DETECTION OF REARRANGEMENT HOTSPOTS AND THEIR IMPLICATION IN COMPLEX HUMAN DISEASES

MOHAMMED UDDIN

DETECTION OF REARRANGEMENT HOTSPOTS AND THEIR IMPLICATION IN COMPLEX HUMAN DISEASES

By

Mohammed Uddin

A thesis submitted to the School of Graduate Studies
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Discipline of Genetics
Faculty of Medicine
Memorial University of Newfoundland
St. John's, Newfoundland and Labrador
January 7th, 2013

ABSTRACT

A segment of DNA can vary in the number of copies between two or more genomes known a copy number variation (CNV). CNVs are a common genomic variant that is now shown to be associated with numerous diseases. A large portion of the genome (approximately 12%) has shown to be vulnerable in producing common or rare CNVs. These genomic regions are often prone to rearrangement due to the underlying molecular mechanisms (i.e. non-allelic homologous recombination (NAHR), non homologous end joining (NHEJ), fork stalling and template switching (FoSTeS) and microhomology-mediated break-induced recombination (MMBIR)) that gives rise to inter-individual genetic differences through copy number changes.

Segmental Duplication (SD) is another type of genomic variant and is at least 1kb in length with >90% sequence identity with other genomic regions, constitutes a significant portion of the human genome. A pair of SD block that are homologous to each other influences the rate of NAHR events, often resulting in CNVs. Large portions of SDs overlap with CNVs, some of which are associated with disease. Although investigating SD is important to the study of evolution, and for inferring the underlying mechanism of nearby CNVs, the structural relationship of SDs and CNVs in complex disease is yet to be elucidated. Detecting SD is notoriously complex and genotyping single nucleotide polymorphism (SNP) within these regions is near impossible with traditional array based approaches.

The primary objectives of this thesis are – i) to detect CNVs arising from complex genomic rearrangements mediated through SDs; ii) to design a custom microarray, targeting rearrangement hotspot regions and iii) to identify novel CNVs associated with complex disease (i.e. Ankylosing Spondylitis (AS) and Tourette Syndrome (TS)) using the custom microarray.

Until the advent of high throughput whole genome sequencing, the primary CNV detection methods included SNP genotyping arrays, comparative genomic hybridization (CGH) arrays, clone-PCR product arrays, and fluorescent *in situ* hybridization (FISH). Each method is reported to have pros and cons show no single method have the capacity to detect the entire CNV content in a genome. In this thesis, the use of SNP microarrays (primarily designed for SNP genotyping) to detect CNVs in a complex disease cohort was explored. This analysis shows the limited capacity of existing SNP arrays for the detection of CNVs. In light of this apparent complexity on the detection of CNVs, a hybrid model was described in the second chapter to detect SDs and CNVs using both whole genome sequencing and microarray technologies. The whole genome sequence analysis was performed to detect approximately 2000 rearrangement hotspots within SDs for an African genome (18x coverage). A high density microarray consisting of 2 X 1million probes was custom designed targeting the hotspots.

The application of the microarray leads to the detection of a large number of CNVs and was applied on two complex (i.e. Ankylosing Spondylitis (AS), and Tourette Syndrome (TS)) diseases to identify novel disease associated CNVS. The third chapter shows the detection of a highly stratified gene *UGT2B17* in copy number and its association with Ankylosing Spondylitis. The fourth chapter of this thesis show the application of the custom array that leads to the detection of a novel locus at 2q21.1-q21.2 for Tourette syndrome. CNVs breakpoint within the locus show atypical micro-duplications includes *C2orf27A* gene that segregates within multiple generations correlates with severe TS phenotypes.

ACKNOWLEDGEMENTS

I would like to thank Dr. Proton Rahman and Dr. Michael Woods for their guidance during my research project and the subsequent writing of our publications and this thesis.

My thanks are also extended to my supervisory committee member Dr. Roger Green. I extend my appreciation to Dr. Darren O'Rielly for his kind guidance and help.

I would like to recognize the Atlantic Innovation Fund (AIF) and Canadian Institutes of Health Research (CIHR), for salary support I received for this project. The financial contribution for my project through operating grants awarded to Dr. Proton Rahman.

I would also like to give thanks to my colleagues at the Population Therapeutic Research Group (PTRG), particularly Mitch Sturge and Catherine Street for their kind support.

Finally, I would like to thank my parents for their constant support. Thanks to my wife Syeda Nusrat Ferdaus for her kind support and, my son Aahiyan Ferdaus for being the joy of my life.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AS	Ankylosing Spondylitis
ASD	Autism Spectrum Disorder
aCGH	Array Comparative Genomic Hybridization
ADHD	Attention Deficit Hyperactivity Disorder
BAC	Bacterial Artificial Clones
BAF	B-Allele Frequency
CNV	Copy Number Variation/Variants
CNVR	Copy Number Variation Region
cDNA	Complementary DNA
dbGaP	The Database of Genotypes and Phenotypes
DD	Developmental Delay
DND	Developmental Neuropsychiatric/cognitive Disorders
DNVs	Duplicated Gene Nucleotide Variants
DP	Dynamic Programming
FoSTeS	Fork Stalling and Template Switching
FISH	Fluorescent *in situ* Hybridization
HapMap	Haplotype Map
HNPCC	Hereditary Nnon-Polyposis Colorectal Cancer
KBp	Kilo Base Pair
LCR	Low Copy Repeat
LINE	Long Interspersed Nuclear Element
mrFAST	Micro Read Fast Alignment Search Tool
mrsFAST	Micro-Read Substitution-only Fast Alignment Search Tool

MBp	Mega Base Pair
MLPA	Multiplex ligation-dependent probe amplification
MAPH	Multiplex Amplifiable Probe Hybridization
MAQ	Multiplex Amplicon Quantification (MAQ)
MAQ	Mapping and Assembly with Quality
MSY	Male Specific Y chromosomal Region
MMBIR	Microhomology-Mediated Break-Induced Recombination
MR	Mental Retardation
NAHR	Non Allelic Homologous Recombination
NHEJ	Non Homologous End Joining
OCD	Obsessive Compulsive Disorder
PsV	Psoriasis Vulgaris
PAR1	Pseudoautosomal Region 1
PCR	Polymerized Chain Reaction
QF-PCR	Quantitative Fluorescent Polymerase Chain Reaction
Q-FISH	Quantitative Fluorescent *in situ* Hybridization
RA	Rheumatoid Arthritis
Rt-PCR	Real-Time Polymerase Chain Reaction
RD	Read Depth
SINE	Short Interspersed Nuclear Element
SV	Structural Variation
SD	Segmental Duplication
SLE	Systemic Lupus Erythematosus
TS	Tourette Syndrome
WTCCC	The Wellcome Trust Case Control Consortium

Chapter 1

INTRODUCTION

1.1 COPY NUMBER VARIATION

Structural variation (SV) was originally defined as insertions, deletions, and inversions greater than 1 kb in size (Feuk et al., 2006). With the high throughput sequencing of human genomes now becoming routine, the detection spectrum of structural variants (SVs) and copy number variation/variants (CNVs) has widened to include much smaller events (for example, those >50 bp in length) (Mills et al., 2011). Genomic CNVs describe a unique DNA segment that differs in copy number between two or more genomes. The earliest evidence of gene CNV found to be associated with a phenotype was reported in 1936, when researchers discovered that the BAR duplicated in Drosophila melanogaster, which was shown to cause the Bar eye phenotype (Bridges, 1936). The evolving definition of CNV now includes a size constraint; as new technology emerges with high resolution detection capacity, the length of a CNV is now considered to be as small as 500bp (Conrad et al., 2010; Mills et al., 2011). Depending on the method employed for assessment, up to about 12% of the human genome is thought to be comprised of CNVs. A significant portion of these regions contain genes that are functionally active in humans. That over 41% of all CNVs identified overlap with known genes suggests that CNVs may play a prominent role in modulating the regulation of genes. Current estimates are that more than 240 Mb of the reference genome is comprised of CNVs and are present in at least 6% of each chromosome (Scherer SW et al., 2007; Zogopoulos et al., 2007; McCarroll et al., 2008; Perry et al., 2008; Henrichsen et al., 2009).

Over the past few years, CNVs have attracted much attention due to the fact that they are common in the human genome and can have dramatic phenotypic consequences, including the capacity to alter gene dosage, disrupt coding segments, and regulate functional genes (Barbara et al., 2007). CNVs are now routinely analyzed to identify disease susceptible loci in the genome. Despite extensive studies, the total number, position, size, gene content, and population distribution of CNVs remain elusive. Moreover, small CNVs (i.e., 0.5–50 kb in size) have been under-ascertained because there has not been an accurate molecular method available to study these smaller rearrangements on a genome-wide scale in different populations (Hurles et al., 2008). The analysis of Conrad et al. is, to date, the most comprehensive catalogue of CNVs; it used 42 million probe aCGH microarrays for three global populations. The 1,000 Genome Project also performed a large CNV detection analysis using high throughput sequencing for major global populations (Mills et al., 2011; Sudmant et al., 2010). These studies used both microarray and sequence technology, providing the most comprehensive common and rare CNV catalogue to date. They identified a set of genes that are highly stratified in terms of copy number within numerous world populations, and they postulated that these CNVs are associated with diseases.

CNVs may have a range of involvement in reference to diseases. In numerous genomic disorders, CNVs manifest with the phenotype as de novo or inherited variants. Within neuropsychiatric disorders (i.e., autism spectrum disorder (ASD), schizophrenia, etc.), CNVs have been shown to be highly penetrant, primarily manifesting as de novo variants (Pinto et al., 2010). Genomic regions containing CNVs may harbor important genes and gene regulatory elements and, therefore, may substantially influence gene expression and, thus, phenotypic diversity (Barbara et al., 2007; Aldred et al., 2005). More importantly, these CNVs within putative genes and their associated pathways could be potential therapeutic targets in the future. At least two distinct models have been proposed with respect to associations between disease and structural variation. The first involves large variants (i.e., typically gains and losses several hundred kilobase pairs in length) that are individually rare in the population (<1%); however, in a control population at least 11% can have large CNVs of at least 400kb in size. A significant fraction of diseases, as seen in reference to some neurological and neurocognitive disorders, can be explained by the presence of such large variants (Stankiewicz et al., 2002; Sebat et al., 2007; Sharp et al., 2006; de Vries et al., 2005). At least 14.2% of developmental delays have been associated with large CNVs (>400kb), which are rare in control populations (Cooper et al., 2011). The second includes multi-copy gene families that are commonly copy number variables and which contribute to complex disease susceptibility, as frequently seen for traits related to immune gene functions (Fellermann et al., 2006; Aitman et al., 2006). It has been reported that dosage sensitive genes may have considerable influence on disease susceptibility (or, alternatively, disease resistance) in humans.

The discovery and genotyping of such small-length CNVs has been central to the understanding of their contribution to complex disease pathogenesis. At present, only the "tip of the CNV iceberg" has been explored in relation to complex trait genetics. In recent years, initiatives involving large-scale population genome analysis have revealed the structural complexity of CNVs. In reference to understanding the contribution of CNVs to gene expression and disease pathogenesis, inferring these complex structural properties is essential. The structure of CNVs is noticeably different from other genomic repeat variants such as retrotransposons (long interspersed elements (LINE) and short interspersed elements (SINE)) and microsatellites (for example, di- or trinucleotide microsatellite tandem repeats). LINEs and SINEs (known as alu sequences in primate genomes) are usually short and repeated elements, whereas the microsatellite DNA is tandemly repeated (Bailey et al., 2002; Bailey et al., 2006). Intuitively, a CNV can overlap or contain retrotransposons as a variant (Asako et al., 2011). However, the presence of repeated elements within CNVs introduces complexities in terms of detection technologies.

Segmental duplication (SD) or low copy repeat (LCR) is another genomic variant that is conceptually similar to CNVs. Segmental duplications are blocks of DNA sequences that reached fixation after duplicating themselves multiple times within the genome through molecular mechanisms during the course of human evolution. Segmental duplications are at least 1kb in length with >90% sequence identity and, unlike CNVs, the majority of the SDs are greater in length. Phylogenetically, SDs were initially unique sequences which, over time, duplicated (often, multiple times) within the genome and reached a static state within the population or species. The current catalogue of SDs comprises approximately 5% of the human genome, encompassing 18% of genes (Bailey et al., 2002; Bailey et al., 2006). Some reports show SDs can vary in copy numbers among individuals or populations. A significant portion of CNVs overlap with SD regions; quantifying copy number for these regions is very difficult. Large-scale studies for detecting CNVs and SDs show that approximately 25% of SD regions overlap with CNVs in a population (Bailey et al., 2006; Redon et al., 2006; Alkan et al., 2011).

1.2 REARRANGEMENT HOTSPOTS AND CNVS

Genomic rearrangement is the mechanism behind the creation of structural variations. Initially, genomic rearrangement was shown to take place within the centromeric and telomeric regions of the chromosomes. These regions, known as "hotspots," are susceptible to rearrangement due to their sequence content. Numerous cases have been identified in which translocation/insertion/deletion/inversion or complex aberrations took place due to genomic rearrangement within these regions. In recent years, numerous mechanisms of rearrangement have been proposed (e.g., non-allelic homologous recombination (NAHR), non-homologous end joining (NHEJ), fork stalling and template switching (FoSTeS), and microhomology-mediated break-induced recombination (MMBIR)) which shed light on the concept of "genome-wide rearrangement hotspots," which are not limited to centromeric/telomeric regions and provide some new insights into non-centromeric/telomeric hotspot regions (Hastings et al., 2009). There is a wide range of functional consequences related to genomic rearrangements, including gene activation, modification of gene expression, and gene fusion.

Genomic rearrangement is an important biological mechanism that can be used to explain some complex human diseases etiology. In recent years, such pathogenic rearrangements have been identified in different types of cancers and in neuropsychiatric diseases. Among the four mechanisms, NAHR and NHEJ are recombination-based methods, whereas the proposed MMBIR and FoSTeS are replication-mediated. Apart from NAHR, the other recombination-based mechanism, NHEJ, is responsible for repair of DNA double-strand breaks. In NHEJ, double-strand breaks are detected; after that, both broken DNA ends are bridged, modified, and, finally, ligated (Weterings et al., 2004). The product of repair often contains additional nucleotides at the DNA end junction, leaving a “molecular scar” (Gu et al., 2008).

Different molecular mechanisms are thought to be responsible for producing genomic rearrangements with certain properties. For example, MMBIR has been shown to be the mechanism that produces smaller-sized SVs than each of the other mechanisms. Additionally, not many studies have provided genome scale evidence of aberrations that originated via MMBIR. In addition to NAHR and NHEJ, fork stalling and template switching (FoSTeS), a mechanism that is based on DNA replication error, has recently been shown to play an important role in the origin of genomic disorder-associated non-recurrent rearrangements that have complex structures (e.g., deletions and/or duplications interrupted by either normal copy numbers or triplicated genomic segments) (Lee et al., 2007; Zhang et al., 2009). In this model, the DNA replication fork stalls; then the lagging strand disengages from the original template and anneals to another replication fork in physical proximity by virtue of microhomology at the 3'end, "priming" or reinitiating DNA synthesis. Depending on the location of the new fork (i.e., upstream or downstream), a deletion or duplication, respectively, will occur. Whether the lagging or leading strand in the new fork is used as a template determines whether the erroneously incorporated fragment from the new replication fork will be in direct or inverted orientation with respect to its original position (Hastings et al., 2009). In this thesis, the NAHR is the primary mechanism that will be discussed and investigated.

Emerging evidence reveals an increasing frequency of deletions and duplications in and around segmental duplications in comparison to the unique portion of the human genome. Recently, the association of SD regions with genomic disorders has been reported frequently, with the vast majority representing novel sites whose genomic architecture is susceptible to disease-causing rearrangements (Sharp et al., 2005). Among the proposed mechanisms, NAHR is the well studied mechanism that produces CNVs mediated through SDs. Especially for NAHR; CNVs that are flanked by two SDs have been found to be near identical in sequence homology (see Figure 1.1). This recombination event takes place due to the presence of segmental duplications where two blocks of SDs that are homologous (completely or partially) to each other recombine. The region between the SDs (see Figure 1.1) which is prone to copy number changes is known as the “critical region.” The two homologous blocks of SDs can reside within the same chromosome or among different chromosomes, creating intra-chromosomal and inter-chromosomal recombination events, respectively. Importantly, NAHR-mediated events occur much more frequently than the rate for the creation of single-nucleotide polymorphism (SNPs; 1x10-4 compared with 1x10-8 per generation) (Hastings et al., 2009; Henrichsen et al., 2009; Shaffer et al., 2000). This molecular mechanism has been shown to be responsible for the vast majority of the common-sized recurrent rearrangements, including reciprocal deletions and duplications, or inversions. However, the complexity of their structural architecture in the human genome and, more importantly, their role in disease pathogenesis remains largely unknown. In the second chapter of this thesis, a proposed solution to the problem of detecting complex SD regions using high throughput sequencing technology is formulated and executed.

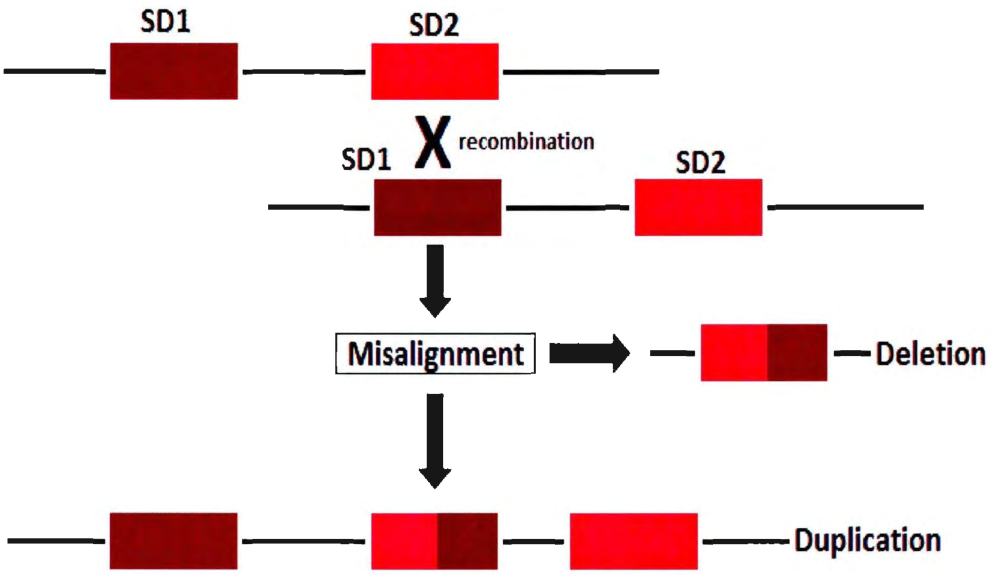


Figure 1.1 NAHR mechanism shows two common outputs from the misalignment of segmental duplications.

The proposed replication-based mechanisms (FoSTeS and MMBIR) have not been thoroughly studied. These mechanisms require population scale analysis using high throughput sequencing from researchers to better understand their contribution to genomic variants, especially in terms of CNVs. In contrast, NAHR and NHEJ have been thoroughly studied. One of the primary focuses of this thesis is to develop techniques to identify genomic regions that are NAHR-prone. The NAHR event and its resultant variant depend vastly on the orientation of the SDs. For example, two inverted SDs will produce an inverted SD; one inverted and one forward-oriented SD will produce a deletion; and two forward-oriented SDs will produce a forward-oriented SD. This recombination mechanism is capable of producing genomic gains or losses through the course of evolution and acting as a mutational mechanism in the human genome (Hastings et al., 2009). In recent years, multiple disorders have been reported to be associated with genomic gains and losses that are due to NAHR. Charcot-Marie-Tooth disease type 1A, Smith-Magenis Syndrome, and Neurofibromatosis type 1 are the classic examples of SD-mediated pathogenic genomic losses/gains that occur through NAHR mechanisms. The presence of the proximal and distal SDs is indicative of a possible NAHR mechanism that caused the pathogenic variants. The location of the proximal and distal SDs can be within a chromosome or between chromosomes. NAHR is the most common mechanism underlying disease-associated genome rearrangements, primarily in reference to neuropsychiatric diseases (Shaw et al., 2004).

There have been very few attempts to detect on a genome-wide scale the total content of SDs that contribute to the formation of CNVs. Moreover, the characterizations of SDs and CNVs have yet to be applied in large complex disease cohorts. The detection of rearrangement hotspots will lead to better understanding of the mechanistic formation of CNVs that are susceptible to various complex diseases. In this thesis, I have proposed a technique to detect rearrangement hotspots within SDs that are likely to produce CNVs, primarily through the NAHR mechanism.

1.3 DETECTION OF COPY NUMBER VARIATIONS (CNVs)

Initially, cytogenetic studies used microscopes to investigate gross chromosomal abnormalities. The first such observation involved the identification of an additional copy of chromosome 21 that is associated with

Down's syndrome. As technology utilizing high resolution detection capacity emerged, an explosion of submicroscopic CNVs were discovered in the human genome. The primary detection methods include microarray, comparative genomic hybridizations, clone-PCR product arrays, and fluorescent in situ hybridization (FISH). In recent years, high-throughput sequencing appears to be the most promising approach to detect genomic variants. This new perspective on human variation has been driven largely by the implementation of whole-genome scanning methods that enable researchers to interrogate the genome at a resolution intermediate between that of cytogenetic analysis using microscopy (>5–10 Mb) and that of DNA sequencing (1–700 bp) (Alkan et al., 2011). Below is a description of the primary detection methods that are commonly used in the detection of CNVs and segmental duplications.

Fluorescent in situ Hybridization (FISH):

Fluorescent in situ Hybridization (FISH), developed in the 1980s, allows visualization of chromosomal regions during the metaphase spread or in the interphase nuclei. The primary aim of this method is to detect a locus or its derivatives within the genome through a fluorescent signal through the use of a fluorescently labeled probe which hybridizes with the targeted locus and its derivatives. Two- and three-color FISH is primarily used to target genomic loci. The use of FISH in clinics is very common for the detection of large aberrations for patients with neuropsychological disorders. The known pathogenic large deletions (usually >40kb) or duplications are routinely tested for in patients with genomic disorders (e.g., Down's syndrome). In recent years, a variety of FISH techniques have been developed with increasing resolution (i.e., fibre FISH, Quantitative FISH or Q-FISH, Multiplex FISH, and flow FISH).

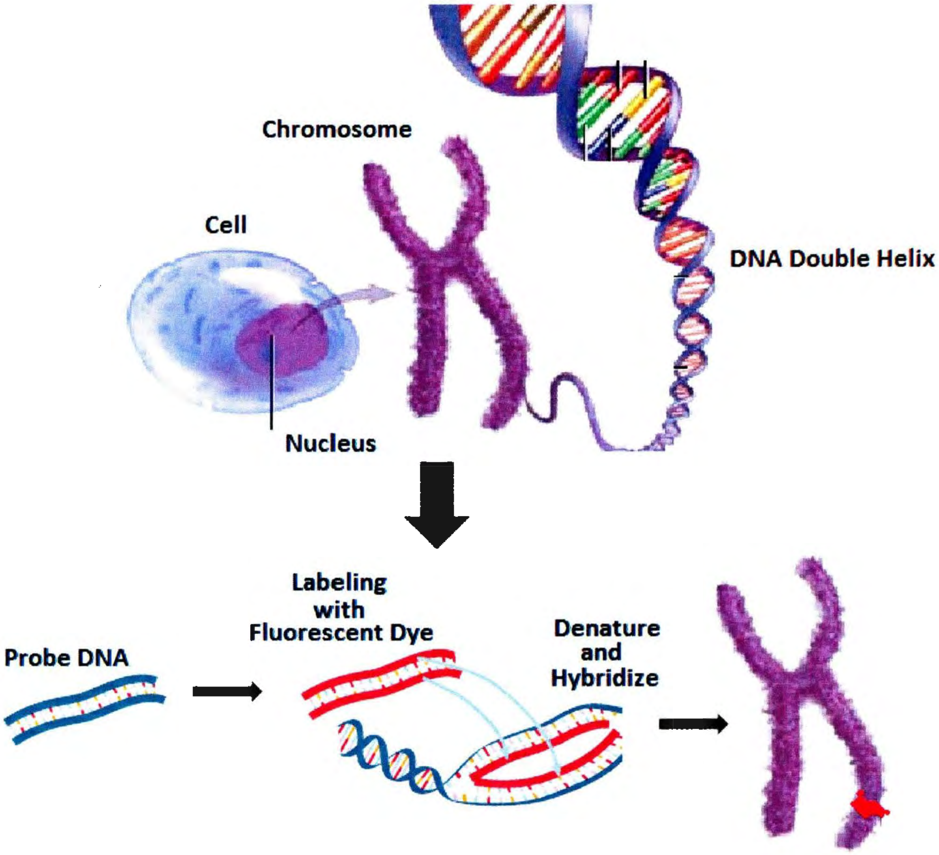


Figure 1.2 The representation of the processes that involve in fluorescent *in situ* hybridization (FISH).

Traditional approaches employ a fluorescently-labeled copy of the probe sequence (a fragment of DNA) that is large enough to produce a hybridized signal (see Figure 1.2). Prior to the hybridization process, both the target and the probe sequence are denatured with an appropriate amount of heat, which allows the formation of hydrogen bonds between the target and the probe sequences.

The probe and the target sequences are then mixed together, and the probe specifically binds to its complementary sequence on the chromosomes (Bayani and Squire, 2004). Probe and the target regions within chromosomes can be detected using a fluorescent microscope. The two major limitations of using the traditional FISH experiment include restrictions in terms of the number of loci as targets and the length of the aberrations that can be detected. The first restriction is due to the fact that introducing too many target loci usually allows for spurious signals which make it difficult to detect true signals using the fluorescent microscope; therefore, the common practice is to use only one or two probes simultaneously. The second restriction is the limitation of the use of a single genome per experiment. These restrictions do not allow detection of aberrations on large cohorts.

Clone and PCR-product Arrays:

The use of microarrays is the most common way of detecting genomic aberrations. Initially, bacterial artificial clones (BAC) were often used as templates to produce DNA microarrays; this method has the lowest noise from among all hybridization techniques. For the use of custom designing arrays, a catalogue of more than 30,000 overlapping clones covering the entire human genome is available. The length varies between 80 to 200 kb, which is a major limitation in reference to detecting CNVs that are less than 50kb in length. More recently, fosmid and cosmid clones (bacterial genomic DNA) were commonly used for array-comparative genome hybridization because it provided an improved CNV detection capacity of at least 20kb in length. There are examples of the use of cDNA (complementary DNA) clones for the construction of array-comparative genome hybridization targeting an entire or partial gene. Using cDNA often produces uneven distribution due to the lack of coverage of the genome, which, in turn, reduces the capacity to detect high resolution breakpoints of CNVs.

PCR and quantitative PCR-based methods provide rapid and accurate cost efficient approaches to detect CNV in a large number of sample sizes. These approaches are limited in terms of the number of loci they can test simultaneously, and so, are appropriate for targeting narrow regions in the genome. Among these methods, multiplex ligation-dependent probe amplification (MLPA) and multiplex amplicon quantification (MAQ) can simultaneously assay a number of loci in a large number of samples. The most commonly used approach is the developing technology that can utilize the PCR product via real-time quantitative PCR (rt-qPCR) for the efficient detection of CNVs located on a single DNA fragment (Schouten et al., 2002; Schaeffeler et al., 2003; Weksberg et al., 2005; Kumps et al., 2010).

Among various molecular approaches, MLPA is interesting because of its ability to detect multiple loci and their copy number changes in one simple PCR reaction, making it an attractive alternative to FISH (Abdool et al., 2010). One of the primary limitations of this method is the restricted number of loci involved, which falls short for genome-wide analysis of CNV detection. MAQ is another approach that has been shown to be a low-cost and high-throughput PCR-based technique that can accurately predict copy number alterations (Kumps et al., 2010). Currently, MAQ allows 40 targets in one reaction, which is sufficient for routine clinical diagnostics, but not enough for genome-wide scale analysis. The most prominent molecular technique is rt-qPCR-based copy number screening, which serves as the method of choice for targeted screening. A large number of samples can be analyzed to detect CNVs for a restricted number of targeted loci. All of these approaches are widely used as validation methods for targeted regions after the analysis of whole genome CNV detection using other microarrays. CNVs that overlap consist of regions with common repeats and segmental duplications which require extra care in reference to detection and analysis.

Genotyping Arrays:

In the last decade, numerous disease loci have been identified using high throughput genotyping arrays for the investigation of single nucleotide polymorphisms (SNPs) associated with a disease. Subsequently, the development of bioinformatics approaches has allowed researchers to detect CNVs by re-using the probe intensity data that was initially designed for SNP genotyping. Two very important sources of genotyping data exists: dbGAP (The database of Genotypes and Phenotypes) and the HapMap project archive. The dbGaP project includes all of the SNP microarray data primarily for common complex diseases (Mailman et al., 2007), whereas the HapMap project produced the catalogue of SNPs for major world populations. The current genotyping platforms (i.e., Affymetrix and Illumina) offer comprehensive coverage of the human genome; recently, these high density data from dbGaP and HapMap have been used in multiple studies to produce CNV catalogues.

The primary difference between other array-based approaches and the SNP genotyping array is that, when using SNP, hybridizations are not performed using cohybridization of two DNA samples, as in array-CGH. Instead, probe hybridization occurs for a single DNA sample. To reduce the signal-to-noise ratio, the DNA is first digested with a restriction enzyme and then ligated with adapters. To minimize the complexity of hybridization, smaller fragments are amplified using universal primers. To detect CNVs using SNP genotyping platforms, the probe signal intensities the match, and mismatch probes are compared with values from another individual (or a group of individuals treated as reference individuals); then the relative copy number per locus is determined (Dennise et al., 2006). The reference individuals have to match ethnically and phenotypically. Although the primary aim is to genotype SNPs, the detection capacity is very high compared to clone-based arrays. The detection capacity is dependent, mostly in terms of the distribution of the probes. For example, the early Affymetric 500k consisted of probes targeting SNPs with a 2.5kb median gap (WTCCC, 2007).

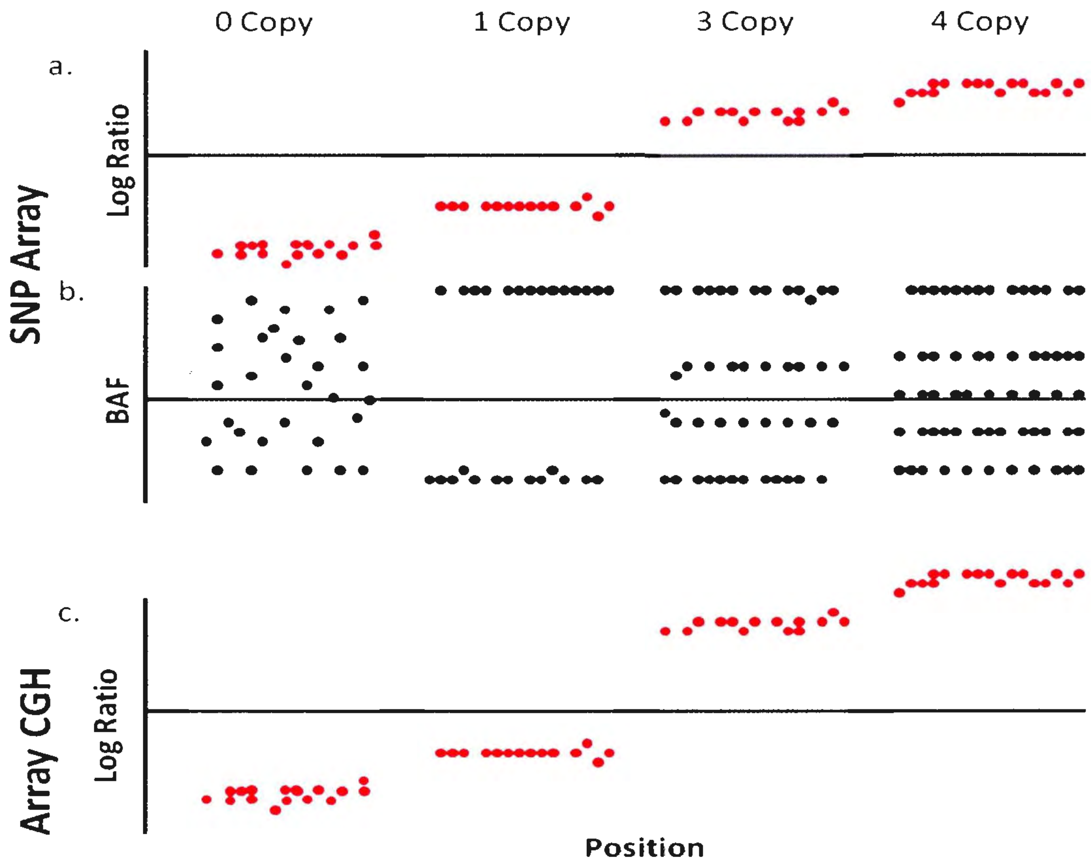


Figure 1.3 Illustration of an SNP microarray and an array CGH copy number detection pattern. In the SNP microarray, **a)** the signal intensities a sample is analyzed with reference to a set of samples from the same population. The B-allele frequency is another metric that **b)** can be calculated as the proportion of the total allele signal (A + B) explained by a single allele (A). In combination with the log2 ratio pattern, the BAF may be used to accurately assign copy numbers from 0 to 4 in diploid regions of the genome. **c)** aCGH generates a similar intensity metric except for the fact that the intensities do not depend on a

reference set. In contrast, the signal is produced by hybridizing a test with a reference sample and the intensity is normalized and converted to a log2 ratio, which acts as a proxy for copy number.

Recent arrays (i.e., Affymetrix 6.0 SNP and Illumina 1M platforms) incorporate better SNP selection criteria for complex regions of the genome and non-polymorphic copy-number probes. Along with log intensity, a second metric, termed BAF (B allele frequency), can be utilized to determine the copy number of a locus (limited to 0 to 4 copies) by calculating the total allele signal (A+B, for allele A and B) as explained by a signal allele A (see Figure 1.3). The BAF has a significantly higher per-probe signal to noise ratio than the log ratio data and can be interpreted as follows: a BAF of 0 represents the genotype (A/A or A/–), whereas 0.5 represents (A/B) and 1 represents (B/B or B/–). Different BAF values occur for AAB and ABB genotypes or more complex genotypes (for example, AAAB, AABB, and BBBA). Homozygous deletions result in a failure of the BAF to cluster (Cooper et al., 2008; Peiffer et al., 2006). Thus, the BAF may be used to accurately assign copy numbers from 0 to 4 in diploid regions of the genome.

Array comparative genome hybridization (aCGH):

Comparative genomic hybridization (CGH) was the first approach used to scan the entire genome comprehensively to detect variations in DNA copy numbers (Pinkel et al., 1998; Iafrate et al., 2004; Conrad et al., 2010). In a typical CGH measurement, total genomic DNA is isolated from test and reference cell populations that are differentially labeled and hybridized to metaphase chromosomes or, more recently, DNA microarrays. The relative hybridization intensity of the test and reference signals at a given location are then (ideally) proportional to the relative copy number of those sequences in the test and reference genomes. If the reference genome is normal, then increases and decreases in the intensity ratio directly indicates DNA copy-number variations in the genome of the test cells. More than two genomes can be compared simultaneously if distinguishable labels are available. Data are typically normalized, so that the modal ratio for the genome is set to some standard value, typically 1.0 on a linear scale or 0.0 on a logarithmic scale (Conrad et al., 2010) (see Figure 1.3). Additional measurements, such as FISH or flow cytometry, can be used to determine the copy number associated with a given ratio level.

Array-comparative genome hybridization (aCGH) technology, developed in the last decade, affords the capacity to examine the whole human genome on a single chip with a level of resolution dependent only on size and distance between the interrogating probes, a process also known as microarray-based cytogenetics (Pinkel et al., 1998). Microarray technology is superior to cytogenetic techniques such as conventional karyotyping and FISH analysis. The resolution afforded by microarray technology is at least 10-fold greater than the best pro-metaphase chromosome analysis obtained via conventional karyotyping, rendering it the most sensitive whole-genome screen for genomic deletions and duplications (Lee et al., 2007).

Most arrays cover a larger segment of the subtelomere and centromere than the existing FISH probes and, subsequently, can identify a greater number of rearrangements (Lee et al., 2007; Iafrate et al., 2004). In addition, they provide more information with respect to the size of the chromosome segment involved, which is often larger than 5 Mb, and may also define complex rearrangements within the subtelomere, which may include both deletion and duplication within the same region (Antonacci et al., 2010). Microarray-based cytogenetic testing characterizes CNV size and genomic location, which facilitates genotype-phenotype correlations. The increased resolution of microarray technology over conventional cytogenetic analysis allows for identification of chromosomal imbalances with greater precision, accuracy, and technical sensitivity; it represents the technical convergence of molecular genetics and cytogenetics and is rapidly revolutionizing modern cytogenetics. The use of aCGH technology for neuropsychiatric patients has led to the identification of atypical duplications and deletions in a growing number of them (Dawson et al., 2010).

Array-based comparative genome hybridization (aCGH) technology offers several advantages over SNP-based approaches, including: 1) coverage of genomic regions with low SNP incidences (i.e., SNP deserts); 2) the use of intra-experimental control rather than use of a pre-established laboratory standard; 3) a higher signal-to-noise ratio; and 4) the availability of custom, high-density probe arrays (Peiffer et al., 2006; Coe et al., 2007). These notable advantages have led to their widespread adoption in clinical diagnostics, essentially replacing karyotype analysis as the primary means of detecting copy-number alterations among children with developmental delays (Miller et al., 2010).

Customized aCGH can be applied to genotype the largest number of CNV loci, as recently demonstrated by the Wellcome Trust Case Control Consortium (WTCCC) (WTCCC, 2010). Microarray-based testing is currently considered to be the first-line test for most patients with neuropsychiatric diseases. Conrad et al. used multiple microarrays consisting of 42 million probes, tilling the entire genome and producing the most comprehensive CNV catalogue with the highest resolution to date. These identified CNVs using microarrays within three major global populations and are now used for association studies for complex disease cohorts. The application of microarray-based cancer cytogenetics has also led to the identification of putative oncogenes and tumor suppressor genes (Tagawa et al., 2005).

High-throughput Sequencing:

Over the past few years, there has been a major shift away from automated Sanger sequencing for the analysis of the genome. The advent of high-throughput sequencing technologies (i.e., Roche/454, Illumina/Solexa, Life/APG's SOLiD, Helicos BioSciences, etc.) promises to revolutionize structural variation and is a potential replacement for microarrays as the platform for discovery and genotyping. The technology differs for each platform in chemistry and the employed molecular approaches used to sequence the entire genome. The output sequence read length also varies among the commonly used platforms. For example, Roche/454, Illumina/Solexa, Life/APG's SOLiD, and Helicos BioSciences produce average sequence read lengths of 330bp, 32-100bp, 50bp, and 32bp, respectively.

However, high-throughput sequencing approaches present substantial computational and bioinformatics challenges. Most of the methods used to detect SVs are computational algorithms that were initially designed for capillary sequence reads or fully sequenced large insert clones (Volik et al., 2003; Tuzun et al., 2005). The algorithms are now improving and can undertake substantially large amounts of data produced by the high throughput sequencer to decipher the SVs from multiple genomes. There are three primary strategies to investigate sequence data in the detection of structural variations in a genome. Each of the strategies has a unique capacity for detection of SVs. Below is a description of the methodologies that are commonly applied on genome-wide sequence data.

Read-pair Technology:

Read-pair technology, instead of utilizing a standard single-read DNA library, facilitates the reading of both the forward and reverse template strands during one paired-end read. Each paired read includes sequence information and positional information that allows for highly precise alignment of reads in the genome. During alignment, the read-pair method computes the span and the orientation of paired-end reads and detects discordant pairs, where either the orientation or the span is not consistent with the reference genome (Tuzun et al., 2005; Kidd et al., 2008). The discordant pairs that indicate the mapping is too far apart infer that a deletion is present. Pairs that appear to map too closely together indicate an insertion, and if the orientations are inconsistent, then it implies a possible inversion or tandem duplication. Novel insertions also can be detected when only one end cluster is mapped.

The unique paired-end sequencing by Illumina usually allows inserts that are within 200 to 500bp in length. A typical paired-end run produces 75bp reads each and allows up to 200 million reads. The detection capacity is in base pair resolution for the detection of deletions, duplications, inversions, and insertions. The major limitations include ambiguous mapping assignments within the repetitive region of the genome, producing inconsistent mapping positions, which may lead to high false positive rates. Based on the fragment size of high throughput sequencing, most of the detected SVs are <1kb in length, and the majority of such SVs are deletions (1000 Genomes Project, 2010; Kidd et al., 2010). The first application of paired-end reads was demonstrated using BAC end sequences generated from cancer cell lines (Volik et al., 2003). Later, this approach was applied to detect germline SVs using a fosmid end sequence library (Tuzun et al., 2005). A wide range of computational algorithms exist to compute and detect SVs from paired-end reads (e.g., PEMer, VariationHunter, BreakDancer, MoDIL, MoGUL, HYDRA, Corona and SPANNER) (Mills et al., 2011; 1000 Genomes Project, 2010).

Read-depth Methods:

Read-depth approaches utilize the mapping distribution of short read sequences from high throughput sequencing in comparison to the reference genome. The read depth is computed for a given region by counting the number of reads that map to that region. Theoretically, a unique region of the genome consists of uniform depth; regions with duplications and deletions show deviated read depths (Alkan et al., 2009). To identify SD blocks, it is crucial to know all of the possible mapping positions for each short read. Unlike unique regions, short reads that belong to SD blocks will return multiple mapping positions. Not all computational aligner algorithms return all mapping positions for short reads in the reference genome. The Micro-read Fast Alignment Search Tool (mrFAST) and the Micro-read Substitution-only Fast Alignment Search Tool (mrsFAST) are the most efficient tools that return all mapping positions for short reads (Alkan et al., 2009; Hach et al., 2010). Using the entire reference genome will return spurious mapping calls within the repeat-rich region of the genome; hence, the repeat masked genome should be used to reduce noise during mapping. Prior to computing the read depth, GC corrections need to be employed to eliminate the GC bias of the sequencing technology. For each user-defined window length, the read depth is computed on GC-corrected data to reduce false positives.

The read-depth-based method is particularly efficient in terms of estimating absolute copy number and detecting segmental duplications in base pair resolution. The detection resolution is positively correlated with the genome coverage, i.e., higher coverage reduces false positive results. The primary benefit of utilizing read depth is the detection of absolute copy number, which is crucial for association analysis. With adequate coverage, it is possible to detect genes that are high in copy number (i.e., DUX4) (Alkan et al., 2009). This is a major advantage compared to microarray-based approaches, which are not capable of computing absolute copy numbers. Read-depth approach algorithms also efficiently detect duplication in a genome with high precision. Moreover, the read-depth approach can identify absolute copy numbers that are embedded or overlapped with segmental duplications. Recently developed tools utilizing read depth methods provide high resolution breakpoints in reference to the discovery of smaller deletions and duplications (Sebat et al., 2004; Yoon et al., 2009).

Read-depth approaches using NGS data were first applied to define rearrangements in cancer and segmental duplication and absolute copy-number maps in human genomes (Campbell et al., 2008; Chiang et al., 2009). Recently, 1000 Genomes data for 159 individuals with low coverage was analyzed for three major populations, and absolute gene CNV was inferred to catalogue their frequencies (Sudmant et al., 2010). The major drawback of this method is the inability to detect inversions and insertions. Unlike paired-end reads, read-depth-based approaches are able to detect only forward-oriented segmental duplications and CNVs, which provides an incomplete catalogue of SVs in the genome.

de novo Assembly:

The microarray and sequence-based approaches described earlier rely on the reference genome for mapping information. de novo assembly eliminates the requirement of a reference genome to construct the genome assembly. In practice, the de novo assembly approach is based purely on computational algorithms that are in their infancy. The approach first assembles the genome, comparing the read sequences, and produces large contigs that are then compared to a reference genome. Although it is a complex task to identify SVs from de novo assembly, this method has the potential to identify thousands of novel variants. Recently, a substantial number of tools have become available to assemble genomes without a reference; this includes EULER-USR, ABySS, SOAPdenovo, Cortex assembler, NovelSeq framework, and ALLPATHS-LG (Chaisson et al., 2009; Simpson et al., 2009; Li et al., 2009; Mills et al., 2011; Gnerre et al., 2011; Hajirasouliha et al., 2010). de novo variant assembly can be done by obtaining different degrees of information from a reference. The Cortex assembler has the ability to assemble multiple genomes and call SVs simultaneously between samples without reference genome information.

The current algorithms are not yet mature enough to provide a full catalogue of the structural variation of the human genome. Thus, these approaches produce an assembly that is 16.2% shorter than the reference genome (including coding regions). A large portion of the missing genome consists of common repeats, and almost the entire content of segmental duplication is missing within these assemblies (Alkan et al., 2011). Although current technologies have substantially improved data production capabilities, accurate genome assembly and correct annotation methodologies remain under development. The common repeat regions and segmental duplication regions remain big challenges in reference to assembly methods.

1.4 NEUROPSYCHIATRIC DISEASES AND COPY NUMBER VARIATIONS

Developmental neurocognitive disorders (DNDs) involve a variety of signs and symptoms, including a range of cognitive impairments from learning disabilities to mental retardation to abnormal behaviors. Several DNDs are now known to be caused by recurrent and non-recurrent genomic rearrangements that are mediated or stimulated by complex regional genomic architectures occurring throughout the human genome. Genetically, the picture is complicated by significant inter-individual heterogeneity, numerous contributing loci, and multiple gene-gene and gene-environment interactions (Persico et al., 2006). Different combinations of the inherited genes might explain the variations in severity or the manifestation of complex neuropsychiatric diseases among family members. However, genetic and phenotypic heterogeneity has made it difficult to detect the DND-associated genetic mechanism(s). Many of these genomic disorders are due to altered gene dosages, or CNV, of one or more dosage-sensitive genes within the rearranged region.

Although DND compose a group of relatively common disorders, a cause is unknown in as many as 60% to 70% of patients with them. Genetically, the picture is complex because of significant inter-individual heterogeneity, multiple genes, and gene-environment interactions (Persico et al., 2006). The presence of different combinations of the inherited genes might explain the variations in severity or the manifestation of DND in siblings and family members. However, phenotypic and genetic heterogeneity has made it difficult to detect the genetic mechanism(s) underpinning DND. Several new genomic disorders caused by CNVs of genes whose dosage is critical for the physiological function of the nervous system have been recently identified. Patients who carry the dup(7)(q11.23) reciprocal duplication of the genomic region deleted in Williams-Beuren syndrome are characterized by prominent speech delays. The phenotypes of Potocki-Lupski syndrome and MECP2 duplication syndrome have been neuropsychologically examined in detail; this has revealed that autism is an endophenotype and a prominent behavioral feature of these disorders. The recent identification of the SHANK gene family (SHANK1, 2, and 3) has revealed extreme variability in copy number for autism spectrum disorder (ASD) (Durand et al., 2006; Berkel et al., 2010; Sato et al., 2012).

Genome-wide studies of large cohorts of patients with mental retardation (MR) have led to the identification of chromosomal aberrations that delineate novel syndromes, some of which occur at frequencies comparable to known recurrent rearrangement disorders. Most notably, three groups (Shaw-Smith et al., 2006; Sharp et al., 2006; Koolen et al., 2006) independently characterized a microdeletion syndrome of 17q21.31 that is associated with MR that arises on parental chromosomes that carry a common inversion of the region. On chromosome 15, three novel syndromes associated with mental retardation have been confirmed at 15q13.3, 15q24, and 15q26.2 (Poot et al., 2007; Sharp et al., 2007; Sharp et al., 2008). Therefore, it is evident that the gene copy number of the genome disrupts neurogenic pathways which manifest in phenotypes with abnormal cognitive function (Lee et al., 2006). In this thesis, we have examined a neuropsychiatric cohort of individuals with Tourette's Syndrome (TS).

Tourette's Syndrome is a developmental neuropsychiatric disorder characterized by the presence of both motor and vocal tics. TS is often accompanied by features associated with obsessive compulsive disorder (OCD), attention deficit hyperactivity disorder (ADHD), and poor impulse control. The co-morbidity of these diseases is complex and the relationship is not clear. The prevalence of Tourette's Syndrome is approximately between 0.3-1% in a population. Although it is very common in a given population, to date, there has not been a single locus identified that segregates within Tourette's-affected individuals in families (Centers for Disease Control and

Prevention, 2009). According to twin studies, the concordance rates are between 50-70% for monozygotic twins compared to 10-23% for dizygotic twins (Price et al., 1985; Walkup et al., 1988; Pauls et al., 1991), and the range is dependent on whether TS manifests alone or with other features, i.e., OCD or ADHD. Similar to other neuropsychiatric disorders (i.e., autism spectrum disorder), TS is a male-dominant disorder (approximately 4:1) (Robertson, 2008). However, studies demonstrate that, within family members with TS, if one includes OCD as the affected status, the risk to male and female relatives of a TS proband approaches 1:1, with female relatives more likely than male relatives to display obsessions and compulsions. The mode of inheritance was initially thought to be autosomal-dominant, but researchers have been unable to detect a TS locus by using the techniques for mapping Mendelian disorders (Pauls et al., 1986; Eapen et al., 1993; State, 2010; State, 2011). Further segregation analysis implicated TS as a complex neuropsychiatric disease (State MW, 2010).

There have been numerous attempts to decipher the genetic cause of TS by linkage and association studies, although no strong common candidate genes have been identified. One candidate locus shows the presence of a de novo chromosome 13 inversion in a TS family. Further analysis revealed that the SLIT and TRK-like family member 1 (SLITRK1) gene that resides in close proximity to the inversions carries a frameshift deletion as well as two independent occurrences of the identical variant in the binding site for microRNA hsa-miR-189 (O'Roak et al., 2010). This gene remains elusive due to the fact that subsequent analysis of different studies found no such mutation. The SLITRK1 gene carries a rare mutation; hence, collecting a large sample size to obtain proper frequency distribution might be an issue to replicate the mutation. Another report found that CNV breakpoints within the NRNX1 gene and a breakpoint within 1q21 locus had previously been implicated in ASD (Senthil et al., 2010). This association is preliminary and requires replication due to the small sample size. The lack of strong evidence of a causative variant in TS provides an opportunity to search for risk-susceptible CNVs.

1.5 COMMON COMPLEX DISEASES AND COPY NUMBER VARIATIONS

Most studies investigating the genetics underpinning complex diseases have focused primarily on SNP-based approaches. They have found that these autoimmune diseases exhibit strong heritability, but linkage and genome-wide SNP-based association studies explain little of the genetic variation (i.e., usually 1–15%), strongly suggesting that other genetic factors are at play (Stranger et al., 2011). For example, genome-wide SNP-based association studies have identified more than 71 genes related to Crohn's disease, which explains only 23.2% of the total heritability (Franke et al., 2010). Hence, SNP-based association analysis identifies only a fraction of the entire genetic burden for complex diseases, leaving a large portion of heritability unresolved. Research into complex diseases is yet to resolve the total genetic contribution involved and it suffers from the problem of missing heritability. Recently, CNV is gaining popularity because of its high frequency in the genome and because of substantial technological developments in terms of the detection of CNVs.

The functional consequences of a CNV event which highlights the potential to significantly alter the expression of a gene shows the importance of CNV studies in complex diseases. The largest CNV association analysis was conducted by the Wellcome Trust Case Control Consortium (WTCCC, 2010). This study analyzed primarily common CNVs, a majority of which can be tagged with a nearby SNP. The main conclusion of that study states that common CNVs that can be detected with existing technology do not contribute to common complex disease susceptibility. Although the negative conclusion is very discouraging, there is a growing list of independent studies that show that CNVs are associated with common complex diseases. A list of CNVs (genic/agenic) that reports associations with complex diseases is given in Table 1.

Table 1.1 A list of CNV loci that is associated with complex diseases.

Disease	**Cytoband (Gene)**	**CNV Status**	**Reference**
Rheumatoid Arthritis (RA)	17q12 (*CCL3L1*)	≥2 copies	(McKinney et al. 2008)
	1q21.3 (*LCE3B, LCE3C*)	Deletion.	(Lu et al. 2011)
Psoriasis (PsV)	8p23.1 (beta-defensin *DEFB* gene family)	≥5 copies	(Hollox et al. 2008)
	1q21.3 (*LCE3B, LCE3C*)	Deletion.	(Cid et al. 2009)
Psoriatic Arthritis	1q21.3 (*LCE3B, LCE3C)*	Deletion.	(Bowes et al. 2010)
Inflammatory Bowel Disease (IBD)	8p23.1 (beta-defensin 2 *DEFB2)*	≤ 2 copies	(Fellermann et al. 2006)
Systemic Lupus Erithmatosis (SLE)	1q23.3 (*FCGR3B)*	<2 copies	(Willcocks et al. 2008)
	17q12 (*CCL3L1*)	≥4 copies	(Mamtani et al. 2008)
Crohn	5q33.1(upstream *IRGM)*	Deletion	(Brest et al. 2011)
Obesity	16p11.2	*Mirror Extreme:* Deletion with Obesity; Duplication with Underweight.	(Jacquemont et al. 2011)
Osteoporosis	*UGT2B17*	Deletion	(Yang et al. 2008)

In this thesis, we have attempted to find novel genetic susceptible loci for Ankylosing Spondylitis (AS), which is the second-most-common cause of inflammatory arthritis worldwide, with a prevalence of 1/1,000–3/1,000 in white populations (Calin, 1998). It is characterized by inflammation in the spine and sacroiliac joints, causing initial bone and joint erosion and subsequent ankylosis.

Arthritis affecting peripheral joints, particularly the hips, occurs in 40% of cases and inflammation may also involve extraarticular sites such as the uvea, tendon insertions, aorta, lungs, and kidneys. Long ago, genetic factors were implicated in the etiology of the disease, with the demonstration of a high degree of disease manifestation within families (de Blecourt et al., 1961). The sibling recurrence-risk ratio is 82 (Brown et al., 2000b) and heritability, assessed by twin studies, is >90% (Brown et al., 1997). The recognition of the association of HLA-B27 with AS confirmed the importance of heritable factors in the disease (Brewerton et al., 1973; Schlosstein et al., 1973) and remains one of the strongest disease associations of any inflammatory human disease.

