Assessing the Impact of Copy Number Variants on miRNA Genes in Autism by Monte Carlo Simulation

Maurizio Marralle¹, Nadia Ninfa Albanese¹, Francesco Calì², Valentino Romano¹,²*

¹ Dipartimento di Fisica e Chimica, Università di Palermo, Palermo, Italy, ² U.O.C. di Genetica Medica Laboratorio di Genetica Molecolare, Associazione Oasi Maria SS. (I.R.C.C.S.), Troina, Italy

Abstract

Autism Spectrum Disorders (ASDs) are childhood neurodevelopmental disorders with complex genetic origins. Previous studies have investigated the role of de novo Copy Number Variants (CNVs) and microRNAs as important but distinct etiological factors in ASD. We developed a novel computational procedure to assess the potential pathogenic role of microRNA genes overlapping de novo CNVs in ASD patients. Here we show that for chromosomes 1, 2 and 22 the actual number of miRNA loci affected by de novo CNVs in patients was found significantly higher than that estimated by Monte Carlo simulation of random CNV events. Out of 24 miRNA genes over-represented in CNVs from these three chromosomes only hsa-mir-4436b-1 and hsa-mir-4436b-2 have not been detected in CNVs from non-autistic subjects as reported in the Database of Genomic Variants. Altogether the results reported in this study represent a first step towards a full understanding of how a dysregulated expression of the 24 miRNAs genes affect neurodevelopment in autism. We also propose that the procedure used in this study can be effectively applied to CNVs/miRNA genes association data in other genomic disorders beyond autism.

Introduction

The Autism Spectrum Disorders (ASDs, MIM: 209850) are a heterogeneous group of childhood diseases characterized by abnormalities in social behaviour and communication, as well as patterns of restricted and repetitive behaviors [1]. The spectrum of autism reflects dimensional variability of each core symptom as well as the occurrence of co-morbid conditions (intellectual disability, epilepsy, dysmorphisms etc.). Like an epidemic, in part due to improved diagnostic tools, the global prevalence of autism and other pervasive developmental disorders has increased over the years reaching the current, impressive, figure of 1/160 children [2]. Twin studies have demonstrated a much higher concordance rates for the disease in monozygotic twins (92%) than in dizygotic twins (10%) [3,4], indicating a strong genetic basis for autism susceptibility, also supported by the presence of autistic features in several monogenic disorders (e.g., Fragile X syndrome, Tuberous sclerosis). However, despite these progresses the identity of genetic factors still remains unknown in the majority of patients and it is likely, that overall, the causes of autism are more complex than previously thought involving an interaction of genetic, epigenetic and environmental factors all interfering with the normal course of neurodevelopment [5–7].

In 2007 Copy Number Variants (CNVs) were for the first time recognized as important genetic factors in ASD [8]. CNVs are a class of inherited or de novo genomic mutations duplicating (“Gains”) or deleting (“Losses”) DNA segments >1 Kb, thus altering the normal dosage of the overlapping genes. According to their frequency they can be divided into unique, rare (<1%) or recurrent. Since 2007, many studies have investigated CNVs in autism, to assess their functional impact, the biological networks where the genes in these CNVs are involved and the general burden of CNVs in these individuals (e.g. see references [9–11]). Extended, multiplex families are more likely to carry heritable risk factors while, in contrast, sporadic ASD families have demonstrated a higher rate of de novo CNVs [8,12]. A striking observation was the high prevalence of de novo CNVs in both sporadic and familial cases of ASD compared with controls. Rare, de novo or inherited CNVs were observed in 5–10% of idiopathic ASD cases.

Studies performed so far have highlighted the pathogenic role of CNVs in terms of dosage change for protein-coding genes [10,13–15], without taking into account the potential involvement of non-coding RNA, particularly microRNAs genes (see reference [16] for a recent exception).

miRNAs are an important class of post-trascriptional regulators each governing the expression of tens or even hundreds of proteins in both differentiated cells and during development. miRNA transcripts undergo several processing steps occurring first in the nucleus, then in the cytoplasm where the mature 22–25 nucleotide-long miRNA [17,18] enacts mRNA translational repression or cleavage by binding to the 3’-untranslated region of their respective target miRNAs [19]. Only a few studies have investigated the ASD transcriptome in post-mortem brain samples, so far [20–22]. Other studies have used miRNA from lymphoblastoid cell lines or peripheral blood cells isolated from patients [23–26]. Disruption of miRNA expression has been repeatedly
reported in several microarray studies and believed to be linked to pathogenesis in autism [27–31]. However, lymphoblastoid cell lines are not the best proxy for neural tissue and only very few miRNAs displayed consistent dysregulation among different studies and patients.

In this paper, we aim to expand our knowledge on the pathogenic role of miRNA in autism by investigating the associations between de novo CNVs and miRNA genes. For this reason we developed a novel and powerful computational procedure based on Monte Carlo randomization [32] and applied it to several published de novo CNV datasets from patients. Our positive findings consist in the identification of 24 miRNA genes over-represented in de novo CNV from chromosomes nos. 1, 2 and 22 and therefore likely to play a pathogenic role in autism.

Results

Study design and preliminary analyses

The general strategy and steps of our study are outlined in the flow-chart of Figure 1. We developed the MAPCNVMIR programme to map all human miRNA genes from miRBase to 178 de novo CNVs from 192 autistic patients (“APL datasets”, see Table S1). Only 64 out of 178 CNVs were found to overlap at least one miRNA gene. In addition, 145 miRNA genes were identified within distinct or partly overlapping de novo CNVs (Table S1) spread over 20 chromosomes. No miRNA genes included in de novo CNVs were detected in chromosomes # 11, 13, 14, Y.

