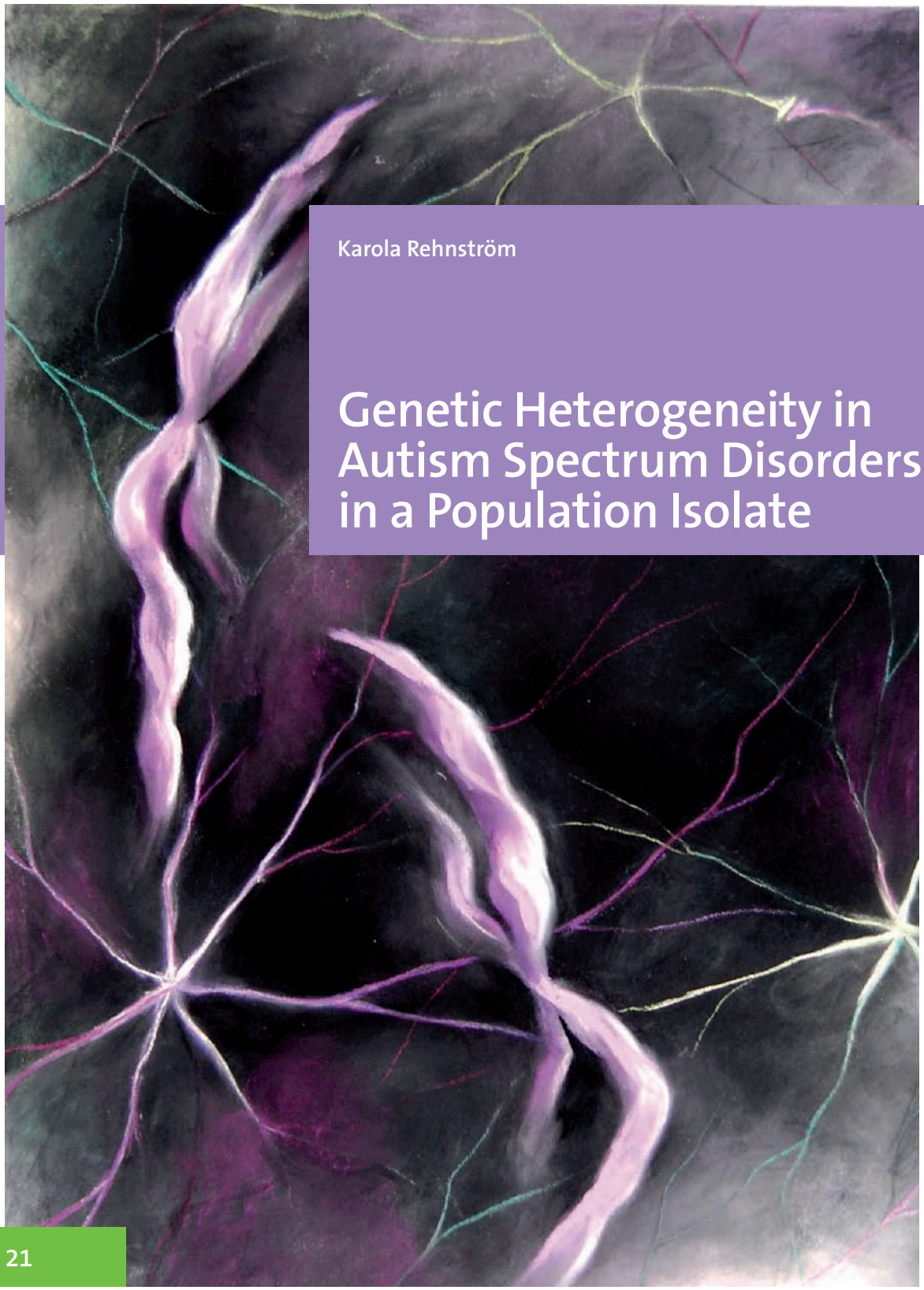
NATIONAL INSTITUTE
FOR HEALTH AND WELFARE



RESEARCH

Karola Rehnström

Genetic Heterogeneity in Autism Spectrum Disorders in a Population Isolate

Karola Rehnström

GENETIC HETEROGENEITY IN AUTISM SPECTRUM DISORDERS IN A POPULATION ISOLATE

A C A D E M I C D I S S E R T A T I O N

To be presented with the permission of the Medical Faculty, University of Helsinki, for public examination in The Small Lecture Hall, Haartman Institute, on October 30th 2009, at 12 noon.

National Institute for Health and Welfare
and
Institute for Molecular Medicine Finland
and
Department of Medical Genetics, University of Helsinki

Helsinki 2009

Helsinki University Biomedical Dissertations No. 126
ISSN 1457-8433

ISBN 978-952-245-132-3 (print)
ISSN 1798-0054 (print)
ISBN 978-952-245-133-0 (pdf)
ISSN 1798-0062 (pdf)

Cover art: Anne Pernaa, 'Lost Chromosomes'. Pastel.
www.annepernaa.fi

Helsinki University Print
Helsinki, Finland 2009

Supervised by
Professor Leena Peltonen-Palotie
Institute for Molecular Medicine Finland FIMM
Helsinki, Finland
Wellcome Trust Sanger Institute
Hinxton, UK

and

Tero Ylisaukko-oja, Ph. D.
National Public Health Institute
Department of Molecular Medicine
Helsinki, Finland

Reviewed by
Professor Jim Schröder
University of Helsinki
Department of Biological and Environmental Sciences
Helsinki, Finland

and

Adjunct Professor Tarja Laitinen
Helsinki University Central Hospital
Clinical Research Unit of Pulmonary Diseases
Helsinki, Finland

Opponent
Professor Kerstin Lindblad-Toh
Vertebrate Genome Biology
The Broad Institute
Cambridge MA, USA
Department of Medical Biochemistry and Microbiology
Uppsala University
Uppsala, Sweden

"What does it matter to Science if her passionate servants are rich or poor, happy or unhappy, healthy or ill? She knows that they have been created to seek and to discover, and that they will seek and find until their strength dries up at its source. It is not in a scientist's power to struggle against his vocation: even on his days of disgust or rebellion his steps lead him inevitably back to his laboratory apparatus."

-From *Madame Curie – A biography* by Eve Curie

Karola Rehnström, Genetic Heterogeneity in Autism Spectrum Disorders in a Population Isolate. National Institute for Health and Welfare, Research 21|2009. 192 pages. Helsinki, Finland 2009. ISBN 978-952-245-132-3 (print); 978-952-245-133-0 (pdf)

ABSTRACT

Positional cloning has made it possible to perform hypothesis-free, genome-wide scans for genetic factors affecting a disorder or trait. Traditionally linkage analysis using microsatellite markers has been used as a first step in this process to identify regions of interest, followed by meticulous fine mapping and candidate gene screening using association methods and subsequent sequence analysis. More recently, genome-wide association analysis has enabled a more direct approach to identify specific genetic variants explaining a part of the variance of the phenotype of interest. In addition, data produced for genome-wide association analysis has also made it possible to assay small, submicroscopic variation on the chromosomes, referred to as copy number variants, which have been shown to confer susceptibility to some complex disorders.

Isolates have proven useful in the identification of genes causing Mendelian, or monogenic, disorders. The Finnish population is genetically homogenous, and has been molded by founder effect, multiple consecutive bottlenecks and genetic drift. These features can be utilized in identification of genetic risk factors for complex disorders, although population isolates have not been shown to be as useful for the genetic mapping of complex traits as for Mendelian disorders. The genetic risk factors for complex disorders in Finland could, however, prove to be less heterogeneous due to the limited range of susceptibility alleles carried into the gene pool by the original settlers.

Autism spectrum disorders (ASDs) are a group of childhood onset neuropsychiatric disorders with shared core symptoms but varying severity. Although a strong genetic component has been established in ASDs, genetic susceptibility factors have largely eluded characterization, despite active research for decades. In this study, we have utilized modern molecular genetics methods combined with the special characteristics of the Finnish population to identify genetic risk factors for ASDs.

The results of this study show that numerous genetic risk factors exist for ASDs even within a population isolate. Stratification based on clinical phenotype resulted in encouraging results, as linkage to 3p14-p24 identified in the Finnish genome-wide linkage scan for Asperger Syndrome (AS) was replicated in an independent family set. The success of linkage mapping of susceptibility regions for AS has interesting

implications for the underlying genetic architecture, suggesting that genetic risk factors for AS could possibly be less heterogeneous than for the wider spectrum of ASDs.

Fine-mapping of the previously identified linkage peak for ASDs at 3q25-q27 revealed association between ASDs and a subunit of the *5-hydroxytryptamine receptor 3C* (*HTR3C*). The 5-hydroxytryptamine pathway has previously been robustly implicated in the etiology of ASDs but this is the first time the 5-hydroxytryptamine receptors on 3q have been evaluated as risk factors for ASDs. However, association to *HTR3C* only accounted for a part of the observed linkage signal, suggesting that other predisposing factors exist at this locus.

As a part of this study, we used dense, genome-wide single nucleotide polymorphism (SNP) data to characterize the population structure. We observed significant population substructure caused by the multiple consecutive bottle-necks experienced by the Finnish population during the population history. We used this information to ascertain a genetically homogenous subset of autism families from Central Finland to identify possible rare, enriched risk variants from dense, genome-wide SNP data. However, no rare enriched genetic risk factors were identified in this dataset, although a subset of families could be genealogically linked to form two extended pedigrees which would suggest shared susceptibility factors. The lack of founder mutations in this isolated population suggests that the majority of genetic risk factors are rare, *de novo* mutations unique to individual nuclear families. We also attempted to use gene ontology (GO)-categories to characterize the biological pathways involved in ASDs, but found significant heterogeneity in identified GO-categories among different genome wide SNP and gene expression datasets.

The results of this study are consistent with other recent studies of genetic risk factors for ASDs. The underlying genetic architecture for this group of complex disorders seems to be highly heterogeneous, with common variants accounting for only a subset of genetic risk. The majority of identified risk factors have turned out to be exceedingly rare, and only explain a subset of the genetic risk in the general population in spite of their high penetrance within individual families. The results of this study, together with other results obtained in this field, indicate that family specific linkage, homozygosity mapping and resequencing efforts are needed to identify these rare genetic risk factors.

Keywords: Autism spectrum disorder, Asperger syndrome, linkage analysis, expression analysis, genome-wide association analysis, isolated population, population genetics, serotonin receptor

Karola Rehnström, Genetic Heterogeneity in Autism Spectrum Disorders in a Population Isolate. Terveyden ja hyvinvoinnin laitos, Tutkimus 21|2009. 192 sivua. Helsinki 2009. ISBN 978-952-245-132-3 (painettu); 978-952-245-133-0 (pdf)

TIIVISTELMÄ

Paikkaan perustuva geenitunnistus eli positionaalinen kloonaus on mahdollistanut hypoteesittomat, koko perimän kattavat tutkimukset joilla voidaan kartoittaa tauteihin ja ominaisuuksiin vaikuttavia perinnöllisiä tekijöitä. Kytkentä-analyysi, sekä sitä seuraava hienokartoitus ja sekvenssianalyysi ovat kuitenkin suuritöisiä ja sairautta aiheuttavan muutoksen tunnistaminen on usein hidasta. Nykyisin perimänlaajuiset assosiaatiomenetelmät tarjoavat oikotien sairautta aiheuttavan muutoksen tai siihen liittyvän pienen kromosomaalisen alueen tunnistamiseen. Lisäksi näistä tutkimuksista saatavaa tietoa voidaan käyttää pienten kromosomaalisten poikkeamien, ns. kopiolukuvarianttien, tunnistamiseen. Näiden kopiolukuvarianttien on osoitettu lisäävän riskiä joillekin sairauksille.

Väestöisolaatit ovat osoittautuneet hyödyllisiksi yhden geenin aiheuttamien sairauksien geenien tunnistamisessa. Suomalainen väestö on geneettisesti yhtenäinen, ja sitä ovat muokanneet perustajavaikutus, genettiset pullonkaulat ja sattuma. Suomalaisen väestön erityispiirteitä voidaan käyttää hyväksi myös tunnistettaessa riskigeenejä monitekijäisille taudeille, vaikka isolaattien hyöty näiden monimutkaisemmin periytyvien tautien geenikartoituksessa ei välttämättä ole yhtä suuri kuin monogeenisten tautien kohdalla. Verrattuna sekaväestöihin monitekijäisten tautien riskigeenen kirjo saattaa olla Suomessa suppeampi, sillä maamme alkuperäiset asuttajat toivat mukanaan vain osan kaikista mahdollisista geneettisistä riskitekijöistä.

Autismikirjon sairaudet ovat ryhmä vakavuudeltaan vaihtelevia, lapsuudessa alkavia neuropsykiatrisia sairauksia. Vaikka perintötekijöiden on todettu vaikuttavan vahvasti autismikirjon sairauksien syntyyn, aktiivisesta tutkimuksesta huolimatta ei sairauksille altistavia geenimuotoja vielä ole tunnistettu kuin kourallinen. Tässä tutkimuksessa olemme käyttääneet molekyyligenetiikan uusimpia menetelmiä hyödyntäen samalla suomalaisen väestön erityispiirteitä tunnistaaksemme näille sairauksille altistavia perinnöllisiä tekijöitä.

Tämän väitöskirjatutkimuksen tulokset osoittavat, että autismikirjon sairauksille altistavat useat eri geenimuodot jopa eristyneessä väestössä. Keskittymällä kliinisesti rajattuun ilmiasuun, eli perheisiin joissa esiintyy vain Aspergerin oireyhtymää (AS), pystyimme toistamaan aikaisemmin raportoimamme kytkennän

kromosomiin 3p14-p24 AS-perheissä. Kytkentäanalyysien tulokset AS:ssä herättävätkin kiinnostavan kysymyksen siitä, onko AS:n perinnöllinen riski yhtenäisempi kuin muissa autismikirjon sairauksissa.

Tutkiessamme tarkemmin aiemmin tunnistamaamme kytkentää kromosomissa 3q25-q27 perheissä, joissa esiintyy autismia ja muita autismikirjon sairauksia tunnistimme assosiaation serotoniinireseptorin 3 alayksikön C (*HTR3C*) ja autismin välillä. Tämä löydös on erityisen kiinnostava, koska serotoniini ja siihen liittyvät biologiset prosessit on aiemminkin liitetty autismikirjon sairauksiin. Assosiaatio *HTR3C*:hen selitti kumminkin vain osan alueella havaitsemastamme kytkennästä, joten on hyvin todennäköistä että alueella sijaitsee myös muita autismille altistavia perintötekijöitä.

Tutkiessamme Suomalaista väestörakennetta löysimme huomattavia eroja eri puolelta Suomea kotoisin olevien ihmisten perimästä. Erot ovat seurausta väestöhistorian aikana tapahtuneista useista pullonkauloista, jotka ovat muokanneet väestön perimää. Käytimme tutkimuksesta saamaamme tietoa valitessamme autismin perimänlaajuista assosiaatiokartoitusta varten geneettisesti yhtenäisen joukon perheitä Keski-Suomesta tunnistaaksemme tähän väestöön rikastuneita, harvinaisia geenimuotoja. Emme kuitenkaan löytäneet tällaisia, vaikka pystyimme yhdistämään osan Keski-Suomalaisista perheistä kahteen suureen sukupuuhun, joka puhuisi yhteisten perinnöllisten riskitekijöiden puolesta. Perustajamutaatioiden puute suomalaisessa väestössä osoittaa, että todennäköisesti autismikirjon perinnöllisistä riskitekijöistä suurin osa on harvinaisia *de novo* mutaatioita, joita esiintyy vain yksittäissä perheissä. Käytimme myös geeniontologia (GO) kategorioita tunnistaaksemme biologisia prosesseja jotka liittyvät autismikirjon sairauksiin. Tunnistimme runsaasti eri GO-katgorioita eri tutkimusaineistoissa, joka omalta osaltaan antavat lisätodisteita siitä että autismille altistavat useat erilaiset biologiset prosessit jotka eivät vältämättä ole samoja eri väestöissä.

Yhteenvetona tutkimuksen tulokset osoittavat että autismikirjon sairauksia aiheuttavat muutokset useissa eri geeneissä, ja yleiset geenimuodot selittävät vain pienen osan perinnöllisestä riskistä. Suurin osa tunnistetuista perinnöllisistä riskitekijöistä on harvinaisia, ja ne selittävät vain pienen osan riskiä väestötasolla, mutta vaikuttavat merkittävästi yksittäisten perheiden riskiin. Tämän tutkimuksen tulokset, yhdessä muualla saatujen tulosten kanssa osoittavat että perinnöllisiä riskejä tulisi etsiä yksittäisistä perheistä käyttäen kytkentä-, homozygotiakartoitus- ja sekvensointimenetelmiä.

Avansanat: Autismikirjon sairaus, Aspergerin oireyhtymä, kytkentäanalyysi, ekspressioanalyysi, genominlaajuinen assosiaatioanalyysi, isolaatti, populaatio-genetiikka, serotoniinreseptori

Karola Rehnström, Genetic Heterogeneity in Autism Spectrum Disorders in a Population Isolate. Institutet för hälsa och välfärd, Forskning 21|2009. 192 sidor. Helsingfors 2009. ISBN 978-952-245-132-3 (tryck); 978-952-245-133-0 (pdf)

ABSTRAKT

Med termen positionell kloning menar man hypotesfria, genomomfattande kartläggningar som har möjliggjort identifikation av genetiska faktorer som inverkar på sjukdomar och egenskaper. Traditionell kopplingsanalys följs upp av finskalig kartläggning och sekvensanalys, som ofta är bade mödosamt och långsamt. Nuförtiden möjliggör genomomfattande associationsanalys ofta en snabbare metod för identifikation av genetiska riskfaktorer. Dessutom möjliggör data från genomomfattande associationsstudier även analys av små kromosomala avvikelser, sk. copy number variants, som medför risk för vissa sjukdomar.

Isolerade populationer har med framgång använts i kartläggning av gener för sjukdomar som orsakas av fel i en gen. Den finländska populationen är genetiskt homogen, och har formats av grundareffekt, flaskhalseffekter och genetisk drift. Den finländska befolkningens särdrag kan också utnyttjas i kartläggning av riskgener för komplexa sjukdomar. Det är dock oklart om isolerade populationer medför lika stor fördel i genkartläggning av komplexa sjukdomar jämfört med monogena sjukdomar. Jämfört med blandade befolkningar är det sannolikt att de genetiska riskfaktorerna för komplexa sjukdomar är färre, eftersom den ursprungliga grundarpopulationen endast bar på en liten del av alla genetiska riskvarianter för en viss sjukdom.

Autismspektrets sjukdomar är en grupp neuropsykiatriska sjukdomar med liknande symptom men varierande svårhetsgrad som manifesterar i tidig barndom. Trots att genetiska faktorer har påvisats spela en viktig roll i risken för dessa sjukdomar, har endast ett fåtal sällsynta genetiska riskfaktorer identifierats trots aktiv forskning inom området. I denna studie har vi använt moderna molekylärgenetiska metoder och samtidigt utnyttjat den finländska populationens särdrag för att identifiera genetiska riskfaktorer för autismspektrets sjukdomar.

Vi använde kopplingsanalys för att identifiera genetiska riskfaktorer i familjer med Aspergers syndrom (AS). Vi har tidigare identifierat koppling till 3p14-p24 i en genomomfattande kopplingsstudie av AS-familjer i Finland, och lyckades nu upprepa detta resultat i nya familjer. Framgången av genetisk kartläggning i familjer med AS väcker intressanta frågor gällande strukturen av de genetiska riskfaktorerna. Kanske är det så att risk för AS styrs av färre riskfaktorer än för autismspektrets sjukdomar i allmänhet.

Vi har tidigare även rapporterat koppling till 3q25-q27 i Finlandska familjer med autism samt andra autismspektrets sjukdomar. I finskalig kartläggning av området identifierade vi association mellan underenhet C av serotoninreceptor 3 (*HTR3C*) och autism. Biologiska processer kopplade till serotonin har tidigare förknippats med autism, vilket gör vårt fynd intressant. Associationsresultaten påvisade dock att association till *HTR3C* endast förklarar en del av kopplingsresultatet, så det är sannolikt att även andra genetiska riskfaktorer för autism är belägna i regionen.

Inom ramen för denna studie ville vi även undersöka populationsstrukturen i Finland, och observerade signifikanta skillnader mellan individer från olika delar av landet. Detta är en följd av de flertal genetiska flaskhalsar den finländska populationen har upplevt under sin historia. Vi använde denna information för att välja ut en genetiskt homogen grupp autism-familjer från Centrala Finland för att identifiera sällsynta genetiska riskfaktorer som anrikats i dessa familjer med hjälp av genomomfattande associationsanalys. Vi identifierade inga anrikade genetiska riskfaktorer, även om vi genom släktforskning kunde koppla ihop en del av familjerna till två stora släktträd, vilket skulle tyda på gemensamma genetiska riskfaktorer i dessa familjer. Avsaknaden av grundarmutationer i den finländska befolkningen tyder på att de genetiska riskfaktorerna för autismspektrets sjukdomar huvudsakligen består av sällsynta *de novo* mutationer som endast förekommer enstaka familjer. Vi analyserade även genontologi (GO) kategorier för att identifiera biologiska processer som är kopplade med autismspektrets sjukdomar. Vi identifierade en heterogen grupp kategorier, som tillsammans med resultaten från associationsanalysen tyder på ett stort antal genetiska riskfaktorer för autism.

Resultaten av denna studie instämmer med resultat från andra studier, och tyder på att ett stort antal genetiska riskfaktorer för autismspektrets sjukdomar finns även inom en isolerad population. Allmänna riskfaktorer inverkar endast i liten mån på risken att insjukna, och de flesta genetiska riskfaktorer som har identifierats är sällsynta. Även om de sällsynta genetiska riskfaktorerna har en väldigt liten inverkan på risk för autismspektrets sjukdomar på populationsnivå, spelar de en betydlig roll inom enstaka familjer. Resultaten av denna studie påvisar att metoder som kopplings-, homozygoti- samt sekvensanalys behövs för att identifiera dessa sällsynta riskfaktorer.

Nyckelord: Autismspektrets sjukdom, Aspergers syndrom, kopplingsanalys, expressionsanalys, genomomfattande associationsanalys, isolerad population, populationsgenetik, serotoninreceptor

CONTENTS

ABBREVIATIONS

aCGH	Array Comparative Genomic Hybridization
ACRD	Autism Chromosome Rearrangement Database
ADI-R	Autism Diagnostic Interview – Revised
AGP	Autism Genome Project
AGRE	Autism Genetic Resource Exchange
AMD	Age-related Macular Degeneration
AS	Asperger Syndrome
ASD	Autism Spectrum Disorder
BAF	B Allele Frequency
bp	Base pair
CD	Crohn‟s disease
CDCV	Common Disease Common Variant
CF	Central Finland
cM	CentiMorgan
CNP	Copy Number Polymorphism
CNV	Copy Number Variant
DNA	Deoxyribonucleic Acid
DSM	Diagnostic and Statistical Manual of Mental Disorders
DWM	Dandy-Walker Malformation
DZ	Dizygotic
F	Inbreeding Coefficient
F_{st}	Fixation index
FXS	Fragile X Syndrome
GEO	Gene Expression Omnibus
GO	Gene Ontology
GWAS	Genome Wide Association Study
HCN	Hyperpolarization-activated Cyclic Nucleotide gated
HFA	High Functioning Autism
HIV	Human Immunodeficiency Virus
HWE	Hardy-Weinberg Equilibrium
IBD	Identity By Descent
IBS	Identity By State
ICD	International Classification of Diseases
IQ	Intelligence Quotient
kb	Kilobase
λ	Genomic inflation factor
LC	Liability Class

LCL	Lymphoblastoid Cell Line
LD	Linkage Disequilibrium
LRR	Log R Ratio
LOD	Logarithm of Odds
M	Morgan
MAF	Minor Allele Frequency
MAGUK	Membrane Associated Guanylate Kinase
Mb	Megabase
MDS	Multidimensional Scaling
MR	Mental Retardation
MS	Multiple Sclerosis
mtDNA	Mitochondrial DNA
MZ	Monozygotic
NFBC66	Northern Finland Birth Cohort 1966
NPL	Non-parametric Linkage
nsSNP	Non-synonymous SNP
OR	Odds Ratio
PCR	Polymerase Chain Reaction
PDD-NOS	Pervasive Developmental Disorder Not Otherwise Specified
QTL	Quantitative Trait Locus
r^2	Square correlation coefficient
RNA	Ribonucleic Acid
ROH	Region Of Homozygosity
RTT	Rett Syndrome
SNP	Single Nucleotide Polymorphism
θ	Recombination fraction
TDT	Transmission Disequilibrium Test
WTCC	Wellcome Trust Case Control Consortium
Z_{max}	Maximum LOD score

In addition, standard one letter abbreviations of nucleotides and amino acids are used.

LIST OF ORIGINAL PUBLICATIONS

This thesis is based on the following original articles referred to in the text by their Roman numerals. In addition, some unpublished data are also presented.

I. **Rehnström K**, Ylisaukko-oja T, Nieminen-von Wendt T, Sarenius S, Källman T, Kempas E, von Wendt L, Peltonen L, Järvelä I. Independent replication and initial fine mapping of 3p21-24 in Asperger syndrome. *J Med Genet.* 2006 43(2):e6.

II. **Rehnström K**, Ylisaukko-Oja T, Nummela I, Ellonen P, Kempas E, Vanhala R, von Wendt L, Järvelä I, Peltonen L. Allelic variants in HTR3C show association with autism. *Am J Med Genet B Neuropsychiatr Genet.* 2009 150B(5):741-6.

III. **Rehnström K**, Ylisaukko-Oja T, Vanhala R, von Wendt L, Peltonen L, Hovatta I. No association between common variants in glyoxalase 1 and autism spectrum disorders. *Am J Med Genet B Neuropsychiatr Genet.* 2008 147B(1):124-7.

IV. Jakkula E*, **Rehnström K***, Varilo T, Pietiläinen OP, Paunio T, Pedersen NL, deFaire U, Järvelin MR, Saharinen J, Freimer N, Ripatti S, Purcell S, Collins A, Daly MJ, Palotie A, Peltonen L. The genome-wide patterns of variation expose significant substructure in a founder population. *Am J Hum Genet.* 2008 83(6):787-94.

V. **Rehnström K*,** Kilpinen H*, Jakkula E, Gaál E, Ylisaukko-oja T, Greco D, Saharinen J, Ripatti S, Daly M, Purcell S, Moilanen I, Varilo T, von Wendt L, Hovatta I, Peltonen L, Integrated Genome-wide Datasets Identify Heterogeneous Biological Processes Affected in Autism. Manuscript.

*These authors contributed equally to this work

1 INTRODUCTION

Positional cloning has provided a hypothesis-free, genome-spanning method for the identification of genes affecting traits and other phenotypes. Usually positional cloning involves identification of families where the same disorder occur, followed by linkage analysis using microsatellites and fine mapping in a larger dataset to improve resolution before performing laborious sequencing and functional analysis. However, recently genome-wide association studies have provided a shortcut to identify finite regions harboring genetic variants affecting traits of interest. The drawback with respect to linkage studies is that a large number of markers, often up to a million, need to be genotyped in a large study sample consisting of thousands of cases and controls. Genome-wide association analysis has revealed a vast number of new risk variants for many disorders. In some studies the positional cloning effort has been facilitated by the innovative use of special populations, such as isolates. Isolated populations have traditionally been linked with successful gene mapping efforts in monogenic diseases, but have recently also proven successful in the identification of risk genes for complex disorders in limited datasets for a subset of phenotypes.

The genetic architecture of traits and diseases vary from monogenic to complex. Monogenic disorders are caused by mutations in just one gene, whereas complex disorders are controlled by a large number of genetic variants with small effects on the phenotype which often interact with environmental factors. Autism spectrum disorders are a group of early onset, highly heritable neuropsychiatric disorders, such as autism and Asperger syndrome. The genetic determinants for these disorders have remained poorly characterized despite decades of intensive research. A fraction of cases have however been linked to syndromic forms of the disorder for which the genetic variants are known, such as Fragile X syndrome and partial duplication of the maternal chromosome 15. Recent large-scale and high resolution efforts using modern methodologies have finally shed some light also on idiopathic forms of these disorders revealing biological processes linked to altered regulation and connectivity at the synapse. In general, most of the identified genetic variants for these disorders seem to be rare, family specific mutations, but incomplete penetrance indicates a role of more common, modifying factors in the etiology as well.

2 REVIEW OF THE LITERATURE

2.1 Genetics of complex traits

The Austrian monk Gregor Mendel was the first scientist to formally describe the laws governing inheritance of traits from parents to offspring. A subset of human traits and disorders, caused by mutations in a single gene, are referred to as Mendelian disorders to highlight the groundbreaking role Mendel‟s insight have had on modern genetics. However, Mendelian disorders are mostly rare, whereas common disorders, which have the most impact on public health, are caused by the interaction of several genetic risk variants. These disorders are often referred to as complex disorders, and are caused by numerous genetic variants. These genetic risk variants individually increase the disease risk only marginally, but together with other genetic and environmental factors result in the disorder. After the structure of deoxyribonucleic acid (DNA) was deduced by James Watson, Francis Crick, Rosalind Franklin and Maurice Wilkins, Mendel‟s observations could finally be explained at the molecular level, and the groundwork was laid for identification of genetics risk factors for human disorders. Although the identification of genetic risk factors for Mendelian disorders has been more successful, recent technological advances as well as greater understanding of the human genome‟s properties have led to breakthroughs in the identification of genetic risk factors involved in complex traits.

2.1.1 Determining the genetic component of a trait

To assess if a trait or disorder is hereditary, family studies are needed to determine if the risk to siblings and relatives of affected individuals is higher than that of the general population. If genetic factors confer risk for the disease, a proband‟s closest relatives, who share the largest part of their DNA, should display increased risk, which decreases toward the population prevalence for more distant relatives. Family studies do not reveal if the mechanism of the familial aggregation is due to environmental or genetic factors. Twin- and adoption studies are needed to elucidate the genetic component. The concordance rate between monozygotic (MZ) twins for a trait or disease compared to that for dizygotic (DZ) twins offers an estimate of the extent of the genetic variation affecting the phenotype. MZ twin concordance also provides a way to estimate the penetrance of complex disorders (Boomsma et al. 2002).

2.1.2 The human genome

The human genome consists of approximately 3.2 billion base pairs (bp), residing on 23 pairs of nuclear chromosomes, and in the small, 16.5 kb ring of mitochondrial DNA (mtDNA). Two teams of scientists, one publicly and another privately funded, competed to finish the sequence of the human genome. The race ended in a tie with both groups publishing the draft sequence in 2001 although it took several years to finalize the drafts (Lander et al. 2001; Venter et al. 2001). The number of genes in the human genome is currently approximated to be around 20 000 (Clamp et al. 2007), but the discovery of novel classes of genes coding for small RNA molecules, such as microRNAs have resulted in a need to further refine the definition of a gene.

Although human genomes are 99.9% identical, there are several forms of variation present between individuals. These consist of repeated sequences of varying length, single nucleotide polymorphisms (SNPs), and copy number variants (CNVs, see section 2.1.4). These variable sites, although representing a small portion of the human genome are responsible for the differences among individuals. These variable sites serve as genetic markers, which can be used in the mapping of genetic factors affecting disorders and traits.

Several classes of repeat polymorphisms exist in the human genome. They can be divided into different classes based on their length and origin. However, only a few are of interest in modern genetic mapping studies. The class of repeat polymorphisms called microsatellites consists of highly polymorphic repeats of 1-10 nucleotides. The microsatellites used in genetic mapping consist of 10-50 copies of di-, tri- or tetranucleotide repeats. The repeat sequences range from tens to hundreds of bps and are flanked by unique DNA sequence and can therefore be amplified by polymerase chain reaction (PCR). The advantage of microsatellites is that they often have up to ten alleles resulting in a high number of different genotypes.

The majority of well characterized variation in the genome is present as SNPs. The International HapMap project, possibly the most important effort in characterizing the variation in the human genome after the determination of the genome sequence itself has discovered millions of SNPs and characterized their frequencies and pairwise correlation in different populations (The International HapMap Consortium 2005; Frazer et al. 2007). The number of SNPs with a minimum allele frequency of 1% in the human genome has been estimated to be at least 10 million resulting in a high density of common SNPs (Kruglyak and Nickerson 2001). Currently, dbSNP contains over 14 million entries, of which 6.5 million are validated. Resequencing of individual genomes is bound to uncover a multitude of rare SNPs, as exemplified by the detection of 1.3 and 0.6 million novel SNPs in the first two sequenced genomes alone (Levy et al. 2007; Wheeler et al. 2008). Only a minority of SNPs,

approximately 135 000 currently listed in dbSNP, potentially impact protein function by directly changing the amino acid sequence. The effect of SNPs in promoters, splice sites and other regulatory regions is still incompletely characterized, but will undoubtedly also have an important impact on the genetic architecture of traits and disorders.

In addition to uncovering the sequence of the human genome, the human genome project advanced the development of new, more efficient sequencing technologies. These have for the first time enabled affordable sequencing of multiple individuals, as well as a multitude of different organisms. Currently the Entrez genome project contains complete genomes of over 950 species, of which 22 are eukaryotes. Another 1180 species have draft assemblies available, and genome sequencing efforts for more than 1000 species are in progress.

The availability of complete genome sequences for a range of species has given rise to a new branch of genetics termed comparative genomics. The aim of this discipline is to establish the relationship of genome structure and function across species. The common evolutionary origin of all species was proposed by Charles Darwin in *On the Origin of Species* exactly 150 years ago. On a molecular scale, the differentiation of species has happened through slow accumulation of changes in their genomes. Therefore, the genomes of more closely related species are more similar compared to more distantly related species. However, comparative genomics can also be used to identify regions which are conserved among species. These regions often correspond to genes and regulatory sequences. Divergence between genomes can be used to determine which genes contribute to the phenotypic differences between species. As an example, comparative analysis between humans and chimpanzees, our closest evolutionary relatives for which the complete genome sequence is available at the moment, have identified genes involved in the evolution of the human brain (Vallender et al. 2008). The sequencing of the Neanderthal genome, which is currently in progress, will add another close relative to aid analyses.

Although the comparison of closely related genomes can be useful for some studies, it is also important that genomes of more distantly related species are present for comparison. It has been shown that mammalian genomes are very similar, and that few novel genes have been introduced in the mammalian lineage. For example, there are only 168 “human specific” genes compared to other mammals. A study of the evolution of the synapse included species from the whole animal kingdom, and concluded that the genes needed for the synaptic machinery are present in all stages of animal evolution. The evolution of the synaptic machinery has instead taken place through the process of paralogous expansion. This means that existing genes have duplicated and the resulting genes have slowly accumulated different variants, leading the encoded proteins to acquire novel functions (Kosik 2009).

2.1.3 Genetic mapping of complex traits

The aim of genetic mapping is to identify a genetic variant which influences the phenotype of interest. Because the predisposing variants are not known, testing can only be performed between a marker genotype and a phenotype. Mapping is easy if strong linkage or linkage disequilibrium (LD) exists between the marker and the variant influencing the phenotype. Mapping is further facilitated if the variant strongly affects the phenotype and has a high penetrance. Factors complicating genetic mapping are the effects of other genes or non-genetic factors, like the environment, on the phenotype (Weiss and Terwilliger 2000).

Primarily, two opposing but complementary hypotheses exist concerning the structure of genetic risk for complex disorders (Reich and Lander 2001). The rare variant hypothesis proposes that the disorder is caused by a small number of variants, each with a large effect on the phenotype with a low frequency (<1%) in the population. Rare variants are often population-specific, as founder effects have established different rare variants in each population. A review of 61 rare variants identified in complex disorders show that rare variants are associated with odds ratios (OR) ranging from 1.1 to over five, with the percentage of variants increasing towards the high end of the OR spectrum, and the mean OR is 3.74 (Bodmer and Bonilla 2008).

The common variant hypothesis proposes that common disorders are caused by the interaction of numerous small effect variants common in the population. The assumption that common variants comprise risk factors for common disorders has been termed the common disease-common variant (CDCV) hypothesis (Lander 1996; Chakravarti 1999). Of the common variants identified to date, most ORs are smaller than two, and the mean OR of 217 common variants from the literature is 1.36 (Bodmer and Bonilla 2008).

The separation of risk factors into common and rare is more than theoretical, as the optimal gene mapping strategy depends on the underlying genetic architecture of the trait of interest. As rare variants are usually family-specific, they can be identified using linkage analysis as linkage analysis investigates the co-segregation of chromosomal loci with the trait within families. To obtain optimal statistical power linkage analysis, large multigeneration families with multiple affected individuals should be ascertained, although methodologies for other family structures have also been developed. If the trait is thought to be caused by common variants, association analysis provides better power to identify genetic risk variants. Although association analysis can also be performed in families, usually it is performed in large case-control cohorts, which are often easier to collect than large pedigrees, especially for

phenotypes with a high age of onset, as parents of index cases might not be available for the study.

Positional cloning makes it possible to locate genetic risk factors for diseases and traits without an *a priori* hypothesis of the genes involved by systematically scanning the entire genome. The process begins by collecting and carefully characterizing a sufficient number of carefully phenotyped families. Subsequently, linkage analysis is performed using a sparse set of genetic markers such as microsatellites or SNPs covering the entire genome to identify large chromosomal regions co-segregating with the trait of interest. The most interesting regions identified in the linkage scan are followed up by a more dense set of microsatellite markers or SNPs. The inclusion of additional families at this stage helps refine the region of interest as more recombinations will be located in the region of interest. Finally, when a sufficiently small region has been identified, *in silico* methods can be utilized to identify candidate genes of interest. This involves data mining of the various publicly available databases which contain for example information of gene expression patterns, gene function, interaction partners, DNA sequence and sequence conservation among species. Before any wet-lab experiments are performed, data is extracted and from the databases and combined to make it possible to predict which genes in the identified regions functionally make the most interesting candidates. These interesting candidate genes can then be analyzed by resequencing to identify possible risk variants. Sequencing efforts can identify a multitude of possible risk variants, and again extensive *in silico* characterization of these variants is usually needed. Subsequently, functional experiments are needed to verify the variant‟s role in the etiology of the disease. Although this approach has been successful for identifying genes underlying Mendelian disorders, the success has been more limited for complex disorders.

To obtain more power to identify common variants predisposing for common disorders, the positional cloning approach has been mostly replaced by large-scale genome-wide association analysis where large cohorts of cases and controls are genotyped for a high-density SNP marker set. Markers where allele frequencies differ between cases and controls identify limited candidate regions, bypassing the need for extensive fine mapping. However, although the identification of limited candidate regions is more straightforward in this approach, the identification of the true risk variant can still be laborious, and resequencing, bioinformatics and functional approaches are needed to evaluate the role of the identified variants.

Previously sequence analysis has only been used after limited regions or genes of interest have been identified through linkage or association analysis to identify risk variants present in cases but not controls. However, second generation sequencing technologies have enabled genome-wide sequencing of large cohorts, and the next

step in the genetic mapping will possibly be the sequencing of the whole genome in cases and controls. Whole genome resequencing produces a massive amount of data, and needs to be followed by bioinformatic analyses and systems biology approaches to separate risk variants from benign polymorphisms.

Linkage analysis

Linkage analysis is used to localize a chromosomal region co-segregating with, or linked to, a nearby marker locus within a pedigree. Recombination during meiosis generates novel allele combinations in each generation, and genetic markers can be used to detect which chromosomal segments have been inherited from each parent. The probability for recombination to occur between any two sites in the genome is related to their physical distance, and described in genetic terms by the recombination fraction (θ) which ranges from 0 corresponding to the two loci always co-segregating, to 0.5 corresponding to independent segregation, as is the case for loci residing on different chromosomes. For linkage studies, physical maps are replaced by genetic maps relating markers by their probability to be separated by recombination. The distance measure used is Morgan (M) which is divided into 100 centiMorgans (cM). A genetic length of one Morgan corresponds to one expected recombination event (crossing over), resulting in one cM corresponding to 1% probability for recombination between loci. Over short distances one cM corresponds roughly to 1 Mb, but the rate varies across the genome and is sex-specific, with higher recombination rates in females than in males. For example chromosome 5 is 260 cM in females and 151 cM in males, corresponding to the expected 2.6 and 1.5 recombination events respectively. The length of the human genome is 44.60 M in females and 25.90 M in males (Kong et al. 2002).

In theory, linkage analysis is simply based on calculating the number of recombinant genotypes in a pedigree. However, as recombinations cannot be directly observed, in reality a likelihood measure is employed to estimate the probability of co-segregation of a genetic marker with the trait of interest compared to the null hypothesis. The hypothesis is tested by a likelihood quotient where the logarithm of odds (LOD) score, Z(θ), is maximized for all values of θ and compared to the null hypothesis that the loci are unlinked, i.e. θ=0.5:

$$Z(\theta) = \log_{10} \frac{L(\theta)}{L(\theta = 0.5)}$$

Different algorithms have been developed for linkage analysis. The Elston-Stewart algorithm (Elston and Stewart 1971) is efficient when big pedigrees but few markers are used whereas the Lander-Green algorithm (Lander and Green 1987) makes it

possible to use many markers but smaller pedigrees. Approximate, simulation-based methods are needed if big pedigrees are analyzed with a large number or markers.

Linkage analysis can be divided in to parametric and non-parametric analyses. In parametric linkage analysis, the mode of inheritance has to be determined using penetrance parameters and allele frequencies. This is possible for Mendelian disorders where the mode of inheritance is known, but is not directly applicable to complex traits. Parametric linkage analysis is very powerful if the correct inheritance model is known, and has been shown to remain robust to model misspecifications as long as the inheritance model is correct (Nyholt 2000). When parametric analysis is performed, data can be analyzed under several genetic models and the highest LOD-score (Z_{max}) can be used as test statistic for linkage (Clerget-Darpoux et al. 1986; Greenberg et al. 1998). The testing of multiple inheritance models increases the likelihood of detecting true linkage, but also increases type I error (false positives) due to multiple testing. However, if maximizing is only performed over one parameter, in this case the inheritance model, and only a few models are used, e.g. one recessive and one dominant, the increase in significance levels is relatively modest (Hodge et al. 1997).

Non-parametric or model-free linkage analysis does not require that the inheritance pattern of the disease is defined, and is therefore more suitable for complex traits. The non-parametric linkage algorithms explore the extent of allele sharing between individuals with the same phenotype. Because the non-parametric linkage methods do not assume any inheritance model they tend to be more robust than parametric analyses, but are less powerful than parametric methods using the correct mode of inheritance (Elston 1998).

A LOD-score of 3 corresponds to an odds ratio of 1000:1 in favor of the alternative hypothesis and has traditionally been considered as a genome-wide threshold for significant linkage. This corresponds to a pointwise significance of $p=1\cdot10^{-4}$ and a genome-wide significance level of 9% (Chotai 1984). To obtain a genome-wide significance level of 5%, corresponding to a point-wise significance of $p=5\cdot10^{-5}$, the LOD score threshold needs to be raised to 3.3 (Lander and Kruglyak 1995).

Replication of positive results in independent datasets is important to rule out false positive linkage findings. Simulations have shown that the variability of the location estimate in linkage studies in complex disorders for a confidence interval of 95% consists of dozens of cMs (Roberts et al. 1999). A failure to replicate linkage results could imply that the effect of the locus has been overestimated in the original study, and a much larger material is needed to replicate the initial finding (Lander and Kruglyak 1995). Replication studies can also lead to contradictory results due to population heterogeneity, diagnostic differences and/or statistical fluctuation.

Because of genetic differences between populations replication is easier in the same population where the initial successful mapping study was performed (Goring et al. 2001). Meta-analysis of several scans can be used to evaluate if the overall evidence for linkage is significant providing that the underlying genetic causes can be assumed to be shared in all populations included in the meta-analysis.

Association analysis and linkage disequilibrium

Association analysis in its simplest form consists of comparing allele frequencies at genetic markers in regions of interest in two cohorts - one consisting of affected individuals and the other of controls. The statistical significance of the difference in allele frequencies in the form of a p-value can be assigned using the χ^2 distribution with one degree of freedom.

Traditionally association studies have been used to test if polymorphisms in a candidate gene are associated with the phenotype of interest. This has required an *a priori* hypothesis concerning the gene of interest, either about the position of the gene in a linkage peak, or about its function in a biological process or pathway related to the phenotype. Although a few groundbreaking studies pioneered the genome-wide association approach in the 1990s (Houwen et al. 1994; Puffenberger et al. 1994), mostly linkage-based approaches were used in genetic mapping at that time. Linkage analysis is efficient for detecting rare highly penetrant variants, and was therefore successful in Mendelian disorders. However, the poor replication of results in complex disorders led to a new hypothesis suggesting that susceptibility to common disorders could be conferred by a large number of common, small risk loci. In a highly influential paper, Risch and Merikangas (1996) proposed that association studies would have more power than linkage studies to identify common variants with a small effect on the phenotype.

As the human genome contains millions of points of variation, testing each of these variants for association with the trait of interest would be highly impractical and result in an overwhelming multiple testing problem. However, not all SNPs are independent, and loci close to each other are often inherited together due to a low probability that recombination will occur between them. In the absence of crossing over, the whole disease-causing chromosome with all its alleles would be transmitted as a block to the next generation. Due to recombination this block, or haplotype, becomes shorter for every generation. If the locus contains a variant which influences the disease phenotype, affected individuals are expected to share the same haplotype (combination of alleles) around the disease locus whereas unaffected individuals have a random assortment of alleles. Markers that are inherited as a group in the described manner are said to be in LD.

