THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF BIOCHEMISTRY AND MOLECULAR BIOLOGY

GENOME-WIDE MICROARRAY ANALYSIS IN A CASE-CONTROL STUDY REVEALS EVELATED LEVEL OF GLOBAL COPY NUMBER BURDEN IN AUTISM

KIAN HUI YEOH
SPRING 2012

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree
in BIOTECHNOLOGY
with honors in BIOTECHNOLOGY

Reviewed and approved* by the following:

Maria Krasilnikova
Research Assistant Professor
Department of Biochemistry and Molecular Biology
Thesis Supervisor and Research Advisor

Scott B. Selleck
Professor and Department Head
Department of Biochemistry and Molecular Biology
Thesis Supervisor and Research Advisor

David S. Gilmour
Professor of Molecular and Cell Biology
Department of Biochemistry and Molecular Biology
Honors Adviser

Wendy Hanna-Rose
Associate Professor of Biochemistry and Molecular Biology
Associate Department Head for Undergraduate Studies

* Signatures are on file in the Schreyer Honors College.

ABSTRACT

Copy number variations (CNV), or structural genomic variations, have recently been shown to be implicated in and associated with numerous human neurodevelopmental disorders such as autism spectrum disorders (ASD), schizophrenia, epilepsy, mental retardations, developmental delays, bipolar disorder, intellectual disability (ID) and many other human diseases. In this study, we used a custom-design DNA microarray targeted at regions flanked by segmental duplications (SD) to evaluate and compare global copy number burden between 274 autistic (AU) patients and 280 typically developing (TD) individuals from the general population. We found that there was an elevated level of global CNV burden in AU individuals (Mann-Whitney Test, U= 23449, p= 0.00185). Additionally, we showed that total average duplication length in AU individuals was also significantly higher than that of controls (Mann-Whitney Test, U= 28129, p= 0.0109) compared to total average deletion length (Mann-Whitney Test, U= 20804, p=0.432). Interestingly, this differential level of global burden between cases and controls was mostly detected in the genomic backbone or non-hotspot region (Mann-Whitney Test, U=7992, p=0.0128), but not the genomic hotspot regions (Mann-Whitney Test, U=28987, p=0.513) as we originally expected. The data from this case-control study demonstrated an elevated level of copy number changes in autistic individuals, which was primarily represented by large duplication events located in the non-hotspot or genomic backbone regions. The findings from this study also implicated that there may be an alternative mechanism other than non-allelic homologous recombination (NAHR) involved in this significant copy number variation in autism.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

ACKNOWLEDGEMENTS

I would like to express my utmost gratitude to my research advisor and thesis supervisor, Dr. Maria Krasilnikova, for her constant, devoted guidance and assistance throughout the last two years I have been working in the lab. Without her, my academic and research achievement would not have been where it is now. She has been a very dedicated and caring professor, and has been very helpful in conducting my own independent research.

I would also like to specially thank Dr. Scott B. Selleck, Professor and Department Head of Biochemistry and Molecular Biology, who is also my research and thesis supervisor, for the precious opportunity he has given me to participate in a research as a student intern in Dr. Evan Eichler's lab at University of Washington, Seattle, over the summer in 2011. I have certainly gained a lot of invaluable experience and advanced research techniques, as an undergraduate student, from the research collaboration I have been involving in. In addition to that, I truly appreciate Dr. Santhosh Girirajan, who is an outstanding senior investigator and my summer research supervisor, and Carl Baker, who is an amazing research scientist, for their guidance and teaching during my summer internship in Eichler Lab.

Last but not least, I would like to thank all my fellow lab colleagues, Qiao Kai Law, Su Jen Khoo, Abhinaya Srikanth, Nari Kim, and Stephen Wellard for all the support and assistance they have given to me over the course of my research in the lab.

INTRODUCTION

Genomic structural variants and human disorders

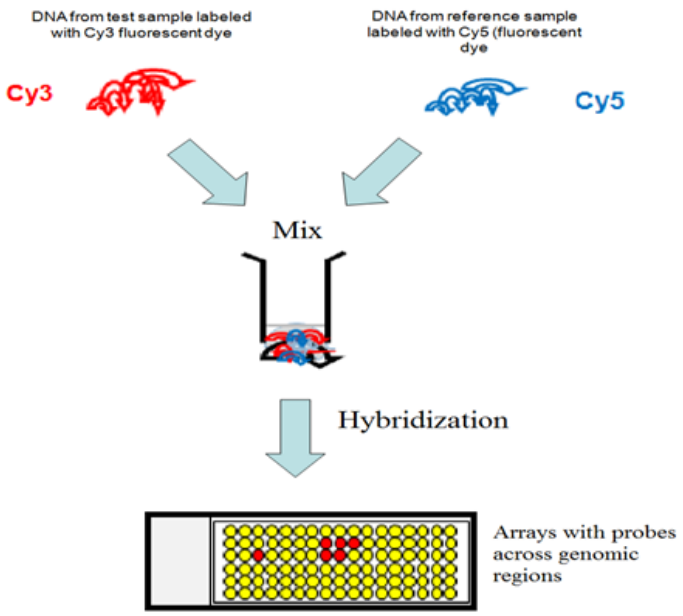
Structural genomic variations or rearrangements are changes in chromosomal structure that could possibly lead to inversions, translocations, duplications and deletions of any parts or sections of the chromosome, which could be as small as one single nucleotide or as large as one whole chromosome. The most pronounced types of genomic variations are duplications and deletions that cumulatively constitute copy number variation (CNV), which results in an atypical number of copies of one or more segments in the human genome. CNV contributes to an estimated portion of 12-15% of the human genome, with each variation ranging from a few thousand base pairs to several million base pairs of DNA, significantly contributing to human genetic heterogeneity [1]. Studies emerging in recent years have shown that CNV was implicated in a wide range of neurodevelopmental disorders, including but not limited to intellectual disability (ID) [2,3], autism [4,5], bipolar disorder [6], epilepsy [7], schizophrenia [8], as well as attention deficit hyperactivity disorder (ADHD) [9]. In addition to that, a study has shown that an increased CNV burden had a positive correlation with disease severity [10], that is, a larger CNV always leads to more severe neurodevelopmental disorders.

Detecting CNVs using microarray-based method

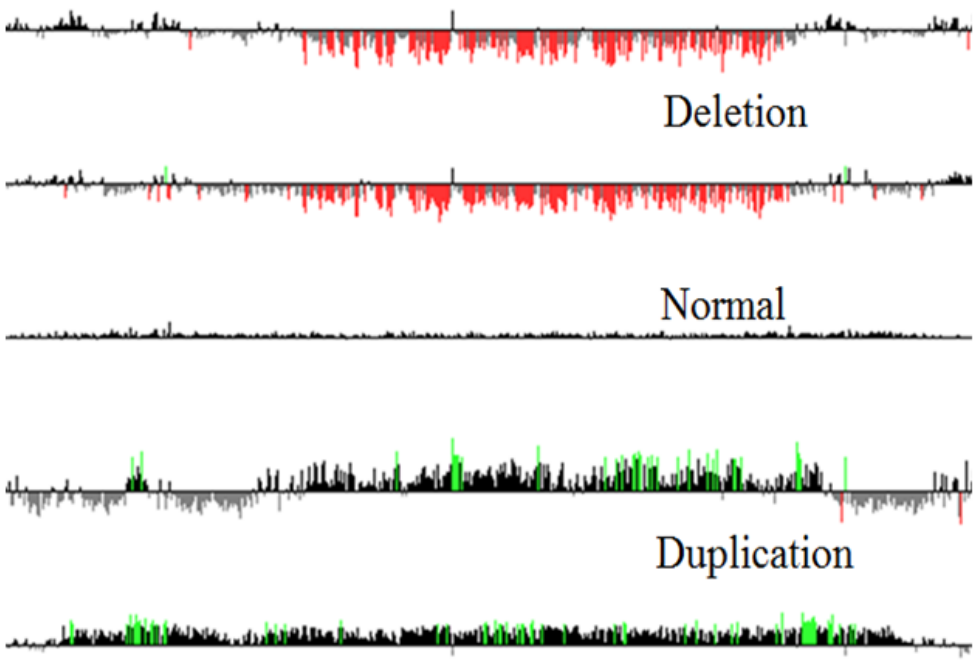
The two most commonly used methods in detecting CNV are SNP-based microarray and array-comparative genomic hybridization (aCGH), even though recent advances in DNA sequencing technology have enabled the use of next-generation whole-genome sequencing, providing a higher accuracy of CNV detections [11]. Both SNP-based microarray and aCGH detect CNV through the comparison of test samples to a reference sample, which is then analyzed by a

computer for producing specific output signals. These signals are then normalized using certain mathematical or statistical algorithms to determine the output signal intensity: increased intensity of the output signal reflects a duplication, which is literally a presence of two or more copies of a particular segment of the genome, whereas a decreased intensity reflects deletions. Figure 1A illustrates the general procedure of aCGH, and Figure 1B shows an example of its generated graph output.

Figure 1. A) Schematic representation of aCGH experimental procedure: DNA samples from both test case and reference control are each labeled with a different fluorescent dye,. After mixing, the labeled DNA samples are then hybridized to an array slide, which has specific probes covering genomic regions attached to its surface. After hybridization, the array is then subject to the computer scanning to produce a signal output (not shown). B) Graph output example produced from UCSC genome browser: Graphs are plotted and colored based on a specific threshold of the log intensity ratio; Red: deletion; Green: duplication; Black: the log intensity ratio falls below the threshold of 1.5. Horizontal axis represents the location across the genome. (Figures Courtesy of Dr. Santhosh Girirajan and Dr. Scott Selleck)


DNA from test sample labeled with Cy3 fluorescent dye
DNA from reference sample labeled with Cy5 (fluorescent dye
Cy3
Cy5
Mix
Hybridization
Arrays with probes across genomic regions

A)


Deletion
Normal
Duplication

B)

Association analysis of variants and neurodevelopmental disorders

There are several approaches to studying how CNV could contribute to pathogenicity in neurodevelopmental disorders. One of those is to search for rare and pathogenic CNVs that occur in affected individuals but not in unaffected controls for each of these disorders using whole genome array studies. These variants are usually overlapping with one or more important and critical genes, where a change in a copy number could potentially result in phenotypic and functional alterations leading to the pathogenesis of a disease. Through this approach of association analysis of rare and functional genic variants, several studies have been able to associate one or more CNVs with a particular disorder. For instance, 15q11.2-13 duplication is associated with Angelman/Prader-Willi syndromes and autism spectrum disorders [12], 7q11.23 deletion causes Williams-Beuren syndrome [13-15] whereas its reciprocal duplication is associated with autism, language or speech delay, and mental retardation [16,17]; 16p12.1 microdeletion is associated with a severe developmental delay [18], and 17q12 duplication has associated risks of intellectual disability, seizures [19] and autism [20]. However, it is important to note that although numerous studies have implicated the association and presence of these CNVs with certain disease phenotypes, these variants are not necessarily responsible for the direct causal mechanism of the disease. Hence, it is crucial to fully understand the underlying mechanism by which CNV causes diseases.

