004). The repeats are in direct orientation and non allelic homologous recombination (NAHR) between these LCRs seems to be a likely explanation for the recurrence of this rearrangement (Finelli P. et al, 2004; Ullmann R. et al, 2007; Ballif BC. et al, 2007). Evidence for the involvement of genes on 16p13 has also come from linkage and association studies that have identified autism susceptibly loci in this region (Lamb JA. et al, 2005; Philippe A. et al, 1999; Weiss LA. et al, 2008). Ulmann et al (2007) proposed two genes (NDE1; nudE nuclear distribution E homolog 1 and NTAN1; N-terminal asparagine amidase) plausibly involved in autism. We recognize the stringency of the analysis is very high. Lowering the strength could was possible to find some relevant candidates located in the region.

In this analysis, the most of outlier genes do not locate within CNVs. We focus, in fact, only on the CNVs potentially causative of ASDs. Further analysis will be performed using the whole CNVs dataset obtained from array-CGH.

5.5 Conclusions

Results highlight different immunological (autoimmune diseases, antigen processing and presentation) and cell adhesion molecular pathways associated to ASDs. GSEA conducted on 22qter subjects suggest an involvement of axon guidance pathway, which contributes to the assembly and function of neural circuitry, confirming that LCLs exhibit genetic biomarkers relevant to autism.

We performed DNA-mRNA correlations based on genomic and gene expression data. We use the outlier gene approach to identify candidate ASD regions. The 16p13.1 duplication contains three expressed genes, all of which are over-expressed in ASD LCLs. The outlier study demonstrates the utility of gene-expression analysis in evaluating the functional consequences of rare functional structural variations in a human neuropsychiatric disease (ASDs). Further analysis will be also extended on the entire CNVs dataset, aiming to identify correlation between gene expression and structural variations. Moreover, analysis designed specifically to examine associations between cases and controls, as well as their family members, will be performed to study the segregation of the genetic factors.

The strength of this study is that transcriptome analysis was performed on a large number of samples using a high-throughput sequencing technology (RNA-Seq). The use of RNA-Seq offers several key advantages: RNA-Seq is not limited to detect known transcripts and can reveal the precise location of transcription boundaries (i.e. spicing junctions), at a single-base resolution. In addition, RNA-Seq can also reveal sequence variations (for example SNPs and indel) in the transcribed regions, gene alleles, differently spliced transcripts, non-coding RNAs and gene fusions. All these features will be used to perform association studies.

This research helps understanding the genetic bases for ASD pathophysiology and unravelling potential new pathways involved in ASDs. Together with studies addressing epigenetic modifications and comprehensive analysis of environmental risk factors, these pieces of information can be better integrated to improve our understanding of the molecular basis of ASDs, and foster the development of early preventive and corrective strategies.

6. REFERENCES

- Abrahams BS, Geschwind DH. Advances in autism genetics: on the threshold of a new neurobiology. *Nat Rev Genet.* 2008 May; 9(5):341-55.

- American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 4th ed, text revision. Washington, DC: *American Psychiatric Association*; 2000.

- Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol*. 2010;11(10):R106.

- Bailey A, Le Couteur A, Gottesman I, Bolton P, Simonoff E, Yuzda E, Rutter M. Autism as a strongly genetic disorder: evidence from a British twin study. *Psychol Med.* 1995 Jan;25(1):63-77.

- Bailey SL, Carpentier PA, McMahon EJ, Begolka WS, Miller SD. Innate and adaptive immune responses of the central nervous system. *Critical Reviews in Immunology*, vol. 26, no. 2, pp. 149–188, 2006.].

- Ballif BC, Hornor SA, Jenkins E, Madan-Khetarpal S, Surti U, Jackson KE, Asamoah A, Brock PL, Gowans GC, Conway RL, Graham JM Jr, Medne L, Zackai EH, Shaikh TH, Geoghegan J, Selzer RR, Eis PS, Bejjani BA, Shaffer LG. Discovery of a previously unrecognized microdeletion syndrome of 16p11.2-p12.2. *Nat Genet*. 2007 Sep;39(9):1071-3.

- Baron CA, Liu SY, Hicks C, Gregg JP. Utilization of lymphoblastoid cell lines as a system for the molecular modeling of autism. *J Autism Dev Disord.* 2006(a) Nov;36(8):973-82.

- Baron CA, Tepper CG, Liu SY, Davis RR, Wang NJ, Schanen NC, Gregg JP. Genomic and functional profiling of duplicated chromosome 15 cell lines reveal regulatory alterations in UBE3A-associated ubiquitin-proteasome pathway processes. *Hum Mol Genet*. 2006(b) ;15:853–869.