LD can be observed as block-like structures across the genome. These blocks consist of regions with low recombination rate and only a few common haplotypes flanked by regions experiencing high recombination rates (The International HapMap Consortium 2005; Frazer et al. 2007). A majority of haplotypes are shared between different populations, and these common haplotypes account for a vast majority of all chromosomes (Jakobsson et al. 2008). This means that within these haplotype blocks one, or a few, SNPs can be selected to capture the information of adjacent SNPs. Only these „tagging" SNPs need to be genotyped, and the common variation in the entire genome can be captured using approximately half a million SNPs or less. This is referred to as an indirect association approach, where only a subset of SNPs need to be genotyped (Collins et al. 1997). The extensive work performed by the International HapMap consortium has characterized the patterns of LD across the genome in four populations, and the publicly available data has made it easy to select tagSNPs for candidate regions or genes of interest. It should be noted that approximately 1% of the SNPs in the second phase of the HapMap project were reported to be untaggable, i.e. not in LD with any surrounding SNP. Therefore, although the LD structure makes it possible to utilize indirect approaches for most of the genome, some parts will require a direct approach where each individual SNP needs to be genotyped (Frazer et al. 2007).

The results of the efforts of the International HapMap consortium, together with the fast progress in high-throughput genotyping technologies have enabled affordable genome-wide SNP genotyping of hundreds of thousands of markers in the last few years, making the ideas proposed by Rich and Merikangas in 1996 a reality. Early genome-wide association studies often failed to identify statistically significant results because of small sample sizes and only replicating known linkage findings. Large international collaborations with thousands, or even tens of thousands of samples, have lead to well replicated findings of small effect common variants.

Among the first success stories of genome-wide association studies (GWASs) was the identification of *complement factor H* (*CFH*) as a risk factor for age-related macular degeneration (AMD). Although *CFH* was identified in a rather small GWAS (96 cases vs. 50 controls) (Klein et al. 2005), the two identified SNPs remained significant after Bonferroni correction for the large number of tests. As the gene was located in a known linkage region, two other studies published simultaneously verified *CFH* as a risk gene for AMD through different approaches (Edwards et al. 2005; Haines et al. 2005). Later meta-analysis have verified the Y402H variant as a risk factor, heterozygous carriers of the risk haplotype have a 2.5 fold increased risk, and homozygous carriers 6 fold increased risk of AMD (Thakkinstian et al. 2006). However, fine mapping studies have revealed other variants showing stronger association to AMD than the non-synonymous variant

suggesting that the identified susceptibility variants in complex disorders are not always similar to those identified in Mendelian disorders (Li et al. 2006). In 2007, the large Wellcome Trust Case Control Consortium (WTCCC) reported results of GWASs for seven phenotypes using 2000 cases and 3000 shared controls across all phenotypes. Interestingly, the results for different phenotypes were very different, with significant peaks identified for coronary artery disease, Crohn˝s disease (CD), rheumatoid arthritis, and type 1 and type 2 diabetes, but much less significant results for bipolar disorder and hypertension suggesting different underlying architecture of the genetic susceptibility variants (WTCCC 2007).

Gibson and Goldstein (2007) outlined three primary goals for GWASs. The first is to predict risk and allow for interventions for individuals with high risk. The second is to improve understanding of the biological processes underlying the disease and the third is to identify subclasses of clinically similar diseases that have a different etiology and therefore respond to different treatment. In light of these goals, right now GWASs fulfill the expectations of the second goal, but have had only limited success with the other two. Also, with regard to the second goal, the identification of novel genetic risk factors is only the beginning of the characterization of the biological processes underlying the disorders, and a wide range of functional studies are needed to characterize the role of these variants in disease etiology. Progress on this part will further the two other goals, as understanding the biological mechanism of action is required for devising new treatment and interventions.

Although GWASs suffer from the same problems as association studies of candidate genes, in that cases and controls need to be well matched for ethnicity to prevent spurious associations, GWASs actually contain their own internal control for this, which could not be obtained using only a small number of markers. In GWASs, the p-values are expected to follow the null distribution, and differ from the expected only for a small number of loci with highly significant p-values. If cases and controls differ significantly from each other, this can be observed as a large number of associated loci, which cannot all be linked to the etiology of the trait of interest. Statistical approaches have been developed to deal with inflation of the test statistic due to population substructure making it possible to combine data from different populations to obtain large sample sizes without inflating the false positive rates of the studies (Devlin and Roeder 1999; Price et al. 2006). However, it should be noted that combination of different cohorts is only beneficial if the underlying genetic risk factors can be assumed to be shared across the populations.

Although both the CDCV theory and the International HapMap project have been criticized, (Pritchard and Przeworski 2001; Terwilliger and Hiekkalinna 2006), results of the project have proven useful for at least a subset of phenotypes. To date, GWASs have identified several hundred novel susceptibility loci for both clinical

conditions as well as disease-related or normal traits (Altshuler et al. 2008). GWASs have uncovered numerous new risk variants for many phenotypes, when large samples providing adequate statistical power are used in concert with appropriate and stringent data analysis suitable for excluding false positives due to population stratification and other artifacts. Properly designed studies have also shown that the susceptibility variants confer at best modest risk, with effect sizes in the range of 1.1-1.5. It is likely, that large meta-analysis and collaborations will reveal further variants with even smaller effects.

2.1.4 Copy number variants

Although the role of chromosomal rearrangements in some disorders, including mental retardation and cancer among many others have been recognized for a long time, small deletions and duplications have only recently been recognized to be a commonly segregating form of genetic variation in the general population. The first studies were performed using array comparative genomic hybridization (aCGH). Despite a relatively low density of probes, these studies revealed numerous differences between phenotypically normal individuals (Iafrate et al. 2004; Sebat et al. 2004). Thereafter, a wave of studies confirmed and extended the original finding using varying methods (Estivill and Armengol 2007).

The term CNV has come to mean a variable site exceeding approximately 1kb in size and ranging up to several Mb (Feuk et al. 2006). Size-wise this places CNVs between chromosomal rearrangements which can be seen by light microscopy, and microsatellites and other small indels which can be assessed by traditional sequence-based approaches. Variant sites with a frequency of 1% in the population are referred to as copy number polymorphisms (CNPs). However, it should be noted that in order to determine allele frequencies of CNVs, large studies and reliable calling algorithms are needed and thus, the determination of allele frequencies is still a work in progress. The majority of CNVs are inherited in a Mendelian manner (Locke et al. 2006; McCarroll et al. 2006). A large scale, high resolution study reported that less than 1% of CNVs did not correspond to a Mendelian inheritance pattern (McCarroll et al. 2008b). The same study also concluded that previous studies have tended to overestimate the sizes of CNVs and that most CNVs fall into discrete CVN regions with defined boundaries.

There are several ways in which CNVs could confer risk for disorders. The most obvious mechanism is gene dosage, as deletions could result in a diminished and duplications to an increased gene dosage. Further, homozygosity for deletion alleles of genes results in null expression of the gene or of the exons encompassed by the deletion. CNVs in regulatory regions of genes could also have an impact on gene

expression levels. A study of the role of CNVs in gene expression concludes that 18% of the variation in expression levels of 15,000 genes is attributable to copy number differences (Stranger et al. 2007). In rare cases, deletion alleles could also uncover deleterious alleles in the remaining gene copy.

Rare, large, often *de novo* CNVs have been linked to varying phenotypes including autism, developmental disorders and schizophrenia (de Vries et al. 2005; Sebat et al. 2007; Stefansson et al. 2008). However, it should be noted that in most studies, data was not collected specifically for the purpose of CNV identification. SNP intensity data from GWASs have been used in several studies to identify CNVs as a spin-off project from the original GWAS. This could introduce biases, notably the lack of control for intelligence quotient (IQ) differences in patients versus controls. As CNVs and larger chromosomal abnormalities have been commonly identified as causative for intellectual disability, the lack of control for patient IQ could result in increased CNV rate estimates as well as frequency estimates for CNP alleles in cases (Joober and Boksa 2009). In a schizophrenia study where IQ data was available, association was detected between CNVs and low IQ (Walsh et al. 2008). In the general population, CNVs exceeding 500kb have been identified in 5-10% of individuals, and CNVs exceeding 1 Mb in 1-2% (Itsara et al. 2009).

Recently, reports have also emerged of CNPs associated with common disorders. A 20kb deletion polymorphism located upstream of *IRGM* and in full LD with the most significantly associated SNP in a GWAS for CD was found to confer susceptibility to CD (McCarroll et al. 2008a). The deletion polymorphism was shown to regulate the expression of *IRGM*, a gene involved in autophagy, a cellular process known to be involved in CD from earlier studies (Rioux et al. 2007). Individuals with a low copy number of the human immunodeficiency virus (HIV)-suppressing chemokine receptor *CCL3L1* have a higher risk for HIV infection and more aggressive disease progress (Gonzalez et al. 2005). High copy number at the same gene has been associated with increased risk for rheumatoid arthritis and type I diabetes (McKinney et al. 2008). A low copy number of *FCGR3B*, a receptor for IgG Fc-fragment, has been associated with systemic organ-specific autoimmune disease (Fanciulli et al. 2007). However, the association results were not consistent across all cohorts, as a French cohort of Wegener‟s granulomatosis showed association to high copy number of *FCGR3B*. High copy number of *DEFB* and lack of *LCE3B* and *LCE3C* have been shown to confer risk for psoriasis (Hollox et al. 2008; de Cid et al. 2009). The high number of disorders involving immunological processes can be explained by the fact that the distribution of CNVs overlapping with genes is not random across the genome. CNVs are enriched in genes encoding secreted, olfactory, and immunity proteins, explaining the strong link between CNVs and immunological disorders (Nguyen et al. 2006). In addition to increased

CNV rates in these genes, significantly elevated synonymous and nonsynonymous nucleotide substitution rates were observed in these genes as well. These genes mediate processes that are involved in the adaptation to new environmental niches, and results suggest that at least CNVs with high frequency have been retained in the human population due to positive selection.

2.1.5 Identification of risk variants through analysis of gene expression

Variants in DNA can modify the encoded protein, if the variants are located in coding regions. However, variants in promoters or other functional elements can modify the level of expression. Gene expression level cannot be directly assayed from the DNA, as each cell contains one copy of the genome. However, the amount of ribonucleic acid (RNA) transcribed from the genes can vary in different individuals and cell types, and correlates with the expression level of proteins. Genes encoding proteins with different expression levels in cases versus controls could influence the phenotype studied.

The expression levels of single genes of interest can be determined using reverse transcription-PCR. However, this method is laborious when there are multiple genes of interest. Furthermore, a prior hypothesis is needed to identify genes linked with the phenotype of interest. Microarrays containing probes for all known genes enable determination of expression levels for all RNA molecules in a single experiment, and have become a mainstream application in molecular genetics over the last decade. Early microarrays used partial complementary DNAs or expressed sequence tags as probes for genes, but virtually all commercially available gene-expression microarrays today contain short synthesized oligonucleotide probes.

As the expression patterns of genes are distinctive for each tissue, expression profiling of disorders involving the brain poses specific problems. In studies of human gene expression, usually only postmortem brain samples are available. These samples often represent the state of the brain years or even decades after disease onset. However, gene expression profiling in animal models, such as mice can be used to identify differentially expressed genes, which can be further investigated in humans. Using this approach an elegant study design was used to identify genes associated with anxiety first in mice and then in man (Hovatta et al. 2005). RNA was extracted from brain regions previously shown to be involved in anxiety from mouse strains differing in anxiety-like behavior. A total of 17 genes with expression patterns correlating with the level of anxiety in the different mouse strains were identified. Two enzymes, *glyoxalase I* (*Glo1*) and *glutathione reductase 1* (*Gsr*) were further investigated, and were shown to influence anxiety-like behavior in the mice. An association study in 13 human homologues of the anxiety-related genes

identified in mice revealed that some of the SNPs in a subset of these genes confer risk for anxiety in humans as well (Donner et al. 2008).

Microarray expression studies produce large amounts of data, as the expression levels of tens of thousands of genes is investigated. For some phenotypes, the identification of transcripts that are differentially expressed in cases and controls using stringent statistical criteria for differential expression can reveal candidate genes. These genes can be studied further using functional assays as well as DNA-sequence-based approaches to reveal new information of the disease pathology or DNA risk variants. However, for complex disorders, the difference in expression levels between cases and controls can be subtle. In these cases few or possibly no genes display significant differences in expression level, but there is a large set of suggestively differentially expressed genes. Often these sets of genes are too large to draw conclusion of unifying biological processes by manual annotation of the identified genes.

Gene Ontology (GO) is a controlled vocabulary describing the functions and relationships of genes and biological function a hierarchical manner (Ashburner et al. 2000). GO can be used to assign functional annotations to genes, such as large lists of differentially expressed genes identified in expression studies. By comparing the relationship of these annotations, large gene lists can be reduced to lists of related functions, facilitating the identification of biological processes related to the genes. Algorithms have been developed to test if differentially expressed genes are enriched in certain GO-categories compared to expected distributions (Breitling et al. 2004a; Breslin et al. 2004). Analysis of enriched GO-categories was used in an expression study of monozygotic twins, discordant for body mass index in order to identify genes involved in acquired obesity. RNA profiling was performed in fat biopsies, and downregulated genes were identified to be enriched in mitochondrial branched chain family amino acid metabolism (Pietilainen et al. 2008). Furthermore, a reduced number of mitochondria and increased serum levels of branched chain amino acids, which increase the secretion of insulin, were identified in the obese co-twin. These results strongly implicate a novel biological process, the mitochondrial amino acid metabolism, in acquired obesity.

2.2 Population substructure in genetic studies

Association analysis requires that the cases and controls included in the study are appropriately matched to avoid false positive results due to the differential distribution of alleles at marker loci in different populations. This effect can remain undetected in candidate gene studies, where only tens or hundreds of SNPs are tested. However in GWASs, a biased selection of cases and controls will be detectable as inflation in the expected distribution of test statistics. Genome-wide association datasets have enabled the characterization of population structure, and recently a multitude of studies investigating the population substructure locally, internationally and even globally have been published. The detailed characterization will allow even better study design of GWASs in the future, and at the same time also provide information concerning historical migrations and the relationship and admixture of current populations.

2.2.1 Finnish population history

Finland has been inhabited since the Pleistocene ice withdrew and the land rose from the sea approximately 10 000 years ago to form the Finnish peninsula. Evidence of both eastern-derived comb ceramic culture as well as western derived corded ware culture has been observed in archeological finds from around 5000-6000 years and 4000 years ago, respectively. These two populations form the major contributions to the genomes of Finns today, and represent classic examples of bottlenecks in the Finnish population history. Later smaller additions to the population have come from Germanic, Scandinavian and Baltic peoples. The population size in Finland has remained small, and immigration during the last thousand years has been low. The main language spoken in Finland today, Finnish, is a member of the Fenno-Ugrian family of languages that includes languages such as Estonian and Hungarian, in contrast to most other European languages that belong to the Indo-European language group. Although linguistic information and genetic origin do not necessarily coincide, this difference provides supporting evidence of an eastern immigration.

For millennia, only southern Finland and the coastal regions, referred to as the early settlement region, were inhabited. In the 16th century, the eastern and northern parts of the country, referred to as the late settlement, were inhabited by an internal migration starting from the South Savo region in southeastern Finland. There are several reasons for this migration, including the taxation benefits bestowed upon individuals who moved to the uninhabited eastern regions to reinforce the border against Russia. In addition, the rapidly growing population needed more land to

cultivate, and burn beating enabled the population to move into new regions. New settlements consisted of a small number of individuals, which together with geographical isolation resulted in severe bottleneck effects and genetic drift. By the end of the 17th century, most of Finland was inhabited and the population had increased to 400 000. However, the great famine during 1696-1968 killed almost a third of the population, forming one more bottleneck in the population. Since then, the population has rapidly expanded from 250 000 to over 5 million inhabitants today (Varilo 1999).

2.2.2 Finnish population genetics

Until recently, population genetics studies have been based on uniparentally inherited mitochondrial- or Y chromosomal markers. The Y chromosome is useful for population genetic studies due to its paternal and haploid mode of inheritance which result in a four times reduced effective population compared to nuclear loci. In addition, the Y chromosome only undergoes recombination in the most telomeric parts of the chromosome, and thereby all observed variation on the chromosome arises as a consequence of new mutations. Two main lineages can be detected in Finnish Y chromosomal studies, one of eastern and one of western origin (Sajantila et al. 1996; Kittles et al. 1998; Lappalainen et al. 2006).

The presence of both eastern and western Y chromosomal haplogroups support the hypothesis that the Finnish population has been founded by at least two major waves, one of western and another of eastern origin. Similar distinction of east and west can be observed in other phenomena, such as dialectal groups and disorder prevalence. The divisions closely follow the historical border between Swedish-ruled Finland and Russia established in 1323 (Norio 2003c) The expansion of the eastern haplogroup has been dated 2000 years prior to the expansion of the western haplogroup, supporting archeological data (Kittles et al. 1998). Finnish Y chromosomes show a reduction in genetic diversity when compared to other populations, including the neighboring populations, such as Estonians and Swedes, suggesting a bottleneck in the population (Sajantila et al. 1996; Lappalainen et al. 2006).

Maternal ancestry can be tracked using mtDNA, which does not undergo recombination and has a higher mutation rate compared to nuclear DNA. The distribution of mitochondrial haplogroups in Finns, excluding the Saami people in Northern Finland, resemble that of the rest of Europe (Sajantila et al. 1995; Kittles et al. 1999; Finnila et al. 2001). In contrast to results obtained in Y chromosomal studies, mtDNA does not display significantly reduced genetic diversity which could be explained by different mutation rates in the Y chromosome and mitochondria, or

by different effective population sizes in males and females (Hedman et al. 2007). However, if only slowly evolving positions are considered, a bottleneck can be dated approximately 4000 years ago using mitochondrial data. This is in agreement with the archeological evidence of the introduction of agriculture 4900-4000 years ago (Sajantila et al. 1996).

The unique population history of Finland has led to enrichment of certain alleles and extinction of others by genetic drift. This has resulted in a unique pattern of disorders consisting of 36 mostly recessive disorders that are overrepresented in Finland compared to any other population. This set of diseases is often referred to as the Finnish disease heritage. The birthplaces of grandparents of individuals affected with these diseases are often not located in areas of high population density, but rather in areas of the late settlement, or in geographically isolated regions (Norio 2003b). Some Mendelian disorders common in European populations, such as phenylketonuria and cystic fibrosis, are absent from Finland, consistent with the principles of genetic drift (Norio 2003a).

2.2.3 Genome-wide SNP data in population genetic studies

The abundance of SNP data available from association studies for various phenotypes has recently enabled population genetic inferences to be drawn from the allele frequencies of dense genetic markers spanning the entire genome. A multitude of studies have recently emerged reporting, in essence, the same main results: Principal componenet-based analyses show that the distribution of samples strongly correlates with geographic distribution and the first two principal components correspond to the east-west and north-south gradients, with the east-west gradient explaining most of the variation of the data (Heath et al. 2008; Lao et al. 2008; Price et al. 2008; Tian et al. 2008; Yamaguchi-Kabata et al. 2008). The dense map of markers has been shown to yield similar analysis in single SNP-based and haplotype-based analysis, suggesting that unphased SNPs provide the same amount of information as phased haplotypes (Jakobsson et al. 2008). It is also evident from these studies that the majority of variation resides within populations and that between-population variation is smaller. The between-population variation reveal an underlying structure which is primarily accounted for by random drift at neutral loci, but also some regions under high selection (Voight et al. 2006).

When the structure of European populations is investigated, Finns cluster separately from the rest of the European populations (Lao et al. 2008; Salmela et al. 2008). Marker heterozygosity decreases from southwest Europe towards northeast, while LD increases over the same trajectory. Therefore, among the European populations Finns display the highest LD and the lowest heterozygosity (Lao et al. 2008). A

study of genome-wide SNP data from Eastern and Western Finland confirm the east-west difference which was originally observed using Y chromosomal haplotypes. Furthermore, identity-by-state (IBS) sharing was higher within Finland compared to other populations in Northern Europe (Salmela et al. 2008).

2.2.4 Isolated populations in disease gene mapping

Isolated populations are defined as populations originating from a small number of founders that experience little immigration and expand primarily through population growth. As a result, they are also known as founder populations. The decreased genetic diversity, which can still be observed in isolates today, is a direct consequence of the genetic bottleneck caused by the limited number of founders. The subsequent rapid population growth is a prerequisite for the population to be useful for genetic mapping, as there are sufficient numbers of individuals affected with disorders that have a low frequency in the population. The same demographic phenomena that have resulted in the decreased genetic diversity have also resulted in extended LD and increased homozygosity. The drastic reduction in population size during the population history has enabled genetic drift to drastically change allele frequencies resulting in distinctive patterns of disorders characteristic for the specific population (Morton et al. 2003; Norio 2003a; Charrow 2004).

In addition to genetic homogeneity, environmental and cultural factors are more homogenous in isolates compared to large, admixed populations. Factors such as shared language and religion increase social cohesion, whereas education, health care and diet minimize the environmental heterogeneity. In many founder populations, careful population registries or church parish records are available for genealogical studies, and allow construction of extended pedigrees with multiple affected individuals.

The genetic architecture of isolated populations has proven to be extremely useful for the identification of genes underlying Mendelian disorders. The bottlenecks experienced by the population have enriched a limited number, sometimes only one genetic variant resulting in the disorder. Identification of this one risk variant, often residing on a large haplotype shared between seemingly unrelated affected individuals is relatively easy. However, the use for isolated populations in mapping of susceptibility genes for complex disorders is more controversial. Recent studies discussed below provide evidence of both successes and failures. However, it seems that even if population isolates have proven successful in identification of novel loci for a variety of phenotypes, they might not be as useful for complex gene mapping as they have been in gene mapping for monogenic disorders. Also, the role of the identified loci still need to be further assessed to determine whether they constitute

isolate-specific risk factors, or if the results of these studies can be extended to outbred populations as well.

Successes in the identification of risk variants for complex traits in isolated populations

Isolated populations display increased LD compared to admixed populations. Originally, one of the primary benefits envisioned for isolates in the identification of genetic risk factors for complex traits was that the increased LD would significantly lower the number of markers needed to tag the entire genome (Shifman and Darvasi 2001; Service et al. 2006). However, today this gain is mostly theoretical, as genotyping is performed using pre-designed SNP sets, which have been chosen to tag the majority of allelic variation in the genomes of outbred European populations. It is therefore not surprising, that tagSNPs chosen using the HapMap European population perform well in several isolates (Service et al. 2007). However, it has been suggested that the tagging power of these SNPs even for rare haplotypes is better in isolated populations, due to the excess of information provided by the off-the-shelf genotyping chips in populations with high LD.

The more significant value of isolated populations is the reduced genetic variability, which has resulted in enrichment of a subset of susceptibility alleles in the population. As isolated populations have undergone bottlenecks, the susceptibility variants often reside on a more homogenous haplotypic background compared to mixed populations. The genes for all but one disorder of the Finnish disease heritage have been identified, and for most disorders, one major mutation accounts for >90% of Finnish cases. However, a multitude of other mutations causing the diseases have been identified in other populations (Norio 2003b). Another well documented example of founder mutations in isolates is the enrichment of only three mutations in *BRCA1* and *BRCA2* conferring risk to familial breast cancer in the Ashkenazi Jewish population (King et al. 2003). The enrichment of certain susceptibility variants will require fewer individuals to find these shared genetic factors.

One of the success-stories of genetic mapping of risk variants for complex disorders in isolates is provided by data from the Icelandic company deCODE genetics. By utilizing extensive genotype and phenotype data combined with genealogical information, an unprecedented myriad of susceptibility variants for complex disorders and traits have been identified. Linkage, and subsequently association mapping have identified common genetic risk variants for various disorders and traits, including multiple types of cancer (Gudmundsson et al. 2007a; Gudmundsson et al. 2007b; Stacey et al. 2007; Goldstein et al. 2008; Gudmundsson et al. 2008; Stacey et al. 2008; Rafnar et al. 2009), pigmentation (Sulem et al. 2008), myocardial infarction and aneurysms (Helgadottir et al. 2007; Helgadottir et al. 2008), glaucoma

(Thorleifsson et al. 2007) and type 2 diabetes (Grant et al. 2006). In these studies, the genomic inflation factor (λ) was used to control for population structure and relatedness (Helgason et al. 2005). In some studies, previous linkage findings in the same population were used to prioritize regions for follow up if no genome-wide significant results were identified in the primary stage of the GWAS. This approach has resulted in the identification of susceptibility variants at previously identified linkage peaks.

An elegant study design combining the benefits of enriched risk alleles within a subisolate with the power of genealogical information was used to identify risk variants for multiple sclerosis (MS). The prevalence of MS is twice as high in an internal subisolate of Finland compared to the rest of the country (Sumelahti et al. 2001). Using genealogical data, and carefully matched controls originating from the same geographically limited region, a haplotype shared identical by descent (IBD) between affected individuals was identified close to *complement component 7* (*C7*) located under a previously identified linkage peak at 5p. The haplotype association was replicated in an independent sample from the same subisolate, and the risk haplotype correlates with increased expression of *C7*. A trend toward association was also observed in other European populations, suggesting that the risk variant has been enriched in the internal isolate, and confers susceptibility to MS in other populations as well (Kallio et al. 2009).

The role of isolate-specific variants for common disorders is still incompletely characterized. A recent GWAS of close to 5000 individuals from the late settlement region of Finland for metabolic traits replicated numerous loci reported in other populations for phenotypes such as blood triglyceride levels, high- and low density lipoprotein, blood glucose and C-reactive protein levels. In addition, for some phenotypes, such as triglycerides, high density lipoprotein, low density lipoprotein and blood insulin levels, novel loci were identified, some of which have shown suggestive, but not genome-wide significant association in other populations (Sabatti et al. 2008). It remains to be shown whether these findings are isolate-specific or if they can be replicated in other populations.

Challenges in genetic mapping in isolates

Although isolated populations provide excellent opportunities for identification of genetic risk factors for genetic disorders, there are some drawbacks using these special populations that need to be considered in study design. Given that isolated populations are often smaller than mixed population, it should be established that a large enough cohort of affected individuals can be assembled from the population given the frequency of the trait of the interest and the size of the population. One of the drawbacks often encountered when the initial region of interest has been

identified is the LD structure. Although extended LD helps identify risk loci, it can simultaneously make it harder to identify causative variants within a block of strong LD. Therefore, if susceptibility factors are shared across populations, it would be beneficial to include a more admixed population in the finemapping stage of the study where LD blocks are shorter and LD between markers is less strong. Replication of initial findings can prove to be challenging, especially if the isolate is small and the disorder is rare, as all available samples have already been used in the initial mapping effort. It should also be noted that less strict requirements of replication might be relevant if the risk variant is significantly enriched in the isolate, and only confers a small risk in the replication population. However, even if the effect of the variant is limited to the isolate, characterization of these risk variants can prove to be beneficial as it could lead to a greater understanding of the trait‟s underlying biology.

The assumption that rare, high impact variants are enriched in founder populations is not always true. A recent study of genetic risk factors with metabolic phenotypes in the Kosrae population revealed a similar spectrum of genetic risk variants compared to other more outbred populations. The Kosrae, a native population of the Federated States of Micronesia, display both decreased genetic diversity and extended LD resulting from founder effects and multiple bottlenecks (Bonnen et al. 2006). In common with many other populations worldwide, the prevalence of obesity and metabolic disorders is increasing in this population (Shmulewitz et al. 2001), and an effort was initiated to identify genes for these disorders in this population with the hypothesis that founder effects, bottlenecks and genetic drift have concentrated a small number of high-impact genetic risk variants in this population. Simulation studies showed that the analytic approaches used in the study provided ample power to identify alleles explaining ≥5% of the variation in the phenotype. Despite this, no high-impact genetic risk factors were identified in this population. The results suggest, that depending on the genetic architecture of the trait of interest, low-impact common variants shared across populations confer susceptibility for common disorders even in isolated populations, where a founder effect could be expected (Lowe et al. 2009).

2.3 Autism Spectrum Disorders

Autism Spectrum Disorders (ASDs), also referred to as pervasive developmental disorders, are a group of childhood-onset neuropsychiatric disorders with shared core deficits but of varying severity. The most common diagnostic categories are autism (F84.0, OMIM 209850) and Asperger Syndrome (AS, F84.5, OMIM608638). Other disorders include disintegrative disorder (F84.3), atypical forms of autism (F84.1), and Rett syndrome (RTT, F84.2). The upcoming new editions of both the International Classification of Diseases (ICD) and Diagnostic and Statistical Manual of Mental Disorders (DSM) will likely contribute substantially modified diagnostic guidelines for ASDs. Suggested changes include treating this group of disorders as a continuum of phenotypes. Another proposed change is to exclude Rett syndrome, a monogenic form of ASDs, affecting predominantly girls and caused by mutations in a single gene (Amir et al. 1999).

2.3.1 Autism

Autism was originally described by American psychiatrist Leo Kanner in 1943. His work is the basis for the modern definition and diagnostic criteria (Kanner 1943). He described 11 children, mostly boys, whose condition differed from mental retardation on the basis of their social isolation. He named the syndrome „infantile autism", because the lack of social interest resembled that described by Swiss psychiatrist Eugene Bleuler and termed „autistic psychopathy".

Diagnostic criteria for autism are outlined in ICD-10 and DSM-IV (World Health Organization 1993; American Psychiatric Association 1994). Abnormalities in a triad of symptoms, consisting of qualitative abnormalities in reciprocal social interaction and communication as well as restricted repetitive and stereotyped patterns of behavior, interests and activities are required for diagnosis (Table 1). Symptoms have to be present before the age of thee years, making autism the most severe of all ASDs. In addition, a diagnosis of other ASDs such as specific developmental disorder of receptive language, mental retardation (MR) or Rett syndrome has to be excluded. MR is present in the majority of individuals with autism. Traditionally, 75-80% of individuals with autism have been reported to have cognitive impairment, but somewhat lower rates have been reported in more recent studies (Chakrabarti and Fombonne 2001; Yeargin-Allsopp et al. 2003).

Table 1. Diagnostic criteria for autism, as outlined in ICD-10 (World Health Organization 1993).

A. Presence of abnormal or impaired development before the age of three years, in at least one out of the following areas:
   (1) receptive or expressive language as used in social communication;
   (2) the development of selective social attachments or of reciprocal social interaction;
   (3) functional or symbolic play.

B. Qualitative abnormalities in reciprocal social interaction, manifest in at least one of the following areas:
   (1) failure adequately to use eye-to-eye gaze, facial expression, body posture and gesture to regulate social interaction;
   (2) failure to develop (in a manner appropriate to mental age, and despite ample opportunities) peer relationships that involve mutual sharing of interests, activities and emotions;
   (3) a lack of socio-emotional reciprocity as shown by an impaired or deviant response to other people's emotions; or lack of modulation of behaviour according to social context, or a weak integration of social, emotional and communicative behaviours.

C. Qualitative abnormalities in communication, manifest in at least two of the following areas:
   (1) a delay in, or total lack of development of spoken language that is not accompanied by an attempt to compensate through the use of gesture or mime as alternative modes of communication (often preceded by a lack of communicative babbling);
   (2) relative failure to initiate or sustain conversational interchange (at whatever level of language skills are present) in which there is reciprocal to and from responsiveness to the communications of the other person;
   (3) stereotyped and repetitive use of language or idiosyncratic use of words or phrases;
   (4) abnormalities in pitch, stress, rate, rhythm and intonation of speech.

D. Restricted repetitive, and stereotyped patterns of behaviour, interests and activities, manifest in at least two of the following areas:
   (1) an encompassing preoccupation with one or more stereotyped and restricted patterns of interest that are abnormal in content or focus; or one or more interests that are abnormal in their intensity and circumscribed nature although not abnormal in their content or focus;
   (2) apparently compulsive adherence to specific, non-functional, routines or rituals;
   (3) stereotyped and repetitive motor mannerisms that involve either hand or finger flapping or twisting, or complex whole body movements;
   (4) preoccupations with part-objects or non-functional elements of play materials (such as their odour, the feel of their surface, or the noise or vibration that they generate);
   (5) distress over changes in small non-functional, details of environment.

E. The clinical picture is not attributable to other varieties of pervasive developmental disorder; specific developmental disorder of receptive language (F80.2) with secondary socio-emotional problems; reactive attachment disorder (F94.1) or disinhibited attachment disorder (F94.2); mental retardation (F70-72) with some associated emotional or behavioural disorder; schizophrenia (F20) of unusually early onset; and Rett's syndrome (F84.2).

2.3.2 Asperger syndrome

In 1944, unaware of Kanner‟s publication, Hans Asperger, a Viennese pediatrician, described what he believed to be a new psychiatric disorder. Asperger described four boys with normal cognitive and verbal skills, but with difficulties in social interactions, unusual and intense interests and motor difficulties (Asperger 1944).

Asperger‟s work was not widely known in the English-speaking world until it was translated into English by Lorna Wing in 1981, who also introduced the term „Asperger syndrome“ (Wing 1981). AS was first granted official recognition in the tenth International Classification of Disease (ICD-10, World Health Organization 1993) and is included as „Asperger‟s disorder“ in DSM-IV (American Psychiatric Association 1994).

Similar to autism, qualitative abnormalities in social interaction and intense circumscribed interests or restricted, repetitive and stereotyped patterns of behavior interests and activities are required for a diagnosis of AS (Table 2).The most significant differences with autism are the lack of delay in spoken or receptive language and normal cognitive development in individuals with AS. However, abnormal use of language, such as formalistic speech, unusual use of inflection patterns and poor modulation of volume are often present (Volkmar and Klin 2000). Motor clumsiness is usual but not required for a diagnosis.

Table 2. Diagnostic criteria for AS, as outlined in ICD-10 (World Health Organization 1993).

A.	A lack of any clinically significant general delay in spoken or receptive language or cognitive development. Diagnosis requires that single words should have developed by two years of age or earlier and that communicative phrases be used by three years of age or earlier. Self-help skills, adaptive behaviour and curiosity about the environment during the first three years should be at a level consistent with normal intellectual development. However, motor milestones may be somewhat delayed and motor clumsiness is usual (although not a necessary diagnostic feature). Isolated special skills, often related to abnormal preoccupations, are common, but are not required for diagnosis.
B.	Qualitative abnormalities in reciprocal social interaction (criteria as for autism).
C.	An unusually intense circumscribed interest or restricted, repetitive, and stereotyped patterns of behaviour, interests and activities (criteria as for autism; however it would be less usual for these to include either motor mannerisms or preoccupations with part-objects or non-functional elements of play materials).
D.	The disorder is not attributable to the other varieties of pervasive developmental disorder; schizotypal disorder (F21); simple schizophrenia (F20.6); reactive and disinhibited attachment disorder of childhood (F94.1 and .2); obsessional personality disorder (F60.5); obsessive-compulsive disorder (F42).

2.3.3 Prevalence of ASDs

The most recent comprehensive meta-analysis of ASD prevalence reports a consensus prevalence estimate of 60-70/10 000 for all ASDs combined (Fombonne 2009). Corresponding figures for autism and AS in the same study are 20 and 6, respectively. However, prevalence studies have yielded highly discordant findings due to methodological differences and differences in diagnostic criteria. Males are affected more often than females with approximately 4 affected males per female, although the ratio decreases when individuals with severe cognitive impairment are included (Fombonne et al. 1997).

A multitude of studies have reported that the prevalence of ASDs is rising, and systematically lower prevalence estimates have been obtained in older studies compared to more recent ones (Newschaffer et al. 2005; Wazana et al. 2007). The increase has mainly been attributed to improved case ascertainment and changes in diagnostic criteria, not to increased incidence. Diagnostic substitution from cognitive impairment and learning disability to ASDs have also been suggested to partially explain the increase in prevalence (Shattuck 2006; Coo et al. 2008). The increase in high-functioning individuals receiving a diagnosis of ASDs has also been associated with the increase in ASD prevalence. In addition, the effect of diagnosis on the quality of care services, especially in the US has been suggested to increase the number of autism diagnoses for high-functioning individuals (Eagle 2004).

There are only two prevalence studies focused specifically on the prevalence of AS (Ehlers and Gillberg 1993; Mattila et al. 2007). In the first study only three affected children were identified, resulting in a prevalence estimate of 28.5/10 000 when ICD-10 diagnostic criteria were employed. The second study included all children in the two most northern provinces of Finland, and a total of 21 individuals affected with AS were identified after comprehensive clinical evaluation. These results indicate a prevalence estimate of 29/10 000, very close to the result obtained in the Swedish study (Ehlers and Gillberg 1993). Prevalence estimates obtained in other studies are usually significantly lower, suggesting that AS would be less common than autism. Whether this reflects ascertainment bias or different diagnostic practices employed in different populations remains unclear. It should also be noted that the two AS studies are both relatively small, and a correlation between smaller study size and higher prevalence estimates has been reported (Fombonne 2005). Furthermore, the age of the group studied is vital when the prevalence of AS is estimated, because AS is identified and diagnosed much later than typical autism and it is possible that surveys of young children underestimate the prevalence of AS (Fombonne 2001).

2.3.4 Mode of inheritance of ASDs

Both Kanner and Asperger reported that parents of individuals with ASDs displayed personality traits similar to those observed in their children, even leading Asperger to propose a genetic basis of the disorder he described. However, a bias in American psychiatry at that time for explaining all psychiatric disorders by deficiencies in parenting led to the hypothesis of autism being caused by emotionally cold mothers. Later studies showed no differences in parenting skills between parents with children affected with autism and healthy children disproving this hypothesis (Cantwell et al. 1979). Discovering the connection between autism and mental retardation as well as the aggregation of affected individuals in families finally established a biological basis of the disorder (Rutter 1968; Lockyer and Rutter 1969).

Family studies have reported a 2-6% rate of ASDs among siblings of individuals with autism (reviewed in Bailey et al. 1998a). This risk is even higher for siblings of female affected individuals, with rates of affected siblings up to 14% having been reported (Ritvo et al. 1989). Even using the most conservative frequency estimates, a clear increase in the rate of ASDs can be observed in families of affected probands compared to the general population.

Twin studies in autism have reported MZ concordance rates ranging from 36% to 91% compared to 0% in DZ twins (Folstein and Rutter 1977; Ritvo et al. 1985; Steffenburg et al. 1989; Bailey et al. 1995). The zero concordance rate for DZ twins can probably be explained by the small number of twin pairs. In all three studies a total of 36 MZ twin pairs and 30 DZ twin pairs were included. If a concordance rate is considered for a broader phenotype including milder cognitive or social deficits, 10-30% of DZ twin pairs are concordant in the same three studies. The heritability of autism has been estimated to be greater than 90% (Bailey et al. 1995; Szatmari et al. 1998). However, as MZ concordance is not 100%, this suggests that environmental factors also contribute to disease susceptibility.

No systematic family or twin studies have been performed in AS. Burgoine and Wing (1983) reported a pair of monozygotic triplets with a diagnosis of AS but who also displayed some characteristics of infantile autism, especially in one of the brothers. Single reports of AS or AS-like features in close relatives of probands with AS have been published supporting familial aggregation of AS (Kerbeshian and Burd 1986; Gillberg 1994; Kracke 1994). A study by Gillberg (1989) found more familial aggregation among families with probands diagnosed with AS compared to families with high-functioning autism (HFA), possibly suggesting a stronger genetic component of AS or a different inheritance pattern in HFA compared to AS. The families in the first genome-wide scan for AS susceptibility loci show strong support

of familial aggregation, and some pedigrees resemble autosomal dominant inheritance (Ylisaukko-oja et al. 2004).

Estimates of the number of genes conferring susceptibility to ASDs are variable: 2-10 (Pickles et al. 1995), 15 (Risch et al. 1999), and as much as 100 (Pritchard 2001; Veenstra-Vanderweele et al. 2003). Further studies are needed to elucidate whether the triad of impairments in ASDs are modified by separate genetic factors or controlled by overlapping risk factors. The correlation between the different domains of dysfunction is low, indicating that individuals with severe dysfunction in one of the domains do not necessarily display severe symptoms in the other domains (Ronald et al. 2006b; Ronald et al. 2006a). Similarly, distinct endophenotypes, often reflecting one of the three primary domains of symptoms, have resulted in non-overlapping linkage regions (Alarcon et al. 2002; Nurmi et al. 2003; Buxbaum et al. 2004; Chen et al. 2006; Schellenberg et al. 2006). The broader phenotype observed in close relatives of probands with ASDs fits best with a model of several interacting risk variants for the distinct domains of dysfunction (Jorde et al. 1991), and quantitative traits reflecting the three domains of dysfunction have resulted in different heritability estimates (Sung et al. 2005). However, other studies have reported a single continuously distributed factor contributing to all primary symptoms of ASDs (Constantino et al. 2004).

A related unsolved question is whether risk for ASD is conferred by multiple common variants or rare, high penetrance mutations. A statistical analysis of autism risk in multiplex families yielded a best fit with a model where ASDs are caused by dominantly acting *de novo* mutations with reduced penetrance in females (Zhao et al. 2007). This is consistent with evidence obtained from linkage studies, chromosomal abnormalities and the limited number of identified risk variants discussed in detail below (see sections 2.3.6-2.3.8). Some studies have also suggested that different genetic risk factors exist in simplex versus multiplex families (Miles et al. 2005; Campbell et al. 2007; Sebat et al. 2007). Others have reported that the broader phenotype is distributed differently in simplex versus multiplex families, with family members of multiplex families displaying a higher frequency of ASD-like traits (Szatmari et al. 2000; Losh et al. 2008; Virkud et al. 2009). As outlined in section 2.1.3, knowledge concerning the genetic architecture of the trait is vital for appropriate selection of gene mapping strategies, and can thereby greatly aid in the identification of genetic risk variants.

2.3.5 Known genetic etiologies of ASDs

In 10-15% of cases, autism is occurs together with conditions of known medical etiologies (Folstein and Rosen-Sheidley 2001; Fombonne 2003). These include

monogenic and complex disorders as well as chromosomal abnormalities, which are discussed in more detail in section 2.3.7. ASDs or ASD-like symptoms are present in these disorders at a higher frequency than expected by population prevalence. Often the symptoms are indistinguishable from idiopathic autism but for some genetic etiologies other distinguishing phenotypic features are also present, which allow for more specific diagnoses and provide possible clues to biological processes of interest.

The most common medical condition co-occurring with ASDs is cognitive impairment, which is present in 75-80% of individuals with ASDs. Some genes have already been identified linking these two conditions, such as mutations in *NLNG 4* (see section 2.3.11). Multiple genes involved in the etiology of cognitive impairment have been identified and they make up an interesting group of possible candidate genes for ASDs.

Epilepsy is present in approximately a third of individuals with ASDs by adulthood (Tuchman and Rapin 2002). The epilepsy can manifest in clinical seizures, or subclinically in epileptiform electroencephalography patterns. Individuals with infantile spasms are particularly likely to develop autism with nondevelopment of language and cognitive impairment (Asano et al. 2001). Stratification based on presence or type of epilepsy could be used as an endophenotype providing in genetically more homogenous groups for gene mapping.

Besides MR and epilepsy, the rest of associated conditions are relatively rare, and present only in a marginal subset of individuals with ASDs. However, ASDs or ASD-like symptoms are present in a significant fraction of individuals affected with these disorders. One of the most common of these is Fragile-X syndrome (FXS). FXS is the most common genetic cause for mental retardation in males, and is caused by mutations in *fragile-X mental retardation 1* (*FMR1*), located on the X chromosome. Approximately 20-30% of individuals with FXS show autistic features and the frequency is higher in males compared to females (Rogers et al. 2001; Hatton et al. 2006). Great variability has been reported in the rate of FXS in individuals diagnosed with autism ranging from 1 to 8% (Muhle et al. 2004).

RTT is currently included in the same diagnostic category with ASDs in ICD-10 and DSM-IV. RTT is the most common inherited cause for MR in girls. The role of RTT as a part of the ASD continuum is currently being debated, as uncertainty exists concerning the true etiological connection between RTT and ASDs. RTT is caused predominantly by *de novo* mutations in *MeCP2*, which silences transcription of methylated genes (Amir et al. 1999). Mutations in *MeCP2* have been linked to disrupted brain development and abnormal dendritic spine structure, closely

resembling anomalies reported in MR syndromes (Kaufmann and Moser 2000; Matarazzo et al. 2004).

An increased rate of ASDs has been reported in a variety of monogenic disorders, including Joubert syndrome, Timothy syndrome, Tuberous Sclerosis Complex, Smith-Lemli-Optiz syndrome and Potocki-Lupski syndrome. The prevalence these disorders in ASDs is hard to determine due to the low frequency of these syndromes, but has been estimated to be less than 1% (Abrahams and Geschwind 2008). As some of these conditions are multi-organ disorders, they serve as an important reminder that if candidate genes are purely identified based on tissue-restricted gene expression patterns some interesting candidates could be missed.

2.3.6 Linkage studies

Multiple genome-wide linkage scans have been performed to identify chromosomal regions linked to ASDs (Yang and Gill 2007). Linkage signals have been obtained on almost every chromosome (Figure 1), but only a few loci are replicated in more than one study, as with many other complex disorders (Altmuller et al. 2001). Differing diagnostic criteria, sample size and statistical analyses used in these studies make comparison of linkage signals difficult.

Significant linkage between autism and the long arm of chromosome 7 has been replicated in several studies. This region was also identified as the most significantly linked locus to ASDs in a meta-analysis (Trikalinos et al. 2006). Several chromosomal rearrangements have been reported in the same region (Folstein and Rosen-Sheidley 2001). Other relatively well replicated linkage regions are on 2q, 16p and 17q.

Most linkage studies have included 50-150 families, providing only limited power to detect linkage regions. However, the Autism Genome Project (AGP) has performed a large linkage study including 1168 families with at least two affected sibs. Surprisingly, this study revealed no genome-wide significant linkage loci suggesting that the lack of replicated linkage loci is not a consequence of a lack of power but instead a consequence of underlying genetic heterogeneity (Szatmari et al. 2007). When all families were combined, 11p12-13 reached suggestive but not genome-wide significance. This locus has not been implicated in previous linkage scans. Stratification of the sample into male-only and female containing families yielded more significant peaks in the female containing families compared to male specific families. This could suggest a stronger genetic component in families with female affected individuals, or reduced genetic heterogeneity as the number of female containing families was 440 compared to 741 male-only families.