Overview of methodology

Whereas numerous studies have associated several rare and pathogenic CNVs with a range of neurobehavioral disorders such as autism, we took an intuitive approach to analyze the relationship between CNV burden and disease phenotypic susceptibility, specifically autism, to

help us understand the relevance of copy number instability across a wide range of these disorders. We performed a systematic analysis on 274 autistic (AU) individuals and 280 typically developing (TD) children from the general population to address the relative contribution of CNV in autism. We hypothesized that there is a greater level of CNV burden in autistic cases compared to controls. To evaluate this, a whole-genome custom microarray was designed to target genomic hotspots for the CGH to detect and identify CNVs in both cases and controls. The data obtained from this microarray was statistically analyzed to assess whether there is an elevated level of copy number burden in AU individuals.

MATERIALS AND METHODS

Ethical Considerations

Patients from each study cohort, including both cases and controls, were recruited in accordance with appropriate human subjects approval and informed consent. Informed consent was also acquired from the patients to obtain DNA samples in conjunction with IRB protocols and guidelines.

Study Cohort and DNA Samples

A total of 553 DNA samples were acquired from the Childhood Autism Risks from Genetics and Environment (CHARGE) Study[21], which was conducted by the Medical Investigations of Neurodevelopmental Disorders (MIND) Institute at University of California, Davis. These DNA samples were obtained from whole blood of the patients who were selected based on complete criteria for Autistic Disorder (OMIM 209850) using ADOS[22], and the Autism Diagnostic Interview, Revised (ADI-R)[23]. On the other hand, individuals with significant impairments in visual, audio, or motor skills, serious birth complications such as extended NICU stay, diagnosis of syndromic autism and known genetic causes of autism including Fragile X Syndrome, were excluded. Evaluation of diagnosis, assessment of cognitive performance, and characterization of phenotypes of each individual as well as data entry and collection were performed at the MIND institute. Aside from the AU individuals, the study cohort also included typically developing (TD) children without any official diagnosis of autism. However, it is worth noting that the TD individuals group obtained from the general population included a small number of individuals with mental retardation (MR)/developmental delay (DD) despite the fact that they had an atypical cognitive performance or other phenotypic considerations non-related to autism. In

summary, the CHARGE study included a total of 273 AU individuals with the following ethnicity breakdown: Caucasian (143), Hispanic (74), Mixed Race (30), Asian (20), and African American (6). A control group consisted of 280 TD individuals with the ethnicities of Caucasian (147), Hispanic (83), Mixed Race (37), Asian (7), African American (5), and Pacific Islander/Hawaii Native (1). Table 1 summarizes the study cohort of both AU and TD individuals, with proportion of males shown in percentage for each group.

Table 1. Summary of the CHARGE study cohort with details of ethnicity breakdown. TD: Typically developing individuals; AU: autistic individuals. % male reflects the percentage of male individuals in each study cohort.

Ethnicity	TD (%male)	AU (% male)	**Total Samples**
Hispanic	83 (0.81)	74 (0.83)	**157**
White	147 (0.83)	143 (0.88)	**290**
Asian	7 (0.66)	20 (0.90)	**27**
African American	5 (0.8)	6 (0.83)	**11**
Mixed	37 (0.58)	30 (0.86)	**67**
Native Hawaiian	1(1.0)	-	**1**
Total samples	**280**	**273**	**553**

Design of the custom oligoarrays for aCGH experiment

As mentioned earlier, array comparative genomic hybridization (aCGH) was used in this study to detect CNV in the AU and TD DNA samples. To perform aCGH on the DNA samples, we used a custom-designed 12-plex oligoarray Hotspot v1.0 (HS1) from Roche Nimblegen Systems, Inc. This array contained a total number of 135,000 probes with two different probe densities coverage: higher density probe coverage in 107 genomic hotspots regions (Table 2) flanked by segmental duplications (SD), and a lower probe density in the genomic backbone that included the rest of the genome. The median probe spacing in the genomic hotspots regions was 2.6 kb, while the genomic backbone region had a mean probe spacing of 36 kb. Although the median density of the chip was designed to be approximately 2.6 kb in the genomic hotspots and 36 kb in genomic backbone, the limitations of the chip design (probe assignment restricted to only up to five mismatches) precluded uniform distribution of the probes throughout the genome. Therefore, the actual probe density varied across regions of the human genome. Figure 2 depicts the distribution of the probe spacing in the Hotspot v1.0 array in details. In addition to Hotspot v1.0 array, another custom targeted 2×400K Agilent chip (Agilent Technologies), with median probe spacing of 500 bp in the genomic hotspots and probe spacing of 14 kb in the genomic backbone, were used in several validation experiments to confirm the rare pathogenic events detected using the Hotspot v1.0 array.

Table 2. The description of the 107 hotspot chromosomal regions (only 87 regions are shown here, since some of the hotspot regions were combined) and probes utilized in the Hotspot v1.0 chip. Note that although the median spacing in the hotspot regions was around 2.6 kb, the average distance between probes was different within each particular region. Some of the regions are known for association of several genomic disorders as indicated. (Table courtesy of Dr. Santhosh Girirajan)

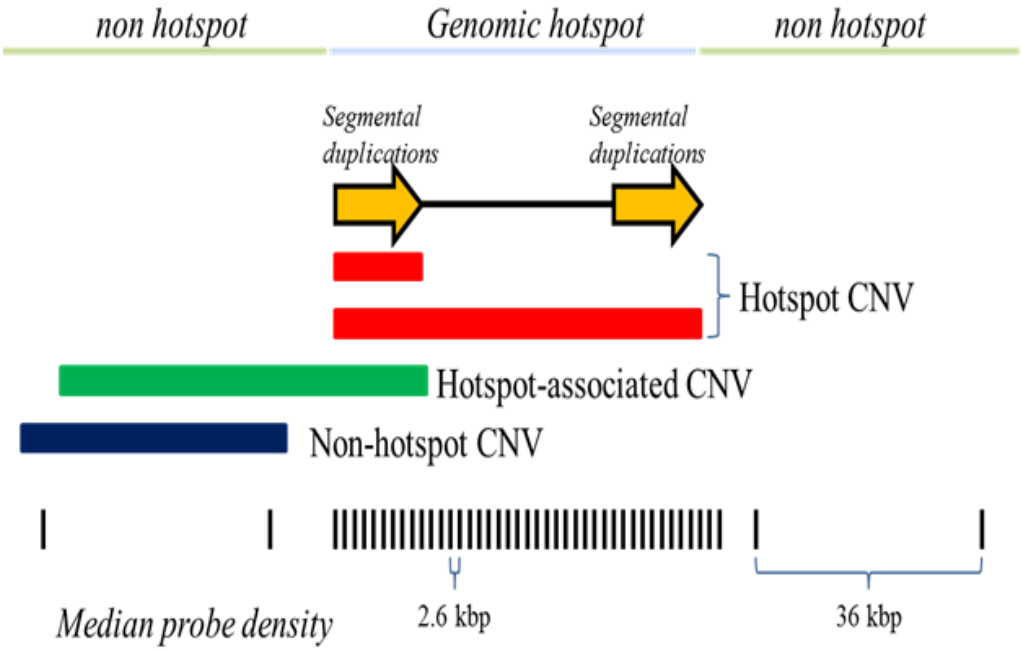
Chr	Start	End	Size	Number of probes	Probe spacing	Chromosomal segments/regions
chr1	39747785	40014937	267,152	101	2650	1q34.2
chr1	47159990	47395404	235,414	89	2650	1p33
chr1	108580322	108727851	147,529	56	2650	1p13.3
chr1	143500000	144800000	1,300,000	491	2650	1q21.1*
chr1	144900000	146437500	1,537,500	580	2650	1q21.1*
chr1	246650862	246900804	249,942	94	2650	1q44
chr2	86978679	89300000	2,321,321	876	2650	2p11.2
chr2	94700000	97580877	2,880,877	1087	2650	2q11.1q11.2
chr2	100063483	112904765	12,841,282	1000	12841	2q11.2q13
chr2	109800000	110900000	1,100,000	415	2650	2q11.2q13
chr2	130535282	131800847	1,265,565	478	2650	2q21.3
chr2	232901972	233023193	121,221	46	2650	2q37.1*
chr3	10065027	11908728	1,843,701	696	2650	3p25.3
chr3	125158409	131242152	6,083,743	1000	6084	3q21.1q21.3
chr3	196864663	198868245	2,003,582	756	2650	3q29*
chr4	3963456	9357304	5,393,848	1000	5394	4p16.2p16.1*
chr4	69713236	69923750	210,514	79	2650	4q13.2
chr4	70056451	70332530	276,079	104	2650	4q13.2
chr4	75501328	75712285	210,957	80	2650	4q13.3
chr4	119740794	120580940	840,146	317	2650	4q26
chr5	288806	1647139	1,358,333	513	2650	5p15.33
chr5	21540354	29492181	7,951,827	1000	7952	5p14.3p14.1
chr5	68886201	70696072	1,809,871	683	2650	5q13.2
chr5	98754229	99738292	984,063	371	2650	5q21.1
chr5	175417851	177414855	1,997,004	754	2650	5q35.2q35.3 (Sotos)*
chr6	167499260	167726129	226,869	86	2650	6q27
chr7	29659104	35197599	5,538,495	1000	5538	7p15.1p14.3
chr7	35951456	45755807	9,804,351	1000	9804	7p14.2p13

chr7	45798959	56428389	10,629,430	1000	10629	7p13p11.2
chr7	64160709	64898405	737,696	278	2650	7q11.21
chr7	66122765	71909367	5,786,602	1000	5787	7q11.22
chr7	72004121	76529725	4,525,604	1000	4526	7q11.23 (Williams)*
chr7	149217715	153493015	4,275,300	1000	4275	7q36.1q36.2
chr8	2167584	2331389	163,805	62	2650	8p23.3p23.2
chr8	6933974	12586975	5,653,001	1000	5653	8p23.1*
chr8	145259368	145429821	170,453	64	2650	8q24.3*
chr9	33372122	47107499	13,735,377	1000	13735	9p13.3p13.2
chr9	66194476	70221504	4,027,028	1000	4027	9q12q13
chr9	85632604	87600000	1,967,396	742	2650	9q21.32q21.33
chr9	89717505	89944931	227,426	86	2650	9q22.1
chr9	96108889	98751432	2,642,543	997	2650	9q22.32
chr10	27600000	28308559	708,559	267	2650	10p12.1
chr10	45491749	51585709	6,093,960	1000	6094	10q11.22
chr10	81129803	89119394	7,989,591	1000	7990	10q23.1*
chr11	48578987	51500000	2,921,013	1000	2921	11p11.2
chr11	54440000	55418071	978,071	369	2650	11q11
chr11	67312280	71193473	3,881,193	1000	3881	11q13.2q13.4
chr12	9327470	9492133	164,663	62	2650	12p13.31
chr12	34012279	37041888	3,029,609	1000	3030	12p11.1q12
chr13	18878590	24453912	5,575,322	1000	5575	13q12.11q12.12
chr15	19600000	20900000	1,300,000	491	2650	15q11.2 (BP1-BP2)*
chr15	20900000	26700000	5,800,000	1000	5800	15q11.2q13.1 (BP2-BP3)*
chr15	26700000	30687000	3,987,000	1000	3987	15q13.1q13.3 (BP3-BP5)*
chr15	70698860	73384192	2,685,332	1000	2685	15q24.1q24.2*
chr15	73760175	75985592	2,225,417	840	2650	15q24.2q24. 3*
chr15	80416103	83597521	3,181,418	1000	3181	15q25.2*
chr16	11926087	21500000	9,573,913	1000	9574	16p13.11p13.3*
chr16	21500000	28800000	7,300,000	1000	7300	16p12.1p11.2*
chr16	28800000	30254369	1,454,369	549	2650	16p11.2*
chr16	31868654	33678258	1,809,604	683	2650	16p11.2
chr16	68534999	73147662	4,612,663	1000	4613	16q22.1q22.3
chr17	2900904	3103469	202,565	76	2650	17p13.3 (Miller-Dieker)*
chr17	14014753	15750000	1,735,247	655	2650	17p12 (CMT1A/HNPP)*
chr17	15750000	18500000	2,750,000	1000	2750	17p11.2p12 (SMS)*
chr17	18500000	20750885	2,250,885	849	2650	17p11.2
chr17	21390699	21863694	472,995	178	2650	GAP
chr17	31430104	33585562	2,155,458	813	2650	17q12 (RCAD)*
chr17	33604213	43026435	9,422,222	1000	9422	17q21*
chr17	55005023	57730605	2,725,582	1000	2726	17q23.1q23.2
chr18	10594201	12221380	1,627,179	614	2650	18p11.2p11.21
chr19	22351817	22654242	302,425	114	2650	19p12
chr19	41455484	42488459	1,032,975	390	2650	19q13.12
chr19	53098555	55333144	2,234,589	843	2650	19q13.32q13.33
chr20	45887205	46571012	683,807	258	2650	20p13