In most populations that have been studied, the prevalence of AS is strongly correlated with the prevalence of the main disease-susceptibility gene, HLA-B27 (B27). In only a few families, has it been reported that AS segregates independently from B27 (van der Linden et al., 1975; Gladman et al., 1986; Deshayes et al., 1987; Woodrow, 1988; Brown et al., 1996; Said-Nahal et al., 2000) and only rare cases of familial B27-negative AS have been reported (Rubin et al., 1994; Skomsvoll et al., 1995), suggesting that B27 is almost essential for the inheritance of AS within families. However, only 1%–5% of B27-positive individuals develop AS, and there is increasing evidence to suggest that other genes must also be involved. B27-positive relatives of AS patients have a recurrence risk of the disease that is 5.6–16 times greater than that of B27-positive individuals in the population at large, implying the presence of non-B27-shared familial risk factors (Calin et al., 1983; van der Linden et al., 1983).

Recurrence-risk modeling in AS rejects single-gene and polygenic models. Oligogenic models between three and nine genes operating in addition to B27 fit the observed pattern of recurrence risks in relatives of patients with AS (Brown et al., 2000b). A major non-B27 contribution to susceptibility to AS is suggested by the greater concordance rate in monozygotic twins (63%) than in B27-positive dizygotic twin pairs (23%) (Brown et al., 1997).

Ankylosing spondylitis is a type of spondyloarthropathy; the pathogenesis of the disease is very poorly understood. Gene expression profiling approaches have been used to determine the molecular mechanisms and pathways of the disease. Elucidating the potential role of non-HLA genes is still being done. Using powerful genome-wide-association study approaches, a growing list of non-HLA genes or intergenic loci with varying effect sizes (*IL23R, RUNX3, KIF21B, 2p15, IL1R2, PTGER4, ERAP1, IL12B, CARD9, TNFR1/ LTBR, TBKBP1*, and 21q22) have been reported, most of which have been subsequently validated (5). The associations were primarily reported by investigating single-nucleotide polymorphism (SNPs) analysis. Recently, apart from single-allele associations, one study revealed the complex interactions between the non-HLA gene *ERAP1* with HLA-B27 (Evans et al., 2011). However, the genetic risk described by HLA or non-HLA genes suggests that other genomic variants might contribute to risk for AS. To date, no CNV has been reported to be associated with AS; this provides a unique opportunity to investigate the contribution of CNVs to the disease pathogenesis of AS.

1.6 RESEARCH HYPOTHESIS AND SPECIFIC OBJECTIVES

In the last decade, SNP-based association studies have successfully identified loci that are associated with complex diseases. These studies are well designed for identifying common variants but are underpowered to detect rare variants. Although the search for common alleles was the focus during the last decade, there has, nonetheless, been a steady parallel effort to evaluate the contribution of rare variants. These have included cytogenetics, parametric linkages in individual pedigrees or isolated populations, targeted sequencing, and analysis of copy number variation. The contribution of copy number variation to most complex diseases is yet to be elucidated. Detection resolution is a key aspect that always plays a big role in the ability to detect disease-susceptible CNVs. There are significant differences in reference to the detection of CNVs in the current technologies. As technological advancements have been made, no single method yet has the capacity to detect the entire content of structural variations in the genome. In light of this detection complexity, identification of CNVs that are susceptible to diseases requires new strategies to evaluate their genetic contribution.

The specific objectives of this thesis are fourfold:

1. To characterize the human genome regions that are susceptible to rearrangements that produce copy number variations and are mediated through the presence of segmental duplications.

2. To design a custom microarray that targets the rearrangement hotspots that have been identified.

3. To identify copy number variation (using the custom-designed microarray) that contributes to the genetic susceptibility of Tourette's Syndrome and Ankylosing Spondylitis.

4. To quantify the detection sensitivity of copy number variations using whole genome SNP-microarray.

Chapter 2

Genome-wide Signatures of 'Rearrangement Hotspots' within Segmental Duplications in Humans

Mohammed Uddin, Mitch Sturge, Lynette Peddle, Darren D O'Rielly, Proton Rahman

Discipline of Genetics, Faculty of Medicine, Memorial University of Newfoundland, St. John's, Newfoundland, Canada.

PREFACE

The primary contribution of this manuscript is the detection rearrangement hotspots using high throughput sequencing technology. The first phase of the analyses includes read depth computation to detect SD and CNV. The next phase of the analysis includes the detection of rearrangement hotspots and characterization of complex genomic regions. A version of this manuscript has already been published in *PLoS ONE* Journal.

All co-authors worked as a team in preparation of the manuscript. The primary author, Mohammed Uddin, conceptualized the problem and with proper guidance from Dr. Proton Rahman, developed the framework of the study design. Mohammed Uddin conducted extensive literature review, and developed the computational algorithms. The application of the detected hotspots in this manuscript is highlighted in chapter 3 and 5.

ABSTRACT

The primary objective of this study was to create a genome-wide high resolution map (i.e., .100 bp) of 'rearrangement hotspots' which can facilitate the identification of regions capable of mediating de novo deletions or duplications in humans. A hierarchical method was employed to fragment segmental duplications (SDs) into multiple smaller SD units. Combining an end space free pairwise alignment algorithm with a 'seed and extend' approach, we have exhaustively searched 409 million alignments to detect complex structural rearrangements within the reference-guided assembly of the NA18507 human genome (18x coverage), including the previously identified novel 4.8 MBp sequence from de novo assembly within this genome. We have identified 1,963 rearrangement hotspots within SDs which encompass 166 genes and display an enrichment of duplicated gene nucleotide variants (DNVs). These regions are correlated with increased nonallelic homologous recombination (NAHR) event frequency which presumably represents the origin of copy number variations (CNVs) and pathogenic duplications/deletions. Analysis revealed that 20% of the detected hotspots are clustered within the proximal and distal SD breakpoints flanked by the pathogenic deletions/duplications that have been mapped for 24 NAHR-mediated genomic disorders. FISH Validation of selected complex regions revealed 94% concordance with in silico localization of the highly homologous derivatives. Other results from this study indicate that intra-chromosomal recombination is enhanced in genic compared with agenic

duplicated regions, and that gene desert regions comprising SDs may represent reservoirs for creation of novel genes. The generation of genome-wide signatures of 'rearrangement hotspots', which likely serve as templates for NAHR, may provide a powerful approach towards understanding the underlying mutational mechanism(s) for development of constitutional and acquired diseases.

2.1 INTRODUCTION

Segmental duplications (SDs) or low-copy repeats are blocks of DNA.1 kbp in size which share a high level of sequence homology (.90%) (Bailey et al. 2006; Redon et al. 2006; Bailey et al. 2002; Alkan et al. 2011). The catalogue of SDs comprises approximately 5% of the human genome encompassing 18% of genes (Bailey et al. 2006; Redon et al. 2006; Bailey et al. 2002; Alkan et al. 2011).. They are considered antecedents to the formation of copy number variants (CNVs) which comprise approximately 12% of the human genome and are responsible for considerable human genetic variation (Redon et al. 2006; Sharp et al. 2005). Emerging evidence suggests that SD regions are frequently associated with known genomic disorders with the vast majority representing novel sites whose genomic architecture is susceptible to disease-causing rearrangements (Sharp et al. 2005). However, the complexity of their structural architecture in the human genome and, more importantly, their role in disease pathogenesis remains largely elusive.

There is a growing body of evidence suggesting the involvement of multiple events in the origin of genomic rearrangements such as non-allelic homologous recombination (NAHR), non-homologous end joining (NHEJ), fork stalling and template switching (FoSTeS), and microhomology-mediated break-induced replication (MMBIR) (Gu et al. 2008; Lieber et al. 2003; Zhang et al. 2008). Although the origins of the aforementioned mechanisms are strongly associated with highly homologous regions residing outside of common repeat elements (e.g., transposons) (Conrad et al. 2010a), the non-random distribution of highly homologous regions within SDs that are susceptible to such mechanisms remain to be fully elucidated. Moreover, evolutionary conservation of these mechanisms complicates the identification of SD breakpoints due to differing levels of sequence homology.

Genomic disorders arising from microdeletions/duplications fail to be adequately explained by a single underlying event. The true contribution of NAHR, NEHJ, MMBIR and FoSTeS events to the origin of genomic rearrangement remains elusive, although large-scale studies are beginning to implicate NAHR as one of the primary events contributing to the origin of these genomic copy number changes (Conrad et al. 2010a; Mills et al. 2011). Genomic DNA situated between distal and proximal SDs represents a critical region often reported to be deleted/duplicated due to misalignment of the SDs between homologous chromosomes (Shaikh et al. 2007). Evidence suggests that the breakpoint architecture of SDs (i.e., distal and proximal) is associated with a higher propensity for NAHR-mediated rearrangement predisposing to an abnormal phenotype (Firth et al. 2009). In other words, the increased frequency of pathogenic rearrangements is often directly correlated with the structural complexity of the local genomic regions involved. This is consistent with numerous reports indicating that highly homologous regions within SDs influence NAHR-mediated rearrangement events (Conrad et al. 2010a; Mills et al. 2011; Turner et al. 2007). Throughout this paper, these highly homologous regions will be referred to as 'rearrangement hotspots'. Classic examples of NAHR-mediated genomic rearrangement include genomic disorders such as 3q29 microdeletion/duplication syndrome, globozoospermia, and Williams-Beuren syndrome (Ballif et al. 2008; Koscinski et al. 2011; Bayes et al. 2003).

In a recent report, the detection and validation of 8,599 CNVs using microarrays (Conrad et al. 2010a) and subsequent targeted sequencing on 1067 of these CNV breakpoints (Conrad et al. 2010b) revealed extreme homologous regions consistent with NAHR-mediated rearrangements as the primary event in the origin of CNVs. In this study, we identified genome-wide 'rearrangement hotspots' within SD regions that often predispose to genomic disorders in humans, mediated predominately by NAHR. We specifically devised a hierarchical approach to detect SD units using an all-hit mapping algorithm, interrogating every 100 bp (GC-corrected read depth window with a 1 bp overlap) excluding common repeat elements. Reference-guided assembly was obtained from reads based on the NA18507 human genome and duplicated sequences were extracted from the assembly using detected breakpoints (Figure. 2.1). The primary objective of this study was to create genome-wide signatures of 'rearrangement hotspots' which can facilitate the detection of genomic regions capable of mediating de novo deletions or duplications in humans. To create a genome-wide high resolution map of 'rearrangement hotspots', we developed an end-space free pairwise alignment algorithm integrating a 'seed and extend' technique which can accurately investigate the complex structural architecture associated with the technically challenging and problematic nature of segmental duplications. The hypothesis of this study is that highly homologous SD regions (i.e., rearrangement hotspots) predispose to genomic rearrangements arising from recombination and replication-based events.

2.2 RESULTS AND DISCUSSION

Detection of Segmental Duplication (SD) Units

Given that SDs intuitively consist of common repeat elements, SDs were fragmented into multiple smaller SD units which did not overlap with known repeat elements during the read depth-based analysis. In this study, 20,237 non-redundant sets of SD units with at least one inter- or intra-chromosomal rearrangement event were identified, representing 16.65 Mbp of SD units residing outside of common repeat elements in the human genome (supplementary 1, chapter 2). At first glance, this total content of SDs may appear small compared with that previously reported (Bailey et al. 2002) and that reported in the database of genomic variants (DGV) which is mainly attributed to methodological differences (i.e., exclusion of common repeats, GC-correction, shorter window length, low read depth threshold). Results from this study and Perry et al (Perry et al. 2008), suggest that previously reported SD breakpoints are overinflated in size, further emphasizing the importance of creating a high-resolution map of'rearrangement hotspots'. Read depth distribution for duplicated and non-duplicated regions throughout the genome produced a distinctive distribution pattern with an approximate 7% error rate (supplementary 1, chapter 2).

Considering CNVs have a tendency to overlap with nearby SD breakpoints, the results of this study were compared with a recent study which identified common CNV breakpoints in three populations (i.e., 57 Yoruba, 48 European and 54 Asian individuals) (Sudmant et al. 2010). The detected autosomal SD units greater than 200 bp shared 82% concordance (i.e., .50% overlap) with common CNV breakpoints using low coverage short-read data (supplementary 1, chapter 2). Moreover, 79% of breakpoints residing within genes with .3 copies as previously reported (Alkan et al. 2009), were located within SD breakpoints identified in this study (supplementary 1 chapter 2).

Comparison with previous read depth-based reports highlights the advantages of our hierarchical strategy which include: 1) the use of a 100 bp read depth window with a 1 bp overlap to detect SD units which enabled the capacity to detect SD units with higher resolution; 2) the use of a lower threshold (i.e., mean +2 standard deviations) than previously reported methods in order to detect homozygous and hemizygous duplications; 3) fragmentation of SDs into smaller SD units in order to separate duplicated regions from common repeated elements while reducing alignment bias for rearrangement analysis and computational time; and 4) integration of end space alignment algorithm with a 'seed and extend' clustering technique to the duplicated region of the reference guided assembly sequences to perform an exhaustive search (i.e., 409 million alignments) to identify rearrangement breakpoints (supplementary 2, chapter 2).

Compared with copy number gains identified using microarray analysis, sequencing data used in this study revealed that autosomal SD unit breakpoints overlapped 54% with copy number gains (Conrad et al. 2010b), which increased to 67% when compared with 436 coverage (supplementary 1, chapter 2) (Sudmant et al. 2010). Discrepancies are attributed to methodical biases, as detection of structural variants can be specific to different methodical approaches and discrepancies between methods can be as high as 80% (Alkan et al. 2011). The rearrangement analysis within the novel sequence revealed multiple hits within the duplicated sequences (i.e., .90% similarity) that were previously uncharacterized (supplementary 2, chapter 2.).

Characterization of Rearrangement Hotspots Within Segmental Duplications

Using 409 million pairwise alignments, we identified 1963 complex SD units or 'rearrangement hotspots' within SDs in the human genome with significantly high distribution of duplicons (p < 1.0 X 10^{-6}) with at least 10 duplicons per SD unit (Figure. 2.2a). Within these regions, an increase in copy number gain (i.e., increase of 62% in copy number gains within hotspots) with at least 50% overlap with SD units and CNV breakpoints has been observed compared with a previous report (Conrad et al. 2010b). Importantly, 25% of these 'rearrangement hotspots' (i.e., 489/1963) overlapped with 166 unique genes (Figure. 2.2b) of which 77% (i.e., 375/489) were contained within 82 genes with increased copy number gain that have been previously validated using microarray analysis (Conrad et al. 2010b). That 25 of these genes are highly variable in copy number within three populations indicates population-specific frequency of the underlying events in the origin of CNVs (Sudmant et al. 2010) which, in turn, implies an increase in frequency of genomic rearrangement events within hotspot regions. However, the extent of gene conversion within the NAHR hotspot is still unknown. In our analysis, we observed a relative increase of gene content transfer within agenic hotspot regions (i.e., approximately 50%) compared with the remainder of agenic non-hotspot duplicated regions (i.e., 32%) (Figure. 2.2c). The finding of elevated levels of gene content transfer is consistent with a previous study which hypothesized such a finding as an apparent feature for hotspots arising from homologous recombination (Gu et al. 2008). Further analysis on duplicated gene variants (DNVs), which is a special type of paralogous sequence variant, was compared between the hotspot and non-hotspot duplicated regions (Ho et al. 2010). We observed a 3-fold increase in DNVs located within hotspots compared with the remainder of the duplicated regions ($p < 0.0001$) which implies greater diversity within hotspot regions. This finding is attributed, in part, to the accumulation of DNV-derived mutations among derivative homologous sequences within hotspot regions. We also observed a strong positive correlation ($R^2 = 0.63$) between the length and the incidence of DNVs within hotspot regions (**Fig. 2.2d**). Genome-wide read depth comparison revealed that a subset of high read depth regions is positively correlated with rearrangement hotspots (**Fig. 2.2e**).

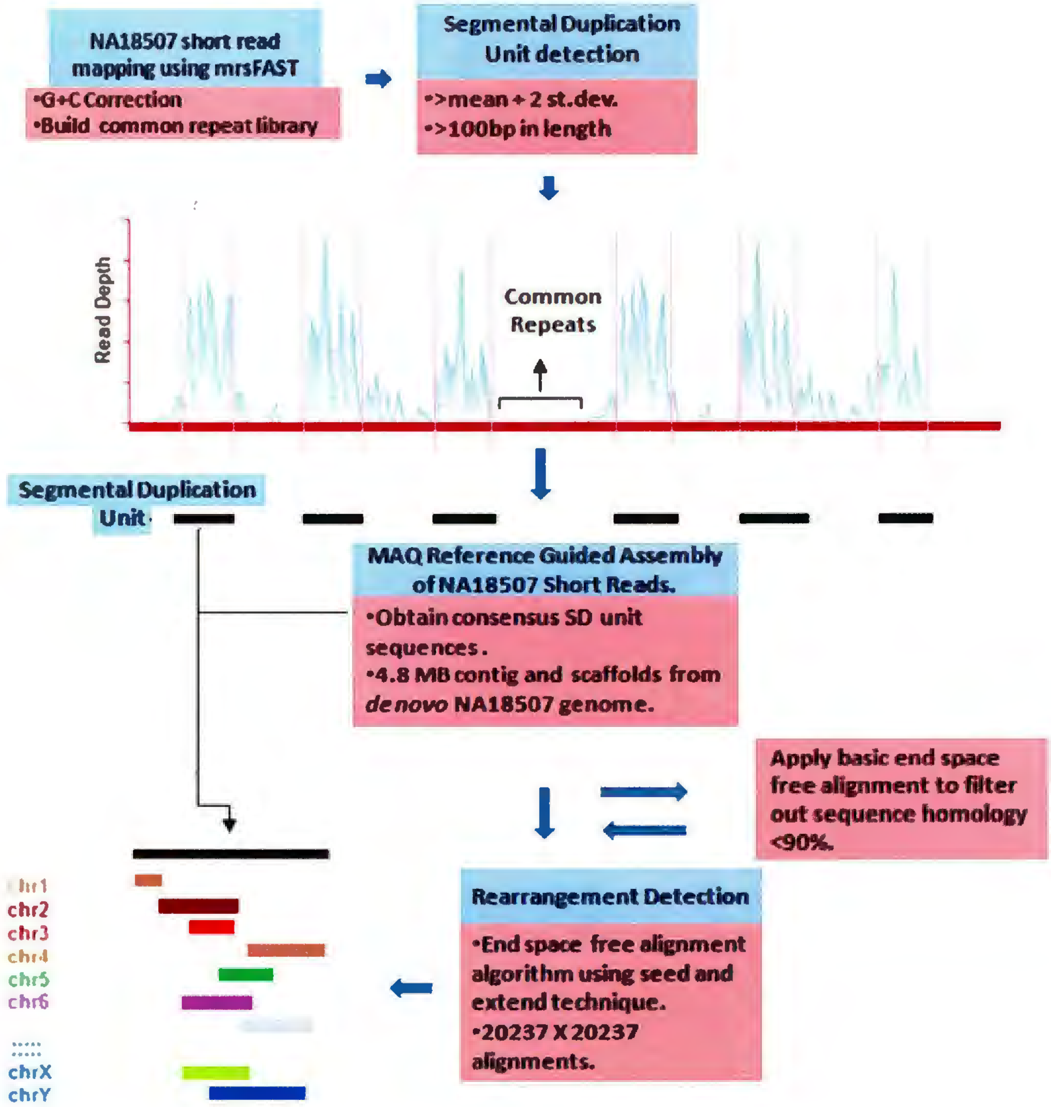

NA18507 short read mapping using mrsFAST
•G+C Correction
•Build common repeat library
Segmental Duplication Unit detection
•>mean + 2 st.dev.
•>100bp in length
Read Depth
Common Repeats
Segmental Duplication Unit
MAQ Reference Guided Assembly of NA18507 Short Reads.
•Obtain consensus SD unit sequences.
•4.8 MB contig and scaffolds from *de novo* NA18507 genome.
Apply basic end space free alignment to filter out sequence homology <90%.
Rearrangement Detection
•End space free alignment algorithm using seed and extend technique.
•20237 X 20237 alignments.
chr1
chr2
chr3
chr4
chr5
chr6
:::::
chrX
chrY

Figure 2.1 A schematic illustrating our hierarchical approach. mrsFAST was used to obtain read depth distribution of the NA18507 human genome with maximum mismatch (n = 2) was allowed against the repeat masked reference human genome (build 36). A mean-based approach was utilized to computationally predict the boundaries of regions associated with excessive read depth. MAQ was used to obtain the consensus genome (mapping quality Q.30 and n = 2) from the NA18507 genome assembly. The consensus

sequence for highly excessive read depth regions was obtained in order to apply a window-based alignment algorithm. The previously identified novel 4.8 MBp sequence from de novo assembly within this genome was also included in the rearrangement analysis. DNVs within hotspot regions (Figure. 2.2d). Genome-wide read depth comparison revealed that a subset of high read depth regions are positively correlated with rearrangement hotspots (Figure. 2.2e).

Distribution of Inter- and Intra-chromosomal Rearrangements

Segmental duplications (SDs) can be categorized according to the location of the rearrangement (supplementary 1, chapter 2) considering that recombination events can occur between homologues (i.e, inter-chromosomal) or by looping out within a single homologue (i.e., intra-chromosomal). Our analysis revealed that 7% of genes (i.e., 1,626/22,159) overlapped with 5,502 nonredundant SD units which represented 73% (i.e., 41/56) of the most highly variable genes previously identified in the human genome within three populations (Sudmant et al. 2010) (Figure. 2.2b). We have identified 91,971 duplicons (i.e., average of 4.5 duplicons per SD unit) with overlapping breakpoints throughout the SD regions. Extreme inter- and intra-chromosomal rearrangements occurred in 10% of genes (i.e., 166/1626) that overlapped with SD units, of which 50% have been previously validated (Conrad et al. 2010b). Further analysis revealed that genic regions were enriched with intra-chromosomal recombination, whereas agenic regions evolved through both inter- and intra-chromosomal recombination (supplementary 1, chapter 2). Such intra-chromosomarecombination within genic SD units may represent conserved genomic organizations subject to gene conversion and concerted evolution (Bailey et al. 2002; Gu et al. 2008; Lieber et al. 2003). Extreme variation, attributed in part, to SDs has been reported in at least 20% of the copy number variable gene families in three human populations (Sudmant et al. 2010).

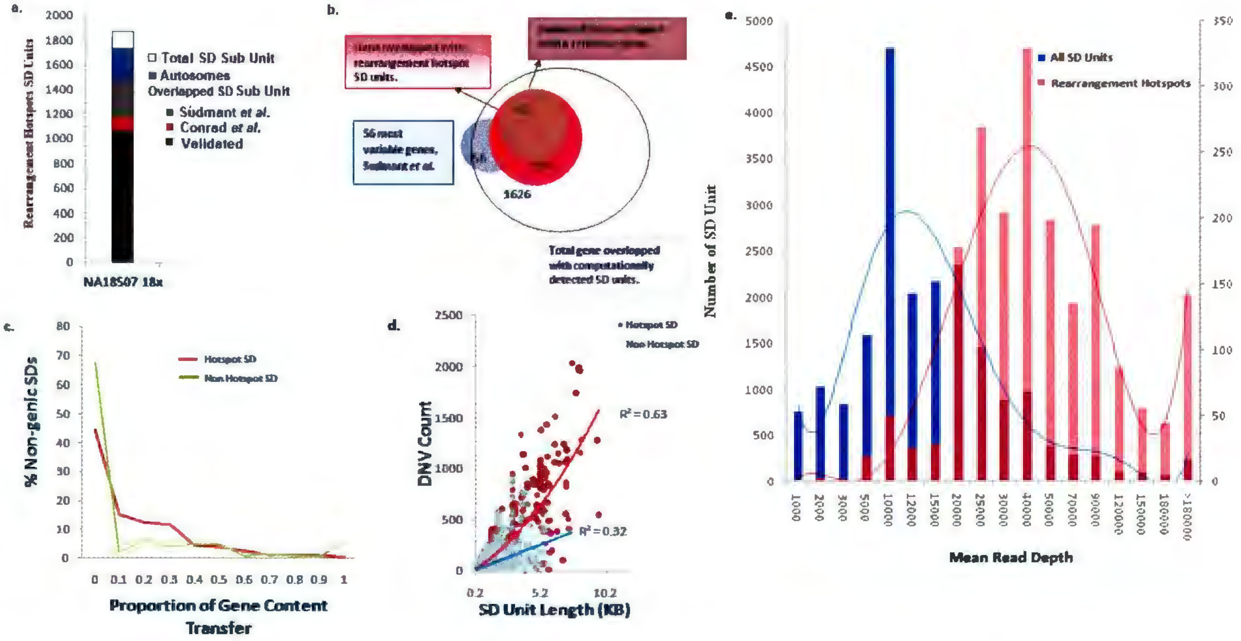


Figure 2.2 Segmental duplication (SD) units which represent the most complex rearrangements within the NA18507 human genome. **a)** A total of 1963 SD complex units (i.e., ≥10 rearrangements) were identified that were significantly different ($p < 1.0 \times 10^{-6}$) compared with the rest of the NA18507 genome duplicated regions. The plot illustrates the concordance of the predicted autosomal complex regions compared with previous studies (Conrad et al. 2010b; Sudmant et al. 2010). **b)** Genes that completely or partially

overlapped with detected SD units in which 73% (41/56) of the most variable genes in three different populations were detected in our analysis of the NA18507 human genome. Among the 1626 genes identified in this study, 10% (i.e., 166/1626) of genes that overlapped with a SD unit revealed extreme inter- and intra-chromosomal rearrangements, 50% of which have been previously validated (Conrad et al. 2010b). **c)** Observed gene content transfer between hotspot and non-hotspot agenic SD units. **d)** scatter plot illustrating DNV count for hotspot and non-hotspot SD units. **e)** A histogram illustrating the mean read depth (RD) of the computationally predicted SD unit breakpoints. The blue bars represent the mean read depth for each of the 20,237 SD unit breakpoints and the red bars represent the mean read depth for hotspot regions.

Previous cytogenetic studies have demonstrated that pericentromeric and subtelomeric SD regions are strikingly polymorphic and both represent hotbeds for genomic rearrangement (Mefford et al. 2002; She et al. 2004). Investigation of recombination within SD units revealed that pericentromeric regions of chromosomes 2, 5, 7, 10, 15, 16, 17, 22 and Y were enriched with inter-chromosomal recombination, whereas only chromosome 11 was associated with intra-chromosomal breakpoints (supplementary 1, chapter 2). Subtelomeric regions of chromosomes 1, 2, 4, 7, 9, 10, 11, 16, 19, 20, 22, and X were enriched with inter-chromosomal recombination, whereas chromosomes 3, 6, 12, 13, 14 and Y were associated with extreme intra-chromosomal breakpoints. This idiosyncratic rearrangement pattern suggests that multiple translocations involving distal regions of chromosomes create complex breakpoints within SDs. This is exemplified by the pseudoautosomal region 1 (PAR1) which displayed extensive inter- and intra-chromosomal tandem duplications, consistent with sex chromosome evolution (supplementary 1, chapter 2). Another complex region where extensive intrachromosomal rearrangements were identified is the distal heterochromatic region of the Y chromosome (i.e., Yq12), housing the male specific (MSY) region (supplementary 1, chapter 2). A comprehensive map of this complex region was generated using PCR analysis in a previous study (Skaletsky et al. 2003). In our analysis, we detected both homozygous and hemizygous duplications using read depth information which represents an extension to previous SD analysis (Sudmant et al. 2010; Alkan et al. 2009) by the inclusion of sex chromosomes (supplementary 1, chapter 2).

An intriguing observation was the identification of complex rearrangements in multiple gene families where rapid evolution of *NBPF, PRAME, RGPD, GAGE, LRRC, TBC1, NPIP* and *TRIM* gene families appear to be predominantly attributed to intrachromosomal gene transfer, whereas other complex gene families (e.g., *ANKRD, OR, GUSB, FAM, POTE, ZNF* and *GOLG*) appear to be more diverse with respect to transfer of gene content, occurring both within and between chromosomes (supplementary 2, chapter 2.). As previously reported (Alkan et al. 2009), the *DUX* family gene was associated with the most copies within the reference genome. The rearrangement analysis of the novel sequence within 10q26.3 region suggests at least 10 additional copies of the *DUX4* gene is specific to novel sequences within the NA18507 human genome. (supplementary 2, chapter 2).

Gene Ontology Analysis within 'Rearrangement Hotspots'

To investigate the impact of genes residing within 'rearrangement hotspot' regions identified in this study and their relation to complex disease, genes were functionally categorized using PANTHER gene ontology analysis (supplementary 1, chapter 2). Genes residing within 'rearrangement hotspot' regions appear to be involved in functions associated primarily with nucleic acid metabolism (22%) and cellular processes (16%), although associations also exist for developmental process (9%), cell cycle (9%), and cell communication (8%). This finding is consistent with a previous report in which copy number gains were associated with genes involved in nucleic acid metabolism and developmental processes, whereas copy number losses were enriched for genes involved in cell adhesion (Park et al. 2010). That genes residing in 'rearrangement hotspot' regions are consistently associated with functions affecting multiple processes important in normal growth and development, further underscores the critical role that rearrangement hotspots play in the genetic etiology of complex disease.

Clinical Relevance of 'Rearrangement Hotspots'

We have produced a genome-wide high resolution map of 'rearrangement hotspots' which likely serve as templates for NAHR and consequently may represent an underlying mechanism for development of constitutional and acquired diseases arising from de novo deletions or duplications. A collection of 24 previously identified genomic disorders predominantly mediated by de novo NAHR events are catalogued in the DECIPHER database (Firth et al. 2009). Comparison of our hotspot regions with pathogenic deletions/duplications breakpoints mapped for those genomic disorders constituting only 15 common genomic loci revealed that 20% of the detected hotspots are clustered within proximal and distal SDs that are flanked by these pathogenic deletions/duplications (Figure. 2.3). This finding indicates a higher rate of NAHR within the genome-wide rearrangement hotspot regions detected in this study.

The rearrangement structure of these hotspots based on our in silico predictions (Figure. 2.4) reveals the complex architecture associated with SDs. To validate the complexity of these hotspots, FISH analysis was performed on selected regions harbouring hotspot clusters demonstrated 94% (i.e., 17/18) concordance with in silico predictions of co-localization (Figure. 2.5a, 2.5b, and 2.6). One example of an identified 'rearrangement hotspot' is a duplication at the 16p12.1 complex region, which contains an S2 inversion (Park et al. 2010), where the alignment localized multiple derivatives of the *NPIPL3* gene within chromosomes 16 and 18 (Figure. 2.5a and supplementary 1, chapter 2). The identified breakpoints revealed the presence of derivative copies of the *NPIPL3* gene within the short arm of chromosomes 16 and 18, possibly attributed to NAHR-mediated recombination, where pathogenic deletions and duplications have been reported in patients with mental retardation and intellectual disability (Antonacci et al. 2010; Nagamani et al. 2011; Heinzen et al. 2010; Weiss et al. 2008; Walters et al. 2010; Ballif et al. 2007; Tokutomi et al. 2009). The derivatives are located within the pathogenic deletion breakpoints among the patients with neurodevelopment disorders. Unfortunately, these studies used methodologies unable to localize derivative copies, and consequently the *NPIPL3* gene was disregarded as a susceptibility gene. A second complex region, 22q11.21, housed a large duplication consisting of two copies, with the 'core duplicon' being copied multiple times in chromosomes 5, 6, 20 and 22 (Figure. 2.5b). Phenotypes attributed to pathogenic deletions and duplications within chromosomes 5 and 22 (Ensenauer et al. 2003; Huang et al. 2010) revealed breakpoint patterns within a 'core duplicon', suggestive of NAHR-mediated duplication.

A third complex region, revealed a previously uncharacterized gene desert within 1q21 indicating a possible harvest region for the NBPF gene family. This 68 Kbp gene desert region revealed extreme intra-chromosomal rearrangement without any signature of inter-chromosomal duplication in our in silico analysis (Figure. 2.6). The gene fragments from

NBPF1,3,9,10,14,15,16,20 and *24* appear to be copied and transferred to 1q21.1 (142867911–142935940) and consequently creating extreme overlapping tandem duplications. The fosmid clone G248P8712C10 covering this region was used on metaphase chromosomes to predict derivative duplicated loci. Multiple signals were obtained within 1p36.12 and 1q21.1 regions, while a weak signal was obtained within the 1p10-p13 region which was not detected by our in silico analysis. The donor region located 2 MBp distal from the gene desert transferred gene content to this 68 kbp region which is associated with recurrent pathogenic deletions and duplications implicated in developmental disorders and neuroblastoma (Firth et al. 2009; Diskin et al. 2009; Brunetti-Pierri et al. 2009). One may speculate that gene deserts may represent reservoirs for creation of novel genes and underscores the necessity to further explore this previously ignored region of the human genome. The complexity of tandem duplications (e.g., 1q21.1) can have a direct impact on estimating copy number for a gene (e.g., *NBPF*). In such cases, the estimation of copy number based solely on read depth may be affected due to the nature of the tandem duplication.

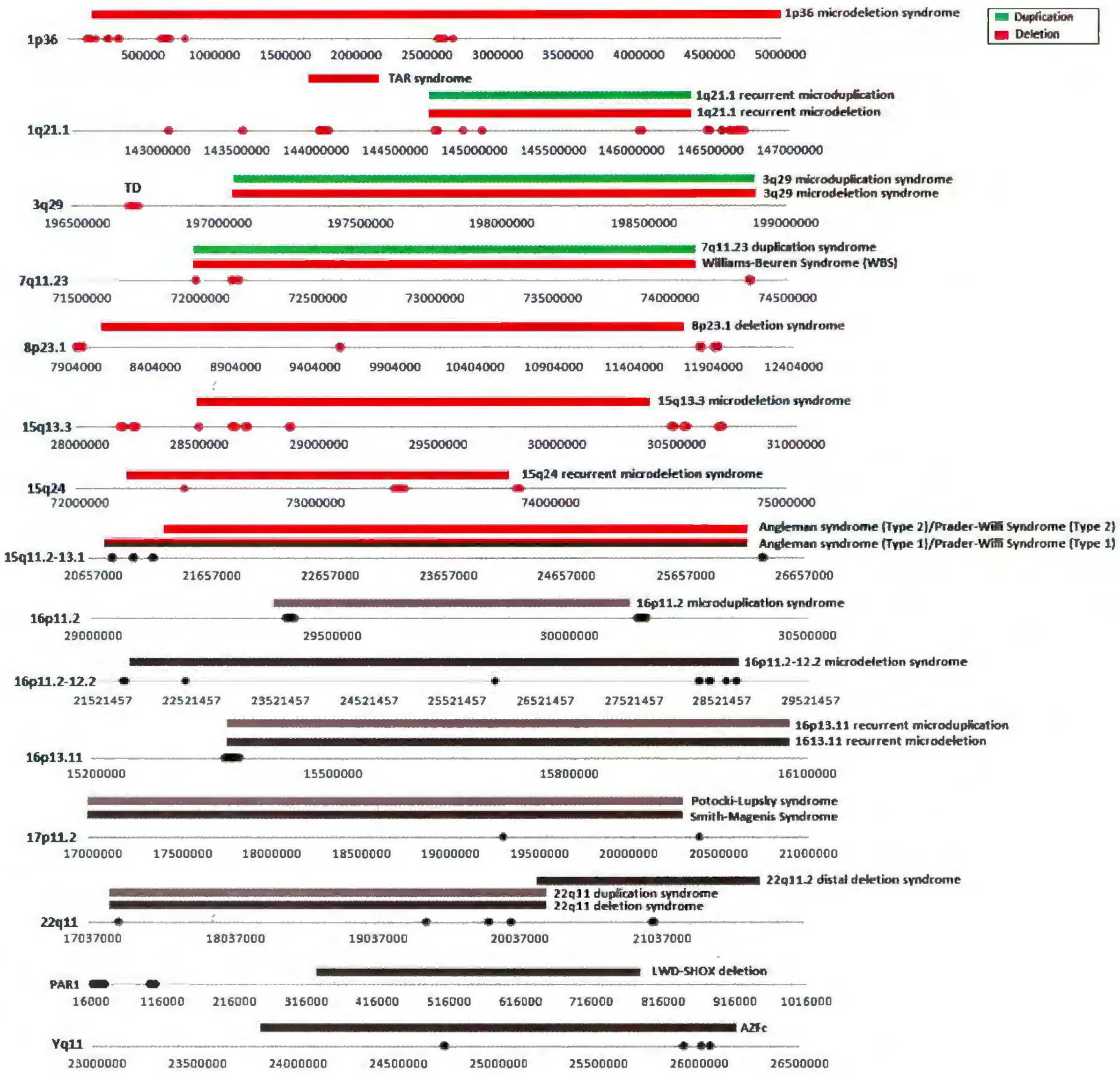


Figure 2.3 The physical position of rearrangement hotspots that has been mapped within the proximal/distal breakpoints of a pathogenic deletion (red horizontal block) or duplication (green horizontal block).

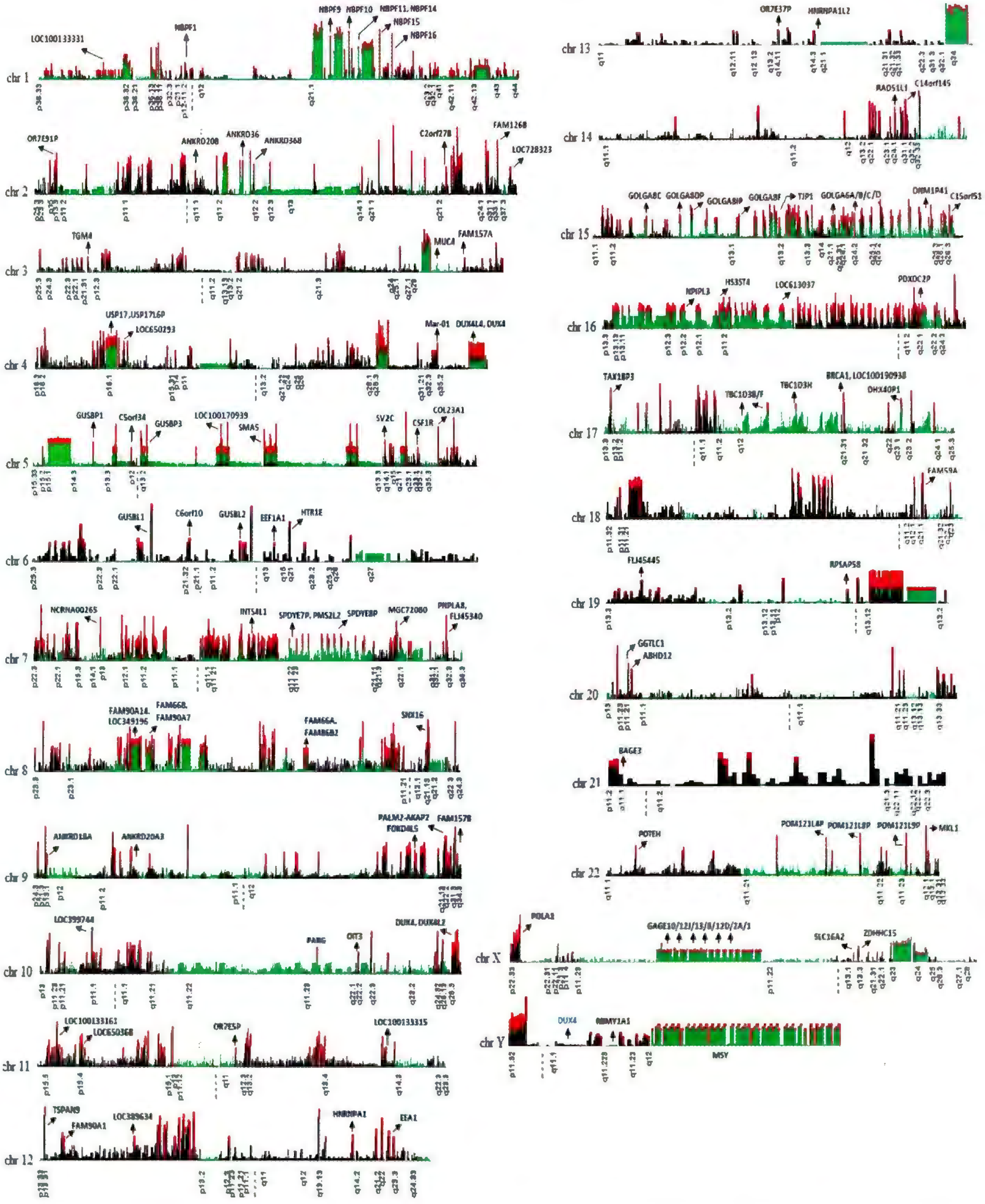


Figure 2.4 Landscape of chromosomal rearrangements in the NA18507 human genome. Chromosomal rearrangements located within duplicated regions are

plotted against the human genome. Green bars representt the signature of intra-chromosomal rearrangements, black bars represent inter-chromosomal rearrangements and red bars represent 'rearrangement hotspots'. Cytobands with duplications for each chromosome and selected genes that completely or partially overlapped with SD units are also indicated.

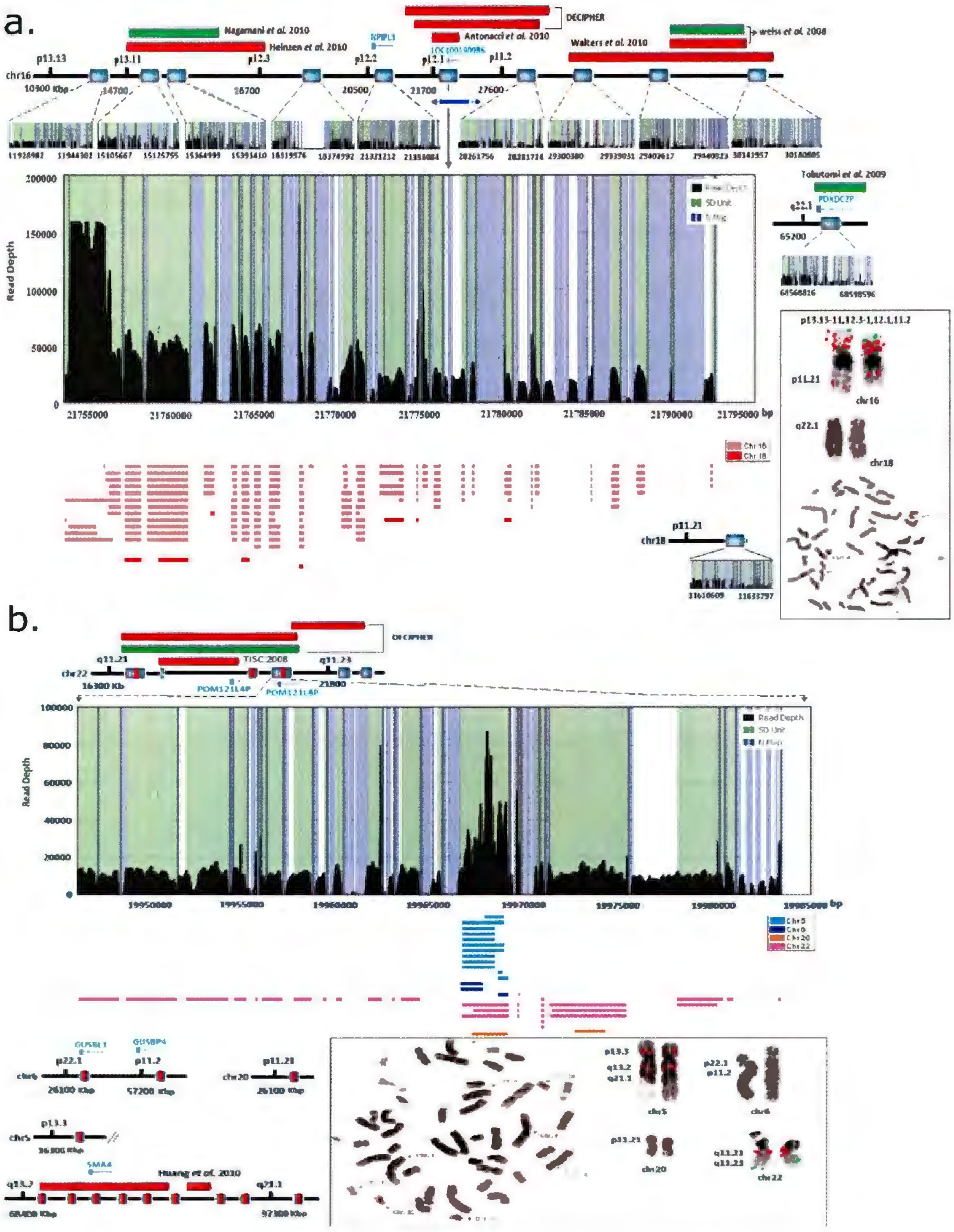

a.
DECIPHER
Nagamani et al. 2010
Heinzen et al. 2010
NPIPL3
Antonacci et al. 2010
LOC100190986
Walters et al. 2010
weiss et al. 2008
p13.13
p13.11
p12.3
p12.2
p12.1
p11.2
chr16
10300 Kbp
14700
16700
20500
21700
27600
11928982
11944302
15105667
15125755
15364999
15393410
18319576
18374992
21321212
21351084
28261756
28281716
29300380
29339031
29402617
29440323
30141957
30180805
200000
150000
100000
50000
0
Read Depth
21755000
21760000
21765000
21770000
21775000
21780000
21785000
21790000
21795000 bp
Read Depth
SD Unit
N Mig
Chr 16
Chr 18
Tokutomi et al. 2009
PDXDC2P
q22.1
65200
68568816
68598596
p13.13-11,12.3-1,12.1,11.2
p11.21
chr16
q22.1
chr18
p11.21
chr18
11610609
11633797
b.
DECIPHER
TISC 2008
q11.21
q11.23
chr22
16300 Kb
21800
POM121L4P
POM121L4P
100000
80000
60000
40000
20000
0
Read Depth
19950000
19955000
19960000
19965000
19970000
19975000
19980000
19985000 bp
Read Depth
SD Unit
N Mig
Chr5
Chr6
Chr20
Chr22
GUSBL1
GUSBP4
p22.1
p11.2
chr6
26100 Kbp
57200 Kbp
p11.21
chr20
26100 Kbp
p13.3
chr5
16300 Kbp
SMA4
Huang et al. 2010
q13.2
q21.1
68400 Kbp
97300 Kbp
p13.3
q13.2
q21.1
chr5
p22.1
p11.2
chr6
p11.21
chr20
q11.21
q11.23
chr22

Figure 2.5 Signature of rearrangement hotspots located at a) 16p12.1 and b) 22q11.21. A 40 KBp region within 16p12.1 is illustrated with its corresponding derivative copies which were localized by hierarchical analysis. This region consists of the *NPIPL3* gene derivatives. The inter- and intra-chromosomal localization of the copies is approximated in the physical map within the chromosome contig (18p11.21).The alignments are color coded for chromosomes (i.e., color coded rectangles below the read depth plot) and FISH validation is illustrated for both inter- and intra-chromosomal localization. The pathogenic deletions and duplications located within these regions (Antonacci et al. 2010; Nagamani et al. 2011; Heinzen et al. 2010; Weiss et al. 2008; Walters et al. 2010; Ballif et al. 2007; Tokutomi et al. 2009) are depicted in red and green bars, respectively The blue bars under the contig represent the approximated inversions previously reported by Antonacci, F. et al (Antonacci et al. 2010). b) Analysis of a 37 KBp duplicated region within 22q11.21 revealed it is comprised of a core 2.7 kbp tandem duplicon copied from different chromosomes. Black lines represent the read depth (x-axis), green shade represent an SD unit, and blue bars represent the region with common repeat elements. The horizontal blocks (color coded according to chromosomes) are the rearrangement (intra/inter) fragments with >90% sequence similarity and >100bp in length.

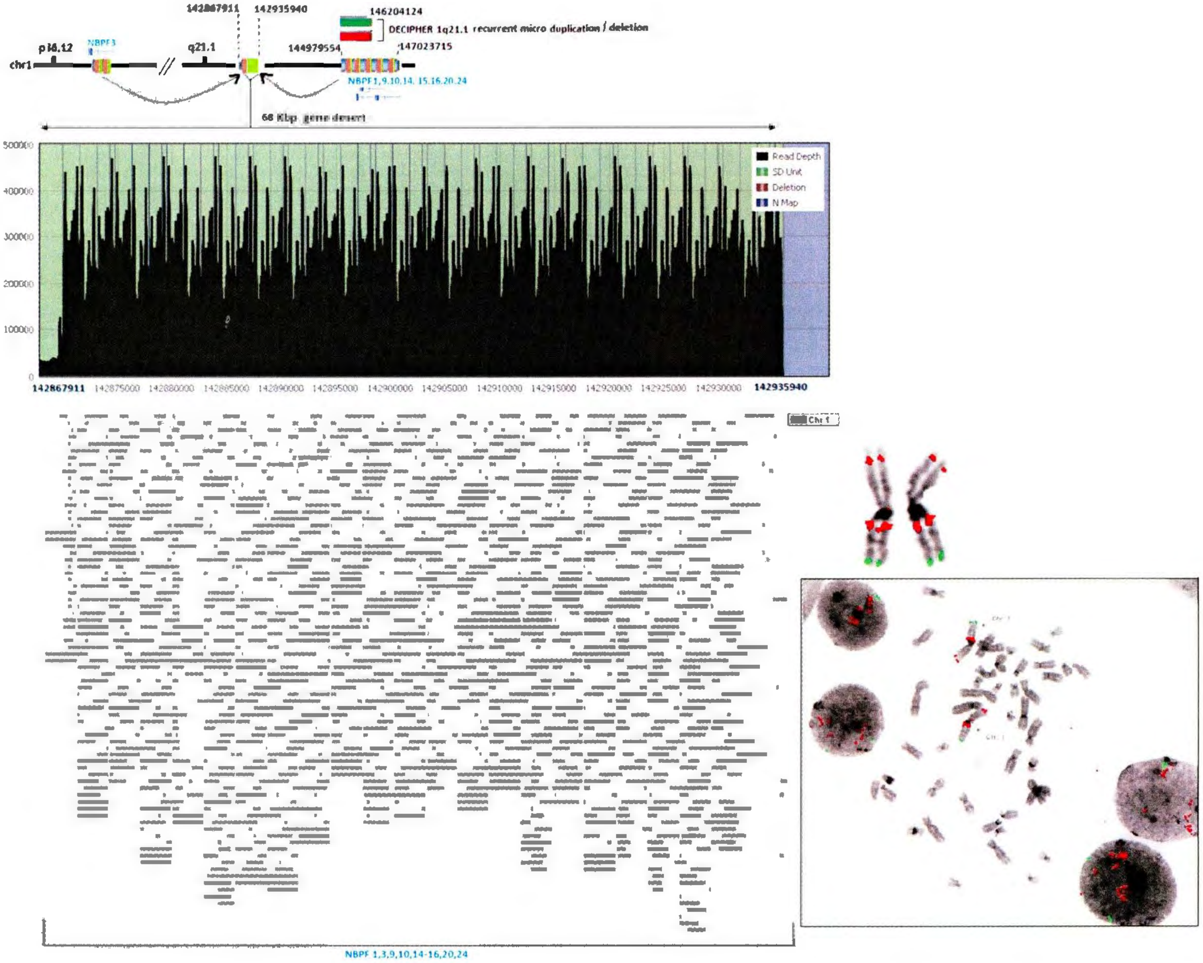


Figure 2.6 Rearrangement hotspots comprising a 68kb gene desert located within 1q21.1 region. Validation of a gene desert where extreme intra-chromosomal rearrangement without any signature of inter-chromosomal duplication observed in our *in silico* predictions. The rearrangement consists of

gene fragments from the *NBPF* gene family located within the p and q arm of chromosome 1.

Limitations

While the results of this study highlight the importance of restricting the number of vulnerable genomic regions that are targeted for clinical application, read depth-based approaches are associated with certain limitations. One of the limitations of our approach was the exclusion of inversions and insertions as the read map algorithm mrsFAST employed in this study was unable to return information regarding the orientation of duplicated loci (Hach et al. 2010) and as a result the map of ‘rearrangement hotspots’ will miss regions with complex orientations. Coverage is another constraint to detect SDs due to the positive correlation between coverage and detection rate (Alkan et al. 2009). A much higher (i.e., .40x) coverage will significantly increase the detection capacity of SD units. It is widely accepted that no single method has the capacity to capture the entire content of structural variants in the genome. For example, read pair and read depth approach overlapped only 20% among the detected variants (Alkan et al. 2011), therefore, a portion of ‘rearrangement hotspots’ will be missed by our analysis. Moreover, a portion of highly duplicated regions (.99% sequence identity) analyzed in this study is reference sequence-specific due to MAQ’s (mapping and assembly with quality) limitation to align short reads within those region precisely (LiHeng, et al. 2008). While the results of this hypothesis-driven in silico study are consistent with limited FISH analysis, additional genome-wide validation is required. In a recent report, it has highlighted the current limitation of de novo assembly approaches that produce a consensus genome with at least 16.2% shorter than the reference genome (Alkan et al. 2011). As de novo assembly progresses with large genome initiatives (i.e., 1000 Genomes Project), integration of comprehensive de novo assembly with our hierarchical approach will afford maximum potential to detect a complete picture of population-specific 'rearrangement hotspots'. Collectively, the results of this study emphasize the complexity of genomic rearrangements and the importance of NAHR-mediated recombination events in the origin of deletions and duplications which underlie the manifestation of germline and somatic disease.