The failure of large-scale genome-wide studies to identify common variants conferring risk for ASDs has prompted the investigation of linkage within individual families to identify rare variants. An elegant example of this approach was a study where homozygosity mapping in consanguineous families was used to identify family specific genetic risk variants for autism (Morrow et al. 2008). In this study 104 consanguineous families, 88 of which had cousin marriages, were genotyped using a dense, genome-wide SNP array as well as an aCGH platform. The SNP data was analyzed for homozygous regions co-segregating with ASDs. Several families showed significant but non-overlapping linkage signals with LOD scores ranging from 2.4 to 2.96, comparable to the highest LOD scores obtained in linkage studies where data from hundreds of unrelated families are pooled. Interestingly, 5 of these families showed rare large inherited homozygous deletions ranging from 18 to over 880 kb within the linkage peaks, depleting whole or parts of genes highly expressed in the brain. Genes within or close to the two largest deletions have been shown to be regulated by neuronal activity or are the targets of transcription factors induced by activity.

If the lack of replicated linkage signals is due to genetic heterogeneity, increasing sample size is not necessarily the best way to increase power in linkage studies. Stratification has been performed using clinical phenotypes such as families with AS only, gender of proband, the presence or absence of spoken language and neurobehavioral features such as regression or obsessive-compulsive behaviors. Stratification based on gender has provided increased evidence for linkage (Stone et al. 2004; Lamb et al. 2005; Szatmari et al. 2007). Replicated linkage to 2q was obtained when a subset of families with delayed onset of phrase speech were analyzed (Buxbaum et al. 2001; Shao et al. 2002).

The use of quantitative traits in linkage analysis have been successful in other complex disorders (Kissebah et al. 2000; Fisher et al. 2002; Weiss et al. 2006a). As family members often display some mild ASD-like traits, the use of quantitative traits such as age of first word provide added power. Analysis of age at first word resulted in the identification of a linkage peak on 7q and association with common variants in *CNTNAP2* in a follow-up study (Alarcon et al. 2002; Alarcon et al. 2008).

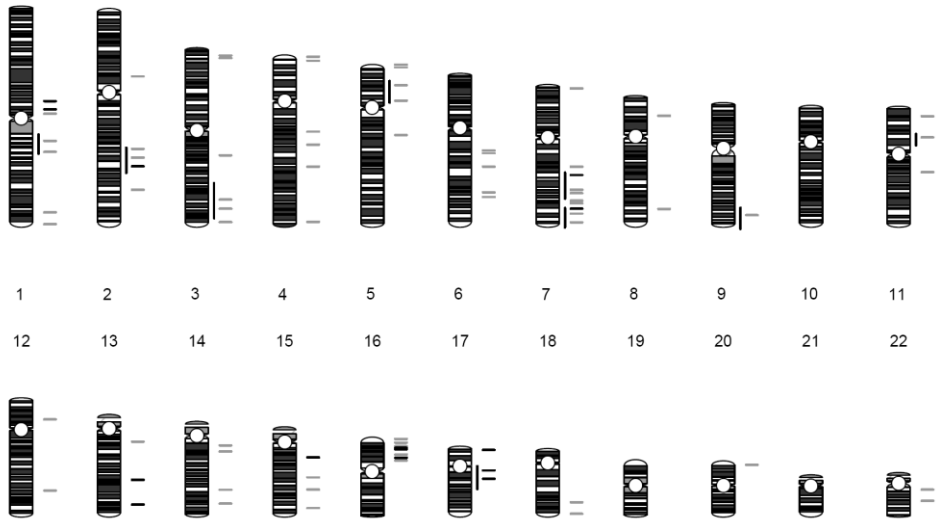


Figure 1. Linkage peaks identified on autosomes in genome-wide scans. *Vertical bars indicate loci where LOD>3 was obtained in at least one study, and LOD>2 in another. Data from Abrahams and Geschwind (2008). Horizontal bars indicate loci where significant or suggestive linkage has been reported in at least one study. Bars in black indicate genome-wide significant linkage (LOD>3.3), suggestive linkage in at least two studies (1<LOD<3.3) or loci where at least one study has resulted in genome-wide significance and another in suggestive linkage. Grey bars indicate markers resulting in suggestive linkage in one study. Data from Yang and Gill (2007).*

2.3.7 Genome-wide association studies in ASDs

Recently, the first GWAS in ASDs was published (Wang et al. 2009). Two datasets were used - a cohort of 780 families from the Autism Genetic Resource Exchange (AGRE) and a case control cohort of 1204 affected individuals and 6491 controls from the US. Although neither cohort alone resulted in any genome-wide significant results, when the two cohorts were analyzed together, a genome-wide significant signal was obtained on 5p14.1, between *cadherin 10* (*CDH10*) and *CDH9*. The results were replicated in two independent datasets, resulting in combined p-values ranging from 10^{-8} to 10^{-10} All SNPs showing significant association to ASDs are located within the same LD block, suggesting they are tagging the same susceptibility variant. Cadherins are neuronal adhesion proteins expressed in the developing brain. *CDH10* was shown to be highly expressed in the frontal cortex and has a similar expression pattern to *CNTNAP2*, which has been recently implicated in ASDs (see section 2.3.9). These results suggest that common variants do play a role in the etiology of ASDs. Further autism GWA studies will be

published in the near future. A GWA study comprising a more extensive set of over 1000 AGRE families has resulted in association to SNPs on 5p15, between *SEMA5A* and *TAS2R1*. The association was replicated in other datasets as well, and the combined p-value for the best marker in the region was $6x10^{-9}$ (Weiss et al. 2009).

2.3.8 Chromosomal aberrations and CNVs

Microscopically visible chromosomal aberrations have been estimated to account for 6-7% of ASD cases. However the frequency of chromosomal aberrations varies widely, ranging from 0-54% (Xu et al. 2004). Higher frequencies have been obtained in studies of syndromic ASD cases compared to idiopathic cases (Jacquemont et al. 2006).

The most common cytogenetic abnormality in ASDs, accounting for approximately 1-2% of cases is maternal duplication of the imprinted 15q11-q13 region (Vorstman et al. 2006). This region overlaps with the deletion causing Angelman syndrome and Prader-Willi syndrome, both of which also have overlapping clinical symptoms with ASDs (Veltman et al. 2004; Bonati et al. 2007). The gene responsible for Angelman syndrome, *ubiquitin protein ligase E3A* (*UBE3A*), has been identified, whereas the specific gene or genes responsible for Prader-Willi syndrome have not been identified. A recent expression study reported that gene copy number does not directly correlate with gene expression in a female with autism, and that epigenetic mechanisms probably controls the expression of genes at the 15q11-13 locus (Hogart et al. 2009) Currently, the *gamma-aminobutyric acid A receptor B3* (*GABARB3*) and *UBE3A* are the most interesting candidate genes for ASDs in the 15q11-q13 region. The GABAergic system has also been linked with epilepsy (Baulac et al. 2001; Wallace et al. 2001). Candidate gene studies of the three GABA receptor genes as well as *UBE3A* have yielded both positive and negative findings in candidate gene studies (Cook et al. 1998; Maestrini et al. 1999; Martin et al. 2000; Buxbaum et al. 2002; Nurmi et al. 2003; McCauley et al. 2004; Kim et al. 2006; Tochigi et al. 2007). Furthermore, smaller CNVs in several genes involved in ubiquitination both at 15q11-q13 as well at other loci have recently been implicated in ASDs (Glessner et al. 2009).

Cytogenetically visible abnormalities have been reported on almost all chromosomes in ASDs. Meta-analysis have identified loci where several overlapping chromosomal aberrations cluster among unrelated cases, although no previous linkage or association results have been reported. These loci were suggested to be novel ASD loci. The loci implicated were 2q37, 5p15, 11q25, 16q22.3, 17p11.2, 18q21.1, 18q23, 22q11.2, 22q13.3 and Xp22.2–p22.3. In addition, regions were identified where deletions or duplications overlap with loci

identified in linkage studies, such as several loci on 7q, 5p14, 10p14 and 16p13.3 (Vorstman et al. 2006). The largest number of chromosomal aberrations was unsurprisingly reported at the 15q11-13 locus. A large number of chromosomal abnormalities were also located at 7q and 22q.

Deletions in 22q11 and 22q13 are both associated with specific syndromic phenotypes (Goizet et al. 2000; Shprintzen 2000). Deletion of 22q11.2 causes velocardiofacial syndrome, which is one of the most significant genetic risk factors for schizophrenia and has recently been linked to autism as well (Murphy et al. 1999; Niklasson et al. 2001). The common deletion at 22q13.3 is characterized by neonatal hypotonia, global developmental delay, absent to severely delayed speech, dysmorphic features and ASD-like behavioral symptoms (Manning et al. 2004). The minimal overlapping region identified in 22q13.3 microdeletion syndrome contains three genes: *ACR*, *RABL2B* and *SHANK3*. *SHANK3* has also been reported to be disrupted by a translocation in an individual with clinical symptoms of 22q13.3 deletion syndrome, implicating *SHANK3* as the causative gene (Bonaglia et al. 2001). However, deletions not overlapping with *SHANK3* indicate that other genes in the region influence language and cognitive development as well (Wilson et al. 2008).

Rare chromosomal abnormalities, identified in only a few individuals, have aided in the identification of genes involved in the etiology of ASDs. At Xp22.3, *de novo* chromosomal deletions were identified in three females with autism (Thomas et al. 1999). A frameshift mutation, resulting in premature termination of the encoded protein was identified in *neuroligin 4* (*NLGN4)*, located in the deleted region (Jamain et al. 2003). The deletion was present in two affected siblings, one with autism and one with AS but not in an unaffected brother or 350 healthy controls. The mutation was also present in the mother. Another mutation was identified in another member of the NLGN family, *NLGN3*, located at Xq13. The mutation resulted in replacement of a conserved Arginine to a Cysteine (R451C) in two affected brothers, one with autism and the other with AS. The mother was a carrier of the mutation, but no grandparents were available for the study. The role of skewed X chromosome inactivation in the mothers was not investigated.

The first systematic study of submicroscopic chromosomal rearrangements in idiopathic ASDs revealed an excess of de novo CNVs in individuals with autism (Sebat et al. 2007). The de novo CNVs were particularly enriched in families with only one affected individual. The overlap between identified CNVs was small, and most CNV were detected only in a single individual. The AGP linkage-study was the first autism study where genome-wide SNP data was assessed for variation in copy number of submicroscopic regions (Szatmari et al. 2007). Despite a sparse marker map, *de novo* CNVs were identified in cases in 10 families. However, the

same CNV was detected in both affected sibs in only three families. In an additional 18 families, CNVs were identified in regions with previously identified chromosomal abnormalities related to autism. The most interesting CNV was a deletion in two female affected sibs spanning 300 kb on 2p16, which was not detected in either parent. The hemizygous deletion removes coding exons of *NRXN1*, an interesting functional candidate interacting with NLGNs at the synapse. A previous study had already reported rare mutations in *NRXN1* in autism, making it a highly relevant candidate gene (Feng et al. 2006).

A deletion spanning almost 600 kb on 16p11.2 was identified in five cases using dense, genome-wide SNP data in the AGRE families (Weiss et al. 2008). In the same study this CNV was also identified in eight individuals from a cohort of individuals with ASDs, developmental delay, mental retardation, or suspected ASD. Furthermore, the reciprocal duplication was also identified as a risk factor for ASDs in the same study. The same CNV was identified in another study using aCGH in two AGRE families (Kumar et al. 2008). The CNV has been detected in unphenotyped population controls with a frequency of approximately 1%, suggesting incomplete penetrance (Weiss et al. 2008).

Large, genome-wide studies are currently producing massive amounts of data concerning the chromosomal rearrangements in ASDs as well as in controls, making the review of the available data painstaking. However, resources summarizing the extensive information obtained in genome-wide studies of large autism datasets are finally emerging, making the use of these large datasets easier. The Autism Chromosome Rearrangement database aims to catalogue all chromosomal abnormalities associated with ASDs, providing an easy user interface to evaluate the significance of novel findings (Xu et al. 2004; Marshall et al. 2008). The DECIPHER database contains information on large, rare CNVs identified in individuals with cognitive impairments. The database also contains detailed phenotype information, allowing for the identification of novel syndromes associated with these genomic rearrangements as cases with overlapping rearrangements are reported (Firth et al. 2009). The database has already facilitated the identification of several novel syndromes, including the 14q11.2, 17q21.3 and 19q13.11 deletion syndromes (Shaw-Smith et al. 2006; Zahir et al. 2007; Malan et al. 2009).

2.3.9 Candidate gene studies

An overwhelming multitude of candidate genes studies using both association- and resequencing methods have been performed in ASDs. Often studies lack power to detect association, and even for the most convincing candidates, both negative and

positive results have been published. This likely reflects both the lack of power of many studies, and the underlying genetic heterogeneity of the genetic risks of ASDs. Some interesting candidate genes are presented in Table 3 and classified according to the available evidence for their role in ASD etiology.

Although approximately half of the candidate genes listed in Table 3 are located at linkage peaks, only a few genes have been identified in follow-up studies of linkage regions. One example, where positional cloning has been successfully used to identify variants influencing risk for ASDs is the identification of *CNTNAP2* as a susceptibility gene for ASDs. Fine-mapping of the language-related quantitative trait locus (QTL)-linkage peak on 7q35 using over 2700 SNPs identified association between four genes in the region and age at first word in 170 families (Alarcon et al. 2002; Alarcon et al. 2005; Alarcon et al. 2008). In the follow-up stage, SNPs in these four genes were tested for association in an independent sample of 304 trios. The only SNP associated with age at first word in this stage was rs2710102 in *CNTNAP2,* and analysis of families containing only male affected individuals resulted in more significant association. In a subsequent study, the link between *CNTNAP2* and language disorders was reinforced as *CNTNAP2* was identified as a target of *FOXP2*, a gene interrupted in rare mendelian speech and language disorders (Vernes et al 2008). Mutations in *CNTNAP2* have been identified in a family with seizures, language regression and ASDs (Strauss et al. 2006). A chromosomal translocation disrupting *CNTNAP2* has also been identified in a family with repetitive behaviors and intellectual disability (Verkerk et al. 2003). Chromosomal rearrangements, rare mutations and identification from a genome-wide linkage study for autism further support the role of *CNTNAP2* in the etiology of ASDs (Arking et al. 2008; Bakkaloglu et al. 2008). Interestingly, *CNTNAP2* is a member of the NRXN gene family implicated in ASDs in earlier studies (see section 2.3.11).

Table 3. Candidate genes with multiple lines of evidence for their role in ASDs

Gene	Pos	Syndr. or mutat.	Repl. assoc.	Analysis of variant	Mouse model	Linkage	CNV in ACRD	Other evidence	Total score	Reference
AVPR1A	12q14	0	0	0	1	0	0	0	1	Wassink et al. (2004)
ITGB3	17q21	0	1	0	0	0	0	0	1	Weiss et al (2006b)
DISC1	1q42	0	0	0	1	1	0	0	2	Kilpinen et al. (2008)
EN2	7q36	0	1	0	1	0	0	0	2	Gharani et al. (2004)
GRIK2	6q16	0	1	0	0	0	0	1; mutation in MR	2	Jamain et al. (2002)
AHI1	6q23	2	0	0	0	1	0	0	3	Ozonoff et al. (1999)
CACNA1C	12p13	2	0	1	0	0	0	0	3	Splawski et al. (2004)
NRXN1	2p16	2	0	0	0	0	1	0	3	Feng et al. (2006)
SLC25A12	2q31	0	1	0	0	1	0	1; expression upregulated in ASD brain	3	Ramoz et al. (2004)
CNTNAP2	7q35	2	1	0	0	1	0	0	4	Strauss et al. (2006)
FMR1	Xq27	2	0	1	1	0	0	0	4	Bailey et al. (1998b)
NLGN3	Xq13	2	0	1	1	0	0	0	4	Jamain et al. (2003)
OXTR	3p25	0	1	0	1	1	0	1; expression reduced in ASD blood	4	Wu et al. (2005)
SLC6A4	17q11	0	1	1	0	1	0	1; clinical benefit from inhibitors	4	Cook et al. (1997)
SHANK3	22q13	2	0	0	0	0	1	1; modulates dendrite spine morphology	4	Durand et al. (2007)
DHCR7	11q13	2	0	1	0	1	0	1; hypocholesterolameia in ASD probands	5	Tierney et al. (2001)
MECP2	Xq28	2	0	1	1	0	0	1; controls expression of UBE3A and GABRB3	5	Amir et al. (1999)

Gene	Pos	Syndr. or mutat.	Repl. assoc.	Analysis of variant	Mouse model	Linkage	CNV in ACRD	Other evidence	Total score	Ref
MET	7q31	0	1	1	0	1	1	1; expression reduced in ASD brains	5	Campbell et al. (2006)
PTEN	10q23	2	0	0	1	1	0	1; mutations result in synaptic abnormalities	5	Butler et al. (2005)
TSC2	16p13	2	0	1	0	1	0	1; regulates dendrite morphology	5	Baker et al. (1998)
CADPS2	7q31	2	0	1	1	1	1	0	6	Sadakata et al. (2007)
GABRB3	15q12	2	1	0	1	0	1	1; expression is dysregulated in ASDs	6	Cook et al. (1998)
NLGN4X	Xp22	2	0	1	1	1	1	0	6	Jamain et al. (2003)
UBE3A	15q11	2	0	1	1	0	1	1; expression is dysregulated in ASDs	6	Peters et al. (2004)
RELN	7q22	2	1	1	1	1	0	1; expression reduced in ASD brains	7	Persico et al. (2001)

NOTE: To emphasize the role of rare mutations in ASDs, genes associated with ASD-linked syndrome or mutation resulted in 2 points, whereas other lines of evidence resulted in one point. To qualify as a mouse model of ASDs, two out of the three clinical core features had to be present. Linkage evidence was scored as present if the best marker from a suggestive linkage peak listed in Yang and Gill (2007) was located within the same cytogenetic band. ACRD – Autism chromosomal rearrangement database, *AH1* – Abelson helper integration site 1, *AVPR1A – arginine vasopressin receptor 1A, CACNA1C – calcium channel voltage-dependent L type alpha 1C subunit, CADPS2 – Ca^{2+}-dependent activator protein for secretion 2, CNTNAP2 – contactin associated protein-like 2, DHCR7 – 7-dehydrocholesterol reductase, DISC1 – disrupted in schizophrenia 1, EN2 – engrailed homeobox 2, FMR1 – fragile mental retardation 1, GABRB3 – gamma-aminobutyric acid A receptor beta 3, GRIK2 – glutamate receptor ionotropic kainite 2 precursor, ITGB3 – integrin beta 3, MECP2 – methyl CpG binding protein 2, MET – met proto-oncogene, NLGN3 – neuroligin 3, NLGN4X – neuroligin 4 X-linked, NRXN1 – neurexin1, OXTR – oxytocin receptor, PTEN – phophatase and tensin homologue, RELN – reelin, SHANK3 – SH3 and multiple ankyrin repeat domains 3, SLC25A12 – solute carrier family 25 member 12, SLC6A4 – solute carrier family 6 member 4, TSC1 – tuberous sclerosis 1, TSC2 – tuberous sclerosis 2, UBE3A – ubiquitin protein ligase E3A*. Modified from Abrahams and Geschwind (2008).

2.3.10 Expression studies in ASDs

To date, only five studies profiling genome-wide gene expression in ASDs have been published (Table 4). Four studies used lymphoblastoid cell lines (LCLs) as models for gene expression in the brain and one study used postmortem brain samples. The value of LCLs as model systems for gene expression remains unclear. Proof-of-principle studies have been performed in animals where gene expression profiles from blood cells were able to differentiate between different neurological diseases (Tang et al. 2001). In a study of bipolar disorder, a subset of differentially expressed genes in postmortem brain tissue were also found to be differentially expressed in LCLs, although the direction of expression changes were not consistent between the two tissues (Iwamoto et al. 2004). In ASDs, LCLs have been reported to differentiate between affected individuals and their healthy siblings (Baron et al. 2006), and individuals having autism with regression from individuals with early onset autism without regression (Gregg et al. 2008). They have also been able to separate individuals with the *FMR1* mutation from individuals with duplication of 15q (Nishimura et al. 2007). Taken together, these results suggest that at least some of the gene expression changes characterizing neurological and neuropsychiatric conditions are reflected in LCLs. Another argument for the use of LCLs is that expression studies could result in a combination of biomarkers which could be used as diagnostic assays in the future. Possible uses of gene expression profiling from blood could be used to confirm diagnosis or to classify affected individuals. The distinction into diagnostic subclasses could have impact on choice of treatment.

Although the use of post-mortem brain samples would seem like a more optimal tissue for expression studies, it also poses problem. Access to ASD brain samples is limited, and as RNA is seldom extracted immediately upon organ retrieval, the quality of RNA is rarely optimal for expression studies. More importantly however, death often occurs long after disease onset and the observed gene expression changes do not necessarily reflect the processes involved at that time.

Of the approximately 900 genes identified as differentially expressed on a suggestive level in these studies, less than 1% have been identified by more than one study. However, for the small group of genes identified in more than one study, there are several genes residing on 15q, including *CYFIP1*, *NIPA2* and *UBE3A*. The group of differentially expressed genes identifies significant enrichment in categories including ubiquitin conjugation, SH3 domain containing proteins, GTPase regulator activity, α-protocadherin genes and alternative splicing (Abrahams and Geschwind 2008). Larger gene-expression studies are needed to verify the role of these genes and biological processes in ASDs.

Table 4. Genome-wide gene expression profiling studies in ASDs.

Tissue	Cases	Ctrls	Platform	Diff. expr. probes	Stat.	Identified pathways	Reference
Brain	9	4	cDNA Atlas Human Neurobiology and UniGEM V2	NS	NS	Glutamate neurotransmission	Purcell et al. (2001)
LCL	10[1]	4	cDNA TIGR 40K Human set	26 up 15 down	|FC|> 1.5	Nervous system development and function, cell death, immune and lymphatic system development and function	Hu et al. (2006)
LCL	3	3[2]	Affymetrix HG-U95Av2	30 up 18 down	p<0.05 & |FC|>1.3	Immune system, signal transduction, cell growth, response to external stimuli, development	Baron et al. (2006)
LCL	15[3]	15	Agilent Whole Human Genome Array G41124A	43 up 25 down	FDR<5% in SAM, RankProd and ANOVA	Cell communication, signal transduction, immune response, defense response	Nishimura et al. (2007)
LCL	35	12	Affymetrix HG-U133Plus 2.0	45 up 10 down	FDR<5%, |FC|>1.5	Keratan sulfate biosynthesis, folate biosynthesis, galactose metabolism, serotonin receptor signaling	Gregg et al. (2008)

NOTE: Ctrls – Controls, Diff. expr. Probes – Number of differentially expressed probes, Stat. – Statistical significance used for differential expression, LCL – Lymphoblastoid cell line, NS – Not stated, FC – Fold change, FDR – False discovery rate, SAM – Significant Analysis of Microarray Analysis (Tusher et al. 2001), RankProd – Rank Product analysis (Breitling et al. 2004b), ANOVA – Analysis of variance, [1] – Five MZ twin pairs, two which were concordant for autism, and two where the co-twin had a Broad Spectrum ASD, [2] – Sibs of the affected individuals, [3] – 8 males with FMR1 mutation and 7 with dup(15q).

2.3.11 Biological pathways identified in ASDs

The numerous findings from linkage, association, expression and chromosomal rearrangement studies in ASDs are almost overwhelming. However, a subset of the findings can be linked to common molecular pathways, revealing biological processes implicated in the etiology of these diseases. Although findings for individual genes are not convincing by themselves, the disruption or dysregulation of several genes in the same biological pathway provide convincing evidence that the process is involved in the etiology of the trait.

Serotonin and autism

One of the most common biological finding in ASDs is elevated blood serotonin in 20-30% of individuals with ASDs as well as in a subset of healthy relatives (Schain and Freedman 1961; Anderson et al. 1987; Abramson et al. 1989; Leboyer et al. 1999). The observed increase in cases compared to controls has ranged from 20% to over 300% (Anderson et al. 1990). However, the prevalence of hyperserotonemia in autism may have been overestimated in these studies due to failure to control for race and pubertal status (McBride et al. 1998). Blood levels of serotonin correlate inversely with verbal IQ and positively with severity of autism (Kuperman et al. 1987; Cook et al. 1990). Position emission tomography have indicated significant differences in serotonin synthesis capacity between children with autism and healthy controls (Chugani et al. 1999).

Pharmacologic studies have further implicated the serotonin system in ASDs. Symptoms of ASDs were ameliorated by the administration of selective serotonin reuptake inhibitors in individuals with autism (McDougle et al. 1996b). Moreover, depletion of tryptophan, a precursor in serotonin synthesis, resulted in the worsening of symptoms in a group of adult ASD patients (McDougle et al. 1996a). A review of studies of selective serotonin reuptake inhibitors in ASDs concluded that most studies demonstrated significant improvement in global functioning and an amelioration of symptoms, especially those linked to anxiety and repetitive behaviors. However, the review also reported methodological weaknesses in several of the clinical trials highlighting the need for additional randomized control trials (Kolevzon et al. 2006).

Serotonin has been shown to play a key role in a variety of behaviours and processes such as behavioural inhibition, appetite, aggression, sleep and mood. It also plays a vital role in neurodevelopment, having critical effects on neurogenesis, morphogenesis and synaptogenesis in the developing brain (Scott and Deneris 2005). Serotonergic neurons are widely distributed across the mammalian brain, and

display rich innervation of limbic areas critical for emotional expression and social behavior, reflecting the core defects in ASDs (Anderson 2002). The serotonin system has also been linked to epilepsy, and variants in the genes involved in the serotonin pathway could explain the co-occurrence of autism and epilepsy (Chugani 2004).

The biological evidence linking the serotonin pathway to autism has led to intense scrutiny of genes involved in serotonin signaling and metabolism. The most intensely investigated gene is the serotonin transporter *SLC6A4* also known as *5-HTT*, which is the site of action of selective serotonin reuptake inhibitors. *SLC6A4* modulates serotonergic neurotransmission by transporting serotonin from the synaptic cleft into presynaptic nerves (Amara and Pacholczyk 1991). Early reports of association between the short allele of a polymorphism in the promoter of *SLC6A4* with autism was followed by a number of replications as well as non-replications (Cook et al. 1997; Devlin et al. 2005). A recent meta-analysis of the studies of the promoter polymorphism as well as another microsatellite marker in intron two of *SLC6A4* concluded that there is no evidence for association between any of these markers and ASDs. However, when only US populations were analyzed, the short allele of the promoter polymorphism was overtransmitted to affected individuals (Huang and Santangelo 2008). A study reported magnesium concentration-dependent genotyping errors at the *SLC6A4* promoter polymorphism resulting in a false increased frequency of the short allele. Corrected genotypes resulted in no association of this locus with ASDs and evidence for deviation from Hardy-Weinberg equilibrium (HWE) in published studies was reported (Yonan et al. 2006). In light of this finding combined with the meta-analysis of *SLC6A4*, the role of this gene in the etiology of ASDs remains highly uncertain.

Tryptophan 2,3-dioxygenase (*TDO2*) is a rate-limiting enzyme in the catabolism of tryptophan. Association between a SNP in the promoter of *TDO2* has been reported in 196 AGRE families with ASDs (Nabi et al. 2004). Decreased activity of this enzyme could slow the metabolism of tryptophan and thereby lead to increased levels of serotonin. No other studies have been published assessing the role of *TDO2* in ASDs.

Another gene involved in the metabolism of serotonin is *monoamine oxidase-A* (*MAO-A*), which regulates the level of serotonin as well as dopamine in the brain. Maternal genotypes of a non-synonymous (ns)SNP in *MAO-A* have been reported to be correlated with IQ in children with autism (Jones et al. 2004). Haplotypes in *MAO-A* have also been associated with ASDs in the Korean population (Yoo et al. 2009). A promoter polymorphism linked to lower activity of *MAO-A* (Sabol et al. 1998; Denney et al. 1999) has been linked with lower IQ, more severe symptoms and larger brain volume in children with ASDs (Yirmiya et al. 2002; Cohen et al.

2003; Davis et al. 2008). The *MAO-A* genotype can be linked with neurobiological findings in autism as early cortical enlargement has been associated with ASDs (Courchesne et al. 2001; Dissanayake et al. 2006; Dawson et al. 2007), and the authors suggested that the effects of *MAO-A* variants on brain volume would be mediated through serotonin and dopamine (Davis et al. 2008). To date, no negative candidate gene studies of *MAO-A* in ASDs have been published.

The *b3-integrin* gene (*ITGB3*) has been identified as a QTL for whole blood serotonin levels in males (Weiss et al. 2004). Association was observed between a coding variant in *ITGB3* and autism, however, the allele associated with autism was associated with lower whole-blood serotonin in population samples, contradictory to reports of elevated serotonin in ASDs (Weiss et al. 2006b). The authors suggested that *ITGB3* is not responsible for hyperserotonemia, but rather perturbation of the serotonin system in general in individuals with ASDs.

Synaptic dysfunction

The role of the serotonin system in autism was originally suggested by the biological finding of elevated whole-blood serotonin, which led to the identification of several candidate genes for ASDs. The opposite route, starting from rare genetic mutations and CNVs merging to reveal a shared biological function, has pointed to synaptic dysfunction in the etiology of ASDs.

CNVs and mutations have implicated NLGNs and NRXNs in the etiology of autism (see section 2.3.7). Following the initial report of one mutation in *NLNG3* and one in *NLGN4X*, several mutations in *NLGN4* have been identified. Mutations have been identified in both *NLGN4X* and the Y chromosomal homolog *NGLN4Y*. *NLGN4* mutations have been linked with a wide range of phenotypes, ranging from autism and Asperger syndrome to mental retardation with or without autism (Laumonnier et al. 2004), and to Tourette syndrome and attention-deficit hyperactivity disorder in another study (Lawson-Yuen et al. 2008). In some families, the carrier mother also had learning disabilities (Yan et al. 2005; Lawson-Yuen et al. 2008). Interestingly, deletion of *NLGN4X* has also been linked with normal intellectual function and mild MR in one study, suggesting that deletion of *NLGN4X* is compatible with normal intellectual function (Macarov et al. 2007).

Altogether these studies along with a number of negative screening studies bring the total number of individuals with ASDs screened for mutations in the coding regions of *NLGN4X* to over 1000. The resulting frequency of mutations in NLGNs is below 0.6%, which is probably an overestimate as not all variants identified in these studies are necessarily causative. Interestingly, although the original report identified a mutation in both *NLGN3* and *NLGN4X*, subsequent reports have only identified mutations in *NLGN4*. *NLGN1*, located at 3q26 has only been investigated

in one study (Ylisaukko-oja et al.). *NLGN2*, which is predominantly located at inhibitory (GABAergic) synapses has not been investigated in ASDs.

When CNVs and mutations were identified in *NRXN1*, a postsynaptic binding partner of NLGNs, great interest was sparked in the role of the synapse in ASDs. Several CNVs spanning the shorter isoform, *α-NRXN1* and the longer *β-NRXN1* isoform have been identified (Kim et al. 2008, Szatmari et al. 2007). Rare variants have also been identified in both isoforms (Kim et al 2008, Feng et al. 2006, Yan et al. 2008). Similarly to NLGN mutations, the *NRXN1* variants result in a wide variety of phenotypes, ranging from ASDs to schizophrenia (Rujescu et al. 2009). Furthermore, the variants do not display complete penetrance, as they have been observed in healthy siblings as well (Feng et al. 2006).

NLGNs and NRXNs are transmembrane proteins, which bind across the synaptic cleft. *NLGN1* is located exclusively at exitatory (glutamatergic) synapses, whereas *NLGN2* is located exclusively at inhibitory (GABAergic) synapses. *NLGN3* might be present in both kinds of synapses. NLGNs are located on the postsynaptic membrane of synapses, whereas NRXNs are located at the presynaptic terminals. NRXNs are encoded by three genes, but multiple promoters and alternative splicing result in thousands of different splice isoforms. The binding affinities between different NRXN isoforms controls binding to different NLGNs at different types of synapses (Südhof 2008).

In vitro experiments have shown that expression of NLGNs and NRNXs in non-neuronal cells induce formation of post- and presynaptic structures, respectively, in neurons. However, knockout-mice have shown that NLGNs are not required *in vivo* for synaptic assembly, but are vital for synaptic function. Mice lacking *NLGN1*, *NLGN2* and *NLGN3* die at birth, and show severe impairment of synaptic transmission despite normal number of synapses. Mice lacking only *NLGN1* or *NLGN2* are viable and fertile, but show significant dysfunction at excitatory and inhibitory synapses respectively (Südhof 2008).

NLGNs and NRXNs bind across the synaptic cleft, and interact with several PDZ-domain containing proteins. The junction formed by NLGNs and NRXNs resemble a tight junction, but is asymmetrical with respect to the binding partners on each side of the synapse. Interestingly, several proteins linked to the NLGN-NRXN complex have been implicated in ASDs or other relevant disorders. NLGNs bind *PSD95*, which recruit glutamate receptors at postsynaptic sites, linking the NLGN-NRXN complex to glutamatergic signaling, which has been previously implicated in ASDs in both expression and candidate gene studies (Sheng et al. 2007, Purcell et al. 2001, Jamain et al 2002). *PSD95* also binds *SHANK3*, which has been implicated in the etiology of ASDs by chromosomal rearrangements as well as point mutations

(Marshall et al 2008, Durand et al 2007, Moessner et al 2007). *PSD95*, also known as *DLG4*, has been shown to be regulated by *FMRP* (Muddashetty et al 2007), and belong to the membrane associated guanylate kinase (MAGUK) family. Another member of the MAGUK family, *DLG3*, has been shown to be mutated in MR (Tarpey 2004), which links several of the downstream binding partners of the NLGN-NRXN complex with phenotypes related to ASDs. A member of the NRXN family, *CNTNAP2*, has recently been implicated in ASDs by multiple lines of evidence (see section 2.3.9).

The interaction of NLGN-NRNX first implicated connectivity problems at synapses in the etiology of autism. The findings at the molecular level serve as a link with more general observations of possible disease mechanisms. The differential expression of NLGN proteins at excitatory and inhibitory synapses modulates exitatory and inhibitory synapse development (Levinson and El-Husseini 2005). This process is critical for brain function, because alteration in the excitatory/inhibitory ratio during brain development can lead to abnormal synaptic connectivity and function, thereby resulting in severe neurological impairments caused by a number of different genes (Rubinstein and Merzenich 2003). Neurobiological findings in ASDs have not identified any major anatomical changes, further supporting the notion that ASDs are associated with abnormal brain function and not structure. The observed synapse-related dysfunction in ASDs could link molecular processes at the synapse with observed disconnection of brain regions involved in higher-order processing (Geschwind and Levitt 2007).

3 AIMS OF THE STUDY

When we started this study, two genome-wide linkage scans had been performed in the Finnish ASD families, one in families with AS but no other ASDs, and another in families with autism, AS and dysphasia. The linkage scans revealed candidate regions for both AS and ASDs. In this study, our aim was to use the existing linkage data in combination with new families and novel high density genotype data to further examine the identified candidate regions to exclude false positive findings, replicate true susceptibility loci and identify novel regions and biological pathways involved in the etiology of ASDs.

The following specific aims were addressed in the studies included in this thesis:

I. To evaluate the linkage regions identified in a previous Finnish AS genome-wide linkage scan in 12 independent families and fine map possible replicated linkage regions.

II. To fine map the 3q26-q29 region previously identified in the Finnish ASD linkage scan and to investigate 11 functionally relevant candidate genes at this locus.

III. To investigate whether genetic variants in *glyoxalase 1* (*GLO1*), which had previously been implicated in the etiology of ASDs, predisposed to ASDs in the Finnish population.

IV. To investigate to what extent the isolated Finnish population displayed population stratification, and what the effects the stratification would have on GWASs.

V. Based on the results of study IV, individuals with ASDs from an internal isolate of Finland was selected in an attempt reduce genetic heterogeneity. Genome-wide SNP data was used to test if any shared haplotypes carrying possible founder mutations could be identified in these distantly related individuals. Further, genome-wide SNP data was combined with genome-wide expression data to test if these complementary methods biological pathways involved in the etiology of ASDs using analysis of GO-categories.

4 MATERIALS AND METHODS

4.1 Methods

The methods used in this study have been described in detail in the original articles (Table 5). Some of the methods are presented in more detail below.

Table 5. Methods used in this study.

Method	Reference	Publication
Laboratory procedures		
Agarose gel electrophoresis		I, II, III
Allele-specific primer extension – based SNP genotyping	Pastinen et al. (2000)	II
DNA extraction	Gentra systems, Minneapolis, MN, US	I, II, III, IV, V
Electrophoresis, ABI377/ABI3730	Applied Biosystems, Foster City, CA, US	I, II
Expression Profiling	Affymetrix, Santa Clara, CA, US	V
Genome-wide SNP genotyping	Illumina, San Diego, CA, US	IV, V
iPLEX SNP genotyping	Sequenom, San Diego, CA, US	II, III
Polymerase Chain Reaction (PCR)	Kleppe et al. (1971)	I, II, III
Sanger Sequencing	Sanger and Coulson (1975)	I, II
Analysis programs		
Genemapper	Applied Biosystems, Foster City, CA, US	I
Illumina BeadStudio	Illumina, San Diego, CA, US	IV, V
Sequencher	Gene Codes, Ann Arbor, MI, US	I, II
SNPSnapper	www.giu.fi/dnn/ Default.aspx?tabid=177	II
Typer Analyzer	Sequenom, San Diego, CA, US	II, III
Statistical methods and software		
ANALYZE	Hiekkalinna et al. (2005)	I, II
Downfreq 2.1	www.genomeutwin.org/member/ cores/stat/linkage/downfreq.html	I, II, III
Eigensoft 2.0	Patterson et al. (2006), Price et al. (2006)	IV
FBAT/HBAT	Horvath et al. (2001)	II, III
Genehunter v.2.1_r5beta	Kruglyak et al. (1996)	I

Method	Reference	Publication
Genehunter-Imprinting	Strauch et al. (2000)	I
Genotype-IBD Sharing Test (GIST)	Li et al. (2004)	II
Homog 3.35	Ott (1986)	I
LDMAP	Maniatis et al. (2002)	IV
Limma (R-package)	Smyth (2004)	V
MLINK/LINKAGE	Lathrop et al. (1984), Lathrop et al. (1986)	I
Pedcheck 1.1	O'Connell and Weeks (1998)	I, III
PennCNV	Wang et al. (2007)	V
PLINK	Purcell et al. (2007)	IV, V
PolyPhen	Ramensky et al. (2002)	II
PSEUDOMARKER	Goring and Terwilliger (2000)	I, II, III, V
SNPiga, Ciga		V
Tagger/Haploview	Barrett et al. (2005), de Bakker (2005)	II, III

DNA extraction

Blood was collected in ethylenediaminetetraacetic acid vials. DNA was extracted from samples using the Puregene DNA purification system (Qiagen, GmbH, Hilden, Germany) according to the manufacturer¨s protocol or using a phenol-chloroform protocol modified from Vandenplas and colleagues (1984).

Sequencing

Regions amplified PCR were sequenced using the BigDye 3.1 terminator (Applied Biosystems) according to manufacturer¨s instructions.

Microsatellite genotyping

PCR amplified microsatellite markers were electrophoresed on an ABI 3730 Automatic DNA sequencer (Applied Biosystems, Foster City, CA, USA) as described in Ylisaukko-oja et al. (2004).

SNP genotyping

SNPs were genotyped using allele-specific primer extension – based SNP genotyping (Pastinen et al. 2000), Sequenom iPLEX technology (Sequenom, San Diego, CA, US) or using Illumina BeadChips (Illumina, San Diego, CA, US) according to manufacturer's instructions.

Expression profiling

Mononuclear lymphocytes were isolated from peripheral blood samples. Total RNA was extracted, purified and treated according to standard procedures, and hybridized to human Affymetrix U133 Plus 2.0 chips (Affymetrix, Santa Clara, CA, US) according to manufacturer‟s instructions.

Linkage analysis

Linkage analyses were performed as described in publication I or using the dominant and recessive analysis options in PSEUDOMARKER (Goring and Terwilliger 2000). In the analysis of the autism-families, a disease allele frequency of 0.001 was used. Imprinting analysis was performed using GENEHUNTER IMPRINTING (Strauch et al 2000).

Association analysis

Association analyses were conducted using PSEUDOMARKER (Goring and Terwilliger 2000) or PLINK (Purcell et al. 2007).

Population genetic analysis

Population genetic analyses were performed as described in publication IV.

Differentially expressed genes

Re-annotation of the Affymetrix probes was performed according to the latest release of the Entrez gene database. Raw data were preprocessed and analyzed using Bioconductor 2.3 implemented in R 2.8.0 software.

Pathway analysis

Pathway analysis was performed with a non-parametric pathway analysis algorithm for global gene expression (CIGA) and GWA data (GWANA) developed in-house. The algorithm utilizes the Gene Ontology (GO) classification and aims to find the optimal regulated pathway compositions without a priori criteria for significance of individual genes. The algorithm uses an iterative cumulative hypergeometric distribution p-value based calculation.

4.2 Study subjects

Families and individuals included in this study are summarized in Figure 2 and Table 6. The datasets are described in more detail below, and in the original articles (I-V).

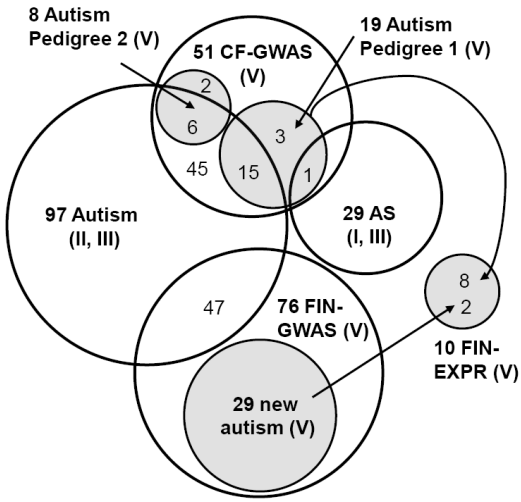


Figure 2. Finnish ASD datasets. *Overview of Finnish ASD families, and the overlap between the datasets used in this study. The studies in which the datasets are used are given in roman numerals in parenthesis.*

4.2.1 Finnish ASD datasets

Autism families (II, III, V)

The families were recruited via Finnish university and central hospitals, mainly Helsinki University Hospital, Jyväskylä Central Hospital and Kuopio University Hospital. All families have at least one child diagnosed with autism and in some families siblings diagnosed with AS. Diagnostic evaluations were made by a multidisciplinary group of clinicians at the neurological department of hospitals. Data was collected from extensive diagnostic examinations including neurological examinations, assessment of developmental history as well as psychological and neuropsychological examinations. Final diagnoses were based on ICD-10 (World Health Organization 1993) and DSM-IV (American Psychiatric Association 1994) diagnostic nomenclatures. Families with known associated medical conditions or chromosomal abnormalities such as fragile X syndrome were excluded from the

Table 6. Summary of family sets and individuals used in this study.

	N fams	N inds	N inds, LC1	N inds, LC2	N inds, LC3	Study
ASD families						
Autism	97	356	118	126	-	II, III
AS combined	29	210	114	136	-	I, III
AS original (part of AS-comb)	17	119	72	82	-	I
AS new (part of AS-comb)	12	91	42	54	-	I
Autism-AS combined	123	540	-	232	-	III
CF-GWAS	51	51	45	49	51	V
Ped1 (part of CF-GWAS)	19	19	15	18	19	V
Ped2 (part of CF-GWAS)	8	8	6	7	8	V
CF GWAS Families	50	160*	52*	58*	63*	V
FIN-GWAS	76	76	76	-	-	V
FIN-GWAS Families	76	258*	90*	96*	106*	V
GWAS Replication AGRE	859	3960	1529	-	-	V
FIN-EXPR	10	10	10	-	-	V
GEO-EXPR	35	35	35	-	-	V
Controls						
Anonymous blood donors	-	100	-	-	-	III
Subset of H2000	-	646	-	-	-	III
Finland- subisolates	-	1093	-	-	-	IV
Sweden - MZ twins	-	302	-	-	-	IV
CF-GWAS controls	-	181	-	-	-	V
FIN-GWAS controls	-	271	-	-	-	V
CNV-controls	-	5400	-	-	-	V
FIN-EXPR controls	-	10	-	-	-	V
GEO-EXPR controls	-	12	-	-	-	V

NOTE: N fams – number of families, N inds – number of individuals, ASD – Autism Spectrum Disorder, AS – Asperger Syndrome, CF – Central Finland, FIN-GWAS – Finnish replication GWAS dataset, FIN-EXPR – Finnish expression dataset, GEO-EXPR – Reference expression dataset GSE6575 LC1 – individuals with a diagnosis of autism in the autism family set and individuals with a diagnosis of AS in the AS-family set. LC2 – LC1 and siblings with a diagnosis of AS in the autism family set, LC1 and individuals not fulfilling all diagnostic criteria for AS in the AS family set and all individuals with a diagnosis of autism or AS in the autism-AS combined family set. LC3 – LC2 and individuals with a diagnosis of PDD-NOS or dysphasia, H2000 – Health 2000 cohort. *Including the individuals from the GWAS.

study. All families were Finnish except for one family where the father was of Turkish origin.

Subsequently, the Autism Diagnostic Interview – Revised (ADI-R) was administered to families willing to continue to participate in the study. The ADI-R based diagnosis has been shown to correlate very well with the consensus diagnosis of autism (Le Couteur et al. 2008). In a subset of the Finnish autism families a 96% concordance rate was observed between hospital and ADI-R diagnosis of autism (Katja Lampi, personal communication) making the Finnish autism families clinically comparable to international family sets used for genetic studies.

Individuals with a diagnosis of autism were classified in the narrow diagnostic category (liability class (LC) 1). Siblings with a diagnosis of AS were assigned to a broader diagnostic category (LC2). A total of 97 families, with one to three individuals affected with autism were included in this study. The number of families and affected individuals varies slightly in the original publications included in this study due to DNA availability limitations, but datasets are fully overlapping in all publications.

