

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

6,900

Open access books available

186,000

International authors and editors

200M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



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](mailto:book.department@intechopen.com)

Numbers displayed above are based on latest data collected.  
For more information visit [www.intechopen.com](http://www.intechopen.com)



---

## Discovering the Genetics of Autism

---

Abdullah K. Alqallaf, Fuad M. Alkoot and  
Mash'el S. Aldabbous

Additional information is available at the end of the chapter

<http://dx.doi.org/10.5772/53797>

---

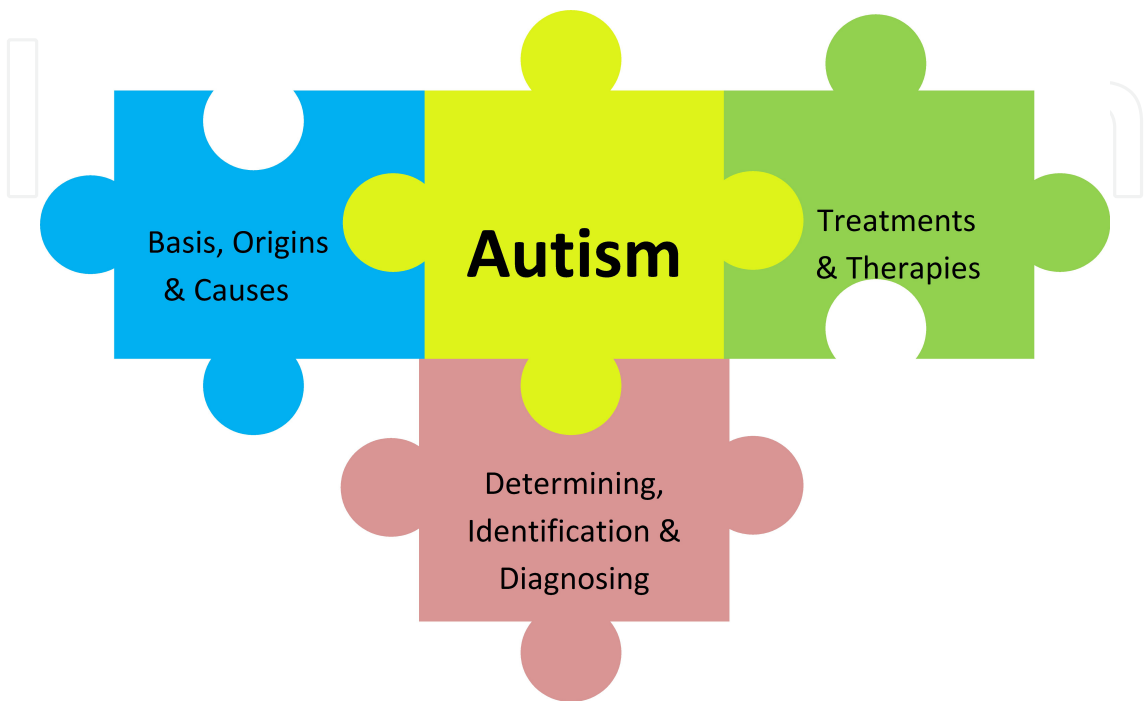
### 1. Introduction

Autism is a complex neurodevelopmental disorder. It is characterized by social isolation, language deficits and repetitive or stereotyped behaviors. Autism spectrum disorder (ASD) has received a great deal of attention in the recent years not only due to the increasing rate of affected children but also because of the social and economical impact of the disorder on their families. Various studies and researches have been proposed to deal with and tackle the ASD. They can be divided into three categories as follows.

1. The basis and causes of the disorder. Different hypotheses have been proposed in an attempt to determine and discover the originality of autism. Genetic risk factors represented by abnormal chromosomal variations and rearrangements, and non-genetic factors represented by environmental agents that have been claimed to contribute to ASD, such as exposure of children to vaccines, infection, certain foods or heavy metals.
2. The methodologies and techniques for characterizing and diagnosing the disorder. Several instrumental diagnostic protocols are commonly used in autism research such as the Autism Diagnostic Interview-Revised (ADI-R) and the Autism Diagnostic Observation Schedule (ADOS). The advances in neuro-imaging techniques such as the functional-Magnetic Resonance Imaging (f-MRI) have allowed scientists to model the structural and functional differences in the human brain tissues of the individuals with ASD. The clinical genetics evaluation provide reliable alternative to the interview-based protocols and screening approaches. It allows geneticists to link an estimate of approximately 40% of the cases to genetic contributors.
3. The treatments and therapies of autistic patients. The available approaches for treatments include applied behavior analysis (ABA), developmental models, structured

teaching, speech and language therapy and social skills therapy. When behavioral treatment fails, many medications are used to treat ASD symptoms.

Figure 1 demonstrates the interaction of the autism spectrum disorders researches and studies.



**Figure 1.** A puzzle-like representation of the interaction process of the researches and studies for autism spectrum disorders.

The advancements of the technologies in the field of genetics provide the opportunities for researchers and scientists to explore in depth the biological information and to convert it into meaningful biological knowledge through computational-based models.

In this chapter, we will investigate the genetics origins of autism and demonstrate the latest techniques and technologies available for diagnosing the complex disorder. We will also propose a robust approach for detecting and identifying the targeted disorder based upon the advantages and strengths of the publically available and commercial approaches while avoiding their weaknesses. The proposed approach is divided into two steps. The preprocessing step is a feature-extraction method used to clearly map and detect the genetic variations and structural rearrangements followed by a statistical-based model as feature-selection to evaluate and measure the statistical and biological significance of the predicted variations. The classification step is to discover the relationship among the tested samples into groups and/or subgroups, and to provide insight into the complex pattern of the genome.

The results suggest that autism is associated with an increased amount of alterations in unstable segments of the genome. The experimental results also show that using high-resolu-

tion custom-tiled samples improve the accuracy of our proposed approach in determining previously reported and new genetic contributors that warrant investigation.