- Belmonte MK, Cook EH, Anderson GM, Rubenstein JL, Greenough WT, Beckel-Mitchener A, Courchesne E, Boulanger LM, Powell SB, Levitt PR, et al. Autism as a disorder of neural information processing: directions for research and targets for therapy. *Mol. Psychiatry.* 2004;9:646-663.

- Benayed R, Gharani N, Rossman I, Mancuso V, Lazar G, Kamdar S, Bruse SE, Tischfield S, Smith BJ, Zimmerman RA, Dicicco-Bloom E, Brzustowicz LM, Millonig JH. Support for the homeobox transcription factor gene ENGRAILED 2 as an autism spectrum disorder susceptibility locus. *Am J Hum Genet*. 2005 Nov;77(5):851-68. Epub 2005 Oct 5.

- Bettelheim B. The empty fortress: infantile autism and the birth of the self. New York (1967).

- Bonaglia MC, Giorda R , Borgatti R , Felisari G , Gagliardi C , Selicorni A , Zuffardi O . Disruption of the ProSAP2 gene in a t(12;22)(q24.1;q13.3) is associated with the 22q13.3 deletion syndrome. *Am J Hum Genet.* 2001;69:261–8.

- Boulanger LM, Shatz CJ. Immune signalling in neural development, synaptic plasticity and disease*. Nature Reviews Neuroscience*, vol. 5, no. 7, pp. 521–531, 2004.

- Brown V, Jin P, Ceman S, Darnell JC, O'Donnell WT, Tenenbaum SA, Jin X, Feng Y, Wilkinson KD, Keene JD, Darnell RB, Warren ST. Microarray identification of FMRP-associated brain mRNAs and altered mRNA translational profiles in fragile X syndrome. *Cell*. 2001;107:477–487.

- Burrows M, Wheeler D. A block sorting lossless data compression algorithm. Technical Report 124. Palo Alto, California: DEC, Digital Systems Research Center; 1994.

- Buysse K, Delle Chiaie B, Van Coster R, Loeys B, De Paepe A, Mortier G, Speleman F, Menten B. Challenges for CNV interpretation in clinical molecular karyotyping: lessons learned from a 1001 sample experience. *Eur J Med Genet*. 2009(a) Nov-Dec;52(6):398-403.

- Buysse K, Reardon W, Mehta L, Costa T, Fagerstrom C, Kingsbury DJ, Anadiotis G, McGillivray BC, Hellemans J, de Leeuw N, de Vries BB, Speleman F, Menten B, Mortier G. The 12q14 microdeletion syndrome: additional patients and further evidence that HMGA2 is an important genetic determinant for human height. *Eur. J. Med. Genet*. 2009(b) ;52, pp. 101–107.

- Campbell DB, Warren D, Sutcliffe JS, Lee EB, Levitt P. Association of MET with social and communication phenotypes in individuals with autism spectrum disorder. *Am J Med Genet B Neuropsychiatr Genet.* 2010 Mar 5;153B(2):438-46.

- Charman T, Baird G. Diagnosis of autism spectrum disorder in 2- and 3-year-old children. *J Child Psychol Psychiatry*. 2002 Mar; 43(3):289-305.

- Chess S. Autism in children with congenital rubella. *Journal of Autism and Childhood Schizophrenia*. 1971;1(1):33–47.

- Chien YL, Wu YY, Chen CH, Gau SS, Huang YS, Chien WH, Hu FC, Chao YL. Association of HLA-DRB1 alleles and neuropsychological function in autism. *Psychiatr Genet*. 2012 Feb;22(1):46-9.

- Cloonan N, Forrest AR, Kolle G, Gardiner BB, Peckham HE, Manning JM, McKernan KJ, Grimmond SM. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods*. 2008 Jul;5(7):613-9.

- Cook EH, Sherer MR. Copy-number variations associated with neuropsychiatric conditions*. Nature*. 2008;455:919–923.

- Comi AM, Zimmerman AW, Frye VH, Law PA, Peeden JN. Familial clustering of autoimmune disorders and evaluation of medical risk factors in autism. *J Child Neurol.* 1999 Jun;14(6):388-94.

- Craddock N, Hurles ME, Cardin N, et al. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls*. Nature*. 2010;464:713–720.

- Cullheim S, Thams S. The microglial networks of the brain and their role in neuronal network plasticity after lesion. *Brain Research Reviews*, vol. 55, no. 1, pp. 89–96, 2007.

- Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol*. 2003;4:3.

- DeStefano F. Vaccines and autism: evidence does not support a causal association. *Clin Pharmacol Ther*. 2007 Dec;82(6):756-9.

- DiCicco-Bloom E, Lord C, Zwaigenbaum L, Courchesne E, Dager SR, Schmitz C, Schultz RT, Crawley J, Young LJ. The developmental neurobiology of autism spectrum disorder. *J Neurosci*. 2006 Jun 28; 26(26):6897-906.

