We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

8

Identifying Variations Within Unstable Regions of the Genome Reveal Autism Associated Patterns

Abdullah Alqallaf and Ali Hajjiah
Electrical Engineering Department, College of Engineering and Petroleum, Kuwait University, Kuwait city, Kuwait

1. Introduction

Autism is a collection of neurodevelopment and abnormal behaviors which can be characterized by social isolation, language deficits and repetitive or stereotyped behaviors. It is a lifelong disorder that starts at early childhood and becomes apparent before three years old up to adulthood that ranging in severity from case to case. Autism spectrum disorder (ASD) has received a great deal of attention in recent years since the apparent prevalence of children with this spectrum of neurological and behavioral deficits is on the rise. It is currently estimated to be approximately 1 in 150 children based on a 14 state survey conducted by the Centers for Disease Control (CDC) in the United States of America (Kuehn, 2007) and it is predominately in males with a ratio of approximately of 4 males to 1 female (Fombonne, 2003).

While it is hotly debated in both the lay and academic communities as to whether ASD incidence is truly increasing and not just a function of increased reporting and changes in diagnostic criteria, it is uncontested that the number of children diagnosed with ASD presents an important pediatric health problem. The social and economical impacts on individuals with ASD and their families as well as the society maybe considerably reduced if early identification and diagnosing can be achieved using simple and accurate approach. Although it is initially described in the 1940s, the exact etiology and pathology of ASD remains rudimentary and challenging. A number of studies have reported links between the development of the ASD and various factors such as genetics, environmental, immunological, nutritional and neurological. It is likely to result from a combination of these factors. Different methodologies have been proposed to identify and diagnose ASD using different criteria. The autism diagnostic observation schedule (ADOS) is a protocol consists of a series of structural tasks that involve social interactions used to diagnose and assess ASD. Others are using functional magnetic resonance imaging (f-MRI) to scan the brain as pattern recognition method of the defected neurons in the autistic individuals. However, these methodologies depend on the interactions between the examiner and the patient. On the other hand, studying the function of the biological system provides alternative way to embrace the complexity of ASD. Although the neurobiological and genetics basis of ASD and related disorders is unclear, multiple lines of evidence have

converged on abnormal brain functions. Using previous knowledge of biological processes and protein interactions of neurological disorders related to ASD, there were able to identify several genes and genetic contributors that had been strongly associated to ASD (Sebat et al., 2007 & Abrahams, 2008). Alterations of these contributors have been proposed as a factor involved in the etiology of ASD.

Understanding the biological mechanisms related to ASD at early stage is essential for identifying and diagnosing the disease and will lead to better treatments. Our main objective in this chapter is to understand the molecular and cellular underpinnings of ASD by identifying the genetic contributors to this set of complex disorders. We are also keenly interested in developing DNA-based methods that can serve to improve our diagnostic evaluation of ASD. Accurate and simple diagnostic methods would go a long ways in promoting early and appropriate interventions. Our research is grounded in recent work showing that deletions and duplications of DNA contribute a very significant degree of genetic variation in human populations. Finally, the work presented in this chapter focuses primarily on determining if DNA copy number changes are associated with ASD.

2. High-resolution genetic data

Data on genome structural and functional features for various organisms is being accumulated and analyzed aiming to explore in depth the biological information and to convert data into meaningful biological knowledge. To date, different experimental technologies such as microarray and DNA sequencing had been proposed to generate high-resolution genetic data and to understand the complex dynamic interactions between complex diseases and the biological system components of genes and genes products. These approaches made it possible to enhance our understanding of biological variations in healthy and diseased organisms through computational-based models. However, these technologies contain many sources of errors. Some types of errors are of our interests that have biological origins. Other types of errors are undesirable and need to be eliminated before further analysis. In particular, these technologies produce certain systematic sources of errors due to the experimental design process used in generating the genetic data such as labeling, printing, and scanning the examined samples. Figure 1 illustrates a simple description of generating DCN data using aCGH technology. Identifying the genomic locations and genetic contributors responsible for these variations is a problem of great importance to biologists. Current estimates indicate that DNA sequence differences due to changes in DNA copy number account for 3-4 fold more variation than that provided by single nucleotide polymorphisms, the most widely studied type of variation. It is also apparent that certain segments of the genome are susceptible to copy number alterations on account of particular sequence features, such as low copy repeats (LCRs).

LCRs are relatively large (>1 Kb), highly related elements (>90% identity) that are typically repeated a modest number of times and frequently found on the same chromosome arm. Many regions of genomic instability are known to be involved in genetic syndromes, termed "genomic disorders", where similar, but not identical, copy number changes produce specific developmental syndromes. It is remarkable that many LCR-rich intervals are located within chromosomal regions where rearrangements are known to be associated with neurobehavioral disorders, including autism (Christian et al., 2008; Marshall et al., 2008; Sebat et al., 2007 & Kirov et al., 2008), mental retardation (Sharp et al., 2006, 2008) and schizophrenia (Cantor et al.; Stefansson et al.; Stone et al. & Walsh et al., 2008). To determine

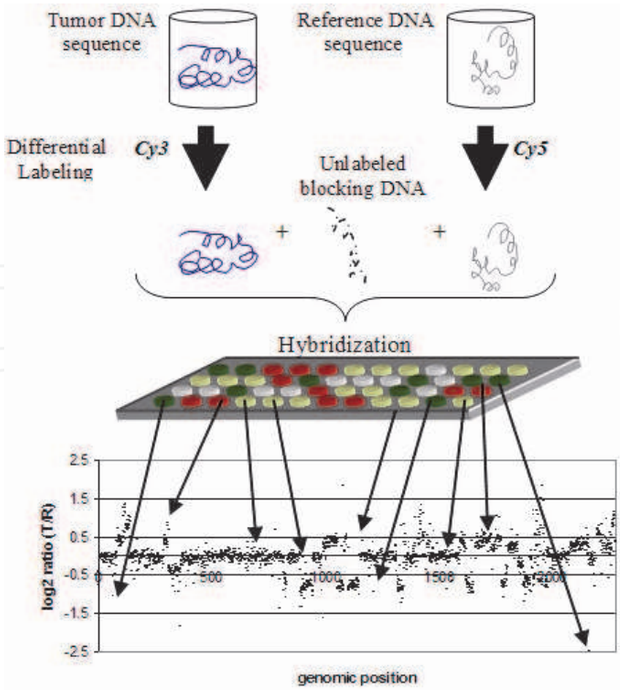


Fig. 1. Illustration of Microarray-based comparative genomic hybridization (array-CGH) process. The tumor and reference DNA are labeled and competitively hybridized to the array together with unlabeled blocking DNA to block repeated sequences. The ratio of the fluorescence intensity for each spot represented as a point in the relative copy number profile.

