Conserved signals of Non-coding RNA across a set of 73 genes associated with Autistic Spectrum Disorders

Theodor B. Rais
Email: Theodor.Rais@utoledo.edu
General Background

- Autism is considered a major developmental disorder which is manifested by lack of adequate social interactions and communication together with stereotyped, rigid and restrictive behaviors.

- 556% increase in the prevalence of Autism in children as measured between 1991 and 1997.

- Autism is not a disorder by itself but a syndrome with multiple causes which could be of genetic nature or not.
General Background

By the term of Autistic Spectrum Disorders we mean a large variation of disorders related to early development and characterized by major impairments in the following domains:

- a) Social interaction and affective expression
- b) Language development and communication skills
- c) Adequate play
- d) Reduced imagination and interests
- e) Narrowed field of activities
A review of the specialty literature published between 1961 and 2007 gives solid evidence about a variable but significant genetic component of ASD’s. However there is clear evidence that different other etiological factors may be responsible for the appearance of ASD’s like: in utero toxic exposure, infections (rubella, cytomegalovirus), etc.

Among the clinical conditions regularly associated with ASD one can find:

- Mental Retardation (75% of cases of classic Kanner’s Autism have an IQ < 70 (Jamain et al, 2003))

- Epilepsy (30% of classic Autism subjects) and macrocephaly (20%) (Jamain et al, 2003)
There are diseases which are very often associated with Autism like Tuberous Sclerosis, Fragile X. However this category of diseases together with other medical or genetic conditions associated with Autism account only for less than 10% of cases (Muhle et al, 2004).

The majority of ASD cases are without a clear evident etiology, “idiopathic” or let’s say “essential” in their nature.

The clinical presentation is variable ranging from cases with dysmorphic appearance, mental retardation and seizures to children with normal appearance and IQ who present only with moderate social insertion and functioning difficulties.
The concerted effort in understanding and modeling the ASD’s for years was to apply linear and parsimonious models of inquiry that tried to establish a cause effect relationship between the clinical expression and specific neurotransmitters, neuronal injuries, and most of the time single genes.
Genetic Disorders associated with Autism

- The cluster of autistic symptoms were very often reported in cases of Tuberous Sclerosis (TSC) and Fragile X Syndrome (FXS) (Baker, 1998, Rogers, 2001).

- TSC is an autosomal dominant neurocutaneous disorder that occurs secondary to two types of mutations: TSC1 on 9q and TSC2 on 16p.

- The disorder is clinically characterized by the appearance of hamartomatous formations throughout the body together with depigmented or ash colored skin lesions. These lesions appear in the brain as well and they are named “tubers”.

FXS is a known chromosome X linked genetic disorder that as said above was reported to be associated with many cases of autism.

The cause for the disease is the appearance of multiple trinucleotide (CGG) repeats in a gene responsible for production of a protein known as the fragile x mental retardation protein. 30% of patients with FXS have ASD’s (Rogers et al, 2001). However this number has been debated and some studies contested the association of FXS with Autism (Fombonne, 1999).

Neurofibromatosis is a fairly common autosomal dominant disease with cutaneous and neurologic manifestations and can associate with ASD (Fombonne, 1998).
Angelman syndrome (AS) and Prader-Willi syndrome (PWS) are the consequence of genetic deletions or one parent disomy of chromosome 15q11-q13 locus associated with pathological imprinting and mutations (Nichols & Knepper, 2001).

Duchene muscular dystrophy patients can present with symptoms of ASD (Komoto et al, 1984).

Several other single defect gene related disorders have been associated with Autism like: Sotos syndrome, Williams syndrome, Hypomelanosis, Moebius syndrome, etc (Gilberg & Coleman, 2000).

Only a minority of ASD’s patients are part of theses associations. Also the association between ASD’s and diverse mentioned conditions is not always present.
Idiopathic Autism

In the cases where the twin who was not affected was reevaluated with different criteria including a more comprehensive autistic phenotype that included communication skills and social interaction problems then the concordance rate in the UK sample went up to 92% in the MZ twins and 10% in the DZ twins (Bailey, 1995).

The existence of obsessive compulsive traits together with social interaction and communication skills milder deficits in the family members of autistic patients (Steffenburg, 1989, Hollander, 2003).
Idiopathic Autism

Altogether these pieces of evidence speak about interactions of multiple genes being at the roots of the autistic phenotypes and also about the heterogenic phenotypic presentation of autism. However despite all the accumulated evidence the genes involved in the idiopathic autism and their exact number is not known so far.
Also the “field” facts do not converge in any specific direction:

Abnormalities of cytogenetic nature were found on every chromosome of patients with autism (Gilberg, 1998).