For each chromosome we compared the number and the length of all human miRNA genes reported so far in miRBase with the length of the chromosome. The fractional length of miRNA genes over the size of the chromosome ($R_1$ ratios) are of the order of $10^{-5}$ for all chromosomes except for chromosome 19 ($>10^{-4}$) (see Table 1). This result shows that the fraction of the length of the chromosome covered by miRNA genes is very similar for all chromosomes (except chr. 19). On the other hand, the fractional length of CNVs over the size of the chromosome ($R_2$ ratios), computed separately for CNV_Gains and CNV_Losses (Table 1) using the MAPCNVMIR software developed by us (see below), largely differ among the various chromosomes, from a value of 0.0006 up to value of 0.2618. This highlights that the fractional length of chromosomes covered by CNVs can be very dissimilar among chromosomes. Indeed, there are several instances (e.g., chr. # 10, 12, 15 and 22 for CNV_Gains and chr. # 7, 12, 16, 18, 21 and 22 for CNV_Losses) of CNVs affecting large regions of a particular chromosome. Therefore, an analysis was performed between the number of miRNA genes overlapping de novo CNVs and the $R_2$ ratios in order to find possible correlations. Furthermore, the number of miRNA genes is also variable. Figures 2a and 2b show that, for many chromosomes, the number of miRNA genes included in CNVs follow a linear trend as function of the total length of both CNV_Gains and CNV_Losses. In order to quantify this kind of correlation, values of correlation coefficients were calculated and were found to be larger for CNV_Gains ($r=0.75765$) than CNV_Losses ($r=0.32732$). This finding is consistent with the larger spread of CNV_Losses compared to CNV_Gains. On the other hand, some chromosomes do not follow the linear trend and are characterized by a large number of miRNA genes associated to CNVs (see Figure 2 and Table 1). For example, chromosome 22 for CNV_Gains and chromosome 2 for CNV_Losses appear as “outliers” since they include a much higher number of miRNA genes in de novo CNVs than expected. However, the latter analysis does not allow us to classify the observed dissimilarities according to a statistically significant criterion.

Monte Carlo randomization

In order to assess which “outliers” with a high number of overlaps are actually significantly different from the other chromosomes, we developed the SIMCNVMIR programme which implements a numerical analysis procedure for the identification of all instances (i.e., chromosomes) where the number of miRNA genes overlapping de novo CNVs is significantly higher than expected in case of random distribution of (simulated) CNVs. The steps of our computational analysis are reported in Figure 3. The results of this analysis are reported in Figure 4 and Table 2 and show that for chromosomes # 1, 2, and 22 the actual number of miRNA loci affected by de novo CNVs in patients is significantly higher (FDR-adjusted p-values <0.05) than that estimated by the simulated random CNV events. Specifically, CNV_Gains in chromosome # 22 and CNV_Losses in chromosomes 1 and 2 display an over-representation of microRNA genes (see Table 2 and Figures 4). In Table 2, note that for CNV_Loss in chromosome # 2 the number of “hits” is higher than the number of distinct miRNA genes (“Unique”), implying that the same miRNA gene is involved. Overall, there are 24 miRNA genes overlapping de novo CNVs in the three positive chromosomes (see Table 3). Only two, hsa-mir-4436b-1 and hsa-mir-4436b-2, have not yet been detected in CNVs from the general population (i.e., the DGV database).

Code scripts of the MAPCNVMIR and SIMCNVMIR programmes are available to readers from the following URL: http://fisicaechimica.unipa.it/cvnmirna/

Pathway analysis

The two autism-specific miRNA genes (hsa-mir-4436 b-1 and hsa-mir-4436b-2), are deleted in a patient who bears one CNV of 332,304 bp. Interestingly, they encode the same 3p and 5p mature microRNAs. According to Mirwalk [33], no validated targets are yet known for these two miRNAs. We therefore performed a
Table 1. Fractional lengths of miRNA genes and CNVs in relation to chromosome’s size.

| Chromosome | # of miRNA genes | Ratio #1 | Gain | Ratio #2 | #hits | Loss | Ratio #2 | #hits |
|------------|------------------|----------|------|----------|-------|------|----------|-------|
| 1          | 127              | 4.05 x 10^{-5} | 0.0241 | 5       | 0.0050 | 5     |
| 2          | 98               | 3.10 x 10^{-5} | 0.0009 | 0       | 0.0314 | 17    |
| 3          | 76               | 3.21 x 10^{-5} | 0.0041 | 1       | 0.0923 | 12    |
| 4          | 56               | 2.31 x 10^{-5} | N/A    | N/A     | 0.0886 | 4     |
| 5          | 67               | 3.03 x 10^{-5} | N/A    | N/A     | 0.0763 | 7     |
| 6          | 54               | 2.56 x 10^{-5} | 0.0019 | 0       | 0.0187 | 2     |
| 7          | 67               | 3.41 x 10^{-5} | 0.0103 | 4       | 0.1425 | 7     |
| 8          | 70               | 3.54 x 10^{-5} | 0.0068 | 0       | 0.0088 | 1     |
| 9          | 71               | 3.95 x 10^{-5} | 0.0371 | 9       | 0.0297 | 1     |
| 10         | 61               | 3.66 x 10^{-5} | 0.0822 | 4       | 0.0005 | 0     |
| 11         | 69               | 4.03 x 10^{-5} | 0.0002 | 0       | 0.0410 | 0     |
| 12         | 57               | 3.48 x 10^{-5} | 0.1691 | 13      | 0.1013 | 3     |
| 13         | 37               | 2.62 x 10^{-5} | 0.0009 | 0       | 0.0169 | 0     |
| 14         | 88               | 6.58 x 10^{-5} | 0.0006 | 0       | N/A    | N/A   |
| 15         | 57               | 4.68 x 10^{-5} | 0.1254 | 39      | 0.0742 | 7     |
| 16         | 48               | 4.40 x 10^{-5} | 0.0320 | 2       | 0.1032 | 9     |
| 17         | 83               | 7.89 x 10^{-5} | 0.0317 | 8       | 0.0256 | 1     |
| 18         | 30               | 2.74 x 10^{-5} | 0.0006 | 0       | 0.2618 | 2     |
| 19         | 108              | 1.47 x 10^{-4} | N/A    | N/A     | 0.0145 | 2     |
| 20         | 40               | 5.16 x 10^{-5} | N/A    | N/A     | 0.0301 | 2     |
| 21         | 15               | 2.65 x 10^{-5} | 0.0074 | 1       | 0.1183 | 3     |
| 22         | 36               | 5.59 x 10^{-5} | 0.0093 | 18      | 0.1473 | 24    |
| X          | 108              | 5.77 x 10^{-5} | 0.0304 | 8       | 0.0406 | 1     |