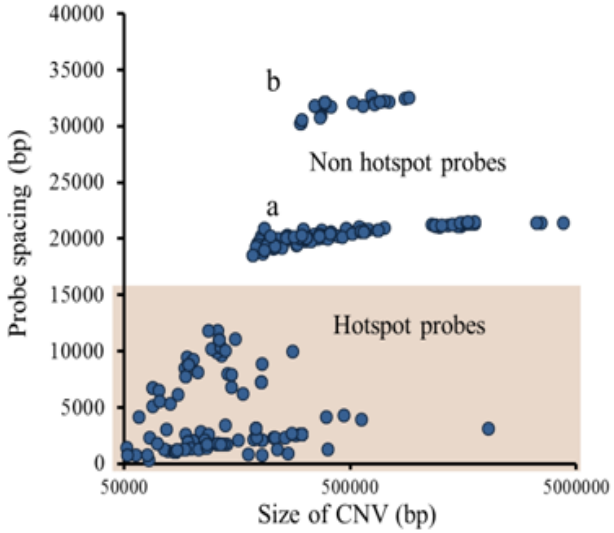
chr22	15385667	23387607	8,001,940	1000	8002	22q11.2q11.23*
chr22	45200000	49500000	4,300,000	1000	4300	22q13*
chrX	36896326	37339751	443,425	167	2650	Xp12.1
chrX	48035332	52772103	4,736,771	1000	4737	Xp11.23p11.22
chrX	57617021	57972819	355,798	134	2650	Xp11.1
chrX	148421847	148866679	444,832	168	2650	Xq28
chrX	152024335	152216219	191,884	72	2650	Xq28
chrX	154219195	154387405	168,210	63	2650	Xq28
chrY	6167831	10362225	4,194,394	1000	4194	Yp11.2
chrY	22077505	23217056	1,139,551	430	2650	Yq11.23
chrY	23566258	26800000	3,233,742	1000	3234	Yq11.23q11.2

*Regions associated with known genomic disorders.

Figure 2. A) The design of the HS1 array: The array has a median probe spacing of 2.6 kb in the genomic hotspot regions (regions flanked by SD), and a median probe spacing of 36 kb in other non-hotspot genomic backbone regions. Hotspot CNV: CNVs that are detected within the genomic hotspot region using the HS1 array; hotspot-associated CNV: part of the CNV detected is located in the genomic hotspot; non-hotspot CNV: CNVs detected in the genomic backbone, outside the genomic hotspot regions. B) Scatter plot of probes distribution: This scatter plot shows the size-distribution of CNVs and the density of array probes targeted to the genomic hotspots and non-hotspot regions (see Figure 2A). Note that non-hotspot genomic backbone regions contain two different probe densities (probe spacing of 20000 bp and 30000 bp, which are labeled (a) and (b)) and the 107 hotspot regions are covered at a probe density ranging from 1000 bp up to 10,000 bp. (Figures courtesy of Dr. Girirajan).


non hotspot
Genomic hotspot
non hotspot
Segmental duplications
Segmental duplications
Hotspot CNV
Hotspot-associated CNV
Non-hotspot CNV
Median probe density
2.6 kbp
36 kbp

A)


40000
35000
30000
25000
20000
15000
10000
5000
0
Probe spacing (bp)
b
Non hotspot probes
a
Hotspot probes
50000
500000
5000000
Size of CNV (bp)

B)

Samples labeling, microarray hybridization and scanning

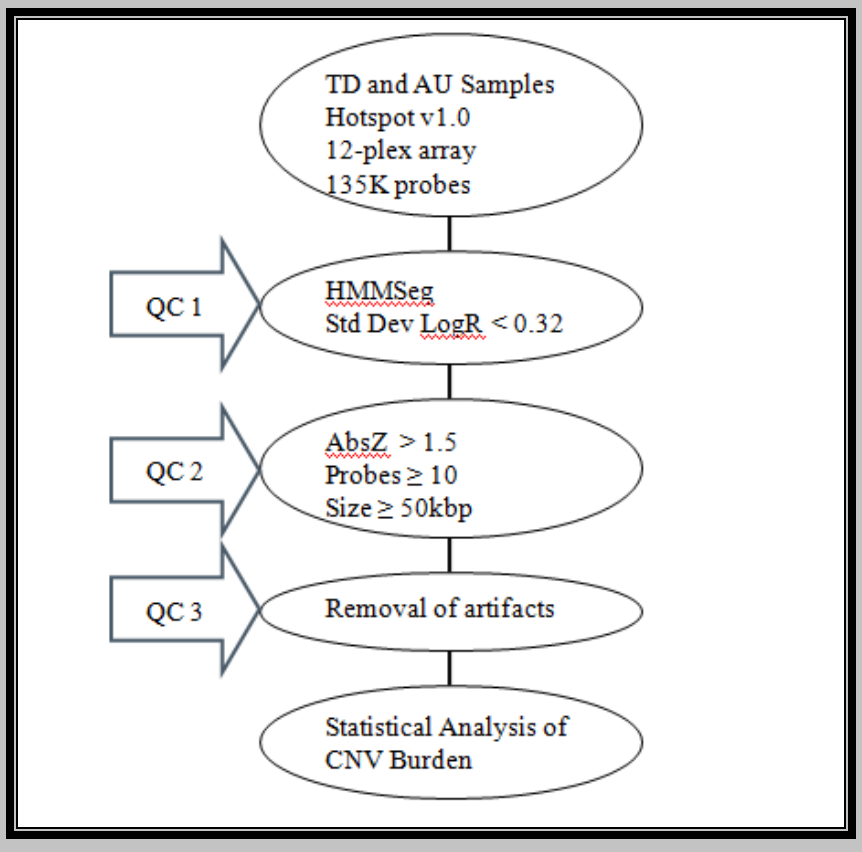
Labeling and microarray hybridization of DNA samples were performed using the Nimblegen array CGH kit (Roche Nimblegen Systems) according to the manufacturer's instructions, and a previously described protocol [24]. All aCGH experiments were performed using a reference DNA sample prepared from whole blood of a single unaffected Caucasian male (Coriell, NA15724). DNA test samples, including both AU and TD individuals, were labeled with Cy3 fluorescent dye, while the reference sample was labeled with Cy5 fluorescent dye. Each of the labeled test samples was mixed with the reference sample and hybridized to the custom-designed HS1 12-plex 135K array using a 12 Bay Hybridization System (Roche Nimblegen) according to manufacturer's recommendations. After hybridization and washing, the arrays were scanned using GenePix 4000B Microarray Scanner (Molecular Devices,) according to the manufacturer's instructions. Whereas for the 2×400K Agilent chip array, hybridization experiments were performed as described previously [10] and according to the manufacturer's recommendations.

Array analysis and CNV callings

Data produced from scanned arrays were extracted, processed and normalized using the NimbleScan software package (Roche Nimblegen). Utilizing the embedded SegMNT algorithm within the software, segments were assigned to the normalized data and log intensity ratio were produced. Log intensity ratios were transformed into z-scores using chromosome-specific means and standard deviations. These z-scores were classified as normal, increased, or decreased copy numbers with the help of HMMSeg algorithm based on a three-state Hidden Markov Model (HMM) [25]. For each sample data, probes were combined into segments if consecutive probes of the same state were less than 50 kb apart; for instance, two probes of the same state (such as

increased in copy number) were merged into a single segment if they were not more than 50 kb from each other. In addition to that, two segments of the same state and the sequences between them would also be called as a single CNV as long as the sequence in between the segments has less than 5 probes and is smaller than 10 kb. For the post-HMM filtering thresholds, we employed stringent quality control (QC) measures to increase the accuracy of our experimental data by reducing false positives. For the first part of the QC measure, which was applied to ensure reliability, we eliminated individual samples that had a standard deviation or Log R ratio larger than 0.32. These samples were excluded because a large standard deviation was typically associated with the noisiness and inconsistency of the data. Samples that passed the first QC were then subjected to the second part of the QC to eliminate false positive CNV calls in each sample. To be considered a real variant, a CNV had to have an absolute z-score larger than 1.5, and contain 10 or more probes, which are empirically estimated thresholds. Since our array had different probe densities in the hotspot and non-hotspot regions, we separated CNVs into three different groups: hotspot CNV, hotspot-associated CNV, and non-hotspot CNV (Figure 2). Hence, we use different threshold of minimum length or size in calling those CNVs. The minimum length for calling the hotspot CNV and hotspot-associated CNV was 20kb; whereas for calling non-hotspot CNV, the CNV length had to be at least 50kb long. In addition to that, a number of artifacts were also removed from the called CNVs; CNVs were considered artifacts if they were located at the centromere/telomere regions, or were not reliable due to the noisiness of the data for these CNV regions. Through the use of these QC criteria, the HMM data outputs could be filtered and scanned for real variants, which were then collected and subjected to further statistical analysis for CNV burden. Figure 3 summarizes the overview of our research pipeline, from the array preparation and hybridization to statistical analysis.

Figure 3. Schematic overview of our research pipeline. 1) All samples were subjected to HS1 array hybridization, scanning and analysis, producing normalized log intensity ratios as well as z-scores as output. 2) These z-scores were then analyzed by HMMSeg to produce CNV calls as output. 3) QC 1: samples with standard deviation larger than 0.32 were eliminated. 4) QC 2: real variants are required to be equal to or larger than 50kb for the non-hotspot CNV, and 20kb or larger for the hotspot and hotspot-associated CNV. They are also required to contain 10 or more probes, and have an absolute z-score of larger than 1.5. QC 3: CNV calls that are artifacts were removed based on their characteristics (such as location near telomeres or centromeres).



Calling for rare, pathogenic CNVs

The detection and calling for rare pathogenic variants were mostly performed using manual curation through utilizing the functionality embedded in the UCSC genome browser. The criteria for calling a rare and potentially pathogenic event were based on the frequency of that particular CNV. In other words, to be considered a rare pathogenic variant, the putative CNV has never been seen in controls or is found in low frequency (10 or less out of 8329 controls), and affects chromosomal regions harboring related functional genes that potentially contribute to the pathogenesis of the disorder. Most of the time, these pathogenic variants had already been reported in many other previous studies. The detected putative pathogenic CNVs were validated using a different platform with relatively much higher probe densities in the array.