Whole genome screens done so far in families with more than one member suffering from ASD show more than 10 genes to be involved in the appearance of clinical symptoms of ASD (Risch, 1999, Pickles, 1995).
The classic example of single mutation genetic heterogeneity of TSC which is caused by TSC1 on 9q and by TSC2 on 16p is a clear example the polymorphous etiology of ADS.

On the other hand cases where twins share 100% of genes show that other factors than genes can concur with the appearance of the illness symptoms.
No single genetic area or marker was identified for ASD yet and is it believed as many as 10 or by other sources 30 genes could be involved in their etiology (Risch, 1999,1990, Stoltenberg, 2000).

It is obvious that other than genetic factors play a role in ASD etiology because the concordance of the disease is less than 100% in monozygotic twins and the phenotypic expression varies as well (Satangelo, 2005).

Linkage studies have pointed toward very diverse areas of the genome and several areas of the brain to be associated with ASD.
Research Goals

1. Review the scientific evidence about the role of microRNA’s in the etiology of ASD

2. Gather the above evidence and formulate a direction of research to further investigate the role of microRNA’s in the pathology of ASD

3. Define a specific approach to achieve goal #2 by appropriate use of objective scientific tools

4. Analyze the results and discuss them
MicroRNAs (miRNAs) are small, RNA molecules encoded in the genomes of plants and animals (Figure 1). These highly conserved, ~21-mer RNAs regulate the expression of genes by binding to the 3'-untranslated regions (3'-UTR) of specific mRNAs.

Each miRNA is thought to regulate multiple genes, and since hundreds of miRNA genes are predicted to be present in higher eukaryotes (Lim 2003b) the potential regulatory circuitry afforded by miRNA is enormous.
http://www.youtube.com/watch?v=StWDQgSg3UE&feature=related
http://www.youtube.com/watch?v=keGPYWnnXJw&feature=related
Micro RNA’s and ASD

Micro RNA’s are known to be regulating the expression of many mRNA’s found in brain. The complex regulatory role of mRNA’s was proved in many brain related diseases like Tourette’s Syndrome, Fragile X syndrome, ASD.

The regulation work of miRNA is done through diverse metabolic and physiologic pathways like:
Its implication in brain synaptic plasticity as in Fragile X and Rett syndrome where the expression of brain related miR-184 is repressed by the binding of MeCP2 (methyl CpG binding protein) to its promoter and upregulated by the release of MeCP2 secondary to depolarization (Nomura, 2008).

GAD 67 mRNA was found to be reduced by 40% in a group of Autistic patients (Yip et al, 2007) causing abnormal regulation of GABA in Purkinje neurons which may affect the cortex cerebellar communication and cause disruption in the motoric and cognitive function which may cause autism.
Kosik KS et al using multiple quantitative PCR (Abu-Elneel, 2003) compared the expression of 466 human miRNA’s taken postmortem from 13 ASD subjects and compared it with a control group of 13 non ASD sample of brains. Most miRNA levels showed no significant variation. However the sample being small it is hard to draw a conclusion about the miRNA expression. 28 miRNA’s were found to be expressed at different levels in the ASD sample. The authors reversed the analysis and compared each non-ASD control to a mean values for each one of the miRNA’s through all the ASD sampled cases. This reduced the number of significantly different miRNA’s to 9. It was significant that some of the targets of these miRNA’s were genes known to be associated with ASD like Neurexin and SHANK3.
Fragile X syndrome is the commonest form of human inherited cognitive deficiency (Bassel, 2008) and probably the best understood cause of Autism. Here the excess and disturbed mRNA translation causes as said above troubles in the synaptic function and loss of protein synthesis dependent plasticity.
Approach to achieve the goals

1. Select the genes involved in ASD

2. Review the criteria for the linkage of specific genes to ASD

3. Select only the genes linked by most objective scientific evidence

4. Because most of the studies linking proposed genes to the idiopathic ASD were not replicated look for any cues/evidence at the level of their introns
Approach to achieve the goals

5. One may hypothesize that candidate gene and linkage studies were not replicated because changes in the structures of introns may have been responsible of certain ASD phenotypes.
Approach (continued)

Why study introns?

Conservation of regions in the genome across species is a common phenomena observed in areas which are supposedly vital for the functioning of the organism.

These areas mainly include: exons of genes, the promoter regions for the genes, splice sites of introns, transcription factor binding sites, etc.
Approach

- Apart from the afore mentioned high conservation segments of the genome there could exist some regions which could code for non-coding RNAs (ncrna) which could include snoRNAs and miRNA.