- Enstrom AM, Lit L, Onore CE et al. Altered gene expression and function of peripheral blood natural killer cells in children with autism. *Brain, Behavior, and Immunity*, vol. 23, no. 1, pp. 124–133, 2009.

- Farber CR, Lusis AJ. Future of osteoporosis genetics: enhancing genome-wide association studies. *J Bone Miner Res.* 2009 Dec;24(12):1937-42.

- Farrington CP, Miller E, Taylor B. MMR and autism: further evidence against a causal association. *Vaccine*. 2001 Jun 14;19(27):3632-5.

- Ferragina P, Manzini G. An experimental study of an opportunistic index; pp. 269–278. Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms. Washington, D.C. USA: 2001.

- Finelli P, Natacci F, Bonati MT, Gottardi G, Engelen JJ, de Die-Smulders CE, Sala M, Giardino D, Larizza L. FISH characterisation of an identical (16)(p11.2p12.2) tandem duplication in two unrelated patients with autistic behaviour. *J Med Genet.* 2004 Jul;41(7):e90.

- Flipsen-ten Berg K, van Hasselt PM, Eleveld MJ., et al. Unmasking of a hemizygous *WFS1* gene mutation by a chromosome 4p deletion of 8.3 Mb in a patient with Wolf-Hirschhorn syndrome. *Eur J Hum Genet*. 2007;15:1132–1138.

- Folstein SE, Rosen-Sheidley B. Genetics of autism: complex aetiology for a heterogeneous disorder*. Nat. Rev. Genet*. 2001;2:943-955.

- Fombonne E. Epidemiological studies of pervasive developmental disorder. In: Volkmar F, Paul R, Klin A, Cohen D, editors. *Handbook of Autism and Pervasive Developmental Disorders*. Hoboken, NJ: Wiley; 2005. pp. 42–69.

- Freitag CM. The genetics of autistic disorders and its clinical relevance: a review of the literature. *Mol Psychiatry.* 2007 Jan; 12(1):2-22.

- Gao J, Luo X, Tang K, Li X, Li G. Epstein-Barr virus integrates frequently into chromosome 4q, 2q, 1q and 7q of Burkitt's lymphoma cell line (Raji). *J Virol Methods*. 2006 Sep;136(1-2):193-9. Epub 2006 Jun 23.

- Gene Ontology Consortium (2006). The Gene Ontology (GO) project in 2006, Nucleic Acids Res. 34, D322-326. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 2000 Jan 1;28(1):27-30.

- Goizet C, Excoffier E, Taine L, Taupiac E, El Moneim AA, Arveiler B, Bouvard M, Lacombe D. Case with autistic syndrome and chromosome 22q13.3 deletion detected by FISH. *Am J Med Genet.* 2000 Dec 4;96(6):839-44. Review.

- Gregg JP, Lit L, Baron CA et al. Gene expression changes in children with autism. *Genomics*, vol. 91, no. 1, pp. 22–29, 2008.

- Gu W, Zhang F, Lupski JR. Mechanisms for human genomic rearrangements. *Pathogenetics*, 2008;1, p. 4.

- Gualandi G, Santolini E, Calef E. Epstein-Barr virus DNA recombines via latent origin of replication with the human genome in the lymphoblastoid cell line RGN1. *J Virol*. 1992 Sep;66(9):5677-81.

- Henrichsen CN, Vinckenbosch N, Zollner S, et al. Segmental copy number variation shapes tissue transcriptomes. *Nat Genet.* 2009;41:424–429.

- Horike S, Cai S, Miyano M, Cheng JF, Kohwi-Shigematsu T. Loss of silent-chromatin looping and impaired imprinting of DLX5 in Rett syndrome. *Nat Genet.* 2005.

- Horton R, Gibson R, Coggill P et al. Variation analysis and gene annotation of eight MHC haplotypes: the MHC haplotype project. *Immunogenetics*, vol. 60, no. 1, pp. 1–18, 2008.

- Hou T, Macmillan H, Chen Z, Keech CL, Jin X, Sidney J, Strohman M, Yoon T, Mellins ED. An insertion mutant in DQA1*0501 restores susceptibility to HLA-DM: implications for disease associations. *J Immunol*. 2011 Sep 1;187(5):2442-52.

- Hu VW, Frank BC, Heine S, Lee NH, Quackenbush J. Gene expression profiling of lymphoblastoid cell lines from monozygotic twins discordant in severity of autism reveals differential regulation of neurologically relevant genes. *BMC Genomics*. 2006 May 18;7:118.