AS families (I, III)

Families were recruited via the Hospital for Children and Adolescents, Department of Child Neurology, Helsinki University Central Hospital, the Helsinki Asperger Center, Helsinki, Finland and Finnish central hospitals. Clinical examination included a structured interview based on ICD-10 (World Health Organization 1993) and DSM-IV (American Psychiatric Association 1994) diagnostic criteria for AS as well as criteria proposed by Gillberg et al. (Gillberg and Gillberg 1989; Ehlers and Gillberg 1993). The interview also included questions on hypersensitivity to external stimuli, prosopagnosia (face blindness), motor clumsiness and sleeping and eating disorders. These are not diagnostic criteria but have been frequently reported to be associated with AS (Nieminen-von Wendt 2004; Nieminen-von Wendt et al. 2005; Paavonen et al. 2008). In some cases the Asperger Syndrome Screening Questionnaire (Ehlers et al. 1999) was used to collect additional information. Interviews were performed by a research nurse with long experience of ASDs, and final diagnosis was established by a child neurologist. The majority of the individuals (80%) with a diagnosis of AS were diagnosed by one of two clinicians involved in the study (Prof. Lennart von Wendt, Taina Nieminen-von Wendt, MD, PhD). The rest were diagnosed by experienced neurologists at Finnish Central hospitals. All consenting family members, including individuals without a diagnosis of AS, were also assessed by the structured interview. All individuals who were not assessed by the clinicians involved in this study or had a hospital diagnosis of AS

were assigned an unknown status in the genetic analyses, but were included in genotyping if DNA was available, to provide phase information.

Individuals fulfilling all ICD-10 criteria for AS were classified in diagnostic category 1 (LC1). Individuals with AS like features who did not fulfill all ICD-10 criteria were classified in a broader diagnostic category (LC2). Families can be divided into two groups. The first consists of 17 extended families included in the Finnish genome-wide linkage scan for AS (Ylisaukko-oja et al. 2004), and are referred to as the original families. The second group consists of 12 independent families, referred to as the new families, collected for this study (Table 6, Figure 3). Autism was present in one family, in two individuals, a pair of male monozygotic twins. None of the families overlap with the autism family set. All families are Finnish.

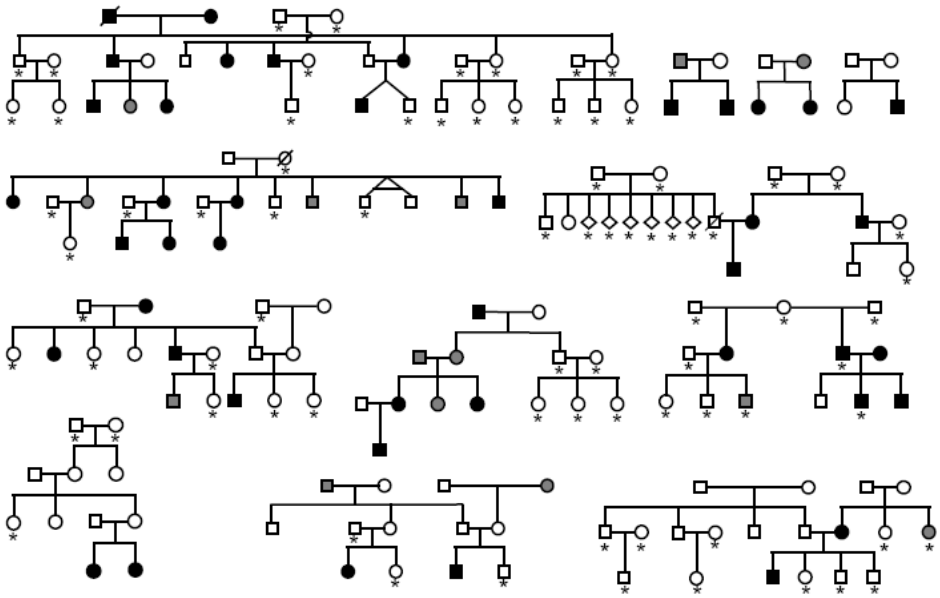


Figure 3. Pedigree structures for the twelve new AS families used in study I. *Black symbols indicate AS (LC1), grey symbols indicate AS like features (LC2) and white symbols indicate unaffected or unknown status. Individuals marked with asterisks were not included in the genotyping.*

Families from Central Finland (V)

From the autism and AS family sets, genealogical studies revealed links between nuclear families living across the country forming two extended pedigrees originating from two municipalities in Central Finland (CF, Figure 4). The first pedigree, referred to as Autism-pedigree 1 consists of 20 nuclear families, of which 16 are included in the autism family set, 1 in the AS family set and 3 in neither. One affected individual from 19 nuclear families was genotyped. The families can be genealogically traced back to two neighboring farms in the 17th century. The other extended pedigree, referred to as Autism-pedigree 2 consists of 9 nuclear families, of which six are included in the autism family set. Three of the families are not included in the autism or AS family set. Two of these had a child diagnosed with ASDs, one with AS and one with pervasive developmental disorder not otherwise specified (PDD-NOS). The third is a family from the Finnish schizophrenia family sample with two individuals affected with schizophrenia. There were no individuals diagnosed with ASDs in this nuclear family, and it was not included in this study. In Autism-pedigrees 1 and 2 individuals with a diagnosis of PDD-NOS were considered affected (LC3). From each family, we selected the individual with the most severe phenotype (autism if present, otherwise AS or PDD-NOS) for genome-wide genotyping using the Illumina HumanHap 300 Beadchip. For families with more than one child affected with autism, we selected the individual with more DNA available for genotyping. For the genome-wide SNP scan we included a subset of 24 autism families with at least two grandparents born in CF. One individual with autism from each of these families were included in the genotyping. The region of CF was defined using population history and linguistic information. The dataset consisting of 51 individuals with ASDs (mostly autism) from CF are referred to as CF-GWAS.

A replication dataset was created consisting of one proband affected with autism from all Finnish autism families that were not genealogically connected to CF (n=76). Ggenome-wide SNP data was produced using the Illumina HumanHap550 beadchip. In addition to individuals with autism from families included in studies II and III (see above) this dataset includes 29 new families in which diagnosis was established using the ADI-R. This dataset is referred to as FIN-GWAS.

For replication of the best SNPs in the GWAS, we genotyped all available family members in the families from which the proband had been included in CF-GWAS. This data was used for family-based association analysis of the best SNPs in the GWAS, which should provide increased power to detect association as some families include more than one affected individual. This dataset is referred to as CF-

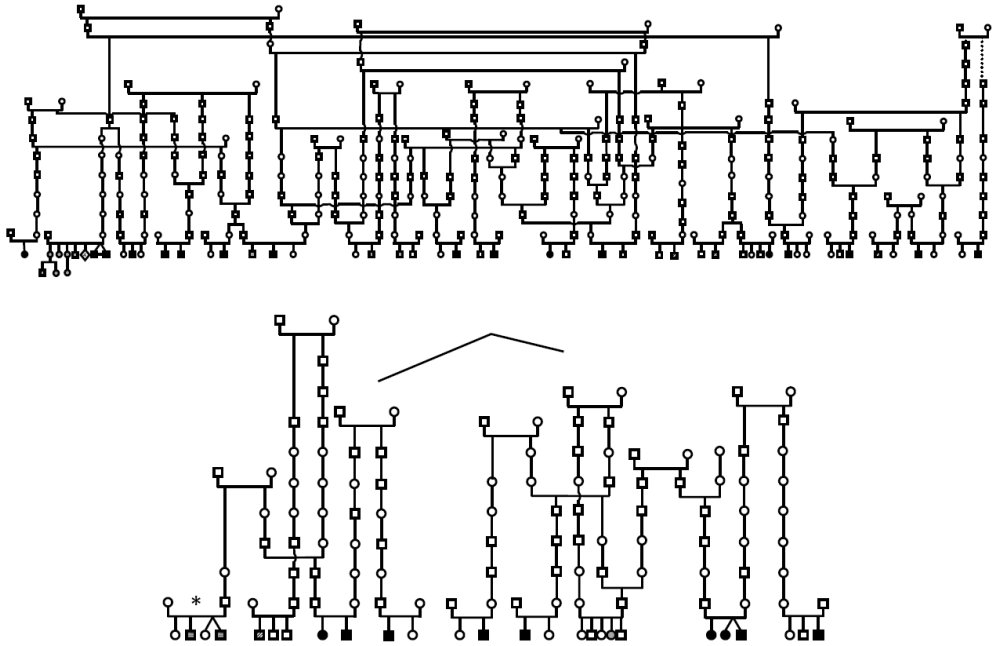


Figure 4. Pedigree structures of the two extended pedigrees identified by genealogical studies used in study V. *Black symbols denote individuals diagnosed with autism, grey symbols denote individual diagnosed with AS, symbols with diagonal lines denote individuals diagnosed with PDD-NOS and symbols with horizontal lines denote individuals with a diagnosis of schizophrenia. The nuclear pedigree indicated with an asterisk is part of the Finnish schizophrenia family collection, and is not included in this study. The top pedigree is referred to as Autism-pedigree 1 and the lower Autism pedigree 2.*

GWAS Families. We also genotyped all available members from the FIN-GWAS families including affected siblings, and this dataset is referred to as the FIN-GWAS families.

For genome-wide expression profiling, all families from the Autism-pedigree 1 with a male individual affected with autism that consented to RNA sample collection were included. This resulted in 8 affected males, which were supplemented by 2 additional samples from males affected with autism that were not part of the large pedigrees. Age at the time of sample collection ranged from 5 years 3 months to 15 years, with an average age of 10 years 9 months.

4.2.2 Non-Finnish ASD datasets

AGRE families (V)

AGRE families were genotyped using the Illumina 550K Beadchip at Children‟s hospital in Philadelphia. The dataset is publicly available on the AGRE website. Only individuals with a diagnosis of autism were included as affected in transmission disequilibrium test (TDT) analysis. This resulted in 859 families with a total of 1529 individuals affected with autism.

Expression dataset (V)

We used an additional expression dataset for pathway analysis. This dataset, GSE6575, was obtained from the Gene Expression Omnibus (GEO) database (Gregg et al. 2008). It consists of 35 individuals with autism and 12 non-autistic controls. Of the cases, 30 were male and 5 were female.

4.2.3 Population stratification study (IV)

Samples were collected at several sites across Finland and Sweden as part of larger studies. The Swedish samples (n=302) are a subset of the GenomeEUtwin study, and consist of female monozygotic twins with no data describing the geographical origin of samples (Pedersen et al. 2002). The majority of Finnish samples were drawn from two large cohorts: Health2000 and the Northern Finland Birth Cohort 1966 (NFBC66). The Health2000 cohort is a nationally representative sample of 10 000 people, originally collected to provide a comprehensive picture of health in the population aged over 18 in Finland in the years 2000-2001 (Aromaa et al. 2004). NFBC66 is a longitudinal birth cohort of individuals born in the two most northern provinces in Finland in 1966 (Rantakallio 1988). In addition, genome-wide SNP data, but only limited information about parents‟ birthplaces was available for samples from South Ostrobothnia (LSW) and Central Finland (LSC). These two groups comprise of individuals living within a particularly geographically limited region, as well as individuals from the population-based cohorts whose parents were born in these regions. Sample groups for the remaining regions were created using samples from NFBC66 and Health2000 limited strictly to individuals with both parents born within the same geographically restricted regions. The total number of Finnish samples included in the study is 1093. All study samples were anonymized with no possibility for identification of individual subjects.

4.2.4 Controls (III, V)

A set of 746 Finnish individuals were used as controls in study III. Of these, 100 were anonymous blood donors from CF and 646 were from the Health2000 cohort. No phenotype data was available for the anonymous blood donors.

For study V, the controls used in association analysis comprise of two sets. The controls used for the CF-GWAS comprise of 181 individuals consisting of anonymous blood donors from CF and individuals from the Health2000 cohorts. The controls have been matched to the cases using multidimensional scaling (MDS) extracted IBS data - no geographical data was available for these individuals. For the FIN-GWAS, we used 371 controls from a large set of Finnish GWA samples genotyped on the Illumina HumanHap 370 Beadchip. Controls were selected by complete linkage agglomerative clustering based on pairwise IBS distance. The individuals in the control database originate from several large population cohorts, which have not been screened for ASDs. However, with a population prevalence of <1%, possible ASD cases in the control cohorts should not result in any inflation of allele frequencies (WTCCC 2007).

For CNV analysis, we used a control dataset comprised of 5400 individuals from the NFBC66 population cohort, which have been genotyped using the Illumina HumanHap 370. In this population, CNVs had been determined for another study following the identical procedure used in this study.

For the Finnish genome-wide expression profiling, ten age-matched males with no diagnosis of ASDs were used as controls. The age of the controls at time of sample collection ranged from 9 years 8 months to 17 years 6 months. The average age was 12 years 3 months.

Controls for the GEO-datasets were included in the same datasets as the cases (GSE6575) and consisted of 12 individuals without a diagnosis of ASDs. The controls were age-matched to the cases, 9 were male and 3 female.

4.3 Ethical considerations

All samples have been collected and studies have been performed in accordance with the Helsinki declaration. The studies have been approved by the appropriate ethics committees. Informed written consent was obtained from all individuals or their legal guardians.

5 RESULTS AND DISCUSSION

5.1 AS Linkage study (I and unpublished data)

The first genome-wide linkage scan in families with AS but no other ASDs was performed using 17 Finnish multi-generation pedigrees with multiple affected individuals (Ylisaukko-oja et al. 2004). Three primary linkage peaks were identified, at 1q21–q22 ($Z_{max\ dom}$=3.58), 3p14–p24 ($Z_{max\ dom}$=2.50) and 13q31–q33 ($Z_{max\ dom}$=1.59).

In this study, a new, independent family set consisting of 12 families ($n_{aff\ LC1}$=41, $n_{aff\ LC2}$= 54) was genotyped using microsatellite markers resulting in Z_{max}>1.5 in the initial stage of the Finnish AS genome-wide linkage scan (Ylisaukko-oja et al. 2004). Two-point analysis under heterogeneity, performed using HOMOG and MLINK implemented in the ANALYZE package, resulted in three markers with maximum LOD scores (Z_{max}) exceeding 1. The highest Z_{max} was obtained using D3S1768 at 3p22.3 (Z_{max}=1.77, LC1, recessive model). In addition, Z_{max}=1.42 (LC2, recessive model) was obtained using D3S3547 located 5 cM distal of D3S1768. Results of two-point analysis for all markers on 3p14-p24 using different diagnostic criteria are presented in Figure 5. The third marker resulting in Z_{max}>1 was D4S3001 (Z_{max}=1.65, LC1, recessive model) on 4p15.

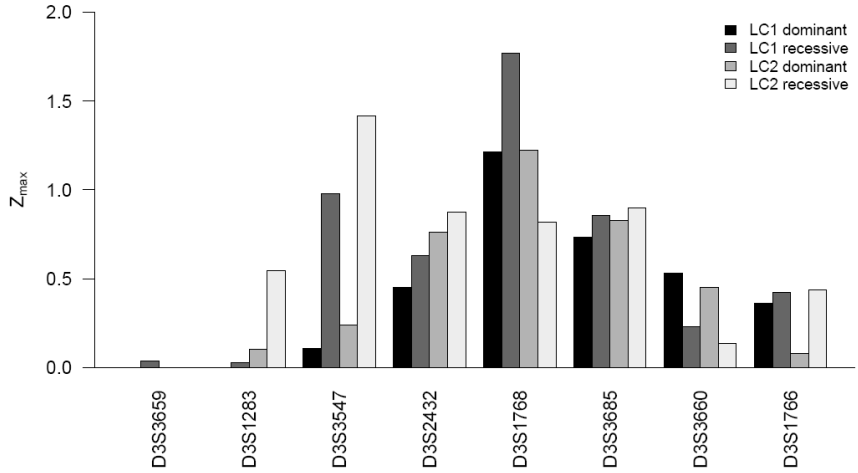


Figure 5. Results of two-point analysis of markers at 3p14-p24 in the 12 new AS families.

The highest multipoint nonparametric linkage (NPL)-score was obtained on chromosome 3p. The highest NPL_{all} was 2.80 (p=0.0055, LC2) obtained at D3S1768 (61.5 cM), the same marker which resulted in the most significant Z_{max} in two-point analysis (Figure 1 in publication I). The most significant multipoint result was obtained only 3.5cM proximal of the peak identified in the original linkage scan (Ylisaukko-oja et al. 2004). NPL_{all} >1 was also observed at 4p. At this locus, maximum NPL_{all}=1.28 (p=0.087, LC1) was observed at D4S3001 (29.47 cM).

The data from the original genome-wide screen (Ylisaukko-oja et al. 2004) and the current study were analyzed jointly (n_{fam}=29, n_{LC1}=114, n_{LC2}=136). The highest two-point LOD score (Z_{max}=3.53, LC1, dominant model) was obtained using D1S484 at 1q23.3 (Table 3, publication I). Also two flanking markers proximal to D1S484 resulted in Z_{max}>2. The new families did not contribute significantly to the linkage at this locus. The second highest two-point maximum LOD score was obtained using D3S2432 (Z_{max}=2.94, LC1, dominant model). Altogether five out of six markers in the flanking region resulted in Z_{max}>1. Both datasets contributed to the linkage signal.

Non-parametric multipoint analysis of the joint family set was performed using GENEHUNTER, and the most significant result, NPL_{all}=3.83 (p=0.0007, LC1), was obtained at 57.92 cM near marker D3S2432 at 3p22.3. The NPL scores of the combined material and the contribution of the new material are shown in (Figure 1 in publication I). The best evidence for linkage at 3p14-24 was obtained using the same marker in the original genome-wide screen (Ylisaukko-oja et al. 2004), and the most significant evidence for linkage using only the new families was obtained at D3S1768, only 3 cM proximal to D3S2432. In addition, NPL_{all}>1 were obtained on chromosomes 1q, 4q and 13q. However, at these loci, the new families contributed only marginally to the linkage evidence.

Family-based association analysis of the combined material was performed with PSEUDOMARKER 0.9.7 beta using the test of association conditional on linkage (LD|linkage) as appropriate in a region where strong linkage has been observed. A total of 11 markers resulted in p<0.05 (Table 7). The best evidence for association was obtained at D1S2815 (p=0.0024, LC1, dominant model). Results from the association analysis show that in general, the best linkage results in this study are not due to significant sharing of alleles across families, which is in agreement with analyses of haplotype sharing. Haplotypes were phased using GENEHUNTER and shared haplotypes on chromosomes 1, 3 and 4 were manually inspected. No common shared haplotypes were identified. In addition, the putative shared haplotype identified in three families from the genome-wide scan on chromosome 1 (Ylisaukko-oja 2002) was not identified in any of the new families. Interestingly, the 1q21-q24 region showing suggestive association in this study also resulted in

significant linkage in the combined analysis. It should be noted however, that the linkage signal is primarily contributed by the families from the original linkage-scan. Linkage to this region has been observed in genome-wide linkage scans for schizophrenia (Brzustowicz et al. 2000; Gurling et al. 2001; Brzustowicz et al. 2002). Finemapping of the linkage peak identified by Brzustowicz et al. (2000) revealed association between schizophrenia and eight markers in the *carboxyl-terminal PDZ ligand of neuronal nitric oxide synthase* (*NOS1AP*) at 1q23.3 (Brzustowicz et al. 2004). *Regulator of G-protein signaling 4* (*RGS4*), another gene associated with susceptibility to schizophrenia is also located in this region (Mirnics et al. 2001; Chowdari et al. 2002; Morris et al. 2004; Williams et al. 2004). The 1q21-24 region could harbor susceptibility variants shared between different neuropsychiatric disorders. Evidence already exists, that there are genes conferring risk to several different clinical entities. One such gene is *disrupted in schizophrenia 1* (*DISC1*), which has been linked to several neuropsychiatric disorders including schizophrenia, bipolar disorder, major depression and AS (St Clair et al. 1990; Hennah et al. 2003; Hennah et al. 2005; Thomson et al. 2005; Hashimoto et al. 2006; Zhang et al. 2006; Kilpinen et al. 2008). *DISC1* is involved in neuronal development, and has been associated with memory functions and other cognitive phenotypes making it an interesting and functionally convincing candidate. *DISC1* provides an encouraging example of the presence of common, shared susceptibility factors for neuropsychiatric disorders by affecting neurobiological processes.

Table 7. Markers resulting in p <0.05 in family-based association analysis of the combined AS family material consisting of 29 families.

Marker	cM	Diagnostic class	Model	p
D1S1595	161.05	LC2	dominant	0.0258
D1S1653	164.09	LC2	recessive	0.0255
D1S1167	168.52	LC2	dominant	0.0324
D1S2815	188.85	LC1	dominant	0.0024
D3S3659	47.44	LC2	dominant	0.0416
D3S2432	57.92	LC1	dominant	0.0280
D3S3583	195.6	LC1	recessive	0.0040
D3S2398	209.41	LC1	dominant	0.0210
D4S3001	49.47	LC1	dominant	0.0430
D4S1090	158.65	LC1	recessive	0.0259
D6S1671	107.88	LC2	recessive	0.0264

NOTE: cM – centiMorgan, LC1 – narrow diagnostic classification, LC2 – broad diagnostic classification. p-values are reported from the PSEUDOMARKER LD|linkage test.

To investigate whether the 3p14-p24 locus is subject to imprinting, we used GENEHUNTER-IMPRINTING (Strauch et al. 2000) In this analysis, two heterozygote penetrance parameters are defined and paternal and maternal origin of the disease allele can be treated differently in terms of probability of expression of the trait. We found no evidence for parent-of-origin effects at this locus, as the maternal imprinting-model resulted in HLOD=0.04 while the paternal imprinting-model resulted in HLOD=1.81 (both LC1, dominant model). Most significant LOD scores were obtained in the original analysis, resulting in multipoint heterogeneity LOD=3.71 between markers D3S2432 and D3S1768 using a model with no imprinting (LC1, dominant model).

In an effort to extract all possible linkage information from this set of families, we finemapped the region on 3p14-p24 using 13 additional markers in all 29 families. Six of the 13 fine mapping markers resulted in $Z_{max}>1$ (Figure 6). Family-based association analysis revealed two adjacent markers D3S2432 (from the first stage) and D3S1619 (finemapping marker) showing suggestive sharing across families using the LC1 and dominant model (LD|linkage p=0.028 and p=0.041, respectively, Figure 7). The region flanked by these two markers extends 3 cM and contains 13 known genes. The fine-map did not significantly increase the evidence of linkage in the region, suggesting that available linkage information has been extracted from the region using the current density of microsatellite markers. Future studies using dense SNPs in the region will reveal the LD structure and the haplotype diversity in the region in these families, hopefully further specifying the region of interest and identifying haplotypes shared between families harboring possible susceptibility variants.

As the best microsatellite marker, D3S2432 is located within the gene *glycerol-3-phosphate dehydrogenase 1-like* (*GPD1L*), we wanted to sequence coding exons and exon-intron junctions to rule out possible coding variants as susceptibility variants for ASDs. *GPD1L* shows 72% identity with *NAD-dependent glycerol-3-phosphate dehydrogenase* (*GPD1*), and is expressed in several tissues including brain (Nagase et al. 1995). Mutations in *GPD1L* have been identified in families with Brugada syndrome, a syndrome that causes cardiac arrest due to ventricular fibrillation in apparently healthy individuals, and therefore, makes for a very unlikely candidate gene for ASDs. We sequenced all 8 coding exons in 13 patients from AS families contributing to the linkage peak at 3p. We identified seven variants which had all been previously detected, and were present in dbSNP (Table 8). Mutation analysis did not reveal any variants likely to contribute to the etiology of ASDs.

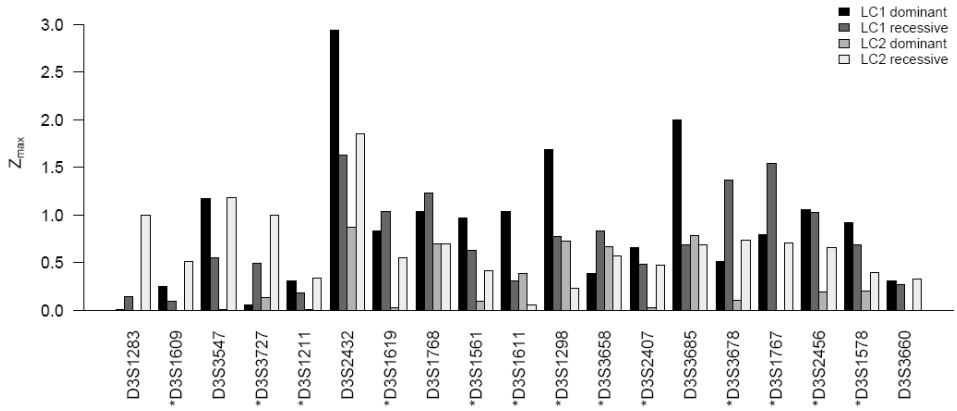


Figure 6. Results of two-point linkage analysis at 3p14-p24 in the combined AS family material consisting of 29 families. *Markers indicated by an asterisk were added in the finemapping stage.*

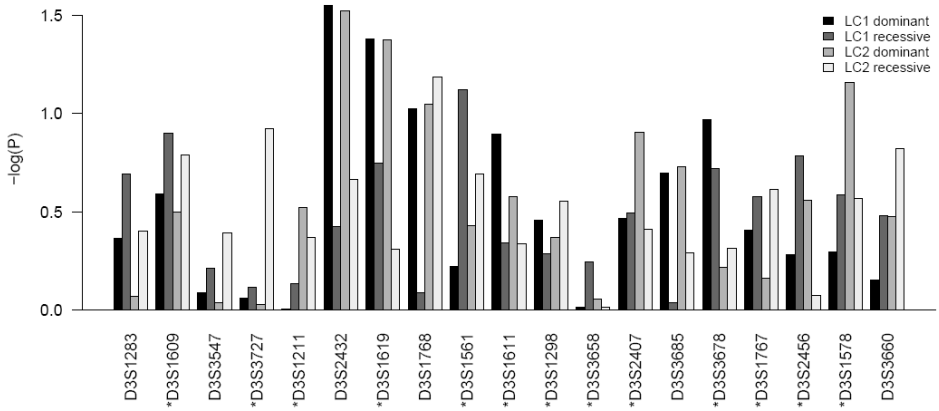


Figure 7. Results of family-based association analysis at 3p14-p24 in the combined AS family material consisting of 29 families. *Results from the PSEUDOMARKER LD|linkage test are reported. Markers indicated by an asterisk were added in the finemapping stage.*

Table 8. Variants identified in the sequencing of* GPD1L *in 13 individuals with AS from families contributing to the linkage signal at 3p.

Exon	Pos	SNP	AA change	N	Sequence	dbSNP
ex1-231	-231	G>A	-	8	GCGCAGCGAC**A**GGGGGTGAGG	rs13095243
ex1-231	-231	G>R	-	4	GCGCAGCGAC**R**GGGGGTGAGG	rs13095243
ex1-168	-168	T>K	-	4	GAGCCGGGCC**K**GGAGGCCGGC	rs6805518
ex1-168	-168	T>G	-	9	GAGCCGGGCC**G**GGAGGCCGGC	rs6805518
ex1-59	-59	C>Y	-	1	TGGCCGCGAC**Y**ATTGGCGGAG	rs1077601
ex4	33 618	C>Y	D136D	2	TCATTTCTGA**Y**ATCATCCGTG	rs9835387
ex5-218	39 863	G>A	-	1	ATCCTATCAT**A**AACATGTTAC	rs9839702
ex5+23	40 106	A>-	-	1	TCAGGGGAGG[]GTTCATCAAG	rs11351972
ex8	59 463	A>R	-	5	ACATGCAAAC**R**TGTGAATGGT	rs6799559
ex8	59 463	A>G	-	2	ACATGCAAAC**G**TGTGAATGGT	rs6799559

NOTE: Pos – Position is given as bp from the start of exon 1 of *GPD1L* (NM_015141.2), corresponding to position chr3:32,123,143 according to NCBI build 36.1. N indicates the number of cases with the variant.

In conclusion, we have replicated the linkage finding to the 3p14-p24 region in an independent set of AS families. Both the original and the replication family sets were relatively small suggesting that the region could contain a susceptibility gene with high individual gene-effect. Genome-wide linkage scans for ASDs have not identified significant linkage overlapping with the region identified in this study.

Suggestive linkage to autism has been obtained in two independent studies, 10.5 and 21.8 cM distal of the best AS region (Auranen et al. 2002; Shao et al. 2002b) (Figure 8). In a combined analysis of the Finnish autism families and publicly available AGRE families from the heterogeneous US population the best linkage signal was observed on 3p24-p26. The highest NPL_{all}=2.20 was obtained at D3S3691 located 28.7 cM distal of the best AS region (Ylisaukko-oja et al. 2006). Simulation studies have shown substantial variation in the location of the best evidence for linkage even when the linkage is the result of a single genetic signal (Roberts et al. 1999). Therefore these non-overlapping but adjacent linkage peaks could be caused by shared underlying genetic risk factors.

In published linkage studies, complementary approaches have been used to untangle the genetic constituents of ASDs. One strategy is to include a large sample set to increase the information content, but this may also lead to increased phenotypic as well as genetic heterogeneity. Another approach is to use individuals with a specific trait or endophenotype, which leads to a decreased sample size, but could benefit

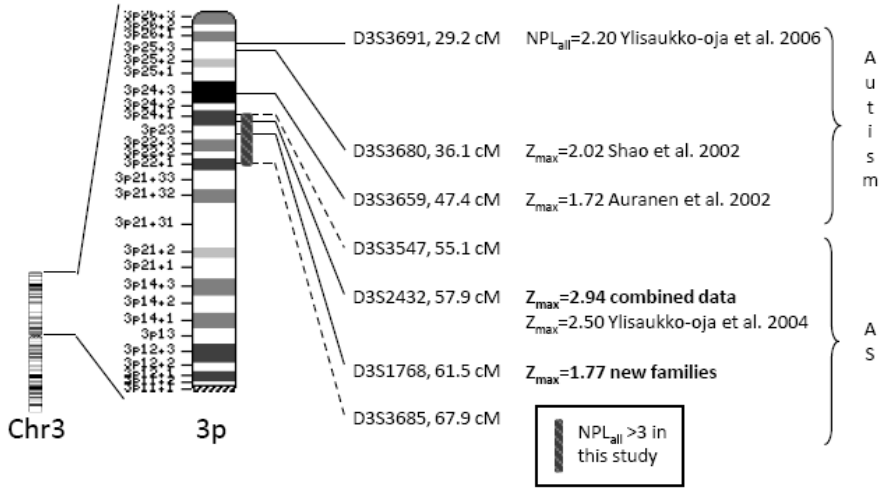


Figure 8. Findings on the short arm of chromosome 3 in autism and AS. *Marker location is given in centiMorgan (cM) from the telomere of the short arm. Results obtained in this study are presented in bold type.*

from the decreased genetic heterogeneity in the smaller sample. This approach has been used successfully in two studies identifying a susceptibility region on chromosome 2q in individuals with autism and delayed phrase speech (Buxbaum et al. 2001; Shao et al. 2002a). In this study we have attempted to decrease the genetic heterogeneity both by selection of AS, an endophenotype of autism, and by performing the study using families from the isolated Finnish population which has been shown to be genetically more homogenous compared to more outbred populations (Sajantila et al. 1996, study IV). Additionally, affected individuals were divided into two diagnostic categories, and in many cases better results were obtained using only a narrow definition of the phenotype.

Although linkage has now been established to 3p14-24 in AS in two independent studies, this locus needs to be further evaluated in other populations to determine if this locus contains genetic variants conferring susceptibility to AS specific to the Finnish population, or if the risk factors are common in families with AS. To date, no family sets comparable to the ones used here and in the original Finnish genome-wide linkage scan have been published. The ASD datasets used in other studies have traditionally been ascertained through probands affected with autism. We propose that large, multigenerational pedigrees affected only with AS serve as a genetically more heterogeneous subgroup of families with ASDs, and encourage the collection of such families in other studies as well.

5.2 3q finemap (II and unpublished data)

In the Finnish ASD linkage scan strongest linkage was observed at 3q25-q27 resulting in a maximum two-point LOD score of 4.31 at D3S3037 (Auranen et al. 2002). Here, we wanted to finemap the region to further restrict the region of interest, and to test functionally interesting candidate genes for association with autism.

5.2.1 Finemapping using microsatellites

We selected microsatellite markers to finemap the distal part of the linkage peak identified in the Finnish ASD genome-wide linkage-scan (Auranen et al. 2002). In a previous study of three neuroligin family genes, we finemapped the proximal part of the linkage peak using six new microsatellites as well as 18 SNPs in and flanking the *NLGN1* gene (Ylisaukko-oja et al. 2005). Here, further finemapping markers were selected from the Marshfield genetic map and UCSC genome browser. Intragenic markers, and markers located in the promoter region of genes were preferred. Two intergenic markers were also included for a more complete coverage of the region. In genes where no microsatellite markers were available, we used the RepeatMasker tool to identify intragenic repeats, and genotyped these in a set of controls to identify the polymorphic repeats. As a result, 11 new microsatellite markers were added in this stage. Together with the microsatellite markers from the original linkage-scan and the *NLGN1* study, a total of 37 microsatellite markers were included in the region between D3S3053 located at 3q26.31 and D3S1311 located at 3q29, covering a region of approximately 25 Mb. The average marker spacing was 1.16 cM. Markers were genotyped in 97 autism families.

Two-point association analyses were performed using both LC1 corresponding to individuals with a diagnosis of autism, and LC2, corresponding to individuals with autism and siblings with AS (Figure 9). We used PSEUDOMARKER to test for association conditional on linkage (LD|linkage) using both recessive and dominant models. Best results were obtained using LC1 and the dominant model, although an almost similar level of significance was obtained using LC2. The results are in agreement with the genome-wide linkage scan, in that the most significant results were obtained using the dominant model. The most significant association was obtained with one of the fine-mapping markers, D3S1521 (p=0.01), located approximately 5.7 Mb distal of D3S3037, which resulted in the highest two-point LOD scores in stage II of the genome-wide linkage scan (Auranen et al. 2002). D3S3037 resulted in the second most significant p-value in PSEUDOMARKER analysis (p=0.02).

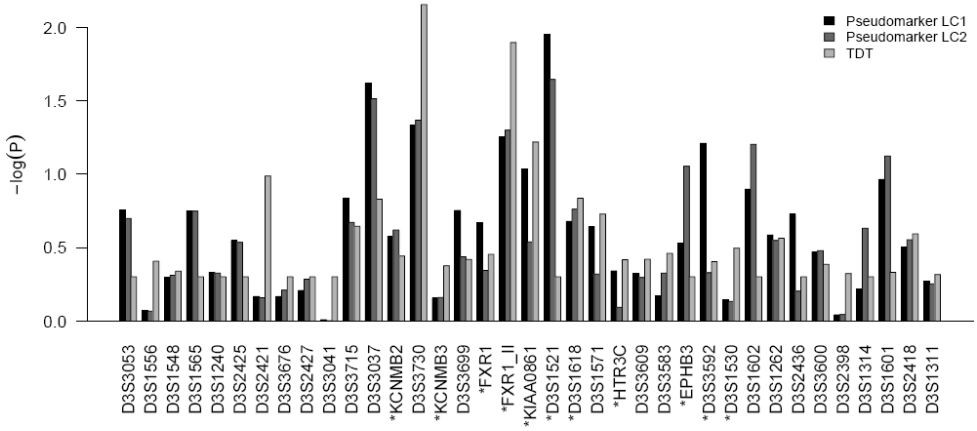


Figure 9. Results of the family-based association analysis of 3q26-q29 in 97 Finnish autism families. *Results from PSEUDOMARKER LD|linkage test using a dominant model and TDT-analysis are reported. Markers marked with an asterisk were added in the finemapping stage.*

When we used the TDT-test, most significant association was obtained with D3S3730, located 1.5 cM distal of D3S3037. However, the results of both tests of association suggest that linkage in this region is not due to significant sharing of susceptibility alleles across families.

5.2.2 Association analysis of 11 candidate genes at 3q26-q27

We used the UCSC genome browser to investigate genes located at the previously identified autism linkage peak at 3q2 *in silico* and selected 11 functionally interesting candidate genes (Table 9). A total of 125 SNPs were polymorphic and were successfully genotyped in 97 families (n_{aff}=112), of which 32 families were included in the original genome-wide linkage scan.

We performed family-based association analysis using PSEUDOMARKER and the LD|linkage test. The dominant model resulted overall in more significant results compared to the recessive model. Best results were obtained using markers in *FXR1* and *HTR3C*, but two SNPs in *KCMNB2* and one SNP in *HTR3E* and *GNB4* respectively, also resulted in suggestive association with autism ($p<0.05$) (Table 10). The most significant association was observed with two nsSNPs in *HTR3C*, rs6766410 (N163K) and rs6807362 (G405A), both resulting in p=0.0012. The non-synonymous variants are common SNPs with minor allele frequencies ranging from

0.42 to 0.49. In agreement with the high frequency of these substitutions, *in silico* analysis did not predict the amino acid changes to result in significant structural or functional changes according to PolyPhen. We used GIST software to test post hoc if the most significantly associated SNPs could partly account for the linkage in the families included in the original linkage scan. Both nsSNPs resulted in non-significant P-values (p=0.15 for rs6766410 and p=0.30 for rs6807362 using a dominant model), suggesting these SNPs alone do not account for the linkage signal.

Table 9. Candidate genes included in the association analysis of 3q26-q27.

Gene symbol	Gene name	Length (kb)	SNPs genotyped
KCNMB2	calcium-activated potassium channel beta 2	308.0	40
PIK3CA	phosphoinositide-3-kinase catalytic alpha	86.2	15
KCNMB3	calcium-activated potassium channel beta 3	24.2	8
GNB4	guanine nucleotide-binding protein beta-4	55.5	17
FXR1	Fragile X mental retardation related, autosomal homolog	64.5	6
HTR3D	5' hydroxytryptamine receptor subunit 3D	7.8	13
HTR3C	5' hydroxytryptamine receptor subunit 3C	7.6	
HTR3E	5' hydroxytryptamine receptor subunit 3E	6.8	
DVL3	dishevelled 3	18.0	5
CHRD	Chordin	9.8	12
EPHB3	ephrin receptor B3	20.6	9

Table 10. Results of the family-based association analysis for all markers resulting in p<0.05 in the candidate gene study of 3q26-q27 in 97 Finnish autism families.

SNP	Position	Gene	Description	P
rs9839672	180034391	*KCNMB2*	Intron	0.026
rs4542999	180045647	*KCNMB2*	Downstream	0.007
rs13087067	180604702	*GNB4*	Intron	0.008
rs1607678	182067412	*FXR1*	Upstream	0.008
rs7630922	182075054	*FXR1*	Upstream	0.006
rs1805567	182121946	*FXR1*	Intron	0.004
rs6766410	185257456	*HTR3C*	NS coding (N163K)	0.001
rs6807362	185260704	*HTR3C*	NS coding (G405A)	0.001
rs9872968	185285553	*HTR3C*	Downstream	0.029
rs7627615	185301110	*HTR3E*	NS coding (A86T)	0.036

NOTE: p-values from PSEUDOMARKER LD|linkage test, dominant model are presented.

We used Haploview to determine the haploblock structure within candidate genes, and HBAT to test for association within the blocks. The only haplotype resulting in suggestive association (p=0.05) was a three-SNP haplotype containing the two nsSNPs in *HTR3C*. The two-marker haplotype consisting only of the two non-synonymous SNPs in this block also showed suggestive association with autism (p=0.020). The haplotype C-C, corresponding to amino acids N163 and A405 was overtransmitted to individuals with autism (p=0.006), whereas the haplotype A-G, corresponding to amino acids K163 and G405 showed suggestive undertransmission to individuals with autism (p=0.014).

To test if the SNPs associated with ASDs in this study tagged other genetic variants, we we sequenced coding regions and splice sites of *FXR1* and *HTR3C* in seven individuals with autism. Individuals were probands from families contributing to the original linkage finding in the Finnish genome-wide linkage scan for ASDs (Auranen et al. 2002). No novel variants were identified in the sequence analysis, suggesting either that the associated variants are the true susceptibility variants, or that other variants exist but are located in intronic regions which were not covered by the sequencing efforts.

5.2.3 Association analysis of *ZIC1* and *ZIC4*

Dandy-Walker malformation (DWM, OMIM #220200) is a congenital malformation of the cerebellum. Affected individuals display motor deficits such as delayed motor development, hypotonia and ataxia, and around half have MR. Recurrence rates of 1-2% suggest that DWM does not follow Mendelian inheritance, and is likely inherited in a polygenic fashion (Murray et al. 1985). Heterozygous deletion of two flanking *Zinc finger in cerebellum* genes, *ZIC1* and *ZIC4*, has been identified in individuals with DWM (Grinberg et al. 2004). In the same study, Grinberg and colleagues showed that heterozygous loss of *Zic1* and *Zic4* in mice is sufficient to cause cerebellar malformations similar to those found in DWM. In some cases, autistic features have been reported as part of DWM (Aldinger 2008). Although there have been no reports on consistent structural brain-abnormalities in ASDs, one of the most common findings is a decrease in the number of Purkinje cells in the cerebellum (Williams et al. 1980; Palmen et al. 2004; Vargas et al. 2005; Whitney et al. 2008). In addition, some studies have reported cerebellar vermal hypoplasia in ASDs, which is also observed in DWM (Courchesne et al. 1988; Hashimoto et al. 1995; Courchesne et al. 2001; Kaufmann et al. 2003). Encouraged by these findings and the location of *ZIC1* and *ZIC4* at 3q24, near the best linkage peak identified in the Finnish ASD linkage-scan, we wanted to investigate if these two genes were involved in the etiology of ASDs in Finnish families. Five rare variants have

Table 11. Rare variants identified in ZIC4 in patients with DWM (Grinberg 2005).

SNP	DNA	Protein	dbSNP	Patients	Controls
ZIC4_SNP1	321C>G	Y16STOP	-	1/176	0/192
ZIC4_SNP2	568G>C	G99R	rs34676558	1/176	2/370
ZIC4_SNP3	779C>G	A169G	-	1/176	2/384
ZIC4_SNP4	1113G>T	S280S	-	1/176	0/192
ZIC4_SNP5	1480delT	3'UTR	-	1/176	0/384

NOTE: DNA – The position of the SNP in the ZIC4 gene, numbering starts from 1 at the transcription start site.

previously been identified in *ZIC4* in patients with DWM (Grinberg 2005) (Table 11), and we included these five variants together with tag-SNPs identified using the Tagger algorithm implemented in Haploview to capture the allelic variation in these two genes. Altogether 13 SNPs were genotyped in 97 ASD families ($n_{aff\ LC1}$=118, $n_{aff\ LC2}$=126). The TDT-test was performed using PLINK and a test of association conditional on linkage was performed using PSEUDOMARKER.

No SNPs showed association with autism ($p<0.05$) in the TDT or PSEUDOMARKER analyses (Table 12, Table 13). Three of the rare variant SNPs were monomorphic: ZIC_SNP1, ZIC_SNP4 and ZIC_SNP5. For ZIC_SNP2, corresponding to rs34676558, a minor allele frequency of 2% was observed in the ASD families. None of the individuals were homozygous for the rare allele, but 5% (18/357) were heterozygotes. Of these, 7 were affected and 11 were unaffected parents. A minor allele frequency of 2% was observed for the ZIC4_SNP3. Again, no homozygotes for the rare allele were observed. A total of 15 individuals heterozygous for the minor allele were identified, 6 were affected with ASDs, 8 were unaffected parents and 1 was an unaffected sib, whose affected sibling was homozygote for the common allele.

To conclude, SNPs in *ZIC1* and *ZIC4* do not show association with ASDs in the Finnish families, and rare variants ZIC4_SNP2 and ZIC4_SNP3 appear to be rare polymorphisms observed in both affected and unaffected individuals. The role of variants ZIC4_SNP1, ZIC4_SNP4 and ZIC4_SNP5 need to be evaluated in larger patient cohorts with DWM to establish the role of these variants in the etiology of DWM. In addition, as DWM was originally reported to be caused by deletion of the *ZIC1/ZIC4* locus, CNVs encompassing this locus should be evaluated in cohorts with ASDs.

Table 12. Results of the TDT analysis of SNPs at the ZIC1-ZIC4 locus in 97 Finnish autism families.

SNP	A1	A2	T	U	OR	CHISQ	P
rs3796128	A	G	13	6	2.17	2.58	0.11
rs2279829	A	G	43	42	1.02	0.01	0.91
ZIC4_SNP1	NA	NA	NA	NA	NA	NA	NA
rs17509456	C	G	10	6	1.67	1.00	0.32
rs34676558	C	G	5	4	1.25	0.11	0.74
ZIC4_SNP3	G	C	6	3	2.00	1.00	0.32
ZIC4_SNP4	NA	NA	NA	NA	NA	NA	NA
rs3852000	C	T	52	60	0.87	0.57	0.45
rs10804719	C	T	31	30	1.03	0.02	0.90
ZIC4_SNP5	NA	NA	NA	NA	NA	NA	NA
rs954735	C	T	23	29	0.79	0.69	0.41
rs7614043	T	C	14	15	0.93	0.03	0.85
rs1394042	G	A	9	10	0.90	0.05	0.82

NOTE: A1 – allele 1, A2 – allele 2, T – transmitted minor allele count, U – untransmitted minor allele count, OR – odds ratio, CHISQ – chi square test statistic, P – p-value. Markers with results indicated by NA were monomorphic.

Table 13. Results of the PSEUDOMARKER analysis of SNPs at the ZIC1-ZIC4 locus in 97 Finnish autism families.