if copy number variants found within unstable segments of the genome are associated with autism susceptibility, we have conducted a high-resolution array CGH analysis of five genomic intervals that are rich in LCRs and where chromosomal rearrangements are associated with neurodevelopmental disorders. These regions include 7q11 (61-82Mb), 10q22.3-23.31 (77-92Mb), 15q11-13 (18-35 Mb), 17p11 (12-22Mb), and 22q11 (14-26Mb). The 7q11 interval spans the segment involved in Williams-Beuren Syndrome, a contiguous gene syndrome that produces a variety of cognitive and adaptive deficiencies (Greer et al., 1997). The reciprocal duplications of the Williams-Beuren deletion interval are associated with language delay and autism (Somerville et al., 2005; Van der Aa et al., 2009 & Depienne et al., 2007), suggesting that duplications in this genomic region are more closely linked to behavioral deficits that fall within the spectrum of autism disorder. Deletions flanked by segmental duplications are associated with language delay, attention deficit hyperactivity disorder (ADHD), and autism for the 10q22-23 interval (Balciuniene et al., 2007), and a balanced translocation affecting the KCNMA1 gene, which encodes a calcium-activated large conductance potassium channel, on 10q22 has also been reported in a child with autism (Laumonnier et al., 2006). Maternally-derived duplications of the 15q11-13 interval are the most common cytogenetic abnormalities associated with autism (Cook et al., 2001), and maternal as well as paternal-derived deletions are responsible for Angelman and Prader-Willi syndromes, respectively. In addition, deletions in the 15q11-13 interval are associated with mental retardation (Sharp et al., 2008), epilepsy (Sharp et al., 2008 & Helbig et al., 2009), and schizophrenia (Stefansson et al., & Stone et al., 2008). LCR-mediated

chromosomal rearrangements within 17p11 result in various nervous system dysfunctions (Lee et al., 2006), including Smith-Magenis and Potocki-Lupski syndromes. Deletions within the 22q11.2 interval are the most frequent interstitial deletions in humans, occurring in approximately 1 in 4000 live births (Papolos et al., 1996). These deletions cause congenital multisystem abnormalities referred to as 22q11 deletion syndrome, and include clinical entities such as Velocardiofacial syndrome, DiGeorge syndrome, and CATCH22 syndrome. Autism spectrum symptoms were reported in 20-50% of patients with 22q11 deletion syndrome; 15-20% of the patients have schizophrenia, and 40% of the patients manifest ADHD (Niklasson et al., 2001; Antonell et al., 2005 & Vorstman et al., 2006). Large deletions within the Velocardiofacial-DiGeorge syndromes critical region of 22q11 are found in patients with schizophrenia at a frequency of less than 1% (Stone et al., 2008). In the next section, we will present a novel methodology for the analysis of genetic data.

3. Method

In this section, we present a framework to evaluate the predictive power of recurrent variations at multiple genomic sites. The section is divided into two main parts. First, as a preprocessing step for feature extraction, a robust methodology based on statistical signal processing techniques is presented to clearly map and detect structural variations in the form of DNA copy number along the genome. Second, as a feature selection method prior to further analysis, a regional evaluation analysis is presented. It includes statistical learning procedures to measure the statistical and biological significance of the predicted variations. Then, classification techniques applied to segregate the tested samples into groups and to provide insight into the complex pattern of the predicted variations as well as discovering the relationship among them. There are three critical elements of our analysis that are novel: 1) we are detecting copy number changes as small as 1000 bp[1] (previous studies provided sensitivity typically hundreds of thousands of bp); this allows us to monitor genetic variants that might contribute incrementally to ASD susceptibility, 2) we are using oligo-arrays as a genotyping tool, performing a case-control association analysis, where copy number changes are the genetic variation being assessed, 3) we are developing algorithms to improve the sensitivity and specificity of array CGH data, assessing false positive and false negative rates.

3.1 Data preprocessing

Microarray data analysis is subject to multiple sources of variation, of which biological sources are of interest whereas most others are due to experimental sources. In other words, the goal of aCGH data analysis is to find the true boundaries of the variant regions (segments) which correspond to chromosomal variations and to remove other variations due to human factors, array printer performance, labeling, and hybridization efficiency (Kallioniemi et al., 1992). It consists of three key steps; 1) data preparation, 2) noise reduction, and 3) variation detection. In the data preparation step, copy number data is generated experimentally through aCGH process and then combined with their genomic positions. The next step is to reduce the experimental errors. This step is generally divided into two parts, data normalization, and data filtering. After normalizing the raw DCN data

[1] base-pairs

and before detecting the variant segments, the necessary step is to filter the normalized data for noise reduction.

3.1.1 Data modeling

According to the data description and properties generated from microarray technologies discussed in the previous section, we approximate a given DCN data sample as a one-dimensional piecewise discrete signal corrupted by additive white Gaussian noise with zero-mean and small variance. A good model for describing DNA copy number data is:

$$y[n] = f[n] + \varepsilon_n, \qquad n=1, 2, \ldots, N. \tag{1}$$

where $y[n]$ and $f[n]$ are the observed and true intensities of the DCN data probe at n^{th} location along the x-axis respectively. Here N is the length of DCN data and ε represents a vector of independent identically distributed (*i.i.d.*) random variables drawn from the Gaussian distribution of zero-mean and small variance (Wang et al., 2007).

3.1.2 Irregular probe position

Most prior works considered the DNA copy number profiles as discrete signals under the assumption that the probes are uniformly distributed along the chromosomes. This assumption may lead to wrong decisions with false positive or/and false negative points. More recent studies (Wang et al., 2007 & Willenbrock et al., 2005) show that considering the nonuniform spacing distance between the probes of the DCN data profiles could be beneficial for detecting and measuring the DNC variations.
Hence, we remodeled the DCN data discussed in the previous section as nonuniformly distributed discrete signals as follows:

$$y[x_n] = f[x_n] + \varepsilon_n, \qquad n=1, 2, \ldots, N. \tag{2}$$

where x_n in this case is the nonuniform distributed probe at n^{th} location along the x-axis. The x_n's are not uniformly distributed and the distance between two adjacent probes x_n and x_{n+1} may vary randomly. The $y[x_n]$ and $f[x_n]$ are the observed and true intensities of the DCN data probe location x_n respectively. The ε_n represent *i.i.d.* random variable from the Gaussian distribution with zero-mean and small variance σ^2.

3.2 Maximum likelihood estimator for genetic variation detection

Generally, Copy Number variations (CNVs) detection techniques fall into two categories: statistical based models and smoothing techniques. In the statistical based models, the noise free signal and noise models are required. Unfortunately, these models are usually unknown or impossible to describe adequately with simple random processes. As a result, the important details (i.e., breakpoints) of the CNVs regions will be included in the segmentation process. In addition, the techniques are computationally costly. Furthermore, most statistical models proposed to analyze array CGH data involve modeling the association between changes in neighboring probes. While this is helpful to find wide changes, it tends to ignore local changes. In the literature, there are various statistical approaches that have been proposed to detect changes in the DCN data.
On the other hand, the smoothing techniques provide alternative methods for processing the DCN data that are characterized by small and long intervals with sharp transitions and

singularities at boundaries edges (breakpoints). The techniques are particularly suitable for denoising DCN data as they do not require a parametric model in finding structures in the data. In these methods, local operators are applied to the noisy data. Only those points in a small local neighborhood are involved in the computation. The main advantage of these techniques is their computational efficiency. They can process the data in parallel without waiting for their neighboring points to be processed.