Diseases arising from structural changes in the human genome are strongly correlated with the local sequence structure in which NAHR appears to be the predominant mechanism producing such vulnerable regions that often predispose to genomic diseases (Gu et al. 2008). Isolating these regions based on high sequence homology will significantly reduce target regions and enable the development of hotspot-specific genotyping assays to capture disease associated deletions/duplication with both higher sensitivity and coverage. The breakpoints previously reported in SDs by aligning nonoverlapping read depth windows of 5 kbp using the reference human genome (Bailey et al. 2002; Alkan et al. 2009) limits the capacity to detect short highly homologous regions vulnerable to NAHR-mediated rearrangement.

In this study, we have identified genome-wide 'rearrangement hotspots' with elevated frequency of pathogenic NAHR mediated events. We have also detected an overwhelming number of overlapping CNV breakpoints, accumulation of DNVs and gene content transfer within hotspot regions. The read depth distribution of these hotspot regions revealed considerably higher read depth compared with the rest of the duplicated regions in the genome. The genome-wide characterization of 'rearrangement hotspots' will enhance the clinical applicability of high resolution genome analysis to uncover uncharacterized genomic disorders. Although current microarray platforms vary in both coverage and sensitivity (Tucker et al. 2011), the generation of a genome-wide 'rearrangement hotspot' map will serve as a powerful tool for a custom design of microarrays targeting regions vulnerable to mutational events that predispose to genomic disorders.

Although NAHR appears to be the dominant mechanism in the origin of pathogenic chromosomal rearrangements, the complete identification of hotspot breakpoints due to NAHR, NHEJ, MMBIR and FoSTeS remains to be fully characterized. The generation of a genome-wide high resolution map of 'rearrangement hotspots', which likely serve as templates for NAHR, represents a risk factor for manifestation of constitutional and acquired diseases as these regions are capable of mediating de novo deletions or duplications. Fine mapping limited to only 20% of detected hotspot regions identified in this study using microarray will detect NAHR-mediated deletions/duplications for 24 known genomic disorders and the remaining 80% will increase the possibility of detecting novel de novo chromosomal loss or gain. We anticipate that discovery of genomic variants using this robust hierarchical approach will translate not into the replacement of microarray-based methods with whole-genome or exome sequencing of patients suspected to have complex disease. Instead, it represents a valuable tool which can be utilized for superior design and selection of probes, and ultimately the creation of a customized microarray chip specifically targeting 'rearrangement hotspot' signatures to detect complex genomic diseases.

2.3 MATERIALS AND METHODS

Data Acquisition and Processing

We have obtained short read data for the NA18507 human genome sequenced using reversible terminator chemistry on an Illumina Genome Analyzer (Bently et al. 2008). The original data consisted of >30X coverage of the genome. We have obtained more than half of the data from the Short Read Archive Provisional FTP (NCBK) site (ftp://ftp.ncbi.nih.gov/pub/TraceDB/ShortRead/SRA000271/) with an average read length of approximately 36 bp. The analysis accuracy of this dataset has been previously described (Bently et al. 2008). The 4.8 MBp novel sequence detected in the NA18507 genome by a previous *de novo* assembly was also integrated in our rearrangement analysis. The length distribution revealed that the contigs/scaffolds are over fragmented and >80% of the sequence length is <1 KBp in length. The NA18507 human genome was selected as it is representative of the ancestral African Euroban population which has been previously shown to contain the most diverse polymorphisms compared with other populations (Sudmant et al. 2010; Alkan et al. 2009), rendering it an ideal sample to generate a 'rearrangement hotspot' map as the majority of the hotspot regions detected should exist within other populations.

Short Read Mapping

We have applied mrsFAST (micro-read substitution only fast alignment search tool - version 2.3.0.2) which implements an all-to-all algorithm unlike other short read mapping algorithms (Hach et al. 2010). Specifically, it is a fast alignment search tool which uses cache oblivious short read mapping algorithm to align short reads in an individual genome against a repeat masked reference human genome within a user-specified number of mismatches. We have mapped our short reads using mrsFAST with a maximum of two mismatches allowed against the repeat masked (UCSC hg18) genome assembly (supplementary 1, chapter 2). The advantage of using mrsFAST is that it returns all possible hits in the genome for a short read, allowing the detection of differential read depth distribution within duplicated regions of the human genome. Using the NA18507 human genome (18x coverage), 1.5 billion short reads were processed with 55.78% (i.e., ~839 million short reads) mapped to the repeat masked human reference genome with the mrsFAST aligner (supplementary 1, chapter 2) which returned all possible mapping locations of a read; a key requirement to accurately predicting the duplicated regions within the reference genome.

GC Correction

There exists a known bias with next generation sequencing technology towards GC-rich and GC-poor regions. Moreover, during library preparation using an Illumina Genome Analyzer, amplification artefacts are introduced in both GC-poor and GC-rich regions producing an uneven distribution of read coverage (Alkan et al. 2009) which has the potential of detecting false positive duplicated regions. We have used a simple GC correction method to reduce this bias. Overlapping windows (i.e., by 1 bp) with length '*l*' was used for read depth computation. Each read was assigned only once by its starting position and read depth was computed for each chromosomal position. The original mean read depth was calculated for each '*l*' length (i.e., 100 bp) block using equation (1). We have computed G+C percentage for every 100 bp window from the reference human genome and the read depth was subsequently interrogated for adjustment. The adjusted read depth was computed using the following equation:

$$RD_i, adjusted = RD_i \times \frac{m1}{m2} \tag{1}$$

where RD_i,adjusted is the read depth after GC correction, RD_i is the original read depth computed for i^{th} window, *m1* is the overall median of all the windows with 100 bp length and *m2* is the mean depth for all windows with same GC percentage. All subsequent analysis was carried out on the GC-corrected read depth.

Read Depth (RD) and Interval Detection

The first step in dissecting SD unit breakpoints using the NA18507 genome from all hit map information was to compute read depth from short read sequence mapping and detect SD intervals that do not overlap with a repeat region of the genome. Read depth was computed for each point after obtaining mapped anchoring positions of the short reads from mrsFAST. We have built a table for each chromosome, each containing coordinates where the common repeats are located. The read depth mean was computed for a chromosome from the genome content excluding common repeat regions. For each window with *l* length (100 bp) an event was determined. Events with excessive read depth and with a deletion were detected using equation (2).

$$event(l) = \begin{cases} 2(ExcessiveRD) & if \sum_{k=0}^{l} RD \geq mean + 2 \times st.d \\ 0(Deletion) & ifRD < 1 \\ 1 & otherwise; \end{cases} \qquad (2)$$

To investigate the interrogating window if it falls within a common repeat elements, we have built a library for the repeat masked regions (masked interspersed repeats, i.e. LINES, SINES, etc.) of the human genome. The mean length of the detected SD units was 822 bp (supplementary 1, chapter 2). The read depth distribution between the detected duplication subunits and the non-duplicated regions of the genome show significant read depth differences with an approximately 7% error rate (supplementary 1, chapter 2).

NA18507 Short Read Reference Guided Assembly

The current version of mrsFAST does not return the quality of the aligned reads within a consensus genome. Instead, we have used MAQ version 0.7.1 (Mapping and Assembly with Quality) which assembles genomes with a specified quality. MAQ searches for the un-gapped match with lowest mismatch score (i.e., maximum of 2) in the first 28 bp. To confidently map alignments, MAQ assigns each alignment a Phred scaled quality score which measures the probability that the true alignment is not the alignment that is detected by MAQ. If a short read maps to multiple positions in the genome, MAQ will randomly pick one position and give the excluded position a mapping quality of zero. We have mapped and assembled the NA18507 genome short reads into the reference genome using MAQ allowing at most 2 mismatches.

Detection of Genomic Re-arrangements

Using read depth as a measure to detect SD unit breakpoints may produce regions that share <90% sequence identity. To reduce false positive and computational burden after detecting SD unit breakpoints, we utilized a basic version of the end space alignment algorithm (without seed and extend approach) and performed pairwise alignment for each of the SD units against the rest of the genome SD units. We included only those SD units for rearrangement analysis described in the following section that contained at least one duplicon >100 bp with >90% sequence identity. We detected 20,237 SD units when every 100 bp window was assessed for a possible rearrangement.

End-Space Free Alignment Algorithm

The ability to detect highly homologous regions between two sequences is essential for duplicon detection. Multiple clusters of non-adjacent duplicons with >90% sequence identity cannot be mapped using basic alignment algorithms. As previously reported, the basic pairwise global alignment algorithm will miss duplicon breakpoints that are non-adjacent within an SD with different thresholds of sequence identity (Gu et al. 2008). Semi-global alignment has a tendency to produce pattern-like alignments (see example below), which are not informative for complex regions with multiple duplications. We have implemented a modified version of the pairwise alignment algorithm where the alignments are scored ignoring end spaces of the two sequences. Adding the option of end spaces in our alignment does not produce pattern-like alignments and therefore accurately pinpoints the breakpoints of the duplicon with an allowed gap that crosses the threshold of >90% sequence identity. The neutral rate of evolutionary decay suggests that 10% sequence divergence is required to accurately detect duplications that are primate-specific (Gu et al. 2008).

Example:

S1:ACGCAATTCGACTAGATCGGGTCGATGATCGATCGATGATCGAGACAGCATAGCAG

S2: CAATTCGACTAGATCGATCGACGATCGATCGAT

Semi-Global Alignment:

S1: ACGCAATTCGACTAGATCGGGTCGATGATCGATCGATGATCGAGACAGCATAGCAG

S2: ***CAATTCGACTAGATC*GATC***GA*CGATC***GAT*****C*G*AT*****

End-Space Free Alignment:

S1: CAATTCGACTAGATCGGGTCGATGATCGATCGAT

S2: CAATTCGACTAGATC*GATCGACGATCGATCGAT

In order to implement the algorithm, a dynamic programming technique was utilized which is a modified version of Smith-Waterman dynamic programming (Smith and Waterman 1981). This approach will detect the pairwise alignment relative to a penalty function corresponding to semi-global alignment. We used the dynamic programming (DP) algorithm to compute the above alignments and used the backtrack pointer to identify the best alignment.

Dynamic Programming Matrix and Recursive Trace Back

As a core searching algorithm, we have implemented a penalty function to complete the dynamic programming matrix *M*. First, we initialized the first column and row with zeroes which provided forgiving spaces at the beginning of the sequences in order to obtain the highest similarity between the interrogated sequences. Our intention was to locate duplicons between a pair of sequences (i.e.., s and t) with >90% identity and alignment with minimal gaps to avoid pattern-like structures. We encoded A with 1, G with 2, C with 3 and T with 4 to construct the $(m+1)\times(n+1)$ DP matrix *M*, where *m* and *n* is the length of two given sequences *s* and *t*, respectively. The algorithm uses a dynamic programming technique to fill a matrix *M* by a look up penalty function from the 5 × 5 matrix *C*. We have introduced penalty function *g(i,j)* for matched alignment with a score of 2. For the mismatches between a pair of bases, we introduced a penalty of -2 for mismatch and -3 for misaligned sequence produced by sequence assembly tools (i.e., MAQ). We used a -3 penalty to reduce the amount of misaligned portions of the sequence into duplicon identification. To allow the algorithm to ignore the end positions of the sequences if it has low similarity, we have performed a trace back from the highest value returned by function *Sim(s,t)* in the matrix *M* (supplementary 1, chapter 2). For any two given sequences (i.e., *s* and *t*), a semi-global alignment is an alignment between a substring (in this case duplicon) of *s* and *t*.

$$a[i,j] = \max\begin{cases} a[i,j-1]-3 \\ a[i-1,j-1]+g(i-1,j-1) \\ a[i-1,j]-3 \end{cases} \qquad (3)$$

$$Sim(s,t) = \max of\ M \qquad (4)$$

The memory requirement to fill out DP matrix *M* is *O(mn)*. The computational time to complete the dynamic programming Matrix *M* and to determine the maximum value in *M* for a given pair of sequence *s* and *t* with nearly similar length is $O(n^2)$ and to trace back starting from the maximum point in the matrix takes *O(m+n)* time to obtain optimal alignment.

Now, it might be apparent that ignoring end spaces might not detect true breakpoints and for long sequences it might produce really short alignments. Considering that majority of the commonly used alignment search methods (i.e., BLAST, BLAT, and SHRiMP) implement a "seed and extend" method to obtain faster sequence comparison (Altschul et al. 1990; Kent, 2002; Yanovsky et al. 2008), this method was also applied in this study. To perform an exhaustive search within the scope of 100 bp windows for any two given segmental unit sequences obtained from NA18507 genome, we have applied the dynamic programming algorithm for each 100 bp window with 10 bp overlaps as "seeds". The highly similar seeds (>90%) went through the "extend" step and the rest was ignored. We acknowledge that this approach might detect the same breakpoints multiple times if multiple seeding events are obtained from a highly duplicated region. Therefore, we have compared the previously extended duplicon breakpoints from the same SD unit and the overlapping "seeds" and only the maximum extended duplicon was kept (supplementary 1, chapter 2). 'Extend' is a recursive procedure which extends bi-directionally by 10 bp and the extend step ceases in each direction when further extension does not cross the sequence identity threshold. As a result, the procedure terminates if any further extension of both directions returns <90% sequence identity.

FISH Validation

Cytogenetic preparations were made from lymphoblastoid culture (obtained from Coriell cell repositories) for the NA18507 sample. The cell suspension was dropped on slides using a thermotone, aged overnight and hybridized with test (i.e., spectrum orange) and control probes. Following post-hybridization washes and 4,6-diamidino-2-phenylindole (DAP1) counterstaining, slides were analyzed using fluorescence microscopy. Pseudocoloring and image editing was performed using Photoshop software. To validate our duplicon rearrangement within SD units, we selected three complex regions in the human genome: 1q21.1, 16p12.1 and 22q11.21. In this study, we used fosmid genomic clones corresponding to a duplicated locus as a probe against chromosomal metaphase. The localization of FISH clones within these regions and the corresponding derivative loci validated >94% (i.e., 17/18) of the *in silico* co-localization predictions. The FISH technique was unable to provide a precise estimate of rearrangement at the level of 100 bp due to resolution limitations (supplementary 1, chapter 2).

Permutation

The basic analyses were conducted using a permutation procedure to assess statistical significance of 1-sided tests. The rearrangement for each SD unit was permuted randomly between the two groups and test statistics was computed in each permutation. All results reported in this study used 1 million permutations to derive an empirical *P*-value.

Gene Ontology Analysis

Gene ontology data analysis was performed using PANTHER (version 7.0) database (Mi et al. 2010). We have analyzed the biological processes of the hotspots genes (supplementary 1, chapter 2).

Chapter 3

Atypical Micro-duplications at 2q21.1-21.2 Co-segregates with Tourette Syndrome in a Three Generation Family

Mohammed Uddin[1], Darren D. O'Reilly[1], Tanya R. Purchase[2], Hubert White[2], Sandra Luscombe[3], Susan J Moore[3], Terry- Lynn Young[1], Proton Rahman[1] Kathy A Hodgkinson[1]

1. Faculty of Medicine, Memorial University,
2. Pediatric Psychiatry, Health Sciences Centre, St. John's, Canada
3. Pediatric Development and Rehabilitation, Health Sciences Centre.

PREFACE

The primary contribution of this manuscript is the detection of microduplications at 2q21.1-21.2 within multigenerational family affected with Tourette syndrome (TS). The detection of these microduplications was carried out by applying a custom array that has been designed targeting the breakpoints identified in chapter 2. The larger block of microduplication consists a gene C2orf27A and the microduplication found to be rare in control population, implicates a link between 2q21.1-21.2 locus with TS pathogenesis. A version of this manuscript has already been submitted in *American Journal of Medical Genetics, Part A*.

The co-authors Drs. Proton Rahman, Darren O'Rielly, Kathy Hodgkinson helped the primary author Mohammed Uddin on the study design. The primary author, Mohammed Uddin, conceptualized the problem and developed the framework of the study design. In addition, the primary author conducted all related computational analysis presented in this manuscript. All the co-authors worked as a team in preparation of the manuscript.

ABSTRACT

Tourette syndrome (TS) is a developmental neuropsychiatric disorder characterized by the presence of both motor and verbal tics. TS may be accompanied by co-morbidities including obsessive compulsive disorder (OCD), attention deficit hyperactivity disorder (ADHD) and poor impulse control. The prevalence of TS is between 0.3-1% and affects males more than females (M:F-3:1). It has a major genetic component with concordance rates for monozygotic and dizygotic twins estimated at 50-70% and 10-23%, respectively. Despite numerous linkage and association studies, no common candidate genes have been identified. Recent analysis of copy number variations (CNVs) in TS reveal an association with genes previously implicated in autism spectrum disorder (ASD) although none have been reported unique to TS.

We have imitated an epidemiological study of families with TS in Newfoundland and Labrador, Canada and have ascertained one multiplex family with five family members clinically diagnosed and another potentially affected (by collateral family history). Recruited family members include four young siblings, their parent and maternal grandfather. Three of the siblings are significantly affected with concomitant OCD, ADHD and anxiety. The fourth has TS with fewer co-morbidities. We designed a custom array CGH targeting genomic hotspots susceptible to rearrangements. The high density array included 2 X 1 million probes with a mean spacing of 270bp. We found two common micro-duplications, one 38KB and the other 131KB in length. The larger duplication comprised the *C2orf27A* gene located at 2q21.1-21.2 that co-segregated with six family

members diagnosed with TS and also detected in an unrelated affected individual. Our finding has been validated using rt-PCR. Addition rt-PCR analysis on 443 control individuals from Newfoundland population revealed rare frequency of the larger block, 4/443. The function of the *C2orf27A* gene is currently unknown. In summary, we have identified CNVs that segregates with TS in a multiplex family. The importance of this and other disease-associated CNVs to the pathogenesis of TS (and associated syndromes) is currently being explored through ascertainment of other multiplex families from the Newfoundland and Labrador population.

3.1 INTRODUCTION

Newfoundland Population Genetics

In this thesis, the genetic predisposition to neuropsychiatric and ankylosing spondylitis was analyzed in the population of Newfoundland and Labrador. As Newfoundland population a well known founder populations (Figure 3.1), it s prudent to beware about the genetic architecture of Newfoundland founder population.

Giovanni Caboto, an Italian navigator, first reached Newfoundland in 1497 and seasonal colonies were first established around 1610. The major influx of immigration to Newfoundland occurred in the mid-1700s, who are the early settlers and included mainly Protestant settlers from the south-west of England and Roman Catholic settlers from the south of Ireland (Rahman et al. 2003) (approximately 20,000 founder). This population grew by natural expansion and settled in other geographic locations in the province by 1850 (Figure 3.2). The population reached approximately 200,000 in 1890(Fernandez 2009). This rate of expansion as a young genetic isolates is comparable with other isolate populations (rahman et al. 1003). The population is genetically isolated and geographically consists of the island of Newfoundland and the mainland Labrador. According to statics Canada, 2011, the current population is approximately 510,000 with 95% of English or Irish descent.

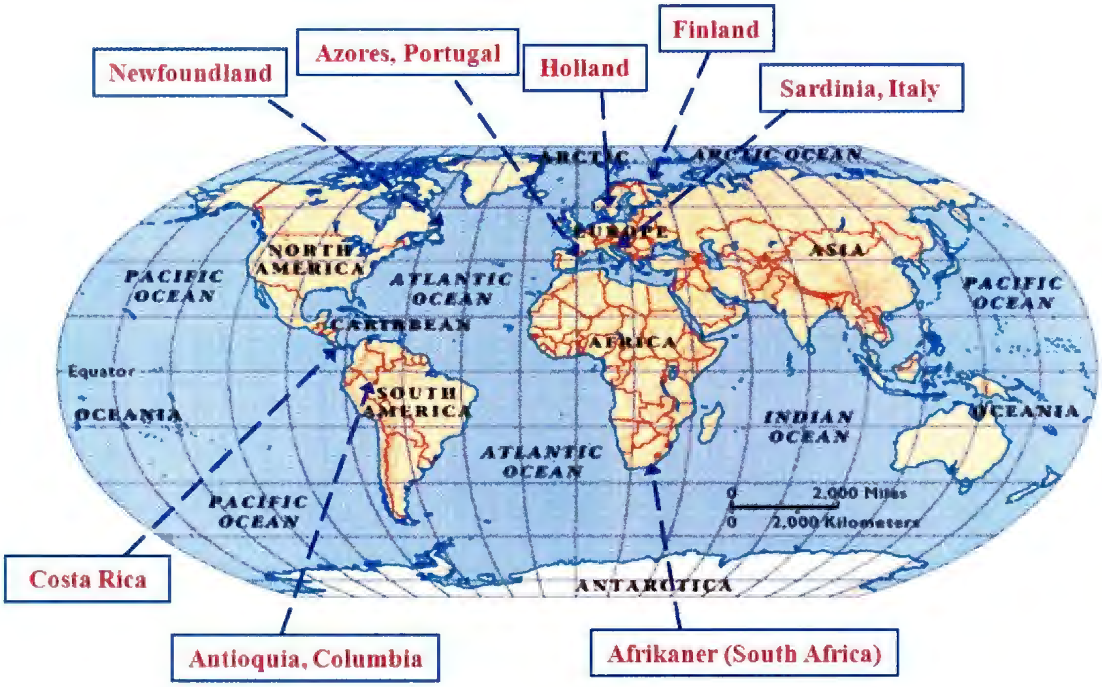


Figure 3.1 Geogrphic locations of world founder populations. (The original image is used from Rahman et al. 2003 with the permission of the corresponding author).

The Newfoundland population inbreeding analysis shows the presence of high degree of relatedness that leads to persistent homogeneity of Newfoundland population (Bear et al. 1987; Bear et al. 1988). This homogeneity of Newfoundland population is directly correlated with the presence of genetic drift. Genetic drift is a phenomenon that refers to the fluctuation in allele frequency (comparing other populations) due to small gene pool contained within a small population. Genetic drift is a strong force that is responsible for the imbalance of disease prevalence within a founder population. It can force the diseases incidence in both extreme and often can lead to an increase incidence of rare diseases or decrease incidence of common disease can be rare. Hence, similar to other isolated populations in the world, there exists a high burden of genetic diseases in Newfoundland (Fernandez 2009). For example, Bardet-Biedle syndrome (BBS) (is a form of retinal dystrophy) is a autosomal recessive disease has a higher incidence rate in Newfoundland population (1/18,000) compared with (1/160,000) admixed Caucasian populations of northern European ancestry (Green et al. 1989).

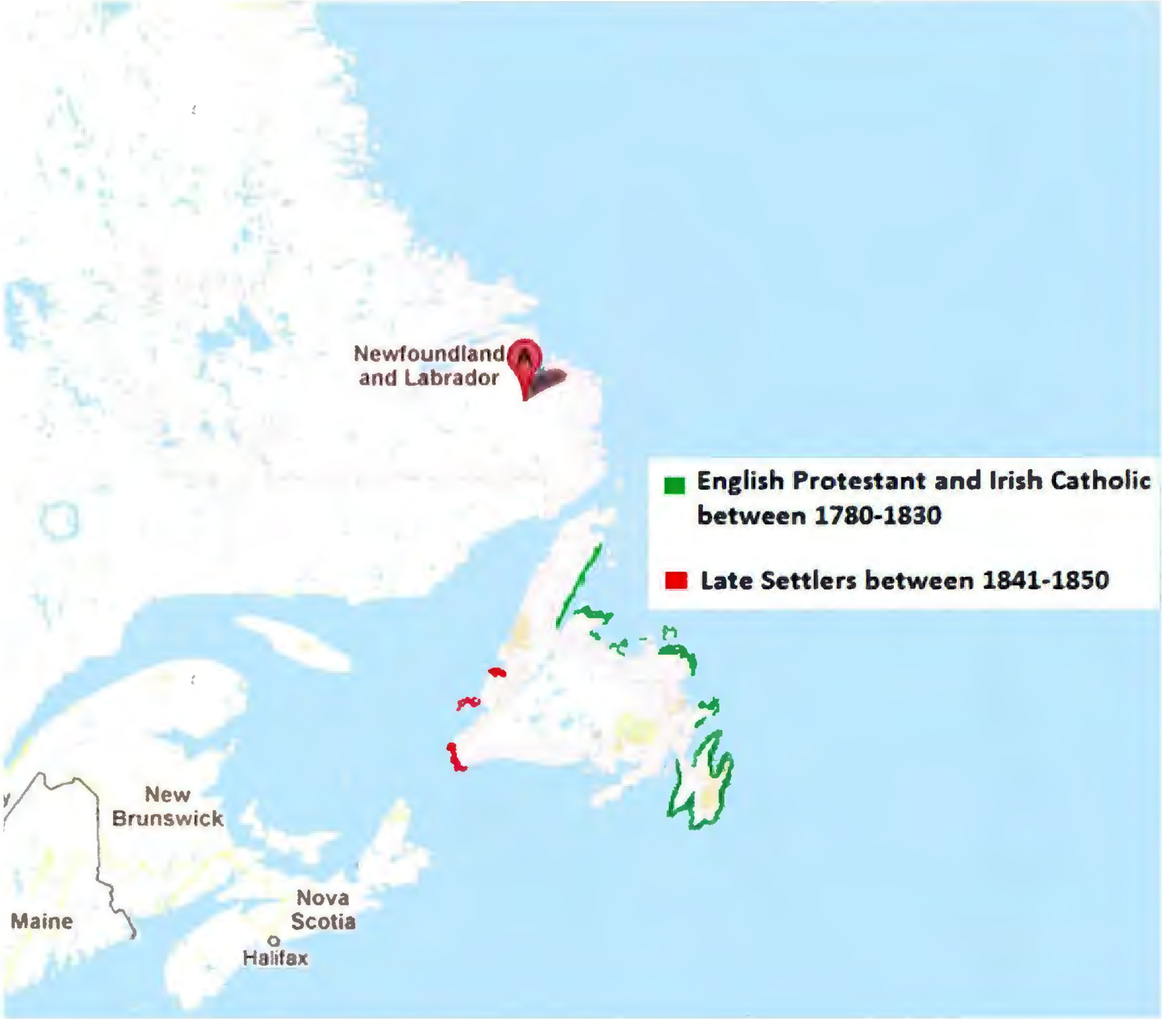

Newfoundland and Labrador
🟩 English Protestant and Irish Catholic between 1780-1830
🟥 Late Settlers between 1841-1850
New Brunswick
Maine
Nova Scotia
Halifax

Figure 3.2 Geographic locations of early and late settlers.

The trend of high incidence rate also observed within complex diseases among Nefoundland population. The incidence rate for Psoriasis is 2 to 3 fold higher compared with most other Caucasian populations (Nall et al. 1999). Juvenile type 1 diabetes in Newfoundland has a incidence rate of 40/100,000 each year, representing the highest known incidence of juvenile type 1 diabetes worldwide. This exceeds the incidence rate in admixed US population (7-15/100,000) and other recognized founder populations i.e. Sardinia and Finland (36.5-36.8/100,000) (Karvonen et al. 2000).

Founder mutation is another concept where one of the original founders carries a rare allele that can be observed higher in frequency after few successive generations. Founder mutations have been noted in multiple diseases in Newfoundland population including multiple endocrine neoplasia type 1, hereditary non-polyposis colorectal cancer syndrome (HNPCC), hemophilia A, and Bardet–Biedl syndrome (Olufemi et al. 1998; Spirio et al. 1999; Xie et al. 2002; Young et al. 1999). Although it is a daunting task to identify founder mutations, a deviation in frequencies is observed in many diseases within Newfoundland population. For example, a single MSH2 mutation has been reported in 50% of Newfoundland families with HNPCC whereas the mutation was only observed 8% of English HNPCC families (Spirio et al. 1999). These observations highlight the stratified nature of disease incidence as well as unique genetic architecture within the Newfoundland population, and provide a unique opportunity to identify disease causal variants.

Tourette Syndrome

Tourette syndrome (TS) is a developmental neuropsychiatric disorder characterized by the presence of motor (simple and/or complex) and verbal tics with a duration longer than one year. TS often manifests with features associated with obsessive compulsive disorder (OCD), attention deficit hyperactivity disorder (ADHD), poor impulse control and other behavioural abnormalities (Robertson, 2012). The pathophysiology of the complex comorbidities associated with TS remains to be elucidated. The prevalence of TS is between 0.3-1 percent in any given population (Robertson et al. 2009; Robertson 2008; Centers for Disease Control and Prevention 2009), and consistently is shown to affect males more than females (Robertson, 2012). Twin studies consistently show higher concordance rates in monozygotic compared with dizygotic twins (Price et al. 1985; Walkup et al. 1988; Pauls et al. 1991) suggestive of a strong genetic component contributing to disease pathogenesis. Although early segregation analyses suggested an autosomal dominant inheritance pattern, recent evidence suggests a heterogeneous complex genetic architecture underpins the pathogenesis of TS (Pauls et al. 1986; Eapen et al. 1993; State 2010; State 2011).

In the last decade, emphasis has focused on identifying common variants (i.e., >1% frequency) associated with common complex diseases. Several recent TS studies highlight the contribution of rare gene mutations and chromosomal abnormalities in disease pathogenesis. For example, a rare mutation (var321) has been identified in the SLIT and NTRK-like family, member 1 (*SLITRK1*) gene

(Abelson et al. 2005). The var321 variant was demonstrated to be highly stratified within multiple populations, however replication analyses have produced conflicting results (Scharf et al. 2008, O'Roak et al. 2010). Another rare functional mutation was identified in a two-generation pedigree in the L-histidine decarboxylase (*HDC*) gene and suggests a role for histaminergic (HA) neurotransmission in the genesis or modulation of tics (Ercan-Sencicek et al. 2010). That mutation was identified as a heterozygous G-to-A transition at position 951 in exon 9 of the *HDC* gene, resulting in a premature termination codon (p.W317X). This nonsense mutation was present in the affected father and all affected offspring within the pedigree and absent in the unaffected mother and in 3000 control samples. Subsequent replication analysis demonstrated the absence of this mutation in a cohort comprising 720 TS and 320 control samples (Ercan-Sencicek et al. 2010). Likewise, replication in a Chinese population also failed to show significant association (Lei et al. 2011). Collectively, this evidence strongly suggests that this nonsense mutation is thus likely to be a private, rare mutation segregating in this family.

Structural variations are a major risk factor for numerous neuropsychiatric diseases. Recent analysis of copy number variations (CNVs) in TS have demonstrated an association with genes previously implicated in autism spectrum disorders (ASD) and other neuropsychiatric disorders (Senthil et al. 2010; Lawson-Yuen et al. 2008; Fernandez et al. 2012). A rare deletion of exons located 5' in the neurexin 1 (*NRXN1*) gene was identified in 2 unrelated TS patients (Senthil et al. 2010). A second deletion in the α-T catenin (*CTNNA3*) gene was identified in two independent TS studies (Senthil et al. 2010; Fernandez et al. 2012). Interestingly, deletions encompassing both *NRXN1* and *CTXN1* gene have been reported in ASD and schizophrenia (SCHZ). Another rare deletion comprising the *NLGN4* gene (i.e. exons 4, 5 and 6) was been previously reported in TS and ASD (Lawson-Yuen et al. 2008). A complex rearrangement (insertion/translocation) between chromosome 2 and 7 was reported to have disrupted the *CNTNAP2* gene in a two generational pedigree (affected father and two children) (Verkerk et al. 2003). The identification of these CNVs that may be implicated in multiple neuropsychiatric disorders complicates genotype-phenotype elucidation. That no single CNV has been reported that is unique or that the segregation is unique to families affected with TS, provides a great opportunity to detect novel CNVs specific to TS which, in turn, may help simplify genotype-phenotype correlation.

3.2 RESULTS

We have designed a high-density array-based comparative genomic hybridization (CGH) experiment by mapping with 2 X 1million probes in regions that are susceptible to genomic rearrangements. We applied our custom 2M microarray on eight members from a three generational family which yielded approximately 2000 aberrations (Figure 3.3, Table 3.1). Comprehensive data analysis revealed the discovery of atypical microduplications located at chromosome 2q21.1 which segregated with the cases in family A, whereas large de novo variants were not detected within the family members affected with TS. Within a 221KB (chr2:132305299-132526804) region, two (2) common blocks of microduplications were identified that segregate together in five (5) out of the six (6) affected individuals with a smaller region of overlap of block2 in the remaining individual (Figure 3.4; ID3003). The blocks,which are separated by 51Kb, are 38KB (block1 - chr2:132305299-132343808) and 131KB block2 - chr2:132395155-132526804) in length. All affected family members carry nearly identical breakpoints except the fourth affected sibling (ID3003) who has a partial 30KBp duplication within block2 (chr2: 132480185-132510827). All affected family members carry a microduplication which comprises the C2orf27A gene. Complex rearrangements were also detected in block2 for individuals ID1002 and ID2002 which contained small microdeletions of 4.1 and 2.4KBp, respectively.

The presence of block2 among five (5) of the six (6) affected family members has been validated using the QF-PCR assay which demonstrated that a relative copy number of 4 within the affected siblings and mother whereas unaffected members have a copy number of 2 or 3 (Figure 3.4). We performed rt-PCR analysis on two additional families and 10 unrelated individuals affected with TS. The block2 micro-duplication with copy number of 4 was detected in one additional affected individual (ID10003), but absent in all other unrelated affected or unaffected samples tested. To quantify the population frequency of these duplication blocks using QF-PCR, we have analyzed a set of 443 individuals where copy number was determined with 99% confidence interval. The frequency of a copy number of 4 was observed in 4/443 (0.009).

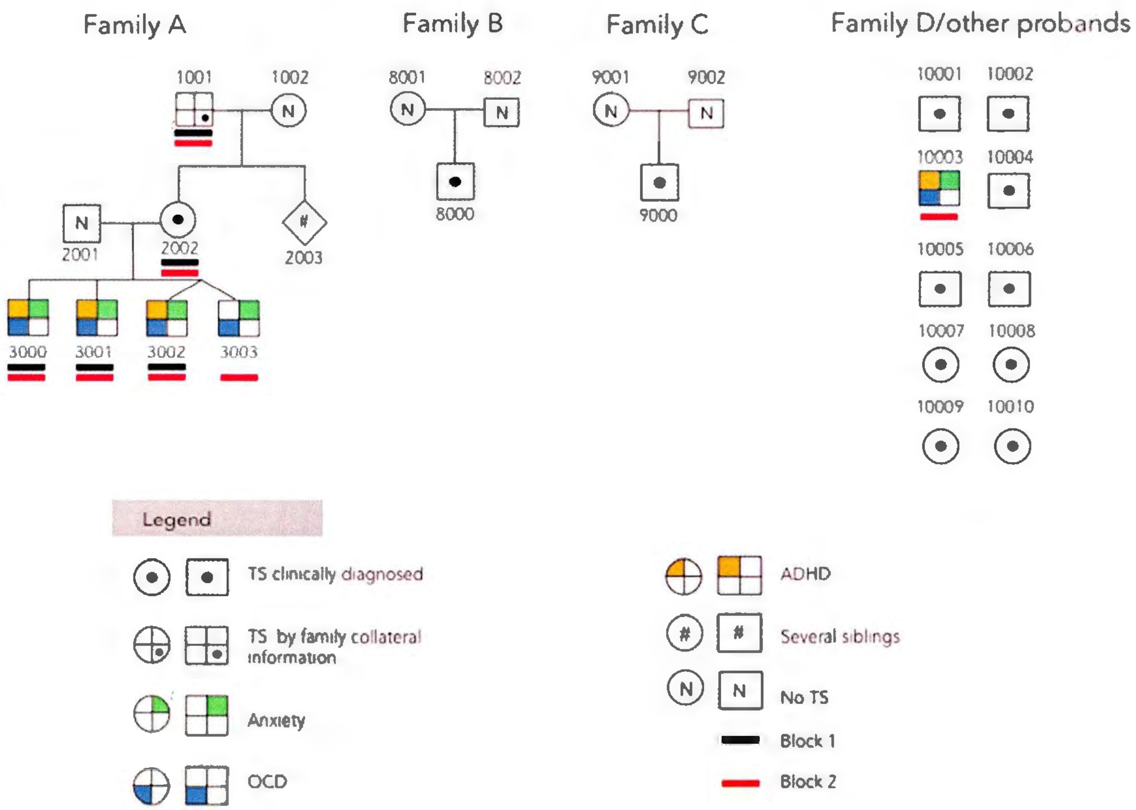


Figure 3.3 Tourette Syndrome samples – Family A. a three generation pedigree, Families B and C: Two trios with affected offspring, D. Ten unrelated TS affected individuals.

Table 3.1 Clinical Features of Family A and Proband B.

ID	Birth Hx	Age at dx.of TS	Presentation and Developmental History	IQ/ Cognitive Profile	Tics	Treatment	Co-Morbidity and age of diagnosis	Other Medical Problems
3000	C section Breech Breast fed 10/12	9y	Normal early milestones. Referred to Developmental Paediatrician (8y) by School Guidance Counsellor for possible ADHD	**WISC (8y)**:Overall high-average IQ. FSIQ 109, VIQ 123, PIQ 93. **Psychoeducational assessment (15y)**: written output learning disability. **WAIS IV (18 y)**:FSIQ 104 (61%); VCI 108 (70%); PRI 116 (86%); WMI 102 (55%); PS 77 (6%) **WIATII (18 y)**: spelling superior, listening comprehension high average, other areas average.	Vocal and motor tics since 6y. Increased tics at 12y. Worst tics reduced by 14y.	Melatonin Risperidone Ritalin Concerta Strattera Prozac Adderall Accutane	ADHD combined type 8y OCD 9y Anxiety 9y school refusal 14y. Addiction to gaming 19y	Allergic to dust mites 9y. Mild asthma 12y. Gynecomastia 13y. Chronic headache 13y. Sleep disorder 14y (difficulty falling asleep, waking early, snoring: sleep study negative). Possible mood disorder, (very apathetic) ongoing assessment with Psychiatrist 19y. Acne 16y
3001	C section Placental failure Breast fed 12/12	6y	Normal early milestones. Referred to Developmental Paediatrician at 4 yrs for assessment of speech problems aggression frustration and fine motor skill delay . Possible ADHD/TS	**WISC (8y)**: overall average: FSIQ 93, VIQ 95, PIQ 111,WMI 86, PS, 88 **WIATII (8y)** composite score: extremely low for reading writing, low average for oral language and borderline for math.	Vocal and motor tics since 6y	Melatonin Risperidone Ritalin Concerta Strattera Clonidine Accutane	ADHD 7y Dyslexia (severe) 9y Sensory integration disorder 9y OCD 9y Auditory processing disorder 14y	Nocturnal enuresis 6y Sleep disorder (? sleep apnea) 8y Acne 13y
3002	C section 2.5 weeks early	8y	Normal early milestones. Referred to developmental paediatrician	**WISC (10 y)**:Overall high-average IQ. FSIQ 117, VIQ 120, PIQ 120. WMI,	Vocal and motor tics since 6y. complex tics at 8y	Melatonin Concerta Strattera Prozac Adderall	ADHD 7y OCD 9y Anxiety 9y	Trace tricuspid and pulmonic regurgitation 9y

	Breast fed 8/12		at 5y re-concerns about fine motor skills, speech development, hyperactivity poor impulse control and aggression.	97, PS 109 Spelling skills, borderline, listening comprehension average		Paxil		Sleep disturbance 5y
3003	C section 2.5 weeks early Breast fed 8/12	9y	Normal early milestones. Referred to developmental paediatrician at 5y re-aggressive behaviour and obsessions. Easily frustrated	Not done	Vocal and motor tics since 6y	Melatonin	OCD 6y Dyslexia 9y Anxiety Fearful of school 10y	Sleep disturbance 7y Sensory issues 5y
2002	NVD	44y	Normal early milestones	Not done	Vocal and motor tics since childhood	Concerta Strattera	ADHD 38y Self manages anxiety and OCD.	Cervical dysplasia 22y Renal colic 25y Sensory issues and Migraine headaches 31y
1001	N/A	N/A	N/A	Not done	Vocal and motor tics* Easily frustrated*		Has OCD traits*	Atrial Fib 69y Basal cell carcimoma 78y Colon Cancer 79y Abdominal aortic aneurysm 80y
10003	NVD	12y	Normal early milestones Referred to pediatrician 12y. History of hyperactivity, ADHD, short attention span, anger issues, sleep problems	Not done	Motor tics 5y Vocal tics10y OCD worsening15y, resolving 19y	Risperidone Strattera Clonidine Seroquel Prozac Celexa Haldol Clonazepam Luvox	ADHD 12y OCD 13y Anxiety NOS 17y	tympanostomy tubes 6y Systolic heart murmer 11y Sleep problems (difficulty falling asleep) 13y Equinus deformity of foot 14y

TS: Tourette Syndrome

ADHD: Attention Deficit and Hyperactivity Disorder

FSIQ: Full Scale Intelligent Quotient

VIQ: Verbal Intelligent Quotient

PIQ: Performance Intelligent Quotient

VCI: Verbal Comprehension Index

PRI: Perceptual Reasoning Index

WMI: Working Memory Index

PS: Processing Speed

y: years

NVD: normal vaginal delivery

N/A: Not available

* Family information only

X/12: number of months

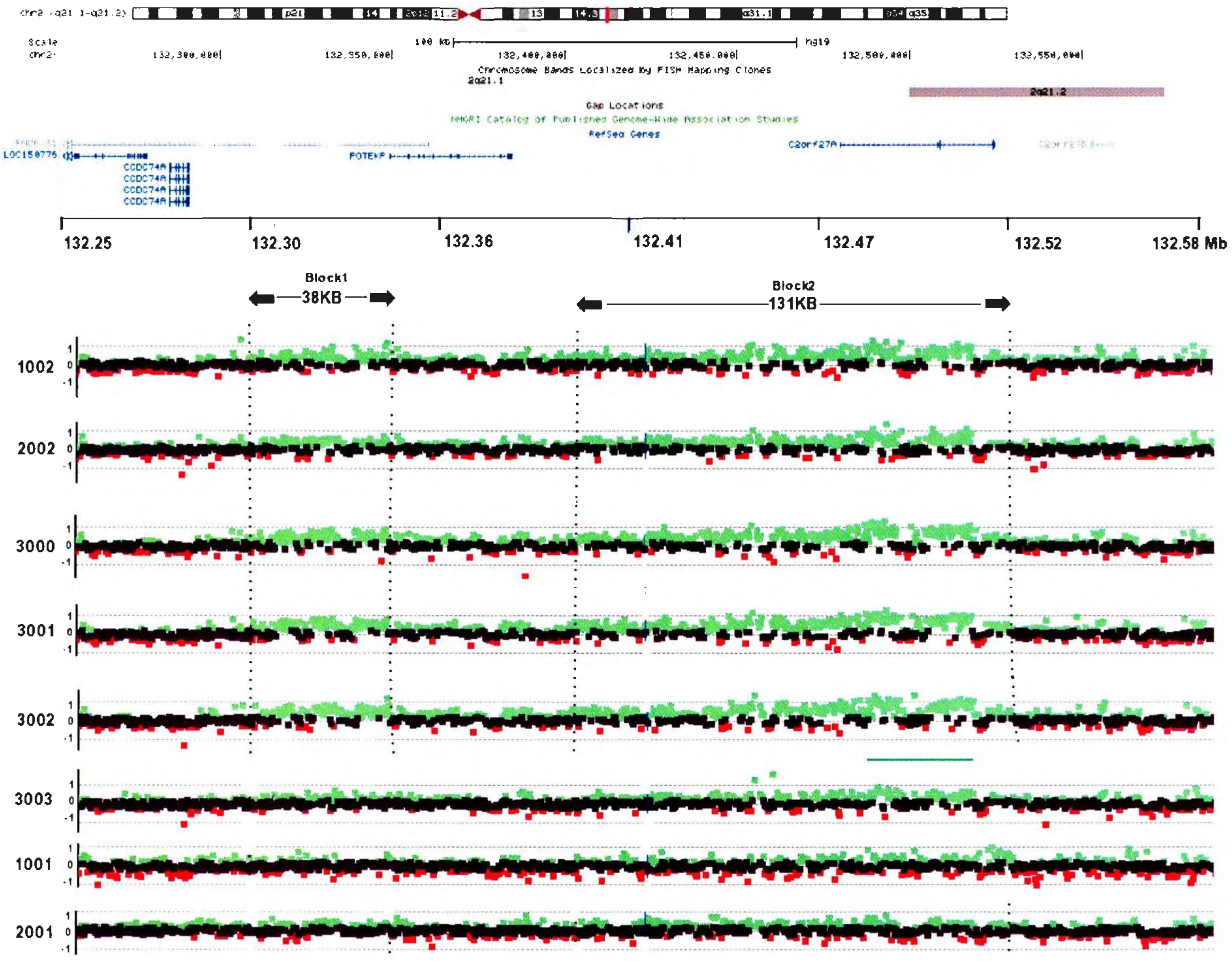


Figure 3.4 Array CGH result from the three generation pedigree of Family A. Each point (red/black/green) represents a probe intensity value. The dotted line shows genomic gains in two distinct genomic blocks within the affected samples. The horizontal geenline for sample ID3002 shows partial gains.

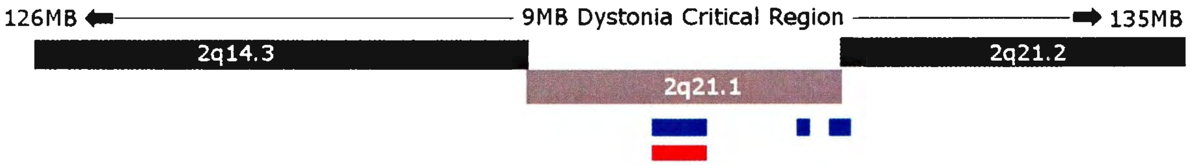


Figure 3.5 Illustration of the 9MB critical region at 2q14.3-q21.2 found in multiplex family with Dystonia. The reciprocal micro-duplications and deletions region found in TS and ADHD samples is also shown in blue and red vertical block, respectively. The micro-duplications reported in this study at 2q21.1-21.2 also shown.

3.3 DISCUSSION

The salient characteristics of TS are manifest in the subjects presenting with the microduplications. These are multiple motor and vocal tics, and common co-morbidities including ADHD, OCD, major depression, anxiety, behavioural problems, and learning disability (Termine et al., 2006). Additionally recognised phenotypic co morbidities include migraine and sleep difficulties (Abelson et al, 2005; Freeman et al, 2000; Kwak, 2003; Lesperance et al., 2004; Singer, 2005) which are also clearly present.

TS appears to segregate with the microduplications described in the manuscript. The morbidity of disease however varies with the individuals represented. The proband (Figure 3.3, 3001) and his three siblings exhibit various features of TS, with the three oldest siblings manifesting the greatest phenotypic morbidity. The mother (2002) and father (2001) of the boys function at a high level socially and occupationally, as do the grandparents. Subjects 1001 and 2002 function well in all objective measures of functionality, and have done so throughout life, although 2002 has been diagnosed with TS, and childhood information suggests that the TS diagnosis would have been made in childhood were she to present to the medical profession as a child today. Subject 1001 has manifested only tics throughout his lifetime. Of early interest in Family A was subject 3003 who presents with the a partial 30KBp duplication within block2 and has a more benign phenotype than his siblings. However, subject 10003 also has the same partial duplication and has a phenotypic presentation similar to the older three siblings in family A, and the common two blocks of microduplications are also present in the mother (2002) and grandfather (1001), neither of whom present with the same burden of disease. Although TS has been considered a single problem, evidence from factor analysis suggests that the disorder can be separated into symptom groupings, with potentially different etiologies (Cavanna et al. 2011). Hence, genotype-phenotype correlation is much more complicated.

We provide here a description of a region containing two micro-duplication blocks that segregates within a family with TS. Unlike previous TS reports on CNV association, these micro-duplications not yet found to be reported with any other neuropsychiatric disorders. This region previously was identified as a locus for distonia by a linkage analysis on a four generation family (Norgren et al. 2011). A 8.9 MBP critical region shown to have the highest LOD score that includes the micro-duplications region identified in this report (Figure 3.5). Extensive analysis was carried out to detect mutation and CNVs within this region applying re-sequencing and Illumina 2 million CNV microarray. This is particularly interesting because of the apparent phenotypic similarities between dystonia and Tourette Syndrome. In a recent report, small recurrent deletions and reciprocal duplications were identified within 2q21.1-21.2 region. The deletions are identified within a patient with ADHD, and the duplications are detected among unrelated individuals with developmental delay (DD)/intellectual disability (ID), ADHD, epilepsy, and other neurobehavioral abnormalities (Dharmadhikari et al. 2012). This region is approximately 500kb downstream from duplication blocks identified in this report (Figure 3.5).

Previously published high-density CNV datasets from Conrad et al. (Conrad et al. 2010b) demonstrated the presence of common deletions interspersed within the region. Very low frequency (between 0.01-0.07) typical duplications have been reported in the Database of Genomic Variants (DGV) within this region (with no validation result). The DGV demonstrated typical CNV gains with low frequencies and of the reported studies, no single individual carries two duplication blocks within this region. This implies the unusual segregation of the two micro-duplication blocks within TS family is highly correlated with TS. These breakpoints are also absent within the large study that investigated 15,767 children with various types of intellectual disability (Cooper et al. 2011).

The finding in this study emphasizes the impact of CNVs with respect to human health and genomic susceptibility to Tourette syndrome. The larger microduplication that segregates within the affected individuals of Family A consists C2orf27A gene. The function of this gene is yet to be elucidated. The rare frequency of this microduplication within the control population indicative of a possible link between 2q21.1-21.2 locus with TS and its pathogenesis. Until a large TS cohort is analyzed, it is difficult to know the actual frequencies of these microduplications within TS individuals in a population. CNV variants within other neuropsychiatric diseases often show the property of manifesting within a family. These variants are found to be extremely rare or absent within the control and the disease group for a population (Sato et al. 2012). Further investigation on large cohorts is required to quantify the frequencies of these microduplications within

TS individuals. The microduplication in this study within 2q21.1-21.2 locus add to the growing list of recurrent CNVs associated with Tourette syndrome and its related features (i.e. ADHD, OCD). Further delineation of the 9MB (Figure 3.5) locus will strengthen the hypothesis of a major risk factor for TS at 2q14.3-2q21.3.

3.4 MATERIALS AND METHODS

Creation of Custom Design Microarray

The genome wide CNV detection was carried out by designing a custom microarray. After the detection of 'rearrangement hotspot' regions in chapter 2, we grouped the distal and proximal breakpoints that are clustered within 1 MBp regions based on physical genomic coordinates. This enabled the ability to detect genomic losses and gains, not only in the rearrangement hotspots but also to investigate the critical region between the hotspots. These critical regions are often overlapped with SDs and are often prone to copy number changes. The importance of the identified breakpoints are evidenced by analysis which revealed that 20% of the detected hotspots are clustered within the proximal and distal SD breakpoints flanked by the pathogenic deletions/duplications that have been mapped for 24 NAHR-mediated genomic disorders (see Figure 2.3). Implementing the custom designed chip using aCGH technology will enhance the ability to detect novel SVs by expanding the scope of coverage by a further 80%. Further analysis revealed that the detected hotspots encompass all of the centromeric and telomeric regions which are often susceptible to genome rearrnagment that are associated with complex disorders (i.e., 3q29 recurrent deletion/duplication syndrome). Selection of Microarray Platform Given that detection of a CNV typically requires a signal from at least 3 to 10 consecutive probes, CGH microarrays can routinely detect anywhere from dozens to several hundred events per genome depending on the platform applied (Alkan et al.

2011). That on the Agilent CGH platform there is little additional genotyping power to be gained from placing more than 10 probes in each CNV. There appears to be a correlation between probe-length and variability, with longer probes giving less variance in fluorescence intensity (Pinto et al. 201). The number of probes required to detect a single-copy alteration varies between platforms, with Agilent Technologies offering the highest per-probe performance and *in silico* probe performance score provides a meaningful metric for prioritizing probes for inclusion in microarrays. Although alterations can, theoretically, be detected with a single probe using the Agilent platform, a more conservative detection limit of three to five probes is advised (Alkan et al. 2011). Of the commercially available CGH platforms, the Agilent platform produced the largest number of CNV calls (Pinto et al. 201). Moreover, higher resolution platforms have been previously demonstrated to more accurately capture the true proportion of genic CNVs and those in close proximity to segmental duplications.

The superiority of array CGH over SNP arrays is attributed to a combination of differences in probe coverage and the type of reference used. Therefore, CGH arrays, and in particular, the Agilent platform, have greater sensitivity to detect small differences in copy number (e.g., four versus five copies); a finding of critical importance in the investigation of structural variants in disease association (Pinto et al. 201). Although CNVs can be better detected by aCGH in terms of resolution, this will undoubtedly change in the coming years with modifications to existing next generation sequencing (NGS) platforms but still economically unfeasible for studies with large cohort. Storage and subsequent analysis of NGS and full exome data is complex, computationally exhaustive, time-consuming and expensive and is therefore considered impractical for large scale investigations of structural variants underlying disease pathogenesis. To alleviate the identified problems and cost associated with NGS approach to studying large cohorts, we have compiled from chapter 2, all of the hotspot breakpoints analyzed from our African individual based on NGS into 2 x 1M microarray chips from Agilent to be used as our custom microarray. We converted the grouped breakpoints from human genome 18 build to human genome 19 build using the Lift-over application provided by University of California, Santa Cruz (UCSC) genome browser.

The eArray platform created by Agilent was used to search for probes to include in the custom array design. As we are focusing on the more complex regions of the genome, there is lack of validated probes located within these regions. To overcome this limitation, we included genome-tiling probes in addition to high-quality, validated probes to achieve maximum coverage in these complex regions. To achieve adequate high resolution throughout the genome, we grouped the rearrangement hotspot breakpoints into four different probe sets with defined probe spacing. The first three probe groups consist of recently discovered CNVs (Conrad et al. 2010b) which represents ~10% of probes within this probe group (i.e., 190 bp spacing for breakpoints <1 KBp; 300 bp spacing for breakpoints within 1-5 KBp; and 450 bp spacing for breakpoints >10 KBp). The final probe group, which comprises 90% of the selected probes, contains all of our detected hotspot breakpoints with an average probe spacing of 280 bp. A primary advantage of using a genomic-tiling approach is that it consists of probes that are uniformly distributed which is likely to cover all exons that are located within the specified breakpoints. Normalization control probes that represent nonaberrant/non-variant regions were included in the design for accurate normalization of the data using Agilent's Feature Extraction software. Replicate probes were also included in the design to compute the reproducibility QC metric in Agilent's Feature Extraction and DNA analysis software.

