European Journal of Human Genetics (2010) 18, 588–595
© 2010 Macmillan Publishers Limited All rights reserved 1018-4813/10 $32.00
www.nature.com/ejhg

ARTICLE

Systematic genotype–phenotype analysis of autism susceptibility loci implicates additional symptoms to co-occur with autism

Jacobine E Buizer-Voskamp[1,2,5], Lude Franke*[,3,5], Wouter G Staal[4], Emma van Daalen[4], Chantal Kemner[4], Roel A Ophoff[2], Jacob AS Vorstman[1], Herman van Engeland[4] and Cisca Wijmenga[3]

Many genetic studies in autism have been performed, resulting in the identification of multiple linkage regions and cytogenetic aberrations, but little unequivocal evidence for the involvement of specific genes exists. By identifying novel symptoms in these patients, enhanced phenotyping of autistic individuals not only improves understanding and diagnosis but also helps to define biologically more homogeneous groups of patients, improving the potential to detect causative genes. Supported by recent copy number variation findings in autism, we hypothesized that for some susceptibility loci, autism resembles a contiguous gene syndrome, caused by aberrations within multiple (contiguous) genes, which jointly increases autism susceptibility. This would result in various different clinical manifestations that might be rather atypical, but that also co-occur with autism. To test this hypothesis, 13 susceptibility loci, identified through genetic linkage and cytogenetic analyses, were systematically analyzed. The Online Mendelian Inheritance in Man database was used to identify syndromes caused by mutations in the genes residing in each of these loci. Subsequent analysis of the symptoms expressed within these disorders allowed us to identify 33 symptoms (significantly more than expected, P=0.037) that were over-represented in previous reports mapping to these loci. Some of these symptoms, including seizures and craniofacial abnormalities, support our hypothesis as they are already known to co-occur with autism. These symptoms, together with ones that have not previously been described to co-occur with autism, might be considered for use as inclusion or exclusion criteria toward defining etiologically more homogeneous groups for molecular genetic studies of autism

European Journal of Human Genetics (2010) **18,** 588–595; doi:10.1038/ejhg.2009.206; published online 25 November 2009

Keywords: autistic disorder; genotype; phenotype; databases; genetic; medical subject headings

INTRODUCTION

Autism spectrum disorder is characterized by deviations and delays in the development of reciprocal social interaction and communication, in combination with restricted and repetitive behaviors and interests.[1] The prevalence of the broad autism spectrum has recently been estimated to be approximately 1% of the childhood population, whereas the prevalence rate for autism is estimated at approximately 4 per 1000 births.[2,3] Phenotypically, autism is very heterogeneous, with varying degrees of severity and associated intellectual functioning.[4] The large variety of neuropathological changes and the variability seen across subjects imply that autism is also etiologically heterogeneous.[5]

Cumulative evidence from family and twin studies suggests that genetic factors have an important role in the pathology of autism.[5,6] The genetic contribution to autism has been estimated to be as high as 90 percent.[4,7–9] Findings of cytogenetic abnormalities and single-gene disorders associated with autism indicate that the disorder is genetically complex, involving multiple (interacting) loci.[4,6,7] Although few susceptibility loci have been consistently replicated, the overlap in linkage findings from genome scans suggests various regions that harbor autism susceptibility genes. Loci found in at least two independent linkage studies are in the regions 2q, 3q25–27, 3p25, 6q14–21, 7q31–36, and 17q11–21.[4] However, each of these loci contains hundreds of genes, of which multiple genes have been implicated in autism. Among these, other genes have been identified from independent association studies,[7–10] but no gene has been unequivocally shown to contribute to autism susceptibility.

The results of all molecular genetic studies point to a model of multiple genetic variants that supposedly can interact in various ways with regard to the phenotypic expression of autism.[4] Recently, evidence appeared that small cytogenetic aberrations, including copy number variations (CNVs), might have important roles in autism.[11,12] Methods to detect these genome-wide provide a powerful alternative to traditional gene-mapping approaches for discovering susceptibility genes in autism.[11,13] Recent CNV studies suggest that lesions at many different loci can contribute to autism, a result consistent both with the findings from cytogenetic studies and with the failure to find causal variants.[11] These CNVs can be recurrent, inherited, and/or arise *de novo*. A recent study showed that *de novo* variants were present in approximately 7% of idiopathic families having at least one child with

[1]Rudolf Magnus Institute of Neurosciences, Department of Psychiatry, University Medical Centre, Heidelberglaan, Utrecht, The Netherlands; [2]Complex Genetics Section, DBG-Department of Medical Genetics, University Medical Centre, Universiteitsweg, Utrecht, The Netherlands; [3]Department of Genetics, University Medical Center Groningen, Groningen, The Netherlands; [4]Department of Child and Adolescent Psychiatry, Rudolf Magnus Institute of Neurosciences, University Medical Centre, Heidelberglaan, Utrecht, The Netherlands
*Correspondence: Dr L Franke, Department of Genetics, University Medical Center Groningen, P.O. Box 30001, 9700 RB, Groningen, The Netherlands. Tel: +31 6 41 54 9962; Fax: +31 50 361 7230; E-mail: lude@ludesign.nl

[5]These authors contributed equally to this work

Received 10 January 2008; revised 18 September 2009; accepted 29 September 2009; published online 25 November 2009

autism spectrum disorder. The mean size of those *de novo* variants was approximately 5 Mb (median ~3 Mb). The mean number of genes encompassed by these rare structural variants was over 30.[14] It has been suggested that these structural variants may account for a larger fraction of the overall genetic risk than was previously assumed.[15] Recently, rare microdeletions and microduplications were found in autistic individuals on 16p11.2, containing 25 annotated genes or transcripts, of which several could be considered good candidates for driving the phenotype on the basis of their expression in the brain or function in neurodevelopment.[16]

When assuming that these loci confer susceptibility to autism, the considerable amount of genes within each CNV suggests that it is possible that some of these rare CNVs represent a contiguous gene syndrome. Williams–Beuren syndrome, a neurodevelopmental disorder attributed to a deletion of 7q11.23, is a prime example of such a contiguous gene syndrome: The deleted region includes more than 20 genes, and it is believed that the characteristic features of this disorder are because of the loss of multiple genes, of which several are presumed to be responsible for a subset of the Williams–Beuren syndrome symptoms.[17]