Ratio #1: The ratio between the sum of lengths of all miRNA genes in a chromosome and the total length of the chromosome; Ratio #2: The ratio between the sum of lengths of all de novo CNVs in a chromosome and the total length of the chromosome; # hits: For each chromosome, the total no. of identical and/or different miRNA genes included in all de novo CNVs detected in patients.

doi:10.1371/journal.pone.0090947.t001
Figure 2. Correlation graph between the no. of miRNA genes in de novo CNVs and the CNV/Chr. lengths ratio (Ratio #2) For each chromosome, the number of miRNA genes associated to CNVs is plotted as a function of the fractional length of CNV over the chromosome’s size for Gains (a) and Losses (b), respectively. The graphs show that whereas the majority of data points lay very close to the best-fit line, indicating that the two variables are positively correlated, few chromosomes instead behave as outliers in which certain CNVs appear to affect a no. of miRNA genes higher than expected (data used for the graphs were taken from Table 1).
doi:10.1371/journal.pone.0090947.g002

functional enrichment analysis on the predicted target mRNAs to highlight the significance of these two microRNAs in ASD pathogenesis. The results of this latter analysis, reported in Table 4, show that several pathways identified by miRPath have been already implicated in autism by previous studies, (in bold in Table 4): Lysine degradation [34], Drug metabolism - cytochrome P450 [35], Notch signaling pathway [31,36], HIF-1 signaling pathway [37], Vasopressin-regulated water reabsorption [38], Natural killer cell mediated cytotoxicity [39], 2NSD1 [40] and AMT [41] have been previously identified as autism candidate genes.

Discussion

In recent years, Copy Number Variants and microRNAs have emerged as potentially important etiological factors in ASD. However, until recently, in nearly all the studies, these two topics have been investigated separately in the context of the research of autism. The pathogenic role of CNVs has been interpreted in terms of their effect on the function of the overlapping protein-coding genes. On the other hand, miRNAs have been studied with the aim of uncovering changes in their level of expression in cells isolated from patients (see the Introduction for references) vs. control cells. In our study the focus is on the potential pathogenic role played by miRNA genes/de novo CNVs, instead. For this reason, we developed a new computational procedure implemented in a Fortran-written programme which allows to detect over-representation of miRNA genes in de novo CNVs in each chromosome. By this computational analysis based on Monte Carlo simulations we found that miRNA loci are under-represented in highly polymorphic and well-validated CNVs from the general population (i.e., the Database of Genomic Variants). Fourthly, in Table 1 the absence of “hits” for both CNV_Gain and CNV_Loss for chromosomes # 11, 13, 14, Y is simply due to the lack of autistic patients bearing de novo CNVs (Loss or Gain) overlapping miRNA genes (see Table S1). In turn, it can be speculated that, the lack of this type of patient may be ascribed to various factors such as the sample size, the use of low-resolution aCGH platforms, the occurrence of “protective” miRNA loci for autism in these chromosomes. Finally, a more suitable analysis could have involved more homogeneous CNV data from subjects (patients AND unaffected individuals) of the same ethnicity analyzed with aCGH platforms with similar resolution. This was indeed the case for the autistic sample (APL) we have used which was homogenous in relation to ethnicity in that all patients from the APL dataset were “Caucasians” (white north Americans and Europeans). In
our study, the use of heterogeneous data concerns instead the different aCGH platforms used (APL and DGV datasets) and the mixed ethnicity of individuals reported in the Database of Genomic Variants. We decided to use such heterogeneous data to increase the chance of collecting a higher number of patients with de novo CNVs. This decision had its strengths and drawbacks. For instance, the use of different aCGH platforms may have caused an under-estimation of the number of CNV/miRNA genes associations in positive chromosomes from the APL dataset. In contrast, the use of samples with mixed ethnicity from the DGV database does not seem to have limited the identification of miRNA genes/CNV association in common between DGV and APL datasets. In conclusion, though the use of heterogeneous CNV data may have limited the identification of additional miRNA/CNV associations, it did not prevent the identification of chromosomes with an enrichment of CNVs overlapping miRNA genes.

In our study, the occurrence of the same 22 deleted or duplicated miRNA genes detected in both patients and unaffected individuals (i.e., DGV) strongly suggest that they are low-penetrant risk factors for autism. Difference in penetrance for such duplicated/deleted miRNA genes would be explained by a variety of factors including: (i) prenatal exposure to environmental risk factors [43], (ii) presence/absence of functional SNPs in susceptibility protein-coding genes of autism [44], (iii) epistasis [45], (iv) epigenetic factors [6], (v) number and type of protein-coding genes co-existing in different CNVs overlapping the same miRNA. It is reassuring that other CNV studies have linked several miRNA
genes from this group to autism, and include: hsa-mir-1306, hsa-mir-185, hsa-mir-1286 and hsa-mir-649 genes [16,46], hsa-mir-200a and hsa-mir-429 [16,47], hsa-mir-200b and hsa-mir-149 [16]. Furthermore, hsa-miR-185 displays an 1.44-fold upregulation in lymphoblastoid cell lines from autistic patients [48]. Interestingly, this latter finding is consistent with the presence of 3 copies of the hsa-mir-185 gene in the 2 patients from our APL dataset.