QC measure was applied and the derivative of log ratio spread (DLRS) <0.25 was considered as a threshold. CNVs were detected using the built-in Aberration Detection Method-2 (ADM-2) algorithm DNA Analytics v.4.0.85 (Agilent Technologies) using the following criteria: at least 5 probes for a CNV call on GC corrected intensity, nested filter was set to 2, log intensity >0.25 for duplication and <-0.25 for deletion was considered. Custom script was applied to detect gene enriched CNVs (overlaps or consist a gene) that segregates (at least three cases) within the affected and absent in the unaffected members of the family.

Study Population

Probands and families (Figure 3.3) were ascertained through a prospective study of TS in Newfoundland and Labrador, from the Janeway tertiary children's hospital psychiatric department (HW, TP) and child development services (SL, SM). This study obtains extended family histories and in depth phenotyping, and is approved by the Human Research Ethics Board (# 07-71). To date 28 probands have been recruited and eight multigenerational family histories completed. DNA samples are collected routinely from all affected subjects, their parents and extended family members (in multiplex pedigrees) following consent and completion of multiple rating scales. To date, one multiplex pedigree has been ascertained (Family A).

Family A

The proband (Figure 3.3, 3000) presented at 9 y to a developmental pediatrician following a referral from a school counsellor and was diagnosed with TS. His three siblings were subsequently diagnosed at 6, 8 and 9 years respectively (Figure 3.3, 3001, 3002, 3003). The mother of these boys had several issues relating to anxiety and obsessive traits and vocal and motor tics since childhood, however, was only diagnosed as having TS at 44y (Table 3.1). The father of the four affected boys has mild anxiety but no tics. The grandfather has no diagnosis, but has manifested both verbal and motor tics through his lifespan according to family information, although these were not obvious during a recent assessment by a psychiatrist. The grandmother was diagnosed with bipolar disorder in her 70s and is treated effectively but has no current tics, nor any history of tics.

Proband 10003

The proband presented at 12 years to a pediatric psychiatrist and was subsequently diagnosed with TS (Table 3.1). The extended family history is in progress, but is complicated by missing information due to adoption.

Families B and C have no known extended history of TS or other co morbidities.

Control Population

To assess the population frequency of the CNV detected using the custom 2M microarray, we used 590 control samples from the NL population with no clinical report of TS and performed real time quantitative polymerase chain reactions (rt-PCR). The recruitment of the controls was approved by the Human Research Ethics Board.

QF-PCR Validation

To confirm the duplication detected using the custom 2M microarray, a Taqman copy number assay Hs03417816_cn (Life Technologies) was performed using the manufacturer's recommended protocol. The assay was performed in quadruplicate on 10ng genomic DNA for each sample in a 96-well plate. The 10 µl reaction mix consisted of 2×µl 2x Taqman Genotyping Master Mix (Life Technologies), 0.5 µl of 20X copy number assay (described above), 0.5 µl TaqMan RNAse P Copy Number Reference Assay (Life Technologies, part 4403326), 2 µl water and 2 µl of 5ng/µl genomic DNA. Cycling conditions for the reaction were 95°C for 10 min, followed by 40 cycles of 95°C for 15 sec and 60°C for 1 min. Samples were analyzed using the ViiA™ 7 Real-Time PCR System (Life Technologies) and analyzed using CopyCaller Software (Life Technologies, PN 4412907). Three reference (calibrator) DNA HapMap samples (NA10851, NA15510 and NA07048 (Coriell Institute)) plus one non-template control were included with the test samples.

Chapter 4

Autosome-wide Copy Number Variation Association Analysis for Rheumatoid Arthritis Using the WTCCC High-density SNP Genotype Data

Mohammed Uddin[1], Robert Inman[2], Walter Maksymowych[3], Dafna Gladman[2], Ramin Yazdani[1], Fawnda Pellett[2], Sean Hamilton[1], Darren O'Rielly[1], Proton Rahman[1]

1. Faculty of Medicine, Memorial University of Newfoundland, St. John's, Newfoundland, Canada.
2. Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada.
3. Department of Medicine, University of Alberta, Edmonton, Alberta, Canada.

PREFACE

This manuscript provides an insight into the challenges how to use existing genome-wide association study (GWAS) SNP array data to detect CNVs and highlights the limitation. Using statistical model, this manuscript shows how to infer genomic gains/losses from a large case-control cohort for rheumatoid arthritis (RA). A version of this manuscript has been published in The Journal of Rheumatology.

The co-authors, Drs. Proton Rahman, Michael O. Wodds provided guidance the primary author, Mohammed Uddin, on the study design. Mitch sturge helped setup computational analysis. Mohammed Uddin conducted extensive literature review, the computational analysis and developed the overall concepts and framework. In addition, the primary author demonstrated the existing limitations and challenges in current techniques to detect CNVs sing SNP arrays.

ABSTRACT

Rheumatoid arthritis (RA) is a complex autoimmune rheumatic disease that is strongly influenced by genetic factors. Numerous genes are convincingly associated with RA, including genes in tumor necrosis factor signaling (TNF) and the nuclear factor-κB pathway. To date, except for genes within the HLA region, no data exist regarding potential copy number variations (CNV) involving RA-associated genes. We set out to identify genes affected by CNV that are associated with RA at a genome-wide level. Data from the Wellcome Trust Case Control Consortium (WTCCC) were used in our analyses. The initial WTCCC cohort genotyped 3004 controls and 1999 RA cases using the GeneChip 500k Mapping Array Set. We performed a comparative intensity analysis using the PennCNV algorithm, which uses a hidden Markov model to detect CNV. A total of 2271 controls and 1572 RA samples passed quality control criteria and were included for association analysis. Association analysis was performed in 2 phases: (1) to identify CNV that are < 1 MBp with a population frequency < 5%; and (2) to identify large CNV that are > 1 MBp. Fishers' exact test was performed to quantify significance of the CNV.

We observed that the genome-wide CNV burden is 2-fold higher in patients with RA compared with controls. We identified 11 rare copy number variable regions with < 5% frequency that had an association with RA that reached a $p < 1 \times 10^{-4}$. These include *TNFAIP3* and *TNIP1*, which has been

implicated in association studies for RA, systemic lupus erythematosus, and psoriasis. We identified CNV involving *IRF1*, which functions as a transcription activator of genes induced by interferons; *ALOX5AP* and *LCP2*, involved in inflammatory mediation; *B2M*, an MHC-class I associated gene; and *PRKCH*, a gene involved in T cell signaling pathways. A 57 KBp deletion with 1% frequency in RA cases at 7p21.3 was also observed. Six of these loci overlap with CNV catalogued in the Database of Genomic Variants. This is the first study to identify non-HLA RA-associated CNV using genome-wide analyses. Validation and functional significance of these deletions/duplications in RA and other autoimmune diseases need to be further investigated.

4.1 INTRODUCTION

Rheumatoid arthritis (RA) is the prototypic seropositive inflammatory arthritis. The heritability of RA is between 50% and 60%, with HLA-DRB1 accounting for 30% of the genetic risk of developing RA (Wordsworth and Bell, 1992). The RA-associated genes can be broadly characterized as being involved in T cell activation (*PTPN22, STAT4*, and *CTLA4*) and the nuclear factor- κB (NF-κB) signaling pathway (*CD40*, *TRAF1, TNFSF14*, and *TNFAIP3*) (Raychaudhuri, et al. 2009). Recently, structural variation of DNA, namely copy number variation (CNV), has been recognized as important for both normal genomic variability and in disease susceptibility. A CNV is a common type of genomic variation ranging in size from 1 kilobase (KBp) to several megabases (MBp) and they cover at least 25% of the human genome (Kidd, et al. 2008). Among different individuals, a single CNV can have different forms (i.e., duplication, deletion). The variable region that contains these different CNV is known as the copy number variation region (CNVR). CNV have the potential to disrupt genes and are associated with many complex diseases, including psoriasis, systemic lupus erythematosus (SLE), Crohn's disease, autism, and osteoporosis (Hollox et al. 2008; Conrad et al. 2010b; Fanciulli et al. 2007; McCarroll et al. 2008; Fernandez et al. 2010; Yang et al. 2008) . Thus, it is important to assess the influence of CNV to obtain a thorough understanding of genetic susceptibility in complex disease.

Previous CNV studies in RA have noted association with chemokine ligand 3-like 1 (*CCL3L1*) and Fc fragment of IgG, low affinity IIIb, receptor (*FCGR3B*). *CCL3L1* binds to several proinflammatory cytokine receptors, including chemokine receptor 5 (*CCR5*). In a recent New Zealand study, a copy number higher than 2 at the *CCL3L1* locus was a risk factor for RA, but this was not replicated in a smaller UK cohort (McKinney et al. 2008). Two independent Dutch studies have reported that the 1q23 region, containing *FCGR3B*, has been associated with CNV that influence susceptibility to RA (McKinney et al. 2010; Thabet et al. 2009). However, no such association was noted in the UK cohort1, nor in the recent WTCCC CNV analysis (The Wellcome Trust Case Control Consortium, 2010). We performed a comparative intensity analysis to detect CNV using the WTCCC single-nucleotide polymorphism (SNP) genotype array data. In addition, we performed a CNV genome-wide association analysis to identify disease-susceptible loci in the WTCCC RA cohort.

4.2 RESULTS

Genome-wide CNV distribution: High-resolution SNP obtained from the WTCCC data center had an average call rate of 99.63%, demonstrating high quality intensity data. The log R spread for the controls and cases showed that ~90% of the samples were within mean standard deviation of 0.15 to 0.22 in controls and 0.13 to 0.20 for the RA cases. After quality control, we identified 5927 CNV in RA cases and 4333 in WTCCC controls that were at least 1 KBp in length. There was almost a 2-fold increase in the total CNV burden of 1 KBp or more in size among RA patients compared to controls. The excessive burden of CNV compared to controls has also been noted in patients with schizophrenia and autism (Dalila et al. 2010; The International Schizophrenia Consortium, 2008). However, without experimental validation of CNV it is difficult to exclude the possibility that this may be due to an experimental artefact. Most of these CNV were found in fewer than 3 samples and thus were omitted from the association analysis. Fifty-four percent of the CNV detected in this study overlapped with CNV in the DGV.

Association analysis

Our association analysis revealed significance for the HLA region chr6 30583394–30889981 ($p < 5 \times 10^{-7}$), which overlaps with DGV ID 3600. There were 11 CNV that achieved significance ($p < 1 \times 10^{-4}$), shown in Table 4.1. The association analysis showed mostly one-sided CNV occurrences (in cases more than in controls). The disease-associated CNV involved numerous genes, including the interferon regulatory factor 1 (*IRF1*) gene; the tumor necrosis factor (TNF)-induced gene *TNFAIP3* and its interacting protein 1 gene TNIP1; 2 autoimmune related genes, lymphocyte cytosolic protein 2 (*LCP2*) and beta-2-micro - globulin (*B2M*); 2 genes, protein kinase C-eta (*PRKCH*) and phosphatidylinositol-3,4,5-trisphosphate-dependent rac exchange factor 1 (*PREX1*), involved in cell-signaling pathways; 2 known inflammatory mediator genes, arachidonate 5-lipoxygenase-activating protein (*ALOX5AP*) and lipopoly - saccharide-induced TNF factor (*LITAF*); and the cell proliferation gene serglycin (*SRGN*). One CNVR contained deletions located in an intergenic region on chromosome 7p21.3, with the nearest gene, thrombospondin type-1 domain-containing protein 7A (THSD7A), located approximately 50 KBp away from the deletion.

Table 4.1 The association results of the most significant CNVRs. The listed genes are contained within, or overlap, with CNV breakpoints. All CNV positions are given in NCBI build 35 coordinates.

Genes	**Chr**	**CNVR Start-End Position**	**Deletion**		**Duplication**		**Fisher Exact *P***	**DGV Id**
			Case	**Control**	**Case**	**Control**		
ALOX5AP	13q12.3	30136207-30264780	0	0	35	2	5.06×10^{-12}	
SRGN	10q22.1	70423581-70542187	6	1	21	1	4.35×10^{-9}	48614
Gene Desert (nearby gene THSD7A)	7p21.3	11696007-11753538	20	0	0	0	1.60×10^{-8}	
LCP2	5q35.1	169605980-169668498	0	0	24	2	5.26×10^{-8}	
PREX1	20q13.13	46676571-46882578	0	0	18	0	9.70×10^{-8}	12195
B2M	15q21.1	42606873-42824209	0	0	26	4	2.71×10^{-7}	3961
LITAF	16p13.13	11506474-11695215	1	0	15	0	5.87×10^{-7}	
PRKCH	14q23.1	60986674-61130641	0	0	20	2	1.39×10^{-6}	49385
IRF1	5q31.1	131735192-131863294	2	0	18	2	1.39×10^{-6}	
TNFAIP3	6q23.3	138064233-138239517	0	0	11	0	5.25×10^{-5}	
TNIP1	5q23.3	150379418-150470998	2	0	18	2	7.59×10^{-5}	6455

The association analysis of large CNV with RA (Table 4.2) revealed a duplication on chromosome 16p11.2 that was 1.46 MBp (*p* < 0.001). Duplications and deletions on 16p11.2 have previously been reported and validated in 2 independent studies (Need et al. 2009; Sebat et al. 2004). However, our data confirmation revealed there was very low probe intensity in this region and therefore this result should be interpreted with caution.

Table 4.2 Association analysis of rare large CNVs that are greater than 1 MBp.

CNVR	CNVR Start-End	Case CNV	Control CNV	Fisher Exact Test
16p11.2	29053928 - 30514652	7	0	0.001
22q11.21	17275227 - 18602641	0	7	0.02

4.3 DISCUSSION

Copy number variation may play a significant role in genetic susceptibility and expression of complex autoimmune disease. An example of this is the role of CNV in beta defensins in Crohn disease and psoriasis. Beta-defensins are small antimicrobial peptides that play an important role in the innate immune system. A cluster of beta-defensin genes contain an extensive array of CNV in humans. It has been demonstrated that individuals with Crohn's disease have significantly lower beta-defensin copy number than that in controls. It is hypothesized that this results in the reduction of the antimicrobial barrier in the gut leading to Crohn disease (Fellermann et al. 2006). Meanwhile, individuals with psoriasis have significantly higher copy numbers than controls. The association analysis suggests 6 or more copies of betadefensin is correlated with susceptibility to psoriasis. Due to the cytokine-like properties of the beta-defensin this could lead to an inappropriate inflammatory response generating psoriatic lesions after minor injury, infection, or other environmental triggers (Hollox et al. 2008).

In our study assessing CNV in RA by analyzing a genome-wide association study, we first report the greater genome-wide burden of CNV in RA patients compared with controls. We found an approximately 2-fold increase in the number of CNV carried by an individual with RA. We also identified 11 rare CNVR, with < 5% frequency that showed evidence of association with RA. Some of the loci identified encoded genes within the

CNVR previously implicated in autoimmune diseases. The 2 most interesting candidates are *TNFAIP3* and *TNIP1*. The products of the *TNFAIP3* and *TNIP1* genes are A20 and ABIN1/Nafla, respectively. These proteins physically interact with each other to influence the ubiquitin-mediated destruction of IKKg (IkB kinase-g)/NEMO, which is an essential nexus of NF-κB signaling (Elder et al. 2009). A20 also regulates the degradation of several other components of the TNF signaling pathway. Polymorphisms of *TNFAIP3* have been associated with RA, SLE, and psoriasis, although the specific variants differ among these auto - immune diseases (Plenge et al. 2007; Graham et al. 2008; Nair et al. 2009; Turer et al. 2008). *TNFAIP3* knockout mice develop multiorgan inflammation and arthritis (Plenge et al. 2007). As well, gene-disruptive CNV in classical Hodgkin's lymphoma are found in *TNFAIP3*. This may be of potential interest as there is increased incidence of lymphoma among patients with moderate to severe RA.

Another candidate of potential interest is interferon regulatory factor 1. Type I interferons are a family of cytokines typically produced during viral infection but their multiple immune-modulatory effects are increasingly being recognized, including up regulation of innate immune receptors, polarization of T cells towards a TH1 phenotype, and activation of B cells. Recent studies provide strong evidence for an association between *IRF-5* gene variants and RA, particularly for patients with RA who are negative for anti-cyclic citrullinated peptide (Sigurdsson et al. 2007). An insertion-deletion polymorphism in the *IRF-5* gene has also been shown to confer risk to inflammatory bowel disease (Dideberg et al. 2007). The most significant CNV detected in our study overlaps with *ALOX5AP*, also known as *FLAP*. Expression of *ALOX5AP* is regulated by the binding of TNF-α to the promoter of *ALOX5AP*, thus modulating its gene expression (Reddy et al. 2003).

In a functional study using collagen-induced arthritis in the DBA/1 mouse, it was shown that the severity of arthritis in *FLAP*-deficient mice was substantially reduced (Griffiths et al. 1997). This implies the potential importance of *ALOX5AP* in arthritis. As well, *LITAF* showed upregulation in mice with glu - cose-6-phosphate isomerase (GPI)-induced arthritis (Inoue et al. 2009). CNV involving 2 other genes reported here, *PRKCH* and *PREX1*, have also been observed in acute myeloid leukemia and gastric cancer cell line, respectively (Bullinger et al. 2010; Takada et al. 2005). *PRKCH* was reported to be associated with RA in a Japanese population (Takata et al. 2007). *LCP2* and B2M are 2 additional autoimmune-related genes (Table 4.1). *LCP2* is important for normal T cell development and *B2M* is an MHC class I associated gene that is associated with Spondylitis (Clements et al. 1998). The CNVR at locus 7p21.3 is of particular interest. First, it contains deletions rather than duplications, and, in general, deletions are known to have higher phenotypic consequences (or penetrance)5. Second, this deletion occurs in at least 1% of RA cases and it is a validated deletion annotated in DGV (ID 3665). Finally, in a recent study, it was shown that CNV can have long-range effects (up to 1 MBp) on gene expression (Henrichsen et al. 2009); therefore, validation of the detected CNVR in the intergenic region near *THSD7A* on 7p21.3 is warranted.

Variation in copy number of *FCGR3B* was shown to influence susceptibility to RA in 2 independent studies in a Dutch population (Takata et al. 2007; Clements et al. 1998). However, in our analysis, no evidence of association was observed due to poor coverage of the regions [*FCGR1A* (chr1: 1146567361–146577146) and *FCGR-A/2A/2B/2C/3A/3B* (chr1: 158,288,275–158,382,415)]. Limitations of our study include the issue of multiple testing and possibly under-estimating the number of common CNV due to inaccurate measurement of CNV breakpoints using SNP microarray data. Technological developments in sequencing will allow efficient identification of these breakpoints. Although we do not know the number of CNV in the human genome, we have done an exploratory analysis detecting disease-associated CNV that are $p < 1.0 \times 10^{-4}$. The false-positive error in the detection of duplications is significantly higher in Affymetrix arrays than in Illumina arrays38. Hence, we used a more restricted LRR standard deviation in our analysis than that used in the previously reported Illumina array CNV analysis using PennCNV (Need et al. 2009). Default parameters were used for the BAF drift and wave factor adjustment. As in a previous study, for the SNP intensity parameter, we have used the criteria where intensities of at least 15 SNP are considered for CNV < 1 MBp and intensities of 50 SNP are considered for CNV > 1 MBp.

In a recent CNV analysis carried out by the WTCCC, there were no associations of 3432 common CNV with common diseases, including RA. That analysis used the same cohort as in our study and a custom-designed Agilent Comparative Genomic Hybridization (CGH) chip for CNV detection, which is much more accurate than using SNP arrays for identifying CNV. However, even with CGH they identified a 15% false-positive rate when detecting duplications. Among the true-positive CNV, 50% had a frequency ≥ 5%, all were > 500 bp in length, and most were tagged by nearby SNP (The Wellcome Trust Case Control Consortium, 2010). Similarly, our analyses also showed no strong associations of common CNV with RA other than those within the HLA region. One possible reason for this observation is that the probes for Affymetrix 500k are not uniformly distributed across the genome (median 2.5 KBp space between SNP), hence the poor coverage affects the detection capacity and the frequency of CNV (Carter et al. 2007).

In summary, we are at the early stages of identification of CNV involved in inflammatory rheumatic diseases. Our study identified 11 rare CNVR associated with RA, and these now need to be verified by additional molecular studies. The functional significance of the validated CNV would then need to be elucidated. We suggest that CNV should be routinely analyzed during a genome-wide association study, which may reveal additional alleles with low to modest disease risk associated with RA.

4.4 MATHERIALS AND METHODS

Cohort

The initial cohort data consist of 3004 shared controls and 1999 individuals with RA obtained from the WTCCC data center. The controls were taken from 2 groups: 1504 samples from the 1958 British Birth Cohort (58C) and 1500 additional controls recruited from the UK Blood Services (UKBS). A set of 500,568 SNP were genotyped using Affymetrix GeneChip 500K Mapping Array Set (Affymetrix Inc., Santa Clara, CA,USA) (The Wellcome Trust Case Control Consortium, 2007).

Quality control

We used PennCNV software to detect high-resolution copy number variation. PennCNV uses a hidden Markov model to detect kilobase resolution detection of CNV with low false positives (Wang et al. 2007). Prior to the analysis, the sample inclusion/exclusion criteria were set based on WTCCC SNP analysis. All subjects that passed quality control criteria of the WTCCC original SNP association analysis were included for CNV analysis. Samples were excluded after a check for contamination, false identity, relatedness, and non-Caucasian ancestry. Affymetrix Power Tool, a BRLMM (Bayesian robust linear model with Mahalanobis distance classifier) algorithm, was used to make genotype calls for NSP and STY arrays separately (Rabbee et al. 2006). The first array uses the NSP I restriction enzyme to genotype 250K

SNP while the second array uses STY I restriction enzyme to genotype 250K SNP. To reduce signal-to-noise ratio we used the WTCCC data to model parameters for canonical genotype clustering, instead of using default canonical genotype clustering information provided by PennCNV, which is used to calculate log R ratio (LRR) and B allele frequency (BAF) values. After separate calculations of LRR and BAF, the 2 array sets, NSP and STY, were combined to make CNV detection calls in the autosomes. Many of the genomic samples can have below-optimum genomic wave quality control values; hence, in our sample inclusion criteria we used the default parameters of PennCNV to restrict case or control samples — LRR standard deviation < 0.25, BAF drift > 0.01, and wave factor > 0.05 or < –0.05. A total of 2271 controls (1123 58C and 1148 UKBS) and 1572 RA samples passed all the quality control criteria and were included for association analysis (supplementary 3, chapter 4).

After exclusion of samples, we performed quality control on CNV calls. PennCNV tends to show false-positive CNV calls on centromeric and telomeric regions. Due to the complex nature of hemizygosity in sex chromosomes, we excluded all CNV detected from sex chromosomes. In addition, human immunoglobulin coding regions showed false-positive CNV calls using PennCNV, therefore we excluded CNV that overlapped at 50% or more of its length with the following immunoglobulin regions: chr2:88.9–89.4 MBp, chr14: 21.1–22.0 MBp, chr15: 17.0–21.0 MBp, chr16:31.8–36.8 MBp, and chr22: 20.7–21.5 MBp. Also, large CNV that overlap with centromeres and telomeres were excluded from the following regions: chr1: 12.0–14.1 MBp, chr10: 46.3–47.1 MBp, chr14: 19.2–19.4 MBp, chr15: 18.4–20.0 MBp, and chr19: 24.2–32.7 MBp (build 35). The boundaries for immunoglobulin, centromere, and telomere regions were obtained from a previous study that used PennCNV (Need et al. 2009).

Association analysis

Association analysis was performed on the detected CNV that were < 1 MBp, and Fisher's exact test was performed to quantify the significance of the CNV with exact breakpoint match or within a CNVR. Association analysis was considered to be suggestively significant if (1) $p < 1.0 \times 10^{-4}$; (2) there were at least 15 SNP within the CNV; and (3) the population frequency was < 5%. To avoid possible plate/batch effects, we manually checked the significant CNV samples and ignored any association if a CNV was identified in > 50% of samples from the same plate/batch. We also excluded CNV with strong *p* values if the LRR standard deviation for that CNV call was between 0.20 and 0.25 to avoid borderline CNV calls (supplementary 3, chapter 4). A second association analysis was performed on large CNV that were > 1 MBp. Large CNVs are not frequent, but any CNV in fewer than 3 samples were ignored. Fisher's exact test was performed to quantify the significance of the CNV. The cutoff value for significance at $p < 1 \times 10^{-4}$ was arbitrary, since we are not aware of the total number of CNV in the human genome. However, this level of significance was also used in the recent CNV analysis by WTCCC that introduced the population frequency threshold for rare CNV of < 5%. The

Database of Genomic Variants (DGV) includes CNV that have been annotated. The detected CNV in our case and control samples were compared with the DGV (Build 35).

Chapter 5

UGT2B17 Copy Number Association with Ankylosing Spondylitis (AS)

Mohammed Uddin[1], Robert Inman[2], Walter Maksymowych[3], Dafna Gladman[2], Ramin Yazdani[1], Fawnda Pellett[2], Sean Hamilton[1], Darren O'Rielly[1], Proton Rahman[1,*]

1. Faculty of Medicine, Memorial University of Newfoundland, St. John's, Newfoundland, Canada.
2. Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada.
3. Department of Medicine, University of Alberta, Edmonton, Alberta, Canada.

PREFACE

The previous chapter shows the challenges that exist in using SNP microarrays in the detection of CNVs for common complex diseases. In this manuscript, we have applied our custom design microarray (see chapter 3 method section) in a well defined multiplex family with Ankylosing Spondylitis (AS). The findings of a proposed candidate gene *UGT2B17* from the initial analysis from the multiplex family and subsequently analyzed in two Canadian AS cohorts. This manuscript illustrates the important correlation of genomic rearrangement hotspots detected in chapter 2 with human complex diseases. A version of this manuscript has been submitted to *BMC Genetics* Journal.

The corresponding author, Dr. Proton Rahman, is a rheumatologists, provided guidance to the primary author, Mohammed Uddin, on the study design. The co-authors, critically reviewed the method and results of this manuscript.

ABSTRACT

Ankylosing spondylitis (AS) is the prototypic spondyloarthritis with strong genetic predisposition. The genetic basis for AS is yet to be elucidated. The primary objective of this study is to identify novel copy number variations (CNVs) associated with familial ankylosing spondylitis (AS). A custom genome-wide microarray was designed to detect CNVs and applied to our multiplex AS family with six (6) out of ten (10) affected family members. CNVs were detected using the built-in DNA analytics aberration detection method-2 (ADM-2) algorithm. Gene enrichment analysis was performed to observe the segregation. Subsequent validation was performed using real time quantitative fluorescence polymerase reaction (QF-PCR). The frequency of copy number variation for the *UGT2B17* gene was then performed on two (2) separate well-defined AS cohorts. Fisher exact test was performed to quantify the association.

Our family-based analysis revealed ten (10) gene-enriched CNVs that segregate with all six (6) family members affected with AS. Based on the proposed function of the *UGT2B17* gene, the *UGT2B17* gene CNV was selected for validated using real time QF-PCR with 100% concordance. The frequency of two (2) copies of the *UGT2B17* gene CNV was 0.41 in the Newfoundland AS cases and 0.35 in the Newfoundland controls (OR = 1.26(0.99-1.59); $p < 0.05$)), whereas the frequency of two (2) copies of the *UGT2B17* gene CNV was 0.40 in the Alberta AS cases and 0.39 in the Alberta controls (OR = 1.05(95% CI: 0.83-1.33); $p < 0.71$)). In conclusion, a genome-wide microarray interrogation of a

large multiplex AS family revealed segregation of the *UGT2B17* gene CNV among all affected family members. The association of the *UGT2B17* CNV with AS is particularly interesting given the recent association of this CNV with osteoporosis and the proposed function as it encodes a key enzyme that inhibits androgens. However, two copies of the *UGT2B17* gene CNV was only marginally significant in a uniplex AS cohort from Newfounland but excess copies was not detected in AS cohort from Alberta.

5.1 INTRODUCTION

Ankylosing spondylitis (AS) is the prototypic spondyloarthritis as it is a chronic inflammatory axial disease with a peak onset between 20 to 30 years (Brown et al. 2000; Brown et al. 1997). Extra-articular features of AS include inflammation of the eyes, skin, bowels, and more rarely the lungs and heart. There is a male preponderance as men are affected 2 to 3 times more frequently than females (Calin et al. 1983; van der Linden et al. 1983).

Genetic factors are of major importance in the susceptibility of AS and, in fact, genetic epidemiological studies suggest that AS represents one of the most heritable complex autoimmune diseases with a heritability greater than 90% and a sibling recurrence ratio of at least 52 (Brown et al. 2000; Calin et al. 1983; van der Linden et al. 1983). *HLA-B27*, which was first recognized in 1973, remains the strongest genetic association signal with AS and it is estimated that *HLA-B27* accounts for 23% of the genetic heritability. Despite this strong association, only a fraction (1-5%) of *HLA-B27* positive individuals develop AS (Brown et al. 2000; Calin et al. 1983; van der Linden et al. 1983). Recent genome-wide association studies (GWAS) in several European populations and in the Hans Chinese population have identified up to 15 high priority genes including *IL23R, RUNX3, KIF21B, 2p15, IL1R2, PTGER4, ERAP1, IL12B, CARD9, TNFR1/ LTBR, TBKBP1* (Evans et al. 2011). These associations were primarily reported based on single nucleotide polymorphisms (SNPs) analysis. Recently, one association study revealed the complex interactions between the non-HLA gene, *ERAP1*, with the *HLA-B27* gene (Evans et al. 2011; Australo-Anglo-American Spondyloarthritis Consortium (TASC) et al. 2010). However, the genetic risk describe by HLA or non-HLA genes suggests that other genomic variants may contribute to the risk factor for AS.

That CNVs contribute to disease pathogenesis is supported by their capacity to disrupt gene expression and to interrupt functional pathways (Cooper et al. 2011; Conrad et al. 2010a,b). In recent years, many CNVs have been associated with complex diseases and the majority of these associations were in autoimmune-mediated diseases, including Rheumatoid Arthritis, Psoriasis, Crohn's disease, and Systemic Lupus Erithmatosis (Conrad et al. 2010b; Hollox et al. 2008; Fellermann et al. 2006; Molokhia et al. 2011). Here, we report a CNV associated specifically with AS using a custom genome-wide microarray using a well-defined multiplex AS family.

5.2 RESULTS

Multiplex family

The quality control (QC) measures for the custom microarray chip were very good as all samples exhibited <0.25 DLRS. Approximately 1700 CNVs were detected in each individual sample using our custom microarray. Segregated gene-centric CNV analysis (i.e., a CNV that consists or overlaps with a gene) revealed that 56 CNVs are enriched in affected family members (at least three) and absent in the unaffected family members. Among these 56 CNVs, we have identified ten (10) gene copy number variation regions (CNVRs) that segregate with at least six (6) affected family members (supplementary 4, chapter 5). Our microarray analysis revealed multiple duplications within the *UGT2B17* gene region that segregate in the affected family members that is absent in unaffected family members. One 7kb duplication (chr4:69423002-69430016) disrupts exon 3 of the *UGT2B17* gene and segregates with six (6) AS affected family members (Figure 5.1). The breakpoint encompassing the *UGT2B17* gene region was covered with probes with 280bp spacing, providing high resolution to detect genomic aberrations. Validation of this targeted region using real time QF-PCR demonstrated 100% concordance with the microarray analysis (Figure 5.2). The real time QF-PCR CNV call (>99% confidence) revealed that affected family members carried two (2) copies of the *UGT2B17* gene, whereas unaffected members carried a single copy. All individual with AS family members were HLA-B27 positive and all six (6) affected family members carried two (2) copies of the *UGT2B17* gene.

Disease prevalence (case control studies)

In the case-control cohorts, real time QF-PCR analysis targeting the identical CNVR (from family analysis) revealed the presence of three (3) alleles (0 copy - homozygous deletion, 1 copy - heterozygous deletion, and 2 copies) within this population for the *UGT2B17* gene (supplementary 5, chapter 5).

In the Newfoundland population, the frequency of the three alleles was 0.13, 0.51, and 0.35 for 0, 1 and 2 copies in controls, respectively. The frequency of two (2) copies of the *UGT2B17* gene CNV was 0.41 in the Newfoundland AS cases and 0.35 in the Newfoundland controls (OR = 1.26 (95% CI: 0.99-1.59); p < 0.05). Similarly, the frequency of two (2) copies of the *UGT2B17* gene in the Alberta population was 0.40 and 0.39 for AS cases and controls, respectively (OR = 1.05 (95% CI: 0.83-1.33); *p* < 0.71). The association analysis for the combined cohort (Newfoundland and Alberta population) show two (2) copies of *UGT2B17* gene frequency 0.40 and 0.37, for the cases and controls, respectively, with association significance *p* < 0.09 (Table 5.1).

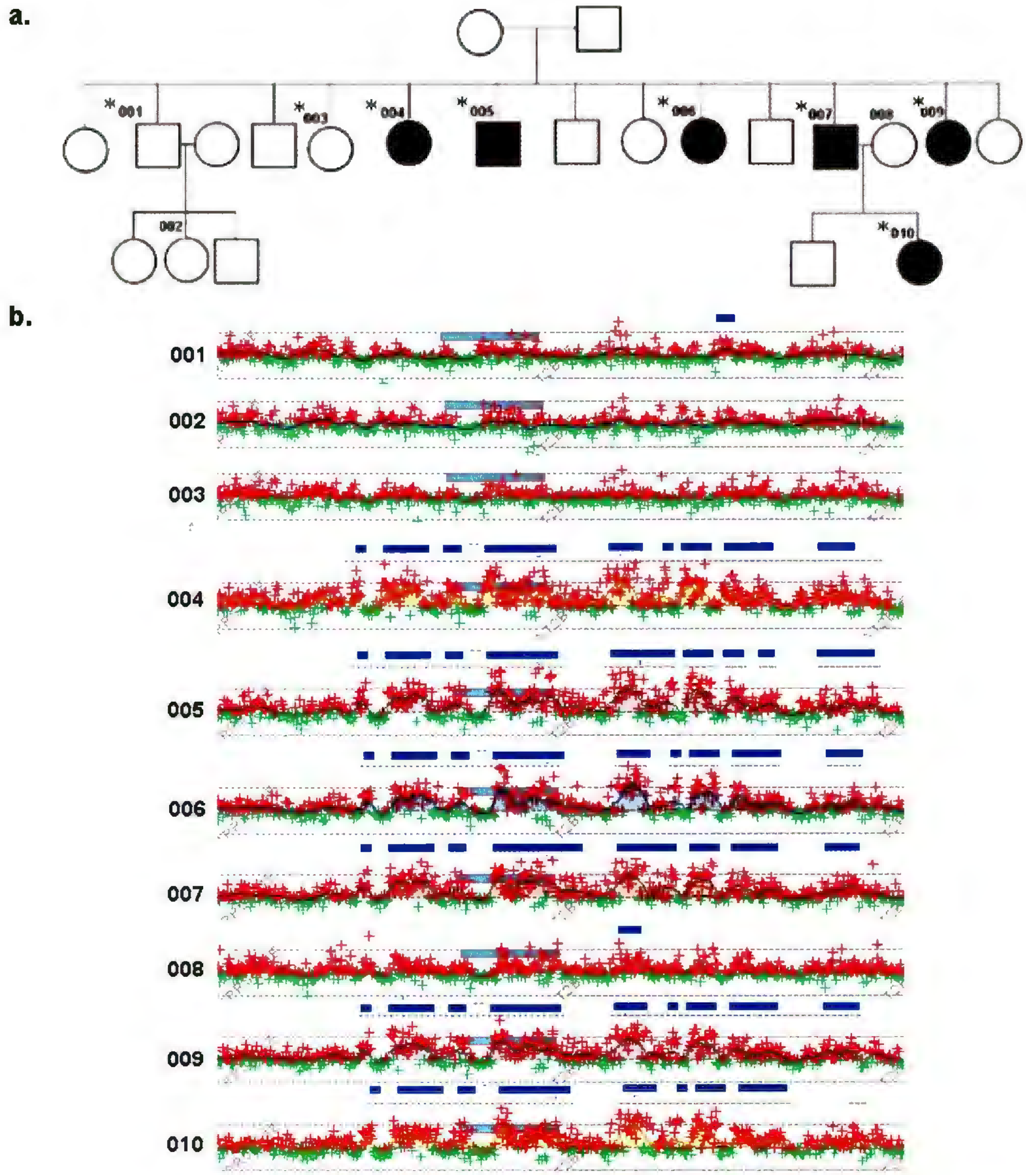


Figure 5.1 a) A three generational pedigree comprising 13 individuals. DNA was available only for the individuals with an identification number (Ids). The affected family members are indicated by solid colors and member 002 has systemic lupus erithmatosis (SLE). **b)** The microarray intensity values for the *UGT2B17* region are indicated below the pedigree for each of the family members. The horizontal dark blue rectangles represent duplications detected by the genome-wide analysis of CNVs.

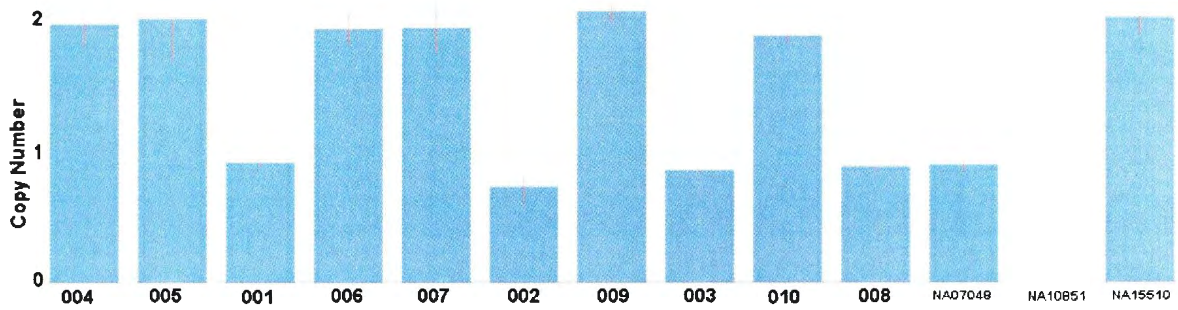


Figure 5.2 *UGT2B17* gene copy number validation using real time rt-PCR on the family members and three reference samples from HapMap.

Table 5.1 *UGT2B17* gene CNV association. The association result for Newfoundland, Alberta and the combined cohort. The significance is observed with Newfoundland population with the higher copy (2 copy) number of *UGT2B17* between AS cases and controls.

Population	Copy	Cases	Controls	OR (95% CI)	*P*
Newfoundland (298 cases, 299 controls)	2	0.41	0.35	1.26(0.99-1.59)	<0.05
Alberta (289 cases, 285 controls)	2	0.40	0.39	1.05(0.83-1.33)	<0.71
Combined (587 cases, 584 controls)	2	0.40	0.37	1.15(0.97-1.36)	<0.09

5.3 DISCUSSIONS

Copy number variations are increasingly being recognized in complex autoimmune diseases as they are capable of altering gene dosage and consequently affect gene function (Henrichsen et al. 2009). The contribution of CNVs to AS pathogenesis remains to be systematically evaluated. The primary goal of this study was to identify CNVs segregating within a multiplex AS family and then subsequently determine the allele frequencies of CNV in a case-control cohorts. The genome-wide CNV analysis performed in this study revealed that increased copy number of the *UGT2B17* gene is a risk factor for AS.

The *UGT2B17* gene represents an excellent candidate gene for susceptibility to AS. *UGT2B17* is a functional gene as it encodes an enzyme that metabolizes steroid hormones including testosterone and selected xenobiotics (Xue et al. 2008; Yang et al. 2008; Karypidis et al. 2008; Olsson et al. 2008). A gene-dosage effect is present as the CNV is associated with urine testosterone level, male insulin sensitivity, fat mass, and prostate-cancer risk (as summarized by Xue et al. 2008). Importantly, copies of the *UGT2B17* gene has recently been associated with osteoporosis (including hip fracture) (Yang et al. 2008). The dose of the *UGT2B17* gene was significantly associated with bone mineral density, cortical thickness, and buckling ratio. These results support the role of *UGT2B17* CNVs in the pathogenesis of osteoporosis. Although AS is characterized by new bone formation, osteoporosis is a well recognized complication of AS and the risk of clinical vertebral fractures is increased in AS patients (Cooper et al. 1994; Vosse et al. 2009).

The deletion polymorphism identified as a risk factor for prostate cancer as it exposes androgen in prostate, whereas individuals with two copy of *UGT2B17* gene is shown to have significantly lower risk of prostate cancer in a well defined Caucasian population from Sweden (Karypidis et al. 2008). The replication of this association failed in a large cohort from the original population. The deletion (0 copy) frequency within the controls of the Swedish prostate cancer study is 4% whereas, the large replication cohort reported frequency is consistent with most European population (Olsson et al. 2008). Replication in other populations on the association of the *UGT2B17* gene in large cohorts is required for both prostate cancer and osteoporosis. Particularly for this gene, failed replication does not undermine the original report of association due to its dynamic nature with respect to population stratification. From an evolutionary perspective, the *UGT2B17* gene in humans is highly stratified which makes it more likely for it to be associated with disease. The deletion allele (or 0 copies) frequency varies significantly among populations. For example, the frequency of the deletion allele is approximately 0.15 in Europeans, ~0.22 in Africans, and ~0.80-0.90 in Asians (Conrad et al. 2010b; Xue et al. 2008). In contrast, the frequency of two (2) copies of the *UGT2B17* gene is rare in Asian populations (~0.01-0.03), ~0.30-0.40 (reports with varying frequencies) in Europeans, and ~0.60 in African populations.

In this study, two (2) copies of the *UGT2B17* gene segregates within a large well-defined AS multiplex family. Given the nature and proposed function of the *UGT2B17* gene (Schulze et al. 2008; Clarke et al. 2009), the *UGT2B17* CNV may account for the susceptibility of AS within this family. Conflicting results were obtained regarding the excess prevalence of this CNV in large uniplex AS cohorts. The population from Newfoundland suggests a trend, however this was not replicated in the Alberta cohort. In this study, the deletion (0 copy) frequency of the *UGT2B17* gene in the Newfoundland population is consistent with the Alberta population. There was a noticeable frequency difference between the Alberta (0.51) and Newfoundland (0.46) populations for a single copy of the *UGT2B17* gene. This stratification may represent a contributing factor for the absence of association.

The analysis of CNVs within multiplex families provides a unique opportunity to detect variants that can be family specific. There are ample reports in psychiatric diseases where family specific CNVs manifest with the disease (Cook et al. 2008). Although it is a daunting task to obtain more multi-generational pedigrees to follow up the segregation of this CNV, the trend observed within the investigated family suggests that segregation analysis into multiple multi-generation pedigrees is promising. The microarray analysis revealed paternal inheritance of the CNV within the sub-family (samples 007, 008, 010). While the multiplex family in this study is a unique family with high penetrance of AS, it is important to assess other large multiplex families to determine if this CNV may be a causal in familial AS.

5.4 MATERIALS AND METHODS

Multiplex AS Family

A multiplex AS family from Newfoundland was identified for this study. All members of the multiplex family were "native Newfoundlanders" of North European ancestry. Ten (10) members from this large family participated in the study. Each individual was assessed clinically including imaging studies. Six (6) of the ten (10) family members had AS as defined by the 1984 modified New York criteria, while the remaining family members were not affected. After consent was obtained, all participating family members had their DNA collected.

Case-control Cohorts

DNA was obtained from two (2) well-defined AS cohorts from the Newfoundland and Alberta populations. All cases and controls were of North European ancestry. We assessed 298 AS cases satisfying the modified New York criteria and 299 ethnically-matched controls from a homogenous population of Newfoundland. The second cohort consisted of 289 AS cases satisfying the modified New York criteria and 285 ethnically-matched controls from Alberta.

Custom Microarray

We designed a custom genome-wide microarray based on genomic hotspot breakpoints previously identified (Uddin et al. 2011) (see Chapter 3 Method section and supplementary 2, chapter 2). The custom microarray consisted of 2 X 1 million probes covering the genome with a mean spacing of

280bp. Prior to CNV analysis, QC measures were applied and the derivative of log ratio spread (DLRS) <0.25 was considered as threshold. CNVs were detected using the built-in Aberration Detection Method-2 (ADM-2) algorithm DNA Analytics v.4.0.85 (Agilent Technologies) using the following criteria: 1) at least 5 probes for a CNV call on GC-corrected intensity; 2) nested filter was set to 2; and 3) log intensity >0.25 for duplication and <-0.25 for deletion. A custom script was applied to detect gene-enriched CNVs (i.e., overlaps or consist a gene) that segregate (at least three cases) within affected AS family members and absent in the unaffected family members.

rt-PCR Validation

To validate the CNV encompassing the *UGT2B17* (UDP glucuronosyltransferase 2 family, polypeptide B17) gene in the family members and in the case-control cohorts, copy number estimation of the *UGT2B17* gene was performed using the Taqman copy number assay Hs03185327_cn (Life Technologies) using the manufacturer's recommended protocol. The assay was performed in quadruplicate using 10ng genomic DNA for each sample in a 96-well plate. The 10µl reaction mix consisted of 2µl of 2x Taqman Genotyping Master Mix (Life Technologies), 0.5µl of 20X copy number assay (described above), 0.5µl of TaqMan RNAse P Copy Number Reference Assay (Life Technologies, part 4403326), 2µl of water, and 2µl of 5ng/µl genomic DNA. Cycling conditions for the reaction were 95°C for 10 min, followed by 40 cycles of 95°C for 15 sec and 60°C for 1 min. Samples were analyzed using the ViiA™ 7 Real-Time PCR System (Life Technologies) and analyzed using CopyCaller

Software (Life Technologies, PN 4412907). Three reference (calibrator) DNA HapMap samples (NA10851, NA15510 and NA07048 (Coriell Institute)) plus one non-template control were included with the test samples. Fisher exact test was performed to determine association of the *UGT2B17* gene copy number with AS in our case-control data.

Chapter 6

GENERAL DISCUSSION

This chapter comprises four major sections which begin with the background of CNVs (i.e., detection methods and disease implications) in the human genome. The latter two subsequent sections discuss the detection of rearrangement hotspots and the creation of a custom microarray. The final section focuses on the application of the custom microarray in two complex diseases and the detection of novel genetic variants.

The human genome consists of multiple variations that range from 1 base pair (bp) to an entire extra chromosome in length. Primarily, we observe six major sub-microscopic types of variation in the human genome including single nucleotide polymorphisms (SNPs), various repeat elements like microsatellites, LINE, SINE; deletions, duplications, insertions, and inversions. This thesis primarily focuses on sub-microscopic structural variants (primarily genomic gains and losses), as they represent the most recently discovered form of genetic variation in the human genome and their role in complex disease etiology largely remains unknown. Genomic gains and losses usually range from 500bp to mega base-pair (MBp) level. Although, these variants are small in size, their overall potential contribution to human genetic disease is suspected to be high as they occur at a relatively high frequency. Although, the first report of gene copy number variation occurred in 1936 (Bridges 1936), it was not until the last decade that significant technological advancements made capturing genomic aberrations much more efficient and more comprehensive.

For the identification of causal variants, it is imperative to investigate all spectrums of mutations in the human genome. Copy number variants are responsible for a large portion of variability observed between human genomes. Deletions, duplications, insertions and complex rearrangements collectively contribute a significant proportion of the total variation in copy numbers among individuals, thus it is prudent to investigate the detection of such variants in disease susceptibility. Applying numerous approaches in large complex disease cohorts, CNVs have been shown to be associated with numerous disease phenotypes and traits. Although, there has been significant progress in assessing CNV content in the human genome, this class of variation has not been systematically studied in complex human genetic studies. CNVs are very prominent for their role in DNDs. These variants also helped to identify previously unrecognized syndromes (Ballif et al. 2007; Ballif et al. 2008). For example, three independent groups characterized a microdeletion syndrome at 17q21.31 (Shaw-Smith et al. 2006; Sharp et al. 2006; Koolen et al. 2006). On chromosome 15, three novel syndromes associated with mental retardation have been confirmed at 15q13.3, 15q24, and 15q26.2 (Poot et al. 2007; Sharp et al. 2007; Sharp et al. 2008). Collectively, these reports strongly suggest the importance of CNVs in the etiology of DNDs. For common complex diseases, the initial major analysis by the WTCCC on CNV association lead to major scepticism (WTCCC 2010) and SNP-based GWAS remained the primary tool to identify variants susceptible to complex diseases. Although GWAS is a prominent approach to identify disease susceptibility SNPs, they can explain only a fraction of the missing heritability for any given complex disease. Hence, investigating the role of CNVs as potential variants to fill the void with respect to the missing heritability represents a logical step forward. In the last few years, numerous reports (see Chapter 1.5, Table 1.1) have showed CNV associations (i.e. beta-defensin *DEFB* gene CNV with Psoriasis, *FCGR3B* and *CCL3L1* gene CNV with SLE, *CCL3L1* CNV and *LCE3C_LCE3B*-deletion with RA, CNV at locus 5q33.1 with Crohn etc.) with common complex diseases (McKinney et al. 2008; Hollox et al. 2008;Willcocks et al. 2008; Mamtani et al. 2008; Brest et al. 2011). The identification of CNVs in complex diseases is a major improvement considering the missing heritability problem associated with SNP-based GWAS.

Recent advances in the detection of causal variants have been made through technological developments that can pinpoint CNV breakpoints with higher precision and accuracy. The refinement of microarrays and high throughput re-sequencing technologies has enhanced the detection capacity to identify CNVs that are below 1kb in length. Multiple large scale population-based analyses on CNVs have produced higher resolution data and smaller variants have been discovered, along with increased precision to identify breakpoints (McCarroll et al. 2008; Kidd et al. 2008; Korbel 2007; Sudmant et al. 2010; Conrad et al. 2010). The work presented in this thesis provides a technique to narrow the target regions in the genome for better detection of CNVs and to analyze their susceptibility to complex diseases. I have used high throughput re-sequencing technology in combination with a custom-design microarray for the detection of CNVs within rearrangement hotspot regions and investigated their susceptibility in two complex diseases.

The accurate detection of genomic aberrations is essential in measuring the effect of gene dosage on a given phenotype. The major detection methods, limitations within the hybridization-based methods include the inability to ensure (see chapter 1.3) exact copy numbers (especially CNV with higher copy number), duplication orientation, the identity of CNV locus within the genome and genome coverage. However, high throughput sequencing is rapidly improving the investigation of human and population genomics. This technology sheds light on these problems by sequencing the entire genome in parallel. Using such large amounts of data, it is possible to achieve higher precision in the detection of genomic variants.

High throughput sequencing technologies typically produce shorter sequences with higher error rates from relatively short insert libraries. The major technological platform for high throughput sequencing includes Roche 454, Illumina, and SOLiD. Each platform produces reads with varying lengths and an array of softwares are available to map the reads. Most mapping tools are developed to produce information with respect to the reference genome by applying various alignment algorithms. This is due to the lack of technological development to produce genome assembly without the reference genome, also known as *de novo* assembly. Most *de novo* assembly methods for short reads are primarily based on computational approaches known as de Bruijin graph and

Eulerian path which are not efficient to assemble complex regions of the genome (Chaisson et al. 2009; Simpson et al. 2009; Li et al. 2009; Mills et al. 2011; Gnerre et al. 2011; Hajirasouliha et al. 2010). For example, a recent analysis showed *de novo* assemblies were 16.2% shorter than the reference genome and that 420.2 MBp pairs of common repeats and 99.1% of validated duplicated sequences were missing from the genome (Alkan et al. 2011). The *de novo* assembly also showed that over 2,377 coding exons were completely missing from the reference genome. Hence, these methods lack the maturity to apply *de novo* approach for genome assembly with high precision.

The limitations mentioned above for each molecular and computational method underscores the importance to adapt a new approach in the detection of “rearrangement hotspots” that are vulnerable to produce CNVs and SDs. It is also imperative to characterize these complex regions within the context of sequence homology to better understand the underlying mechanisms in the origin of CNVs and SDs. To achieve the three objectives specified in chapter 1.6, this thesis is organized in two major parts – first, the detection of complex genomic regions (i.e., “rearrangement hotspots” – see chapter 2) and secondly, the analysis of these regions in complex diseases to detect novel CNV variants (see chapters 3,4 and 5).

In chapter 2 of this thesis, I applied novel algorithm techniques to identify SDs and gene CNVs using high throughput sequencing technology and subsequently validated them by using molecular experiments. One of the primary objectives of my thesis was to detect and characterize rearrangement hotspots.

In chapter 2, I applied algorithmic techniques that analyze mapped short reads (using reference specific or *de novo* assembly) to detect SD regions and gene CNVs. One of the findings arising from this work was the characterization of rearrangement hotspots within SDs that are prone to produce CNVs. We have processed short read data that was sequenced using reversible terminator chemistry on an Illumina Genome Analyzer. A novel hierarchical approach was applied to the NA18507 human genome (i.e., an Yoruban male) where, 1.5 billion short reads (18x coverage) were processed with 55.78% (i.e., ~839 million short reads) mapped to the repeat masked human reference genome with the mrsFAST aligner, which returned all possible mapping locations of a read, which is a key requirement to accurately predicting the absolute copy number of duplicated sequences (Alkan et al. 2009). After applying median-based GC correction to reduce the known bias and associated amplification artefacts with next generation sequencing technology towards GC-rich and GC-poor regions (Yoon et al. 2009), a mean + 2 S.D. (standard deviation) was applied to the read depth as a threshold for each overlapping window of 100 bp across the human genome in order to assess the breakpoints associated with SD units.

One limitation of mrsFAST includes the inability to generate a consensus genome from short reads. Hence, after the duplicated breakpoints were identified, the short reads were reprocessed and the NA18507 genome was assembled using MAQ by mapping the short reads against the reference (hg18) human genome. Once the consensus sequence was obtained using MAQ, low copy sequences with <90% sequence identity were removed through filtering.