To this end, the proposed smoothing techniques provide efficient run-time speed and they are well suited to predict the variations in the discontinuous nature of such data. However, the smoothing techniques suffer from two main drawbacks. First, the breakpoints of the variation regions are involved in the smoothing process and these techniques exhibit artifacts in the neighborhood of these discontinuities that tend to blur the variation edges. Second, they did not consider the physical distances between the adjacent probes and simply assumed that they were uniformly spaced. This simplification will lead to suboptimal results. In this section, we propose a robust method based on maximum likelihood principle (Alqallaf et al., 2009) to clearly map and detect structural variation in the form of DNA copy number along the human genome. We apply dynamic programming to compute the DNA copy number estimates and reduce the computational complexity. Furthermore, we employ the minimum description length approach to estimate the number of unknown parameters. To evaluate our proposed method, we examine and compare the ability to reliably predict variations using molecular test, quantitative polymerase chain reaction. We take the comparison a step further by conducting two experiments designed specifically to assess the sensitivity and specificity of our proposed methods using high-density oligonucleotide array that have been examined by a number of different platforms and laboratories. Using well-characterized cell lines and custom tiled arrays, we show that the proposed method outperforms other popular commercial software and published algorithms in terms of detection performance and computational complexity.

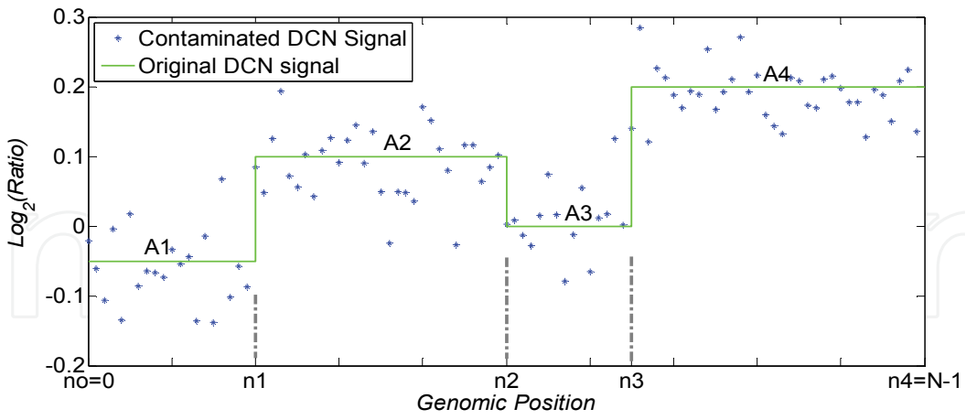


Fig. 2. Illustration of the observed DNA copy number data modeling of (3) with 4 segments.

As described in the previous section, the DNA copy number observations can be modeled as one-dimensional discrete time series with multilevel and jumps at unknown transition times, corrupted by additive white Gaussian noise (AWGN) of zero-mean and small variance σ^2. Figure 2 displays a graphical representation of the observed DNA copy number data modeling with 3 segments. Here we define $f[n]$ is the true piece-wise constant DCN signal to be estimated. Then, we define

$$f[n]=\sum_{i=1}^{M}A_i\left[u[n_{i-1}]-u[n_i]\right]$$
$$=\begin{cases}A_1 & n=0,\ 1,\ \cdots,\ n_1-1,\\ A_2 & n=n_1,\ n_1+1,\ \cdots,\ n_2-1,\\ \vdots & \\ A_M & n=n_{M-1},\ n_{M-1}+1,\ \cdots,\ N-1.\end{cases} \tag{3}$$

where n_0=0<n_1<n_2< … <n_{M-1}<n_M=N and $u[n]$ is the unit step function. Here A_i and n_i are the intensity level and the length of the i^{th} variant segment, respectively, with a total of M segments.

Based on the data assumption, we wish to design a detector to detect or equivalently estimate the uknown parameters. To do so, we first apply dynamic programming (DP) (Larson & Castie, 1982) to estimate the minimum number of the variant regions M using the minimum description principle (MDL) technique (Rissanen, 1978). Next, we apply the principle of maximum likelihood (ML) to estimate the values of breakpoints locations and intensity levels corresponding to these regions. Assuming that the number of variant regions M is known, then the i^{th} variant region can be characterized by the probability density function (PDF) $p_i([y[n_{i-1}]:y[n_i\text{-}1]];A_i)$, where A_i and n_i are the unknown parameters representing the intensity level and the breakpoint of the i^{th} variant segment, respectively. Moreover, each variant region is assumed to be statistically independent of all other regions. Hence, the PDF of the entire data record can be written as

$$p(\boldsymbol{y};\boldsymbol{A},\boldsymbol{n})=\prod_{i=1}^{M}p_i(y[n_{i-1}:n_i-1];A_i)=\frac{1}{(2\pi\sigma^2)^{N/2}}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{M}\left[\sum_{n=n_{i-1}}^{n_i-1}\left(y[n]-A_i\right)^2\right]\right]. \tag{4}$$

The DP algorithm can also be applied here to reduce the computational complexity to a more manageable level that is linearly proportional with the number of variant regions M.

3.3 Comparison study

In this section, we conduct two experiments to compare our proposed method with recent approaches to improve the sensitivity and specificity of array CGH data (assessing false positive and false negative rates).

3.3.1 Self-self hybridization experiment

In this experiment, we compare the performance of our proposed method MLE (Alqallaf et al., 2009) with Circular Binary Segmentation (CBS) algorithm (Venkatraman et al., 2007) and Copy Number Professional software package (BioDiscovery) Nexus algorithm by direct measurement of false positives. The same DNA sample is used as both the test and reference and hence any copy number variant assigned by an algorithm is incorrect and a false positive. In other words, we compare the DNA sample with itself in the aCGH process to generate the DCN data as described in section 2. In the ideal case, the intensity level, the difference between the tested sample and a known reference measured in $\log_2$ ratio, should equal to zero. However, due to the experimental noise, we expect to detect segments with relatively small intensity level value that are below cut-offs criteria. Otherwise, the detected segments would be considered as false positives. As shown in Table 1, the average number

of events detected by CBS algorithm is lower than the events detected by other algorithms. However, the average length of the events detected by our proposed algorithm MLE is relatively shorter than the average length of the events detected by CBS and Nexus.

Array#	Nexus		CBS		MLE	
	E	*L*	*E*	*L*	*E*	*L*
1	10	40336	2	790	6	8481
2	30	391030	4	625	19	27866
3	100	2894494	7	4130039	55	42440
Avg.	47	1108620	5	1377151	27	26262

Table 1. Comparison of the proposed algorithms using number of detected events, *E*, and their length, *L*, in base-pair for the three tested array samples.

3.3.2 Duplicated dye-swap experiments for two HapMap samples

Here we take the comparison a step further by conducting experiments designed specifically to assess our proposed algorithm, MLE, using high-density oligonucleotide array CGH. In this experiment, replicate dye-swap experiments were conducted comparing DNA samples from two hapmap (Redon et al., 2006) subjects that have been examined by a number of different platforms and laboratories, NA15510 and NA10851, for a total of four arrays. The relative intensities differences are measured and reported. It should be noted that the directionality of any detected variant is expected to be opposite when the dyes are swapped. That is, deletions with the first array will appear to be duplications with the second array. This is due to the convention of reporting the $\log_2$ ratios as described in section 2. This experiment allows us to assess the sensitivity of the proposed algorithms. Table 2 shows that the number of CNVs detected by the MLE is considerably higher than those detected using CBS (a range of 4.5% to 36% more for the 4 different array experiments). Our results show that applying the averaging window of 2Kb allow the algorithms to be well suited for detecting variations in high-density oligonucleotide array aCGH.