To avoid an over-interpretation of our results we have adopted a stringent, conservative criterion according to which we propose that the 22 miRNA genes shared by unaffected individuals and patients should be considered as provisional candidates miRNA genes in ASD. On the other hand, hsa-mir-4436b-1 and hsa-mir-4436b-2, appear at the present time as strong pathogenic candidates in ASD. Unfortunately, no validated targets have yet been identified for these two miRNAs. However, functional annotation analysis carried out on predicted mRNA targets for these two miRNAs revealed, that 43% (6/14) of the statistically significant KEGG pathways obtained with hsa-miR-4436b-3p have been already implicated in autism in previous studies (referenced in Table 4), a finding which supports a pathogenic role for this miRNA.

During the preparation of our manuscript, Vaishnavi et al published an article also addressing the impact of microRNAs present in autism-associated Copy Number Variants [16]. Despite, several differences (methodological, type of CNV data used) distinguishing our study from that of Vaishnavi et al., it is worth noting that 8 miRNA genes have been found in common between the two studies (chr. 1: hsa-mir-429, hsa-mir-200a, hsa-mir-200b; chr. 2: hsa-mir-149; chr. 22: hsa-mir-185, hsa-mir-1306, hsa-mir-1286, hsa-mir-649).

Conclusions

Summing up, positive findings of our study include the identification of 24 miRNA genes over-represented in *de novo* CNVs from 3 chromosomes. Two miRNA genes from this group, hsa-mir-4436b-1 and hsa-mir-4436b-2, are likely to play a significant pathogenic role in autism since they have not been found in CNVs from unaffected individuals. We hope these results will lead experimental research towards a better understanding on the role played by miRNAs in autism. Finally, we propose that the novel procedure used in this study can be effectively applied to CNV/miRNA genes association data from other genomic disorders beyond autism.

---

Table 2. *de novo* CNVs from autistic patients with an overrepresented no. of miRNA genes.

| chr | GAIN | LOSS |
|-----|------|------|
|     | Unique | Hits | FDR-adjusted p-value | Unique | Hits | FDR-adjusted p-value |
| 1   | 5     | 5    | 0.45204               | 5     | 5    | 0.02309               |
| 2   | 0     | 0    | 0.84211               | 10    | 17   | 0.00449               |
| 3   | 1     | 1    | 0.45642               | 10    | 12   | 0.64896               |
| 4   | N/A   | N/A  | N/A                   | 4     | 4    | 0.98516               |
| 5   | N/A   | N/A  | N/A                   | 7     | 7    | 0.73942               |
| 6   | 0     | 0    | 0.84211               | 2     | 2    | 0.70911               |
| 7   | 2     | 4    | 0.17818               | 6     | 7    | 1                     |
| 8   | 0     | 0    | 0.84211               | 1     | 1    | 0.70911               |
| 9   | 9     | 9    | 0.17818               | 1     | 1    | 0.98516               |
| 10  | 3     | 4    | 0.84211               | 0     | 0    | 1                     |
| 11  | 0     | 0    | 0.84211               | 0     | 0    | 1                     |
| 12  | 13    | 13   | 0.46568               | 3     | 3    | 1                     |
| 13  | 0     | 0    | 0.84211               | 0     | 0    | 1                     |
| 14  | 0     | 0    | 0.84211               | N/A   | N/A  | N/A                   |
| 15  | 11    | 39   | 0.17818               | 6     | 7    | 0.70911               |
| 16  | 2     | 2    | 0.84211               | 8     | 9    | 0.70911               |
| 17  | 8     | 8    | 0.13920               | 1     | 1    | 1                     |
| 18  | 0     | 0    | 0.84211               | 2     | 2    | 1                     |
| 19  | N/A   | N/A  | N/A                   | 2     | 2    | 0.64896               |
| 20  | N/A   | N/A  | N/A                   | 2     | 2    | 0.64896               |
| 21  | 1     | 1    | 0.17818               | 3     | 3    | 0.64896               |
| 22  | 9     | 18   | 0.04349               | 9     | 24   | 0.07785               |
| X   | 8     | 8    | 0.19599               | 1     | 1    | 1                     |

Unique: the no. of distinct miRNA genes overlapping *de novo* CNVs in patients. Hits: the no. of identical or distinct miRNA genes overlapping *de novo* CNVs in patients. FDR-adjusted p-value: probability of obtaining a number of miRNA genes overlapping randomly-distributed CNVs larger than hits after correction for multiple testing. In bold, chromosomes displaying significant FDR-adjusted p-values (<0.05). N/A: no *de novo* CNVs are present in the APL dataset.

doi:10.1371/journal.pone.0090947.t002

MicroRNAs and Copy Number Variants in Autism: A Whole-Genome Analysis

PLOS ONE | www.plosone.org

March 2014 | Volume 9 | Issue 3 | e90947
Table 3. List of 24 microRNA genes overrepresented in de novo CNVs.

