12-17-2012 12:00 AM

Comprehensive Assessment of CNV Calling Algorithms: A Family Based Study Involving Monozygotic Twins Discordant for Schizophrenia

Sujit Maiti
The University of Western Ontario

Supervisor
Dr. Shiva M. Singh
The University of Western Ontario

Graduate Program in Biology
A thesis submitted in partial fulfillment of the requirements for the degree in Master of Science

COMPREHENSIVE ASSESSMENT OF CNV CALLING ALGORITHMS: A FAMILY BASED STUDY INVOLVING MONOZYGOTIC TWINS DISCORDANT FOR SCHIZOPHRENIA

(Spine title: CNV in monozygotic twins)

(Thesis format: Monograph)

by

Sujit Maiti

Graduate Program in Biology

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science

The School of Graduate and Postdoctoral Studies
The University of Western Ontario
London, Ontario, Canada

THE UNIVERSITY OF WESTERN ONTARIO
SCHOOL OF GRADUATE AND POSTDOCTORAL STUDIES

CERTIFICATE OF EXAMINATION

Supervisor

Dr. Shiva M. Singh

Supervisory Committee

Dr. Kathleen Hill

Dr. Mark Daley

Examiners

Dr. Kathleen Hill

Dr. Susanne Kohalmi

Dr. Graham Thompson

The thesis by

Sujit Maiti

entitled:

COMPREHENSIVE ASSESSMENT OF CNV CALLING ALGORITHMS: A FAMILY BASED STUDY INVOLVING MONOZYGOTIC TWINS DISCORDANT FOR SCHIZOPHRENIA

is accepted in partial fulfillment of the
requirements for the degree of
Master of Sciencce

Date

Chair of the Thesis Examination Board

ABSTRACT

Genetic variability is essential to human individuality. Genetic variation includes differences in sequence at the single nucleotide level to structural variations of large segments of DNA called copy number variations (CNVs). CNVs within a genome can be identified using microarray technology; however, the analysis of microarray results resulting in the "calling" of CNVs is not always precise. The research included in this manuscript describes the identification and analysis of CNVs using three commercially-available packages, Affymetrix® Genotyping Console™, Partek® Genome Suite™ and PennCNV, that are most commonly used in the analysis of SNP and CNV data. Specifically, this research assessed the ability of these platforms to successfully analyze Affymetrix® Genome-Wide Human SNP Array 6.0 data for CNVs within two families, each with a set of monozygotic twins discordant for schizophrenia. Results show that the three methods identified a set of CNVs in each individual, but the specific sets identified were not identical between softwares. Affymetrix® Genotyping Console™ detected a wide variety of sizes of CNVs while the other two methods were able to identify only CNVs greater than 1 Mb in size. Interestingly, all platforms showed that monozygotic twins differ for some CNVs, a difference that may be acquired during their somatic development. This suggests that CNV differences between monozygotic twins may offer an explanation for discordance of phenotype, such as schizophrenia. Also, this analysis of CNVs within related individuals may identify previously unreported unusual features, including the repeated CNVs on chromosome 13q observed in the father of family 2. Such results support the use of CNV in familial studies, but argue for a careful assessment of CNVs including a careful selection of analysis tools and the necessity of independent confirmation.

Key words: Copy number variants, microarray, monozygotic twin, CNV calling algorithms

ACKNOWLEDGMENTS

I would like to sincerely acknowledge my supervisor, Dr. Shiva Singh, for believing in me and giving me the opportunity to complete this research. Exposure to his approach that seeks to ask the right scientific questions and extend science from bench to bedside has been a great learning experience for me. I am thankful for all his support and patience, and the valuable time he spent to answer all my questions. His friendly and welcoming attitude is always encouraging. I have a great training experience under his supervision.

I am thankful to all my labmates, who offered tremendous support, assistance and encouragement. I would like to specially thank Morgan Kleiber, Haroon Sheikh and Aniruddho Chokroborty Hoque for their support during hard times.

I appreciate the support of David Carter from the London Regional Genomics Center at Robarts Research Institute for his assistance during Partek® Genome Suite™ data analysis. Thanks to Dr. Richard O'Reilly who provided all biological samples. Without his support this research could not be possible. I wish to acknowledge the patients who graciously offered their time and DNA for this research project. I am thankful to my advisory committee members Dr. Kathleen Hill and Dr. Mark Daley for their tremendous support.

Lastly, but certainly not the least, I thank my parents and in-laws who believed in me and provided tremendous support. I cannot find enough words to express my thanks to my wife, Mousumi, for her constant support and encouragement. She is instrumental to my desire to achieve my goals. Finally, thanks to my daughter Mithika, who is the inspiration to complete this research. Life is always easy being surrounded by family like you.

TABLE OF CONTENTS

CHAPTER 3: RESULTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BAF: B allele frequency

bp: Base pair

.cel file: Cell intensity file, a single intensity value is calculated for every probe on the chip

CN: Copy number

CNV: Copy Number Variation

DSM-IV: Diagnostic and Statistical Manual of Mental Disorders IV

GTC: Affymetrix® Genotyping Console™

HMM: Hidden Markov Model

Kb: Kilobase

LRR: Log R Ratio

Mb: Megabase

NAHR: Non allelic Homologous Recombination

NHEJ: Non Homologous End Joining

PGS: Partek® Genome Suite™

QC: Quality Control

SNC: Single Nucleotide Changes

SNP: single nucleotide polymorphism

SCZ: Schizophrenia

Chapter 1: Introduction

The principles of biological evolution provide the foundation for modern biology. They have been instrumental in revolutionizing our understanding of the relationships between all life forms and can explain the extensive variability seen both between and within all species. This variability has evolved over time and is encoded in DNA. Some differences represent part of the evolutionary processes that have occurred over time and others are newly arisen and may or may not be passed on to the next generation. The degree, type and source (common or rare, polymorphic, ancient or *de novo*) of genetic variation that exists is not fully understood for most species, information that could be used for understanding a species' past and reflecting on its future in terms of survival, reproduction and fitness. This is particularly true for humans where an ever-increasing number of individual genomes are being assessed using DNA microarrays and complete genome sequencing. These data are organized and have been made available to researchers, but this rich and valuable resource remains to be analyzed and fully understood. Briefly, data available to date show that the human genome carries extensive genetic variability. One may argue that no two individuals are identical. Also, this variability exists in the form of a number of mutational events that include single nucleotide polymorphisms (SNPs), variable number of tandem repeats (VNTRs such as mini- and microsatellites), presence or absence of transposable elements (e.g. *Alu* elements) and structural alterations such as deletions, duplications and inversions including copy number variations (CNVs) that cover large (1Kb to 5Mb) DNA segments. Although it has been possible to identify specific events by focussing on individual genes and genomic regions, a full evaluation of the extent to which they may occur across an entire genome has become possible only recently. This opportunity has its origin in the complete human genome sequences generated during early 2000s (Istrail *et. al.,* 2004). The availability of human genome sequence has allowed the development of novel technologies, including DNA microarrays that can be used to interrogate the entire genome (Struan *et. al.,* 2008). The development of these technologies now allows an assessment of the degree and nature of DNA variability across a genome and between

individuals. Further, although human DNA microarray technology started with a relatively small number of SNP markers, modern chips now include an ever-increasing number of markers (>1 million) that are evenly spaced and cover the entire genome (Bierut *et al.,* 2010). This allows the quantification of different types of genomic variations that occur on an individual basis. Of all types of genomic variations identified, SNPs and CNVs are probably the most commonly evaluated in most experiments involving humans. This is based on two facts: first, they are among the most common forms of genetic variation in humans, and second, they are thought to represent functional variations that may contribute to differences in human phenotypes such as diseases. In this research, I have chosen to focus on CNVs.

1.1 Copy Number Variations (CNVs)

CNVs represent differences in the numbers of copies of relatively large genomic regions (1Kb to 5 Mb) that may include a number of genes. Two fundamental studies in 2004 pioneered the establishment of copy number variations as an important source of genetic differences in humans (Iafrate *et al.,* 2004; Sebat *et al.,* 2004). They concluded that CNVs are responsible for approximately 18% of variation in gene expression and, subsequently, may play a vital role in the determination of complex traits (Strange *et al.,* 2007). The origin of CNVs at a given genomic site is not fully understood, however the mechanism may involve non-homologous end joining (NHEJ) or non-allelic homologous recombination (NAHR) (Zhang *et. al.,* 2009). CNVs have been found to vary across populations in frequency distribution, with copy number polymorphisms (CNPs) occurring in more than 1% of a given population (Scherer *et al.,* 2007). CNVs represent a novel insight into the extent of structural variation that may occur in the human genome, offering a new tool to understand complex traits (Li *et al.,* 2009) and investigate the role of gene functions in disease (McCaroll *et al.,* 2007). Currently, several studies have successfully associated human phenotypes or diseases with copy number variation. In fact, copy number variation (CNV) may have extensive role in genetic variability and human disease (Lonita-Laza *et al.,* 2009). CNVs have been shown to functionally contribute to a number of autoimmune diseases (Lee *et al.,* 2012) and cancers (Pylkas *et al.,* 2012). In addition, they can determine the onset of various Mendelian diseases.

Interestingly, CNVs have also been shown to be associated with the variability observed in human complex traits, such as susceptibility to HIV infection, autism, and schizophrenia (Zhang *et al.*, 2009).

Several unique characteristics of CNVs sustain their role in human disease genetics. First, although less abundant than SNPs, it has been suggested that the thousands of bases included in a CNV locus may result in more nucleotide variation than SNPs for their size, given that they often encompass and may even interrupt the function of coding DNA sequences (Tuzun *et al.*, 2005). Second, CNVs appear to be an enhancement of "environmental sensor" genes that are not essentially vital for early embryonic development but rather help humans to adapt to an ever-changing environment, such as olfactory receptors, immune and inflammatory response and ion channels genes (Sebat *et al.*, 2004; Tuzun *et al.*, 2005).

1.2 Assessment of Genome-wide Copy Number Variations (CNVs):

Identification and evaluation of CNVs begins with an assessment of the dosage of a given region of genomic DNA present in the genome. This process may use any region-specific dosage assessment technology if the region of the genome affected is already known or hypothesized. However, region-specific technologies are not practical for identification of CNVs across the genome, or the identification of novel CNVs. Therefore, genome-wide studies of CNVs typically utilize genomic arrays. This leaves us with the option of two technologies: comparative genomic hybridization (CGH) (Pinkel *et. al.*, 1998), commonly used in the detection of gross structural changes associated with cancer tissues, or commercially-available SNP arrays (Huang *et. al.*, 2004). In both cases, the identification of CNVs relies on the signal intensity of a stretch of DNA or a linear sequence of SNP markers. The SNP arrays in particular generate a large number of SNP (>1 million) signal intensities that are expected to be proportional to their dose. These results are then normalized for the amount of DNA used, experimental conditions, image analysis, plate and batch effects, and other experimental conditions. This is followed by a number of analytic methods that are expected to identify specific CNVs. However, the accuracy of

CNV detection is not currently ideal and results differ depending on the detection programs used and their parameters.

Published literature attests that SNP arrays, particularly the Affymetrix® Genome-Wide Human SNP Array 6.0, have become the major tool for CNV analysis in humans. These arrays have been designed to allow the simultaneous genotyping of a very high number of polymorphisms. SNP arrays typically screen from thousands to millions of SNPs for a given sample and have been used extensively and successfully for genome-wide association studies (Ragoussis, 2009). However, recent interest in the structural variation of the genome has led to the use of these arrays to satisfy other purposes. Intensity data from SNP arrays can be used to detect copy number (CN) changes and a number of methods have been developed to assist with this procedure. Initially, the arrays were optimised only for genotyping SNPs, but recent array designs now include specially-designed probes to enable the detection of sites of known structural or copy number variation. These copy number polymorphisms (CNPs) can be assayed in parallel with other SNPs to produce a clearer picture of genome variation. At the current time, the Affymetrix® Genome-Wide Human SNP Array 6.0 has 1.8 million probes. Since CNVs are events found in less than 1% of the population (Feuk *et. al*., 2006), this increased overall genomic coverage and assay density allows the detection of rare CNV changes.

It is also apparent from the literature that different analytical methods may yield different sets of CNVs from the same raw (.cel file) data source. Also, attempts to confirm these results using independent methods including quantitative PCR are not always successful. Such results argue that there is a need to refine detection and analysis methods so that common .cel files will yield reliable and reproducible results. In this research, I have used a set of novel Affymetrix® Genome-Wide Human SNP Array 6.0-generated .cel files generated from the analysis of related individuals in order to test the suitability of a set of commonly used analytical methods in the identification of CNVs. The three analytical algorithms assessed in this research are, i) Affymetrix® Genotyping Console™, ii) Partek® Genome Suite™ and iii) PennCNV, all of which can use the same raw intensity files to filter and investigate CNVs. Three algorithms are based on a Hidden Markov Model (HMM) and use different copy numbers as hidden states to detect changes using

different approaches. The objective of this study is to evaluate these three calling algorithms including their biases, their stringency, and what commonalities they may share.

1.3 Analytical Challenges

There are two goals of this analysis. The first is to identify regions of the genome that have a higher (3+) or lower (0-1) intensity of marker probes across a given genome. Such results allow the identification of genomic regions where a number of continuous markers have either higher or lower intensity of signals as compared to the other related individuals or than would be expected based on known genomic data. Such regions, if they are between 1Kb to 5Mb in size and involve more than 20 consecutive markers, are flagged as regions existing in more than 3 copies (a gain) or less than 2 copies (a loss) within the genome and are categorized as copy number variation. The next set of analyses seeks to identify CNVs from Affymetrix platform (Genome-Wide Human SNP Array 6.0 using different algorithms or packages as listed below.

1) Affymetrix® Genotyping Console™ (GTC) (http://www.affymetrix.com/).
2) Partek® Genome Suite™ (PGS) (http://www.partek.com/).
3) PennCNV (http://www.openbioinformatics.org/penncnv/)
4) dChip (http://biosun1.harvard.edu/complab/dchip/snp.htm).
5) QuantiSNP (http://groups.google.co.uk/group/quantisnp).
6) Nexus Biodiscovery (http://www.biodiscovery.com/index/nexus).

Choosing an algorithm for the detection for copy number changes is dependent on the platform that generated the data. It is also important to take into account the factors in the dataset, since some algorithms provide more flexibility for adjusting parameters. Further, by analysing with more than one algorithm, the user can take advantage of different

features and strengths of individual tools, and the results can be compared to confirm events of interest. This study uses the first three algorithms (Affymetrix® Genotyping Console™, Partek and PennCNV) as they are most commonly used in studies examining human copy number variation.

1.3.1 Affymetrix® Genotyping Console™ (GTC)

GTC is a software designed for the accurate and thorough detection of single nucleotide polymorphisms (SNP) as well as both common and rare copy number (CNV) variants. It is comprised of a four-stage analytical framework for deriving integrated and mutually consistent copy number and SNP genotypes. These four stages are i) genotyping analysis using a Birdseed algorithm, ii) copy number/loss of heterozygosity analysis using a BRLMM-P+ algorithm, iii) common or known CNV detection using the Canary algorithm, and iv) rare CNV detection with the Birdseye algorithm. Birdseed is a two-dimensional Gaussian Mixture Model (GMM). It uses normalized intensities from a set of .cel files that then create a N by 2*M matrix where N is the number of samples and M is the number of SNPs. The Canary algorithm, on the other hand, identifies regions of known copy-number variation. Canary is a one-dimensional Gaussian Mixture Model (GMM). Birdseye is specialized to discover rare or *de novo* regions of variable copy number using a Hidden Markov Model (HMM) (Korn *et. al*., 2008, McCarroll *et. al*., 2008).

1.3.2 Partek® Genome Suite™ (PGS)

PGS is able to use Affymetrix .cel files for copy number analysis. The first step of CNV analysis is to estimate the number of copies of each marker (allele). This estimation is done by comparing each marker with reference from either paired or unpaired samples. A paired design assumes that each sample has its own reference sample, whereas unpaired samples use a common reference. PGS provides three options for selecting unpaired samples references: i) an existing reference file provided by Partek™ ii) a file created from a set of normal samples and ii) a file generated from a subset or all samples within the current project. The next step is the detection of regions with copy number variation. Starting with copy number estimates for each marker, PGS derives a list of regions where

adjacent markers share the same copy number using one of two offered algorithms for region detection: Genomic Segmentation or Hidden Markov Model (HMM). Essentially, both algorithms examine trends across multiple adjacent markers; the genomic segmentation algorithm identifies breakpoints in the data, i.e., changes in copy number between two neighbouring regions, while the HMM algorithm looks for discrete changes of copy number states (e.g., 0, 1, 2… with no upper limit) and will find regions with those numbers of copies. Therefore, the HMM model performs better in cases of homogenous samples when the copy numbers can be anticipated. Genomic segmentation is preferred for heterogeneous samples with unpredictable copy numbers. The number of copies of each marker detected in the previous step is then used to detect the genomic regions with copy number variation, i.e., to identify amplifications and deletions across the genome. Genome segmentation itself is divided into two steps. In the first step, each region is compared to an adjacent region in order to tell whether both have the same average copy number and if a breakpoint can be inserted by using a two-tailed *t*-test, which compares the average intensities of regions and if the corresponding cut-off *p*-value is below the *p*-value threshold. The genomic size of a region is defined by the number of genomic markers included in the region ("minimum genomic markers"), while the magnitude of significant difference between two regions is controlled by the signal to noise ratio (more simply, this could be thought of as the difference in copy numbers between the regions). If the *t*-test is significant, it can be concluded that the region differs significantly from its nearest neighbours with respect to copy number. However, a second step is needed to identify the exact nature of the difference, i.e., whether the difference is an amplification or deletion. In this stage, two one-tailed *t*-tests are used to compare the mean copy number in the region with the expected (normal) diploid copy number (Diskin *et. al*., 2008, Ramakrishna *et. al*., 2010, Yamamoto *et. al*., 2007).

1.3.3 PennCNV

PennCNV creates precise models for CNV analysis using log R Ratio and B Allele frequency. This develops more realistic models for state transition between different copy number states to better reflect the distribution of the intensity data. In addition, PennCNV incorporates the population allele frequency for each SNP and the distance

between adjacent SNPs. The raw signal intensity values measured for the A and B alleles are then subject to normalization using the signal intensity of all SNPs. This procedure produces the two values for each SNP, representing the experiment-wide normalized signal intensity of the A and B alleles. Two additional measures are then calculated for each SNP: total signal intensity and relative allelic signal intensity ratio. As a normalized measure of total signal intensity, the log R Ratio (LRR) value for each SNP is then calculated as $\log_2$ ratio of $R_{observed}$ and $R_{expected}$, where $R_{expected}$ is computed from linear interpolation of canonical genotype clusters. The B Allele Frequency (BAF) is calculated using a normalized measure of the relative signal intensity ratio of the B and A alleles where AA, AB, and BB are the values for the three canonical genotype clusters generated from a large set of reference samples. PennCNV uses a Hidden Markov model (HMM) for CNV detection. HMM provides a natural statistical framework for modeling dependence structures between copy numbers at nearby SNPs. To detect CNVs, PennCNV uses the HMM that assumes that the hidden copy number state at each SNP depends only on the copy number state of the most preceding SNP and the values of log R Ratio and B allele frequency are independent (Wang *et. al*., 2007).

1.4 Hypothesis:

The three analytical methods of identification of CNVs from the Affymetrix Genome-Wide Human SNP Array 6.0 data will generate similar copy number variation changes in monozygotic twins. Also, the first degree relatives of these individuals (mother or father) will share approximately 50% of their CNVs.

1.5 Specific Objectives

a) To identify the nature of copy number variation in eight individuals representing two families using three CNV calling algorithms, i) GTC, ii) PGS and iii) PennCNV, from .cel files generated using Affymetrix® Genome-Wide Human SNP Array 6.0. This characterization will include i) CNV size, ii)

chromosome distribution, iii) losses or gains and iv) *de novo* versus *inherited* changes.

b) To use CNV results (above) to establish any differences between monozygotic twins discordant for schizophrenia in two unrelated families.

c) To compare CNV results across platforms.

d) To assess the suitability of the platform combination for the reliable identification of human CNVs.

Chapter 2: Materials and Methods

2.1 Ethics Review

Research examining genetic discordance in monozygotic twins was approved by the Committee on Research Involving Human Subjects at the University of Western Ontario, in London, Ontario, Canada (Review Number : 09390E)

2.2 Sample Collection

The patients and families included in this study were identified and clinically assessed by Dr. Richard O'Reilly, Psychiatrist, Regional Health Care, London, ON. Informed consent was obtained from all participants, and relevant medical records were reviewed by Dr. O'Reilly. Participants were interviewed using the Structured Clinical Interview for DSM-IV (for personality disorders). Whole blood samples from each participant were collected by Dr. O'Reilly and immediately stored at -80°C. All blood samples were collected prior to the initiation of my research.

2.2.1 Sample background

Two families were assessed in this study. Each family consists of four participants, one twin pair and both parents, for a total of eight participants (Figure 1). Demographic information and relevant clinical history are listed in Table 1.