- Huh GS, Boulanger LM, Du H, Riquelme PA, Brotz TM, Shatz CJ. Functional requirement for class I MHC in CNS development and plasticity. *Science*, vol. 290, no. 5499, pp. 2155–2159, 2000.

- Hviid A, Stellfeld M, Wohlfahrt J, Melbye M. Association Between Thimerosal-Containing Vaccine and Autism. *JAMA*. 2003;290(13):1763-1766.

- Iwamoto K, Kakiuchi C, Bundo M, Ikeda K, Kato T. Molecular characterization of bipolar disorder by comparing gene expression profiles of postmortem brains of major mental disorders. *Mol Psychiatry*. 2004;9:406–416.

- Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 2000 Jan 1;28(1):27-30.

- Kanner L. Autistic disturbances of affective contact. *Nervous Child.* 1943;2, 217-250.

- Kassahn KS, Waddell N, Grimmond SM. Sequencing transcriptomes in toto. *Integr Biol (Camb*). 2011 May;3(5):522-8.

- Kleefstra T, Brunner HG, Amiel J, Oudakker AR, Nillesen WM, Magee A, Genevieve D, Cormier-Daire V, van Esch H, Fryns JP, Hamel BC, Sistermans EA, de Vries BB, van Bokhoven H. Loss-of-function mutations in euchromatin histone methyl transferase 1 (EHMT1) cause the 9q34 subtelomeric deletion syndrome. *Am. J. Hum. Genet*. 2006;79, pp. 370–377.

- Lam KSL, Aman MG, Arnold LE. Neurochemical correlates of autistic disorder: A review of the literature. *Res Dev Disabi*. 2006;27:254–289.

- Lamb JA, Barnby G, Bonora E, Sykes N, Bacchelli E, Blasi F, Maestrini E, Broxholme J, Tzenova J, Weeks D, Bailey AJ, Monaco AP; International Molecular Genetic Study of Autism Consortium (IMGSAC). Analysis of IMGSAC autism susceptibility loci: evidence for sex limited and parent of origin specific effects. *J Med Genet*. 2005 Feb;42(2):132-7.

- Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 2009;10(3):R25.

- Lauritsen MB, Pedersen CB, Mortensen PB. Effects of familial risk factors and place of birth on the risk of autism: a nationwide register-based study. *J Child Psychol Psychiatry*. 2005 Sep;46(9):963-71.

- Le Couteur A, Rutter M, Lord C, Rios P, Robertson S, Holdgrafer M, McLennan J. Autism diagnostic interview: a standardized investigator-based instrument. *J Autism Dev Disord.* 1989 Sep; 19(3):363-87.

- Leenman EE, Panzer-Grümayer RE, Fischer S, Leitch HA, Horsman DE, Lion T, Gadner H, Ambros PF, Lestou VS. Rapid determination of Epstein-Barr virus latent or lytic infection in single human cells using in situ hybridization. *Mod Pathol*. 2004 Dec;17(12):1564-72.

- Lestou VS, De Braekeleer M, Strehl S, Ott G, Gadner H, Ambros PF. Non-random integration of Epstein-Barr virus in lymphoblastoid cell lines. *Genes Chromosomes Cancer*. 1993 Sep;8(1):38-48.

- Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011 Aug 4;12:323.

- Lord C, Risi S, Lambrecht L, Cook EH Jr, Leventhal BL, DiLavore PC, Pickles A, Rutter M. The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. *J Autism Dev Disord*. 2000 Jun; 30(3):205-23.

- Luo R, Sanders SJ, Tian Y, Voineagu I, Huang N, Chu SH, Klei L, Cai C, Ou J, Lowe JK, Hurles ME, Devlin B, State MW, Geschwind DH. Genome-wide transcriptome profiling reveals the functional impact of rare de novo and recurrent CNVs in autism spectrum disorders. *Am J Hum Genet.* 2012 Jul 13;91(1):38-55.

- Ma DQ, Whitehead PL, Menold MM, Martin ER, Ashley-Koch AE, Mei H, Ritchie MD, Delong GR, Abramson RK, Wright HH, Cuccaro ML, Hussman JP, Gilbert JR, Pericak-Vance MA. Identification of significant association and gene-gene interaction of GABA receptor subunit genes in autism. *Am J Hum Genet*. 2005 Sep;77(3):377-88.

- Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* 2008 Sep;18(9):1509-17.

- Marshall CR, Noor A, Vincent JB, et al. Structural variation of chromosomes in autism spectrum disorder. *Am J Hum Genet.* 2008;82:477–488.

- Meloni I, Muscettola M, Raynaud M, Longo I, Bruttini M, Moizard MP, Gomot M, Chelly J, des Portes V, Fryns JP, Ropers HH, Magi B, Bellan C, Volpi N, Yntema HG, Lewis SE, Schaffer JE, Renieri A. FACL4, encoding fatty acid-CoA ligase 4, is mutated in nonspecific X-linked mental retardation. *Nat Genet.* 2002;30:436–440.