Although it can be that within each autism susceptibility locus only one single gene raises the susceptibility to autism, we hypothesize here that for some of these loci autism resembles a contiguous gene syndrome, caused by (*de novo*) aberrations within multiple (contiguous) genes. These rare *de novo* structural mutations can result in the disruption of multiple biological functions as multiple genes reside within these loci, resulting in phenotypes that can easily be related to the autistic phenotype, but that can also be rather atypical.[18]

To substantiate the hypothesis of autism as a (partly) contiguous gene syndrome, we sought evidence that certain atypical symptoms co-occur with autism: For a set of loci that have already been implicated in autism[12] we systematically investigated all positional candidate genes and determined what symptoms are usually caused by aberrations within each of these genes. This analysis was performed by text mining the Online Mendelian Inheritance in Man (OMIM) database. In OMIM, considerable information is present describing rare monogenic mutations that cause rare—most often serious—diseases. We hypothesized that these rare diseases with serious phenotypes can be informative for complex diseases with a more subtle phenotype, such as autism, that have not been described as extensively within OMIM. Part of the symptoms caused by rare monogenic variants may be found with more subtle presentation in complex disorders that have other less deleterious variants within the same genes. This hypothesis is supported by recently identified genetic variants that influence adult height variation.[19–21] For several of the genes to which these variants map, monogenic mutations are known (and reported within OMIM) that lead to rare syndromes and symptoms that affect skeletal development, such as skeletal dysplasia. Comparable findings now exist, for example, diabetes and lipid levels. As such, we argue that a study of rare syndromes and symptoms of genes in loci, implicated in autism, might be useful in identifying clinical features that are common to autism spectrum disorders.

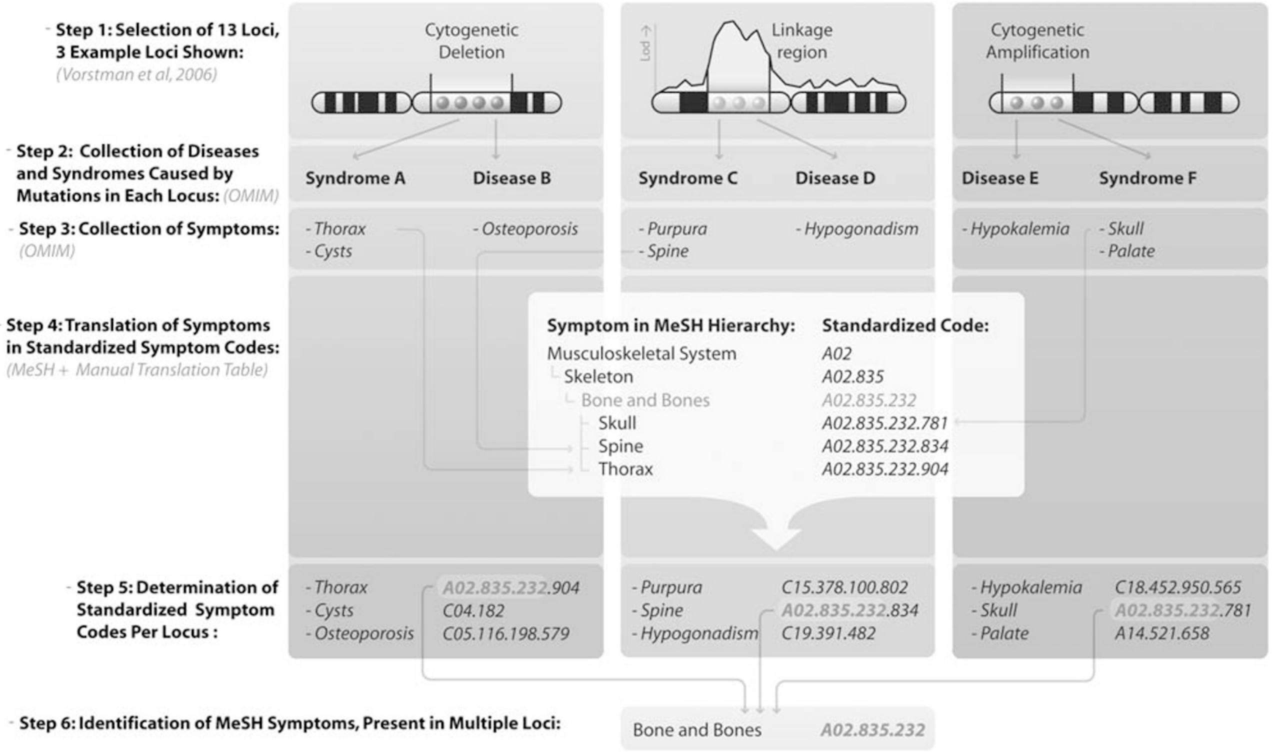

Figure 1 Overview of identification of overrepresented symptoms in autism loci. For many loci, associated with autism, no genes have been unequivocally shown to be associated. We have assumed that a systematic analysis of symptoms caused by aberrations in positional candidate genes in these loci might reveal symptoms that are present more often than one would expect by chance, and thus might co-occur with autism. First, loci identified through cytogenetic and linkage analysis are used as input (Step 1). In this example three loci have been identified, each of them causing known diseases or syndromes when mutated (Step 2). For each disorder, we subsequently determine the associated symptoms (Step 3). MeSH is then used for the generation of standardized codes, which are hierarchically organized, allowing to be both specific and generic at the same time (eg, 'spine' is specific, 'bone and bones' is generic) (Step 4). Once all the symptoms have been recoded, it can be determined what symptoms are expressed per locus (Step 5), allowing for the identification of overrepresented ones (eg, 'bone and bones') (Step 6).