Statistical analysis of CNV burdens

The comparison of CNV burden between AU and TD individuals were done mainly by using two approaches: 1) Comparison of a total average length of CNVs per individual, including duplications and deletions, in autistic and control groups; 2) Comparison of total numbers of CNV per individual, in autistic and control group. This analysis was performed separately for duplications and deletions. CNV length burden was normalized on the total number of kids in each group.

We also employed two different approaches to comparing the CNV burdens between AU and TD individuals to investigate whether genomic instability was more pronounced at the hotspot regions or the genomic backbone. First, we performed a global CNV burden analysis in the entire genome, which included the high-density probe regions (the SD or hotspot regions) as well as the low-density probe regions (other unique genomic or non-hotspot regions). Second,

we performed a CNV burden comparison between autistic and controls on the hotspot, hotspot-associated, and non-hotspot regions separately. In addition to analysis of CNV burden on hotspot and non-hotspot regions of the genome, we have also performed copy number burden analysis on both sample cohorts after removing rare, potentially pathogenic CNV calls as detected previously from the HMM calls.

Significances of all the differences were tested statistically using Mann-Whitney U-Test. We were not able to use the T-test due to the non-Gaussian distribution characteristic of the data sets. A maximum p-value threshold of 0.05 and differential threshold of 5% were expected for the differences to be statistically significant.

RESULTS

Qualification control of samples for CNV callings

Through utilization of QC parameters mentioned above for calling CNVs in each individual, samples that did not meet the qualification criteria for reliable data, that is, the standard deviation was larger than 0.32, were eliminated. Out of the 553 analyzed samples, 54 did not pass QC. Table 3 below summarizes the ethnicity breakdown of the 499 samples that passed QC.

Table 3. Summary of the number of samples that passed QC (having a standard deviation or Log R ratio < 0.32). Out of the initial number of 553 samples, 54 did not pass QC.

Ethnicity	TD (# of samples not passing QC)	AU (# of samples not passing QC)	**Total**
Hispanic	81 (2)	57 (17)	**138**
White	134 (13)	133 (10)	**267**
Asian	7	20	**27**
African American	4 (1)	6	**10**
Mixed	30 (7)	27 (3)	**57**
Native Hawaiian	0 (1)	-	**0**
Total	**256**	**243**	**499**

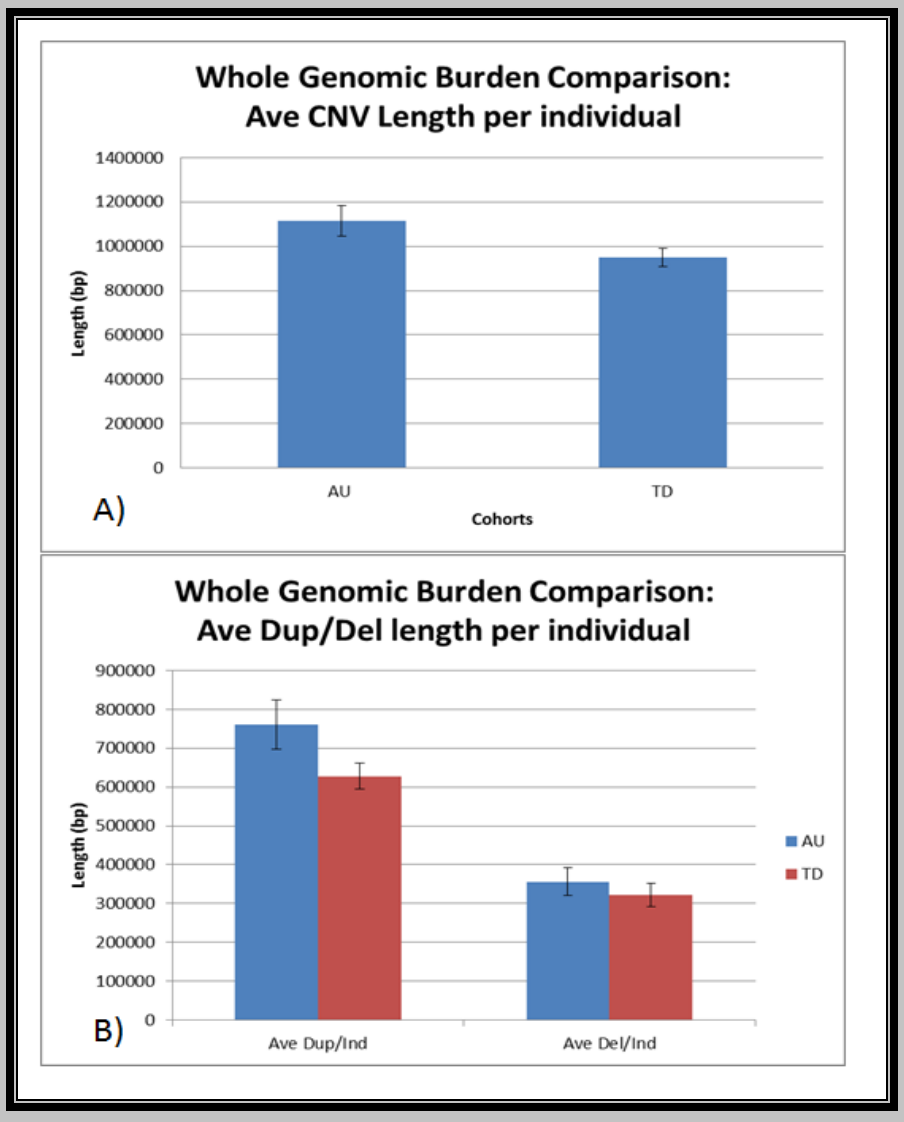
Elevated level of global CNV burden in autistic individuals

Analysis of the CNV burden in the entire genome, including hot spot and non-hot spot regions, showed that the total average CNV burden length per individual was 1116235bp in AU patients (total average duplication length: 760660 bp; total average deletion length: 355575bp) and 950038bp in TD controls (total average duplication length: 627708bp; total average deletion length: 322329bp) (Table 4). As shown in Figure 4A, the AU patients group showed a significantly increased level of global CNV burden compared to that of TD (Mann-Whitney Test, U= 23449, p= 0.00185). The total average duplication length in AU individuals was also significantly higher (Mann-Whitney Test, U= 28129, p= 0.0109) (Figure 4B). On the other hand, there was no significant difference in the average total length of deletions between AU and TD individuals (Mann-Whitney Test, U= 20804, p=0.432) (Figure 4B).

Table 4. CNV burden lengths (duplications and deletions) in each AU and TD cohort

	AU individuals	TD individuals
Average Duplication Length per individual (bp)	760660	627709
Average Deletion Length per individual (bp)	355575	322329
Total CNV Burden Length per individual (bp)	**1116235**	**950038**

Figure 4. CNV burden comparison in whole genome regions. A) The AU patients group showed a significantly increased level of global CNV burden compared to that of TD (Mann-Whitney Test, U= 23,449, p= 0.00185). B) The total average duplication length in AU individuals was significantly higher (Mann-Whitney Test, U= 28,129, p= 0.0109); whereas there was no significant difference in total average deletion length between AU individuals and TD individuals (Mann-Whitney Test, U= 20,804, p=0.432).



Elevated level of CNV burden was largely found in genomic backbone

As mentioned earlier, we also performed analysis on the genomic backbone region separately to compare the non-hotspot CNV burden level in AU and TD individuals. The burden comparisons on the genomic backbone shows that the total average CNV burden length per individual was 692,446 bp in AU patients (total average duplication length: 489,378 bp; total average deletion length: 203,068 bp) and 542,985 bp in TD controls (total average duplication length: 372,248 bp; total average deletion length: 170,737 bp). As shown in Figure 5A, the AU patients group showed a significantly increased level of CNV burden compared to that of TD (Mann-Whitney Test, U=7,992, p=0.0128). Similar to the data obtained from the whole genome region analysis, the total average duplication length in AU individuals was significantly higher (Mann-Whitney Test, U=19,986, p=0.0209) (Figure 5B), but there was no significant difference in the comparison of average total length of deletions between AU and TD individual (Mann-Whitney Test, U=2,920.5, p=0.483) (Figure 5B). On the other hand, CNV burden analysis on the hotspot CNVs revealed no significant difference between AU individuals and TD individuals, in terms of total average duplication size (Figure 6B) (Mann-Whitney Test, U=24,022, p=0.944), total average deletion size (Figure 5B) (Mann-Whitney Test, U=24,049, p=0.575), as well as total CNV burden size (Figure 6A) (Mann-Whitney Test, U=28,987, p=0.513). Similar results were also obtained when burden comparison was estimated combining the hotspot CNV and hotspot-associated CNV: there was no statistically significant difference in CNV burden level of AU and TD individuals. The total average duplication size (Figure 7B) (Mann-Whitney Test, U=24,334, p=0.931), total average deletion size (Figure 7B) (Mann-Whitney Test, U=23,045, p=0.830), and total CNV burden size (Figure 7A) (Mann-Whitney Test, U=28,662, p=0.719) were comparable for autistic and control groups.

Figure 5. CNV burden comparison in non-hotspot genomic regions. A) Total average length of CNV burden comparison between AU and TD individuals. The AU patients group showed a significantly increased level of CNV burden compared to that of TD (Mann-Whitney Test, U= ?, p=0.0128). B) Total average duplication and deletion lengths comparison between AU and TD individuals.

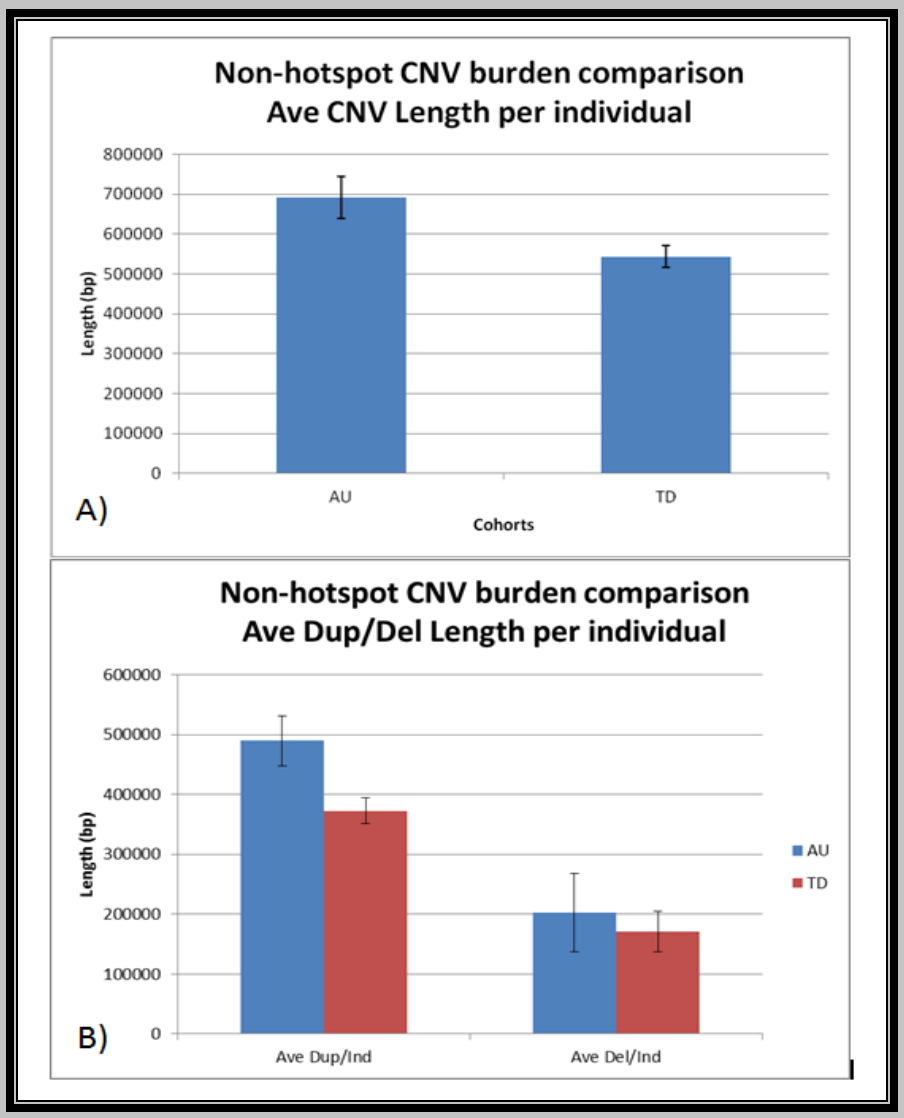


Figure 6. CNV burden comparison in genomic hotspot regions. A) Total average length of CNV burden comparison between AU and TD individuals. B) Total average duplication and deletion lengths comparison between AU and TD individuals.