Marker	LC1, dominant	LC1, recessive	LC2, dominant	LC2, recessive
rs3796128	0.51	0.26	0.19	0.16
rs2279829	0.84	0.54	0.96	0.73
ZIC4_SNP1	NA	NA	NA	NA
rs17509456	0.24	0.03	0.30	0.05
rs34676558	0.52	0.96	0.67	0.61
ZIC4_SNP3	0.81	0.78	0.74	0.71
ZIC4_SNP4	NA	NA	NA	NA
rs3852000	0.89	0.40	0.94	0.62
rs10804719	0.96	0.60	0.77	0.94
ZIC4_SNP5	NA	NA	NA	NA
rs954735	0.26	0.36	0.58	0.82
rs7614043	0.45	0.41	1.00	0.77
rs1394042	0.71	0.40	1.00	1.00

NOTE: LC1 – narrow diagnostic classification corresponding to autism only, LC2 – broad diagnostic classification corresponding to autism and AS. All results are p-values from LD|linkage statistic. Markers with results indicated by NA were monomorphic.

5.2.4 Sequence analysis of *PEX5L*

Hyperpolarization-activated cyclic nucleotide-gated (HCN) channels are homo- or heterotetramers consisting of four subunits (HCN1-HCN4). These voltage-gated channels mediate the hyperpolarization-activated current (I_h), and characterize the integration of incoming synaptic input as well as determine the pattern of action potential output (reviewed in Wahl-Schott and Biel 2009). The differential expression patterns of the subunits with different biophysical properties contribute to differential neuronal excitability in different cell types and brain regions (Notomi and Shigemoto 2004). HCN family members interact with different scaffolding proteins, among others the tetratricopeptide repeat (TPR)-containing Rab8b-interacting protein (TRIP8b). TRIP8b results in altered trafficking and reduction in cell surface expression of HCN channels (Santoro et al. 2004). It has also been recently shown that alternatively spliced isoforms of TRIP8b differentially control HCN channel localization function (Lewis et al. 2009). The TRIP8b protein is encoded by the peroxisomal biogenesis factor 5-like (*PEX5L*) gene at 3q26.33, located only 2 Mb distal of the linkage peak identified in the Finnish ASD genome-wide linkage scan (Auranen et al. 2002). As TRIP8b has been shown to regulate normal neural function and excitability associated with neurological disease we wanted to investigate whether *PEX5L* contributed to etiology of ASDs in Finnish families.

We sequenced a total of 17 coding exons, of which the first two were alternative first exons, present in different splice variants in 83 probands with autism and 11 controls. The success rate was >97%, ranging from 90.5%-100% per exon. A total of 12 variants were identified, of which 3 were not present in dbSNP (Table 14). All identified variants were located in introns.

We identified three novel SNPs, of which one was also identified in controls. Two of the variants that were not present in dbSNP were identified in cases but not in controls. One variant was identified 39 bp downstream of the end of exon 1a, and was found in one case as a heterozygous SNP, and in another case as a homozygous C>A SNP. The other SNP identified only in cases was present in three cases as a heterozygous SNP. It was located 36 bp upstream of the start of exon 7. As the number of controls in this study was very small (n=11) these two SNPs should be evaluated in a larger control cohort for accurate estimates of allele frequencies.

Although we did not identify any SNPs which are likely to modify the protein structure, *PEX5L* remains a positional and functionally interesting candidate gene for ASDs. Future work should include the evaluation of different isoforms of *PEX5L* mRNA in autism cases and healthy controls in different brain regions. In addition, the locus should be tested for CNVs compassing *PEX5L*.

Table 14. Variants identified in the sequencing of coding exons of PEX5L in 83 probands from the Finnish autism families.

Exon	Position*	SNP	Case	Contr	Sequence	dbSNP
ex1a+39	190	C>A	1	0	ATTCTTGGGG**A**CTAATTTAAG	NA
ex1a+39	190	C>R	1	0	ATTCTTGGGG**R**CTAATTTAAG	NA
ex1c	63 424	C>T	3	0	CTGTGGTTCT**T**TGGTCAGCAA	rs12054457
ex1c	63 424	C>Y	29	2	CTGTGGTTCT**Y**TGGTCAGCAA	rs12054457
ex2+17	65 153	G>A	3	0	GAATACTTAC**A**TACATTTTCT	rs3774257
ex2+17	65 153	G>R	29	2	GAATACTTAC**R**TACATTTTCT	rs3774257
ex5-75	148 871	G>T	3	2	TCATTTTGTT**T**CAAGGCTGCA	rs41265425
ex5-75	148 871	G>K	35	6	TCATTTTGTT**K**CAAGGCTGCA	rs41265425
ex6-68	156 539	C>T	4	0	CTGTACCCAA**T**TGAAACCTAT	rs13067789
ex6-68	156 539	C>Y	15	1	CTGTACCCAA**Y**TGAAACCTAT	rs13067789
ex7-87	161 166	G>R	33	5	TGCACCACAC**R**CTCCACCCTC	rs2302743
ex7-87	161 166	G>A	22	2	TGCACCACAC**A**CTCCACCCTC	rs2302743
ex7-36	161 217	A>R	3	0	CCGCCCTGCC**R**CCACCTGCCT	NA
ex8-29	162 278	G>A	27	2	TCTAGTGTAC**A**TTTAATTTCC	rs2339914
ex8-29	162 278	G>R	37	6	TCTAGTGTAC**R**TTTAATTTCC	rs2339914
ex8+62	162 465	C>T	66	9	AGCATAAATG**T**TTTATTTTCA	rs1609981
ex8+62	162 465	C>Y	16	2	AGCATAAATG**Y**TTTATTTTCA	rs1609981
ex10-64	216 691	A>R	5	1	TAGGGGGGGA**R**GAAAAACACA	rs1001668
ex11-34	220 693	A>R	6	1	AGAGAAATCA**R**TGGCTCATGC	rs17690759
ex14+34	228 494	G>R	5	1	AAACCACAGC**R**TTCTCTGTAA	NA

NOTE: Position is given as bp from the star of *PEX5L* exon 1 (NM_016559.1), corresponding to position chr3:181,237,211 according to NCBI build 36.1. Case and Contr columns indicate the number of cases or controls carrying the variant.

5.2.5 Discussion

In this study, we wanted to finemap the linkage peak at 3q2 identified in the Finnish ASD scan. We used microsatellites in an attempt to limit the region of interest, and after that, investigated a number of functionally interesting candidate genes to identify possible susceptibility genes in the region. For candidate gene analysis, we employed both association- and sequence analysis. Together with two previously published candidate gene studies, one of *NLGN1* located 8 Mb proximal of the linkage peak at 3q26.31 (Ylisaukko-oja et al. 2005), and the other of *ATP13A4* located 16 Mb distal of the linkage peak at 3q29, (Kwasnicka-Crawford et al. 2005), we have now evaluated 15 genes at 3q2 for association with ASDs.

Finemapping using microsatellites did not significantly help limit the region of interest, although suggestive evidence of allele sharing between families was obtained using flanking markers D3S1521 and D3S3037. Of the candidate genes tested here, this region contains *KCMNB2, PIK3CA, KCMNB3, GNB4, PEX5L* and *FXR1*.

The linkage finding to 3q2 originally reported in Finnish families has been replicated in other populations, making it a highly interesting candidate region. In a genome-wide linkage scan of one large ASD family, the most significant linkage signal was observed at rs1402229 (p=0.0003), only 2.6 Mb from the best marker (D3S3037) in the Finnish genome-wide scan (Coon et al. 2005). A sequence analysis of *FXR1,* which is located in the linkage region, did not reveal any variants likely to contribute to ASDs, in line with the results observed in our study. However, we observed suggestive association of three SNPs in the promoter and introns of *FXR1* (p=0.04-0.08), suggesting that possible intronic risk-variants could exist in *FXR1*.

Further support for linkage between autism and 3q was obtained in a study of language QTLs in autism. The strongest evidence for "age at first word" (p<0.001) was at 147cM, around 40cM from the best findings in the Finnish genome-wide scan (Alarcon et al. 2005). In the same study suggestive evidence for linkage of "age of first phrase" (p=0.04) was obtained at 180 cM, only 10 cM proximal to our peak at 190cM.

The most promising association results in this study were obtained with two common, coding SNPs in *HTR3C*. This gene is located only 6 Mb distal of D3S3037 and is the most interesting region identified in the microsatellite finemapping. The two SNPs, rs6766410 (N163K) and rs6807362 (G405A), both resulted in p=0.0012 in family-based association analysis. The SNPs are in high LD (D'=0.96 in the HapMap CEU data) and are located within the same haplotype block. In addition, the haplotype consisting of these two SNPs resulted in suggestive evidence for association (p=0.02).

Most serotonin receptors are G-protein-coupled binding proteins and the 5-HT3 receptor is the only ligand gated ion channel in the family. Subunits *HTR3A* and *HTR3B* of this receptor are well characterized (Miyake et al. 1995; Davies et al. 1999) but *HTR3C, HTR3D* and *HTR3E*, forming a cluster on 3q27, have only been recently discovered and are incompletely characterized (Karnovsky et al. 2003; Niesler et al. 2003; Niesler et al. 2008). Subunits C-E display identities of roughly 70% between each other and a 26-29% identity to *HTR3A* and *HTR3B* at the protein level. Expression of *HTR3E* and *HTR3D* is limited to some internal organs, whereas *HTR3C* displays a wider expression pattern that includes the brain, and therefore,

resembles *HTR3A* and *HTR3B* more closely. Exon structure of *HTR3C* is identical to *HTR3A* and *HTR3B*, while the two other subunits display different exon structures (Niesler et al. 2003). It has recently been shown that that although *HTR3C, HTR3D* or *HTR3E* cannot form homomeric receptors, these subunits are able to form heteromeric receptors with *HTR3A* and modulate receptor function and properties in a similar manner to *HTR3B* (Davies et al. 1999; Niesler et al. 2007). *HTR3C, HTR3D* and *HTR3E* are absent in rodents making it necessary to use human cells to investigate the function of these subunits (Karnovsky et al. 2003). It has recently been shown that *HTR3C*, *HTR3D* and *HTR3E* cannot alone form functional receptors. They are, however, able to form heteromeric receptors together with *HTR3A*, and modulate functional properties of these receptors compared with homomeric *HTR3A* receptors (Niesler et al. 2007).

Elevated whole blood serotonin in individuals with autism and their close relatives is the most consistent biological finding in autism (see section 2.3.11 for a summary of studies of the serotonergic system in ASDs). Although the HTR3 receptors cannot be directly linked to levels of blood serotonin, variants in *HTR3A* have interestingly been shown to modulate neural activation of the amygdala during a face recognition task (Iidaka et al. 2005), a process which has repeatedly been implicated as abnormal in individuals with ASDs (Klin et al. 1999; Schultz 2005; Kleinhans et al. 2008)It should be noted, that the results of the linkage analysis do not necessarily support the hypothesis of common variants in this region contributing to ASD susceptibility. Linkage studies only investigate the co-segregation of a haplotype within families, but do not require the haplotype be shared between families as the total LOD-score is obtained by summing up the LOD scores in each family. Although a haplotype shared between two families at this locus was reported in the original linkage study (Auranen et al. 2002), genotyping a dense map of microsatellites revealed that the haplotypes differed between the two families (data not shown). Therefore it is possible, that the linkage-peak is a not a result of a shared susceptibility variant, but rather of multiple rare risk variants not shared between families in this region. These kinds of susceptibility factors would not be detectable in an association study. Rare variants could all be located in the same gene. It is even possible, though less probable, that rare risk factors reside in several genes in the region. Most risk factors identified in ASDs to date are rare variants present in only one or a few families (Jamain et al. 2003; Feng et al. 2006; Durand et al. 2007; Marshall et al. 2008).

5.3 GLO1 association study (III)

In a previous study, a proteomics approach was used to identify protein abnormalities in *glyoxalase I* (*GLO1*) in brains of individuals with autism (Junaid et al. 2004). *GLO1*, residing at 6p21.2, is part of the glyoxalase system involved in the detoxification of methylglyoxal, a cytotoxic byproduct of glycolysis (Thornalley 2003). A nsSNP (rs2736654, Ala111Glu) resulting in a protein product with lower enzymatic activity was identified in the *GLOI* gene, which was also reported to be associated with autism (p=0.0079). In addition to autism, *GLO1* has been linked to anxiety-like behavior (Hovatta et al. 2005; Kromer et al. 2005), diabetes (Thornalley 2003), Alzheimer˝s disease (Chen et al. 2004), and the regulation of theta oscillations during sleep (Tafti et al. 2003). Oxidative stress has also been implicated in neuropsychiatric disorders such as schizophrenia and depression (Bilici et al. 2001; Yao et al. 2001). The purpose of this study was to carry out a genetic analysis to see whether allelic variants of the *GLO1* gene predispose to autism or AS in the Finnish population.

We used the PSEUDOMARKER joint test of linkage and association in addition to linkage of six tag SNPs in *GLO1*, which were chosen to capture the majority of genetic variation in this gene. AS and autism families were analyzed separately and together. We did not observe statistically significant ($p<0.05$) linkage or association for any SNP in *GLO1* in the family based analyses (Table 15 and Table 1 in publication III). For a more detailed analysis of the functional SNP, one affected proband was randomly selected from each family. The frequency of the A allele, which was reported to be the risk allele in the original study (Junaid et al. 2004), was 0.62 for the combined autism and AS cases as well as for controls (Table 16) and the frequency of the AA genotype was 0.37 in both cases and controls.

The nsSNP is located in a block of high LD incorporating three other SNPs included in this study. To test if any specific haplotype of this block would confer susceptibility to autism, we tested whether haplotypes of these SNPs were associated with ASDs. However, no statistically significant association ($p<0.05$) was observed for any haplotype in any set of families using HBAT. The most common haplotype for these markers is A-G-G-G, with a frequency of approximately 0.5 in autism and AS families, as well as in controls.

The allele frequencies of the *GLO1* nsSNP have been studied in many populations, with only small variability across populations observed (Table 16). The frequency found in the controls of the Junaid et al. (2004) study is lower than in any other population. In addition, controls in that study were related: 33 of the controls were from only 6 extended families suggesting that the association arose from a low A-allele frequency in controls and not an increased frequency in cases. This

observation is in agreement with frequencies obtained in population controls and negative candidate studies of *GLO1*. Interestingly, in another study of *GLO1* in ASDs, no association was observed between markers in *GLO1* and ASD, and in addition, a protective effect of the A allele in siblings of probands was observed (Sacco et al. 2007). We were not able to test if a similar effect was seen in our sample, because with only 25 families having genotypes for unaffected siblings, power was too low to detect possible overtransmission.

The results of this study suggest that *GLO1* does not represent a susceptibility gene for ASDs in the Finnish autism families. We also provide evidence that the study by Junaid et al. (2004) overestimated the association between *GLO1* and ASDs due to a small number of controls that resulted in an underestimation of the A-allele frequency.

Table 15. Results of the family based association analysis of SNPs in GLO1 in 97 Finnish autism families, 29 Finnish AS families and the combined analysis of these two family materials.

SNP ID	AS, LC1		Autism, LC1		Combined	
	LD | linkage	LD + Linkage	LD | linkage	LD + Linkage	LD | linkage	LD + Linkage
rs7604	0.39	0.36	0.22	0.34	0.62	0.71
rs2736654	0.99	0.20	0.13	0.23	0.70	0.08
rs9394523	0.4	0.26	0.53	0.68	0.77	0.75
rs6932648	0.36	0.24	0.62	0.75	0.67	0.70
rs10484854	0.98	0.11	0.66	0.78	0.68	0.19
rs1937780	1.00	0.34	0.81	0.26	0.47	0.25

NOTE: All results are p-values from PSEUDOMARKER analysis.

Table 16. A allele frequencies of the nsSNP rs2736654 in different populations.

Population	ASD case		Control		Reference
	n	freq	n	freq	
Finland	120	0.62	740	0.62	This study
HapMap CEU	NA	NA	60	0.53	Frazer et al. (2007)
Italy	371	0.58	171	0.54	Sacco et al. (2007)
US	71	0.61	49	0.44	Junaid et al. (2004)
African American	NA	NA	24	0.65	dbSNP (Coriell Cell repository)

NOTE: ASD – autism spectrum disorder, NA – not available.

5.4 Population stratification study (IV)

In addition to enabling the identification of novel genetic determinants to complex disorders and traits, the genome-wide, high-density SNP data generated for GWASs also makes it possible to assess population structure in more detail than ever before. Most studies of Finnish population genetics have been performed using Y chromosomal or mitochondrial data that give only a limited picture of the genomic landscape. Here we wanted to explore, in high detail, the Finnish genome using an extensive dataset for the first time. Our aim was to characterize whether the Finnish population, which has traditionally been thought to have decreased genetic heterogeneity, would display stratification similar to other populations, and consequently, how this would affect the design of GWASs in the Finnish population.

We used all available Finnish GWAS data from both large population cohorts and smaller disease-specific cohorts. According to parental birthplace information samples representing different geographic regions were ascertained. The groups were chosen to represent geographical areas which have been inhabited at different stages during the population history (Table 17, Figure 11a). We had SNP genotype data generated using different Illumina 300K BeadChips, and after rigorous quality control, a set of approximately 230K SNPs of high quality were used in this study.

Table 17. Summary of populations used in the population stratification study (IV).

Group	N	% females	Chip type	Success rate	N SNPs with HWE p<0.001
SWE	302	100	318K duo	0.998	0
HEL	162	56	370K	0.999	166
ESS	73	41	318K duo, 370K	0.999	126
ESW1	179	53	370K, 318K duo	0.998	152
ESW2	145	65	370K, 318K duo	0.998	163
ESN	76	46	370K	0.998	115
LSW	48	50	317K, 318K duo, 370K	0.997	115
LSC	46	52	317K, 370K	0.998	69
LSN	107	48	370K	0.998	119
ISS	96	46	318K duo, 370K	0.998	135
ISC	78	45	318K duo, 370K	0.998	138
ISN	53	57	370K	0.998	94

NOTE: SWE – Sweden, HEL – Helsinki, ESS – Early Settlement South, ESW1 – Early Settlement West 1, ESW2 – Early Settlement West 2, ESN – Early Settlement North, LSW – Late Settlement West, LSC – Late Settlement Central, LSN – Late Settlement North, ISS – Isolate South, ISC – Isolate Central, ISN – Isolate North, 317K – Illumina HumanHap 300, 318K duo – Illumina HumanHap 300 duo, 370K – Illumina HumanCNV370-duo, HWE – Hardy Weinberg Equilibrium.

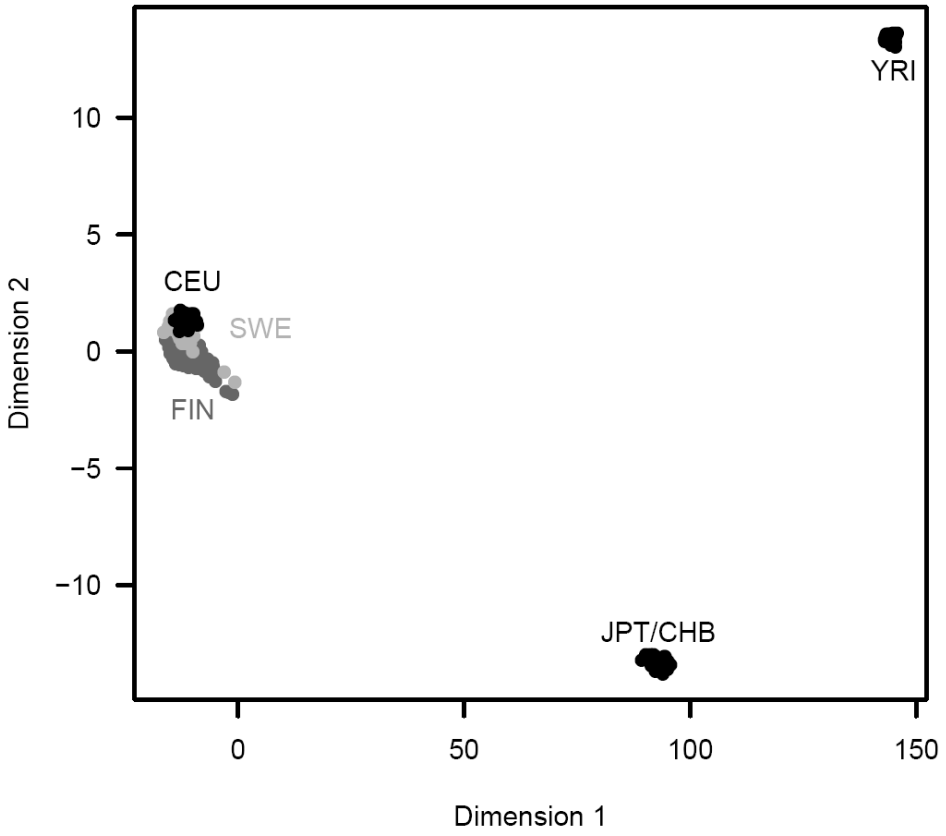


Figure 10. The first two dimensions of variation in Finns, Swedes and the HapMap populations from genome-wide SNP data. *CEU – CEPH individuals of European Ancestry, CHB – Han Chinese, FIN – All Finnish individuals included in this study, JPT – Japanese, SWE – Sweden, YRI – Yoruba.*

To obtain a measure of distance between samples, we calculated IBS sharing between all pairs of individuals and used MDS to extract the primary dimensions of the IBS sharing matrix. Finns cluster close to, but not fully overlapping with, the European HapMap population and Swedes, Finns also extend toward the HapMap Asian populations (HCB, JPT) possibly reflecting the eastern immigration during the original settling of the country (Figure 10). Finns have clearly been shown to cluster apart from other European populations (Lao et al. 2008).

When the HapMap African and Asian samples were removed, the two primary dimensions of the MDS correspond closely to east-west and north-south directions. The distribution of individuals closely resembles the geographic distribution according to grandparents‟ birthplaces (Figure 11). Particularly fine scale separation can be achieved in the youngest subisolates in northeastern Finland (Figure 12). This is in agreement with studies where similar approaches have been used to delineate the relationships among other populations (Heath et al. 2008; Price et al. 2008; Salmela et al. 2008), and in a limited region in western Finland (Saxena et al. 2007). However, results of a recent study proposed that gradients correlating with continuous geographical variation might be artifacts introduced by principal component analysis and should be interpreted with caution (Novembre and Stephens 2008).

We used the population differentiation test implemented in Eigensoft (Table 18) and estimated IBS sharing within and among groups to show that the subgroups differ significantly from each other. The variation among groups is mediated by small differences in allele frequencies across the entire genome, and not a small number of distinct loci (Figure 13). We estimated the fixation index (F_{st}), and in agreement with the MDS-analysis, the largest F_{st} was observed between the most eastern and western subgroups (ISC vs. ESN and ISC vs. LSW both resulting in F_{st}=0.006, Table 18). The F_{st}-value is significantly larger than those obtained in a study of the Icelandic population, which is also an isolate originating from a small number of founders, (Helgason et al. 2005). The F_{st}-value observed between eastern and western Finland is of similar magnitude to those observed between South-Western and North-Eastern Europeans (Price et al. 2008)

To illustrate the effect the observed population substructure would have on a case-control GWAS where cases were from an isolate and controls from the general Finnish population, we estimated the genomic inflation factor (λ) for each subisolate versus the Helsinki-population (Figure 14). All subisolates, except for the southern coastal “early settlement south” subset resulted in significant inflation (λ>1.05). As certain complex disorders, such as schizophrenia and MS (Hovatta et al. 1997; Sumelahti et al. 2001), are overrepresented in certain subisolates compared with the general Finnish population, the proper selection of controls based on careful genealogical data could be imperative in preventing spurious false positive association signals.

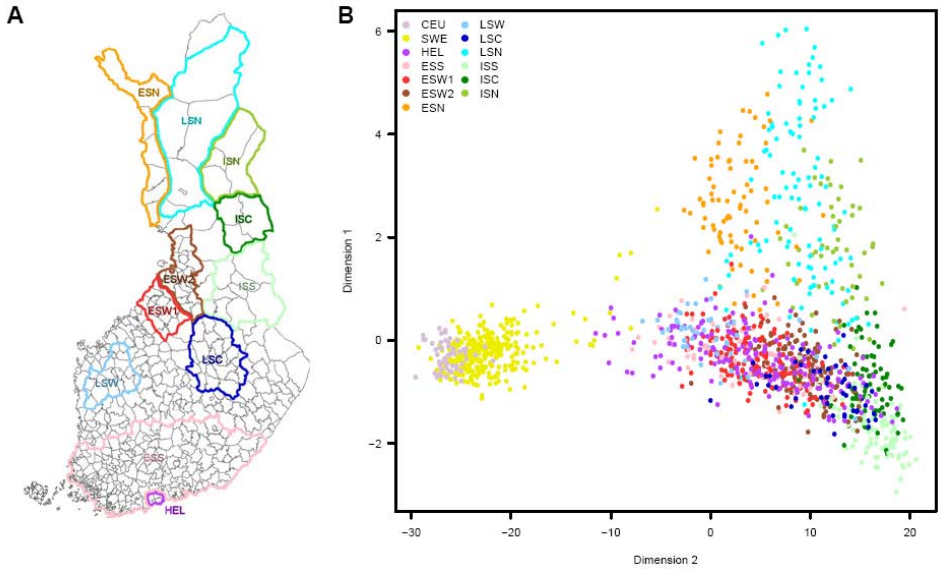


Figure 11. Finnish subpopulations. *A) Geographic distribution of subisolates representing different stages in the inhabitation of the country. B) The first two dimensions of variation in the Finnish subgroups compared to Swedes (SWE, yellow) and HapMap CEU (grey). Group abbreviations as in Table 17.*

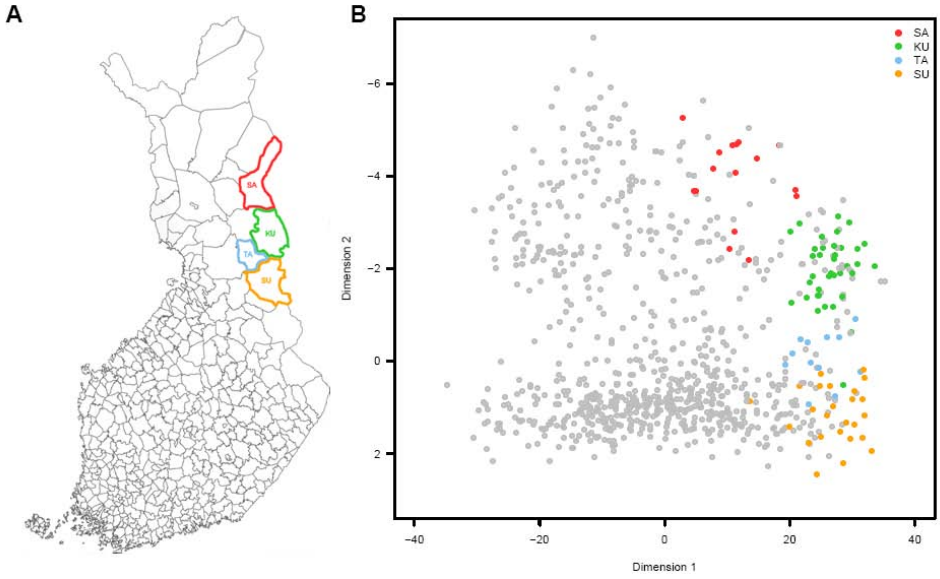


Figure 12. Fine-scale resolution of IBS-similarities in the youngest subisolates.

Table 18. F_{st} values and p-values for the test of population differentiation implemented in Eigensoft for the Finnish subpopopulations used in this study. *F_{st} values are indicated above the diagnonal, and approximate p-values from the population differentiation test, a summary of Anova statistics across the 10 first eigenvectors, are displayed below the diagnonal. Group abbreviations as in Table 17.*

	HEL	ESS	ESW1	ESW2	ESN	LSW	LSN	LSC	ISS	ISC	ISN
HEL		0.000	0.001	0.001	0.002	0.001	0.002	0.001	0.002	0.004	0.003
ESS	0.031		0.001	0.001	0.003	0.002	0.002	0.001	0.003	0.004	0.004
ESW1	10^{-142}	10^{-103}		0.001	0.003	0.002	0.003	0.001	0.003	0.004	0.004
ESW2	10^{-93}	10^{-93}	10^{-128}		0.003	0.002	0.002	0.001	0.002	0.003	0.003
ESN	10^{-197}	10^{-133}	10^{-273}	10^{-213}		0.003	0.002	0.004	0.005	0.006	0.005
LSW	10^{-61}	10^{-45}	10^{-116}	10^{-151}	10^{-121}		0.004	0.003	0.005	0.006	0.005
LSN	10^{-123}	10^{-83}	10^{-151}	10^{-89}	10^{-102}	10^{-101}		0.003	0.004	0.004	0.002
LSC	10^{-27}	10^{-41}	10^{-103}	10^{-57}	10^{-126}	10^{-86}	10^{-57}		0.002	0.003	0.004
ISS	10^{-168}	10^{-149}	10^{-257}	10^{-133}	10^{-207}	10^{-133}	10^{-131}	10^{-76}		0.003	0.004
ISC	10^{-217}	10^{-154}	10^{-283}	10^{-227}	10^{-161}	10^{-151}	10^{-147}	10^{-136}	10^{-173}		0.004
ISN	10^{-195}	10^{-132}	10^{-225}	10^{-159}	10^{-160}	10^{-131}	10^{-60}	10^{-104}	10^{-165}	10^{-146}	

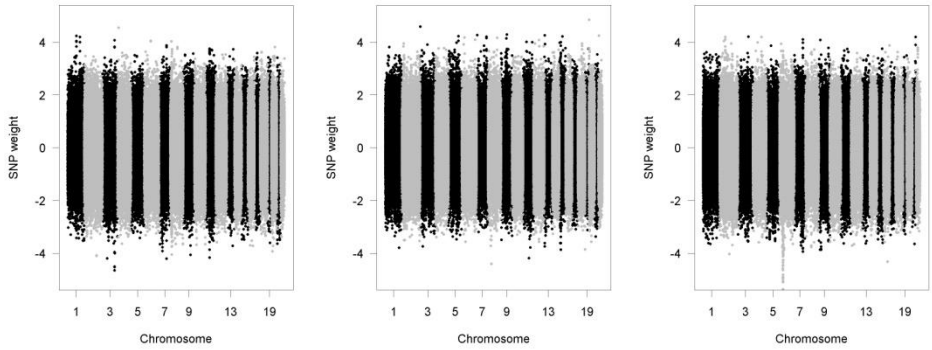


Figure 13. SNP weights for the first three eigenvectors*. All Finnish subpopulations were included in the analysis.*

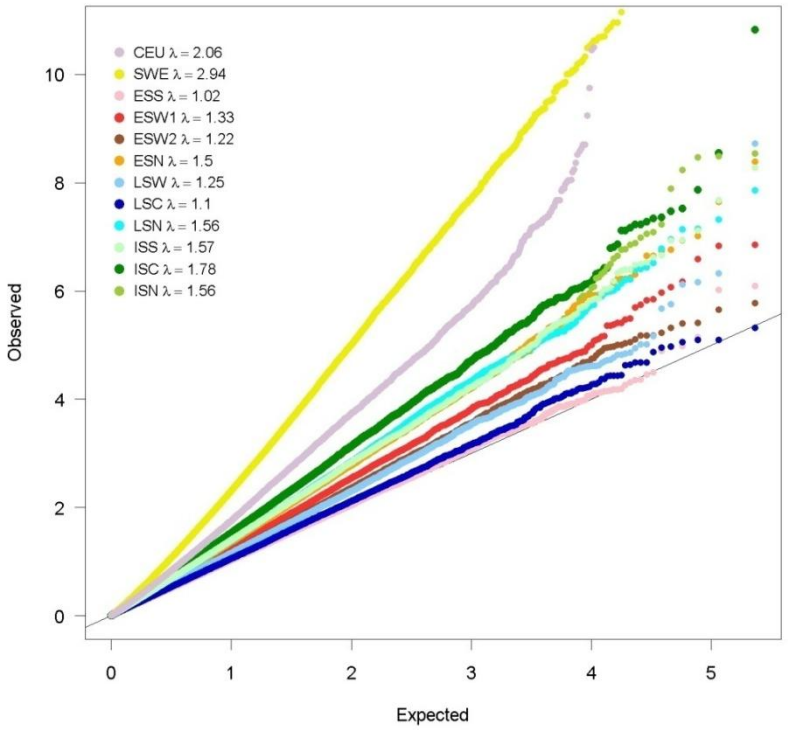


Figure 14. Quantile-quantile plot of the association analyses of each subisolate versus the Helsinki (HEL) group. *The genomic inflation factor (λ) for each subgroup is indicated in the legend. Group abbreviations as in Table 17.*

We used two approaches to characterize LD structure. Chromosome 22 was taken as an example as it has been used in a previous study of world-wide population isolates (Service et al. 2006). We estimated the pairwise LD for all pairs of SNPs in 70 SNP windows using the square correlation coefficient (r^2). The lowest number of SNP pairs in high LD was observed in the oldest populations, gradually increasing toward the younger populations (Figure 15a). To quantitatively describe the effects of recombination on LD over the whole chromosome, we constructed LD maps and again observed the longest maps, corresponding to least LD in the oldest populations with map lengths decreasing gradually towards the younger northeastern subisolates (Figure 15b). In addition, we observed that map lengths obtained in this study closely correlate with map lengths in another study where corresponding population groups were included (Service et al. 2006). We propose that LD map lengths can be considered characteristic for a given population.

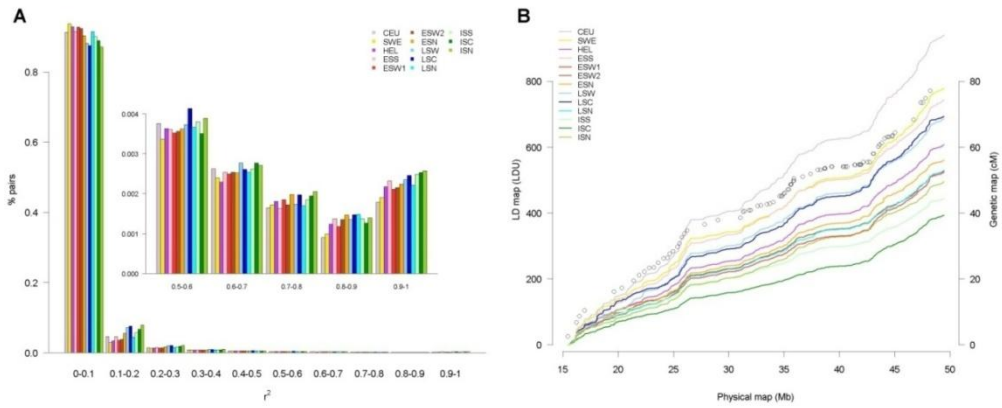


Figure 15. Properties of linkage disequilibrium across chromosome 22. *A) Proportion of SNP pairs in different LD bins for the different subgroups. B) A linkage disequilibrium map across chromosome 22. The genetic map is indicated by open circles, and is scaled according to the y-axis on the right-hand side of the figure. Group abbreviations as in Table 17.*

We also determined the lengths and numbers of extended regions of homozygosity (ROHs) exceeding 1 Mb and 100 SNPs with a SNP density of at least 1 SNP per 50 kb. Again, both the length and number of ROHs were lowest in the old subpopulations and highest in the younger populations (Figure 4, publication IV). This is probably due to autozygosity, reflecting the low number of founders and subtle increases in the relatedness of individuals from the younger subisolates. The effect was even more pronounced for ROHs exceeding 5Mb, observed in approximately 50% of individuals in the early settlement groups, 50-70% of individuals in the late settlement groups and up to 80-90% of individuals in ISS, ISC and ISN. Similar ROHs were observed in only 20% of Swedes, and 9.5% of individuals from the US (Simon-Sanchez et al. 2007). In the HapMap samples, 20% had ROHs exceeding 3Mb (Frazer et al. 2007). Although the lengths of homozygous segments are higher in Finns than in more outbred populations, they are significantly lower than in populations where consanguineous marriages are common. Individuals with Pakistani and Arab origins whose parents were first cousins have a mean genome homozygosity of 11% compared to an average of 1-2% in Finland (Woods et al. 2006). Inbreeding coefficients (F) were similar for all Finnish subgroups: only 1.65% of individuals corresponding to first cousin marriage (F=0.025–0.07), whereas 8.7% of individuals had F corresponding to second-cousin marriage (F=0.01–0.025), suggesting that the number and length of ROHs is a more sensitive measure of relatedness compared to F.

The results of this study show that on a genome-wide level, there are significant differences in the genomic landscapes of individuals originating from different parts of the country, as seen in other population isolates (Helgason et al. 2005; Yamaguchi-Kabata et al. 2008). Although this study was not designed to show that the genomic properties observed here are specifically the results of founder bottlenecks and drift, the results from several complementary analytic approaches agree remarkably well with the carefully documented population history of the country, which has been characterized by multiple founding bottlenecks and severe isolation. The groups in this study were selected using information on population history as well as linguistics and were shown to represent true genetically distinctive subgroups using analysis of eigenvectors and IBS similarities.

Although the population structure reported here would have significant effects on GWASs, proper correction of the substructure using quantitative measures such as the primary MDS dimensions has proven to be an efficient approach to tackle the genetic variability within Finland. In a study of metabolic traits in the NFBC66 cohort, from which a subset of samples in this study are drawn, the first two MDS dimensions were used to correct for population substructure (Sabatti et al. 2008).

The Sabatti et al. (2008) study replicated most previously reported associations for the traits tested and identified several interesting novel risk variants for high- and low density lipoprotein and insulin levels. However, p-values obtained in an alternative analysis not corrected for population substructure resulted in p-values of similar magnitude. It should be noted that we only included individuals with both parents born in the defined geographic regions and these individuals represent "extremes" of the genetic variability within Finland. Most individuals have parents born in different parts of the country, and thereby, display significantly less population stratification. Similar effects were reported by the WTCCC, where north-south gradients were observed for a subset of SNPs, but the authors concluded that the effect of population structure is negligible and did not correct for it in association analysis (WTCCC 2007).

Detailed knowledge of the genomic variability within Finland has already proven useful in identifying genetic determinants for complex disorders, as exemplified by the study of metabolic traits mentioned above. Another example is a GWASs aiming to identify susceptibility variants for MS. In these studies, cases from an internal isolate revealed two functionally relevant risk factors including *complement component 7* (Kallio et al. 2009).

5.5 GWA and expression study (V)

Genealogical data has revealed that a subset of the Finnish ASD families, currently living across Finland, can be linked to form two extended pedigrees originating from two municipalities in CF in the 18th century (Figure 4). We wanted to use these unique extended families to identify possible enriched rare genetic susceptibility factors shared IBD using dense, genome-wide SNP data. In addition, we included individuals diagnosed with autism, and at least 2 grandparents born in Central Finland that were not genealogically linked to the extended pedigrees. In study IV, we showed that individuals within subisolates in Finland are genetically homogenous, and that systematic genetic differences can be observed among subisolates.,Therefore we hypothesized that genetic risk factors in the CF autism families not linked to the pedigrees could be shared with the families in the extended pedigrees. The initial phase of the GWAS was followed up by replication in two datasets. Because genes for ASDs have eluded identification using traditional methods, we took a systems biology approach in order to combine the power of genome-wide genotype data with expression profiling so that biological pathways involved in the etiology of ASDs could be identified (Figure 16).

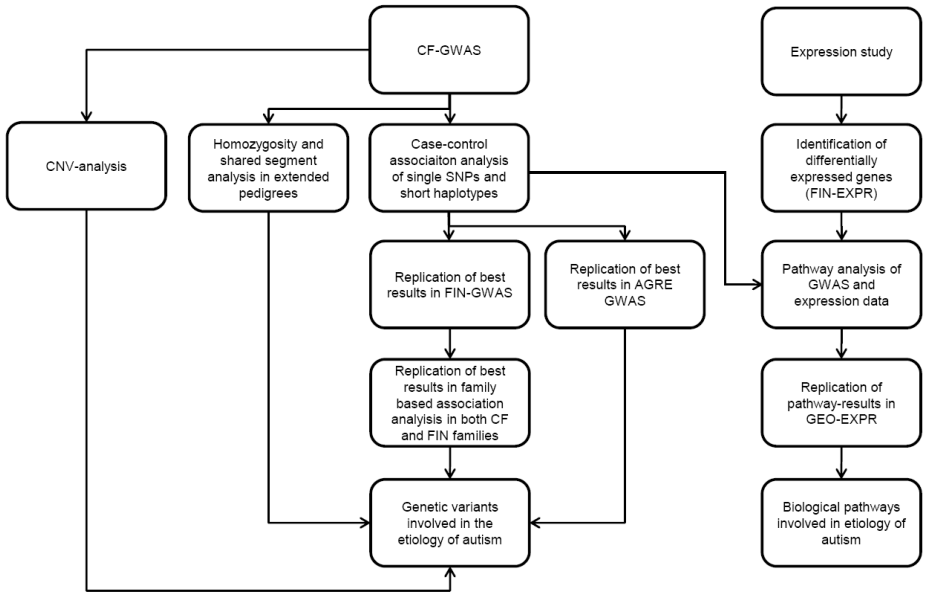


Figure 16. Outline for study V.

5.5.1 Quality control

We used Illumina HumanHap 300 and HumanHap 300 Duo beadchips to genotype approximately 317 000 SNPs in 51 cases and 181 controls (CF-GWAS). Of the cases, 19 were part of Autism-pedigree 1 and 8 were part of Autism pedigree 2. We excluded SNPs and individuals with a success rate below 90%. PLINK was used to estimate IBD sharing within the genealogically connected pedigrees and for the rest of families which were not genealogically linked but had at least two grandparents born in Central Finland (Figure 17). Higher IBD sharing was observed within pedigree 2 compared to pedigree 1 (mean IBD sharing 0.0057 and 0.0036, respectively), in agreement with the more distant genealogical links in pedigree 1. Surprisingly, IBD sharing was higher between the non-genealogically linked families than within the families in pedigree 1 (mean IBD sharing 0.0046), and as expected IBD sharing was lowest in controls (mean IBD sharing 0.0024).

5.5.2 Homozygosity mapping

First we analyzed the extended pedigrees for shared ROHs to monitor possible recessive susceptibility variants. ROHs exceeding 100kb were identified and assessed for overlap between individuals. A minimum overlap cutoff of 50 kb was used. Only ROHs shared among more than half of the affected individuals in each pedigree(n=5 and n=10 for pedigrees 1 and 2 respectively) were used providing that the individuals were homozygous for the same haplotype. We used liberal cutoffs for both segment length and frequency to avoid missing regions of interest. To rule

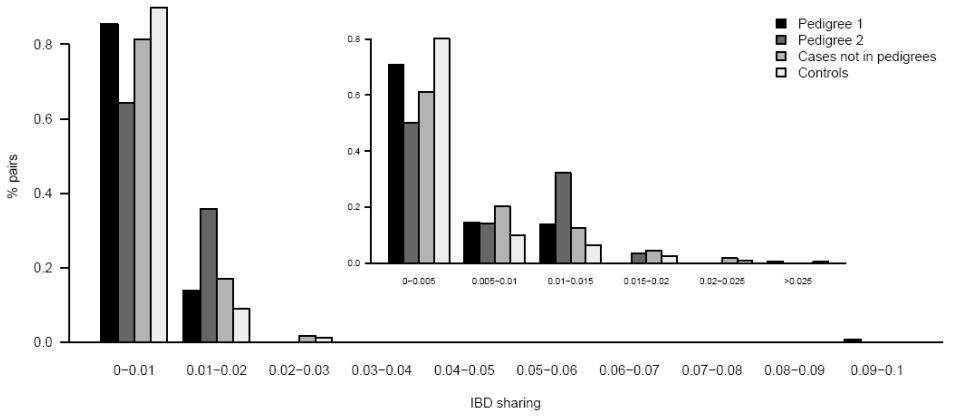


Figure 17. IBD sharing in samples included in the CF-GWAS. *The inset shows a more detailed view of the distribution of individuals sharing up to 2.5% of their genome IBD.*

out false positive findings due to the liberal cut-off criteria, we assessed the regions of interest in the 180 regionally matched controls as well.

In the smaller pedigree, we did not identify any ROHs where 6 or more individuals would be homozygous for the same haplotype at any locus. We identified 5 shared ROHs exceeding 50kb where 5 individuals were homozygous for the same haplotype. In the larger pedigree we identified three regions were 11-18 individuals were homozygous for the same haplotype (Table 19). However, the ROHs identified in cases in both pedigrees were all common in controls as well (frequency 0.3-0.81), suggesting that none of the regions contain high-penetrance risk variants. Homozygosity analysis did not implicate any loci harboring enriched, recessive haplotypes conferring risk for ASDs in either extended pedigree.

5.5.3 Shared segment analysis

To identify shared, enriched risk variants inherited in a dominant-like fashion, we used the shared segment analysis implemented in PLINK (Purcell et al. 2007). The underlying hidden IBD state is estimated by using the observed IBS sharing across chromosomes by a Hidden Markov Model. This differs from traditional haplotype estimation of sharing for rare and long regions, even for pairs of distantly related

Table 19. Regions of homozygosity identified in the extended pedigrees.

Sample	N_{case}	$Freq_{case}$	N_{ctrl}	$Freq_{ctrl}$	Chr	From-kb	To-kb	Size	N_{SNP}
Ped2	10(18)*	0.53(0.95)	137	0.76	4	33720	34094	373	23
Ped1	5	0.63	74	0.41	6	28629	28738	109	6
Ped1	5	0.63	91	0.50	7	117503	117943	440	48
Ped1	5	0.63	57	0.32	8	51513	52109	596	38
Ped2	11	0.58	53	0.30	11	48105	48573	467	29
Ped1	5	0.63	88	0.49	15	69889	70501	611	47
Ped2	11	0.58	88	0.49	15	69899	70794	894	57
Ped1	5(7)*	0.63(0.88)	146	0.81	18	64802	64929	126	37

NOTE: Ped1 – Autism pedigree 1, Ped2 – Autism pedigree 2, N_{case} – number of cases homozygous in the region, $Freq_{case}$ – Frequency of homozygous region in cases, N_{ctrl} – number of controls homozygous in the region, $Freq_{ctrl}$ – Frequency of homozygous region in controls, Chr – Chromosome, From-kb – Start (in kb from ptel) of overlapping homozygous region, To-kb – End (in kb from ptel) of shared homozygous region, Size – Size of shared homozygous region in kb, N_{SNP} – Number of SNPs in the shared homozygous region. *The number of individuals in parenthesis include individuals with an homozygous segment <100kb in the region.

individuals. For the shared segment analysis, only SNPs in linkage equilibrium are used, as regions with high LD result in false positive calls of IBD sharing.