Table 1 Autism susceptibility loci included in the analysis

Locus		*Chromosomal location (basepair positions)*		*No of genes*	*Evidence for inclusion*
1q42.2	(D1S1656)	217 212 087	237 212 087	138	Buxbaum *et al.*[26]
2q31.1	(D2S2188)	165 430 238	185 430 238	109	IMGSAC[27]
2q37		233 875 000	243 020 000	77	Vorstman *et al.*[12]
3q26.32	(D3S3037)	168 924 373	188 924 373	120	Auranen *et al.*[25]
5p15		0	16 900 000	65	Vorstman *et al.*[12]
7q22.1	(D7S477)	90 370 231	110 370 231	193	IMGSAC[27]
7q36.2	(D7S2462)	142 999 403	162 999 403	105	Auranen *et al.*[25]
15q11-14		18 940 000	31 390 000	65	Vorstman *et al.*[12]
17q11.2	(5-HTTLPR)	15 406 471	35 406 471	272	McCauley *et al.*[28]
18q21-23		41 800 000	73 160 000	117	Vorstman *et al.*[12]
22q11.2		16 970 000	20 830 000	74	Vorstman *et al.*[12]
22q13.3		44 555 000	49 550 000	51	Vorstman *et al.*[12]
Xp22		0	24 700 000	122	Vorstman *et al.*[12]
			Total number of genes	1508	

The chromosomal location is provided for each locus. If a locus has been identified through linkage analysis, the microsatellite markers are given in brackets. The chromosomal location provides base-pair boundaries for each locus. The total number of genes for the loci are given, as well as the total number of genes in all the loci and the linkage results/cytogenetic regions of interest.

Once we had determined what syndromes and symptoms could be caused by mutations in genes residing in the autism loci, we assessed for each identified symptom whether more than one of the autism susceptibility loci could cause this symptom and determined whether the amount of loci that could cause this symptom (through affected genes within these loci) was significantly higher than expected by chance (Figure 1).

The identification of certain symptoms, reported more often in these loci than expected, would substantiate this hypothesis, and additionally might help to identify symptoms that have not yet been described to co-occur with autism and which could be relevant for the clinic.

Knowledge on these symptoms is relevant for genetic research as well: When assuming that a symptom is expressed in only a subset of patients, it might be worthwhile to condition on this symptom. Although such symptoms can arise because of aberrations within multiple loci, it can be that within these patients additional (shared) susceptibility loci exist that help cause this particular symptom (as for complex diseases, different genetic aberrations can result in identical symptoms and phenotypes through disruptions within the same biological cascades,[17] the concept of convergence). By grouping autism individuals sharing such symptoms, it might subsequently be possible to increase statistical power to identify those additional susceptibility loci that are shared among these individuals. In a recent study, three schizophrenia patients were identified with deletions in *CNTNAP2*, a known epilepsy susceptibility locus. It turned out that these individuals also had epilepsy.[22]

Taken together, defining subgroups on the basis of clinical presentation (representing disruptions within the same biological pathway) could be useful for follow-up research.[23,24]

MATERIALS AND METHODS

Definition of susceptibility loci

Loci for autism were selected on the basis of evidence from both linkage and cytogenetic studies. Of all the linkage studies that we had previously analyzed,[12] four studies that had at least one locus with a multipoint logarithm of the odds score above 3.0 were included in the analysis.[25–28] Given that no unequivocal method of defining the extent of the region is provided in the literature, and the fact that information about a logarithm of the odds-1 drop region was not always present, boundaries of the linkage regions were pragmatically defined at both ends of a 20 MB base-pair block centered around the most significantly linked marker in each locus.

Definition of the Cytogenetic Regions of Interest (CROIs) was based on criteria that have been previously described.[12] In short, regions on the human genome where multiple overlapping cytogenetic abnormalities co-occurred with an autism phenotype were identified through extensive literature search. Only CROIs that contained more than five overlapping cases were included for analysis. Cases involving chromosomal mosaicism or well-described gene mutation as the most likely genetic cause for autism were excluded (for example, patients with fragile X syndrome caused by *Fmr1* mutations). In total, we defined 13 loci, of which six were based on linkage data and seven were based on cytogenetic data (Table 1). The NCBI V35 assembly was used to physically map all markers, probes, and banding information.

Identification of syndromes and subsequent symptoms caused by aberrations

The OMIM database catalogs the majority of all known diseases that have genetic components providing extensive information on both clinical aspects and the genetic basis of these syndromes. We determined which syndromes were caused by aberrations that were (partly) overlapping with each of the 13 loci (Figure 1). OMIM provides a clinical synopsis describing the core symptoms caused by each disorder. As this information is both well organized and extensive, we chose this repository as the basis for collecting symptom information for each syndrome. We included only the core clinical manifestation information, and not the entries contained in the 'miscellaneous,' 'molecular basis,' and 'inheritance' sections, because these never describe actual symptoms. For each entry, only the complete text was used to prevent subsets of phrases being incorrectly attributed (eg, 'spot quality assessment' was taken as a whole, because 'spot' can be interpreted to be a symptom in 'Exanthema').

Subsequently, the Medical Subject Headings (MeSH) vocabulary[29] was used to code these symptoms displayed within disorders in a standardized way. This transformation could be applied as the MeSH ontology is hierarchically organized, allowing one to describe specific symptoms (eg, 'spine,' MeSH code 'A02.835.232.834'), but be generic at the same time ('spine' is part of the parent MeSH term 'bone and bones,' MeSH code 'A02.835.232'). As such, slightly different but related symptoms (eg, 'skull,' 'spine,' and 'thorax') all share a more generic parent MeSH term ('bone and bones'), which enabled us to associate these symptoms with each other through a common parent term.

To ensure that the automatic assignment of clinical synopsis information to MeSH terms was performed with high accuracy, we also manually assigned all the symptoms for the syndromes contained in the 13 loci to MeSH terms. This manual curation resulted in a conversion table, which maps clinical synopsis entries to known MeSH terms (Supplementary Table S1). This allowed for the automatic extraction of information, by text mining[30–33] OMIM, and MeSH, and through the conversion table it increased the yield of clinical synopsis assignments to MeSH.