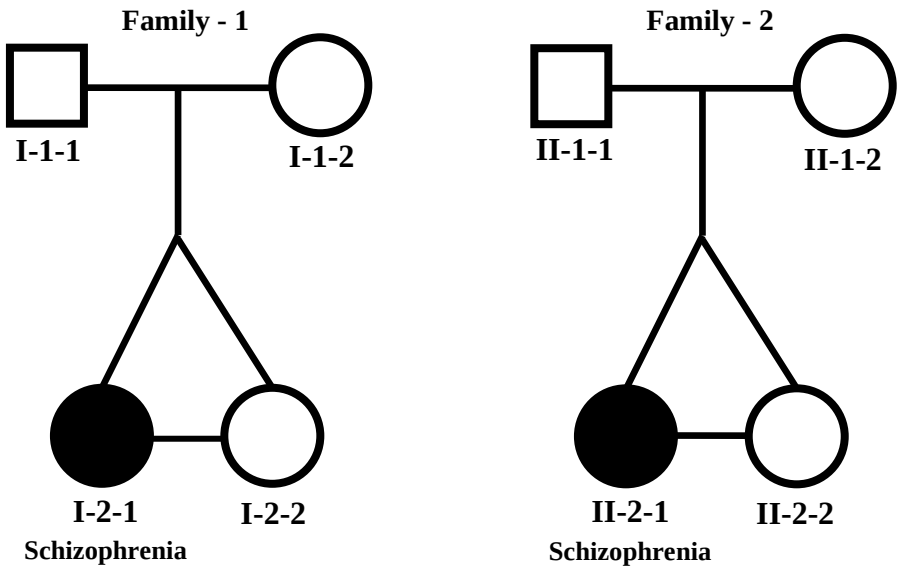


Figure 1: Pedigree of two families with monozygotic twins discordant for schizophrenia. Members of the family one are indicated with (I-) and members of the family two are indicated with (II-). The designations included in this figure are followed in subsequent figures and tables.

Table 1. Demography and Clinical History of monozygotic (MZ) twins discordant for Schizophrenia (SCZ).

	Family 1				Family 2			
	I-1-1	**I-1-2**	**I-2-1**	**I-2-2**	**II-1-1**	**II-1-2**	**II-2-1**	**II-2-2**
Age (yrs.) at assessment	82	74	53	53	N/A	N/A	43	43
Sex	Male	Female	Female	Female	Male	Female	Female	Female
Declared Race	Afro-American				Caucasian			
Psychiatric features	Compulsive Personality Disorder	N/A	Schizophrenia, Paranoid Type, onset age 22	Bipolar I Disorder, onset age 52	Major depression and panic disorder for 6 months after cardiac surgery, onset age 73	N/A	Schizoaffective Disorder, onset age 27	Single episode of Major Depression, fully remitted, onset age 18

Family one is indicated with (I-), family two is indicated with (II-). N/A = Not Applicable.

2.3 Genomic DNA extraction

Genomic DNA was extracted from white blood cells of all eight participants using the PerfectPure DNA Blood Kit (5 Prime, Inc., Gaithersburg, MD, USA) following manufacturer protocol. Genomic DNA was extracted prior to the initiation of my research.

2.4. Microarray data generation using the Affymetrix® Genome Wide Human SNP Array 6.0

Genomic DNA was hybridized to Affymetrix® Genome Wide Human SNP Array 6.0 at the London Regional Genomics Centre (London, ON) following manufacturer's protocol. Approximately 5 µg of genomic DNA was used from each sample. These data were provided to me for this research.

2.4.1 Affymetrix® Genome-Wide Human SNP Array 6.0

There are 1.8 million genetic markers represented on the Affymetrix Genome Wide Human SNP Array 6.0. Among these, there are more than 906,600 probes testing for single nucleotide polymorphisms (SNPs) and more than 946,000 probes for the detection of copy number variation.

Copy number variation probes include:

i) 5,677 CNV regions from the Database of Genomic Variants (http://projects.tcag.ca/variation/) are covered by 202,000 probes.

ii) 744,000 probes are evenly distributed along the human genome.

iii) Non-polymorphic copy number probes are included.

iv) Genotyping Console™ (GTC) software, developed by the Broad Institute (Boston, MA, USA), is available and uses a high-resolution reference map and a copy number polymorphism (CNP)-calling algorithm for the detection of CNVs.

Single nucleotide polymorphisms (SNPs) probes include:

i) Unbiased selection of 482,000 SNPs.
ii) Selection of additional 424,000 SNPs.
iii) Tag SNPs.
iv) SNPs from chromosomes X and Y.
v) Mitochondrial SNPs.
vi) New SNPs added to the dbSNP database.
vii) SNPs located in recombination hotspots.

2.5 Selection of packages for Microarray data analysis

Currently, there are several packages available to analyze microarray data. For the Affymetrix® Genome Wide Human SNP Array 6.0 data, the most commonly used packages are (i) Affymetrix Genotyping Console 4.1 (GTC) (Affymetrix, Santa Clara, CA, USA), (ii) Partek® Genomics Suite™ (PGS) (Partek Inc., St. Louis, MO) and (iii) PennCNV (Jeanette E Eckel-Passow *et. al.,* 2011). This study utilized all three packages to analyze the same set of raw array data.

2.5.1 Affymetrix® Genotyping Console™

The Affymetrix® Genotyping Console™ 4.1 software (GTC) (http://www.affymetrix.com/) provides a way to create genotype calls for .cel files. GTC generates genotype, copy number and Loss of Heterozygosity (LOH), copy number segments data, and copy number variation data for Genome-Wide Human SNP Array 6.0. For these analyses, GTC uses Birdseed V2, BRLMM-P+, and Canary algorithms respectively. An overview of GTC workflow for the detection of these genomic features using array data is shown in Figure 2.

Before performing analysis of the data generated by the Genome-Wide SNP Array 6.0, GTC requires some initial setup, the steps of which are outlined below:

i] Creation of new workspace and data set where you can load the microarray data.

ii] Import the array data into the data set. Input data files are:

A] Sample files (.arr/.xml files for each sample)

B] Intensity data files (.cel files for each sample)

iii] Perform the intensity quality control (QC) test for the quality of each sample intensity file. This process allows the determination of the input data quality. GTC has developed different control features for this process, including the QC Call Rate matrix and Contrast QC matrix:

a) *QC Call Rate matrix*: GTC creates a QC call rate matrix table with QC values for each sample. Default QC Call Rate Threshold is ≥ 86. This value may be changed depending on the user's requirements. Data in this study were analyzed using the default value. *Contrast QC matrix*: Contrast QC is the QC matrix within GTC software. The default threshold is ≥ 0.4 for each sample. It uses 10,000 random SNPs detected by the Genome-Wide Human SNP Array 6.0 to calculate the contrast QC for each sample.

Depending on the QC value for each sample, GTC places each call into one of three groups: All, In Bounds, and Out of Bounds.

After the initial setup, three different types of analyses can be performed: Genotyping, Copy Number and Loss of Heterozygosity (CN/LOH) analysis, and Copy Number Variation analysis. Furthermore, from the CN/LOH analysis we can do Copy Number Segment Analysis.

i. *Genotyping:* Birdseed's assign AA, AB, and BB genotype calls for each sample with two copies of a SNP, then characterize genotype and allele-specific probe responses for each SNP. Birdseed uses normalized and summarized intensities from a set of .cel files for each SNP allele. This takes the form of a matrix with N columns and 2*M rows, where N is the number of samples to be analyzed, and M

is the number of SNPs. It is a 2d Gaussian Mixture Model (GMM) that clusters diploid samples into the canonical SNP genotype AA, AB, and BB.

This analysis creates following results for each SNP probe set for each sample:

a) SNP Call Rate

b) SNP Call (AA, AB, BB)

c) Minor Allele Frequency

d) Hardy-Weinberg p-value

ii. *Copy Number & Loss of Heterozygosity:* GTC creates Copy Numbers and Loss of Heterozygosity using this analysis. Segment analysis then creates the copy number segment report, which allows the user to see the chromosomal break points and gene overlap for the copy number variation using the GTC browser. This analysis creates an output file extension with CN5.cnchp, which contains both copy number and loss of heterozygosity values. This analysis can be performed in two different ways:

a) Reference Model File Creation and Analysis (Batch Sample Mode): This mode first creates a Reference Model file using the .cel files for the selected samples. Then each sample file is compared with this reference file.

b) Analysis with a Previously Created Reference Model File (Single Sample Mode).

In this analysis, each sample file is compared to a previously created Reference Model file. This reference file was created using HapMap270 sample files supplied by Affymetrix.

Both these methods use BRLMM-P+ algorithm for the analysis. Copy number segment analysis creates segment data files and segment summary files.

iii. *Copy Number Variation Analysis*: Copy Number Variation (CNV) Analysis uses the Canary algorithm to make a CNV state call (0, 1, 2, 3, 4) for copy number variations in the genome. A region with copy number variation may contain a few or too many CN/SNP probe sets. Canary algorithm was developed by the Broad Institute for the purpose of making copy number state calls for genomic regions with copy number variations. Canary algorithm identifies regions of known copy-number variation, especially suited for regions of common variation. Canary is a 1d Gaussian Mixture Model (GMM). It uses HapMap model file to determine the region of known CNV. Another algorithm called Birdseye discovers regions of variable copy number, especially those that are rare or de novo. It is a Hidden Markov Model (HMM) to find regions of variable copy number in a sample.

The following support files are necessary to use for all above mentioned data analysis in GTC:

a) Library file sets for genotyping, copy number/LOH/CN segment and copy number variation analysis.

b) Reference Model files for SNP 6.0 single sample Copy Number/LOH analysis

SNP lists (provided by Affymetrix).

c) Browser Annotation files.

Using the GTC browser the user is able to observe the copy number analysis results; specifically, Genome View, Chromosome Wise View, Segment Report Table, Copy Number Variation with Gene Overlap, log2 Ratio values and CN state values for each individual.

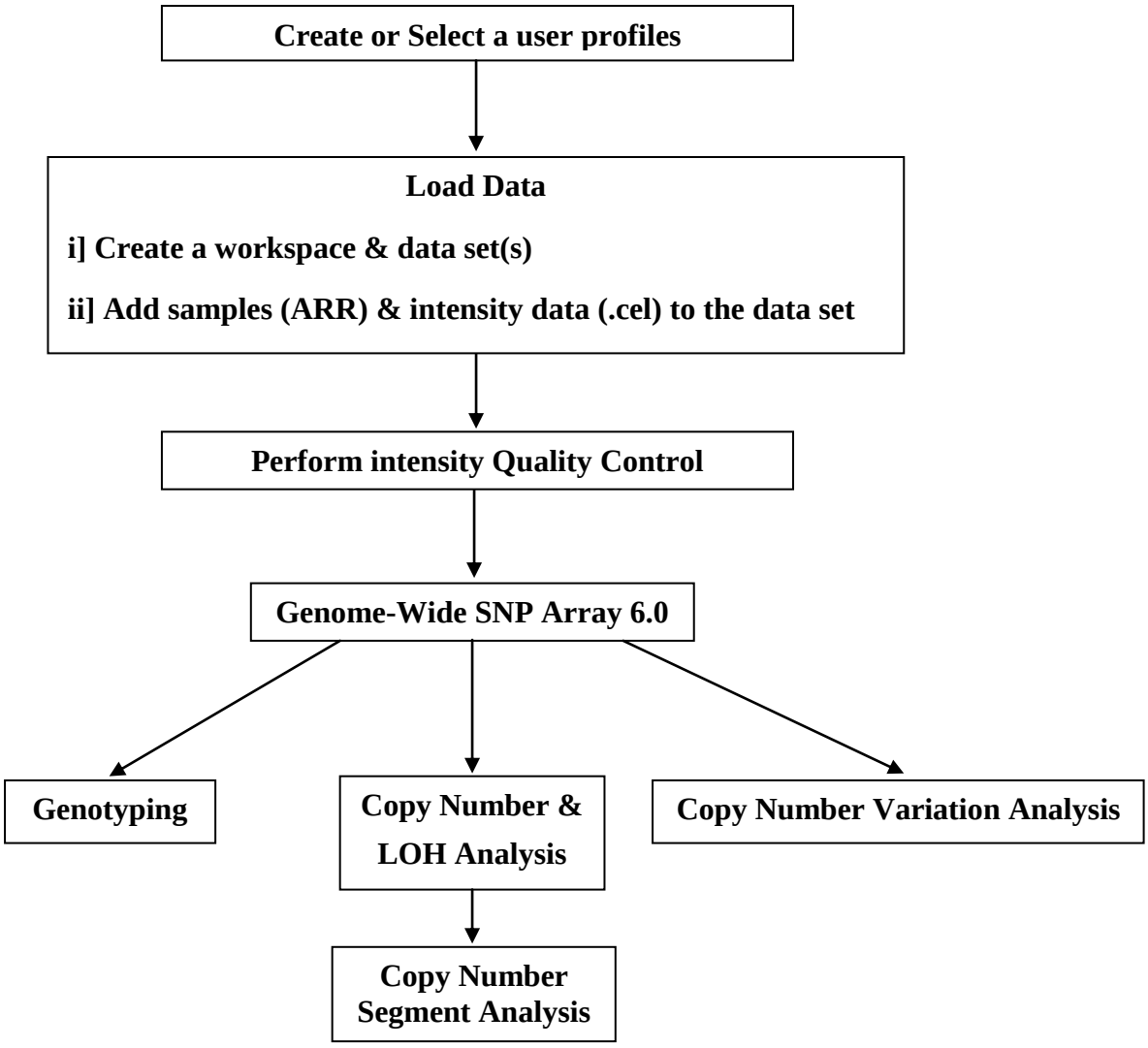


Figure 2: Overview of Affymetrix® Genotyping Console™ workflow

2.5.2 Partek® Genomics Suite™

Partek® Genomics Suite™ (PGS) (http://www.partek.com/) is a software that is capable of analyzing SNP-genotyping arrays with closely spaced genomic markers (Affymetrix® and Illumina®), to detect amplified or deleted regions in the genome within and shared across samples. An overview of copy number variation analysis in PGS is shown in Figure 3.

PGS uses Affymetrix .cel and CHP files as input data. PGS creates a principal component analysis (PCA) scatter plot to visually assess the intensity distribution of the dataset including any outliers. By observing the PCA scatter plot we can assess the appropriateness of the variation within and between samples. Next, the software estimates the number of copies of each marker (allele). For this step, PGS uses a paired and unpaired method. In the paired method, PGS assumes that each sample has its own reference sample. In the unpaired design, PGS uses a common reference sample for all samples in the dataset. Our data was analyzed using the unpaired method. Next, the software was used to detect regions with copy number variation for each marker, with the goal of deriving a list of regions where adjacent markers share the same copy number. PGS offers two different algorithms for region (CNV) detection.

At their core, both algorithms examine trends across multiple adjacent markers. The genomic segmentation algorithm identifies breakpoints in the data, changes in copy number between two neighbouring regions. This method is used to find segmentation according to three criteria:

i. Criteria one: Neighbouring regions should have statistically significantly different average intensity. This is calculated using two-tailed *t*-test and cut-off *p*-value defined by user (default *p*-value is 0.001).

ii. Criteria two: Determine breakpoints (region boundary), chosen to give optimal statistical significance (as defined by a lower *p*-value).

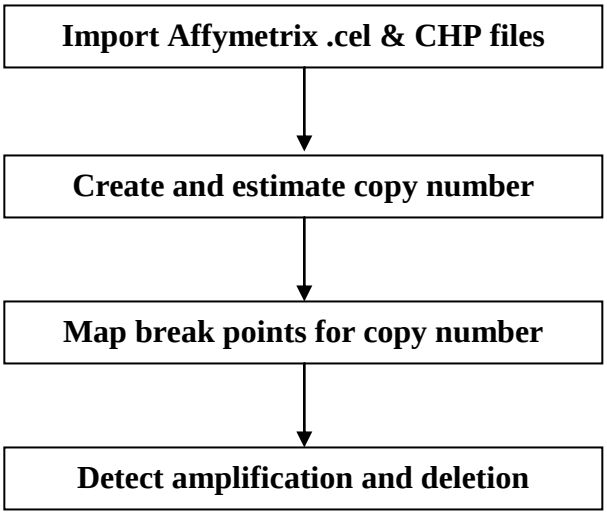


Figure 3: Overview of Partek® Genome Suite™ workflow

iii) Criteria three: Define the minimum number of genomic markers (with a default value of 10) that constitutes a "region". The HMM algorithm identifies discrete changes of whole number copy number states (e.g., 0, 1, 2 …) and attempts to find regions with those numbers of copies. We have used this segmentation model for our data analysis.

Using the above information, the next goal is to determine the amplifications and deletions across the genome. Two one-sided *t*-tests are used to compare the mean copy number variation predicted in the region with the expected normal copy number.

Finally PGS generates an output file entitled "segmentation.txt " containing the above mentioned "all values" for copy number variation. These "all values" are chromosome number, start position, stop position, cytoband, and copy number state. This study has used this output file for further analysis.

2.5.3 PennCNV

PennCNV (http://www.openbioinformatics.org/penncnv/) is a copy number variation detection software package that used SNP array data for variation identification. It analyses intensity data files from Affymetrix and Illumina arrays and applies the PennCNV Hidden Markov Model (HMM) algorithm to identify CNV calls for individuals that are different from segmentation-based algorithms.

PennCNV does not use Affymetrix raw .cel files directly. Affymetrix raw .cel files need a data processing protocol that converts it into an intensity file. PennCNV has its own data pre-processing protocol called PennCNV-Affy. that converts Affymetrix raw .cel files into a signal intensity file. Signal intensity files contain Log R Ratio (LRR) and B Allele Frequency (BAF) values that are used by PennCNV for CNV calling. The flowchart in Figure 4 is an overview of the PennCNV workflow.

There are two major steps in PennCNV for calling CNVs within data for an individual.

Step I: Conversion of raw Affymetrix .cel files into intensity data files

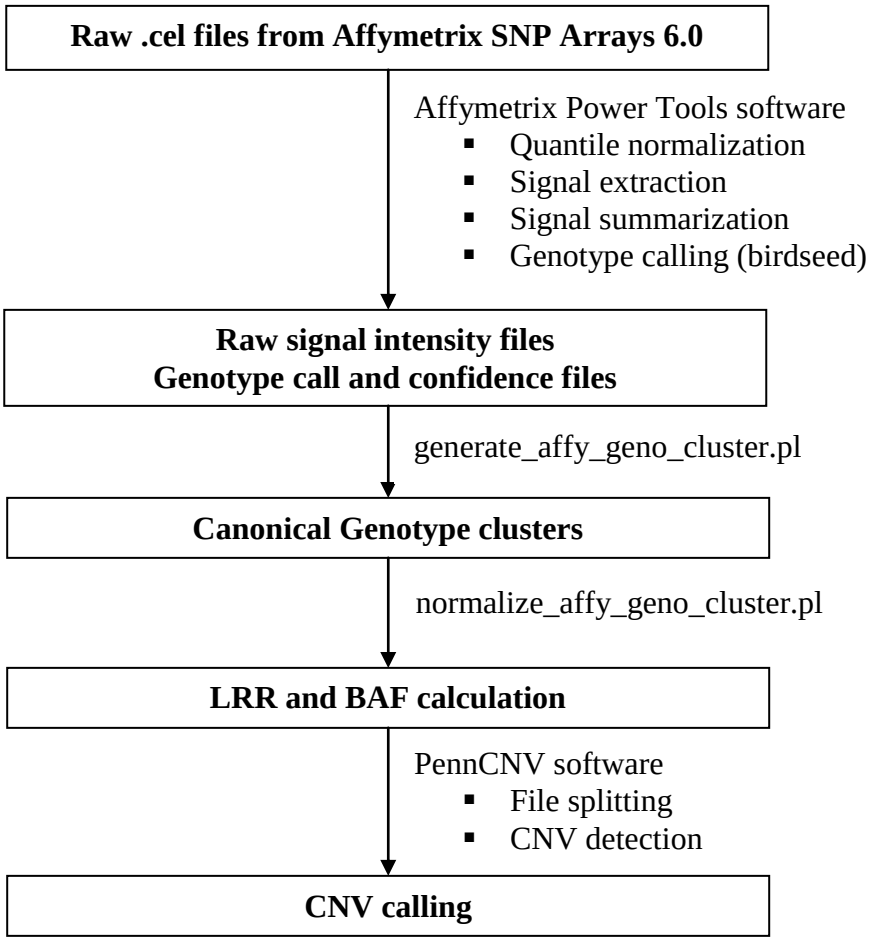


Figure 4: Overview of PennCNV workflow.

The main goal of this step is to generate normalized signal intensity text files from each .cel file which can be used subsequently by PennCNV. This step is divided into four sub steps:

i) Affymetrix Power Tools (APT) is used to generate genotyping calls from the raw .cel files using the Birdseed algorithm. For this step, we also need particular library files and annotation files for Affymetrix genome-wide 6.0 array.

ii) Extractions of allele-specific signal values for each SNP probe using APT. Applying quantile normalization algorithm generates a signal intensity value for A allele and B allele for each SNP probe. PennCNV-Affymetrix package use HapMap samples for quantile normalization. For this normalization process, our samples are more comparable to each others.

iii) Generation of canonical clustering information for each SNP and CNV marker called generates a genotype clusters file. The PennCNV- Affymetrix package uses annotation files that contain marker positions in the hg18 (NCBI build 36) human genome assembly. It also uses the sample's sex information to identify chromosome X and chromosome Y markers. The clustering file is also used to calculate Log R Ratio (LRR) and B Allele Frequency (BAF) values.

iv) This software creates Log R Ratio (LRR) and B Allele Frequency (BAF) value files for each marker for each individual. These files are used for CNV calling for each individual.

Step II: CNV calling using intensity data files:

Using Log R Ratio (LRR) and B Allele Frequency (BAF) values, PennCNV gives CNV calls for each individual. In this step, PennCNV uses a Hidden Markov Model (HMM) algorithm.

2.6 Summarizing the findings:

Using the above mentioned software all samples has been analyzed and results are presented as:

- Chromosome-wise distribution of copy number variations
- Size-wise of distribution of copy number variations
- Frequency of copy number variations: loss (deletion) and gain (duplication)
- Frequency of *de novo* copy number variation, of them frequency of Mitotic and Meotic CNVs
- Frequency of inherited copy number variations.