For shared segment analysis, all SNPs with a genotyping rate <0.99 and a minor allele frequency (MAF)<0.05 were filtered out. SNPs were pruned by estimating LD in 100 SNP windows, excluding all SNPs with r^2>0.2. This resulted in a dataset consisting of roughly 56 000 SNPs. Segments larger than 1000 kb were included.

We required IBD sharing between more than half of the individuals in each pedigree, and that the same alleles would be shared between all pairs in the region, as this would suggest inheritance IBD from a common ancestor in the risk region. No regions of interest were identified using this approach. Regions where more than three pairs in Pedigree 1 and two pairs in Pedigree 2 shared chromosomal segments IBD are indicated in Table 20. However, in none of these regions did more than three individuals share the same allelic combination IBD.

To ensure we did not miss any shared regions we also performed the same test for shorter regions, requiring that the shared segments only span 100kb and 20 SNPs. This did not result in any further regions of interest. In summary, this method did not identify any shared segments where genetic risk factors enriched in these pedigrees would reside.

5.5.4 Association analysis

We did not identify any enriched rare risk variants in homozygosity or shared segment analysis. Therefore we wanted to test for the presence of common susceptibility variants in the entire CF dataset. We used all 54 affected individuals and 181 controls in order to identify possible common variants enriched in this subisolate. We identified a slight excess of significant p-values suggesting that common shared susceptibility variants do exist, but unsurprisingly, our limited dataset, designed to identify rare variants did not have sufficient power to detect these shared common variants at a genome-wide significant level (Figure 1, publication V). A total of 7 p-values less than 0.00001 were observed, compared to the 3 expected, and 37 p-values below 0.0001 were observed compared to 31 expected. There was not significant inflation in the overall distribution of p-values (λ = 1.05). One marker, rs9309326 reached genome-wide significance ($p=8.00 \cdot 10^{-09}$). The minor allele was significantly enriched in cases (MAF=0.28) compared to controls (MAF=0.07). The SNP is located at 2p16.1, 60 kb from the start of *BCL11A*. All seven SNPs resulting in p<0.00001 showed much higher minor allele frequencies in cases compared to controls, thereby suggesting the enrichment of rare variants (Table 21).

Table 20. Segments shared IBD between at least four pairs of individuals in Autism pedigree 1 or three pairs of individuals in Autism pedigree 2.

PED	ID1	ID2	CHR	FROM-BP	TO-BP	KB	NSNP	GRP
PED1	152	110	2	124193101	129117426	4924.32	75	1
PED1	152	138	2	123182056	130174939	6992.88	98	1
PED1	110	138	2	120913571	129108475	8194.9	131	1
PED1	150	113	2	122832236	130174939	7342.7	107	2
PED1	SHARED	SHARED	2	124193101	129108475	4915.37	74	NA
PED1	5	25	4	172932609	175514110	2581.5	54	1
PED1	5	138	4	172561090	177328037	4766.95	93	1
PED1	152	113	4	166392779	177120896	10728.1	202	2
PED1	150	182	4	173090726	177086340	3995.61	86	3
PED1	SHARED	SHARED	4	173090726	175514110	2423.38	53	NA
PED1	115	138	7	108116795	117023250	8906.45	108	1
PED1	152	115	7	106790674	114162055	7371.38	94	1
PED1	110	112	7	105545460	126072141	20526.7	214	2
PED1	30	129	7	106170809	114393860	8223.05	99	3
PED1	SHARED	SHARED	7	108116795	114162055	6045.26	72	NA
PED2	209	218	9	68190904	100219712	32028.8	649	1
PED2	209	1	9	68190904	72418000	4227.1	68	1
PED2	204	3	9	68190904	75855194	7664.29	141	2
PED2	SHARED	SHARED	9	68190904	72418000	4227.1	68	NA
PED1	150	5	10	226452	2419307	2192.86	61	1
PED1	150	182	10	226452	2079982	1853.53	50	1
PED1	110	112	10	226452	10228122	10001.7	335	2
PED1	152	138	10	1433016	2190634	757.618	39	3
PED1	SHARED	SHARED	10	1433016	2079982	646.966	34	NA
PED2	1	3	16	62312584	71875058	9562.47	101	1
PED2	132	1	16	69105798	74396097	5290.3	96	2
PED2	218	3	16	71026964	73066203	2039.24	60	3
PED2	SHARED	SHARED	16	71026964	71875058	848.094	21	NA

NOTE: PED – PEDIGREE, ID1 – ID of first individual, ID2 – ID of second individual, CHR – Chromosome, FROM-BP – Start of shared segment, TO-BP – End of shared segment, KB – length of shared segment in kb, NSNP – Number of SNPs in shared segment, GRP – Pairs sharing the same homozygous haplotype IBD, SHARED – the regions shared between all pairs of individuals.

Table 21. SNPs resulting in p<0.00001 in single-SNP association analysis in CF-GWAS.

CHR	SNP	POSITION	F_A	F_U	CHISQ	P	OR
2	rs9309326	60478443	0.2843	0.07418	33.28	8.00E-09	4.958
2	rs4076769	192292451	0.4314	0.1951	23.94	9.95E-07	3.131
5	rs1863984	4264882	0.1569	0.03297	21.66	3.26E-06	5.457
5	rs1077606	32135698	0.1373	0.02198	23.54	1.23E-06	7.08
7	rs4724793	6281541	0.2857	0.08659	27.08	1.95E-07	4.219
7	rs12701440	35617230	0.3235	0.1291	21.18	4.19E-06	3.226
11	rs10501399	68979039	0.1373	0.02747	19.66	9.27E-06	5.632

NOTE: CHR – Chromosome, F_A – Minor allele frequency in cases, F_U – Minor allele frequency in controls, CHISQ – Chi-square test statistic, P – p-value, OR – Odds ratio.

5.5.5 Haplotype analysis

To increase our power to detect shared common risk variants, we performed an association analysis using 5 SNP haplotypes in sliding windows across all chromosomes. Only two regions resulted in $p<1\cdot10^{-5}$ in the global test. The first region is located at 2p16.1 and overlaps with the best association result from the pointwise SNP analysis. Five consecutive 5-SNP haplotypes resulted in $2.02\cdot10^{-6}<p<8.24\cdot10^{-6}$. One specific haplotype had a frequency of approximately 27% in cases and 7% in controls suggesting that the haplotype association result is primarily driven by rs9309326. The number of haplotypes, indicated by the degrees of freedom of the test statistic, suggests that the associated haplotype is located between two recombination hotspots (Figure 18). The other region highlighted by haplotype analysis was 2q33, covering parts of *ALS2CR11* and *ALS2CR4*. In the region, four consecutive 5-SNP haplotypes resulted in $2.95\cdot10^{-7}<p<5.24\cdot10^{-6}$. The SNPs in the region are in high LD in the HapMap CEU data, and the haplotype in this study set ends in a possible recombination hotspot indicated by a region of low LD and a high number of different haplotypes (Figure 19). The best single-SNP p-value in the region was observed with rs12464623 ($p=3.26\cdot10^{-5}$), resulting from an enrichment of the minor allele in cases (frequency 0.14 in cases versus 0.03 in controls). It should be noted that none of the haplotype findings remain significant after correction for multiple testing.

Analysis of single haplotypes revealed four additional loci of interest. Single haplotypes resulting in $p<10^{-6}$ were identified on 2q24.3, 2q32.3, 7p22.1 and 20q13.2 (Table 22). However, global p-values in these regions all exceeded $p=1\cdot10^{-5}$.

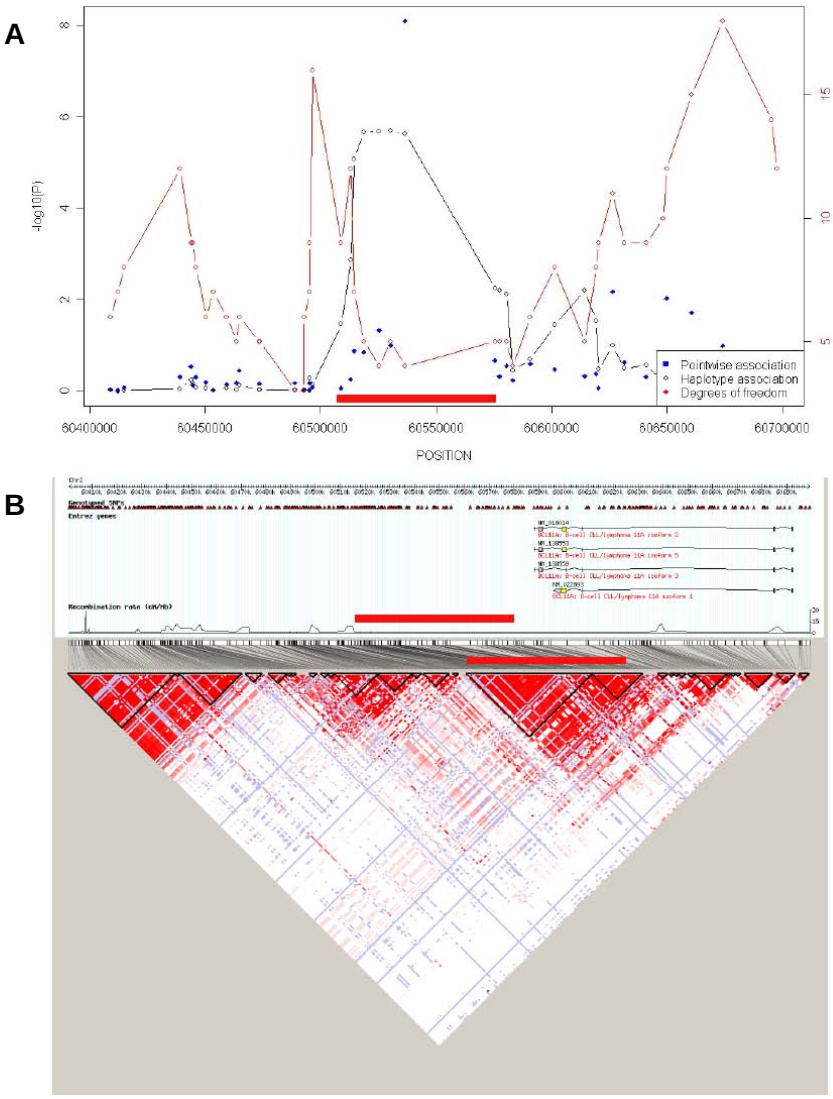


Figure 18. A region on 2p16.1 showing association to autism both in single SNP and haplotype association analyses. *A) Results of association analyses for both single SNP and 5 SNP sliding window haplotype tests. The red lines and the y-axis on the right hand side indicate the degrees of freedom for the haplotype association test, which correspond to the number of haplotypes. B) The LD structure of the region in the HapMap CEU population. Red bars indicate the corresponding regions in panels A and B.*

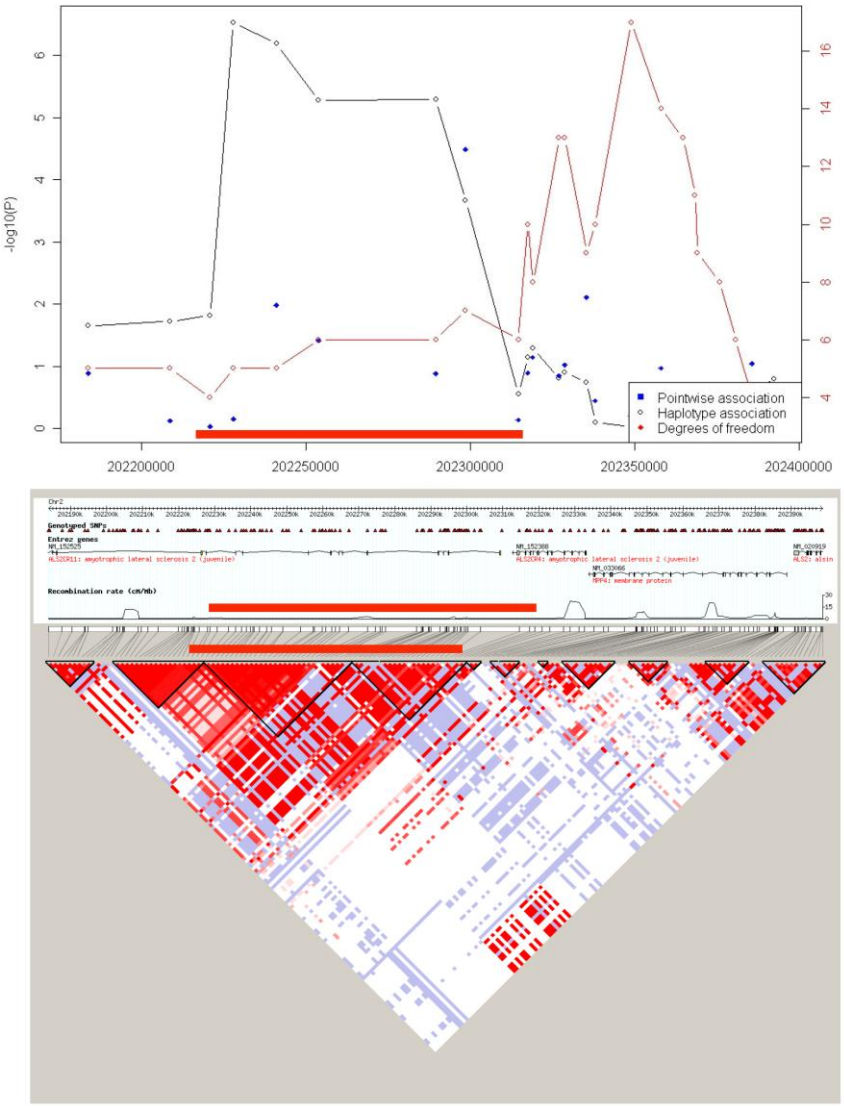


Figure 19. A region on 2q33 showing association to autism both in single SNP and haplotype association analysis. *A) Results of association analyses for both single SNP and 5 SNP sliding window haplotype tests. The red lines and the y-axis on the right hand side indicate the degrees of freedom for the haplotype association test, which correspond to the number of haplotypes. B) The LD structure of the region in the HapMap CEU population. Red bars indicate the corresponding regions in panels A and B.*

Table 22. Results of the haplotype association analysis in the CF-GWAS.

LOCUS	HAPLOTYPE	F_A	F_U	P
rs243034|rs243027|rs243081|rs243073|rs9309326				**8.2E-06**
2p16.1	*GAGGA*	*0.07*	*0.02*	*0.012*
2p16.1	*AAGGA*	*0.22*	*0.06*	***4.5E-07***
2p16.1	*GAGAG*	*0.18*	*0.27*	*0.049*
2p16.1	*AAGAG*	*0.10*	*0.09*	*0.775*
2p16.1	*GAGGG*	*0.00*	*0.01*	*0.243*
2p16.1	*GCAGG*	*0.06*	*0.07*	*0.547*
2p16.1	*ACAGG*	*0.35*	*0.41*	*0.221*
2p16.1	*AAAGG*	*0.04*	*0.06*	*0.317*
rs243027|rs243081|rs243073|rs9309326|rs8179712				**2.2E-06**
2p16.1	*AGGGA*	*0.00*	*0.01*	*0.233*
2p16.1	*CAGGA*	*0.06*	*0.08*	*0.648*
2p16.1	*AAGGA*	*0.04*	*0.07*	*0.412*
2p16.1	*AGGAG*	*0.28*	*0.07*	***8.7E-09***
2p16.1	*AGAGG*	*0.27*	*0.36*	*0.104*
2p16.1	*CAGGG*	*0.33*	*0.41*	*0.175*
rs243081|rs243073|rs9309326|rs8179712|rs1011407				**2.1E-06**
2p16.1	*AGGGG*	*0.10*	*0.08*	*0.643*
2p16.1	*AGGAA*	*0.11*	*0.14*	*0.354*
2p16.1	*GGAGA*	*0.28*	*0.07*	***3.2E-08***
2p16.1	*GAGGA*	*0.28*	*0.37*	*0.102*
2p16.1	*AGGGA*	*0.24*	*0.33*	*0.075*
rs243073|rs9309326|rs8179712|rs1011407|rs10490071				**2.0E-06**
2p16.1	*GGGAA*	*0.25*	*0.30*	*0.341*
2p16.1	*GGGGG*	*0.09*	*0.09*	*0.818*
2p16.1	*GGAAG*	*0.11*	*0.16*	*0.236*
2p16.1	*GAGAG*	*0.27*	*0.07*	***1.8E-08***
2p16.1	*AGGAG*	*0.28*	*0.36*	*0.116*
2p16.1	*GGGAG*	*0.00*	*0.03*	*0.095*
rs9309326|rs8179712|rs1011407|rs10490071|rs1012585				**2.3E-06**
2p16.1	*GGGGA*	*0.09*	*0.08*	*0.814*
2p16.1	*GAAGA*	*0.11*	*0.16*	*0.241*
2p16.1	*GGAAG*	*0.25*	*0.30*	*0.318*
2p16.1	*AGAGG*	*0.27*	*0.07*	***3.1E-08***
2p16.1	*GGAGG*	*0.28*	*0.39*	*0.049*
rs7424428|rs10497238|rs4668422|rs10199735|rs16848651				4.13E-05
2q24.3	*AAAAA*	*0.07*	*0.10*	*0.305*
2q24.3	*GCAAA*	*0.01*	*0.02*	*0.627*
2q24.3	*GCGAG*	*0.04*	*0.07*	*0.270*
2q24.3	*AAAAG*	*0.05*	*0.06*	*0.587*
2q24.3	*ACAAG*	*0.08*	*0.12*	*0.250*
2q24.3	*AAAGG*	*0.10*	*0.01*	***2.4E-07***
2q24.3	*GCAGG*	*0.66*	*0.62*	*0.546*
rs4076769|rs6434503|rs4241279|rs7598944|rs6718527				1.2E-04
2q32.3	*GGAGG*	*0.04*	*0.04*	*0.955*
2q32.3	*GAAGG*	*0.02*	*0.06*	*0.099*
2q32.3	*GAGGG*	*0.01*	*0.02*	*0.519*
2q32.3	*GAGGA*	*0.06*	*0.06*	*0.970*
2q32.3	*GGGAA*	*0.00*	*0.03*	*0.094*
2q32.3	*AAGAA*	*0.44*	*0.20*	***9.2E-07***
2q32.3	*GAGAA*	*0.43*	*0.59*	*0.004*

LOCUS	HAPLOTYPE	F_A	F_U	P
rs10931953|rs4675170|rs4675195|rs10931963|rs12464623				**3.0E-07**
2q33	*GGGAA*	*0.00*	*0.02*	*0.134*
2q33	*GAGGA*	*0.08*	*0.01*	*8.4E-06*
2q33	*AGAGA*	*0.03*	*0.00*	*0.007*
2q33	*GAGGG*	*0.53*	*0.44*	*0.136*
2q33	*AGAGG*	*0.25*	*0.25*	*0.941*
2q33	*GGAGG*	*0.11*	*0.27*	*0.001*
rs4675170|rs4675195|rs10931963|rs12464623|rs17384203				**6.5E-07**
2q33	*AGGGG*	*0.07*	*0.07*	*0.995*
2q33	*GGAAA*	*0.00*	*0.02*	*0.132*
2q33	*AGGAA*	*0.07*	*0.00*	*2.9E-05*
2q33	*GAGAA*	*0.06*	*0.01*	*5.7E-04*
2q33	*AGGGA*	*0.45*	*0.37*	*0.175*
2q33	*GAGGA*	*0.35*	*0.52*	*0.003*
rs4675195|rs10931963|rs12464623|rs17384203|rs2540450				**5.2E-06**
2q33	*GGAAA*	*0.06*	*0.00*	*1.6E-04*
2q33	*GGGAA*	*0.36*	*0.32*	*0.463*
2q33	*GGGGG*	*0.07*	*0.07*	*0.917*
2q33	*GAAAG*	*0.00*	*0.02*	*0.133*
2q33	*AGAAG*	*0.07*	*0.01*	*4.7E-04*
2q33	*GGGAG*	*0.09*	*0.06*	*0.197*
2q33	*AGGAG*	*0.36*	*0.52*	*0.004*
rs10931963|rs12464623|rs17384203|rs2540450|rs1914261				**5.2E-06**
2q33	*GAAGA*	*0.07*	*0.01*	*3.0E-04*
2q33	*GGAGA*	*0.34*	*0.50*	*0.006*
2q33	*GAAAC*	*0.06*	*0.00*	*1.3E-04*
2q33	*GGAAC*	*0.36*	*0.33*	*0.537*
2q33	*GGGGC*	*0.07*	*0.07*	*0.963*
2q33	*AAAGC*	*0.00*	*0.02*	*0.134*
2q33	*GGAGC*	*0.10*	*0.08*	*0.340*
rs10262194|rs1559560|rs10085518|rs7790267|rs4724793				5.0E-05
7p22.1	*GGAGA*	*0.02*	*0.01*	*0.448*
7p22.1	*AGGGA*	*0.24*	*0.07*	***9.4E-07***
7p22.1	*AGGAG*	*0.05*	*0.06*	*0.882*
7p22.1	*GGAGG*	*0.05*	*0.09*	*0.233*
7p22.1	*AGAGG*	*0.05*	*0.03*	*0.279*
7p22.1	*GGGGG*	*0.00*	*0.04*	*0.074*
7p22.1	*AGGGG*	*0.59*	*0.72*	*0.016*
rs7267819|rs6097659|rs6022839|rs2904376|rs6068717				1.0E-04
20q13.2	*AGGAA*	*0.04*	*0.05*	*0.640*
20q13.2	*GGGAA*	*0.14*	*0.13*	*0.807*
20q13.2	*GGGGA*	*0.01*	*0.02*	*0.379*
20q13.2	*GGGAC*	*0.16*	*0.03*	***1.7E-07***
20q13.2	*GAAGC*	*0.04*	*0.05*	*0.724*
20q13.2	*GAGGC*	*0.05*	*0.03*	*0.365*
20q13.2	*AGGGC*	*0.32*	*0.41*	*0.139*
20q13.2	*GGGGC*	*0.24*	*0.29*	*0.313*

Note: For each locus, the first line indicates the SNPs included in the haplotype, and a p-value for the global association test for that haplotype. Indicated in italics are the frequencies of each haplotype allele and p-value for association tests for specific haplotypes. F_A – Frequency in cases, F_U – Frequency in controls, P – p-value. Indicated in bold are haplotypes resulting in $p<1\cdot10^{-5}$ in the global test or $p<1\cdot10^{-6}$ for specific haplotypes.

5.5.6 Replication of CF-GWAS in two datasets

We selected the best SNPs from the single SNP and haplotype association analyses for replication in two datasets. Our first replication set consisted of one affected individual from the Finnish autism families who did not have genealogical roots in Central Finland (n=76, FIN-GWAS) along with 271 matched controls. The second control dataset was the large AGRE family set from the admixed US population, whose genotypes are publicly available from AGRE (AGRE-GWAS). We also genotyped all available family members of the individuals included in the CF-GWAS and the Finnish replication set in order to perform a family-based association study, considering some families had more than one affected individual. We selected all SNPs resulting in $p<1\cdot10^{-4}$ in single SNP association analysis, haplotypes resulting in $p<1\cdot10^{-5}$ in a global test, and specific haplotypes resulting in $p<1\cdot10^{-6}$ in the CF-GWAS for replication and family-based analysis. A total of 70 SNPs passed these criteria.

Case-control analysis for the Finnish replication set was performed by logistic regression. The position of each individual on the first three dimensions of MDS extracted IBS sharing data was used as covariates in the analysis to control for the population structure that was evident in the dataset. The TDT-test was used to test for association in the AGRE dataset as genotypes for parents were available. PLINK was used for all analysis. None of the SNPs resulting in $p<10^{-4}$ in the CF-GWAS resulted in suggestive association in either of the replication study sets (Table 23). However, the Finnish replication dataset does not provide sufficient power to detect variants with small effects on the phenotype. As family-based association studies provide more power to detect association, we used PSEUDOMARKER to perform a family-based, nonparametric association analysis using all available family members and controls to further assess the role of these variants in autism. In the family-based association analysis, suggestive association was observed in the replication and combined Finnish datasets for a subset of SNPs (Table 24). Three SNPs, rs2302561 rs17053011and rs6935315 resulted in $p<0.05$ in both Finnish datasets.

Table 23. Results for the replication of best single-SNP results from the CF-GWAS.

				CF-GWAS			FIN-GWAS		AGRE-GWAS		
SNP	**CHR**	**POSITION**	**GENE**	**F_A**	**F_U**	**P**	**OR**	**P**	**T**	**U**	**P**
rs6696977	1	41128813		0.31	0.54	4.0E-05	1.19	0.35	264	246	0.43
rs7539694	1	48822262	*AGBL4*	0.25	0.09	4.7E-05	1.15	0.61	148	139	0.60
rs10911628	1	182916126		0.16	0.04	2.8E-05	NA	NA	NA	NA	NA
rs1354369	1	193082712		0.45	0.25	6.3E-05	1.18	0.47	214	205	0.66
rs4658971	1	230174437	*DISC1*	0.10	0.01	2.0E-05	0.78	0.77	59	63	0.72
rs12472764	2	338470		0.30	0.53	4.3E-05	0.32	0.75	268	290	0.35
rs9309326	2	60478443		0.28	0.07	8.0E-09	0.99	0.97	165	NA	NA
rs2302561	2	71075330	*TEX261*	0.43	0.22	6.9E-05	0.73	0.19	232	274	0.06
rs4076769	2	192292451		0.43	0.20	1.0E-06	0.84	0.42	126	136	0.54
rs4483984	2	192366430		0.51	0.29	4.8E-05	0.97	0.90	235	265	0.18
rs12464623	2	202181323	*ALS2CR11*	0.14	0.03	3.3E-05	NA	NA	46	60	0.17
rs3819340	3	150097977		0.08	0.01	5.3E-05	1.58	0.29	48	60	0.25
rs2968671	4	5025571		0.24	0.09	8.1E-05	0.84	0.58	163	157	0.74
rs1471928	4	140319218		0.37	0.18	7.1E-05	1.15	0.55	178	165	0.48
rs1863984	5	4264882		0.16	0.03	3.3E-06	0.95	0.88	104	86	0.19
rs10066063	5	32126051	*PDZD2*	0.11	0.02	1.5E-05	0.59	0.36	99	88	0.42
rs1077606	5	32135448	*PDZD2*	0.14	0.02	1.2E-06	0.83	0.69	111	89	0.12
rs17053011	5	55293694	*SGCD*	0.39	0.20	8.8E-05	1.36	0.17	133	118	0.34
rs1422859	5	166021965		0.35	0.17	4.6E-05	1.10	0.71	126	104	0.15
rs2988681	6	87016654		0.12	0.31	9.2E-05	0.76	0.23	188	177	0.56
rs6935315	6	99459993	*FBXL4*	0.18	0.40	6.5E-05	1.15	0.49	231	248	0.44
rs4724793	7	6281291		0.29	0.09	2.0E-07	NA	NA	73	75	0.87
rs12701440	7	35617230		0.32	0.13	4.2E-06	1.19	0.47	198	164	0.07
rs4717992	7	63219204		0.19	0.41	4.6E-05	1.34	0.15	272	262	0.67

				CF-GWAS			FIN-GWAS		AGRE-GWAS		
SNP	**CHR**	**POSITION**	**GENE**	**F_A**	**F_U**	**P**	**OR**	**P**	**T**	**U**	**P**
rs7009615	8	18617115	*PSD3*	0.24	0.09	7.3E-05	1.33	0.31	114	113	0.95
rs10501399	11	68979039		0.14	0.03	9.3E-06	0.47	0.10	87	81	0.64
rs2564580	12	30288653		0.13	0.03	2.1E-05	1.34	0.58	NA	NA	NA
rs12229563	12	89984555		0.21	0.07	3.8E-05	NA	NA	67	56	0.32
rs9576952	13	39626828		0.25	0.09	2.9E-05	0.86	0.59	149	118	0.06
rs4587895	14	27532119		0.16	0.04	1.7E-05	NA	NA	NA	NA	NA
rs10484187	14	46802461	*MDGA2*	0.13	0.02	1.2E-05	0.68	0.38	92	69	0.07
rs1959033	14	63405339	*SYNE2*	0.16	0.04	6.7E-05	0.92	0.84	71	84	0.30
rs16967173	16	16003127	*ABCC1*	0.16	0.04	7.8E-05	0.13	0.72	84	63	0.08
rs2042005	17	5674282		0.51	0.30	9.2E-05	1.05	0.83	287	284	0.90
rs11661310	18	27781882		0.22	0.08	6.3E-05	0.84	0.60	63	56	0.52
rs5906714	X	48569590		0.46	0.20	3.9E-05	0.98	0.94	87	81	0.64
rs952422	X	50660814		0.25	0.07	3.8E-05	0.97	0.93	54	44	0.31

NOTE : SNPs resulting in $p<1 \cdot 10^{-4}$ in single-SNP association analysis in the CF-GWAS were included. Analysis was performed using PLINK. Case-control association analysis was performed in the CF-GWAS and Finnish replication study sets, whereas TDT analysis was performed in the AGRE data set. Position is given according NCBI build 36.1. CHR – Chromosome, F_A – Allele frequency in cases, F_U – Allele frequency in controls, P – p-value, T –transmitted minor allele count, U – untransmitted minor allele count. p-values <0.05 are indicated in bold.

Table 24. Results for the linkage and family-based association analysis of replication SNPs in Finnish families.

					CF-GWAS FAMILIES		FIN-GWAS FAMILIES		COMBINED	
SNP	**CHR**	**BP**	**GENE**	**REP**	**Linkage**	**LD|Linkage**	**Linkage**	**LD|Linkage**	**Linkage**	**LD|Linkage**
rs6696977	1	41128813		SS	0.321	**0.015**	0.065	0.904	0.082	**0.043**
rs7539694	1	48822262	*AGBL4*	SS	0.470	**1.2E-04**	0.488	0.655	0.470	**0.037**
rs1354369	1	182916126		SS	0.306	**0.001**	0.477	0.349	0.470	**0.001**
rs4658971	1	230174437	*DISC1*	SS	0.418	0.155	0.349	**0.014**	0.319	**0.002**
rs12472764	2	338470		SS	0.288	**0.003**	0.040	0.428	0.064	**0.014**
rs243034	2	60456396		GH, SH	0.059	**0.048**	0.180	0.553	**0.037**	0.111
rs243073	2	60472194		GH, SH	0.153	**0.050**	0.446	0.922	0.263	0.474
rs9309326	2	60478443		SS	0.335	**8.1E-08**	0.500	0.173	0.500	**0.026**
rs2302561	2	71075330	*TEX261*	SS	0.500	**0.002**	0.500	**0.019**	0.500	0.556
rs4076769	2	192292451		SS, SH	0.500	**3.9E-05**	0.237	0.651	0.385	**0.002**
rs4241279	2	192317911		SH	**0.002**	0.160	0.500	0.716	0.075	0.873
rs4483984	2	192366430		SS	0.446	**1.5E-04**	0.343	0.957	0.337	**0.021**
rs10931963	2	202172222	*ALS2CR11*	GH	0.382	**0.033**	0.382	0.452	0.336	0.082
rs12464623	2	202181323	*ALS2CR11*	SS, GH	0.365	**0.014**	0.378	0.242	0.333	**0.023**
rs3819340	3	150097977		SS	**0.020**	**0.003**	0.500	0.182	0.161	0.123
rs2968671	4	5025571		SS	0.184	**0.001**	0.384	0.239	0.171	0.212
rs1471928	4	140319218		SS	0.466	**0.001**	0.474	0.399	0.458	0.245
rs1863984	5	4264882		SS	0.296	**9.0E-06**	0.140	0.219	0.140	**4.0E-04**
rs10066063	5	32126051	*PDZD2*	SS	0.500	**6.7E-05**	0.321	0.289	0.500	0.136
rs1077606	5	32135448	*PDZD2*	SS	0.500	**6.0E-06**	0.500	0.374	0.500	0.075
rs17053011	5	55293694	*SGCD*	SS	0.500	**9.0E-06**	0.238	**0.005**	0.500	**1.4E-05**
rs1422859	5	166021965		SS	0.471	**1.5E-04**	0.500	0.153	0.479	**0.001**
rs2988681	6	87016654		SS	0.493	**1.4E-04**	0.448	0.185	0.448	**0.001**
rs6935315	6	99459993	*FBXL4*	SS	0.500	**2.7E-04**	0.455	**0.049**	0.472	0.259

					CF-GWAS FAMILIES		FIN-GWAS FAMILIES		COMBINED	
SNP	**CHR**	**BP**	**GENE**	**REP**	**Linkage**	**LD|Linkage**	**Linkage**	**LD|Linkage**	**Linkage**	**LD|Linkage**
rs10262194	7	6241110	*CYTH3*	SH	0.500	0.113	0.500	**0.013**	0.500	0.268
rs12701440	7	35617230		SS	0.500	**1.1E-05**	0.446	0.723	0.449	**0.012**
rs4717992	7	63219204		SS	0.466	**0.002**	0.195	0.956	0.458	**0.037**
rs7009615	8	18617115	*PSD3*	SS	0.500	**1.0E-05**	0.472	0.587	0.500	**0.003**
rs10501399	11	68979039		SS	0.207	**5.1E-05**	0.472	0.195	0.468	**0.001**
rs2564580	12	30288653		SS	0.488	**9.1E-05**	0.479	0.924	0.492	**0.006**
rs4587895	14	27532119		SS	**0.047**	0.496	0.495	0.084	0.318	0.459
rs10484187	14	46802461	*MDGA2*	SS	0.476	**1.5E-05**	**0.020**	0.744	0.430	**0.005**
rs1959033	14	63405339	*SYNE2*	SS	0.500	**0.001**	0.500	0.519	0.500	**0.008**
rs2042005	17	5674282		SS	0.075	**4.9E-04**	0.459	0.716	0.455	**0.011**
rs11661310	18	27781882		SS	0.500	**1.0E-04**	0.464	0.667	0.500	**0.005**
rs2904376	20	51955127		SH	0.466	**0.020**	0.477	0.074	0.500	**0.007**

NOTE: Only SNPs resulting in $p<0.05$ in either linkage- or family-based association analyses are shown. In addition to the autism families, controls were included in the PSEUDOMARKER analysis. For the CF-GWAS families, the 181 GWAS controls were included, and for the replication families, 492 controls from all of Finland were included. Position is given according NCBI build 36.1. CHR – chromosome, CRIT – Criteria for inclusion in replication dataset, SS – Single SNP association analysis resulting in $p<1\cdot10^{-4}$, GH – Global haplotype association analysis resulting in $p<1\cdot10^{-5}$, SH – Single haplotype association analysis resulting in $p<1\cdot10^{-6}$, Linkage – p-value for nonparametric linkage analysis, LD|linkage – p-value for recessive PSEUDOMARKER association test conditional on linkage. p-values <0.05 are indicated in bold.

The first SNP showing suggestive association in family based analysis, rs2302561, resulted in p=0.002 in the CF families, and p=0.019 in the Finnish replication families. However, although the ancestral G allele is significantly enriched in the CF data set (frequency 0.43 in cases versus 0.22 in controls), the G allele frequency is actually higher in controls than cases in the replication data set (frequency 0.22 in cases versus 0.29 in controls). The SNP also resulted in the most significant evidence for association (p=0.06) of all replication SNPs in the AGRE families. In the AGRE dataset, the derived G allele is slightly overtransmitted to affected individuals in TDT analysis, in agreement with the CF-GWAS results. Frequencies in the Finnish controls are similar to those in the European HapMap populations, where the derived T allele is the major allele, and the ancestral G allele frequency is around 30%. The rs2302561 SNP is a synonymous SNP located in the last coding exon of *TEX261,* a gene linked to NMDA receptors.

The second SNP, rs17053011, resulted in $p=9\cdot10^{-6}$ in the family-based association analysis of the CF families and in p=0.005 in the replication families. The SNP is located in *SGCD* and encodes a component of the dystrophin-glycoprotein complex, which links the F-actin cytoskeleton to the extracellular matrix. As the protein is expressed mainly in skeletal and cardiac muscle and has been linked with autosomal recessive limb-girdle muscular dystrophy and dilated cardiomyopathy, it is an unlikely candidate gene for ASDs (Nigro et al. 1996).

The third SNP resulting in $p<0.05$ in the family-based association analysis in both the CF and replication families is rs6935315. The SNP is located in an intron of *FBXL4* encoding a member of the F-box protein family. The F-box proteins constitute one of the four subunits of a ubiquitin protein ligase complex, which functions in phosphorylation-dependent ubiquitination and is involved in the control of the cell cycle (Cenciarelli et al. 1999; Winston et al. 1999). In the CF families, the ancestral allele A is overrepresented in cases, whereas in the replication data set the derived G allele is more common in cases than controls.

The SNP rs4658971 did not result in significant evidence for association in the family-based association analysis of the CF families, but resulted in $p=2\cdot10^{-5}$ in the case-control association analysis in the CF-GWAS. It also resulted in suggestive association in the replication families (p=0.014) as well as in the combined CF and replication families (p=0.002). This SNP is located in *DISC1*, which has been shown to be involved in several neuropsychiatric disorders and related traits (see also 5.1).

We also attempted to replicate the haplotype findings, and performed family-based association analysis using FBAT in all the Finnish replication families. We selected all SNPs residing in haplotypes resulting in $p<1\cdot10^{-5}$ in the global association test and $p<1\cdot10^{-6}$ for single haplotypes in the CF-GWAS. The same haplotypes were also

analyzed in the AGRE data using the haplotype-TDT test implemented in PLINK. Haplotype analysis in the replication data sets provided further support for the role of 2p16.1 in ASDs. Three 5-SNP haplotypes were analyzed in this region in the Finnish replication families, and the most significant evidence for association (global test p=0.017) was obtained for a haplotype comprised of rs243034, rs243027, rs243081, rs243073 and rs8179712. Using the same SNPs, a single haplotype allele resulted in p=0.004 in the AGRE dataset. The adjacent 5-SNP haplotype, comprised of rs243027, rs243081, rs243073, rs8179712 and rs1011407 also resulted in suggestive evidence for association. The Finnish replication families resulted in p=0.038 in the global test and one haplotype allele at this locus resulted in p=0.009 in the AGRE families. As mentioned previously, the haplotype association results in the CF-GWAS are probably largely due to the effects of the large allele frequency differences at rs9309326. However, it should be noted that this SNP did not result in significant p-values in the replication data-set, suggesting that rs9309326 serves as a proxy for the true susceptibility variant in the CF-dataset, that is more efficiently tagged by the 5-SNP haplotype in the replication dataset.

No haplotypes outside the 2p16.1 locus resulted in $p<0.05$ in the Finnish replication families. In the AGRE dataset, one haplotype at the 2q33 locus resulted in p=0.013 and another at 20q13.2 in p=0.024. The haplotype alleles associated with ASDs in the AGRE dataset are not the same as those observed in the CF-GWAS.

To summarize the association analysis, both the single SNP and haplotype association analysis revealed common variants enriched in the CF dataset compared to regionally matched controls. To further assess the role of these variants in the etiology of ASDs, we used two replication datasets, one from Finland and another from the US. A subset of single SNP and haplotype findings showed suggestive evidence for association in the replication datasets, and encouragingly, some of the haplotype findings resulted in suggestive p-values in both replication datasets. However, for some susceptibility variants, the replication findings were not obtained for the same allele as in the CF-GWAS, and these results should be interpreted with caution. The haplotype alleles associated with ASDs in the original and replication datasets are different, but this could be a result of the presence of different susceptibility variants in the different populations.

5.5.7 CNV analysis

In addition to genotype information, the SNP probe intensity values from genome-wide SNP microarrays can be used to detect amplified or deleted segments. We used PennCNV (Wang et al. 2007) to identify CNV regions in the entire autism dataset, consisting of both the CF-GWAS and FIN-GWAS. PennCNV implements a hidden

Markov model to infer CNV states using both the total probe intensity (Log R ratio, LRR) and the relative ratio of the fluorescent signals between two alleles at each SNP (B allele frequency, BAF). PennCNV also implements a correction for wavelike patterns in the intensity spectra by modeling the GC content in the region (Diskin et al. 2008).

As the frequency and distribution of CNVs in the general population still need to be further characterized, we wanted to adopt a conservative threshold defining CNV regions of interest (CNV-ROIs) and to focus on CNVs in regions which have not been reported in controls. In addition to the controls included in this study, we used a set of approximately 5400 Finnish individuals from the NFBC66 cohort, which have been genotyped using the Illumina HumanHap 370 beadchip, and assessed for CNVs using PennCNV. We discarded all samples with large variability in the intensity measurements as they did not meet quality control criteria (LRR_SD>0.3, BAF_DRIFT>0.0025, WF>0.04 or WF<-0.04). QC resulted in the exclusion of 4 cases and 848 controls, resulting in a total of 124 cases and 4608 controls for CNV analysis.

We defined CNV-ROIs as CNVs identified in cases but with no more than a 50% overlap with a CNV reported in the Database of Genomic Variants. We also excluded all regions identified in 2 or more controls in our large Finnish control dataset. We excluded all CNV calls containing <10 SNPs, as the variability in the intensity data results in a large probability of false positive calls in small regions. Only CNVs spanning genes were included in the analysis. A total of 35 CNV-ROIs were identified (Table 25).

We identified five large CNVs, three at 15q11-13, one at 1q and one at 9q. We genotyped available family members to investigate the inheritance of these CNVs using Illumina HumanHap 370 Beadchip. Both the gain on chromosome 1 as well as the gain on chromosome 9 were inherited from the father, who did not have a diagnosis of ASDs. Interestingly, in the family with the 1q duplication, the nephew of the affected father is undergoing assessment for ASDs at three years of age. In the other family, the father has had delayed speech development and difficulties in learning to read and write. Two of the 15q11-13 duplications were inherited from an unaffected mother. The third 15q11-13 duplication, estimated to have a copy number of four and is *de novo*. In addition to these loci, several smaller regions of interest were identified. These include a deletion spanning part of *NRXN1*, and a duplication of 22q13 that includes *SHANK3*. Nine regions overlapped with CNVs reported in the Autism Chromosome Rearrangement Database (ACRD), supporting the role of these loci in the etiology of ASDs. In addition, several novel loci were identified.

Table 25. CNV regions of interest identified in the CF-GWAS and FIN-GWAS.

CHR	From-kb	To-kbP	N_{SNP}	Length	CN	N_{ind}	ACRD	Genes
1	198 896	199 007	26	110	3	1	No	DDX59, CAMSAP1L1
1	226 859	229 343	577	2 482	3	1	No	18 genes
2	37 031	37 165	30	134	1	1	No	STRN, HEATR5B
2	47 455	47 569	16	115	1	1	No	EPCAM, MSH2
2	51 068	51 189	27	121	1	1	Loss	NRXN1
2	197 079	197 200	16	121	3	1	No	HECW2
2	208 742	208 808	11	66	1	1	Gain	C20orf80
2	230 582	230 633	10	51	1	1	No	FBX036, SLC16A14
2	238 078	238 159	17	70	3	2	Loss	MLPH, PRLH, RAB17
3	129 114	129 172	10	43	1	2	No	KLHDC6
5	43 037	43 248	22	211	1	1	No	C5orf39, ZNF131, MGC42105
5	59 759	59 795	15	37	3	1	No	PDE4D
6	7 927	8 003	13	76	1	2	No	MUTED
6	10 871	10 937	10	66	1	1	No	MAK
6	72 989	73 382	95	393	3	1	No	RIMS1
6	116 723	116 808	14	85	1	1	Loss	DSE
7	33 098	33 154	10	56	3	1	No	RP9, BBS9
9	87 214	88 935	225	1 721	3	1	No	AGTPBP1, MAK10, GOLM1, C9orf153, ISCA1, ZCCHC6, GAS1
9	106 297	106 874	187	576	3	1	No	8 olfactory receptors,, NIPSNAP3A, NIPSNAP3B, ABCA1
9	135 459	135 533	42	74	3	1	No	DBH, SARDH
10	11 367	11 531	34	164	3	1	No	CUGBP2
10	51 456	51 828	63	372	3	3	Gain	FAM21A, FAM21B, ASAH2, SGMS1
10	72 253	72 310	14	57	1	2	No	SGPL1
13	18 514	18 613	13	98	3	1	No	DKFNp686A1627
14	34 999	35 148	15	150	1	1	No	INSM2, GARNL1
15	18 421	25 688	572	7 266	4	1	Yes	26 genes
15	20 307	26 209	620	5 902	3	1	Yes	22 genes
15	21 240	26 209	566	4 969	3	1	Yes	14 genes
15	90 016	90 381	96	365	3	1	No	SLCO3A1

CHR	From-kb	To-kbP	N_{SNP}	Length	CN	N_{ind}	ACRD	Genes
19	12 868	12 942	11	73	1	1	No	GCDH, SYCE2, FARSA, CALR, RAD23A, GADD45GIP1, DAND5
20	41 032	41 177	36	145	1	1	Gain	PTPRT
22	48 777	49 525	141	748	1	1	Gain and loss	32 genes including SHANK3

NOTE: CHR – Chromosome, From-kb – Start of CNV in kilobases from ptel, To-kb – End of CNV in kilobases from pten. Positions according to NCBI build 36 NSNP – Number of SNPs in the CNV region, Length – Length of CNV in kilobases, CN – Copy number, ACRD – Presence of CNV in Autism Chromosome Rearrangement database.