Table 2 Overview of difficulties for text mining in OMIM and using MeSH

Difficulty	*Description*
Difficulties for text mining in OMIM	
Clinical synopsis not designed to be easily machine interpretable	Sometimes, the clinical synopsis contains symptoms such as 'Heart: prolonged QTc interval; T-wave abnormalities'. Having computers interpret this as something which has to do with MeSH term 'ECG abnormalities', is difficult.
Non-standardized method for describing phenotypes	In some cases, limited clinical synopsis field is present, whereas various symptoms are described in the 'clinical features' part of the full-text OMIM record. In addition, the clinical synopsis field is not consistent in describing phenotypes. Sometimes different phrasing exists for nearly identical symptoms, such as 'Height: short stature' and 'height: adult height reduced; final adult height less than 152 cm'.
Minor spelling errors within OMIM	In the clinical synopsis sometimes spelling errors are present, such as 'hypereflexia' instead of 'hyperreflexia', 'congential' instead of 'congenital', and 'defeciency' instead of 'deficiency'.
Difficulties with utilizing MeSH	
Symptoms not present in MeSH	Various symptoms are not present in MeSH, such as 'short stature', 'broad nasal bridge' or 'striae'.
Idiosyncrasies in MeSH	'Microcephaly' (C05.660.207.620) is present in MeSH as a member of 'Craniofacial abnormalities' (C05.660.207). However, 'Macrocephaly' is not present in MeSH. The only solution for including this symptom is to assign it to the generic term 'Craniofacial abnormalities'.
Differences in extensiveness of MeSH	MeSH is not equally extensive for all medical subjects: The 'respiratory tract diseases' (C08) tree contains many highly specific terms, whereas the 'mental disorders' (F03) tree only contains terms on the level of individual diseases but not that much on specific psychiatric symptoms. Therefore symptoms such as 'impaired social smile' can only be assigned to the generic term 'child behavior disorders'.

Although text mining and natural language processing have gained attention recently, there are still numerous practical problems to deal with in OMIM and MeSH. Commonly observed difficulties, along with examples, are shown.

Analysis of over-represented symptoms in loci

We then traversed all MeSH terms, including those that had been explicitly mentioned, along with their more generic parent and grandparent MeSH terms, and determined in how many loci each MeSH term was reported at least once. Once this was assessed, we determined whether any of these MeSH terms had been described within more loci than expected by performing a 10 000 fold permutation analysis on the data. In each permutation, the 13 loci were shuffled randomly across the genome and the text mining analysis was performed again on these permuted loci. For each MeSH term, the number of shuffled loci in which this term had been described was determined and this number was compared with the original number of loci in which this term had been described. Consequently, after these 10 000 permutations, for each MeSH term an empiric *P*-value could be determined.

To identify potentially common symptoms, we only followed up MeSH terms that were present in at least four loci. As our strategy was to determine potentially relevant novel symptoms in autism, we deemed a symptom interesting when its empirically determined *P*-value was below 0.05. We assessed whether the number of identified symptoms with an empiric *P*-value below 0.05 was significantly more than expected by a 1000-fold permutation analysis. We shuffled the loci randomly across the genome and determined for each permutation how many terms had a *P*-value below 0.05, using the same filtering as we had applied for the original CROIs. This enabled us to empirically determine whether the amount of nominally significantly identified symptoms was more than expected.

RESULTS

An overview of the 13 selected loci is shown in Table 1, along with the evidence for their inclusion (linkage results or cytogenetic region of interest). To ensure that the loci that were identified through cytogenetic analyses were potentially specific to autism and were not commonly deleted or duplicated, we investigated each locus in the Database of Common Genetic Variations.[34] None of these loci were known to contain aberrations in healthy individuals as extensive as the ones observed within autism patients.

Once these loci had been defined, OMIM was assessed to determine, which known syndromes are caused by mutations in each of these loci. Subsequent analysis of the clinical synopsis information for each syndrome and mapping to MeSH terms allowed us to extract a standardized set of symptoms. Assignment of symptoms through the use of the manually curated conversion table (Supplementary Table S1) resulted in the assignment of over 500 extra symptoms to MeSH terms. Although this increase of assignment was considerable and as accurate as possible, mining OMIM and mapping of symptoms to MeSH terms were sometimes problematic, as outlined in Table 2.

Once all syndromes had been processed, we assessed per MeSH term the number of loci in which this term was mentioned. To establish whether any term was over-represented, that is, present in more loci than expected by chance, a permutation analysis was performed, which allowed for the determination of an empiric *P*-value for each term (Figure 1; Supplementary Table S2). As we had manually translated the clinical synopsis information for the syndromes that mapped within our 13 loci to MeSH terms, we wanted to ensure that clinical information for syndromes residing outside of these loci could also be mapped using this translation table. If a slightly different phrasing of symptoms had been used in syndromes that we had not manually assessed, as they mapped outside of our 13 loci, this could influence the accuracy of the empirically determined *P*-value. This was, however, not the case, as the results from an analysis that relied entirely on the automatic translation of clinical synopsis symptoms to MeSH terms (Supplementary Table S2) gave comparable results to an analysis that included the manual assignment (Supplementary Table S3).

As autism, Asperger's disorder, and RETT syndrome had already been described (OMIM numbers 209850, 608636, 607373, 300495, 312750, 608638, and 300497) in four out of the 13 loci, this allowed for an initial validation of our method. Symptoms mentioned for these syndromes could be attributed to the MeSH term 'Child behavior disorders,' for which the empirically determined *P*-value was 0.01 (Supplementary Table S3). To prevent a bias toward autism symptoms already described in OMIM, we excluded these syndromes, along with autism-related syndromes that were defined within OMIM (OMIM numbers 606053, 609378, 611015, 611016, 605309, 608049, 610676, 610836, 300425, 610838, 300496, 610908, 300672, 300624, 608631, 300494, 609954, 608781)—but which mapped outside of our 13 loci—from further analyses (Table 3).