2.7 Finding the overlap with segmental duplication

I have used the Database of Genomic Variants (http://projects.tcag.ca/variation/) to calculate the overlap between copy number variation and segmental duplication.

2.8 Comparing different packages result

Finally, I have made input files, taking all output from different packages, to assess the comparison between these files in Affymetrix's Genotyping Console browser. I have tried to compare my results from different packages with different algorithms and have presented a comparison of these data using different methods, with the goal of assessing the accuracy of the detection of copy number variation algorithms that are commercially available.

Chapter 3: Results

In this study I used eight individual's microarray data for data analysis. Microarray data was generated using the Affymetrix Genome Wide Human SNP Array 6.0. The same set of microarray data was analysed by three different methods or algorithms. The methods are (i) Affymetrix® Genotyping Console™ (GTC), (ii) Partek® Genomics Suite™ (PGS) and (iii) PennCNV. All the outcomes from these three methods are presented below. The next step was to identify the overlap of outcomes among these three methods.

3.1 Method one: Genotyping Console 4.1 (GTC)

In the first method we used Affymetrix analysing package called Genotyping Console. These results are described one by one.

3.1.1 Distribution of CNV among family members according to size

Table 2 summarizes the size-wise distribution of CNVs for all eight individuals of the two families (Figure 1). It was observed that in both families, CNVs of size 100 to 200 kb were most common. There are only 3 CNVs identified with sizes less than 100 kb. On the other hand only 9 CNVs were greater than 20 Mb (large size). Only one member in the first family (I-1-1) has CNVs sized between 10 to 20 Mb. Not surprisingly most CNVs were in the range of 100 kb – 200 kb to be followed by 200 kb to 400 kb in every individual. Probably the most unusual observation was in II-1-1 who carried 119 CNVs in the rage of 100-200 kb.

Table 2. Distribution of CNVs among family members according to size using GTC

CNV Size	Family 1				Family 2			
	I-1-1	I-1-2	I-2-1	I-2-2	II-1-1	II-1-2	II-2-1	II-2-2
<=100 kb	0	0	0	0	2	0	0	1
>100 to 200 kb	17	18	15	20	119	50	24	24
>200 to 300 kb	11	6	4	10	25	6	13	9
>300 to 400 kb	5	5	6	5	11	3	1	4
>400 to 500 kb	2	0	2	2	6	1	1	2
>500 to 1000 kb	6	2	4	7	7	2	5	2
>1 to 10 Mb	9	4	5	3	5	2	4	5
>10 to 20 Mb	5	0	0	0	0	0	0	0
>20 Mb	3	0	0	0	2	0	2	2
Total	**58**	**35**	**36**	**47**	**177**	**64**	**50**	**49**

Numerical values in each cell of the table indicate how many CNVs of that particular size range were observed in that particular individual. Interestingly all eight individuals had an average number of CNVs in the range of 35-65, the father in family 2 had exceptionally high number of CNVs (177). This exception is characterised by 119 CNVs in this individual in this group characterised by a size range group of 100-200kb.

3.1.2 Origin of copy number variation across family members

Gains of CNVs were higher than loss for every individual in both families. Most of the CNVs identified were already reported in DGV (http://projects.tcag.ca/variation/) but few novel CNVs were also observed (Table 3). Father in family 2 had a large number of (125) of CNVs that were gain and also carried 42 novel CNVs.

3.1.3 Chromosome wise distribution of CNVs

Table 4 represents chromosome wise distribution of CNVs for all eight individuals studied. Interestingly there was no CNV in chromosome number 13 in family 1. The unusual exception in this analysis was the father (II-1-1) in family 2 who carries approximately 40 CNVs in chromosome 13. In fact this individual had a total of 177 CNVs that are distributed to a general increase in CNVs in most chromosomes.

Table 3. Copy number variation across family members using GTC

CNVs	Family 1				Family 2			
	I-1-1	**I-1-2**	**I-2-1**	**I-2-2**	**II-1-1**	**II-1-2**	**II-2-1**	**II-2-2**
No. of Loss	21	6	5	6	52	11	6	4
No. of Gain	37	29	31	41	125	53	44	45
Novel (absent in DGV)	1	1	0	2	42	6	1	0
Present in DGV	57	34	36	45	135	58	49	49
Total (for the individual)	**58**	**35**	**36**	**47**	**177**	**64**	**50**	**49**

Frequency of CNVs which are losses (deletion) or gains (duplication) and characterisation as present or absent from The Database of Genomic Variants (DGV).

Table 4. Chromosome wise distribution of CNV using GTC

Chr No.	Family 1				Family 2			
	I-1-1	**I-1-2**	**I-2-1**	**I-2-2**	**II-1-1**	**II-1-2**	**II-2-1**	**II-2-2**
1	4	2	2	2	11	2	6	6
2	4	2	5	5	11	2	2	3
3	1	4	2	3	7	4	4	2
4	4	3	2	4	8	1	3	2
5	0	0	0	0	12	0	2	1
6	0	0	0	0	10	2	0	0
7	2	6	3	4	10	5	3	4
8	1	0	3	3	7	3	3	3
9	1	1	2	3	4	4	3	4
10	2	1	1	1	3	5	1	1
11	3	1	2	1	5	1	2	2
12	0	1	0	1	4	3	0	1
13	0	0	0	0	40	1	1	0
14	4	6	3	5	4	6	5	3
15	4	3	1	3	2	6	6	8
16	2	1	2	3	9	1	1	1
17	4	1	2	3	8	2	2	1
18	0	0	0	0	1	0	0	1
19	0	1	0	0	7	3	2	1
20	0	0	0	0	0	0	0	1
21	1	2	2	3	2	3	2	1
22	3	0	3	2	3	1	1	1
X	18	0	1	1	9	9	1	2
Total	**58**	**35**	**36**	**47**	**177**	**64**	**50**	**49**

Chromosome specific distribution of *de novo* (present in twin(s) and absent in parents) and *inherited* (present in at least one parent) CNVs in family 1 and family 2. Chr. No = Chromosome number

3.1.4 *De novo* CNVs identified with GTC:

De novo CNVs were defined as those that were not present in parents but present in either or both member of twin pairs. Table 5 describes the total number of *de novo* CNVs detected by GTC in twins of family 1 and 2. This method detected 15 CNVs in family 1 and 25 CNVs in family 2. Twin 1 in family 1 and 2 had 3 and 8 *de novo* CNVs respectively. Same way twin 2 of both families had 7 and 8 *de novo* CNVs respectively. Both twin share 5 and 9 CNVs in family 1 and 2 respectively. Of these two novel CNVs in family 1 in chromosomal position 8q11.21 and 14q32.11 were identified in family 1 and linked with three genes KIAA0146, PSMC1 and C14orf102. (Appendix Table 1). There was only one novel *de novo* CNV detected in family 2 which was in chromosomal location 19q13.41 and in linkage with two genes ZNF331 and DPRX. (Appendix Table 2)

Table 5: Identified *de novo* CNVs using GTC

***De novo* CNV present in**	**Family 1**	**Family 2**
Twin1 (only)	3	8
Twin2 (only)	7	8
Twin1 & Twin2	5	9
Total	15	25

3.1.5 Probe assessment of *de novo* mitotic CNV in family 1

Table 6 represents the probe distribution pattern to identify *de novo* CNVs. There were 10 *de novo* CNVs present in only one of the two twins in family 1. Most of the *de novo* CNVs were identified with both SNP and CN probes except two CNVs. CNVs at chromosomal locations 8p23.1 and 9p11.2were detected with CN probes only (asterisk in the first column). Identified CNVs had more CN probes than SNP probes. CNV at 14q32.11 contain 32 SNP probes and 27 CN probes on that region. The signals at those probes had allowed identification of gains/losses listed in the table.

Table 6: Probe distribution across the *de novo* mitotic CNV region in Family 1

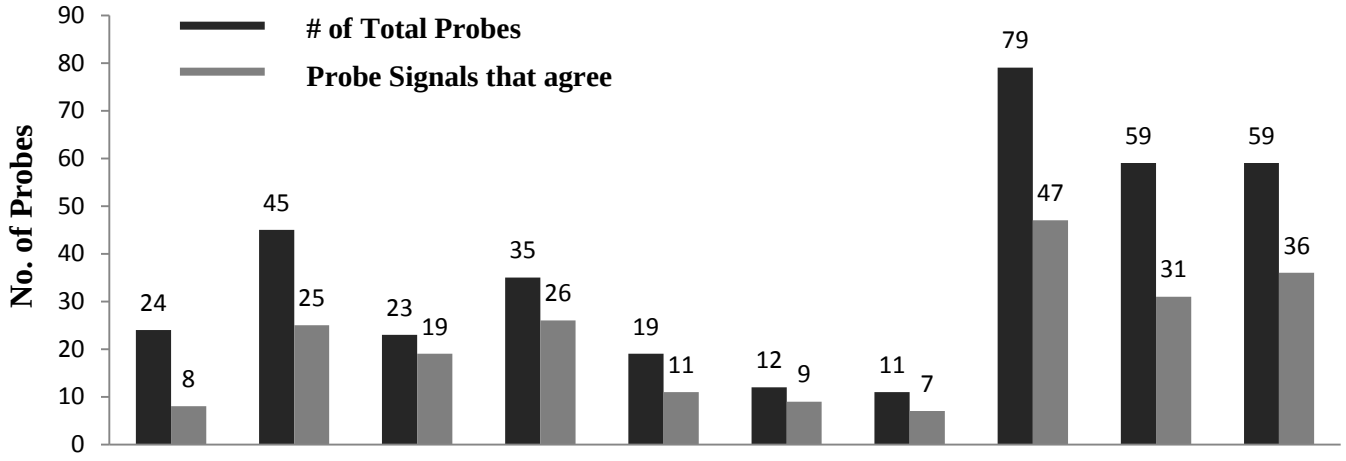
Sl. No.	Present in	CNV Position	Status	# of Total Probe	# of SNP Probe	# of CN Probe
1	I-2-1	1p36.13	Gain	24	2	22
2	I-2-2	4q28.3	Gain	45	11	34
3*	I-2-1	8p23.1	Loss	23	0	23
4	I-2-2	8q11.21	Loss	35	5	30
5*	I-2-1	9p11.2	Gain	19	0	19
6	I-2-2	9p13.1	Gain	12	3	9
7	I-2-2	9q12	Gain	11	4	7
8	I-2-2	12p13.31	Gain	79	24	55
9	I-2-2	14q32.11	Loss	59	32	27
10	I-2-2	21q11.2	Gain	59	11	48

3.1.6 Distribution of probes used to detect CNVs and number of probe signals those agreed with loss or gain calls in family 1

Figure 5 is the graphical presentation of the distribution of probes those were used in Affymetrix array to detect CNV (black bar) and the number of probes that followed the expected pattern of *de novo* CNVs (grey bar). The table underneath of the chart shows the percentage of the probes that satisfies the presence of CNVs along with chromosomal location. This is a graphical presentation of Table 6. It was interesting to observe that except one CNV at 1p36.13, all other CNVs were detected by more than 50% of the probes used.

3.1.7 Graphical representation of probe intensity in log2 ratio for family 1

The following line charts graphically presents the gain or loss of particular CNV present in one twin but absent from co-twin. Two line charts (Figure 6 (a) and (b)) represent the log2 ratio intensity values of each probe in the twin sample at the chromosomal locations 8p23.1 and 9p13.1 respectively. The chart represents the probe intensity values for loss and gain of CNVs in twin two in family number 1. The line charts characterised by black continuous lines represent the probes intensity for twin (I-2-1**)** with either loss or gain of copy number variation and light dotted line represent co-twin (I-2-2**)** probe intensity value.



Location	1p36.13	4q28.3	8p23.1	8q11.21	9p11.2	9p13.1	9q12	12p13.31	14q32.11	21q11.2
Percent	33.33	55.56	82.61	74.29	57.89	75	63.64	59.49	52.54	61.02

Figure 5: Graphical representation of the distribution of total number of probes used to detect CNVs and number of probe signals that agreed with loss or gain calls in family 1

8p23.1 – Loss (continuous line)

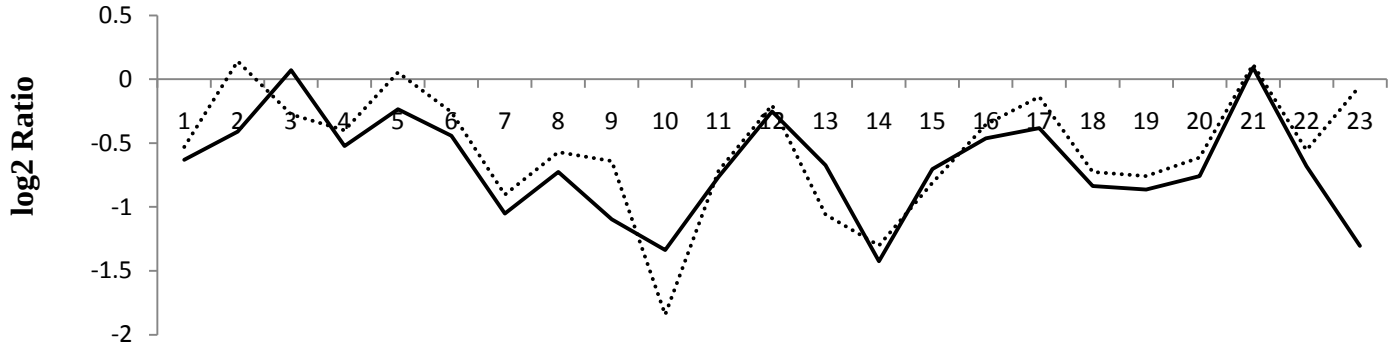


Figure 6(a): Pattern of distribution of the log2 ratio between twins for a loss of CNV. Dotted line represents data for twin 2 and continuous line represents data for twin 1 with loss of CNV.

9p13.1 – Gain (continuous line)

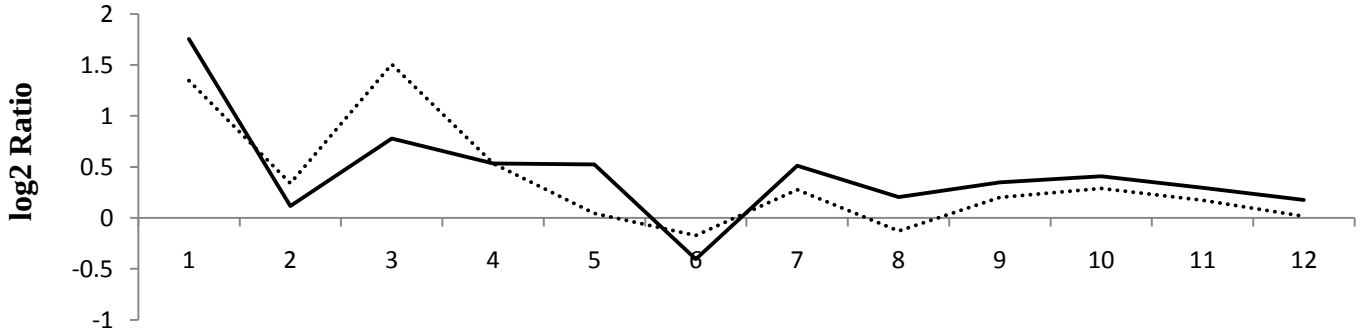


Figure 6(b): Pattern of distribution of Log2 ratio between twins for a gain of CNV. Dotted line represents data for twin 1 and continuous line represents data for twin 2 with gain of CNV.

3.1.8 Probe assessment of *de novo* mitotic CNV in family 2

Table 7 represents the probe distribution pattern used to detect *de novo* CNVs in family 2. There were 16 (present only one twin) *de novo* CNVs in family 2, identified with several copy number (CN) and single nucleotide change (SNP) probes. Most of the *de novo* CNVs were identified with both SNP and CN probes except one CNV. CNVs at chromosomal locations 15q13.2 were detected with CN probes only (asterisk in the first column). All identified CNVs had more CN probes than SNP probes (almost twice) excluding one CNV (15q13.1).

Table 7: Probe distribution across the *de novo* mitotic CNV region in Family 2

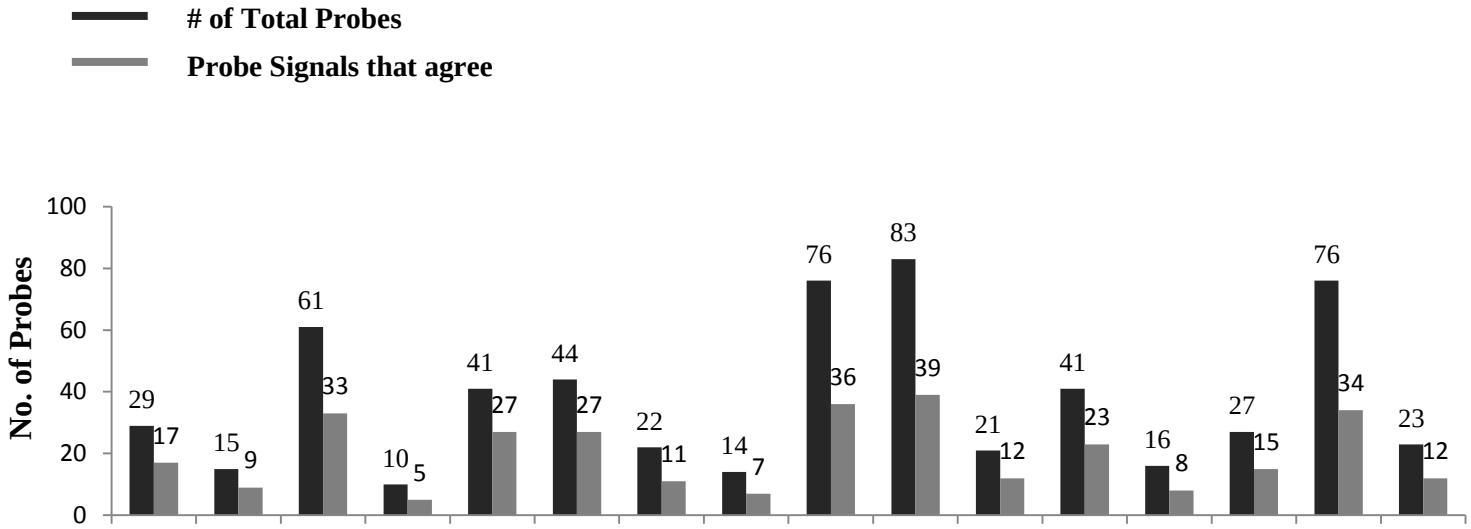
Sl. No.	Present in	CNV Position	Status	# of Total Probe	# of SNP Probe	# of CN Probe
1	II-2-2	1q21.1	Gain	29	1	28
2	II-2-2	1q43	Gain	15	4	11
3	II-2-1	3q21.2	Gain	61	16	45
4	II-2-1	4p11	Gain	10	2	8
5	II-2-1	5p13.3	Gain	41	9	32
6	II-2-1	7q11.21	Gain	44	1	43
7	II-2-2	7q35	Gain	22	1	21
8	II-2-2	9q12	Gain	14	2	12
9	II-2-2	12p13.31	Gain	76	24	52
10	II-2-1	13q11	Gain	83	24	59
11	II-2-2	15q11.1	Gain	21	2	19
12	II-2-2	15q13.1	Gain	41	21	20
13*	II-2-1	15q13.2	Gain	16	0	16
14	II-2-1	17p11.1	Gain	27	4	23
15	II-2-1	19q13.41	Gain	76	25	51
16	II-2-2	20q11.1	Gain	23	6	17

3.1.9 Distribution of probes used to detect CNVs and number of probe signals those agreed with loss or gain calls in family 2

The following chart is the graphical presentation of Table 5. Figure 7 presents the distribution of probes that were used in Affymetrix array to detect CNV (black bar) and the number of probes that agree or confirm the presence of that particular CNV (grey bar). The attached table represents the percentage of the same probe distribution that satisfies the presence of CNVs along with chromosomal location. It was interesting to observe that each CNV had been detected by more than 44% of the probes.

3.1.10 Graphical representation of probe intensity in log2 ratio:

The following line chart (Figure 8 (a) and (b)) is the graphical presentation of gain or loss of particular CNV present in one twin but absent from co-twin. Two line charts representing the log2 ratio value of each probe present in copy number variation in one twin sample which located at chromosomal locations 5p13.3 and 7q11.21 respectively. Both were gain of copy number variations present in one twin (I-2-1) between monozygotic twin pair and represented in black continuous line whereas light dotted line represent co-twin (I-2-2) probe intensity value. This graph supports the presence of gain of copy number variation depending of log2 ration intensity value.



Location	1q21.1	1q43	3q21.2	4p11	5p13.3	7q11.21	7q35	9q12	12p13.31	13q11	15q11.1	15q13.1	15q13.2	17p11.1	19q13.41	20q11.1
Percent	58.62	60	54.1	50	65.85	61.36	50	50	47.37	46.99	57.14	56.1	50	55.56	44.74	52.17

Figure 7: Probe frequencies and percent of probes that support the loss or gain in *de novo* CNVs in family 2

5p13.3 – Gain (Continuous line)

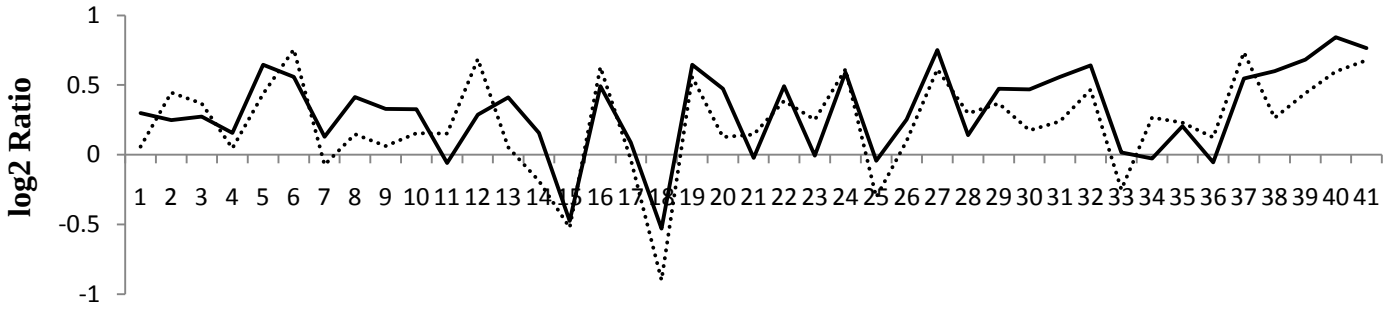


Figure 8(a): Pattern of distribution of the log2 ratio between twins in family 2. Dotted line represents data for twin 2 and continuous line represents data for twin 1 with gain of CNV.