5.5.8 Genome-wide expression profiling

We used Affymetrix U133 2.0 Plus chips to establish genome-wide expression profiles for 10 male individuals with autism (FIN-EXPR). Age-matched male individuals with no diagnosis of ASDs were used as controls. After reannotation of the probes according to Entrez and normalization, we identified a total of 325 differentially expressed genes at the p=0.01 level with fold change ranges between -2.24 to +4.14 (Figure 2a, publication V). Of these, 92 were upregulated and 233 downregulated. At p=0.05 significance level, a total of 1286 genes were differentially expressed. Hierarchical clustering using the differentially expressed genes resulted in the clear separation of autism cases and controls, which were split into two clusters (Figure 2b, publication V). Due to the limited number of samples available for gene expression profiling, we chose not to evaluate the differentially expressed genes in traditional ways. Instead, we chose to use the results for pathway analysis to evaluate the data together with the GWAS data.

5.5.9 Pathway-analysis of GWA and expression data

We used non-parametric pathway algorithm developed in-house to identify GO-categories which were enriched in differentially expressed or associated genes. Pathway analysis of gene expression data was performed by ranking all of the genes on the absolute fold change. SNP data was ranked based on the uncorrected p-value from the association analysis. All three GWA-datasets (CF-GWAS, FIN-GWAS and AGRE-GWAS), and two expression datasets (the Finnish expression dataset (FIN-EXPR) and a publicly available dataset (GEO-EXPR)) were included in the pathway analyses. Very little overlap was observed between the GO-categories from the different datasets (Table 26).

Table 26. Top 10 regulated GO-categories in the datasets included in this study.

GO-category	Category description	p
CF-GWAS		
GO:0016528	sarcoplasm	0.001
GO:0016529	sarcoplasmic reticulum	0.001
GO:0043193	positive regulation of gene-specific transcription	0.001
GO:0032583	regulation of gene-specific transcription	0.001
GO:0001525	angiogenesis	0.003
GO:0048514	blood vessel morphogenesis	0.003
GO:0008168	methyltransferase activity	0.003
GO:0001568	blood vessel development	0.004
GO:0016741	transferase activity, transferring one-carbon groups	0.004
GO:0035107	appendage morphogenesis	0.006
FIN-GWAS		
GO:0005764	lysosome	0.0038
GO:0000323	lytic vacuole	0.0038
GO:0005773	vacuole	0.0069
GO:0031968	organelle outer membrane	0.0073
GO:0019888	protein phosphatase regulator activity	0.0078
GO:0019867	outer membrane	0.0081
GO:0046483	heterocycle metabolic process	0.0088
GO:0016810	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds	0.0119
GO:0019208	phosphatase regulator activity	0.0139
GO:0018193	peptidyl-amino acid modification	0.0171
AGRE-GWAS		
GO:0032870	cellular response to hormone stimulus	0.0001
GO:0005819	spindle	0.0002
GO:0000278	mitotic cell cycle	0.0019
GO:0007067	mitosis	0.002
GO:0043434	response to peptide hormone stimulus	0.0023
GO:0000087	M phase of mitotic cell cycle	0.0025
GO:0000279	M phase	0.0031
GO:0051329	interphase of mitotic cell cycle	0.0033
GO:0005179	hormone activity	0.0037
GO:0009725	response to hormone stimulus	0.0059
FIN-EXPR		
GO:0051640	organelle localization	0.0006
GO:0000242	pericentriolar material	0.0007
GO:0004715	non-membrane spanning protein tyrosine kinase activity	0.0008
GO:0031532	actin cytoskeleton reorganization	0.0011
GO:0001515	opioid peptide activity	0.0012
GO:0010573	vascular endothelial growth factor production	0.0021
GO:0008156	negative regulation of DNA replication	0.0022
GO:0010574	regulation of vascular endothelial growth factor production	0.0022

GO-category	Category description	p
GO:0034464	BBSome	0.0024
GO:0032350	regulation of hormone metabolic process	0.0026
GEO-EXPR		
GO:0007019	microtubule depolymerization	0.0003
GO:0031646	positive regulation of neurological system process	0.0004
GO:0050806	positive regulation of synaptic transmission	0.0005
GO:0031111	negative regulation of microtubule polymerization or depolymerization	0.0006
GO:0050000	chromosome localization	0.0007
GO:0051963	regulation of synaptogenesis	0.0008
GO:0051303	establishment of chromosome localization	0.0009
GO:0046459	short-chain fatty acid metabolic process	0.0009
GO:0051971	positive regulation of transmission of nerve impulse	0.0011
GO:0050807	regulation of synapse organization and biogenesis	0.0011

NOTE: CF-GWAS – Central Finland GWAS, FIN-GWAS – Finnish replication GWAS, AGRE-GWAS – Autism Genetic Research Exchange GWAS, GEO-EXPR – Gene Expression Omnibus expression dataset GSE6575.

5.5.10 Discussion

Previous studies have shown that the genetic architecture of ASDs is highly heterogeneous. We have previously shown that the Finnish population is stratified, with allele frequencies differing drastically between subisolates. To reduce genetic heterogeneity in our sample, we limited our study sample to families from the internal isolate of CF and a subset of individuals in this dataset can be genealogically linked to two large pedigrees. We have previously performed a dense microsatellite linkage study in the larger extended pedigree, in which no haplotypes shared among nuclear families were identified. In this study we wanted to follow up these findings using a denser set of markers as well as an increased sample size.

First, we used dense, genome-wide SNP data to identify regions shared IBD in these distantly related individuals. We sought regions resembling either a dominant or recessive model within the extended pedigrees. However, no regions of interest were identified, suggesting that genetic risk factors are not shared across individuals in these pedigrees. The lack of shared haplotypes seems surprising, and even more surprising, IBD sharing data suggest that these families share less of their genome IBD compared to the other families used who were not linked genealogically but had two grandparents from CF. These results are in line with findings in other studies where identified mutations have often been *de novo* and specific to nuclear families. Homozygosity mapping has been primarily successful in closely consanguineous families where parents have been first or second cousins (Morrow

et al. 2008). The lack of increased IBD sharing between individuals in the large pedigrees agree with the lack of highly penetrant rare alleles enriched in these pedigrees. This data, combined with the results of the homozygosity- and shared segment analyses strongly agree with the hypothesis of *de novo*, family-specific risk variants. The methodology used in this study is poorly suited for identification of rare mutations unique to single families. These variants can only be revealed by genome-wide resequencing of affected individuals.

We tested for common genetic risk factors within the CF dataset, and tried to replicate findings in two other datasets - one from the rest of Finland and the other from the US population. Suggestive support from both Finnish datasets was observed for a total of four variants, of which the functionally most interesting are located in *DISC1* and *TEX261*. *DISC1* has been linked to schizophrenia and neurocognitive phenotypes, and we have previously shown that markers in *DISC1* are associated with ASD in the Finnish families (Kilpinen et al. 2008). It should be noted, however, that the families used in the *DISC1* study comprise a significant subset of the families in the CF-GWAS and replication material, and so the association observed in *DISC1* here cannot be viewed as an independent replication. The best results in our previous *DISC1* study were obtained with a microsatellite marker located only 77 kb from the SNP associated with ASDs in this study.

TEX261 is reported to be induced by NMDA stimulation in the mouse hippocampus, modulating the NMDA receptor-mediated signaling cascade and resulting in cell death (Taniura et al. 2007). The NMDA receptors, located at many excitatory glutamate synapses in the central nervous system, are interesting in autism studies as they are involved in biological processes linked to brain development, excitatory neurotransmission and synaptic plasticity. Overexpression of NMDA receptors in the forebrain of mice results in enhanced learning and memory capabilities in various behavioral tasks, and have therefore even been suggested to be a possible target for treating learning and memory disorders (Tang et al. 1999).

All GWA datasets, including the AGRE dataset, showed evidence for suggestive association at 2p16.1 in haplotype analysis. The haplotype does not overlap with any coding genes, and is located close to the end of the *BCL11A* gene, which is involved in B and T lymphocyte development, and is expressed in erythroid cells (Satterwhite et al. 2001; Sankaran et al. 2008) Mice deficient in *Bcl11a* lack B cells and have alterations in several types of T cells (Liu et al. 2003). The closest gene upstream is *FANCL*, located over 2 Mb from rs9309326. The gene encodes an ubiquitin ligase implicated in Fanconi anemia etiology (Meetei et al. 2003). Mice deficient in Pog, the mouse homolog of *FANCL*, display deficient proliferation of germ-cells (Agoulnik et al. 2002). The haplotype could modify the expression of these genes, or overlap with unknown small non-protein coding genes.

It should be noted that for all of these loci except for the marker in *DISC1*, different alleles or haplotypes showed association within the different datasets. The inconsistent allele frequency findings in the CF and replication datasets could indicate either that the findings are false positives, or that the SNPs tag causal variants present on different haplotypes in the different datasets and are not the risk variants themselves. As only a limited number of common susceptibility variants have been identified in autism, susceptibility variants could differ among families. Although these datasets are limited, recent studies have shown that if the phenotype of interest is controlled by common risk factors enriched in an isolate, they can be identified in a study design identical to the one used here, where a limited number of distantly related individuals and carefully matched population controls are used (Kallio et al. 2009).

We also assessed CNVs in these datasets. No enriched CNVs were identified in the Central Finland families, but two large duplications were identified in the extended pedigrees, further supporting the role of rare, family specific variants even within these sets of distantly related individuals. Interestingly, many of the CNV-ROIs identified here have been previously reported in ACRD. The ACRD uses a similarly conservative approach for the inclusion of putative ASD related CNVs where variants only identified in ASD cases but not in controls are included. Although the distribution of CNVs in normal populations still needs to be more carefully characterized, the presence of identical CNVs in our dataset and the ACRD further support the role of these CNVs in the etiology of ASDs.

The data from the GWA analysis, together with findings from other ASD studies, support the role of several rare, family specific variants. However, we wanted to take the GWAS dataset a step further from traditional GWA data analysis, which is usually confined to association or haplotype analysis. We attempted to identify GO-categories that link the "grey zone" of results, which are suggestive results that do not reach genome-wide significance. As we also had access to two expression datasets, we were able to rank the evidence from these analyses based on the number of datasets supporting the role of the identified biological processes. Although different genetic variants in these datasets could cause disruptions of separate genes that result in a similar phenotype, the genes should affect common biological pathways. However, very little overlap was observed between the GO-categories.

It should be noted that the analysis of GO-categories also presents limitations. The GO-categories provide a far from complete picture of true biological function, because the majority of gene functions are still incompletely characterized. However, GO-categories comprise the most extensive systematic annotation of available data concerning gene function, and despite its limitations, it is currently the best option available for system-biology analysis. More thoroughly annotated

pathway-collections such as the Kyoto Encyclopedia of Genes and Genomes (KEGG)-database exist, but they are mainly focused on metabolic pathways, and provide only limited information of the neuronal processes, of greatest interest in this study.

In summary, the primary result of the GWA analysis is that there are no enriched high impact mutations in the ASD-families originating from a subisolate in a founder population. This is in agreement with other studies, where rare, family-specific mutations have been identified and where common variants have only been identified in large samples (Wang et al. 2009). Further, the combination of GWAS and expression data did not yield evidence for overlapping biological processes among the most associated GO-categories.

6 CONCLUDING REMARKS AND FUTURE PROSPECTS

When this study was initiated, knowledge of the genetics of idiopathic ASDs had been obtained from linkage studies and observation of large, microscopically detectable chromosomal aberrations. Linkage studies typically comprised of 50-150 families or affected sib pairs, and resulted in candidate regions on almost every chromosome (Yang and Gill 2007). Potential reasons for the heterogeneity of results are the different clinical categories used, various statistical approaches taken, and partial overlap of study samples. Although some of the linkage regions will prove to be false positives caused by small samples sizes, the findings most likely also represent the substantial genetic heterogeneity underlying ASDs. The success of risk variant identification within linkage peaks has been poor in general, but some interesting candidate genes have been identified. These include *EN2* and *CNTNAP2* at 7q36 (Gharani et al. 2004; Benayed et al. 2005; Cheh et al. 2006; Alarcon et al. 2008; Arking et al. 2008; Bakkaloglu et al. 2008; Turunen et al. 2008), *RELN* and *MET* at 7q21-22 (Persico et al. 2001; Bonora et al. 2003; Skaar et al. 2005; Campbell et al. 2006; Campbell et al. 2007), and *HTR3C* at 3q27 in this study. The majority of families included in the linkage studies have been included in association studies, but so far these studies have only been used to test for association between ASDs and common variants. To evaluate the role of linkage regions in these families, the linkage regions should be further characterized using the genome-wide SNP data through homozygosity and haplotype sharing analysis within families. As generation of SNP data has been both laborious and costly, all possible approaches should be used to utilize this data as effectively as possible.

The chromosomal abnormalities reported in ASDs further support the hypothesis of genetic heterogeneity in these disorders. Deletions and duplications have been reported on almost every chromosome. Rare chromosomal rearrangements have resulted in the identification of risk variants in functionally interesting genes, such as NLGNs, *NRXN1* and *SHANK3*, which in turn implicate synaptic dysfunction in the etiology of ASDs. A significant step forward in the genetic characterization of ASDs was taken when genome-wide BAC arrays made it possible to identify submicroscopic lesions. It soon became evident that CNVs contribute to autism susceptibility (Sebat et al. 2007; Szatmari et al. 2007; Marshall et al. 2008), but again shared CNVs were the exception rather than the rule. Genome-wide SNP genotyping and simultaneously offered the opportunity to test for association and the presence of CNVs resulting in numerous novel loci. The results of linkage studies, rare mutations and numerous large and small chromosomal aberrations point to rare, high penetrance variants unique to single families.

However, some candidate gene studies, together with the recently published GWASs in ASDs point to the presence of common variants in the etiology of the disorders as well. In the first published GWAS, variants located between *CDH9* and *CDH10* on 5p14.1 were identified and the finding was replicated in two independent datasets. This agrees well with findings in other complex disorders where common variants have low risk ratios and a large number of individuals are needed to reliably identify susceptibility variants. Furthermore, the cadherins being neuronal cell-adhesion molecules fit well with the emerging evidence of the role of cell-adhesion molecules, such as NLGNs and NRXNs in ASDs. In another GWAS, replicated association was identified to an adjacent region, 5p15, close to *SEMA5A* (Weiss et al. 2009). In that study, the genome-wide SNP data was also used for linkage analysis and significant linkage was identified on 6q27 and 20p13. The fact that the linkage peaks did not overlap with the association peak could suggest that the linkage peaks contain rare, highly penetrant risk alleles. This agrees with results obtained in this study where strong linkage has been identified, but the GWAS did not reveal association at the linkage peaks. The presence of common variants associated with autism in *HTR3C* identified in this study only explain a part of the observed linkage, which suggests that the linkage peaks could contain rare and common risk variants. Accumulating evidence from both this study, and others strongly suggests that there are only few common variants predisposing for ASDs. This would agree with multiple studies reporting *de novo* genetic variants as predisposing factors. Further, a minority of individuals with ASDs have children, resulting in strong negative selection against these *de novo* variants in the population. The presence of common variants could however help explain the observed incomplete penetrance of rare variants and CNVs. The common variants could modify the expressivity of the observed rare, high penetrance mutations resulting in apparent nonsegregation of these variants in families and the presence of these variants in unaffected parents.

In this study, we wanted to utilize the advantages offered by a well characterized family set from an isolated population to pinpoint genetic risk factors for ASDs. We have focused on clinically well characterized small subgroups to limit genetic heterogeneity. This approach has proved successful at least in the case of AS where the linkage signal on 3p14-24 has been replicated in an independent family set. The family set used in the AS study was the smallest in this thesis, but it consisted of large families with multiple affected individuals providing substantial power for linkage analysis. The results of the population stratification study (IV) showed that even though Finns display reduced genetic heterogeneity compared to other European populations, significant substructure can be observed within the population. To take advantage of this substructure, in study V we concentrated on a family set originating from an internal isolate in an attempt to enrich shared genetic

risk factors. We have shown using GWA data that even if shared ancestors can be genealogically linked, this does not necessarily mean that families share any high impact genetic risk variants or haplotypes. In fact, our data suggest the opposite in that the individuals in the two genealogically linked extended pedigrees share no more of their genome IBD than healthy controls collected from the same geographic region. There are no systematic studies of the prevalence of ASDs in different parts of Finland, but there is evidence available of specific geographic regions with higher prevalence. The ASD prevalence in Finland is also similar to that in other countries (Kielinen et al. 2000). The lack of increased IBD sharing together with the expected frequency of ASDs in Finland provide evidence against founder mutations enriched in this isolate. These facts, together with the results of study V provide convincing evidence that even in the Finnish population, genetic susceptibility factors for ASDs are *de novo*, undergo strong negative selection and are unique in nuclear families and not shared among distantly related individuals.

Linkage and association analyses detect different predisposing factors. Linkage analysis evaluates the co-segregation of haplotypes with phenotypes within families, whereas association analysis in the simplest form compares allele frequencies between cases and controls. Linkage analysis is more efficient for detection of rare variants, whereas association analysis is more suited for the identification of common variants. The results obtained by these methods are dependent on the makeup of the genetic risk factors for the phenotype under study. The results of the WTCCC‟s landmark paper on the association analysis of seven disorders nicely illustrate the effects that the underlying genetic architecture has on the probability of identifying risk variants (WTCCC 2007). For all phenotypes, 2000 cases and 3000 shared controls were used. Some disorders, such as coronary artery disease, type 1 diabetes and rheumatoid arthritis, resulted in clear candidate regions. For other disorders, such as bipolar disorder and hypertension, no genome-wide significant variants were identified. It is probably not a coincidence that one of the phenotypes resulting in only suggestive association signals was a neuropsychiatric disorder. Phenotypes such as coronary artery disease and type 1 diabetes have clear biological markers which can be used to establish a diagnosis. For neuropsychiatric phenotypes, the definition of the phenotype is based on observations of the patient, and although diagnostic criteria are clearly outlined in diagnostic manuals, the subjective opinions of clinicians may play a role in the diagnosis. It is also obvious that the GWA approach will be fruitful for phenotypes where many common, shared variants confer susceptibility to the disorder even if risk alleles individually confer only a marginal increase in risk. For such phenotypes, increasing sample-sizes will reveal new, even smaller effect variants, as has been done for many phenotypes including metabolic phenotypes, height and weight (Lettre et al. 2008; Weedon et al. 2008; Aulchenko et al. 2009; Thorleifsson et al. 2009). Several studies, including

this study, have shown that the genetic risk factors for autism are not predominantly of this type, but consist more prominently of rare, high impact alleles unique to small subsets of families.

The studies of the genetics of ASDs, including this one, have usually contributed individually only small pieces of the puzzle. However, when put together in the right context, the genetic picture of autism has never looked as clear as it does today. For idiopathic autism, the available data suggests that rare, family specific variants might compose the most significant genetic risk factors and a smaller, although apparent role is attributed to common, possibly modifying genetic factors. Existing knowledge that includes known genetic syndromes, chromosomal abnormalities, CNVs and coding mutations account for 10-20% of ASD cases (Abrahams and Geschwind 2008). Developing methodologies for more sophisticated CNV analysis and the gathering of data containing less variability will make it possible to identify smaller CNVs. Next-generation sequencing technologies will enable massive resequencing projects of genes involved in neuronal development, thereby making it possible to identify rare variants present in only one or a few families. The bioinformatic challenges will, however, increase tremendously as larger amounts of datapoints for each individual are produced. The differentiation of rare benign variants from variants increasing risk of disorders will require massive, well-characterized control data, which will become available from projects such as the 1000 genome project. A proof-of-principal study of the feasibility of massive resequencing has already provided new genes associated with X-linked MR (Tarpey et al. 2009). However, the same study also identified loss of function mutations in individuals with no clinical phenotype. Therefore future resequencing studies should be well designed to be able to correctly interpret the large amount of data and rule out benign polymorphisms, even if they are rare in the population.

The small number of identified mutations suggests that the majority of genetic risk factors for ASDs remain to be discovered. Several aspects of genetic analysis remain largely uncharacterized, such as methylation and microRNA profiling as well as pathway and transcription factor analysis. These could provide important new insight into the genetic architecture of ASDs in a similar way that the groundbreaking studies of the role of CNVs have done recently. Finally, it should be noted that the identification of a genetic variant is a mere first step in the understanding of the underlying biology. Identifying genetic factors for ASDs is only the beginning of the long path of validating the findings functionally in order to identify biologically relevant processes. The identification of these processes and pathways will advance the understanding of the pathogenesis of ASDs and hopefully in the future also provide diagnostic markers or even serve as a starting point for providing a cure for these disorders.

7 ACKNOWLEDGEMENTS

This study was carried out at the National Public Health Institute, National Institute for Health and Welfare (THL) and Institute for Molecular Medicine Finland (FIMM). Director General of THL Pekka Puska, the head of the Public Health Genomics Unit Anu Jalanko and Director of FIMM Olli Kallioniemi are acknowledged for providing excellent research facilities.

This study has been financially supported by the Helsinki Biomedical Graduate School, The Medical Society of Finland (Finska Läkaresällskapet), the Academy of Finland, the Alma & KA Snellman Foundation, the Biomedicum Helsinki Foundation, the Emil Aaltonen Foundation, Helsinki University Funds, the Maud Kuistila Foundation and the Päivikki and Sakari Sohlberg Foundation. All families who have participated in this study are warmly thanked for their collaboration, without which this study would not be possible.

Professor Kerstin Lindblad-Toh is thanked for accepting the role of Opponent. Professor Jim Schröder and Adjunct Professor Tarja Laitinen are thanked for valuable comments during the revision of this thesis. Professor Schröder is also warmly thanked for his inspirational lectures in human genetics during my undergraduate studies. I greatly appreciate his expertise in the field of human genetics, but also the fact that his door is always open for his students. Adjunct Professor Laitinen is also warmly thanked for being a part of my thesis committee and for help and support during this project. Adjunct Professor Marjo Kestilä, also a member of my thesis committee, is warmly thanked for help with so many practical issues, and for always supporting me especially through the hard times. Peter Wagner is thanked for language revision of this thesis.

I want to thank my two supervisors, Professor Leena Peltonen and Tero Ylisaukko-oja, Ph. D. for their guidance and support during this all these years. Leena‟s vision and enthusiasm have been the driving forces behind this project. Leena has also given me the opportunity to freely take my research in any direction I have wanted and been there to point me back on track when I have got it wrong. Tero is thanked for patiently spending hours and hours teaching me about molecular genetics, statistical analysis, manuscript writing and for listening to me nervously practicing my presentations over and over again. Tero also set the positive and open working atmosphere in our autism group which has remained to this day.

Molecular genetics would not be possible without the efforts of outstanding clinicians. The autism clinical team was headed by Professor Lennart von Wendt whose enormous work and expertise made these studies possible and tied together

the different ASD research projects in Finland. Reija Alen, Marko Kielinen, Tiia Källman, Irma Moilanen, Taina Nieminen-von Wendt, Esko Pulkkinen, Susan Sarenius and Raija Vanhala are acknowledged for their invaluable help with sample collection and characterization as well as help and encouragement along the way.

I have been lucky to work with the wonderful autism-team. Elli Kempas has produced a large part of the data included in this study. Her encouragement and support have meant so much. Helena Kilpinen is thanked for putting up with me for all these years we have collaborated. I have been fortunate there is such an ambitious, talented and highly organized scientist in our team, who has taken on the huge task of bringing the autism project from *in silico* mapping to "real biology". Mari Rossi, Emília Gaál and Ilona Kotala are thanked for help and collaboration with autism research as well as for their friendship.

Teppo Varilo is thanked for valuable help with genealogical data, and helpful revision of several manuscripts. Iiris Hovatta is thanked for our efficient collaboration and numerous helpful scientific discussions. Samuli Ripatti is thanked for help with statistical analysis, and for explaining them in a way even a blond geneticist understands. Markus Perola is acknowledged for help with ethical review board applications. Juha Saharinen is thanked for efficient and swift help with various bioinformatics problems. Irma Järvelä is thanked for introducing me to the field of autism genetics, and for giving me my first job in the lab when I had no experience of research work. Mark Daly and Shaun Purcell are acknowledged for their help with the GWA studies, and for making time to meet with me and help out when I have been visiting the Broad. Andrew Collins is warmly thanked for help with the LD-map analysis. Dane Chetkovich is thanked for collaboration with the PEX5L study, and Kimberly Aldinger for coming up with the idea for the ZIC-study. Dario Greco is thanked for help with the expression data analysis. Joe Terwilliger and Tero Hiekkalinna are thanked for help with statistical analysis. Tiina Paunio is acknowledged for valuable help with the population stratification study, as well as for help and support along the way.

A warm thanks to Pekka Ellonen and everybody else in the SeqLab for invaluable help during all these years. Heli Keränen is especially thanked for teaching me how to run sequencing and genotyping gels back in the good old days. Outi Törnwall and Minttu Jussila are warmly thanked for help with DNA related matters. Anne Nyberg is warmly thanked for help in the lab, which has always been accompanied by a smile. Everybody at GIU are thanked for helping with computer-related problems. Sari Kivikko is warmly thanked for her patience and help with a multitude of problems ranging from scheduling a meeting with Leena to improving the quality of coffee at BM2. Kind help from Tuija Koski, Mika Kivimäki, Sanna Tossavainen,

Sisko Lietola, and Liisa Penttilä with various practical matters has always been valuable.

I want to thank my peer-support group, Eveliina Jakkula and Jonna Tallila for sharing my joys and frustrations in the lab. Jonna is thanked for always sharing her positive attitude, and for believing in me even when I didn‟t. Eve is thanked for support which can only come from a person who has also been frustratedly struggling with GWA data, as well as for hosting wonderful visits in Boston and serving as my personal travel agency. Olli Pietiläinen is thanked for fruitful scientific collaboration as well as interesting and funny scientific and not-so-scientific discussions. I would also like to thank all my other colleagues in the lab for wonderful company, good parties and congress trips. There are so many of you, and I am sure I am forgetting some, but Annika, Annu, Ansku, Antti, Anu, Emma, Emmi, Hanski, Heidi, Heli, Henna, Ida, Jarkko, Jenni, Johannes, Jonas, Joni, Juho, Jussi, Kaisu, Kati, Laura, Liisa, Marika, Marine, Markus, Mervi, Mikko K, Mikko M, Minna, Nabil, Nora, Pia, P-P, Siv, Suvi, Tiia, Virpi and Will: If „Nauru pidentää ikää‟ (laughter makes you live longer) I will probably live a very long life because of you all. I want to thank the IR group (the best friends a slightly antisocial scientist can have) Anne, Hanna K, Hanna N, Heli, Ilona, Mari, Sini, Susanna and Suvi for dragging me out of the lab to party. The IR group tyky-days and nights are unforgettable. Anne is especially thanked for the wonderful artwork on the cover of this thesis and Ilona is thanked for bringing the little bundle of joy, Laurus, and his family into my life.

Finally I want to thank my family for their love and support. I want to thank my parents for always making sure I have been able to pursue my ever-changing ambitions (and the Mum&Dad foundation is gratefully acknowledged for supporting me financially when my decisions have resulted in a negative bank balance). I also want to thank my godparents Peter and Krisse for being my extra parents in Helsinki and for giving me a roof over my head all these years. I want to thank my wonderful baby sister Myran for always being there for me, supporting me when I have needed it and kicked my behind when I have deserved it.

Helsinki, August 2009

8 ELECTRONIC DATABASE INFORMATION

Autism Chromosome Rearrangement Database; http://projects.tcag.ca/autism/

Autism Genetic Resource Exchange; http://www.agre.org/

Database of Genomic Variants; http://projects.tcag.ca/variation/

DECIPHER; https://decipher.sanger.ac.uk/application/

Entrez SNP; http://www.ncbi.nlm.nih.gov/projects/SNP/

Gene Expression Omnibus; http://www.ncbi.nlm.nih.gov/geo/

Gene Ontology; http://www.geneontology.org/

International HapMap Project; http://www.hapmap.org/

Marshfield Clinic Mammalian Genotyping Service;
http://research.marshfieldclinic.org/genetics/home/index.asp

Online Mendelian Inheritance in Man; http://www.ncbi.nlm.nih.gov/omim

Primer3; http://primer3.sourceforge.net/

PubMed; http://www.ncbi.nlm.nih.gov/pubmed/

UCSC Genome Browser; http://genome.ucsc.edu/

9 REFERENCES

Abrahams BS, Geschwind DH (2008) Advances in autism genetics: on the threshold of a new neurobiology. Nat Rev Genet 9(5): 341-355.

Abramson RK, Wright HH, Carpenter R et al. (1989) Elevated blood serotonin in autistic probands and their first-degree relatives. J Autism Dev Disord 19(3): 397-407.

Agoulnik AI, Lu B, Zhu Q et al. (2002) A novel gene, Pog, is necessary for primordial germ cell proliferation in the mouse and underlies the germ cell deficient mutation, gcd. Hum Mol Genet 11(24): 3047-3053.

Alarcon M, Cantor RM, Liu J et al. (2002) Evidence for a language quantitative trait locus on chromosome 7q in multiplex autism families. Am J Hum Genet 70(1): 60-71.

Alarcon M, Yonan AL, Gilliam TC et al. (2005) Quantitative genome scan and Ordered-Subsets Analysis of autism endophenotypes support language QTLs. Mol Psychiatry 10(8): 747-757.

Alarcon M, Abrahams BS, Stone JL et al. (2008) Linkage, Association, and Gene-Expression Analyses Identify CNTNAP2 as an Autism-Susceptibility Gene. Am J Hum Genet 82(1): 150-159.

Aldinger KA (2008) Identification of chromosome 6p25 genes involved in Dandy-Walker malformation: the role of FOXC1 in cerebellar development and implications for cerebellar genes in autism Chicago: University of Chicago. 197 p.

Altmuller J, Palmer LJ, Fischer G et al. (2001) Genomewide scans of complex human diseases: true linkage is hard to find. Am J Hum Genet 69(5): 936-950.

Altshuler D, Daly MJ, Lander ES (2008) Genetic mapping in human disease. Science 322(5903): 881-888.

Amara SG, Pacholczyk T (1991) Sodium-dependent neurotransmitter reuptake systems. Curr Opin Neurobiol 1(1): 84-90.

American Psychiatric Association (1994) Diagnostic and Statistical Manual of Mental Disorders (4th edn) (DSM-IV). Washington, DC: APA.

Amir RE, Van den Veyver IB, Wan M et al. (1999) Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. Nat Genet 23(2): 185-188.

Anderson GM (2002) Genetics of childhood disorders: XLV. Autism, part 4: serotonin in autism. J Am Acad Child Adolesc Psychiatry 41(12): 1513-1516.

Anderson GM, Horne WC, Chatterjee D et al. (1990) The hyperserotonemia of autism. Annals of the New York Academy of Sciences 600: 331-340; discussion 341-332.

Anderson GM, Freedman DX, Cohen DJ et al. (1987) Whole blood serotonin in autistic and normal subjects. J Child Psychol Psychiatry 28(6): 885-900.

Arking DE, Cutler DJ, Brune CW et al. (2008) A Common Genetic Variant in the Neurexin Superfamily Member CNTNAP2 Increases Familial Risk of Autism. Am J Hum Genet 82(1): 160-164.

Asano E, Chugani DC, Muzik O et al. (2001) Autism in tuberous sclerosis complex is related to both cortical and subcortical dysfunction. Neurology 57(7): 1269-1277.

Ashburner M, Ball CA, Blake JA et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25(1): 25-29.

Asperger H (1944) Die 'Autistischen Psychopathen' im Kindesalter. Archiv fur Psychiatrie und Nervenkrankheiten 117: 76-136. Translated to swedish in Frith U (ed) Autism och Aspergers syndrom. Liber, Stockholm 1998. pp 1953-1122.

Aulchenko YS, Ripatti S, Lindqvist I et al. (2009) Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. Nat Genet 41(1): 47-55.

Auranen M, Vanhala R, Varilo T et al. (2002) A genomewide screen for autism-spectrum disorders: evidence for a major susceptibility locus on chromosome 3q25-27. Am J Hum Genet 71(4): 777-790.

Bailey A, Palferman S, Heavey L et al. (1998a) Autism: the phenotype in relatives. J Autism Dev Disord 28(5): 369-392.

Bailey A, Le Couteur A, Gottesman I et al. (1995) Autism as a strongly genetic disorder: evidence from a British twin study. Psychol Med 25(1): 63-77.

Bailey DB, Jr., Mesibov GB, Hatton DD et al. (1998b) Autistic behavior in young boys with fragile X syndrome. J Autism Dev Disord 28(6): 499-508.

Baker P, Piven J, Sato Y (1998) Autism and tuberous sclerosis complex: prevalence and clinical features. J Autism Dev Disord 28(4): 279-285.

Bakkaloglu B, O'Roak BJ, Louvi A et al. (2008) Molecular Cytogenetic Analysis and Resequencing of Contactin Associated Protein-Like 2 in Autism Spectrum Disorders. Am J Hum Genet 82(1): 165-173.
Baron CA, Liu SY, Hicks C et al. (2006) Utilization of lymphoblastoid cell lines as a system for the molecular modeling of autism. J Autism Dev Disord 36(8): 973-982.
Barrett JC, Fry B, Maller J et al. (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics (Oxford, England) 21(2): 263-265.
Baulac S, Huberfeld G, Gourfinkel-An I et al. (2001) First genetic evidence of GABA(A) receptor dysfunction in epilepsy: a mutation in the gamma2-subunit gene. Nat Genet 28(1): 46-48.
Benayed R, Gharani N, Rossman I et al. (2005) Support for the homeobox transcription factor gene ENGRAILED 2 as an autism spectrum disorder susceptibility locus. Am J Hum Genet 77(5): 851-868.
Bilici M, Efe H, Koroglu MA et al. (2001) Antioxidative enzyme activities and lipid peroxidation in major depression: alterations by antidepressant treatments. Journal of affective disorders 64(1): 43-51.
Bodmer W, Bonilla C (2008) Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet 40(6): 695-701.
Bonaglia MC, Giorda R, Borgatti R et al. (2001) 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 69(2): 261-268.
Bonati MT, Russo S, Finelli P et al. (2007) Evaluation of autism traits in Angelman syndrome: a resource to unfold autism genes. Neurogenetics 8(3): 169-178.
Bonnen PE, Pe'er I, Plenge RM et al. (2006) Evaluating potential for whole-genome studies in Kosrae, an isolated population in Micronesia. Nat Genet 38(2): 214-217.
Bonora E, Beyer KS, Lamb JA et al. (2003) Analysis of reelin as a candidate gene for autism. Mol Psychiatry 8(10): 885-892.
Boomsma D, Busjahn A, Peltonen L (2002) Classical twin studies and beyond. Nat Rev Genet 3(11): 872-882.
Breitling R, Amtmann A, Herzyk P (2004a) Iterative Group Analysis (iGA): a simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. BMC bioinformatics 5: 34.
Breitling R, Armengaud P, Amtmann A et al. (2004b) Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS Lett 573(1-3): 83-92.
Breslin T, Eden P, Krogh M (2004) Comparing functional annotation analyses with Catmap. BMC bioinformatics 5: 193.
Burgoine E, Wing L (1983) Identical triplets with Asperger's syndrome. Br J Psychiatry 143: 261-265.
Butler MG, Dasouki MJ, Zhou XP et al. (2005) Subset of individuals with autism spectrum disorders and extreme macrocephaly associated with germline PTEN tumour suppressor gene mutations. J Med Genet 42(4): 318-321.
Buxbaum JD, Silverman J, Keddache M et al. (2004) Linkage analysis for autism in a subset families with obsessive-compulsive behaviors: evidence for an autism susceptibility gene on chromosome 1 and further support for susceptibility genes on chromosome 6 and 19. Mol Psychiatry 9(2): 144-150.
Buxbaum JD, Silverman JM, Smith CJ et al. (2001) Evidence for a susceptibility gene for autism on chromosome 2 and for genetic heterogeneity. Am J Hum Genet 68(6): 1514-1520.
Buxbaum JD, Silverman JM, Smith CJ et al. (2002) Association between a GABRB3 polymorphism and autism. Mol Psychiatry 7(3): 311-316.
Campbell DB, D'Oronzio R, Garbett K et al. (2007) Disruption of cerebral cortex MET signaling in autism spectrum disorder. Ann Neurol 62(3): 243-250.
Campbell DB, Sutcliffe JS, Ebert PJ et al. (2006) A genetic variant that disrupts MET transcription is associated with autism. Proc Natl Acad Sci U S A 103(45): 16834-16839.
Cantwell DP, Baker L, Rutter M (1979) Families of autistic and dysphasic children. I. Family life and interaction patterns. Arch Gen Psychiatry 36(6): 682-687.
Cenciarelli C, Chiaur DS, Guardavaccaro D et al. (1999) Identification of a family of human F-box proteins. Curr Biol 9(20): 1177-1179.
Chakrabarti S, Fombonne E (2001) Pervasive developmental disorders in preschool children. Jama 285(24): 3093-3099.
Chakravarti A (1999) Population genetics--making sense out of sequence. Nat Genet 21(1 Suppl): 56-60.
Charrow J (2004) Ashkenazi Jewish genetic disorders. Familial cancer 3(3-4): 201-206.
Cheh MA, Millonig JH, Roselli LM et al. (2006) En2 knockout mice display neurobehavioral and neurochemical alterations relevant to autism spectrum disorder. Brain Res 1116(1): 166-176.
Chen F, Wollmer MA, Hoerndli F et al. (2004) Role for glyoxalase I in Alzheimer's disease. Proc Natl Acad Sci U S A 101(20): 7687-7692.

Chen GK, Kono N, Geschwind DH et al. (2006) Quantitative trait locus analysis of nonverbal communication in autism spectrum disorder. Mol Psychiatry 11(2): 214-220.
Chotai J (1984) On the lod score method in linkage analysis. Ann Hum Genet 48 (Pt 4): 359-378.
Chugani DC (2004) Serotonin in autism and pediatric epilepsies. Ment Retard Dev Disabil Res Rev 10(2): 112-116.
Chugani DC, Muzik O, Behen M et al. (1999) Developmental changes in brain serotonin synthesis capacity in autistic and nonautistic children. Ann Neurol 45(3): 287-295.
Clamp M, Fry B, Kamal M et al. (2007) Distinguishing protein-coding and noncoding genes in the human genome. Proc Natl Acad Sci U S A 104(49): 19428-19433.
Clerget-Darpoux F, Bonaiti-Pellie C, Hochez J (1986) Effects of misspecifying genetic parameters in lod score analysis. Biometrics 42(2): 393-399.
Cohen IL, Liu X, Schutz C et al. (2003) Association of autism severity with a monoamine oxidase A functional polymorphism. Clin Genet 64(3): 190-197.
Collins FS, Guyer MS, Charkravarti A (1997) Variations on a theme: cataloging human DNA sequence variation. Science 278(5343): 1580-1581.
Constantino JN, Gruber CP, Davis S et al. (2004) The factor structure of autistic traits. J Child Psychol Psychiatry 45(4): 719-726.
Coo H, Ouellette-Kuntz H, Lloyd JE et al. (2008) Trends in autism prevalence: diagnostic substitution revisited. J Autism Dev Disord 38(6): 1036-1046.
Cook EH, Jr., Leventhal BL, Heller W et al. (1990) Autistic children and their first-degree relatives: relationships between serotonin and norepinephrine levels and intelligence. The Journal of neuropsychiatry and clinical neurosciences 2(3): 268-274.
Cook EH, Jr., Courchesne R, Lord C et al. (1997) Evidence of linkage between the serotonin transporter and autistic disorder. Mol Psychiatry 2(3): 247-250.
Cook EH, Jr., Courchesne RY, Cox NJ et al. (1998) Linkage-disequilibrium mapping of autistic disorder, with 15q11-13 markers. Am J Hum Genet 62(5): 1077-1083.
Coon H, Matsunami N, Stevens J et al. (2005) Evidence for linkage on chromosome 3q25-27 in a large autism extended pedigree. Hum Hered 60(4): 220-226.
Courchesne E, Yeung-Courchesne R, Press GA et al. (1988) Hypoplasia of cerebellar vermal lobules VI and VII in autism. N Engl J Med 318(21): 1349-1354.
Courchesne E, Karns CM, Davis HR et al. (2001) Unusual brain growth patterns in early life in patients with autistic disorder: an MRI study. Neurology 57(2): 245-254.
Davies PA, Pistis M, Hanna MC et al. (1999) The 5-HT3B subunit is a major determinant of serotonin-receptor function. Nature 397(6717): 359-363.
Davis LK, Hazlett HC, Librant AL et al. (2008) Cortical enlargement in autism is associated with a functional VNTR in the monoamine oxidase A gene. Am J Med Genet B Neuropsychiatr Genet 147B(7): 1145-1151.
Dawson G, Munson J, Webb SJ et al. (2007) Rate of head growth decelerates and symptoms worsen in the second year of life in autism. Biol Psychiatry 61(4): 458-464.
de Bakker PI, Yelensky R, Pe'er I et al. (2005) Efficiency and power in genetic association studies. Nat Genet 37(11): 1217-1223.
de Cid R, Riveira-Munoz E, Zeeuwen PL et al. (2009) Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. Nat Genet 41(2): 211-215.
de Vries BB, Pfundt R, Leisink M et al. (2005) Diagnostic genome profiling in mental retardation. Am J Hum Genet 77(4): 606-616.
Denney RM, Koch H, Craig IW (1999) Association between monoamine oxidase A activity in human male skin fibroblasts and genotype of the MAOA promoter-associated variable number tandem repeat. Hum Genet 105(6): 542-551.
Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55(4): 997-1004.
Devlin B, Cook EH, Jr., Coon H et al. (2005) Autism and the serotonin transporter: the long and short of it. Mol Psychiatry 10(12): 1110-1116.
Diskin SJ, Li M, Hou C et al. (2008) Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. Nucleic Acids Res 36(19): e126.
Dissanayake C, Bui QM, Huggins R et al. (2006) Growth in stature and head circumference in high-functioning autism and Asperger disorder during the first 3 years of life. Development and psychopathology 18(2): 381-393.

Donner J, Pirkola S, Silander K et al. (2008) An association analysis of murine anxiety genes in humans implicates novel candidate genes for anxiety disorders. Biol Psychiatry 64(8): 672-680.
Durand CM, Betancur C, Boeckers TM et al. (2007) Mutations in the gene encoding the synaptic scaffolding protein SHANK3 are associated with autism spectrum disorders. Nat Genet 39(1): 25-27.
Eagle RS (2004) Commentary: Further commentary on the debate regarding increase in autism in California. J Autism Dev Disord 34(1): 87-88.
Edwards AO, Ritter R, 3rd, Abel KJ et al. (2005) Complement factor H polymorphism and age-related macular degeneration. Science 308(5720): 421-424.
Ehlers S, Gillberg C (1993) The epidemiology of Asperger syndrome. A total population study. J Child Psychol Psychiatry 34(8): 1327-1350.
Elston RC (1998) Methods of linkage analysis--and the assumptions underlying them. Am J Hum Genet 63(4): 931-934.
Elston RC, Stewart J (1971) A general model for the genetic analysis of pedigree data. Hum Hered 21(6): 523-542.
Estivill X, Armengol L (2007) Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. PLoS Genet 3(10): 1787-1799.
Fanciulli M, Norsworthy PJ, Petretto E et al. (2007) FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. Nat Genet 39(6): 721-723.
Feng J, Schroer R, Yan J et al. (2006) High frequency of neurexin 1beta signal peptide structural variants in patients with autism. Neurosci Lett 409(1): 10-13.
Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. Nat Rev Genet 7(2): 85-97.
Finnila S, Lehtonen MS, Majamaa K (2001) Phylogenetic network for European mtDNA. Am J Hum Genet 68(6): 1475-1484.
Firth HV, Richards SM, Bevan AP et al. (2009) DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. Am J Hum Genet.
Fisher SE, Francks C, Marlow AJ et al. (2002) Independent genome-wide scans identify a chromosome 18 quantitative-trait locus influencing dyslexia. Nat Genet 30(1): 86-91.
Folstein S, Rutter M (1977) Infantile autism: a genetic study of 21 twin pairs. J Child Psychol Psychiatry 18(4): 297-321.
Folstein SE, Rosen-Sheidley B (2001) Genetics of autism: complex aetiology for a heterogeneous disorder. Nat Rev Genet 2(12): 943-955.
Fombonne E (2001) What is the prevalence of Asperger disorder? J Autism Dev Disord 31(3): 363-364.
Fombonne E (2003) Epidemiological surveys of autism and other pervasive developmental disorders: an update. J Autism Dev Disord 33(4): 365-382.
Fombonne E (2005) Epidemiology of autistic disorder and other pervasive developmental disorders. J Clin Psychiatry 66 Suppl 10: 3-8.
Fombonne E (2009) Epidemiology of pervasive developmental disorders. Pediatric research.
Fombonne E, Du Mazaubrun C, Cans C et al. (1997) Autism and associated medical disorders in a French epidemiological survey. J Am Acad Child Adolesc Psychiatry 36(11): 1561-1569.
Frazer KA, Ballinger DG, Cox DR et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449(7164): 851-861.
Gharani N, Benayed R, Mancuso V et al. (2004) Association of the homeobox transcription factor, ENGRAILED 2, 3, with autism spectrum disorder. Mol Psychiatry 9(5): 474-484.
Gibson G, Goldstein DB (2007) Human genetics: the hidden text of genome-wide associations. Curr Biol 17(21): R929-932.
Gillberg C (1989) Asperger syndrome in 23 Swedish children. Dev Med Child Neurol 31(4): 520-531.
Gillberg C (1994) Kilniska och neurobiologiska aspekter av Aspergers syndrom i sex familjestudier. In: Frith U, editor. Autism och Aspergers Syndrom. Stockholm: Liber. pp. 158-187.
Glessner JT, Wang K, Cai G et al. (2009) Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. Nature 459(7246): 569-573.
Goizet C, Excoffier E, Taine L et al. (2000) Case with autistic syndrome and chromosome 22q13.3 deletion detected by FISH. Am J Med Genet 96(6): 839-844.
Goldstein AM, Stacey SN, Olafsson JH et al. (2008) CDKN2A mutations and melanoma risk in the Icelandic population. J Med Genet 45(5): 284-289.
Gonzalez E, Kulkarni H, Bolivar H et al. (2005) The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science 307(5714): 1434-1440.