Table 3 Significantly over-represented symptoms mentioned in at least four loci

	MeSH number	*MeSH description*	*Number mentioned in different syndromes in 13 CROIs*	*Empiric* P*-value*
1.	C23.888.885	Skin manifestations	11	0.00369963
2.	C13.371.852	Uterine diseases	5	0.00519948
3.	C07.793.494	Malocclusion	6	0.00629937
4.	A17.360	Hair	12	0.01359864
5.	A14.521.658	Palate	8	0.01409859
6.	C23.550.291.812	Facies	8	0.01479852
7.	C05.660.207.620	Microcephaly	11	0.01579842
8.	C23.300.175	Calculi	4	0.01659834
9.	C07.650	Stomatognathic system abnormalities	13	0.01979802
10.	C10.597.617	Pain	6	0.02319768
11.	C19.391.482	Hypogonadism	6	0.02349765
12.	C17.800	Skin diseases	13	0.02349765
13.	C05.116.099.343	Dwarfism	12	0.02379762
14.	C13.371	Genital diseases, female	10	0.02419758
15.	C17	Skin and connective tissue diseases	13	0.02479752
16.	C05.116.099	Bone diseases, developmental	13	0.02729727
17.	C17.800.946	Sweat gland diseases	4	0.02929707
18.	C05.660.207	**Craniofacial abnormalities**	13	0.03089691
19.	C05.116.198.579	Osteoporosis	8	0.03109689
20.	A04.623.557	Nasopharynx	6	0.03169683
21.	C18.452.950.565	Hypokalemia	4	0.03189681
22.	C05.500.460	Jaw abnormalities	11	0.03279672
23.	C06.130.564	Gallbladder diseases	4	0.03309669
24.	C10.292	Cranial nerve diseases	13	0.03559644
25.	C10.228.140.490.631	**Seizures**	13	0.04059594
26.	A05.810	Urinary tract	9	0.04109589
27.	A05	Urogenital system	9	0.04139586
28.	C10.228.140.490	**Epilepsy**	13	0.04269573
29.	C04.588	Neoplasms by site	8	0.04379562
30.	C05.116.099.370.894	Synostosis	7	0.04479552
31.	C07	Stomatognathic diseases	13	0.04809519
32.	C05.116.099.370.894.819	Syndactyly	6	0.04809519
33.	C06.405.469	Intestinal diseases	9	0.04859514

Over-represented symptoms mentioned in at least four loci with an empiric *P*-value < 0.05 are shown. Symptoms indicated in bold are significantly over-represented and already known to be involved in autism.

Although 33 over-represented symptoms were observed, it should be noted that many different symptoms had been assessed within this analysis, requiring us to control for multiple testing issues. To do this, we performed a permutation analysis to determine whether the number of 33 over-represented symptoms was significantly higher than expected. This was indeed the case (empiric *P*-value=0.037), indicating that some of the reported symptoms are likely to reflect true-positive findings. To ensure the robustness of this analysis, we assessed whether the number of genes within the CROIs differed from the average number of genes in the permuted loci, but did not observe a difference (1508 genes map within the CROIs, opposed to on average 1569 genes within the permuted loci, empiric *P*-value=0.47). Subsequent inspection of the most significantly over-represented symptoms (Table 3) suggests that some of these are related (Table 4). Notable are epilepsy/seizures and craniofacial abnormalities, as these have been previously implicated in autism.[35,36] Furthermore, the results indicate that most of these symptoms affect tissues that are of ectodermal origin (Figure 2a). They develop in the first and second trimester of pregnancy and affect many organs (Figure 2b).

DISCUSSION

Through text mining of syndromes caused by aberrations in 13 linkage regions and CROIs, this study suggests that various symptoms co-occur with autism that have not yet been widely studied or previously described: We found 33 symptoms that were present in these regions more often than expected by chance (nominal empiric *P*-value < 0.05). Through subsequent permutation analyses, we observed that this number of 33 over-represented symptoms is higher than expected. These observations support our hypothesis that autism might partly be a contiguous gene syndrome, in which the function of multiple positional candidate genes within the susceptibility loci is affected. This would result in various different clinical manifestations that might be quite atypical, but jointly might also be able to cause autism-like features, which is supported by reports on Xp22.3 deletions, in which patients show the variable association of apparently unrelated clinical manifestations.[37,38] Jointly, the multiple genes with their resulting clinical phenotypes could increase the probability of developing autism. Additional support for autism as a partly contiguous gene syndrome, and the probable existence of different subtypes, comes from CNV studies identifying rare variants covering multiple genes of different function in the etiology of autism.[18] For some of the co-occurring symptoms, evidence already exists that they indeed have a role in autism. The most prominent are epilepsy/seizures and craniofacial abnormalities, which have been previously mentioned as possible genetically informative phenotypes in autism.[35,36]

Table 4 Clustering of significantly over-represented symptoms in at least four loci

				Diseases affecting the organs		*Hormone-related diseases*	
Face & Skull	*Skin diseases*	*Bone diseases*	*Brain and nerve dysfunction*	*Digestive system diseases*	*Urogenital diseases*	*Metabolic disorders*	*Endocrine diseases*
Malocclusion	Skin manifestations	Pain[a]	Cranial nerve diseases[a]	Gallbladder diseases	Uterine diseases	Hypokalemia	Hypogonadism
Hair	Pain[a]	Dwarfism	**Seizures**	Neoplasms by site[b]	Calculi		
Palate	Skin diseases	Bone diseases, developmental	**Epilepsy**	Intestinal diseases	Genital diseases, female		
Facies	Skin/connective tissue diseases	Osteoporosis			Urinary tract		
Microcephaly	Sweat gland diseases	Synostosis			Urogenital system		
Stomatognathic system abnormalities		Syndactyly					
Craniofacial abnormalities							
Nasopharynx Jaw Abnormalities Stomatognathic diseases							
References[42–46,48]	*References* [53–55]	*References*[56,57]	*References Cranial nerve:*[45,52] *Epilepsy:*[39-41]	*References* [49–51]			

Relationships between different over-represented symptoms are shown. Symptoms indicated in bold are already known to be involved in autism.
Some closely related symptoms were combined (indicated by a and b).
Original OMIM symptoms: bone pain; back pain; burning of skin.
Original OMIM symptoms: neoplasms in colon, liver, biliary tract, and gastrointestinal tract.

Epilepsy is one of the best known and validated associations with autism.[39–41] It is much more common in people with autism than in the general population and, vice versa, it appears that autism and autistic-like conditions are more common in people with epilepsy. Recent studies suggest that more than one-third of the children with autism develop epilepsy.[39,41] About 15–20% of all people with autism had seizures before the age of 3 years.[40] The prevalence rates of epilepsy and the types of seizures seem to depend on the level of mental retardation, age, and incidence of regression.[39,41] Not surprisingly, this comorbidity led to researchers proposing that these diseases share common pathophysiological mechanisms.[41] The observed over-representation of seizures within this study supports these hypotheses because our method assumes that the same genetic background can yield both autism and other symptoms.