This chapter aims at utilizing research to bring benefits to individuals and families affected by autism spectrum disorders and to improve the quality of their life. And this can be done by clear mapping and identifying the biomarkers associated with ASD at the early childhood stages which are essential to provide better treatments and therapies. Finally, the proposed approach presented in this chapter is broadly applicable to case-control studies of genetic diseases beyond the ASD.

The chapter is organized as follows. In section 2, we demonstrate the genetic data generating techniques, data modeling and chromosomal variations that are associated with the targeted disorder, ASD. Section 3 is devoted for the methods used to analyze the genetic data trying to discover the variant regions along the genome and to identify the tested samples. In section 4, we apply molecular test to evaluate the predictive power of the proposed approach. Finally, discussion and conclusion based on the results are presented in section 5.

2. Genetic data

2.1. Genomic structural variations and ASD susceptibility

Genetic alterations in the form of chromosomal rearrangements are genomic structural variations that lead to changes in the DNA copy number such as duplications and deletions of the DNA copies. However, copy number changes do not include other genomic structural variations such as inversions, insertions and reciprocal translocations. Figure 2 demonstrates different types of chromosomal rearrangements.

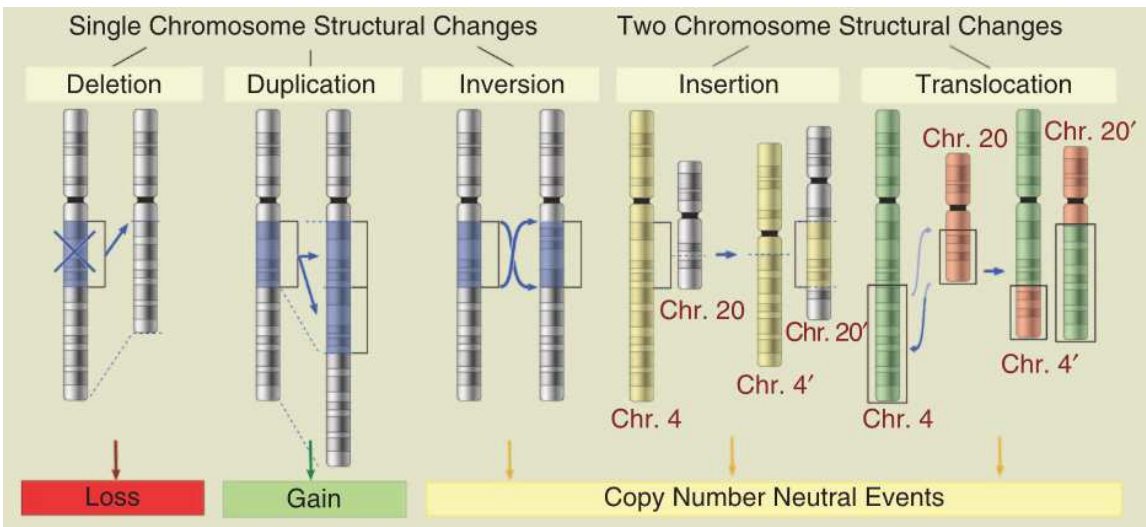


Figure 2. a schematic representation of types of chromosomal rearrangements [67].

	Chromosome region	Gene	Phenotype
Mendelian Syndromes	6q23.3	AHI1	Joubert syndrome
	7q35-q36.1	CNTNAP2	Recessive EPI syndrome, ASD, ADHD, TS, OCD
	9q34.13	TSC1	Tuberous Sclerosis type I
	10q23.31	PTEN	Cowden disease*
	11q13.4	DHCR7	Smith-Lemli-Opitz syndrome
	12p13.33	CACNA1C	Timothy syndrome
	15q11.2	UBE3A	Angelman syndrome
	16p13.3	TSC2	Tuberous Sclerosis type II
	17q11.2	NF1	Neurofibromatosis
	Xp21.2	DMD	Duchenne muscular dystrophy
	Xp21.3	ARX	LIS, XLID, EPI, ASD
	Xp22.13	CDKL5	X-linked infantile spasm syndrome
	Xq27.3	FMR1	Fragile X syndrome
	Xq28	MECP2	Rett syndrome
Rare Variants	1q21.1	NBPF9	ASD, ID, SCZ, ADHD, EPI
	2p16.3	NRXN1	ASD, ID, language delay, SCZ.
	3p13	FOXP1	ID, ASD, SLI
	6q16.3	GRIK2	Recessive ID
	7q11.23	FKBP6/CLIP2	ASD, ID, language delay
	7q31.1	FOXP2	SLI
	11q13.3-q13.4	SHANK2	ASD, ID
	15q11-q13	MAGEL2/ NDN	ASD, EPI, ID
	16p11.2	VPS35/ORC6	ASD, ADHD, ID, EPI, SCZ
	16p13.3	A2BP1	ID, ASD, EPI, SCZ, ADHD
	17q11.2	SLC6A4	ASD, OCD
	17q12	ACCN1/PNMT	ASD, SCZ, EPI
	22q11.21		DiGeorge syndrome, SCZ, ASD, ID.BPAD
	22q13.33	SHANK3	ASD, Phelan McDermid syndrome**
	Xq13.1	NLGN3	ASD
	Xp22.11	PTCHD1	ASD, ID
	Xp22.32-p22.31	NLGN4X	ASD, ID, TS, ADHD
Common Alleles	1q42.2	DISC1	SCZ,BPAD
	2q31.1	SLC25A12	ASD
	3p25.3	OXTR	ASD
	7q31.2	MET	ASD, Diabetes II
	7q22.1	RELN	ASD
	7q36.3	EN2	ASD
	12q14.2	AVPR1A	ASD
	17q21.32	ITGB3	ASD