Array #	CBS	MLE
1	14	20
2	10	20
3	20	21
4	13	21

Table 2. List of the number of events (CNVs) detected by CBS and MLE algorithms.

4. Statistical significance

After filtering multiple DCN datasets of normal control and test samples, we need to apply a statistical analysis to reveal the randomness and to classify the genes or genomic locations that are involved or play roles in the targeted disease, ASD. In this section, we present two statistical approaches to measure the significance of common CNVs across the samples and especially in the complex LCRs regions. First, we measure the relative frequency at each genomic position within the LCRs regions. Second, based on the relative frequency, a

regional evaluation scheme is used to measure the significance of the overlapping recurrent CNVs and to classify the tested DCN samples.

4.1 Statistical-based model

In summary, most of the proposed algorithms in the literature did not consider the statistical and biological significance of the analysis of multiple DCN data samples. In particular, they did not address the task of identifying common variations that overlap a set or subset of the study samples to reveal the randomness of the predicted CNVs. Indeed, few studies have addressed class discovery across multiple samples of DCN data (Grant et al., 1999 & Diskin et al., 2006). However, they did not consider denoising the data prior to applying the statistical analysis. Although these are effective methods for searching statistically for common variations across multiple samples, it suffers from two main issues which can be summarized as follows: First, it does not take into considerations that different variation types (gain and loss) may occur within the same genomic locations. They simply discard these locations and indicate them as missing values. This will lead to decreases in the data resolution. Second, it does not differentiate between the intensity levels. This is an important issue for characterizing the variations in the complex areas of low copy repeats (LCR). For this, we propose in our statistical analysis to identify nonrandom gains and losses across multiple samples with the consideration of these issues.

To reveal the randomness and identify the genes or genomic locations that are involved or play roles in the targeted disease, we apply a statistical analysis to measure the significance of recurrent CNVs including those in the complex regions of LCRs. Here, we plot the frequency of the occurrence of the predicted CNVs (deletions and/or duplications) that are overlapped across multiple case samples with respect to control samples. Suppose that a set of M filtered DCN samples each with N probes, then the normalized frequency at the n^{th} position can be measured as

$$G[n] = \frac{\sum_{s \in M} v_{s,n}}{M}. \qquad n = 1,\ 2,\ ...,\ N \tag{5}$$

where s represents the sample of the same variation type and $v_{s,n}$ is a binary number which equals to 1 if the variation is present and 0 otherwise. Figure 3 shows the differences in the frequency of occurrence of the gains and losses between 71 normal control and 71 autistic samples of chromosome 7. The differences suggest further analysis to discover the relationship between the predicted CNVs and to classify the tested samples.

4.2 Putative recurrent CNVs classification

Although the predicted variant segments of each aCGH profile have their own importance, finding recurrent copy number variations that overlap and share the same type adds another dimension to link them with the targeted disease. The size of our aCGH profiles is relatively large and many of the variants regions of the same type (deletions/duplications) are found in both cases and controls. We therefore include a filtering step by removing these CNVs to make it easier to find the interesting variations and reduce the number of data points to some subset of concatenated CNVs.

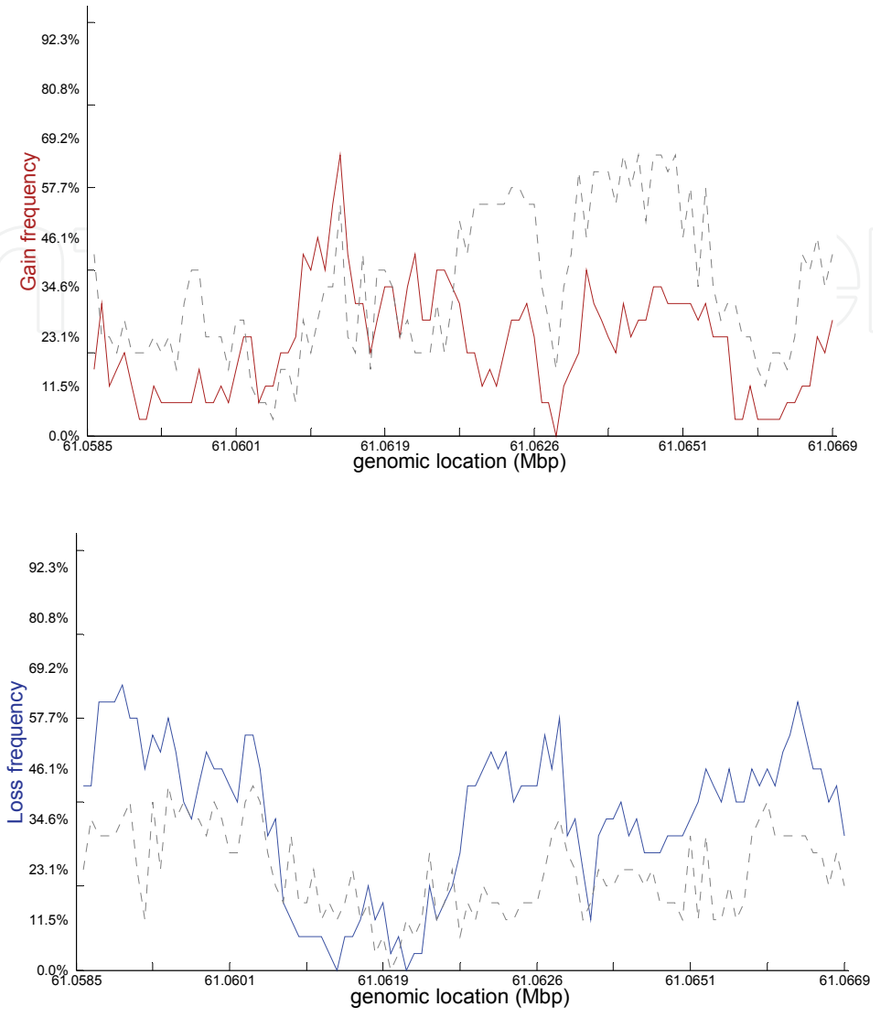


Fig. 3. Frequency plots of the filtered samples. A) The solid red and the dash gray lines represent the gain frequency of the typically developed (TD) and autistic (AU) samples, respectively. B) The solid blue and the dash gray lines represent the loss frequency of the TD and AU samples, respectively.