| miRNA gene name | Chr | Cytoband | Coordinates (GRCh37) | strand | Size (bp) | Clustered miRNA genes | DGV no. of Patients | Gain/Loss in Patients | Gain/Loss in positive chr. | Intra/Intergenic miRNA loci |
|-----------------|-----|----------|----------------------|--------|-----------|------------------------|----------------------|----------------------|--------------------------|-----------------------------|
| hsa-mir-2682    | 1   | 1p21.3   | 98510798 98510907    |        | 110       | hsa-mir-137; hsa-mir-2682. | 1 Loss/1 Gain       | L/L                  | none (miRNA gene)           | none                        |
| hsa-mir-137     | 1   | 1p21.3   | 98511066 98511277    |        | 102       | hsa-mir-2682.           | 1 Loss/1 Gain       | L/L                  | none (miRNA gene)           | none                        |
| hsa-mir-200b    | 1   | 1p36.33  | 1102344 1102532      | +      | 95        | hsa-mir-200b; hsa-mir-200a; hsa-mir-429. | 1 Loss/1 Gain       | L/L                  | Intergenic                | none                        |
| hsa-mir-200a    | 1   | 1p36.33  | 1102494 1102532      | +      | 95        | hsa-mir-200b; hsa-mir-200a; hsa-mir-429. | 1 Loss/1 Gain       | L/L                  | Intergenic                | none                        |
| hsa-mir-429     | 1   | 1p36.33  | 1102548 1102578      | +      | 95        | hsa-mir-200b; hsa-mir-200a; hsa-mir-429. | 1 Loss/1 Gain       | L/L                  | Intergenic                | none                        |
| hsa-mir-2467    | 2   | 2q37.3   | 240273419 240273499   | +      | 81        | -                      | 2/3 Gain/Loss       | L/G, L              | none                     | none                        |
| hsa-mir-3133    | 2   | 2q37.3   | 240273419 240273499   | +      | 81        | -                      | 2/3 Gain/Loss       | L/G, L              | none                     | none                        |
| hsa-mir-4436b-1| 2   | 2q13     | 110844010 110844100   | +      | 91        | -                      | 2/3 Gain/Loss       | L/G, L              | none                     | none                        |
| hsa-mir-4436b-2| 2   | 2q13     | 110844010 110844100   | +      | 91        | -                      | 2/3 Gain/Loss       | L/G, L              | none                     | none                        |
| hsa-mir-4267    | 2   | 2q37.3   | 240273419 240273499   | +      | 81        | -                      | 2/3 Gain/Loss       | L/G, L              | none                     | none                        |
| hsa-mir-4269    | 2   | 2q37.3   | 240273419 240273499   | +      | 81        | -                      | 2/3 Gain/Loss       | L/G, L              | none                     | none                        |
| hsa-mir-149     | 2   | 2q13     | 110844010 110844100   | +      | 91        | -                      | 2/3 Gain/Loss       | L/G, L              | none                     | none                        |
| hsa-mir-4761    | 2   | 2q13     | 110844010 110844100   | +      | 91        | -                      | 2/3 Gain/Loss       | L/G, L              | none                     | none                        |
| hsa-mir-1306    | 2   | 2q13     | 110844010 110844100   | +      | 91        | -                      | 2/3 Gain/Loss       | L/G, L              | none                     | none                        |
| hsa-mir-3618    | 2   | 2q13     | 110844010 110844100   | +      | 91        | -                      | 2/3 Gain/Loss       | L/G, L              | none                     | none                        |
| hsa-mir-130b    | 2   | 2q13     | 110844010 110844100   | +      | 91        | -                      | 2/3 Gain/Loss       | L/G, L              | none                     | none                        |
| hsa-mir-4786    | 2   | 2q13     | 110844010 110844100   | +      | 91        | -                      | 2/3 Gain/Loss       | L/G, L              | none                     | none                        |
| hsa-mir-4269    | 2   | 2q37.3   | 240273419 240273499   | +      | 81        | -                      | 2/3 Gain/Loss       | L/G, L              | none                     | none                        |
| hsa-mir-149     | 2   | 2q13     | 110844010 110844100   | +      | 91        | -                      | 2/3 Gain/Loss       | L/G, L              | none                     | none                        |
| hsa-mir-4761    | 2   | 2q13     | 110844010 110844100   | +      | 91        | -                      | 2/3 Gain/Loss       | L/G, L              | none                     | none                        |
| hsa-mir-1306    | 2   | 2q13     | 110844010 110844100   | +      | 91        | -                      | 2/3 Gain/Loss       | L/G, L              | none                     | none                        |
| hsa-mir-3618    | 2   | 2q13     | 110844010 110844100   | +      | 91        | -                      | 2/3 Gain/Loss       | L/G, L              | none                     | none                        |
| hsa-mir-130b    | 2   | 2q13     | 110844010 110844100   | +      | 91        | -                      | 2/3 Gain/Loss       | L/G, L              | none                     | none                        |
| hsa-mir-4761    | 2   | 2q13     | 110844010 110844100   | +      | 91        | -                      | 2/3 Gain/Loss       | L/G, L              | none                     | none                        |

Information on miRNA gene coordinates, size, gene clustering, intergenic/intragenic loci are as reported in miRBase; no. of patients: number of patients bearing a CNV (including a given miRNA gene); CNVs among patients may have identical or different start/end; Gain/Loss in patients: CNV type overlapping miRNA genes detected in patients (from Table S1); Gain/Loss in positive chr.: CNV type overlapping miRNA genes detected in simulation.

DOI: 10.1371/journal.pone.0090947.t003
Methods

Data and databases used

Data on de novo CNVs detected in autistic patients were downloaded from three different sources: (i) 71 CNVs from the Autism Chromosome Rearrangements Database [49,50]; (ii) 51 CNVs from Suppl. Table 8 of [10] and (iii) 75 CNVs from Table S1 (document S2) of [13]. Throughout our paper the combined three above sets of data will be named “APL datasets”. CNV (“APL datasets”) data used in this study are reported in the Table S1. CNVs and indels detected in individuals of the general population were downloaded from the Database of Genomic Variants (DGV). miRBase software was used to look for experimentally validated mRNA targets for miRNAs [33], miRPath [55,56] was used to identify statistically significant KEGG pathways enriched in the list of predicted miRNA targets (p<0.05; p-values were corrected to account for the False Discovery Rate).