The primary reason for applying MAQ assembly was to obtain all consensus sequences for the breakpoints obtained by read depth-based analysis. When the consensus sequences were obtained for those high read depth breakpoints, a new variant of semi-global algorithm, end space free pair wise alignment coupled with seed and extend technique was implemented to accurately pinpoint duplicon breakpoints (or homologues). The consensus sequences were analyzed to detect rearrangement hotspots by investigating each 100 bp window for possible duplicons with >90% sequence identity. It is noteworthy that both mapping tools applied in this thesis were unable to return information regarding the orientation of the duplicated loci; therefore analysis excluded detection of inversions.

One of the highlights of this study (presented in chapter 2) was the inference of all possible homologues (i.e., intra or inter chromosomal) for each SDs that leads to the characterization of complex genomic regions (see chapter 2). While intra- and inter-chromosomal recombination events cannot be determined by asserting copy number estimation from read depth approaches, an integrating alignment algorithm was described. Also in chapter 2, we determined the breakpoint information required to characterize the complexity of a region in cases of excessive tandem duplication or where duplications within a duplication occurs within a chromosome or between chromosomes. Using 409 million pairwise alignments, the identification of 1963 complex SD units or 'rearrangement hotspots' in the human genome showed an increase in copy number gain (i.e., increase of 62% in copy number gains within hotspots). Further analysis showed that 25% of these 'rearrangement hotspots' (i.e., 489/1963) overlapped with 166 unique genes (Fig. 2b) of which 77% were contained within 82 genes which were previously shown to have an increased copy number gain by using microarray analysis (Chid et al. 2009). Twenty five of these genes are highly variable in copy number within three populations indicating a population-specific frequency of the underlying events in the origin of CNVs (McCarroll et al. 2008) which, in turn, implies an increase in frequency of genomic rearrangement events within hotspot regions.

Chapter 2 also highlighted that genic regions were enriched with intra-chromosomal recombination, whereas agenic regions evolved through both inter- and intra-chromosomal recombination. Such intra-chromosomal recombination within genic SD units may represent conserved genomic organizations subject to gene conversion and concerted evolution (Shevell et al. 2003; VanderWeele et al. 2004; Dawson 2010). Extreme variation, attributed in part, to SDs has been reported in at least 20% of the copy number variable gene families in three human populations (McCarroll et al. 2008). We have produced a genome-wide high resolution map of 'rearrangement hotspots' which likely serve as templates primarily for NAHR mechanism. The characterization of pathogenic regions by our approach (i.e. 16p11-p12, 22q11.21 etc.) and subsequent validation with FISH experiment helps understand the possibility of intra/inter chromosomal NAHR as a disease mechanism within those regions. A collection of 24 previously identified genomic disorders predominantly mediated by *de novo* NAHR events are catalogued in the DECIPHER database (Figure 2.3). Analysis of the detected hotspot regions revealed that 20% of the hotspots were located at the end/start of those 24 pathogenic deletions/duplications breakpoints (see chapter 2). This finding indicates a higher rate of NAHR that produces *de novo* aberrations within the detected genome-wide rearrangement hotspot regions.

One of the limitations of our approach was the exclusion of inversions and insertions into the identification of “rearrangement hotspots”. The method employed a read map algorithm, mrsFAST, which was unable to return information regarding the orientation of duplicated loci. Hence, the hotspot regions which are located within inversions are beyond the scope of the proposed work. The work described in chapter 2 indicates that NAHR represents an underlying mechanism for the creation of hotspots within the SD regions. Hence, this thesis largely excluded hotspot regions that are vulnerable to genomic aberrations through mechanisms other than NAHR (i.e. NHEJ, MMBIR, FoSTeS). Although the aforementioned limitations require further improvement, the overall strength outweighs the limitations and weaknesses. The genome-wide characterization of ‘rearrangement hotspots’ will enhance the clinical applicability of high resolution genome analysis to uncover uncharacterized genomic disorders.

In order to identify these complex regions, applying high throughput sequencing with high coverage will allow the detection of CNVs in a comprehensive manner. A similar approach described in chapter 2 can be applied with high detection accuracy for CNV breakpoints. Although the method showed promising results in a personal genome, it is impractical to apply high throughput sequencing in large cohorts due to the extreme computational burden and high cost associated with this new technology (Alkan et al. 2011). That these limitations associated with CNV detection using high throughput sequencing are tangible is evidenced by fact that the most comprehensive CNV profile published to date using high throughput sequencing analyzed only 159 individuals (Sudmant et al. 2010). The next commonly applied approach towards CNV detection is the use of microarray-based platforms. The most comprehensive CNV catalogue using aCGH was published using 20 genomic tilling arrays consisting of 42 million probes (Conrad et al. 2010). Similar to high throughput sequencing, the cost and time associated with processing and analyzing 20 microarrays per individual is not realistic for large disease cohorts as indicated by the inclusion of only 40 individuals in the study conducted by Conrad et al.

A more practical approach is to design a custom microarray to capture the aberrations within the detected hotpot regions (see chapter 2). There are two primary technologies that can be used – SNP-based microarray or array comparative genomic hybridization (aCGH). In contrast with SNP-based microarray, aCGH is primarily used for clinical characterization of CNVs. Modern molecular cytogenetics has begun to characterize CNVs responsible for mental retardation, autism spectrum disorder, and developmental delay, identify causal genes for known genetic conditions, and define new genomic syndromes on the basis of common genomic aberrations. Shaffer et al. 2008, reported on a series of 8789 clinical cases of children with a variety of developmental problems who were tested with aCGH (Shaffer et al. 2005). The results of that study indicated that 6.9% of children with abnormal results of microarray-based CGH exhibited clinically relevant abnormalities, 1.2% exhibited a polymorphism, and 3.9% manifested changes of unclear clinical significance.

Microarray-based cytogenetics is currently considered to be the first line test for patients with mental retardation, autism spectrum disorder, or developmental delay (Miller et al. 2010). It is crucial to select an optimum platform with highest sensitivity and accuracy. CGH arrays generally show better signal-to-noise ratios compared to SNP-based arrays, probably as a consequence of non-uniform probe distribution on the former platform (Pinto et al. 2011). The number of probes required to detect a single-copy alteration varies between platforms, with Agilent Technologies offering the highest per-probe performance (Coe et al. 2007; Pinto et al. 2011). Moreover, Agilent's *in silico* probe performance score is a meaningful metric for prioritizing probes for inclusion in microarrays. Although alterations can, theoretically, be detected with a single probe using the Agilent platform, a more conservative detection limit of three to five probes is advised (Pinto et al. 2011). There are also examples of studies where the analysis was a combination of array CGH and SNP-based platforms to offer higher confidence in CNV detection (Redon et al. 2006; McCarroll et al. 2008; Winchester et al. 2009). These studies also implicated the need for a better single detection technique that will provide higher precision in CNV detection. We chose Agilent technology and designed a custom microarray which consisting of 2 million probes with an average spacing of 280bp for the hotspot regions (see chapter 3 method section). For algorithmic prediction, it has been shown that platform-specific algorithms perform best in comparison with platform-independent algorithms (Pinto et al. 2011). We used an Agilent platform-specific algorithm (i.e., aberration detection module 2 - ADM-2) to detect CNVs from the custom microarray.

The detected hotspots show convincing evidence that these regions are highly variable and are often found to be associated with disease. Such complex regions often are the harbour ground for *de novo* or atypical aberrations that are frequent in patients with neuropsychiatric disease. The application of the custom microarray described in chapter 3, revealed the detection of novel atypical microduplications within the 2q21.1-21.2 locus for Tourette syndrome. TS is a common complex neuropsychiatric disease and the genetic basis for this disease remains unknown. TS is characterized by involuntary motor and vocal tics. Other conditions (i.e., ADHD, OCD) are often found to be manifesting with TS and the exact cause for such co-morbidity is unknown. The findings in chapter 3 highlight the impact of CNVs with respect to human health and genomic susceptibility to Tourette syndrome. The analysis detected co-occurrences of two atypical microduplication blocks that segregated in a multiplex family.

The co-morbidity is very complex issue for the genetic analysis in any neuropsychiatric disease. The co-morbidity of diseases within Family A showed variability within the affected members. In-depth phenotypic analysis showed three siblings within the family with various features of TS. The oldest siblings demonstrated extreme phenotypic morbidity of TS. The mother and the grandfather also exhibited high functioning TS (Table 3.1). Five of the six affected members within Family A showed segregation of two microduplication blocks.

The custom microarray analysis revealed the presence of a partial duplication of the larger block on the fourth sibling. The large microduplication that segregated within the family is rare within the ethnically-matched control population. This indicates that the 2q21.1-21.2 locus is possibly linked with the TS etiology in this particular family.

The larger microduplication that segregated within the affected individuals of Family A contains the *C2orf27A* gene. There is no report with respect to the function of the *C2orf27A* gene which complicates to hypothesize the exact role of this gene in TS pathogenesis. We have analyzed a well characterized TS cohort which consisted of a multi-generational pedigree, two trios and additional unrelated individuals with a history of TS. Although we have determined the frequency of the larger microduplication block within a large control population, due to the lack of a larger TS sample set, it was impossible to determine the true frequency within TS patients. The rare frequency of this microduplication within the control population is indicative of a possible link between the 2q21.1-21.2 locus and TS pathogenesis. Until a large TS cohort is analyzed, it is not possible to determine the actual frequencies of these microduplications among TS individuals.

Family-specific CNVs are often associated with numerous neuropsychiatric diseases. These variants have the property of manifesting within affected members of a family and are extremely rare or absent within the control and other affected pedigree (Sato et al. 2012). Although the analysis identified the presence of the larger duplication block in an unrelated TS sample, it was absent within the additional two trios and nine other unrelated TS samples. Previously reported neuropsychiatric disease studies using aCGH have showed the absence of these two microduplications (Cooper et al. 2011). However, it is premature to state that this locus is TS-specific until a more comprehensive study has been completed.

In previous reports, several regions showed complex genotype-phenotype correlation where a locus consisted of multiple microduplications/microdeletions reported to be associated with several neuropsychiatric diseases (see Figure 2.5) (Antonacci et al. 2010; Ensenauer et al. 2003; Shaw-Smith et al. 2006). The exact cause for such complex relationships between genomic variants and neuropsychiatric diseases remains unclear. This complexity is also apparent with the identified locus 2q21.1-21.2 being a part of a broader 9MB locus 2q14.3-2q21.3. This 9MB locus was previously implicated in a study where linkage analysis on a multi-generational pedigree with the history of Dystonia showed the high LOD score within this region (Norgren et al. 2011). In a more recent study, multiple patients with developmental delay (DD)/intellectual disability (ID), ADHD, epilepsy were reported to have recurrent microduplications and microdeletions within 2q21.1-21.2 (Dharmadhikari et al. 2012). Our findings of microduplications at 2q21.1-21.2 within a family with a history of TS (and ADHD/OCD) adds to the growing list of genomic variants within that 9MB region. Screening for CNVs within this region (2q14.3-2q21.3) on a large cohort with TS and related conditions is warranted.

We have detected novel microduplications in one of our complex diseases using the custom microarray. The second disease of interest in this thesis is a common complex autoimmune disease, specifically Ankylosing Spondylitis (AS). Common complex diseases have been investigated extensively with SNP-based microarrays for the last decade through numerous genome-wide association studies. The National Human Genome Research Institute (NHGRI) published a catalogue of more than 237 GWAS studies on common complex diseases that produced a huge amount of SNP-based microarray data (Hindorffa et al. 2009). Until very recently, the development of microarrays showed promising results and became the ultimate experimental workhorse for CNV discovery and genotyping. SNP-based microarrays have proved popular in CNV-detection studies, historically as complements to array CGH platforms for fine-mapping regions and currently in the large-scale discovery of CNVs in a broad variety of populations (Itsara et al. 2009; McCarroll et al. 2008; Cooper et al. 2008; Jakobsson et al. 2008; Gusev et al. 2009). Early SNP-based microarrays demonstrated poor coverage of CNV regions and were specifically designed for SNP-based genome wide association studies. CNV detection reusing GWAS data is becoming routine through the development of algorithms that can analyze SNP probe intensities.

In chapter 4, CNV association analysis was performed using SNP probe intensities for rheumatoid arthritis (RA), which is an inflammatory polyarthritis involving many different pathways including aberrations in the innate and the adaptive immune system (Firestein 2003). RA can affect the articular structures particularly small joints of the hand and feet (and also extra articular structures such as rheumatoid nodules) (Isenberg 2004). The WTCCC performed association analysis on a large cohort from a homogenous British population. In that study, the Affymetrix 500K platform was used to identify RA susceptible loci. The use of SNP-based microarrays were demonstrated to be useful in detecting CNVs where there is enough SNP-specific probe coverage to compute hybridization intensities for a CNV call. This suggests that SNP-based microarrays suffer from the lack of a back bone covering the genome with uniform probe spacing. Regions dense with SNP probes will have high coverage which creates the problem of insufficient probe coverage to make a CNV call for genomic loci with low SNP probe density. Probe coverage is strongly correlated with CNV detection resolution and therefore detecting small CNVs with SNP-based microarrays is problematic and often introduces higher false positives. Hence, for complex disease CNV association studies, one of the major limitations includes uniform distribution of probes to form a consistent backbone for genome wide coverage. Our analysis also failed to replicate and validate findings due to the absence of a rheumatoid arthritis cohort. To overcome the limitation of inadequate genome coverage, recent arrays (such as the Affymetrix 6.0 SNP and Illumina 1M platforms) incorporate better SNP selection criteria along with custom selection of probes specific to CNV detection in complex regions of the genome and non-polymorphic copy-number probes (McCarroll et al. 2008; Cooper et al. 2008; Winchester et al. 2009).

The association analysis of the detected CNVs in chapter 4 was performed in two distinct CNV groups with a population frequency of < 5% – CNV that are <

1 MBp; and large CNV that are > 1 MBp. The primary goal of this study was to identify rare CNV variants that are susceptible to RA. The analysis model introduced two CNV groups of varying length. The association for <1 MBp reached significance within the cohort for the following gene CNVs – *IRF1, TNFAIP3, TNIP1, LCP2, B2M, PRKCH, PREX1, ALOX5AP, LITAF* and *SRGN*. Among these genes, *TNFAIP3* and *TNIP1* are implicated in autoimmune pathways and are also associated with RA in GWAS studies. The other two important genes were *ALOX5AP* and *LITAF*; both inflammatory mediator genes. The non-genic CNVs that reached significance included deletions located on chromosome 7p21.3, with the nearest gene (located approximately 50 kb away from the deletion) being thrombospondin type-1 domain-containing protein 7A (*THSD7A*). The association of the second group of CNVs with >1MB length includes 16p11.2 which has frequently been previously associated with neurodevelopment disorders.

A recently validated CNV for rheumatoid arthritis was the deletion of the Late Cornified Envelope *(LCE) 3B* and *3C* genes (*LCE3C_LCE3B*-del) (Lu et al. 2011; Judith et al. 2012). Another non-*HLA* CNV found to be associated with RA is the *FCGR3B* gene copy number variant that is reported to be associated with susceptibility to RA. The presence of inversions/genomic gaps does not necessarily provide the true picture for CNV associations. In the case of the chemokine ligand 3-like 1 (*CCL3L1*) gene copy number association (McKinney,C. 2008), there exists contradicting evidence due to the fact that within this region a large inversion is also present. We were unable to replicate previously implicated

CNVs in RA pathogenesis - *LCE3C_LCE3B*-del and *FCGR3B* which is primarily due to poor coverage of the region. The validation and replication was not performed due to the lack of an additional RA cohort.

As we have discussed the pros and cons of using SNP-based microarrays from existing GWAS data, it is clear that for comprehensive detection of CNVs, a uniform and targeted array will provide higher accuracy. To overcome these limitations and to achieve higher detection precision, the custom hotspot microarray have been created to detect CNVs within the Ankylosing Spondylitis (AS) cohort in chapter 5. We analyzed a unique multi-generation family with a high incidence of AS and back pain. The custom microarray analysis revealed that 56 CNVs that overlap or consist of a gene are enriched in at least three affected members and are absent within the unaffected family members. Among these 56 genes, ten CNVs segregated with all the affected members of the family. After scrutinizing all ten CNVs based on potential function, further analysis was performed on the association of the *UGT2B17* gene in large AS cohorts. The primary reason for investigating this gene was the functional relevance with a bone-related condition (i.e., osteoporosis). The affected individuals were all *HLA-B27* positive and they all carried two copies of the gene that was validated using QF-PCR.

The *UGT2B17* gene appears to be highly stratified within multiple populations however the reason for stratification is not yet known. It is often suspected that stratified genes are correlated with diseases and consequently hard to replicate into other populations. The association of this gene was conducted in two geographically distant Canadian Caucasian populations from Newfoundland and Alberta. The association showed a trend of significance (p < 0.05) in the Newfoundland population but replication failed in the Alberta population. In the global population, this gene is highly stratified, where deletion is much more prevalent within Asian populations, and rare in African populations. Although the cause of such varying selection pressure of these types of genes in different populations is unknown, they are often thought to be good candidates for disease susceptibility. One hypothesis stated that stratified genes are a more general risk factor with particular manifestations sensitive to genetic modifiers or environmental effects (Itsara et al. 2009). This requires more investigation to elucidate the cause that underlies the stratification.

Functionally, the *UGT2B17* gene has been demonstrated to be relevant in two different diseases, osteoporosis and prostate cancer. Although, there are conflicting reports on *UGT2B17* association, this gene is highly expressed in the prostate. The association reported with osteoporosis is much more relevant to the AS study presented in chapter 5. *UGT2B17* encodes a key enzyme responsible for glucuronidation of androgens and their metabolites in humans. For estrogen, androgen is a major source, and both have direct effects for stimulating bone formation. This is relevant with the osteoporosis association and *UGT2B17* because the gene product influences serum androgen or estrogen concentrations. The authors hypothesised that *UGT2B17* gene dosage through CNV is strongly linked with androgen or estrogen concentrations. AS is a disease characterized by bone overgrowth, whereas osteoporosis is the loss of bone density that causes the diseases. Although the primary effect on bone for these two diseases (AS and OP) appears to be in opposite directions, it has been acknowledged that a significant portion of AS patients develop osteoporosis (Magrey and Khan 2010). This correlation supports the role of *UGT2B17* gene dosage in the regulation of bone density within AS patients.

A comprehensive map of CNVs is presently not available and it will take an unprecedented amount of data and work to unfold the true content of structural variants in the human genome. The methods described in this thesis needs further validation, especially the characterization of rearrangement hotspots within multiple complex diseases that requires large scale investigation. This thesis did not attempt to capture all structural variants, and excluded new insertions relative to the reference genome and inversions. The method of detection of insertions and inversions is still in the developmental phase. More research is required to improve the detection resolution and to elucidate their role in disease pathogenesis. The analyses of CNVs in complex diseases are not mature enough because determining a technique to make an absolute copy number call is not yet straight forward. Hence, the effect of gene dosage within disease pathogenesis is still not clear. High throughput technology holds the promise to overcome these complexities in the near future.

For the identification of novel variants and to better understand the role of these hotspots, additional complex diseases need to be investigated. In the case of Tourette syndrome, in-depth evaluation of the 2q14.3-2q21.3 critical region within large multiplex families and well defined case-control TS cohort is warranted. Functional analysis on the *C2orf27A* gene is required to unravel the potential molecular correlation of *C2orf27A* gene in Tourette syndrome. Due to multiple co-morbidities in this family, it is also of interest to assess genes specific to the various phenotypes (such as OCD and ADHD). Finally, a large well-phenotyped cohort of TS patients will allow us to assess the generalizability of our findings and allow subset analysis on CNV association that will eventually lead to more direct genotype-phenotype correlations.

Regarding Ankylosing Spondylitis, the copy number of the *UGT2B17* gene warrants analysis in additional datasets within the SPARCC (Spondylitis Research Consortium of Canada) or NASC (North American Spondylitis Consortium) cohorts that have already been collected and in additional large multiplex families. The remaining nine CNVs (chapter 5, supplementary 1) that segregated within the large multiplex family should also be assessed in multiplex AS families and in a large AS cohort (similar to what is being proposed for the *UGT2B17* analysis).

Family-specific association with CNVs (atypical or rare) are very common in developmental neurocognitive disorders (Dharmadhikari et al. 2012; Cooper et al. 2011; Cook et al. 2008). This property of CNV segregation has yet to be systematically assessed for common complex diseases (i.e. RA, AS, SLE, Crohns, etc.). The primary two diseases (TS - chapter 3; AS - chapter 5) investigated in this thesis used multiplex families as a starting point in the detection of novel CNV variants. The present strategy being employed for gene identification in common complex diseases involve recruitment of a large number of unrelated case-control samples for GWAS (WTCCC 2007; WTCCC 2010). However, this approach overlooks the informative multiplex families that are important to assess rare highly penetrant variants. As a result, we feel that the optimal approach for gene identification is to combine the interrogation of multigenerational pedigrees along with large population-based cohorts. This hypothesis of family-specific causal variants for common complex diseases requires strong validation through extensive research. As high throughput sequencing technology is becoming available as a cost effective approach, genomic research will shed light on family-specific segregation of causal genomic variants.

The effort being put into genomic discovery is expanding exponentially, especially with respect to structural variations. The results described in this thesis will prove increasingly important for the elucidation of complex regions and their underlying mechanism into the etiology of complex diseases. The future direction of this research should focus predominantly on the characterization of these hotspot regions in human complex diseases. Improvement in high throughput sequencing technology will provide an edge over all other existing technologies to better understand the origin and mechanism of these hotspot regions in disease pathogenesis. We have explored these complex genomic regions specifying primarily as a hotbed for the NAHR mechanism. A comprehensive analysis will shed light on the contribution of other proposed mechanisms (i.e. NHEJ, MMBIR, FoSTeS, etc.). Another important factor is the environmental effect on genotype-phenotype correlation that is beyond the scope of this thesis. For future research, along with genomic variants, analyzing environmental covariates will increase the possibility to identify the apparent missing heritability of complex diseases.

In this thesis, an attempt was made to characterize complex genomic regions through computational and molecular approaches; subsequent analyses also investigated the role of such regions in complex diseases. The primary contributions of this thesis include: 1) the development of computational algorithms to characterize complex regions; 2) custom design microarray targeting those detected hotspot regions; and 3) the application of custom design microarray in two complex diseases. The application of the custom microarray detected a novel locus for Tourette syndrome and a candidate CNV for AS and represent a major contribution to this thesis. There are three major contributions arising from this thesis. First, a computational technique using high throughput sequencing was developed that is able to decipher complex regions (specifically, the relationship between SDs) that alleviate all previous microarray-based approaches. Secondly, the detection and characterization of rearrangement hotspots to better elucidate the underlying mechanism for pathogenic CNV variants was performed. Thirdly, molecular experiments targeting these regions were effectively utilized to detect novel variants, as was demonstrated in two disease cohorts.

Although it is ideal to analyze the entire spectrum of mutation variants, CNV was the primary variant type that was investigated in chapter 3, 4 and 5. For future research, a complete genome sequence analysis for ‘Family A’ (analyzed in chapter 3), using high throughput sequencing technology, with at least 40x coverage, will be undertaken. This will allow us to detect and analyze all spectrums of genomic mutations (i.e. SNP, CNV etc). Similar sequencing analysis will be applied to the AS multiplex family described in chapter 5, with the anticipation of detecting a rare familial AS susceptibility locus.

REFERENCES

1000 GENOMES PROJECT CONSORTIUM, 2010. A map of human genome variation from population-scale sequencing. *Nature,* **467**(7319), pp. 1061-1073.

ABDOOL, A., DONAHUE, A.C., WOHLGEMUTH, J.G. and YEH, C.H., 2010. Detection, analysis and clinical validation of chromosomal aberrations by multiplex ligation-dependent probe amplification in chronic leukemia. *PloS one,* **5**(10), pp. e15407.

ABELSON, J.F., KWAN, K.Y., O'ROAK, B.J., BAEK, D.Y., STILLMAN, A.A., MORGAN, T.M., MATHEWS, C.A., PAULS, D.L., RASIN, M.R., GUNEL, M., DAVIS, N.R., ERCAN-SENCICEK, A.G., GUEZ, D.H., SPERTUS, J.A., LECKMAN, J.F., DURE, L.S.,4TH, KURLAN, R., SINGER, H.S., GILBERT, D.L., FARHI, A., LOUVI, A., LIFTON, R.P., SESTAN, N. and STATE, M.W., 2005. Sequence variants in SLITRK1 are associated with Tourette's syndrome. *Science (New York, N.Y.),* **310**(5746), pp. 317-320.

ALBIN, R.L., YOUNG, A.B. and PENNEY, J.B., 1995. The functional anatomy of disorders of the basal ganglia. *Trends in neurosciences,* **18**(2), pp. 63-64.

ALKAN, C., KIDD, J.M., MARQUES-BONET, T., AKSAY, G., ANTONACCI, F., HORMOZDIARI, F., KITZMAN, J.O., BAKER, C., MALIG, M. and MUTLU, O., 2009. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature genetics,* **41**(10):1061-1067.

ALKAN, C., COE, B.P. and EICHLER, E.E., 2011. Genome structural variation discovery and genotyping. *Nature reviews.Genetics,* **12**(5), pp. 363-376.

ALKAN, C., SAJJADIAN, S. and EICHLER, E.E., 2011. Limitations of next-generation genome sequence assembly. *Nature methods,* **8**(1), pp. 61-65.

ALTSCHUL, S.F., GISH, W., MILLER, W., MYERS, E.W. and LIPMAN, D.J., 1990. Basic local alignment search tool. *Journal of Molecular Biology,* **215**(3), pp. 403-410.

ANTONACCI, F., KIDD, J.M., MARQUES-BONET, T., TEAGUE, B., VENTURA, M., GIRIRAJAN, S., ALKAN, C., CAMPBELL, C.D., VIVES, L., MALIG, M., ROSENFELD, J.A., BALLIF, B.C., SHAFFER, L.G., GRAVES, T.A., WILSON, R.K., SCHWARTZ, D.C. and EICHLER, E.E., 2010. A large and complex structural polymorphism at 16p12.1 underlies microdeletion disease risk. *Nature genetics,* **42**(9), pp. 745-750.

AUSTRALO-ANGLO-AMERICAN SPONDYLOARTHRITIS CONSORTIUM (TASC), REVEILLE, J.D., SIMS, A.M., DANOY, P., EVANS, D.M., LEO, P., POINTON, J.J., JIN, R., ZHOU, X., BRADBURY, L.A., APPLETON, L.H., DAVIS, J.C., DIEKMAN, L., DOAN, T., DOWLING, A., DUAN, R., DUNCAN, E.L., FARRAR, C., HADLER, J., HARVEY, D., KARADERI, T., MOGG, R., POMEROY, E., PRYCE, K., TAYLOR, J., SAVAGE, L., DELOUKAS, P., KUMANDURI, V., PELTONEN, L., RING, S.M., WHITTAKER, P., GLAZOV, E., THOMAS, G.P., MAKSYMOWYCH, W.P., INMAN, R.D., WARD, M.M., STONE, M.A., WEISMAN, M.H., WORDSWORTH, B.P. and BROWN, M.A., 2010. Genome-wide association study of ankylosing spondylitis identifies non-MHC susceptibility loci. *Nature genetics,* **42**(2), pp. 123-127.

BAILEY, J.A. and EICHLER, E.E., 2006. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nature reviews.Genetics,* **7**(7), pp. 552-564.

BAILEY, J.A., GU, Z., CLARK, R.A., REINERT, K., SAMONTE, R.V., SCHWARTZ, S., ADAMS, M.D., MYERS, E.W., LI, P.W. and EICHLER, E.E., 2002. Recent segmental duplications in the human genome.*Science (New York, N.Y.),* **297**(5583), pp. 1003-1007.

BALLIF, B.C., HORNOR, S.A., JENKINS, E., MADAN-KHETARPAL, S., SURTI, U., JACKSON, K.E., ASAMOAH, A., BROCK, P.L., GOWANS, G.C., CONWAY, R.L., GRAHAM, J.M.,JR, MEDNE, L., ZACKAI, E.H., SHAIKH, T.H., GEOGHEGAN, J., SELZER, R.R., EIS, P.S., BEJJANI, B.A. and SHAFFER, L.G., 2007. Discovery of a previously unrecognized microdeletion syndrome of 16p11.2-p12.2. *Nature genetics,* **39**(9), pp. 1071-1073.

BALLIF, B.C., THEISEN, A., COPPINGER, J., GOWANS, G.C., HERSH, J.H., MADAN-KHETARPAL, S., SCHMIDT, K.R., TERVO, R., ESCOBAR, L.F., FRIEDRICH, C.A., MCDONALD, M., CAMPBELL, L., MING, J.E., ZACKAI, E.H., BEJJANI, B.A. and SHAFFER, L.G., 2008. Expanding the clinical phenotype of the 3q29 microdeletion syndrome and characterization of the reciprocal microduplication. *Molecular cytogenetics,* **1**, pp. 8.

BAYANI, J. and SQUIRE, J.A., 2004. Fluorescence in situ Hybridization (FISH). *Current protocols in cell biology / editorial board, Juan S.Bonifacino ...[et al.],* **Chapter 22**, pp. Unit 22.4.

BAYES, M., MAGANO, L.F., RIVERA, N., FLORES, R. and PEREZ JURADO, L.A., 2003. Mutational mechanisms of Williams-Beuren syndrome deletions. *American Journal of Human Genetics,* **73**(1), pp. 131-151.

BEAR, J.C., NEMEC, T.F., KENNEDY, J.C., MARSHALL, W.H., POWER, A.A., KOLONEL, V.M. and BURKE, G.B., 1988. Inbreeding in outport Newfoundland. *American Journal of Medical Genetics,* **29**(3), pp. 649-660.

BEAR, J.C., NEMEC, T.F., KENNEDY, J.C., MARSHALL, W.H., POWER, A.A., KOLONEL, V.M. and BURKE, G.B., 1987. Persistent genetic isolation in outport Newfoundland. *American Journal of Medical Genetics,* **27**(4), pp. 807-830.

BENTLEY, D.R., BALASUBRAMANIAN, S., SWERDLOW, H.P., SMITH, G.P., MILTON, J., BROWN, C.G., HALL, K.P., EVERS, D.J., BARNES, C.L., BIGNELL, H.R., BOUTELL, J.M., BRYANT, J., CARTER, R.J., KEIRA CHEETHAM, R., COX, A.J., ELLIS, D.J., FLATBUSH, M.R., GORMLEY, N.A., HUMPHRAY, S.J., IRVING, L.J., KARBELASHVILI, M.S., KIRK, S.M., LI, H., LIU, X., MAISINGER, K.S., MURRAY, L.J., OBRADOVIC, B., OST, T., PARKINSON, M.L., PRATT, M.R., RASOLONJATOVO, I.M., REED, M.T., RIGATTI, R., RODIGHIERO, C., ROSS, M.T., SABOT, A., SANKAR, S.V., SCALLY, A., SCHROTH, G.P., SMITH, M.E., SMITH, V.P., SPIRIDOU, A., TORRANCE, P.E., TZONEV, S.S., VERMAAS, E.H., WALTER, K., WU, X., ZHANG, L., ALAM, M.D., ANASTASI, C., ANIEBO, I.C., BAILEY, D.M., BANCARZ, I.R., BANERJEE, S., BARBOUR, S.G., BAYBAYAN, P.A., BENOIT, V.A., BENSON, K.F., BEVIS, C., BLACK, P.J., BOODHUN, A., BRENNAN, J.S., BRIDGHAM, J.A., BROWN, R.C., BROWN, A.A., BUERMANN, D.H., BUNDU, A.A., BURROWS, J.C., CARTER, N.P., CASTILLO, N., CHIARA E CATENAZZI, M., CHANG, S., NEIL COOLEY, R., CRAKE, N.R., DADA, O.O., DIAKOUMAKOS, K.D., DOMINGUEZ-FERNANDEZ, B., EARNSHAW, D.J., EGBUJOR, U.C., ELMORE, D.W., ETCHIN, S.S., EWAN, M.R., FEDURCO, M., FRASER, L.J., FUENTES FAJARDO, K.V., SCOTT FUREY, W., GEORGE, D., GIETZEN, K.J., GODDARD, C.P., GOLDA, G.S., GRANIERI, P.A., GREEN, D.E., GUSTAFSON, D.L., HANSEN, N.F., HARNISH, K., HAUDENSCHILD, C.D., HEYER, N.I., HIMS, M.M., HO, J.T., HORGAN, A.M., HOSCHLER, K., HURWITZ, S., IVANOV, D.V., JOHNSON, M.Q., JAMES, T., HUW JONES, T.A., KANG, G.D., KERELSKA, T.H., KERSEY, A.D., KHREBTUKOVA, I., KINDWALL, A.P., KINGSBURY, Z., KOKKO-GONZALES, P.I., KUMAR, A., LAURENT, M.A., LAWLEY, C.T., LEE, S.E., LEE, X., LIAO, A.K., LOCH, J.A., LOK, M., LUO, S., MAMMEN, R.M., MARTIN, J.W., MCCAULEY, P.G., MCNITT, P., MEHTA, P., MOON, K.W., MULLENS, J.W., NEWINGTON, T., NING, Z., LING NG, B., NOVO, S.M., O'NEILL, M.J., OSBORNE, M.A., OSNOWSKI, A., OSTADAN, O., PARASCHOS, L.L., PICKERING, L., PIKE, A.C., PIKE, A.C., CHRIS PINKARD, D., PLISKIN, D.P., PODHASKY, J., QUIJANO, V.J., RACZY, C., RAE, V.H., RAWLINGS, S.R., CHIVA RODRIGUEZ, A., ROE, P.M., ROGERS, J., ROGERT BACIGALUPO, M.C., ROMANOV, N., ROMIEU, A., ROTH, R.K., ROURKE, N.J., RUEDIGER, S.T., RUSMAN, E., SANCHES-KUIPER, R.M., SCHENKER, M.R., SEOANE, J.M., SHAW, R.J., SHIVER, M.K., SHORT, S.W., SIZTO, N.L., SLUIS, J.P., SMITH, M.A., ERNEST SOHNA SOHNA, J., SPENCE, E.J., STEVENS, K., SUTTON, N., SZAJKOWSKI, L., TREGIDGO, C.L., TURCATTI, G., VANDEVONDELE, S., VERHOVSKY, Y., VIRK, S.M., WAKELIN, S., WALCOTT, G.C., WANG, J., WORSLEY, G.J., YAN, J., YAU, L., ZUERLEIN, M., ROGERS, J., MULLIKIN, J.C., HURLES, M.E., MCCOOKE, N.J., WEST, J.S., OAKS, F.L., LUNDBERG, P.L., KLENERMAN, D., DURBIN, R. and SMITH, A.J., 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature,* **456**(7218), pp. 53-59.

BERGBOER, J.G., UMICEVIC-MIRKOV, M., FRANSEN, J., DEN HEIJER, M., FRANKE, B., VAN RIEL, P.L., SCHALKWIJK, J., COENEN, M.J. and NIJMEGEN BIOMEDICAL STUDY PRINCIPAL INVESTIGATORS, 2012. A replication study of the association between rheumatoid arthritis and deletion of the late cornified envelope genes LCE3B and LCE3C. *PloS one,* **7**(2), pp. e32045.

BERKEL, S., MARSHALL, C.R., WEISS, B., HOWE, J., ROETH, R., MOOG, U., ENDRIS, V., ROBERTS, W., SZATMARI, P., PINTO, D., BONIN, M., RIESS, A., ENGELS, H., SPRENGEL, R., SCHERER, S.W. and RAPPOLD, G.A., 2010. Mutations in the SHANK2 synaptic scaffolding gene in autism spectrum disorder and mental retardation. *Nature genetics,* **42**(6), pp. 489-491.

BOWES, J., FLYNN, E., HO, P., ALY, B., MORGAN, A.W., MARZO-ORTEGA, H., COATES, L., MCMANUS, R., RYAN, A.W., KANE, D., KORENDOWYCH, E., MCHUGH, N., FITZGERALD, O., PACKHAM, J., BRUCE, I.N. and BARTON, A., 2010. Variants in linkage disequilibrium with the late cornified envelope gene cluster deletion are associated with susceptibility to psoriatic arthritis. *Annals of the Rheumatic Diseases,* **69**(12), pp. 2199-2203.

BREST, P., LAPAQUETTE, P., SOUIDI, M., LEBRIGAND, K., CESARO, A., VOURET-CRAVIARI, V., MARI, B., BARBRY, P., MOSNIER, J.F., HEBUTERNE, X., HAREL-BELLAN, A., MOGRABI, B., DARFEUILLE-MICHAUD, A. and

HOFMAN, P., 2011. A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn's disease. *Nature genetics,***43**(3), pp. 242-245.

BRIDGES, C.B., 1936. The Bar "Gene" a Duplication. *Science (New York, N.Y.),* **83**(2148), pp. 210-211.

BROWN, M.A., KENNEDY, L.G., MACGREGOR, A.J., DARKE, C., DUNCAN, E., SHATFORD, J.L., TAYLOR, A., CALIN, A. and WORDSWORTH, P., 1997. Susceptibility to ankylosing spondylitis in twins: the role of genes, HLA, and the environment. *Arthritis and Rheumatism,* **40**(10), pp. 1823-1828.

BROWN, M.A., LAVAL, S.H., BROPHY, S. and CALIN, A., 2000. Recurrence risk modelling of the genetic susceptibility to ankylosing spondylitis. *Annals of the Rheumatic Diseases,* **59**(11), pp. 883-886.

BRUNETTI-PIERRI, N., BERG, J.S., SCAGLIA, F., BELMONT, J., BACINO, C.A., SAHOO, T., LALANI, S.R., GRAHAM, B., LEE, B., SHINAWI, M., SHEN, J., KANG, S.H., PURSLEY, A., LOTZE, T., KENNEDY, G., LANSKY-SHAFER, S., WEAVER, C., ROEDER, E.R., GREBE, T.A., ARNOLD, G.L., HUTCHISON, T., REIMSCHISEL, T., AMATO, S., GERAGTHY, M.T., INNIS, J.W., OBERSZTYN, E., NOWAKOWSKA, B., ROSENGREN, S.S., BADER, P.I., GRANGE, D.K., NAQVI, S., GARNICA, A.D., BERNES, S.M., FONG, C.T., SUMMERS, A., WALTERS, W.D., LUPSKI, J.R., STANKIEWICZ, P., CHEUNG, S.W. and PATEL, A., 2008. Recurrent reciprocal 1q21.1 deletions and duplications associated with microcephaly or macrocephaly and developmental and behavioral abnormalities. *Nature genetics,* **40**(12), pp. 1466-1471.

BULLINGER, L., KRONKE, J., SCHON, C., RADTKE, I., URLBAUER, K., BOTZENHARDT, U., GAIDZIK, V., CARIO, A., SENGER, C., SCHLENK, R.F., DOWNING, J.R., HOLZMANN, K., DOHNER, K. and DOHNER, H., 2010. Identification of acquired copy number alterations and uniparental disomies in cytogenetically normal acute myeloid leukemia using high-resolution single-nucleotide polymorphism analysis.*Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K,* **24**(2), pp. 438-449.

CALIN, A., MARDER, A., BECKS, E. and BURNS, T., 1983. Genetic differences between B27 positive patients with ankylosing spondylitis and B27 positive healthy controls. *Arthritis and Rheumatism,***26**(12), pp. 1460-1464.

CAMPBELL, P.J., STEPHENS, P.J., PLEASANCE, E.D., O'MEARA, S., LI, H., SANTARIUS, T., STEBBINGS, L.A., LEROY, C., EDKINS, S., HARDY, C., TEAGUE, J.W., MENZIES, A., GOODHEAD, I., TURNER, D.J., CLEE, C.M., QUAIL, M.A., COX, A., BROWN, C., DURBIN, R., HURLES, M.E., EDWARDS, P.A., BIGNELL, G.R., STRATTON, M.R. and FUTREAL, P.A., 2008. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature genetics,* **40**(6), pp. 722-729.

CARTER, N.P., 2007. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nature genetics,* **39**(7 Suppl), pp. S16-21.

CENTERS FOR DISEASE CONTROL AND PREVENTION (CDC), 2009. Prevalence of diagnosed Tourette syndrome in persons aged 6-17 years - United States, 2007. *MMWR.Morbidity and mortality weekly report,* **58**(21), pp. 581-585.

CHAISSON, M.J., BRINZA, D. and PEVZNER, P.A., 2009. De novo fragment assembly with short mate-paired reads: Does the read length matter? *Genome research,* **19**(2), pp. 336-346.

CHIANG, D.Y., GETZ, G., JAFFE, D.B., O'KELLY, M.J., ZHAO, X., CARTER, S.L., RUSS, C., NUSBAUM, C., MEYERSON, M. and LANDER, E.S., 2009. High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature methods,* **6**(1), pp. 99-103.

CLARKE, B.L. and KHOSLA, S., 2009. Androgens and bone. *Steroids,* **74**(3), pp. 296-305.

CLEMENTS, J.L., YANG, B., ROSS-BARTA, S.E., ELIASON, S.L., HRSTKA, R.F., WILLIAMSON, R.A. and KORETZKY, G.A., 1998. Requirement for the leukocyte-specific adapter protein SLP-76 for normal T cell development. *Science (New York, N.Y.),* **281**(5375), pp. 416-419.

COE, B.P., YLSTRA, B., CARVALHO, B., MEIJER, G.A., MACAULAY, C. and LAM, W.L., 2007. Resolving the resolution of array CGH. *Genomics,* **89**(5), pp. 647-653.

CONRAD, D.F., BIRD, C., BLACKBURNE, B., LINDSAY, S., MAMANOVA, L., LEE, C., TURNER, D.J. and HURLES, M.E., 2010a. Mutation spectrum revealed by breakpoint sequencing of human germline CNVs.*Nature genetics,* **42**(5), pp. 385-391.

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, Fitzgerald T, Hu M, Ihm CH, Kristiansson K, Macarthur DG, Macdonald JR, Onyiah I, Pang AW, Robson S, Stirrups K, Valsesia A, Walter K, Wei J; Wellcome Trust Case Control Consortium, Tyler-Smith C, Carter NP, Lee C, Scherer SW, Hurles ME. 2010b. Origins and functional impact of copy number variation in the human genome. *Nature* **464**(7289), pp. 704-12.

COOK, E.H.,JR and SCHERER, S.W., 2008. Copy-number variations associated with neuropsychiatric conditions. *Nature,* **455**(7215), pp. 919-923.

COOPER, C., CARBONE, L., MICHET, C.J., ATKINSON, E.J., O'FALLON, W.M. and MELTON, L.J.,3RD, 1994. Fracture risk in patients with ankylosing spondylitis: a population based study. *The Journal of rheumatology,* **21**(10), pp. 1877-1882.

COOPER, G.M., COE, B.P., GIRIRAJAN, S., ROSENFELD, J.A., VU, T.H., BAKER, C., WILLIAMS, C., STALKER, H., HAMID, R., HANNIG, V., ABDEL-HAMID, H., BADER, P., MCCRACKEN, E., NIYAZOV, D., LEPPIG, K., THIESE, H., HUMMEL, M., ALEXANDER, N., GORSKI, J., KUSSMANN, J., SHASHI, V., JOHNSON, K., REHDER, C., BALLIF, B.C., SHAFFER, L.G. and EICHLER, E.E., 2011. A copy number variation morbidity map of developmental delay. *Nature genetics,* **43**(9), pp. 838-846.

COOPER, G.M., ZERR, T., KIDD, J.M., EICHLER, E.E. and NICKERSON, D.A., 2008. Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nature genetics,* **40**(10), pp. 1199-1203.

DALMA-WEISZHAUSZ, D.D., WARRINGTON, J., TANIMOTO, E.Y. and MIYADA, C.G., 2006. The affymetrix GeneChip platform: an overview. *Methods in enzymology,* **410**, pp. 3-28.

DAWSON, G., 2010. Recent advances in research on early detection, causes, biology, and treatment of autism spectrum disorders. *Current opinion in neurology,* **23**(2), pp. 95-96.

DE CID, R., RIVEIRA-MUNOZ, E., ZEEUWEN, P.L., ROBARGE, J., LIAO, W., DANNHAUSER, E.N., GIARDINA, E., STUART, P.E., NAIR, R., HELMS, C., ESCARAMIS, G., BALLANA, E., MARTIN-EZQUERRA, G., DEN HEIJER, M., KAMSTEEG, M., JOOSTEN, I., EICHLER, E.E., LAZARO, C., PUJOL, R.M., ARMENGOL, L., ABECASIS, G., ELDER, J.T., NOVELLI, G., ARMOUR, J.A., KWOK, P.Y., BOWCOCK, A., SCHALKWIJK, J. and ESTIVILL, X., 2009. Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. *Nature genetics,* **41**(2), pp. 211-215.

DHARMADHIKARI, A.V., KANG, S.H., SZAFRANSKI, P., PERSON, R.E., SAMPATH, S., PRAKASH, S.K., BADER, P.I., PHILLIPS, J.A.,3RD, HANNIG, V., WILLIAMS, M., VINSON, S.S., WILFONG, A.A., REIMSCHISEL, T.E., CRAIGEN, W.J., PATEL, A., BI, W., LUPSKI, J.R., BELMONT, J., CHEUNG, S.W. and STANKIEWICZ, P., 2012. Small rare recurrent deletions and reciprocal duplications in 2q21.1, including brain-specific ARHGEF4 and GPR148. *Human molecular genetics,* **21**(15), pp. 3345-3355.

DIDEBERG, V., KRISTJANSDOTTIR, G., MILANI, L., LIBIOULLE, C., SIGURDSSON, S., LOUIS, E., WIMAN, A.C., VERMEIRE, S., RUTGEERTS, P., BELAICHE, J., FRANCHIMONT, D., VAN GOSSUM, A., BOURS, V. and SYVANEN, A.C., 2007. An insertion-deletion polymorphism in the interferon regulatory Factor 5 (IRF5) gene confers risk of inflammatory bowel diseases. *Human molecular genetics,* **16**(24), pp. 3008-3016.

DISKIN, S.J., HOU, C., GLESSNER, J.T., ATTIYEH, E.F., LAUDENSLAGER, M., BOSSE, K., COLE, K., MOSSE, Y.P., WOOD, A., LYNCH, J.E., PECOR, K., DIAMOND, M., WINTER, C., WANG, K., KIM, C., GEIGER, E.A., MCGRADY, P.W., BLAKEMORE, A.I., LONDON, W.B., SHAIKH, T.H., BRADFIELD, J., GRANT, S.F., LI, H., DEVOTO, M., RAPPAPORT, E.R., HAKONARSON, H. and MARIS, J.M., 2009. Copy number variation at 1q21.1 associated with neuroblastoma. *Nature,* **459**(7249), pp. 987-991.

DURAND, C.M., BETANCUR, C., BOECKERS, T.M., BOCKMANN, J., CHASTE, P., FAUCHEREAU, F., NYGREN, G., RASTAM, M., GILLBERG, I.C., ANCKARSATER, H., SPONHEIM, E., GOUBRAN-BOTROS, H., DELORME, R., CHABANE, N., MOUREN-SIMEONI, M.C., DE MAS, P., BIETH, E., ROGE, B., HERON, D., BURGLEN, L., GILLBERG, C., LEBOYER, M. and BOURGERON, T., 2007. Mutations in the gene encoding the synaptic scaffolding protein SHANK3 are associated with autism spectrum disorders. *Nature genetics,* **39**(1), pp. 25-27.

EAPEN, V., PAULS, D.L. and ROBERTSON, M.M., 1993. Evidence for autosomal dominant transmission in Tourette's syndrome. United Kingdom cohort study. *The British journal of psychiatry : the journal of mental science,* **162**, pp. 593-596.

EGAN, C.M., SRIDHAR, S., WIGLER, M. and HALL, I.M., 2007. Recurrent DNA copy number variation in the laboratory mouse. *Nature genetics,* **39**(11), pp. 1384-1389.

ELDER, J.T., 2009. Genome-wide association scan yields new insights into the immunopathogenesis of psoriasis. *Genes and immunity,* **10**(3), pp. 201-209.

ENSENAUER, R.E., ADEYINKA, A., FLYNN, H.C., MICHELS, V.V., LINDOR, N.M., DAWSON, D.B., THORLAND, E.C., LORENTZ, C.P., GOLDSTEIN, J.L., MCDONALD, M.T., SMITH, W.E., SIMON-FAYARD, E., ALEXANDER, A.A., KULHARYA, A.S., KETTERLING, R.P., CLARK, R.D. and JALAL, S.M., 2003. Microduplication 22q11.2, an emerging syndrome: clinical, cytogenetic, and molecular analysis of thirteen patients. *American Journal of Human Genetics,* **73**(5), pp. 1027-1040.

ERCAN-SENCICEK, A.G., STILLMAN, A.A., GHOSH, A.K., BILGUVAR, K., O'ROAK, B.J., MASON, C.E., ABBOTT, T., GUPTA, A., KING, R.A., PAULS, D.L., TISCHFIELD, J.A., HEIMAN, G.A., SINGER, H.S., GILBERT, D.L., HOEKSTRA, P.J., MORGAN, T.M., LORING, E., YASUNO, K., FERNANDEZ, T., SANDERS, S., LOUVI, A., CHO, J.H., MANE, S., COLANGELO, C.M., BIEDERER, T., LIFTON, R.P., GUNEL, M. and STATE, M.W., 2010. L-histidine decarboxylase and Tourette's syndrome. *The New England journal of medicine,* **362**(20), pp. 1901-1908.

EVANS, D.M., SPENCER, C.C., POINTON, J.J., SU, Z., HARVEY, D., KOCHAN, G., OPPERMANN, U., DILTHEY, A., PIRINEN, M., STONE, M.A., APPLETON, L., MOUTSIANAS, L., LESLIE, S., WORDSWORTH, T., KENNA, T.J., KARADERI, T., THOMAS, G.P., WARD, M.M., WEISMAN, M.H., FARRAR, C., BRADBURY, L.A., DANOY, P., INMAN, R.D., MAKSYMOWYCH, W., GLADMAN, D., RAHMAN, P., SPONDYLOARTHRITIS RESEARCH CONSORTIUM OF CANADA (SPARCC), MORGAN, A., MARZO-ORTEGA, H., BOWNESS, P., GAFFNEY, K., GASTON, J.S., SMITH, M., BRUGES-ARMAS, J., COUTO, A.R., SORRENTINO, R., PALADINI, F., FERREIRA, M.A., XU, H., LIU, Y., JIANG, L., LOPEZ-LARREA, C., DIAZ-PENA, R., LOPEZ-VAZQUEZ, A., ZAYATS, T., BAND, G., BELLENGUEZ, C., BLACKBURN, H., BLACKWELL, J.M., BRAMON, E., BUMPSTEAD, S.J., CASAS, J.P., CORVIN, A., CRADDOCK, N., DELOUKAS, P., DRONOV, S., DUNCANSON, A., EDKINS, S., FREEMAN, C., GILLMAN, M., GRAY, E., GWILLIAM, R., HAMMOND, N., HUNT, S.E., JANKOWSKI, J., JAYAKUMAR, A., LANGFORD, C., LIDDLE, J., MARKUS, H.S., MATHEW, C.G., MCCANN, O.T., MCCARTHY, M.I., PALMER, C.N., PELTONEN, L., PLOMIN, R., POTTER, S.C., RAUTANEN, A., RAVINDRARAJAH, R., RICKETTS, M., SAMANI, N., SAWCER, S.J., STRANGE, A., TREMBATH, R.C., VISWANATHAN, A.C., WALLER, M., WESTON, P., WHITTAKER, P., WIDAA, S., WOOD, N.W., MCVEAN, G., REVEILLE, J.D., WORDSWORTH, B.P., BROWN, M.A., DONNELLY, P., AUSTRALO-ANGLO-AMERICAN SPONDYLOARTHRITIS CONSORTIUM (TASC) and WELLCOME TRUST CASE CONTROL CONSORTIUM 2 (WTCCC2), 2011. Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. *Nature genetics,* **43**(8), pp. 761-767.

FANCIULLI, M., NORSWORTHY, P.J., PETRETTO, E., DONG, R., HARPER, L., KAMESH, L., HEWARD, J.M., GOUGH, S.C., DE SMITH, A., BLAKEMORE, A.I., FROGUEL, P., OWEN, C.J., PEARCE, S.H., TEIXEIRA, L., GUILLEVIN, L., GRAHAM, D.S., PUSEY, C.D., COOK, H.T., VYSE, T.J. and AITMAN, T.J., 2007. FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nature genetics,* **39**(6), pp. 721-723.

FELLERMANN, K., STANGE, D.E., SCHAEFFELER, E., SCHMALZL, H., WEHKAMP, J., BEVINS, C.L., REINISCH, W., TEML, A., SCHWAB, M., LICHTER, P., RADLWIMMER, B. and STANGE, E.F., 2006. A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. *American Journal of Human Genetics,* **79**(3), pp. 439-448.

FERNANDEZ, B., 2009. *The burden of genetic disease among hospitalized children in Newfoundland.*, Memorial University of Newfoundland.

FERNANDEZ, B.A., ROBERTS, W., CHUNG, B., WEKSBERG, R., MEYN, S., SZATMARI, P., JOSEPH-GEORGE, A.M., MACKAY, S., WHITTEN, K., NOBLE, B., VARDY, C., CROSBIE, V., LUSCOMBE, S., TUCKER, E., TURNER, L., MARSHALL, C.R. and SCHERER, S.W., 2010. Phenotypic spectrum associated with de novo and inherited deletions and duplications at 16p11.2 in individuals ascertained for diagnosis of autism spectrum disorder. *Journal of medical genetics,* **47**(3), pp. 195-203.

FERNANDEZ, T.V., SANDERS, S.J., YURKIEWICZ, I.R., ERCAN-SENCICEK, A.G., KIM, Y.S., FISHMAN, D.O., RAUBESON, M.J., SONG, Y., YASUNO, K., HO, W.S., BILGUVAR, K., GLESSNER, J., CHU, S.H., LECKMAN, J.F., KING, R.A., GILBERT, D.L., HEIMAN, G.A., TISCHFIELD, J.A., HOEKSTRA, P.J., DEVLIN, B., HAKONARSON, H., MANE, S.M., GUNEL, M. and STATE, M.W., 2012. Rare copy number variants in tourette syndrome disrupt genes in histaminergic pathways and overlap with autism. *Biological psychiatry,* **71**(5), pp. 392-402.