Before we make a decision on the predicted segments, in this section, we extend our method by imposing cutoff criteria based on regional genetic information as an optimal feature selection. The reasons for performing this procedure are as follows. First, we seek the genetic structure and thus the genetic mechanisms responsible for the progression of the disease. Second, we would like to remove or eliminate the irrelevant features (e.g., CNVs) from the classification and hence, to increase the run-time speed and to improve the accuracy of the classification. After ranking the CNVs, a suitable set is identified and declared as an optimal feature set to be used for classification analysis. Although the feature selection step is a major step attempting to discover and reveal genetic mechanisms, it can not be claimed to discover the true biological relationship without further experimental evaluation. The extension accounts for the minimal number of probes within in each

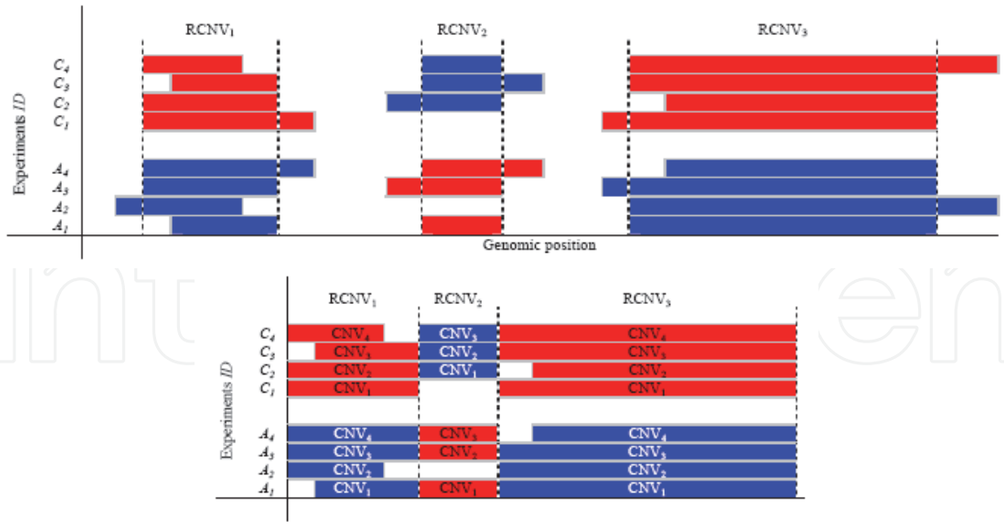


Fig. 4. Schematic representation of 3 recurrent copy number variations (RCNVs) with different lengths (Top) and concatenation vectors of the predicted combinations of RCNVs (Bottom). The *x*-axis represents the genomic position and the *y*-axis represents the experiments indices, C_i for normal control samples and A_i for autistic samples, respectively. The vertical dashed lines represent the RCNVs boundaries. The red and blue bars represent duplication and deletion for the corresponding positions.

segment, the intensity level represented by the $\log_2$ ratio value, and the repeat content of the region where the CNV is located.

Each segment that met biological and statistical cut-off criteria is considered a CNV and assembled into a segmentation table for further biological analysis. Figure 4 is an illustration of three RCVNs with different sizes of filtered DCN data for multiple samples of normal control, C_i's, and autistic, A_i's, individuals, respectively.

With this setup, we apply the traditional clustering algorithms (Fuzzy c-means and k-means) to the concatenation vectors of the predicted combinations of RCNVs to classify the DCN data samples and to provide insight into the pattern of the variations using the concatenated recurrent CNVs that are statistically significant.

In the next section, we will investigate the classification performance using the predicted combinations of multiple RCNV sites of different chromosomes produced by the regional evaluation method presented in this section that may have direct role in the targeted disease, ASD.

5. Visualization and pattern recognition

To visualize the microarray data, we apply agglomerative hierarchical clustering algorithm to decide the level or scale of clustering that is most appropriate for our clustering analysis. It provides a graphical representation of the samples to explore the number of ways to look for relationships between the samples and to provide insight into the pattern of the recurrent CNVs. The algorithm groups the data samples based on the defined measure of the distances between the samples elements using similarities functions to create the clusters. It starts from each single sample as a cluster and it merge the samples into clusters

(groups or subgroups) based on the updated similarity measures (linkage), where clusters at one level are joined as clusters at the next level. The definition of the similarity measures depends on the clustering algorithm and the biological meaning of similarity. For example, a correlation distance, $d_p(\boldsymbol{x},\boldsymbol{y})$, based on Pearson's correlation (6) may bring together samples whose probes intensity levels are different, but have a similar behavior, and which would be considered different by the Euclidean distance $d_e(\boldsymbol{x},\boldsymbol{y})$ (7) which is suitable for discovering the common CNVs. Specifically,

$$d_p(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{x})^2\sum_{i=1}^{N}(y_i-\bar{y})^2}}, \tag{6}$$

$$d_e(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{N}(x_i-y_i)^2}, \tag{7}$$

where $\bar{x}$ and $\bar{y}$ are the sample mean values of the two data vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ with N data points, respectively.

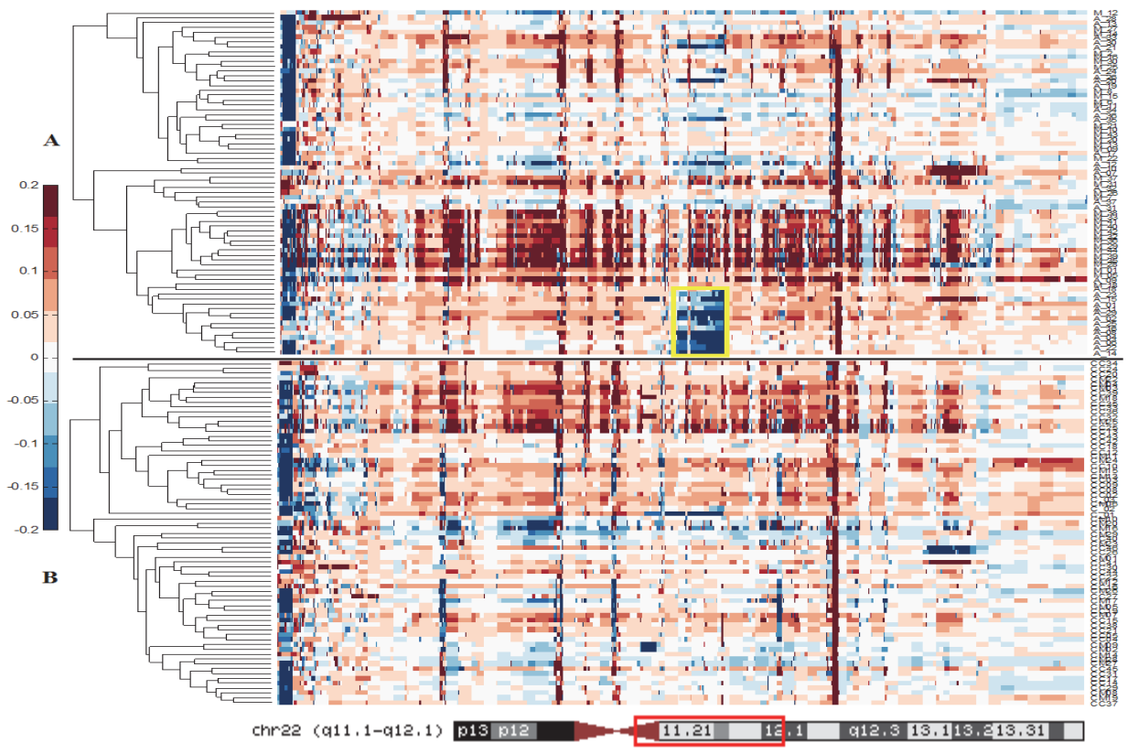


Fig. 5. Hierarchical clustering of chromosome 22 using 142 samples. A) 71 autistic (AU) samples. B) 71 typically developed (TD) samples. Dark red represents duplications and dark blue represents deletions. The solid black line used to separate the AU and TD samples. Yellow square represents deletion region within the AU subgroup.