MAPCNVMIR program (Python)

Data pre-processing included: (i) computation of the total DNA length accounted for by all miRNA genes (L1) and CNVs (L2) in each chromosome; (ii) computation of R1 (R1 = L1/chromosome length) and R2 (R2 = L2/chromosome length); (iii) counts of the number of miRNA loci overall included in distinct or overlapping CNVs (“hits”). This analysis is performed separately for CNV_Gains and CNV_Losses. Thus, for a given chromosome, “hits” may consist of distinct and/or identical miRNA genes associated to CNVs; on the other hand, we indicate as “unique” the distinct miRNA genes overlapping de novo CNVs in patients.

We developed the MAPCNVMIR programme in Python language to achieve a two-fold task: (i) to calculate for each chromosome the total length of DNA corresponding to the de novo CNV regions and (ii) to map the microRNA genes within the de novo CNVs detected in patients using their genomic coordinates (Build 37). In order to achieve the first task this programme compares its initial (C1) and last (C2) nucleotide positions (’start’ and ’end’) of the first CNV (of the total list of CNVs reported in Table S1) into a sub-array. Afterwards, the code considers the overlapping DNA regions of different CNVs once only. The programme first initializes for each chromosome an empty array and put the numeric values corresponding to the first (C1) and last (C2) nucleotide positions with

### Table 4. KEGG pathways enriched for targets of miRNAs hsa-mir-4436b-3p and -5p identified by mirPath1.

| KEGG PATHWAY                        | p-value     | # genes | Target Genes                      |
|-------------------------------------|-------------|---------|-----------------------------------|
| Lysine degradation                  | 3.30×10^-6  | 5       | NSD1, KMT2D, PIPOX, KMT2A, KMT2E.|
| Drug metabolism - cytochrome P450   | 6.73×10^-6  | 1       | CYP3A43.                          |
| Other glycan degradation            | 0.000197    | 1       | NEU1.                             |
| Glycerophospholipid metabolism     | 0.00029     | 4       | PNLPL7, PGS1, LPCAT3, MBOAT2.     |
| Notch signaling pathway             | 0.000942    | 4       | CTBP1, APH1A, NOTCH2, MAML1.      |
| HIF-1 signaling pathway             | 0.006618    | 5       | PGF, EGLN3, PLCG1, ENO3, SLC2A1. |
| Vasopressin-regulated water reabsorption | 0.00847    | 1       | AQP2.                             |
| Glycine, serine and threonine metabolism | 0.01361    | 3       | SHMT2, AMT, PIPOX.               |
| Renal cell carcinoma                | 0.016804    | 3       | PGF, EGLN3, SLC2A1.              |
| One carbon pool by folate           | 0.01935     | 2       | SHMT2, AMT.                      |
| Pathogenic Escherichia coli infection | 0.04271   | 3       | TUBA1A, TUBA8, TUBA1C.           |
| Graft-versus-host disease           | 0.04271     | 3       | KIR3DL2, KIR2DL3, KIR3DL1.        |
| Antigen processing and presentation | 0.044381    | 5       | NFYA, NEF8, KIR3DL2, KIR2DL3, KIR3DL1. |
| Natural killer cell mediated cytotoxicity | 0.048633   | 5       | KIR3DL2, KIR2DL3, TNFRSF10B, PLCG1, KIR3DL1. |
| Prion diseases                      | 2.16×10^-73 | 1       | PRNP.                             |
| Sulfur relay system                 | 2.63×10^-9  | 1       | NFS1.                             |
| Thiamine metabolism                | 0.002145    | 1       | NF51.                             |
| Lysine degradation                  | 0.008932    | 2       | WHSC1.                            |
| Transcriptional misregulation in cancer | 0.023567   | 2       | NCO1, WHSC1.                     |

1In bold pathways or genes previously implicated in autism, see main text.

doi:10.1371/journal.pone.0090947.t004

In Table 4, the KEGG pathways enriched for targets of miRNAs hsa-mir-4436b-3p and -5p identified by mirPath1. The p-values are corrected to account for the False Discovery Rate.
1. the first CNV is totally included in the second one (i.e. \( C_{2} < C_{1} \) and \( C_{2} > C_{4} \)) and the values inside the sub-array are replaced by these new ones;
2. the second CNV is partially or totally included in the first one [i.e. \( C_{2} < C_{1} \) and \( C_{2} < C_{4} \)] or \( (C_{1} < C_{2} \) and \( C_{4} < C_{2} \)) or \( (C_{4} < C_{2} \) and \( C_{2} > C_{4} \)) and the sub-array is composed of the minimum between \( C_{2} \) and \( C_{4} \) and the maximum between \( C_{1} \) and \( C_{2} \);
3. the second CNV does not overlap the first one and in this case a new sub-array with the \( C_{2} \) and \( C_{2} \) is added to the initial array.

Then another CNV is analyzed and its \( C_{3} \) and \( C_{3} \) values are compared with the values of the first and second CNVs and the values of the array are modified according the above-described procedure. This procedure is carried out for all CNVs and finally an array with the “start” and “end” values of non-overlapping DNA regions covered by different CNVs is achieved. The total length of CNVs in each chromosome is the sum of the lengths of the corresponding DNA regions. Regarding the mapping of the microRNA genes within the de novo CNVs detected in patients, for each chromosome, the programme first initializes a variable to microRNA genes within the corresponding DNA regions. Then another CNV is analyzed and its \( C_{3} \) and \( C_{3} \) values are compared with the values of the first and second CNVs and the values of the array are modified according the above-described procedure. This procedure is carried out for all CNVs and finally the programme then computes the number of miRNA genes overlapping a CNV is obtained for each chromosome.