FEUK, L., CARSON, A.R. and SCHERER, S.W., 2006. Structural variation in the human genome. *Nature reviews.Genetics,* **7**(2), pp. 85-97.

FIRESTEIN, G.S., 2003. Evolving concepts of rheumatoid arthritis. *Nature,* **423**(6937), pp. 356-361.

FIRTH, H.V., RICHARDS, S.M., BEVAN, A.P., CLAYTON, S., CORPAS, M., RAJAN, D., VAN VOOREN, S., MOREAU, Y., PETTETT, R.M. and CARTER, N.P., 2009. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *American Journal of Human Genetics,* **84**(4), pp. 524-533.

FRANKE, A., MCGOVERN, D.P., BARRETT, J.C., WANG, K., RADFORD-SMITH, G.L., AHMAD, T., LEES, C.W., BALSCHUN, T., LEE, J., ROBERTS, R., ANDERSON, C.A., BIS, J.C., BUMPSTEAD, S., ELLINGHAUS, D., FESTEN, E.M., GEORGES, M., GREEN, T., HARITUNIANS, T., JOSTINS, L., LATIANO, A., MATHEW, C.G., MONTGOMERY, G.W., PRESCOTT, N.J., RAYCHAUDHURI, S., ROTTER, J.I., SCHUMM, P., SHARMA, Y., SIMMS, L.A., TAYLOR, K.D., WHITEMAN, D., WIJMENGA, C., BALDASSANO, R.N., BARCLAY, M., BAYLESS, T.M., BRAND, S., BUNING, C., COHEN, A., COLOMBEL, J.F., COTTONE, M., STRONATI, L., DENSON, T., DE VOS, M., D'INCA, R., DUBINSKY, M., EDWARDS, C., FLORIN, T., FRANCHIMONT, D., GEARRY, R., GLAS, J., VAN GOSSUM, A., GUTHERY, S.L., HALFVARSON, J., VERSPAGET, H.W., HUGOT, J.P., KARBAN, A., LAUKENS, D., LAWRANCE, I., LEMANN, M., LEVINE, A., LIBIOULLE, C., LOUIS, E., MOWAT, C., NEWMAN, W., PANES, J., PHILLIPS, A., PROCTOR, D.D., REGUEIRO, M., RUSSELL, R., RUTGEERTS, P., SANDERSON, J., SANS, M., SEIBOLD, F., STEINHART, A.H., STOKKERS, P.C., TORKVIST, L., KULLAK-UBLICK, G., WILSON, D., WALTERS, T., TARGAN, S.R., BRANT, S.R., RIOUX, J.D., D'AMATO, M., WEERSMA, R.K., KUGATHASAN, S., GRIFFITHS, A.M., MANSFIELD, J.C., VERMEIRE, S., DUERR, R.H., SILVERBERG, M.S., SATSANGI, J., SCHREIBER, S., CHO, J.H., ANNESE, V., HAKONARSON, H., DALY, M.J. and PARKES, M., 2010. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature genetics,* **42**(12), pp. 1118-1125.

FRIEDBERG, E.C., 2008. A brief history of the DNA repair field. *Cell research,* **18**(1), pp. 3-7.

GNERRE, S., MACCALLUM, I., PRZYBYLSKI, D., RIBEIRO, F.J., BURTON, J.N., WALKER, B.J., SHARPE, T., HALL, G., SHEA, T.P., SYKES, S., BERLIN, A.M., AIRD, D., COSTELLO, M., DAZA, R., WILLIAMS, L., NICOL, R., GNIRKE, A., NUSBAUM, C., LANDER, E.S. and JAFFE, D.B., 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences of the United States of America,* **108**(4), pp. 1513-1518.

GONZALEZ, E., KULKARNI, H., BOLIVAR, H., MANGANO, A., SANCHEZ, R., CATANO, G., NIBBS, R.J., FREEDMAN, B.I., QUINONES, M.P., BAMSHAD, M.J., MURTHY, K.K., ROVIN, B.H., BRADLEY, W., CLARK, R.A., ANDERSON, S.A., O'CONNELL, R.J., AGAN, B.K., AHUJA, S.S., BOLOGNA, R., SEN, L., DOLAN, M.J. and AHUJA, S.K., 2005. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science (New York, N.Y.),* **307**(5714), pp. 1434-1440.

GRAHAM, R.R., COTSAPAS, C., DAVIES, L., HACKETT, R., LESSARD, C.J., LEON, J.M., BURTT, N.P., GUIDUCCI, C., PARKIN, M., GATES, C., PLENGE, R.M., BEHRENS, T.W., WITHER, J.E., RIOUX, J.D., FORTIN, P.R., GRAHAM, D.C., WONG, A.K., VYSE, T.J., DALY, M.J., ALTSHULER, D., MOSER, K.L. and GAFFNEY, P.M., 2008. Genetic variants near TNFAIP3 on 6q23 are associated with systemic lupus erythematosus. *Nature genetics,* **40**(9), pp. 1059-1061.

GREEN, J.S., PARFREY, P.S., HARNETT, J.D., FARID, N.R., CRAMER, B.C., JOHNSON, G., HEATH, O., MCMANAMON, P.J., O'LEARY, E. and PRYSE-PHILLIPS, W., 1989. The cardinal manifestations of Bardet-Biedl syndrome, a form of Laurence-Moon-Biedl syndrome. *The New England journal of medicine,* **321**(15), pp. 1002-1009.

GRIFFITHS, R.J., SMITH, M.A., ROACH, M.L., STOCK, J.L., STAM, E.J., MILICI, A.J., SCAMPOLI, D.N., ESKRA, J.D., BYRUM, R.S., KOLLER, B.H. and MCNEISH, J.D., 1997. Collagen-induced arthritis is reduced in 5-lipoxygenase-activating protein-deficient mice. *The Journal of experimental medicine,* **185**(6), pp. 1123-1129.

GU, W., ZHANG, F. and LUPSKI, J.R., 2008. Mechanisms for human genomic rearrangements. *PathoGenetics,* **1**(1), pp. 4.

GUSEV, A., LOWE, J.K., STOFFEL, M., DALY, M.J., ALTSHULER, D., BRESLOW, J.L., FRIEDMAN, J.M. and PE'ER, I., 2009. Whole population, genome-wide mapping of hidden relatedness. *Genome research,***19**(2), pp. 318-326.

HACH, F., HORMOZDIARI, F., ALKAN, C., HORMOZDIARI, F., BIROL, I., EICHLER, E.E. and SAHINALP, S.C., 2010. mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nature methods,* **7**(8), pp. 576-577.

HAJIRASOULIHA, I., HORMOZDIARI, F., ALKAN, C., KIDD, J.M., BIROL, I., EICHLER, E.E. and SAHINALP, S.C., 2010. Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. *Bioinformatics (Oxford, England),* **26**(10), pp. 1277-1283.

HASTINGS, P.J., IRA, G. and LUPSKI, J.R., 2009. A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS genetics,* **5**(1), pp. e1000327.

HASTINGS, P.J., LUPSKI, J.R., ROSENBERG, S.M. and IRA, G., 2009. Mechanisms of change in gene copy number. *Nature reviews.Genetics,* **10**(8), pp. 551-564.

HEINZEN, E.L., RADTKE, R.A., URBAN, T.J., CAVALLERI, G.L., DEPONDT, C., NEED, A.C., WALLEY, N.M., NICOLETTI, P., GE, D., CATARINO, C.B., DUNCAN, J.S., KASPERAVICIUTE, D., TATE, S.K., CABOCLO, L.O., SANDER, J.W., CLAYTON, L., LINNEY, K.N., SHIANNA, K.V., GUMBS, C.E., SMITH, J., CRONIN, K.D., MAIA, J.M., DOHERTY, C.P., PANDOLFO, M., LEPPERT, D., MIDDLETON, L.T., GIBSON, R.A., JOHNSON, M.R., MATTHEWS, P.M., HOSFORD, D., KALVIAINEN, R., ERIKSSON, K., KANTANEN, A.M., DORN, T., HANSEN, J., KRAMER, G., STEINHOFF, B.J., WIESER, H.G., ZUMSTEG, D., ORTEGA, M., WOOD, N.W., HUXLEY-JONES, J., MIKATI, M., GALLENTINE, W.B., HUSAIN, A.M., BUCKLEY, P.G., STALLINGS, R.L., PODGOREANU, M.V., DELANTY, N., SISODIYA, S.M. and GOLDSTEIN, D.B., 2010. Rare deletions at 16p13.11 predispose to a diverse spectrum of sporadic epilepsy syndromes. *American Journal of Human Genetics,* **86**(5), pp. 707-718.

HENRICHSEN, C.N., CHAIGNAT, E. and REYMOND, A., 2009. Copy number variants, diseases and gene expression. *Human molecular genetics,* **18**(R1), pp. R1-8.

HENRICHSEN, C.N., VINCKENBOSCH, N., ZOLLNER, S., CHAIGNAT, E., PRADERVAND, S., SCHUTZ, F., RUEDI, M., KAESSMANN, H. and REYMOND, A., 2009. Segmental copy number variation shapes tissue transcriptomes. *Nature genetics,* **41**(4), pp. 424-429.

HILL, A.V., ALLSOPP, C.E., KWIATKOWSKI, D., ANSTEY, N.M., TWUMASI, P., ROWE, P.A., BENNETT, S., BREWSTER, D., MCMICHAEL, A.J. and GREENWOOD, B.M., 1991. Common west African HLA antigens are associated with protection from severe malaria. *Nature,* **352**(6336), pp. 595-600.

HINDORFF, L.A., SETHUPATHY, P., JUNKINS, H.A., RAMOS, E.M., MEHTA, J.P., COLLINS, F.S. and MANOLIO, T.A., 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America,* **106**(23), pp. 9362-9367.

HO, M.R., TSAI, K.W., CHEN, C.H. and LIN, W.C., 2011. dbDNV: a resource of duplicated gene nucleotide variants in human genome. *Nucleic acids research,* **39**(Database issue), pp. D920-5.

HOLLOX, E.J., HUFFMEIER, U., ZEEUWEN, P.L., PALLA, R., LASCORZ, J., RODIJK-OLTHUIS, D., VAN DE KERKHOF, P.C., TRAUPE, H., DE JONGH, G., DEN HEIJER, M., REIS, A., ARMOUR, J.A. and SCHALKWIJK, J., 2008. Psoriasis is associated with increased beta-defensin genomic copy number. *Nature genetics,* **40**(1), pp. 23-25.

HUANG, X.S., XIAO, L., LI, X., XIE, Y., JIANG, H.O., TAN, C., WANG, L. and ZHANG, J.X., 2010. Two neighboring microdeletions of 5q13.2 in a child with oculo-auriculo-vertebral spectrum. *European journal of medical genetics,* **53**(3), pp. 153-158.

HUGOT, J.P., CHAMAILLARD, M., ZOUALI, H., LESAGE, S., CEZARD, J.P., BELAICHE, J., ALMER, S., TYSK, C., O'MORAIN, C.A., GASSULL, M., BINDER, V., FINKEL, Y., CORTOT, A., MODIGLIANI, R., LAURENT-PUIG, P., GOWER-ROUSSEAU, C., MACRY, J., COLOMBEL, J.F., SAHBATOU, M. and THOMAS, G., 2001. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature,***411**(6837), pp. 599-603.

IAFRATE, A.J., FEUK, L., RIVERA, M.N., LISTEWNIK, M.L., DONAHOE, P.K., QI, Y., SCHERER, S.W. and LEE, C., 2004. Detection of large-scale variation in the human genome. *Nature genetics,* **36**(9), pp. 949-951.

INOUE, A., MATSUMOTO, I., TANAKA, Y., IWANAMI, K., KANAMORI, A., OCHIAI, N., GOTO, D., ITO, S. and SUMIDA, T., 2009. Tumor necrosis factor alpha-induced adipose-related protein expression in experimental arthritis and in rheumatoid arthritis. *Arthritis research & therapy,* **11**(4), pp. R118.

INTERNATIONAL SCHIZOPHRENIA CONSORTIUM, 2008. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature,* **455**(7210), pp. 237-241.

ISENBERG, D.A., ALLEN, E., FAREWELL, V., EHRENSTEIN, M.R., HANNA, M.G., LUNDBERG, I.E., ODDIS, C., PILKINGTON, C., PLOTZ, P., SCOTT, D., VENCOVSKY, J., COOPER, R., RIDER, L., MILLER, F. and INTERNATIONAL MYOSITIS AND CLINICAL STUDIES GROUP (IMACS), 2004. International consensus outcome measures for patients with idiopathic inflammatory myopathies. Development and initial validation of myositis activity and damage indices in patients with adult onset disease. *Rheumatology (Oxford, England),* **43**(1), pp. 49-54.

ITSARA, A., COOPER, G.M., BAKER, C., GIRIRAJAN, S., LI, J., ABSHER, D., KRAUSS, R.M., MYERS, R.M., RIDKER, P.M., CHASMAN, D.I., MEFFORD, H., YING, P., NICKERSON, D.A. and EICHLER, E.E., 2009. Population analysis of large copy number variants and hotspots of human genetic disease. *American Journal of Human Genetics,* **84**(2), pp. 148-161.

JACQUEMONT, S., REYMOND, A., ZUFFEREY, F., HAREWOOD, L., WALTERS, R.G., KUTALIK, Z., MARTINET, D., SHEN, Y., VALSESIA, A., BECKMANN, N.D., THORLEIFSSON, G., BELFIORE, M., BOUQUILLON, S., CAMPION, D., DE LEEUW, N., DE VRIES, B.B., ESKO, T., FERNANDEZ, B.A., FERNANDEZ-ARANDA, F., FERNANDEZ-REAL, J.M., GRATACOS, M., GUILMATRE, A., HOYER, J., JARVELIN, M.R., KOOY, R.F., KURG, A., LE CAIGNEC, C., MANNIK, K., PLATT, O.S., SANLAVILLE, D., VAN HAELST, M.M., VILLATORO GOMEZ, S., WALHA, F., WU, B.L., YU, Y., ABOURA, A., ADDOR, M.C., ALEMBIK, Y., ANTONARAKIS, S.E., ARVEILER, B., BARTH, M., BEDNAREK, N., BENA, F., BERGMANN, S., BERI, M., BERNARDINI, L., BLAUMEISER, B., BONNEAU, D., BOTTANI, A., BOUTE, O., BRUNNER, H.G., CAILLEY, D., CALLIER, P., CHIESA, J., CHRAST, J., COIN, L., COUTTON, C., CUISSET, J.M., CUVELLIER, J.C., DAVID, A., DE FREMINVILLE, B., DELOBEL, B., DELRUE, M.A., DEMEER, B., DESCAMPS, D., DIDELOT, G., DIETERICH, K., DISCIGLIO, V., DOCO-FENZY, M., DRUNAT, S., DUBAN-BEDU, B., DUBOURG, C., EL-SAYED MOUSTAFA, J.S., ELLIOTT, P., FAAS, B.H., FAIVRE, L., FAUDET, A., FELLMANN, F., FERRARINI, A., FISHER, R., FLORI, E., FORER, L., GAILLARD, D., GERARD, M., GIEGER, C., GIMELLI, S., GIMELLI, G., GRABE, H.J., GUICHET, A., GUILLIN, O., HARTIKAINEN, A.L., HERON, D., HIPPOLYTE, L., HOLDER, M., HOMUTH, G., ISIDOR, B., JAILLARD, S., JAROS, Z., JIMENEZ-MURCIA, S., HELAS, G.J., JONVEAUX, P., KAKSONEN, S., KEREN, B., KLOSS-BRANDSTATTER, A., KNOERS, N.V., KOOLEN, D.A., KROISEL, P.M., KRONENBERG, F., LABALME, A., LANDAIS, E., LAPI, E., LAYET, V., LEGALLIC, S., LEHEUP, B., LEUBE, B., LEWIS, S., LUCAS, J., MACDERMOT, K.D., MAGNUSSON, P., MARSHALL, C., MATHIEU-

DRAMARD, M., MCCARTHY, M.I., MEITINGER, T., MENCARELLI, M.A., MERLA, G., MOERMAN, A., MOOSER, V., MORICE-PICARD, F., MUCCIOLO, M., NAUCK, M., NDIAYE, N.C., NORDGREN, A., PASQUIER, L., PETIT, F., PFUNDT, R., PLESSIS, G., RAJCAN-SEPAROVIC, E., RAMELLI, G.P., RAUCH, A., RAVAZZOLO, R., REIS, A., RENIERI, A., RICHART, C., RIED, J.S., RIEUBLAND, C., ROBERTS, W., ROETZER, K.M., ROORYCK, C., ROSSI, M., SAEMUNDSEN, E., SATRE, V., SCHURMANN, C., SIGURDSSON, E., STAVROPOULOS, D.J., STEFANSSON, H., TENGSTROM, C., THORSTEINSDOTTIR, U., TINAHONES, F.J., TOURAINE, R., VALLEE, L., VAN BINSBERGEN, E., VAN DER AA, N., VINCENT-DELORME, C., VISVIKIS-SIEST, S., VOLLENWEIDER, P., VOLZKE, H., VULTO-VAN SILFHOUT, A.T., WAEBER, G., WALLGREN-PETTERSSON, C., WITWICKI, R.M., ZWOLINKSI, S., ANDRIEUX, J., ESTIVILL, X., GUSELLA, J.F., GUSTAFSSON, O., METSPALU, A., SCHERER, S.W., STEFANSSON, K., BLAKEMORE, A.I., BECKMANN, J.S. and FROGUEL, P., 2011. Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. *Nature,* **478**(7367), pp. 97-102.

JAKOBSSON, M., SCHOLZ, S.W., SCHEET, P., GIBBS, J.R., VANLIERE, J.M., FUNG, H.C., SZPIECH, Z.A., DEGNAN, J.H., WANG, K., GUERREIRO, R., BRAS, J.M., SCHYMICK, J.C., HERNANDEZ, D.G., TRAYNOR, B.J., SIMON-SANCHEZ, J., MATARIN, M., BRITTON, A., VAN DE LEEMPUT, J., RAFFERTY, I., BUCAN, M., CANN, H.M., HARDY, J.A., ROSENBERG, N.A. and SINGLETON, A.B., 2008. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature,* **451**(7181), pp. 998-1003.

JARVINEN, P., 1995. Occurrence of ankylosing spondylitis in a nationwide series of twins. *Arthritis and Rheumatism,* **38**(3), pp. 381-383.

KALANITHI, P.S., ZHENG, W., KATAOKA, Y., DIFIGLIA, M., GRANTZ, H., SAPER, C.B., SCHWARTZ, M.L., LECKMAN, J.F. and VACCARINO, F.M., 2005. Altered parvalbumin-positive neuron distribution in basal ganglia of individuals with Tourette syndrome. *Proceedings of the National Academy of Sciences of the United States of America,* **102**(37), pp. 13307-13312.

KARVONEN, M., VIIK-KAJANDER, M., MOLTCHANOVA, E., LIBMAN, I., LAPORTE, R. and TUOMILEHTO, J., 2000. Incidence of childhood type 1 diabetes worldwide. Diabetes Mondiale (DiaMond) Project Group. *Diabetes care,* **23**(10), pp. 1516-1526.

KARYPIDIS, A.H., OLSSON, M., ANDERSSON, S.O., RANE, A. and EKSTROM, L., 2008. Deletion polymorphism of the UGT2B17 gene is associated with increased risk for prostate cancer and correlated to gene expression in the prostate. *The pharmacogenomics journal,* **8**(2), pp. 147-151.

KATAOKA, Y., KALANITHI, P.S., GRANTZ, H., SCHWARTZ, M.L., SAPER, C., LECKMAN, J.F. and VACCARINO, F.M., 2010. Decreased number of parvalbumin and cholinergic interneurons in the striatum of individuals with Tourette syndrome. *The Journal of comparative neurology,* **518**(3), pp. 277-291.

KENT, W.J., 2002. BLAT--the BLAST-like alignment tool. *Genome research,* **12**(4), pp. 656-664.

KIDD, J.M., COOPER, G.M., DONAHUE, W.F., HAYDEN, H.S., SAMPAS, N., GRAVES, T., HANSEN, N., TEAGUE, B., ALKAN, C., ANTONACCI, F., HAUGEN, E., ZERR, T., YAMADA, N.A., TSANG, P., NEWMAN, T.L., TUZUN, E., CHENG, Z., EBLING, H.M., TUSNEEM, N., DAVID, R., GILLETT, W., PHELPS, K.A., WEAVER, M., SARANGA, D., BRAND, A., TAO, W., GUSTAFSON, E., MCKERNAN, K., CHEN, L., MALIG, M., SMITH, J.D., KORN, J.M., MCCARROLL, S.A., ALTSHULER, D.A., PEIFFER, D.A., DORSCHNER, M., STAMATOYANNOPOULOS, J., SCHWARTZ, D., NICKERSON, D.A., MULLIKIN, J.C., WILSON, R.K., BRUHN, L., OLSON, M.V., KAUL, R., SMITH, D.R. and EICHLER, E.E., 2008. Mapping and sequencing of structural variation from eight human genomes. *Nature,* **453**(7191), pp. 56-64.

KOOLEN, D.A., VISSERS, L.E., PFUNDT, R., DE LEEUW, N., KNIGHT, S.J., REGAN, R., KOOY, R.F., REYNIERS, E., ROMANO, C., FICHERA, M., SCHINZEL, A., BAUMER, A., ANDERLID, B.M., SCHOUMANS, J., KNOERS, N.V., VAN KESSEL, A.G., SISTERMANS, E.A., VELTMAN, J.A., BRUNNER, H.G. and DE VRIES, B.B., 2006. A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. *Nature genetics,* **38**(9), pp. 999-1001.

KORBEL, J.O., URBAN, A.E., AFFOURTIT, J.P., GODWIN, B., GRUBERT, F., SIMONS, J.F., KIM, P.M., PALEJEV, D., CARRIERO, N.J., DU, L., TAILLON, B.E., CHEN, Z., TANZER, A., SAUNDERS, A.C., CHI, J., YANG, F., CARTER, N.P., HURLES, M.E., WEISSMAN, S.M., HARKINS, T.T., GERSTEIN, M.B., EGHOLM, M. and SNYDER, M., 2007. Paired-end mapping reveals extensive structural variation in the human genome.*Science (New York, N.Y.),* **318**(5849), pp. 420-426.

KOSCINSKI, I., ELINATI, E., FOSSARD, C., REDIN, C., MULLER, J., VELEZ DE LA CALLE, J., SCHMITT, F., BEN KHELIFA, M., RAY, P.F., KILANI, Z., BARRATT, C.L. and VIVILLE, S., 2011. DPY19L2 deletion as a major cause of globozoospermia. *American Journal of Human Genetics,* **88**(3), pp. 344-350.

KUMPS, C., VAN ROY, N., HEYRMAN, L., GOOSSENS, D., DEL-FAVERO, J., NOGUERA, R., VANDESOMPELE, J., SPELEMAN, F. and DE PRETER, K., 2010. Multiplex Amplicon Quantification (MAQ), a fast and efficient method for the simultaneous detection of copy number alterations in neuroblastoma. *BMC genomics,* **11**, pp. 298.

LAWSON-YUEN, A., SALDIVAR, J.S., SOMMER, S. and PICKER, J., 2008. Familial deletion within NLGN4 associated with autism and Tourette syndrome. *European journal of human genetics : EJHG,* **16**(5), pp. 614-618.

LEE, C., IAFRATE, A.J. and BROTHMAN, A.R., 2007. Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. *Nature genetics,* **39**(7 Suppl), pp. S48-54.

LEE, J.A., CARVALHO, C.M. and LUPSKI, J.R., 2007. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell,* **131**(7), pp. 1235-1247.

LEE, J.A. and LUPSKI, J.R., 2006. Genomic rearrangements and gene copy-number alterations as a cause of nervous system disorders. *Neuron,* **52**(1), pp. 103-121.

LEI, J., DENG, X., ZHANG, J., SU, L., XU, H., LIANG, H., HUANG, X., SONG, Z. and DENG, H., 2012. Mutation screening of the HDC gene in Chinese Han patients with Tourette syndrome. *American journal of medical genetics.Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics,* **159B**(1), pp. 72-76.

LESAGE, S., ZOUALI, H., CEZARD, J.P., COLOMBEL, J.F., BELAICHE, J., ALMER, S., TYSK, C., O'MORAIN, C., GASSULL, M., BINDER, V., FINKEL, Y., MODIGLIANI, R., GOWER-ROUSSEAU, C., MACRY, J., MERLIN, F., CHAMAILLARD, M., JANNOT, A.S., THOMAS, G., HUGOT, J.P., EPWG-IBD GROUP, EPIMAD GROUP and GETAID GROUP, 2002. CARD15/NOD2 mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease. *American Journal of Human Genetics,* **70**(4), pp. 845-857.

LI, H., RUAN, J. and DURBIN, R., 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research,* **18**(11), pp. 1851-1858.

LI, R., ZHU, H., RUAN, J., QIAN, W., FANG, X., SHI, Z., LI, Y., LI, S., SHAN, G., KRISTIANSEN, K., LI, S., YANG, H., WANG, J. and WANG, J., 2010. De novo assembly of human genomes with massively parallel short read sequencing. *Genome research,* **20**(2), pp. 265-272.

LIANG, J.S., SHIMOJIMA, K. and YAMAMOTO, T., 2008. Application of array-based comparative genome hybridization in children with developmental delay or mental retardation. *Pediatrics and neonatology,* **49**(6), pp. 213-217.

LIEBER, M.R., MA, Y., PANNICKE, U. and SCHWARZ, K., 2003. Mechanism and regulation of human non-homologous DNA end-joining. *Nature reviews.Molecular cell biology,* **4**(9), pp. 712-720.

LIU, Y., JIANG, L., CAI, Q., DANOY, P., BARNARDO, M.C., BROWN, M.A. and XU, H., 2010. Predominant association of HLA-B*2704 with ankylosing spondylitis in Chinese Han patients. *Tissue antigens,***75**(1), pp. 61-64.

LU, X., GUO, J., ZHOU, X., LI, R., LIU, X., ZHAO, Y., ZHU, B., LIU, X., XU, J., ZHU, P., WU, X., HE, J., LIU, X., ZHANG, H. and LI, Z., 2011. Deletion of LCE3C_LCE3B is associated with rheumatoid arthritis and systemic lupus erythematosus in the Chinese Han population. *Annals of the Rheumatic Diseases,* **70**(9), pp. 1648-1651.

MAGREY, M. and KHAN, M.A., 2010. Osteoporosis in ankylosing spondylitis. *Current rheumatology reports,* **12**(5), pp. 332-336.

MAILMAN, M.D., FEOLO, M., JIN, Y., KIMURA, M., TRYKA, K., BAGOUTDINOV, R., HAO, L., KIANG, A., PASCHALL, J., PHAN, L., POPOVA, N., PRETEL, S., ZIYABARI, L., LEE, M., SHAO, Y., WANG, Z.Y., SIROTKIN, K., WARD, M., KHOLODOV, M., ZBICZ, K., BECK, J., KIMELMAN, M., SHEVELEV, S., PREUSS, D., YASCHENKO, E., GRAEFF, A., OSTELL, J. and SHERRY, S.T., 2007. The NCBI dbGaP database of genotypes and phenotypes. *Nature genetics,* **39**(10), pp. 1181-1186.

MAMTANI, M., ROVIN, B., BREY, R., CAMARGO, J.F., KULKARNI, H., HERRERA, M., CORREA, P., HOLLIDAY, S., ANAYA, J.M. and AHUJA, S.K., 2008. CCL3L1 gene-containing segmental duplications and polymorphisms in CCR5 affect risk of systemic lupus erythaematosus. *Annals of the Rheumatic Diseases,* **67**(8), pp. 1076-1083.

MCCARROLL, S.A., HUETT, A., KUBALLA, P., CHILEWSKI, S.D., LANDRY, A., GOYETTE, P., ZODY, M.C., HALL, J.L., BRANT, S.R., CHO, J.H., DUERR, R.H., SILVERBERG, M.S., TAYLOR, K.D., RIOUX, J.D., ALTSHULER, D., DALY, M.J. and XAVIER, R.J., 2008. Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nature genetics,* **40**(9), pp. 1107-1112.

MCCARROLL, S.A., KURUVILLA, F.G., KORN, J.M., CAWLEY, S., NEMESH, J., WYSOKER, A., SHAPERO, M.H., DE BAKKER, P.I., MALLER, J.B., KIRBY, A., ELLIOTT, A.L., PARKIN, M., HUBBELL, E., WEBSTER, T., MEI, R., VEITCH, J., COLLINS, P.J., HANDSAKER, R., LINCOLN, S., NIZZARI, M., BLUME, J., JONES, K.W., RAVA, R., DALY, M.J., GABRIEL, S.B. and ALTSHULER, D., 2008. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature genetics,* **40**(10), pp. 1166-1174.

MCKINNEY, C., FANCIULLI, M., MERRIMAN, M.E., PHIPPS-GREEN, A., ALIZADEH, B.Z., KOELEMAN, B.P., DALBETH, N., GOW, P.J., HARRISON, A.A., HIGHTON, J., JONES, P.B., STAMP, L.K., STEER, S., BARRERA, P., COENEN, M.J., FRANKE, B., VAN RIEL, P.L., VYSE, T.J., AITMAN, T.J., RADSTAKE, T.R. and MERRIMAN, T.R., 2010. Association of variation in Fcgamma receptor 3B gene copy number with rheumatoid arthritis in Caucasian samples. *Annals of the Rheumatic Diseases,* **69**(9), pp. 1711-1716.

MCKINNEY, C., MERRIMAN, M.E., CHAPMAN, P.T., GOW, P.J., HARRISON, A.A., HIGHTON, J., JONES, P.B., MCLEAN, L., O'DONNELL, J.L., POKORNY, V., SPELLERBERG, M., STAMP, L.K., WILLIS, J., STEER, S. and MERRIMAN, T.R., 2008. Evidence for an influence of chemokine ligand 3-like 1 (CCL3L1) gene copy number on susceptibility to rheumatoid arthritis. *Annals of the Rheumatic Diseases,* **67**(3), pp. 409-413.

MEFFORD, H.C. and TRASK, B.J., 2002. The complex structure and dynamic evolution of human subtelomeres. *Nature reviews.Genetics,* **3**(2), pp. 91-102.

MI, H., DONG, Q., MURUGANUJAN, A., GAUDET, P., LEWIS, S. and THOMAS, P.D., 2010. PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic acids research,* **38**(Database issue), pp. D204-10.

MILLER, D.T., ADAM, M.P., ARADHYA, S., BIESECKER, L.G., BROTHMAN, A.R., CARTER, N.P., CHURCH, D.M., CROLLA, J.A., EICHLER, E.E., EPSTEIN, C.J., FAUCETT, W.A., FEUK, L., FRIEDMAN, J.M., HAMOSH, A., JACKSON, L., KAMINSKY, E.B., KOK, K., KRANTZ, I.D., KUHN, R.M., LEE, C., OSTELL, J.M., ROSENBERG, C., SCHERER, S.W., SPINNER, N.B., STAVROPOULOS, D.J., TEPPERBERG, J.H., THORLAND, E.C., VERMEESCH, J.R., WAGGONER, D.J., WATSON, M.S., MARTIN, C.L. and LEDBETTER, D.H., 2010. Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *American Journal of Human Genetics,* **86**(5), pp. 749-764.

MILLS, R.E., WALTER, K., STEWART, C., HANDSAKER, R.E., CHEN, K., ALKAN, C., ABYZOV, A., YOON, S.C., YE, K., CHEETHAM, R.K., CHINWALLA, A., CONRAD, D.F., FU, Y., GRUBERT, F., HAJIRASOULIHA, I., HORMOZDIARI, F., IAKOUCHEVA, L.M., IQBAL, Z., KANG, S., KIDD, J.M., KONKEL, M.K., KORN, J., KHURANA, E., KURAL, D., LAM, H.Y., LENG, J., LI, R., LI, Y., LIN, C.Y., LUO, R., MU, X.J., NEMESH, J., PECKHAM, H.E., RAUSCH, T., SCALLY, A., SHI, X., STROMBERG, M.P., STUTZ, A.M., URBAN, A.E., WALKER, J.A., WU, J., ZHANG, Y., ZHANG, Z.D., BATZER, M.A., DING, L., MARTH, G.T., MCVEAN, G., SEBAT, J., SNYDER, M., WANG, J., YE, K., EICHLER, E.E., GERSTEIN, M.B., HURLES, M.E., LEE, C., MCCARROLL, S.A., KORBEL, J.O. and 1000 GENOMES PROJECT, 2011. Mapping copy number variation by population-scale genome sequencing. *Nature,* **470**(7332), pp. 59-65.

MOLOKHIA, M., FANCIULLI, M., PETRETTO, E., PATRICK, A.L., MCKEIGUE, P., ROBERTS, A.L., VYSE, T.J. and AITMAN, T.J., 2011. FCGR3B copy number variation is associated with systemic lupus erythematosus risk in Afro-Caribbeans. *Rheumatology (Oxford, England),* **50**(7), pp. 1206-1210.

NAGAMANI, S.C., EREZ, A., BADER, P., LALANI, S.R., SCOTT, D.A., SCAGLIA, F., PLON, S.E., TSAI, C.H., REIMSCHISEL, T., ROEDER, E., MALPHRUS, A.D., ENG, P.A., HIXSON, P.M., KANG, S.H., STANKIEWICZ, P., PATEL, A. and CHEUNG, S.W., 2011. Phenotypic manifestations of copy number variation in chromosome 16p13.11. *European journal of human genetics : EJHG,* **19**(3), pp. 280-286.

NAIR, R.P., DUFFIN, K.C., HELMS, C., DING, J., STUART, P.E., GOLDGAR, D., GUDJONSSON, J.E., LI, Y., TEJASVI, T., FENG, B.J., RUETHER, A., SCHREIBER, S., WEICHENTHAL, M., GLADMAN, D., RAHMAN, P., SCHRODI, S.J., PRAHALAD, S., GUTHERY, S.L., FISCHER, J., LIAO, W., KWOK, P.Y., MENTER, A., LATHROP, G.M., WISE, C.A., BEGOVICH, A.B., VOORHEES, J.J., ELDER, J.T., KRUEGER, G.G., BOWCOCK, A.M., ABECASIS, G.R. and COLLABORATIVE ASSOCIATION STUDY OF PSORIASIS, 2009. Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. *Nature genetics,* **41**(2), pp. 199-204.

NALL, L., GULLIVER, W., CHARMLEY, P. and FARBER, E.M., 1999. Search for the psoriasis susceptibility gene: the Newfoundland Study. *Cutis; cutaneous medicine for the practitioner,* **64**(5), pp. 323-329.

NEED, A.C., GE, D., WEALE, M.E., MAIA, J., FENG, S., HEINZEN, E.L., SHIANNA, K.V., YOON, W., KASPERAVICIUTE, D., GENNARELLI, M., STRITTMATTER, W.J., BONVICINI, C., ROSSI, G., JAYATHILAKE, K., COLA, P.A., MCEVOY, J.P., KEEFE, R.S., FISHER, E.M., ST JEAN, P.L., GIEGLING, I., HARTMANN, A.M., MOLLER, H.J., RUPPERT, A., FRASER, G., CROMBIE, C., MIDDLETON, L.T., ST CLAIR, D., ROSES, A.D., MUGLIA, P., FRANCKS, C., RUJESCU, D., MELTZER, H.Y. and GOLDSTEIN, D.B., 2009. A genome-wide investigation of SNPs and CNVs in schizophrenia. *PLoS genetics,* **5**(2), pp. e1000373.

NORGREN, N., MATTSON, E., FORSGREN, L. and HOLMBERG, M., 2011. A high-penetrance form of late-onset torsion dystonia maps to a novel locus (DYT21) on chromosome 2q14.3-q21.3.*Neurogenetics,* **12**(2), pp. 137-143.

OLSSON, M., LINDSTROM, S., HAGGKVIST, B., ADAMI, H.O., BALTER, K., STATTIN, P., ASK, B., RANE, A., EKSTROM, L. and GRONBERG, H., 2008. The UGT2B17 gene deletion is not associated with prostate cancer risk. *The Prostate,* **68**(5), pp. 571-575.

OLUFEMI, S.E., GREEN, J.S., MANICKAM, P., GURU, S.C., AGARWAL, S.K., KESTER, M.B., DONG, Q., BURNS, A.L., SPIEGEL, A.M., MARX, S.J., COLLINS, F.S. and CHANDRASEKHARAPPA, S.C., 1998. Common ancestral mutation in the MEN1 gene is likely responsible for the prolactinoma variant of MEN1 (MEN1Burin) in four kindreds from Newfoundland. *Human mutation,* **11**(4), pp. 264-269.

O'ROAK, B.J., MORGAN, T.M., FISHMAN, D.O., SAUS, E., ALONSO, P., GRATACOS, M., ESTIVILL, X., TELTSH, O., KOHN, Y., KIDD, K.K., CHO, J., LIFTON, R.P. and STATE, M.W., 2010. Additional support for the association of SLITRK1 var321 and Tourette syndrome. *Molecular psychiatry,* **15**(5), pp. 447-450.

PARK, H., KIM, J.I., JU, Y.S., GOKCUMEN, O., MILLS, R.E., KIM, S., LEE, S., SUH, D., HONG, D., KANG, H.P., YOO, Y.J., SHIN, J.Y., KIM, H.J., YAVARTANOO, M., CHANG, Y.W., HA, J.S., CHONG, W., HWANG, G.R., DARVISHI, K., KIM, H., YANG, S.J., YANG, K.S., KIM, H., HURLES, M.E., SCHERER, S.W., CARTER, N.P., TYLER-SMITH, C., LEE, C. and SEO, J.S., 2010. Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. *Nature genetics,* **42**(5), pp. 400-405.

PAULS, D.L., HURST, C.R., KRUGER, S.D., LECKMAN, J.F., KIDD, K.K. and COHEN, D.J., 1986. Gilles de la Tourette's syndrome and attention deficit disorder with hyperactivity. Evidence against a genetic relationship. *Archives of General Psychiatry,* **43**(12), pp. 1177-1179.

PAULS, D.L. and LECKMAN, J.F., 1986. The inheritance of Gilles de la Tourette's syndrome and associated behaviors. Evidence for autosomal dominant transmission. *The New England journal of medicine,* **315**(16), pp. 993-997.

PAULS, D.L., RAYMOND, C.L., STEVENSON, J.M. and LECKMAN, J.F., 1991. A family study of Gilles de la Tourette syndrome. *American Journal of Human Genetics,* **48**(1), pp. 154-163.

PAULS, D.L., TOWBIN, K.E., LECKMAN, J.F., ZAHNER, G.E. and COHEN, D.J., 1986. Gilles de la Tourette's syndrome and obsessive-compulsive disorder. Evidence supporting a genetic relationship.*Archives of General Psychiatry,* **43**(12), pp. 1180-1182.

PEIFFER, D.A., LE, J.M., STEEMERS, F.J., CHANG, W., JENNIGES, T., GARCIA, F., HADEN, K., LI, J., SHAW, C.A., BELMONT, J., CHEUNG, S.W., SHEN, R.M., BARKER, D.L. and GUNDERSON, K.L., 2006. High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome research,* **16**(9), pp. 1136-1148.

PERRY, G.H., BEN-DOR, A., TSALENKO, A., SAMPAS, N., RODRIGUEZ-REVENGA, L., TRAN, C.W., SCHEFFER, A., STEINFELD, I., TSANG, P., YAMADA, N.A., PARK, H.S., KIM, J.I., SEO, J.S., YAKHINI, Z., LADERMAN, S., BRUHN, L. and LEE, C., 2008. The fine-scale and complex architecture of human copy-number variation. *American Journal of Human Genetics,* **82**(3), pp. 685-695.

PERSICO, A.M. and BOURGERON, T., 2006. Searching for ways out of the autism maze: genetic, epigenetic and environmental clues. *Trends in neurosciences,* **29**(7), pp. 349-358.

PINKEL, D., SEGRAVES, R., SUDAR, D., CLARK, S., POOLE, I., KOWBEL, D., COLLINS, C., KUO, W.L., CHEN, C., ZHAI, Y., DAIRKEE, S.H., LJUNG, B.M., GRAY, J.W. and ALBERTSON, D.G., 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature genetics,* **20**(2), pp. 207-211.

PINTO, D., DARVISHI, K., SHI, X., RAJAN, D., RIGLER, D., FITZGERALD, T., LIONEL, A.C., THIRUVAHINDRAPURAM, B., MACDONALD, J.R., MILLS, R., PRASAD, A., NOONAN, K., GRIBBLE, S., PRIGMORE, E., DONAHOE, P.K., SMITH, R.S., PARK, J.H., HURLES, M.E., CARTER, N.P., LEE, C., SCHERER, S.W. and FEUK, L., 2011. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nature biotechnology,* **29**(6), pp. 512-520.

PINTO, D., PAGNAMENTA, A.T., KLEI, L., ANNEY, R., MERICO, D., REGAN, R., CONROY, J., MAGALHAES, T.R., CORREIA, C., ABRAHAMS, B.S., ALMEIDA, J., BACCHELLI, E., BADER, G.D., BAILEY, A.J., BAIRD, G., BATTAGLIA, A., BERNEY, T., BOLSHAKOVA, N., BOLTE, S., BOLTON, P.F., BOURGERON, T., BRENNAN, S., BRIAN, J., BRYSON, S.E., CARSON, A.R., CASALLO, G., CASEY, J., CHUNG, B.H., COCHRANE, L., CORSELLO, C., CRAWFORD, E.L., CROSSETT, A., CYTRYNBAUM, C., DAWSON, G., DE JONGE, M., DELORME, R., DRMIC, I., DUKETIS, E., DUQUE, F., ESTES, A., FARRAR, P., FERNANDEZ, B.A., FOLSTEIN, S.E., FOMBONNE, E., FREITAG, C.M., GILBERT, J., GILLBERG, C., GLESSNER, J.T., GOLDBERG, J., GREEN, A., GREEN, J., GUTER, S.J., HAKONARSON, H., HERON, E.A., HILL, M., HOLT, R., HOWE, J.L., HUGHES, G., HUS, V., IGLIOZZI, R., KIM, C., KLAUCK, S.M., KOLEVZON, A., KORVATSKA, O., KUSTANOVICH, V., LAJONCHERE, C.M., LAMB, J.A., LASKAWIEC, M., LEBOYER, M., LE COUTEUR, A., LEVENTHAL, B.L., LIONEL, A.C., LIU, X.Q., LORD, C., LOTSPEICH, L., LUND, S.C., MAESTRINI, E., MAHONEY, W., MANTOULAN, C., MARSHALL, C.R., MCCONACHIE, H., MCDOUGLE, C.J., MCGRATH, J., MCMAHON, W.M., MERIKANGAS, A., MIGITA, O., MINSHEW, N.J., MIRZA, G.K., MUNSON, J., NELSON, S.F., NOAKES, C., NOOR, A., NYGREN, G., OLIVEIRA, G., PAPANIKOLAOU, K., PARR, J.R., PARRINI, B., PATON, T., PICKLES, A., PILORGE, M., PIVEN, J., PONTING, C.P., POSEY, D.J., POUSTKA, A., POUSTKA, F., PRASAD, A., RAGOUSSIS, J., RENSHAW, K., RICKABY, J., ROBERTS, W., ROEDER, K., ROGE, B., RUTTER, M.L., BIERUT, L.J., RICE, J.P., SALT, J., SANSOM, K., SATO, D., SEGURADO, R., SEQUEIRA, A.F., SENMAN, L., SHAH, N., SHEFFIELD, V.C., SOORYA, L., SOUSA, I., STEIN, O., SYKES, N., STOPPIONI, V., STRAWBRIDGE, C., TANCREDI, R., TANSEY, K., THIRUVAHINDRAPDURAM, B., THOMPSON, A.P., THOMSON, S., TRYFON, A., TSIANTIS, J., VAN ENGELAND, H., VINCENT, J.B., VOLKMAR, F., WALLACE, S., WANG, K., WANG, Z., WASSINK, T.H., WEBBER, C., WEKSBERG, R., WING, K., WITTEMEYER, K., WOOD, S., WU, J., YASPAN, B.L., ZURAWIECKI, D., ZWAIGENBAUM, L., BUXBAUM, J.D., CANTOR, R.M., COOK, E.H., COON, H., CUCCARO, M.L., DEVLIN, B., ENNIS, S., GALLAGHER, L., GESCHWIND, D.H., GILL, M., HAINES, J.L., HALLMAYER, J., MILLER, J., MONACO, A.P., NURNBERGER, J.I.,JR, PATERSON, A.D., PERICAK-VANCE, M.A., SCHELLENBERG, G.D., SZATMARI, P., VICENTE, A.M., VIELAND, V.J., WIJSMAN, E.M., SCHERER, S.W., SUTCLIFFE, J.S. and BETANCUR, C., 2010. Functional impact of global rare copy number variation in autism spectrum disorders. *Nature,* **466**(7304), pp. 368-372.

PLENGE, R.M., COTSAPAS, C., DAVIES, L., PRICE, A.L., DE BAKKER, P.I., MALLER, J., PE'ER, I., BURTT, N.P., BLUMENSTIEL, B., DEFELICE, M., PARKIN, M., BARRY, R., WINSLOW, W., HEALY, C., GRAHAM, R.R., NEALE, B.M., IZMAILOVA, E., ROUBENOFF, R., PARKER, A.N., GLASS, R., KARLSON, E.W., MAHER, N., HAFLER, D.A., LEE, D.M., SELDIN, M.F., REMMERS, E.F., LEE, A.T., PADYUKOV, L., ALFREDSSON, L., COBLYN, J., WEINBLATT, M.E., GABRIEL, S.B., PURCELL, S., KLARESKOG, L., GREGERSEN, P.K., SHADICK, N.A., DALY, M.J. and ALTSHULER, D., 2007. Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. *Nature genetics,* **39**(12), pp. 1477-1482.

POOT, M., ELEVELD, M.J., VAN 'T SLOT, R., VAN GENDEREN, M.M., VERRIJN STUART, A.A., HOCHSTENBACH, R. and BEEMER, F.A., 2007. Proportional growth failure and oculocutaneous albinism in a girl with a 6.87 MBp deletion of region 15q26.2-->qter. *European journal of medical genetics,* **50**(6), pp. 432-440.

PRICE, R.A., KIDD, K.K., COHEN, D.J., PAULS, D.L. and LECKMAN, J.F., 1985. A twin study of Tourette syndrome. *Archives of General Psychiatry,* **42**(8), pp. 815-820.

PRINGSHEIM, T., FREEMAN, R. and LANG, A., 2007. Tourette syndrome and dystonia. *Journal of neurology, neurosurgery, and psychiatry,* **78**(5), pp. 544.

RABBEE, N. and SPEED, T.P., 2006. A genotype calling algorithm for affymetrix SNP arrays. *Bioinformatics (Oxford, England),* **22**(1), pp. 7-12.

RAHMAN, P., JONES, A., CURTIS, J., BARTLETT, S., PEDDLE, L., FERNANDEZ, B.A. and FREIMER, N.B., 2003. The Newfoundland population: a unique resource for genetic investigation of complex diseases. *Human molecular genetics,* **12 Spec No 2**, pp. R167-72.

RAYCHAUDHURI, S., THOMSON, B.P., REMMERS, E.F., EYRE, S., HINKS, A., GUIDUCCI, C., CATANESE, J.J., XIE, G., STAHL, E.A., CHEN, R., ALFREDSSON, L., AMOS, C.I., ARDLIE, K.G., BIRAC CONSORTIUM, BARTON, A., BOWES, J., BURTT, N.P., CHANG, M., COBLYN, J., COSTENBADER, K.H., CRISWELL, L.A., CRUSIUS, J.B., CUI, J., DE JAGER, P.L., DING, B., EMERY, P., FLYNN, E., HARRISON, P., HOCKING, L.J., HUIZINGA, T.W., KASTNER, D.L., KE, X., KURREEMAN, F.A., LEE, A.T., LIU, X., LI, Y., MARTIN, P., MORGAN, A.W., PADYUKOV, L., REID, D.M., SEIELSTAD, M., SELDIN, M.F., SHADICK, N.A., STEER, S., TAK, P.P., THOMSON, W., VAN DER HELM-VAN MIL, A.H., VAN DER HORST-BRUINSMA, I.E., WEINBLATT, M.E., WILSON, A.G., WOLBINK, G.J., WORDSWORTH, P., YEAR CONSORTIUM, ALTSHULER, D., KARLSON, E.W., TOES, R.E., DE VRIES, N., BEGOVICH, A.B., SIMINOVITCH, K.A., WORTHINGTON, J., KLARESKOG, L., GREGERSEN, P.K., DALY, M.J. and PLENGE, R.M., 2009. Genetic variants at CD28, PRDM1 and CD2/CD58 are associated with rheumatoid arthritis risk. *Nature genetics,* **41**(12), pp. 1313-1318.

REDDY, K.V., SERIO, K.J., HODULIK, C.R. and BIGBY, T.D., 2003. 5-lipoxygenase-activating protein gene expression. Key role of CCAAT/enhancer-binding proteins (C/EBP) in constitutive and tumor necrosis factor (TNF) alpha-induced expression in THP-1 cells. *The Journal of biological chemistry,* **278**(16), pp. 13810-13818.

REDON, R., ISHIKAWA, S., FITCH, K.R., FEUK, L., PERRY, G.H., ANDREWS, T.D., FIEGLER, H., SHAPERO, M.H., CARSON, A.R. and CHEN, W., 2006. Global variation in copy number in the human genome.*Nature,* **444**(7118), pp. 444-454.

ROBERTSON, M.M., 2008. The prevalence and epidemiology of Gilles de la Tourette syndrome. Part 2: tentative explanations for differing prevalence figures in GTS, including the possible effects of psychopathology, aetiology, cultural differences, and differing phenotypes. *Journal of psychosomatic research,* **65**(5), pp. 473-486.

ROBERTSON, M.M., EAPEN, V. and CAVANNA, A.E., 2009. The international prevalence, epidemiology, and clinical phenomenology of Tourette syndrome: a cross-cultural perspective. *Journal of psychosomatic research,* **67**(6), pp. 475-483.

SATO, D., LIONEL, A.C., LEBLOND, C.S., PRASAD, A., PINTO, D., WALKER, S., O'CONNOR, I., RUSSELL, C., DRMIC, I.E., HAMDAN, F.F., MICHAUD, J.L., ENDRIS, V., ROETH, R., DELORME, R., HUGUET, G., LEBOYER, M., RASTAM, M., GILLBERG, C., LATHROP, M., STAVROPOULOS, D.J., ANAGNOSTOU, E., WEKSBERG, R., FOMBONNE, E., ZWAIGENBAUM, L., FERNANDEZ, B.A., ROBERTS, W., RAPPOLD, G.A., MARSHALL, C.R., BOURGERON, T., SZATMARI, P. and SCHERER, S.W., 2012. SHANK1 Deletions in Males with Autism Spectrum Disorder. *American Journal of Human Genetics,* **90**(5), pp. 879-887.

SCHAEFFELER, E., SCHWAB, M., EICHELBAUM, M. and ZANGER, U.M., 2003. CYP2D6 genotyping strategy based on gene copy number determination by TaqMan real-time PCR. *Human mutation,* **22**(6), pp. 476-485.

SCHARF, J.M., MOORJANI, P., FAGERNESS, J., PLATKO, J.V., ILLMANN, C., GALLOWAY, B., JENIKE, E., STEWART, S.E., PAULS, D.L. and TOURETTE SYNDROME INTERNATIONAL CONSORTIUM FOR GENETICS, 2008. Lack of association between SLITRK1var321 and Tourette syndrome in a large family-based sample. *Neurology,* **70**(16 Pt 2), pp. 1495-1496.

SCHMITZ, R., HANSMANN, M.L., BOHLE, V., MARTIN-SUBERO, J.I., HARTMANN, S., MECHTERSHEIMER, G., KLAPPER, W., VATER, I., GIEFING, M., GESK, S., STANELLE, J., SIEBERT, R. and KUPPERS, R., 2009. TNFAIP3 (A20) is a tumor suppressor gene in Hodgkin lymphoma and primary mediastinal B cell lymphoma. *The Journal of experimental medicine,* **206**(5), pp. 981-989.

SCHOUTEN, J.P., MCELGUNN, C.J., WAAIJER, R., ZWIJNENBURG, D., DIEPVENS, F. and PALS, G., 2002. Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic acids research,* **30**(12), pp. e57.

SCHULZE, J.J., LUNDMARK, J., GARLE, M., SKILVING, I., EKSTROM, L. and RANE, A., 2008. Doping test results dependent on genotype of uridine diphospho-glucuronosyl transferase 2B17, the major enzyme for testosterone glucuronidation. *The Journal of clinical endocrinology and metabolism,* **93**(7), pp. 2500-2506.

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, T.C., TRASK, B., PATTERSON, N., ZETTERBERG, A. and WIGLER, M., 2004. Large-scale copy number polymorphism in the human genome. *Science (New York, N.Y.),* **305**(5683), pp. 525-528.

SHAFFER, L.G. and AMERICAN COLLEGE OF MEDICAL GENETICS PROFESSIONAL PRACTICE AND GUIDELINES COMMITTEE, 2005. American College of Medical Genetics guideline on the cytogenetic evaluation of the individual with developmental delay or mental retardation. *Genetics in medicine : official journal of the American College of Medical Genetics,* **7**(9), pp. 650-654.

SHAFFER, L.G., COPPINGER, J., ALLIMAN, S., TORCHIA, B.A., THEISEN, A., BALLIF, B.C. and BEJJANI, B.A., 2008. Comparison of microarray-based detection rates for cytogenetic abnormalities in prenatal and neonatal specimens. *Prenatal diagnosis,* **28**(9), pp. 789-795.

SHAFFER, L.G. and LUPSKI, J.R., 2000. Molecular mechanisms for constitutional chromosomal rearrangements in humans. *Annual Review of Genetics,* **34**, pp. 297-329.