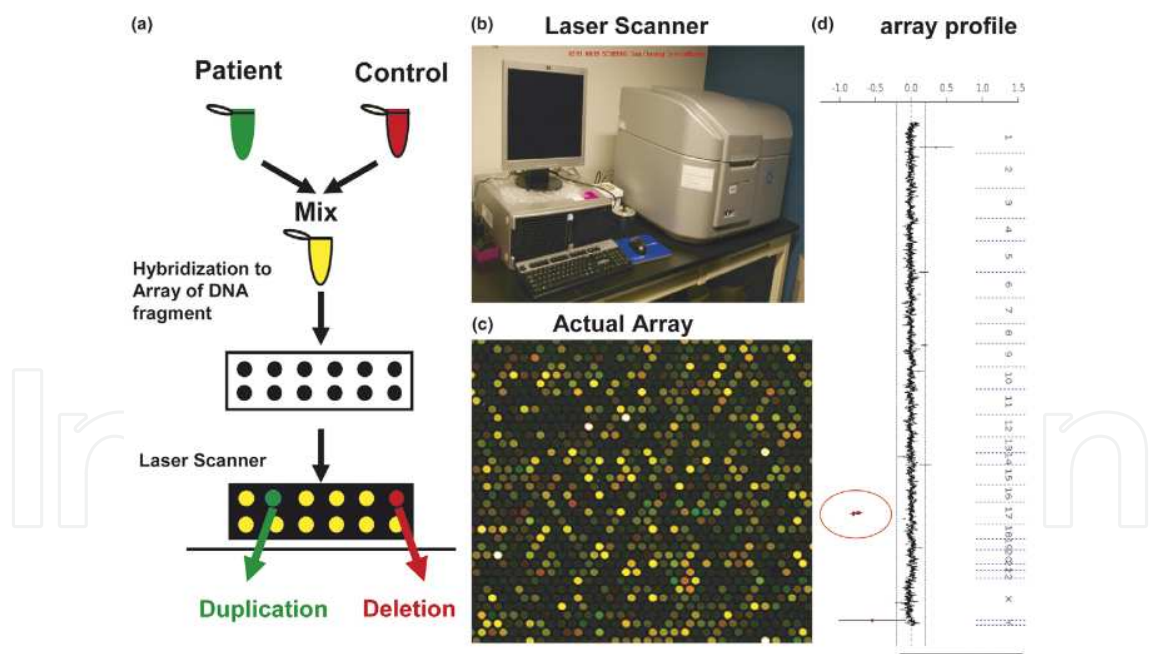
**Table 1.** Chromosomal regions and genes that are implicated in risk for ASD, and associated genetic disorders and syndroms [68& 69].Abbreviations: LTD, long-term depression; LTP, long-term potentiation; PPI, prepulse inhibition; E/I, excitatory/inhibitory; PSD, postsynaptic density; ASD, autism spectrum disorders; SCZ, schizophrenia; ADHD, attention deficit hyperactivity disorder; ID, intellectual disability; XLID, X-linked intellectual disability; LIS, lissencephaly; EPI, epilepsy; OCD, obsessive compulsive disorder; TS, Tourette syndrome; SLI, speech and language impairment; USV, ultrasonic vocalization; TF, transcription factor; ECM, extracellular matrix; GPCR, G-protein-coupled receptor;BPAD, Bipolar affective disorder. \*A rare autosomal dominant inherited disorder characterized by multiple tumor-like growths, increased risk of certain forms of cancer, and diverse clinical features including neurologic features such as

autism and Lhermitte Duclos disease [39& 40].\*\* A genetic syndrome caused by disruption of the SHANK3 gene which codes for the shank3 protein. The protein most important role is in the brain. It is involved in processes crucial for learning and memory. It also has an important role in brain development. It is also known as 22q13.3 deletion syndrome and is highly associated with autism. Human (Homo sapiens) Genome Browser Gateway, <http://genome.ucsc.edu/cgi-bin/hgGateway>.

A set of chromosomal regions and genes that are implicated with ASD are listed in Table 1. Some of the regions are associated with known Mendelian syndromes. In some individuals affected with these syndromes, ASD occurs as a secondary diagnosis. In other regions and genes, genetic variations causing ASD include a wide range of possibilities each with very low frequency among the cases (rare variants). In some cases the rare variants are found only once in the population. In contrast to rare variants we see that in other chromosomal regions and genes only few common genetic variations (common alleles) account for ASD susceptibility.

2.2. Data Generating

Figure 3 illustrates the process of generating DNA copy number data using Microarray-based comparative genomic hybridization (array CGH) technology.

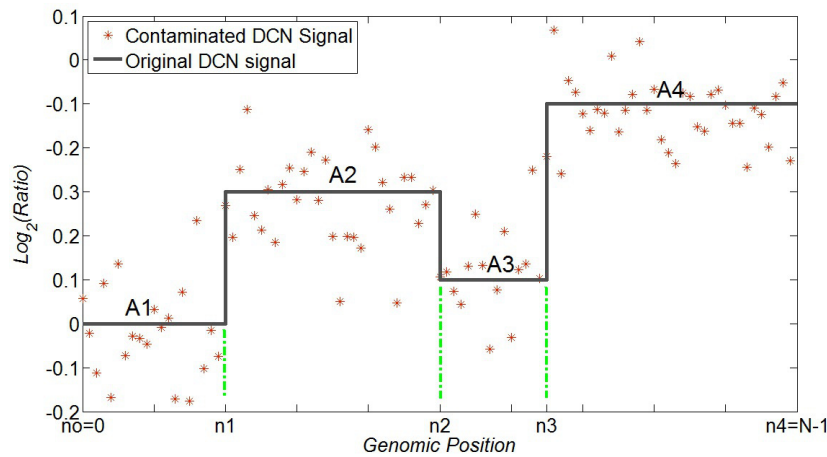


**Figure 3.** Principles of the aCGH technology. (a) DNA from the sample to be tested and reference DNA are labeled with a green fluorescence dye (Cy3) and red (Cy5), respectively, and competitively co-hybridized to an array containing genomic DNA targets that have been spotted on a glass slide. The resulting ratio of the fluorescence intensities is proportional to the ratio of the copy numbers of DNA sequences in the test and reference genomes measured in a logarithmic scale. (b) The slides are scanned using a specific microarray scanner shown in (c). (d) The output of the scanning process is the ratio of the fluorescence intensities for each spot represented as a point in the relative copy number profile [66].