- Menten B, Buysse K, Zahir F, Hellemans J, Hamilton SJ, Costa T, Fagerstrom C, Anadiotis G, Kingsbury D, McGillivray BC, Marra MA, Friedman JM, Speleman F, Mortier G. Osteopoikilosis, short stature and mental retardation as key features of a new microdeletion syndrome on 12q14. *J. Med. Genet*. 2007;44, pp. 264–268.
- Minshew NJ, Williams DL. The new neurobiology of autism: cortex, connectivity, and neuronal organization. *Arch Neurol*. 2007 Jul; 64(7):945-50.

- Morgan JT, Chana G, Pardo CA et al. Microglial activation and increased microglial density observed in the dorsolateral prefrontal cortex in autism. *Biological Psychiatry,* vol. 68, no. 4, pp. 368–376, 2010.

- Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*. 2008 Jul;5(7):621-8.

- Mostafa GA, Shehab AA, Al-Ayadhi LY. The link between some alleles on human leukocyte antigen system and autism in children. *J Neuroimmunol*. 2013 Feb 15;255(1-2):70-4.

- Muhle R, Trentacoste SV, Rapin I. The genetics of autism. *Pediatrics.* 2004 May;113(5):e472-86.

- Nishimura Y, Martin CL, Vazquez-Lopez A, Spence SJ, Alvarez-Retuerto AI, Sigman M, Steindler C, Pellegrini S, Schanen NC, Warren ST, Geschwind DH. Genome-wide expression profiling of lymphoblastoid cell lines distinguishes different forms of autism and reveals shared pathways. *Hum Mol Genet*. 2007;16:1682–1698.

- Ohtsuka M, Inoko H, Kulski JK, Yoshimura S. Major histocompatibility complex (Mhc) class Ib gene duplications, organization and expression patterns in mouse strain C57BL/6. *BMC Genomics*, vol. 9, p. 178, 2008.

- Pagnamenta AT, Bacchelli E, de Jonge MV, et al. Characterization of a family with rare deletions in *CNTNAP5* and *DOCK4* suggests novel risk loci for autism and dyslexia. *Biol Psych*. 2010;68:320–328.

- Pagnamenta AT, Khan H, Walker S, et al. Rare familial 16q21 microdeletions under a linkage peak implicate cadherin 8 (*CDH8*) in susceptibility to autism and learning disability. *J Med Genet*. 2011;48:48–54.

- Pardo CA, Eberhart CG. The neurobiology of autism. *Brain Pathol*. 2007 Oct; 17(4):434-47.

- Perry GH, Dominy NJ, Claw KG, et al. Diet and the evolution of human amylase gene copy number variation. *Nat Genet*. 2007;39:1256–1260.

- Phelan MC, Rogers RC, Saul RA, Stapleton GA, Sweet K, McDermid H, Shaw SR, Claytor J, Willis J, Kelly DP. 22q13 deletion syndrome. *Am J Med Genet*. 2001 Jun 15;101(2):91-9.

- Philippe A, Martinez M, Guilloud-Bataille M, Gillberg C, Råstam M, Sponheim E, Coleman M, Zappella M, Aschauer H, Van Maldergem L, Penet C, Feingold J, Brice A, Leboyer M. Genome-wide scan for autism susceptibility genes. Paris Autism Research International Sibpair Study. *Hum Mol Genet*. 1999 May;8(5):805-12.

- Prevalence of Autism Spectrum Disorders - Autism and Developmental Disabilities Monitoring Network, 14 Sites, United States, 2008. Surveillance Summaries. March 30, 2012 / 61(SS03);1-19

- Ramalingam A, Zhou XG, Fiedler SD, Brawner SJ, Joyce JM, Liu HY, Yu S. 16p13.11 duplication is a risk factor for a wide spectrum of neuropsychiatric disorders. *J Hum Genet.* 2011 Jul;56(7):541-4.

- Ramocki MD, Zoghbi HY. Failure of neuronal homeostasis results in common neuropsychiatric phenotypes. *Nature.* 2008;455:912–918.

- Redon R, Ishikawa S, Fitch KR, et al. Global variation in copy number in the human genome. *Nature.* 2006;444:444–454.

- Rimland B. On the objective diagnosis of infantile autism. *Acta Paedopsychiatr.* 1968;35(4):146-61.

- Schaefer GB, Mendelsohn NJ. Genetics evaluation for the etiologic diagnosis of autism spectrum disorders. *Genet Med.* Review 2008 Jan; 10(1):4-12.

- Schultz RT. Developmental deficits in social perception in autism: the role of the amygdala and fusiform face area. *Int J Dev Neurosci*. 2005 Apr-May; 23(2-3):125-41.