7q11.21 – Gain (Continuous line)

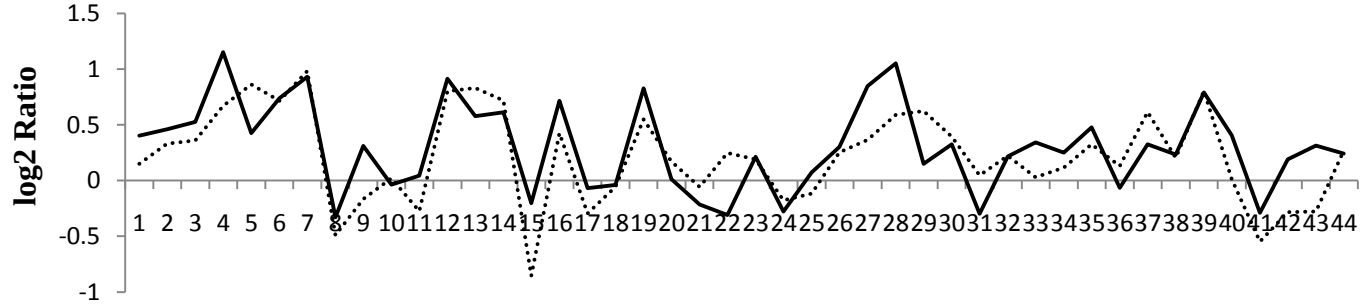


Figure 8(b): Pattern of distribution of Log2 ratio between twins in family 2. Dotted line represents data for twin 2 and continuous line represents data for twin 1 with gain of CNV.

3.1.11 *Inherited* CNV

Table 8 summarises the exact number of inherited CNVs in twins of both families. The table describes the pattern of inheritance or the number of CNVs twins inherited exclusively from either father or mother or from both parents. The later does not carry much information since those CNVs were common to all members in the family. It was interesting to observe that twin 1 in family 1 and twin 2 in family 2 inherited their CNVs exclusively from father. The overlapped genes with these CNVs are described in Appendices tables 3 and 4 for family 1 and 2 respectively.

3.1.12 Summary of GTC analysis

1. Frequencies of CNVs for individuals in two monozygotic twin families were between 35 and 64 (except one individual).
2. Most of the CNVs were of size range 100 and 200 kb.
3. CNVs were distributed over all whole genome (chromosomes, 1 to X) for all 8 individuals.
4. Observed CNVs were more gain than loss in all members.
5. Twin did not inherit identical CNVs from parents in both families and most of the *de novo* CNVs are present only in one twin.

Table 8: Inherited CNV using GTC

	Family 1			Family 2		
Inherited in	**Father (only)**	**Mother (only)**	**Father & Mother**	**Father (only)**	**Mother (only)**	**Father & Mother**
Twin1 (only)	2	0	0	2	3	1
Twin2 (only)	2	2	2	3	0	1
Twin1 & Twin2	8	6	11	11	4	10

3.2 Method two: Partek® Genome Suite™ (PGS)

In the second method we used PGS for the CNV analysis for same eight individual's data from two families. All outcomes were tabulated below.

3.2.1 Size wise distribution of CNVs among family members

Table 9 represents the size distribution of CNVs detected by PGS in all eight individuals in family 1 and 2. It was interesting to observe that PGS can detect large (>500kb -20 Mb) CNVs. CNV of size less than 500 kb was almost absent in both families. Father of family 2 had exceptionally large number of CNVs characterised by 198 CNVs of size 10-20 Mb and 541 CNVs of size more than 20 Mb. The Mother of the same family also had 113 CNVs of size more than 20 Mb.

Table 9: Size distribution of CNV using PGS

CNVs Size	Family 1				Family 2			
	I-1-1	I-1-2	I-2-1	I-2-2	II-1-1	II-1-2	II-2-1	II-2-2
<=100 kb	0	0	0	0	0	0	0	0
>100 to 200 kb	0	0	0	0	0	0	0	0
>200 to 300 kb	0	0	0	0	0	0	0	0
>300 to 400 kb	0	0	0	0	0	0	0	0
>400 to 500 kb	0	0	0	0	0	0	0	0
>500 to 1000 kb	0	0	1	0	1	1	1	1
>1 to 10 Mb	26	22	16	15	31	31	17	22
>10 to 20 Mb	14	10	14	15	198	40	8	13
>20 Mb	24	17	24	21	541	113	30	29
Total	**64**	**49**	**55**	**51**	**771**	**185**	**56**	**65**

3.2.2 Origin of copy number variation across individual family members

Table 10 represents origin of CNVs identified in all family members of family 1 and 2. Interestingly PGS detected higher gains of CNVs in father and mother of family 2; in them, the father contained 691 and mother contained 134 CNVs respectively.

3.2.3 Chromosome wise distribution of CNVs

Table 11 represents chromosome wise distribution of CNVs in all eight individuals of family 1 and 2. PGS detected almost same number of CNVs in all individuals except in family 2 where the father contained exceptionally high number of CNVs (771) and the mother had 185 CNVs. Another observation was that the father of the same family contained 54 CNVs only in chromosome X which were significantly more than any other individual in both families.

Table 10: Copy number variation across family members using PGS

CNVs	Family 1				Family 2			
	I-1-1	**I-1-2**	**I-2-1**	**I-2-2**	**II-1-1**	**II-1-2**	**II-2-1**	**II-2-2**
No. of Loss	39	35	33	27	80	51	36	46
No. of Gain	25	14	22	24	691	134	20	19
Total (for the individual)	**64**	**49**	**55**	**51**	**771**	**185**	**56**	**65**

Table 11: Chromosome wise CNV distribution using PGS

Chr No.	Family 1				Family 2			
	I-1-1	**I-1-2**	**I-2-1**	**I-2-2**	**II-1-1**	**II-1-2**	**II-2-1**	**II-2-2**
1	3	6	8	8	62	19	5	6
2	5	2	6	4	66	15	5	6
3	8	4	4	4	53	9	3	3
4	7	5	6	6	54	10	4	6
5	1	5	4	4	45	16	4	3
6	4	2	2	1	51	17	6	7
7	3	3	3	3	38	10	2	3
8	4	5	4	4	45	12	2	5
9	2	1	1	2	33	3	1	0
10	2	1	3	2	40	10	2	4
11	2	1	2	2	31	5	3	4
12	4	3	1	1	38	12	3	2
13	1	0	1	2	38	5	1	1
14	0	2	2	2	23	8	2	1
15	3	2	1	1	18	7	1	1
16	3	1	1	1	23	7	3	2
17	2	2	0	0	15	2	1	1
18	1	1	2	0	16	1	1	1
19	0	1	1	0	10	1	1	2
20	0	0	0	0	12	6	2	2
21	0	0	0	0	7	1	0	0
22	2	1	1	1	2	5	1	3
X	7	1	2	3	51	4	3	2
Total	**64**	**49**	**55**	**51**	**771**	**185**	**56**	**65**

3.2.4 *De novo* CNVs identified with PGS

Table 12 describes the total number of *de novo* CNVs detected by PGS in twins of family 1 and 2. Twin 1 of both families had 6 *de novo* CNVs exclusive to them; on the other hand twins of family 1 had 4 and in family 2, 12 CNVs respectively. Both twins shared 14 and 11 CNVs in family 1 and 2 respectively. Out of 24 in family 1 and 29 in family 2 there were few known genes mapped near or within the CNVs of both families (Appendices Table 5 and 6).

3.2.5 *Inherited* CNVs identified using PGS

Table 13 describes the pattern of inheritance and total number of inherited CNVs in eight individuals of family 1 and 2. The pattern of inheritance detected by PGS was quite different. This method detected twin 2 of family 1 and twin 1 in family 2 inherited their CNVs exclusively from father and mother respectively.

Table 12: Identified *de novo* CNVs using PGS

***De novo* CNV present in**	**Family 1**	**Family 2**
Twin1 (only)	6	6
Twin2 (only)	4	12
Twin1 & Twin2	14	11
Total	24	29

Table 13: Identified *inherited* CNVs using PGS

	Family 1			Family 2		
Inherited in	**Father (only)**	**Mother (only)**	**Father & Mother**	**Father (only)**	**Mother (only)**	**Father & Mother**
Twin1 (only)	2	1	0	0	2	0
Twin2 (only)	2	0	0	1	1	1
Twin1 & Twin2	12	9	2	12	18	5

3.2.6 Summary of PGS analysis

1. Almost in all individuals except two, frequencies of CNVs were between 35 and 64.
2. PGS detected CNVs of larger size (>1 Mb) in all individuals. Smaller size CNVs were almost absent.
3. Detected CNVs were distributed in whole genome (chromosomes, 1 to X) for all 8 individuals.
4. There were more losses than gains of CNVs in all family members.
5. Twin did not inherit identical CNVs from parents in both families and *de novo* CNVs observed in twins varied significantly from co-twin.

3.3 Method three: PennCNV

In the third method we use PennCNV. The same raw data were analysed by PennCNV and results are tabulated below.

3.3.1 Distribution of CNVs among family members according to size

Table 14 represents the size wise distribution of CNVs in all eight individuals of family 1 and 2. Like PGS, PennCNV also detected drastically higher number of CNVs of size 1-20 Mb. Very few CNVs of size range 200-1000 kb were detected. With this method as well father in family 2 had very large number (131) of CNVs of size more than 20 Mb. All other individuals carried total number CNVs of range 31-49.

Table 14: Size distribution of CNV using PennCNV

CNVs Size	Family 1				Family 2			
	I-1-1	I-1-2	I-2-1	I-2-2	II-1-1	II-1-2	II-2-1	II-2-2
<=100 kb	0	0	0	0	0	0	0	0
>100 to 200 kb	0	0	0	0	0	0	0	0
>200 to 300 kb	0	0	0	0	0	1	0	0
>300 to 400 kb	1	0	1	1	2	0	1	2
>400 to 500 kb	0	0	0	1	0	0	0	0
>500 to 1000 kb	0	0	0	0	0	0	1	1
>1 to 10 Mb	17	13	10	11	10	13	14	17
>10 to 20 Mb	3	3	4	3	35	3	3	4
>20 Mb	23	18	16	14	161	32	12	15
Total	**44**	**34**	**31**	**30**	**208**	**49**	**31**	**39**

3.3.2 Origin of copy number variation across family members

Table 15 summarises the origin of CNVs in all members of family 1 and 2. Gain or loss of CNVs detected with PennCNV was almost similar in all members of both families except the father in family 2. This observation was characterised by 66 losses and 142 gains of CNVs. Twin 1 of family 2 also had more loss of CNVs (22) than gain (9).

3.3.3 Chromosome wise distribution of CNVs

Table 16 summarises the chromosome wise distribution of CNVs obtained with PennCNV analysis. Like GTC, PennCNV also detected on exceptionally higher number (49) of CNVs in chromosome 13 in the father of family 2. Another interesting observation is that PennCNV does not identify any CNV in the X chromosome for any member of both families.

Table 15: Copy number variation across family members using PennCNV

CNVs	Family 1				Family 2			
	I-1-1	**I-1-2**	**I-2-1**	**I-2-2**	**II-1-1**	**II-1-2**	**II-2-1**	**II-2-2**
No. of Loss	21	11	14	18	66	29	22	29
No. of Gain	23	23	17	12	142	20	9	10
Total (for the individual)	**44**	**34**	**31**	**30**	**208**	**49**	**31**	**39**

Table 16: Chromosome wise CNV distribution using PennCNV

Chr No.	Family 1				Family 2			
	I-1-1	**I-1-2**	**I-2-1**	**I-2-2**	**II-1-1**	**II-1-2**	**II-2-1**	**II-2-2**
1	8	3	4	5	14	8	3	3
2	3	3	4	3	13	3	2	2
3	4	5	2	3	10	3	1	1
4	8	1	3	3	12	2	2	4
5	0	5	2	2	13	4	0	1
6	1	0	0	0	11	4	4	6
7	2	2	4	3	16	3	3	3
8	2	2	2	3	10	2	2	4
9	1	1	1	0	10	1	0	0
10	0	0	1	0	10	3	1	1
11	2	2	2	3	5	2	2	2
12	3	2	0	0	8	1	1	0
13	1	0	0	0	49	0	1	1
14	1	2	2	1	6	2	1	2
15	3	1	1	2	5	2	1	1
16	2	0	2	1	3	3	2	2
17	1	2	0	0	3	2	0	0
18	1	1	0	0	1	0	1	1
19	0	1	0	0	1	2	3	2
20	0	1	0	0	5	2	1	2
21	0	0	0	0	2	0	0	0
22	1	0	1	1	1	0	0	1
X	0	0	0	0	0	0	0	0
Total	**44**	**34**	**31**	**30**	**208**	**49**	**31**	**39**

3.3.4 *De novo* CNVs

Table 17 describes the total number of *de novo* CNVs detected by PennCNV in twins of family 1 and 2. Twin 1 of both families had 3 and 2 *de novo* CNVs exclusively and twin 2 of both families had 4 and 8 CNVs respectively. Both twins shared 6 and 7 CNVs in family 1 and 2 respectively. Out of 13 in family 1 and 17 in family 2 there were few known genes mapped near or in the CNVs of both families (Appendices Table 9 and 10). CNV in 8p23.1 in twin 2 of family 2 contain exceptionally high number of gene of beta defensin (DEFB) family (Appendix Table 10).

3.3.5 *Inherited* CNVs identified with PennCNV

Table 18 summarises the total number of inherited CNVs and their pattern of inheritance in both families. Like method 2, PennCNV also detected that twin 2 in family 1 and twin 1 in family two inherited their CNVs from corresponding father and mother.

Table 17: Identified *de novo* CNVs using PennCNV

***De novo* CNV present in**	**Family 1**	**Family 2**
Twin1 (only)	3	2
Twin2 (only)	4	8
Twin1 & Twin2	6	7
Total	13	17

Table 18: Identified *inherited* CNVs using PennCNV

	Family 1			Family 2		
Inherited in	**Father (only)**	**Mother (only)**	**Father & Mother**	**Father (only)**	**Mother (only)**	**Father & Mother**
Twin1 (only)	4	1	0	1	0	0
Twin2 (only)	2	0	0	2	1	0
Twin1 & Twin2	5	7	4	8	12	0

3.3.6 Summary of PennCNV analysis

1. Observed frequencies of CNVs in all individuals range from 35-64 except one individual.
2. PennCNV detected CNVs of larger size (>1 Mb) in all individuals. Smaller size CNVs were almost absent.
3. PennCNV was unable to detect any CNV in X chromosome.
4. This method detected almost equal loss and gain of CNVs in all individuals.
5. Twin did not inherited identical CNVs from parents in both families and *de novo* CNVs observed in twin varied significantly from co-twin.

3.4 Comparison of CNV results obtained with three different methods:

Data generated and analysed with Affymetrix platform were reanalysed with multiple methods in order to

I) validate the results of Affymetrix® Genotyping Console™
II) identify any new structural variation
III) compare the similarities and dissimilarities among three CNV calling algorithms

3.4.1 Frequency of copy number variations using three different packages

Same raw data were analysed with three different packages GTC, PGS and PennCNV. The total numbers of CNVs obtained with each packages for all individuals were tabulated in Figure 9(a) and 9(b). It was interesting to observe that total number of CNVs obtained with PGS scored maximum and with PennCNV scored minimum in all cases except individual II-1-1 in family 2.

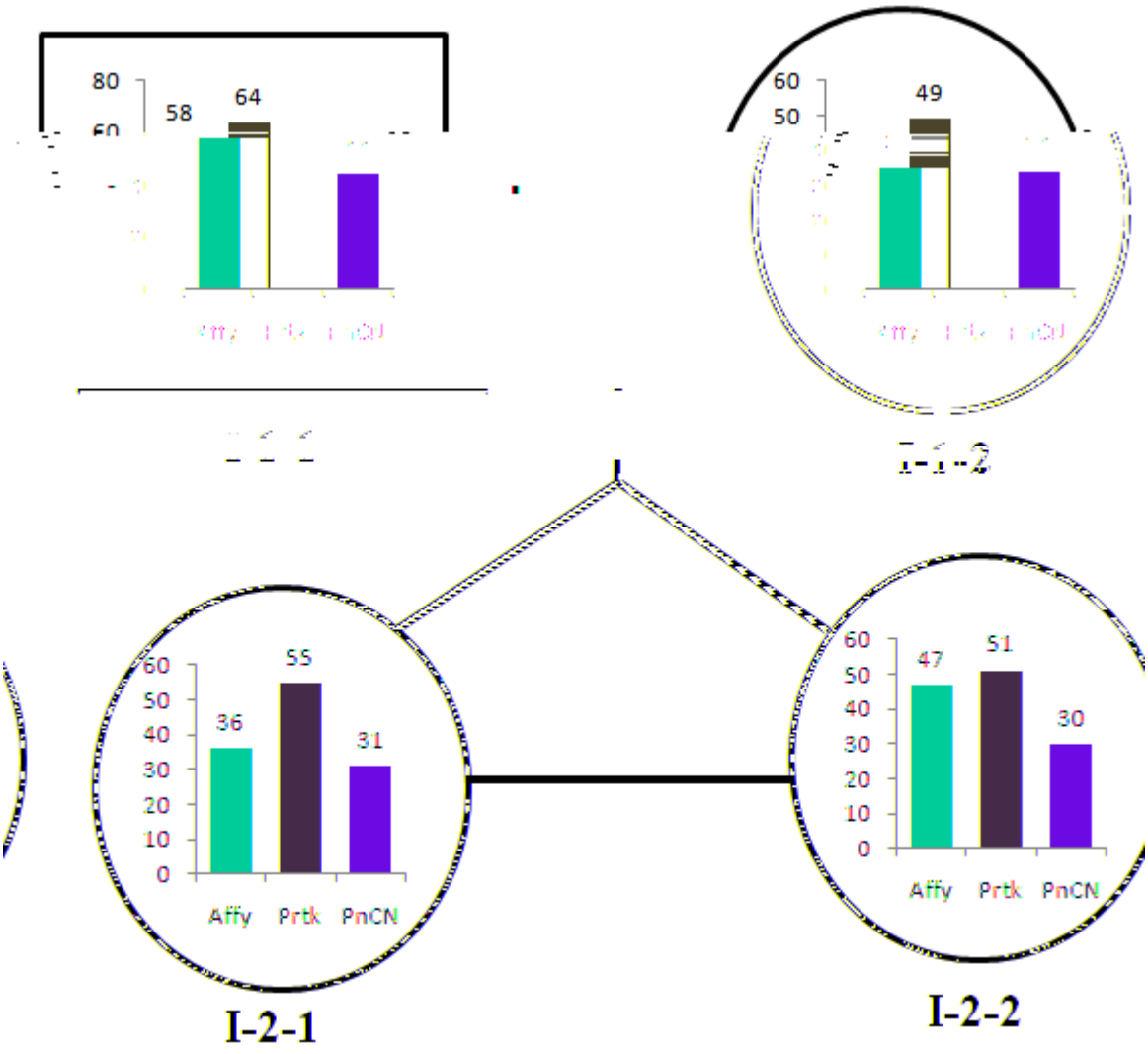


Figure 9(a): Total number of copy number variation identified by three different CNV calling algorithms for all family members of family 1. Purple, olive and orange color bar represents data originated by GTC, PGS and PennCNV respectively.