### 2.3. Data Modeling

As illustrated in Figure 3, aCGH technology is an experimental approach for genome-wide scanning of differences in DCN samples. It provides a high-resolution method to map and measure relative changes in DCN simultaneously at thousands of genomic loci. In a biological experiment, unknown (test) and reference (normal) DNA samples are labeled with fluorescent dyes Cy3 and Cy5, respectively. Then, they are combined and competitively co-hybridized to an array containing genomic DNA targets that have been spotted on a glass slide. The resulting ratio of the fluorescence intensities is proportional to the ratio of the copy numbers of DNA sequences in the test and reference genomes measured in a logarithmic scale for a certain genomic location. These intensity ratios are informative about DNA copy number changes. We expect to see duplication (gain) for positive ratio, deletion (loss) for negative ratio and normal state for neutral ratio. Due to the logarithmic scale and the probes performance, the data can be approximated as a piecewise function of short and long intervals with different intensity levels that are not equally-spaced along the genome. Moreover, microarray experiments suffer from many sources of error due to human factors, array printer performance, labeling, and hybridization efficiency.



**Figure 4.** Graphical representation of the generated data using aCGH technology. The red stars represent the raw data as described in (1). The grey solid line represents the true value of 4 variant segments that need to be estimated with intensity levels  $A_i$  measured in  $\log_2(\text{ratio})$  and bounded by the breakpoints  $n_{i-1}$  and  $n_i$ , respectively.

According to the data description and properties generated by microarray technology, the DCN cell line can be approximated as a one-dimensional piecewise constant (PWC) discrete-time signal contaminated with some error. A good model of the genetic data generated by the aCGH technology can be model as follows.

$$y[n] = x[n] + \varepsilon_n, \quad n = 1, 2, \dots, N. \quad (1)$$

where  $y[n]$  is the contaminated genetic signal and  $x[n]$  is the true value of the genetic variation to be estimated at genomic location  $n$  of the length  $N$ .  $\varepsilon_n$  is assumed to be modeled as additive white Gaussian noise with zero mean and some variance  $\sigma^2$ .

As described in (1), Figure 4 illustrates the genetic data in the form of DNA copy number generated by aCGH technology where 4 variant segments are presented with different intensity levels.

### 3. Methods

#### 3.1. Data Filtering

Although the recent advancement in microarray technologies and sequencing now make it easy to measure the genetic variations with high-resolution through scanning large number of samples, small changes, particularly at the low copy repeat (LCRs) regions, remain difficult to detect due to different noise conditions. Thus, the challenging problem is to differentiate between the true biological signaling and the noise measurements.

Various methods have been proposed as preprocessing techniques to tackle this problem. These methods have been motivated by either well-known signal processing techniques or statistical-based models.

METHOD	COMPUTATIONAL
SMOOTHING TECHNIQUES	COMPLEXITY
SIGMA FILTERING (Alqallaf et al., 2007)	$O(N)$
SMOOTHING AND EDGE DETECTION (Huang et al., 2004)	$O(N)$
WAVELETS (Hsu et al., 2005)	$O(N \log N)$
STATISTICAL-BASED MODELS	
CIRCULAR BINARY SEGMENTATION (Olshen et al., 2004)	$O(N^2)$
HIDDEN MARKOV MODELS (Fridlyand et al., 2004)	$O(C^2N)$
SPARSE BAYESIAN LEARNING (Pique-Regi et al., 2008)	$O(N \log N)$

**Table 2.** Comparison based on the computational complexity of the proposed denoising techniques.

In Table 2, we present a comparison study based on the computational cost of the most recent and successful approaches. As can be noticed that the smoothing techniques are well suited to process very large amount of data such as the genetic signals compared to the statistical-based models. However, these techniques include important features such as the variant regions boundaries in the smoothing process.

Here we present our previously proposed method (Alqallaf et al., 2007), Sigma filter (SF). It is a nonlinear method used as a feature extraction to detect the variant segments edges and

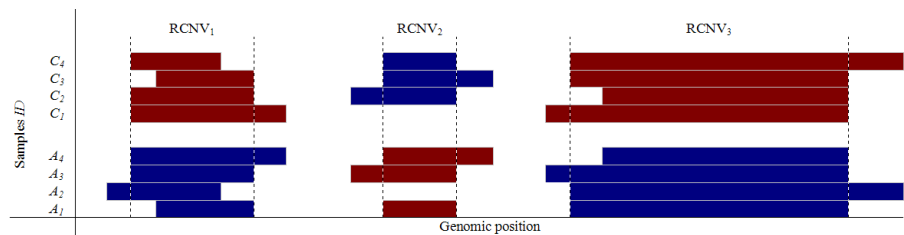


to smooth the rest of the genetic data. The filter is conceptually simple but effective noise smoothing algorithm. Based on the assumption of the aCGH data modeling, the SF algorithm is well suited to denoise the tested samples before further analysis. SF algorithm is motivated by the sigma probability of the Gaussian distribution, and it smooths the noise by averaging only those neighborhood variant segments which have the intensities within a fixed sigma range of the center data point. Consequently, variant segments edges are preserved, and subtle details are retained.

3.2. Statistical significance

Few studies in the literature have addressed the power of class discovery of the recurrent copy number variations (CNVs) across multiple samples of the genetic data [52& 53]. However, they did not consider denoising the data prior to applying the statistical analysis.