SHAIKH, T.H., O'CONNOR, R.J., PIERPONT, M.E., MCGRATH, J., HACKER, A.M., NIMMAKAYALU, M., GEIGER, E., EMANUEL, B.S. and SAITTA, S.C., 2007. Low copy repeats mediate distal chromosome 22q11.2 deletions: sequence analysis predicts breakpoint mechanisms. *Genome research,* **17**(4), pp. 482-491.

SHARP, A.J., HANSEN, S., SELZER, R.R., CHENG, Z., REGAN, R., HURST, J.A., STEWART, H., PRICE, S.M., BLAIR, E., HENNEKAM, R.C., FITZPATRICK, C.A., SEGRAVES, R., RICHMOND, T.A., GUIVER, C., ALBERTSON, D.G., PINKEL, D., EIS, P.S., SCHWARTZ, S., KNIGHT, S.J. and EICHLER, E.E., 2006. Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nature genetics,* **38**(9), pp. 1038-1042.

SHARP, A.J., LOCKE, D.P., MCGRATH, S.D., CHENG, Z., BAILEY, J.A., VALLENTE, R.U., PERTZ, L.M., CLARK, R.A., SCHWARTZ, S., SEGRAVES, R., OSEROFF, V.V., ALBERTSON, D.G., PINKEL, D. and EICHLER, E.E., 2005. Segmental duplications and copy-number variation in the human genome. *American Journal of Human Genetics,* **77**(1), pp. 78-88.

SHARP, A.J., MEFFORD, H.C., LI, K., BAKER, C., SKINNER, C., STEVENSON, R.E., SCHROER, R.J., NOVARA, F., DE GREGORI, M., CICCONE, R., BROOMER, A., CASUGA, I., WANG, Y., XIAO, C., BARBACIORU, C., GIMELLI, G., BERNARDINA, B.D., TORNIERO, C., GIORDA, R., REGAN, R., MURDAY, V., MANSOUR, S., FICHERA, M., CASTIGLIA, L., FAILLA, P., VENTURA, M., JIANG, Z., COOPER, G.M., KNIGHT, S.J., ROMANO, C., ZUFFARDI, O., CHEN, C., SCHWARTZ, C.E. and EICHLER, E.E., 2008. A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. *Nature genetics,***40**(3), pp. 322-328.

SHARP, A.J., SELZER, R.R., VELTMAN, J.A., GIMELLI, S., GIMELLI, G., STRIANO, P., COPPOLA, A., REGAN, R., PRICE, S.M., KNOERS, N.V., EIS, P.S., BRUNNER, H.G., HENNEKAM, R.C., KNIGHT, S.J., DE VRIES, B.B., ZUFFARDI, O. and EICHLER, E.E., 2007. Characterization of a recurrent 15q24 microdeletion syndrome. *Human molecular genetics,* **16**(5), pp. 567-572.

SHAW, C.J. and LUPSKI, J.R., 2004. Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. *Human molecular genetics,* **13 Spec No 1**, pp. R57-64.

SHAW-SMITH, C., PITTMAN, A.M., WILLATT, L., MARTIN, H., RICKMAN, L., GRIBBLE, S., CURLEY, R., CUMMING, S., DUNN, C., KALAITZOPOULOS, D., PORTER, K., PRIGMORE, E., KREPISCHI-SANTOS, A.C., VARELA, M.C., KOIFFMANN, C.P., LEES, A.J., ROSENBERG, C., FIRTH, H.V., DE SILVA, R. and CARTER, N.P., 2006. Microdeletion encompassing MAPT at chromosome 17q21.3 is associated with developmental delay and learning disability. *Nature genetics,* **38**(9), pp. 1032-1037.

SHE, X., HORVATH, J.E., JIANG, Z., LIU, G., FUREY, T.S., CHRIST, L., CLARK, R., GRAVES, T., GULDEN, C.L., ALKAN, C., BAILEY, J.A., SAHINALP, C., ROCCHI, M., HAUSSLER, D., WILSON, R.K., MILLER, W., SCHWARTZ, S. and EICHLER, E.E., 2004. The structure and evolution of centromeric transition regions within the human genome. *Nature,* **430**(7002), pp. 857-864.

SHEVELL, M., ASHWAL, S., DONLEY, D., FLINT, J., GINGOLD, M., HIRTZ, D., MAJNEMER, A., NOETZEL, M., SHETH, R.D., QUALITY STANDARDS SUBCOMMITTEE OF THE AMERICAN ACADEMY OF NEUROLOGY and PRACTICE COMMITTEE OF THE CHILD NEUROLOGY SOCIETY, 2003. Practice parameter: evaluation of the child with global developmental delay: report of the Quality Standards Subcommittee of the American Academy of Neurology and The Practice Committee of the Child Neurology Society. *Neurology,* **60**(3), pp. 367-380.

SIGURDSSON, S., PADYUKOV, L., KURREEMAN, F.A., LILJEDAHL, U., WIMAN, A.C., ALFREDSSON, L., TOES, R., RONNELID, J., KLARESKOG, L., HUIZINGA, T.W., ALM, G., SYVANEN, A.C. and RONNBLOM, L., 2007. Association of a haplotype in the promoter region of the interferon regulatory factor 5 gene with rheumatoid arthritis. *Arthritis and Rheumatism,* **56**(7), pp. 2202-2210.

SIMPSON, J.T., WONG, K., JACKMAN, S.D., SCHEIN, J.E., JONES, S.J. and BIROL, I., 2009. ABySS: a parallel assembler for short read sequence data. *Genome research,* **19**(6), pp. 1117-1123.

SKALETSKY, H., KURODA-KAWAGUCHI, T., MINX, P.J., CORDUM, H.S., HILLIER, L., BROWN, L.G., REPPING, S., PYNTIKOVA, T., ALI, J., BIERI, T., CHINWALLA, A., DELEHAUNTY, A., DELEHAUNTY, K., DU, H., FEWELL, G., FULTON, L., FULTON, R., GRAVES, T., HOU, S.F., LATRIELLE, P., LEONARD, S., MARDIS, E., MAUPIN, R., MCPHERSON, J., MINER, T., NASH, W., NGUYEN, C., OZERSKY, P., PEPIN, K., ROCK, S., ROHLFING, T., SCOTT, K., SCHULTZ, B., STRONG, C., TIN-WOLLAM, A., YANG, S.P., WATERSTON, R.H., WILSON, R.K., ROZEN, S. and PAGE, D.C., 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature,* **423**(6942), pp. 825-837.

SMITH, T.F. and WATERMAN, M.S., 1981. Identification of common molecular subsequences. *Journal of Molecular Biology,* **147**(1), pp. 195-197.

SPIRIO, L., GREEN, J., ROBERTSON, J., ROBERTSON, M., OTTERUD, B., SHELDON, J., HOWSE, E., GREEN, R., GRODEN, J., WHITE, R. and LEPPERT, M., 1999. The identical 5' splice-site acceptor mutation in five attenuated APC families from Newfoundland demonstrates a founder effect. *Human genetics,* **105**(5), pp. 388-398.

STATE, M.W., 2011. The genetics of Tourette disorder. *Current opinion in genetics & development,* **21**(3), pp. 302-309.

STATE, M.W., 2010. The genetics of child psychiatric disorders: focus on autism and Tourette syndrome. *Neuron,* **68**(2), pp. 254-269.

STRANGER, B.E., STAHL, E.A. and RAJ, T., 2011. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics,* **187**(2), pp. 367-383.

SUDMANT, P.H., KITZMAN, J.O., ANTONACCI, F., ALKAN, C., MALIG, M., TSALENKO, A., SAMPAS, N., BRUHN, L., SHENDURE, J., 1000 GENOMES PROJECT and EICHLER, E.E., 2010. Diversity of human copy number variation and multicopy genes. *Science (New York, N.Y.),* **330**(6004), pp. 641-646.

SUNDARAM, S.K., HUQ, A.M., WILSON, B.J. and CHUGANI, H.T., 2010. Tourette syndrome is associated with recurrent exonic copy number variants. *Neurology,* **74**(20), pp. 1583-1590.

TAGAWA, H., SUGURO, M., TSUZUKI, S., MATSUO, K., KARNAN, S., OHSHIMA, K., OKAMOTO, M., MORISHIMA, Y., NAKAMURA, S. and SETO, M., 2005. Comparison of genome profiles for identification of distinct subgroups of diffuse large B-cell lymphoma. *Blood,* **106**(5), pp. 1770-1777.

TAKATA, Y., HAMADA, D., MIYATAKE, K., NAKANO, S., SHINOMIYA, F., SCAFE, C.R., REEVE, V.M., OSABE, D., MORITANI, M., KUNIKA, K., KAMATANI, N., INOUE, H., YASUI, N. and ITAKURA, M., 2007. Genetic association between the PRKCH gene encoding protein kinase Ceta isozyme and rheumatoid arthritis in the Japanese population. *Arthritis and Rheumatism,* **56**(1), pp. 30-42.

THABET, M.M., HUIZINGA, T.W., MARQUES, R.B., STOEKEN-RIJSBERGEN, G., BAKKER, A.M., KURREEMAN, F.A., WHITE, S.J., TOES, R.E. and VAN DER HELM-VAN MIL, A.H., 2009. Contribution of Fcgamma receptor IIIA gene 158V/F polymorphism and copy number variation to the risk of ACPA-positive rheumatoid arthritis. *Annals of the Rheumatic Diseases,* **68**(11), pp. 1775-1780.

TOKUTOMI, T., WADA, T., NAKAGAWA, E., SAITOH, S. and SASAKI, M., 2009. A de novo direct duplication of 16q22.1 --> q23.1 in a boy with midface hypoplasia and mental retardation. *American journal of medical genetics.Part A,* **149A**(11), pp. 2560-2563.

TUCKER, T., MONTPETIT, A., CHAI, D., CHAN, S., CHENIER, S., COE, B.P., DELANEY, A., EYDOUX, P., LAM, W.L., LANGLOIS, S., LEMYRE, E., MARRA, M., QIAN, H., ROULEAU, G.A., VINCENT, D., MICHAUD, J.L. and FRIEDMAN, J.M., 2011. Comparison of genome-wide array genomic hybridization platforms for the detection of copy number variants in idiopathic mental retardation. *BMC medical genomics,***4**, pp. 25.

TURER, E.E., TAVARES, R.M., MORTIER, E., HITOTSUMATSU, O., ADVINCULA, R., LEE, B., SHIFRIN, N., MALYNN, B.A. and MA, A., 2008. Homeostatic MyD88-dependent signals cause lethal inflamMation in the absence of A20. *The Journal of experimental medicine,* **205**(2), pp. 451-464.

TURNER, D.J., MIRETTI, M., RAJAN, D., FIEGLER, H., CARTER, N.P., BLAYNEY, M.L., BECK, S. and HURLES, M.E., 2008. Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. *Nature genetics,* **40**(1), pp. 90-95.

TUZUN, E., SHARP, A.J., BAILEY, J.A., KAUL, R., MORRISON, V.A., PERTZ, L.M., HAUGEN, E., HAYDEN, H., ALBERTSON, D., PINKEL, D., OLSON, M.V. and EICHLER, E.E., 2005. Fine-scale structural variation of the human genome. *Nature genetics,* **37**(7), pp. 727-732.

UDDIN, M., STURGE, M., PEDDLE, L., O'RIELLY, D.D. and RAHMAN, P., 2011. Genome-wide signatures of 'rearrangement hotspots' within segmental duplications in humans. *PloS one,* **6**(12), pp. e28853.

UDDIN, M., STURGE, M., RAHMAN, P. and WOODS, M.O., 2011. Autosome-wide copy number variation association analysis for rheumatoid arthritis using the WTCCC high-density SNP genotype data.*The Journal of rheumatology,* **38**(5), pp. 797-801.

VAN DER LINDEN, S., VALKENBURG, H. and CATS, A., 1983. The risk of developing ankylosing spondylitis in HLA-B27 positive individuals: a family and population study. *British journal of rheumatology,***22**(4 Suppl 2), pp. 18-19.

VEENSTRA-VANDERWEELE, J. and COOK, E.H.,JR, 2004. Molecular genetics of autism spectrum disorder. *Molecular psychiatry,* **9**(9), pp. 819-832.

VERKERK, A.J., MATHEWS, C.A., JOOSSE, M., EUSSEN, B.H., HEUTINK, P., OOSTRA, B.A. and TOURETTE SYNDROME ASSOCIATION INTERNATIONAL CONSORTIUM FOR GENETICS, 2003. CNTNAP2 is disrupted in a family with Gilles de la Tourette syndrome and obsessive compulsive disorder. *Genomics,* **82**(1), pp. 1-9.

VERMEIRE, S., WILD, G., KOCHER, K., COUSINEAU, J., DUFRESNE, L., BITTON, A., LANGELIER, D., PARE, P., LAPOINTE, G., COHEN, A., DALY, M.J. and RIOUX, J.D., 2002. CARD15 genetic variation in a Quebec population: prevalence, genotype-phenotype relationship, and haplotype structure. *American Journal of Human Genetics,* **71**(1), pp. 74-83.

VOLIK, S., ZHAO, S., CHIN, K., BREBNER, J.H., HERNDON, D.R., TAO, Q., KOWBEL, D., HUANG, G., LAPUK, A., KUO, W.L., MAGRANE, G., DE JONG, P., GRAY, J.W. and COLLINS, C., 2003. End-sequence profiling: sequence-based analysis of aberrant genomes. *Proceedings of the National Academy of Sciences of the United States of America,* **100**(13), pp. 7696-7701.

VOSSE, D., LANDEWE, R., VAN DER HEIJDE, D., VAN DER LINDEN, S., VAN STAA, T.P. and GEUSENS, P., 2009. Ankylosing spondylitis and the risk of fracture: results from a large primary care-based nested case-control study. *Annals of the Rheumatic Diseases,* **68**(12), pp. 1839-1842.

WALKUP, J.T., LECKMAN, J.F., PRICE, R.A., HARDIN, M., ORT, S.I. and COHEN, D.J., 1988. The relationship between obsessive-compulsive disorder and Tourette's syndrome: a twin study.*Psychopharmacology bulletin,* **24**(3), pp. 375-379.

WALTERS, R.G., JACQUEMONT, S., VALSESIA, A., DE SMITH, A.J., MARTINET, D., ANDERSSON, J., FALCHI, M., CHEN, F., ANDRIEUX, J., LOBBENS, S., DELOBEL, B., STUTZMANN, F., EL-SAYED MOUSTAFA, J.S., CHEVRE, J.C., LECOEUR, C., VATIN, V., BOUQUILLON, S., BUXTON, J.L., BOUTE, O., HOLDER-ESPINASSE, M., CUISSET, J.M., LEMAITRE, M.P., AMBRESIN, A.E., BRIOSCHI, A., GAILLARD, M., GIUSTI, V., FELLMANN, F., FERRARINI, A., HADJIKHANI, N., CAMPION, D., GUILMATRE, A., GOLDENBERG, A., CALMELS, N., MANDEL, J.L., LE CAIGNEC, C., DAVID, A., ISIDOR, B., CORDIER, M.P., DUPUIS-GIROD, S., LABALME, A., SANLAVILLE, D., BERI-DEXHEIMER, M., JONVEAUX, P., LEHEUP, B., OUNAP, K., BOCHUKOVA, E.G., HENNING, E., KEOGH, J., ELLIS, R.J., MACDERMOT, K.D., VAN HAELST, M.M., VINCENT-DELORME, C., PLESSIS, G., TOURAINE, R., PHILIPPE, A., MALAN, V., MATHIEU-DRAMARD, M., CHIESA, J., BLAUMEISER, B., KOOY, R.F., CAIAZZO, R., PIGEYRE, M., BALKAU, B., SLADEK, R., BERGMANN, S., MOOSER, V., WATERWORTH, D., REYMOND, A., VOLLENWEIDER, P., WAEBER, G., KURG, A., PALTA, P., ESKO, T., METSPALU, A., NELIS, M., ELLIOTT, P., HARTIKAINEN, A.L., MCCARTHY, M.I., PELTONEN, L., CARLSSON, L., JACOBSON, P., SJOSTROM, L., HUANG, N., HURLES, M.E., O'RAHILLY, S., FAROOQI, I.S., MANNIK, K., JARVELIN, M.R., PATTOU, F., MEYRE, D., WALLEY, A.J., COIN, L.J., BLAKEMORE, A.I., FROGUEL, P. and BECKMANN, J.S., 2010. A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. *Nature,* **463**(7281), pp. 671-675.

WANG, K., LI, M., HADLEY, D., LIU, R., GLESSNER, J., GRANT, S.F., HAKONARSON, H. and BUCAN, M., 2007. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome research,* **17**(11), pp. 1665-1674.

WEKSBERG, R., HUGHES, S., MOLDOVAN, L., BASSETT, A.S., CHOW, E.W. and SQUIRE, J.A., 2005. A method for accurate detection of genomic microdeletions using real-time quantitative PCR. *BMC genomics,* **6**, pp. 180.

WELLCOME TRUST CASE CONTROL CONSORTIUM, 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature,* **447**(7145), pp. 661-678.

WELLCOME TRUST CASE CONTROL CONSORTIUM, CRADDOCK, N., HURLES, M.E., CARDIN, N., PEARSON, R.D., PLAGNOL, V., ROBSON, S., VUKCEVIC, D., BARNES, C., CONRAD, D.F., GIANNOULATOU, E., HOLMES, C., MARCHINI, J.L., STIRRUPS, K., TOBIN, M.D., WAIN, L.V., YAU, C., AERTS, J., AHMAD, T., ANDREWS, T.D., ARBURY, H., ATTWOOD, A., AUTON, A., BALL, S.G., BALMFORTH, A.J., BARRETT, J.C., BARROSO, I., BARTON, A., BENNETT, A.J., BHASKAR, S., BLASZCZYK, K., BOWES, J., BRAND, O.J., BRAUND, P.S., BREDIN, F., BREEN, G., BROWN, M.J., BRUCE, I.N., BULL, J., BURREN, O.S., BURTON, J., BYRNES, J., CAESAR, S., CLEE, C.M., COFFEY, A.J., CONNELL, J.M., COOPER, J.D., DOMINICZAK, A.F., DOWNES, K., DRUMMOND, H.E., DUDAKIA, D., DUNHAM, A., EBBS, B., ECCLES, D., EDKINS, S., EDWARDS, C., ELLIOT, A., EMERY, P., EVANS, D.M., EVANS, G., EYRE, S., FARMER, A., FERRIER, I.N., FEUK, L., FITZGERALD, T., FLYNN, E., FORBES, A., FORTY, L., FRANKLYN, J.A., FREATHY, R.M., GIBBS, P., GILBERT, P., GOKUMEN, O., GORDON-SMITH, K., GRAY, E., GREEN, E., GROVES, C.J., GROZEVA, D., GWILLIAM, R., HALL, A., HAMMOND, N., HARDY, M., HARRISON, P., HASSANALI, N., HEBAISHI, H., HINES, S., HINKS, A., HITMAN, G.A., HOCKING, L., HOWARD, E., HOWARD, P., HOWSON, J.M., HUGHES, D., HUNT, S., ISAACS, J.D., JAIN, M., JEWELL, D.P., JOHNSON, T., JOLLEY, J.D., JONES, I.R., JONES, L.A., KIROV, G., LANGFORD, C.F., LANGO-ALLEN, H., LATHROP, G.M., LEE, J., LEE, K.L., LEES, C., LEWIS, K., LINDGREN, C.M., MAISURIA-ARMER, M., MALLER, J., MANSFIELD, J., MARTIN, P., MASSEY, D.C., MCARDLE, W.L., MCGUFFIN, P., MCLAY, K.E., MENTZER, A., MIMMACK, M.L., MORGAN, A.E., MORRIS, A.P., MOWAT, C., MYERS, S., NEWMAN, W., NIMMO, E.R., O'DONOVAN, M.C., ONIPINLA, A., ONYIAH, I., OVINGTON, N.R., OWEN, M.J., PALIN, K., PARNELL, K., PERNET, D., PERRY, J.R., PHILLIPS, A., PINTO, D., PRESCOTT, N.J., PROKOPENKO, I., QUAIL, M.A., RAFELT, S., RAYNER, N.W., REDON, R., REID, D.M., RENWICK, RING, S.M., ROBERTSON, N., RUSSELL, E., ST CLAIR, D., SAMBROOK, J.G., SANDERSON, J.D., SCHUILENBURG, H., SCOTT, C.E., SCOTT, R., SEAL, S., SHAW-HAWKINS, S., SHIELDS, B.M., SIMMONDS, M.J., SMYTH, D.J., SOMASKANTHARAJAH, E., SPANOVA, K., STEER, S., STEPHENS, J., STEVENS, H.E., STONE, M.A., SU, Z., SYMMONS, D.P., THOMPSON, J.R., THOMSON, W., TRAVERS, M.E., TURNBULL, C., VALSESIA, A., WALKER, M., WALKER, N.M., WALLACE, C., WARREN-PERRY, M., WATKINS, N.A., WEBSTER, J., WEEDON, M.N., WILSON, A.G., WOODBURN, M., WORDSWORTH, B.P., YOUNG, A.H., ZEGGINI, E., CARTER, N.P., FRAYLING, T.M., LEE, C., MCVEAN, G., MUNROE, P.B., PALOTIE, A., SAWCER, S.J., SCHERER, S.W., STRACHAN, D.P., TYLER-SMITH, C., BROWN, M.A., BURTON, P.R., CAULFIELD, M.J., COMPSTON, A., FARRALL, M., GOUGH, S.C., HALL, A.S., HATTERSLEY, A.T., HILL, A.V., MATHEW, C.G., PEMBREY, M., SATSANGI, J., STRATTON, M.R., WORTHINGTON, J., DELOUKAS, P., DUNCANSON, A., KWIATKOWSKI, D.P., MCCARTHY, M.I., OUWEHAND, W., PARKES, M., RAHMAN, N., TODD, J.A., SAMANI, N.J. and DONNELLY, P., 2010. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature,* **464**(7289), pp. 713-720.

WETERINGS, E. and VAN GENT, D.C., 2004. The mechanism of non-homologous end-joining: a synopsis of synapsis. *DNA repair,* **3**(11), pp. 1425-1435.

WILLCOCKS, L.C., LYONS, P.A., CLATWORTHY, M.R., ROBINSON, J.I., YANG, W., NEWLAND, S.A., PLAGNOL, V., MCGOVERN, N.N., CONDLIFFE, A.M., CHILVERS, E.R., ADU, D., JOLLY, E.C., WATTS, R., LAU, Y.L., MORGAN, A.W., NASH, G. and SMITH, K.G., 2008. Copy number of FCGR3B, which is associated with systemic lupus erythematosus, correlates with protein expression and immune complex uptake. *The Journal of experimental medicine,* **205**(7), pp. 1573-1582.

WINCHESTER, L., YAU, C. and RAGOUSSIS, J., 2009. Comparing CNV detection methods for SNP arrays. *Briefings in functional genomics & proteomics,* **8**(5), pp. 353-366.

WORDSWORTH, B.P. and BELL, J.I., 1992. The immunogenetics of rheumatoid arthritis, *Springer seminars in immunopathology* 1992, Springer, pp. 59-78.

XIE, Y.G., ZHENG, H., LEGGO, J., SCULLY, M.F. and LILLICRAP, D., 2002. A founder factor VIII mutation, valine 2016 to alanine, in a population with an extraordinarily high prevalence of mild hemophilia A. *Thrombosis and haemostasis,* **87**(1), pp. 178-179.

XUE, Y., SUN, D., DALY, A., YANG, F., ZHOU, X., ZHAO, M., HUANG, N., ZERJAL, T., LEE, C., CARTER, N.P., HURLES, M.E. and TYLER-SMITH, C., 2008. Adaptive evolution of UGT2B17 copy-number variation. *American Journal of Human Genetics,* **83**(3), pp. 337-346.

YANG, T.L., CHEN, X.D., GUO, Y., LEI, S.F., WANG, J.T., ZHOU, Q., PAN, F., CHEN, Y., ZHANG, Z.X., DONG, S.S., XU, X.H., YAN, H., LIU, X., QIU, C., ZHU, X.Z., CHEN, T., LI, M., ZHANG, H., ZHANG, L., DREES, B.M., HAMILTON, J.J., PAPASIAN, C.J., RECKER, R.R., SONG, X.P., CHENG, J. and DENG, H.W., 2008. Genome-wide copy-number-variation study identified a susceptibility gene, UGT2B17, for osteoporosis. *American Journal of Human Genetics,* **83**(6), pp. 663-674.

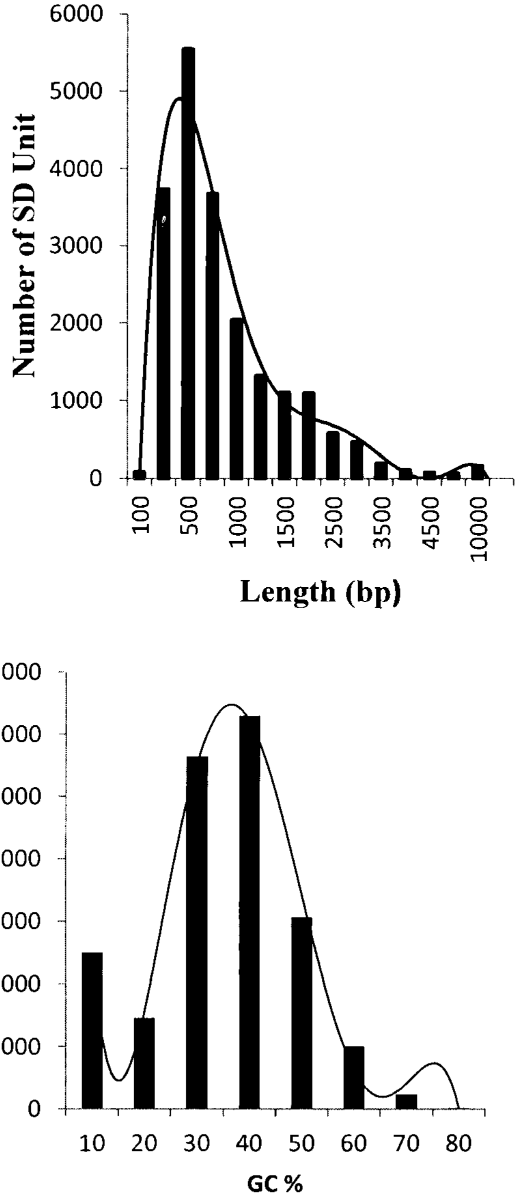
YOON, S., XUAN, Z., MAKAROV, V., YE, K. and SEBAT, J., 2009. Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome research,* **19**(9), pp. 1586-1592.

YOUNG, T.L., PENNEY, L., WOODS, M.O., PARFREY, P.S., GREEN, J.S., HEFFERTON, D. and DAVIDSON, W.S., 1999. A fifth locus for Bardet-Biedl syndrome maps to chromosome 2q31. *American Journal of Human Genetics,* **64**(3), pp. 900-904.

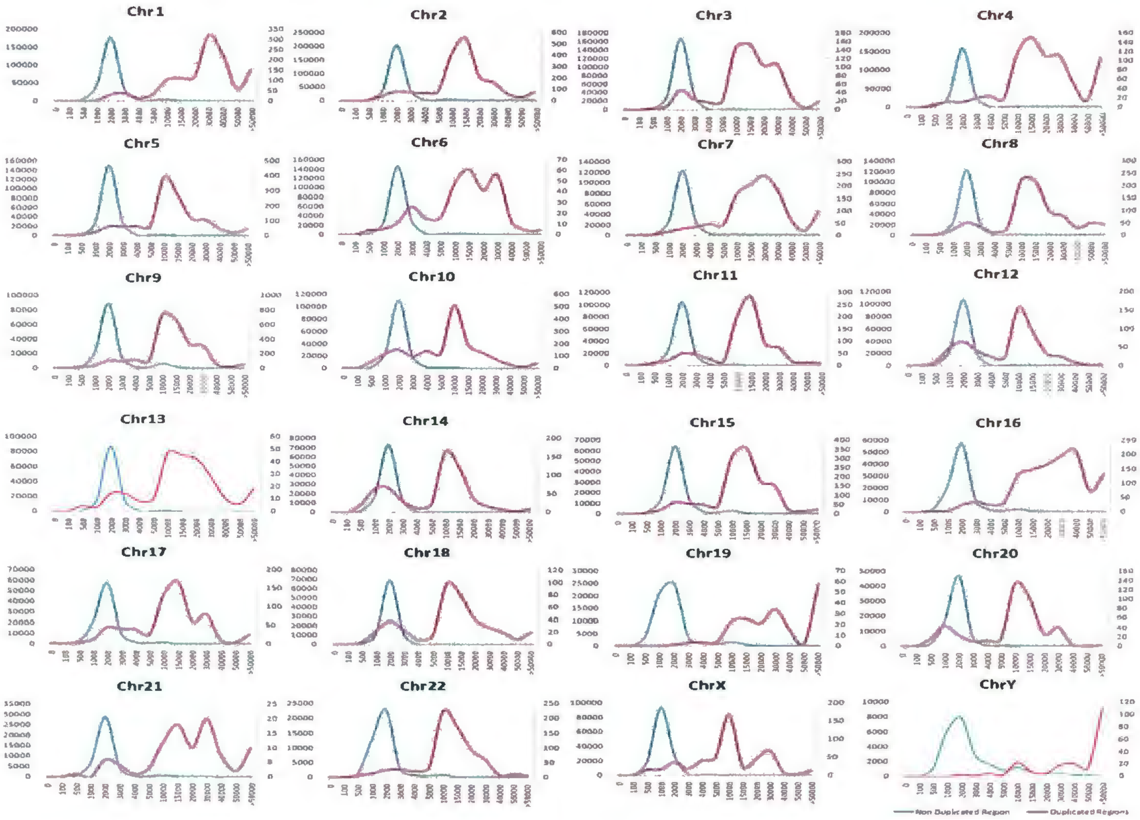
ZHANG, F., GU, W., HURLES, M.E. and LUPSKI, J.R., 2009. Copy number variation in human health, disease, and evolution. *Annual review of genomics and human genetics,* **10**, pp. 451-481.

ZHANG, F., KHAJAVI, M., CONNOLLY, A.M., TOWNE, C.F., BATISH, S.D. and LUPSKI, J.R., 2009. The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nature genetics,* **41**(7), pp. 849-853.

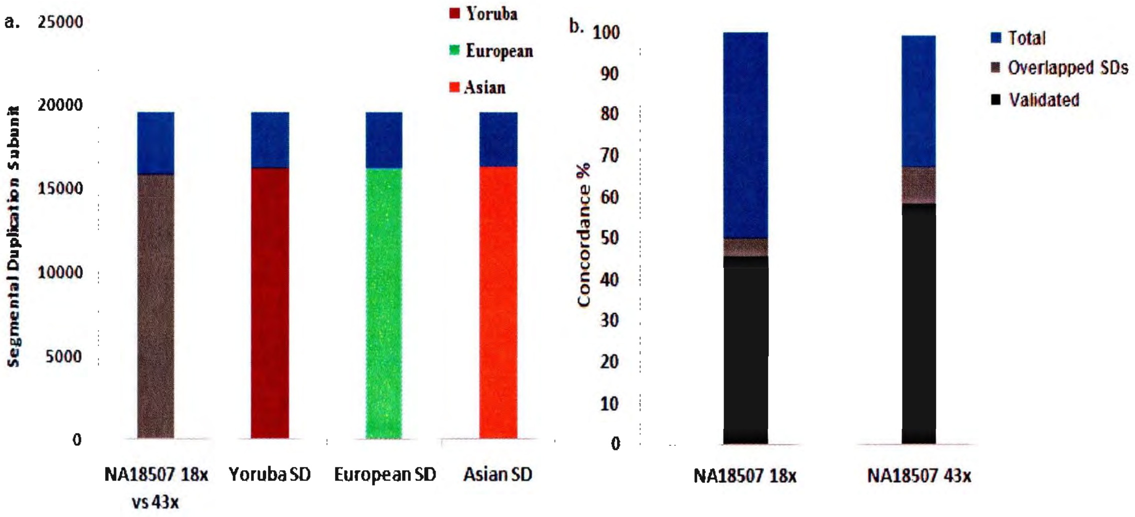
SUPPLEMENTARY 1 CHAPTER 2



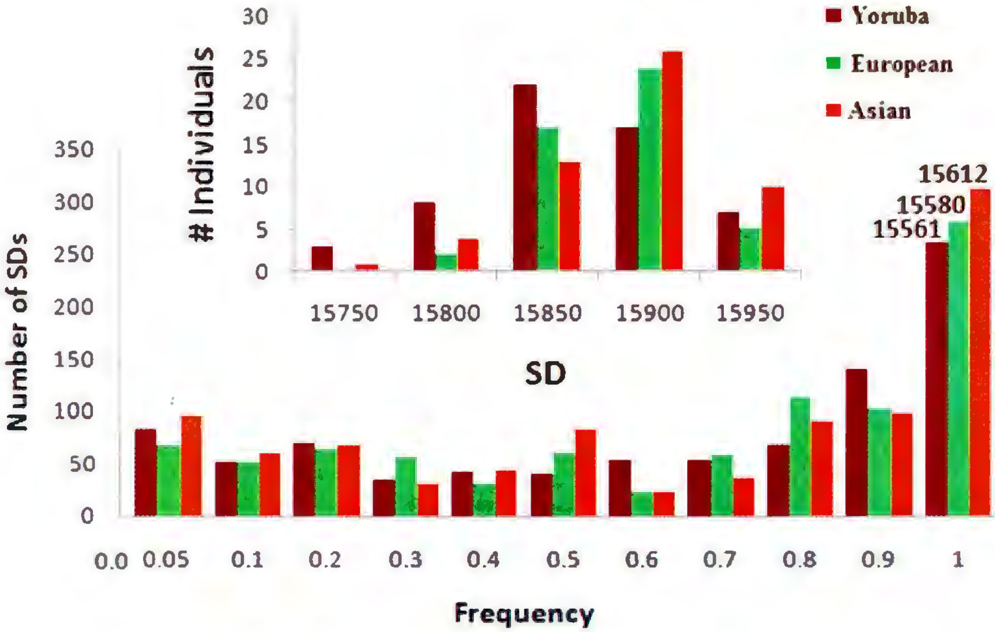
Supplementary Figure S1.1 Length and GC distribution of segmental duplication (SD) units detected by computational prediction. **a)** The mean size was 822 bp after exclusion of common repeat elements from duplicated loci. **b)** Interestingly, the majority (i.e., approximately >80%) of these loci were located within the 30th to 70th G+C percentile.



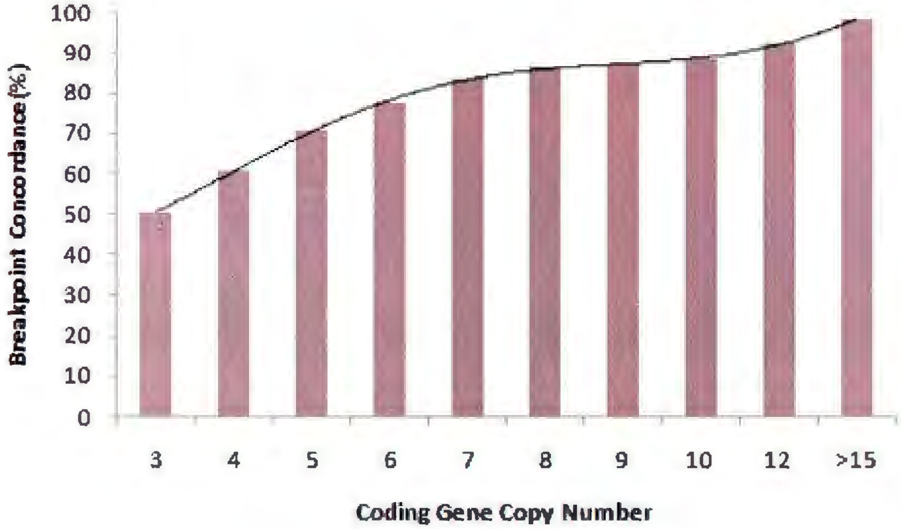
Supplementary Figure S1.2 Read depth distribution of the complete human genome segmented between duplicated (right Y-axis - red line) and non-duplicated regions (left Y-axis - blue line). The distribution illustrates a distinctive distribution pattern between duplicated and non-duplicated regions with an approximate 7% error rate. These discrepancies are largely attributed to micro-deletion polymorphisms that are located within the SD units as previously reported.



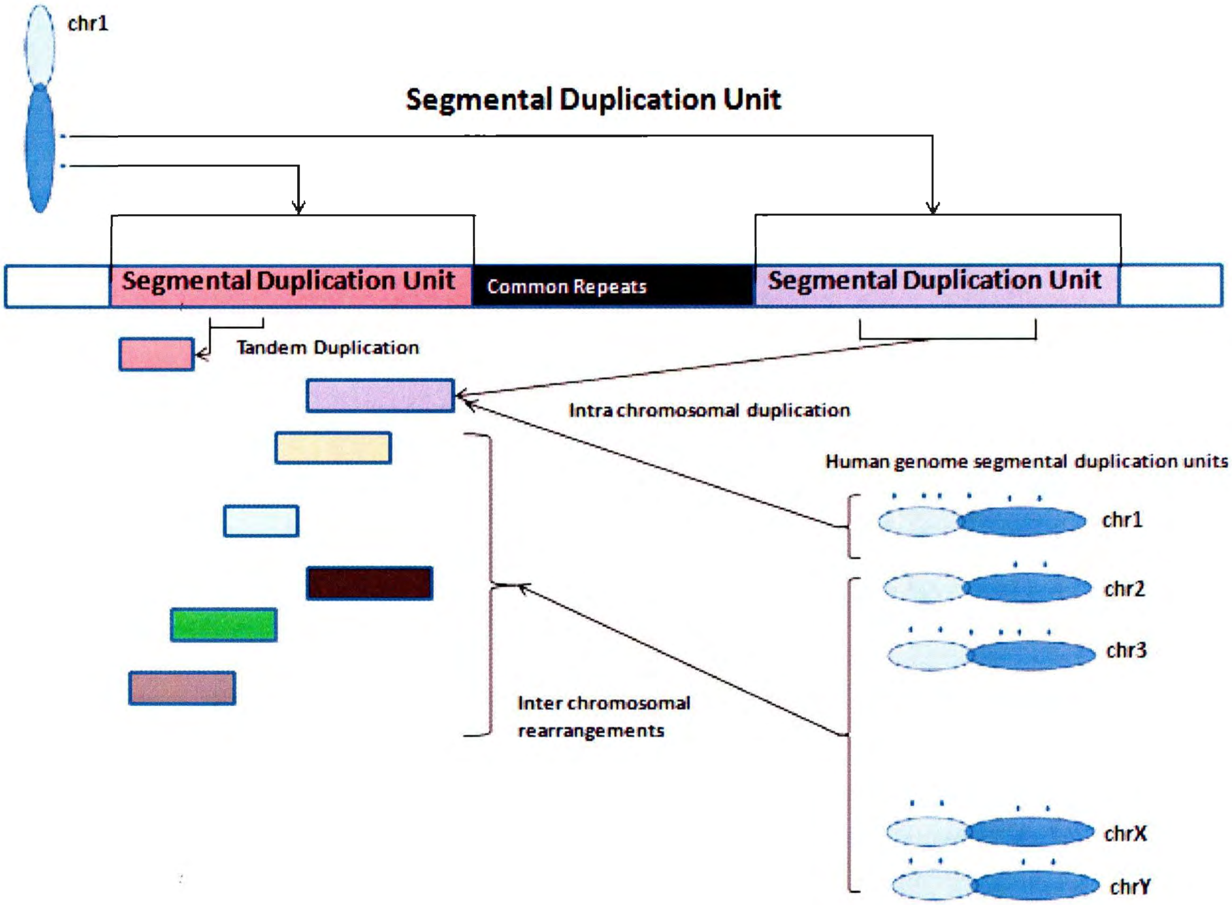
Supplementary Figure S1.1 Population comparison of detected SD units. **a)** The bar chart illustrates the comparison between detected duplicated regions which are common between 18x and 43x coverage of the NA18507 human genome. The SD regions detected using low coverage short read data in three different populations (i.e., 57 Yoruba, 48 European and 54 Asian individuals) illustrate a concordance of 82.95%, 83.03% and 83.19%, respectively. **b)** The first bar (NA18507 18X) represents a comparison between this study and Conrad et al. 2010b study while the second bar (NA18507 43X) represents a comparison between studies by Sudmant *et al.* 2010 and Conrad *et al.* 2010.



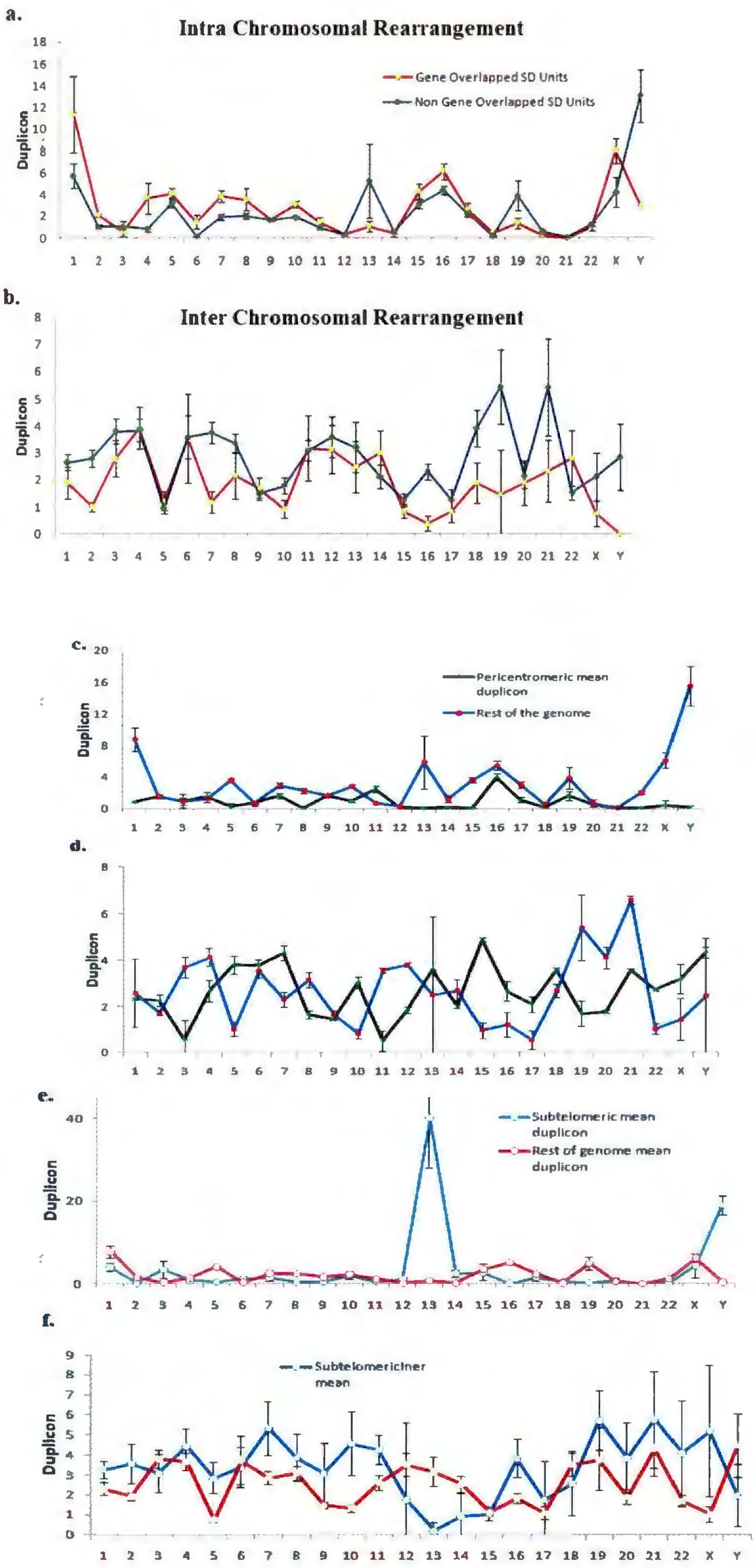
Supplementary Figure S1.2 Concordance of autosomal SD unit in three populations are depicted in color-coded histograms. More than 90% of the concordant SD units are common within these three populations. Note that the average read depth for most of the individuals varies from 1.5 to 7x. The Asian population is associated with a higher error rate which is attributed to the lowest coverage.



Supplementary Figure S1.3 Breakpoint comparison (i.e., >50% overlap) of highly variable genes detected by Alkan *et al.* 2009, illustrates that 79% of these genes are within our detected segmental duplication (SD) breakpoints.


chr1
Segmental Duplication Unit
Segmental Duplication Unit
Common Repeats
Segmental Duplication Unit
Tandem Duplication
Intra chromosomal duplication
Human genome segmental duplication units
chr1
chr2
chr3
Inter chromosomal rearrangements
chrX
chrY

Supplementary Figure S1.4 Criteria for inter- and intra-chromosomal rearrangements. In this scenario, two mechanistic evolutionary forces are imposed (intra-chromosomal rearrangements including tandem duplications, and inter-chromosomal rearrangements) together to illustrate the criteria for rearrangements.


a.
Intra Chromosomal Rearrangement
Gene Overlapped SD Units
Non Gene Overlapped SD Units
Duplicon
0
2
4
6
8
10
12
14
16
18
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
b.
Inter Chromosomal Rearrangement
Duplicon
0
1
2
3
4
5
6
7
8
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
c.
Pericentromeric mean duplicon
Rest of the genome
Duplicon
0
4
8
12
16
20
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
d.
Duplicon
0
2
4
6
8
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
e.
Subtelomeric mean duplicon
Rest of genome mean duplicon
Duplicon
0
20
40
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
f.
Subtelomericlner mean
Duplicon
0
1
2
3
4
5
6
7
8
9
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y

Supplementary Figure S1.5 Inter- and intra-chromosomal rearrangement distribution for the NA18507 genome depicting the landscape of human genic and agenic region rearrangements with a 99% confidence interval. **a)** Intra-chromosomal rearrangement enrichment and **b)** low inter-chromosomal rearrangement was observed within genic regions as compared with agenic regions. For Y chromosome the analysis shows no inter chromosomal rearrangement within the genic duplicated region (excluding DUX family region from the analysis). Mean duplicons that define pericentromeric duplications with **c)** intra-chromosomal and **d)** inter-chromosomal rearrangements. **e)** The mean duplicon within the telomeric region with intra-chromosomal rearrangement depicts chromosome 13q34 and Yq12 as an outlier with extensive intra-chromosomal rearrangement. **f)** Illustrates that inter-chromosomal rearrangement is dominant within the telomeric region of NA18507 human genome.

Supplementary Table S1.1 Rearrangement analysis of SD units within genic, subtelomeric and pericentromeric regions of the human genome.

Chromosome	Genic / Non-Genic		Subtelomeric		Pericentromeric	
	Intra $P^{Ŧ}$	Inter $P^{Ŧ}$	Intra $P^{Ŧ}$	Inter $P^{Ŧ}$	Intra $P^{Ŧ}$	Inter $P^{Ŧ}$
1	1.0 X 10^{-6}	0.995	0.999	1.0 X 10^{-6}	1	0.841
2	1.0 X 10^{-6}	1	1	1.0 X 10^{-6}	0.479	3.5 X 10^{-4}
3	0.843	0.996	1.0 X 10^{-6}	0.939	0. 93	0.999
4	1.0 X 10^{-6}	0.407	0.974	9.5 X 10^{-3}	4.8 X 10^{-2}	0.999
5	1.0 X 10^{-6}	4.5 X 10^{-2}	1	1.0 X 10^{-6}	1	1.0 X 10^{-5}
6	1.0 X 10^{-6}	0.526	3.3 X 10^{-4}	0.695	0.234	0.405
7	1.0 X 10^{-6}	1	0.992	1.0 X 10^{-6}	1	1.0 X 10^{-6}
8	8.0 X 10^{-6}	0.996	0.999	3.3 X 10^{-2}	1	0.986
9	0.261	0.106	1	5.9 X 10^{-5}	0.649	0.869
10	1.0 X 10^{-6}	0.999	0.766	1.0 X 10^{-6}	1	1.0 X 10^{-6}
11	3.2 X 10^{-3}	0.418	1	1.0 X 10^{-6}	1.0 X 10^{-6}	1
12	0.166	0.827	1.0 X 10^{-6}	0.770	0.985	0.999
13	0.997	0.908	1.0 X 10^{-6}	1	1	1.9 X 10^{-2}
14	0.784	9.9 X 10^{-4}	1.0 X 10^{-6}	0.990	1	0.955
15	1.3 X 10^{-5}	0.997	0.779	0.652	1	1.0 X 10^{-6}
16	1.0 X 10^{-6}	1	1	1.0 X 10^{-6}	1	1.0 X 10^{-6}
17	6.3 X 10^{-2}	0.951	0.965	0.123	1	1.0 X 10^{-6}
18	3.2 X 10^{-4}	0.999	2.9 X 10^{-2}	0.865	0.999	8.4 X 10^{-2}
19	0.999	0.999	1	8.0 X 10^{-3}	0.998	0.999

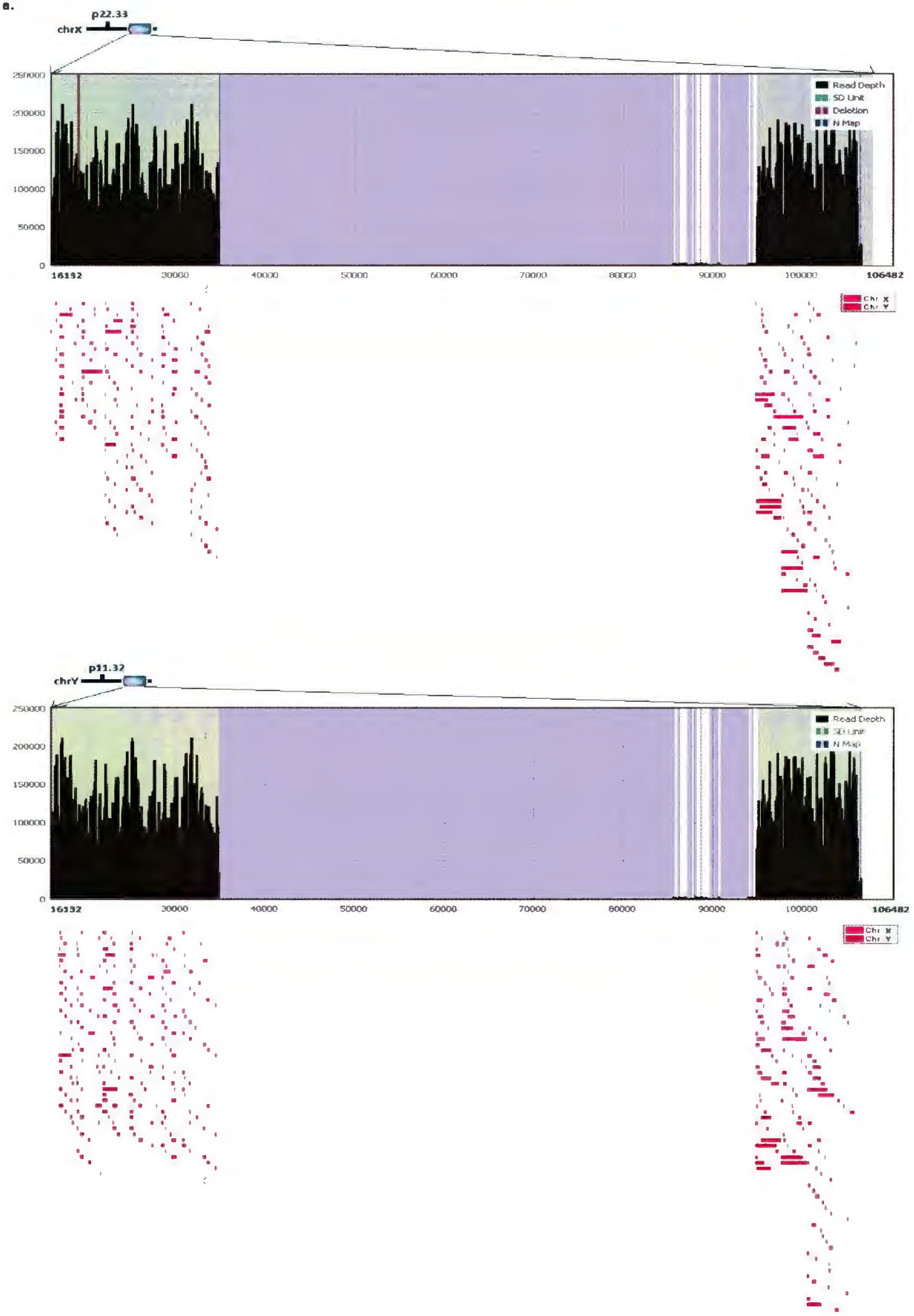
20	0.999	0.714	9.2 X 10^{-2}	5.6 X 10^{-5}	0.832	0.999
21	1	0.999	1	6.5 X 10^{-2}	0.933	0.995
22	0.865	9.0 X 10^{-6}	0.998	1.4 X 10^{-3}	1	1.0 X 10^{-6}
X	1.0 X 10^{-6}	0.999	0.913	1.0 X 10^{-6}	0.997	1.6 X 10^{-2}
Y	N/A*	N/A*	1.0 X 10^{-6}	0.999	1	8.6 X 10^{-3}

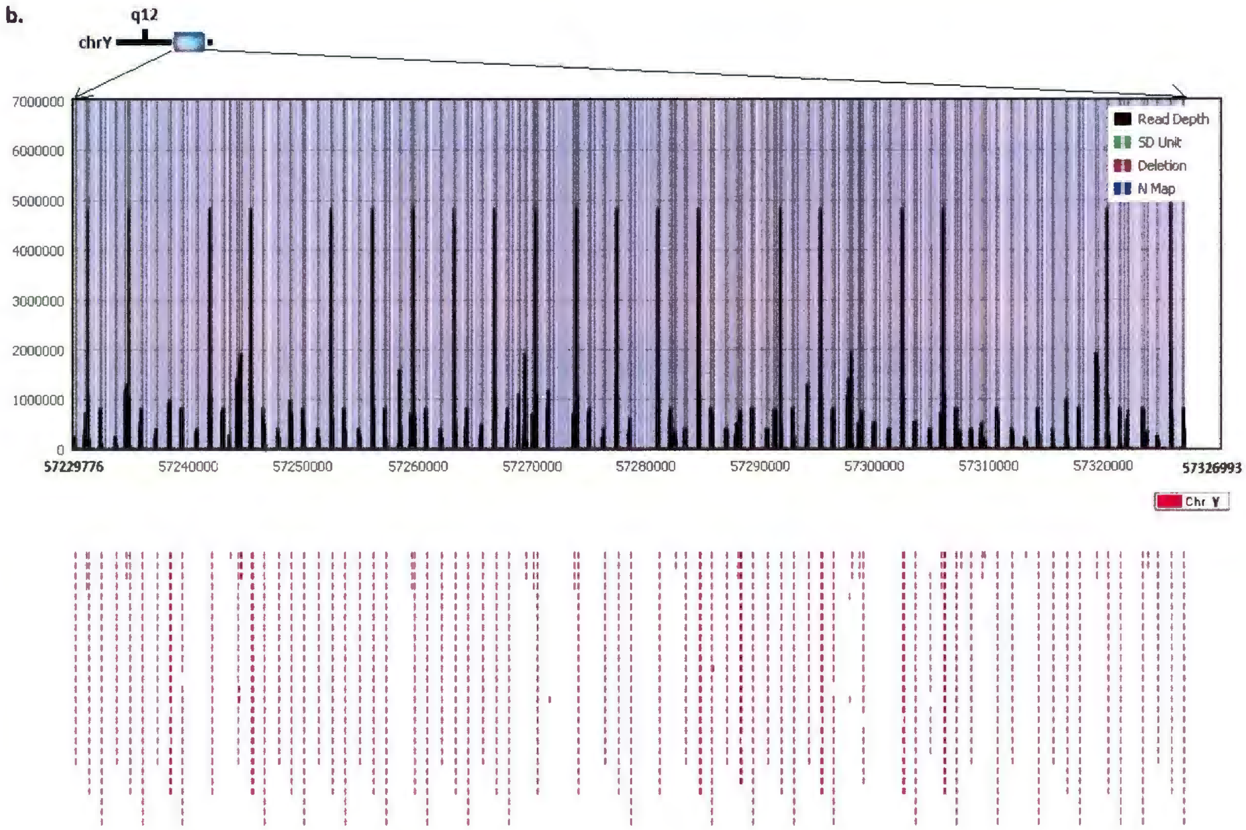
Note: The SD units that overlap with a reference gene was analyzed for each chromosome based on the captured inter- and intra-chromosomal rearrangements against the rest of the human genome. The inter and inter-chromosomal rearrangement of the segmental duplications within subtelomeric and pericentromeric regions were also analyzed against the rest of the duplicated regions in the human genome.