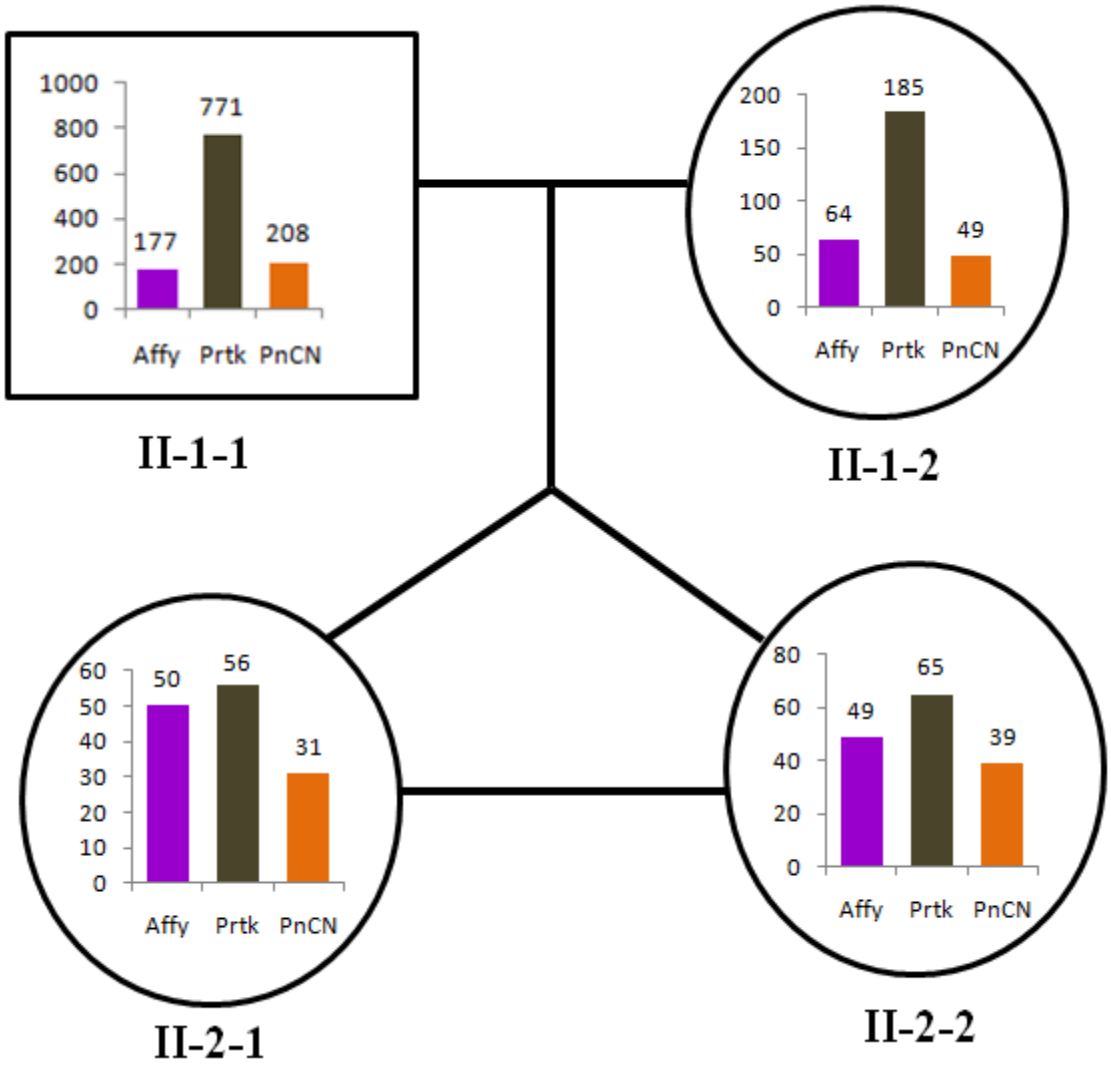


Figure 9(b): Total number of copy number variation identified by three different CNV calling algorithms for all family members of family 2. Purple, olive and orange color bar represents data originated by GTC, PGS and PennCNV respectively.

3.4.2 Comparison of percent distribution of Gain and Loss of CNV

Comparison of percentage of gain and loss of copy number variation for each individual in two families (Figure. 10(a) and (b)) represented in bar carts within pedigree. It was interesting to observe that gain of CNVs was maximally detected with GTC compared to PGS and PennCNV.

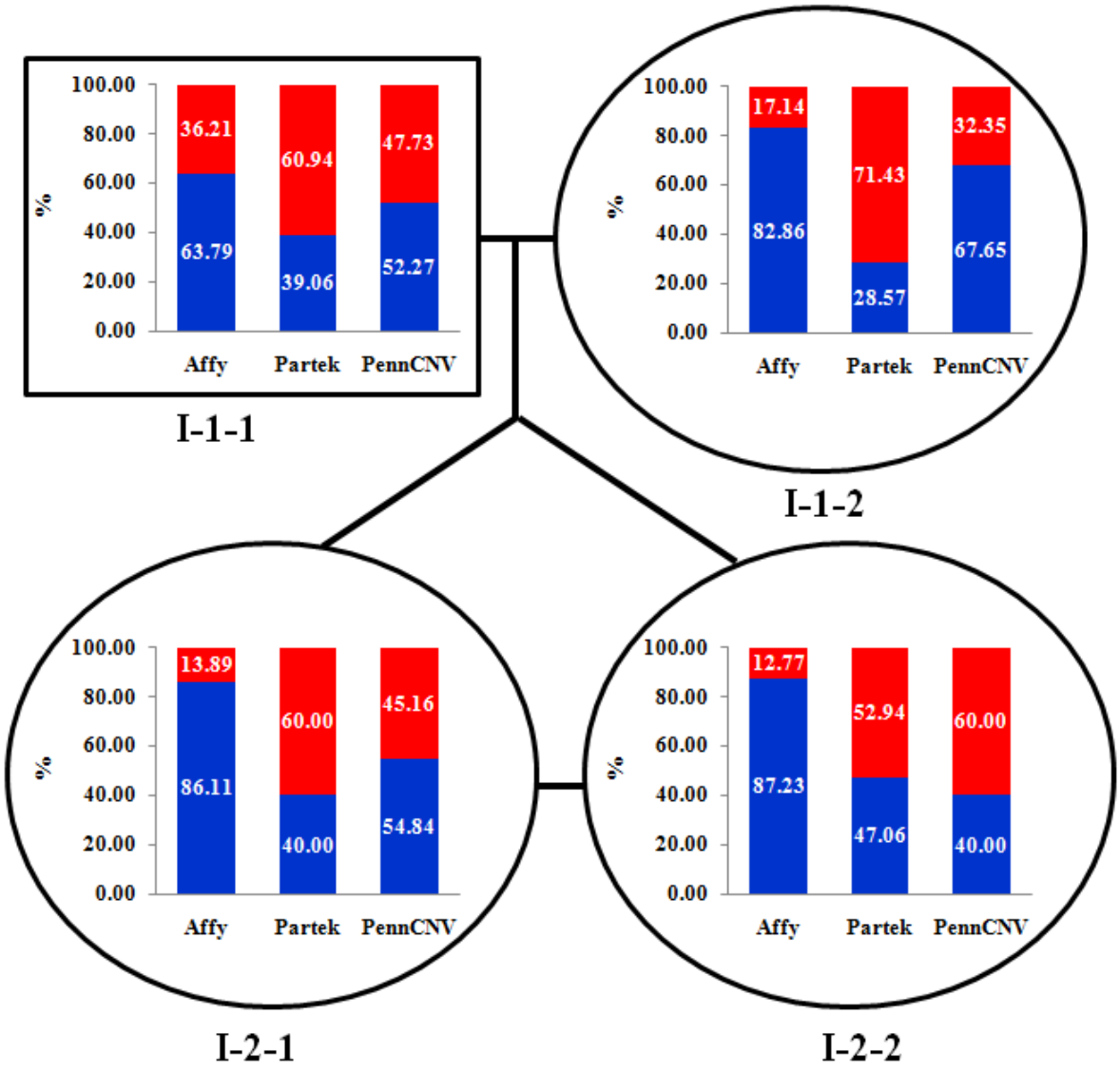


Figure 10(a) : Percentages of gains and losses of CNVs obtained by three different CNV calling algorithms in family 1. Blue represents gain and red represents loss of CNV.

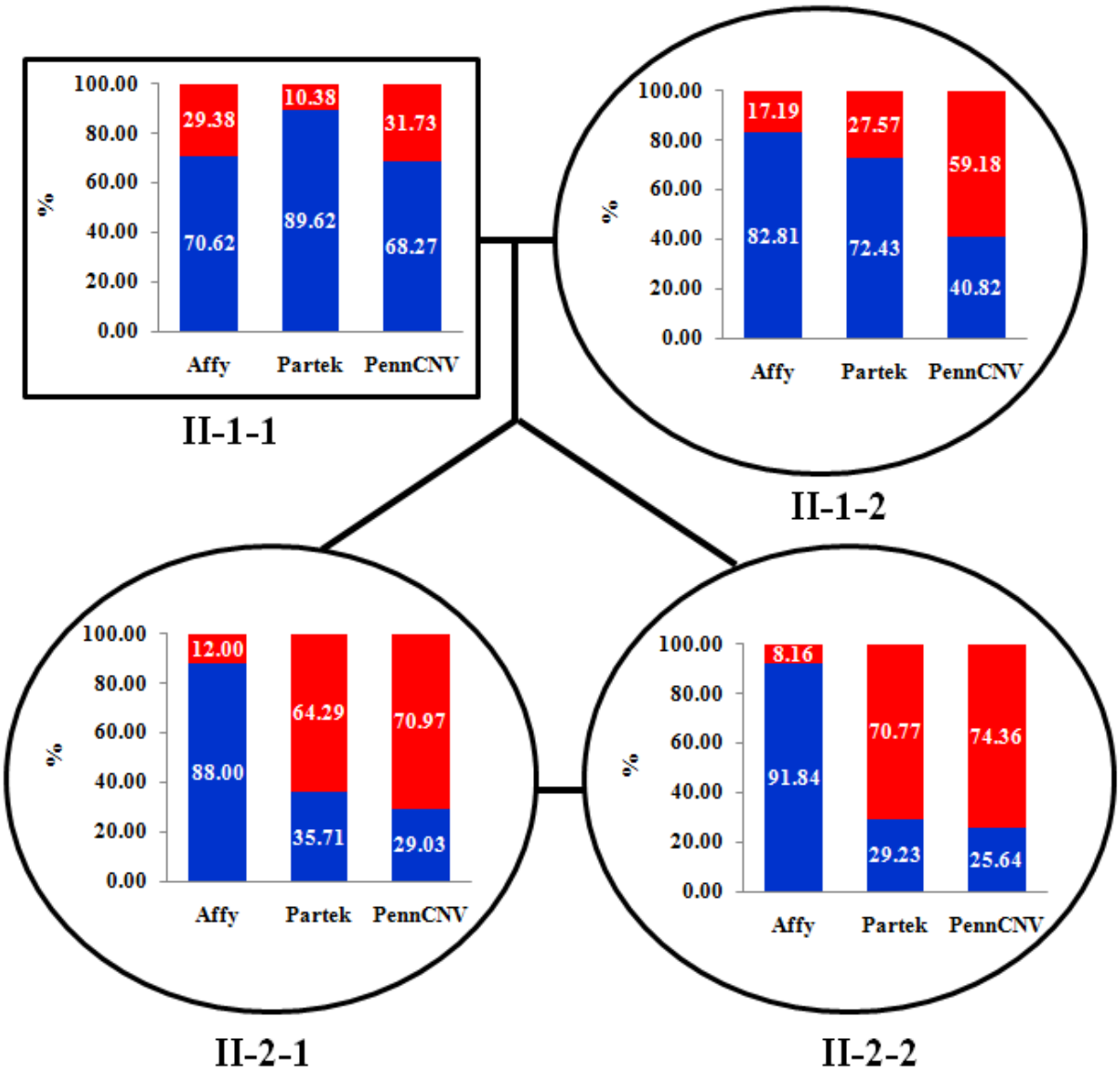


Figure 10(b) : Percentages of gains and losses of CNVs obtained with three different CNV calling algorithms in family 2. Blue represents gain and red represents loss of CNV.

3.4.3 Size distribution of copy number variations identified with different algorithms for each individual

Size of CNVs identified with different algorithms or packages differ significantly from each other. It was interesting to observe that both in family 1 (Figure 11(a)) and family 2 (Figure 11(b)), in all individuals, CNV of large size (1-20mb) were identified mostly by PGS and PennCNV; on the other hand smaller CNV (100-500 kb) identified significantly by GTC.

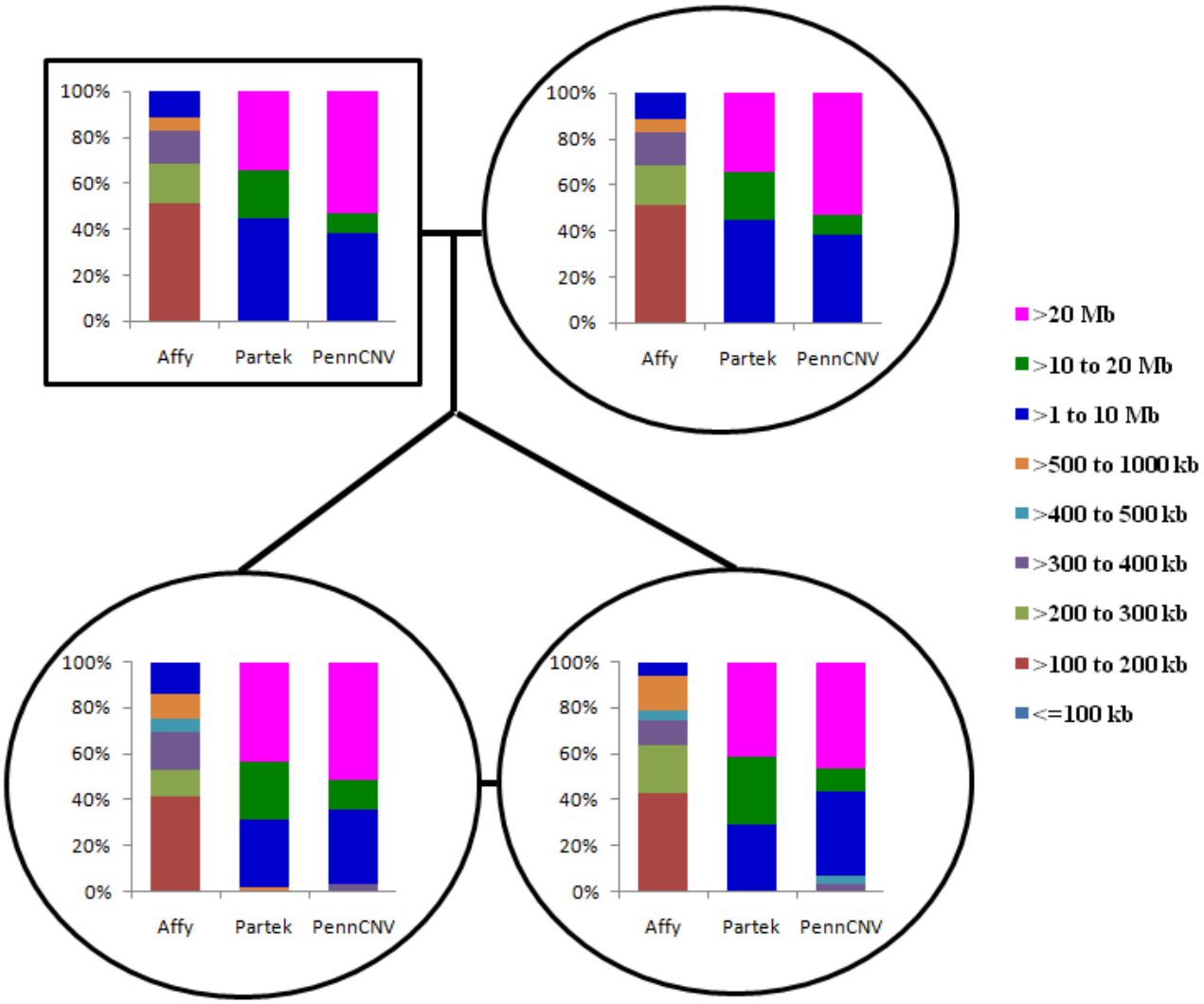


Figure 11(a): Size distribution of copy number variations identified with three different algorithms for each individual in family 1 presented in pedigree.

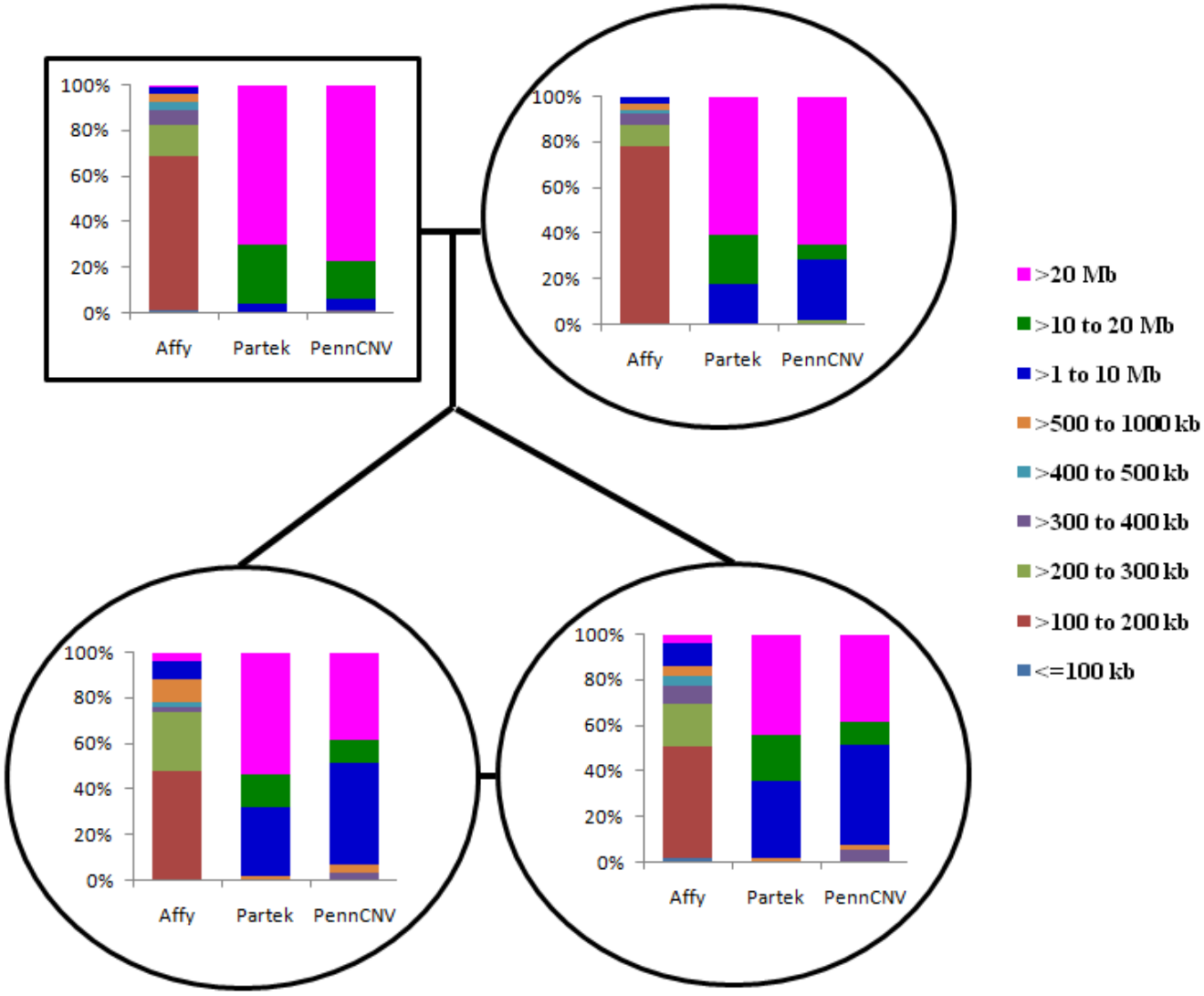


Figure 11(b): Size distribution of copy number variations identified with three different algorithms for each individual in family 2 presented in pedigree

3.5 Summary of CNVs identified with three different methods:

It was important to list any exclusive CNV identified with any particular package, or CNV identified with more than one package. Later CNVs are very crucial because they were captured with multiple softwares.

3.5.1 Summary of results in family 1:

Total, shared and exclusive number of *de novo* and *inherited* CNVs are presented with Venn diagram (Figure 12) and tabulated in Table 19. In total six CNVs which are approximately 14 percent were confirmed by all packages.

Table 19 is the list of 6 CNVs commonly identified by all three methods for family 1. All CNVs excluding (2p11.2) were overlapped with genomic regions. All these were inherited CNVs.

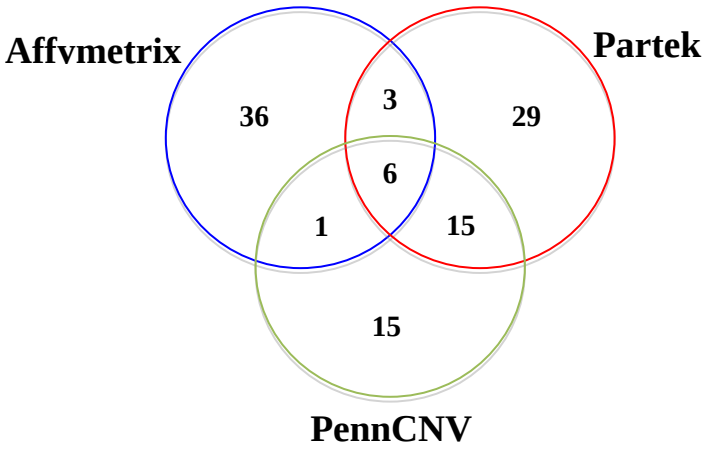


Figure 12: Venn diagram showing exclusive and overlapped CNVs identified with Affymetrix, Partek and PennCNV in family 1. Shared CNVs are shown in overlapped portion of diagram.

Table 19: Common CNVs identified by three methods for family 1

Sl. No.	Affymetrix-Partek-PennCNV	
	Location	Gene
1	2p11.2	
2	7q11.21	INTS4L1, ZNF92
3	8p23.1	LOC349196, DEF109P1B, DEFB103A, DEFB103B, SPAG11B, DEFB104B, DEFB104A, DEFB106B, DEFB106A, DEFB105B, DEFB105A, DEFB107A, DEFB107B, FAM90A7, SPAG11A, DEFB4, FAM66E, DEF109P1B
4	11p15.4	OR51A4, OR51A2
5	14q32.33	KIAA0125, ADAM6
6	22q11.21	RIMBP3C, RIMBP3B, HIC2, PI4KAP2, RIMBP3C, UBE2L3, POM121L8P

3.5.2 Summary of results in family 2:

Like family 1, total, shared and exclusive number of *de novo* and *inherited* CNVs scored with three different packages are presented with Venn diagram (Figure 16). Six CNVs which is around 11 percent were common to all packages in family 2.