To explore the dataset before imposing the cut-off criteria, we perform unsupervised hierarchical clustering with the Euclidean distance as a distance metric to calculate the pair-wise distances between the tested samples and centroid linkage method to create linkage between the clusters tree. The heat map is used to represent thousands of $\log_2$ ratios (intensity values) of the probes of each sample it uses two-color of a matrix of colored cells, red for duplication, where the $\log_2$ ratio is positive, and blue for deletion, where the $\log_2$ ratio is negative. The rows represent the tested samples and the columns represent the probes positions, and the brightness of the cells is proportional to their intensity levels. For this analysis, our case-control population of study consists of 71 individuals with a diagnosis of autism compared to 71 typically developing controls matched for gender and ethnicity. Figure 5 shows an example of one of the five chromosomal regions used in this study. By simple comparison between the recurrent CNVs detected in the entire or subset of the autistic samples (Figure 5. A) and those detected in the typically developed samples (Figure 5. B), we can detect patterns of variations that are exclusively or selectively represented in one or the other group (see for example the deletions noted with yellow box). The yellow square show long deletion region within the AU subgroup compared to the other members in the AU individuals and TD controls.

6. Conclusion

In this chapter, we presented an overview for the analysis of genetic variations in the form of DNA copy number changes and their association with the targeted disease, autism spectrum disorder. Our study shows that our proposed algorithm, MLE, is computationally efficient and it can achieve even better detection capabilities by considering the effect of the nonuniform genomic spacing distance between the biomarkers. Moreover, to enhance our algorithm's ability to map and identify regions of variation across multiple samples, we preformed statistical analysis on the filtered samples searching for common variations. The potential impact of the statistical analysis is to provide insight into the patterns of the variations by characterizing and classifying the samples that are involved in the targeted diseases. Indeed, the high frequency of variants (duplications and/or deletions) detected in these regions across the samples allowed the assembly of a copy number map of both typically developed and diseased individuals. The mapping approach reveals patterns of copy number change along these chromosomal intervals that are not currently represented in the assembly of genomic variants compiled from relatively low-resolution genome-wide platforms. Our findings indicate that Low copy repeat-rich intervals, known to be relatively susceptible to copy number changes and sequence rearrangement, show a greater degree of copy number alteration in diseased compared to typically developed individuals. A larger contribution of variations detected (duplications and/or deletions) in the total copy number burden differences have been reported to be associated with different genetic diseases. Our findings also show ethnicity is an important consideration that should be integrated into case-control study design. The findings suggest that autism is associated with an increased amount of copy number alteration in unstable segments of the genome. The experimental results also show that using high-resolution custom-tiled oligonucleotide array comparative genomic hybridization samples, improve the accuracy of the proposed methods to detect the true amount of structural variations of the human genome including previously reported variations with known biological and clinical relevance and new variations that warrant further investigated. To explore the idea that patterns of relatively common copy number

variations can increase the power of discrimination between autistic and typically developing patients, a set of recurrent variants that are statistically differed between the two groups is identified and presented. The findings suggest that combinations of copy number variations could provide the basis for discriminating autistic and typically developing groups and potentially identifying distinct subgroups within the phenotypic heterogeneity of autism spectrum disorder. Finally, the analysis presented in this chapter is broadly applicable to case-control studies of genetic diseases beyond the targeted disease, autism.

7. References

Freeman, JL.; Perry, GH.; Feuk, L.; Redon, R.; McCarroll, SA.; Altshuler, DM.; Aburatani, H.; Jones, KW.; Tyler-Smith, C. & Hurles, ME. (2006). Copy number variation: new insights in genome diversity. *Genome Research*, Vol.16, No.8, pp. 949-961.

Lee, JA. & Lupski, JR. (2006). Genomic rearrangements and gene copy-number alterations as a cause of nervous system disorders. *Neuron*, Vol.52, No.1, pp. 103-121.

Stankiewicz, P. & Lupski, JR. (2002). Molecular-evolutionary mechanisms for genomic disorders. *Current Opinion in Genetics & Development on ScienceDirect*, Vol.12, No.3, pp. 312-319.

Christian, SL.; Brune, CW.; Sudi, J.; Kumar, RA.; Liu, S.; Karamohamed, S.; Badner, JA.; Matsui, S.; Conroy, J. & McQuaid, D. (2008). Novel submicroscopic chromosomal abnormalities detected in autism spectrum disorder. *Biological Psychiatry*, Vol.63, No.12, pp. 1111-1117.

Marshall, CR.; Noor, A.; Vincent, JB.; Lionel, AC.; Feuk, L.; Skaug, J.; Shago, M.; Moessner, R.; Pinto, D. & Ren, Y. (2008). Structural variation of chromosomes in autism spectrum disorder. *American Journal of Human Genetics*, Vol.82, No.2, pp. 477-488.

Sebat, J.; Lakshmi, B.; Malhotra, D.; Troge, J.; Lese-Martin, C.; Walsh, T.; Yamrom, B.; Yoon, S.; Krasnitz, A. & Kendall, J.; (2007) Strong association of de novo copy number mutations with autism. *Science*, Vol.316, No.5823, pp. 445-449.

Kirov, G.; Gumus, D.; Chen, W.; Norton, N.; Georgieva, L.; Sari, M.; O'Donovan, MC.; Erdogan, F.; Owen, MJ. & Ropers, HH. (2008). Comparative genome hybridization suggests a role for NRXN1 and APBA2 in schizophrenia. *Human Molecular Genetics*, Vol.17, No.3 pp. 458-465.

Sharp, AJ.; Hansen, S.; Selzer, RR.; Cheng, Z.; Regan, R.; Hurst, JA.; Stewart, H.; Price, SM.; Blair, E. & Hennekam, RC. (2006). Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nature Genetics*, Vol.38, No.9, pp. 1038-1042.

Sharp, AJ.; Mefford, HC.; Li, K.; Baker, C.; Skinner, C.; Stevenson, RE.; Schroer, RJ.; Novara, F.; DeGregori, M. & Ciccone, R. (2008). A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. Nature Genetics, Vol.40, No.3, pp. 322-328.

Cantor, RM. & Geschwind, DH. (2008). Schizophrenia: genome, interrupted. *Neuron*, Vol.58, No.2, pp. 165-167.

Stefansson, H.; Rujescu, D.; Cichon, S.; Pietilainen, OP.; Ingason, A.; Steinberg, S.; Fossdal, R.; Sigurdsson, E.; Sigmundsson, T. & Buizer-Voskamp, JE. (2008). Large recurrent microdeletions associated with schizophrenia. Nature, Vol.455, No.7210, pp. 232-236.

Stone, JL.; O'Donovan, MC.; Gurling, H.; Kirov, GK.; Blackwood, DH.; Corvin, A.; Craddock, NJ.; Gill, M.; Hultman, CM. & Lichtenstein, P. (2008). Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature, Vol.455, No.7210, pp. 237-241.