Correlation analysis

On the data obtained by the MAPCNVMIR programme (see above) we performed an analysis to evaluate possible correlation between the number of miRNA genes overlapping de novo CNVs (“hits”) and the fractional length of CNVs in relation to the size of the chromosome \( (R_2 \) ratios). Briefly, for each chromosome the number of “hits” was plotted against the \( R_2 \) ratios, thus obtaining the linear best fit functions and correlation coefficients were calculated. These analyses were performed separately for CNV_Gains and CNV_Losses.

SIMCNVMIR program (FORTRAN)

The SIMCNVMIR programme was developed to perform Monte Carlo randomization analyses, separately for CNV_Gains and CNV_Losses in each chromosome. These analyses included: (i) simulation of random CNV events, (ii) generation of a frequency distribution of “hits” in simulated CNVs, (iii) computation of \( p \)-values and FDR-adjusted \( p \)-values and (iv) selection of chromosomes displaying over-representation of miRNA genes in de novo CNVs from patients.

The analysis aimed at evaluating an over-representation of miRNA genes in CNVs was carried out by means of a computational simulation procedure implemented by a homemade written FORTRAN code. The null hypothesis underlying our investigation is that the distribution of CNVs within the chromosome is absolutely random, that is, they can occur anywhere throughout the whole length of a chromosome. Therefore, for each chromosome, the sizes (number of nucleotides) of all CNVs reported in the APL datasets (Table S1) are computed and various realizations of random distributions of these CNV regions within each chromosome are simulated. Once a new distribution of CNVs is obtained, the programme then computes the number of miRNA genes overlapping the simulated CNVs. This procedure was repeated \( 10^6 \) times for each chromosome. Data are then plotted as histograms displaying the occurrence frequency of miRNA genes associated to each CNV. These histograms provide information on the number of times a certain number of miRNAs genes are found to overlap CNVs randomly-distributed in each chromosome. These distributions take into account many factors such as the size of the chromosome, the number and positions of miRNA genes inside the chromosomes, and the size of CNVs. For each chromosome, the number of miRNA genes overall associated to the experimentally observed CNVs (i.e., the CNVs reported in the APL datasets) was compared to the corresponding histogram obtained by the simulation. In order to evaluate whether the number of miRNA genes included in de novo CNV in patients is significantly larger than expected with a random distribution of CNVs in each chromosome, we estimated the probability \( (p\text{-value}) \) of obtaining a number of miRNA genes associated to the simulated CNVs larger than that seen with experimental CNVs. This probability is calculated by summing the area under the histogram for a number of miRNA genes included in CNVs larger than or equal to the experimental value. The \( p \)-value is very small if the number of miRNA genes included in experimental CNVs is much larger than the mean value. This means that if the distribution of the CNVs on a given chromosome was random (see the above-mentioned null hypothesis) we would have a low probability of finding a greater number of miRNA genes associated to CNVs. In other words, in autistic patients, CNVs tend to be more frequent in chromosomal regions where the miRNA genes are present than in other regions of the chromosome. The analyses described above have been performed twice: (i) for CNV_Gains and (ii) for CNV_Losses respectively. In our analysis we used a false discovery rate procedure (as multiple hypothesis testing) developed by Benjamini and Hochberg to control the expected proportion of incorrectly rejected null hypotheses [57]. In particular, we exploited a spreadsheet available on-line [58] which calculates FDR-adjusted \( p \)-values from the knowledge of the \( p \)-values for the various chromosomes. We set acceptable FDR 0.05 as a maximum (which is the default value of the spreadsheet).

Supporting Information

Table S1 The APL dataset of de novo CNVs and the overlapping miRNA genes.

Author Contributions

Conceived and designed the experiments: MM VR. Performed the experiments: MM NNA FC. Analyzed the data: MM VR. Contributed reagents/materials/analysis tools: MM. Wrote the paper: VR.

References

1. Abrahams B, Geschwind D (2008) Advances in autism genetics: On the threshold of a new neurobiology. Nature Reviews Genetics 9: 341–355.
2. Elabbash M, Dvian G, Koh YJ, Kim Y, Kaushali S, et al. (2012) Global Prevalence of Autism and Other Pervasive Developmental Disorders. Autism Research 5: 160–179.
3. Bailey A, Le Couteur A, Gottesman I, Bolton P, Simonoff E, et al. (1995) Autism as a strongly genetic disorder: Evidence from a British twin study. Psychological Medicine 25: 63–77.

4. Steffenburg S, Gillberg C, Hellgren L, Andersson L, Gillberg I, et al. (1989) A twin study of autism in Denmark, Finland, Iceland, Norway and Sweden. Journal of Child Psychology and Psychiatry and Allied Disciplines 30: 405–416.

5. Schaaf C, Zoghbi H (2011) Solving the Autism Puzzle A Few Pieces at a Time. Neuron 70: 806–808.

6. Miyake K, Hirasaawa T, Kode T, Kubota T (2012) Epigenetics in autism and other neurodevelopmental diseases. Advances in Experimental Medicine and Biology 724: 91–98.

7. LaBelle J (2011) A genomic point-of-view on environmental factors influencing the human brain: Epigenetics. Epigenetics 6: 892–899.

8. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, et al. (2007) Strong association of de novo copy number mutations with autism spectrum disorder. Science 315: 44–49.

9. Glesner J, Wang K, Cai G, Koivatsos O, Kimm C, et al. (2009) Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. Nature 459: 569–72.