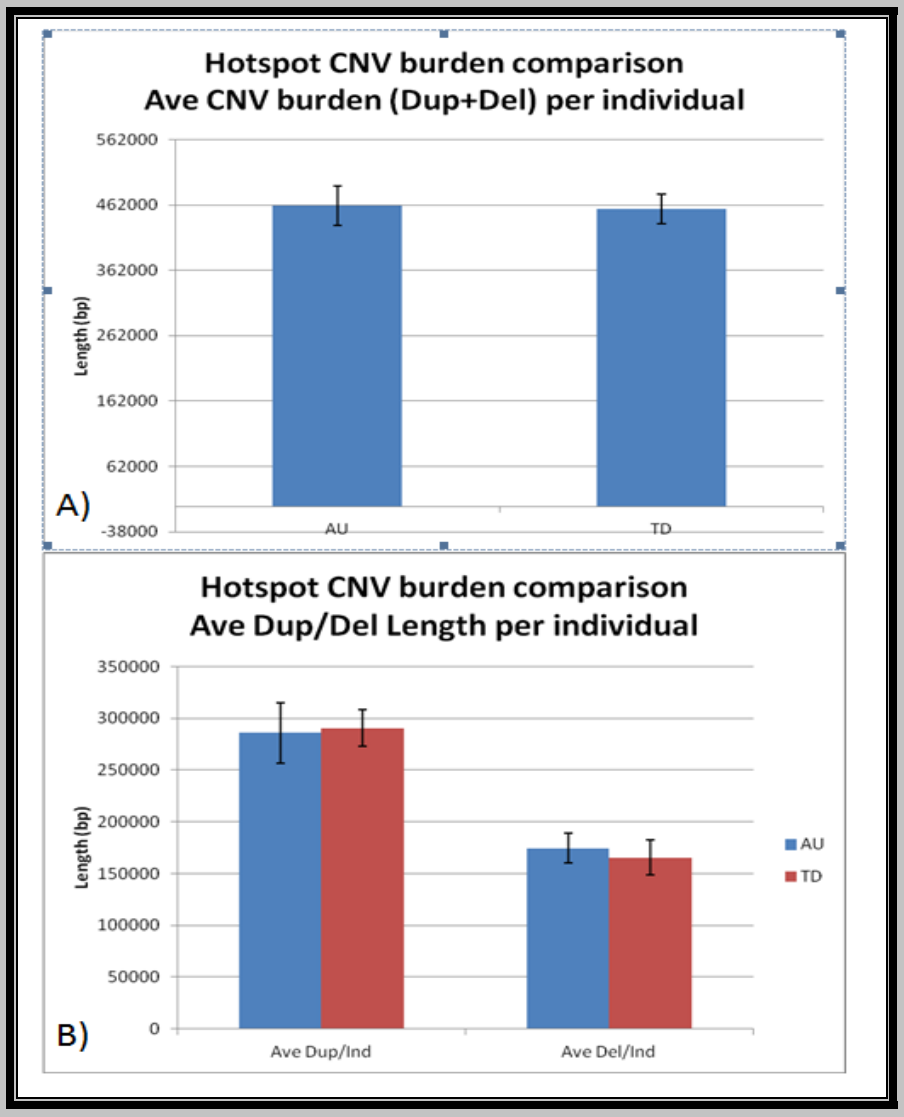
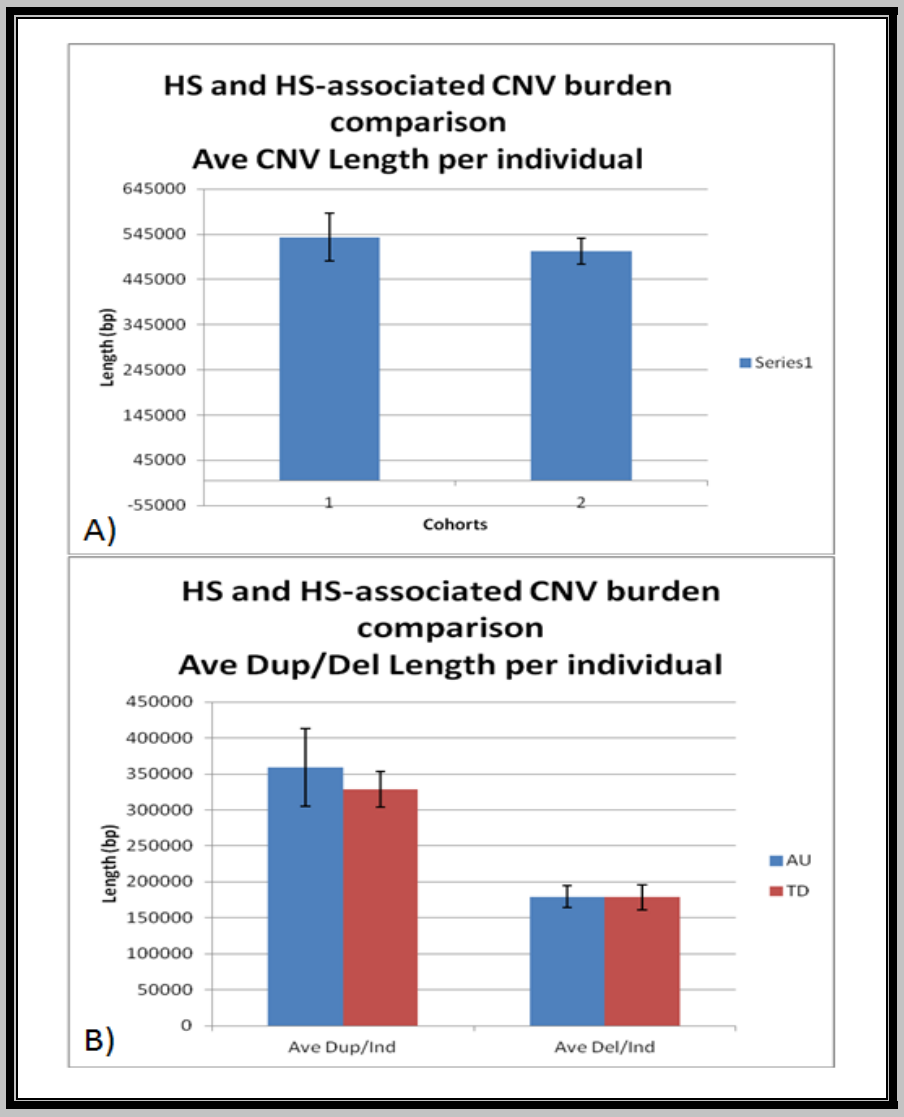


Figure 7. CNV burden comparison in genomic hotspot and hotspot-associated regions A) Total average length of CNV burden comparison between AU and TD individuals. B) Total average duplication and deletion lengths comparison between AU and TD individuals.



Elevated level of non-rare and nonpathogenic copy number burden

In order to investigate whether the elevated level of copy number burden found in AU individuals was largely due to rare, potentially pathogenic events, we have performed analysis and calculation on the CNV burden after removing and eliminating those functional genic variants (Table 5) from the event calls. This burden comparison was performed on the entire genome regions (including hotspot and non-hotspot regions). Our statistical analysis showed a similar trend to the previously obtained results: the AU cohort showed a significantly increased level of CNV burden compared to that of TD (Mann-Whitney Test, U= 24232.000, p=0.008) (Figure 8A). Similarly, the total average duplication length in AU individuals was significantly higher (Mann-Whitney Test, U= 21,493, p= 0.022) (Figure 8B), but there was no significant difference in the comparison of average total length of deletions between AU and TD individual (Mann-Whitney Test, U= 20,819.5, p= 0.477) (Figure 8B).

Figure 8. Non-pathogenic CNV burden comparisons for the whole genome regions after rare and potentially pathogenic events were removed. A) Total average length of CNV burden comparison between AU and TD individuals. B) Total average duplication and deletion lengths comparison between AU and TD individuals.