Table 20 represents the list of 6 CNVs commonly identified by all three methods for family 2. First CNV (5p13.3) in the table was a *de novo* but rest were inherited CNVs. All of them were overlapped with genic regions.

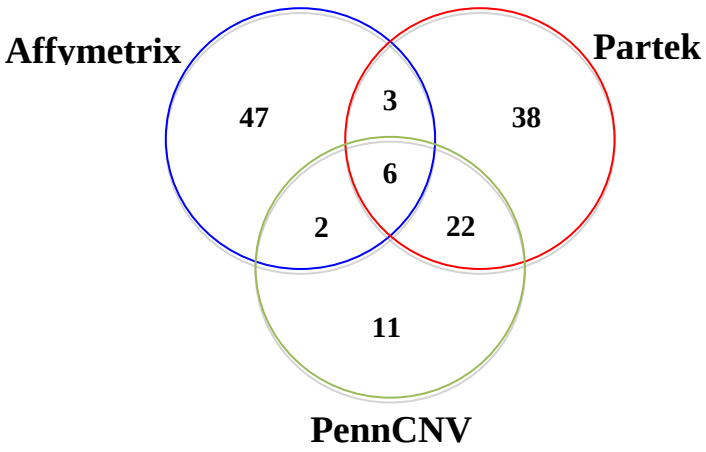


Figure 13: Venn diagram showing exclusive and overlapped CNVs identified with Affymetrix, Partek and PennCNV in family 2. Shared or commonly identified CNVs are shown in overlapped Venn diagrams.

Table 20: Common CNVs identified by three methods for family 2

.

Sl. No.	Affymetrix-Partek-PennCNV	
	Location	Gene
1	5p13.3	PDZD2, GOLPH3
2	1p21.1	AMY2B, LOC648740, AMY2A, AMY1A, AMY1C, AMY1B
3	3q12.2	GPR128, TFG
4	11p15.4	LOC650368, OR51A4, OR51A2
5	14q32.33	KIAA0125, ADAM6
6	15q25.3	AKAP13

3.5.3 Comparison between two families

This study identified six copy number variations those were confirmed by all the three methods for both families. One more interesting observation was that among these six CNVs, there were two common CNVs found in both families and both were inherited CNVs (Table 21). In family 1, CNV (11p15.4) were overlapped with gene region and common to all family members. That means this CNV was inherited from both parents to both twin pairs. But in case of family 2, it was inherited only from father to both twins. The second CNV (14q32.33) was also overlapped with the gene region and present in all family members for the both families.

Table 22 describes gene functions of the commonly identified gene for both families. Except olfactory receptor family 51 genes (OR51A2 and OR51A4) all other were pseudo genes.

Table 21: Common CNV identified in two families

Sl. No.	Affymetrix-Partek-PennCNV	
	Location	**Gene**
1	11p15.4	LOC650368, OR51A4, OR51A2
2	14q32.33	KIAA0125, ADAM6

Table 22: Functions of genes located in commonly identified CNVs

Gene Name	Gene Function
OR51A2 and **OR51A4** (olfactory receptor, family 51, subfamily A, member 2 and 4)	Olfactory receptors are responsible to initiate a neuronal response that triggers the perception of a smell. The olfactory receptor proteins are members of G-protein-coupled receptors (GPCR) family and encoded by single coding-exon. Olfactory receptors are trans membrane and are responsible for the recognition and G protein-mediated transduction of odorant signals. The olfactory receptor gene family is the largest in the genome.
LOC650368 (asparagine-linked glycosylation 1-like pseudogene), **KIAA0125**, **ADAM6** (ADAM metallopeptidase domain 6, pseudo gene)	Pseudo genes

3.6 A special observation

Interestingly only family number 2 has loss of CNV in chromosome 13 in ‘q’ arm. The distribution of the CNV including start and end point is presented in the tables (Table 23, 24 and 25). This deletion ranges the genomic region between 13q14.2 to 13q21.33. All three different methods (GTC, PGS and PennCNV) were able to identify these CNV deletions. But number and size of CNVs identified with PGS differed from rest two methods. First and third methods results were more or less same. Affymetrix and Partek identified 38 and 39 loss of CNVs respectively in the genomic region. They had some difference of CNV sizes. On the other hand second method covered the same genomic region with only 3 large size CNVs losses. A graphical comparison of this deletion among three methods is shown in figure 14. Detailed information about the CNVs was listed in the tables 23, 24 and 25 respectively for three different methods.

Table 23: Identity of CNV present in chromosome 13 in father of family 2 using GTC

Sl. No.	Start Position	End Position	Size(kb)	Loss/Gain
1	47237551	47666738	429	Loss
2	48148998	48696505	548	Loss
3	48751953	49312876	561	Loss
4	49343727	51639572	2296	Loss
5	52437155	52763266	326	Loss
6	52826196	52932732	107	Loss
7	53057893	53192116	134	Loss
8	53256041	53590407	334	Loss
9	54680266	54790556	110	Loss
10	54935168	55098944	164	Loss
11	55750401	55874339	124	Loss
12	55974036	56272375	298	Loss
13	56398440	56585146	187	Loss
14	56708289	56827502	119	Loss
15	56934307	57038858	105	Loss
16	57282053	57980283	698	Loss
17	58340087	58443546	103	Loss
18	58676802	58970079	293	Loss
19	59126155	59350332	224	Loss
20	59419017	59781724	363	Loss
21	60075937	60201120	125	Loss
22	60685046	61083734	399	Loss
23	61379792	61483739	104	Loss
24	61650921	61897775	247	Loss
25	62086051	62225287	139	Loss
26	62279951	62386580	107	Loss
27	63059363	63205520	146	Loss
28	63386740	63794997	408	Loss
29	63956144	64073049	117	Loss
30	64177520	64483758	306	Loss
31	64835495	65128868	293	Loss
32	65196691	65485151	288	Loss
33	65671222	66142551	471	Loss
34	66444238	66561990	118	Loss

Table 23 (cont'd)

35	66767856	67064534	297	Loss
36	67105010	67226610	122	Loss
37	67389818	67536058	146	Loss
38	67647031	67858955	212	Loss

Table 24: Identity of CNV present in chromosome 13 in father of family 2 using PGS

Sl. No.	Start Position	End Position	Size(kb)	Loss/Gain
1	49489783	50379977	890194	Loss
2	50379977	67073644	16693667	Loss
3	67094698	68375606	1280908	Loss

Table 25: Identity of CNV present in chromosome 13 in father of family 2 using PennCNV

Sl. No.	Start Position	End Position	Size(kb)	Loss/Gain
1	47819459	47951563	132	Loss
2	48465537	48669985	204	Loss
3	49049927	50068586	1019	Loss
4	50070025	50100433	30	Loss
5	50103430	50495232	392	Loss
6	50648626	50708007	59	Loss
7	50728714	51227759	499	Loss
8	51337276	51377929	41	Loss
9	52307395	52386159	79	Loss
10	52510619	52701944	191	Loss
11	53265283	53335594	70	Loss
12	53454796	53501480	47	Loss
13	53672584	53733768	61	Loss
14	54841867	55045560	204	Loss
15	55207842	55221845	14	Loss
16	55756707	56740628	984	Loss
17	56893714	57553306	660	Loss
18	57801407	57941247	140	Loss
19	58340099	58603660	264	Loss
20	58799930	58923940	124	Loss
21	60107596	60180155	73	Loss
22	60731828	61245543	514	Loss
23	61290320	61476014	186	Loss
24	61675048	61776378	101	Loss
25	62278487	62386592	108	Loss
26	62604846	62642623	38	Loss
27	63060191	63153394	93	Loss
28	63395249	63829212	434	Loss
29	63913286	64065226	152	Loss
30	64204974	64696387	491	Loss
31	64835507	65055390	220	Loss
32	65106479	65301715	195	Loss
33	65344417	65551994	208	Loss
34	65724319	65991852	268	Loss

Table 25 (cont'd)

35	66315311	66399797	84	Loss
36	66592236	66738575	146	Loss
37	66836797	67056831	220	Loss
38	67114843	67497129	382	Loss
39	67568305	67858955	291	Loss

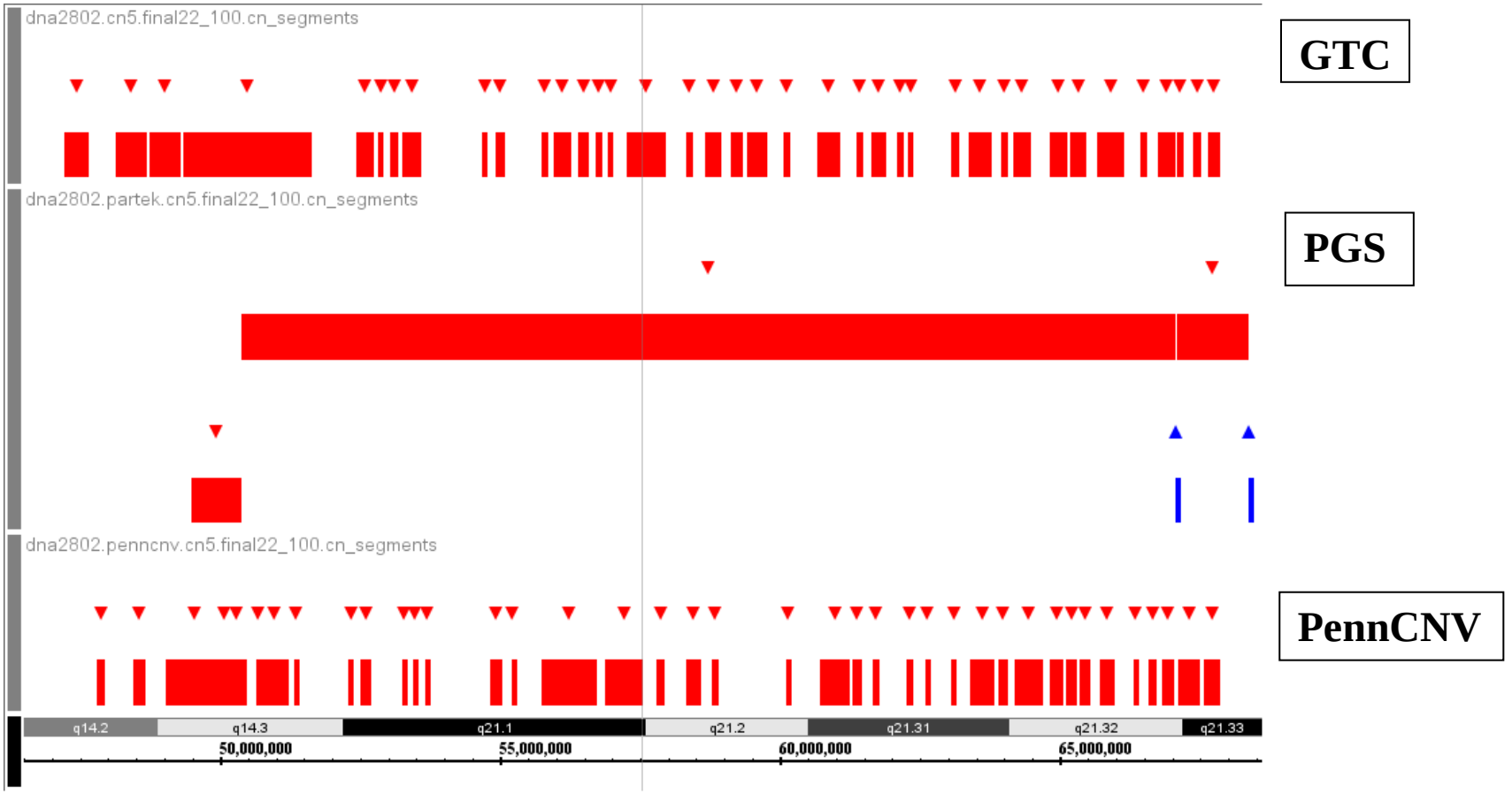


Figure 14: Graphical representation from Genotyping Console Browser identifying the 13q deletion for father of family 2 with three different methods. From the top, first one was identified by GTC, second one with PGS and bottom last one with PennCNV.

Chapter 4: Discussion

4.1 Summary

The identification and characterization of genetic differences across individuals represents the next major challenge in human biology. This is currently being attempted using two sets of results, those from DNA microarrays and complete genome sequences. In both cases, it has proved easier to generate the data than to make sense of the results. In this research, I have attempted to assess the analytical challenges associated with the generation of whole-genome DNA microarray data. Specifically, I have tested three commonly used methods (Affymetrix® Genotyping Console™, Partek® Genome Suite™ and PennCNV) for their ability to accurately and reliably detect copy number variation from array data generated by the Affymetrix® Genome-Wide Human SNP Array 6.0.

The dataset used in this analysis assesses three degrees of genetic relatedness: unrelated individuals representing two different families, first degree relatives representing parents and offspring, and monozygotic twin pairs. The results obtained have offered a number of conclusions, discussed below.

1. In addition to single nucleotide differences such as SNPs, copy number variations (CNV) contribute to significant genetic variation across individuals:

   a) *Familial distribution of copy number variation*:

   The number of CNVs per individual in both families analyzed ranges from 35 to 65 variations, with the exception of one individual who is described later. This is similar to the number of CNVs per subject reported from many other studies that have used the Affymetrix® Genome-Wide Human SNP Array 6.0 (Ku *et. al.*, 2010). This range is also comparable with the number of CNVs found in Venter's genome (62), established using his complete genome sequence (Levy *et. al.*, 2007). Most CNVs identified were in the range of 100 to 200 kb, also consistent with the size distribution of CNVs reported in the literature (Ku *et. al.*, 2010). The

majority of CNVs observed were copy number gains (78.5%). Further, the chromosomal distribution of CNVs was comparable across individuals with the exception of the father in family 2 who had consistently higher CNVs that affected most of this individual's chromosomes. Further, more than 50 percent of the identified CNVs overlapped with known RefSeq genes.

b) *Genetic difference between monozygotic twins*:

This study has been able to establish genome-wide copy number variations (CNV) as exist between monozygotic twins that are discordant.

The monozygotic twins in both families genetically differ from each other via different forms of the structural variations. The CNVs present in the monozygotic twins were classified into two groups, *de novo* and *inherited*. The monozygotic twins in family 1 share only 33 percent *de novo* CNV as compared to her co-twin, while the twins in family 2 share 36 percent. This study also observed inherited CNV differences between monozygotic twins for the both families. The twins in family 1 show more than 9 percent inherited CNVs that they do not share with their co-twin and, in family 2, nearby 8.5 percent of their inherited CNVs differ. Interestingly, the inherited patterns of CNVs for both twins were not same for both families. In family 1, twin one inherited 37%, 22% and 41% CNVs from her father (exclusively), mother (exclusively), and both parents respectively, but twin two inherited 32%, 26% and 42% respectively. In family two these numbers are 42%, 23% and 35% for twin one and 48%, 14% and 38% for twin two respectively. This study suggests that discordant monozygotic twins also vary genetically from each other, data which is supported by recent other studies (Bruder *et. al.*, 2008).

c) De novo events may contribute the individual differences

A novelty of this study is that we were able to classify observed CNVs into two groups based on their absence or presence in one of the parents. CNVs that were found in one or both twins and not seen in either parent were classified as *de novo.*

If a *de novo* CNV was present in both twins, it was considered to have originated during meiosis or parental development and when present in only one of the two twins, it was assumed to have originated in mitosis during development. This classification allowed us to identify 15 and 25 *de novo* CNVs in family 1 and family 2 respectively. The results include the genomic locations as well as individual-specific break points, which allow for the assessment of regions of overlap that have been previously reported in the Database of Genomic Variants (Toronto, Ontario). The results also suggest that the rate of mitotically-derived CNVs is approximately three times higher than the rate of CNVs generated during parental meiosis. Some of this data is has been previously published in Maiti *et al.,* 2011.

2. ***This study observed an exceptional 13q deletion containing 38 CNVs at a single genomic location present in the father in family 2 (II-1-1).***

An exceptional and unexplained finding of this study was the CNV quantity and distribution found in the genome of the father in family 2 (II-1-1) who was found to harbour a rare chromosome 13q deletion. This deletion starts from cytoband position 13q14.2 and ends at 13q21.32 and contains a loss of 38 CNVs. Although this finding is beyond the scope of this study, it is important to note that II-1-1 underwent chemotherapy treatment and that the samples utilized in this study were obtained towards the end of this treatment (Maiti *et al.*, 2011).

3. ***Different analytical tools on same data may result in similar but different results.***

The underlying statistical models for the several copy number variation calling methods evaluated in this study diverged by varying degrees. The primary raw data used for calling the CNVs from the Affymetrix® Genome-Wide Human SNP Array 6.0 were the SNP intensity signals measured by logR ratios, however some of the methods also used B allele frequencies. Principally, all data analysis packages mainly used Hidden Markov Model (HMM) for calling copy number variation but how this model was used by each platform differed slightly. Affymetrix® Genotyping Console™ uses median log2 ratio values of all contiguous SNPs in the given HMM copy number state segment. Partek®

Genome Suite™ also uses HMM algorithm but evaluates discrete changes of whole number copy number states. The PennCNV program uses the combined logR ratio and B allele frequency. GTC developed their algorithm for analyzing the Affymetrix® SNP array data, whereas non-Affymetrix platforms such as the PGS and PennCNV have developed their algorithms for use with both Illumina and Affymetrix® SNP arrays. This study observed significant of variability among the results obtained using Affymetrix, Partek and PennCNV algorithms. The observed CNV frequency distribution was found to be different in family members for both families, and highly dependent on platform. GTC identified more than 86 and 87 percent gains in the twins of family 1 whereas PGS and PennCNV identified only 40, 47 percent and 54, 40 percent respectively. Similarly, in family 2, GTC found 88 and 91 percent gains in twin pairs one and two, whereas PGS and PennCNV found 36 and 29 percent, and 29 and 26 percent, respectively. In the case of CNV size distribution (Dalila Pinto *et. al.*, 2011), this study observed that GTC was able to identify relatively small CNVs, primarily 100 to 200 kb in size. PGS and PennCNV identified only larger CNVs between 1 Mb to 20 Mb in size for both families. Surprisingly, among all copy number variation identified, there were only 6 CNVs commonly identified with all platforms and in both families.

4.2 Caveats

Microarrays are a hybridization method that currently needs further validation with resequencing. At the very least, data should be regenerated using one more hybridization method. Other than whole genome sequencing, all kinds of array-based methods identify only known variants in the human genome. As far as data analysis of the raw array data files, each analytical tool has their own algorithm that is based on specific hypotheses. As a result, depending on what questions or null hypothesis is asked by a given investigator, the final outcome may be quite different. So, analysis of same microarray raw data using multiple packages does not nullify the necessity of a discovery algorithm that meets the criteria of a researcher's specific interest. Lastly, *in silico* findings require confirmation by biological approaches and should be revalidated by methods such as quantitative PCR.

4.3 Future studies

Follow-up studies can be planned using the same set of individual samples and data in multiple ways. First, the microarray data can be generated using another platform available from other array providers, such as Illumina Inc. (San Diego, CA, USA). Raw microarray data generated from these platforms can be analyzed using the same set of CNV-calling tools and the results can be compared for percent overlap of results. Secondly, using the same set of samples, we can generate data using whole genome sequencing. The outcome of these results can be used to detect and confirm copy number variation that had previously been determined using microarray data. Nevertheless, given the state of the current technology, it is prudent to utilize more than a single detection method and CNV-calling algorithm to test any result for consistency and reproducibility.

4.4 Novelty of the study

We (and others) have shown that monozygotic twins are not genetically identical (Maiti *et al.,* 2011, Bruder *et. al*., 2008). Copy number variations may contribute the phenotypic discordance observed even in monozygotic twins. These structural variants are very fluid and thus may play a major role in evolution. In this study, we have also evaluated and revalidated findings obtained with Affymetrix technology to examin the reliability of the obtained results. Here, the result of an Affymetrix array and analysis platform has been re-tested with Partek and PennCNV, two alternative packages that utilize very different algorithms than Affymetrix and each other. To provide further validity to the detected CNVs, the same raw data analyzed with different packages and algorithms, which resulted in six CNVs confirmed by all three. These results highlight these CNVs for future study, suggesting that they may be good candidates for further research towards the identification of a genetic basis of the twins' discordance for schizophrenia.

Every microarray data analysis platform possesses its own algorithm that is limited by how it evaluates a specific null hypothesis, thus making one package restricted in its ability to identify losses or gains of CNVs of a specific size or other criteria. This study demonstrates the scientific hazards of interpreting results scored with single package. The familial inheritance of CNVs varies from one family to another. Further, although all

three algorithms have demonstrated that monozygotic twins do not share genetic homogeneity, exactly how they may differ requires careful and precise identification, particularly in the case of examining the causes of discordance for genetic diseases.

4.5 Original contribution

This study identified several *de novo* CNVs that differ between co-twins. This study also identified three novel CNVs which were not previously listed in the Database of Genomic Variants (Toronto, Ontario) (now submitted). Of these three CNVs, two were gains and one was a loss of a CNV in chromosomal locations 8q11.21, 19q13.41 and 14q32.11 respectively.

4.6 Conclusions

In this study, the genomic DNA of discordant monozygotic twins and their families were analyzed to identify and characterize the genomic structural variation (copy number variation) that existed between monozygotic twins and their parents. The study of copy number variation in families containing monozygotic twin pairs is a new and rapidly-evolving field. Microarray data are becoming a very powerful source to indentify copy number variation. Different algorithms use different statistical methods to identify and classify copy number variations. The goal of this study was to evaluate the existence and potential role of genetic variants in the monozygotic twins, and stress the importance of statistical methods or algorithms to identify these genetic variants.