Goring HH, Terwilliger JD (2000) Linkage analysis in the presence of errors IV: joint pseudomarker analysis of linkage and/or linkage disequilibrium on a mixture of pedigrees and singletons when the mode of inheritance cannot be accurately specified. Am J Hum Genet 66(4): 1310-1327.

Goring HH, Terwilliger JD, Blangero J (2001) Large upward bias in estimation of locus-specific effects from genomewide scans. Am J Hum Genet 69(6): 1357-1369.

Grant SF, Thorleifsson G, Reynisdottir I et al. (2006) Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. Nat Genet 38(3): 320-323.

Greenberg DA, Abreu P, Hodge SE (1998) The power to detect linkage in complex disease by means of simple LOD-score analyses. Am J Hum Genet 63(3): 870-879.

Gregg JP, Lit L, Baron CA et al. (2008) Gene expression changes in children with autism. Genomics 91(1): 22-29.

Grinberg I (2005) Genetic analysis of Dandy Walker malformation and the role of Zic1 and Zic4 genes in cerebellar development Chichago: University of Chicago. 204 p.

Grinberg I, Northrup H, Ardinger H et al. (2004) Heterozygous deletion of the linked genes ZIC1 and ZIC4 is involved in Dandy-Walker malformation. Nat Genet 36(10): 1053-1055.

Gudmundsson J, Sulem P, Manolescu A et al. (2007a) Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. Nat Genet 39(5): 631-637.

Gudmundsson J, Sulem P, Rafnar T et al. (2008) Common sequence variants on 2p15 and Xp11.22 confer susceptibility to prostate cancer. Nat Genet 40(3): 281-283.

Gudmundsson J, Sulem P, Steinthorsdottir V et al. (2007b) Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. Nat Genet 39(8): 977-983.

Haines JL, Hauser MA, Schmidt S et al. (2005) Complement factor H variant increases the risk of age-related macular degeneration. Science 308(5720): 419-421.

Hashimoto T, Tayama M, Murakawa K et al. (1995) Development of the brainstem and cerebellum in autistic patients. J Autism Dev Disord 25(1): 1-18.

Hatton DD, Sideris J, Skinner M et al. (2006) Autistic behavior in children with fragile X syndrome: prevalence, stability, and the impact of FMRP. Am J Med Genet A 140A(17): 1804-1813.

Heath SC, Gut IG, Brennan P et al. (2008) Investigation of the fine structure of European populations with applications to disease association studies. Eur J Hum Genet 16(12): 1413-1429.

Hedman M, Brandstatter A, Pimenoff V et al. (2007) Finnish mitochondrial DNA HVS-I and HVS-II population data. Forensic science international 172(2-3): 171-178.

Helgadottir A, Thorleifsson G, Manolescu A et al. (2007) A common variant on chromosome 9p21 affects the risk of myocardial infarction. Science 316(5830): 1491-1493.

Helgadottir A, Thorleifsson G, Magnusson KP et al. (2008) The same sequence variant on 9p21 associates with myocardial infarction, abdominal aortic aneurysm and intracranial aneurysm. Nat Genet 40(2): 217-224.

Helgason A, Yngvadottir B, Hrafnkelsson B et al. (2005) An Icelandic example of the impact of population structure on association studies. Nat Genet 37(1): 90-95.

Hiekkalinna T, Terwilliger JD, Sammalisto S et al. (2005) AUTOGSCAN: powerful tools for automated genome-wide linkage and linkage disequilibrium analysis. Twin Res Hum Genet 8(1): 16-21.

Hodge SE, Abreu PC, Greenberg DA (1997) Magnitude of type I error when single-locus linkage analysis is maximized over models: a simulation study. Am J Hum Genet 60(1): 217-227.

Hogart A, Leung KN, Wang NJ et al. (2009) Chromosome 15q11-13 duplication syndrome brain reveals epigenetic alterations in gene expression not predicted from copy number. J Med Genet 46(2): 86-93.

Hollox EJ, Huffmeier U, Zeeuwen PL et al. (2008) Psoriasis is associated with increased beta-defensin genomic copy number. Nat Genet 40(1): 23-25.

Horvath S, Xu X, Laird NM (2001) The family based association test method: strategies for studying general genotype--phenotype associations. Eur J Hum Genet 9(4): 301-306.

Houwen RH, Baharloo S, Blankenship K et al. (1994) Genome screening by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. Nat Genet 8(4): 380-386.

Hovatta I, Terwilliger JD, Lichtermann D et al. (1997) Schizophrenia in the genetic isolate of Finland. Am J Med Genet 74(4): 353-360.

Hovatta I, Tennant RS, Helton R et al. (2005) Glyoxalase 1 and glutathione reductase 1 regulate anxiety in mice. Nature 438(7068): 662-666.

Hu VW, Frank BC, Heine S et al. (2006) Gene expression profiling of lymphoblastoid cell lines from monozygotic twins discordant in severity of autism reveals differential regulation of neurologically relevant genes. BMC Genomics 7(1): 118.

Huang CH, Santangelo SL (2008) Autism and serotonin transporter gene polymorphisms: a systematic review and meta-analysis. Am J Med Genet B Neuropsychiatr Genet 147B(6): 903-913.
Iafrate AJ, Feuk L, Rivera MN et al. (2004) Detection of large-scale variation in the human genome. Nat Genet 36(9): 949-951.
Iidaka T, Ozaki N, Matsumoto A et al. (2005) A variant C178T in the regulatory region of the serotonin receptor gene HTR3A modulates neural activation in the human amygdala. J Neurosci 25(27): 6460-6466.
Itsara A, Cooper GM, Baker C et al. (2009) Population Analysis of Large Copy Number Variants and Hotspots of Human Genetic Disease. Am J Hum Genet.
Iwamoto K, Kakiuchi C, Bundo M et al. (2004) Molecular characterization of bipolar disorder by comparing gene expression profiles of postmortem brains of major mental disorders. Mol Psychiatry 9(4): 406-416.
Jacquemont ML, Sanlaville D, Redon R et al. (2006) Array- based comparative genomic hybridization identifies high frequency of cryptic chromosomal rearrangements in patients with syndromic autism spectrum disorders. J Med Genet.
Jakobsson M, Scholz SW, Scheet P et al. (2008) Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451(7181): 998-1003.
Jamain S, Betancur C, Quach H et al. (2002) Linkage and association of the glutamate receptor 6 gene with autism. Mol Psychiatry 7(3): 302-310.
Jamain S, Quach H, Betancur C et al. (2003) Mutations of the X-linked genes encoding neuroligins NLGN3 and NLGN4 are associated with autism. Nat Genet 34(1): 27-29.
Jones MB, Palmour RM, Zwaigenbaum L et al. (2004) Modifier effects in autism at the MAO-A and DBH loci. Am J Med Genet B Neuropsychiatr Genet 126(1): 58-65.
Joober R, Boksa P (2009) A new wave in the genetics of psychiatric disorders: the copy number variant tsunami. J Psychiatry Neurosci 34(1): 55-59.
Jorde LB, Hasstedt SJ, Ritvo ER et al. (1991) Complex segregation analysis of autism. Am J Hum Genet 49(5): 932-938.
Junaid MA, Kowal D, Barua M et al. (2004) Proteomic studies identified a single nucleotide polymorphism in glyoxalase I as autism susceptibility factor. Am J Med Genet A 131(1): 11-17.
Kallio SP, Jakkula E, Purcell S et al. (2009) Use of a Genetic Isolate to Identify Rare Disease Variants: C7 on 5p associated with MS. Hum Mol Genet.
Kanner L (1943) Autistic disturbances of affective contact. Nervous Child 2: 217-250.
Karnovsky AM, Gotow LF, McKinley DD et al. (2003) A cluster of novel serotonin receptor 3-like genes on human chromosome 3. Gene 319: 137-148.
Kaufmann WE, Moser HW (2000) Dendritic anomalies in disorders associated with mental retardation. Cereb Cortex 10(10): 981-991.
Kaufmann WE, Cooper KL, Mostofsky SH et al. (2003) Specificity of cerebellar vermian abnormalities in autism: a quantitative magnetic resonance imaging study. J Child Neurol 18(7): 463-470.
Kerbeshian J, Burd L (1986) Asperger's syndrome and Tourette syndrome: the case of the pinball wizard. Br J Psychiatry 148: 731-736.
Kielinen M, Linna SL, Moilanen I (2000) Autism in Northern Finland. Eur Child Adolesc Psychiatry 9(3): 162-167.
Kilpinen H, Ylisaukko-Oja T, Hennah W et al. (2008) Association of DISC1 with autism and Asperger syndrome. Mol Psychiatry 13(2): 187-196.
Kim SA, Kim JH, Park M et al. (2006) Association of GABRB3 polymorphisms with autism spectrum disorders in Korean trios. Neuropsychobiology 54(3): 160-165.
King MC, Marks JH, Mandell JB (2003) Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. Science 302(5645): 643-646.
Kissebah AH, Sonnenberg GE, Myklebust J et al. (2000) Quantitative trait loci on chromosomes 3 and 17 influence phenotypes of the metabolic syndrome. Proc Natl Acad Sci U S A 97(26): 14478-14483.
Kittles RA, Bergen AW, Urbanek M et al. (1999) Autosomal, mitochondrial, and Y chromosome DNA variation in Finland: evidence for a male-specific bottleneck. American journal of physical anthropology 108(4): 381-399.
Kittles RA, Perola M, Peltonen L et al. (1998) Dual origins of Finns revealed by Y chromosome haplotype variation. Am J Hum Genet 62(5): 1171-1179.
Klein RJ, Zeiss C, Chew EY et al. (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308(5720): 385-389.

Kleinhans NM, Richards T, Sterling L et al. (2008) Abnormal functional connectivity in autism spectrum disorders during face processing. Brain 131(Pt 4): 1000-1012.
Kleppe K, Ohtsuka E, Kleppe R et al. (1971) Studies on polynucleotides. XCVI. Repair replications of short synthetic DNA's as catalyzed by DNA polymerases. Journal of molecular biology 56(2): 341-361.
Klin A, Sparrow SS, de Bildt A et al. (1999) A normed study of face recognition in autism and related disorders. J Autism Dev Disord 29(6): 499-508.
Kolevzon A, Mathewson KA, Hollander E (2006) Selective serotonin reuptake inhibitors in autism: a review of efficacy and tolerability. J Clin Psychiatry 67(3): 407-414.
Kong A, Gudbjartsson DF, Sainz J et al. (2002) A high-resolution recombination map of the human genome. Nat Genet 31(3): 241-247.
Kosik KS (2009) Exploring the early origins of the synapse by comparative genomics. Biology letters 5(1): 108-111.
Kracke I (1994) Developmental prosopagnosia in Asperger syndrome: presentation and discussion of an individual case. Dev Med Child Neurol 36(10): 873-886.
Kromer SA, Kessler MS, Milfay D et al. (2005) Identification of glyoxalase-I as a protein marker in a mouse model of extremes in trait anxiety. J Neurosci 25(17): 4375-4384.
Kruglyak L, Nickerson DA (2001) Variation is the spice of life. Nat Genet 27(3): 234-236.
Kruglyak L, Daly MJ, Reeve-Daly MP et al. (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet 58(6): 1347-1363.
Kumar RA, Karamohamed S, Sudi J et al. (2008) Recurrent 16p11.2 microdeletions in autism. Hum Mol Genet 17(4): 628-638.
Kuperman S, Beeghly J, Burns T et al. (1987) Association of serotonin concentration to behavior and IQ in autistic children. J Autism Dev Disord 17(1): 133-140.
Kwasnicka-Crawford DA, Carson AR, Roberts W et al. (2005) Characterization of a novel cation transporter ATPase gene (ATP13A4) interrupted by 3q25-q29 inversion in an individual with language delay. Genomics 86(2): 182-194.
Lamb JA, Barnby G, Bonora E et al. (2005) Analysis of IMGSAC autism susceptibility loci: evidence for sex limited and parent of origin specific effects. J Med Genet 42(2): 132-137.
Lander E, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11(3): 241-247.
Lander ES (1996) The new genomics: global views of biology. Science 274(5287): 536-539.
Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci U S A 84(8): 2363-2367.
Lander ES, Linton LM, Birren B et al. (2001) Initial sequencing and analysis of the human genome. Nature 409(6822): 860-921.
Lao O, Lu TT, Nothnagel M et al. (2008) Correlation between genetic and geographic structure in Europe. Curr Biol 18(16): 1241-1248.
Lappalainen T, Koivumaki S, Salmela E et al. (2006) Regional differences among the Finns: a Y-chromosomal perspective. Gene 376(2): 207-215.
Lathrop GM, Lalouel JM, White RL (1986) Construction of human linkage maps: likelihood calculations for multilocus linkage analysis. Genet Epidemiol 3(1): 39-52.
Lathrop GM, Lalouel JM, Julier C et al. (1984) Strategies for multilocus linkage analysis in humans. Proc Natl Acad Sci U S A 81(11): 3443-3446.
Laumonnier F, Bonnet-Brilhault F, Gomot M et al. (2004) X-linked mental retardation and autism are associated with a mutation in the NLGN4 gene, a member of the neuroligin family. Am J Hum Genet 74(3): 552-557.
Lawson-Yuen A, Saldivar JS, Sommer S et al. (2008) Familial deletion within NLGN4 associated with autism and Tourette syndrome. Eur J Hum Genet 16(5): 614-618.
Leboyer M, Philippe A, Bouvard M et al. (1999) Whole blood serotonin and plasma beta-endorphin in autistic probands and their first-degree relatives. Biol Psychiatry 45(2): 158-163.
Lettre G, Jackson AU, Gieger C et al. (2008) Identification of ten loci associated with height highlights new biological pathways in human growth. Nat Genet 40(5): 584-591.
Lewis AS, Schwartz E, Chan CS et al. (2009) Alternatively spliced isoforms of TRIP8b differentially control h channel trafficking and function. J Neurosci 29(19): 6250-6265.
Levy S, Sutton G, Ng PC et al. (2007) The diploid genome sequence of an individual human. PLoS Biol 5(10): e254.

Li C, Scott LJ, Boehnke M (2004) Assessing whether an allele can account in part for a linkage signal: the Genotype-IBD Sharing Test (GIST). Am J Hum Genet 74(3): 418-431.
Li M, Atmaca-Sonmez P, Othman M et al. (2006) CFH haplotypes without the Y402H coding variant show strong association with susceptibility to age-related macular degeneration. Nat Genet 38(9): 1049-1054.
Liu P, Keller JR, Ortiz M et al. (2003) Bcl11a is essential for normal lymphoid development. Nature immunology 4(6): 525-532.
Locke DP, Sharp AJ, McCarroll SA et al. (2006) Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. Am J Hum Genet 79(2): 275-290.
Lockyer L, Rutter M (1969) A five- to fifteen-year follow-up study of infantile psychosis. Br J Psychiatry 115(525): 865-882.
Losh M, Childress D, Lam K et al. (2008) Defining key features of the broad autism phenotype: a comparison across parents of multiple- and single-incidence autism families. Am J Med Genet B Neuropsychiatr Genet 147B(4): 424-433.
Lowe JK, Maller JB, Pe'er I et al. (2009) Genome-wide association studies in an isolated founder population from the Pacific Island of Kosrae. PLoS Genet 5(2): e1000365.
Macarov M, Zeigler M, Newman JP et al. (2007) Deletions of VCX-A and NLGN4: a variable phenotype including normal intellect. J Intellect Disabil Res 51(Pt 5): 329-333.
Maestrini E, Lai C, Marlow A et al. (1999) Serotonin transporter (5-HTT) and gamma-aminobutyric acid receptor subunit beta3 (GABRB3) gene polymorphisms are not associated with autism in the IMGSA families. The International Molecular Genetic Study of Autism Consortium. Am J Med Genet 88(5): 492-496.
Malan V, Raoul O, Firth HV et al. (2009) 19q13.11 deletion syndrome: a novel clinically recognizable genetic condition identified by array-CGH. J Med Genet.
Maniatis N, Collins A, Xu CF et al. (2002) The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. Proc Natl Acad Sci U S A 99(4): 2228-2233.
Manning MA, Cassidy SB, Clericuzio C et al. (2004) Terminal 22q deletion syndrome: a newly recognized cause of speech and language disability in the autism spectrum. Pediatrics 114(2): 451-457.
Marshall CR, Noor A, Vincent JB et al. (2008) Structural variation of chromosomes in autism spectrum disorder. Am J Hum Genet 82(2): 477-488.
Martin ER, Menold MM, Wolpert CM et al. (2000) Analysis of linkage disequilibrium in gamma-aminobutyric acid receptor subunit genes in autistic disorder. Am J Med Genet 96(1): 43-48.
Matarazzo V, Cohen D, Palmer AM et al. (2004) The transcriptional repressor Mecp2 regulates terminal neuronal differentiation. Molecular and cellular neurosciences 27(1): 44-58.
Mattila ML, Kielinen M, Jussila K et al. (2007) An epidemiological and diagnostic study of Asperger syndrome according to four sets of diagnostic criteria. J Am Acad Child Adolesc Psychiatry 46(5): 636-646.
McBride PA, Anderson GM, Hertzig ME et al. (1998) Effects of diagnosis, race, and puberty on platelet serotonin levels in autism and mental retardation. J Am Acad Child Adolesc Psychiatry 37(7): 767-776.
McCarroll SA, Hadnott TN, Perry GH et al. (2006) Common deletion polymorphisms in the human genome. Nat Genet 38(1): 86-92.
McCarroll SA, Huett A, Kuballa P et al. (2008a) Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. Nat Genet.
McCarroll SA, Kuruvilla FG, Korn JM et al. (2008b) Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet 40(10): 1166-1174.
McCauley JL, Olson LM, Delahanty R et al. (2004) A linkage disequilibrium map of the 1-Mb 15q12 GABA(A) receptor subunit cluster and association to autism. Am J Med Genet B Neuropsychiatr Genet 131B(1): 51-59.
McDougle CJ, Naylor ST, Cohen DJ et al. (1996a) Effects of tryptophan depletion in drug-free adults with autistic disorder. Arch Gen Psychiatry 53(11): 993-1000.
McDougle CJ, Naylor ST, Cohen DJ et al. (1996b) A double-blind, placebo-controlled study of fluvoxamine in adults with autistic disorder. Arch Gen Psychiatry 53(11): 1001-1008.
McKinney C, Merriman ME, Chapman PT et al. (2008) Evidence for an influence of chemokine ligand 3-like 1 (CCL3L1) gene copy number on susceptibility to rheumatoid arthritis. Annals of the rheumatic diseases 67(3): 409-413.

Meetei AR, de Winter JP, Medhurst AL et al. (2003) A novel ubiquitin ligase is deficient in Fanconi anemia. Nat Genet 35(2): 165-170.
Miles JH, Takahashi TN, Bagby S et al. (2005) Essential versus complex autism: definition of fundamental prognostic subtypes. Am J Med Genet A 135(2): 171-180.
Miyake A, Mochizuki S, Takemoto Y et al. (1995) Molecular cloning of human 5-hydroxytryptamine3 receptor: heterogeneity in distribution and function among species. Molecular pharmacology 48(3): 407-416.
Morrow EM, Yoo SY, Flavell SW et al. (2008) Identifying autism loci and genes by tracing recent shared ancestry. Science 321(5886): 218-223.
Morton DH, Morton CS, Strauss KA et al. (2003) Pediatric medicine and the genetic disorders of the Amish and Mennonite people of Pennsylvania. Am J Med Genet C Semin Med Genet 121C(1): 5-17.
Muhle R, Trentacoste SV, Rapin I (2004) The genetics of autism. Pediatrics 113(5): e472-486.
Murphy KC, Jones LA, Owen MJ (1999) High rates of schizophrenia in adults with velo-cardio-facial syndrome. Arch Gen Psychiatry 56(10): 940-945.
Murray JC, Johnson JA, Bird TD (1985) Dandy-Walker malformation: etiologic heterogeneity and empiric recurrence risks. Clin Genet 28(4): 272-283.
Nabi R, Serajee FJ, Chugani DC et al. (2004) Association of tryptophan 2,3 dioxygenase gene polymorphism with autism. Am J Med Genet B Neuropsychiatr Genet 125(1): 63-68.
Newschaffer CJ, Falb MD, Gurney JG (2005) National autism prevalence trends from United States special education data. Pediatrics 115(3): e277-282.
Nguyen DQ, Webber C, Ponting CP (2006) Bias of selection on human copy-number variants. PLoS Genet 2(2): e20.
Niesler B, Frank B, Kapeller J et al. (2003) Cloning, physical mapping and expression analysis of the human 5-HT3 serotonin receptor-like genes HTR3C, HTR3D and HTR3E. Gene 310: 101-111.
Niesler B, Kapeller J, Hammer C et al. (2008) Serotonin type 3 receptor genes: HTR3A, B, C, D, E. Pharmacogenomics 9(5): 501-504.
Niesler B, Walstab J, Combrink S et al. (2007) Characterization of the novel human serotonin receptor subunits 5-HT3C,5-HT3D, and 5-HT3E. Molecular pharmacology 72(1): 8-17.
Nigro V, de Sa Moreira E, Piluso G et al. (1996) Autosomal recessive limb-girdle muscular dystrophy, LGMD2F, is caused by a mutation in the delta-sarcoglycan gene. Nat Genet 14(2): 195-198.
Niklasson L, Rasmussen P, Oskarsdottir S et al. (2001) Neuropsychiatric disorders in the 22q11 deletion syndrome. Genet Med 3(1): 79-84.
Nishimura Y, Martin CL, Lopez AV et al. (2007) Genome-wide expression profiling of lymphoblastoid cell lines distinguishes different forms of autism and reveals shared pathways. Hum Mol Genet 16(14): 1682-1689.
Norio R (2003a) Finnish Disease Heritage I: characteristics, causes, background. Hum Genet 112(5-6): 441-456.
Norio R (2003b) The Finnish Disease Heritage III: the individual diseases. Hum Genet 112(5-6): 470-526.
Norio R (2003c) Finnish Disease Heritage II: population prehistory and genetic roots of Finns. Hum Genet 112(5-6): 457-469.
Notomi T, Shigemoto R (2004) Immunohistochemical localization of Ih channel subunits, HCN1-4, in the rat brain. The Journal of comparative neurology 471(3): 241-276.
Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. Nature genetics 40(5): 646-649.
Nurmi EL, Dowd M, Tadevosyan-Leyfer O et al. (2003) Exploratory subsetting of autism families based on savant skills improves evidence of genetic linkage to 15q11-q13. J Am Acad Child Adolesc Psychiatry 42(7): 856-863.
Nyholt DR (2000) All LODs are not created equal. Am J Hum Genet 67(2): 282-288.
O'Connell JR, Weeks DE (1998) PedCheck: a program for identification of genotype incompatibilities in linkage analysis. Am J Hum Genet 63(1): 259-266.
Ott J (1986) Linkage probability and its approximate confidence interval under possible heterogeneity. Genet Epidemiol Suppl 1: 251-257.
Ozonoff S, Williams BJ, Gale S et al. (1999) Autism and autistic behavior in Joubert syndrome. J Child Neurol 14(10): 636-641.
Palmen SJ, van Engeland H, Hof PR et al. (2004) Neuropathological findings in autism. Brain 127(Pt 12): 2572-2583.

Pastinen T, Raitio M, Lindroos K et al. (2000) A system for specific, high-throughput genotyping by allele-specific primer extension on microarrays. Genome Res 10(7): 1031-1042.
Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2(12): e190.
Persico AM, D'Agruma L, Maiorano N et al. (2001) Reelin gene alleles and haplotypes as a factor predisposing to autistic disorder. Mol Psychiatry 6(2): 150-159.
Peters SU, Beaudet AL, Madduri N et al. (2004) Autism in Angelman syndrome: implications for autism research. Clin Genet 66(6): 530-536.
Pickles A, Bolton P, Macdonald H et al. (1995) Latent-class analysis of recurrence risks for complex phenotypes with selection and measurement error: a twin and family history study of autism. Am J Hum Genet 57(3): 717-726.
Pietilainen KH, Naukkarinen J, Rissanen A et al. (2008) Global transcript profiles of fat in monozygotic twins discordant for BMI: pathways behind acquired obesity. PLoS medicine 5(3): e51.
Price AL, Patterson NJ, Plenge RM et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38(8): 904-909.
Price AL, Butler J, Patterson N et al. (2008) Discerning the ancestry of European Americans in genetic association studies. PLoS Genet 4(1): e236.
Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69(1): 124-137.
Pritchard JK, Przeworski M (2001) Linkage disequilibrium in humans: models and data. Am J Hum Genet 69(1): 1-14.
Puffenberger EG, Kauffman ER, Bolk S et al. (1994) Identity-by-descent and association mapping of a recessive gene for Hirschsprung disease on human chromosome 13q22. Hum Mol Genet 3(8): 1217-1225.
Purcell AE, Jeon OH, Zimmerman AW et al. (2001) Postmortem brain abnormalities of the glutamate neurotransmitter system in autism. Neurology 57(9): 1618-1628.
Purcell S, Neale B, Todd-Brown K et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3): 559-575.
Rafnar T, Sulem P, Stacey SN et al. (2009) Sequence variants at the TERT-CLPTM1L locus associate with many cancer types. Nat Genet 41(2): 221-227.
Ramensky V, Bork P, Sunyaev S (2002) Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30(17): 3894-3900.
Ramoz N, Reichert JG, Smith CJ et al. (2004) Linkage and association of the mitochondrial aspartate/glutamate carrier SLC25A12 gene with autism. Am J Psychiatry 161(4): 662-669.
Reich DE, Lander ES (2001) On the allelic spectrum of human disease. Trends Genet 17(9): 502-510.
Rioux JD, Xavier RJ, Taylor KD et al. (2007) Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. Nat Genet 39(5): 596-604.
Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273(5281): 1516-1517.
Risch N, Spiker D, Lotspeich L et al. (1999) A genomic screen of autism: evidence for a multilocus etiology. Am J Hum Genet 65(2): 493-507.
Ritvo ER, Freeman BJ, Mason-Brothers A et al. (1985) Concordance for the syndrome of autism in 40 pairs of afflicted twins. Am J Psychiatry 142(1): 74-77.
Ritvo ER, Jorde LB, Mason-Brothers A et al. (1989) The UCLA-University of Utah epidemiologic survey of autism: recurrence risk estimates and genetic counseling. Am J Psychiatry 146(8): 1032-1036.
Roberts SB, MacLean CJ, Neale MC et al. (1999) Replication of linkage studies of complex traits: an examination of variation in location estimates. Am J Hum Genet 65(3): 876-884.
Rogers SJ, Wehner DE, Hagerman R (2001) The behavioral phenotype in fragile X: symptoms of autism in very young children with fragile X syndrome, idiopathic autism, and other developmental disorders. J Dev Behav Pediatr 22(6): 409-417.
Ronald A, Happe F, Price TS et al. (2006a) Phenotypic and genetic overlap between autistic traits at the extremes of the general population. J Am Acad Child Adolesc Psychiatry 45(10): 1206-1214.
Ronald A, Happe F, Bolton P et al. (2006b) Genetic heterogeneity between the three components of the autism spectrum: a twin study. J Am Acad Child Adolesc Psychiatry 45(6): 691-699.
Rutter M (1968) Concepts of autism: a review of research. J Child Psychol Psychiatry 9(1): 1-25.
Sabatti C, Service SK, Hartikainen AL et al. (2008) Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. Nat Genet.

Sabol SZ, Hu S, Hamer D (1998) A functional polymorphism in the monoamine oxidase A gene promoter. Hum Genet 103(3): 273-279.
Sacco R, Papaleo V, Hager J et al. (2007) Case-control and family-based association studies of candidate genes in autistic disorder and its endophenotypes: TPH2 and GLO1. BMC Med Genet 8(1): 11.
Sadakata T, Washida M, Iwayama Y et al. (2007) Autistic-like phenotypes in Cadps2-knockout mice and aberrant CADPS2 splicing in autistic patients. J Clin Invest 117(4): 931-943.
Sajantila A, Salem AH, Savolainen P et al. (1996) Paternal and maternal DNA lineages reveal a bottleneck in the founding of the Finnish population. Proc Natl Acad Sci U S A 93(21): 12035-12039.
Sajantila A, Lahermo P, Anttinen T et al. (1995) Genes and languages in Europe: an analysis of mitochondrial lineages. Genome Res 5(1): 42-52.
Salmela E, Lappalainen T, Fransson I et al. (2008) Genome-wide analysis of single nucleotide polymorphisms uncovers population structure in Northern Europe. PLoS ONE 3(10): e3519.
Sanger F, Coulson AR (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. Journal of molecular biology 94(3): 441-448.
Sankaran VG, Menne TF, Xu J et al. (2008) Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor BCL11A. Science 322(5909): 1839-1842.
Santoro B, Wainger BJ, Siegelbaum SA (2004) Regulation of HCN channel surface expression by a novel C-terminal protein-protein interaction. J Neurosci 24(47): 10750-10762.
Satterwhite E, Sonoki T, Willis TG et al. (2001) The BCL11 gene family: involvement of BCL11A in lymphoid malignancies. Blood 98(12): 3413-3420.
Saxena R, Voight BF, Lyssenko V et al. (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 316(5829): 1331-1336.
Schain RJ, Freedman DX (1961) Studies on 5-hydroxyindole metabolism in autistic and other mentally retarded children. The Journal of pediatrics 58: 315-320.
Schellenberg GD, Dawson G, Sung YJ et al. (2006) Evidence for multiple loci from a genome scan of autism kindreds. Mol Psychiatry 11(11): 1049-1060, 1979.
Schultz RT (2005) Developmental deficits in social perception in autism: the role of the amygdala and fusiform face area. Int J Dev Neurosci 23(2-3): 125-141.
Scott MM, Deneris ES (2005) Making and breaking serotonin neurons and autism. Int J Dev Neurosci 23(2-3): 277-285.
Sebat J, Lakshmi B, Troge J et al. (2004) Large-scale copy number polymorphism in the human genome. Science 305(5683): 525-528.
Sebat J, Lakshmi B, Malhotra D et al. (2007) Strong Association of De Novo Copy Number Mutations with Autism. Science 316(5823): 445-449.
Service S, Sabatti C, Freimer N (2007) Tag SNPs chosen from HapMap perform well in several population isolates. Genet Epidemiol 31(3): 189-194.
Service S, Deyoung J, Karayiorgou M et al. (2006) Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. Nat Genet 38(5): 556-560.
Shao Y, Raiford KL, Wolpert CM et al. (2002) Phenotypic homogeneity provides increased support for linkage on chromosome 2 in autistic disorder. Am J Hum Genet 70(4): 1058-1061.
Shattuck PT (2006) The contribution of diagnostic substitution to the growing administrative prevalence of autism in US special education. Pediatrics 117(4): 1028-1037.
Shaw-Smith C, Pittman AM, Willatt L et al. (2006) Microdeletion encompassing MAPT at chromosome 17q21.3 is associated with developmental delay and learning disability. Nat Genet 38(9): 1032-1037.
Shifman S, Darvasi A (2001) The value of isolated populations. Nat Genet 28(4): 309-310.
Shmulewitz D, Auerbach SB, Lehner T et al. (2001) Epidemiology and factor analysis of obesity, type II diabetes, hypertension, and dyslipidemia (syndrome X) on the Island of Kosrae, Federated States of Micronesia. Hum Hered 51(1-2): 8-19.
Shprintzen RJ (2000) Velo-cardio-facial syndrome: a distinctive behavioral phenotype. Ment Retard Dev Disabil Res Rev 6(2): 142-147.
Simon-Sanchez J, Scholz S, Fung HC et al. (2007) Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals. Hum Mol Genet 16(1): 1-14.
Skaar DA, Shao Y, Haines JL et al. (2005) Analysis of the RELN gene as a genetic risk factor for autism. Mol Psychiatry 10(6): 563-571.
Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical applications in genetics and molecular biology 3: Article3.

Splawski I, Timothy KW, Sharpe LM et al. (2004) Ca(V)1.2 calcium channel dysfunction causes a multisystem disorder including arrhythmia and autism. Cell 119(1): 19-31.
Stacey SN, Gudbjartsson DF, Sulem P et al. (2008) Common variants on 1p36 and 1q42 are associated with cutaneous basal cell carcinoma but not with melanoma or pigmentation traits. Nat Genet 40(11): 1313-1318.
Stacey SN, Manolescu A, Sulem P et al. (2007) Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. Nat Genet 39(7): 865-869.
Stefansson H, Rujescu D, Cichon S et al. (2008) Large recurrent microdeletions associated with schizophrenia. Nature.
Steffenburg S, Gillberg C, Hellgren L et al. (1989) A twin study of autism in Denmark, Finland, Iceland, Norway and Sweden. J Child Psychol Psychiatry 30(3): 405-416.
Stone JL, Merriman B, Cantor RM et al. (2004) Evidence for sex-specific risk alleles in autism spectrum disorder. Am J Hum Genet 75(6): 1117-1123.
Stranger BE, Forrest MS, Dunning M et al. (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315(5813): 848-853.
Strauch K, Fimmers R, Kurz T et al. (2000) Parametric and nonparametric multipoint linkage analysis with imprinting and two-locus-trait models: application to mite sensitization. Am J Hum Genet 66(6): 1945-1957.
Strauss KA, Puffenberger EG, Huentelman MJ et al. (2006) Recessive symptomatic focal epilepsy and mutant contactin-associated protein-like 2. N Engl J Med 354(13): 1370-1377.
Sulem P, Gudbjartsson DF, Stacey SN et al. (2008) Two newly identified genetic determinants of pigmentation in Europeans. Nat Genet 40(7): 835-837.
Sumelahti ML, Tienari PJ, Wikstrom J et al. (2001) Increasing prevalence of multiple sclerosis in Finland. Acta neurologica Scandinavica 103(3): 153-158.
Sung YJ, Dawson G, Munson J et al. (2005) Genetic investigation of quantitative traits related to autism: use of multivariate polygenic models with ascertainment adjustment. Am J Hum Genet 76(1): 68-81.
Szatmari P, Jones MB, Zwaigenbaum L et al. (1998) Genetics of autism: overview and new directions. J Autism Dev Disord 28(5): 351-368.
Szatmari P, MacLean JE, Jones MB et al. (2000) The familial aggregation of the lesser variant in biological and nonbiological relatives of PDD probands: a family history study. J Child Psychol Psychiatry 41(5): 579-586.
Szatmari P, Paterson AD, Zwaigenbaum L et al. (2007) Mapping autism risk loci using genetic linkage and chromosomal rearrangements. Nat Genet 39(3): 319-328.
Tafti M, Petit B, Chollet D et al. (2003) Deficiency in short-chain fatty acid beta-oxidation affects theta oscillations during sleep. Nat Genet 34(3): 320-325.
Tang Y, Lu A, Aronow BJ et al. (2001) Blood genomic responses differ after stroke, seizures, hypoglycemia, and hypoxia: blood genomic fingerprints of disease. Ann Neurol 50(6): 699-707.
Tang YP, Shimizu E, Dube GR et al. (1999) Genetic enhancement of learning and memory in mice. Nature 401(6748): 63-69.
Taniura H, Iijima S, Kambe Y et al. (2007) Tex261 modulates the excitotoxic cell death induced by N-methyl-D-aspartate (NMDA) receptor activation. Biochem Biophys Res Commun 362(4): 1096-1100.
Terwilliger JD, Hiekkalinna T (2006) An utter refutation of the "Fundamental Theorem of the HapMap". Eur J Hum Genet 14(4): 426-437.
Thakkinstian A, Han P, McEvoy M et al. (2006) Systematic review and meta-analysis of the association between complement factor H Y402H polymorphisms and age-related macular degeneration. Hum Mol Genet 15(18): 2784-2790.
The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437(7063): 1299-1320.
Thomas NS, Sharp AJ, Browne CE et al. (1999) Xp deletions associated with autism in three females. Hum Genet 104(1): 43-48.
Thorleifsson G, Magnusson KP, Sulem P et al. (2007) Common sequence variants in the LOXL1 gene confer susceptibility to exfoliation glaucoma. Science 317(5843): 1397-1400.
Thorleifsson G, Walters GB, Gudbjartsson DF et al. (2009) Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. Nat Genet 41(1): 18-24.
Thornalley PJ (2003) Glyoxalase I--structure, function and a critical role in the enzymatic defence against glycation. Biochem Soc Trans 31(Pt 6): 1343-1348.

Tian C, Plenge RM, Ransom M et al. (2008) Analysis and application of European genetic substructure using 300 K SNP information. PLoS Genet 4(1): e4.
Tierney E, Nwokoro NA, Porter FD et al. (2001) Behavior phenotype in the RSH/Smith-Lemli-Opitz syndrome. Am J Med Genet 98(2): 191-200.
Tochigi M, Kato C, Koishi S et al. (2007) No evidence for significant association between GABA receptor genes in chromosome 15q11-q13 and autism in a Japanese population. J Hum Genet 52(12): 985-989.
Trikalinos TA, Karvouni A, Zintzaras E et al. (2006) A heterogeneity-based genome search meta-analysis for autism-spectrum disorders. Mol Psychiatry 11(1): 29-36.
Tuchman R, Rapin I (2002) Epilepsy in autism. Lancet Neurol 1(6): 352-358.
Turunen JA, Rehnstrom K, Kilpinen H et al. (2008) Mitochondrial Aspartate/Glutamate Carrier SLC25A12 Gene Is Associated With Autism. Autism Res 1: 189-192.
Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 98(9): 5116-5121.
Wahl-Schott C, Biel M (2009) HCN channels: structure, cellular regulation and physiological function. Cell Mol Life Sci 66(3): 470-494.
Wallace RH, Marini C, Petrou S et al. (2001) Mutant GABA(A) receptor gamma2-subunit in childhood absence epilepsy and febrile seizures. Nat Genet 28(1): 49-52.
Vallender EJ, Mekel-Bobrov N, Lahn BT (2008) Genetic basis of human brain evolution. Trends Neurosci 31(12): 637-644.
Walsh T, McClellan JM, McCarthy SE et al. (2008) Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science 320(5875): 539-543.
Wang K, Li M, Hadley D et al. (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 17(11): 1665-1674.
Vargas DL, Nascimbene C, Krishnan C et al. (2005) Neuroglial activation and neuroinflammation in the brain of patients with autism. Ann Neurol 57(1): 67-81.
Varilo T (1999) The age of the mutations in the Finnish disease heritage; a genealogical and linkage equilibrium study. Helsinki: National Public Health Institute.
Wassink TH, Piven J, Vieland VJ et al. (2004) Examination of AVPR1a as an autism susceptibility gene. Mol Psychiatry 9(10): 968-972.
Wazana A, Bresnahan M, Kline J (2007) The autism epidemic: fact or artifact? J Am Acad Child Adolesc Psychiatry 46(6): 721-730.
Weedon MN, Lango H, Lindgren CM et al. (2008) Genome-wide association analysis identifies 20 loci that influence adult height. Nat Genet 40(5): 575-583.
Veenstra-Vanderweele J, Cook E, Jr., Lombroso PJ (2003) Genetics of childhood disorders: XLVI. Autism, part 5: genetics of autism. J Am Acad Child Adolesc Psychiatry 42(1): 116-118.
Weiss KM, Terwilliger JD (2000) How many diseases does it take to map a gene with SNPs? Nat Genet 26(2): 151-157.
Weiss LA, Arking DE, Gene Discovery Project of Johns Hopkins and the Autism Consortium (2009) A genome-wide linkage and association scan reveals novel loci for autism. Nature in press.
Weiss LA, Pan L, Abney M et al. (2006a) The sex-specific genetic architecture of quantitative traits in humans. Nat Genet 38(2): 218-222.
Weiss LA, Kosova G, Delahanty RJ et al. (2006b) Variation in ITGB3 is associated with whole-blood serotonin level and autism susceptibility. Eur J Hum Genet.
Weiss LA, Veenstra-Vanderweele J, Newman DL et al. (2004) Genome-wide association study identifies ITGB3 as a QTL for whole blood serotonin. Eur J Hum Genet 12(11): 949-954.
Weiss LA, Shen Y, Korn JM et al. (2008) Association between Microdeletion and Microduplication at 16p11.2 and Autism. N Engl J Med 358(7): 667-675.
Veltman MW, Thompson RJ, Roberts SE et al. (2004) Prader-Willi syndrome--a study comparing deletion and uniparental disomy cases with reference to autism spectrum disorders. Eur Child Adolesc Psychiatry 13(1): 42-50.
Venter JC, Adams MD, Myers EW et al. (2001) The sequence of the human genome. Science 291(5507): 1304-1351.
Verkerk AJ, Mathews CA, Joosse M et al. (2003) CNTNAP2 is disrupted in a family with Gilles de la Tourette syndrome and obsessive compulsive disorder. Genomics 82(1): 1-9.
Wheeler DA, Srinivasan M, Egholm M et al. (2008) The complete genome of an individual by massively parallel DNA sequencing. Nature 452(7189): 872-876.

Whitney ER, Kemper TL, Bauman ML et al. (2008) Cerebellar Purkinje cells are reduced in a subpopulation of autistic brains: a stereological experiment using calbindin-D28k. Cerebellum (London, England) 7(3): 406-416.
Williams RS, Hauser SL, Purpura DP et al. (1980) Autism and mental retardation: neuropathologic studies performed in four retarded persons with autistic behavior. Arch Neurol 37(12): 749-753.
Wilson HL, Crolla JA, Walker D et al. (2008) Interstitial 22q13 deletions: genes other than SHANK3 have major effects on cognitive and language development. Eur J Hum Genet 16(11): 1301-1310.
Wing L (1981) Asperger's syndrome: a clinical account. Psychol Med 11(1): 115-129.
Winston JT, Koepp DM, Zhu C et al. (1999) A family of mammalian F-box proteins. Curr Biol 9(20): 1180-1182.
Virkud YV, Todd RD, Abbacchi AM et al. (2009) Familial aggregation of quantitative autistic traits in multiplex versus simplex autism. Am J Med Genet B Neuropsychiatr Genet 150B(3): 328-334.
Voight BF, Kudaravalli S, Wen X et al. (2006) A map of recent positive selection in the human genome. PLoS Biol 4(3): e72.
Volkmar FR, Klin A (2000) Diagnostic Issues in Asperger Syndrome. In: Klin A, Volkmar FR, Sparrow SS, editors. Asperger Syndrome. New York: The Guilford Press. pp. 25-71.
Woods CG, Cox J, Springell K et al. (2006) Quantification of homozygosity in consanguineous individuals with autosomal recessive disease. Am J Hum Genet 78(5): 889-896.
World Health Organization (1993) The ICD-10 Classification of Mental and Behavioural Disorders. Diagnostic Criteria for Research. Geneva: WHO.
Vorstman JA, Staal WG, van Daalen E et al. (2006) Identification of novel autism candidate regions through analysis of reported cytogenetic abnormalities associated with autism. Mol Psychiatry 11(1): 1, 18-28.
WTCCC (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447(7145): 661-678.
Wu S, Jia M, Ruan Y et al. (2005) Positive association of the oxytocin receptor gene (OXTR) with autism in the Chinese Han population. Biol Psychiatry 58(1): 74-77.
Xu J, Zwaigenbaum L, Szatmari P et al. (2004) Molecular Cytogenetics of Autism. Curr Genomics 5(4): 347-364.
Yamaguchi-Kabata Y, Nakazono K, Takahashi A et al. (2008) Japanese population structure, based on SNP genotypes from 7003 individuals compared to other ethnic groups: effects on population-based association studies. Am J Hum Genet 83(4): 445-456.
Yan J, Oliveira G, Coutinho A et al. (2005) Analysis of the neuroligin 3 and 4 genes in autism and other neuropsychiatric patients. Mol Psychiatry 10(4): 329-332.
Yang MS, Gill M (2007) A review of gene linkage, association and expression studies in autism and an assessment of convergent evidence. Int J Dev Neurosci 25(2): 69-85.
Yao JK, Reddy RD, van Kammen DP (2001) Oxidative damage and schizophrenia: an overview of the evidence and its therapeutic implications. CNS drugs 15(4): 287-310.
Yeargin-Allsopp M, Rice C, Karapurkar T et al. (2003) Prevalence of autism in a US metropolitan area. Jama 289(1): 49-55.
Yirmiya N, Pilowsky T, Tidhar S et al. (2002) Family-based and population study of a functional promoter-region monoamine oxidase A polymorphism in autism: possible association with IQ. Am J Med Genet 114(3): 284-287.
Ylisaukko-oja T, Nieminen-von Wendt T, Kempas E et al. (2004) Genome-wide scan for loci of Asperger syndrome. Mol Psychiatry 9(2): 161-168.
Ylisaukko-oja T, Rehnstrom K, Auranen M et al. (2005) Analysis of four neuroligin genes as candidates for autism. Eur J Hum Genet 13(12): 1285-1292.
Yonan AL, Palmer AA, Gilliam TC (2006) Hardy-Weinberg disequilibrium identified genotyping error of the serotonin transporter (SLC6A4) promoter polymorphism. Psychiatr Genet 16(1): 31-34.
Yoo HJ, Lee SK, Park M et al. (2009) Family- and population-based association studies of monoamine oxidase A and autism spectrum disorders in Korean. Neurosci Res 63(3): 172-176.
Zahir F, Firth HV, Baross A et al. (2007) Novel deletions of 14q11.2 associated with developmental delay, cognitive impairment and similar minor anomalies in three children. J Med Genet 44(9): 556-561.
Zhao X, Leotta A, Kustanovich V et al. (2007) A unified genetic theory for sporadic and inherited autism. Proc Natl Acad Sci U S A.