Minor physical anomalies, such as craniofacial abnormalities, in association with autism, have also been mentioned frequently.[42–46] Numerous case reports of thalidomide-induced autism suggest abnormal development very early in the gestation, resulting in craniofacial abnormalities.[4,42,45,46] Although most of these physical anomalies are also sometimes observed in other developmental disorders and in normally developing children as well,[47] craniofacial abnormalities might be potentially interesting because of their higher frequencies in autistic patients.[48]

Many parents report gastrointestinal symptoms in their autistic child,[49] in line with the digestive system disease symptoms we report. Although gastrointestinal problems are also fairly common in normally developing children, it has been estimated that they affect 46–84% of autism patients.[50,51] Chronic diarrhea, increased bile fluid output, constipation, and increased intestinal permeability are the most frequently mentioned abnormalities in autistic children.[49–51]

Limited evidence is available for the involvement of cranial nerves in autism. Although not convincing, a few studies on thalidomide-induced autism have suggested that the cranial nerves could be dysfunctional.[45,52] In this form of autism, individuals showed abnormalities in eye movement and facial expression. Other support comes from the observation that the exposure period for thalidomide autistic individuals is during days 20 and 24 of gestation. Few neurons form in this period, but the motor neurons of the cranial nerves are a notable exception. Interestingly, these nerves operate the muscles of the ears, jaw, throat, tongue, face, and eyes.[45]

For other symptoms, such as skin, bone, and urogenital problems, hypokalemia, and hypogonadism, there is little evidence of association with autism. Although skin symptoms, such as eczema,[53–55] and occasionally bone problems[56,57] have been reported, evidence for the presence of other symptoms is not available.

We should emphasize that although we observed a higher number of over-represented symptoms than expected (P=0.037, between 7 and 36 symptoms had been found in the 1000 permutations), some of these symptoms are likely to be false positives as none of these symptoms individually attained significance after stringent Bonferroni correction. But, taken together, some of these symptoms might have a role in autism. They could be considered as inclusion or exclusion criteria for future research, defining etiologically more homogeneous subgroups of autistic patients, by disrupting the same biological pathways, either caused by the same genetical aberration or not.

Limitations of our study

Although this method has identified various symptoms that are likely to co-occur with autism, we are aware of a number of limitations in our methodology. One important issue is that this study does not unequivocally prove that these symptoms, for which there is no

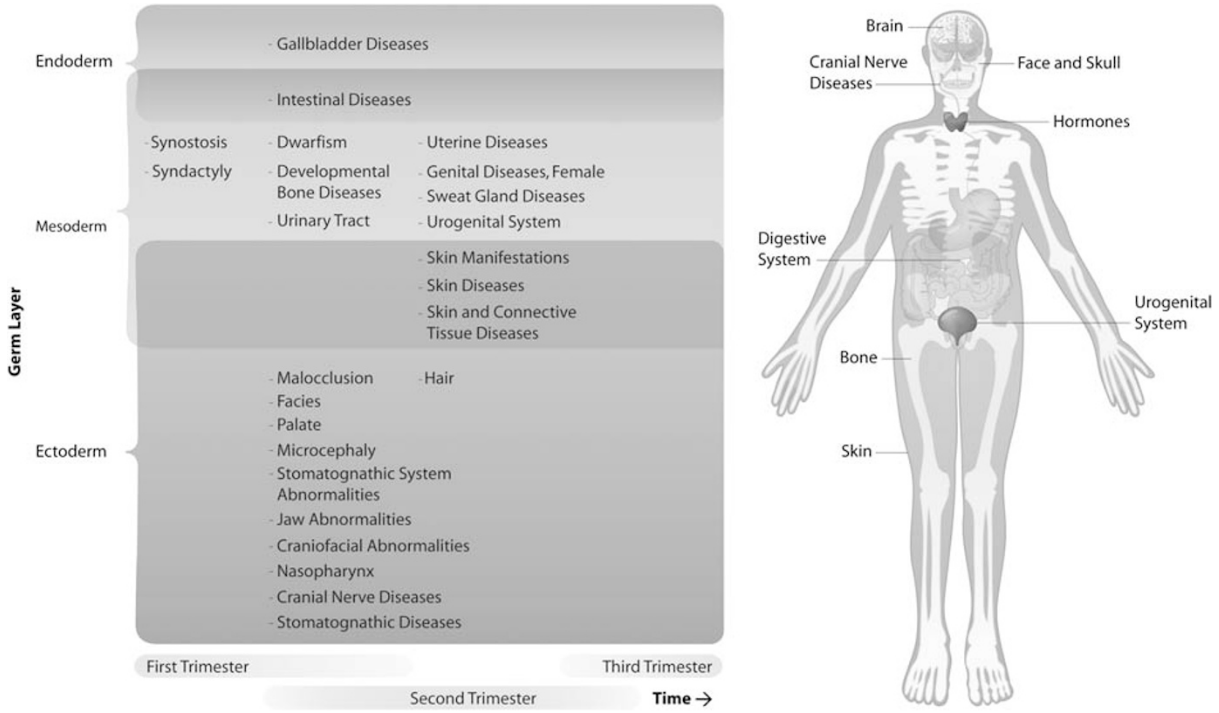

Figure 2 Overview of overrepresented symptoms in autism loci. Overrepresented symptoms (empiric P-value < 0.05), present in at least four loci, are shown along with the responsible organs. When possible, symptoms were assigned to a trimester of pregnancy and germ layer. The majority of symptoms are of ectodermal origin, while the majority of affected organs develop in the first to second trimesters.

evidence in the literature, are truly associated with autism. It could also be that they have never been studied, as in the clinical setting most attention is usually devoted to a triad of features: social impairments, communication impairments, and restricted repetitive behaviors and interests.

Another issue is how to determine what are the appropriate criteria for including a susceptibility locus. We tried to do this as carefully as possible, but it is possible that some are false positives. Other loci may well have been overlooked. Apart from these statistical power issues, there are no clear definitions on how to determine the exact boundaries of linkage regions and there is no consensus on whether to include only linkage regions that have shown significant linkage, or to also include loci that were suggestive of linkage. The cytogenetic regions of interest show comparable problems: how many overlapping cases are required to consider regions interesting is somewhat arbitrary, and again, how to define the boundaries of these loci accurately is open to discussion.