The central outcomes from this research are listed below:

1. Monozygotic twins differ in copy number variation (CNV) and are not 100% genetically identical, as expected. CNVs may be either inherited from their parents or acquired during their development. It is the *de novo* differences, however, that may cause genetic differences between monozygotic twins and may be responsible for phenotypic differences.
2. Different analytical algorithms for calling CNVs may yield variable results from the same microarray data.

3. This study observed an exceptional 13q deletion containing 38 CNVs at a single genomic location, present in the father in family 2 (II-1-1).
4. There is yet a need to develop reliable analytical methods to identify genetic variations from microarray results.

REFERENCES

Bierut LJ, Agrawal A, Bucholz KK, Doheny KF, Laurie C, Pugh E, Fisher S, Fox L, Howells W, Bertelsen S, Hinrichs AL, Almasy L, Breslau N, Culverhouse RC, Dick DM, Edenberg HJ, Foroud T, Grucza RA, Hatsukami D, Hesselbrock V, Johnson EO, Kramer J, Krueger RF, Kuperman S, Lynskey M, Mann K, Neuman RJ, Nöthen MM, Nurnberger JI Jr, Porjesz B, Ridinger M, Saccone NL, Saccone SF, Schuckit MA, Tischfield JA, Wang JC, Rietschel M, Goate AM, Rice JP; Gene, Environment Association Studies Consortium (2010) A genome-wide association study of alcohol dependence. Proc Natl Acad Sci U S A 107(11):5082-5087

Bruder CE, Piotrowski A, Gijsbers AA, Andersson R, Erickson S, Diaz dS, Menzel U, Sandgren J, von Tell D, Poplawski A, Crowley M, Crasto C, Partridge EC, Tiwari H, Allison DB, Komorowski J, van Ommen GJ, Boomsma DI, Pedersen NL, den Dunnen JT, Wirdefeldt K, Dumanski JP (2008) Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. Am J Hum Genet 82: 763-771

Diskin SJ, Li M, Hou C, Yang S, Glessner J, Hakonarson H, Bucan M, Maris JM, Wang K (2008) Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. Nucleic Acids Res 36(19):e126.

Eckel-Passow JE, Atkinson EJ, Maharjan S, Kardia SL, de Andrade M (2011) Software comparison for evaluating genomic copy number variation for Affymetrix 6.0 SNP array platform. BMC Bioinformatics 12:220.

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

Grant SF, Hakonarson H (2008) Microarray technology and applications in the arena of genome-wide association. Clin Chem 54(7):1116-24.

Huang J, Wei W, Zhang J, Liu G, Bignell GR, Stratton MR, Futreal PA, Wooster R, Jones KW, Shapero MH (2004) Whole genome DNA copy number changes identified by high density oligonucleotide arrays. Hum Genomics 1(4):287-99.

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C (2004) Detection of large-scale variation in the human genome. Nat Genet 36: 949-951.

Ionita-Laza I, Rogers AJ, Lange C, Raby BA, Lee C (2009)Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis. Genomics 93(1):22-6.

Istrail S, Sutton GG, Florea L, Halpern AL, Mobarry CM, Lippert R, Walenz B, Shatkay H, Dew I, Miller JR, Flanigan MJ, Edwards NJ, Bolanos R, Fasulo D, Halldorsson BV, Hannenhalli S, Turner R, Yooseph S, Lu F, Nusskern DR, Shue BC, Zheng XH, Zhong F, Delcher AL, Huson DH, Kravitz SA, Mouchard L, Reinert K, Remington KA, Clark AG, Waterman MS, Eichler EE, Adams MD, Hunkapiller MW, Myers EW, Venter JC (2004) Whole-genome shotgun assembly and comparison of human genome assemblies.Proc Natl Acad Sci U S A 101(7):1916-21.

Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K, Lee C, Nizzari MM, Gabriel SB, Purcell S, Daly MJ, Altshuler D (2008) Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet 40(10):1253-60.

Ku CS, Pawitan Y, Sim X, Ong RT, Seielstad M, Lee EJ, Teo YY, Chia KS, Salim A (2010) Genomic copy number variations in three Southeast Asian populations. Hum Mutat 31(7):851-7.

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, MacDonald JR, Pang AW, Shago M, Stockwell TB, Tsiamouri A, Bafna V, Bansal V, Kravitz SA, Busam DA, Beeson KY, McIntosh TC, Remington KA, Abril JF, Gill J, Borman J, Rogers YH, Frazier ME, Scherer SW, Strausberg RL, Venter JC (2007) The diploid genome sequence of an individual human. PLoS Biol 5: e254.

Li J, Yang T, Wang L, Yan H, Zhang Y, Guo Y, Pan F, Zhang Z, Peng Y, Zhou Q, He L, Zhu X, Deng H, Levy S, Papasian CJ, Drees BM, Hamilton JJ, Recker RR, Cheng J, Deng HW (2009) Whole genome distribution and ethnic differentiation of copy number variation in Caucasian and Asian populations.PLoS One 4(11):e7958.

Machado LR, Hardwick RJ, Bowdrey J, Bogle H, Knowles TJ, Sironi M, Hollox EJ (2012) Evolutionary history of copy-number-variable locus for the low-affinity Fcγ receptor: mutation rate, autoimmune disease, and the legacy of helminth infection. Am J Hum Genet 90(6):973-85.

Maiti S, Kumar KH, Castellani CA, O'Reilly R, Singh SM (2011) Ontogenetic de novo copy number variations (CNVs) as a source of genetic individuality: studies on two families with MZD twins for schizophrenia. PLoS One 6(3):e17125.

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

McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, Shapero MH, de Bakker PI, Maller JB, Kirby A, Elliott AL, Parkin M, Hubbell E, Webster T, Mei R, Veitch J, Collins PJ, Handsaker R, Lincoln S, Nizzari M, Blume J, Jones KW, Rava R, Daly MJ, Gabriel SB, Altshuler D (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet 40(10):1166-74

Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, Dairkee SH, Ljung BM, Gray JW, Albertson DG (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet 20(2):207-11.

Pinto D, Darvishi K, Shi X, Rajan D, Rigler D, Fitzgerald T, Lionel AC, Thiruvahindrapuram B, Macdonald JR, Mills R, Prasad A, Noonan K, Gribble S, Prigmore E, Donahoe PK, Smith RS, Park JH, Hurles ME, Carter NP, Lee C, Scherer SW, Feuk L (2011) Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nat Biotechnol 29(6):512-20.

Pylkäs K, Vuorela M, Otsukka M, Kallioniemi A, Jukkola-Vuorinen A, Winqvist R (2012) Rare copy number variants observed in hereditary breast cancer cases disrupt genes in estrogen signaling and TP53 tumor suppression network. PLoS Genet 8(6):e1002734.

Ragoussis J (2009) Genotyping technologies for genetic research. Annu Rev Genomics Hum Genet 10:117-33.

Ramakrishna M, Williams LH, Boyle SE, Bearfoot JL, Sridhar A, Speed TP, Gorringe KL, Campbell IG (2010) Identification of candidate growth promoting genes in ovarian cancer through integrated copy number and expression analysis. PLoS One 8;5(4):e9983.

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M (2004) Large-scale copy number polymorphism in the human genome. Science 305: 525-528.

Stephen W Scherer, Charles Lee, Ewan Birney, David M Altshuler, Evan E Eichler, Nigel P Carter, Matthew E Hurles & Lars Feuk (2007) Challenges and standards in integrating surveys of structural variation. Nat Genet 39(7 Suppl):S7-15.

Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavaré S, Deloukas P, Hurles ME, Dermitzakis ET (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315(5813):848-53.

Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, Olson MV, Eichler EE (2005) Fine-scale structural variation of the human genome. Fine-scale structural variation of the human genome. Nat Genet 37(7):727-32.

Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, Bucan M (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 17(11):1665-74.

Yamamoto G, Nannya Y, Kato M, Sanada M, Levine RL, Kawamata N, Hangaishi A, Kurokawa M, Chiba S, Gilliland DG, Koeffler HP, Ogawa S (2007) Highly sensitive method for genomewide detection of allelic composition in nonpaired, primary tumor specimens by use of Affymetrix single-nucleotide-polymorphism genotyping microarrays. Am J Hum Genet 81(1):114-26.

Zhang F, Khajavi M, Connolly AM, Towne CF, Batish SD, Lupski JR (2009) The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. Nat Genet 41: 849-853.

Appendix Table 1: *de novo* CNVs in Family 1 identified by GTC.

Sl. No	Location	Family 1						Status	Meiosis	Mitosis	Novel	Genes (Overlapping or Nearby)	SD
		I-2-1	Size (kb)	Breakpoints	I-2-2	Size (kb)	Breakpoints						
1	**1p36.13**	Yes	112	16724089...16835888				**Gain**		**Yes**		NBPF1, NBPF10	1
2	**2p25.3**	Yes	152	1407209...1559511	Yes	152	1407209...1559511	**Gain**	**Yes**			TPO	0
3	**2p11.2**	Yes	1147	89862331...91008912	Yes	1159	89850279...91008912	**Loss**	**Yes**				0
4	**4q28.3**				Yes	191	132801221...132992517	**Gain**		**Yes**			1
5	**7q11.21**	Yes	118	64706066...64823721	Yes	118	64704377...64822216	**Loss**	**Yes**				1
6	**8p23.1**	Yes	126	7847289...7973253				**Loss**		**Yes**			1
7	**8q11.1**	Yes	336	47045602...47381308	Yes	250	47131383...47381308	**Gain**	**Yes**				0
8	**8q11.21**				Yes	154	48178242...48332398	**Loss**		**Yes**	**Yes**	**KIAA0146**	0
9	**9p11.2**	Yes	569	45361389...45929992				**Gain**		**Yes**		FAM27A	1
10	**9p13.1**				Yes	141	38777481...38918566	**Gain**		**Yes**			1
11	**9q12**				Yes	861	65412415...66273526	**Gain**		**Yes**			1
12	**12p13.31**				Yes	196	8303317...8499801	**Gain**		**Yes**		CLEC6A	1
13	**14q32.11**				Yes	103	89780137...89883415	**Loss**		**Yes**	**Yes**	**PSMC1, C14orf102**	0
14	**21q11.2**				Yes	119	13891136...14009908	**Gain**		**Yes**		ANKRD21, LOC441956	1
15	**Xp11.23**	Yes	149	47917899...48066856	Yes	149	47917899...48066856	**Gain**	**Yes**			SSX5, SSX1, SSX9	0

Identity of *de novo* CNVs found in Family 1 (Table 5) and the gene regions (overlapping or nearby). *De novo* CNVs are defined as those that are present in either or both twins but not found in parents. SD displays the percentage of overlap with segmental duplications, '0' indicates no overlap between the CNV and segmental duplication and '1' indicates 90-100% overlap. The table includes genomic locations as well as twin specific breakpoints which allow for the assessment of regions of overlap with the Database of Genomic Variants (Toronto, ntario). SI No. = Serial number. Novel indicates a CNV which is not present in The Database of Genomic Variants (DGV).

Appendix Table 2: *de novo* CNVs in Family 2 by GTC.

Sl. No	Location	Family 2						Status	Meiosis	Mitosis	Novel	Genes (Overlapping or Nearby)	SD
		II-2-1	Size (kb)	Breakpoints	II-2-2	Size (kb)	Breakpoints						
1	**1q21.1**				Yes	120	143867807...143987616	**Gain**		**Yes**		NOTCH2NL	0
2	**1q21.1**	Yes	104	147353175...147456930	Yes	104	147353175...147456930	**Loss**	**Yes**				0
3	**1q43**				Yes	119	241230453...241349107	**Gain**		**Yes**			0
4	**3q21.2**	Yes	155	126958012...127112518				**Gain**		**Yes**			1
5	**4p11**	Yes	299	48986100...49285347				**Gain**		**Yes**			0
6	**5p15.33**	Yes	101	770367...871743	Yes	107	770367...877436	**Gain**	**Yes**			ZDHHC11	1
7	**5p13.3**	Yes	151	34119387...34269887				**Gain**		**Yes**			1
8	**7q11.21**	Yes	202	61761008...61962936				**Gain**		**Yes**			0
9	**7q11.21**	Yes	116	64588316...64704125	Yes	125	64579322...64704125	**Gain**	**Yes**				1
10	**7q35**				Yes	100	142956516...143056637	**Gain**		**Yes**		LOC441294, FAM139A	1
11	**8p23.1**	Yes	220	12071704...12291845	Yes	220	12071704...12291845	**Gain**	**Yes**			FAM86B1, DEFB130	0
12	**9p12**	Yes	2720	41465094...44184864	Yes	1901	42249132...44149779	**Gain**	**Yes**			ANKRD20A2, ANKRD20A3, FOXD4L4, FOXD4L2	1
13	**9q12**				Yes	250	67416254...67665974	**Gain**		**Yes**		ANKRD20A1, ANKRD20A3	1
14	**11q13.2**	Yes	267	67239223...67505822	Yes	139	67239223...67378031	**Gain**	**Yes**				1
15	**12p13.31**				Yes	189	8310909...8499801	**Gain**		**Yes**			1
16	**13q11**	Yes	208	18138676...18346383				**Gain**		**Yes**			1
17	**14q11.1**	Yes	601	18072112...18672662	Yes	601	18072112...18672662	**Gain**	**Yes**			OR11H12, ACTBL1	1
18	**15q11.1**				Yes	106	18276329...18382609	**Gain**		**Yes**			1
19	**15q11.2**	Yes	203	19882763...20085783	Yes	221	19864583...20085783	**Gain**	**Yes**			OR4M2, OR4N4, LOC650137	1
20	**15q13.1**				Yes	227	26808083...27035216	**Gain**		**Yes**		APBA2	0
21	**15q13.2**	Yes	110	28452853...28563274				**Gain**		**Yes**		CHRFAM7A	1
22	**17p11.1**	Yes	199	22127012...22326425				**Gain**		**Yes**			0
23	**19q13.41**	Yes	109	58847652...58957090				**Gain**		**Yes**	**Yes**	**ZNF331, DPRX**	0
24	**20q11.1**				Yes	118	28147331...28264860	**Gain**		**Yes**			1
25	**21p11.2**	Yes	3480	10106540...13586186	Yes	3814	9758730...13572586	**Gain**	**Yes**			BAGE2, BAGE4, BAGE	0

Identity of *de novo* CNVs found in Family 2 (Table 7) and the gene regions (overlapping or nearby). *De novo* CNVs are defined as those that are present in either or both twins but not found in parents. SD displays the percentage of overlap with segmental duplications, '0' indicates no overlap between the CNV and segmental duplication and '1' indicates 90-100% overlap. The table includes genomic locations as well as twin specific breakpoints which allow for the assessment of regions of overlap with the Database of Genomic Variants (Toronto, Ontario). SI No. = Serial number. Novel indicates a CNV which is not present in The Database of Genomic Variants (DGV).

Appendix Table 3: *Inherited* CNVs in Family 1 by GTC.

Sl. No	Location	Family 1												Status	Novel	Genes (Overlapping or Nearby)	SD
		I-1-1	Size(kb)	Breakpoints	I-1-2	Size(kb)	Breakpoints	I-2-1	Size(kb)	Breakpoints	I-2-2	Size(kb)	Breakpoints				
1	1p36.33	Yes	167	51586...218557				Yes	707	51586...758644	Yes	707	51586...758644	**Gain**		OR4F5, OR4F3, OR4F16, OR4F29	1
2	1q21.1	Yes	765	147303136...148068045	Yes	734	147311699...148045353				Yes	577	147381253...147958358	**Gain**		PPIAL4, FCGR1A, HIST2H2BF	1
3	2p11.2	Yes	338	88917155...89254935	Yes	322	88914734...89236978	Yes	325	88917155...89242149	Yes	327	88914734...89242149	**Gain**			1
4	2p11.1				Yes	138	91017077...91154841	Yes	143	91017077...91160399	Yes	137	91017077...91154463	**Gain**			1
5	2q21.2	Yes	236	132597824...132833718				Yes	222	132597824...132819911	Yes	222	132597824...132819911	**Gain**			1
6	3p12.3	Yes	306	75677859...75984129	Yes	380	75597086...75977210	Yes	182	75583442...75764996	Yes	402	75582277...75984129	**Gain**			1
7	3q21.2				Yes	132	126907150...127039328				Yes	185	126907150...127091652	**Gain**			1
8	3q21.3				Yes	170	131198515...131368353	Yes	166	131213377...131379054	Yes	176	131214431...131389948	**Gain**			1
9	4p16.2				Yes	196	4040542...4236511	Yes	363	3873500...4236511	Yes	366	3870638...4236511	**Gain**			0
10	4p11	Yes	436	48849363...49285347				Yes	497	48788531...49285347	Yes	497	48788531...49285347	**Gain**			1
11	4q35.2	Yes	232	191021837...191254119	Yes	195	191059369...191254119				Yes	223	191031042...191254119	**Gain**		FRG1, TUBB4Q, FRG2, DUX4	0
12	7p11.1	Yes	231	57523223...57753919	Yes	101	57640100...57741512	Yes	315	57640100...57954861	Yes	117	57640100...57757406	**Gain**			1
13	7q11.21				Yes	112	61365830...61477958				Yes	111	61365830...61476918	**Gain**			0
14	7q11.21				Yes	253	64320173...64573380	Yes	415	64204380...64619667	Yes	385	64204380...64589253	**Loss**		ZNF92	0
15	8p23.1	Yes	694	7209579...7903560				Yes	215	7027251...7242508	Yes	270	7021193...7291135	**Loss**		DEFB103A, DEFB103B, SPAG11B, DEFB104B, DEFB104A, DEFB106B, DEFB106A, DEFB105B, DEFB105A, DEFB107B, DEFB107A, SPAG11A, DEFB4	2
16	9q12				Yes	690	68115006...68805366	Yes	1141	68115006...69256300	Yes	694	68115006...68809437	**Gain**		FOXD4L6, CBWD6, ANKRD20A4, CCDC29	1
17	10q11.1	Yes	183	41972779...42155347	Yes	105	41934430...42039743	Yes	183	41972779...42155347	Yes	239	41934430...42173117	**Gain**			1
18	11p15.4	Yes	175	3405799...3580813	Yes	131	3430789...3561991	Yes	139	3430789...3569305	Yes	156	3406002...3561991	**Gain**			1
19	11q13.2	Yes	227	67239223...67466368				Yes	193	67273413...67466368				**Gain**			1
20	14q11.1	Yes	1322	18138794...19460382	Yes	1103	18072112...19175240	Yes	705	18072112...18776746	Yes	705	18072112...18776746	**Gain**		OR11H12, ACTBL1, OR4Q3, OR4M1, OR4N2, OR4K5	0
21	14q32.33	Yes	126	105265510...105391419	Yes	167	105100670...105268160	Yes	632	105190672...105822317	Yes	181	105149735...105331052	**Gain**			0
		Yes	156	105413825...105569826	Yes	213	105289618...105502685	Yes	178	105827891...106005581	Yes	261	105341035...105601720	**Gain**			0
		Yes	205	105612786...105818132	Yes	279	105508896...105788389				Yes	280	105612786...105892769	**Gain**			0
22	15q11.1	Yes	471	18370252...18841457	Yes	178	18522238...18700540	Yes	1223	18845990...20068512	Yes	562	18276329...18838423	**Gain**		LOC283755, POTE15, OR4M2	1
		Yes	1177	18845990...20022565	Yes	344	18845990...19189673				Yes	1078	18845990...19923712	**Gain**		OR4N4, LOC650137	1
					Yes	264	19303160...19566863							**Gain**			1
23	15q11.2	Yes	189	22026287...22214843							Yes	174	22026287...22200408	**Gain**			1
24	16p11.2	Yes	1217	32303108...33520394	Yes	1297	32538757...33836128	Yes	1142	32538757...33680554	Yes	249	32538757...32787273	**Gain**		LOC729355, TP53TG3	1
											Yes	752	32910319...33662480	**Gain**			1
25	16p11.2	Yes	250	34374795...34624994				Yes	249	34375533...34624994	Yes	249	34375533...34624994	**Gain**			1
26	17p11.2	Yes	140	20559979...20700133							Yes	164	20538867...20703365	**Gain**			1
27	17q21.31	Yes	229	41521621...41750183				Yes	123	41521621...41644356	Yes	123	41521621...41644356	**Gain**		KIAA1267, LRRC37A	0
28	17q21.31	Yes	351	41756820...42107467	Yes	296	41811739...42107467	Yes	392	41700624...42092926	Yes	302	41700624...42002447	**Gain/Loss**		ARL17, LRRC37A2, NSF	1
29	21p11.2				Yes	204	9758730...9962501	Yes	204	9758730...9962501	Yes	204	9758730...9962501	**Gain**		TPTE	0
30	21p11.1	Yes	3411	10106540...13517603	Yes	3419	10106540...13525448	Yes	3419	10106540...13525448	Yes	3477	10106540...13583117	**Gain**		BAGE2, BAGE4, BAGE	2
31	22q11.1	Yes	339	14435171...14774593				Yes	320	14435207...14754960	Yes	320	14435207...14754960	**Gain**		ACTBL1	1
32	22q11.21	Yes	124	20051708...20175282				Yes	136	20145854...20281562	Yes	136	20145854...20281562	**Gain**		HIC2, UBE2L3	1
33	22q11.22	Yes	240	21292462...21532509				Yes	127	21327799...21454509				**Gain**		GGTL4	0

Appendix Table 4: *Inherited* CNVs in Family 2 by GTC.