¥ - The table shows a series of 1-tail analysis of genomic rearrangements using 1 million permutations of mean differences to obtain empirical *P*-values.

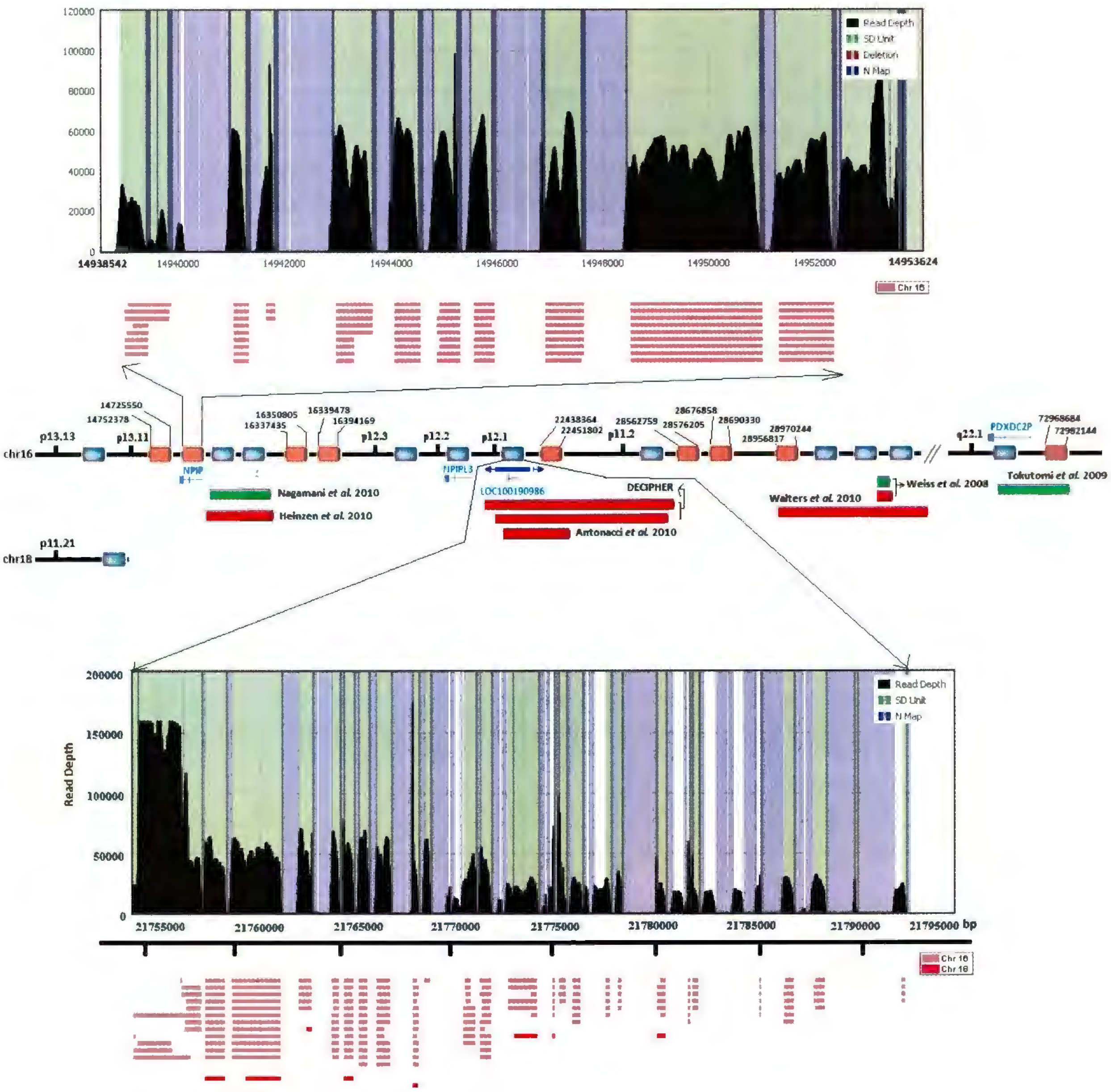
*Chromosme Y contained only two genes (i.e., *RBMY1A1* and *RBMY1F*) which revealed intra-chromosomal rearrangements, whereas the *DUX4* gene was associated with inter-chromosomal rearrangement. Thus, we have excluded chromosome Y from permutation test.

a.


p22.33
chrX
250000
200000
150000
100000
50000
0
16132
30000
40000
50000
60000
70000
80000
90000
100000
106482
Read Depth
SD Unit
Deletion
N Map
Chr X
Chr Y
p11.32
chrY
250000
200000
150000
100000
50000
0
16132
30000
40000
50000
60000
70000
80000
90000
100000
106482
Read Depth
SD Unit
N Map
Chr X
Chr Y



Supplementary Figure S1.6 Rearrangement within PAR (pseudohomologous region). **a)** The pseudohomologous region of the sex chromosomes contain a 90 kbp duplication between chromosomes X and Y where extensive inter- and intra-chromosomal tandem duplications were observed. **b)** A 90 kbp region within Yq12 locus illustrating extreme tandem duplication with a duplicon of 100 bp in length.

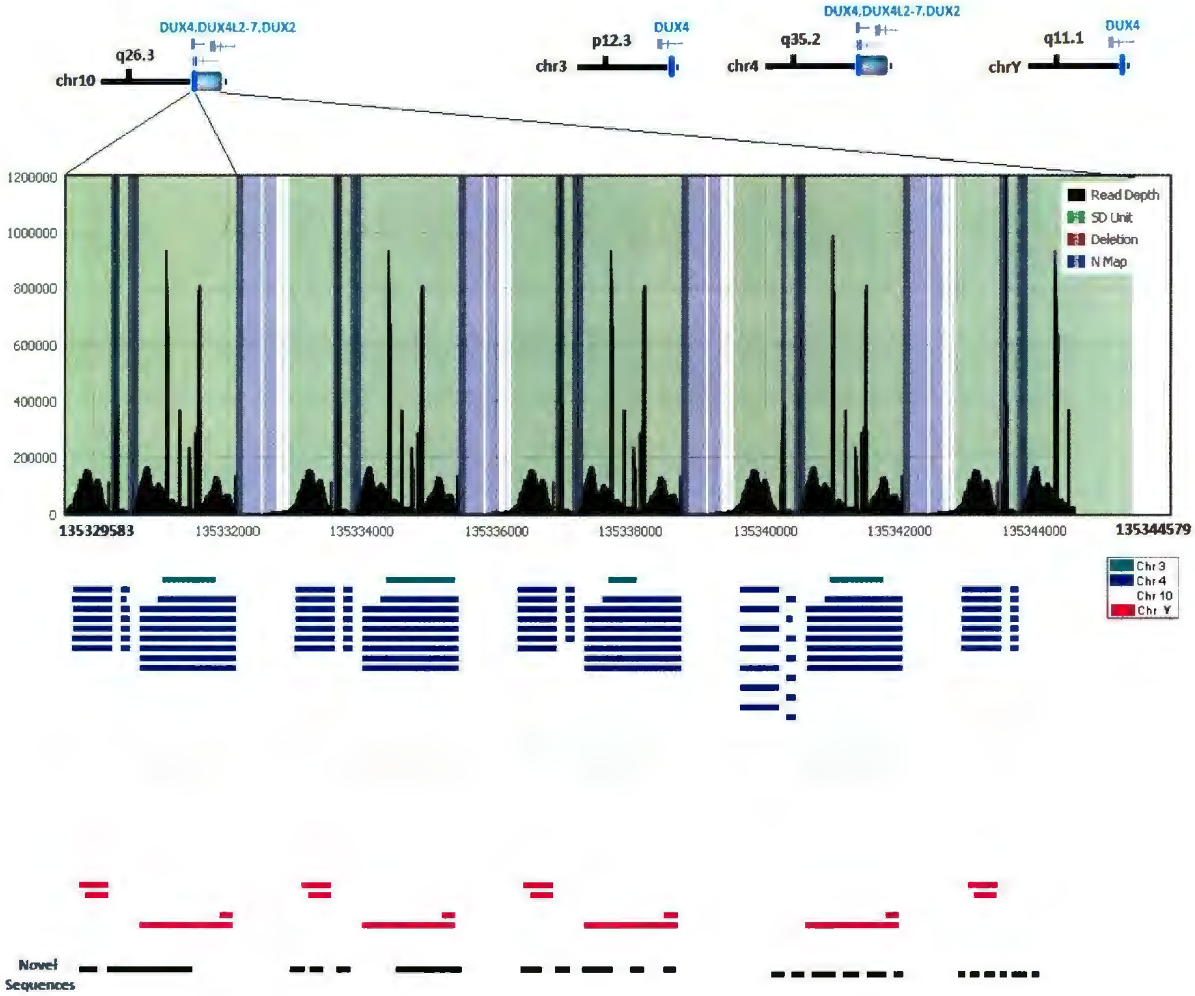


Supplementary Figure S1.9 Localization of the *NPIP* and *NPIPL3* gene derivatives. The alignment of the read depth plot is approximated in the chromosome contig. Multiple copies of the *NPIP* gene is located within close proximity of the *NPIPL3* gene derivatives. Our analysis revealed the localization for the *NPIP* gene is within chromosome 16, whereas the *NPIPL3* derivatives are localized in both chromosomes 16 and 18, which has been confirmed by FISH.

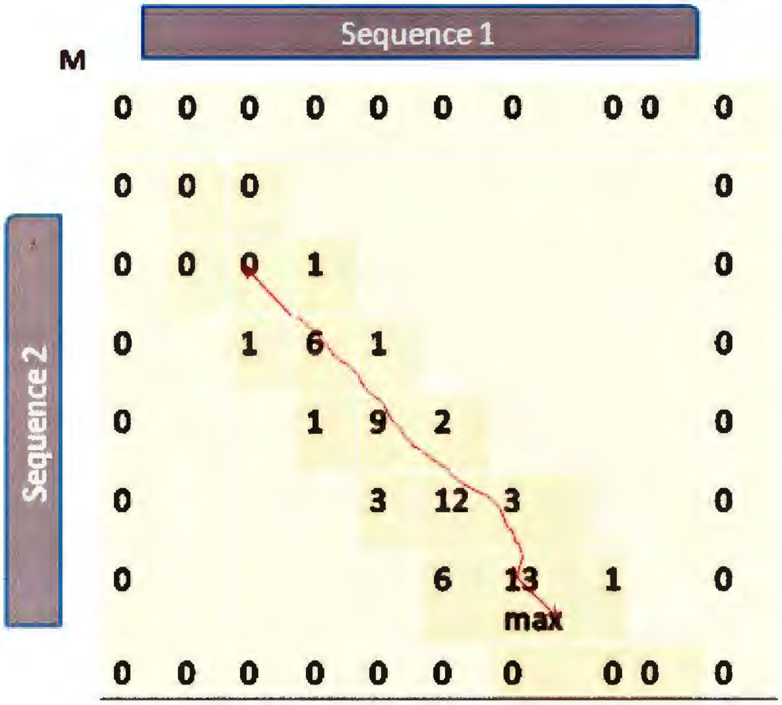
Supplementary Table S1.2 Summary of FISH analysis.

Number of cells counted	**Chromosome modal number**	**Probes used in hybridization** **(Probe label in brackets)**	**Chromosome hybridization of probes**	**Signal Summary** **(approximated)**
10	46 10 cells with 46 chromosomes	G248P8712C10 (SpectrumOrange)	Chr 1 of both homologs	Signals observed at three localizations on both chr 1 homologs: 1) at band 1p36: >one signal per homolog; 2) at band
10	46 10 cells with 46 chromoso	G248P8661F9 (SpectrumOrange)	Chr 16 and 18 of both homologs	Signals observed at three localizations on both chr 16 and chr 18
17	46 17 cells with 46 chromosomes	G248P80054G1 (SpectrumOrange)	Chr 1, 5, 20, and 22 of both homologs	Signals observed at six localizations on both chr 1, chr 5, chr 6, chr 20, and chr 22 homologs: 1) at band 1p13: one signal per homolog;

		RP11-1113I2 (SpectrumGreen)	Chr 22 of both homologs	2) at band 5p13: >one signals per homolog; 3) at band 6p: >one to two signals per homolog; 4) at band 5q21: one signal per homolog. 5) at band 20p11.2 - one signal per homolog; 6) at band 22q11.2 - ~four to five signals per homolog. Two signals present, one on each homolog of chr 22 at band 22q13.

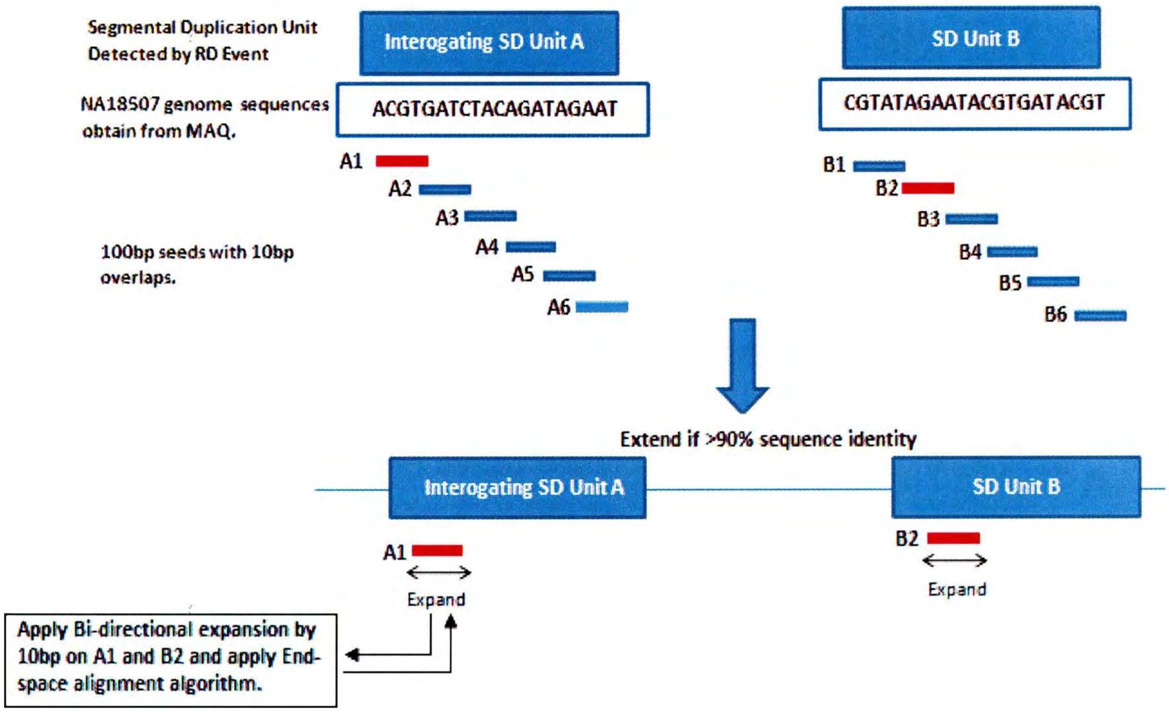


Supplementary Figure S1.10 The localization of the *DUX* gene family illustrating copies at the base pair level in chromosomes 3, 4, 10, Y and within the novel *de novo* assembly sequences. The *DUX4* gene has been previously reported to possess the most number of copies within the human genome, however the location of this gene family remains incompletely characterized. We have revealed the position of each copy within the human genome, indicating the presence of the *DUX* gene family on chromosome 3.



0	0	0	0	0	0	0	0	0	0
0	0	0							0
0	0	0	1						0
0		1	6	1					0
0			1	9	2				0
0				3	12	3			0
0					6	13 max	1		0
0	0	0	0	0	0	0	0	0	0

Supplementary Figure S1.11 End space free alignment algorithm. The DP matrix populates according to the cost function and only trace back occurs from the maximum position of the matrix to 0. The gap is minimized by introducing a high penalty rate (i.e., -2 or -3).


Segmental Duplication Unit Detected by RD Event
Interogating SD Unit A
SD Unit B
NA18507 genome sequences obtain from MAQ.
ACGTGATCTACAGATAGAAT
CGTATAGAATACGTGATACGT
A1
A2
A3
A4
A5
A6
B1
B2
B3
B4
B5
B6
100bp seeds with 10bp overlaps.
Extend if >90% sequence identity
Interogating SD Unit A
SD Unit B
A1
B2
Expand
Expand
Apply Bi-directional expansion by 10bp on A1 and B2 and apply End-space alignment algorithm.

Supplementary Figure S1.12 The ‘seed and extend’ mechanism to detect optimal seeds with a 100 bp window of alignment. The seeds with <90% sequence identity was ignored for extend step. The extend step is a recursive procedure which does not stop until it crosses the predetermined threshold of 90%. If there are multiple overlapping seeds, only the seeds with maximum expansion were kept.

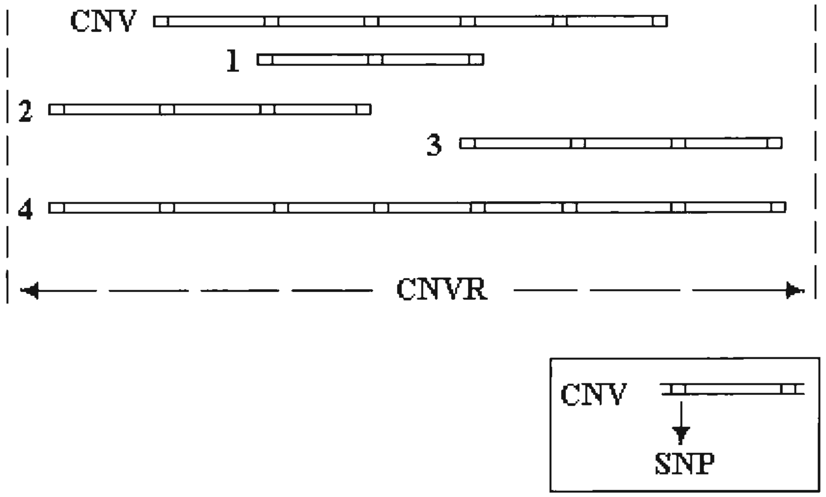
Supplementary Table S1.3 Short reads map statistics. Short reads were mapped using mrsFAST with an average read length of 36 bp in the NA18507 human genome with approximately 56% being mapped against the repeat masked reference genome.

Summary	Statistics
Genome	NA18507 (Hapmap Yoruba sample)
Platform	Illumina Genome Analyzer
Number of Short Reads	1,504,002,272
Total Mapped Short Reads	839,039,591

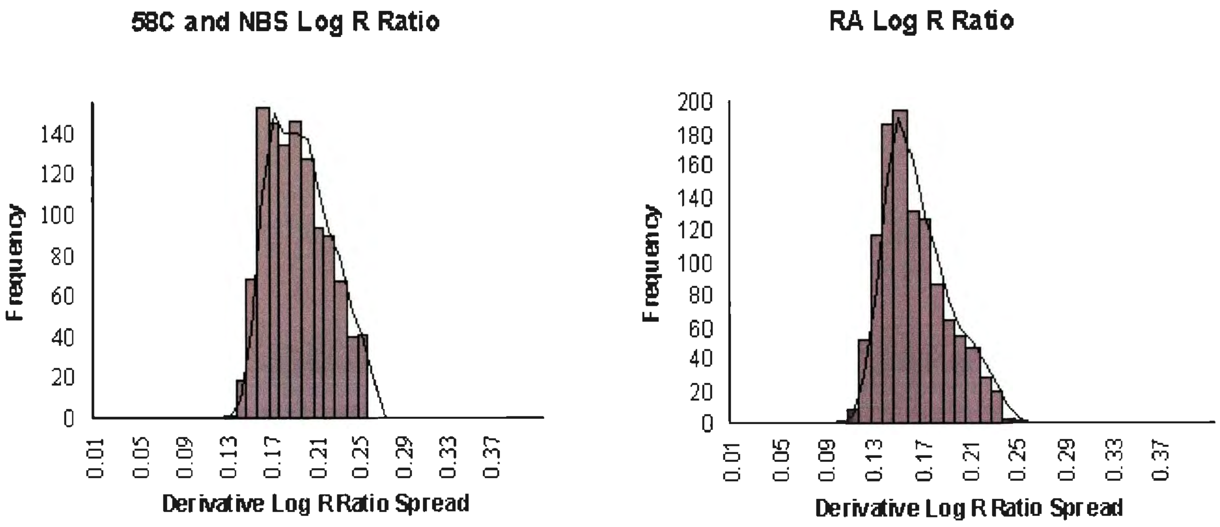
SUPPLEMENTARY 2 CHAPTER 2

Please find the attached CD.

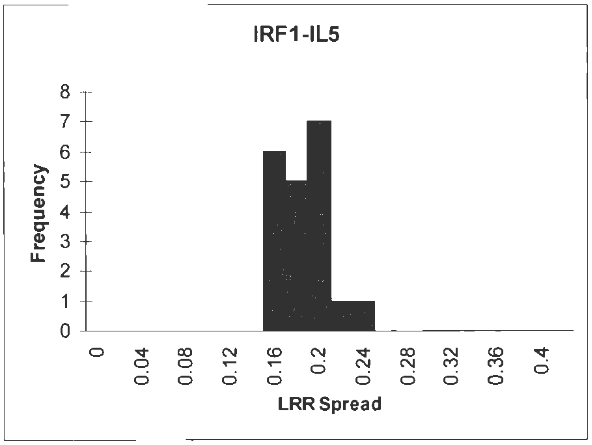
SUPPLEMENTARY 3 CHAPTER 4

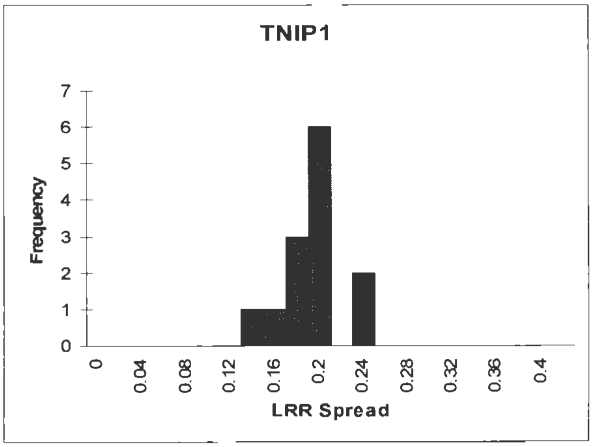


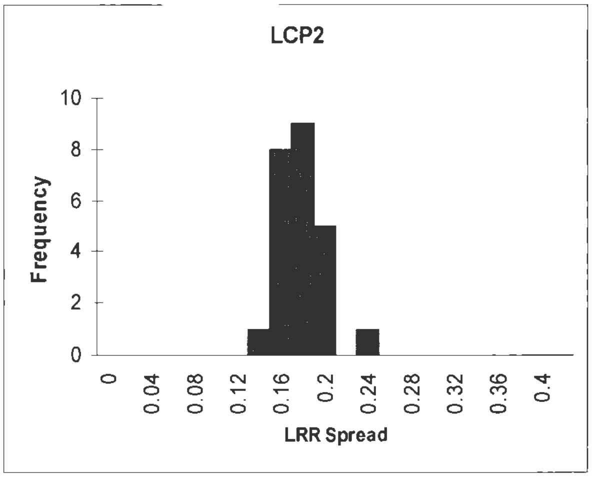
Supplementary Figure S3.1: Criteria to consider CNVR breakpoints where multiple CNVs that overlap with each other in a region. The first case indicates where CNV1 is contained within CNV, the second and third cases show the overlap between boundaries. The fourth case shows where a CNV is contained within a larger CNV. The CNVR start point is defined with the start point of the left most CNV and end point is defined as the end point of the right most CNV.

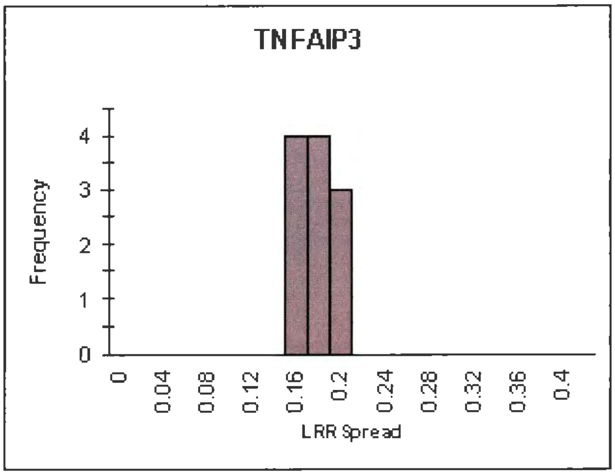


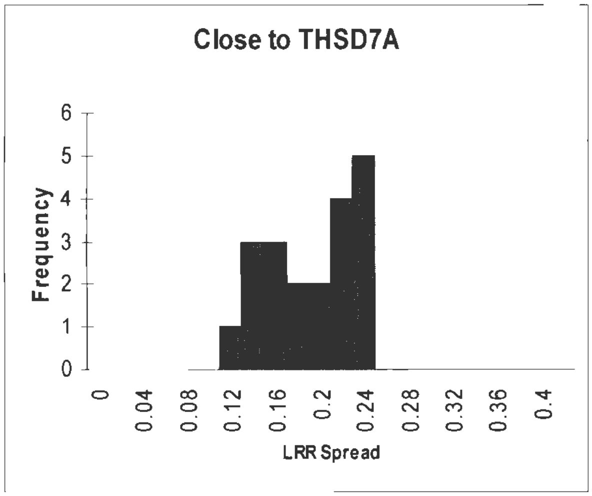
Supplementary Figure S3.2 Histogram of copy number signals plotted for controls and rheumatoid arthritis samples.

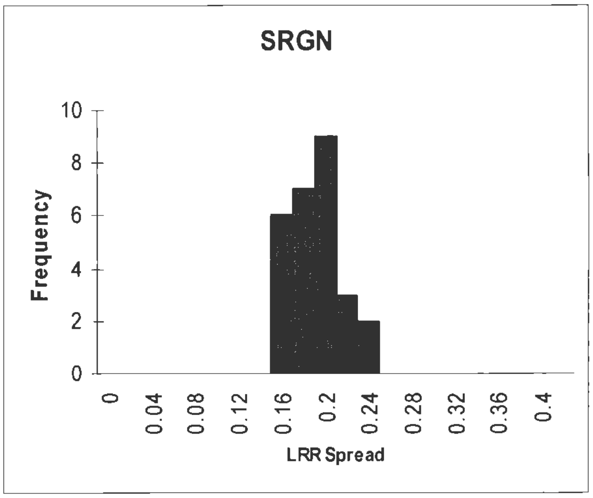

IRF1-IL5
Frequency
0
1
2
3
4
5
6
7
8
LRR Spread
0
0.04
0.08
0.12
0.16
0.2
0.24
0.28
0.32
0.36
0.4

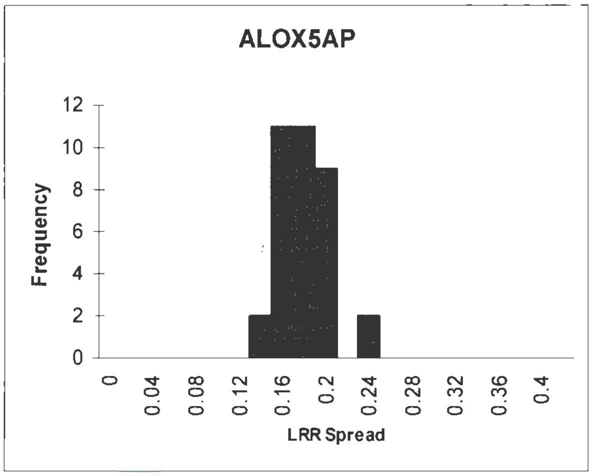

TNIP1
Frequency
0
1
2
3
4
5
6
7
LRR Spread
0
0.04
0.08
0.12
0.16
0.2
0.24
0.28
0.32
0.36
0.4

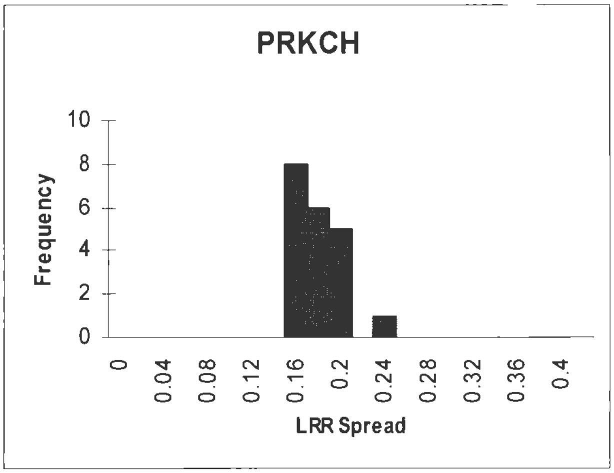

LCP2
Frequency
0
2
4
6
8
10
LRR Spread
0
0.04
0.08
0.12
0.16
0.2
0.24
0.28
0.32
0.36
0.4

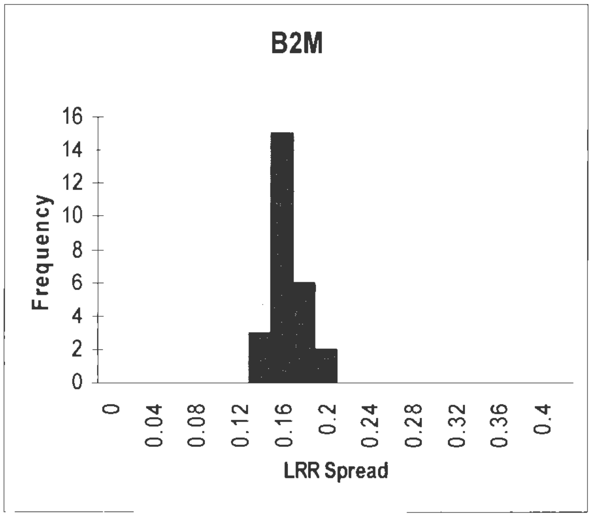

TNFAIP3
Frequency
0
1
2
3
4
LRR Spread
0
0.04
0.08
0.12
0.16
0.2
0.24
0.28
0.32
0.36
0.4

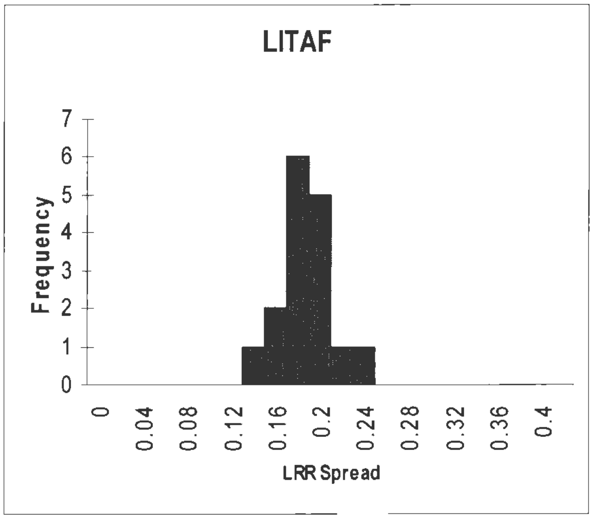

Close to THSD7A
Frequency
0
1
2
3
4
5
6
LRR Spread
0
0.04
0.08
0.12
0.16
0.2
0.24
0.28
0.32
0.36
0.4

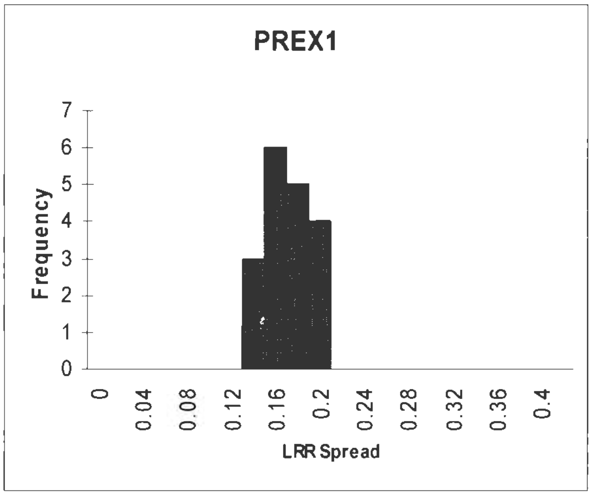

SRGN
Frequency
0
2
4
6
8
10
LRR Spread
0
0.04
0.08
0.12
0.16
0.2
0.24
0.28
0.32
0.36
0.4


ALOX5AP
Frequency
12
10
8
6
4
2
0
0
0.04
0.08
0.12
0.16
0.2
0.24
0.28
0.32
0.36
0.4
LRR Spread

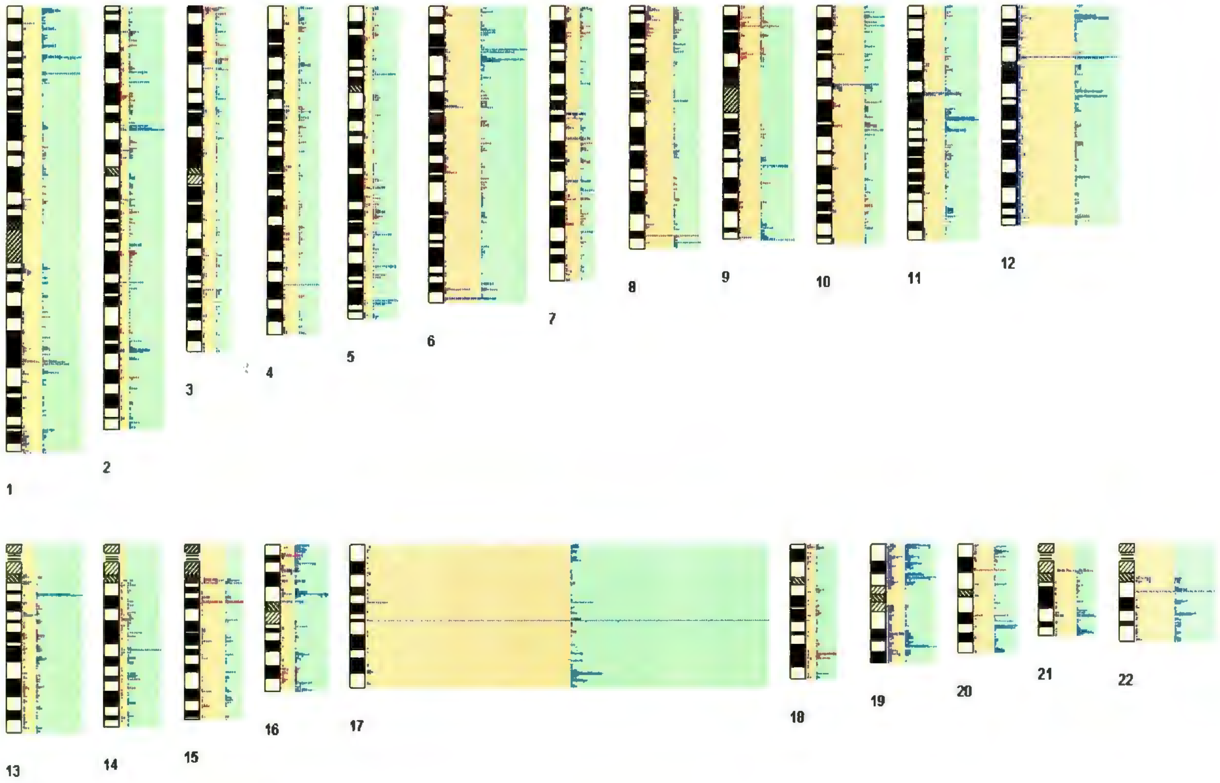

PRKCH
Frequency
10
8
6
4
2
0
0
0.04
0.08
0.12
0.16
0.2
0.24
0.28
0.32
0.36
0.4
LRR Spread


B2M
Frequency
16
14
12
10
8
6
4
2
0
0
0.04
0.08
0.12
0.16
0.2
0.24
0.28
0.32
0.36
0.4
LRR Spread

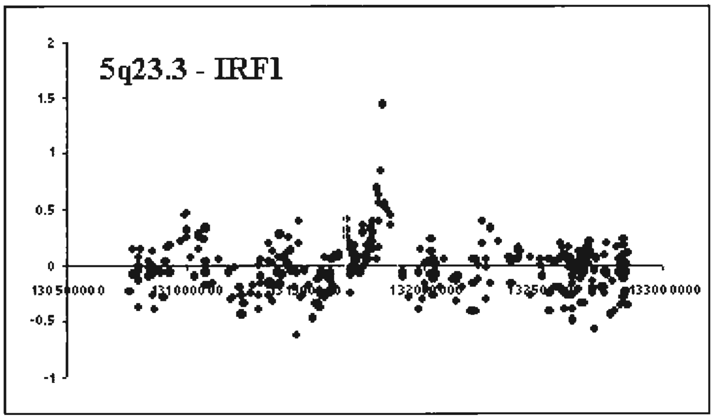

LITAF
Frequency
7
6
5
4
3
2
1
0
0
0.04
0.08
0.12
0.16
0.2
0.24
0.28
0.32
0.36
0.4
LRR Spread

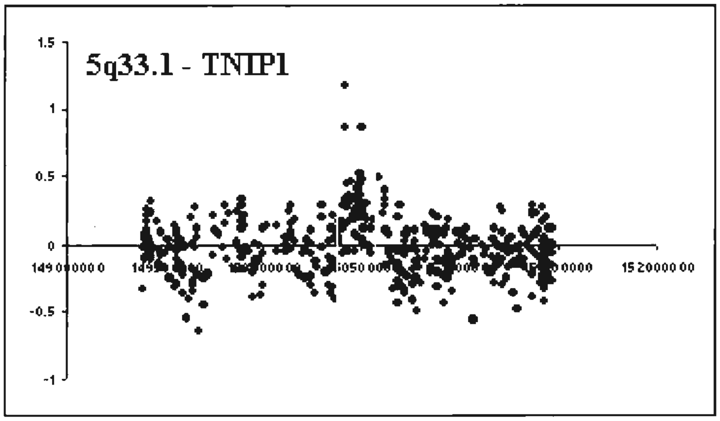

PREX1
Frequency
7
6
5
4
3
2
1
0
0
0.04
0.08
0.12
0.16
0.2
0.24
0.28
0.32
0.36
0.4
LRR Spread

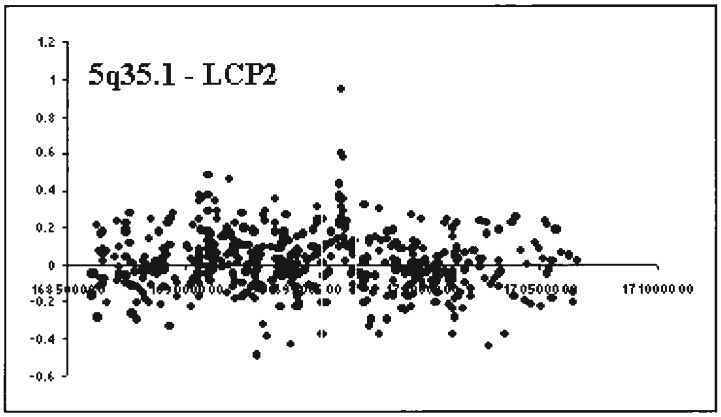
Supplementary Figure S3.3: LRR Spread histograms of RA samples for the 11 selected genic regions.

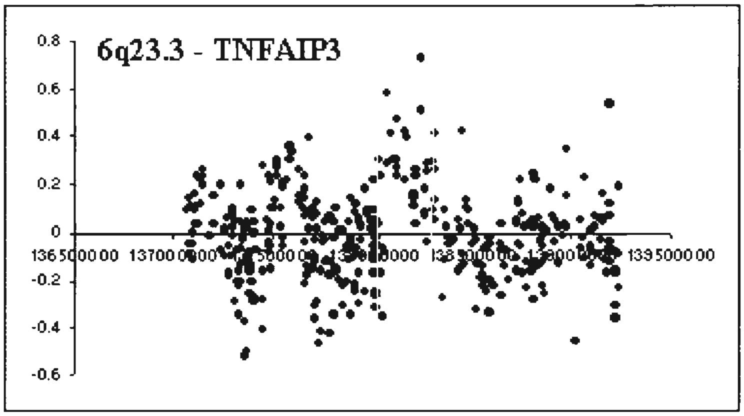


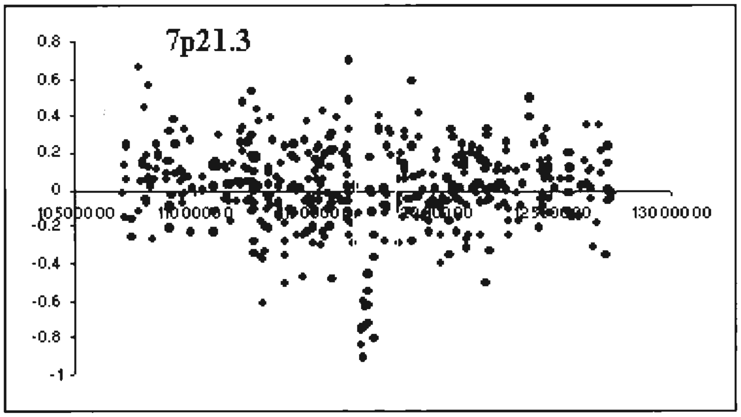
Supplementary Figure S3.4: Plot of all identified CNVs. Blue lines represent copy number duplications and red lines represent deletions. CNVs in orange are from WTCCC controls and CNVs from RA cases are in green.

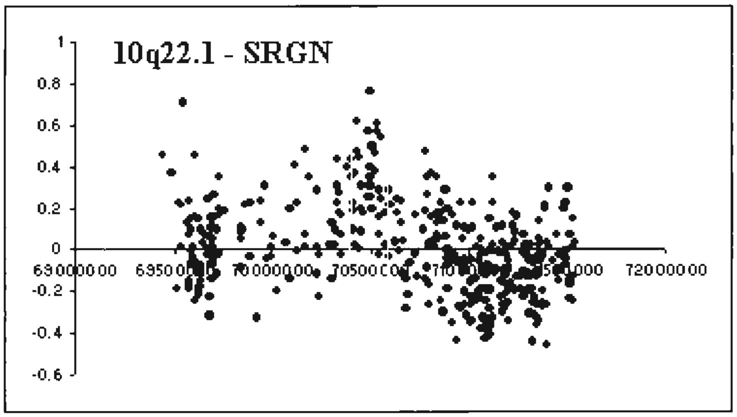

5q23.3 - IRF1
2
1.5
1
0.5
0
-0.5
-1
130500000
131000000
131500000
132000000
132500000
133000000

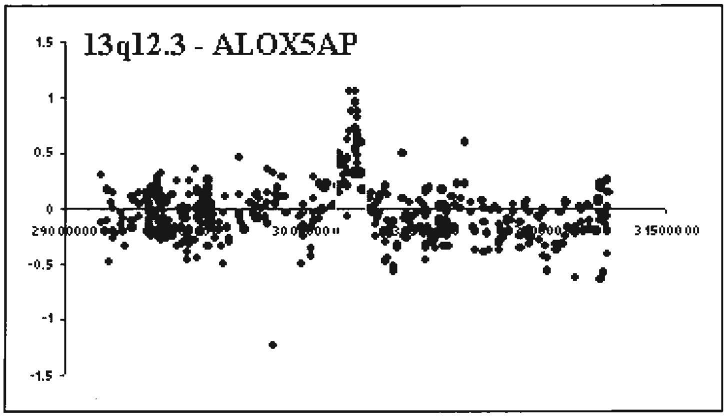

5q33.1 - TNIP1
1.5
1
0.5
0
-0.5
-1
149000000
149500000
150000000
150500000
151000000
152000000

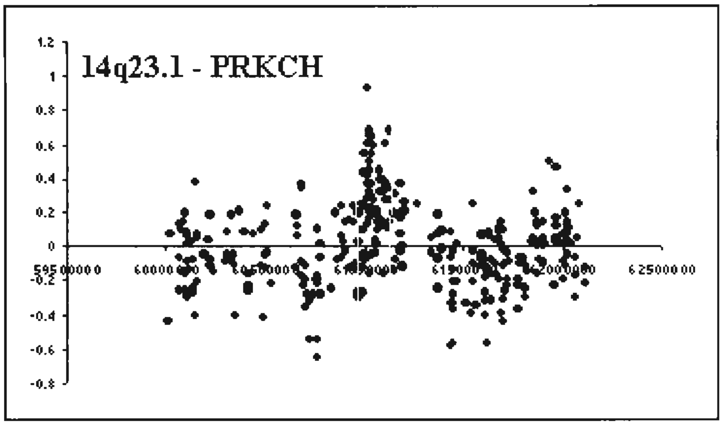

5q35.1 - LCP2
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
168500000
169000000
169500000
170000000
170500000
171000000

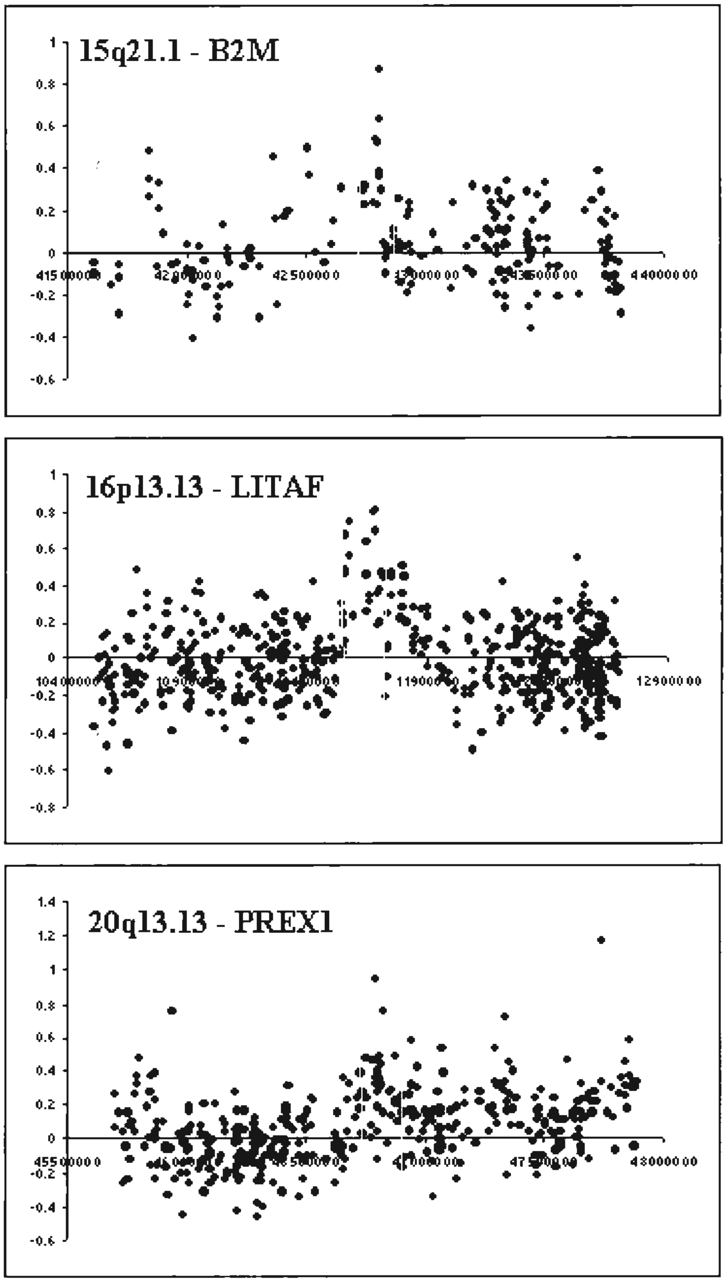

6q23.3 - TNFAIP3
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
136500000
137000000
137500000
138000000
138500000
139000000
139500000


7p21.3
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
105000000
110000000
115000000
120000000
125000000
130000000


10q22.1 - SRGN
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
69000000
69500000
70000000
70500000
71000000
71500000
72000000


13q12.3 - ALOX5AP
1.5
1
0.5
0
-0.5
-1
-1.5
29000000
30000000
31500000


14q23.1 - PRKCH
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
59500000
60000000
60500000
61000000
61500000
62000000
62500000



Supplementary Figure S3.5: CNVs detected by comparative intensity analysis. The y-axis is the log R ratios of probes from the 11 loci with the greatest significant difference in signal intensity for duplications and deletions.

SUPPLEMENTARY 4 CHAPTER 5

Supplementary Table S4.1. A list of gene CNVs that segregates at with six AS affected individuals within the family.

Chr	Start	End	Length (bp)	# of Probes	Status	Gene(s)
4	69518934	69530196	11262	40	gain	*UGT2B15*
9	84533087	84558449	25362	56	gain	*FLJ43950, FLJ43859, FLJ44082*
10	48747724	49389683	641959	1857	gain	*PTPN20B, PTPN20A, FRMPD2L1, FRMPD2L2, BMS1P5, BMS1P1, FAM25B, FAM25C, FAM25G, LOC399753, FRMPD2*
14	36327173	36329122	1949	7	loss	*BRMS1L*
17	43589332	43597242	7910	27	loss	*LRRC37A4*
2	3440855	3441900	1045	6	loss	*TTC15*
5	677193	852102	174909	592	loss	*TPPP, ZDHHC11*
5	69534796	69559366	24570	86	gain	*LOC653391*
13	77871860	77874499	2639	10	gain	*MYCBP2*

SUPPLEMENTARY 5 CHAPTER 5

Please find the attached CD.

APPENDIX A

ACHIEVEMENTS DURING THE DOCTORAL STUDY

ADDITIONAL PEER REVIEWED JOURNAL PUBLICATIONS:

Chan-Bum Choi, Tae-Hwan Kim, Jae-Bum Jun, Hye-Soon Lee, Seung Cheol Shim, Bitnara Lee, Angela Pope, **Mohammed Uddin**, Proton Rahman and Robert D Inman (2009) ARTS1 polymorphisms are associated with ankylosing spondylitis in Koreans. Annals of Rheumatic Diseases. 69:582-584, 2011.

Proton Rahman, Nicole M. Roslin, Mathieu Lemire, Celia M.T. Greenwood, Joseph Beyene, Angela Pope, Andrew D. Paterson, **Mohammed Uddin**, Dafna D. Gladman. High resolution mapping in the MHC region identifies multiple independent novel loci for psoriatic arthritis. Annals of Rheumatic Diseases, 70(4): 690-694, 2010.

AWARDS:

2012	Canadian Rheumatology Association (CRA) Basic Science Award, Canada.
2012	Best PhD. Student publication award, Discipline of Human Genetics, Faculty of Medicine, Memorial University of Newfoundland, Canada.
2011-12	Janeway Trainee Research Grant, Child Based Research Grant.
2011	Colman Graduate Student Award

2011 Best Ph.D. Research Presentation, Discipline of Human Genetics, Faculty of Medicine, Memorial University of Newfoundland, Newfoundland, Canada.

2010 Canadian Rheumatology Association (CRA) Basic Science Award, Canada.

PATENTS:

1. A provisional Patent has been filed for the custom hotspot microarray that was developed in chapter 3.
2. A provisional patent has been filed for the identification of the TS locus.