- We hypothesized that small non-coding RNA (ncRNA), like miRNA, siRNA or orphan snoRNAs, could be involved in the pathogenesis of ASD.
Approach

- The complex regulatory roles of miRNA’s have been implicated in diseases like Tourette’s syndrome and Fragile X syndrome.

- Since ncRNAs have been found abundantly in introns, we set out to find evolutionarily conserved areas in the introns of ASD associated genes, which may play a causal role in some clinical phenotypic forms of ASD.
Our task was to try and fundament the role of ncRNA’s in ASD by using Bioinformatics’ tools. The working hypothesis was that taken the possible role of miRNA’s in the etiology of Autism we may find well conserved areas in the introns of ASD associated genes as a potential proof for the role that the ncRNA may have in causing some clinical phenotypic forms of ASD.
Method of work

1. Perform a comprehensive search for ASD genes
2. Select the ASD genes by most objective criteria
3. Use orthologues to obtain the most reliable alignments within appropriate degree of conservation
4. Extract the introns from the orthologue genes
5. Perform the introns alignments and analyze them
6. Elaborate an algorithm to analyze the introns alignments for conservation.

7. Repeat the procedure in step 1-6 with a control group

8. Choose a matched group of genes that is minimally expressed in the brain

9. Elaborate a method of mathematical transformation of the data which will allow classic statistical analysis

10. Analyze the data and discuss the results
Working Protocol

- We performed a search in the NCBI databases using the Key word “Autism” and “Autistic Spectrum Disorders”.

- The most comprehensive and inclusive database was found at OMIM. All the hits were reviewed and on the first selection 124 relevant genes were identified.
On the second review we retained only those genes proved to be associated with ASD by:

- Linkage studies
- Presence of biological markers representing physiological or metabolic pathways like evidence of SNP’s or haplotypes
- Candidate genes association studies

After careful scrutiny we retained only 73 relevant genes.
ORTHOLOGUES AND
MULTIPLE ALIGNMENTS

Orthologues of the respective genes related proteins in dog, mouse and rat were identified.

The following organisms were used in such way that existed a significant distance between the organisms and the human in the evolutionary time scale. Using for example chimp or macaque would result in too much conservation, while using invertebrates could show very few areas of conservation.
The amino acid sequence for the human protein was used to find the orthologues in the other organisms using BLAST and a “Home made” PERL script.

After procuring the orthologous proteins in the four organisms we used the intron exon database at UT website (http://hsc.utoledo.edu/bioinfo/eid/index.html), to extract the introns for these orthologous genes.
A PERL program which was called atsm_intsk_clustv6.pl, created four-way alignments for these introns of different organisms, the method essentially used clustalw on a UNIX platform to render the alignments.
ANALYSIS FOR CONSERVATION

- The alignments of the corresponding introns among the 73 genes of autism needed to be analyzed for conservation.

- For this another program in PERL was developed called, stats_atsm_conservn4.pl. This program analyzes each intron alignment in an orderly fashion using a scanning algorithm.
The scanning is done by analyzing 50 nucleotides at a time in a “window”, and then analyzing the window for the percentage conservation. It allows the user to input a percentage of conservation they want to analyze for, and searches for areas in the intron with equal to or greater than the user defined percentage of conservation.
The window of 50 nucleotides uses a “STRETCHY” algorithm where once a 50 nucleotide segment with at least the user defined percentage of conservation is found, it then progresses one nucleotide at a time until the conservation for that segment drops below the queried percentage. This scanning is done in all reading frames of the intron.
Some of the considerations while selecting interesting sites are, firstly the ends of the intron are generally more conserved as they are adjacent to the exons and thus could be part of alternately spliced exons or could be splice sites.
Secondly segments were chosen such that they have frame shift mutations among the four organisms just to be doubly sure that the conservation we see will very unlikely be from alternatively spliced exons.

Through this protocol regions which are well conserved within the introns of the organisms for the 73 genes were isolated
CONTROL GROUP

To see if the conservation seen in the autism specific genes was by accident and could be seen all through the human genome irrespective of their gene, we studied a group of 75 proteins which were minimally expressed in the brain so as to not have any overlap with the autism related genes.
The 75 proteins were procured from NCBI protein search, (http://www.ncbi.nlm.nih.gov/sites/entrez?db=protein), and their expression values were searched for using UNIGENE (http://www.ncbi.nlm.nih.gov/unigene)
The control group was similarly processed like the group for autism specific genes and multiple alignments for human, dog, mouse and rat were created for these genes and analyzed using the same PERL program.