To reduce the dimensionality of the detected variant regions, we apply a simple statistical-based approach to measure the significance of the candidate genomic regions. The approach is based on the frequency difference between the case and control samples at each genomic location. It is used as a feature selection algorithm to select a small subset of variant segments as features for classification. Figure 5 is an illustration of three RCVNs with different sizes of filtered DCN data for multiple samples of normal control ( $C_i$ ) and autistic ( $A_i$ ) individuals, respectively. After selecting the informative segments of the genome, we then applied comparative classification algorithms on the reduced data.



**Figure 5.** Schematic representation of 3 recurrent copy number variant segments (RCNVs) with different lengths. The x-axis represents the genomic position and the y-axis represents the indices of the samples,  $C_i$  is for normal-control samples and  $A_i$  is for autistic samples, respectively. The vertical dashed lines represent the RCNVs boundaries. The dark red and dark blue bars represent duplication and deletion for the corresponding chromosomal regions.

3.3. Data classification

Based on the collected and processed genetic data, we apply a system of classifiers that are used to identify autistic individuals based on their genetic information. This system will help improve detection, identification and diagnosis of autism, which will benefit both the patients and the society in general and will lead to early diagnosis and treatment.

Generally classifiers are used by researchers faced with the task of classification based on a given data. Classifiers are mathematical models that are able to perform the task of classifi-

cation or decision making, based on a previously provided data. Classifier's ability to spot trends and relationships in large data sets makes it well suited for many applications. In the field of medicine classifiers can be used to classify accurately diseases, genes, tumors, and other medical phenomena [54; 55; 56; 57; 58; 59& 60]. Although some attempts were made to use classifiers in genetics [61]. Our attempt is to use three comparative classifiers, namely,  $k$ -Nearest Neighbor, Neural Network, and Support Vector Machine, to help in diagnosing patients with ASD.

The leave-one-out cross-validation (LOOCV) is applied to evaluate the proposed classifiers by measuring the classification performance to accurately identify the association between the tested samples and the targeted disorder, ASD. The LOOCV involves using a single variant segment from the original sample as the validation data, and the remaining segments as the training data. This is repeated such that each variant segment in the sample is used once as the validation data.

### 3.3.1. $k$ -Nearest Neighbor Classifier

The  $k$ -Nearest Neighbor ( $k$ -NN) classifier [64] is a well known nonparametric classifier. To classify a new input  $x$ , the  $k$  nearest neighbors are retrieved from the training data. The input  $x$  is then labeled with the majority class label corresponding to the  $k$  nearest neighbors. For the  $k$ -NN classifier, we used the Euclidean distance as the distance metric, and the best  $k$  between 1 and 10 was found by performing LOOCV on the training data.

### 3.3.2. Neural Network

Neural networks are another type of classifier or mathematical models used for classification, regression or decision making. Their structure is inspired by the human neural system and brain. It consists of many neurons, interconnected at different stages. The direction of flow of information is usually from the input stage to the output stage. Each neuron has an input and an output, where an activation function converts a neurons input to its output. The output of each neuron is connected to the next stage through a weighted connection. A learning function determines the value of the weights of all the connections. The weights are updated based on a mathematical function that relates the network together. Therefore, a neural network is considered as an adaptive network that changes its structure during the learning or training phase, based on mathematical functions that relate input data to the corresponding class labels. The sum of all neurons at the different layers and the weighted interconnections make up a complex network that is commonly referred to as a black box.

Before its use to classify a test sample, the neural network is trained on a given data set with known classes or labels. During the training phase the weights are updated to minimize the output error. The selected value of the minimum acceptable error determines when the training stops. For a difficult data where it is impossible to reach the set minimum error, the maximum number of epochs is used as criteria for stopping the training process.

3.3.3. Support Vector Machine

The Support Vector Machine (SVM) belongs to a new generation of learning system based on recent advances in statistical learning theory [65]. A linear SVM, which is used in our system, aims to find the separating hyper-plane with the largest margin, defined as the sum of the distances from a hyper-plane (implied by a linear classifier) to the closest positive and negative exemplars. The expectation is that the larger the margin, the better the generalization of the classifier. In a non-separable case, a linear SVM seeks a trade-off between maximizing the margin and minimizing the number of errors.



**Figure 6.** Comparison study of the performance of the three tested classifiers. The x-axis represents the number of segments and the y-axis represents the percentage average LOOCV accuracy.

Figure 6 illustrates the LOOCV classification accuracies using the tested classifiers, *k*-NN, NN, and SVM. The *x*-axis is associated with the number of selected top-ranked variant segments and the *y*-axis shows the average LOOCV accuracy.

4. Validation of the predicted variant segments

To evaluate our predictive power of our method in detecting and identifying patients with ASD, we use molecular test, quantitative Polymerase Chain Reaction (qPCR). It is a very sensitive and precise tool used for the quantification of nucleic acids. It can detect and quantify very small amounts of specific nucleic acid sequence. It is based on the method of PCR, developed by Kary Mullis in the 1980s. It allows the amplification of specific nucleic acid sequence (DNA) more than a billion-fold. Using qPCR allows scientists to quantify the starting amount of a specific DNA sequence in the sample before the amplification by PCR method [62].

Quantitative PCR is an indispensable tool for researchers in various fields including fundamental biology, molecular diagnostics, biotechnology, and forensic sciences. Critical points and limitations of qPCR-based assays must be considered to increase the reliability of the obtained data. For the detection of qPCR four technologies are commonly used all of which are based on the measurement of fluorescence during the PCR. One principle is based on intercalation of double-stranded DNA-binding dyes (simplest and cheapest). The other three principles are based on the introduction of an additional fluorescence-labeled oligonucleotide (probe). Detectable fluorescence are only released either after cleavage of the probe (hydrolysis probes) or during hybridization of one (molecular beacon) or two (hybridization probes) oligonucleotides to the amplicon. The introduction of an additional probe increases the specificity of the quantified PCR product and allows the development of multiplex reactions. Other technologies have been described for the detection of qPCR [63].