Not without their own problems are the use of OMIM and MeSH: as OMIM was designed to be interpreted by humans, there was no immediate need to use a standardized system for coding phenotypes. Consequently, when performing automated text mining in OMIM, various problems became apparent (as described in Table 2). Although manual curation partly overcame these problems, as it enabled the assignment of a substantial number of extra symptoms to MeSH terms, a more standardized method for describing symptoms in OMIM and in MeSH is desired, as previously suggested.[30,33,58]

Recently, some studies have been published that also use text mining to associate different types of information, a few of which also take OMIM and MeSH into account: Van Driel *et al.*[33] associated different phenotypes with each other using OMIM and MeSH; Butte *et al.*[30] associated phenotypes with expression data, and Lage *et al.*[31] have associated syndromes with protein complexes through text mining of OMIM and protein–protein interaction studies. However, as far as we are aware, no study has used OMIM and MeSH to assess whether there are any symptoms over-represented in multiple loci that have been implicated in complex diseases, such as autism, to provide leads for the involvement of unreported symptoms in these disorders.

Although much work remains to be performed to validate the actual co-occurrence of these symptoms in autistic patients, this study might be useful in pointing to ways for better characterizing patients, thereby providing new avenues for biologically informative phenotypes, which could lead to the identification of etiologically more homogeneous groups in patients and increase the statistical power to detect genetic associations. In addition, this method can easily be applied to other psychiatric disorders, as the input for our method consists solely of a set of susceptibility loci and an optional OMIM 'Clinical Synopsis to MeSH term' conversion table. This method will allow researchers to gain insight into the potential involvement of unreported symptoms associated with other psychiatric disorders as well.

ACKNOWLEDGEMENTS

We thank Jackie Senior, Sasha Zhernakova, Madelien van de Beek, members of the Complex Genetics Section, the Department of Human Genetics, and the Department of Child & Adolescent Psychiatry for a critical reading of the paper. This study was supported by a Fellowship from the Netherlands Genomics Initiative and a grant from the Celiac Disease Consortium, an innovative cluster approved by the Netherlands Genomics Initiative and partially funded by a Dutch government Grant (BSIK03009).

1 American Psychiatric Association: *Diagnostic and Statistical Manual of Mental disorders DSM-IV-TR*. 4th edn, text revision: 2000. Washington DC: USA.