Walsh, T.; McClellan, JM.; McCarthy, SE.; Addington, AM.; Pierce, SB.; Cooper, GM.; Nord, AS.; Kusenda, M.; Malhotra, D. & Bhandari, A. (2008). Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science, Vol.320, No.5875, pp. 539-543.

Estivill, X. & Armengol, L. (2007). Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. *PLoS Genetics*, Vol.3, No.10, pp. 1787-1799.

Greer, MK.; Brown, FR.; Pai, GS.; Choudry, SH. & Klein, AJ. (1997). Cognitive, adaptive, and behavioral characteristics of Williams syndrome. *American Journal of Human Genetics*, Vol.74, No.5, pp. 521-525.

Somerville, MJ.; Mervis, CB.; Young, EJ.; Seo, EJ.; del Campo, M.; Bamforth, S.; Peregrine, E.; Loo, W.; Lilley, M. & Perez-Jurado, LA. (2005). Severe expressive-language delay related to duplication of the Williams-Beuren locus. *The New England Journal of Medicine*, Vol.353, No.16, pp. 1694-1701.

Van der Aa, N.; Rooms, L.; Vandeweyer, G.; van den Ende, J.; Reyniers, E.; Fichera, M.; Romano, C.; Delle Chiaie, B.; Mortier, G. & Menten, B. (2009). Fourteen new cases contribute to the characterization of the 7q11.23 microduplication syndrome. *European Journal of Medical Genetics*, Vol.52, No.2-3, pp. 94-100.

Depienne, C.; Heron, D.; Betancur, C.; Benyahia, B.; Trouillard, O.; Bouteiller, D.; Verloes, A.; LeGuern, E.; Leboyer, M. & Brice, A. (2007). Autism, language delay and mental retardation in a patient with 7q11 duplication. *Journal of Medical Genetics*, Vol.44, No.7, pp. 452-458.

Balciuniene, J.; Feng, N.; Iyadurai, K.; Hirsch, B.; Charnas, L.; Bill, BR.; Easterday, MC.; Staaf, J.; Oseth, L. & Czapansky-Beilman, D. (2007). Recurrent 10q22-q23 deletions: a genomic disorder on 10q associated with cognitive and behavioral abnormalities. *American Journal of Human Genetics*, Vol.80, No.5, pp. 938-947.

Laumonnier, F.; Roger, S.; Guerin, P.; Molinari, F.; M'Rad, R.; Cahard, D.; Belhadj, A.; Halayem, M.; Persico, AM. & Elia, M. (2006). Association of a functional deficit of the BKCa channel, a synaptic regulator of neuronal excitability, with autism and mental retardation. *American Journal of Psychiatry*, Vol.163, No.9, pp. 1622-1629.

Cook, EH. Jr. (2001). Genetics of autism. *Child and Adolescent Psychiatric Clinics of North America*, Vol.10, No.2, pp. 333-350.

Helbig, I.; Mefford, HC.; Sharp, AJ.; Guipponi, M.; Fichera, M.; Franke, A.; Muhle, H.; de Kovel, C.; Baker, C. & von Spiczak, S. (2009). 15q13.3 microdeletions increase risk of idiopathic generalized epilepsy. *Nature Genetics*, Vol.41, No.2, pp. 160-162.

Papolos, DF.; Faedda, GL.; Veit, S.; Goldberg, R.; Morrow, B.; Kucherlapati, R. & Shprintzen, RJ. (1996). Bipolar spectrum disorders in patients diagnosed with velo-cardio-facial syndrome: does a hemizygous deletion of chromosome 22q11 result in bipolar affective disorder? *American Journal of Psychiatry*, Vol.153, No.12, pp. 1541-1547.

Niklasson, L.; Rasmussen, P.; Oskarsdottir, S. & Gillberg, C. (2001) Neuropsychiatric disorders in the 22q11 deletion syndrome. *Genetics in Medicine*, Vol.3, No.1, pp. 79-84.

Antonell, A.; de Luis, O.; Domingo-Roura, X. & Perez-Jurado, LA. (2005). Evolutionary mechanisms shaping the genomic structure of the Williams-Beuren syndrome chromosomal region at human 7q11.23. *Genome Research*, Vol.15, No.9, pp. 1179-1188.

Vorstman, JA.; Morcus, ME.; Duijff, SN.; Klaassen, PW.; Heineman-de Boer, JA.; Beemer, FA.; Swaab, H.; Kahn RS. & van Engeland, H. (2006). The 22q11.2 deletion in children: high rate of autistic disorders and early onset of psychotic symptoms. *Journal of the American Academy of Child & Adolescent Psychiatry*, Vol.45, No.9, pp. 1104-1113.

Hertz-Picciotto, I.; Croen, LA.; Hansen, R.; Jones, CR.; van de Water, J. & Pessah, IN. (2006). The CHARGE study: an epidemiologic investigation of genetic and environmental factors contributing to autism. *Environmental Health Perspectives*, Vol.114, No.7, pp. 1119-1125.

Redon, R.; Ishikawa, S.; Fitch.; KR.; Feuk, L.; Perry, GH.; Andrews, TD.; Fiegler, H.; Shapero, MH.; Carson, AR. & Chen, W. (2006). Global variation in copy number in the human genome. *Nature*, Vol.444, No.7118, pp. 444-454.

Shen F, Huang J, Fitch KR, Truong VB, Kirby A, Chen W, Zhang J, Liu G, McCarroll SA, Jones KW (2008). Improved detection of global copy number variation using high density, non-polymorphic oligonucleotide probes. *BMC Genetics*, Vol.9 No.27.

Conrad, DF.; Pinto, D.; Redon, R.; Feuk, L.; Gokcumen, O.; Zhang, Y.; Aerts, J.; Andrews, TD.; Barnes, C. & Campbell, P. (2010). Origins and functional impact of copy number variation in the human genome. *Nature*, Vol.464, No.7289, pp. 704-712.

Perry, GH.; Ben-Dor, A.; Tsalenko, A.; Sampas, N.; Rodriguez-Revenga, L.; Tran, CW.; Scheffer, A.; Steinfeld, I.; Tsang, P. & Yamada, NA. (2008). The fine-scale and complex architecture of human copy-number variation. *American Journal of Human Genetics*, Vol.82, No.3, pp 685-695.

Berg, JS.; Brunetti-Pierri, N.; Peters, SU.; Kang, SH.; Fong, CT.; Salamone, J.; Freedenberg, D.; Hannig, VL.; Prock, LA. & Miller, DT. (2007). Speech delay and autism spectrum behaviors are frequently associated with duplication of the 7q11.23 Williams-Beuren syndrome region. *Genetics in Medicine*, Vol.9, No.7, pp. 427-441.

Potocki, L.; Bi, W.; Treadwell-Deering, D.; Carvalho, CM.; Eifert, A.; Friedman, EM.; Glaze, D.; Krull, K.; Lee, JA. & Lewis, RA.; (2007). Characterization of Potocki-Lupski syndrome (dup(17)(p11.2p11.2)) and delineation of a dosage-sensitive critical interval that can convey an autism phenotype. *American Journal of Human Genetics*, Vol.80, No.4, pp. 633-649.