- Sebat J, Lakshmi B, Malhotra D, Troge J, Geschwind DH, Gilliam TC, Ye K, Wigler M. Strong association of de novo copy number mutations with autism. *Science.* 2007;316(no. 5823):445–449.

- She X, Jiang Z, Clark RA, Liu G, Cheng Z, Tuzun E, Church DM, Sutton G, Halpern AL, Eichler EE. Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature*. 2004 Oct 21;431(7011):927-30.

- Slavotinek AM. Novel microdeletion syndromes detected by chromosome microarrays. *Hum. Genet.* 2008;124, pp. 1–17.

- Sparrow SS, Cicchetti DV. Diagnostic uses of the Vineland Adaptive Behavior Scales. *J Pediatr Psychol*. 1985 Jun;10(2):215-25

- Stranger BE, Forrest MS, Dunning M, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science.* 2007;315:848–853.

- Stefansson H, Ophoff RA, Steinberg S et al. Common variants conferring risk of schizophrenia. *Nature*, vol. 460, no. 7256, pp. 744–747, 2009.

- Suda S, Iwata K, Shimmura C, Kameno Y, Anitha A, Thanseem I, Nakamura K, Matsuzaki H, Tsuchiya KJ, Sugihara G, Iwata Y, Suzuki K, Koizumi K, Higashida H, Takei N, Mori N. Decreased expression of axon-guidance receptors in the anterior cingulate cortex in autism. *Mol Autism*. 2011 Aug 22;2(1):14.

- Sultan M, Schulz MH, Richard H, Magen A, Lehrach H, Yaspo ML. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science.* 2008 Aug 15;321(5891):956-60.

- Sung H, Kang SH, Bae YJ, Hong JT, Chung YB, Lee CK, Song S. PCR-based detection of Mycoplasma species*. J Microbiol*. 2006 Feb;44(1):42-9.

- Sweeten TL, Posey DJ, McDougle CJ. High blood monocyte counts and neopterin levels in children with autistic disorder. *American Journal of Psychiatry*, vol. 160, no. 9, pp. 1691–1693, 2003(a).

- Sweeten TL, Bowyer SL, Posey DJ, Halberstadt GM, McDougle CJ. Increased prevalence of familial autoimmunity in probands with pervasive developmental disorders. *Pediatrics*. 2003(b) Nov;112(5):e420.

- Tabuchi K, Blundell J, Etherton M, et al. A neuroligin-3 mutation implicated in autism increases inhibatory synaptic transmission in mice. *Science.* 2007;318:71–76.

- Tang J, Hu M, Lee S, Roblin R. A polymerase chain reaction based method for detecting Mycoplasma/Acholeplasma contaminants in cell culture. *J Microbiol Methods*. 2000 Jan;39(2):121-6.

- Tang Y, Nee AC, Lu A, Ran R, Sharp FR. Blood genomic expression profile for neuronal injury. *J Cereb Blood Flow Metab*. 2003 Mar;23(3):310-9.

- Tang Y, Lu A, Ran R, Aronow BJ, Schorry EK, Hopkin RJ, Gilbert DL, Glauser TA, Hershey AD, Richtand NW, Privitera M, Dalvi A, Sahay A, Szaflarski JP, Ficker DM, Ratner N, Sharp FR. Human blood genomics: distinct profiles for gender, age and neurofibromatosis type 1*. Brain Res Mol Brain Res*. 2004 Dec 20;132(2):155-67.

- Torres AR, Maciulis A, Stubbs EG, Cutler A, Odell D. The transmission disequilibrium test suggests that HLA-DR4 and DR13 are linked to autism spectrum disorder. *Hum Immunol*. 2002 Apr;63(4):311-6.

- Ullmann R, Turner G, Kirchhoff M, Chen W, Tonge B, Rosenberg C, Field M, Vianna-Morgante AM, Christie L, Krepischi-Santos AC, Banna L, Brereton AV, Hill A, Bisgaard AM, Müller I, Hultschig C, Erdogan F, Wieczorek G, Ropers HH. Array CGH identifies reciprocal 16p13.1 duplications and deletions that predispose to autism and/or mental retardation. *Hum Mutat*. 2007 Jul;28(7):674-82.

- Valicenti-McDermott M, McVicar K, Rapin I, Wershil BK, Cohen H, Shinnar S. Frequency of gastrointestinal symptoms in children with autistic spectrum disorders and association with family history of autoimmune disease*. J Dev Behav Pediatr.* 2006 Apr;27(2 Suppl):S128-36.

- Veenstra-Vanderweele J, Christian SL, Cook EH. Autism as a paradigmatic complex genetic disorder. *Annu Rev Genomics Hum Genet*. 2004;5:379–405.