Sl. No	Location	Family 2											Status	Novel	Genes (Overlapping or Nearby)	SD
		II-1-1	Size(kb)	Breakpoints	II-1-2	Size(kb)	Breakpoints	II-2-1	Size(kb)	Breakpoints	II-2-2	Size(kb)	Breakpoints				
1	1p36.33	Yes	707	51586...758644				Yes	537	218557...755132				Gain		OR4F5, OR4F3, OR4F16, OR4F29	1
2	1p36.13	Yes	117	16718622...16835888				Yes	167	16718622...16885360	Yes	345	16718622...17063437	Gain		NBPF1, NBPF10	1
3	1p21.1				Yes	130	103910749...104041200	Yes	127	103931691...104058426				Gain		AMY2B, AMY2A, AMY1A, AMY1C, AMY1B	1
4	1p11.2	Yes	21680	121045307...142725034				Yes	21725	121045307...142770353	Yes	21725	121045307...142770353	Gain			0
5	1q23.3	Yes	121	159775403...159896554				Yes	116	159780383...159896554	Yes	121	159775403...159896554	Loss		FCGR3A, FCGR2C, FCGR3B	1
6	2p11.2	Yes	401	88925215...89326446	Yes	759	88914227...89673147	Yes	935	88926972...89861763	Yes	450	88914734...89365010	Gain			1
7	2p11.1	Yes	160	91017077...91176948				Yes	268	91017077...91285520	Yes	1275	89879561...91154463	Gain			1
8	2q21.2	Yes	260	132593436...132853218	Yes	183	132597824...132780848				Yes	222	132597824...132819911	Gain			1
9	3p12.3	Yes	260	75583442...75843060	Yes	226	75538978...75764996	Yes	260	75583442...75843060	Yes	182	75583442...75764996	Gain			1
10	3q12.2	Yes	108	101822746...101930873				Yes	102	101822746...101925168				Gain		GPR128, TFG	0
11	3q21.3				Yes	141	131198817...131339424	Yes	195	131194669...131389948	Yes	199	131198515...131397648	Gain			1
12	4p16.2	Yes	407	3870638...4278016	Yes	180	3964803...4144453	Yes	366	3870638...4236511	Yes	357	3870638...4227503	Gain		OTOP1	0
13	4q35.2	Yes	200	191053845...191254119				Yes	226	191028537...191254119	Yes	158	191052245...191210542	Gain		FRG1, TUBB4Q, FRG2, DUX4	0
14	7p22.1	Yes	157	6838697...6995298							Yes	155	6840798...6995298	Gain			1
15	7p11.1	Yes	117	57640100...57757406	Yes	114	57640100...57753919	Yes	149	57604989...57753919	Yes	123	57597399...57720623	Gain			1
16	8p23.1	Yes	203	12415742...12618442				Yes	199	12415742...12614748	Yes	136	12415742...12551430	Gain			1
17	8p11.23	Yes	139	39349470...39488053	Yes	151	39354748...39506110	Yes	133	39354748...39488053	Yes	151	39354748...39506110	Loss			0
18	9p11.2	Yes	21937	44336683...66273526				Yes	21015	45258754...66273526	Yes	21015	45258754...66273526	Gain		FAM27A, FAM75A7	0
	9q12	Yes	861	68352238...69213455	Yes	103	68115006...68218485	Yes	1180	68076544...69256300	Yes	1099	68115006...69213671	Gain		FOXD4L6, CBWD6, ANKRD20A4, CCDC29	1
					Yes	457	68352238...68809437							Gain			
19	10q11.1	Yes	129	41974796...42103488	Yes	236	41934430...42170853	Yes	105	41934430...42039743	Yes	239	41934430...42173117	Gain			1
20	11p15.4	Yes	206	3383178...3588946				Yes	205	3376078...3580813	Yes	131	3430789...3561991	Gain			1
21	14q11.2				Yes	186	21602854...21788783	Yes	172	21625813...21787161				Loss			0
22	14q11.2				Yes	226	21804698...22030660	Yes	226	21804698...22030660				Loss			0
23	14q32.33	Yes	386	105190672...105576359	Yes	253	105345270...105597999	Yes	411	105190672...105601397	Yes	336	105265510...105601397	Gain			0
24	14q32.33	Yes	137	105760582...105897672	Yes	173	105645593...105818132	Yes	150	105638133...105788389	Yes	182	105640496...105822317	Gain			0
25	15q11.2				Yes	156	18682380...18838423	Yes	183	18655531...18838423	Yes	156	18682380...18838423	Gain		LOC283755, POTE15	1
					Yes	1067	18861808...19928521	Yes	572	18850029...19422452	Yes	344	18845990...19189673	Gain		OR4M2, OR4N4, LOC650137	1
											Yes	624	19207088...19835514	Gain			1
26	15q25.3				Yes	161	83524791...83685356	Yes	228	83524791...83752853	Yes	228	83524791...83752450	Gain		AKAP12	0
27	15q25.3				Yes	123	83784507...83907801	Yes	157	83784507...83941483	Yes	159	83790259...83949305	Gain		AKAP13	0
28	16p11.2	Yes	118	31882658...32000323				Yes	1131	32531735...33662480	Yes	1377	32303108...33680554	Gain		LOC729355, TP53TG3	1
		Yes	294	32088275...32382422										Gain			1
		Yes	185	32538757...32723310										Gain			1
		Yes	474	32962147...33436245										Gain			1
		Yes	211	33451476...33662480										Gain			1
29	17q21.31	Yes	586	41521621...42107467	Yes	198	41521621...41719935	Yes	586	41521621...42107467	Yes	407	41700624...42107467	Gain		KIAA1267, LRRC37A, ARL17, LRRC37A2, NSF	1
30	18p11.21	Yes	1545	15262486...16807594							Yes	130	15218647...15348836	Gain		ROCK1	1
31	19q13.31	Yes	116	47991257...48107552				Yes	133	47991257...48123857	Yes	235	47986218...48221228	Loss		PSG1, PSG6, PSG7, PSG11	1
32	21p11.2	Yes	204	9758730...9962501	Yes	204	9758730...9962501	Yes	204	9758730...9962501				Gain		TPTE	0
33	22q11.22	Yes	148	21300127...21448190				Yes	200	21298324...21498767	Yes	178	21298324...21476564	Gain		GGTL4	0
34	Xp11.23	Yes	112	47917899...48029446	Yes	269	47917899...48186708	Yes	184	47917899...48102337	Yes	257	47935225...48192383	Gain		SSX5, SSX1, SSX9, SSX3	1
35	Xq13.1	Yes	185	71869375...72054837							Yes	185	71869375...72054837	Gain	Yes	DMRTC1	1

Identity of *inherited* CNVs found in Family 1, Family 2 and the gene regions which they overlap. *Inherited* CNVs are those which are present in either or both parents and transmitted to either or both twins. All size is in kb. SD indicates the percentage of overlap between segmental duplications and CNVs. '0' means there is no overlap between CNV and segmental duplication, '1' means 90-100% and '2' means 50-90% overlap. Parental CNVs not transmitted to offspring were not included in *inherited* and *de novo* Table so the total number of CNVs present in Table 2-4 was not same as *inherited* and *do novo* Table. The table includes genomic locations as well as individual specific break points which allow for the assessment of regions of overlap with the Database of Genomic Variants (Toronto, Ontario). SI No. = Serial number. Novel indicates a CNV which is not present in The Database of Genomic Variants (DGV).

Appendix Table 5: *de novo* CNVs in Family 1 identified by PGS.

Sl. No.	Location	I-2-1	I-2-2	Gene
1	1p36.11	Gain	Gain	RHD
2	1p21.1	Loss		
3	2p25.3	Gain	Gain	TPO
4	2q22.3	Loss		
5	3p24.1	Loss	Loss	
6	3p	Loss	Loss	EPHA3
7	4q13.2	Gain	Gain	
8	4q28.3		Loss	
9	4q34.1	Gain	Gain	GALNTL6
10	4q34.1		Gain	GALNTL6
11	4q34.3	Loss		LOC285501
12	5p	Gain	Gain	
13	6p21.33	Gain	Gain	
14	7q34	Gain	Gain	
15	8p23.3	Loss	Loss	
16	9q21.13		Gain	
17	10p12.1	Loss		
18	10q26.12	Loss	Loss	
19	12q12	Loss	Loss	
20	13q21.31		Loss	OR7E156P
21	15q11.2	Loss	Loss	
22	16q22.1	Loss	Loss	PDPR, CLEC18C
23	18q22.1	Loss		
24	19q13.2	Gain		

Appendix Table 6: *de novo* CNVs in Family 2 identified by PGS.

Sl. No.	Location	II-2-1	II-2-2	Gene
1	1p31.1		Loss	
2	1p13.3		Loss	GSTM1
3	1q31.1	Loss		
4	2p23.3	Loss	Loss	
5	2q33.3	Loss	Loss	
6	3q21.2	Loss	Loss	
7	4q12		Gain	
8	5p15.2	Gain		
9	5p13.3	Gain	Gain	
10	5q14.2	Loss		ATG10
11	6p21.33	Loss	Loss	
12	6q14.1		Loss	
13	7q33	Loss	Loss	
14	7q33		Loss	
15	8p11.23	Loss	Loss	ADAM5P, ADAM3A
16	8p11.23	Loss		ADAM3A
17	8q23.3		Loss	
18	8q23.3		Loss	
19	9p21.3	Gain		
20	10q11.22		Loss	
21	10q21.3	Gain	Gain	
22	10q26.13		Loss	DMBT1
23	11p15.1		Gain	MRGPRX1
24	11p11.12	Loss	Loss	LOC440040
25	12p	Loss	Loss	
26	16q12.2	Gain		CES4
27	18q12.3	Loss	Loss	
28	19q13.12		Loss	
29	22q11.23		Gain	GSTT2B, GSTT2, DDTL, DDT, GSTTP1

Appendix Table 7: *Inherited* CNVs in Family 1 by PGS.

Sl. No.	Location	I-1-1	I-1-2	I-2-1	I-2-2	Gene
1	1p34.3		Loss	Loss	Loss	
2	1p21.1		Loss	Loss	Loss	
3	1q21.3	Loss		Loss	Loss	LCE1E, LCE1D
4	1q31.3	Loss			Loss	
5	2p11.2	Loss		Loss	Loss	
6	2q11.2	Gain			Gain	ANKRD36B
7	3q21.2	Loss		Loss	Loss	
8	3q26.1		Loss	Loss	Loss	
9	4q12	Loss		Loss		HOPX
10	4q13.2	Gain		Gain	Gain	UGT2B17
11	4q34.1	Gain		Gain	Gain	
12	5p15.2		Gain	Gain	Gain	
13	5q11.2	Loss		Loss	Loss	RAB3C
14	5q23.1		Loss	Loss	Loss	COMMD10
15	6q14.1	Gain		Gain		
16	7q11.21		Loss	Loss	Loss	INTS4L1, ZNF92
17	7q34	Gain		Gain	Gain	TRY6
18	8p23.1	Loss		Loss	Loss	DEFB103A, DEFB103B, SPAG11B, DEFB104B, DEFB104A, DEFB106B, DEFB106A, DEFB105B, DEFB105A, DEFB107A, DEFB107B, FAM90A7, SPAG11A, DEFB4, FAM66E, DEF109P1B
19	8p21.2	Gain		Gain	Gain	
20	8p11.23	Gain	Gain	Gain	Gain	ADAM5P ADAM3A
21	9p21.3		Loss	Loss	Loss	
22	10q21.3	Gain		Gain	Gain	
23	11p15.4	Gain	Gain	Gain	Gain	OR51A2
24	11p15.4		Loss	Loss	Loss	OR52N5, OR52N1
25	13q21.33	Loss		Loss	Loss	
26	14q32.33		Gain	Gain	Gain	
27	18q22.1		Loss	Loss		
28	22q11.21	Gain		Gain	Gain	PI4KAP2

Appendix Table 8: *Inherited* CNVs in Family 2 by PGS.

Sl. No.	Location	II-1-1	II-1-2	II-2-1	II-2-2	Gene
1	1p31.1		Gain	Gain	Gain	
2	1p21.1		Gain	Gain	Gain	AMY2A, AMY1A, AMY1C, AMY1B
3	1q21.1		Loss	Loss	Loss	
4	1q21.3		Loss	Loss	Loss	LCE3C
5	2p22.3		Gain	Gain	Gain	
6	2p16.3		Loss	Loss	Loss	
7	2q33.1		Loss	Loss	Loss	
8	2q37.1	Loss			Loss	
9	3p26.1	Gain		Gain	Gain	
10	3q12.2	Gain		Gain	Gain	GPR128, TFG
11	4q13.2	Loss		Loss	Loss	UGT2B28
12	4q32.2	Loss	Loss	Loss	Loss	
13	4q34.1	Loss	Loss	Loss	Loss	
14	4q34.1	Loss		Loss	Loss	GALNTL6
15	5p15.1	Loss	Loss	Loss	Loss	
16	5q11.2	Gain	Gain		Loss	
17	6q13	Loss		Loss	Loss	
18	6q14.1		Loss	Loss	Loss	
19	6q16.3	Gain		Gain	Gain	
20	6q21		Loss	Loss	Loss	
21	6q27		Gain	Gain	Gain	RPS6KA2
22	7q21.13	Loss		Loss	Loss	
23	8p21.2		Loss	Loss	Loss	
24	10p12.1	Loss	Loss	Loss	Loss	KIAA1217
25	11p15.4		Loss	Loss	Loss	OR51A4
26	11q22.3	Loss		Loss	Loss	CASP12
27	12p13.31	Gain	Gain	Gain	Gain	
28	12p11.22		Loss	Loss		
29	13q13.3	Gain		Gain	Gain	
30	14q		Loss	Loss		POTEG, P704P, OR4Q3, OR4M1, OR4N2, OR4K5, OR4K1
31	14q32.33	Loss		Loss	Loss	
32	15q25.3		Gain	Gain	Gain	AKAP13
33	16q23.1		Gain	Gain	Gain	CHST5, FLJ22167
34	16q23.1		Loss	Loss	Loss	WWOX
35	17q21.31	Gain		Gain	Gain	ARL17, LRRC37A2,

						ARL17P1, NSF
36	19q13.31	Loss		Loss	Loss	PSG10, PSG1, PSG6, PSG7, PSG11
37	20p13		Loss	Loss	Loss	SIRPB1
38	20q12		Loss	Loss	Loss	PTPRT
39	22q11.23		Loss		Loss	GSTTP1
40	22q11.23		Loss	Loss	Loss	GSTT1, GSTTP2

Appendix Table 9: *de novo* CNVs in Family 1 identified by PennCNV.

Sl. No.	Location	I-2-1	I-2-2	Gene
1	1p31.1	Gain	Gain	
2	2q22.3	Loss	Loss	
3	5p	Gain		
4	5q11.2		Loss	RAB3C
5	7q33	Gain	Gain	
6	7q34	Gain	Gain	
7	8p23.1	Loss	Loss	FAM66E, DEF109P1B
8	10p12.1	Gain		KIAA1217
9	11q22.3		Loss	CWF19L2
10	15q11.2	Loss	Loss	
11	15q26.3		Loss	OR4F4
12	16q23.1	Gain		WWOX
13	22q11.23		Gain	LOC391322, GSTT1, GSTTP2

Appendix Table 10: *de novo* CNVs in Family 2 identified by PennCNV.

Sl. No.	Location	II-2-1	II-2-2	Gene
1	2p23.3		Loss	
2	2q33.3	Loss	Loss	
3	4q34.1		Loss	
4	5p13.3		Gain	PDZD2, GOLPH3
5	6p21.33	Loss		
6	6p21.32		Loss	
7	6q16.3	Gain	Gain	
8	6q21		Loss	
9	7q33	Loss	Loss	
10	8p23.1		Gain	DEFB103A, DEFB103B, SPAG11B, DEFB104B, DEFB104A, DEFB106B, DEFB106A, DEFB105B, DEFB105A, DEFB107A, DEFB107B, FAM90A7
11	11p15.1	Gain	Gain	MRGPRX1
12	12p11.22	Loss		
13	13q12.12	Loss	Loss	
14	14q11.2		Loss	OR4K2
15	18q12.3	Loss	Loss	
16	19q13.31	Loss	Loss	PSG10, PSG1, PSG6, PSG7
17	22q11.23		Loss	GSTTP1, LOC391322, GSTTP2

Appendix Table 11: *Inherited* CNVs in Family 1 by PennCNV.

Sl. No.	Location	I-1-1	I-1-2	I-2-1	I-2-2	Gene
1	1p34.3		Loss	Loss	Loss	
2	1p21.1	Gain	Loss	Loss	Loss	
3	1q21.3	Loss		Loss	Loss	
4	2p16.3		Gain	Gain	Gain	
5	2p11.2	Loss		Loss	Loss	
6	3p		Loss	Loss	Loss	EPHA3
7	3q26.1		Loss	Loss	Loss	
8	4q13.2	Gain		Gain	Gain	TMPRSS11E, TMPRSS11E2, UGT2B17
9	4q32.2	Gain		Gain		
10	4q34.1	Gain		Gain	Gain	
11	4q34.1	Gain			Gain	GALNTL6
12	5q23.1		Loss	Loss	Loss	COMMD10
13	7q11.21	Gain	Loss	Loss	Loss	INTS4L1, ZNF92
14	7q34	Gain		Gain		TRY6
15	8p23.1	Loss			Loss	DEF109P1B, DEFB103A, DEFB103B, SPAG11B, DEFB104B, DEFB104A, DEFB106B, DEFB106A, DEFB105B, DEFB105A, DEFB107A, DEFB107B, FAM90A7, SPAG11A, DEFB4
16	8p11.23	Gain	Gain	Gain	Gain	ADAM5P ADAM3A
17	9p21.3		Loss	Loss		
18	11p15.4	Gain	Gain	Gain	Gain	OR51A4, OR51A2
19	11p15.4		Loss	Loss	Loss	OR52N5, OR52N1
20	14q	Gain		Gain		POTEG, P704P, OR4Q3, OR4M1, OR4N2, OR4K2, OR4K5, OR4K1
21	14q32.33		Gain	Gain	Gain	
22	16p11.2	Gain		Gain	Gain	LOC283914, LOC283914, LOC146481
23	22q11.21	Gain		Gain		POM121L8P, RIMBP3C, RIMBP3B, HIC2, PI4KAP2

Appendix Table 12: *Inherited* CNVs in Family 2 by PennCNV.

Sl. No.	Location	II-1-1	II-1-2	II-2-1	II-2-2	Gene
1	1p31.1		Gain	Gain	Gain	
2	1p21.1		Gain	Gain	Gain	
3	1q21.1		Loss	Loss	Loss	
4	2p22.3	Loss		Gain		
5	3q12.2	Gain		Gain	Gain	GPR128, TFG
6	4q13.2	Loss		Loss	Loss	UGT2B28
7	4q32.2		Loss	Loss	Loss	
8	4q34.1	Loss			Loss	GALNTL6
9	6q14.1		Loss	Loss	Loss	
10	6q27		Gain	Gain	Gain	RPS6KA2
11	7q21.13	Loss		Loss	Loss	
12	7q34	Loss		Loss	Loss	
13	8p21.2		Loss	Loss	Loss	
14	8p11.23	Loss		Loss	Loss	ADAM5P, ADAM3A
15	8q23.3	Loss			Loss	
16	10p12.1	Loss		Loss	Loss	KIAA1217
17	11p15.4		Loss	Loss	Loss	OR51A2
18	14q32.33	Loss		Loss	Loss	
19	15q25.3		Gain	Gain	Gain	AKAP13
20	16q23.1		Gain	Gain	Gain	CHST5, FLJ22167
21	16q23.1		Loss	Loss	Loss	WWOX
22	19q13.12	Loss		Loss	Loss	
23	20p13		Loss	Loss	Loss	SIRPB1
24	20q12		Loss		Loss	PTPRT

VITA

NAME: Sujit Maiti

POST-SECONDARY EDUCATION AND DEGREES:
Vidyasagar University, India
1994-1997 H.B.Sc. (Physics)
Indian Statistical Institute, India
1997-1999 Post Graduate Diploma (Computer Science)
Annamalai University, India
2003-2005 M.Sc. (Masters of Computer Application)
The University of Western Ontario,
London, Ontario, Canada
2010-2012 M.Sc. (Human Molecular Genetics)

HONORS AND AWARDS:
Western Graduate Research Scholarships (WGRS)
2010-2012

RELATED WORK EXPERIENCE:
Graduate Teaching Assistant
2009-2012 BIOL/STATS 2244A
Bioinformatics Fellow
2009-2010 University of Western Ontario, Canada
Bioinformatics Fellow
1999-2009 Indian Statistical Institute, India

PUBLICATIONS:
Partha P. Majumder, Neeta Sarkar Roy, Herman Staats, T. Ramamurthy, **Sujit Maiti**, Goutam Chowdhury, Carol Whisnant, K. Narayanasamy, Diane Wagener (2012). Genomic correlates of variability in immune response to an oral cholera vaccine. **European Journal of Human Genetics** (in press).

Maiti Sujit, Kumar KH, Castellani CA, O'Reilly R, Singh SM. Ontogenetic de novo copy number variations (CNVs) as a source of genetic individuality: studies on two families with MZD twins for schizophrenia. **PLoS One. 2011** Mar 2;6(3):e17125.

Majumder PP, Staats HF, Sarkar-Roy N, Varma B, Ghosh T, **Maiti Sujit**, Narayanasamy K, Whisnant CC, Stephenson JL, Wagener DK. Genetic determinants of immune-response to a polysaccharide vaccine for typhoid. **Hugo J. 2009** Dec;3(1-4):17-30. Epub 2010 Mar 11.

Bairagya BB, Bhattacharya P, Bhattacharya SK, Dey B, Dey U, Ghosh T, **Maiti Sujit**, Majumder PP, Mishra K, Mukherjee S, Mukherjee S, Narayanasamy K, Poddar S, Roy NS, Sengupta P, Sharma S, Sur D, Sutradhar D, Wagener DK. Genetic variation and haplotype structures of innate immunity genes in eastern India. **Infect Genet Evol. 2008** May;8(3):360-6.