2 Baird G, Simonoff E, Pickles A *et al*: Prevalence of disorders of the autism spectrum in a population cohort of children in South Thames: the Special Needs and Autism Project (SNAP). *Lancet* 2006; **368**: 210–215.
3 CDC: Prevalence of autism spectrum disorders—autism and developmental disabilities monitoring network sites United States 2002. *MMWR* 2007; **56**: 12–28, (SS–1).
4 Freitag C: The genetics of autistic disorders and its clinical relevance: a review of the literature. *Mol Psychiatry* 2007; **12**: 2–22.
5 Herbert M, Russo J, Yang S *et al*: Autism and environmental genomics. *Neurotoxicology* 2006; **27**: 671–684.
6 Coon H: Current perspectives on the genetic analysis of autism. *Am J Med Genet Part C* 2006; **142C**: 24–32.
7 Bacchelli E, Maestrini E: Autism spectrum disorders: molecular genetic advances. *Am J Med Genet Part C* 2006; **142C**: 13–23.
8 Folstein SE, Rosen-Sheidley B: Genetics of autism: complex aetiology for a heterogeneous disorder. *Nat Rev Genet* 2001; **2**: 943–955.
9 Wassink TH, Brzustowicz LM, Bartlett CW, Szatmari P: The search for autism disease genes. *Ment Retard Dev Disabil Res Rev* 2004; **10**: 272–283.
10 Muhle R, Trentacoste SV, Rapin I: The genetics of autism. *Pediatrics* 2004; **113**: 472–486.
11 Sebat J, Lakshmi B, Malhotra D *et al*: Strong association of *de novo* copy number mutations with autism. *Science* 2007; **316**: 445–449.
12 Vorstman JAS, Staal WG, van Daalen E, van Engeland H, Hochstenbach PFR, Franke L: Identification of novel autism candidate regions through analysis of reported cytogenetic abnormalities associated with autism. *Mol Psychiatry* 2006; **11**: 18–28.
13 Zhao X, Leotta A, Kustanovich V *et al*: A unified genetic theory for sporadic and inherited autism. *Proc Natl Acad Sci USA* 2007; **104**: 12831–12836.
14 Marshall CR, Noor A, Vincent JB *et al*: Structural variation of chromosomes in autism spectrum disorder. *Am J Hum Genet* 2008; **82**: 477–488.
15 Stefansson H, Rujescu D, Cichon S *et al*: Large recurrent microdeletions associated with schizophrenia. *Nature* 2008; **455**: 232–236.
16 Weiss L, Shen Y, Korn J *et al*: Association between microdeletion and microduplication at 16p11.2 and autism. *N Engl J Med* 2008; **358**: 667–675.
17 Yeung-Courchesne R, Courchesne E: From impasse to insight in autism research: From behavioral symptoms to biological explanations. *Dev Psychopathol* 1997; **9**: 389–419.
18 Abrahams B, Geschwind D: Advances in autism genetics: on the threshold of a new neurobiology. *Nat Rev Genet* 2008; **9**: 341–355.
19 Gudbjartsoon D, Walters G, Thorleifsson G *et al*: Many sequence variants affecting diversity of adult human height. *Nat Genet* 2008; **40**: 609–615.
20 Visscher P: Sizing up human height variation. *Nat Genet* 2008; **40**: 489–490.
21 Weedon M, Lango H, Lindgren C *et al*: Genome-wide association analysis identifies 20 loci that influence adult height. *Nat Genet* 2008; **40**: 575–583.
22 Friedman J, Vrijenhoek T, Markx S *et al*: *CNTNAP2* gene dosage variation is associated with schizophrenia and epilepsy. *Mol Psychiatry* 2008; **13**: 261–266.
23 Bearden CE, Freimer NB: Endophenotypes for psychiatric disorders: ready for primetime? *Trends Genet* 2006; **22**: 306–313.
24 Szatmari P, Maziade M, Zwaigenbaum L *et al*: Informative phenotypes for genetic studies of psychiatric disorders. *Am J Med Genet Part B* 2007; **144B**: 581–588.
25 Auranen M, Vanhala R, Varilo T *et al*: A genomewide screen for autism-spectrum disorders: evidence for a major susceptibility locus on chromosome 3q25–27. *Am J Hum Genet* 2002; **71**: 777–790.
26 Buxbaum JD, Silverman J, Keddache M *et al*: 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* 2004; **9**: 144–150.
27 IMGSAC: A genomewide screen for autism: strong evidence for linkage to chromosomes 2q, 7q, and 16p. *Am J Hum Genet* 2001; **69**: 570–581.
28 McCauley JL, Olson LM, Dowd M *et al*: Linkage and association analysis at the serotonin transporter (SLC6A4) locus in a rigid-compulsive subset of autism. *Am J Med Genet Part B* 2004; **127B**: 104–112.
29 National Institutes of Health: Medical Subject Headings (MeSH(R)). *National Library of Medicine* 2007, Available from URL http://www.nlm.nih.gov/mesh.
30 Butte AJ, Kohane IS: Creation and implications of a phenome-genome network. *Nat Biotechnol* 2006; **24**: 55–62.
31 Lage K, Karlberg EO, Storling ZM *et al*: A human phenome-interactome network of protein complexes implicated in genetic disorder. *Nat Biotechnol* 2007; **25**: 309–316.
32 Tiffin N, Kelso JF, Powell AR, Pan H, Bajic VB, Hide WA: Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Res* 2005; **33**: 1544–1552.
33 van Driel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JAM: A text-mining analysis of the human phenome. *Eur J Hum Genet* 2006; **14**: 535–542.
34 The centre for applied genomics: Database of Genomic Variants; Department of Genetics and Genomic Biology, MaRS Centre,Canada 2006(Build 36 (Mar 2006))Available from: URL: http://projects.tcag.ca/variation/.
35 Steiner CE: On macrocephaly, epilepsy, autism, specific facial features, and mental retardation. *Am J Med Genet Part A* 2003; **120A**: 564–565.
36 Veenstra-VanderWeele J, Cook EH: Molecular genetics of autism spectrum disorder. *Mol Psychiatry* 2004; **9**: 819–832.
37 Lonardo F, Parenti G, Luquetti D *et al*: Contiguous gene syndrome due to an interstitial deletion in Xp22.3 in a boy with ichthyosis, chondrodysplasia punctata, mental retardation and ADHD. *Eur J Med Genet* 2007; **50**: 301–308.
38 Macarov M, Zeigler M, Newman J *et al*: Deletions of VCX-A and NLGN4: a variable phenotype including normal intellect. *J Intellect Disabil Res* 2007; **51**: 329–333.
39 Canitano R: Epilepsy in autism spectrum disorders. *Eur Child Adolesc Psychiatry* 2006; **16**: 61–66.
40 Danielsson S, Gillberg IC, Billstedt E, Gillberg C, Olsson I: Epilepsy in young adults with autism: a prospective population-based follow-up study of 120 individuals diagnosed in childhood. *Epilepsia* 2005; **46**: 918–923.
41 Tuchman R, Rapin I: Epilepsy in autism. *Lancet Neurol* 2002; **1**: 352–358.
42 Arndt TL, Stodgell CJ, Rodier PM: The teratology of autism. *Int J Dev Neurosci* 2005; **23**: 189–199.
43 Hardan A, Keshavan MS, Sreedhar S, Vemulapalli M, Minshew NJ: An MRI study of minor physical anomalies in autism. *J Autism Dev Disord* 2006; **36**: 607–611.
44 Lauritsen MB, Mors O, Mortensen PB, Ewald H: Medical disorders among inpatients with autism in Denmark according to ICD-8 a nationwide register-based study. *J Autism Dev Disord* 2002; **32**: 115–119.
45 Rodier PM: 2003 Warkany lecture: autism as a birth defect. *Clin Mol Teratology* 2004; **70**: 1–6.
46 Wier ML, Yoshida CK, Odouli R, Grether JK, Croen LA: Congenital anomalies associated with autism spectrum disorders. *Dev Med Child Neurol* 2006; **48**: 500–507.
47 Merks JHM, zgen HM, Cluitmans TLM *et al*: Normal values for morphological abnormalities in school children. *Am J Med Genet Part A* 2006; **140A**: 2091–2109.
48 Hultman CM, Sparén P, Cnattingius S: Perinatal risk factors for infantile autism. *Epidemiology* 2002; **13**: 417–423.
49 Horvath K, Papadimitriou JC, Rabsztyn A: Gastrointestinal abnormalities in children with autistic disorder. *J Pediatr* 1999; **135**: 559–563.
50 Erickson CA, Stigler KA, Corkins MR, Posey DJ, Fitzgerald JF, McDougle CJ: Gastrointestinal factors in autistic disorder: a critical review. *J Autism Dev Disord* 2005; **35**: 713–727.
51 Kuddo T, Nelson KB: How common are gastrointestinal disorders in children with autism? *Curr Opin Pediatr* 2003; **15**: 339–343.
52 Miller MT, Strömland K, Ventura L, Johansson M, Bandim JM, Gillberg C: Autism associated with conditions characterized by developmental errors in early embryogenesis: a mini review. *Int J Dev Neurosci* 2005; **23**: 201–219.
53 Gurney JG, McPheeters ML, Davis MM: Parental report of health conditions and health care use among children with and without autism. *Arch Pediatr Adolesc Med* 2006; **160**: 825–830.
54 Titomanlio L, Marzano MG, Rossi E *et al*: Case of Myhre syndrome with autism and peculiar skin histological findings. *Am J Med Genet* 2001; **103**: 163–165.
55 Whiteley P: Developmental, behavioral and somatic factors in pervasive developmental disorders: preliminary analysis. *Child Care Health Dev* 2004; **30**: 5–11.
56 Bolton P, Powell JE, Rutter M *et al*: Autism, mental retardation, multiple exostoses and short stature in a female with 46,X,t(X;8)(p22.13;q22.1). *Psychiatr Genet* 1995; **5**: 51–55.
57 Lohiya GS, Tan-Figueroa L, Iannucci A: Identification of low bone mass in a developmental center: finger bone mineral density measurement in 562 residents. *Am J Med Dir Assoc* 2004; **5**: 371–376.
58 Biesecker LG: Mapping phenotypes to language: a proposal to organize and standardize the clinical descriptions of malformations. *Clin Genet* 2005; **68**: 320–326.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)