There were 552 introns in the control group after the selection based on the nucleotide length.
The PERL program which scanned the multiple alignments for degree of conservation and isolated the chunks of conserved sequences within the intron was also used to calculate statistics for the purpose of quantification of the data so as to be able to make a substantial claim that conservation in autism specific introns are greater than the control group.
The program calculates the length of the conserved chunks and the length of the intron in human, and the degree of conservation in the conserved chunk equals or exceeds a user defined lower limit for the percentage of conservation.
In order to normalize the conserved chunks based on their presence in varying lengths of introns, the following formula was used:

\[ C_{\text{normalized}} = \frac{C}{L} \times \left(\frac{TP}{100}\right) \]

- \( C_{\text{normalized}} \) -> The normalized chunk length
- \( C \) -> The actual chunk length in nucleotides
- \( L \) -> The length of the human intron in nucleotides
- \( TP \) -> Threshold percentage (user defined lower limit percentage for conservation)
Using the following formula the total length of normalized chunks for a given gene was calculated by adding up the chunks for respective introns.

This was done for both autism specific genes and the genes which are minimally expressed in the brain. The results are depicted in the graph below.
RESULTS Review and Analysis
GRAPH 01: indicates the $C_{\text{normalized}}$ in the Y axis and the genes in the X axis. The graph depicts conservation in autism specific introns.
GRAPH 02: indicates the $C_{\text{normalized}}$ in the Y axis and the genes in the X axis. The graph depicts conservation in introns of genes minimally expressed in the brain.
## Results

<table>
<thead>
<tr>
<th></th>
<th>ASD-related sample</th>
<th>Non-Brain-specific sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of genes</td>
<td>53</td>
<td>75</td>
</tr>
<tr>
<td>Total number of introns</td>
<td>450</td>
<td>522</td>
</tr>
<tr>
<td>Total number of conserved intronic segments</td>
<td>1763</td>
<td>705</td>
</tr>
<tr>
<td>Avg. number of conserved segments per intron</td>
<td>3.92</td>
<td>1.35</td>
</tr>
<tr>
<td>Median length of introns (bp)</td>
<td>3,062</td>
<td>1,173</td>
</tr>
<tr>
<td>Median length of conserved regions (bp)</td>
<td>65</td>
<td>64</td>
</tr>
<tr>
<td>Average length of introns (bp)</td>
<td>16,812</td>
<td>1,845</td>
</tr>
<tr>
<td>Average length of conserved regions (bp)</td>
<td>202</td>
<td>127</td>
</tr>
</tbody>
</table>
Results

The GRAPH 01 and GRAPH 02 are on the same scale and show the striking finding that autism specific introns at least in five genes have a conservation of greater than 30% while the highest conservation seen in the nonbrain introns is at about 22%. And looking at the raw data for the number of conserved chunks in the two groups there were:

- Autism specific introns: 186 conserved chunk count: 534
  Average chunks/intron: 2.871

- Non brain introns: 522 conserved chunk count: 705
  Average chunks/intron: 1.351
The next question will be to find a statistical method to prove a significant difference between the 2 groups.

- On the analysis of the two above graphs we can see that the gathered data is probably not following a normal distribution. I decided to analyze the data with two main goals in my mind:
  - Use objective statistical testing to demonstrate the data distribution
  - Find a statistical test to compare the two group based on the resolution of the first goal.
AUTISM set of genes Results

The kurtosis scores (far less than 3 which is expected for a normal distribution) and the skewness expected to be close to 0) are descriptive parameters indicating distribution is probably not normal.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Variance</td>
<td>171.996</td>
</tr>
<tr>
<td>Std. Deviation</td>
<td>13.114713</td>
</tr>
<tr>
<td>Minimum</td>
<td>.0000000</td>
</tr>
<tr>
<td>Maximum</td>
<td>49.838187</td>
</tr>
<tr>
<td>Range</td>
<td>7</td>
</tr>
<tr>
<td>Interquartile Range</td>
<td>7</td>
</tr>
<tr>
<td>Skewness</td>
<td>1.337</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>.899</td>
</tr>
</tbody>
</table>
## Tests of Normality

<table>
<thead>
<tr>
<th></th>
<th>Kolmogorov-Smirnov(a)</th>
<th>Shapiro-Wilk</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Statistic</td>
<td>df</td>
</tr>
<tr>
<td>C: l<em>100</em>0.5</td>
<td>.193</td>
<td>270</td>
</tr>
</tbody>
</table>

a Lilliefors Significance Correction
The Normality tests results are shown above. The KS test is used as a non-parametric test of equality of one-dimensional probability distributions which is generally used to compare a sample with a reference probability distribution. The KS statistics quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution. The null hypothesis is that the sample is drawn from the reference distribution. Basically the null hypothesis assumes normality of the sample to study. The Sig. score of 0.000 shows clearly that we can reject the null hypothesis of normality.