- Vissers LE, van Ravenswaaij CM, Admiraal R, Hurst JA, de Vries BB, Janssen IM, van der Vliet WA, Huys EH, de Jong PJ, Hamel BC, Schoenmakers EF, Brunner HG, Veltman JA, van Kessel AG. Mutations in a new member of the chromodomain gene family cause CHARGE syndrome. *Nat. Genet.* 2004;36 pp. 955–957.

- Voineagu I, Wang X, Johnston P, Lowe JK, Tian Y, Horvath S, Mill J, Cantor RM, Blencowe BJ, Geschwind DH. Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature*. 2011 May 25;474(7351):380-4.

- Volkmar FR, Pauls D. Autism. *Lancet.* 2003 Oct 4; 362(9390):1133-41.

- Vorstman JA, van Daalen E, Jalali GR, et al. A double hit implicates *DIAPH3* as an autism risk gene. *Mol Psychiatry*. 2011;16:442–451.

- Walsh T, McClellan JM, McCarthy SE, et al. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science*. 2008;320:539–543.

- Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, Fossdal R, Saemundsen E, Stefansson H, Sklar P, Wu BL, Daly MJ. the Autism Consortium. Association between Microdeletion and Microduplication at 16p11.2 and Autism. *The New England journal of medicine.* 2008.

- Weiss LA, Arking DE; Gene Discovery Project of Johns Hopkins & the Autism Consortium, Daly MJ, Chakravarti A. A genome-wide linkage and association scan reveals novel loci for autism. *Nature.* 2009 Oct 8;461(7265):802-8.

- White JF. Intestinal pathophysiology in autism. *Exp Biol Med (Maywood).* 2003 Jun;228(6):639-49.

- World Health Organization, The International Classification of Diseases (ICD-10), Classification of mental and behavioral disorders: clinical descriptions and diagnostic guidelines. (Geneva: World Health Organization), 1992.

- Xiao BG, Link H. Immune regulation within the central nervous system. *Journal of the Neurological Sciences*, vol. 157, no. 1, pp. 1–12, 1998.

- Zafeiriou DI, Ververi A, Vargiami E. Childhood autism and associated comorbidities. *Brain Dev.* 2007 Jun; 29(5):257-72.

- Zhou X, Chen Q, Schaukowitch K, Kelsoe JR, Geyer MA. Insoluble DISC1-Boymaw fusion proteins generated by *DISC1* translocation. *Mol Psych.* 2010;15:669–672.

7. APPENDIX

Appendix 1

Official gene symbol	Description
ABCD4	ATP-binding cassette, sub-family D (ALD), member 4
AC004410.1	signal peptide peptidase like 2B
AC005280.1	pseudogene
AC005696.2	N/A
ACTB	actin, beta
ACTG1	actin, gamma 1
AKAP12	A kinase (PRKA) anchor protein 12
AKAP17A	A kinase (PRKA) anchor protein 17A
AP003068.6	Putative protein coding
BCL2	B-cell CLL/lymphoma 2
BRD1	bromodomain containing 1
BRD2	bromodomain containing 2
C10orf103	chromosome 10 open reading frame 103
CA2	carbonic anhydrase II
CCL4L2	chemokine (C-C motif) ligand 4-like 1; chemokine (C-C motif) ligand 4-like 2
CD74	CD74 molecule, major histocompatibility complex, class II invariant chain
CERCAM	cerebral endothelial cell adhesion molecule
CHKB	choline kinase beta
DOCK9	dedicator of cytokinesis 9
DSC3	desmocollin 3
EEF1A1	eukaryotic translation elongation factor 1 alpha-like 7; eukaryotic translation elongation factor 1 alpha-like 3; eukaryotic translation elongation factor 1 alpha 1
EEF2	eukaryotic translation elongation factor 2
EIF3CL	eukaryotic translation initiation factor 3, subunit C-like
ENO1	enolase 1, (alpha)
FOXO3	forkhead box O3; forkhead box O3B pseudogene
FRMD4B	FERM domain containing 4B
FTL	similar to ferritin, light polypeptide; ferritin, light polypeptide
FXYD2	FXYD domain containing ion transport regulator 2
GAPDH	glyceraldehyde-3-phosphate dehydrogenase
GMPR	guanosine monophosphate reductase
HIST1H1C	histone cluster 1, H1c
HLA-B	major histocompatibility complex, class I, C; major histocompatibility complex, class I, B
HLA-DMA	major histocompatibility complex, class II, DM alpha
HLA-DOB	major histocompatibility complex, class II, DO beta
HLA-DQA1	major histocompatibility complex, class II, DQ alpha 1
HLA-DQA2	major histocompatibility complex, class II, DQ alpha 2
HLA-DQB1	major histocompatibility complex, class II, DQ beta 1; similar to major histocompatibility complex, class II, DQ beta 1
HLA-DQB1-AS1	*HLA-DQB1* antisense RNA 1
HLA-DRA	major histocompatibility complex, class II, DR alpha
HLA-DRB1	major histocompatibility complex, class II, DR beta 4; major histocompatibility complex, class II, DR beta 1
HLA-H	major histocompatibility complex, class I, H (pseudogene)