The qPCR method quickly became the first choice when it comes to quantitative analysis of nucleic acid because of many reasons. It is highly sensitive and it allows the detection of less than five copies (one copy in some cases) of a target sequence. It has good reproducibility. In addition, it has broad dynamic quantification range, at least 5 log units. It is also easy to use and has reasonable good value for money (low consumable and instrumentation costs).

For the purpose of this chapter, we are focusing on one of the many applications of qPCR, which is indispensable for research and diagnostics, the genetic variations.

Array	CBS	SF
1	14	20
2	10	20
3	20	21
4	13	21

**Table 3.** Representation of the number of events (CNVs) detected by the circular binary segmentation (CBS) and sigma filtering methods, respectively, for 22 qPCR confirmed CNVs.

Table 3 shows that the number of qPCR-confirmed CNVs detected by the sigma filtering (SF) method is considerably higher than those detected using the circular binary segmentation (CBS), ranging from 4.5% to 36% more for 4 different array experiments. The results show that applying the averaging window of 2Kb allow the algorithms to be well suited for detecting variations in high-density microarray data, especially at the LCR-rich regions.

5. Conclusion

The etiology of Autism spectrum disorders involves genetic and environmental risk factors. In this chapter, we have discussed the genetics basis of the complex disorder, autism. With the recent advances in the new screening technologies to investigate the entire genome such

as array comparative genomic hybridization (aCGH) and whole genome sequencing, provide the opportunities insight into the pattern of the genetic variations and reveal their roles in the genetic diseases. In this study, we have demonstrated an overview for the analysis of genetic variations in the form of DNA copy number changes and their association with autism susceptibility.

Through mathematical-based models and computational-based approaches, we analyze the genetic data trying to discover and identify the relationship between the structural chromosomal rearrangements along the genome and the targeted disorder, ASD. In conclusion, the results show strong evidence that the genetic variations contribute in the complex disorder, autism.

## Author details

Abdullah K. Alqallaf<sup>1\*</sup>, Fuad M. Alkoot<sup>2</sup> and Mash'el S. Aldabbous<sup>3</sup>

\*Address all correspondence to: al.qallaf@ku.edu.kw

1 Electrical Engineering Department, College of Engineering and Petroleum, Kuwait University, Kuwait

2 The Higher Institute of Telecommunication and Navigation, Public Authority for Applied Education and Training, Kuwait

3 Department of Biological Sciences, College of Science, Kuwait University, Kuwait

## References

- [1] Freeman, J. L., Perry, G. H., Feuk, L., Redon, R., Mc Carroll, S. A., Altshuler, D. M., Aburatani, H., Jones, K. W., Tyler-Smith, C., & Hurles, . (2006). Copy number variation: new insights in genome diversity. *Genome Research*, 16(8), 949-961.
- [2] Lee, J. A., & Lupski, J. R. (2006). Genomic rearrangements and gene copy-number alterations as a cause of nervous system disorders. *Neuron*, 52(1), 103-121.
- [3] Stankiewicz, P., & Lupski, J. R. (2002). Molecular-evolutionary mechanisms for genomic disorders. *Current Opinion in Genetics & Development on ScienceDirect*, 12(3), 312-319.
- [4] Christian, S. L., Brune, C. W., Sudi, J., Kumar, R. A., Liu, S., Karamohamed, S., Badner, J. A., Matsui, S., Conroy, J., & Mc Quaid, D. (2008). Novel submicroscopic chromosomal abnormalities detected in autism spectrum disorder. *Biological Psychiatry*, 63(12), 1111-1117.



- [5] Marshall, C. R., Noor, A., Vincent, J. B., Lionel, A. C., 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*, 82(2), 477-488.
- [6] 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*, 316(5823), 445-449.
- [7] Kirov, G., Gumus, D., Chen, W., Norton, N., Georgieva, L., Sari, M., O'Donovan, M. C., Erdogan, F., Owen, , & Ropers, H. H. (2008). Comparative genome hybridization suggests a role for NRXN1 and APBA2 in schizophrenia. *Human Molecular Genetics*, 17(3), 458-465.
- [8] Sharp, A. J., Hansen, S., Selzer, R. R., Cheng, Z., Regan, R., Hurst, J. A., Stewart, H., Price, S. M., Blair, E., & Hennekam, R. C. (2006). Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nature Genetics*, 38(9), 1038-1042.
- [9] Sharp, A. J., Mefford, H. C., Li, K., Baker, C., Skinner, C., Stevenson, R. E., Schroer, R. J., Novara, F., De Gregori, M., & Ciccone, R. (2008). A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. *Nature Genetics*, 40(3), 322-328.
- [10] Cantor, R. M., & Geschwind, D. H. (2008). Schizophrenia: genome, interrupted. *Neuron*, 58(2), 165-167.
- [11] Stefansson, H., Rujescu, D., Cichon, S., Pietilainen, O. P., Ingason, A., Steinberg, S., Fossdal, R., Sigurdsson, E., Sigmundsson, T., & Buizer-Voskamp, J. E. (2008). Large recurrent microdeletions associated with schizophrenia. *Nature*, 455(7210), 232-236.
- [12] Stone, J. L., O'Donovan, M. C., Gurling, H., Kirov, G. K., Blackwood, D. H., Corvin, A., Craddock, N. J., Gill, M., Hultman, C. M., & Lichtenstein, P. (2008). Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature*, 455(7210), 237-241.
- [13] Walsh, T., Mc Clellan, J. M., Mc Carthy, S. E., Addington, A. M., Pierce, S. B., Cooper, G. M., Nord, , Kusenda, M., Malhotra, D., & Bhandari, A. (2008). Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science*, 320(5875), 539-543.
- [14] Estivill, X., & Armengol, L. (2007). Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. *PLoS Genetics*, 3(10), 1787-1799.
- [15] Greer, M. K., Brown, F. R., Pai, G. S., Choudry, S. H., & Klein, A. J. (1997). Cognitive, adaptive, and behavioral characteristics of Williams syndrome. *American Journal of Human Genetics*, 74(5), 521-525.
- [16] Somerville, M. J., Mervis, C. B., Young, E. J., Seo, E. J., del Campo, M., Bamforth, S., Peregrine, E., Loo, W., Lilley, M., & Perez-Jurado, L. A. (2005). Severe expressive-lan-