10. Parnell D, Pagamnenta A, Kiss I, Arnay R, Merico D, et al. (2010) Functional impact of global rare copy number variation in autism spectrum disorders. Nature 466: 368–372.

11. Sanders S, Eccan-Sencicek A, Hsu V, Luo R, Murtha M, et al. (2011) Multiple Recurrent De Novo CNVs, Including Duplications of the 7q11.23 Williams Syndrome Region, Are Strongly Associated with Autism. Neuron 70: 865–875.

12. Morrow E, Yoo SY, Flavell S, Kim TK, Lin Y, et al. (2008) Identifying autism loci and genes by tracing recent shared ancestry. Science 321: 220–223.

13. Levy D, Ronemus M, Yamron E, Lee YH, Llao A, et al. (2011) Rare De Novo and Transmitted Copy-Number Variation in Autism Spectrum Disorders. Neuron 70: 886–897.

14. Curcio I, Medrano A, Gener B, Vilaridel M, Gallastegui F, et al. (2009) Autism-specific copy number variants further implicate the phosphatidylinositol signaling pathway and the glutamatergic synapse in the etiology of the disorder. Human Molecular Genetics 18: 1795–1804.

15. Rornemus M, Issafo A, Levy D, Wgler M (2014) The role of de novo mutations in the genetics of autism spectrum disorders. Nature Reviews Genetics 15: 133–141.

16. Vaishnavi V, Manikandan M, Tiswary BK, Munirajan AK (2013) Insights on the Functional Impact of MicroRNAs Present in Autism-Associated Copy Number Variants. PLoS ONE 8(6): e67601.1–13.

17. Bartel D (2004) MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. Cell 116: 281–297.

18. Lee Y, Ahn C, Han J, Choi H, Kim J, et al. (2003) The nuclear RNAase III Drosha initiates microRNA processing. Nature 425: 415–419.

19. van den Berg A, Mols J, Han J (2008) RISC-target interaction: Cleavage and translational suppression. Biochimica et Biophysica Acta - Gene Regulatory Mechanisms 1779: 668–677.

20. Abou-Elneel K, Liu T, Gazazinta F, Nishimura Y, Wall D, et al. (2008) Heterogeneous dysregulation of microRNAs across the autism spectrum. Neurogenetics 9: 153–161.

21. Garbett K, Ebert P, Mitchell A, Lintas C, Manzi B, et al. (2008) Immune transcriptome alterations in the temporal cortex of subjects with autism. Human Neurobiology 179: 689–699.

22. Bressler JP, Gillin PK, O’Driscoll C, Kiihl S, Solomon M, et al. (2012) Maternal antibody reactivity to lymphocytes of offspring with autism. Pediatric Neurology 47: 337–340.

23. Rowsbaum JD, Cai G, Nguyen G, Chaste P, Delorime R, et al. (2007) Mutation analysis of the ml1 gene in patients with autism spectrum disorders and macrocephaly. BMC medical genetics 8: 68.

24. Yu TW, Chahrour MH, Gouler ME, Jirzlaneon P, Okamura-Ikeda K, et al. (2013) Using whole-exome sequencing to identify inherited causes of autism. Neuron 77: 239–273.

25. Marcinkowska M, Szymanski M, Krzyzosiak W, Kozlowski P (2011) Copy number variation of microRNA genes in the human genome. BMC Genomics 12: 1–13.

26. Gardner H, Spiegelman D, Buka S (2009) Prenatal risk factors for autism: Comprehensive meta-analysis. British Journal of Psychiatry 195: 7–14.

27. Sanders S, Murtha M, Gupta A, Murdoh J, Rabaoson M, et al. (2012) De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature 484: 237–241.

28. Coutinho A, Sousa I, Martins M, Correia C, Morgadinho T, et al. (2007) Evidence for epistasis between SLCO6A and ITGB3 in autism etiology and in the determination of platelet serotonin levels. Human Genetics 121: 243–256.

29. Xu B, Karayorgou M, Gogos JA (2010) MicroRNAs in psychiatric and neurodevelopmental disorders. Brain research 1338: 78–88.

30. Qiao Y, Badduce C, Mercer E, Lewis SM, Pavlidis P, et al. (2013) mirna and mirna target genes in copy number variations occurring in individuals with intellectual disability. BMC genomics 14: 344.

31. Sarachana T, Zhou R, Chen G, Manji H, Hu VW (2010) Investigation of post-transcriptional gene regulatory networks associated with autism spectrum disorders by microRNA expression profiling of lymphoblastoid cell lines. Genome Med 2: 23.

32. Marshall C, Noor A, Vincent J, Lionel A, Feuk L, et al. (2008) Structural Variation of Chromosomes in Autism Spectrum Disorder. American Journal of Human Genetics 82: 477–480.

33. Autism Chromosome Rearrangements Database (version 2012). Available: http://users.ox.ac.uk/.

34. Marshall C, Noor A, Vincent J, Lionel A, Feuk L, et al. (2008) Structural Variation of Chromosomes in Autism Spectrum Disorder. American Journal of Human Genetics 82: 477–480.

35. Autism Chromosome Rearrangements Database (version 2012). Available: http://projects.tcg.ca/autism/1.7.

36. Zhang J, Frick L, Duggan G, Khaja R, Scherer S (2006) Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. Cytogenetic and Genome Research 115: 205–214.

37. Griffiths-Jones S, Grocock R, von Dangos B, Bateman A, Enright A (2006) MicroRNA version 20. Available: http://diana.imis.athena-innovation.gr.

38. Sysmex M, Szymanski M, Krzyzosiak W, Kozlowski P (2011) Copy number variation of microRNA genes in the human genome. BMC Genomics 12: 1–13.