HMGA2	high mobility group AT-hook 2
HPDL	4-hydroxyphenylpyruvate dioxygenase-like
IGHA1	immunoglobulin heavy constant alpha 1
IGHV1-2	immunoglobulin heavy variable 1-2
IL12RB2	interleukin 12 receptor, beta 2
IQCG	IQ motif containing G
ITGB5	integrin, beta 5
JUB	jub, ajuba homolog (Xenopus laevis)
KIAA0430	KIAA0430
LDOC1L	leucine zipper, down-regulated in cancer 1-like
MANEAL	mannosidase, endo-alpha-like
MAPK12	mitogen-activated protein kinase 12
MARCH6	membrane-associated ring finger (C3HC4) 6
MXRA7	matrix-remodelling associated 7
MYL9	myosin, light chain 9, regulatory
NINJ2	ninjurin 2
NUDT1	nudix (nucleoside diphosphate linked moiety X)-type motif 1
PADI2	peptidyl arginine deiminase, type II
PDXDC1	pyridoxal-dependent decarboxylase domain containing 1
PHBP9	prohibitin pseudogene 9
PHF1	PHD finger protein 1
PIM3	pim-3 oncogene
PKD1P6	polycystic kidney disease 1 (autosomal dominant) pseudogene 6
PKMYT1	protein kinase, membrane associated tyrosine/threonine 1
PLTP	phospholipid transfer protein
PNCK	pregnancy up-regulated non-ubiquitously expressed CaM kinase
PNRC2	proline-rich nuclear receptor coactivator 2; similar to hCG1728885
POFUT1	protein O-fucosyltransferase 1
POU3F1	POU class 3 homeobox 1
POU5F1	POU class 5 homeobox 1
PPP1R18	protein phosphatase 1, regulatory subunit 18
PRRC2A	proline-rich coiled-coil 2A
PXT1	peroxisomal, testis specific 1
RGPD5	RANBP2-like and GRIP domain containing 5
RGS2	regulator of G-protein signaling 2, 24kDa
RNF183	ring finger protein 183
RP11-106J23.2	Known processed transcript
RP11-290F5.2	Putative processed transcript
RP11-366L20.2	N/A
RP11-368J21.1	nuclear pore complex interacting protein pseudogene (LOC613037), non-coding RNA
RP11-632C17__A.1	pseudogene
RP11-726G1.1	pseudogene
RPL3	ribosomal protein L3; similar to 60S ribosomal protein L3 (L4)
RPLP0	ribosomal protein, large, P0 pseudogene 2; ribosomal protein, large, P0 pseudogene 3; ribosomal protein, large, P0 pseudogene 6; ribosomal protein, large, P0
SBF1	SET binding factor 1
SCAMP5	secretory carrier membrane protein 5
SFRP1	secreted frizzled-related protein 1
SHANK3	SH3 and multiple ankyrin repeat domains 3
SHROOM3	shroom family member 3

SLC9A7P1	solute carrier family 9, subfamily A (NHE7, cation proton antiporter 7), member 7 pseudogene 1
SUPT3H	suppressor of Ty 3 homolog (S. cerevisiae)
SYNM	synemin, intermediate filament protein
TIAM1	T-cell lymphoma invasion and metastasis 1
TLN2	talin 2
TMSB4X	thymosin-like 2 (pseudogene); thymosin-like 1 (pseudogene); thymosin beta 4, X-linked
TNFRSF14	tumor necrosis factor receptor superfamily, member 14 (herpesvirus entry mediator)
TNK1	tyrosine kinase, non-receptor, 1
TNRC18	trinucleotide repeat containing 18
TRABD	TraB domain containing
TRIM26	tripartite motif-containing 26
TRIM27	tripartite motif-containing 27
VANGL2	vang-like 2 (van gogh, Drosophila)
VCAN	versican
VPS37B	vacuolar protein sorting 37 homolog B (S. cerevisiae)
ZBED4	zinc finger, BED-type containing 4
ZFAND5	similar to zinc finger, AN1-type domain 5; zinc finger, AN1-type domain 5
ZNF185	zinc finger protein 185 (LIM domain)
ZNRD1	zinc ribbon domain containing 1

Table 10. List of genes mentioned in this study. Official gene symbol and description are shown.