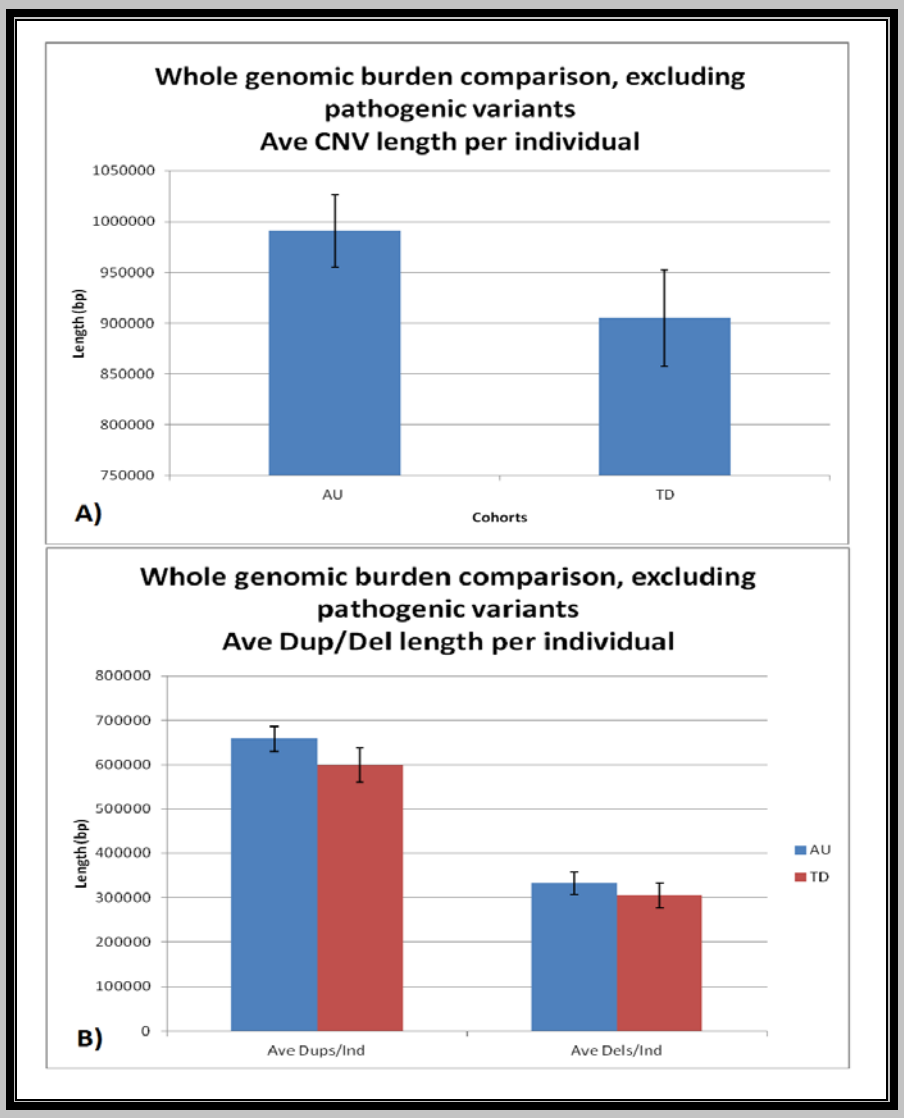


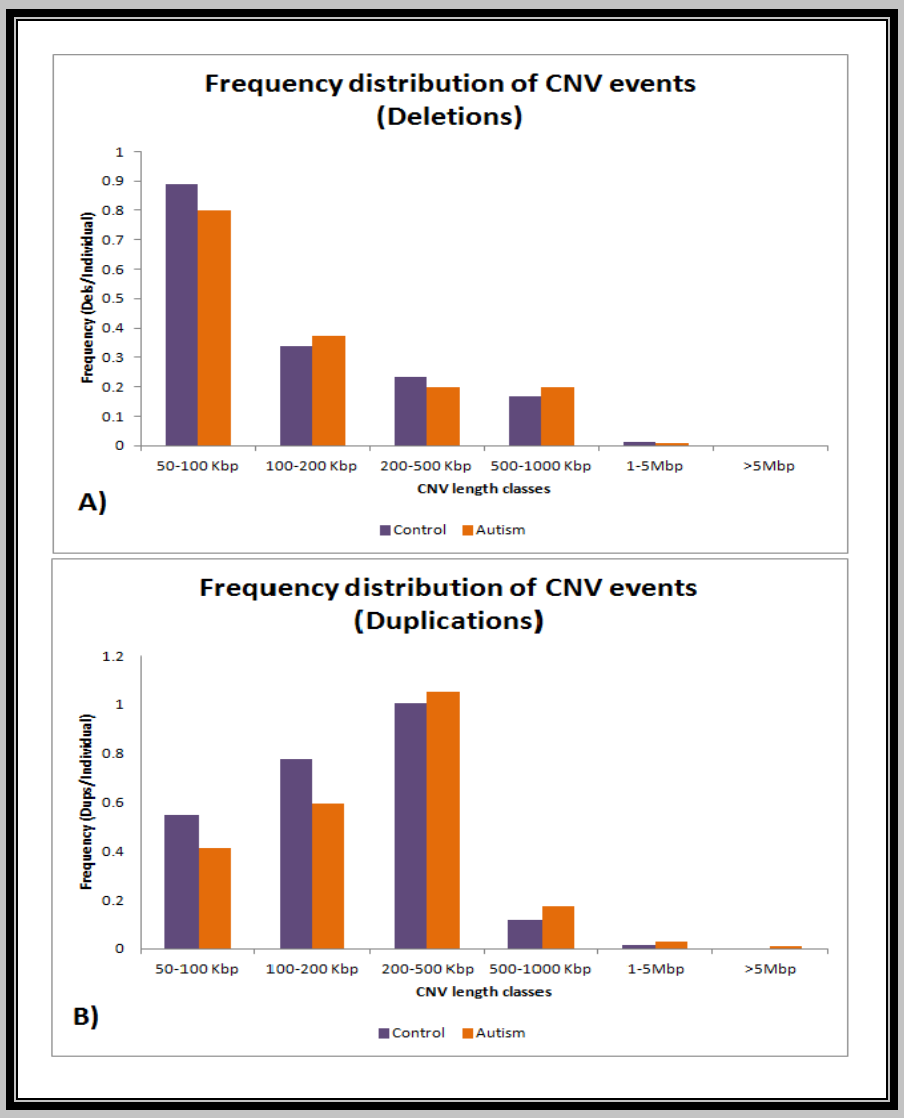
Table 5. A list of rare, potentially pathogenic events that were removed from the HMM calls for non-pathogenic CNV burden analysis. These events were found in 10 or less of 8329 controls. Some of these variants have been previously described.

Chr	Section	CNV	Start	End	Size	Samples ID	Controls
chr4	4q32.3	Deletion	166240300	169876274	3635974	458-04-101669	0/8329
chr15	15q11.2q12	Duplication	18259360	26226639	7967279	193-05-103869	0/8329
chr15	15q11.2q12	Duplication	20812153	26225840	5413687	172-09-111199	0/8329
chr15	15q11.2q12	Duplication	21226843	26242596	5015753	609-05-104617	0/8329
chr15	15q13.3	Duplication	28913092	30246254	1333162	091-06-105625	3/8329
chr15	15q13.3	Duplication	29224635	29439783	215148	156-07-108080	0/8329
chr7	7q11.23	Deletion	72238679	73828887	1590208	393-08-110603	0/8329
chr17	17q12	Duplication	32539340	32627777	88437	082-03-100466	0/8329
chr16	16p11.2	Deletion	28521912	29253758	731846	157-09-111165	0/8329
chr16	16p11.2	Duplication	28521912	29253758	731846	694-05-105404	0/8329
chr16	16p11.2	Duplication	28624768	29080269	455501	082-03-100466	0/8329
chr1	1q21.1	Deletion	143665017	144529218	864201	006-07-107598	0/8329
chr17	17p11.2	Deletion	19920507	19937782	17275	089-05-102822	0/8329
chr17	17p11.2	Deletion	19437254	19476889	39635	065-03-100309	10/8329
chr4	4q13.1	Duplication	62808305	64394393	1586088	381-08-110540	0/8329
chr7	7p12.3	Duplication	47247609	47761896	514287	703-05-105431	0/8329
chr10	10q11.23	Deletion	50094947	50388501	293554	087-05-102962	0/8329
chr7	7q11.22	Deletion	68528129	69368926	840797	521-04-101908	0/8329
chr6	6q22.31	Duplication	120625209	122799034	2173825	298-06-106231	0/8329
chr6	6q23.2	Deletion	131677976	132324687	646711	033-06-105584	2/8329
chr6	6q26	Duplication	162565112	163564122	999010	326-07-108609	0/8329
chr12	12p11.1	Duplication	33372513	34695616	1323103	129-07-107979	10/8329
chr22	22q11.22	Duplication	20175001	22125000	1949999	623-05-104660	0/8329
chr15	15q13.1	Duplication	26692914	28176512	1483598	558-09-112365	1/8329
chr17	17q12	Deletion	31872319	33329702	1457383	011-09-110676	2/8329
chr22	22q11.21	Duplication	19034683	19813839	779156	369-09-111849	6/8329

Large events constitute the elevated level of duplication

When we separated each of the CNVs into several different classes based on their lengths to assess the frequency distribution of the CNV events in each cohort, interestingly, we found that the elevated level burden of duplications we observed earlier in the autistic individuals was mostly due to the large size events, with lengths of CNVs more than 200 kb (Figure 9B). This suggested that the elevated level of duplications in AU individuals compared to that of controls was principally contributed by large duplicated segments of CNVs. On the other hand, there was no such observable trend in the deletion events, which corresponded to our expectations since we found no significant difference in the burden level of deletion between cases and controls. The test for the significance of the data was not performed since the frequency plot was graphed mainly for the purpose of observing the trends on what types of CNVs (in terms of size) contributed to the elevated level of burden we observed earlier.

Figure 9. Frequency distribution plots of CNV events in both cohorts, separated into different classes based on sizes. A) Deletions: There is no observable trend in the distribution of deletion events in the comparison between AU and TD individuals. B) Duplications: The duplications events found in AU individuals mostly fall into the large CNV classes (≥ 200 kbp), leading to the elevated level of duplication we observed earlier.



DISCUSSION

In the design of our Hotspot v1.0 array, the segmental duplication architecture of the human genome [26] was utilized to custom-design a DNA oligonucleotide microarray enriched for genomic hotspots, which are regions surrounded by high-identity level segmental duplications. Because the hotspot regions were the main target with higher probe density designed to cover those regions, this array actually had more powerful detection sensitivity within hotspot and hotspot-associated regions compared to several other available commercial arrays. In fact, recurrent events detected in the genomic hotspots were twenty-five times more frequent comparing to the rest of the genomic backbone regions [27].

Many studies have been focused on discovering the large, rare pathogenic CNVs that lead to a variety of neurodevelopmental diseases such as autism, mental retardation, schizophrenia, intellectual disability as well as Prader-Willi/Angelman syndrome, in which those structural variations often involved copy number changes in genomic segments that harbor large functional cluster of genes, leading to the pathogenesis of these diseases [28, 29]. However, some other studies focused on a disease association with copy number polymorphism, which included the events that were much more common in the population than rare pathogenic CNVs [30]. However, very few of these studies have actually evaluated the relationship between global CNV burden (not limited to rare, genic CNVs) and pathogenesis of a particular disease. Therefore, the question whether there is a higher level of genomic instability in autism has been left unanswered for decades. In this study, we were able to at least partially address that question by demonstrating the increased level of genomic instability in autistic individuals.

Autistic individuals have higher level of global CNV burden

Our array and statistical results showed that there was a significantly higher level of overall global CNV burden in autistic cases compared to that of the controls. This result is in agreement with a previous study where the authors reported an increased level of copy number burden in AU individuals using the whole genome studies focused on large *de novo* events [31]. Another study of genome-wide assessment via single-nucleotide polymorphism microarrays and karyotyping has also found that structural variants or unbalanced CNVs were present in relatively higher frequency in individuals with autism spectrum disorders (ASD) [34]. In a more recently published work, an elevated level of copy number burden was detected for rare pathogenic variants primarily associated with genes previously implicated in ASD or intellectual disability (ID) [32]. In addition to that, numerous studies have also discovered high level of genomic imbalance in other diseases or psychiatric disorders such as schizophrenia [35-37], bipolar disorder [38], epilepsy [39,40], as well as mental retardation [41] and developmental delay/intellectual disability [42,43]. However, these studies only focused on examining the burden level of *de novo* events of pathogenic or functional CNVs [33], but did not include those common CNVs or events that are not pathogenic (not harboring genes that might influence disease susceptibility), and have not investigated whether the increased instability occurred in hotspot or non-hotspot genomic regions. Nevertheless, these findings are in agreement with our principal result that there is indeed an increased level of copy number burden in autistic individuals compared to unaffected controls. The increase of global CNV burden in autism demonstrates the important role of genomic structural variants in autism.

Increased level of CNV burden in autism is manifested by large duplications

We have also shown that the elevated level of CNV burden was mainly represented by large duplications ranging from 200kb to 5Mb based on the frequency distribution of CNV in different size classes. This is in agreement with many other studies that have reported large duplicated CNVs such as those located in chromosomal segments associated with autism: 7q11.23 [60], 15q11.2-13.1 [61,62], as well as 16p11.2 [63,64]. We speculate that from an evolutionary point of view a large duplication of a chromosomal segment was relatively more tolerable compared to a deletion in the human genome, since deletions totally disrupt and eliminate some functional genes that might be critical and essential for the survival of the organism. Hence, large duplications in the genome provide a higher chance for an organism to survive, but at the same time, could affect some functional genes' dosages leading to certain observable disease phenotype such as autism.