- guage delay related to duplication of the Williams-Beuren locus. *The New England Journal of Medicine*, 353(16), 1694-1701.
- [17] van der , 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*, 52(2-3), 94-100.
  - [18] Depienne, C., Heron, D., Betancur, C., Benyahia, B., Trouillard, O., Bouteiller, D., Verloes, A., Le Guern, E., Leboyer, M., & Brice, A. (2007). Autism, language delay and mental retardation in a patient with 7q11 duplication. *Journal of Medical Genetics*, 44(7), 452-458.
  - [19] Balciuniene, J., Feng, N., Iyadurai, K., Hirsch, B., Charnas, L., Bill, B. R., Easterday, M. C., 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*, 80(5), 938-947.
  - [20] Laumonnier, F., Roger, S., Guerin, P., Molinari, F., M'Rad, R., Cahard, D., Belhadj, A., Halayem, M., Persico, A. M., & 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*, 163(9), 1622-1629.
  - [21] Cook, E. H., & Jr , . (2001). Genetics of autism. *Child and Adolescent Psychiatric Clinics of North America*, 10(2), 333-350.
  - [22] Helbig, I., Mefford, H. C., Sharp, A. J., 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*, 41(2), 160-162.
  - [23] Papolos, D. F., Faedda, G. L., Veit, S., Goldberg, R., Morrow, B., Kucherlapati, R., & Shprintzen, R. J. (1996). Bipolar spectrum disorders in patients diagnosed with velocardio-facial syndrome: does a hemizygous deletion of chromosome 22q11 result in bipolar affective disorder? *American Journal of Psychiatry*, 153(12), 1541-1547.
  - [24] Niklasson, L., Rasmussen, P., Oskarsdottir, S., & Gillberg, C. (2001). Neuropsychiatric disorders in the 22q11 deletion syndrome. *Genetics in Medicine*, 3(1), 79-84.
  - [25] Antonell, A., de Luis, O., Domingo-Roura, X., & Perez-Jurado, L. A. (2005). Evolutionary mechanisms shaping the genomic structure of the Williams-Beuren syndrome chromosomal region at human 7q11.23. *Genome Research*, 15(9), 1179-1188.
  - [26] Vorstman, J. A., Morcus, M. E., Duijff, S. N., Klaassen, P. W., Heineman-de, Boer. J. A., Beemer, F. A., Swaab, H., Kahn, R. S., & 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*, 45(9), 1104-1113.
  - [27] Hertz-Picciotto, I., Croen, L. A., Hansen, R., Jones, C. R., van de Water, J., & Pessah, I. N. (2006). The CHARGE study: an epidemiologic investigation of genetic and envi-

- ronmental factors contributing to autism. *Environmental Health Perspectives*, 114(7), 1119-1125.
- [28] Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., & Chen, W. (2006). Global variation in copy number in the human genome. *Nature*, 444(7118), 444-454.
- [29] Shen, F., Huang, J., Fitch, K. R., Truong, V. B., Kirby, A., Chen, W., Zhang, J., Liu, G., Mc Carroll, S. A., & Jones, K. W. (2008). Improved detection of global copy number variation using high density, non-polymorphic oligonucleotide probes. *BMC Genetics*, 9(27).
- [30] Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T. D., Barnes, C., & Campbell, P. (2010). Origins and functional impact of copy number variation in the human genome. *Nature*, 464(7289), 704-712.
- [31] Perry, G. H., Ben-Dor, A., Tsalenko, A., Sampas, N., Rodriguez-Revena, L., Tran, C. W., Scheffer, A., Steinfeld, I., Tsang, P., & Yamada, N. A. (2008). The fine-scale and complex architecture of human copy-number variation. *American Journal of Human Genetics*, 82(3), 685-695.
- [32] Berg, J. S., Brunetti-Pierri, N., Peters, S. U., Kang, S. H., Fong, C. T., Salamone, J., Freedenberg, D., Hannig, V. L., Prock, L. A., & Miller, D. T. (2007). Speech delay and autism spectrum behaviors are frequently associated with duplication of the 7q11.23 Williams-Beuren syndrome region. *Genetics in Medicine*, 9(7), 427-441.
- [33] Potocki, L., Bi, W., Treadwell-Deering, D., Carvalho, C. M., Eifert, A., Friedman, E. M., Glaze, D., Krull, K., Lee, J. A., & Lewis, R. A. (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*, 80(4), 633-649.
- [34] Carvalho, C. M., & Lupski, J. R. (2008). Copy number variation at the breakpoint region of isochromosome 17q. *Genome Research*, 18(11), 1724-1732.
- [35] Mukaddes, N. M., & Herguner, S. (2007). Autistic disorder and 22q11.2 duplication. *World Journal of Biological Psychiatry*, 8(2), 127-130.
- [36] Barabash, A., Marcos, A., Ancin, I., Vazquez-Alvarez, B., de Ugarte, C., Gil, P., Fernandez, C., Encinas, M., Lopez-Ibor, J. J., & Cabranes, J. A. (2009). APOE, ACT and CHRNA7 genes in the conversion from amnesic mild cognitive impairment to Alzheimer's disease. *Neurobiol Aging*, 30(8), 1254-1264.
- [37] Miller, D. T., Shen, Y., Weiss, L. A., Korn, J., Anselm, I., Bridgemohan, C., Cox, G. F., Dickinson, H., Gentile, J., & Harris, D. J. (2009). Microdeletion/duplication at 15q13.2q13.3 among individuals with features of autism and other neuropsychiatric disorders. *Journal of Medical Genetics*, 46(4), 242-248.
- [38] Shinawi, M., Schaaf, , Bhatt, S. S., Xia, Z., Patel, A., Cheung, S. W., Lanpher, B., Nagl, S., Herding, H. S., & Nevinsky-Stickel, C. (2009). A small recurrent deletion within

15q13.3 is associated with a range of neurodevelopmental phenotypes. *Nature Genetics*, 41(12), 1269-1271.