The Shapiro-Wilk test is in general testing the null hypothesis that a sample X1,...,Xn came from a normally distributed population. To accept the null hypothesis of normality the statistic needs to be close to 1. The statistic score of 0.828 and the Sig. of 0.000 shows clearly we need to reject the null hypothesis that assumes normality. The two tests concur in rejecting the null hypothesis of normality.
The box plots show the asymmetrical pattern with the right skewed distribution. The 25% and 75% are not symmetric and the mean and median are far from each other.
Non Brain Genes set

- The normality tests both of them reject the null hypothesis of assumed normality

<table>
<thead>
<tr>
<th></th>
<th>Kolmogorov-Smirnov(a)</th>
<th>Shapiro-Wilk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistic</td>
<td>df</td>
<td>Sig.</td>
</tr>
<tr>
<td>C:i<em>100</em>0.5</td>
<td>.281</td>
<td>.000</td>
</tr>
</tbody>
</table>

a Lilliefors Significance Correction
THE BOX PLOTS ARE CHARACTERISTIC FOR A SET OF DATA NOT NORMALLY DISTRIBUTED.
Mann Whitney test – It is a non parametric test for the sets of data which are not normally distributed. It was done to compare the two sets of data: The Autism set of genes and the non brain set of genes to find out if there is a statistical significant difference between them.

The Mann – Whitney test shows there is a clear and statistically significant difference between the two sets of data.

<table>
<thead>
<tr>
<th>Mann-Whitney Test</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ranks</strong></td>
</tr>
<tr>
<td>group</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>CI1000.5</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

**Test Statistics(a)**

<table>
<thead>
<tr>
<th>CI1000.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mann-Whitney U</td>
</tr>
<tr>
<td>Wilcoxon W</td>
</tr>
<tr>
<td>Z</td>
</tr>
<tr>
<td>Asymp. Sig. (2-tailed)</td>
</tr>
</tbody>
</table>

a Grouping Variable: group
Knowing what the chunks were could provide a vital clue towards their function and importance in causing the ASD.

Thus four genes from the autism specific genes which had higher than 30% conservation according to the graph above were selected and BLASTed against any knows ncRNA databases (http://biobases.ibch.poznan.pl/ncRNA/).

The results indicated that the conserved chunks did contain some known ncRNAs.
Critique of Structure Analysis

- Difficult to prove the results have objective meaning
- Fair chance of false positives
FIG 01: Secondary structures found within the conserved sequences

However current knowledge about ncRNAs is very much in premature state. We decided to look for local secondary structures within the conserved sequences.

From these secondary structures from (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi) we could hypothesize that the conserved chunks may be microRNAs and possibly may serve a function in the genetic causation of ASD.
Disscusion

- The results showed that the average number of conserved segments per intron in Autism specific genes was 3.9.
- While in the case of the genes minimally expressed in the brain it was 1.4.
- A non-parametric statistical test (Mann-Whitney) showed a significant difference in the degree of conservation between the two groups.
- Application of RNA Vienna package (for prediction of RNA secondary structures) showed that few conserved segment sequences have stable RNA structures.
- They also have cross-species evolutionary conservation.
Discussion

- The density of conservation is higher in the ASD genes selected group

- The weak points of the study relate to the very nature of the Idiopathic Autism:

- Selection of candidate genes based on the involved neurotransmitters when we do not clearly understand the etiology of a disease

- The non replication of the linkage studies (possible cohort effect)
Discussion

- Inability to check the strength of the samples from the quoted studies in the OMIM
- The strength of the work:
  - Solid design that proves clear difference between the groups
  - Reliable sample of minimally expressed brain genes
  - Some reinforcement from structure analysis
Discussion – Future Directions

- Computer mining may demonstrated that among our set of possible intronic ncRNAs sequences some are already present in the largest databases of experimentally verified ncRNAs (Fantom-3 RNAdb and fRNA database).

- Several hundreds of functional ncRNAs may be inside introns of the 73 ASD-associated genes. Based on these observations, the experimental focus is on finding novel mutations inside intronic ncRNAs of ASD patients.

- One can anticipate that these mutations will provide reliable biomarkers for autistic diseases.
These findings may help to open a new direction of inquiry on some possible mechanisms of development and genetic susceptibility of Autistic Spectrum Disorders.

“No! This is not the end.

It is not even the Beginning of the End

It is perhaps the End of the Beginning”

- Sir Winston Churchill