Remarkably, the increased level of CNV burden in AU individuals was still statistically significant even after the rare, pathogenic CNVs were excluded from our analysis. This indicates that the increased genomic instability we have observed in affected individuals is not exclusively caused by the rare, pathogenic CNVs. This could serve as a starting point for a future research to confirm that genomic instability and copy number burden in certain disorders, at least in autism, may not be limited to the rare and large genic events.

Copy number burden is elevated in non-hotspot genomic regions

When we performed the CNV burden level analysis separately on the genomic hotspot and non-hotspot regions, unexpectedly, we found that the elevated level of copy number burden was mainly localized in the non-hotspot regions. This finding question the current paradigm that assumes that majority of the specific genetic diseases are caused by homologous recombination and recurrent chromosomal aberrations, influenced by the presence of segmental duplications [48-50]. Hence, it is certainly reasonable to think that mechanisms other than non-allelic homologous recombination (NAHR) could be involved in the elevated level of non-hotspot CNV burden we observed in the AU individuals. One of the possible factors that contributes to genomic instability may be the presence of Alu elements, which were considered crucial for chromosomal rearrangements and aberrations in the human genome in the last two decades [51,52], and have been thought to mediate genomic instability in diseases such as breast cancer [53,54], colon cancer [55], leukemia [56], and many other human diseases [57]. In fact, a recently published study [58] has shown that a disruption of an important neuronal gene associated with ASD, *CNTN4* [58,59], was a result of Alu Y-mediated unequal recombination. It would be certainly reasonable to consider other mechanisms not mediated by segmental duplications, including microhomology-mediated break induced replication that could potentially lead to a higher genomic instability and an increased level of CNV burden reported by this and other studies.

Genomic instability, copy number variations and autism

One interesting implication of our result is that an increased level of copy number changes could result from a large number and variety of genetic alterations, only part of which may be involved in pathogenesis of the disorder. It is possible that some specific external or internal factors may exist that promotes this higher overall instability. For instance, environmental factors that affect genomic stability could play a role, contributing to the pathogenesis of autism. The idea of connection between environmental factors and genomic instability is certainly not new; many studies reported the defined and clear relationship between these two factors for many other diseases including cancer [44] and many complex genetic disorders [45,46]. Hence, it would certainly be interesting to investigate how various environmental factors affect the genome, which in turn may lead to genomic structural variants and ultimately the observable disease phenotype.

As previously mentioned in many other studies and literature reviews, even though most common CNVs do not necessarily have a negative effect on health (normally, CNVs cover approximately 12% of the human genome [48]), when their amount reaches a certain threshold, they could have a number of important implications in the causal mechanism of genetic disorders.

The association of neurological disorders such as autism with the increased CNV burden suggests an important role of the non-Mendelian inheritance in the emergence of the disease. An individual's genetic code may not be simply the addition of the genetic contributions from both of the individual’s parents. The mechanisms such as unequal crossover events, which occur during the production of sperm and eggs, are responsible for formation of novel CNVs in the

genome. As a result, the progeny may lose or gain additional copies of genetic information that were not present in either of their parents' genetic code. For many years, geneticists believed that a disease is the result of inherited genetic variants from the previous generations, where a mal-functional copy of a genomic segment was passed on to the next generation. However, as more and more studies started to discover and recognize other mechanisms that can lead to genetic disorders, we are in an era where we begin to understand and appreciate that structural variants may provide the genetic basis for increasing the risk of common and complex diseases such as mental retardation, schizophrenia, and autism.

Limitations of the Hotspot v1.0 array

The Hotspot v1.0 array certainly has its limitations in terms of detecting CNVs in the human genome, due to the differential probe density across the genomic hotspot and non-hotspot regions. As mentioned earlier, the array was originally designed to target genomic hotspot regions flanked by segmental duplications by increased density of probes in these regions (median probe spacing 2.6 kb) compared to the genomic backbone. Consequently, the array design was biased towards targeting hotspot regions in the genome, and has its limitations in detecting CNVs that are located in the non-hotspot regions. Specifically, it was not possible to locate smaller and shorter CNVs due to the low density of probes in those regions (the minimum probe spacing about 20 kb), and the requirement for a CNV to contain at least 10 probes. Hence, the non-hotspot CNVs that were smaller than 200 kb completely missed our detection in the non-hotspot regions. In other words, we would inadvertently overlook those smaller and intragenic as well as exonic CNVs in the genomic backbone that might also influence the expression of other autism candidate genes.

Future directions and implications of this study

Due to the current limitations of our microarray and experimental design, there are many other aspects and areas that we need to consider and evaluate before we could have a better insight in the underlying causal mechanism of genomic instability and its association with autism. To address the limiting resolution power of our Hotspot v1.0 array, it would certainly be plausible to design a finely-tiled array with a much higher density of probes allocated across the human genome to detect the presence of any CNVs of any sizes as completely and accurately as possible. On the other hand, next-generation sequencing technology or whole-genome sequencing [65] is an extremely promising methodology to characterize CNV under maximal resolution across the genome, which could allow us to detect previously unreported variants. We are also very interested in analyzing the relationship between the level of copy number burden and disease phenotypes, as well as its interaction with the environmental factors. For example, is there a clear relationship between CNV burden and severity of the disease? Do environmental factors have a role in affecting the level of copy number burden? These questions are well worth for further investigation for our study in the future since there are available phenotypic and environmental data for each of the 553 individual kids we have analyzed.

Through the use of our custom-design microarray platform, we were able to show that rare, pathogenic events are not solely responsible for the pathogenesis of autism; global genomic instability, and possibly its interaction with environmental factors, plays an important role in this neurodegenerative disorder. This sparks our thoughts that we should focus not only on looking for specific genes that cause a disorder, because many of them especially complex diseases do not follow the conventional Mendelian genetics.

REFERENCE:

1. Merikangas AK, Corvin AP, Gallagher L. (2009). Copy-number variants in neurodevelopmental disorders: promises and challenges. Trends Genet 25(12):536-44.

2. Sharp AJ, Hansen S, Selzer RR, Cheng Z, Regan R, et al. (2006). Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. Nat Genet 38: 1038-1042.

3. de Vries BB, Pfundt R, Leisink M, Koolen DA, Vissers LE, et al. (2005). Diagnostic genome profiling in mental retardation. Am J Hum Genet 77: 606-616.

4. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, et al. (2007). Strong association of de novo copy number mutations with autism. Science 316: 445-449.

5. Marshall CR, Noor A, Vincent JB, Lionel AC, Feuk L, et al. (2008). Structural variation of chromosomes in autism spectrum disorder. Am J Hum Genet 82: 477-488.

6. Grozeva D, Kirov G, Ivanov D, Jones IR, Jones L, et al. (2010). Rare copy number variants: a point of rarity in genetic risk for bipolar disorder and schizophrenia. Arch Gen Psychiatry 67: 318-327.

7. Helbig I, Mefford HC, Sharp AJ, Guipponi M, Fichera M, et al. (2009). 15q13.3 microdeletions increase risk of idiopathic generalized epilepsy. Nat Genet 41: 160-162.

8. Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, et al. (2008) Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science 320: 539-543.

9. Elia J, Gai X, Xie HM, Perin JC, Geiger E, et al. (2010). Rare structural variants found in attention-deficit hyperactivity disorder are preferentially associated with neurodevelopmental genes. Mol Psychiatry 15: 637-646.

10. Girirajan S, Brkanac Z, Coe BP, Baker C, Vives L, et al. (2011). Relative burden of large CNVs on a range of neurodevelopmental phenotypes. PLoS Genet 7(11):e1002334.

11. Yoon S, Xuan Z, Makarov V, Ye K, Sebat J. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res 19(9):1586-92.

12. Veltman MW, Craig EE, Bolton PF. (2005). Autism spectrum disorders in Prader-Willi and Angelman syndromes: a systematic review. Psychiatr Genet 15(4):243-54.

13. Urbán Z, Helms C, Fekete G, Csiszár K, Bonnet D, et al. (1996). 7q11.23 deletions in Williams syndrome arise as a consequence of unequal meiotic crossover. Am J Hum Genet 59(4): 958–962.

14. Ewart AK, Morris CA, Atkinson D, Jin W, Sternes K, et al. (1993). Hemizygosity at the elastin locus in a developmental disorder, Williams syndrome. Nat Genet 5(1): 11–16.

15. Osborne LR. (1999). Williams-Beuren syndrome: unraveling the mysteries of a microdeletiondisorder. Mol Genet Metab 67(1): 1–10.

16. Depienne C, Heron D, Betancur C, Benyahia B, Trouillard O, et al. (2007). Autism, language delay and mental retardation in a patient with 7q11 duplication. J Med Genet 44(7):452-8.

17. Berg JS, Brunetti-Pierri N, Peters SU, Kang SH, Fong CT, et al. (2007). Speech delay and autism spectrum behaviors are frequently associated with duplication of the 7q11.23 Williams-Beuren syndrome region. Genet Med 9(7):427-41.

18. Girirajan S, Rosenfeld JA, Cooper GM, Antonacci F, Siswara P, et al. (2010). A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. Nat Genet 42(3):203-9.

19. Nagamani SC, Erez A, Shen J, Li C, Roeder E, et al. (2010) Clinical spectrum associated with recurrent genomic rearrangements in chromosome17q12. Eur. J. Hum. Genet 18:278-84.

20. Moreno-De-Luca D, SGENE Consortium, Mulle JG, Simons Simplex Collection Genetics Consortium, Kaminsky EB, et al. (2010). Deletion 17q12 is a recurrent copy number variant that confers high risk of autism and schizophrenia. Am J Hum Genet 87(5):618-30.

21. 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. Environ Health Perspect 114(7):1119-1125.

22. Lord C, Risi S, Lambrecht L, Cook EH, Jr., Leventhal BL, et al. (2000). The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. J Autism Dev Disord 30: 205-223.

23. Lord C, Rutter M, Le Couteur A. (1994). Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. J Autism Dev Disord 24: 659-685.

24. Selzer RR, Richmond TA, Pofahl NJ, Green RD, Eis PS, et al. (2005). Analysis of chromosome breakpoints in neuroblastoma at sub-kilobase resolution using fine-tiling oligonucleotide array CGH. Genes Chromosomes Cancer 44: 305-319.

25. Day N, Hemmaplardh A, Thurman RE, Stamatoyannopoulos JA, Noble WS. (2007). Unsupervised segmentation of continuous genomic data. Bioinformatics 23: 1424-1426.

26. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV. (2002). Recent segmental duplications in the human genome. Science 297(5583):1003-7.

27. Itsara A, Cooper GM, Baker C, Girirajan S, Li J, et al. (2009) Population analysis of large copy number variants and hotspots of human genetic disease. Am J Hum Genet 84: 148-161.

28. Mefford HC, Cooper GM, Zerr T, Smith JD, Baker C, et al. (2009). A method for rapid, targeted CNV genotyping identifies rare variants associated with neurocognitive disease. Genome Res 19(9):1579-85

29. Gilman SR, Iossifov I, Levy D, Ronemus M, Wigler M, et al. (2011). Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. Neuron 70(5):898-907.

30. McCarroll SA, Altshuler DM. (2007). Copy-number variation and association studies of human disease. Nat Genet 39(7 Suppl):S37-42.

31. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, et al. (2007). Strong association of *de novo* copy number mutations with autism. Science 316(5823):445-449.

32. Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, et al. (2010). Functional impact of global rare copy number variation in autism spectrum disorders. Nature 466(7304):368-372.

33. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, et al. (2010). Origins and functional impact of copy number variation in the human genome. Nature 464(7289):704-712.

34. Marshall CR, Noor A, Vincent JB, Lionel AC, Feuk L, et al. (2008). Structural variation of chromosomes in autism spectrum disorder. Am J Hum Genet 82(2):477-88.

35. Xu B, Roos JL, Levy S, van Rensburg EJ, Gogos JA, et al. (2008). Strong association of de novo copy number mutations with sporadic schizophrenia. Nat Genet 40(7):880-5.

36. Stefansson H, Rujescu D, Cichon S, Pietiläinen OP, Ingason A, et al. (2008). Large recurrent microdeletions associated with schizophrenia. Nature 455(7210):232-6.

37. Stone JL, O'Donovan MC, Gurling H, Kirov GK, Blackwood DH, et al. (2008). Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature 455(7210):237-41.

38. Zhang D, Cheng L, Qian Y, Alliey-Rodriguez N, Kelsoe JR, et al. (2009). Singleton deletions throughout the genome increase risk of bipolar disorder. Mol Psychiatry 14(4):376-80.

39. Mefford HC, Muhle H, Ostertag P, von Spiczak S, Buysse K, et al. (2010). Genome-wide copy number variation in epilepsy: novel susceptibility loci in idiopathic generalized and focal epilepsies. PLoS Genet 6(5): e1000962.

40. Striano P, Coppola A, Paravidino R, Malacarne M, Gimelli S, et al. (2011). Clinical Significance of Rare Copy Number Variations in Epilepsy: A Case-Control Survey Using Microarray-Based Comparative Genomic Hybridization. Arch Neurol Epub.

41. Aradhya S, Manning MA, Splendore A, Cherry AM. (2007). Whole-genome array-CGH identifies novel contiguous gene deletions and duplications associated with developmental delay, mental retardation, and dysmorphic features. Am J Med Genet A 143A(13):1431-41.

42. Shoukier M, Klein N, Auber B, Wickert J, Schröder J. (2012). Array CGH in patients with developmental delay or intellectual disability: are there phenotypic clues to pathogenic copy number variants. Clin Genet Epub.

43. Christofolini DM, de Paula Ramos MA, Kulikowski LD, da Silva Bellucco FT, Belangero SI. (2010). Subtelomeric rearrangements and copy number variations in people with intellectual disabilities. J Intellect Disabil Res 54(10):938-42.

44. Halazonetis TD, Gorgoulis VG, Bartek J. (2008). An oncogene-induced DNA damage model for cancer development. Science 319(5868):1352-1355.

45. Bell CG, Beck S. (2010). The epigenomic interface between genome and environment in common complex diseases. Brief Funct Genomics 9(5-6):477-85.

46. Gohlke JM, Thomas R, Zhang Y, Rosenstein MC, Davis AP. (2009). Genetic and environmental pathways to complex diseases. BMC Syst Biol 3:46.

47. Feuk L, Carson AR, Scherer SW. (2006). Structural variation in the human genome. Nat Rev Genet 7(2):85-97.

48. Emanuel BS, Shaikh TH. (2001). Segmental duplications: an 'expanding' role in genomic instability and disease. Nat Rev Genet 2(10):791-800.

49. Bailey JA, Eichler EE. (2006). Primate segmental duplications: crucibles of evolution, diversity and disease. Nat Rev Genet 7(7):552-6.

50. Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, et al. (2005). Segmental duplications and copy-number variation in the human genome. Am J Hum Genet 77(1):78-88.

51. Calabretta B, Robberson DL, Barrera-Saldaña HA, Lambrou TP, Saunders GF. (1982). Genome instability in a region of human DNA enriched in Alu repeat sequences. Nature 296(5854):219-25.

52. Stenger JE, Lobachev KS, Gordenin D, Darden TA, Jurka J. (2001). Biased distribution of inverted and direct Alus in the human genome: implications for insertion, exclusion, and genome stability. Genome Res 11(1):12-27.

53. Fazza AC, Sabino FC, de Setta N, Bordin NA Jr, da Silva EH, et al. (2009). Estimating genomic instability mediated by Alu retroelements in breast cancer. Genet Mol Biol 32(1):25-31.

54. Miki Y, Katagiri T, Kasumi F, Yoshimoto T, Nakamura Y. (1996). Mutation analysis in the BRCA2 gene in primary breast cancers. Nature Genet 13:245–247.

55. Jingshan C, Barbara GH, Leonard HA. (1995). Presence and instability of repetitive elements in sequences the altered expression of which characterizes risk for colonie cancer. Cancer Res 55:174-180.

56. Mattarucchi E, Guerini V, Rambaldi A, Campiotti L, Venco A. (2008). Microhomologies and interspersed repeat elements at genomic breakpoints in chronic myeloid leukemia. Genes Chromosomes Cancer 47(7):625-32.

57. Deininger PL, Batzer MA. (1999). Alu repeats and human disease. Mol Genet Metab 67(3):183-93.

58. Roohi J, Montagna C, Tegay DH, Palmer LE, DeVincent C. (2009). Disruption of contactin 4 in three subjects with autism spectrum disorder. J Med Genet 46(3):176-82.

59. Fernandez T, Morgan T, Davis N, Klin A, Morris A. (2004). Disruption of contactin 4 (CNTN4) results in developmental delay and other features of 3p deletion syndrome. Am J Hum Genet 74(6):1286-93.

60. Sanders SJ, Ercan-Sencicek AG, Hus V, Luo R, Murtha MT. (2011). Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. Neuron 70(5):863-85.

61. Frye RE. (2009). 15q11.2-13 duplication, mitochondrial dysfunction, and developmental disorders. J Child Neurol 24(10):1316-20.

62. Koochek M, Harvard C, Hildebrand MJ, Van Allen M, Wingert H. (2006). 15q duplication associated with autism in a multiplex family with a familial cryptic translocation t(14;15)(q11.2;q13.3) detected using array-CGH. Clin Genet 69(2):124-34.

63. Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT. (2008). Association between microdeletion and microduplication at 16p11.2 and autism. N Engl J Med 358(7):667-75.

64. Shinawi M, Liu P, Kang SH, Shen J, Belmont JW. (2010). Recurrent reciprocal 16p11.2 rearrangements associated with global developmental delay, behavioural problems, dysmorphism, epilepsy, and abnormal head size. J Med Genet 47(5):332-41.

65. Xi R, Hadjipanayis AG, Luquette LJ, Kim TM, Lee E. (2011). Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. Proc Natl Acad Sci U S A 108(46):E1128-36.

Academic Vita

Kian Hui, Yeoh
Undergraduate Honors Student

518 University Drive,
Apartment 111,
State College, PA 16801.

Tel: 814-753-0301
Email: kiy5025@psu.edu

Education and Research Experience

Pennsylvania State University, State College, PA

- Bachelor of Science in Biotechnology, 2009 – 2012
- Minor in Microbiology; Minor in Biochemistry and Molecular Biology, 2009 – 2012

Department of Biochemistry and Molecular Biology, Pennsylvania State University, State College, PA

- Undergraduate honors research with Dr. Maria Krasilnikova, 2010 – 2012
- Research project: Deciphering the mechanism of the first replication cycle in mammalian embryonic cells.

Department of Biochemistry and Molecular Biology, Pennsylvania State University, State College, PA

- Undergraduate research with Dr. Scott Selleck, 2011-2012
- Research project: Copy number variations burden in autism.

Department of Genome Sciences, University of Washington, Seattle, WA

- Undergraduate summer research with Dr. Santhosh Girirajan, Summer 2011
- Research project: Structural variation and large-scale genomic rearrangements in human genome leading to various genomic disorders such as autism.

Honors, Awards and Academic Achievements

- Summer Undergraduate Research Fellowship 2011 - John Lapinski Summer Scholar Award & Paul and Mildred Berg Endowment for Eberly College of Science Summer Research (Awarded for excellence in undergraduate research, Biochemistry and Molecular Biology Department, Pennsylvania State University).
- The Evan Pugh Scholar Award (seniors) (Awarded for outstanding academic achievement, Pennsylvania State University).
- Student Marshall representing Biotechnology majors for Spring Commencement 2012, Biochemistry and Molecular Biology Department, Pennsylvania State University.
- An outstanding member of the Schreyer Honors College performing honor thesis research, 2010 – present, Pennsylvania State University.
- Summer Undergraduate Research Fellowship 2010 Finalist, Biochemistry and Molecular Biology Department, Pennsylvania State University.

- Enlisted in Dean's list for excellent academic achievement, Pennsylvania State University, 2009 – 2012

Seminar/Extracurricular Activities

- Presenter/Speaker, Eberly College of Science Undergraduate Research Fair 2012 Poster Session, Pennsylvania State University, 2012.
- Presenter/Speaker, Undergraduate Research Poster Exhibition, Department of Biochemistry and Molecular Biology, Pennsylvania State University, 2011
- THON Chair of Malaysian Students Club for THON 2012, Pennsylvania State University, 2011-2012
- Sub-committee and cultural performer of Malaysian Cultural Night 2012, Malaysian Student Clubs, Pennsylvania State University, 2011-2012
- Participated in Summer Discovery Grant for undergraduate research, Pennsylvania State University, 2011
- A participant in PNC Leadership Assessment Center, Pennsylvania State University, 2011.
- A volunteer in Schreyer Day of Service, Schreyer Honors College, Pennsylvania State University, 2011.

Teaching Experience

- Tutor of BMB 400 Molecular Biology of the Gene for the Student Support Services Program, Pennsylvania State University, 2011.
- Tutor of BMB 402 General Biochemistry for the Student Support Services Program, Pennsylvania State University, 2011.

Laboratory Skills/Techniques

- Bacterial and mammalian cells culturing
- PCR and qPCR techniques
- 2-D gel electrophoresis and Southern Blot Hybridization
- Chromatin immunoprecipitation assay
- NMR and UV spectroscopic method.
- DNA microarray/Array comparative genomic hybridization (aCGH).
- Methods in bio-fermentation and bioprocessing.
- Protein expression (β-galactosidase) assays

References

Maria Krasilnikova, Ph.D. (honors thesis research advisor)
Research Assistant Professor
Department of Biochemistry and Molecular Biology
Pennsylvania State University
Life Sciences Building, 206C
University Park, PA 16802
Email: muk19@psu.edu
Tel: (814) 863-5555

Scott Selleck, MD, PhD (undergraduate research advisor)
Professor and Head
Department of Biochemistry and Molecular Biology
Pennsylvania State University
Life Sciences Building, 206D
University Park, PA 16802
Email: sbs24@psu.edu
Tel: (814) 867-3274

Santhosh Girirajan, MBBS, Ph.D. (summer research supervisor)
Senior Fellow
Department of Genome Sciences
University of Washington
Foege Building S413A
3720 15th Ave NE
Seattle, WA 98195
Email: sangi@u.washington.edu
Tel: (206) 685-7336

David Gilmour, Ph.D. (honors adviser)
Professor of Molecular and Cell Biology
Department of Biochemistry and Molecular Biology
Pennsylvania State University
465A North Frear Laboratory
University Park, PA 16802
Email: dsg11@psu.edu
Work: (814) 863-8905