- [39] Tsuchiya, K. D., Wiesner, G., Cassidy, S. B., Limwongse, C., Boyle, J. T., & 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*, 21(2), 113-118.
- [40] Zhou, X. P., 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*, 69(4), 704-711.
- [41] Fombonne, E. (2003). Epidemiological Surveys of Autism and Other Pervasive Developmental Disorders: An Update. *Journal of Autism and Developmental Disorders*, 33(4), 365-382.
- [42] Kuehn, B. (2007). CDC: Autism Spectrum Disorders Common. *Journal of the American Medical Association*, 297(940).
- [43] Abrahams, B., & Geschwind, D. (2008). Advances in autism genetics: on the threshold of a new neurobiology. *Nature Reviews Genetics*, 9(5), 341-355.
- [44] Alqallaf, A., Tewfik, A., & Selleck, S. (2009). Genetic variation detection using maximum likelihood estimator. *IEEE International Workshop on Genomic Signal Processing and Statistics*, 978-1-42444-761-9, Minnesota, USA.
- [45] 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*, 258(5083), 818-821.
- [46] 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*, 3(2), 206-222.
- [47] Willenbrock, H., & Fridlyand, J. (2005). A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics*, 21(22), 4084-4091.
- [48] Rissanen, J. (1978). Modeling by Shortest Data Description. *Automatica*, 14, 465-471.
- [49] Larson, R., & Castie, J. (1982). Principles of Dynamic Programming. 1-2, Marcel Dekker Inc., NY.
- [50] BioDiscovery Inc. (2012). Nexus: Copy Number Professional software package (BioDiscovery), Inc., El Segundo, CA) available from: <http://www.biodiscovery.com>.
- [51] Venkatraman, E., & Olshen, A. (2007). A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, 23(6), 657-663.

- [52] Grant, G., Manduchi, E., Cheung, V., & Ewens, W. (1999). Significance testing for direct identity-by-descent mapping. *Annals of Human Genetics*, 63(5), 441-454.
- [53] 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*, 16(9), 1149-1158.
- [54] Rohlfing, T., Russakoff, D., & Maurer, C. (2004). Performance based classifier combination in atlas based image segmentation using expectation maximization parameter estimation. *IEEE transaction on medical imaging*, 23(8), 983-994.
- [55] Guler, I., & Ubeyli, E. (2005). ECG beat classifier designed by combined neural network model. *Pattern Recognition*, 38, 199-208.
- [56] Sampat, M. ., Bovik, A., Aggarwal, J., & Castleman, K. (2005). Supervised parametric and non parametric classification of chromosome images. *Pattern Recognition*, 38, 1209-1223.
- [57] Bruijne, M., & Nielson, M. (2004). Shape particle filtering for image segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 1, 168-175.
- [58] Shin, H., & Sohn, S. (2005). Selected tree classifier combination based on both accuracy and error diversity. *Pattern Recognition*, 38, 191-197.
- [59] Tsymbal, A., Cunningham, P., Pechenizkiy, M., & Puuronen, S. (2003). Search strategies for ensemble feature selection in medical diagnosis. 16th IEEE symposium on computer based medical systems,, 124-129.
- [60] Kook, H., Gupta, L., Molfese, D., & Fadem, K. (2005). Multi-stimuli multi-channel data and decision fusion strategies for dyslexia prediction using neonatal ERPS. *Pattern Recognition*, 38(11), 2174-2184.
- [61] Moon, H., Ahn, H., Kodell, R., Lin, C., Baek, S., & Chen, J. (2006). Classification methods for the development of genomic signatures from high-dimensional data. *Genome Biology*, 7(12), 121.
- [62] Valasek, M., & Repa, J. (2005). The power of real-time PCR,. *Advances in Physiology Education*, 29(3), 151-159.
- [63] Klein, D. (2002). Quantification using real-time PCR technology: applications and limitations. *Trends in Molecular Medicine*, 8(6), 257-260.
- [64] Dasarathy, B. (1991). Nearest Neighbor Norms: NN Pattern Classification Techniques. IEEE. Computer Society Press.
- [65] Vapink, V. (1998). Statistical learning Theory. Wiley-Interscience.
- [66] Shinawi, M., & Cheung, S. (2008). The array CGH and its clinical applications. *Drug Discovery Today*, 13(17-18), 760-770.

- [67] Schwab, M., Alitalo, K., Klempnauer, K., Varmus, H., Bishop, J., Gilbert, F., Brodeur, G., Goldstein, M., & Trent, J. (1983). Amplified DNA with limited homology to MYC cellular oncogene is shared by human neuroblastoma cell lines and a neuroblastoma tumour. *Nature*, 305(5931), 245-248.
- [68] Aldinger, K., Plummer, J., Qiu, S., & Levitt, P. (2011). SnapShot: Genetics of Autism. *Neuron*, 72, 418.
- [69] Sanders, S., et al. (2011). Multiple Recurrent De Novo CNVs, Including Duplications of the 7q11.23 Williams Syndrome Region, Are Strongly Associated with Autism. *Neuron*, 70, 863-885.