Carvalho, CM. & Lupski, JR. (2008). Copy number variation at the breakpoint region of isochromosome 17q. *Genome Research*, Vol.18, No.11, pp. 1724-1732.

Mukaddes, NM. & Herguner, S. (2007). Autistic disorder and 22q11.2 duplication. *World Journal of Biological Psychiatry*, Vol.8, No.2, pp. 127-130.

Barabash, A.; Marcos, A.; Ancin, I.; Vazquez-Alvarez, B.; de Ugarte, C.; Gil, P.; Fernandez, C.; Encinas, M.; Lopez-Ibor, JJ. & Cabranes, JA. (2009) APOE, ACT and CHRNA7

genes in the conversion from amnestic mild cognitive impairment to Alzheimer's disease. *Neurobiol Aging*, Vol.30, No.8, pp. 1254-1264.

Miller, DT.; Shen, Y.; Weiss, LA.; Korn, J.; Anselm, I.; Bridgemohan, C.; Cox, GF.; Dickinson, H.; Gentile, J. & Harris, DJ. (2009). Microdeletion/duplication at 15q13.2q13.3 among individuals with features of autism and other neuropsychiatric disorders. *Journal of Medical Genetics*, Vol.46, No.4 pp. 242-248.

Shinawi, M.; Schaaf, CP.; Bhatt, SS.; Xia, Z.; Patel, A.; Cheung, SW.; Lanpher, B.; Nagl, S.; Herding, HS. & Nevinny-Stickel, C. (2009). A small recurrent deletion within 15q13.3 is associated with a range of neurodevelopmental phenotypes. *Nature Genetics*, Vol.41, No.12, pp. 1269-1271.

Tsuchiya, KD.; Wiesner, G.; Cassidy, SB.; Limwongse, C.; Boyle, JT. & Schwartz, S. (1998). Deletion 10q23.2-q23.33 in a patient with gastrointestinal juvenile polyposis and other features of a Cowden-like syndrome. *Genes, Chromosomes & Cancer*, Vol.21, No.2, pp. 113-118.

Zhou, XP.; Woodford-Richens, K.; Lehtonen, R.; Kurose, K.; Aldred, M.; Hampel, H.; Launonen, V.; Virta, S.; Pilarski, R. & Salovaara, R. (2001). Germline mutations in BMPR1A/ALK3 cause a subset of cases of juvenile polyposis syndrome and of Cowden and Bannayan-Riley-Ruvalcaba syndromes. *American Journal of Human Genetics*, Vol.69, No.4, pp. 704-711.

Fombonne, E. (2003). Epidemiological Surveys of Autism and Other Pervasive Developmental Disorders: An Update. *Journal of Autism and Developmental Disorders*, Vol.33, No.4, pp. 365–382.

Kuehn, B. (2007). CDC: Autism Spectrum Disorders Common. *Journal of the American Medical Association*, Vol.297, No.940.

Abrahams, B. & Geschwind, D. (2008). Advances in autism genetics: on the threshold of a new neurobiology. *Nature Reviews Genetics*. Vol.9, No.5, pp. 341–355.

Alqallaf, A.; Tewfik, A. & Selleck, S. (2009). Genetic variation detection using maximum likelihood estimator. *IEEE International Workshop on Genomic Signal Processing and Statistics*. ISBN: 9781424447619, Minnesota, USA.

Kallioniemi, A.; Kallioniemi, O.; Sudar, D.; Rutovitz, D.; Gray, J.; Waldman, F. & Pinkel, D. (1992). Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science, Vol.258, No.5083, pp. 818-821.

Wang, Y. & Wang, S. (2007). A novel stationary wavelet denoising algorithm for array-based DNA copy number data. International Journal of Bioinformatics Research and Applications, Vol.3, No.2, pp. 206–222.

Willenbrock, H. & Fridlyand, J. (2005). A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics*, Vol.21, No.22, pp. 4084–4091.

Rissanen J. (1978). Modeling by Shortest Data Description. *Automatica*, Vol.14, pp. 465-471.

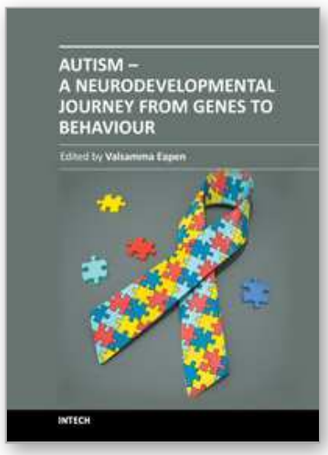
Larson, R. & Castie, J. (1982). Principles of Dynamic Programming. Vol.1–2, *Marcel Dekker* Inc., NY.

Nexus: Copy Number Professional software package (BioDiscovery), Inc., El Segundo, CA) BioDiscovery Inc. available from: http://www.biodiscovery.com.

Venkatraman, E. & Olshen, A. (2007). A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics, Vol.23, No.6 pp. 657-663.

Grant, G.; Manduchi, E.; Cheung, V. & Ewens, W. (1999). Significance testing for direct identity-by-descent mapping. Annals of Human Genetics, Vol.63, No.5, pp. 441–454.

Diskin, S.; Eck, T.; Greshock, J.; Mosse, Y.; Naylor, T.; Stoeckert, C.; Weber, B.;Maris, J. & Grant, G. (2006). STAC: A method for testing the significance of DNA copy-number aberrations across multiple array-CGH experiments. Genome Research, Vol.16, No.9, pp. 1149–1158.



Autism - A Neurodevelopmental Journey from Genes to Behaviour
Edited by Dr. Valsamma Eapen

ISBN 978-953-307-493-1
Hard cover, 484 pages
Publisher InTech
Published online 17, August, 2011
Published in print edition August, 2011

The book covers some of the key research developments in autism and brings together the current state of evidence on the neurobiologic understanding of this intriguing disorder. The pathogenetic mechanisms are explored by contributors from diverse perspectives including genetics, neuroimaging, neuroanatomy, neurophysiology, neurochemistry, neuroimmunology, neuroendocrinology, functional organization of the brain and clinical applications from the role of diet to vaccines. It is hoped that understanding these interconnected neurobiological systems, the programming of which is genetically modulated during neurodevelopment and mediated through a range of neuropeptides and interacting neurotransmitter systems, would no doubt assist in developing interventions that accommodate the way the brains of individuals with autism function. In keeping with the multimodal and diverse origins of the disorder, a wide range of topics is covered and these include genetic underpinnings and environmental modulation leading to epigenetic changes in the aetiology; neural substrates, potential biomarkers and endophenotypes that underlie clinical characteristics; as well as neurochemical pathways and pathophysiological mechanisms that pave the way for therapeutic interventions.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:

Abdullah Alqallaf and Ali Hajjiah (2011). Identifying Variations Within Unstable Regions of the Genome Reveal Autism Associated Patterns, Autism - A Neurodevelopmental Journey from Genes to Behaviour, Dr. Valsamma Eapen (Ed.), ISBN: 978-953-307-493-1, InTech, Available from: http://www.intechopen.com/books/autism-a-neurodevelopmental-journey-from-genes-to-behaviour/identifying-variations-within-unstable-regions-of-the-genome-reveal-autism-associated-patterns

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen