# DNA Copy Number Variation in a Cohort of Healthy Australian Women

Nicole Leisa Nisbet Chia

March 2016

A thesis submitted for the degree of

Doctor of Philosophy

Australian National University

I hereby certify that the work embodied in this thesis is the result of original research and has not been submitted for a degree to any other University or Institution

.....

Nicole Leisa Nisbet Chia

# Acknowledgements

I would like to acknowledge the efforts of many people who have assisted in achieving the completion of this project. Firstly the support and guidance provided by Professor Julia Potter, as Chair of the Supervisory Panel and as a mentor, has been pivotal in the successful completion of this project. I thank Professor Howard Slater for accepting the role of supervisor in the middle of the project. His direction and advice has been a valuable contribution to the project.

I thank the members of the "Aussie Normals" Project for accepting the project submission and providing the samples that form the basis of this study. In particular I acknowledge Professor Juleen Cavanaugh for starting the project with me and providing expert knowledge and assistance in molecular genetic techniques. What I learned of molecular genetic techniques during the first phase of the project has been an invaluable contribution to my professional development.

I express thanks to my colleagues past and present for their understanding and support, and in particular to Gary Koerbin. It has been a long journey and the many listeners along the way have helped to provide clarity and reason.

To my father, Ray Nisbet, I can only aspire to your strength and resilience. To my mother Vivienne Moya Nisbet, who passed away before knowing that I was taking this journey, I know that you are with me.

I dedicate this journey to my beautiful daughters, Kimberley and Maddison Chia. They supported me for so many years and provided the strength and courage that I needed to make it to the end. For this I am eternally grateful.

One thing that I have learnt in taking this journey is that it takes conviction, determination, persistence and support to make it to the end.

**In loving memory of**

**Ray Nisbet**

**22/8/1927-16/5/2016**

*I byd it* **(I endure)**

# Abstract

The investigation and characterisation of genetic variation has progressed rapidly in the past decade. Concomitant with this is the advancement in technology facilitating the interrogation of the genome for different types of genetic variants. The global aim is to develop a complete map of genetic variation.

Several fundamental studies have provided the impetus for the investigation of structural variation. However early research was hindered by study design, differing quality control criteria, technical limitations and poorly defined "normal" populations culminating in low levels of reproducibility and inconsistency between studies. Few studies documented genetic variation in general population cohorts.

Reported here is a study of 64 Australian women representing a Western European descendent population using high resolution SNP microarray. This cohort represents a "healthy" general population that is clinically documented differentiating it from other studies. The focus of this thesis is to characterise the properties, chromosome distribution, gene content and identify rare and common events of copy number variation (CNV) and long contiguous stretches of homozygosity (LCSH). The cause and effect of technical limitations on the detection and reporting of these variants is considered with respect to the impact for diagnostic interpretation.

The findings in this thesis demonstrate a non-random chromosome distribution of CNV and predominance of small deletion (<10Kb) although the relative contribution of duplication and deletion equalised in CNV >100kb. Comparison

against clinical cohorts revealed a predominance of duplication, in particular CNV <250kb in the clinical cohort, a concept not previously reported. The breakpoint sequence signature of selected benign CNV identified base pair microhomology and investigation of the intervening genomic architecture provided the structure indicative of a replication repair mechanism.

Long contiguous stretches of homozygosity (LCSH) has roles in genetic diversity and pathogenesis of disease. The contribution of LCSH to the genome of outbred populations is reported to be greater than first indicated and population variation is apparent. This thesis describes the LCSH chromosomal landscape, identifies population specific events and illustrates the technical challenges of LCSH detection. The ascertainment of the background level of LCSH in the local population (0.7% of the genome) enabled determination of appropriate thresholds for interpretation in clinical diagnostics.

Documenting CNV and LCSH in a cohort of the general population contributes to the common goal of defining genetic variation in the human genome. Comparison against clinical cohorts identifies similarities and differences which will provide insight for the direction of further investigation. Furthermore the documentation of common and rare variants in a general population cohort provides evidence for the interpretation in clinical diagnostics.

# TABLE OF CONTENTS

# List of Figures

# List of Tables

# CHAPTER 1

# Introduction

## 1.1 Genetic Variation in the Human Genome

Structural variation is well described in the human genome. The spectrum of variants, abundance, distribution, mechanism of formation and role in genetic heterogeneity are yet to be fully characterised. Copy number variation (CNV) is a structural variant that is defined as the change in copy number of a segment of DNA larger than 1000 base pairs (bp) in length, when compared to a reference DNA. This may be represented as a gain (duplication) or loss (deletion) and has been described as the most common form of genetic variation representing 12% of the genome (1, 2). The discovery of CNV will contribute to further development of the map of human genetic variation.

The concept of duplication and deletion of DNA was recognised at the gross morphological level, with disease associations such as Down Syndrome from the gain of the whole chromosome 21 and Smith Magenis Syndrome with an interstitial deletion on chromosome 17 just visible on high resolution chromosome analysis. The seminal studies of Sebat et al. 2004, using genome-wide low resolution bacterial artificial clones, identified copy number variation (CNV) of segments of DNA in the normal population at a level of resolution not previously visible (3). Several studies rapidly followed, identifying duplication and deletion throughout the genome (1, 4).

By 2010 the significant contribution to the documentation of CNV was based on studies of cell lines derived from the International HapMap Consortium or meta-analysis of previously published data (4-8). The appreciation for the contribution of CNV to the genome was complicated by the registration of CNV in global

databases from studies based on different design, quality control parameters, acceptance criteria and limited, poorly defined cohorts (8). The International HapMap Phase 1 and 2 represent individuals from 270 individuals from 4 populations (9). However there is no medical or phenotypic data collected for these individuals. At this time there was limited studies of CNV in general population studies and few have utilized genomic DNA (6).

The predominant mechanism of derivation of recurrent duplications and deletions, with consistent breakpoints, has recently been described. However at the time of this study, the contribution of mechanisms to the derivation of common benign and rare CNV was not well understood. The initial focus of early reports was CNV discovery and few studies characterised breakpoints at the base pair level of resolution. Characterisation of the genomic architecture surrounding breakpoints and the intervening sequence is required to postulate the mutational mechanisms leading to the derivation of CNV. It is postulated that not one mechanism is involved and numerous interacting factors contribute to CNV formation (10).

Major advancements in technology culminating in the development of a range of microarray platforms facilitated CNV discovery. The technology evolved rapidly with improvements in marker design, density and genomic distribution (2, 5). CNV detection programs with built-in customised algorithms were developed to estimate DNA copy number from data generated from microarray platforms. Discordance of CNV detection and low levels of reproducibility was reported among studies (6, 11). In response to this and in accordance with improved resolution and platform design, previously reported CNV were reviewed and re-

defined (7). Comparative studies illustrated that not one algorithm can detect all variants. The impact to diagnostic interpretation has not been evaluated.

Another form of genetic variation that contributes to genetic heterogeneity and has roles in disease pathogenesis is long contiguous stretches of homozygosity (LCSH). The development of SNP microarray platforms accelerated the characterisation of LCSH in the human genome. Early linkage disequilibrium studies identified an increase of shared haplotypes in the offspring of consanguineous relationships. However the appreciation of the extent of LCSH in the outbred population was in its infancy prior to 2010 (12). Population studies revealed variation of the degree of LCSH and non-random population specific LCSH that reflected the ancestral and cultural background of populations (13-15). Studies of European populations identified variation of the degree of homozygosity consistent with migration patterns, geographic isolation, founder effects and localised consanguinity (14, 15). This was shown to be reflected in the LCSH length and number of events per individual. Characterisation of the degree and properties of LCSH in the Western European, Australian population has not been previously described. An appreciation of background levels of LCSH will assist in the clinical diagnostic laboratory with respect to establishing thresholds for analysis. The knowledge of common population specific LCSH events facilitates the targeted investigation of rare events for candidate genes.

## 1.2 Thesis Aims

The intention of this thesis is to perform a comprehensive investigation to document the contribution of CNV and LCSH to the individual genomes of a cohort of healthy Australian females of Western European descendent. It is anticipated that this investigation will contribute to the knowledge of human genetic variation and is the first study performed on a cohort of the Australian population. This study also represents the first to investigate the contribution of copy number change and long contiguous stretches of homozygosity to genetic variation in the same individuals of a general population.

A global ambition is to develop a map of human genetic variation. The investigation of general population cohorts such as this will contribute to the knowledge of human genetic variation in numerous ways. Firstly determination of copy number change of benign significance identifies regions of the genome and genes where alteration can be tolerated. The characterisation of the contribution of CNV and LCSH to an individual genome, in particular rare and novel CNV, will assist in evidence based interpretation in clinical diagnostics. The documentation of specific CNV and LCSH in general population studies will contribute to the chromosomal map of genetic variation in the human genome. Next, comparison against clinical cohorts illustrates the similarities and differences of benign CNV and LCSH properties, genomic position and gene content to provide evidence for further investigation or determination of pathogenic significance. Finally, knowledge of the mechanisms of derivation of benign CNV will provide insight of the contribution of mechanisms involved in genomic diversity.

## 1.3 Thesis Overview

The objective of this thesis is to perform a comprehensive analysis of a cohort of Australian females for genetic variants, CNV and LCSH, using data derived from high resolution SNP microarray. The cohort represented here are volunteers in the "Aussie Normals" community based study (16). All volunteers that passed initial screening criteria were interviewed, medical history recorded and nearly 100 chemical pathology tests performed. The investigation reported in this thesis began in 2010 when the knowledge of CNV and LCSH of benign significance in general populations was limited. At this time the application of microarray technology for the first tier investigation in clinical cohorts was recommended highlighting the need for evidence based interpretation.

This study focuses on the investigation of CNV and LCSH on chromosomes 1 to 22. It begins with the interrogation of one chromosome describing in detail the contribution of CNV and comparison with published control and clinical cohorts. The mechanisms of derivation of benign CNV are then considered from the genomic architecture of the CNV breakpoints of selected deletions. The properties of CNV and LCSH events on all autosomes are investigated and the application of general population cohorts demonstrated by comparison against clinical cohorts. In the final section the limitation of CNV detection is demonstrated and the implications to diagnostic application presented.

The thesis begins in Chapter 2 by providing a review of the knowledge of genetic variation that was documented at the time of initiation of this project with consideration of future directions outlined in throughout the thesis.

Chapter 3 presents comprehensive investigation of CNV on Chromosome 18. This chromosome was selected based on chromosomal length, number of CNV and paucity of published reports in the literature. This chapter provides a comprehensive analysis and stratification of benign CNV. The population frequency of selected CNV is determined using molecular based techniques. The similarities and differences of CNV to HapMap based population studies are considered and a framework for the investigation of CNV in the remaining autosomes is presented. Finally the CNV described in this chapter, which includes 3 novel variants, have contributed to the current knowledge of genetic variation by inclusion in the Database of Genomic Variants (DGV).

Chapter 4 contributes to the thesis in two ways. Firstly selected CNV are confirmed using molecular-based methods at the base pair level and secondly the information gained by the methodology employed provides the opportunity to consider the mechanism of formation of CNV. At the time of the study few CNV were characterised at the base pair level. This chapter details the DNA sequencing that was performed to validate selected deletions, characterise the genomic architecture of the breakpoints and postulate the mechanism of CNV formation.

In Chapter 5 the model of investigation developed in chapter 3 is employed to evaluate the contribution of CNV in chromosomes 1-22. This chapter contributes to a global initiative to describe, characterise and catalogue regions of copy number variation in the human genome. In this chapter I investigate the incidence, distribution, length and genomic content of benign CNV. The documentation of novel and rare CNV in this study will assist by establishing evidence for interpretation of CNV in clinical diagnostics. To demonstrate the utility of general

population cohorts I have compared the CNV properties identified in the "Aussie Normals" cohort with clinical cohorts.

The contribution of long contiguous stretches of homozygosity (LCSH) to the individual genomes of a cohort of the Australian population is investigated in Chapter 6. Homozygosity mapping, using methods such as linkage disequilibrium, has identified shared haplotypes in the human genome. This concept was used in gene mapping studies for the identification of candidate genes with autosomal recessive inheritance patterns. The identification of significant differences of patterns of distribution and genomic contribution of homozygosity in normal individuals provided evidence of the ancestral and cultural history of populations. The data generated from SNP microarray platforms can be investigated for evidence of homozygosity. To determine the contribution of LCSH to the local population and potential significance for diagnostic testing, I have evaluated the LCSH length, contribution to the genome and chromosomal distribution pattern. Review of LCSH events identified population specific LCSH. I have evaluated the limitation of LCSH detection in this study and reviewed thresholds that can be applied to the interpretation of LCSH in diagnostic laboratories.

In Chapter 7 I have demonstrated the application of general population cohorts such as the "Aussie Normals". In this chapter I have investigated CNV and LCSH in essential hypertension to illustrate the utility of general population cohorts in the investigation of common complex disease.

At the time of implementation of this study in 2010, significant discordance was apparent between studies. Chapter 8 reviews the limitations of CNV detection programs and associated detection algorithms. Here I illustrate the differences of

CNV detected by different algorithms from the same raw data. This is the first study to investigate the properties of discordant calls, providing further insight to the cause of differences of CNV detection by algorithms and software programs. I then apply this information to assess the potential impact on clinical diagnostics.

Presented in Appendix 2 is an alternative method of confirmation of duplications using fluorescence in-situ hybridisation (FISH) methodology. I have applied this method to enhance currently used methods of FISH investigation for the detection of small and novel CNV and demonstrate its application in the diagnostic laboratory.

Finally, the map of genetic variants for chromosomes 1-22, representing the chromosomal landscape for CNV, CNVR, CNV >100kb, proportion and genomic location of duplication and deletion and association with segmental duplication and repetitive elements in a cohort of "Aussie Normals" is presented in Appendix 1.

## 1.4 References

1. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number in the human genome. Nature. 2006;444(7118):444-54.
2. Carter N. Methods and Strategies for analyzing copy number variation using DNA microarrays. Nature Genetics. 2007;39(Supplement):5.
3. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, et al. Large-scale copy number polymorphism in the human genome. Science. 2004;305(5683):525-8.
4. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, et al. Origins and functional impact of copy number variation in the human genome. Nature. 2010;464(7289):704-12.
5. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet. 2009;12(5):363-76.
6. Pinto D, Marshall C, Feuk L, Scherer SW. Copy-number variation in control population cohorts. Hum Mol Genet. 2007;16 Spec No. 2:R168-73.
7. Matsuzaki H, Wang PH, Hu J, Rava R, Fu GK. High resolution discovery and confirmation of copy number variants in 90 Yoruba Nigerians. Genome Biol. 2009;10(11):R125.
8. Scherer SW, Lee C, Birney E, Altshuler DM, Eichler EE, Carter NP, et al. Challenges and standards in integrating surveys of structural variation. Nat Genet. 2007;39(7 Suppl):S7-15.
9. Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467(7311):52-8.
10. Bacolla A, Cooper DN, Vasquez KM. DNA structure matters. Genome Med. 2013; 5(6):51.
11. Pinto D, Darvishi K, Shi X, Rajan D, Rigler D, Fitzgerald T, et al. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nat Biotechnol. 2011;29(6):512-20.
12. Gibson J, Morton NE, Collins A. Extended tracts of homozygosity in outbred human populations. Hum Mol Genet. 2006;15(5):789-95.
13. Alkuraya FS. Autozygome decoded. Genet Med. 2010; Dec;12(12):765-71.
14. McQuillan R, Leutenegger AL, Abdel-Rahman R, Franklin CS, Pericic M, Barac-Lauc L, et al. Runs of homozygosity in European populations. Am J Hum Genet. 2008 ;83(3):359-72.
15. Nothnagel M, Lu TT, Kayser M, Krawczak M. Genomic and geographic distribution of SNP-defined runs of homozygosity in Europeans. Hum Mol Genet. 2010;19(15):2927-35.
16. Koerbin G, Cavanaugh JA, Potter JM, Abhayaratna WP, West NP, Glasgow N, et al. 'Aussie normals': an a priori study to develop clinical chemistry reference intervals in a healthy Australian population. Pathology. 2015;47(2):138-44.

# CHAPTER 2

# Copy Number Variation and Contribution to Genomic Diversity

## 2.1 Introduction

The understanding of human genetic variation has evolved rapidly over the past decade. This progress is attributed to technological developments such as high-throughput genome-wide microarray and will continue to evolve with next generation sequencing. The development of these technologies and application to research and clinical diagnosis has enabled the mining of normal population and pathogenic case cohorts, defining a spectrum of types of genetic variation (1-5). These genetic variations have been described in the form of single nucleotide polymorphism (SNP), short insertion/deletion variant (indel), tandem repeats, segmental duplication and structural variation that includes microscopically visible inversion, translocation, duplication and deletion (1, 3-8).

In 2004 the first studies using low resolution whole genome microarray detected sub-microscopic duplication and deletions in "normal individuals" (9, 10). This was later defined as variation in DNA copy number when compared to a reference genome (2, 11, 12). Development of SNP microarray platforms expanded the appreciation of genetic variation to include the detection of copy neutral loss of heterozygosity, culminating in an insight into ancestral population history and consanguinity (13-22) .

Several studies have contributed to the knowledge of human genetic heterogeneity (1, 2, 9-11, 23). These studies combined with the International HapMap Consortium, Wellcome Trust Case Consortium and 1000 Genome Project have provided insight to the scope of genetic diversity. The objective of the initial International HapMap Project (Phase I and II) was to identify common SNP variants with a focus on facilitating expansion of genome-wide association studies

(3). HapMap 3, released in late 2010, expanded the number of individuals in the cohort from 270 to 1,184 and the number of populations from 4 to 11 (3). Of importance, however, is the inclusion of copy number variation (CNV) detection. This study provided CNV discovery with high accuracy and established a public resource of cell lines and database of CNV (3, 24).

Many of the studies of "normal" population cohorts prior to 2010, and initial CNV recorded in global databases such as The Database of Genomic Variants (DGV, http://dgv.tcag.ca), are derived largely from HapMap Phase I and II (2, 3, 11, 25-28) and poorly defined control cohorts (6). There are few studies of CNV in non-HapMap general populations (29, 30).

A number of limitations of the early studies are apparent. Lymphoblastoid cell lines are derived from transformed B cells. Accordingly CNV estimations may not be indicative of the true frequency of CNV (4, 6, 30). The study by Wellcome Trust Case Consortium in 2010 demonstrated the potential for false positive disease associations where transformed cell lines are used as the primary sample (4). Next, the aim of this phase of the HapMap project was for the detection of common SNPs and mining for relative copy number was not the intended purpose. As a result the data outputs may not reflect CNV indicative of the general population (30). Furthermore, the frequency of some CNV appeared to differ among populations in the HapMap cohort (1-4, 23, 25, 26, 30, 31). When used as a reference source in CNV detection, population specific CNV may lead to incorrect estimation of CNV incidence (30). Finally, no phenotype or medical information on the HapMap individuals was collated and representation of a "healthy" general population is unknown (6, 30).

CNV discovery was limited in early studies by platform design and restricted to high confidence large and simple biallelic variants (30). Some studies only reported on deletion and not duplication (2, 11) leading to ascertainment bias in the initial entries in DGV. With the release of higher resolution platforms subsequent studies reviewed and redefined CNV from early studies of HapMap cohorts (25, 26), or provided a meta-analysis of data available in the public domain (25, 28). Furthermore differences in platform and CNV detection algorithm design have influenced the estimation of CNV properties resulting in inconsistencies between studies (30, 32-36). This is further impacted by differences in study design and the level of pre- and post-analytical screening and CNV confirmation, resulting in high false positive and negative rates (6). As such the contribution of CNV to genomic diversity is poorly characterised in early studies (30, 37-40).

The application to clinical diagnostics for the detection of chromosome abnormalities was demonstrated in a study by Slater et al. 2005. The authors evaluated the utility of SNP based microarray on a range of samples referred to a routine diagnostic service, some of which failed to yield a result on conventional cytogenetic analysis (14). The transition of microarray technology to clinical diagnostic application was rapid and by 2010 was considered the “gold standard” for diagnostic cytogenetic analysis and recommended as “first tier” for the investigation of patients with developmental delay, mental retardation, congenital anomalies and autism spectrum disorders (14, 21, 41, 42). At this time the cataloguing of CNV was still in the discovery phase. Of particular concern was the paucity of data for general population studies from which comparison and determination of benign significance could be made (24, 30, 39).

The appreciation of the clinical significance of CNV developed over the initial years. Association of duplication and deletion of DNA involving genes with syndromes such as velo-cardio-facial syndrome (VCFS), Prader Willi Syndrome, Smith Magenis Syndrome is well established (24, 43, 44). However gene involvement in benign CNV and contribution to phenotypic diversity is evolving. The potential role in the predisposition to common complex disease such as hypertension, diabetes and hyper-cholesterolemia is yet to be fully characterised (25, 33).

The mechanisms by which CNV are formed have been the subject of scrutiny since the realisation of the extent of this variation in the human genome. Non allelic homologous recombination (NAHR) is a well described mechanism of CNV formation (7, 11, 12, 31, 37, 45). Other mechanisms have been proposed for CNV with and without homology at breakpoints, the evidence suggesting a minimum of 2bp microhomology is required for some of these mechanisms (46, 47). But the relative contribution of these and the role of genomic instability to the prevalence of CNV in the genome are yet to be characterised (27, 29, 48, 49).

This chapter reviews previous studies of the properties and prevalence of CNV, the mechanisms that lead to the formation of CNV and highlights the need for, and applications of, "healthy" general population studies. It discusses the factors that influence CNV detection and provides an understanding of the knowledge of CNV at the time of this study.

## 2.2 What is CNV ?

CNV is defined as the change in copy number of a segment of DNA, measuring >1kb in length, when compared to a reference genome (6, 33). This variant is described

in terms of gain (duplication) and loss (deletion) of a segment when compared to the reference (6) (Figure 1). Other acronyms include copy number polymorphism (CNP)(9), a term that reflects the frequency in the population (6).

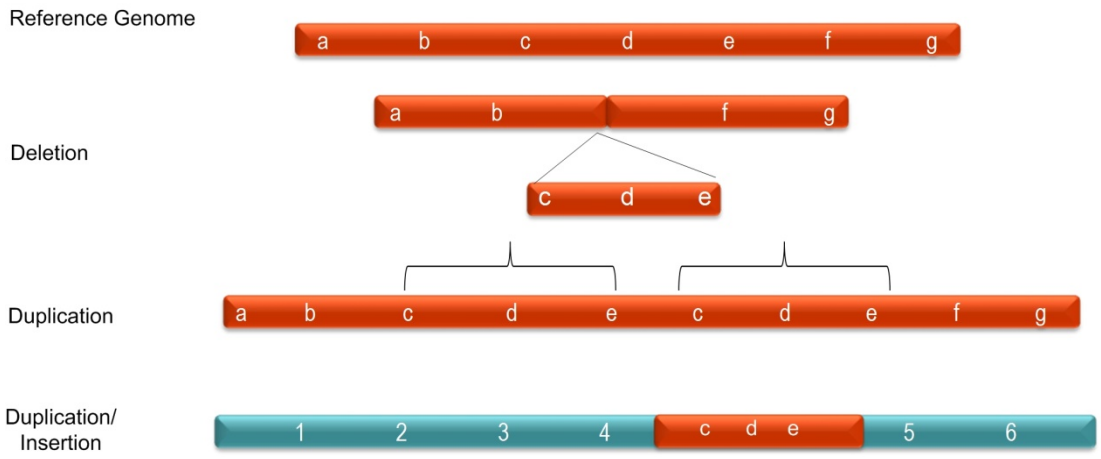



**Figure 1**. Schematic diagram of CNV showing deletion of a segment of DNA (cde) compared to the reference genome and duplication of DNA (cde) in tandem orientation and as an insertion to another location within the genome.

## 2.3 Functional Impact of CNV

The effect of duplication and deletion on gene expression may be due to direct involvement of genes or by indirect mechanisms that impact gene function (6, 8, 23, 33, 50). Early association of copy number change with pathogenic conditions were made by the correlation of microscopically detectable chromosomal

rearrangements with candidate genes and postulation of pathways that involved structural and functional aspects of gene expression (29, 33, 40, 51).

Dosage sensitive genes are genes that require the presence of both copies to maintain normal function. Deletion of a sequence of DNA that harbours dosage sensitive genes resulting in haploinsufficiency is well described in the pathogenesis of dominant pathogenic conditions (51). These genes are enriched in regions of well described abnormal phenotypes and may reflect a network of genes associated with developmental functions (51).

CNV may confer pathogenesis where deletions or loss of gene function alterations unmask autosomal recessive gene mutations on the remaining chromosome (8, 29, 52). Alternatively, duplication of a DNA segment involving an un-mutated gene may convey an advantage by masking the mutation on the homologous chromosome (29). Genes located at the breakpoints may form fusion genes subsequent to the deletion or duplication leading to disease conditions. This has been demonstrated in malignancy (53-55) and other disease states (29, 56).

Disruption of gene coding sequences by duplication or deletion of DNA segment may indirectly lead to alteration of gene function (8, 29, 33, 50, 52). Protein-coding genes consist of coding exons that are translated to proteins after the excision of the non-coding intervening sequence, introns (50) (Figure 2). Gene regulatory elements can be located up to 1 Mb either 5' or 3' of the transcriptional unit (8, 50). These comprise of promoters, of which the core promoter has a role in assembling the protein complex for RNA synthesis. Some genes also have regulatory promoters, upstream of the core promoters, as transcription factor binding sites. Enhancers are *cis*-acting regulatory units with a role in spatial conformation of

chromatin, making the chromatin available for transcription and gene expression (33, 50). These are distantly located and may be situated within an intronic region of a neighbouring gene. Some genes may be complex in structure with alternating exon/enhancer configuration (50).

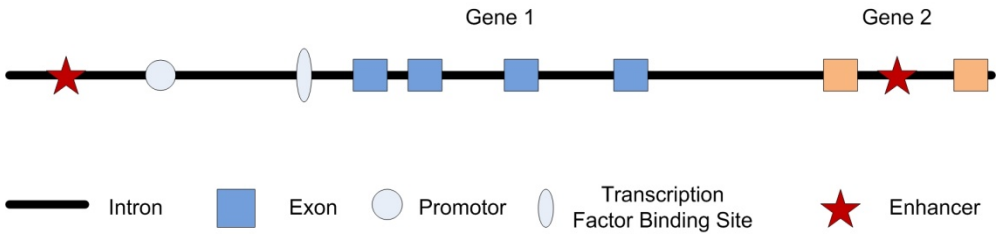



**Figure 2**. Schematic diagram of components of a protein coding gene. Elements that have a role in gene function and regulation of expression are shown (8).

Alteration to any aspect of gene structure may lead to loss of gene function or altered gene expression (Figure 3) (8, 29, 33, 50). Deletion of the DNA segment between the regulatory elements will affect gene function by preventing access of the transcription factors to the transcription binding site. Likewise deletion of regulatory elements will prevent the completion of formation of the transcription factor unit. Duplication or deletion resulting in relocation of the transcription unit away from or within close proximity of *cis*- regulators will modify gene expression by potentially modifying the chromatin structure that is required for transcription. DNA that is compacted is not "open" to proceed to transcription and translation and the enhancer has a role in this process (8, 29, 33, 50).

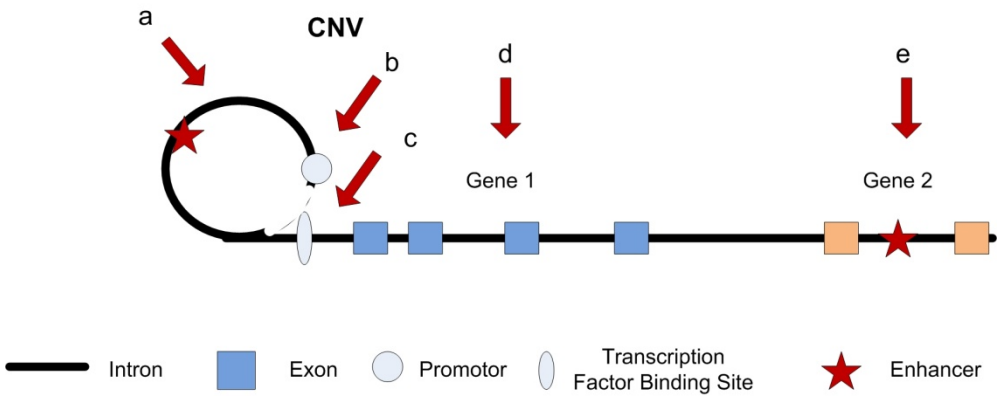



**Figure 3**. Ways in which structural rearrangement can alter gene expression, a) CNV can alter chromatin structure or dissociate the regulatory elements, b-c, e) loss of the regulatory elements prevents formation of the transcription unit and d) CNV involving dosage sensitive genes can lead to changes in gene expression (8).

Examples of the disruption of long range control of gene function is illustrated by Williams Beuren Syndrome where the deletion impacts on expression of neighbouring genes that are not involved in the deleted sequence (33). Intronic deletions of *CNTNAP2* gene (7q35) involving the regulatory element located in intron 1 have been reported in association with autism spectrum disorder and language delay (57). Aniridia (absence of iris) is known to be associated with deletion resulting in haploinsufficiency of pair box/homeodomain gene (*PAX6*) at 11p13. Although a dosage sensitive gene, Aniridia is reported without deletions of *PAX6*. Here deletions 200kb downstream of the transcription unit involving *cis*-regulatory elements were detected (50). Other examples of genetic conditions with evidence of disruption of long range gene control include campomelic

dysplasia (SOX9), eye development (*FOXC1*) and holoprosencephaly (*SIX3* and *SHH*) (50). The genetic basis for some of these conditions has been identified historically by chromosome rearrangements (50). It is anticipated that the implementation of high-resolution microarray will identify more examples of perturbation of long range gene control as a cause of pathogenic conditions.

## 2.4 The Development of Microarray Technology

Chromosome studies with an estimated resolution of 5-10Mb, where metaphase spreads are microscopically analysed for changes to the pattern of banding compared to the expected band pattern as illustrated in idiograms, was the gold standard for clinical cytogenetic testing (23, 38). Mutations, indels and variable number tandem repeat (VNTR) are detected using DNA sequencing methods, at an effective resolution of 1-700bp (38). The discovery of microarray technology provided whole genome coverage for analysis of DNA between these ranges. Numerous array platforms have been developed over a short time period. The rapid change of these platforms was driven by the need for improved resolution to elucidate the contribution to genetic variation and characterise CNV properties.

Comparative genomic hybridisation (CGH), first developed in 1993 (58) provided the foundation for future technological developments. This method compared a differentially labelled test DNA and a normal reference DNA using fluorescence in-situ hybridisation (FISH) which are applied to metaphase spreads of a normal individual. The presence of gain or loss for a segment of DNA in the patient sample

compared to the reference sample was determined by measuring the ratio of fluorescence. However, the resolution of this method was limited at 5-10 Mb (38).

The first of many developments was the introduction of array based CGH (aCGH) where large-insert clones are assembled into over-lapping contigs and immobilized onto glass slides using robotic technology (38). The resolution of aCGH is dependent on the size and number of clones labelled onto the glass slide. Copy number is measured by comparing the ratio of fluorescence between the differentially labelled patient and reference DNA when hybridised to the clones on the slide. The clone position can be mapped to the human genome to provide the position of detected copy number change (38).

Iafrate et al. 2004 and Sebat et al. 2004 reported on the detection of copy number change in healthy individuals using aCGH technology. Bacterial artificial clones (BAC) measuring 80-200kb in length and located across the whole genome with an inter-spacing of 1Mb between clones, provided whole genome analysis of copy number change at limited resolution (10, 23, 38). Smaller fosmid measuring 40kb in length were later applied to this technology, providing improved resolution and signal-noise ratio (23, 37, 38, 59). The interpretation here is based on the assumption that pair-end sequences for each fosmid should align to the reference genome at 40kb intervals and deviation from this size is indicative of loss or gain of intervening DNA sequence (37, 59).

Oligonucleotide aCGH, utilizes synthesized oligonucleotides that are spotted onto slides and the resolution is determined by the density of oligonucleotides. This methodology improved detection of copy number change but had a poor signal to noise ratio. Representational oligonucleotide microarray analysis (ROMA)

provided improved signal to noise ratio and probe inter-spacing of 35kb. This method is based on digestion and amplification of fragments (up to 1.2kb) and hybridization to a slide labelled with pre-selected matching oligonucleotides (9). Oligonucleotide aCGH has the benefit of development to high resolution with the potential for improved probe density but is not without limitations in CNV detection, particularly with respect to the absence of probe coverage in repeat sequences of the genome (38).

High through-put genotyping array platforms that utilised the whole genome distribution of single nucleotide polymorphisms (SNP) were developed for the determination of SNP genotypes. This information was applied in genome-wide disease association studies (3, 23, 38, 60, 61) and clinical diagnostic applications (14). The application of this platform for CNV detection by measuring intensity information to determine genomic copy number (38, 62) or by inferring copy number change from SNP genotypes has been documented in numerous studies (23, 38, 60, 61). McCarroll et al. 2006 based this inference on genotyping errors and deviation from Hardy Weinberg equilibrium (23, 38, 61) whereas Conrad et al. 2006 predicted CNV from errors in Mendelian inheritance (23, 38, 60). Both of these studies were applied to the detection of deletions rather than duplications (23, 38, 60, 61).

Commercial availability has made genotyping microarrays universally accessible by research and diagnostic laboratories (14, 38). To overcome the limitations of early oligonucleotide arrays, non-polymorphic probes were included in the next generation SNP array platforms, to improve probe coverage within regions of low SNP density such as highly repetitive regions of the genome (30, 38). Concomitant

with this is the customisation of marker distribution for CNV detection with the inclusion of probes that encompassed regions of the genome where there is CNV of established clinical relevance (63).

## 2.5 CNV Detection Algorithms

Computational analysis of SNP genotyping and aCGH data is required to detect changes in copy number (64, 65). Numerous informatic tools have been developed for the analysis of raw data for CNV detection including defined thresholds (66), statistical approaches (67, 68, 123) and adaptation of these in customised programs (64, 69).

In principle the process of detection of copy number change includes pre-processing, data alignment, statistical analysis and post processing of input data (62, 65). During pre-processing the LogR data is normalised against the sample mean or median for each hybridisation and adjusts for systematic error. The data alignment and statistical analysis varies among algorithms and can be tailored for specific applications. In essence these methods identify the genomic location of LogR mean changes and estimate breakpoint locations. Additional statistical analysis is required to assign copy number state. The assigning of biological meaning to the data is done during post processing (62, 65).

### 2.5.1 The Algorithms

The commonly used algorithms are Hidden Markov Model and Circular Binary Segmentation calculation (36, 65, 67, 68, 123). In brief, the Hidden Markov Model

works on the assumption that the observed intensities are related to the hidden copy number state at each locus. The copy number state of each locus is dependent on that of adjacent loci. A transition matrix is applied to predict the probability of changing between states. Bayesian and non-Bayesian calculations can be applied to set parameters from the data. The copy number state at a given locus is then predicted using the Viterbi algorithm (36, 64, 69, 123).

Circular Binary Segmentation calculations are applicable to aCGH data analysis although they have been adapted for SNP array analysis (36, 65, 67, 68). This model divides the chromosome into regions and converts the intensity values to reflect equal copy number. The algorithm continues along the chromosome comparing the copy number state in contiguous segments to identify a region with a copy number state that differs from adjacent regions (36, 65, 67, 68). The significance is then determined using a permutation reference distribution (65).

Optimisation of these algorithms has resulted in the development of numerous CNV analysis tools (36, 62, 64, 65, 69), proprietary and customised, that have different strengths and weaknesses. For example QuantiSNP, a publicly available free program uses the Hidden Markov Model and includes an Objective Bayes analysis (64). In this model the parameters are learnt from the data using a re-sampling framework and the maximum likelihood of hidden states can be inferred. Each region of copy number change is assigned a Bayes factor which indicates the call confidence in comparison to the null hypothesis, thereby enabling the filtering of data based on the probability score. This algorithm is applicable to high-throughput analysis and permits comparison of CNV across many samples and within chromosomes (36, 64, 69). The authors used previously characterised

samples and demonstrated improved robustness and breakpoint accuracy of CNV detected (64).

PennCNV uses an integrated Hidden Markov Model that incorporates information from individual SNP allelic intensity, distance between SNPs and B allele population frequency. In contrast to QuantiSNP, PennCNV uses an emission probability of LogR ratio and B allele frequency in the hidden state to determine the likelihood of transition of copy number change between adjacent SNPs. The Viterbi algorithm is incorporated for the purpose of CNV calling by inferring the most probable state for all SNPs along a chromosome (69, 70). This algorithm uses "state specific and distance dependent transition" to predict changes in copy number state and uses a large population reference source to index the B allele frequency, the latter providing more accurate estimation of genotypes (69). However, the method employed by PennCNV restricts the calling of multiple changes in copy number state within a defined region and only the most probable outcome is reported. Accordingly PennCNV is not the preferred CNV detection program to detect segmental copy number change of heterogeneous cell populations in somatic tumours (70). Xu et al. 2011 in a study comparing the CNV outputs from PennCNV and QuantiSNP reported higher concordance of CNV outputs from PennCNV than QuantiSNP when applied to single data. The authors presented a further optimisation of the integrated HMM model used in PennCNV and included a regression model to test CNV associations with disease (70).

## 2.5.2 Factors affecting CNV detection

Biological factors may affect data quality and impact on the power of algorithms to detect CNV. Wave artefact is such an example and may be present as small fluctuations in LogR ratio (38, 62, 71, 72). Wave artefacts may impact CNV calling thresholds and the ability of CNV detection algorithms to detect copy number change leading to false calls (38, 71, 72). Marioni et al. 2007 in an investigation of the aetiology of this anomaly demonstrated a strong correlation of peaks and troughs of LogR values with the GC content of the clones (71). A subsequent study by Diskin et al. 2008, investigated genomic and technical factors that may influence the quality of signal intensity data generated by SNP array platforms. They demonstrated that GC content is a significant factor in the appearance of genomic waves. To compensate, wave correction calculations involving normalisation of datasets has been included in CNV detection algorithm software programs (38, 71, 72).

## 2.5.3 Comparison of algorithms

Algorithms were adapted for the purpose of improved CNV detection, resolution (64, 69) or for a specific application (36, 70, 73, 74). Several studies have evaluated sensitivity, specificity, reproducibility and accuracy of CNV outputs from multiple CNV detection algorithms (32, 34-36, 65, 74, 75). The results show that CNV detected by programs using the same DNA source yielded inconsistent results (6, 30, 32, 45). Discordance is reported for CNV incidence suggesting a significant

false discovery rate (30, 32, 65). With regard to CNV detection, a CNV is considered as a false positive when a copy number of 0, 1, 3 or 4 is assigned to a region that is actually diploid (copy number 2). A false negative is when a region is assigned copy number 2 when it is actually 0, 1, 3 or 4 (74).

Pinto et al. 2011 performed a comprehensive analysis of the performance of 11 platforms and 5 CNV detection algorithms, including CNV Partition (Illumina, Inc.) and PennCNV, using the same raw data. The authors demonstrated <50% concordance among the algorithms and <70% reproducibility of CNV detected between replicate experiments. The study illustrates that large CNV are less likely to be affected, however regions of complex architecture are vulnerable to inconsistency in CNV detection (32). Zhang et al. 2011 compared the performance of 4 CNV detection algorithms (Birdsuite, Partek, HelixTree and PennCNV) and evaluated the performance against the detection of rare and common CNV (75). This study demonstrated that the number of markers within a CNV correlates with recovery rate. Dellinger et al. 2010 compared the statistical power and false positive rates for 7 CNV detection algorithms. Using statistical methods the authors ranked the performance of each algorithm for a number of factors (34). They concluded that HMM based models performed better and that QuantiSNP as the best performing CNV detection algorithm. This contradicts the finding of Xu et al. 2011 which found that PennCNV out performed QuanitSNP achieving a higher concordance of CNV calls (70).

The consistent finding among all studies is that not one program will detect all types of structural variation or accurately identify all CNV. The power of CNV detection algorithms can vary on the same platform and is relative to the quality of

raw data generated (32). These studies highlight the variability among algorithms and demonstrate the need to assess the limitation and performance characteristics of CNV detection programs for use in clinical diagnostic applications. The performance of CNV detection programs can be improved by using a high quality sample data set to standardise parameters prior to full analysis (34). Further improvement in commercially available platforms would be the provision of more than one CNV detection algorithm (32, 65). The benefit of this is to allow the technician to select the algorithm appropriate for the experiment and CNV outputs can be compared or merged for more robust analysis and confirmation of CNV calls.

## 2.6 QC Metrics and Confirmation of CNV

The QC metric applied to CNV data outputs is essential to ensure confidence of CNV calls and minimize the likelihood of reporting false positive CNV. Experiment design, quality filtering strategies and modes of CNV confirmation vary among early studies and may have contributed to the inconsistencies of estimation of CNV prevalence (6, 7, 30).

Experimental quality control to optimise sensitivity and specificity varied among studies (Table 1). The strategies employed include definition of the number of markers within a CNV, calculation of the standard deviation of the normalised log intensity ratios (LogR) and the threshold of LogR values for duplication and deletion. CNV not falling within these defined thresholds were excluded. The determination of false negatives was not emphasized and is a limitation of the scope of microarray technology.

The difference in the strategies applied for the confirmation of detected CNV is another anomaly of early studies (6, 7, 30) (Table 1). Some studies based the confirmation of large numbers of CNV on detection in previous reports or if the CNV is present in multiple samples (2, 27, 31, 39). Fluorescent in-situ hybridisation (FISH), multiplex ligation dependent probe amplification (MLPA) and quantitative PCR (qPCR) are used for confirmation of some CNV, but these methods are size or sample dependent and not suitable for confirmation of all CNV (9, 21, 23, 45). Metaphase or interphase FISH on cultured cells is well described for deletion > 150kb (21, 45) but the demonstration of small duplications is challenging (45). An alternative approach was taken by Pinto et al. 2007 where the authors used multiple CNV detection algorithms and a CNV call was considered "real" when detected in 2 or more of the CNV detection algorithms. Confirmation of a subset of the study cohort by an alternative microarray platform was a method used by Conrad et al. 2010, Komura et al. 2006 and Redon et al. 2006. Characterisation of copy number change by sequencing (3, 5, 11, 37, 76, 77), and comparison against the reference genome was performed by few studies (9, 11).

The standardisation of quality parameters is recommended in future studies including a reference data set to provide benchmarking and validation (6). This would ensure accurate cataloguing of structural variants that can be used for a wide variety of research and diagnostic purposes (6, 37).

**Table 1.** Comparison of platforms, methods of confirmation and QC metrics in published studies reported prior to 2010.

| Reference | Platform | Confirmation | QC Metric |
|---|---|---|---|
| Baross et al. 2007 | Affymetrix 100k | dCHIP, CNAG, CNAT<br>FISH<br>Karyotyping of LCLs<br>qPCR (single deletion) | 4 markers |
| Bruno et al. 2009 | Affymetrix 250k | FISH >150kb losses on MLPA <150kb gains and losses | 7 markers<br>sd <0.26 |
| Conrad et al. 2010 | NimbleGen 2.1M<br>Agilent 105k<br>Illumina Infinium | Alternate platform<br>Comparison pooled data against discovery data | 10 markers |
| Itsara et al. 2009 | Illumina platforms | Alternate platform<br>Compared to previous documented CNVs | 3-10 markers<br>sd <0.25 |
| Komura et al. 2006 | 500k EA Array<br>Whole Genome Sampling Assay (WGSA) | qPCR<br>Mass spectrometry<br>PCR | 4 markers |
| Perry et al. 2008 | Agilent 60k | Repeat analysis<br>Comparison with previous studies | 2 markers<br>sd <0.25 |
| Pinto et al. 2007 | Affymetrix 500k | Detected by minimum of 2 algorithms from dCHIP, CNAG, GEMCA | Not specified |
| Redon et al. 2006 | Affymetrix 500k<br>Whole Genome Tile Path (WGTP) Bac | FISH<br>Alternate platform<br>CNV calls compared to database<br>qPCR | 4 markers<br>sd <0.22 (Affy)<br>sd < 0.047 (WGTP) |
| Sebat et al. 2004 | ROMA | FISH<br>Repeat testing with different restriction conditions | Not specified |
| Shaikh et al. 2009 | Illumina 550k | qPCR <10 SNP<br>Affymetrix 6.0<br>FISH MLPA | 2 Markers |
| Sharp et al. 2005 | Bac clones | Compared to previous documented CNVs<br>FISH | Not specified |

## 2.7 CNV Discovery

The mining of genome-wide investigations of "healthy" individuals has revealed the substantial contribution of copy number variation (11, 31). Subsequent to the studies by Sebat et al. 2004 and Iafrate et al. 2004, documentation of CNV flooded the literature. Incremental improvements in the resolution and robustness of microarray platforms have lead to a better understanding of the extent of CNV in the genome. The association of large and sub-microscopic duplication and deletions to well described syndromes has been previously established (21, 24, 41, 43, 44, 63, 78-82). However the advent of genome-wide microarray technology has identified the breadth of CNV in the "healthy" population.

### 2.7.1 HapMap based CNV discovery

The HapMap Phase 3 project is designed primarily for genome-wide association studies and comprised of individuals recruited from northern and Western Europeans living in Utah, USA (CEU); Han Chinese from Beijing (CHB); Japanese, Tokyo (JPT); Yoruba, Nigeria (YRI); African ancestry, USA (ASW); Chinese, Denver USA (CHD); Gujarti Indians Texas, USA (GIH); Luhya, Kenya (LWK); Maasi Kenya (MKK); Mexican ancestry, USA (MXL); Tuscan, Italy (TSI). These cell lines are made possible by The Human Genome Project, the SNP Consortium and The International HapMap Project (3). Early studies utilised the accessibility of HapMap cell lines for the investigation of benign CNV and provided initial evidence for differences in structural variation among populations (2, 3, 11, 25-28, 37).

Experimental applications of the HapMap cell lines include use as a control cohort for the investigation of CNV enrichment in pathogenic cohorts (44) and to demonstrate the robustness of the microarray platform or algorithm under investigation (25, 62, 69, 83).

HapMap cell lines have been extensively investigated for structural genetic variants (2, 3, 5, 11, 25-28, 37) (Table 2). Collectively these studies contribute to the appreciation of the contribution of benign CNV to genetic diversity, which is estimated at approximately 12% to under 20% of the human genome (6, 11, 30, 31). CNV was found to be non-unique with up to 50% detected in more than one individual some of these sharing consistent breakpoints (2, 11, 25). When considering total CNV, deletions outweighed duplications, though some studies demonstrated that the reverse applied for small CNV (2, 3, 11, 25, 26, 28, 62, 83). RefSeq and OMIM genes are encompassed in benign CNV (2, 3, 11) and duplication is enriched when compared to deletion (26). Early studies reported a predominance of CNV <100kb (9, 62) whereas later studies, utilizing high resolution platforms, provided limited CNV length analysis and CNV <10kb predominated (25, 26, 28, 83) although some variation in reports pertained (27, 30).

**Table 2.** Published studies of CNV discovery using microarray technology

| Reference | Platform | Population | Summary Findings |
|---|---|---|---|
| **HapMap Samples** | | | |
| Conrad et al. 2010 | NimbleGen 2.1M<br>Agilent 105k<br>Illumina Infinium | 270 HapMap Phase II | Median 29kb; Max length 1.28Mb; 51% non unique;<br>Loss (77%); 13.4% overlapped RefSeq; NAHR proposed for large CNV |
| HapMap3 2010 | Affymetrix<br>Illumina | 1,184 HapMap Phase III | Median length 7.2kb; Loss 92%; 33.5% overlap RefSeq genes;<br>Gain>loss for gene content |
| Ju et al. 2010 | Agilent 24M<br>Pair End seq. | NA10851 | Characterised CNV in one sample. Median length 2.7kb;<br>Loss > gain and size dependent;<br>Overlap SD in 10.9% (Loss) and 43% (Gain) |
| Komura et al. 2006 | 500k EA Array Whole Genome Sampling Assay (WGSA) | 270 HapMap Phase II | Compared to Redon et al. 2006; Max 3.4Mb; Median 71kb;<br>24 CNV per individual;<br>Loss >Gain (1.2:1) |
| Matsuzaki et al. 2009 | Affymetrix<br>custom array | 90 (YRI) | Re-analysed samples previously studied; Median length 4.9kb;<br>33% novel; singletons (44%);<br>67% overlapped DGV;<br>2-41% concordant with other HapMap studies; Loss > gain |
| Park et al. 2010 | Agilent 24 M | 30 (KOR,CHB,JPT) | Population specific CNV; 72.6% loss;<br>Gain>loss with gene content; CNV <10kb predominated; |
| Pang et al. 2010 | Agilent 24 M<br>Nimblegen 42 M<br>Affymetrix 6.0<br>Illumina 1M | Venter Genome | Platform comparison CNV <1Mb; 65.6% gain; ratio of dup vs del varied depending on platform; Loss > in CNV <10kb; Median 1.2kb-42.7kb; CNV incidence and length platform dependent |
| Perry et al. 2008 | Agilent 60k | 30 (YRI), (CEPH), (ASN) | Compared calls to Redon et al. 97% concordance;<br>Size difference between studies noted;<br>CNV<100kb predominate; |
| Redon et al. 2006 | Affymetrix 500k Whole Genome Tile Path (WGTP) Bac | 270 HapMap Phase II | 12% genome; Median length 81-228kb; ~ 50% non unique; Loss> gain (2:1); 24% overlap SD; 58% overlap RefSeq genes |

**Table 3.** CNV discovery in published population based studies

| **Population Based Samples** | | | |
|---|---|---|---|
| Itsara et al. 2009 | Illumina platforms | 2500 "Wellness" clinics | 16% genome; Loss >gain in CNV <100kb; CNV <100kb in 65-80% of samples; CNV <500kb in 5-10% and >1Mb in 1% of cohort; Enrichment CNV with SD; 94% non unique; 71% CNV >100kb are rare; > 500kb are singletons |
| Pinto et al. 2007 | Affymetrix 500k | 506 PopGen Individuals | Median 185kb; Gain >Loss (2.3 fold); 50-61% not in DGV; population distribution |
| Qiao et al. 2006 | aCGH Bac, Spectral Genomcs | 27 unrelated healthy individuals | 1 Mb resolution; FISH confirmation; 38% novel |
| Shaikh et al. 2009 | Illumina 550k Affymetrix 6.0 | 2026 paediatric (CHOP) | 78% non unique; 26.9 per individual; 19.4% genome; Median 7.2 kb; Loss (88%); CNV <10kb predominate |

## 2.7.2 Population based CNV discovery

Investigation of copy number variation in control populations provides evidence upon which comparison can be made in clinical diagnostics. Many studies of the HapMap population using transformed cell lines defined a baseline of copy number variation (2, 3, 5, 11, 25-28, 37). However, there is no phenotypic information available for the HapMap cohort and there are few studies of non-HapMap control populations (Table 3) (6, 29-31, 39). In a comprehensive study of phenotypically normal non-HapMap individuals, Itsara et al. 2009 analysed the CNV of lymphoblastoid cell lines of 2,500 individuals. This study demonstrated a high incidence of low frequency CNV and a correlation of CNV length with prevalence in

the healthy population (31). Shaikh et al. 2009 used genomic DNA from 2026 "disease-free" children recruited from The Children's Hospital of Philadelphia HealthCare Network (CHOP). They reported an average of 26.9 CNV per individual with an average length of common CNV of 38.3kb. Heterozygous deletions predominated and accounted for 81.2% of CNV and common CNV (observed in two or more individuals) represented 77.8% of CNV detected (39). In recognising the limitation of previous studies Pinto et al. 2007 investigated 506 unrelated healthy individuals who were enrolled in the PopGen Project. This study revealed a high detection rate of novel CNV and an average CNV length of 369kb. (30).

## 2.8 CNV Properties

Few studies have provided a stratified investigation of the properties of benign CNV. Limited analysis of CNV properties has demonstrated an enrichment of CNV <10kb and deletions predominate in this size category (3, 11, 25, 26, 28, 31, 62). The incidence of CNV negatively correlated with CNV length (26, 28, 31, 39) with few exceptions (27). Chromosomal CNV distribution, illustrated in a few studies, shows that CNV is ubiquitous in the genome (11, 26, 31, 83). These studies considered the chromosomal distribution (83), association with genomic architecture (11, 31) and the chromosomal location of population specific CNV (26). An association of CNV with telomeric and centromeric regions (7) and association with segmental duplication was proposed (7, 12, 76).

The relative proportion of duplication and deletion varies among studies of "normal" individuals. Duplication exceeded deletion in the study by Pinto et al. 2007 representing a 2.3 fold occurrence of duplication compared to deletion. However many studies have reported a higher proportion of deletion than duplication (2, 3, 11, 27, 31, 60) when considering total incidence of CNV, but duplication exceeded deletion in large CNV (31). The comparison among studies is complicated by technical factors. Bias may have been introduced in the early studies due to the technical challenge of detecting and confirming duplications (2, 3, 11, 60). Concomitant with this is potential ascertainment bias introduced by study design. For example CNV length categories are not indicated in all studies (30). The proportion of duplication and deletion has been shown to correlate with CNV length (31). The over-representation of duplication in large CNV is reported to be due selection pressures against deletions and this may be due to the deleterious effect of haploinsufficiency of dosage sensitive genes (1, 11, 31).

## 2.9 Gene Content of Benign CNV

Benign CNV overlap RefSeq and OMIM genes (2, 7, 11, 26, 29, 31). Gene function associated with cell adhesion, transduction, sensory perception and immunological response area common finding of these reports (1, 7, 11, 26, 29). CNV associated with segmental duplication were enriched in immunological and sensory genes, while genes with roles in cell adhesion are not dependent on genomic architecture (7, 11). Of note is the absence of dosage sensitive genes and genes with roles in signalling, proliferation and kinase categories in benign CNV (2, 11).

## 2.10 Interpretation and Classification of CNV

The documented evidence of CNV has expanded rapidly in recent years. The early contribution of disease free control cohorts originated from research studies and evidence inferred with little genotype/phenotype correlation (6, 30). Sufficient evidence was established for the benefit of whole genome microarray analysis that lead to the recommendation of its use as the first tier investigation in the clinical diagnosis of neurodevelopmental and congenital anomalies, replacing conventional karyotyping (41, 80, 82). However, CNV discovery and documentation was still in its infancy, and evidenced based interpretation of CNV remained challenging (63).

Guidelines and workflows were developed to provide assistance for the interpretation and reporting of CNV in the clinical diagnostic setting (63, 80, 82). The American College of Medical Genetics (ACMG) published guidelines for the interpretation and reporting of CNV in clinical diagnostic services (63, 80). These and other guidelines provide a definition of categories of clinical significance: benign, uncertain, unknown and pathogenic (63, 80). Benign CNV also sometimes referred to as copy number polymorphism , are CNV that are well supported by published studies and registration in curated databases and occur commonly (>1%) in the general population (80). The “uncertain significance” category is applied to CNV where there is insufficient supporting evidence to categorize the CNV as either pathogenic or benign (80). The definition of “unknown” pertains to

CNV where there is an absence of documented evidence of genotype/phenotype correlation in either published studies or curated databases. This category would apply for the reporting of a novel CNV. Well established evidence of clinical association applies to the assignment of "pathogenic" clinical significance (80). Although the categories are well defined and have been universally adopted, the process of assigning the category based on evidence of gene content is not well defined and may result in inconsistency of reporting CNV among laboratories (63).

The interpretation of CNV in the clinical setting is challenging where there is no or limited evidence upon which to base a decision. The identification of novel and rare CNV and characterisation of known CNV in non-disease general population and pathogenic populations will contribute to refining an overall genotype/phenotype map of CNV. Early studies such as Shaikh et al. 2009, identified numerous novel and rare variants, accounting for 51.5% of the CNV detected in the cohort (39). A high proportion of novel CNV compared to that reported in the DGV was reported by Pinto et al. 2007. These early findings may be indicative of the discovery phase or may reflect improvements in platform resolution and algorithm design rather than representation of mutation rates.

Rare CNV (<0.05-0.5% of population) (HapMap3) may represent low frequency benign CNV or rare clinically significant variants (31), which must be differentiated in clinical diagnostics. CNV occurring in one individual (singleton) are enriched in CNV > 500 kb in disease free cohorts (31). Due to the low frequency of these CNV, broad studies of healthy control cohorts irrespective of cohort size but with well documented phenotype criteria are required in a continued effort to document

rarely occurring CNV to make clear association with a benign outcome (24, 29-31, 63).

## 2.11 Population Diversity of CNV

Population studies have been facilitated by technological advances and accessibility to high resolution microarray platforms. Global variation of the prevalence, gene content, burden and mutational mechanisms of CNV has been reported (1, 2, 11, 25, 26, 37). The frequency and gene content of specific CNV have been shown to vary among populations suggesting adaptation to environment pressures and in some examples, influencing risk of disease (1, 29). Further studies of diverse populations will expand the knowledge of structural variation and ascertain influences of selective pressures and mutational mechanisms in generating genetic diversity.

## 2.12 Mechanisms of Derivation of CNV

In the early study by Sebat et al. 2004, the authors observed a non-random distribution of CNV and prevalence in pericentromeric and subtelomeric regions. The authors also observed an occurrence of CNV within or adjacent to recurring pathogenic rearrangements. They then demonstrated an enrichment of segmental duplication (SD) within a third of their reported CNV (9). This finding is supported by the studies of Shaw et al. 2004 that illustrated the correlation of segmental duplication and microdeletion/microduplication syndromes. The authors

postulated that unstable regions within the genome may contribute to the mechanism of CNV derivation (84).

Subsequent studies expanded the number of individuals investigated and the number of CNV identified. The non-random distribution of CNV and correlation with genomic features such as segmental duplication and transposable elements (7, 12, 30) was a consistent finding in all studies. These regions have high sequence homology and may be associated with secondary DNA structures predisposing the DNA segment to breakage and subsequent repair by non-allelic homologous recombination (NAHR). However these and subsequent studies also concluded that breakpoints near segmental duplication and transposable elements may account for the formation of some but not all CNV (7, 27, 30, 45, 85).

Mechanisms described in genomic rearrangements include non homologous repair of double and single stranded DNA breaks and replication/repair mechanisms (2, 23, 46, 47, 85-87). In a review by Gu et a. 2008, they proposed that these mechanisms may explain non-recurrent chromosomal rearrangements, including CNV, and are believed to occur in the germ line but also explain somatic rearrangements in cancer (85). To appreciate the contribution of mechanism to CNV properties, prevalence and distribution, a review of the mechanisms is provided.

### 2.12.1 Genome instability - the catalyst for CNV formation

Genomic instability underlying the formation of CNV is underpinned by the DNA sequence and its conformation. Repetitive DNA with particular sequence motifs

can form non-B DNA structures including hairpin, cruciform, G quadruplexes, H-DNA and left-handed Z DNA and are purported to provide the substrate for the generation of single strand (ssDNA) and double stranded breaks (DSB) (86-89) Table 4.

**Table 4**. DNA conformation is consistent with the underlying sequence motif

| Non B structure | Sequence Motif |
|---|---|
| Cruciform and hairpin structures | Inverted or direct repeats |
| Left handed Z | Alternating pyrimidine/purine bases (e.g. $GT_n$) |
| G Quadruplexes | 4 array of 2-4 guanine bases (e.g.TTAGGG) |
| Triplex and H DNA | Polypurine/polypyrimidine mirror repeat |

Non-B structures and the repeat sequence motifs accountable for their formation are abundant in the genome. These DNA sequences are conserved in evolution and the structures formed have roles in the formation of non recurrent CNV, constitutional translocations (49, 88), cancer mutations (49, 89, 90), fragile sites (89, 90), triplet repeat expansion (89) and immune system antibody formation (49) inferring a significant role in genetic diversity.

The presence of non-B DNA sequence motifs results in a transition of the right handed B-form helical duplex DNA conformation (canonical) as described by Watson and Crick, to left handed non-canonical structures. In the human genome, DNA is predominantly in canonical conformation, however can transition to the non-canonical form leading to genomic instability (88, 91, 92).

The relationship of non-B DNA structures and genomic instability is complex with several varied mechanisms proposed and this being dependent on the type of structure, DNA sequence and cellular response mechanism. "DNA structure induced" instability is replication independent and may apply to H DNA and Z DNA with distortion of the double helix or at junctions of hairpins and cruciform structures. In this model the alteration to the DNA configuration by non-canonical configuration may be perceived as DNA damage initiating the repair process (49, 88, 89). Repair proteins are recruited and cleavage initiated. Wang and Vasques 2009 postulated that the cell cycle is slowed during the repair process and mismatch repair (MMR) or nucleotide excision (NER) proteins are recruited resulting in cleavage in small loops or cleavage and excision of large loops with the non-loop strand acting as a template during synthesis (89). Alternatively the presence of these structures is associated with high energy torsional stress due to negative supercoiling (88). This may lead to weakened chemical stability such as oxidative damage at or near the non-B structure resulting in DNA breakage and subsequent repair (88).

"Replication dependent" genome instability can occur when the stable DNA duplex unwinds from the nucleosome and separates to single strand DNA in preparation for replication. Non-B DNA sequences within the single strand DNA has the propensity to form non-B structures. Schwabb et al. 2013 proposed that secondary structures form obstacles in the replication process in regions where there is reduced fork velocity (93). Non-B structures may form on the lagging strand and culminate in fork stalling and collapse leading to double strand break (49, 89, 94). Likewise Chen et al. 2015 reported an enrichment of CNV in genomic regions that

reflect slowed replication (95). Chromosome breakage "hotspots" have also been associated with non-palindromic formation of non-B structures (85). Telomeric regions reportedly CNV enriched (7, 78, 79) are hotspots for the formation of G quadruplexes (49, 85).

DNA damage and repair are active processes of the cell cycle. Events of DNA damage can result in double or single DNA strand breaks. Exogenous or endogenous stimuli recruit DNA repair proteins and enzymes. Cleavage of DNA may occur within, or at a distance from, non-B DNA sequences (89). The resulting ssDNA can form non-B DNA structures such as hairpins that may assist in the repair process by joining the ends of the complementary strand (89). Subsequent completion of replication and DNA synthesis will result in amplification, deletion or translocation. "Repair induced" instability is a model that applies to some forms of non-B DNA, but does not apply for conformations that involve torsional stress from negative supercoiling such as G quadruplexes (89). Arlt et al. 2009 demonstrated by using aphidicholine on fibroblast cells that CNV formation can be induced in mitosis by replication stress. However the authors reported no evidence of non-B DNA sequence motifs at breakpoints or within 30 base pairs (96). This may represent an under-estimation of the participation of non-B structures in induced CNV derivation, since its role in the repair process may involve DSB and non-B DNA several kilobases (kb) away from the CNV breakpoints (90).

The sequences predisposed to non canonical DNA conformation are present throughout the genome and not all form non-B structures or lead to DNA instability. This suggests that a regulation or control mechanism is involved to

resolve non-B structures without interference of the replication process (94). DNA helicases such as FANCJ and RecQ have a role in unwinding duplex and non canonical DNA structures, assisting in replication fork progression (49, 92, 94). RecQ proteins have a significant role in genome maintenance, being recruited at sites of double and single strand breaks to repair damaged DNA, DNA resection and re-anneal double strand breaks. They are recruited to remove endogenous damage including non-B DNA structures that has caused distortion of DNA and act to preserve replication fork integrity (92). The roles of these enzymes in the genome is varied (91, 92) and together with the presence of non-B DNA sequences may account for CNV formation and distribution.

## 2.12.2 CNV formation

There are two pathways for the mechanism of CNV derivation: homology directed repair (HR) that includes mechanisms such as NAHR (48) and non homologous (NH) including non homologous end joining (NHEJ), microhomology mediated end joining (MMEJ) and microhomology mediated break induced repair (MMBIR) (27, 46-48, 85) (Table 5, Figure 4). Recombination by NAHR accounts for the formation of recurrent CNV, whereas mechanisms for example NHEJ, MMEJ and replication errors such as fork stalling and template switching (FoSTeS) have been shown to contribute to non-recurrent CNV (27, 85, 96).

**Table 5.** Glossary of terms for mechanisms of derivation of CNV

| Acronym | |
|---|---|
| NAHR | Non Allelic Homologous Recombination |

| | |
|---|---|
| BIR | Break Induced Recombination |
| NHEJ | Non Homologous End Joining |
| MMEJ | Microhomology Mediated End Joining |
| FoSTeS | Fork Stalling and Template Switch |
| MMBIR | Microhomology Mediated Break Induced Recombination |

### 2.12.2.1 Homology directed repair mechanisms

Segmental duplication and short and long interspersed nuclear elements (SINES and LINES) are DNA repeat sequences of extensive homology measuring over 1kb in length. These repeat sequences provide the architectural structure to mediate homologous recombination (HR) during repair of DSB (47, 87, 97). Recurrent CNV have consistent start and end breakpoints and are frequently associated with low copy repeat (LCR) sequences (47, 87). In homologous recombination, strand exchange proteins catalyse the invasion of a homologous sequence at the 3' end of DSB with the complimentary strand in the DNA duplex (47) (Table 6). A segment of at least 200bp of homology is required and the presence of this at CNV breakpoints is a hallmark of HR mechanisms (47). Repair of DSB between homologous sequences of sister chromatids or homologues at the exact location will remain uneventful. Whereas deletion and duplication of the intervening DNA segment will occur as a consequence of repair between misaligned direct repeat sequences. This is not limited to the extensive homology of segmental duplication but SINE-*Alu* sequence-mediated NAHR has been reported previously to cause large events of duplication and deletion (98). These retrotransposons are believed to have a propensity for breakage and single strand formation which could initiate recombination events or template switch during replication (47, 86-88, 99-101). In a recent study Startek et al. 2014 investigated the role of LINE/LINE mediated

NAHR as a predicator of CNV derivation and identified LINE/LINE sequences at the breakpoint of CNV in both pathogenic and polymorphic CNV (97).

### 2.12.2.2 Non Homologous repair mechanisms

Non recurrent CNV have variable start and end positions and have microhomology or no homology at the breakpoint junction. Although these CNV are not mediated by blocks of LCR they may be located within regions that are in close proximity to LCR, direct and inverted repeat sequences (47).

Several DSB repair mechanisms are reported to involve small elements of microhomology at the sites of the breakpoint mediating the formation of the CNV (table 6) (47, 87). Non-homologous end joining (NHEJ) mechanisms involve small elements of microhomology at the sites of the DSB and may result in no change to the sequence or 1-4bp deletions or insertion at the breakpoint junction (46, 47, 87, 96). These mechanisms anneal two-ended DSB formed from exogenous factors or breaks in replication forks (47). Microhomology mediated end joining (MMEJ) is a double stranded repair mechanism between regions of <5bp microhomology resulting in deletion of the intervening sequence (47, 87, 96).

### 2.12.2.3 Replication mechanisms

FoSTeS, MMBIR and replication slippage are replication error/repair mechanisms (46, 47) (Table 7). FoSTeS involves the stalling of a fork in replication and invasion of a nearby fork by the lagging strand to a site of microhomology. This can occur

repeatedly and depending on the number of invasion events may result in a complex signature of inversion and translocation at the breakpoints (46, 47). Other replication errors such as MMBIR and replication slippage are microhomology mediated and occur where there is either fork collapse or single strand break due to reduced replication velocity (95) and may be dependent on the genomic architecture of the local sequence (85, 95).

**Table 6.** Overview of characteristics and mechanisms of derivation of CNV

| Model | Mechanism | Characteristics | Reference |
|---|---|---|---|
| **Homologous Recombination (HR)** | | | |
| NAHR | Invasion of homologous sequence by 3' end of ssDNA. CNV results if different location of homologous sequences | DSB required, > 200bp direct repeat sequence, > 97% homology | Hastings et al. 2009 , Vissers et al. 2009, Conrad et al. 2010 |
| BIR | NAHR using ectopic homology to repair breaks and continue replication | Somatic recombination, extensive homology, | Hastings et al.2009 |
| Single strand annealing | Double strand break with 5' resection on each strand till complimentary single strand reached. Annealing repairs the DSB | Deletion of intervening sequence and one set of direct repeats e.g. Alu. Small CNV only < 1kb | Hastings et al. 2009 |
| **Non Homologous (NH)** | **Non replicative** | | |
| NHEJ | Repairs DSB ends exactly or 1-4bp deletions | DSB required, insertion of non templated sequence, no homology or 1-4bp microhomology | Conrad et al 2010, Vissers et al 2009 |
| MMEJ | Intervening sequence between microhomologies deleted | Microhomology >5bp (5-25bp), deletion rather than insertion | Conrad et al 2010 |

**Table 7.** Overview of replicative mechanisms of derivation of CNV

| **Replicative** | | | |
|---|---|---|---|
| FoSTeS | Interruption of fork progression can lead to the 3' end invading a single strand template in another fork | 4-15bp microhomology at breakpoint junctions, associated with fork stalling, presence of nonB structures , single strand break. | Hastings et al. 2009 |
| MMBIR | Collapse of replication fork , the 3' tail anneals with a homologous ssDNA sequence in another fork. Replication continues, multiple events of template switching | Insertion of short sequences from within the genome or the same chromosome. Minimal microhomology required, may result in complex changes. LCSH may result adjacent to repair . Single strand break | Hastings et al. 2009b, Conrad et al. 2010, Vissers et al. 2009 |
| Replication slippage | Deletion or duplication of a sequence between short lengths of homology, when these occur within a single strand sequence i.e Okazaki fragment (1-2kb) | Inverted repeats , small CNV only 1-2kb | Hastings et al. 2009b |

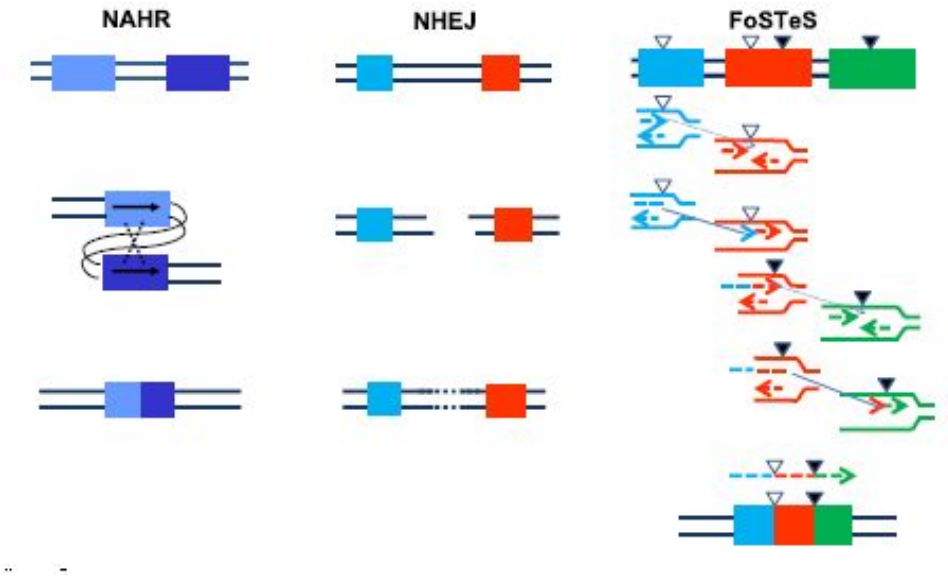



**Figure 4**. Illustration of the mechanisms that contribute to CNV formation (Gu et al. 2008)(85).

## 2.12.3 Contribution of mechanisms to CNV formation

Several studies have provided evidence for the contribution of homologous, non-homologous and replication repair mechanisms to the formation of CNV (45-48, 85). However, few of the studies prior to 2010 investigated CNV breakpoints at the base pair level which is necessary to determine the contribution of these mechanisms to CNV formation (48).

NAHR is the mechanism described for recurrent CNV with consistent breakpoints. Dependent on genomic architecture, these are non-random in distribution and are formed by the misalignment of sequences of high homology resulting in deletion and duplication in the successive cell division (85). The contribution of this mechanism to the prevalence of CNV is influenced by the requirement for and proximity of blocks of homologous sequences (7, 85). Prior to the development of high resolution whole genome microarray, NAHR was proposed as the mechanism

for CNV formation (31, 37). Perry et al. 2008 proposed that only a minority of the CNV reported by Redon et al. in 2006 overlap segmental duplication and further decreases with decreasing size of CNV (30). This is further supported in the study of mutational mechanisms in a HapMap cohort by Conrad et al. 2010 in which the authors demonstrated that NAHR accounts for the derivation of large CNV (48).

Non-homology mediated and replication mechanisms are postulated to account for non-recurrent CNV formation (23, 48, 85). In an early study by Cooper et al. 2007, the authors demonstrated the association of segmental duplication with NAHR mechanisms and CNV formation on chromosome 16 but reported that low frequency CNV are not associated with segmental duplication (7). Kaminsky et al. 2011 compared the frequency of recurrent and non-recurrent CNV of a large database compiled by the International Standards for Cytogenomic Arrays consortium and reported that up to 76% of CNV are non-recurrent (24).

Several mechanisms of non-homology directed and replication repair have been implicated in the derivation of CNV (46, 47, 87, 89). Arlt et al. 2009 induced CNV formation using aphidicoline in cell cultures found that replication stress induced CNV are derived from non homologous end joining mechanisms and alternate end joining such as microhomology mediated end joining or replication errors (96). Recently Chen et al. 2015 described influences of fork progression in replication and potential for fork stalling mechanisms (95).

The non-homologous mechanism may be inferred from the sequence signature surrounding CNV breakpoints (46, 47). It has been proposed that local DNA structure rather than sequence per se may underly the initiation of genomic instability leading to repair and coinciding CNV formation (48, 49, 85-89).

However the sequence surrounding these structures does not facilitate prediction of predisposition to CNV formation, akin to the NAHR mechanism driven by flanking LCR (48). Other factors such as genome maintenance regulators (49, 92, 94), and endogenous and exogenous stimuli may influence genomic stability and CNV formation (86).

More studies of the breakpoints of different sizes of CNV at base pair resolution are required to ascertain the contribution of mechanisms to formation, prevalence, distribution and properties of CNV in the human genome.

## 2.13 Long Contiguous Stretches of Homozygosity (LCSH)

SNP microarray platforms detect copy number change and the genotyping data (B allele frequency or B allele difference) provides evidence of sequences of homozygous alleles (22). The presence of segments of uninterrupted sequences of homozygous allelic markers in the human genome is well established (102), but the advent of high resolution SNP microarray platforms has recently provided insight to the extent in the outbred and "healthy" population (17, 19, 103, 104). These segments are ubiquitous and non-randomly distributed in the genome (15, 17-20, 105, 106). Differences in the length of long contiguous stretches of homozygosity (LCSH) and number of events per individual have been demonstrated among populations and this being indicative of ancestral history including geographical differences, migration patterns, bottlenecks, founder effects and cultural beliefs (15, 18-20, 105).

Homozygosity mapping and disease association studies have provided evidence on the clinical implication of segments of homozygous markers (103, 104, 107-109). The detection of segments of autozygosity is not diagnostic of a clinical outcome, and would usually be of benign significance (22). However, autosomal recessive mutations may be "unmasked" in regions of homozygosity and the chance inheritance of monogenic disorders increases with the degree of parental relatedness (22, 103, 109). An increased risk of common complex disorders such as hypertension, hypercholesterolaemia, and diabetes is reported in association with consanguineous kindred (20, 103, 108).

The purpose of this section is to review the incidence and properties of autozygosity in the human genome, with particular focus on the differences among populations and the potential mechanisms of variability. These properties, together with the clinical significance of autozygosity are considered, to illustrate the impact and potential role of general population studies in the interpretation of LCSH events in clinical diagnostics.

## 2.13.1 Population diversity of patterns of LCSH

Long contiguous stretches of homozygosity (LCSH) are indicative of identical bi-allelic markers that are inherited from a common ancestor (autozygosity). One mechanism of autozygosity is consanguinity. Assuming larger regions of shared haplotypes are present between closely related individuals (103, 107, 109), cross-over during meiotic recombination reduces the length of LCSH over successive

generations resulting in shorter segments in outbred populations (15, 18-20, 103, 105). Recent studies have illustrated LCSH events to be globally widespread and the presence of shared haplotypes in outbred populations exemplifies continental distribution patterns (15, 20, 105).

There is variation in the LCSH length, number of events and overall contribution of homozygosity to an individual and population genome (15, 18-20, 105). A higher prevalence of short LCSH events and increased overall genomic contribution is apparent in association with population dispersal from East Africa (15, 20, 105). Gibson et al. 2006 investigated the genotypes from the HapMap phase 1 data by linkage disequilibrium and reported the average number of events >1 Mb per individual as 4.4 (YRI), 5.3 (CHB), 8.3 (CEU) and 8.4 (JPT) (15). A review of the Human Genome Diversity Project by Kirin et al. 2010 revealed enrichment of long LCSH in South and Western Asian populations, enrichment of short LCSH events in the Oceania population and both short and long LCSH events in the Native American population consistent with dispersal out of Africa and past or recent practices of inbreeding (20, 105). Higher levels of long LCSH events and overall contribution to the total genome is apparent in populations from the Middle East, Central and South Asia, Oceania and the Americas suggestive of higher levels of recent parental relatedness (20, 105). These findings are supported in the comprehensive study of the SNP data for 64 populations of the Human Genome Diversity Panel- CEPH Cell line Panel (HGDP-CEPH) and Phase 3 of the International Haplotype Map (HapMap3) by Pemberton et al. 2012. The authors provided evidence of consistency of cumulative length of LCSH among continental groups and demonstrated an increase in small to intermediate length of LCSH

events (overall less than several Mb) consistent with migration patterns of population groups out of Africa. They explained that this is consistent with the reduction in population size with each successive phase of migration, increasing the chance of shared haplotypes (105).

In a review of European populations and sub-populations McQuillan et al. 2008 report LCSH events shorter than 1.5 Mb are detected in all individuals and events up to 4 Mb are not uncommon (19). By comparing the LCSH properties from populations of geographically isolated regions or rural and urban Scotland they suggest a correlation of genomic contribution of autozygosity with population size and prolonged geographical isolation (19). Nothnagel et al. 2010 analysed the genotyping data of 23 European subpopulations for the number and cumulative length of LCSH events and demonstrated a South to North Europe gradient consistent with ancestral migration patterns. Individuals representing Romanian populations recorded a median number of events > 1Mb per individual of 32.55, whereas Finland and Norway reported 48.04 and 42.16 respectively (18).

The finding of autozygosity in the clinical diagnostic setting must be considered with caution given the differences among populations and this is indicative of population history or cultural practices. The inbreeding coefficient (proportion of the contribution of homozygosity to the genome), varies from 0.21% in European populations (LCSH events > 5Mb) to 3.93% in the Americas (20). There are numerous mechanisms that can influence the level of autozygosity in the genome. One of which is the founder effect where the ancestry of a population can be traced to a small number of founders as a consequence of a bottleneck, reduction in population, or establishment of a new population from a limited pool rather than

by consanguinity (103, 107). Li et al. 2006 refer to an increased level of autozygosity with LCSH events ranging in length of 2.46Mb to 19.43Mb, in Taiwanese aboriginal tribes where consanguinity is outlawed but tribal isolation leads to common ancestors (17). The Finnish population is well known as a founder population and the inbreeding coefficient is significantly higher than that of the outbred population (103, 107). The median number of events per individual of 48.04 +/- 7.34 is notably higher when compared to the European average (23 subpopulations) of 38.74 +/- 6.6 (18).

Inter-individual variation of the contribution of long LCSH events may also represent consanguinity. Endogamous populations, where cultural practices encourage consanguinity, have an elevated population average for LCSH length and proportion of homozygosity in the genome (19, 103, 105). This is evident in Middle East, Central and South Asia, Oceania and populations of the Americas (20, 103, 105). Incidental findings of elevated levels of autozygosity in 2 individuals in the JPT HapMap Phase 1 data (15) and events of consanguinity reported by McGuillan et al. 2008 in a low proportion of Scottish samples have been reported (19). Other examples include Li et al. 2006 where the authors reviewed events of homozygosity in unrelated Han Chinese population and identified LCSH events consistent with a degree of consanguinity in 6% of the study cohort. LCSH lengths ranged from 2.94Mb to 26.27 Mb (17). Similarly 1.6% of the European population recorded the proportion of autozygosity in the autosomal genome of 3.3% (18) compared to the population average of 0.21% (20). Firstly these findings suggest that incidental individual events of consanguinity are not uncommon in outbred populations and secondly, elevated levels of autozygosity may be associated with a

benign outcome. However the lack of phenotypic data available for some study individuals, for example the HapMap, means that this is an assumption rather than confirmed (15).

### 2.13.2 Non-Random chromosomal distribution of common LCSH

A consistent finding of all studies is the non-random chromosomal distribution of LCSH events (15, 17-20, 105, 106). Nothnagel et al. 2010 investigated events of homozygosity in 23 European sub-populations and identified chromosomal regions where LCSH events are present at high frequency (18). LCSH events occurring in more than one individual or up to 50% of study cohorts have been reported in numerous population studies (15, 17-19, 105). These regions are interstitial and the chromosomal distribution is similar among individuals (17, 18). The centromeric region is noted to be associated with an increased incidence of LCSH and may reflect the low recombination rate in the chromosomal region (106). However this is not supported by all studies due to the low SNP density, poor SNP integrity and reduced marker coverage within the region (15, 19, 20).

Common LCSH events are apparent between published population cohorts (15, 17-20, 105). An LCSH event at chromosome 1p33 and 1q24.3 reported in McQuillan et al. 2008 in the Scottish population was also observed in the European population studied by Nothnagel et al. 2010 (18, 19). Li et al. 2006 identified 17 common LCSH events in the Han Chinese population that overlapped events reported by Gibson et al. 2006 in the HapMap study (15, 17). It is noted however that common LCSH

events in outbred populations occur in co-located chromosomal regions whereas LCSH events reflecting parental relatedness tend to have a random chromosomal distribution (19).

### 2.13.2.1 Mechanisms of non-random distribution

There are numerous mechanisms that influence the distribution patterns of LCSH. Gibson et al. 2006, performed linkage disequilibrium studies on HapMap Phase 1 SNP genotype data. Linkage disequilibrium (LD) is the inheritance of alleles together, which is not explicable by random segregation (15). The presence of linkage disequilibrium is reduced by recombination events. Gibson et al. 2006 demonstrated an association of LCSH events with extensive linkage disequilibrium and low recombination rates and proposed a highly similar pattern of linkage disequilibrium among populations (15, 18, 110). The presence of long tracts of homozygosity co-localised in chromosomal regions is indicative of segments of low recombination flanked by recombination "hotspots" and exemplifies current and ancestral recombination patterns (15, 19, 105).

There are regions of high and low incidence of LCSH throughout the genome (17, 105) and not all are explained by linkage disequilibrium and recombination events (17, 18). Common LCSH represent regions of the genome where homozygosity is tolerated. This may be due to the absence of deleterious genes or a region that has survived selection pressures (105). Conversely regions of high recombination rates and low incidence of LCSH occur near deleterious genes, where

heterozygosity has the selective advantage (105). In a comprehensive study by Pemberton et al. 2012, they demonstrated common LCSH in regions of known positive selective pressures and correlated these regions with geographical backgrounds. Selection pressure alters haplotype diversity impacting the potential to form extended homozygosity (105).

Regions of extended haplotypes that occur at high frequency in a population may also be inherited by chance alone accounting for common LCSH events (19, 20, 106). Linkage disequilibrium reduces haplotype diversity increasing the probability of inheriting the same haplotype from both parents (105). This is particularly relevant to small LCSH events (<2Mb) which is ubiquitous in all populations (20, 105, 106).

### 2.13.3 Clinical implication of LCSH

The association of extensive autozygosity with disease is well established (22, 103-105, 107-109, 111-115). An increased risk for the inheritance of autosomal recessive disease is based on the assumption that both parents carry the gene mutation (103, 104, 107, 113). Recessive alleles inherited in this way are generally of low population frequency and the probability of inheriting the mutated gene is proportional to the level of autozygosity (103, 104). As such there is an increased risk for the inheritance of autosomal recessive disease in conjunction with the extensive autozygosity associated with consanguinity.

The progress of disease gene mapping is restricted in early investigations due to the technological limitation of methods based on microsatellites and recruitment of disease patients from consanguineous kindred (103, 104). The investigation of

the autosomal recessive disease for candidate genes has been largely facilitated by founder populations such as the Finnish population where up to 40 autosomal recessive diseases have been characterised (107, 113). The implementation of SNP microarray has identified LCSH in the outbred population and the improved resolution has enabled the detection of small and intermediate lengths of LCSH. Hildbrandt et al. 2009 and Sund et al. 2013 reported no limitation of LCSH length that harbour disease related genes (104, 109).

It is proposed that the same principle may apply to an increased risk of common complex disease such as hypertension (112), increased LDL cholesterol, and diabetes (20, 103, 108, 114, 115). The prevalence of common and late onset adult disease is elevated in consanguineous populations (108, 112, 114-116). Homozygosity for weak recessive alleles that are under less negative selection pressure may have a role in adult onset disease (103). In contrast to monogenic autosomal recessive disease, common complex diseases may be mediated by many common recessive variants located throughout the genome that have individually small but cumulative effect, leading to disease pathogenesis (20, 103, 112). An alternative model considers that these common diseases are due to rare recessive variants but abundant in the genome and the disease risk is increased in the presence of autozygosity (19, 103).

Homozygosity mapping for disease gene discovery is based on the premise that the disease causing gene is located within a region of homozygous alleles (103, 104). Once a potential region is identified, targeted sequencing of the candidate gene can confirm the allelic variant (103, 113). Correlation of genotype and phenotype in a number of individuals leads to the identification of disease causing genes.

Until recently homozygosity mapping has been performed where there is evidence of elevated levels of homozygosity as observed in consanguinity or founder populations. The advent of SNP microarray technology and its application to the detection of homozygosity has expanded the potential to include the affected individuals in the outbred population (103-105). Homozygosity mapping of outbred populations will assist in disease gene discovery in numerous ways. Firstly, the number of affected individuals is increased impacting both discovery and confirmation (104). Secondly, the LCSH length in outbred populations is generally shorter (<4mb) compared to that in consanguineous and founder populations allowing a more targeted investigation of genes within the candidate region (104, 105) and finally the exclusion of regions with common and population specific LCSH events will prioritise the investigation of regions of low LCSH frequency for rare disease variants (105).

There are few studies defining LCSH properties in healthy general populations. Population variation of LCSH properties is clearly demonstrated in HapMap studies and reflects geographical, cultural and ancestral backgrounds of the population. Due to this variability further studies of a range of general populations and in particular definition in the local population is required. The description of LCSH events in general populations will assist in the characterisation of common and population specific homozygous regions for which comparison can be made in the clinical setting. It has the potential to accelerate discovery of rare recessive disease genes and biological causes of common complex diseases.

## 2.14 Database Repositories and Web-based Resources

Evidenced based assessment of CNV in clinical diagnostics is reliant on CNV registered in curated databases and other related web-based resources. These are available for non-disease and clinical cohorts (8).

## 2.14.1 Database of Genomic Variants (DGV)

The Database of Genomic Variants was established as a repository for structural variants reported in healthy unaffected individuals and was established in response to initial studies in 2004 (9, 10, 117). Registered entries grew rapidly. However the early stages of curation to the DGV was imperfect, with the registration of unpublished CNV data of undefined phenotypic background, assignment according to the reference assembly at the time of discovery and no standardisation of reporting (6, 23, 117). Many of these CNV were generated from studies of HapMap individuals where there is no phenotypic information (30) and the SNP data was originally generated for disease association (8).

In an effort to amend this limitation the curation process was modified in 2008. All entries from DGV were archived as DGVa (CNV) or dbVar for sequence variants. Submission of new entries for inclusion in the DGV is performed by direct upload to DGVa. Each CNV entry is reviewed and those that pass strict quality criteria and filters are assigned a unique accession number; esv for DGV variants and nsv for dbVar variants and included for display in the DGV or as DGV structural variants tracks in UCSC Genome Browser (Figure 5 and 6)(117, 118).

In accordance with ongoing appraisal of the curation process, some of the early CNV have been removed and this will continue as variants are re-defined with CNV datasets from subsequent studies (117). The entries in DGV represent HapMap and unique cohorts representing populations from diverse geographical backgrounds (117). The intention of the curation process is to provide a robust and high quality database of structural variant in healthy unaffected individuals for use in clinical diagnostics and research settings (117).

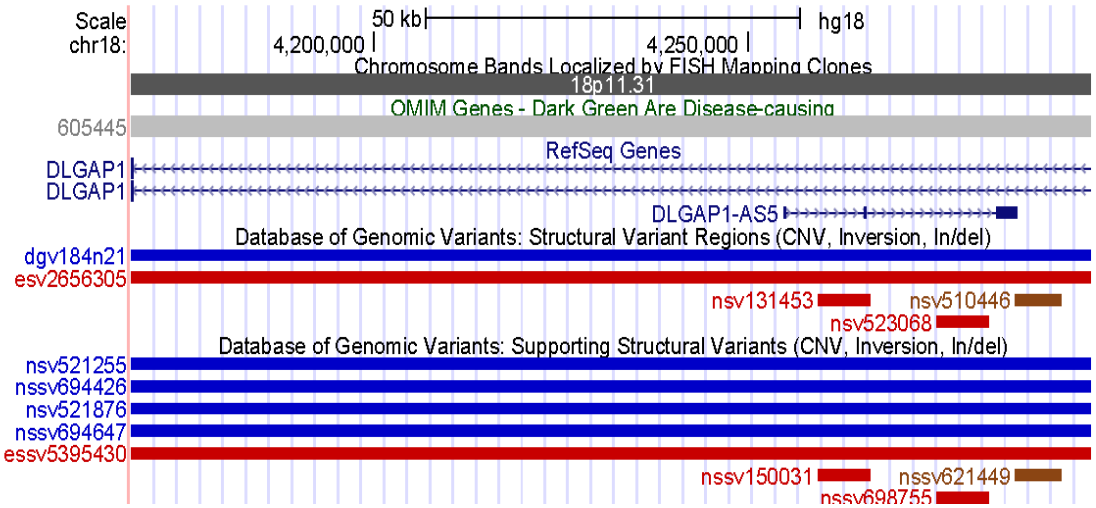



**Figure 5.** DGV entries visualised as tracks in UCSC Genome Browser for CNV at 18p11.31 (117-119).

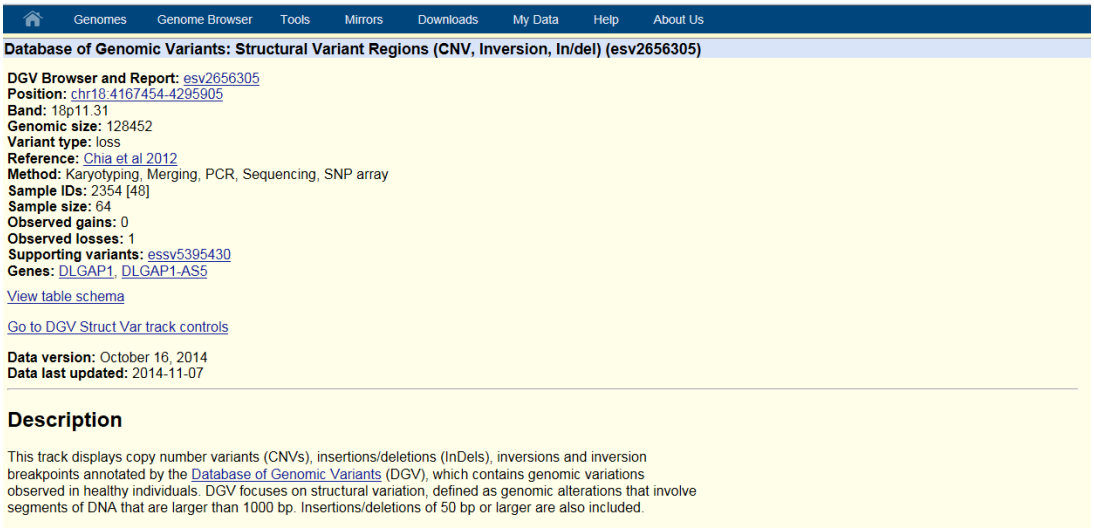



**Figure 6**. Selection of variant esv2656305 provides curation information including the link to the publication, incidence, length, copy number state and gene content (117, 118).

## 2.14.2 International Standards for Cytogenomic Arrays (ISCA)

The International Standards for Cytogenomic Arrays (ISCA), more recently known as Clinical Genome Resource (ClinGen), represents a consortium of diagnostic laboratories and medical and research professionals formed with the objective to provide a "standardized interpretation of array data" and produce a "CNV Atlas" of genomic variation (63, 120). The resulting database provides a comprehensive correlation of phenotype with genotype of the region of the genome, and is applicable to evidence based approach to CNV interpretation (24, 63, 120). This data will also contribute to the development of a database which has practical application such as the inclusion in customised whole genome microarray

platforms. It will assist in the development of a whole genome map of CNV and dosage sensitive genes (24, 63, 120). The ISCA tracks can be selected in the UCSC Genome Browser (http://genome.ucsc.edu) (118) and indicate the assigned clinical significance of a region based current evidence.

### 2.14.3 Other Resources

Shaikh et al. 2009 documented CNV in 2026 "disease free" individuals generated using Illumina Infinium II Hap550 bead chip. These CNV were published on a publicly available database in a Web-based format (http://cnv.chop) and the CNV can be downloaded for investigation. The CNV outputs can also be visualised in a track on the UCSC Genome Browser (http://genome.ucsc.edu) (39). The information attained includes copy number state, incidence, chromosomal location, CNV length and number of SNP markers. The authors assessed the clinical utility of the database by examination of two previously reported CNV of pathogenic significance. The findings re-affirmed a pathogenic significance in one CNV (15q13.3/*CHRNA7*), but refuted the second CNV (15q11.2) inferring caution with regard to assignment of clinical significance (39).

Resources for the investigation of clinical association are available from DECIPHER and ClinVar. DECIPHER (Database of Chromosomal Imbalance and Phenotype using Ensembl Resources https://decipher.sanger.ac.uk) was established to record genomic imbalances in patients with a clinical phenotype, to assist in determination of genetic aetiology (23, 121). ClinVar provides

genotype/phenotype correlations for structural allelic variants that have been determined at the sequence level and entries are sourced from dbVar (122).

## 2.15 Conclusion

Over the past decade microarray technology has contributed significantly to the appreciation of the scope of human genetic variation, consideration of the mechanisms that contribute to the formation of these variants and manner in which they affect human development. The rapid progression of this technology and implementation as a diagnostic tool has contributed to the knowledge of disease associations and defined gene involvement not previously detected by conventional methods. The aetiology of common complex and late onset diseases, believed to be associated with these forms of genetic variation, is yet to be fully characterised.

HapMap studies provided the initial opportunity to compare CNV and LCSH properties among populations and consider evidence of ancestral structural and genetic variation. As illustrated here, several studies have investigated these genetic variants in transformed cell lines from individuals recruited from the HapMap project. However there are few studies of genomic DNA from general populations and of these the phenotypic information is limited or incomplete. Definition of the incidence, length and gene content of CNV and LCSH from general populations allows more informed evaluation in the clinical setting.

Another limitation of the studies prior to 2010 is the difference in technical and experimental design and study objectives. Subsequently there is evidence of discrepancy of reported CNV properties and incidence among studies, even when analysing the same data or sample source. Pang et al. 2010 compared SNP and aCGH platforms and proposed platform variation as one contributing factor (28). Several studies after 2010 noted the effect of CNV detection programs and the statistical algorithm employed on CNV detection and characterisation of breakpoints. Another contributing factor is the progressive improvements in resolution resulting in the refinement of breakpoints and detection of smaller CNV. These factors highlight the need for continued and robust review of the information from published studies and inclusion of a broad range of both case and control cohorts in the development of a map of human genetic variation.

## 2. 16 References

1. Conrad DF, Hurles ME. The population genetics of structural variation. Nat Genet. 2007;39(7 Suppl):S30-6.
2. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, et al. Origins and functional impact of copy number variation in the human genome. Nature. 2010 ;464(7289):704-12.
3. Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467(7311):52-8.
4. Craddock N, Hurles ME, Cardin N, Pearson RD, Plagnol V, Robson S, et al. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature. 2010;464(7289):713-20.
5. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012 ;491(7422):56-65.
6. Scherer SW, Lee C, Birney E, Altshuler DM, Eichler EE, Carter NP, et al. Challenges and standards in integrating surveys of structural variation. Nat Genet. 2007;39(7 Suppl):S7-15.

7. Cooper GM, Nickerson DA, Eichler EE. Mutational and selective effects on copy-number variants in the human genome. Nat Genet. 2007;39(7 Suppl):S22-9.
8. Feuk L, Marshall CR, Wintle RF, Scherer SW. Structural variants: changing the landscape of chromosomes and design of disease studies. Hum Mol Genet. 2006;15 Spec No 1:R57-66.
9. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, et al. Large-scale copy number polymorphism in the human genome. Science. 2004;305(5683):525-8.
10. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, et al. Detection of large-scale variation in the human genome. Nat Genet. 2004;36(9):949-51.
11. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number in the human genome. Nature. 2006;444(7118):444-54.
12. Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, et al. Segmental duplications and copy-number variation in the human genome. Am J Hum Genet. 2005 ;77(1):78-88.
13. Ting JC, Ye Y, Thomas GH, Ruczinski I, Pevsner J. Analysis and visualization of chromosomal abnormalities in SNP data with SNPscan. BMC Bioinformatics. 2006;7:25.
14. Slater HR, Bailey DK, Ren H, Cao M, Bell K, Nasioulas S, et al. High-resolution identification of chromosomal abnormalities using oligonucleotide arrays containing 116,204 SNPs. Am J Hum Genet. 2005;77(5):709-26.
15. Gibson J, Morton NE, Collins A. Extended tracts of homozygosity in outbred human populations. Hum Mol Genet. 2006;15(5):789-95.
16. Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, Garcia F, et al. High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. Genome Res. 2006;16(9):1136-48.
17. Li LH, Ho SF, Chen CH, Wei CY, Wong WC, Li LY, et al. Long contiguous stretches of homozygosity in the human genome. Hum Mutat. 2006;27(11):1115-21.
18. Nothnagel M, Lu TT, Kayser M, Krawczak M. Genomic and geographic distribution of SNP-defined runs of homozygosity in Europeans. Hum Mol Genet. 2010;19(15):2927-35.
19. McQuillan R, Leutenegger AL, Abdel-Rahman R, Franklin CS, Pericic M, Barac-Lauc L, et al. Runs of homozygosity in European populations. Am J Hum Genet. 2008;83(3):359-72.
20. Kirin M, McQuillan R, Franklin CS, Campbell H, McKeigue PM, Wilson JF. Genomic runs of homozygosity record population history and consanguinity. PLoS One 2010; 5(11):e13996.
21. Bruno DL, Ganesamoorthy D, Schoumans J, Bankier A, Coman D, Delatycki M, et al. Detection of cryptic pathogenic copy number variations and constitutional loss of heterozygosity using high resolution SNP microarray analysis in 117 patients referred for cytogenetic analysis and impact on clinical practice. J Med Genet. 2009;46(2):123-31.
22. Kearney HM, Kearney JB, Conlin LK. Diagnostic implications of excessive homozygosity detected by SNP-based microarrays: consanguinity, uniparental disomy, and recessive single-gene mutations. Clin Lab Med. 2011;31(4):595-613, ix.
23. Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, et al. Copy number variation: new insights in genome diversity. Genome Res. 2006 Aug;16(8):949-61.
24. Kaminsky EB, Kaul V, Paschall J, Church DM, Bunke B, Kunig D, et al. An evidence-based approach to establish the functional and clinical significance of copy number variants in intellectual and developmental disabilities. Genet Med. 2011;13(9):777-84.
25. Matsuzaki H, Wang PH, Hu J, Rava R, Fu GK. High resolution discovery and confirmation of copy number variants in 90 Yoruba Nigerians. Genome Biol. 2009;10(11):R125.
26. Park H, Kim JI, Ju YS, Gokcumen O, Mills RE, Kim S, et al. Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. Nat Genet. 2010;42(5):400-5.

27. Perry GH, Ben-Dor A, Tsalenko A, Sampas N, Rodriguez-Revenga L, Tran CW, et al. The fine-scale and complex architecture of human copy-number variation. Am J Hum Genet. 2008;82(3):685-95.
28. Pang AW, MacDonald JR, Pinto D, Wei J, Rafiq MA, Conrad DF, et al. Towards a comprehensive structural variation map of an individual human genome. Genome Biol.2010;11(5):R52.
29. de Smith AJ, Walters RG, Froguel P, Blakemore AI. Human genes involved in copy number variation: mechanisms of origin, functional effects and implications for disease. Cytogenet Genome Res. 2008;123(1-4):17-26.
30. Pinto D, Marshall C, Feuk L, Scherer SW. Copy-number variation in control population cohorts. Hum Mol Genet. 2007;16 Spec No. 2:R168-73.
31. Itsara A, Cooper GM, Baker C, Girirajan S, Li J, Absher D, et al. Population analysis of large copy number variants and hotspots of human genetic disease. Am J Hum Genet. 2009 ;84(2):148-61.
32. Pinto D, Darvishi K, Shi X, Rajan D, Rigler D, Fitzgerald T, et al. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nat Biotechnol. 2011;29(6):512-20.
33. Henrichsen CN, Chaignat E, Reymond A. Copy number variants, diseases and gene expression. Hum Mol Genet. 2009;18(R1):R1-8.
34. Dellinger AE, Saw SM, Goh LK, Seielstad M, Young TL, Li YJ. Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. Nucleic Acids Res. 2010;38(9):e105.
35. Ritchie ME, Liu R, Carvalho BS, Irizarry RA. Comparing genotyping algorithms for Illumina's Infinium whole-genome SNP BeadChips. BMC Bioinformatics.2011;12:68.
36. Winchester L, Yau C, Ragoussis J. Comparing CNV detection methods for SNP arrays. Brief Funct Genomic Proteomic. 2009;8(5):353-66.
37. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, et al. Mapping and sequencing of structural variation from eight human genomes. Nature. 2008 ;453(7191):56-64.
38. Carter N. Methods and Strategies for analyzing copy number variation using DNA microarrays. Nature Genetics. 2007;39(Supplement):5.
39. Shaikh TH, Gai X, Perin JC, Glessner JT, Xie H, Murphy K, et al. High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications. Genome Res. 2009;19(9):1682-90.
40. de Smith AJ, Trewick AL, Blakemore AI. Implications of copy number variation in people with chromosomal abnormalities: potential for greater variation in copy number state may contribute to variability of phenotype. Hugo J. 2010;4(1-4):1-9.
41. Miller DT, Adam MP, Aradhya S, Biesecker LG, Brothman AR, Carter NP, et al. Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. Am J Hum Genet. 2010;86(5):749-64.
42. Kearney HM, South ST, Wolff DJ, Lamb A, Hamosh A, Rao KW. American College of Medical Genetics recommendations for the design and performance expectations for clinical genomic copy number microarrays intended for use in the postnatal setting for detection of constitutional abnormalities. Genet Med. 2011;13(7):676-9.
43. Sanders SJ, Ercan-Sencicek AG, Hus V, Luo R, Murtha MT, Moreno-De-Luca D, et al. Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. Neuron. 2011;70(5):863-85.
44. Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, et al. A copy number variation morbidity map of developmental delay. Nat Genet. 2011;43(9):838-46.

45. Qiao Y, Liu X, Harvard C, Nolin SL, Brown WT, Koochek M, et al. Large-scale copy number variants (CNVs): distribution in normal subjects and FISH/real-time qPCR analysis. BMC Genomics. 2007;8:167.
46. Hastings PJ, Ira G, Lupski JR. A microhomology-mediated break-induced replication model for the origin of human copy number variation. PLoS Genet. 2009;5(1):e1000327.
47. Hastings PJ, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. Nat Rev Genet. 2009;10(8):551-64.
48. Conrad DF, Bird C, Blackburne B, Lindsay S, Mamanova L, Lee C, et al. Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. Nat Genet. 2010;42(5):385-91.
49. Bacolla A, Cooper DN, Vasquez KM. DNA structure matters. Genome Med.2013 5(6):51.
50. Kleinjan DA, van Heyningen V. Long-range control of gene expression: emerging mechanisms and disruption in disease. Am J Hum Genet. 2005;76(1):8-32.
51. Huang N, Lee I, Marcotte EM, Hurles ME. Characterising and predicting haploinsufficiency in the human genome. PLoS Genet. 2010;6(10):e1001154.
52. Girirajan S, Eichler EE. Phenotypic variability and genetic susceptibility to genomic disorders. Hum Mol Genet. 2010;19(R2):R176-87.
53. Howarth KD, Pole JC, Beavis JC, Batty EM, Newman S, Bignell GR, et al. Large duplications at reciprocal translocation breakpoints that might be the counterpart of large deletions and could arise from stalled replication bubbles. Genome Res. 2011;21(4):525-34.
54. Jaju RJ, Fidler C, Haas OA, Strickson AJ, Watkins F, Clark K, et al. A novel gene, NSD1, is fused to NUP98 in the t(5;11)(q35;p15.5) in de novo childhood acute myeloid leukemia. Blood. 2001;98(4):1264-7.
55. Adams SA, Coppinger J, Saitta SC, Stroud T, Kandamurugu M, Fan Z, et al. Impact of genotype-first diagnosis: the detection of microdeletion and microduplication syndromes with cancer predisposition by aCGH. Genet Med. 2009;11(5):314-22.
56. Holt R, Sykes NH, Conceicao IC, Cazier JB, Anney RJ, Oliveira G, et al. CNVs leading to fusion transcripts in individuals with autism spectrum disorder. Eur J Hum Genet. 2012;20(11):1141-7.
57. Rodenas-Cuadrado P, Ho J, Vernes SC. Shining a light on CNTNAP2: complex functions to complex disorders. Eur J Hum Genet. 2014;22(2):171-8.
58. Kallioniemi OP, Kallioniemi A, Sudar D, Rutovitz D, Gray JW, Waldman F, et al. Comparative genomic hybridization: a rapid new method for detecting and mapping DNA amplification in tumors. Semin Cancer Biol. 1993;4(1):41-6.
59. Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, et al. Fine-scale structural variation of the human genome. Nat Genet. 2005;37(7):727-32.
60. Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK. A high-resolution survey of deletion polymorphism in the human genome. Nat Genet. 2006;38(1):75-81.
61. McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC, et al. Common deletion polymorphisms in the human genome. Nat Genet. 2006;38(1):86-92.
62. Komura D, Shen F, Ishikawa S, Fitch KR, Chen W, Zhang J, et al. Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. Genome Res. 2006;16(12):1575-84.
63. Riggs ER, Church DM, Hanson K, Horner VL, Kaminsky EB, Kuhn RM, et al. Towards an evidence-based process for the clinical interpretation of copy number variation. Clin Genet. 2012;81(5):403-12.
64. Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P, et al. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res. 2007;35(6):2013-25.

65. Karimpour-Fard A, Dumas L, Phang T, Sikela JM, Hunter LE. A survey of analysis software for array-comparative genomic hybridisation studies to detect copy number variation. Hum Genomics. 2010;4(6):421-7.
66. Vermeesch JR, Melotte C, Froyen G, Van Vooren S, Dutta B, Maas N, et al. Molecular karyotyping: array CGH quality criteria for constitutional genetic diagnosis. J Histochem Cytochem. 2005;53(3):413-22.
67. Venkatraman ES, Olshen AB. A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics. 2007;23(6):657-63.
68. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics. 2004;5(4):557-72.
69. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. 2007;17(11):1665-74.
70. Xu Y, Peng B, Fu Y, Amos CI. Genome-wide algorithm for detecting CNV associations with diseases. BMC Bioinformatics. 2011; 12:331.
71. Marioni JC, Thorne NP, Valsesia A, Fitzgerald T, Redon R, Fiegler H, et al. Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. Genome Biol. 2007;8(10):R228.
72. Diskin SJ, Li M, Hou C, Yang S, Glessner J, Hakonarson H, et al. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. Nucleic Acids Res. 2008;36(19):e126.
73. Barnes C, Plagnol V, Fitzgerald T, Redon R, Marchini J, Clayton D, et al. A robust statistical method for case-control association testing with copy number variation. Nat Genet. 2008;40(10):1245-52.
74. Wineinger NE, Tiwari HK. The impact of errors in copy number variation detection algorithms on association results. PLoS One.2012 7(4):e32396.
75. Zhang D, Qian Y, Akula N, Alliey-Rodriguez N, Tang J, Gershon ES, et al. Accuracy of CNV Detection from GWAS Data. PLoS One.2011 6(1):e14511.
76. Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, et al. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet. 2009;41(10):1061-7.
77. Kim JI, Ju YS, Park H, Kim S, Lee S, Yi JH, et al. A highly annotated whole-genome sequence of a Korean individual. Nature. 2009;460(7258):1011-5.
78. Knight SJ, Flint J. Perfect endings: a review of subtelomeric probes and their use in clinical diagnosis. J Med Genet. 2000;37(6):401-9.
79. Knight SJ, Lese CM, Precht KS, Kuc J, Ning Y, Lucas S, et al. An optimized set of human telomere clones for studying telomere integrity and architecture. Am J Hum Genet. 2000;67(2):320-32.
80. Kearney HM, Thorland EC, Brown KK, Quintero-Rivera F, South ST. American College of Medical Genetics standards and guidelines for interpretation and reporting of postnatal constitutional copy number variants. Genet Med. 2011;13(7):680-5.
81. Girirajan S, Rosenfeld JA, Coe BP, Parikh S, Friedman N, Goldstein A, et al. Phenotypic heterogeneity of genomic disorders and rare copy-number variants. N Engl J Med. 2012;367(14):1321-31.
82. Gijsbers AC, Lew JY, Bosch CA, Schuurs-Hoeijmakers JH, van Haeringen A, den Hollander NS, et al. A new diagnostic workflow for patients with mental retardation and/or multiple congenital abnormalities: test arrays first. Eur J Hum Genet. 2009;17(11):1394-402.
83. Ju YS, Hong D, Kim S, Park SS, Lee S, Park H, et al. Reference-unbiased copy number variant analysis using CGH microarrays. Nucleic Acids Res. 2010;38(20):e190.

84. Shaw CJ, Withers MA, Lupski JR. Uncommon deletions of the Smith-Magenis syndrome region can be recurrent when alternate low-copy repeats act as homologous recombination substrates. Am J Hum Genet. 2004;75(1):75-81.
85. Gu W, Zhang F, Lupski JR. Mechanisms for human genomic rearrangements. Pathogenetics. 2008;1(1):4.
86. Bacolla A, Jaworski A, Larson JE, Jakupciak JP, Chuzhanova N, Abeysinghe SS, et al. Breakpoints of gross deletions coincide with non-B DNA conformations. Proc Natl Acad Sci U S A. 2004;101(39):14162-7.
87. Vissers LE, Bhatt SS, Janssen IM, Xia Z, Lalani SR, Pfundt R, et al. Rare pathogenic microdeletions and tandem duplications are microhomology-mediated and stimulated by local genomic architecture. Hum Mol Genet. 2009;18(19):3579-93.
88. Bacolla A, Wells RD. Non-B DNA conformations, genomic rearrangements, and human disease. J Biol Chem. 2004;279(46):47411-4.
89. Wang G, Vasquez KM. Models for chromosomal replication-independent non-B DNA structure-induced genetic instability. Mol Carcinog. 2009;48(4):286-98.
90. Bose P, Hermetz KE, Conneely KN, Rudd MK. Tandem repeats and G-rich sequences are enriched at human CNV breakpoints. PLoS One.2014; 9(7):e101607.
91. Zhao J, Bacolla A, Wang G, Vasquez KM. Non-B DNA structure-induced genetic instability and evolution. Cell Mol Life Sci. 2010;67(1):43-62.
92. Bernstein KA, Gangloff S, Rothstein R. The RecQ DNA helicases in DNA repair. Annu Rev Genet.2010; 44:393-417.
93. Schwab RA, Nieminuszczy J, Shin-ya K, Niedzwiedz W. FANCJ couples replication past natural fork barriers with maintenance of chromatin structure. J Cell Biol. 2013;201(1):33-48.
94. Sharma S. Non-B DNA Secondary Structures and Their Resolution by RecQ Helicases. J Nucleic Acids.2011;724215.
95. Chen L, Zhou W, Zhang C, Lupski JR, Jin L, Zhang F. CNV instability associated with DNA replication dynamics: evidence for replicative mechanisms in CNV mutagenesis. Hum Mol Genet. 2015;24(6):1574-83.
96. Arlt MF, Mulle JG, Schaibley VM, Ragland RL, Durkin SG, Warren ST, et al. Replication stress induces genome-wide copy number changes in human cells that resemble polymorphic and pathogenic variants. Am J Hum Genet. 2009;84(3):339-50.
97. Startek M, Szafranski P, Gambin T, Campbell IM, Hixson P, Shaw CA, et al. Genome-wide analyses of LINE-LINE-mediated nonallelic homologous recombination. Nucleic Acids Res. 2015;43(4):2188-98.
98. Abeysinghe SS, Chuzhanova N, Krawczak M, Ball EV, Cooper DN. Translocation and gross deletion breakpoints in human inherited disease and cancer I: Nucleotide composition and recombination-associated motifs. Hum Mutat. 2003;22(3):229-44.
99. Chen JM, Ferec C, Cooper DN. Transient hypermutability, chromothripsis and replication-based mechanisms in the generation of concurrent clustered mutations. Mutat Res. 2011;750(1):52-9.
100. Liu P, Erez A, Nagamani SC, Dhar SU, Kolodziejska KE, Dharmadhikari AV, et al. Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements. Cell. 2011;146(6):889-903.
101. Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell. 2011;144(1):27-40.
102. Broman KW, Weber JL. Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain. Am J Hum Genet. 1999;65(6):1493-500.
103. Alkuraya FS. Autozygome decoded. Genet Med. 2010;12(12):765-71.

104. Hildebrandt F, Heeringa SF, Ruschendorf F, Attanasio M, Nurnberg G, Becker C, et al. A systematic approach to mapping recessive disease genes in individuals from outbred populations. PLoS Genet. 2009;5(1):e1000353.
105. Pemberton TJ, Absher D, Feldman MW, Myers RM, Rosenberg NA, Li JZ. Genomic patterns of homozygosity in worldwide human populations. Am J Hum Genet. 2012;91(2):275-92.
106. Curtis D, Vine AE, Knight J. Study of regions of extended homozygosity provides a powerful method to explore haplotype structure of human populations. Ann Hum Genet. 2008;72(Pt 2):261-78.
107. Wang SR, Agarwala V, Flannick J, Chiang CW, Altshuler D, Hirschhorn JN. Simulation of Finnish population history, guided by empirical genetic data, to assess power of rare-variant tests in Finland. Am J Hum Genet. 2014;94(5):710-20.
108. Campbell H, Carothers AD, Rudan I, Hayward C, Biloglav Z, Barac L, et al. Effects of genome-wide heterozygosity on a range of biomedically relevant human quantitative traits. Hum Mol Genet. 2007;16(2):233-41.
109. Sund KL, Zimmerman SL, Thomas C, Mitchell AL, Prada CE, Grote L, et al. Regions of homozygosity identified by SNP microarray analysis aid in the diagnosis of autosomal recessive disease and incidentally detect parental blood relationships. Genet Med. 2013;15(1):70-8.
110. Wang S, Haynes C, Barany F, Ott J. Genome-wide autozygosity mapping in human populations. Genet Epidemiol. 2009;33(2):172-80.
111. Lencz T, Lambert C, DeRosse P, Burdick KE, Morgan TV, Kane JM, et al. Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. Proc Natl Acad Sci U S A. 2007;104(50):19942-7.
112. Rudan I, Smolej-Narancic N, Campbell H, Carothers A, Wright A, Janicijevic B, et al. Inbreeding and the genetic complexity of human hypertension. Genetics. 2003;163(3):1011-21.
113. Chong JX, Ouwenga R, Anderson RL, Waggoner DJ, Ober C. A population-based study of autosomal-recessive disease-causing mutations in a founder population. Am J Hum Genet. 2010;91(4):608-20.
114. Jaber L, Shohat T, Rotter JI, Shohat M. Consanguinity and common adult diseases in Israeli Arab communities. Am J Med Genet. 1997;70(4):346-8.
115. Ben Halim N, Ben Alaya Bouafif N, Romdhane L, Kefi Ben Atig R, Chouchane I, Bouyacoub Y, et al. Consanguinity, endogamy, and genetic disorders in Tunisia. J Community Genet. 2013;4(2):273-84.
116. Wright A, Charlesworth B, Rudan I, Carothers A, Campbell H. A polygenic basis for late-onset disease. Trends Genet. 2003;19(2):97-106.
117. MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW. The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res. 2014;42(Database issue):D986-92.
118. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002;12(6):996-1006.
119. Chia NL, Bryce M, Hickman PE, Potter JM, Glasgow N, Koerbin G, et al. High-resolution SNP microarray investigation of copy number variations on chromosome 18 in a control cohort. Cytogenet Genome Res. 2013; 141:16-25.
120. Riggs ER, Jackson L, Miller DT, Van Vooren S. Phenotypic information in genomic variant databases enhances clinical care and research: the International Standards for Cytogenomic Arrays Consortium experience. Hum Mutat. 2012;33(5):787-96.
121. Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, et al. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. Am J Hum Genet. 2009;84(4):524-33.

122. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;42(Database issue):D980-5.
123. Fridlyand J, Snijders AM, Pinkel D et al. Hidden Markov models approach to the analysis of array CGH data, Journal of Multivariate Anal. 2004; 90: 132-153

# CHAPTER 3

# Copy Number Variation in Chromosome 18

Parts of this chapter were published in

High resolution SNP microarray investigation of copy number variation on chromosome 18 in a control cohort.

Nicole Chia, Michaela Bryce, Peter Hickman, Julia Potter,
Nicholas Glasgow, Gus Koerbin, Patrick Danoy,
Matthew A Brown and Juleen Cavanaugh

Cytogenetics and Genome Research 2013 141:16-25

## 3.1 Abstract

Copy number variation (CNV) has been recently reported to contribute to genetic heterogeneity. Previous studies have described copy number variation using lymphoblastoid cell lines or by application of specifically developed algorithms to interrogate previously described data. However, the full extent of CNV in a healthy population remains unclear. Using high density SNP microarray a comprehensive investigation of chromosome 18 is presented for CNV discovery and to describe CNV distribution and association with genomic architecture. 399 CNVs were detected, of which loss represents 98%, 58% are less than 2.5kb in size and 71% are intergenic. Intronic deletions account for the majority of copy number change with gene involvement. Furthermore, repetitive sequences are not uncommon within the duplicated or deleted DNA segment. This study provides a comprehensive investigation of CNV on chromosome 18 and establishes a framework for the investigation of all autosomes.

## 3.2 Introduction

Most studies of normal populations have been performed in the context of providing controls against which a pathogenic cohort is assessed in an effort to assign disease associations. Efforts then turned to describing the distribution and extent of variation in copy number in normal populations. Shaikh et al. 2009 investigated 2026 disease free children and parents using the Illumina Infinium II Human Hap 550 BeadChip (1). Two studies reported by Altshuler et al. 2010 and

Conrad et al. 2010 investigated subjects sourced from the International HapMap Consortium (2, 3).

The improvement of high resolution microarray has led to a redefining of known CNV (4-6). Few of the CNV identified in earlier studies were validated by alternative platform or molecular methods (7). Studies such as Matsuzaki et al. 2009 revised the type and size of known CNV regions, some of which are registered in the Database of Genomic Variants (DGV), to address this deficiency (5). While much of the research has focused on CNV discovery few investigators have provided precise definition of breakpoints by sequencing. Several studies have reported an association of CNV with segmental duplications and low copy repeat sequences (2, 6-13). More importantly, characterisation of specific breakpoints at the sequence level will enable investigators to interrogate this association and infer mechanisms of CNV formation.

In this study CNV in a cohort of a Western European descendent female population are investigated. The purpose of this chapter is to explore the incidence, distribution and copy number state of CNV on chromosome 18. Chromosome 18 was selected for the first phase of a wider study. CNV are investigated for association with chromosome morphological structures, genomic architecture and gene density. The investigation workflow is presented as a model to explore CNV in other chromosomes.

## 3.3 Materials and Methods

### 3.3.1 Sample collection

The "Aussie Normal" Collection is a community based collection of healthy Australians who are recruited from the Australian Electoral Roll. All participants have given informed written consent. In addition to the 64 individuals used in this study, a further 354 "Aussie Normal" samples were used to assess CNV frequency. An additional 82 DNA samples from anonymous Australian Red Cross blood donors were also used in the population analysis to a total of 500 samples. All investigations were performed with approval of the Australian National University Human Research Ethics Committee.

### 3.3.2 Cytogenetic analysis

Lymphocyte cultures were established for all 64 samples. Peripheral blood is cultured in RPMI medium and HAMs F10 containing phytohaemagglutinin (PHA) for 72-96 hrs. The cells are arrested in metaphase using Colcemid (40ug/ml) for 9 min, 0.075M KCL for 13 min, 5% acetic acid wash and 4 changes of Carnoy fixative. Metaphase spreads at the 550-700 bands per haploid set were analysed for constitutional structural rearrangements. No constitutional cytogenetic abnormality was detected in any of the 64 samples.

### 3.3.3 DNA extraction

DNA was extracted from whole blood using standard organic separation and ethanol precipitation. Peripheral blood is transferred to a 50ml falcon tube and lysis buffer added to a volume of 50ml. The tube is shaken vigorously and centrifuged at 1800rpm. After repeating the process the pellet is resuspended in STE buffer (NaCl/EDTA). Proteinase K and 10% SDS are added and incubated overnight. Phenol/chloroform is added to the solution and the organic and aqueous layers separated. Cold ethanol is added to the aqueous suspension for DNA precipitation. The DNA is dried through repeated ethanol rinses and air dried. The DNA "threads" are resuspended in TE pH 8.0 buffer for storage.

### 3.3.4 Analysis of microarray data

Genome-wide analysis was performed on 64 apparently healthy females of Western European descent. CNV discovery was performed using Illumina Human Omni1-Quad, which has a median probe interspacing of 1.2kb and overall resolution of 5 kb. The standard Illumina cluster file was used for the analysis. To minimise false positives, two different CNV detection algorithms, CNV Partition (Illumina, Inc.) and PennCNV (14) were applied to the raw data generated by Illumina Omni1-Quad. CNV calls detected by both algorithms were considered "real" and included for further investigation.

All samples achieved a genotype call rate of 99.8%. The robustness of the assay was confirmed by calculating the standard deviation of the LogR value for

autosomal probes for each sample. The assay achieved an average LogR deviation of 0.13 and BAF deviation of 0.029.

A CNV is defined here as >4 consecutive probes with LogR <-0.3 for deletion and >0.15 for duplication. This threshold was determined by calculating the mean LogR value of 40 normal samples in 4 regions and compared with the LogR values of a total of 30 regions that recorded a copy number of 0, 1, 3, 4.

Overlapping CNV are classified into CNV regions (CNVR). CNVR is defined here as a region where there is a shared start and or end position, greater than 50% overlap, or the CNV is within a larger CNV region (Figure 1). All CNV calls were cross referenced against the CNV registered in the Database of Genomic Variants (DGV) and Children's Hospital of Pennsylvania (CHOP) repositories and a CNV was considered novel if there was no record of registration in either database.

For the purpose of this study, common CNV is defined as an incidence of >5 individuals in the cohort, low frequency CNV is 2-5 individuals and private variants occur in one individual in the cohort.

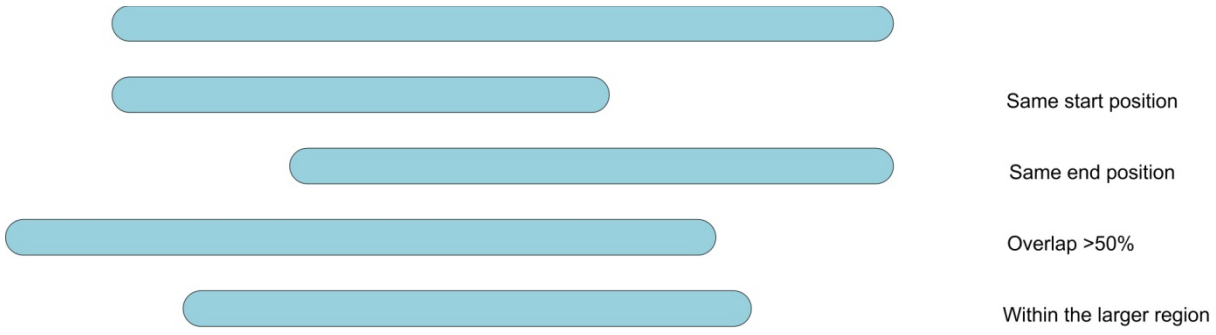



**Figure 1.** Definition of a copy number region (CNVR) as applied in this study.

### 3.3.5 CNV on chromosome 18

Chromosome 18 was selected in the first phase of the investigation of CNV in the study here. This chromosome was selected on the basis of chromosomal length, CNV incidence and paucity of reports of CNV in the literature.

The CNVR on chromosome 18 were issued with a unique identifier. This was done in consecutive order of CNV location, from the distal short arm to the distal long arm of chromosome 18. To determine the distribution of CNV on chromosome 18, the CNV incidence was recorded for 5 Mb sections along chromosome 18.

### 3.3.6 Molecular confirmation

Two CNVR (18q12.3 CNV18020:38308078–38311652 and 18q21.2 CNV18027:49390404-49391772)(NCBI38/hg18) were selected for experimental confirmation, breakpoint estimation and population frequency estimation (Figure 2). Characteristics of the two CNVR are shown in Table 1. CNV 18020 was called as a deletion in 16 samples and CNV18027 called as a deletion in 8 samples. Selection criteria for these CNVR were based on

1) Incidence (more than one sample)
2) CNVR detected by both CNV detection programs,
3) CNV type,
4) Start and end position consistent with CNVR

Homozygous and heterozygous deletions were confirmed by PCR. Primers were selected to flank the putative breakpoints and within the CNV using Primer 3 V4.0 and default parameters applied. A PCR tiling path was designed to refine the breakpoints. To achieve this, DNA primers were selected proximal and distal to the putative start and end positions and within the intervening sequence (Figure 3). Primer pair combinations were tested to obtain an optimal pair that produced a PCR product consistent with the expected copy number change (Figure 4). Samples known to carry the CNV were tested with the optimised forward and reverse primers along with an equivalent number of normal diploid controls. PCR products were visualised on 1.5% agarose and measured against a 1kb ladder (New England Biolabs).

PCR conditions were optimised for CNV18.020 consisting of 1 cycle of 94˚C for 90s and 40 cycles of 94˚C for 30s, 57.5˚C for 60s and 72˚C for 3.0 mins. PCR optimised for CNV18.027 consisted of 1 cycle of 94˚C for 30s and 40 cycles of 57.5˚C for 20s followed by 72˚C for 40s.

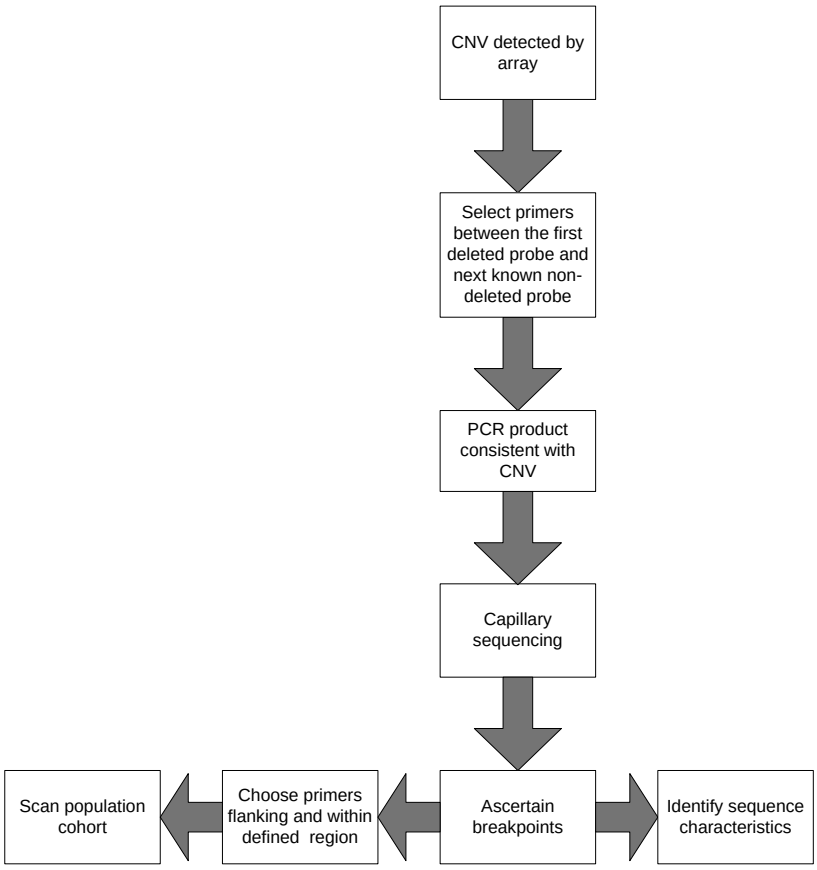



**Figure 2.** Pipeline of molecular confirmation and population screening of selected CNV.

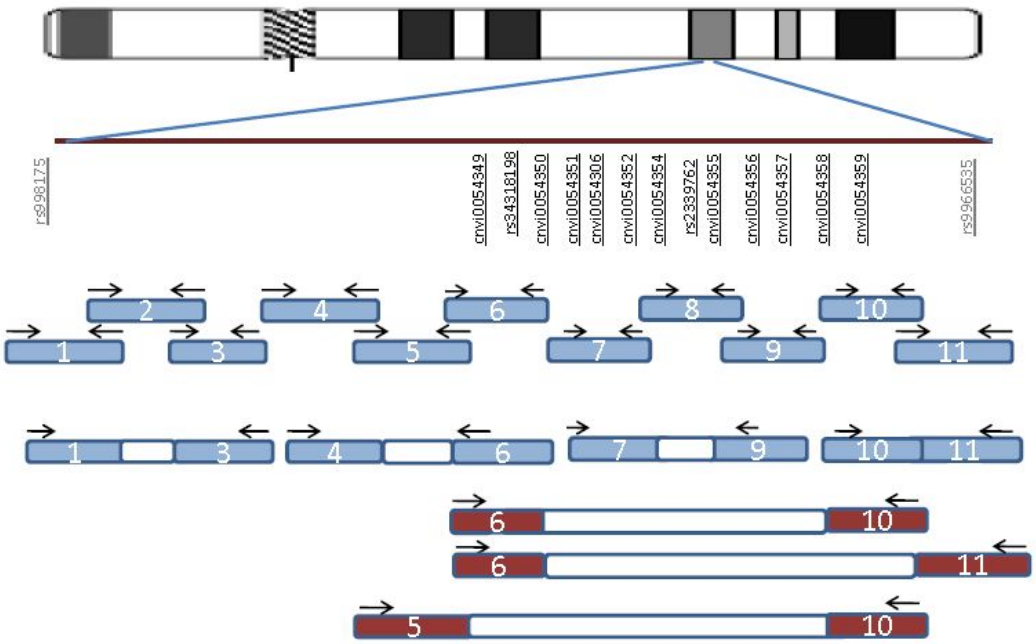



**Figure 3.** Experiment design to determine the optimised set of forward and reverse primers.

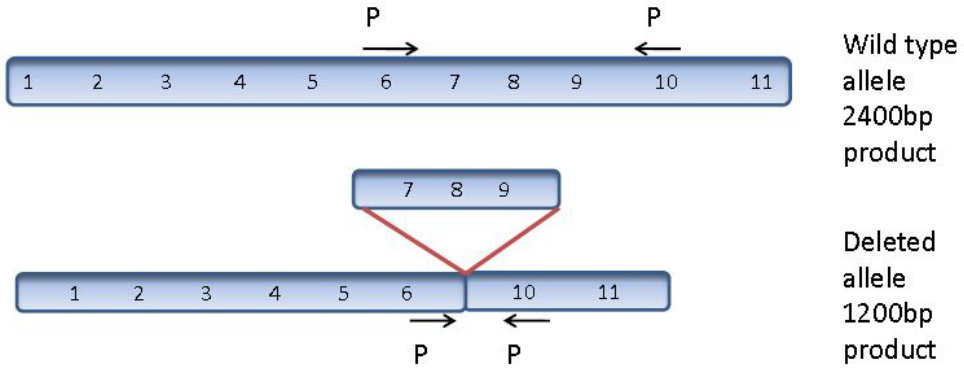



**Figure 4.** Optimised PCR product for CNV18027 (18q21.2: 49390404-49391772)

**Table 1a**. Call characteristics and confirmation of CNV18020.

| Sample id. | CNV Partition | | PennCNV | | Actual | | Size bp | Ave.LogR | Validation |
|---|---|---|---|---|---|---|---|---|---|
| **CNV18020** | **Start** | **End** | **Start** | **End** | **Start** | **End** | | | |
| AN2301 | Not called | | 38308102 | 38311523 | | | 3575 | -0.33 | PCR |
| AN2305 | 38308102 | 38311523 | 38308102 | 38311523 | 38308078 | 38311652 | 3575 | -3.88 | PCR/Sequence |
| AN2308 | Not called | | 38308102 | 38309912 | | | 3575 | -0.44 | PCR |
| AN2310 | 38308102 | 38311523 | 38308102 | 38309912 | | | 3575 | -0.40 | PCR |
| AN2311 | 38308102 | 38311523 | 38308102 | 38309912 | 38308078 | 38311652 | 3575 | -0.39 | PCR/Sequence |
| AN2313 | 38308102 | 38311523 | 38308102 | 38311523 | | | 3575 | -0.43 | PCR |
| AN2321 | Not called | | 38308709 | 38311523 | | | 3575 | -0.38 | PCR |
| AN2322 | 38308102 | 38311523 | 38308102 | 38311523 | | | 3575 | -0.41 | PCR |
| AN2327 | 38308102 | 38311523 | 38308102 | 38311523 | | | 3575 | -0.38 | PCR |
| AN2333 | Not called | | 38308102 | 38310534 | | | 3575 | -0.40 | PCR |
| AN2334 | Not called | | 38308266 | 38311238 | | | 3575 | -0.38 | PCR |
| AN2336 | 38308102 | 38311523 | 38308102 | 38311523 | | | 3575 | -0.39 | PCR |
| AN2342 | Not called | | 38308102 | 38311523 | 38308078 | 38311652 | 3575 | -0.38 | PCR/Sequence |
| AN2344 | Not called | | 38308266 | 38311523 | | | 3575 | -0.37 | PCR |
| AN2349 | Not called | | 38308484 | 38309912 | | | 3575 | -0.46 | PCR |
| AN2358 | 38308102 | 38311523 | 38308102 | 38311523 | 38308078 | 38311652 | 3575 | -4.08 | PCR/Sequence |

**Table 1b**. Call characteristics and confirmation of CNV 18027

| Sample id. | CNV Partition | | PennCNV | | Actual | | Size bp | Ave LogR | Validation |
|---|---|---|---|---|---|---|---|---|---|
| **CNV18027** | **Start** | **End** | **Start** | **End** | **Start** | **End** | | | |
| AN2244 | 49390524 | 49391736 | 49390524 | 49391736 | | | 1369 | -0.49 | PCR |
| AN2313 | 49390524 | 49391736 | 49390524 | 49391736 | | | 1369 | -0.53 | PCR |
| AN2321 | 49390524 | 49391736 | 49390524 | 49391736 | | | 1369 | -0.42 | PCR |
| AN2339 | 49390524 | 49391736 | 49390524 | 49391736 | 49390404 | 49391772 | 1369 | -0.49 | PCR/Sequence |
| AN2340 | Not called | | 49390524 | 49391736 | | | 1369 | -0.39 | PCR |
| AN2345 | 49390524 | 49391736 | 49390524 | 49391736 | 49390404 | 49391772 | 1369 | -0.55 | PCR/Sequence |
| AN2355 | 49390524 | 49391736 | 49390524 | 49391736 | | | 1369 | -0.42 | PCR |
| AN2361 | Not called | | 49390524 | 49391736 | | | 1369 | -0.40 | PCR |

## 3.3.7 Breakpoint accuracy

To assess the accuracy of breakpoint estimation capillary sequencing was performed for CNV18020 and CNV18027 using the optimised forward and reverse primer pair. The PCR conditions were 1 cycle of 94˚C for 90s and 40 cycles of 94˚C for 30s, 57.5˚C for 30s and 72˚C for 2.0 mins using 20ng of template. The sequence was then compared to the reference genome to determine the breakpoint (http://genome.ucsc.edu; NCBI36/hg18; accessed 18th October, 2011).

## 3.3.8 Population screen

A PCR assay was developed for population screening to estimate the frequency of the CNV in the healthy population. The PCR was optimised to permit multiplexing of three primers in the following ratios, 2:1:1 (forward primer: reverse primer 1: reverse primer 2). Forward and reverse primers were positioned flanking the breakpoint and a second reverse primer located within the breakpoint region (Figure 5). This protocol provided differential separation of products for wild-type, heterozygous and homozygous deletions (Figure 6).

PCR conditions were optimised for CNV18027 for the population screen and consisted of 1 cycle of 94˚C for 20s and 35 cycles of 57.5˚C for 15s followed by 72˚C for 15s. Screening was undertaken on 12% polyacrylamide gels (Table 2).

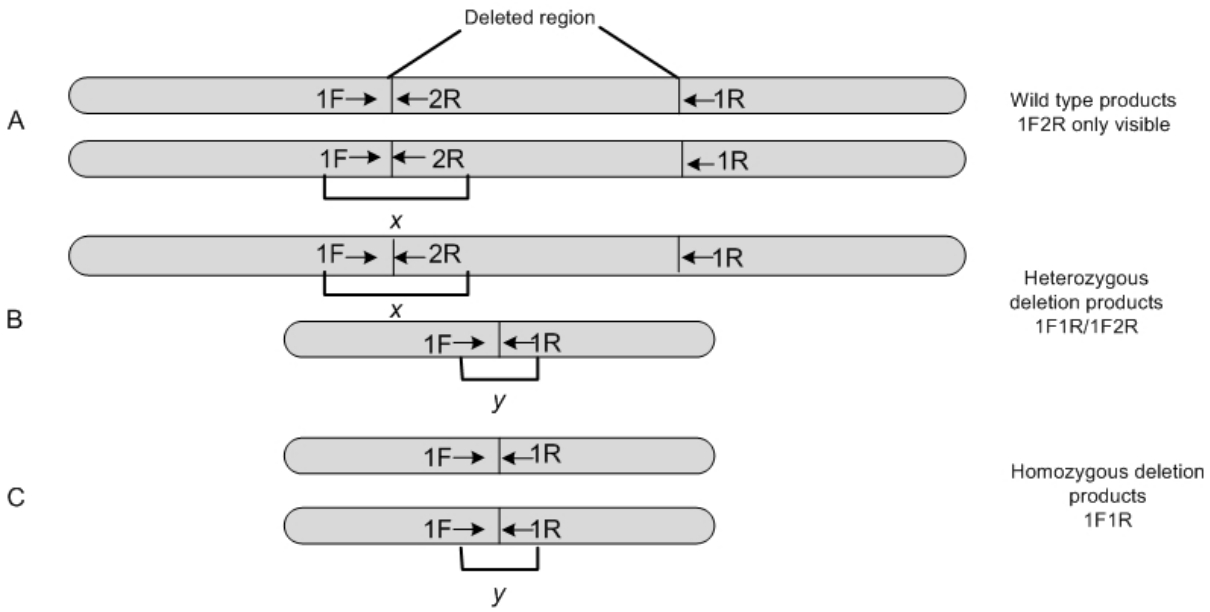



**Figure 5.** Experimental design of multiplex PCR used to distinguish genotypes in the population screen. The PCR products are represented by *x* and *y.* PCR

conditions are optimised to achieve products using the primers 1F2R for wild type, 1F1R and 1F2R for heterozygous loss and 1F1R product for homozygous loss.

**Table 2.** The expected products for the population screen of CNV 18027.

| | Wild Type | Heterozygous deletion | Homozygous deletion |
|---|---|---|---|
| **PCR Product** | F1R2 | F1R1/F1R2 | F1R1 |
| **Expected Result** | 197 bp | 197/131 bp | 131 bp |
| **Gel Product** | | | |

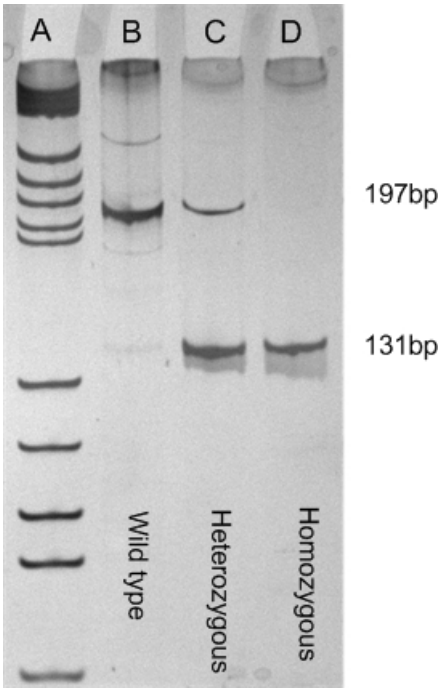



**Figure 6.** Results of a genotype screen. The ladder A) PBR/MSP1. Products for the (B) wild type (197bp) (C) heterozygous deletion (197/131bp) and (D) homozygous deletion (131bp) are demonstrated.

### 3.3.9 Website investigations

CNV were analysed using the UCSC Browser (http://genome.ucsc.edu/ hg18) for chromosome band location, gene content and repetitive elements. Ensembl (www.ensembl.org) was used for gene content and gene density estimations. Gene functional clusters were analysed using PANTHER classification system at http://www.pantherdb.org/. Sequence correlation with the reference genome was done using UCSC Blat (http://genome.ucsc.edu/ hg18). Secondary structures were investigated using Mfold (http://mfold.rna.albany.edu/ (15). The statistical analyses were performed using free access software at Graphpad Quick Calcs www.graphpad.com.

## 3.4 Results

### 3.4.1 Microarray investigations

A total of 399 CNV events representing 52 CNVR were detected on chromosome 18 by both CNV detection algorithms. Losses represent 98% of calls (390/399) while gains recorded 2% of calls (9/399). The majority (58%) of calls are less than 2.5kb in size with losses accounting for 99% of this size category. Copy number gains were more prevalent in the 2.5kb to 10kb size category representing 8/9 CNV gains (Figure 7).

The majority of CNVR (72%) are common and observed in more than one unrelated individual, while 28% are calls in a single individual. At the time of investigation 3 novel CNV were detected on chromosome 18. These are unique

events and may represent either false positive calls or private variants. All 3 novel CNV were called by both algorithms and passed QC parameters. Two novel CNV are >100kb and involve 1-3 RefSeq genes (NCBI36/hg18).

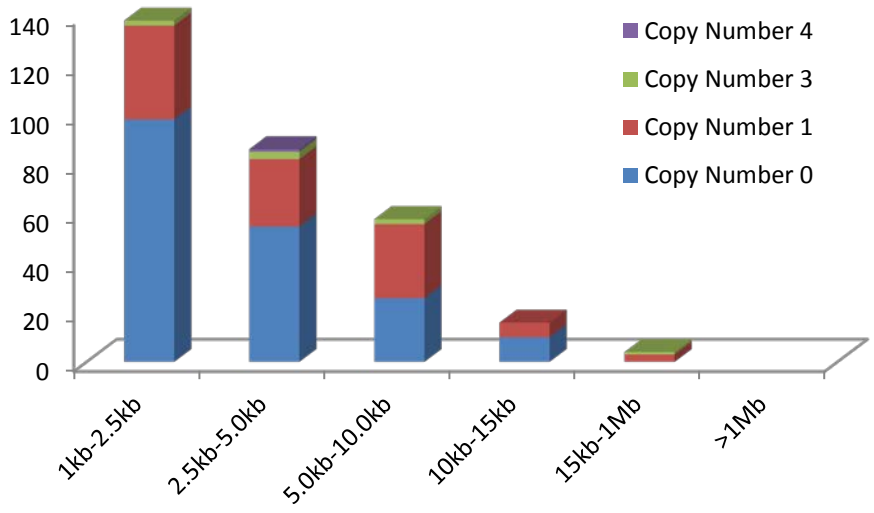



**Figure 7.** The proportion of CNV length and contribution of gain and loss to the total incidence of CNV on chromosome 18 is shown.

## 3.4.2 Chromosome 18 CNV landscape

The distribution of CNV and CNVR on chromosome 18 was investigated for the parameters of CNV length, copy number state and frequency. To do this the incidence of CNV and CNVR were recorded for 5Mb non-overlapping sections (Figure 8).

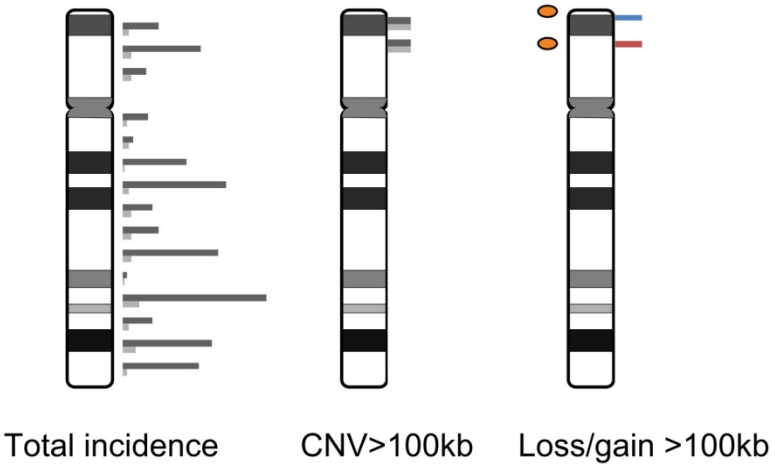



**Figure 8.** Schematic representation of CNV on chromosome 18 is shown for total incidence and CNV >100kb. The two CNV >100kb are located in the short arm representing loss (red) and gain (blue). Repetitive sequence (orange oval) is associated with these CNV.

Intra-chromosomal variation of CNV incidence was observed. For example, sections show high incidence whereas others such as 55-60Mb recorded a low incidence of CNV. To exclude the bias caused by high incidence common CNV the proportion of CNVR was analysed and the intra-chromosomal variation pertained (Figure 9).

The subtelomeric and pericentromeric regions have been previously identified as hotspots of genomic instability and associated with an increase in CNV incidence (7). This theory was assessed for chromosome 18. There was no evidence of increased CNV incidence in the subtelomeric regions of chromosome 18. The short and long arm subtelomere recorded 6% and 4% of total CNVR respectively, while there were no CNV recorded for the region immediately adjacent to the

centromere at 15-20Mb. The latter is most likely due to the low probe density in the centromeric region.

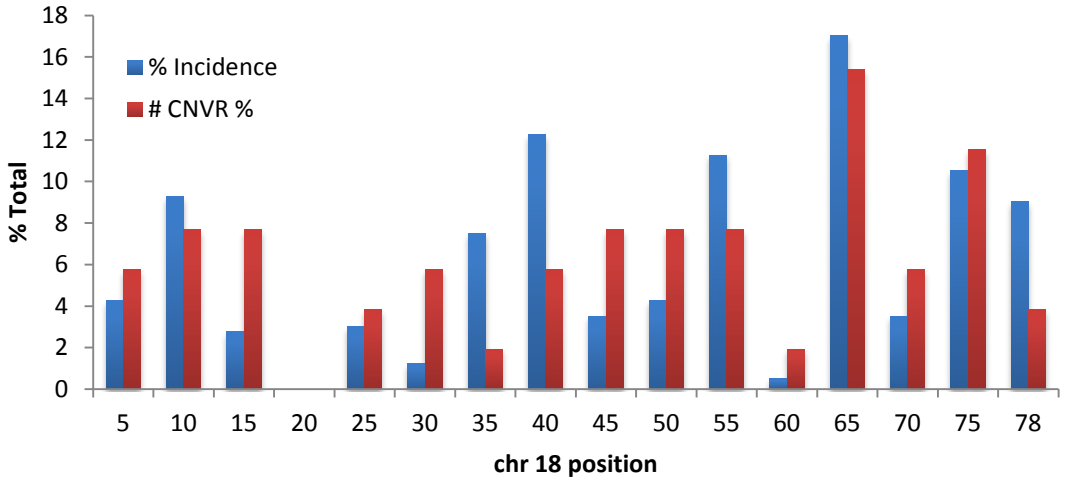



**Figure 9.** The proportion of CNV and CNVR in 5mb windows for chromosome 18 is shown. Intra-chromosomal variation is apparent, however not associated with chromosomal morphological features. High incidence common CNV are enriched in the regions adjacent to the subtelomere but the proportion of CNVR is not elevated. The pericentromeric region at 10-15Mb shows enrichment of CNVR but there is no evidence of enrichment of CNV.

Analysis of CNV length showed CNV <15kb distributed over the entire chromosome whereas CNV >50kb are located in the short arm. The chromosomal distribution of copy number state was analysed to determine if there is a correlation with the intra-chromosomal variation. Gain accounted for 9/399 of CNV on chromosome 18. Due to the low incidence of copy number gain evaluation

of distribution has not been made. Common CNV in the long arm accounts for 8/9 gains and the remaining CNV gain is a private variant >500kb located in the short arm.

### 3.4.3 Comparison to published control studies

To determine if the variation in CNV distribution is representative of chromosome 18 or isolated to the study performed here, the CNV output was cross referenced against previous reports of CNV on chromosome 18 (2, 5, 7). This also provides comparison of a cohort from a general population with that using DNA from transformed cell line of the HapMap cohort. A similar trend of CNV distribution was observed in the data presented for chromosome 18 demonstrating that the intra-chromosomal variation was observed by all studies. Matsuzaki et al. 2009 recorded similar results for the subtelomeric regions with approximately 2% (short arm subtelomere), 4% (long arm subtelomere) and 3% (centromere) respectively (Figure 10).

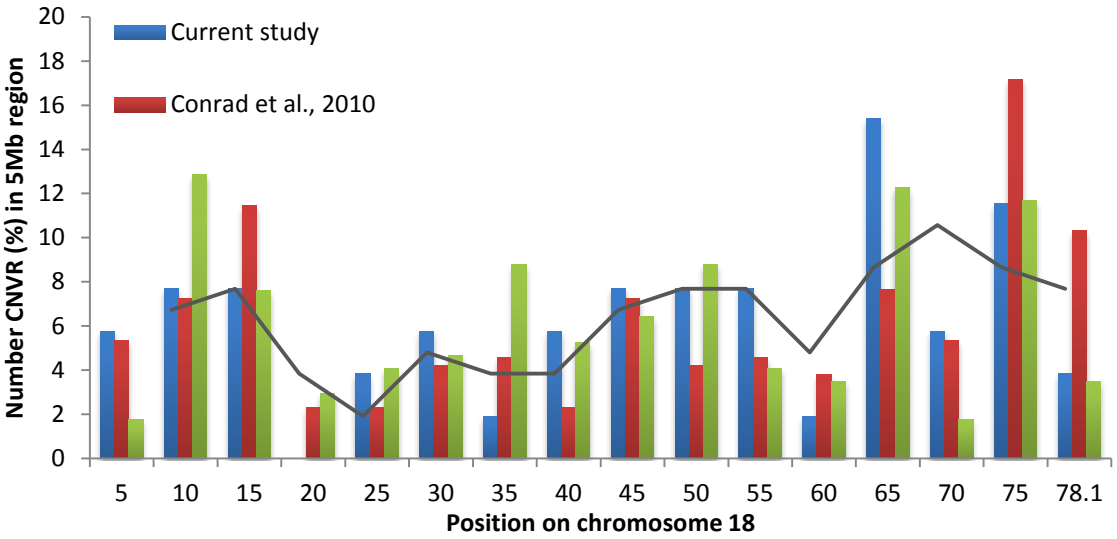



**Figure 10.** Comparison of the proportion of CNVR recorded in the current study (left), Conrad et al. 2010 (middle) and Matsuzaki et al .2009 (right). A consistent pattern of distribution of CNV regions along the length of chromosome 18 is apparent among the studies as illustrated by the trend line.

## 3.4.4 Population frequency of two confirmed CNVR

The PCR design applied in this study demonstrates that the judicious selection of primers can yield PCR products that are amenable to identification of individuals of each genotype. The results of sequence analysis and the population screen of 500 healthy individuals for CNV18020 and CNV18027 are presented in Table 3. CNV18020 located on chromosome 18q12.3 at 38308078–38311652 bp has been reported previously (2, 3, 5, 14). All of these latter studies have been performed using lymphoblastoid cell lines used in the HapMap investigations. Losses have

been reported by Conrad et al. 2010, Matsuzaki et al. 2009 and Aultsuhler et al. 2010, whereas in the report by Wang et al. 2007 using PennCNV, a gain event was reported in a single sample from the 112 HapMap individuals. An allele frequency of 0.032 and 0.039 was reported by Conrad et al. 2010 and Aultsuhler et al. 2010 respectively (2, 3). In the study here the frequency is higher at 0.094 in 500 healthy Caucasian controls. This finding may be representative of true variance between populations or a reflection of the screening method. As demonstrated screening by PCR minimises the high false negative rate that is observed in CNV detection algorithms that are applied to array data.

CNV18027 located on chromosome 18q21.2 at 49390404-49391772bp has been reported previously by genome-wide investigation of HapMap individuals (2, 5, 18-20) (Table 3). Both losses and gains have been reported in this region. Matsuzaki et al. 2009 reported a deletion in 13 of 90 Yoruba Nigerians from the HapMap project. Using ultra-high resolution aCGH platform, Park et al. 2010 investigated 30 HapMap individuals from three Asian populations to identify putative Asian specific CNVs and reported a deletion in 19/30 individuals (19). This frequency is significantly higher than the allele frequency of 0.061 reported here in 500 healthy Caucasians. Conversely, Conrad et al. 2010 reported a gain in the same region in 23/40 HapMap individuals (European and West African Ancestry)(2). This finding may represent a population specific variant.

**Table 3**. Review of the literature for the two CNVs investigated in this study

| Reference | DGV id # | Start | End | Size | Cohort | Incidence | Allele freq | Type |
|---|---|---|---|---|---|---|---|---|
| Current CNV18.027 | | 49390404 | 49391772 | 1369 | Healthy Australian whole blood | 59/500 | 0.061 | Loss |
| Kim et al 2009 | Var_58835 | 49390456 | 49391785 | 1330 | AK1 genomic DNA | 1 | | Gain |
| Matsuzaki et al 2009 | Var_8850 | 49390537 | 49391615 | 1078 | HapMap cell line DNA | 13/90 | | Loss |
| Conrad et al 2009 | Var_73150 | 49390443 | 49391768 | 1325 | HapMap cell line DNA | 23/40 | | Gain |
| Ju et al 2010 | Var_104354 | 49390426 | 49391842 | 1417 | HapMap cell line DNA | 1 | | Loss |
| Park et al 2010 | Var_114594 | 49390426 | 49391842 | 1416 | 10 Korean, 20 HapMap Asian whole blood and LCL | 19/30 | | Loss |
| Current CNV18.020 | | 38308077 | 38311652 | 3575 | Australian control whole blood | 77/500 | 0.094 | Loss |
| Conrad et al 2009 | Var_67345 | 38308078 | 38311677 | | HapMap cell line DNA | 28/450 | 0.032 | Loss |
| Matsuzaki et al 2009 | Var_88814 | 38308305 | 38311765 | 3185 | HapMap cell line DNA | 4/90 | | Loss |
| Aultshuler et al 2010 | Var_104144 | 38309785 | 38311765 | | HapMap cell line DNA | 88/1184 | 0.039 | Loss |
| Wang et al | Var_9780 | 38309785 | 38312891 | | HapMap cell line DNA | 1/112 | | Gain |

## 3.4.5 CNV association with gene density

The distribution of CNV was compared to the distribution of genes on chromosome 18 to investigate putative correlations. The number of genes for each 5Mb non-overlapping segment was scored from www.NCBI.nlm.nih.gov (NCBI36/hg18) and compared to the incidence of CNV (Figure 11). In general, high incidence common CNV are more frequently associated with low gene density whilst high gene density is observed where there is a low incidence of CNV including low frequency and private variants.

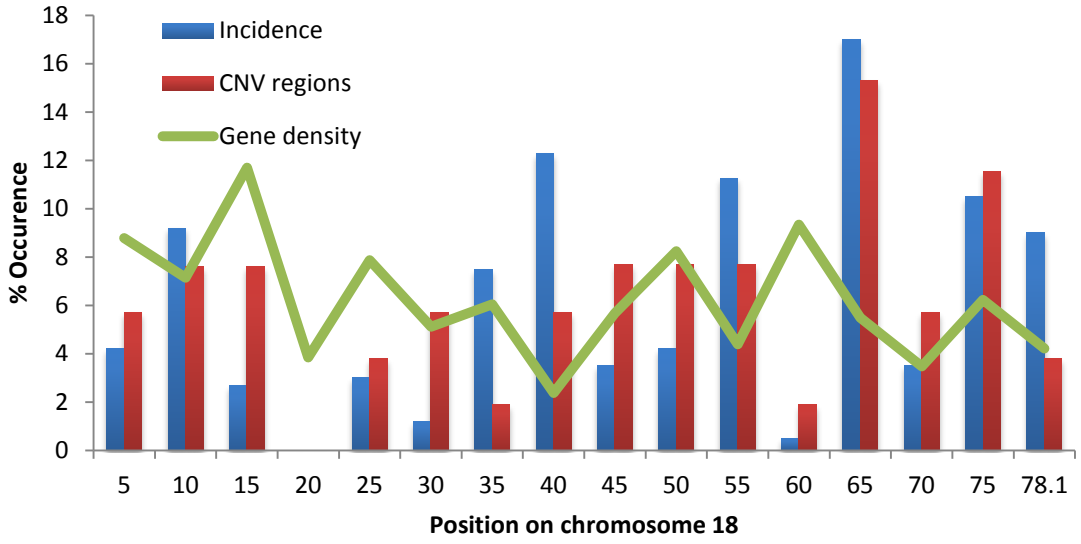



**Figure 11.** CNV distribution on chromosome 18 compared to gene density. The incidence (number of individuals) is compared to the number of CNV regions and number of genes represented within the 5Mb segment and is displayed as percent of the total. Common and low frequency CNV reside in both gene rich and regions of low gene density.

The adjacent segments between 55-60Mb and 60-65Mb exemplify the apparent correlation of CNV incidence and gene density (Figure 11). There is a single private variant recorded in one sample for the segment located at 55-60 Mb and 68 CNV representing 5 high incidence CNVRs and 3 low frequency CNVRs in the 60-65Mb segment. Inspection of gene distribution within the 5Mb sections reveals differences in the number of genes, 51 genes (9%) in the 55-60Mb region compared to 30 genes (5%) in the 60-65 Mb region. In addition gene deserts were observed in both regions measuring <1Mb and 1.5Mb respectively (NCBI36/hg18, accessed 13/06/2011) (Figure

12). There is gene involvement in only one of the 9 CNV represented within this 10Mb region and this is an intronic deletion in *CDH19*.

Investigation of the gene ontology of this region suggests that gene function does not appear to relate to CNV incidence. At the time of investigation both 5Mb regions were seen to contain cancer related genes e.g. *MALT1* and *BCL2*. Pseudogenes are located within the 55-60Mb region whereas serpin peptidase inhibitor genes essential for the attenuation of peptidase (16) and cadherin, calcium channel cell adhesion genes are located within the region 60-65Mb. These findings suggest that for these adjacent 5Mb regions on chromosome 18, CNV incidence does not appear to correlate with gene function, although correlation with gene density and distribution has not been excluded.

The gene content was reviewed for each CNV on chromosome 18. The results show that 71% of CNVR are intergenic and 27% of CNVR showed involvement of a single gene with the latter group showing a predominance of deletions within introns. Two novel CNV >100kb were detected on chromosome 18. CNV18002.1 located at 3486030-4445266 involves an intronic deletion of *DLGAP1.* CNV18006.1 located at 7101033-7648679 is a 547kb duplication involving exon 1 of *PTPRM* and *LAMA1* and full involvement of *LRRC30* (www.Ensembl.org). This is the only CNV involving multiple genes (http://genome.ucsc.edu).

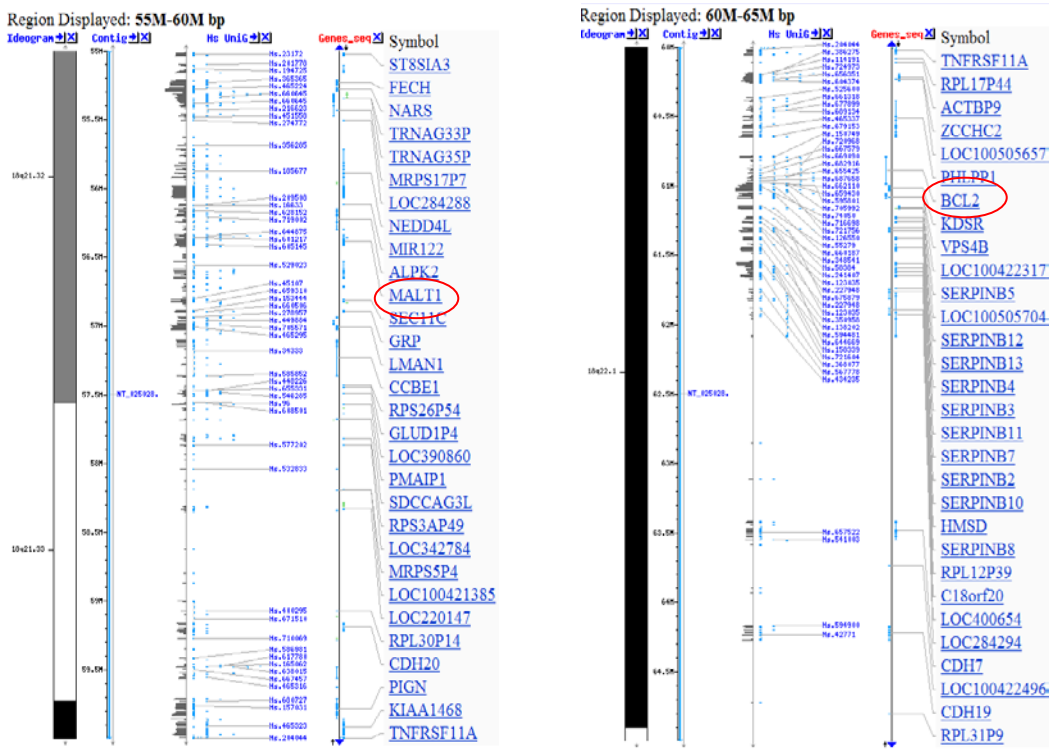



**Figure 12.** The gene distribution is shown for the segments 55-60Mb and 60-65Mb. These correspond with a private variant in 55-60Mb and several common CNVR at 60-65Mb (www.ensembl.org/ accessed 28/03/2011).

Gene ontology investigations of the genes encompassed by CNV on chromosome 18, using PANTHER analysis tools, reveals a predominance of genes involved in cell growth and differentiation, signalling and cell adhesion pathways. CNV involving DNA transcription factors or protein metabolic pathways were only observed in private variants. These findings are consistent with previous reports adding support to the concept that common and low frequency CNV are unlikely to be involved in disease associations (17).

## 3.5 Discussion

The incidence, distribution and association of copy number change with morphological and genomic structures on chromosome 18 are presented. The results of the study here show that the distribution of CNV shows no evidence of association with chromosome morphological structures. However there is intra-chromosomal variation in the distribution of CNV and CNVR on chromosome 18. Two thirds of the CNV are intergenic, while intronic deletions predominantly account for gene disruption. Breakpoints within repeat and non-repeat sequences provide evidence for multiple mechanisms of derivation.

### 3.5.1 Detection of copy number variation

Algorithms used for CNV detection can produce varied results depending on calling criteria and preparation of reference samples (21-24). Two CNV detection algorithms are used in this study with the prime aim of providing a greater degree of confidence of true calls and minimizing false positive calls. The employment of this stringent approach improves the robustness of the assay and confidence of the calls.

The excess of deletions over duplications observed in this study is in accord with that observed in other genome-wide CNV studies (1-3). The common finding in the studies from Conrad et al. 2010, Shaikh et al. 2009 and Altshuler et al. 2010 is a significantly higher number of deletions than duplications, ranging from a ratio of 5:1 to 7:1 and as high as 11.5:1 respectively (1, 3). The elevated incidence of

deletions in comparison to duplications has been attributed in part to the reduced power of algorithms to detect small duplications (2, 3, 25). Alternatively, the higher incidence of copy number loss may reflect true biology and be a consequence of the mechanism of formation of CNVs (26). A wider investigation of copy number change for all chromosomes will indicate if the ratio of deletion to duplication reported for chromosome 18 is consistent with others or an indicator of architectural or selection factors.

A review of CNV length on chromosome 18 revealed that 58% are less than 2.5kb. Some of these may represent false positive calls. However the risk of false positives is reduced by the application of two CNV detection algorithms (22, 24). Experimental confirmation of two CNVR measuring 1369bp (CNV18027) and 3575bp (CNV18020) provide evidence of the robustness of the process.

## 3.5.2 Distribution

Analysis of CNV on chromosome 18 demonstrated intra-chromosomal variation. There was no evidence of association with chromosomal morphological landmarks. The definition of subtelomere applied here as described by Knight et al. 2000, is the region of repetitive DNA proximal to the telomere that shares sequence similarity between numerous chromosomes (27). The size of the subtelomere varies for each chromosome but is <1mb from the telomere sequence (28). With this in mind there was no evidence of enrichment at the subtelomere or pericentromeric regions for chromosome 18. However marker coverage is not supported in this region due to high level of sequence homology. The first CNV on

chromosome 18p is at position chr 18:1898848-1964966 and 18q at position 75746951-75747880. Enrichment of CNV was observed in the 5Mb segments adjacent to, but not within, the subtelomere and pericentromeric regions.

### 3.5.3 CNV and gene density correlations

An increased incidence of CNV with regions of high gene density has been previously reported (7). This correlation was tested in this study for chromosome 18. Gene involvement was observed in 16/52 CNVR, although intronic deletions accounted for 15/16 CNVR. Overall, common, high incidence CNV occurred within regions of low gene density, while regions of high gene density correlated with low frequency and private variants. An example of this is the region encompassing 55-60Mb (18q21.33) and the adjacent region 60-65Mb (18q22.1). These adjacent regions reflect contrasting levels of CNV incidence with gene density. Examination of gene distribution shows fewer genes and larger segments of gene deserts associated with the high incidence CNV. Gene ontology investigations suggested that there was no correlation with gene function, as both regions contain oncogenes and cell adhesion genes. However analysis of the same region using the UCSC genome browser (http://genome.ucsc.edu/ hg18) indicated a ”shift” in the position of genes compared to the original analysis using www.ensembl.org. As such *BCL2* is located at 58.9-59.1Mb in the same segment as *MALT1* consistent with the occurrence of low frequency CNV in this region. The serpin peptidase genes are also located in the 55-60Mb segment which has low CNV incidence.

Despite the incidence of CNV in regions of high gene density in some 5Mb sections on chromosome 18, close inspection revealed that these CNV are intergenic. Investigations of gene ontology reveals a predominance of genes involved in cell growth and differentiation, signalling and cell adhesion pathways in low frequency and common copy number change. Deletions involving DNA transcription factors or protein metabolic pathways were only observed in private variants. These findings are consistent with previous reports adding support to the concept that common and low frequency CNV are unlikely to be involved in disease associations (17). Albeit the significance of intronic copy number changes on gene function and hence disease association is yet to be fully ascertained. In this study a deletion was observed in the "deleted in colorectal cancer" (DCC) tumour suppressor gene in 7/64 samples. Intronic deletions and disruption to tumour suppressor genes has been shown to confer altered risk in cancer (29).

### 3.5.4 Limitations

The study performed here investigates the characteristics of CNV on chromosome 18. Experimental confirmation of 24 CNV <5kb provided evidence of real calls for these CNV. However, due to the presence of only 2 CNV>100kb, confirmation of large CNV on chromosome 18 was not performed. Likewise the CNV investigated by PCR methods are deletions and investigation of the characteristics of duplications has not been performed. To explore duplications the positional information must also be determined and is not suitable for PCR methods.

During the course of investigation the genome build was updated. To ensure continuity the genome build is maintained at NCBI36/hg18 for the duration of this study. However changes to gene position with the release of new builds (GRCh37/hg19 Feb 2009 and GRCh38/hg38 December 2013), have made correlation of CNV with gene content difficult. In addition variation is observed between genome browsers e.g. http://genome.ucsc.edu/ and www.Ensembl.org , with regard to gene content and chromosomal position. The details reported here are in accordance with the information available at the time of investigation.

### 3.5.5 Further Investigations

Information gained from sequence analysis will contribute to the understanding of the mechanism of derivation and chromosomal distribution of CNV. Identification of exact breakpoints and sequence motif for 2 CNVR on chromosome 18 is presented in chapter 4. Proposed mechanisms are discussed for these CNV.

Following the investigation of CNV on chromosome 18, the model of investigation will be applied to chromosome 1-22 to describe CNV in a cohort of females from a Western European descendant population. Copy number gains were not characterised for chromosome 18, along with CNV >100kb. The study workflow will be expanded to include these investigations (Figure 13).

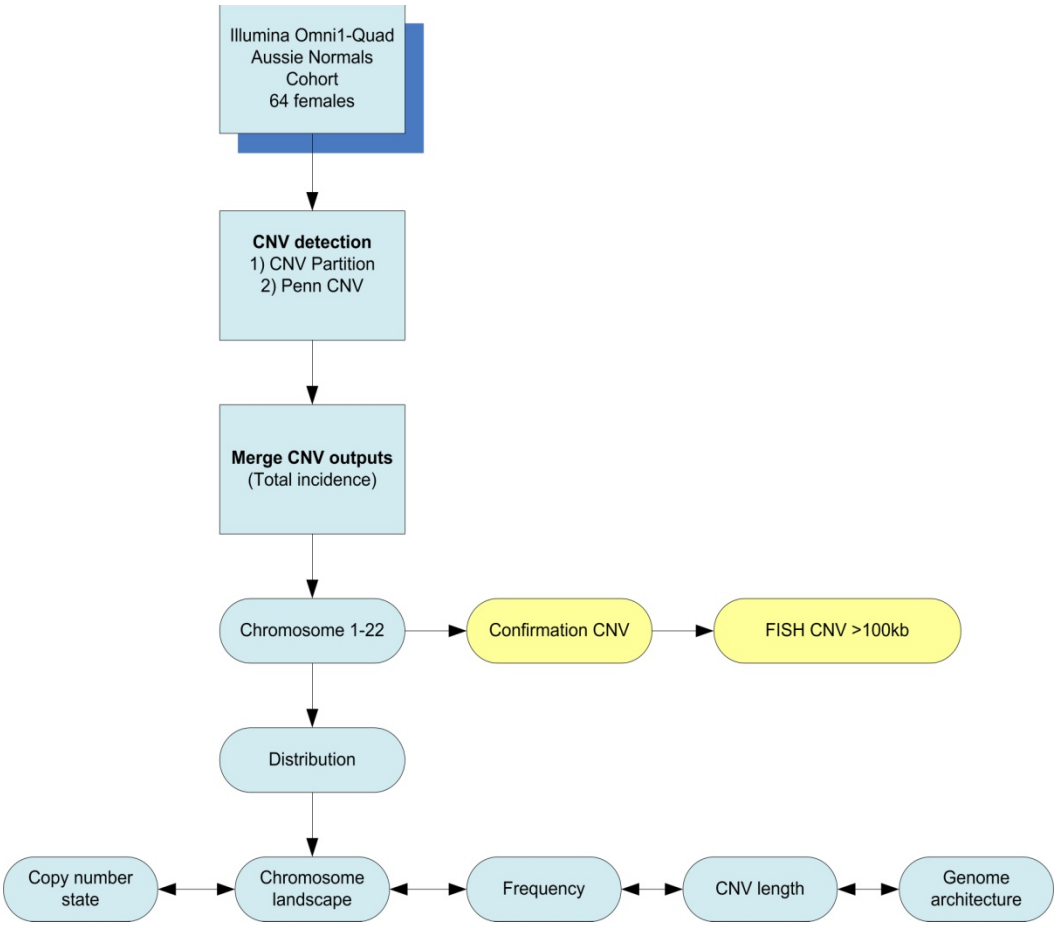



**Figure 13.** Pipeline of investigation for chromosomes 1-22, based on the example developed for chromosome 18.

## 3.6 Conclusion

Copy number variation on chromosome 18 is investigated in a cohort of healthy females from a Western European descendant population. The CNV reported here have been selected for inclusion in the Database of Genomic Variants. Comparison with the CNV outputs from HapMap studies confirmed consistency of results and identified population variation for the incidence of 2 CNVR. Small CNV

predominated on chromosome 18 with only 2 singletons >100kb detected and 80% less than 5kb in size. The cause for intra-chromosomal variation of CNV distribution for some regions on chromosome 18 suggested a bias away from gene involvement in the healthy population. Further investigation at the base pair level is required to ascertain the mechanism of CNV derivation and basis for CNV distribution. Furthermore a model of investigation developed for chromosome 18 can be applied to chromosome 1-22 to characterise CNV in this general population cohort. Finally this study contributes to the global aim of developing a map of genetic variation.

## 3. 7 Data Access

The CNV data presented herein is published at http://www.ebi.ac.uk/dgva/data-download: Accession number estd198

## 3.8 References


1. Shaikh TH, Gai X, Perin JC, Glessner JT, Xie H, Murphy K, et al. High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications. Genome Res. 2009;19(9):1682-90.
2. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, et al. Origins and functional impact of copy number variation in the human genome. Nature. 2010;464(7289):704-12.
3. Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467(7311):52-8.
4. McCarroll SA. Copy number variation and human genome maps. Nat Genet. 2010;42(5):365-6.
5. Matsuzaki H, Wang PH, Hu J, Rava R, Fu GK. High resolution discovery and confirmation of copy number variants in 90 Yoruba Nigerians. Genome Biol. 2009;10(11):R125.
6. Perry GH, Ben-Dor A, Tsalenko A, Sampas N, Rodriguez-Revenga L, Tran CW, et al. The fine-scale and complex architecture of human copy-number variation. Am J Hum Genet. 2008;82(3):685-95.
7. Cooper GM, Nickerson DA, Eichler EE. Mutational and selective effects on copy-number variants in the human genome. Nat Genet. 2007;39(7 Suppl):S22-9.
8. Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, et al. Segmental duplications and copy-number variation in the human genome. Am J Hum Genet. 2005;77(1):78-88.
9. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number in the human genome. Nature. 2006;444(7118):444-54.
10. Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, et al. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet. 2009;41(10):1061-7.
11. Hastings PJ, Ira G, Lupski JR. A microhomology-mediated break-induced replication model for the origin of human copy number variation. PLoS Genet. 2009 ;5(1):e1000327.
12. Hastings PJ, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. Nat Rev Genet. 2009;10(8):551-64.
13. Abeysinghe SS, Chuzhanova N, Krawczak M, Ball EV, Cooper DN. Translocation and gross deletion breakpoints in human inherited disease and cancer I: Nucleotide composition and recombination-associated motifs. Hum Mutat. 2003;22(3):229-44.
14. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. 2007;17(11):1665-74.
15. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 2003;31(13):3406-15.
16. Rawlings ND, Tolle DP, Barrett AJ. Evolutionary families of peptidase inhibitors. Biochem J. 2004;378(Pt 3):705-16.
17. Craddock N, Hurles ME, Cardin N, Pearson RD, Plagnol V, Robson S, et al. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature. 2010;464(7289):713-20.
18. Kim JI, Ju YS, Park H, Kim S, Lee S, Yi JH, et al. A highly annotated whole-genome sequence of a Korean individual. Nature. 2009;460(7258):1011-5.

19. Park H, Kim JI, Ju YS, Gokcumen O, Mills RE, Kim S, et al. Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. Nat Genet. 2010;42(5):400-5.
20. Ju YS, Hong D, Kim S, Park SS, Lee S, Park H, et al. Reference-unbiased copy number variant analysis using CGH microarrays. Nucleic Acids Res. 2010;38(20):e190.
21. Baross A, Delaney AD, Li HI, Nayar T, Flibotte S, Qian H, et al. Assessment of algorithms for high throughput detection of genomic copy number variation in oligonucleotide microarray data. BMC Bioinformatics. 2007;8:368.
22. Pinto D, Marshall C, Feuk L, Scherer SW. Copy-number variation in control population cohorts. Hum Mol Genet. 2007;16 Spec No. 2:R168-73.
23. Diskin SJ, Li M, Hou C, Yang S, Glessner J, Hakonarson H, et al. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. Nucleic Acids Res. 2008;36(19):e126.
24. Pinto D, Darvishi K, Shi X, Rajan D, Rigler D, Fitzgerald T, et al. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nat Biotechnol. 2011;29(6):512-20.
25. Itsara A, Cooper GM, Baker C, Girirajan S, Li J, Absher D, et al. Population analysis of large copy number variants and hotspots of human genetic disease. Am J Hum Genet. 2009;84(2):148-61.
26. Quinlan AR, Clark RA, Sokolova S, Leibowitz ML, Zhang Y, Hurles ME, et al. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. Genome Res. 2010;20(5):623-35.
27. Knight SJ, Flint J. Perfect endings: a review of subtelomeric probes and their use in clinical diagnosis. J Med Genet. 2000;37(6):401-9.
28. Knight SJ, Lese CM, Precht KS, Kuc J, Ning Y, Lucas S, et al. An optimized set of human telomere clones for studying telomere integrity and architecture. Am J Hum Genet. 2000;67(2):320-32.
29. Ramburan A, Chetty R, Hadley GP, Naidoo R, Govender D. Microsatellite analysis of the DCC gene in nephroblastomas: pathologic correlations and prognostic implications. Mod Pathol. 2004;17(1):89-95.

# CHAPTER 4

# CNV Distribution and Genomic Architecture:

# Mechanisms of CNV Derivation

Parts of this chapter were published in

High resolution SNP microarray investigation of copy number variation on chromosome 18 in a control cohort.

Nicole Chia, Michaela Bryce, Peter Hickman, Julia Potter, Nicholas Glasgow, Gus Koerbin, Patrick Danoy, Matthew A Brown and Juleen Cavanaugh

Cytogenetics and Genome Research 2013 141:16-25

## 4.1 Abstract

SNP microarray investigation of a cohort of healthy females from a Western European descendant population has demonstrated intra-chromosomal variation consistent with non-random distribution of CNV. Multiple factors participate in the formation of CNV. Characteristics of genomic sequence predispose DNA to breakage, instigation of DNA repair and have a role in mediation of cell cycle processes (1-5). DNA breakage may be induced by exogenous or endogenous factors and can be reflected in the chromosomal distribution and frequency of CNV. The CNV on chromosome 18 were investigated for predictors of the derivation pathways. The aim is to provide insight into the contribution of the pathways involved in the formation of benign and polymorphic CNV and their role in the distribution patterns of CNV on chromosome 18. It is proposed that the non-allelic homologous recombination (NAHR) mechanism, known to be involved in the formation of recurrent and pathogenic CNV, is not the predominant pathway for the formation of CNV on chromosome 18 in this "normal" cohort. The findings illustrate the presence of repeat sequences within the intervening sequence of all CNVR. The sequence signature of selected deletions infer that replication errors or mechanisms such as non homologous end joining (NHEJ) and microhomology mediated break induced repair (MMBIR) are the major contributors of CNV on chromosome 18. The sequence signatures of CNVR characterised at the base level demonstrate the role of the genomic architecture in CNV formation and chromosomal landscape.

## 4.2 Introduction

Genomic architecture plays a significant role in the distribution of CNV (3, 6-9). Early microarray platforms were of low resolution and reported low accuracy of breakpoint estimation (9-13). Subsequent studies using more refined platforms and CNV detection algorithms improved breakpoint estimation (9, 12-16). However the initial focus was on CNV discovery and few studies investigated CNV to the base pair level. Sequence determination flanking the breakpoints and within the intervening sequence provides information about the mutational mechanisms leading to the derivation of CNV. Interrogation of these regions at the sequence level has shown that there are signatures reminiscent of the mechanism of derivation. This information has shown heterogeneity in the derivation of CNV with not one mechanism being responsible for CNV formation (1-8, 17-22).

Hotspots of CNV incidence were noted in early studies of case cohorts and found to occur in regions of segmental duplication (SD) (7, 17, 22-24). Recurrent CNV have defined breakpoints and are mediated by blocks of low copy repeat (LCR) sequences. These LCRs share a high level (>97%) of sequence homology and have the potential to mediate non allelic homologous recombination (NAHR) (5, 9, 17, 22). This mechanism accounts for the formation of deletion and duplication of DNA segments and is well described in pathogenic cohorts including deletion 22q11.2q11.2 (DiGeorge syndrome), 17p11.1p11.2 (Smith-Magenis syndrome) and 15q12q13 (Prader-Willi/Angelman syndromes) (25, 26). Likewise, the high copy homologous sequences of short interspersed repeats (SINES) and long

interspersed repeat (LINES) mediate NAHR repair mechanisms (6, 7, 22) leading to the consideration of NAHR as the prime mechanism of CNV formation.

Increasing platform resolution resulted in the detection of smaller CNV and a comparatively low proportion of CNV was found in segmental duplication and regions of extensive homology (9, 18, 27). Non-recurrent CNV, accounting for a large proportion of disease causing CNV, rare CNV and polymorphic CNV, have non-specific breakpoints and are generated by mechanisms indicative of mitotic repair such as non-homologous end joining (NHEJ) (9, 24, 27).

The genomic sequence provides the framework that impacts CNV distribution in a number of ways. DNA conformation, enrichment of tandem repeats and formation of non-B DNA structures provide the substrate for genomic instability leading to the formation of double strand breaks (2, 3, 7-9, 28). The sequence signature at CNV breakpoint junctions provides a clue to the mutational mechanism. This may be characterised by the presence of microhomology, sequence motif or insertions adjacent to the breakpoint (4, 5, 7, 29).

More recently Koren et al. 2012 using sequencing data generated from the 1000 Genome Project investigated the correlation of CNV distribution with replication timing (30). The authors concluded that NAHR mechanisms are enriched in regions of the genome that replicate early in the S phase whereas non homologous mechanisms such as NHEJ occur in regions that are late replicating.

The relative contributions of meiotic recombination and mitotic double strand break (DSB) repair mechanisms and replication errors to the derivation of pathogenic and polymorphic CNV is yet to be fully determined (7, 9, 24, 28). Whilst

large bodies of work have described NAHR mechanism few studies have reviewed the contribution of replication based mechanisms to CNV derivation (31).

To gain more insight into the mechanisms contributing to the formation and distribution of polymorphic and rare benign CNV in a cohort of the general population, the flanking and intervening sequence of all CNVR on chromosome 18 are investigated for DNA properties. In addition the sequence properties of two common CNVR, described in chapter 3, are investigated and the hypothesis presented for the putative break and repair mechanisms.

## 4.3 Materials and Methods

### 4.3.1 Molecular confirmation

Two CNVR (18q12.3 CNV18020:38308078–38311652 and 18q21.2 CNV18027: 49390404-49391772)(NCBI36/hg18) were selected for molecular confirmation, breakpoint estimation and sequence analysis. These CNVR were detected in 16 and 8 individuals respectively. The same breakpoint was predicted by microarray analysis for CNV18027 whereas there was variation in breakpoint predictions between the individuals for CNV18020.

### 4.3.2 Sequence determination

Capillary sequencing was performed for CNV18020 and CNV18027 using the optimised forward and reverse primer pair described previously. The PCR

conditions were 1 cycle of 94˚C for 90s and 40 cycles of 94˚C for 30s, 57.5˚C for 30s and 72˚C for 2.0 mins using 20ng of template. The sequence was then compared to the reference genome to determine the breakpoint (http://genome.ucsc.edu; NCBI36/hg18; accessed 18th October, 2011).

### 4.3.3 Website investigations

All CNV on chromosome 18 were analysed using numerous freely available computational tools. The UCSC Browser (http://genome.ucsc.edu/ hg18) was used to analyse CNV for chromosome band location and repetitive elements. To confirm the origin of the sequence and identify breakpoint insertions the sequence obtained from the PCR product was checked against the reference genome using UCSC Blat (http://genome.ucsc.edu/ hg18).

The potential to form secondary DNA structures was investigated using Mfold (http://www.bioinfo.rpi.edu/applications/mfold)(32). A 120bp sequence centred on the breakpoints was entered into the program and the putative secondary structures checked for the breakpoint location.

RepeatMasker v3.3.0 (www.repeatmasker) was used to calculate the proportion and type of repeat sequences. The sequences flanking the defined breakpoints were entered into repeat masker and “repeats in lower” case selected. The relative contribution of repetitive elements was recorded and the location of the breakpoint confirmed in the masked file sequence.

## 4.4 Results

To ascertain the correlation of the genomic sequence and mutational mechanisms of CNV on chromosome 18, the surrounding and intervening sequence was investigated for DNA elements that may provide evidence of the mechanism of derivation. Microhomology, inserted sequences, repeat sequences and the formation of secondary structures at breakpoint junctions have been reported previously as predictors of the mechanism of formation of CNV (4, 5, 7, 9, 17, 28, 33, 34). In order to assess the role of these features in the derivation of polymorphic and rare benign CNV, DNA properties were investigated for the 300 bp region centred on the putative breakpoints and the intervening sequence. In addition the precise breakpoints and sequence properties were determined for two common CNVR on chromosome 18. The specific breakpoints were identified by sequence analysis for CNV18020 and CNV18027. Sequence data were obtained on four samples for CNV18020 and two samples for CNV18027 and compared to the reference sequence using the UCSC Genome Browser (NCBI36/hg18).

### 4.4.1 The role of sequence properties in CNV formation

#### 4.4.1.1 Repetitive sequences within the CNVR

Copy number variation has previously been reported to be prevalent in regions of segmental duplication and repetitive elements (9, 17, 21, 23, 33, 35). The distribution of CNV on chromosome 18 was evaluated to determine if there is correlation with genome sequence signatures. Assessment of segmental

duplication was performed using UCSC Genome Browser (NCBI36/hg18). The results reported here show that 3.8% (2/52) of CNVR on chromosome 18 overlaps a region of segmental duplication with 90-98% similarity. RepeatMasker v3.3.0 (www.repeatmasker.org, accessed 20th September, 2011) was used to identify the types and density of repetitive elements within CNV and within 300 bp flanking the putative breakpoints. With the exception of 5 CNVR, repeat sequences were present within all CNVs at a density range of 7%-98.15% with a median density of 40%. The average density for short interspersed nuclear elements (SINES) is 7.39%, long interspersed nuclear elements (LINES) (17.89%) and long terminal repeats (LTRs) (12.16%) with simple and low complexity repeats measuring a density of 1-2% within the CNVs (Table 1). There was a single (low frequency) loss that showed no evidence of repeat sequence within the CNV however, an *Alu* sequence is located 200bp distal of the end breakpoint. The remaining 4 had simple repeat sequences and SINE/LTR within the flanking region.

The density of repetitive sequences for common, low frequency and rare variants was compared to explore association of CNV incidence with genomic sequence. There was no apparent difference between the total proportion of repeat sequences within common and private variants scoring 38.5 and 38.2% respectively (Table 2). Further examination shows that common CNV occur in regions of LINE sequences, whereas low frequency and private variants are associated with LINE and LTR sequences.

The GC content of DNA was scored within the CNV and no difference was observed for common and low frequency (39.5 and 39.1% respectively), although elevated for private variants (45.1%).

**Table 2**. Examination of repeat sequences in CNV of varying incidence

| | % Ave. Repeat | % Median Repeat | % Ave. GC | % SINE | % LINE | % LTR | % DNAE | % SIMPLE | % Low Complexity |
|---|---|---|---|---|---|---|---|---|---|
| **Common** | 38.5 | 28.4 | 39.5 | 6.2 | 20.5 | 7.7 | 1.8 | 0.8 | 0.6 |
| **Low Frequency** | 52.9 | 50.7 | 39.1 | 9.5 | 18.6 | 18.5 | 2.2 | 3.5 | 0.6 |
| **Rare/Private** | 38.2 | 38.3 | 45.1 | 6.9 | 13.5 | 11.7 | 2.7 | 1.2 | 2.0 |

## 4.4.1.2 Association of breakpoints with repetitive sequence

Investigation of the breakpoints showed 26.5% of CNV on chromosome 18 with both putative breakpoints within repetitive sequences, 30% with one breakpoint within a repetitive sequence and 12.5% within the 300bp flanking region of a start or end position. Furthermore, 31% (126/399) of CNV calls did not have putative breakpoints within repeat sequences.

**Table 1.** Repeat sequences within the intervening sequences of CNV on chromosome 18

| Start | End | CNVR | Incidence | CN State | GC% | RPT% | % SINE | % LINE | % LTR | % DNA E | % SIMPLE | % Low Complexity |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1898848 | 1964966 | 18001 | 1 | 1 | 38.6 | 57.4 | 4 | 25.9 | 22 | 0 | 0 | 0 |
| 4167454 | 4295905 | 18002 | 1 | 1 | 37 | 38.47 | 7.86 | 24.22 | 4.29 | 2.11 | 0 | 0 |
| 4522500 | 4526648 | 18003 | 15 | 1,0 | 35.77 | 46.07 | 0 | 0 | 41.97 | 0 | 0 | 4.1 |
| 5199569 | 5200656 | 18004 | 2 | 0 | 37.5 | 51.19 | 0 | 0 | 45.04 | 0 | 3.31 | 2.85 |
| 5313289 | 5316221 | 18005 | 33 | 0 | 44 | 97.88 | 0 | 97.88 | 0 | 0 | 0 | 0 |
| 7101033 | 7648679 | 18006 | 1 | 3 | 42.98 | 47.47 | 18.15 | 16.6 | 8.32 | 4.33 | 0 | 0 |
| 7722141 | 7751056 | 18007 | 1 | 1 | 40.91 | 37.86 | 13.36 | 6.67 | 11.32 | 6.51 | 0.44 | 0.35 |
| 11498904 | 11500794 | 18008 | 2 | 0 | 37.23 | 98.15 | 16.5 | 81.65 | 0 | 0 | 0 | 0 |
| 11500046 | 11500900 | 18009 | 3 | 0 | 38.13 | 93.68 | 7.95 | 85.73 | 0 | 0 | 0 | 0 |
| 13133415 | 13134101 | 18010 | 2 | 1 | 42.65 | 91.27 | 10.63 | 0 | 80.64 | 0 | 0 | 0 |
| 14270949 | 14285231 | 18011 | 4 | 0 | 46.33 | 93.43 | 0 | 0 | 93.43 | 0 | 0 | 0 |
| 22018233 | 22019611 | 18012 | 1 | 1 | 41.33 | 57.43 | 43.87 | 0 | 13.56 | 0 | 0 | 0 |
| 22825242 | 22826188 | 18013 | 11 | 0 | 42.87 | 81.94 | 0 | 81.94 | 0 | 0 | 0 | 0 |
| 25743059 | 25744733 | 18014 | 1 | 1 | 29.43 | 91.94 | 0 | 85.61 | 0 | 0 | 6.33 | 0 |
| 26068846 | 26069966 | 18015 | 1 | 1 | 39.52 | 65.12 | 0 | 0 | 65.12 | 0 | 0 | 0 |
| 28142501 | 28143201 | 18016 | 3 | 1,0 | 35.91 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 33555925 | 33560607 | 18017 | 30 | 0 | 47.88 | 10.46 | 4.74 | 3.82 | 0 | 0 | 0.92 | 0.98 |
| 36514151 | 36520673 | 18018 | 2 | 0 | 33.25 | 60.69 | 5.95 | 40.38 | 13.44 | 0 | 0 | 0 |
| 37116145 | 37127830 | 18019 | 39 | 0,4 | 36.87 | 95.08 | 11.92 | 69.05 | 14.11 | 0 | 0 | 0 |
| 38308102 | 38311523 | 18020 | 8 | 1,0 | 39.51 | 23.09 | 6.17 | 0 | 13.68 | 0 | 2.54 | 0.7 |
| 40019676 | 40030578 | 18021 | 1 | 1 | 33.03 | 35.14 | 2.21 | 18.12 | 14.23 | 0.58 | 0 | 0 |
| 40229642 | 40235957 | 18022 | 10 | 1 | 33.84 | 17.91 | 5.79 | 8.95 | 0 | 1.28 | 0.47 | 1.42 |
| 41090885 | 41092686 | 18023 | 1 | 0 | 38.62 | 20.37 | 11.76 | 0 | 0 | 8.6 | 0 | 0 |
| 41091870 | 41100304 | 18024 | 2 | 0 | 41 | 25.95 | 25.95 | 0 | 0 | 0 | 0 | 0 |
| 45943238 | 45952266 | 18025 | 2 | 0 | 42.45 | 44.22 | 15.54 | 21 | 6 | 0 | 0 | 0 |
| 48716550 | 48716914 | 18026 | 7 | 0 | 33.42 | 13.97 | 0 | 13.97 | 0 | 0 | 0 | 0 |
| 49390524 | 49391736 | 18027 | 6 | 1 | 29.35 | 77.58 | 0 | 75.93 | 0 | 0 | 1.65 | 0 |
| 49450390 | 49464477 | 18028 | 2 | 1,0 | 41.99 | 39.8 | 17.24 | 0 | 14.86 | 6.03 | 0 | 1.67 |
| 50542568 | 50545111 | 18029 | 6 | 1 | 37.85 | 46.97 | 17.49 | 14.15 | 14.11 | 0 | 1.22 | 0 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 52081026 | 52081998 | 18030 | 2 | 1 | 29.5 | 69.99 | 0 | 16.14 | 0 | 0 | 51.18 | 2.67 |
| 53090436 | 53099515 | 18031 | 35 | 0 | 42.98 | 20.45 | 3.13 | 4.47 | 11.62 | 0.34 | 0.89 | 0 |
| 54082151 | 54088089 | 18032 | 2 | 1 | 40.41 | 21.37 | 6.08 | 5.42 | 9.18 | 0.69 | 0 | 0 |
| 56819912 | 56823680 | 18033 | 2 | 1 | 36.67 | 55.05 | 0 | 1.49 | 26.96 | 26.61 | 0 | 0 |
| 61352708 | 61357481 | 18034 | 4 | 1,3 | 36.41 | 50.17 | 12.86 | 34.69 | 1.74 | 0 | 0.88 | 0 |
| 61875651 | 61882719 | 18035 | 15 | 1,0 | 35.11 | 32.28 | 0 | 31.38 | 0 | 0 | 0.57 | 0.34 |
| 61912073 | 61920046 | 18036 | 13 | 0,3 | 36.06 | 7.27 | 7.27 | 0 | 0 | 0 | 0 | 0 |
| 62058501 | 62062587 | 18037 | 6 | 1,0 | 37.12 | 21.8 | 5.92 | 0 | 14.61 | 0 | 1.27 | 0 |
| 62370755 | 62371870 | 18038 | 1 | 0 | 35.93 | 18.01 | 0 | 0 | 0 | 18.01 | 0 | 0 |
| 62928255 | 62930313 | 18039 | 11 | 1,0 | 32.88 | 20.59 | 14.18 | 6.41 | 0 | 0 | 0 | 0 |
| 64379881 | 64380304 | 18040 | 11 | 0 | 42.22 | 50.94 | 0 | 0 | 50.94 | 0 | 0 | 0 |
| 64895288 | 64907327 | 18041 | 7 | 1 | 32.08 | 28.36 | 5.18 | 20.62 | 0 | 0 | 0.43 | 1.65 |
| 65359701 | 65362926 | 18042 | 4 | 1 | 38.04 | 39.9 | 24.21 | 6.96 | 5.03 | 1.19 | 0.85 | 1.66 |
| 66104701 | 66106210 | 18043 | 9 | 1,0 | 41.52 | 50.26 | 47.42 | 2.85 | 0 | 0 | 0 | 0 |
| 67579536 | 67590954 | 18044 | 1 | 1 | 37.51 | 63.26 | 3 | 24.65 | 36.62 | 0 | 0 | 0 |
| 72234235 | 72235724 | 18045 | 1 | 1 | 71.88 | 40.47 | 0 | 0 | 0 | 0 | 11.54 | 28.93 |
| 73386134 | 73396956 | 18046 | 24 | 0 | 32.8 | 47.53 | 0 | 0 | 0 | 35.81 | 7.53 | 4.19 |
| 73510959 | 73512121 | 18047 | 1 | 1 | 63.48 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 73810158 | 73810801 | 18048 | 1 | 1 | 62.9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 74763474 | 74767073 | 18049 | 5 | 1 | 48.53 | 12.47 | 8.33 | 4.14 | 0 | 0 | 0 | 0 |
| 74874961 | 74880140 | 18050 | 10 | 1 | 49.73 | 17.78 | 0 | 0 | 0 | 0 | 0 | 0 |
| 75411150 | 75413066 | 18051 | 35 | 0 | 66.14 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 75746951 | 75747880 | 18052 | 1 | 1 | 63.52 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

## 4.4.2 Sequence analysis

### 4.4.2.1 CNV 18020 18q12.3

Capillary sequencing of the optimised PCR product is described in Chapter 3. The output sequence for each sample was entered into UCSC BLAT and the origin confirmed to correspond to 18q12.3. The precise breakpoint was the same for each sample tested and determined to be chr18:38308078 and 38311652, confirming a 3575bp deletion. A 2bp microhomology for the bases, GG, is present at start and end breakpoints and no evidence of insertion is detected using UCSC BLAT tools (Table 3).

The sequence flanking and within the intervening sequence was analysed for repetitive DNA using RepeatMasker v3.3.0. The start breakpoint of CNV18.020 (38308078) is located within an LTR/ERV1 sequence measuring 142bp and repetitive sequences were not detected at the distal breakpoint (38311652) (www.repeatmasker). Within the intervening sequence LTR accounts for 17.4%, SINE/Alu 211bp (5.51%), LINE 108bp (2.83%), low complexity 210bp (4.98%) and simple tandem repeats $(GT)_{23}$ and $(A)_{17}$ account for 2%. In addition repeats $(GT)_{14}$ are located 280bp proximal to the distal breakpoint (Figure 1). The GC content for the CNV is 39%.

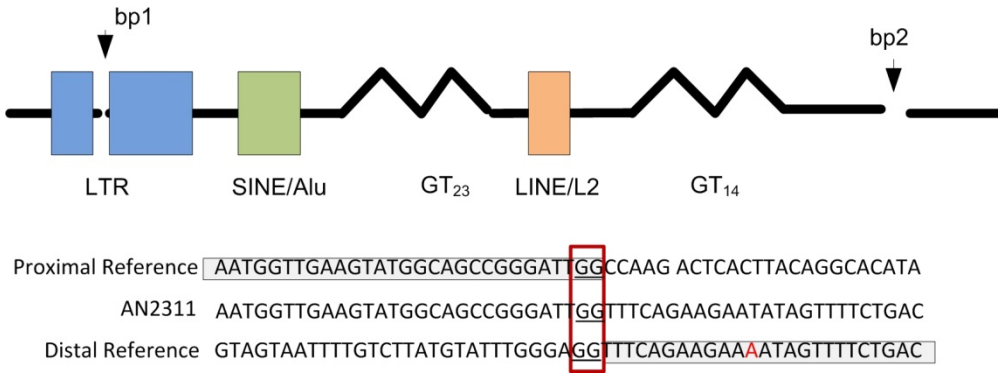



**Figure 1.** The respective location of repeat sequences and potential non-B conformations in relation to the breakpoints is shown for the 3575bp deletion, CNV18020. The 2bp microhomology at the breakpoint junction is indicated in the box. The zigzag line represents the potential Z DNA conformation formed by $GT_n$ repeats. A SNP (A/T) is located at position 38311664 and is annotated in red.

## 4.4.2.2 CNV18027 18q21.2

The origin of the sequence from the optimised PCR product corresponded to 18q21.2 and the precise breakpoints were determined to be 49390404 and 49391772 confirming a 1369bp deletion in the samples tested. A 2bp microhomology for the bases CT is present at both breakpoints and no evidence of insertion was detected using UCSC BLAT tools (Table 3).

The start breakpoint of CNV18027 (49390404) is located within an LINE (L1) sequence measuring 1224bp. Repetitive sequences were not detected at the distal breakpoint (49391772) (www.repeatmasker). A simple repeat sequence $(A)_{16}$ is located 113bp distal to the end breakpoint (49391772) and $(AC)_6$ is located 83bp

proximal to the start breakpoint (49390404) (Figure 2). An LTR sequence is located 250bp distal to the end breakpoint. The GC content for the CNV is 30.18%.

**Table 3.** The breakpoint sequence of two CNVR in this study.

| Sample id. | CNV id | Size | Breakpoint 1 | Breakpoint 2 |
|---|---|---|---|---|
| **AN2305** | CNV18.020 | 3575 | GCCGGGATTGG/*CCAAGACT* | *TTTGGGAGG*/TTTCAGAAGAA |
| **AN2311** | CNV18.020 | 3575 | GCCGGGATTGG/*CCAAGACT* | *TTTGGGAGG*/TTTCAGAAGAA |
| **AN2342** | CNV18.020 | 3575 | GCCGGGATTGG/*CCAAGACT* | *TTTGGGAGG*/TTTCAGAAGAA |
| **AN2358** | CNV18.020 | 3575 | GCCGGGATTGG/*CCAAGACT* | *TTTGGGAGG*/TTTCAGAAGAA |
| **AN2339** | CNV18.027 | 1369 | GACAGAAGCT/*CATTTATTTT* | *AAAAGTAACT*/ACACAAAGAA |
| **AN2345** | CNV18.027 | 1369 | GACAGAAGCT/*CATTTATTTT* | *AAAAGTAACT*/ACACAAAGAA |

The 2bp microhomology at the breakpoints of both deletions is indicated by the underline. The italicised bases represent the deleted sequence.

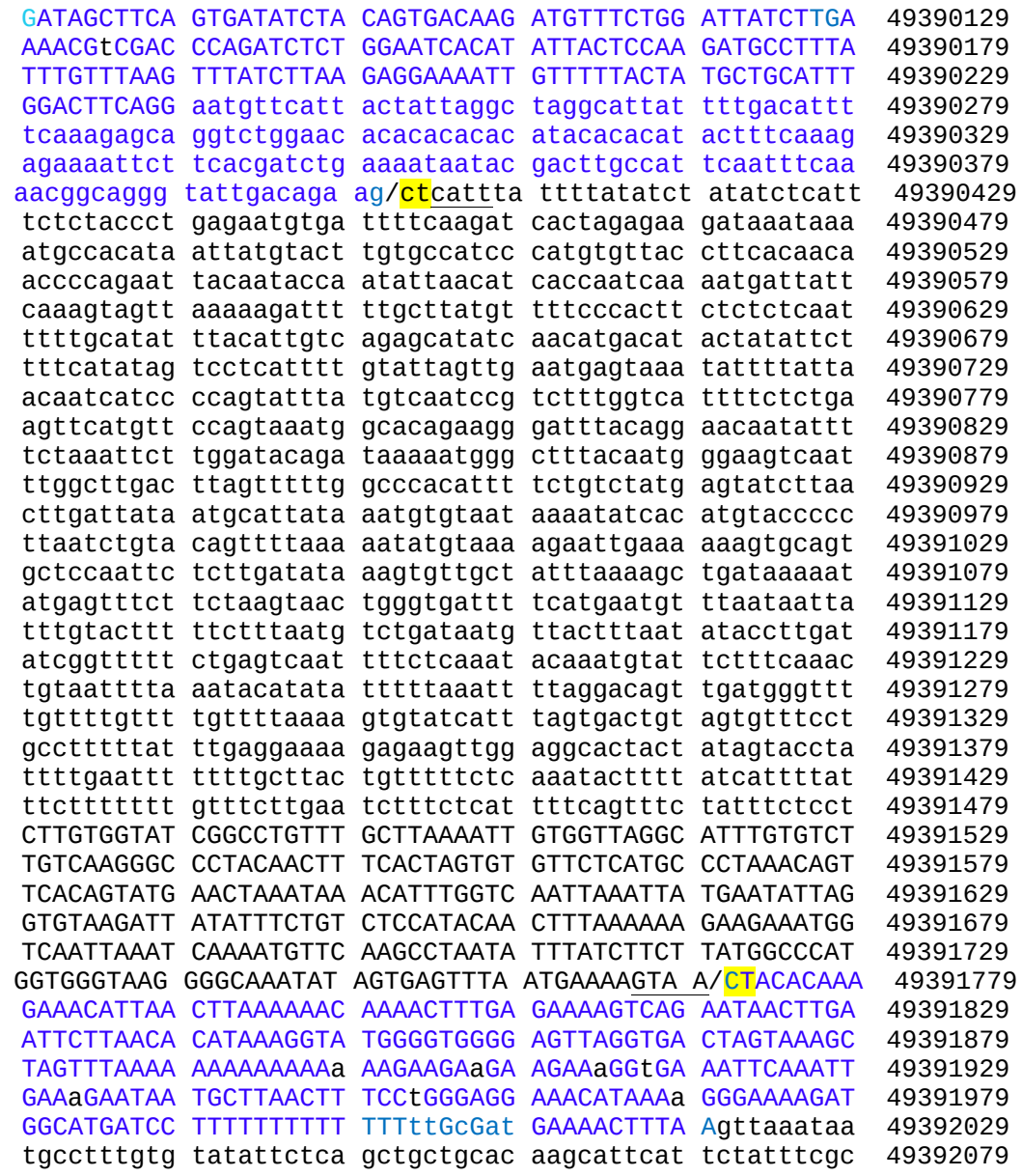



**Figure 2.** The sequence for CNV 18027 compared to the consensus sequence. The sequence for sample AN2345 is in blue with the intervening sequence deleted. The 2bp microhomology (ct) at the breakpoints is highlighted in yellow. The LINE sequence is in lower case with unique sequence in capital. The breakpoints are indicated by a slant line (/) (http://genome.ucsc.edu; NCBI36/hg18; accessed 18th October, 2011).

### 4.4.3 Formation of non B Structures

The formation of secondary structures may provide a substrate for the formation of double strand breaks (DSB) and replication errors (1-3, 36). The 120bp sequence centred at the start and end breakpoints was analysed for the potential to form non-B secondary structures using free software MFold (32). The formation of non-B structures as a mechanism for DNA instability was excluded for CNV18027. Whilst, the potential to form extended loop and hairpin loops were observed for sequences flanking both breakpoints for CNV18020. The start breakpoint is at the junction of a hairpin and the end breakpoint is at the junction of an extended loop (Figure 3).

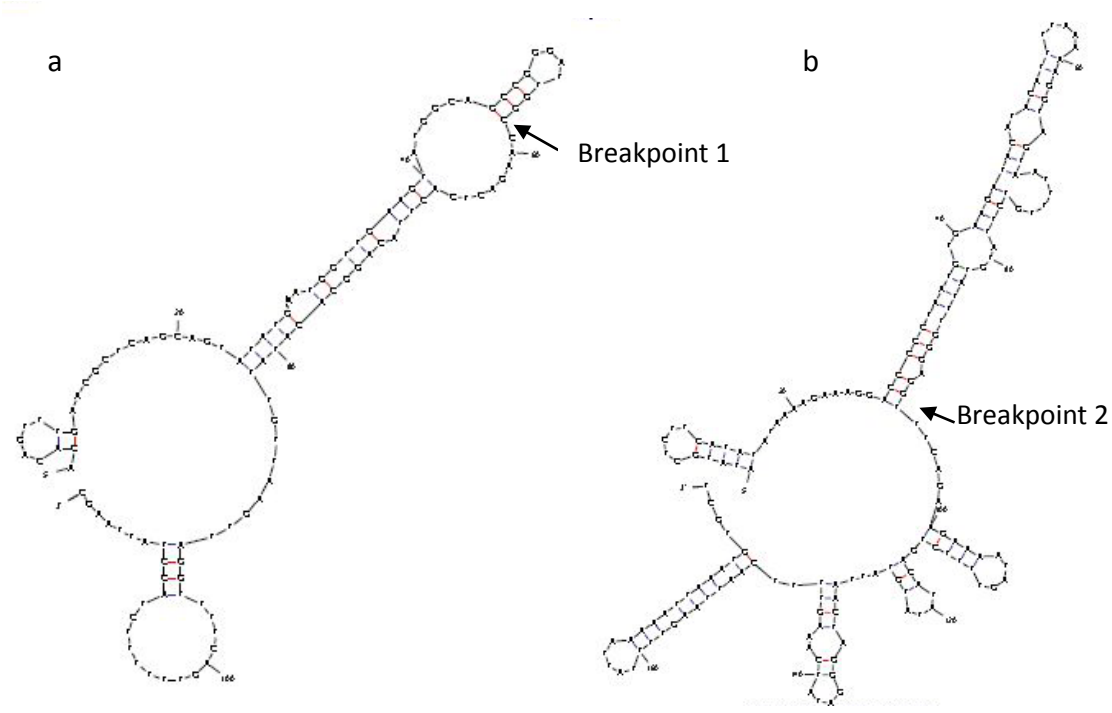



**Figure 3.** The sequence for CNV 18020 forms potential slipped structures that may provide the template for replication errors for the breakpoint at a) start position 38308078 and b) end position 38311652. (MFold version 3.5, http://mfold.rna.albany.edu/; Zucker et al. 2003, accessed 17/10/11)(32).

### 4.4.4 DNA repair and genomic instability

CNV can be formed as a consequence of repair mechanisms of single and double strand breaks in somatic cells (4, 5, 28, 37). These breaks may be induced by endogenous or exogenous stimuli. An example of this is age-related processes, with early reports indicating an accumulation of the effects of oxidative damage, one of the causes of DNA breakage, or compromised integrity of the homologous repair

(HR) mechanisms with concomitant accumulation of mutations and cytogenetically visible rearrangements (37). This hypothesis is evaluated for enrichment of CNV, a product of NHEJ and replication repair mechanisms, in the cohort of female individuals in this study. This was done by comparing the average number of CNV per individual within age categories. Somatic mosaicism was excluded by visual assessment of plots. The age range is 23 years old to 84 years old. The limitation of this comparison is the small size of the cohort and individual age categories with the smallest being 5-6 individuals for the age categories <40, 41-50 and > 80. The age categories 51-60 and 71-80 had 13 individuals in each category and the remaining 21 individuals are aged 61-70 years old. Due to the low incidence of CNV on chromosome 18 the total incidence for the autosomes is evaluated. No apparent difference in the incidence of autosomal CNV >1kb was detected among the age categories in this cohort (Figure 4). CNV in females <50 years of age were slightly lower with an average of 155 CNV per individual compared to 174 for females >50 years of age (<2 standard deviation).

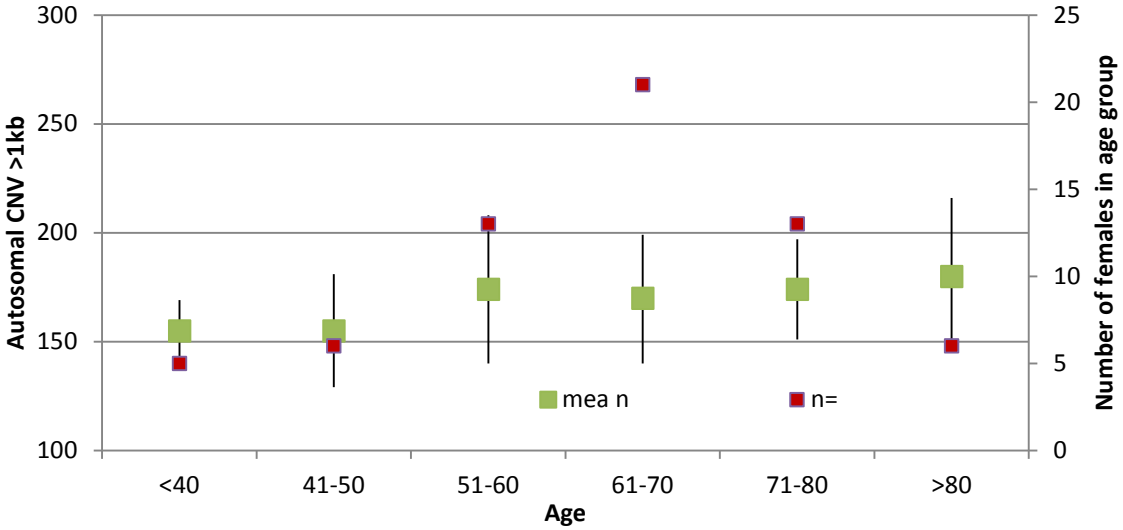



**Figure 4.** The incidence of autosomal CNV>1kb across age categories shows minimal variation. The numbers of individuals in each category is indicated by the red marker.

## 4.5 Discussion

The investigation of CNV in a female cohort of a Western European descendent population shows a pattern of intra- and inter-chromosomal variation in the distribution and incidence of CNV. Chromosome 18 showed no evidence of association with subtelomeric or pericentromeric regions and CNV enrichment correlated with gene function and low gene density in some locations on chromosome 18. However this did not explain all of the intra-chromosomal variation. To investigate the hypothesis that the chromosomal distribution of CNV is associated with the genomic architecture and cell cycle processes, all CNVR on chromosome 18 were analysed for DNA elements to gain insight to the contribution of the DNA sequence to CNV mutation mechanisms. In addition two

common CNVR on chromosome 18 were analysed at the base pair level and compared to the reference genome. Microhomology is observed for the two CNVR however the DNA composition infers different mechanisms of derivation and potential for genome instability.

## 4.5.1 Derivation of CNV on chromosome 18

The results reported here suggest that CNV activity on chromosome 18 does not appear to be enriched in regions of segmental duplication. In a study of the HapMap cohort by Cooper et al. 2007, it is reported that regions of the genome with a high density of CNV also have a high density of segmental duplication (25%) as compared to the genome-wide density of 4-5%. Itsara et al. 2009, also report on hotspots with a 25-fold genomic enrichment of CNV for flanking homologous segmental duplications. In the study by Perry et al. 2008, however, the authors report that only a minority of CNV overlap segmental duplication.

This investigation of CNV in a cohort of healthy female Western descendent population show that 3% of the CNV identified on chromosome 18 involve segmental duplications with the remainder associated with repetitive DNA such as SINES, LINES and LTR. This finding is not unexpected as chromosome 18 has the lowest genomic contribution of intra-chromosomal segmental duplication compared to other autosomes, 0.3% compared to 3.21% (38). Although this is located predominantly in the pericentromeric and subtelomeric regions the segmental duplication length (230kb) is small compared to other autosomes with

an average of 3.65Mb (38, 39). This also accounts for the low contribution of duplication to the total incidence of CNV on chromosome 18 (18).

When compared with the whole genome, chromosome 18 has overall low GC density and is not enriched in repeat sequences (38, 39), the architecture that is reported to mediate chromosomal rearrangements and provide the substrate for CNV formation by NAHR mechanisms (5, 13, 15, 22). Putative breakpoints involving both start and end potentially occurred in segmental duplication, LINE and SINE sequences in 26.5% of calls suggesting that NAHR or BIR based mechanisms may contribute to some but not all of the CNV reported here. As such the genomic architecture of chromosome 18 provides the opportunity to investigate the contribution of replication and repair mechanisms to the formation of CNV.

### 4.5.2 Heterogeneity of mechanisms involved in CNV formation

CNV formation is generated by a combination of meiotic homologous recombination, non-homologous repair mechanisms and replication errors. Determination of the contribution of mechanisms requires breakpoint analysis at the base pair level with determination of sequence signatures.

Differences in the estimated contribution of mechanisms between studies are apparent (7, 9, 19, 28, 40). Previous reports have shown that NAHR mechanisms account for 10-15% of CNV, VNTR 10-15%, retrotransposition (1%) and the remainder being NHEJ and alternate NHEJ mechanisms such as MMEJ, FoSTeS,

MMBIR mechanisms (7). Similarly, Mills et al. 2011 in the study of HapMap cell lines in the 1000 Genomes Project reported a predominance of non homologous mechanisms and this being independent of CNV size (40). Whilst CNV derived from NAHR predominated in recombination "hotspot" regions. In contrast, in a study of CNV >7kb, Kidd et al. 2008 reported relative contributions of NAHR(38%), NHEJ (39%) and retrotransposition and VNTRs account for 22% (19). These differences may be due to platform resolution, marker position and CNV detection algorithm. The sample type may also influence the estimation of the contribution of meiotic and mitotic mechanisms with many studies investigating CNV in transformed cell lines (15, 41).

The high resolution platform used here provides the opportunity to investigate putative breakpoints for large blocks of homologous segments associated with the signatures of NAHR. Non homologous mechanisms can be inferred however the distinction between NHEJ, MMEJ, MMBIR, FoSTeS and other replication mechanisms requires base pair level resolution.

In the study reported here the putative breakpoints and 300bp segments centred on the breakpoints of 52 CNVR on chromosome 18 were analysed for repetitive DNA. It is of interest that 98.9% (395/399) of CNV on chromosome 18 are <15kb in size, strongly suggesting an association with non homologous or replication repair mechanisms for the formation of CNV on chromosome 18. The paucity of CNV >100kb on chromosome 18 is consistent with the low proportion of segmental duplication and low copy repeat sequences that are known to mediate NAHR rearrangements (39).

CNV with both the start and end breakpoints within segmental duplication, LINE or SINE repetitive DNA, represents 26.5% of CNV. In this category there is a predominance of common CNV of the length range of 1kb-15kb. These CNV have the basis for NAHR and BIR mechanisms. There was no evidence of repetitive DNA observed at the breakpoints or within 300bp of the breakpoints of 31% of CNV. These CNV range in length from 1kb to 10kb and represent common, low frequency and private variants and are suggestive of NHEJ or replication based mechanisms. The remaining 42.5% of CNVR accounting for common, low frequency and private variants up to 550kb in length have repeat sequences located at one of the breakpoints or within the flanking 300bp region. These are non-recurrent CNV suggestive of non homologous mechanisms of derivation, however further characterisation of breakpoints is needed to exclude NAHR and BIR mechanisms.

## 4.5.3 Derivation of CNV 18020

The sequence properties of two copy number regions are presented providing an insight into the derivation mechanisms in these CNVR.

The potential formation of non-B DNA was observed in the region flanking both breakpoints of CNV18020. A simple triplet repeat (CCG) present at the start breakpoint can base pair with itself to form a hairpin conformation (Figure 3). The end breakpoint potentially forms an extended loop conformation and since there is no repeat DNA sequence detected, this is formed by non-palindromic pairing. In addition two segments of purine-pyrimidine-repeats, $(GT)_{14}$ and $(GT)_{23}$ are located within the deleted sequence and can assume a left handed Z conformation and

'zigzag' configuration of the DNA backbone (36). This would potentially be recognized as damaged DNA and initiate the "structure induced" repair process, with resection of the DNA between the loop structures. The combination of potential non-B DNA structures at the breakpoints and at least two non-B sequence motifs within the deleted sequence suggests that this DNA segment is predisposed to instability and the deletion is likely to be formed by "repair induced" model involving cleavage and repair. The location of both breakpoints at the junction of the non B-DNA structures suggests a model for the formation of breaks at both junctions resulting in deletion of the intervening sequence and repair mechanisms such as MMEJ or FoSTes between the 2bp microhomologous sequences. In FoSTes the non-B structures may interfere with the replication fork progress, the 3' end of ssDNA in the fork invades another replication fork that shares microhomology resulting in a deletion when in the forward direction (5). There is no evidence of insertion at or near the breakpoint junction suggesting a single template switch.

### 4.5.4 Derivation of CNV 18027

The proximal breakpoint for CNV 18027 is located within a 1224bp LINE sequence. The end breakpoint is in unique sequence but is located 250bp proximal to a 358bp LTR sequence (Figure 2) and a 2bp microhomology was observed at the breakpoint junction. This finding is not suggestive of NAHR mechanisms. Investigation of a 300bp sequence centred on the breakpoints also revealed no evidence for the formation of potential non-B structures at the breakpoints or within the intervening sequence. The finding of 2bp of microhomology (CT) at the

breakpoint junction and the absence of complexity or inserted sequences, the deletion is strongly suggestive of NHEJ mechanisms or FoSTeS in a single replication fork. There was no evidence of non-B DNA sequence that may cause fork stalling, suggesting an alternative reason for reduced fork velocity. Breakpoint analysis uncovered complimentary sequences at the start and end breakpoints. An alternative hypothesis is single strand annealing (SSA) whereby resection of sequences on the 5'ends of a DSB until the homologous and the complimentary single strand sequence is reached. The sequence then anneals and ligation occurs resulting in deletion of the intervening sequence and one set of repeats (5). This is consistent with the observed sequence pattern in CNV18027. This mechanism has been described between identical *Alu* repeats but has not been reported for short (4bp) complimentary sequence or microhomology (5).

### 4.5.5 Genomic instability, repair mechanism and inter-individual variation

The predominance of CNV <15kb on chromosome 18 may be indicative of genomic instability and repair processes. CNV can be formed during repair mechanisms of DNA breakage incurred by replication stress (28) and oxidative damage (37) in the germ line and in somatic cells (5). The effects of genomic instability are shown to be accumulative over time and may be reflected in the aging process (37) and age-related diseases such as malignancy (4).

The cohort in this study represents the age range 23-84 years old providing the opportunity to compare the relative load and length of CNV across age categories.

Due to the low incidence of CNV on chromosome 18 the incidence of autosomal CNV >1kb was assessed. The CNV load was analysed to test the hypothesis that somatic CNV accumulates with age (5). The results indicate that there is no evidence of enrichment of CNV with increasing age. This finding is consistent with Keller et al. 2013, and reflects apparently stable levels of CNV load indicative of germ-line CNV formation (42). It also implies that no bias was introduced to the evaluation of CNV load and CNV derivation mechanisms by the age of the females in the study. However, this interpretation is limited by the size of the cohort and the number of individuals in each age category. In addition, the CNV represents the incidence at the time of collection and it does not account for changes in the incidence with the aging process or investigate for tissue specific variation. Further studies are required to investigate the associations of age with drivers of genomic instability and CNV derivation mechanisms.

## 4.5.6 Distribution of CNV on chromosome 18

The distribution of CNV is influenced in numerous ways, by genomic instability, the presence of repetitive DNA and response to cell cycle processes such as replication timing (1, 6, 23, 28, 30, 31). Repeat sequences have important roles in CNV mutagenesis by providing the intra or inter-chromosomal homology to mediate rearrangement but also in genomic instability leading to replication errors (1-4, 8, 22). The lack of CNV enrichment with chromosomal morphological features on chromosome 18 such as the subtelomeres and pericentromeric regions is explained in part by the relatively low proportion and small size of segmental duplication (38, 39). Similarly the scarcity of large CNV that are mediated by blocks

of LCR in NAHR mechanisms is consistent with the genomic architecture on chromosome 18.

The sequences of the two common CNVR investigated at the base pair level and demonstration of microhomology at the breakpoint junctions of both CNVR is consistent with NH or replication mechanisms, such as FoSTeS. The absence of insertions at the breakpoint junctions is analogous with a single template switch. Sequence properties such as sequence motifs, non-B DNA sequences and direct or inverted repeats provide the substrate for DNA breakage. This is observed in one CNVR (CNV18020), demonstrating the roles of repeat sequences in genomic instability and subsequent association with chromosome distribution.

## 4.6 Conclusion

Genomic architecture has been shown to provide the substrate for CNV formation however, the molecular basis for CNV derivation is yet to be defined. Segmental duplication is not a contributing factor in CNV distribution for chromosome 18. In contrast, repetitive DNA sequences are present within the intervening sequence of most of the CNV reported here and may contribute to the distribution of CNV on chromosome 18.

It is apparent that multiple mechanisms are involved in the aetiology of CNV. It is proposed that the predominant mechanism of CNV on chromosome 18 in a cohort of healthy females is replication and repair based mechanisms. Repair of single and double strand DNA breaks formed by either exogenous or endogenous factors culminating in CNV formation during the repair process. This accounts for the

predominance of small CNV, 98.9% less than 15kb, and low contribution of duplication. Although, the possibility that homologous mechanism such as NAHR requiring small segments of homology (<200kb) and microhomology, although not previously described, cannot be excluded.

This study has contributed to an insight into the contribution of replication based mechanism to the derivation of CNV. However more needs to be understood about the endogenous and exogenous causes for the formation of double and single strand breaks and regulation of repair processes as the distribution of CNV suggests that these are non-random events.

## 4.7 Data Access

The CNV data presented herein is published at http://www.ebi.ac.uk/dgva/data-download: Accession number estd198

## 4.8 References

1. Bacolla A, Cooper DN, Vasquez KM. DNA structure matters. Genome Med.2013 5(6):51.
2. Bacolla A, Jaworski A, Larson JE, Jakupciak JP, Chuzhanova N, Abeysinghe SS, et al. Breakpoints of gross deletions coincide with non-B DNA conformations. Proc Natl Acad Sci U S A. 2004;101(39):14162-7.
3. Bacolla A, Wells RD. Non-B DNA conformations, genomic rearrangements, and human disease. J Biol Chem. 2004;279(46):47411-4.
4. Hastings PJ, Ira G, Lupski JR. A microhomology-mediated break-induced replication model for the origin of human copy number variation. PLoS Genet. 2009;5(1):e1000327.
5. Hastings PJ, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. Nat Rev Genet. 2009;10(8):551-64.
6. Chen L, Zhou W, Zhang L, Zhang F. Genome architecture and its roles in human copy number variation. Genomics Inform. 2014;12(4):136-44.
7. Conrad DF, Bird C, Blackburne B, Lindsay S, Mamanova L, Lee C, et al. Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. Nat Genet. 2010; 42(5):385-91.
8. Bose P, Hermetz KE, Conneely KN, Rudd MK. Tandem repeats and G-rich sequences are enriched at human CNV breakpoints. PLoS One. 2014; 9(7):e101607.
9. Perry GH, Ben-Dor A, Tsalenko A, Sampas N, Rodriguez-Revenga L, Tran CW, et al. The fine-scale and complex architecture of human copy-number variation. Am J Hum Genet. 2008;82(3):685-95.
10. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, et al. Large-scale copy number polymorphism in the human genome. Science. 2004;305(5683):525-8.
11. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, et al. Detection of large-scale variation in the human genome. Nat Genet. 2004;36(9):949-51.
12. Park H, Kim JI, Ju YS, Gokcumen O, Mills RE, Kim S, et al. Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. Nat Genet. 2010 May;42(5):400-5.
13. Matsuzaki H, Wang PH, Hu J, Rava R, Fu GK. High resolution discovery and confirmation of copy number variants in 90 Yoruba Nigerians. Genome Biol. 2009; 10(11):R125.
14. Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467(7311):52-8.
15. Pinto D, Marshall C, Feuk L, Scherer SW. Copy-number variation in control population cohorts. Hum Mol Genet. 2007;16 Spec No. 2:R168-73.
16. Conrad DF, Hurles ME. The population genetics of structural variation. Nat Genet. 2007;39(7 Suppl):S30-6.
17. Cooper GM, Nickerson DA, Eichler EE. Mutational and selective effects on copy-number variants in the human genome. Nat Genet. 2007;39(7 Suppl):S22-9.
18. Chia NL, Bryce M, Hickman PE, Potter JM, Glasgow N, Koerbin G, et al. High-resolution SNP microarray investigation of copy number variations on chromosome 18 in a control cohort. Cytogenet Genome Res. 2013; 141:16-25.
19. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, et al. Mapping and sequencing of structural variation from eight human genomes. Nature. 2008;453(7191):56-64.
20. Sharma S. Non-B DNA Secondary Structures and Their Resolution by RecQ Helicases. J Nucleic Acids. 2011; 1-15.

21. Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, et al. Segmental duplications and copy-number variation in the human genome. Am J Hum Genet. 2005;77(1):78-88.
22. Startek M, Szafranski P, Gambin T, Campbell IM, Hixson P, Shaw CA, et al. Genome-wide analyses of LINE-LINE-mediated nonallelic homologous recombination. Nucleic Acids Res. 2015;43(4):2188-98.
23. Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, et al. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet. 2009;41(10):1061-7.
24. Vissers LE, Bhatt SS, Janssen IM, Xia Z, Lalani SR, Pfundt R, et al. Rare pathogenic microdeletions and tandem duplications are microhomology-mediated and stimulated by local genomic architecture. Hum Mol Genet. 2009;18(19):3579-93.
25. Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, et al. A copy number variation morbidity map of developmental delay. Nat Genet. 2011;43(9):838-46.
26. Kaminsky EB, Kaul V, Paschall J, Church DM, Bunke B, Kunig D, et al. An evidence-based approach to establish the functional and clinical significance of copy number variants in intellectual and developmental disabilities. Genet Med. 2011;13(9):777-84.
27. van Binsbergen E. Origins and breakpoint analyses of copy number variations: up close and personal. Cytogenet Genome Res. 2011; 135(3-4):271-6.
28. Arlt MF, Mulle JG, Schaibley VM, Ragland RL, Durkin SG, Warren ST, et al. Replication stress induces genome-wide copy number changes in human cells that resemble polymorphic and pathogenic variants. Am J Hum Genet. 2009;84(3):339-50.
29. McCarroll SA. Copy number variation and human genome maps. Nat Genet. 2010;42(5):365-6.
30. Koren A, Polak P, Nemesh J, Michaelson JJ, Sebat J, Sunyaev SR, et al. Differential relationship of DNA replication timing to different forms of human mutation and variation. Am J Hum Genet. 2012;91(6):1033-40.
31. Chen L, Zhou W, Zhang C, Lupski JR, Jin L, Zhang F. CNV instability associated with DNA replication dynamics: evidence for replicative mechanisms in CNV mutagenesis. Hum Mol Genet. 2015;24(6):1574-83.
32. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 2003;31(13):3406-15.
33. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, et al. Origins and functional impact of copy number variation in the human genome. Nature. 2010;464(7289):704-12.
34. Abeysinghe SS, Chuzhanova N, Krawczak M, Ball EV, Cooper DN. Translocation and gross deletion breakpoints in human inherited disease and cancer I: Nucleotide composition and recombination-associated motifs. Hum Mutat. 2003;22(3):229-44.
35. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number in the human genome. Nature. 2006;444(7118):444-54.
36. Wang G, Vasquez KM. Models for chromosomal replication-independent non-B DNA structure-induced genetic instability. Mol Carcinog. 2009;48(4):286-98.
37. Lombard DB, Chua KF, Mostoslavsky R, Franco S, Gostissa M, Alt FW. DNA repair, genome stability, and aging. Cell. 2005;120(4):497-512.
38. Zhang L, Lu HH, Chung WY, Yang J, Li WH. Patterns of segmental duplication in the human genome. Mol Biol Evol. 2005;22(1):135-41.
39. Nusbaum C, Zody MC, Borowsky ML, Kamal M, Kodira CD, Taylor TD, et al. DNA sequence and analysis of human chromosome 18. Nature. 2005;437(7058):551-5.
40. Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, et al. Mapping copy number variation by population-scale genome sequencing. Nature. 2011; 470 (7332):59-65.

41. Scherer SW, Lee C, Birney E, Altshuler DM, Eichler EE, Carter NP, et al. Challenges and standards in integrating surveys of structural variation. Nat Genet. 2007;39 (7 Suppl):S7-15.

42. Keller M, Glessner J, Resnick E, Perez E, Chapel H, Lucas M, et al. Burden of copy number variation in common variable immunodeficiency. Clin Exp Immun. 2013; 177: 269-271.

# CHAPTER 5

# DNA Copy Number Variation in a Cohort of Healthy Australian Women

## 5.1 Abstract

Genome-wide investigation has identified uiquitous polymorphic rare and private benign copy number variation (CNV) in the healthy population. Published studies have analysed CNV properties in case cohorts and HapMap individuals but few studies have documented this in general populations. A high resolution genome-wide investigation of 64 healthy females of a Western European descendent population is presented here. The CNV incidence, length and relative proportions of duplication and deletion are assessed and compared to previous reports of HapMap based studies. In this female cohort, 3% of CNV are larger than (>) 100 kb in length. Intra-chromosomal variation of the CNV distribution is observed and in contrast to previous reports, this distribution did not reveal enrichment at the subtelomeric or pericentromeric regions.

At the time of investigation, 31 novel and several rare CNV were identified, some of the latter coincided with CNV currently classified in a diagnostic setting as variants of uncertainty or unknown significance. A comparative analysis of CNV properties from clinical material referred for fetal demise and published studies of neurological disorders demonstrates the differences in CNV properties including CNV load. This study contributes to a common initiative to describe, characterise and catalogue regions of copy number variation in a general population cohort.

## 5.2 Introduction

Many large scale genome-wide investigations have described and characterised copy number variation (CNV) since the first description by Sebat et al. 2004 and Iafrate et al. 2004. Several studies of the HapMap population using transformed cell lines described CNV in control cohorts (1-6). Documentation of CNV has been refined with improving platform resolution and with the contribution of sequencing data the derivation mechanisms of specific CNV can be inferred (7-9). However, inconsistencies in frequency, CNV load, copy number state and chromosome landscape exist between these studies. Biological contributions such as population variation (1, 4, 10, 11), discordance relative to the reference genome, or technical variation derived from the use of different platforms and CNV detection algorithms (12-15) have been shown to contribute to significant variation in CNV calls therefore undermining our understanding of the background incidence and properties of CNVs. Likewise the CNV calls identified in transformed cell lines may differ to those in other sources of DNA, adding a further bias to these studies.

Prior evidence has associated copy number imbalance with complex disease and disease predisposition (16-20). Evidence-based assessment of the significance of mutation in the genes involved in copy number change is required for ascertainment of clinical significance. However, the diagnostic interpretation of rare and low frequency variants in clinical referrals is complicated by the paucity of evidence which makes assignment of clinical significance challenging. As such the reporting of novel and rare benign CNV in a healthy cohort is applicable to

establishing such evidence. This is even more apparent in the prenatal setting where there is limited phenotypic information. Furthermore, recent studies have demonstrated an increased load of large and rare CNV in individuals with developmental delay, intellectual disabilities and congenital anomalies (16, 20-24), although the concept of contribution of CNV load to clinical significance is yet to be fully established.

In this chapter the CNV load, CNV length, copy number state and chromosome distribution are investigated for all autosomes in accordance with the model previously described for chromosome 18. The CNV properties are cross referenced against control cohort studies of transformed cell lines of HapMap individuals to identify the similarities and differences between these studies. The aim of this study is to describe the characteristics of CNV in a general population cohort and document rare and novel variants associated with an apparent benign outcome. Further comparison with the characteristics of CNV in clinical referrals of fetal loss and published studies of autism (20) demonstrates the contribution of general population studies to the investigation of the clinical significance of copy number variation.

## 5.3 Materials and Methods

### 5.3.1 Sample collection

The female control cohort was recruited from the "Aussie Normal" Collection representing a Western European descendent population collected in Canberra,

Australia. This is a community based study of Australians recruited from the Australian Electoral Roll. Written consent was provided by the participants and ethics approval was initiated through the Australian National University Human Research Ethics committee.

CNV data were collected from 106 samples of fetal demise referred to The Canberra Hospital, Australia. The products of conception were analysed as part of routine diagnostic molecular investigation of fetal loss ranging in gestational age from 10 to 38 weeks. The analysis of de-identified data was undertaken with permission from the ACT Health Research and Ethics Committee. Referrals included fetal death, congenital anomalies and spontaneous miscarriage. Copy number changes associated with monosomy, trisomy, triploidy and molar pregnancies were excluded from the investigation.

## 5.3.2 Cytogenetic investigation

Lymphocyte cultures were established for all 64 samples. Peripheral blood was cultured in RPMI medium and HAMs F10 containing phytohaemagglutinin (PHA) for 72-96 hrs. The cells were arrested in metaphase using Colcemid (40ug/ml) for 9 min, 0.075M KCL for 13 min, 5% acetic acid wash and 4 changes of Carnoy fixative. Two drops from a glass pasteur pipette was placed separately on the slide for spreading and drying. The slides were checked using a phase microscope for mitotic index and optimal spread of the metaphase without overlaps or cell breakage. The slides were dried in the 90° C oven for 90 min and 60° C oven

overnight. G–banded slides were prepared from 1-2 slides with trypsin and counterstained with Leishman stain.

Ten metaphase spreads at the 550-700 bands per haploid set were analysed on each sample for constitutional structural rearrangements. The bands on each homologous pair are matched and compared against the expected band pattern of the idiograms (25).

## 5.3.3 DNA extraction

DNA was extracted from whole blood using standard organic separation and ethanol precipitation. The method is described in chapter 3 and briefly described here. All samples were extracted manually, using mechanical extraction in lysis buffer. The suspension is left overnight in proteinase K and separation by phenol/chloroform is then performed. The DNA threads are precipitated using cold ethanol, air dried and stored in TE buffer.

## 5.3.4 Microarray investigation

The microarray investigation, molecular confirmation and quality control criteria were reported previously (Chapter 3). In brief, DNA was extracted using standard protocols from blood samples collected from 64 females who are enrolled in the "Aussie Normals" Collection. Genome-wide analysis was performed using Illumina Human Omni1-Quad. This platform has a median probe interspacing of 1.2kb and overall resolution of 5kb (Illumina, Inc.). A minimum of 4 consecutive markers

with Log R <-0.3 (loss) or >0.15 (gain) was required to detect a genuine call. The standard Illumina clusterfile was used for determination of genotypes. Two CNV detection algorithms, CNV Partition v2.3.4 (Illumina, Inc.) and PennCNV (13) were applied to the raw data and only the CNV calls common to both CNV detection programs were included in the study.

Molecular karyotyping of products of conception was performed using Illumina CytoSNP12. This platform is used for routine diagnostic services and has 300,000 markers, with an average interspacing of 10kb. A minimum of 20 consecutive markers with Log R <-0.3 (loss) and >0.15 (gain) was required to detect a genuine call. The standard Illumina clusterfile was used in the analysis and CNV Partition v2.3.4 was applied for the detection of CNV.

### 5.3.5 Molecular confirmation

Two deletions detected in 8/64 and 16/64 individuals, measuring less than 5kb on chromosome 18 and were confirmed by PCR and sequencing. The method has been described previously in Chia et al. 2013 and in Chapter 3 (8). In brief, putative breakpoints were refined using a PCR tiling path. Primers were designed using Primer 3 V4.0 and the position optimised to differentiate between wild type, heterozygous and homozygous deletions. Capillary sequencing was performed using the optimised forward and reverse primer pair. The precise breakpoint was identified by comparing the sequence to the reference genome (NCBI36/hg18).

## 5.3.6 FISH confirmation

A selection of novel, rare and common copy number variation regions (CNVR) on chromosomes 1, 2, 7, 12, 16, and 17, measuring greater than 100kb was confirmed using fluorescence in-situ hybridisation (FISH) of metaphase spreads or interphase nuclei obtained from lymphocyte cell preparations.

Fluorescence-labelled bacterial artificial chromosome clones were designed using the UCSC Genome Browser (http://genome.ucsc.edu/ hg18)(26). A custom file of the CNV calls was applied and RP11 clones were optimised for the CNV region. To demonstrate gains, probes were selected within the CNV or where possible two RP11 clones were selected in dual colour. To optimise signal interpretation of duplications a pre-treatment with lysis buffer was performed. FISH signals were analysed and captured using a Zeiss Axioscope microscope and Metasystem digital imaging software. Metaphase spreads were analysed to confirm the location of the RP11 clone to the correct chromosome and band assignment and to identify intra or inter-chromosomal insertions. Interphase nuclei were scored for the number and pattern of signals and compared to a negative control.

## 5.3.7 Website investigations

The UCSC genome browser (http://genome.ucsc.edu/ hg18) including the tracks for RefSeq, OMIM gene content, repetitive elements, segmental duplication, chromosome band location, International Standard for Cytogenomic Arrays (ISCA) and Database of Genomic Variants (DGV)(26). Assessment of segmental

duplication 300kb proximal and distal to the putative breakpoints was performed using UCSC Genome Browser (NCBI36/hg18) (26). ISCA tracks were applied to ascertain previous registration in benign, curated benign, uncertain and curated pathogenic categories. Genes were investigated using UCSC (http://genome.ucsc.edu/ hg18), OMIM and Panther (http://www.pantherdb.org/) websites. Statistical analysis was performed using Graphpad Quick Calcs (www.graphpad.com).

## 5.4 Results

### 5.4.1 Conventional karyotype

Cytogenetic analysis was performed for all samples to identify constitutional chromosomal rearrangements. Conventional karyotyping did not detect balanced reciprocal translocations or unbalanced rearrangements in the 64 females. Inversion 9qh, a common polymorphic variant is observed here in 1/64 females.

### 5.4.2 Determining the properties of CNV

To ascertain the characteristics of copy number variation in a cohort of Western European descendent females, the CNV were analysed for length, relative contribution of duplication and deletion and chromosome distribution. The focus

of the study here pertains to the autosomes, excluding the complex nature of detection of CNV on the X chromosome.

## 5.4.3 Incidence of CNV in the cohort

A total of 17,292 CNV >1kb for chromosome 1-22 were detected by PennCNV whilst CNV Partition detected 13,422. A total of 10,671 CNV >1kb on chromosomes 1-22 are included in the study after merging of the CNV outputs and QC metrics were applied. A comprehensive investigation identified 452 CNV calls >100kb representing 115 CNVR. There is an average of 166 autosomal CNV per individual and 7 CNV >100kb per individual. The largest CNV is 1.4Mb and CNV larger than 500kb are singletons representing private variants that were detected in 8/64 individuals. The overall average CNV size >100kb is 214kb.

The number of CNV events correlated with the size of each chromosome with few exceptions (Figure 1). Chromosomes 6, 16, 19 and 22 have higher than the average incidence of CNVs whilst chromosomes 15, 20 and 21 recorded comparatively lower incidence of CNVs per megabase (Mb) (Figure 2). Chromosome 14 recorded a high incidence which correlates with the hyper-variable immunoglobulin gene region at 105Mb.

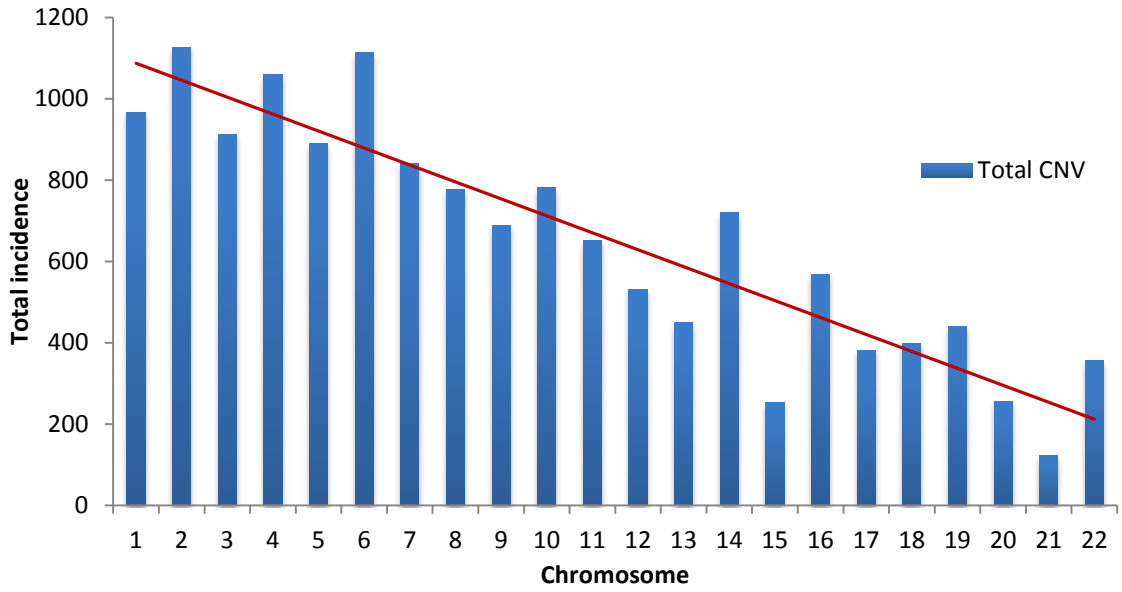



**Figure 1.** The number of CNV events apparently correlates with the size of each chromosome with the exception of chromosome 6, 14, 15, 21 and 22.

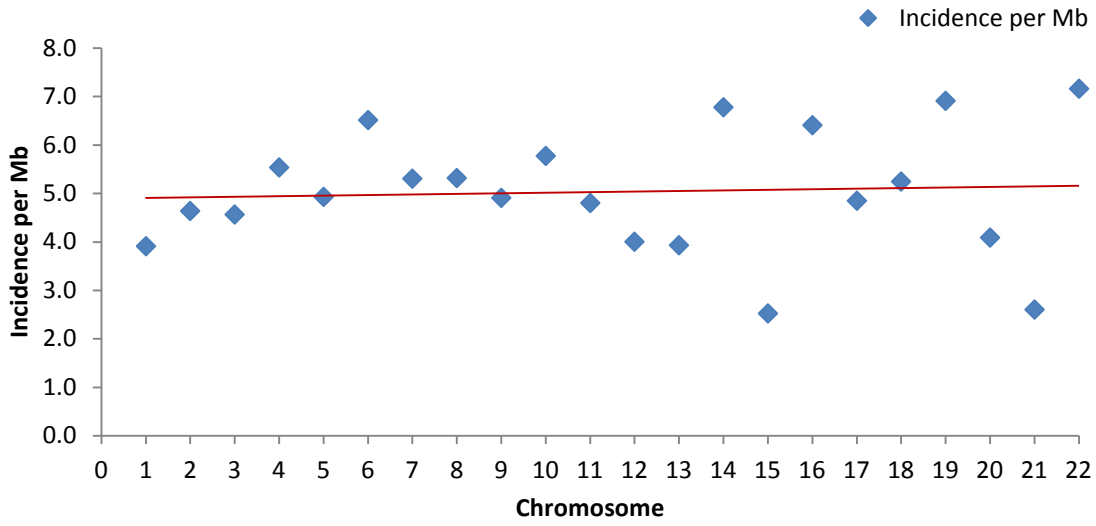



**Figure 2**. The incidence of CNV per Mb of chromosome length shows uniformity for CNV incidence among the autosomes with the exception of chromosomes 6, 15, 16, 19, 21 and 22.

In contrast the incidence for CNV>100kb did not correlate with chromosome length and demonstrates a marked variation in the CNV load between chromosomes (Figure 3). Chromosome 2, 8 and 16 recorded a high incidence whilst a low incidence was reported for chromosomes 5, 13, 18 and 21. To determine the contribution of common CNV compared to low frequency and private variant the number of CNV is considered relative to the number of CNVR for each chromosome. The results indicate that the inter-chromosomal variation for CNV >100kb may be explained in part by the presence of high incidence common CNV resulting in ascertainment bias such as seen here for chromosome 2 and 8. However, evidence for inter-chromosomal variation remains when the CNVR alone is reviewed (Figure 3).

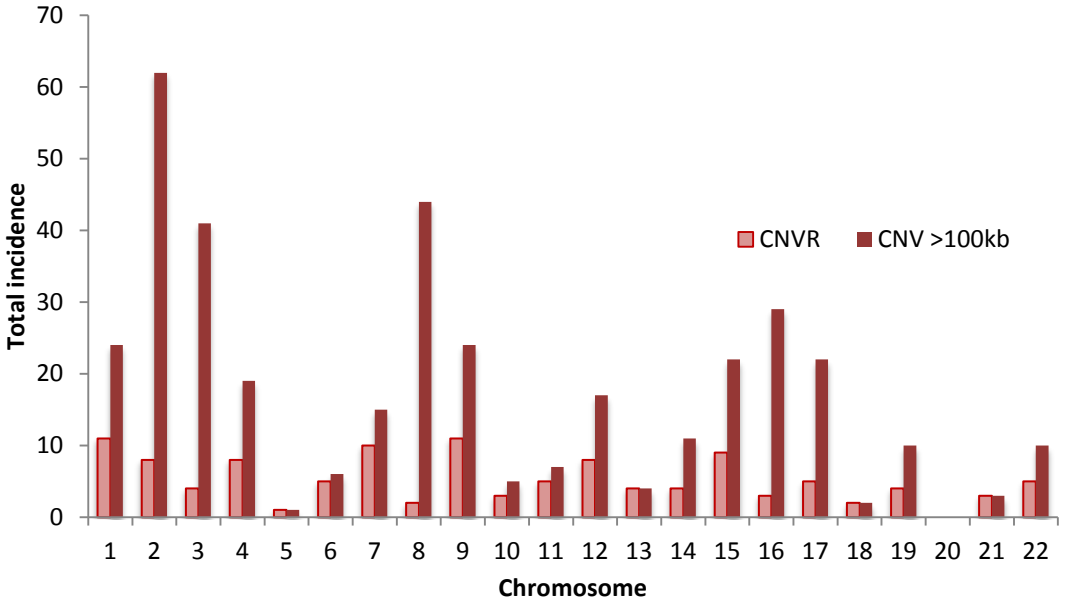



**Figure 3.** Inter- chromosomal variation is apparent when the incidence of CNV >100 kb is plotted for each chromosome. The incidence of CNV relative to the number of CNVR is consistent with events of common, low frequency or private variants. For example chromosomes 2 and 8 have a high number of CNV in a low

number of regions, whereas chromosome 13 and 18 are indicative of private variants.

## 5.4.4 Distribution of CNV length in the cohort

All CNV are categorised according to CNV length. The association of CNV events and length is skewed with 76% (8119/10,671) of CNV less than 10kb and 4.2% (452/10,671) of calls larger than 100kb (Figure 4).

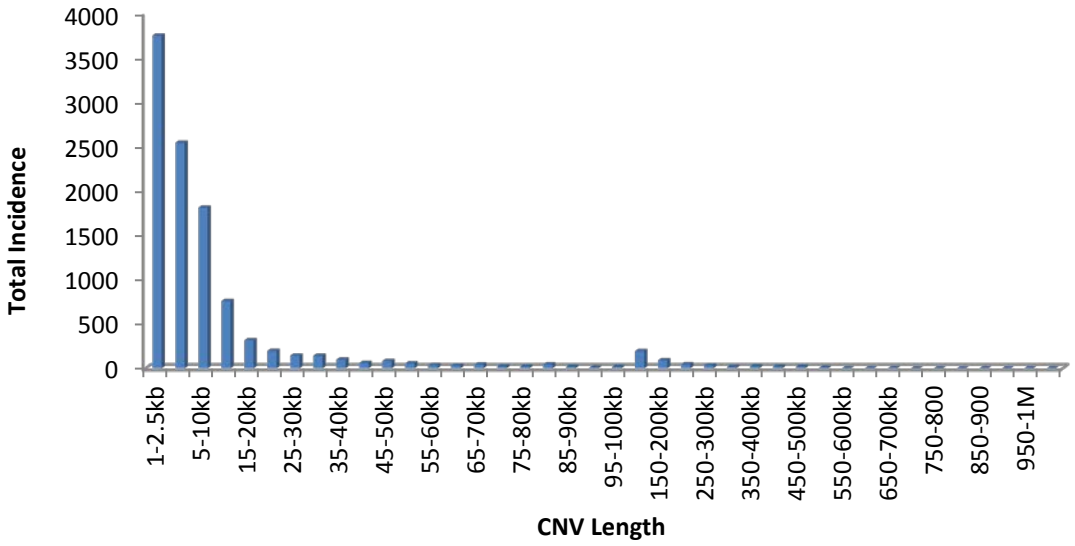



**Figure 4**. The length of benign CNV is skewed with an over-representation of CNV in the 1kb-50kb size category.

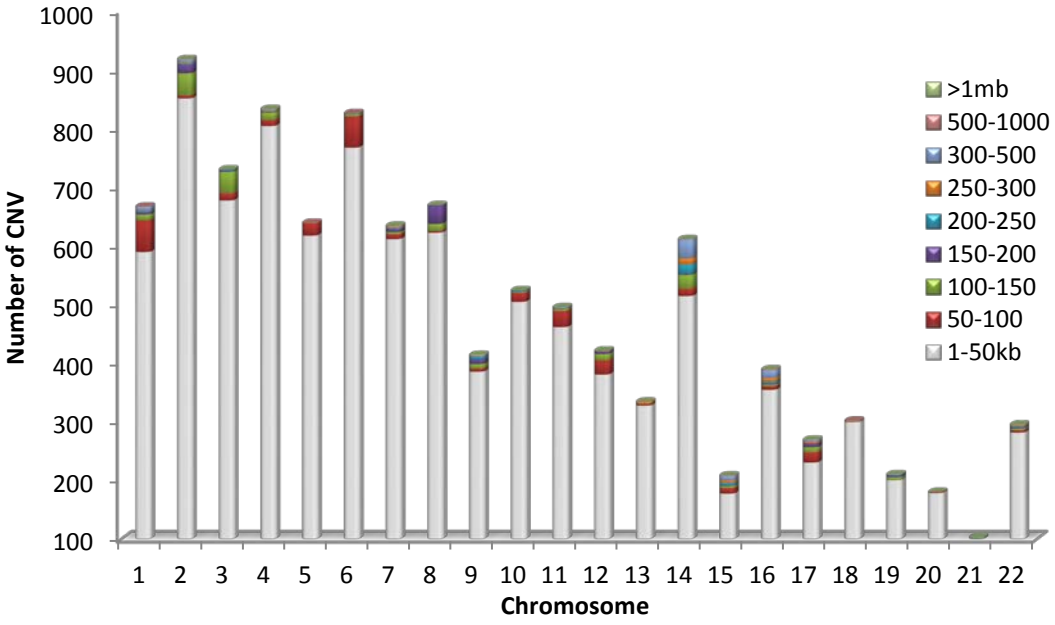



**Figure 5**. Inter-chromosomal comparison of CNV length shows that all chromosomes have high proportion of CNV up to 50 kb in length. Chromosomes 2, 3 and 8 show the greatest contribution of CNV > 100kb whilst chromosome 5, 13, 18, 20 and 21 have the least.

The CNV are plotted for each chromosome with respect to CNV length (Figure 5). All chromosomes recorded CNV < 50kb accounting for 85-99% of the total number of CNV for each chromosome. Inter-chromosomal variation in CNV lengths is observed. For example chromosome 2 recorded a comparatively low number of CNV in the size range 50 -100kb but the highest number of CNV > 100kb. Chromosome 20 recorded no CNV larger than 100kb and chromosome 21 recorded 99/103 (96%) of CNV < 50kb in length, with 3/103 (2.9%) larger than 100kb in length. Chromosome 18 recorded the highest proportion of CNV <50kb (99.01%) and a contribution of CNV >100kb of 0.06%.

## 5.4.5 Contribution of gain and loss in the cohort

The total CNV were further investigated for the copy number state to determine the contribution of duplication and deletion to benign CNV. Assessment of the relative proportion of gain and loss, shows that copy number loss predominates being 93.7% (9,995/10,671) of calls. Evaluation of CNV length together with copy number change revealed that deletions outweighed duplications at a ratio of 27:1 in CNV smaller than 100kb, but in CNV larger than 100kb the deletion to duplication to ratio shifted to 1:1.2.

The proportion of duplication and deletion for CNV >100kb was investigated for each chromosome. Although the overall loss: gain ratio is 1:1.2 there are striking differences between chromosomes. Chromosomes 1, 8, 9, 16 and 19 demonstrated predominantly loss, whereas a high incidence of gain is observed on chromosomes 2, 3 and 12 (Figure 6).

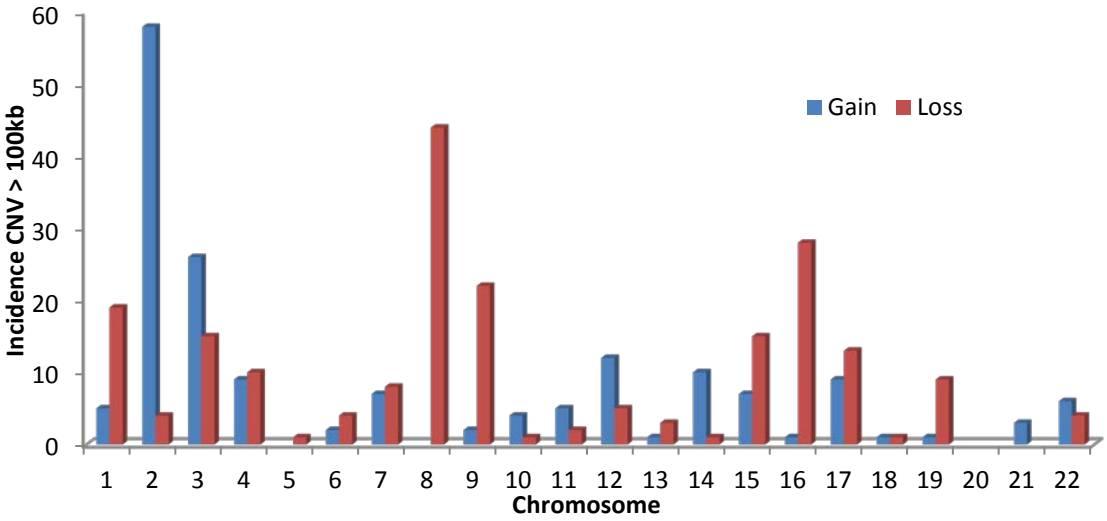



**Figure 6.** The proportion of duplication and deletion is compared for CNV >100kb for each chromosome and demonstrates inter-chromosomal variation with respect to copy number state.

## 5.4.6 CNV chromosomal landscape

The incidence of CNV and CNVR were scored in 5Mb windows and mapped along each chromosome. The distribution was examined to investigate correlation of CNV with gross morphological structures (Figure 7). All chromosomes showed CNV distributed for the entire length, but not all chromosomes were consistent with the general correlation of CNV with chromosome structures such as subtelomeric and pericentromeric regions. To investigate this further CNV are scored for the terminal 1Mb for each chromosome based on the documented length of the chromosome (NCBI36/hg18). The subtelomere is defined as the region within 1Mb from the telomere, in accordance with the reports of Knight et

al. 2000 (27). The association with subtelomeric regions is found to be confined to specific chromosomes such as the short and long arm subtelomeres of chromosomes 1, 5 and 11. With the exception of chromosome 21, all chromosomes scored CNV >1kb in at least one subtelomere. However the proportion of the total incidence of CNV located at the subtelomeres ranged from none (2p, 18p, 20p, 9q, 16q, 18q, 19q) to 7% (8p). CNV >100kb were distributed along all chromosomes with few recording CNVs in the pericentromeric region (Chromosomes 2, 6, 8, 15, 16 and 19) and in subtelomeric regions (chromosomes 2q, 5p, 6p, 8p, 9p and 14q) (Appendix 1). Intra-chromosomal variation for the distribution of CNV is apparent with some chromosomal regions recording high incidence and others CNV "deserts".

## 5.4.7 Chromosomal distribution and sequence properties

To investigate the intra-chromosomal variation of CNV distribution, the start and end positions for CNV >100kb was investigated for correlation with genomic architecture (Figure 7 and Appendix 1). An equal number of CNVR are associated with segmental duplications (55/115) and repetitive elements (55/115) and both were located at the breakpoints in 5/115 CNVR. However the results indicate that common CNV are more frequently associated with segmental duplication (288 CNV in 55 CNVR) and low frequency and private variants with repetitive elements (141 CNV in 55 CNVR). Likewise, there is no significant difference in the proportion of gain or loss with respect to breakpoints in segmental duplications or repetitive elements (p=0.58; Fisher's Exact test).

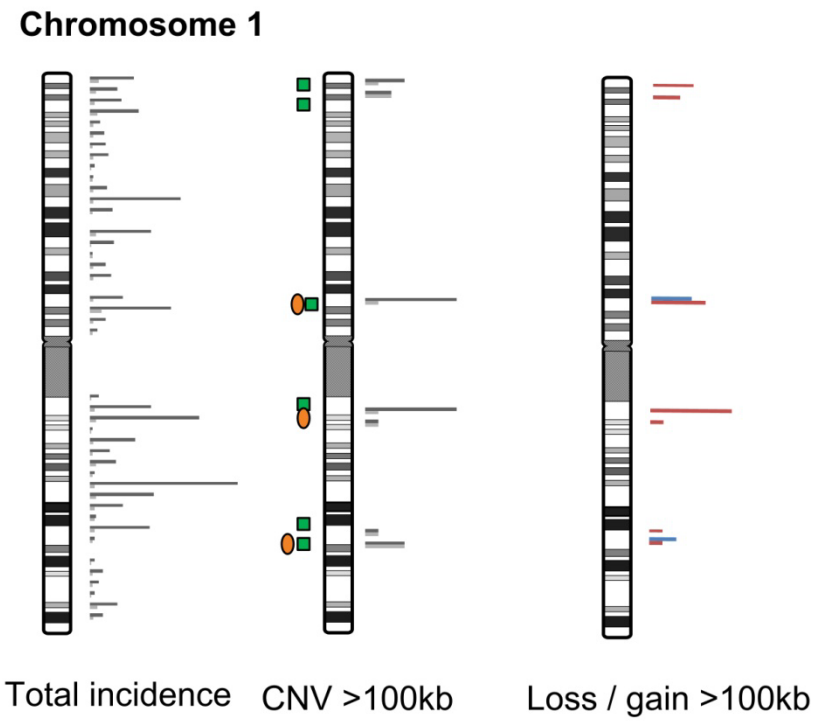



**Figure 7**. The positional location of CNV was mapped for chromosome 1. (a) The incidence of CNV (dark bar) and CNVR (pale bar) demonstrates that CNV covers the entire length of the chromosome with no apparent concentration at the subtelomeric or pericentromeric region. (b) CNV >100kb (dark bar) and CNVR (pale bar) show the presence of high incidence common CNV at 1p and 1q with others representing private and low frequency variants. Mapping of breakpoint regions with segmental duplication (green box) and repetitive elements (orange oval) shows a correlation of the common CNVR with segmental duplication for chromosome 1. Further investigation of copy number type (c) shows a higher incidence of gain (blue) for chromosome 1 than loss (red), and no correlation of copy number type with either genomic architecture tor common/private variants.

## 5.4.8 Gene content of benign CNV

CNV >100kb were further analysed for gene content to determine the contribution to benign CNV (Table 1). To do this, CNVR >100kb were analysed in UCSC Browser (http://genome.ucsc.edu; NCBI36/hg18). RefSeq genes are involved in 79% (91/115) CNVR. Although loss was found to account for 46% (208/452) of CNV >100kb, genes are involved in 79% (164/208) of loss compared to 48% (118/244) of gains (p=0.0001, Fisher's Exact test).

The function of genes involved in benign CNV was investigated using the PANTHER analysis tools (28). There is an over-representation of genes involved in immunity, signal transduction and unknown processes in the benign CNV of this cohort. Investigation of the genic contribution to deletion and duplication shows that signalling pathways, 53/164 (32.3%), and undefined biological processes are involved in 70/164 (42.6%) of deletions, whereas genes involved in immunological roles accounted for 39/118 (33%) of gains (http://pantherdb.org) (Figure 8).

**Table 1.** RefSeq genes encompassed by CNV > 100kb

| Chr.band | Start | End | Size | Loss/gain | Incidence | RefSeq genes |
|---|---|---|---|---|---|---|
| 1p36.21 | 12761104 | 12897871 | 136767 | Loss | 3 | *PRAMEF1-PRAMEF10* |
| 1p36.13 | 16886251 | 16988068 | 101817 | Loss | 1 | *ESPNP; MST1L* |
| 1p36.13 | 17050549 | 17152057 | 101508 | Loss | 1 | *MIR3675; CROCC* |
| 1p21.1 | 102434268 | 102618735 | 184467 | Loss | 1 | *MIR548AI* |
| 1p21.1 | 103902434 | 104131400 | 228966 | Loss/Gain | 6 | *ACTG1P4-AMY1C* |
| 1q21.1 | 147167498 | 147664565 | 497067 | Loss | 7 | *LOC645166;FCGR1C* |
| 1q22 | 153775506 | 153914057 | 138551 | Loss | 1 | *ASH1L;MSTO1;YY1AP1* |
| 1q31.3 | 195059520 | 195168376 | 108856 | Loss | 1 | *CFHR1;CFHR4* |
| 1q32.1 | 202165699 | 202389494 | 223795 | Gain | 1 | *LINC00303;SOX13;ETNK2* |
| 1q32.1 | 202455140 | 202838694 | 383554 | Gain | 1 | *PLEKHA6; LINC00628* |
| 1q32.1 | 204476945 | 204643116 | 166171 | Loss | 1 | *CTSE;SRGAP2* |
| 2p11.2 | 86149308 | 86363012 | 213704 | Gain | 1 | *POLR1A- REEP1* |
| 2p11.2 | 87240102 | 87450741 | 210639 | Gain | 1 | *MIR4771* |
| 2p11.2 | 87512143 | 88041630 | 529487 | Gain | 1 | *LINC00152-RGPD2* |
| 2q11.2 | 97516593 | 97643522 | 126929 | Gain | 1 | *ANKRD36B; ACTR1B* |
| 2q13 | 110214098 | 110342804 | 128706 | Loss | 1 | *MALL;NPHP1* |
| 2q23.1 | 153718969 | 154187910 | 468941 | Loss | 1 | *RPRM* |
| 2q37.3 | 242512605 | 242750507 | 237902 | Loss | 1 | *LOC728323* |
| 3p12.3 | 75502426 | 75730431 | 228005 | Loss | 3 | *FAM86DP* |
| 3q29 | 196823043 | 196962869 | 139826 | Loss | 3 | *SDHAP2-MUC4* |
| 4p16.1 | 8934014 | 9081875 | 147861 | Loss/Gain | 2 | *USP17L15-DEFB131* |
| 4p12 | 47186827 | 47596233 | 409406 | Gain | 1 | *ATP10D;CORIN;NFXL1* |
| 4q13.2 | 69064675 | 69226833 | 162158 | Loss/Gain | 11 | *UGT2B17; UGT2B15* |
| 4q25 | 110220124 | 110356061 | 135937 | Gain | 1 | *COL25A1* |
| 4q26 | 117366455 | 117620842 | 254387 | Gain | 1 | *MIR1973* |
| 4q28.3 | 135139215 | 135402787 | 263572 | Loss | 1 | *PABPC4L* |
| 4q35.2 | 187301547 | 187609258 | 307711 | Loss | 1 | *FAM149A-LOC285441* |
| 5p15.33 | 801621 | 932134 | 130513 | Loss | 1 | *ZDHHC11;BRD9* |
| 6p25.3 | 210760 | 315830 | 105070 | Loss | 2 | *DUSP22* |
| 6q15 | 89316093 | 89631810 | 315717 | Gain | 1 | *RNGTT* |
| 6q27 | 168121697 | 168290121 | 168424 | Gain | 1 | *KIF25;FRMD1* |
| 7q11.23 | 76243712 | 76477775 | 234063 | Gain | 2 | *DTX2P1-UPK3BP1-PMS2P11* |
| 7q11.23 | 76626383 | 76746626 | 120243 | Gain | 1 | *CCDC146; FGL2* |
| 7q21.3 | 94176222 | 94446001 | 269779 | Gain | 1 | *PPP1R9A* |
| 7q31.3 | 110618016 | 110770811 | 152795 | Loss | 1 | *IMMP2L* |
| 7q31.32 | 121033877 | 121359954 | 326077 | Gain | 1 | *PTPRZ1* |
| 7q35 | 143506472 | 143701573 | 195101 | Loss | 1 | *CTAGE4;ARHGEF35,ARHGEF5; LOC728377* |
| 8p11.23 | 39353198 | 39506336 | 153138 | Loss | 41 | *ADAM5P,ADAM3A* |

| | | | | | | |
|---|---|---|---|---|---|---|
| 9p24.3 | 181843 | 339557 | 157714 | Gain | 1 | *DOCK8,C9orf66* |
| 9p24.3 | 754128 | 865364 | 111236 | Loss | 1 | *DMRT1* |
| 9p12 | 41857190 | 42084257 | 227067 | Loss | 1 | *MGC21881;LOC643648;KG FLP2* |
| 9p11.2 | 43444850 | 43599650 | 154800 | Loss | 1 | *SPATA31A6* |
| 9p11.2 | 43606156 | 44481665 | 875509 | Loss | 1 | *CNTNAP3B; LOC643648* |
| 9q31.1 | 104152898 | 105593849 | 1440951 | Loss | 1 | *CYLC2* |
| 9q31.1 | 105609407 | 106307173 | 697766 | Loss | 1 | *SMC2,OR13F* |
| 9q34.3 | 137288987 | 137449867 | 160880 | Gain | 1 | *C9orf62* |
| 10q11.2 | 47004551 | 47223842 | 219291 | Gain | 3 | *ANXA8L2* |
| 10q23.2 | 88977114 | 89103066 | 125952 | Loss | 1 | *FAM22A; LOC728190;LOC439994* |
| 10q26.3 | 135090909 | 135239466 | 148557 | Gain | 1 | *LOC19207;CYP2E;SYCE1;SP RNP1* |
| 11p15.4 | 4038080 | 4179554 | 141474 | Gain | 1 | *STIM1,RRM1, LOC100506082* |
| 11p11.2 | 50961054 | 51450453 | 489399 | Gain | 1 | *OR4A5, OR4C46* |
| 11q11 | 54998132 | 55153248 | 155116 | Loss | 1 | *OR4C15-OR4C11* |
| 11q11.1 | 55122337 | 55271506 | 149169 | Gain | 3 | *OR4C11-OR4C6* |
| 12p13.33 | 1727243 | 1888251 | 161008 | Gain | 1 | *ADIPOR2;CACNAD4; LRTM2* |
| 12p13.31 | 7888814 | 8014573 | 125759 | Loss/Gain | 7 | *SLC2A14, SLC2A3* |
| 12p13..31 | 9446466 | 9630716 | 184250 | Loss | 2 | *DDX12P* |
| 12p12.3 | 19360587 | 19464917 | 104330 | Gain | 1 | *PLEKHA5* |
| 12q13.11 | 45519568 | 45772649 | 253081 | Gain | 1 | *AMIGO2* |
| 12q14.2 | 62221022 | 62426466 | 205444 | Gain | 2 | *DPY19L2* |
| 13q21.31 | 60516854 | 60785883 | 269029 | Loss | 1 | *MIR3169* |
| 14q11.2 | 19133955 | 19492423 | 358468 | Gain | 11 | *OR11H2 - OR4K1* |
| 14q13.2 | 34665459 | 34772551 | 107092 | Loss | 1 | *KIAA0391* |
| 14q32.33 | 105161788 | 105307401 | 145613 | Loss | 1 | *ELK2AP* |
| 15q11.2 | 18492589 | 18839283 | 346694 | Loss | 1 | *CHEK2P2* |
| 15q11.2 | 18850028 | 19831153 | 981125 | Loss/Gain | 5 | *HERC2P3-NF1P2* |
| 15q11.2 | 19788901 | 20096944 | 308043 | Loss/Gain | 8 | *LOC72792-REREP3* |
| 15q13.1 | 25807809 | 25975632 | 167823 | Loss | 1 | *OCA2* |
| 15q13.2 | 28251094 | 28641378 | 390284 | Loss | 1 | *CHRFAM7A* |
| 15q13.3 | 30231488 | 30695727 | 464239 | Loss | 1 | *CHRNA7 ARHGAP11* |
| 15q14 | 32505886 | 32625184 | 119298 | Loss | 2 | *GOLGA8A;GOLGA8B* |
| 15q15.3 | 41681110 | 41816819 | 135709 | Loss | 2 | *STRC-CATSPER2P1* |
| 16p13.2 | 6855484 | 7031955 | 176471 | Loss | 1 | *RBFOX1* |
| 16p11.2 | 32071836 | 32570401 | 498565 | Loss | 10 | *LOC* |
| 16p11.2 | 33204993 | 33783127 | 578134 | Loss | 16 | *RNU6-76 prov* |
| 16p11.2 | 34336806 | 34614585 | 277779 | gain | 1 | *LOC283914, LOC100130700* |
| 17q12 | 31456238 | 31885980 | 429742 | Loss | 5 | *CCL4-TBC1D3G* |
| 17p11.2 | 21293006 | 21417020 | 124014 | Gain | 1 | *C17orf51* |
| 17q12 | 33347981 | 33658764 | 310783 | Loss | 1 | *TBC1D3;TBC1D3F* |
| 17q21.31 | 41519627 | 41713128 | 193501 | Gain | 8 | *KANSL1* |

| | | | | | | |
|---|---|---|---|---|---|---|
| 17q21.32 | 42003374 | 42153004 | 149630 | Loss | 5 | *ARL17A - NSF* |
| 18p11.31 | 4167454 | 4295905 | 128451 | Loss | 1 | *DLGAP1* |
| 18p11.23 | 7101033 | 7648679 | 547646 | Gain | 1 | *LAMA1-PTPRM* |
| 19p12 | 20404485 | 20596353 | 191868 | Loss | 2 | *ZNF737* |
| 19p12 | 20627130 | 20788764 | 161634 | Gain | 1 | *ZNF626* |
| 19q13.33 | 48062619 | 48435900 | 373281 | Loss | 2 | *PSG1-PSG6* |
| 21q22.3 | 45590493 | 45707581 | 117088 | Gain | 1 | *COL18A1,COL18A1-AS1* |
| 22q11.1 | 15430088 | 15639954 | 209866 | Gain | 1 | *CCT8L2; ANKRD62P1 PARP4P3* |
| 22q11.2 | 17054349 | 17258027 | 203678 | Loss | 3 | *GGT3P* |
| 22q11.21 | 17258409 | 17388108 | 129699 | Gain | 3 | *DGCR6;PRODH; DGCR9* |
| 22q11.23 | 23999569 | 24255760 | 256191 | Gain | 2 | *IGLL3P; CRYBB2p1* |

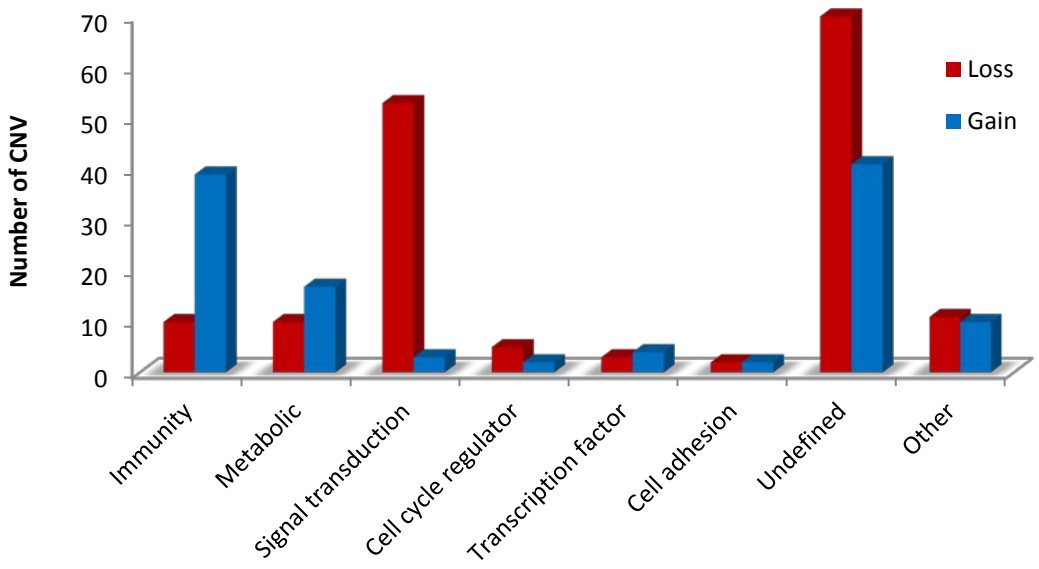



**Figure 8**. Immunity and signal transduction genes are involved in benign CNV. The role of a large proportion of genes involved in benign CNV is yet to be determined. Deletion is more frequently associated with genes involved in signal transduction, while duplication is more frequent of genes with roles in immunity.

## 5.4.9 Determination of clinical significance

The CNV calls were analysed in UCSC Browser (http://genome.ucsc.edu; NCBI36/hg18), with tracks applied for International Standards for Cytogenomic Arrays (ISCA) (curated benign, uncertain and curated pathogenic) and Database of Genomic Variants (DGV) to determine the category of clinical significance. This analysis is complicated by the overlapping of benign, uncertain and pathogenic categories in the ISCA database for a given chromosomal position. For the purpose of this study a CNVR is considered of benign clinical significance where it overlaps a registration in the DGV and curated benign in the ISCA database. CNVR of benign clinical significance is represented in 62/115 CNVR >100kb and CNVR that did not involve genes occurred in 24/115. Novel CNVR was observed for 31/115 CNVR >100kb, 8 of which did not involve genes and 6/115 overlapped uncertain or pathogenic categories.

## 5.4.10 Novel variants

At the time of investigation, 31 novel CNVR > 100kb, a selection of which were confirmed by FISH, were identified on all chromosomes with the exception of 5, 8, 10, 16, 17, 19 and 20. All of these are singletons apart from one CNV at 6q27 that was represented in two samples. Gene involvement was observed in 66% (20/31) of the novel CNVR including 5 OMIM listed disease associated genes**,** such as 4p12 (CORIN; OMIM# 605236) and 12p13.33 (CACNA2D4; OMIM # 608171) ( Table 2).

**Table 2**. List of CNVR that are reported as novel at the time of investigation (http://genome.ucsc.edu; NCBI36/hg18; accessed May 2012)

| Band | Start | End | Size | Copy number | RefSeq Genes | OMIM |
|---|---|---|---|---|---|---|
| 1q32.1 | 202165699 | 202389494 | 223795 | 3 | *LINC00303; SOX13; ETNK2* | No |
| 1q32.1 | 202455140 | 202838694 | 383554 | 3 | *PLEKHA6;PPP1R15B; PIK3C2B; MDM4* | No |
| 1q32.1 | 204476945 | 204643116 | 166171 | 1 | *CTSE;SRGAP2* | No |
| 2q23.1 | 153718969 | 154187910 | 468941 | 1 | *RPRM* | No |
| 3q26.1 | 167268104 | 167427320 | 159216 | 4 | No | No |
| 4p15.1 | 29695582 | 29924403 | 228821 | 1 | No | No |
| 4p12 | 47186827 | 47596233 | 409406 | 3 | *ATP10D;CORIN;NFXL1* | CORIN |
| 4q25 | 110220124 | 110356061 | 135937 | 3 | *COL25A1* | No |
| 4q26 | 117366455 | 117620842 | 254387 | 3 | putative microRNA | No |
| 4q35.2 | 187301547 | 187609258 | 307711 | 1 | *FAM149A;FLJ38576;CYP4V2; KLKB1;F11;LOC285441* | CYP4V2; KLKB1;F11 |
| 6q15 | 89316093 | 89631810 | 315717 | 3 | *RNGTT* | No |
| 6q27 | 167755609 | 167879332 | 123723 | 1 | No | No |
| 7q11.23 | 76626383 | 76746626 | 120243 | 3 | *CCDC146; FGL2* | No |
| 7q21.3 | 94176222 | 94446001 | 269779 | 3 | *PPP1R9A* | No |
| 7q31.32 | 121033877 | 121359954 | 326077 | 3 | *PTPRZ1* | No |
| 7q35 | 145085148 | 145332649 | 247501 | 1 | No | No |
| 9q31.1 | 104152898 | 105593849 | 1440951 | 1 | *CYLC2* | No |
| 9q31.1 | 105609407 | 106307173 | 697766 | 1 | *SMC2,OR13F* | No |
| 11p11.2 | 50961054 | 51450453 | 489399 | 3 | *OR4A5, OR4C46* | No |
| 12p13.3 | 1727243 | 1888251 | 161008 | 3 | *ADIPOR2;CACNA2D4;LRTM2* | CACNA2 |
| 12q13.1 | 45519568 | 45772649 | 253081 | 3 | *AMIGO2* | No |
| 13q21.1 | 56646935 | 56784440 | 137505 | 3 | No | No |
| 13q21.2 | 60057476 | 60308192 | 250716 | 1 | No | No |
| 13q21.3 | 60516854 | 60785883 | 269029 | 1 | MIR | No |
| 14q13.2 | 34665459 | 34772551 | 107092 | 1 | *KIAA0391* | No |
| 15q13.1 | 25807809 | 25975632 | 167823 | 1 | *OCA2* | OCA2 |
| 18p11.3 | 4167454 | 4295905 | 128451 | 1 | *DLGAP1* | No |
| 18p11.2 | 7101033 | 7648679 | 547646 | 3 | *LAMA1,LRRC30,PTPRM* | No |
| 21q21.3 | 28432764 | 28661799 | 229035 | 3 | No | No |
| 21q22.3 | 45790608 | 45917586 | 126978 | 3 | No | No |
| 22q13.3 | 47775187 | 48186380 | 411193 | 3 | No | No |

## 5.4.11 Comparison with published control cohorts

The properties of CNV load per chromosome, CNV length and composition of copy number state reported here are compared with previously reported control cohorts including HapMap studies to determine if the CNV properties are consistent or isolated to this study. The CNV properties for chromosome 18 are presented in chapter 3 and a consistent pattern is demonstrated when compared to HapMap based studies (8). In this chapter the properties of CNV in all autosomes is compared with the study reported by Park et al. 2010. In Park et al. 2010 the authors performed CNV analysis of HapMap transformed cell line using a high resolution CGH array platform (24x1M Agilent). The results are determined to be comparable for CNV length distribution, overall CNV load and contribution of duplication and deletion (table 3). The study reported an average CNV size of 212kb (range: 100kb-1.06Mb) (4) which is comparable to 214kb (range 100kb-1.4Mb) reported in the cohort reported here. A review of the distribution of CNV length showed a similar skewed distribution with 69% measuring 100-200kb in size compared to 63% in the current study (Figure 9). Analysis of the relative contribution of copy number change in CNV >100kb, shows that these are consistent between the studies with Park et al. 2010 reporting a loss: gain ratio of 1:1.4 (4).

A review of the CNV load for each chromosome was performed to determine if the inter-chromosomal variation observed in the current study is replicated in the study by Park et al. 2010 of HapMap individuals. A similar pattern of inter-

chromosomal variation is apparent for all chromosomes, although the proportions differed for chromosomes 1, 2, 3, 8, 9 and 10 (Figure 10).

**Table 3.** Comparison of the contribution of CNV>100kb to control and pathogenic cohorts

| | Current | Park et al | p= | Fetal demise | p= | Sanders et al | p= |
|---|---|---|---|---|---|---|---|
| **Total >100kb** | 452 | 863 | | 189 | | 1480 | |
| **Gain** | 244 | 497 | 0.2192 | 142 | <0.0001 | 926 | 0.0012 |
| **Loss** | 208 | 366 | 0.2192 | 47 | <0.0001 | 554 | 0.0012 |
| **100-200** | 286 | 598 | 0.0304 | 71 | <0.0001 | 844 | 0.0288 |
| **200-500** | 154 | 209 | 0.0002 | 72 | 0.3647 | 415 | 0.0156 |
| **500-1Mb** | 11 | 38 | 0.0909 | 20 | <0.0001 | 124 | <0.0001 |
| **>1Mb** | 1 | 18 | 0.0059 | 26 | <0.0001 | 87 | <0.0001 |

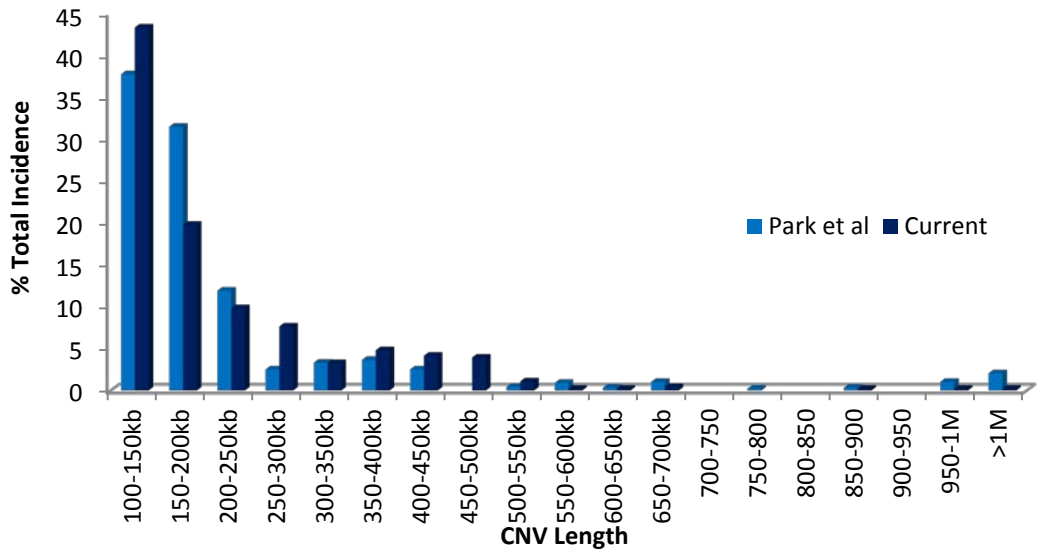



**Figure 9.** The relative contribution of CNV >100kb is shown to be consistent between the Park et al.2010 and the study reported here.

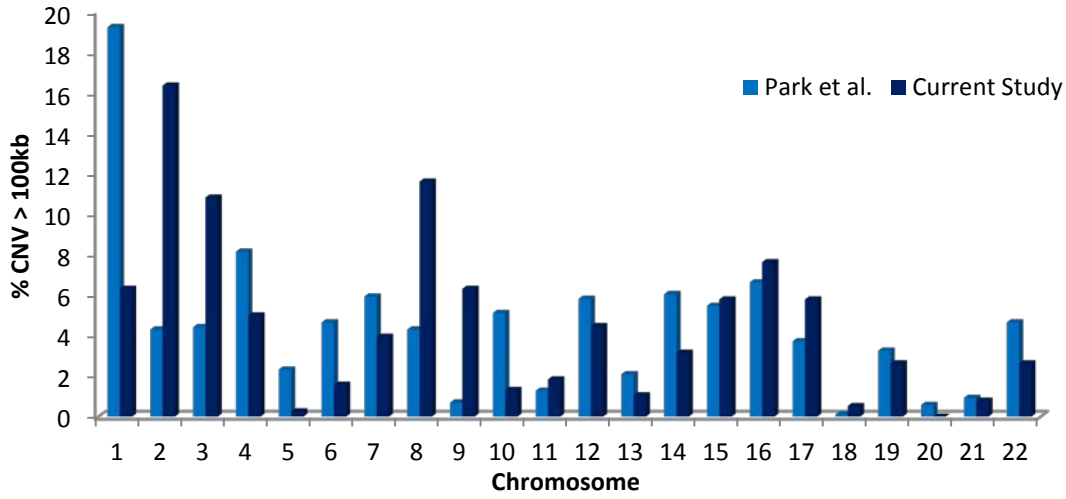



**Figure 10.** Inter-chromosomal variation of CNV load > 100kb is apparent for both studies.

## 5.4.12 Comparison with pathogenic cohorts

### 5.4.12.1 Autism Spectrum Disorder Study

The parameters of CNV length and proportion of duplication and deletion in the cohort reported here were compared with those identified in a genome-wide analysis of an autism spectrum disorder (ASD) cohort (20). In the study by Sanders et al. 2011, the Illumina 1M platform was used and CNV detection was performed with PennCNV and QuantiSNP. By applying the same model of investigation, the CNV length, and proportion of duplication and deletion was analysed and compared to the CNV properties in the cohort presented here. An over-representation of CNV measuring 10kb-250kb accounting for 55% compared to 16.7% in the current study ($p<0.0001$, $\chi^2$) was observed. The proportion of CNV is

comparable between the studies for the length range 200kb-500kb and enriched in Sanders et al. 2011 in CNV >500kb (Table 3). Evaluation of the contribution of copy number showed that the incidence of gain is increased across all CNV length ranges when compared to the current study (p=0.0012; Fishers Exact) and outweighed deletion at 2.3:1 compared to the current study of 1.2:1 (Figure 11).

Inter-chromosomal variation of CNV >100kb is demonstrated in the study by Sanders et al. 2011. This finding is consistent with the current study, although in contrast to the healthy population cohort, enrichment of CNV >100kb in chromosomes 5, 10, 11, 18, 20 and 21 is recorded in the autism study (Figure 12). This trend pertains with evaluation of the contribution of duplication (Figure 13).

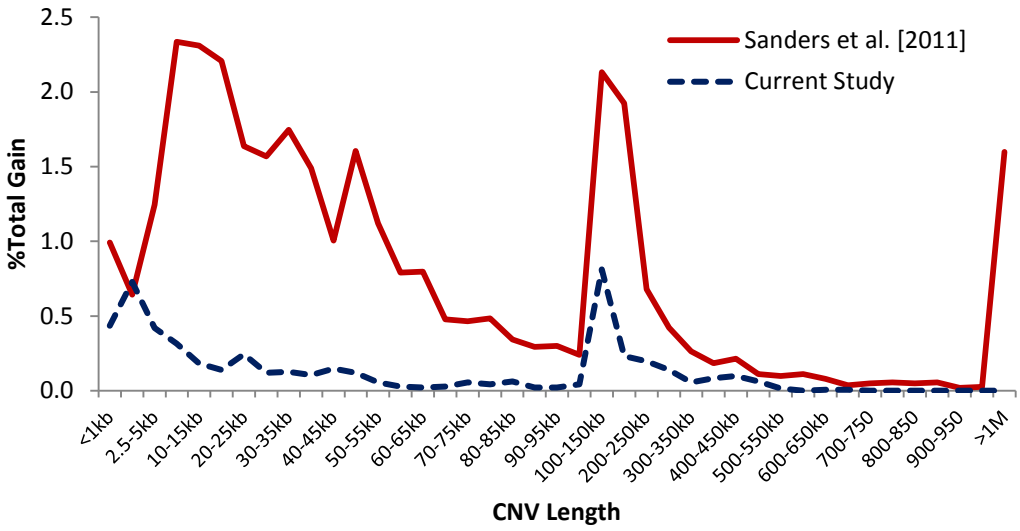



**Figure 11**. The proportion of gain is greater across all ranges of CNV length in Sanders et al. 2011 (solid line) compared to the current study (dashed line).

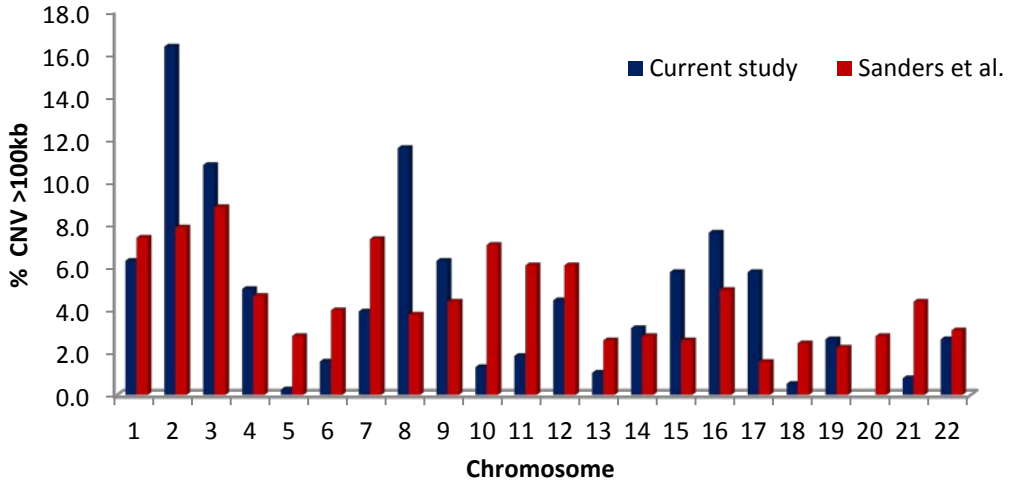



**Figure 12.** Inter-chromosomal variation of CNV load >100kb is apparent however the chromosome bias differs between the studies.

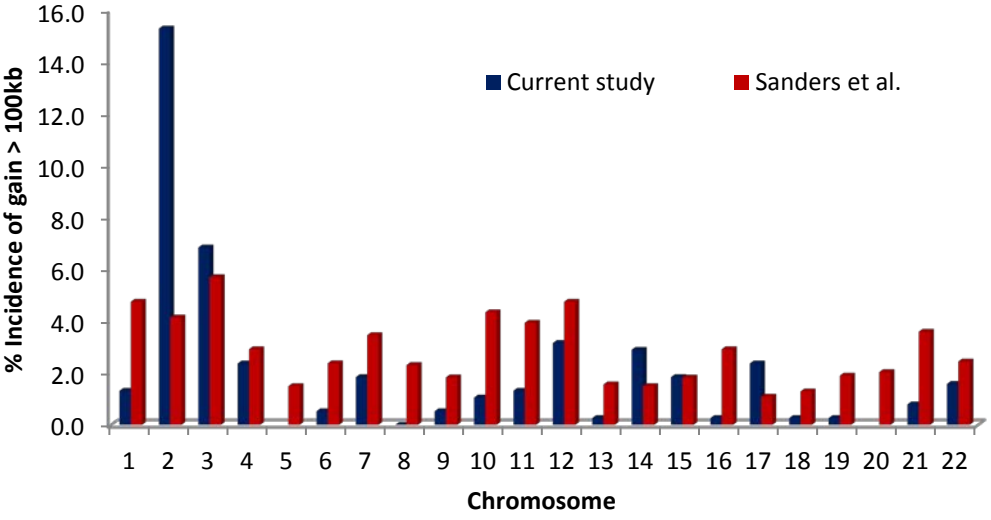



**Figure 13.** Inter-chromosomal variation of CNV gain compared to Sanders et al. 2011.

### 5.4.12.2 Fetal demise cohort

CNV >100kb was evaluated in 111 samples referred for fetal demise. After exclusion of molecular karyotypes with whole chromosome aneuploidy or triploidy, 189 CNV were included in the study. Due to differences in platform resolution between the studies, only CNV >100kb are considered for comparative analysis. The healthy population cohort is enriched with respect to the proportion of CNV in the size range 100kb-200kb (Figure14). Whereas, CNV >500kb is over-represented in fetal demise accounting for 24.4% compared to 2.6% ($p<0.0001$, Fishers' Exact) in the healthy population cohort (Table 3). The average CNV size is larger in the fetal demise cohort compared to the current study (541kb and 214kb respectively). In addition enrichment of duplication in the fetal demise samples is recorded, with a duplication to deletion ratio of 3:1 which is higher than the healthy cohort presented here ($p<0.0001$; Fisher's Exact) (Figure 15). Furthermore comparison of the average length of duplication demonstrates an enrichment of duplication in the fetal demise cohort compared to the healthy cohort (604kb and 210kb respectively).

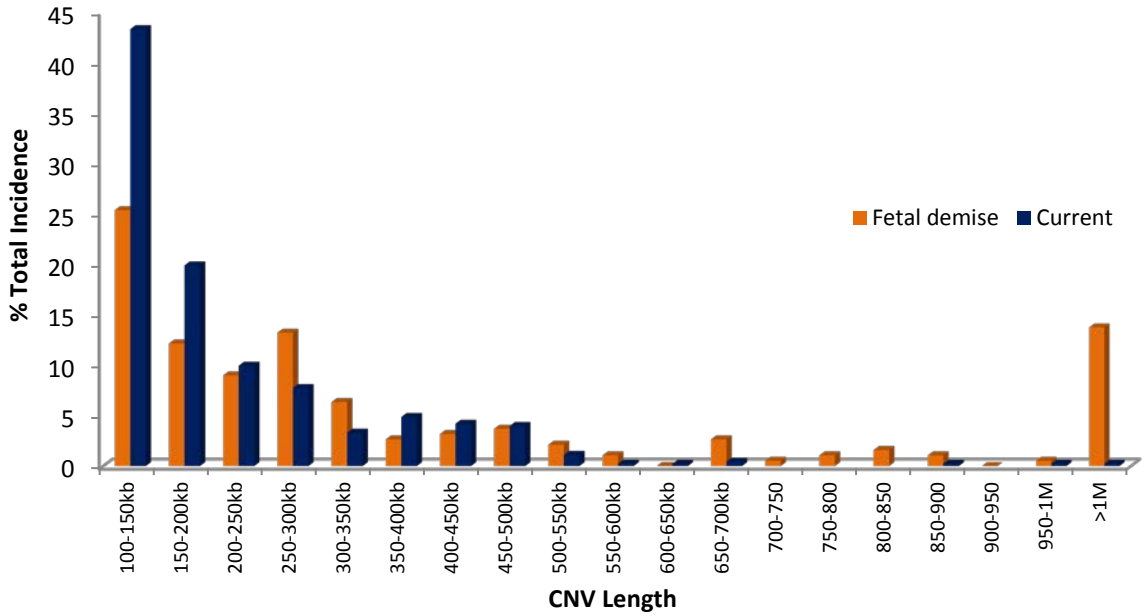



**Figure 14**. The CNV length in the fetal demise cohort compared to the current study shows a higher proportion of CNV > 500kb in the fetal demise cohort.

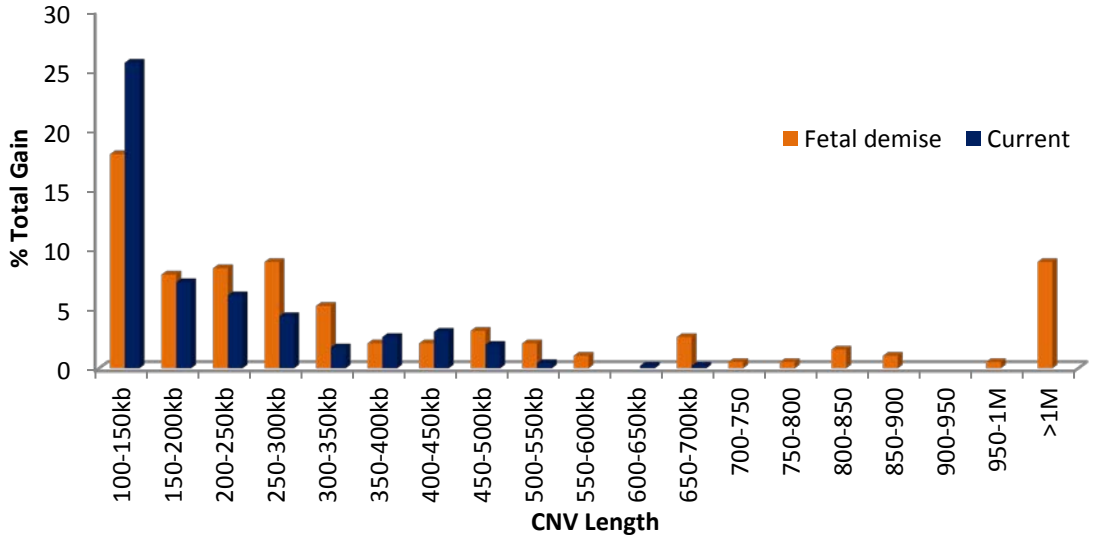



**Figure 15.** The relative proportion of duplication in fetal demise compared to the current study. The fetal demise cohort shows enrichment of duplication in CNV > 500kb.

The CNV load for each chromosome was recorded to identify chromosome bias in the fetal demise cohort compared to the healthy population. Inter-chromosomal variation of CNV load is apparent as previously demonstrated in the study reported by Sanders et al. 2011 of a cohort of autistic children and the healthy cohort presented here. However there are differences in the chromosome bias. In the fetal demise cohort there is CNV enrichment in chromosomes 6, 10, 12, 14, 17, 18, 20 and 22 compared to the cohort of healthy females (Figure 16).

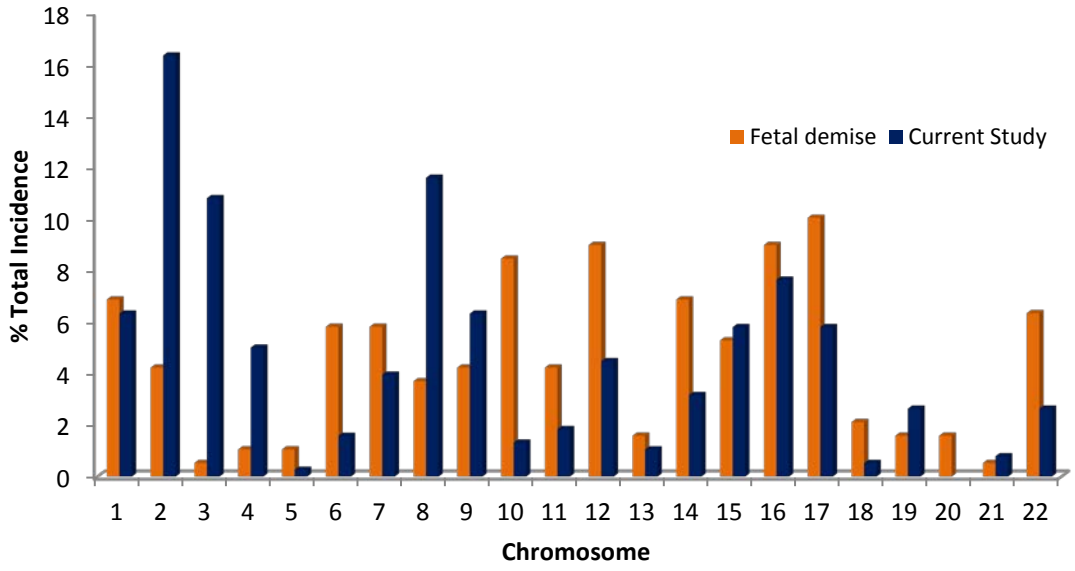



**Figure 16.** Inter-chromosomal variation of CNV load >100kb is apparent between the studies however there are differences in the chromosome bias.

### 5.4.13 Clinical application of healthy cohort studies

To demonstrate a clinical application of CNV studies of healthy individuals the CNV outputs were cross-referenced to the output from the fetal demise cohort for CNVR common to both. In the fetal demise group 87/189 CNV >100kb were shared with the current study. Common benign CNV defined as those occurring with an incidence of >5/64 in the current study and recorded in the DGV, accounted for 70/87. The remaining 17/87 partially overlapped common and low frequency CNV but were 100kb to 3Mb larger and encompassed more genes in the fetal demise cohort. Of interest is the homozygous deletions in the fetal demise cohort which account for 7% of shared calls (6/87) but these are reported as heterozygous deletions or duplications in the healthy population cohort.

## 5.5 Discussion

Several published studies have described the stratification of CNV in HapMap individuals but general population studies have rarely been described (15, 19, 29, 30). Factors such as platform type, marker density and study population have complicated the appreciation of the incidence and significance of CNV in the human genome (1, 30, 31). To gain insight into the contribution of benign CNV to genetic variation a comprehensive investigation of CNV properties was performed on a cohort of healthy Western European descendent female individuals.

A model of investigation was applied to determine the CNV properties of the cohort presented here. These were established for CNV load, length characteristics and relative contribution of duplication and deletion. The results show an average of 166 autosomal CNV per person of the average CNV length of 214kb and a predominance of small CNV with only 3% of CNV >100kb. The largest CNV is 1.4Mb and all CNV > 500kb are private variants. There is an enrichment of deletion in CNV < 100kb whereas the contribution of duplication is equivalent in CNV >100kb. It is proposed that the predominance of CNV <50kb and enrichment of deletion in the healthy cohort reported here is reflective of the mechanism of the derivation of CNV and indicative of replication repair mechanisms as described in Chapter 4.

Previous published studies suggest an association of CNV with gross morphological structures such as subtelomeres and pericentromeric regions (5, 11, 32). This hypothesis was evaluated in the current study by mapping CNV incidence in 5Mb windows along the chromosome and within 1000kb of the telomeres for evaluation of the subtelomeric region. CNV were observed along the entire length of each chromosome with the exception of CNV >100kb. Intra-chromosomal variation was apparent with enrichment of CNV in some regions. Chromosomes 1, 5 and 11 showed enrichment in the short and long arm subtelomeric regions. CNV >1kb accounting for 3% of the total incidence for the chromosome was observed in the subtelomeres of chromosomes 1p, 8p, 12p, 11q, and 22q.

The findings here do not show evidence that CNV are enriched in subtelomeric regions. However marker location in this region may be compromised due to the

complexity of homologous sequences in the subtelomeric region (27, 33, 34) and the marker density at the subtelomeric regions may differ among platforms. In addition the authors of published studies have not defined the region considered to be subtelomeric for their analysis (5). The study here defines the subtelomeric region in accordance with that reported by Knight et al. 2000 and as such few chromosomes show enrichment of CNV in the subtelomeric regions.

### 5.5.1 Comparison with other studies

Chromosome 18 was previously shown to have CNV properties consistent with those described by Conrad et al. 2010 and Matsuzaki et al. 2009. The same model of investigation was applied here for chromosomes 1-22 and the properties were determined to be comparable with those of published control studies of HapMap individuals (1, 4, 8, 10). Furthermore the contribution of CNV length and proportion of duplication and deletion are consistent between the studies.

Comparison of CNV load >100kb in the study by Park et al. 2010 demonstrated a similar pattern of inter-chromosomal variation although differences in chromosome bias were observed. For example, the high incidence loss reported on chromosome 2 at 88-89Mb by Park et al. 2010 coincides with the high incidence gain reported in the current study. This finding may represent a race variant (1, 4, 8) however technical factors such as platform specificity should also be considered.

## 5.5.2 Contribution of CNV load and duplication

Comparison of the CNV properties described here with previously published examples from pathogenic cohorts (20, 24) and a sample of referred tests for fetal demise has demonstrated differences in the CNV length, proportion of gains and loss and CNV load.

The contribution of CNV length in pathogenic studies differs to that observed in the benign cohort reported here. Sanders et al. 2011 and Liu et al. 2013 reported a load of large CNV calls >500kb in the pathogenic population for autism and Parkinson disease respectively (20, 24). Likewise in a large study of patients with developmental delay, Cooper et al. 2011 reported that CNV load was apparent from CNVs >250kb (21). Analysis of copy number variation in fetal demise in the study here shows an increased load of CNV >500kb (24%) compared to 2.6% and CNV >1Mb was significantly increased, accounting for 13.76% in fetal demise compared to 0.22% in the healthy female cohort. In addition the average CNV length is greater in fetal demise studies and encompasses more genes when compared to CNV in the healthy cohort.

Previous published reports suggest that deletion is more frequent in pathogenic cohorts (19, 20, 24, 35, 36). However a higher contribution of duplication was observed in the pathogenic cohort, including fetal demise, when compared to that in the healthy cohort. In the CNV reported by Sanders et al. 2011, comparison against the CNV gain detected in the healthy female cohort showed a load of duplication that was apparent across all ranges of CNV length, and is emphasised in the CNV 1kb- 250kb length range. In the fetal demise cohort although only CNV

>100kb were considered due to limited microarray resolution, load of duplication is observed with the greatest proportion of gain recorded for CNV lengths of 100kb-150kb and >1Mb. The contribution of duplication to pathogenic cohorts and in particular the load in duplication measuring 1kb-250kb in CNV reported by Sanders et al. 2011, has not been demonstrated previously.

The findings here suggest that the cumulative effect of copy number change should be considered and that a level of duplication can be deleterious. However, this data may be limited by the sample size and additional comparisons are required to ascertain the contribution of duplication to pathogenic cohorts. Consideration must also be given to the different resolution of the platforms used between the studies. This may contribute to the differences in CNV length, duplication bias and start and end positions for CNV in the fetal demise cohort compared to the general population sample.

### 5.5.3 Inter-chromosomal variation of CNV

All studies demonstrated inter-chromosomal variation of CNV load. Apart from chromosome bias as a result of the contribution of high incidence CNV for chromosome 1, 2, 3 and 8, the results in the study here are comparable to the CNV in the HapMap study reported by Park et al.2010. When compared to the CNV load in the neurological cohort reported by Sanders et al. 2011 and the fetal demise cohort presented here, the chromosome bias differs. Chromosomes 10, 18 and 20 show enrichment in the CNV > 100kb reported by Sanders et al. 2011 and the fetal

demise cohort compared to the healthy cohort. This contrasts to the findings in the benign CNV reported here with the lowest proportion of CNV >100kb recorded for chromosomes 18 and 20. This finding may provide the scope for further investigation of the contribution of these chromosomes to microdeletion and microduplication syndromes and the mechanisms involved in the derivation of pathogenic CNV.

Inter-chromosome variation of CNV load may be explained in part by differences in the genomic sequence of the chromosomes. Low copy repeats (LCR) have been shown to mediate chromosomal rearrangement leading to segments of duplication or deletion (23, 37, 38) and contribute to pathogenic CNV (16, 19-22, 39). In a study of the proportion of low copy repeats in chromosomes, Zhang et al. 2006 identified differences among the chromosomes ranging from 0.3% of chromosomal DNA for chromosome 18 to 8.2% for chromosome 16 (34). The comparatively low incidence of CNV and relative paucity of CNV >100kb in benign CNV on chromosome 18 may be due to the scarcity of low copy repeats. Conversely chromosomes 7 and 10 had a high proportion of CNV > 100kb in the study by Sanders et al. 2011 this being consistent with the level of segmental duplication, 6.1% and 4.8% respectively (34). However in the study here, CNV mediated by low copy repeats is estimated to account for 26.5% of CNV, the remainder formed by mechanisms not requiring large blocks of homologous sequences. Whilst the presence and location of segmental duplication may contribute to the formation and distribution of benign and pathogenic CNV, the findings here suggest that the source of inter and intra-chromosomal variation is likely to be multi-factorial and

may reflect a combination of biological factors including selection pressures (5, 6, 36, 40), gene content (5, 11, 36, 40) and DNA composition (7, 38, 41).

## 5.5.4 Gene ontology

PANTHER analysis tool (28) was used to investigate the genes encompassed in benign CNV observed in this study. In a study of CNV detected in HapMap individuals, Cooper et al. 2007 described an enrichment of genes involved in sensory perception and immune response in CNV associated with segmental duplication. CNV not overlapping segmental duplication were noted to be associated with genes involved in signalling pathways, although an earlier study by Redon et al. 2006 reported a paucity of cell signalling genes in benign CNV (11). In the current study genes involved in signalling pathways accounted for one –third of CNV and genes involved in immunological roles accounted for one-third of duplications (www.Panther). As described previously in chapter 4, the latter finding is consistent with the DNA structure and subsequent mechanism of derivation of CNV associated with the hyper-variable regions of the immunoglobulin genes. It is of interest that gene with undefined biological function were involved in 70/164 (42.6%) of deletions. This finding highlights here that the role of a large proportion of benign CNV is yet to be determined.

## 5.5.5 Clinical application of general population studies

### 5.5.5.1 Evidence from rare benign CNV

Documentation of rare variants in case and control studies is required to establish sufficient evidence to ascertain the role in disease. Interpretation of the clinical significance of CNV can be challenging. CNV are considered to be of uncertain clinical significance where there is limited evidence in the literature, where the literature has not established an association or where it is not registered in copy number variant repositories. In the absence of any evidence in the literature or databases the CNV is categorised as unknown (42-44).

An example of this in the cohort reported here is a 161kb duplication at 12p13.33 involving *CACN2D4* (exons 5-37) and *ADIPOR* (exons 2-8). Disruption within genes including duplication breakpoints may result in loss of function of the gene (45). *CACNA2D4* is a regulatory subunit that forms a complex with *CACNA1C*, a nearby gene (46). Van Den Bossche et al. 2012 proposed that deletion in *CACNA2D4* alters the VDCC complex formed and the concentration of free intra-cellular calcium (Ca2+) impacting calcium homeostasis (46).

There are few reports of a deletion at 12p13.33 presenting with a clinically variable phenotype and included mental retardation, developmental delay, facial dysmorphic features and congenital anomalies (47, 48). Rooryck et al. 2009 proposed *WNT5B, ADIPOR2, CACN2D4, DCP1B* and *CACNA1C* as candidate genes in a microdeletion syndrome. Deletions of part of *CACN2D4* with a minimal critical region involving exons 19-26, have been implicated in studies of ASD and

psychiatric disorders (46, 48). However at the time of investigation there were no reports of CNV duplication involving *CACN2D4*. Although late onset psychiatric disorder cannot be excluded in the individual reported here, the detection and FISH confirmation of duplication in the current study supports a classification of rare benign variant.

### 5.5.5.2 Contribution to CNV classification

Candidate genes were identified in a recent genome-wide association study of autism spectrum disorder (ASD), by the Autism Genome Project Consortium (49). Holt et al. (2012) in a study of the role of fusion genes as a mechanism in ASD investigated a copy number gain resulting in fusion of *POL1RA* and *REEP1* genes. Similarly, in an investigation of rare variants associated with ASD, Sanders et al. 2011 detected a gain at 2q11.2 (*POL1RA* and *REEP1*) in 2 probands and 3 unaffected parents in a study of the Simon Simplex Collection. This duplication was detected in 1/64 in the current study of healthy female individuals. Review of the UCSC browser with the ISCA and DGV tracks applied revealed several cases of a 200kb gain at 2p11.2 involving *POLR1A* to *REEP1* genes in the uncertain and unknown categories. The findings here in a healthy female cohort, provide further supportive evidence for this CNV be considered a rare benign variant.

### 5.5.6 Homozygous and heterozygous deletions

The cataloguing of regions of duplication and deletion with an apparent benign outcome is ongoing (2, 3, 29), but regions of nullisomy that can be tolerated are poorly documented. Hemizygous deletions associated with a benign outcome have been shown previously to have clinical significance as a homozygous deletion (50). An example of this is the common 2q13 heterozygous deletion involving *NPHP1* (Nephrocystin 1) observed as a private variant in the cohort reported here (1/64), but when present as a homozygous deletion is associated with juvenile nephronophthisis and Joubert syndrome (51). Homozygous deletion 2q13 involving *NPHP1* was detected in one fetal demise sample.

The contribution of nullisomy was investigated further in the fetal demise cohort to ascertain regions of interest. The CNV in the fetal demise cohort that are shared in the healthy cohort (n=87) were evaluated for regions of nullisomy. It was noted that homozygous deletions occurred in 7% of calls (6/87) that are observed as heterozygous deletions in the general population sample. Of these a 193kb homozygous deletion at 1q21.1 involving *NBPF25P* (neuroblastoma breakpoint family) represented 50% of these (3/6). *NBPF* genes function as DNA-binding transcription factors and are located in the cell nucleus (52). The encoded proteins contain 'highly conserved domains of unknown function (*DUF1220*)(53) and are reported to be expressed in fetal brain and fetal sympathetic nervous tissue (53, 54). Although present in the general population, Dumas et al. 2012 proposed an association *DUF1220* proteins and *NBPF* genes with normal and pathological brain

size and the clinical presentation of microcephaly in the 1q21.1 microdeletion syndrome (54).

Although this CNV is observed as a heterozygous deletion in 7/64 of the general population sample reported here and reported as a benign CNV on both ISCA and DGV websites, there are no published reports of nullisomy for the region. This finding may represent a benign CNV which becomes of clinical significance as a homozygous deletion. Further studies of fetal demise are required to investigate this hypothesis. However the question can be raised, are there regions of the genome where nullisomy (homozygous CNV deletion) represent a significant risk in fetal development?

## 5.6 Conclusion

The findings here represent a comprehensive investigation of CNV in a cohort of healthy Western European descendent females. The CNV load, CNV length and contribution of duplication and deletion are investigated for chromosomes 1-22. Comparison with previous studies of control cohorts, in particular those of HapMap individuals, has demonstrated consistency in CNV properties, with enrichment of CNV <50kb and a predominance of deletion in CNV <100kb. CNV are distributed at variable frequency over the entire chromosome although enrichment associated with gross morphological structures was not apparent for all chromosomes.

CNV reflect the functional attributes of DNA. In the healthy cohort CNV encompass immunity and genes with roles in signal transduction, although the roles of many genes encompassed by benign CNV are yet to be defined. The role of genomic sequence is apparent with all putative breakpoints in CNV >100 kb located either within or near segmental duplications or repetitive elements. There is no apparent difference for the occurrence of CNVR > 100kb with breakpoints in segmental duplication or repetitive elements, however high incidence common CNV associate more frequently with segmental duplication and low frequency and private variants are associated with repetitive elements. Together these contributing factors suggest that the properties of CNV in the healthy population including CNV length, chromosomal landscape, copy number, and chromosomal variation incorporates structural, functional and molecular attributes of DNA.

The CNV properties determined in this cohort are compared with pathogenic populations to ascertain similarities and differences. This study highlighted a comparatively higher CNV load, in particular duplication and larger average CNV length in the clinical cohort. In particular this comparison identified an enrichment of duplications in CNV 1kb-250kb, a CNV size not previously considered of clinical significance in studies of CNV load in pathogenic cohorts. Further comparisons of general population cohorts with pathogenic cohorts are recommended to confirm this finding and determine the relevance of CNV load and duplication.

To evaluate a clinical application of general population cohorts, the CNV output from the cohort here is analysed for novel CNV. Several were identified inferring the dynamic process of CNV formation. This finding highlights the need to review

general population cohorts and to continue to document benign CNV in registered databases. This is necessary to establish evidence to assist in assigning clinical significance and in refining the candidate genes in CNV that involve several genes.

## 5.7 References

1. Matsuzaki H, Wang PH, Hu J, Rava R, Fu GK. High resolution discovery and confirmation of copy number variants in 90 Yoruba Nigerians. Genome Biol. 2009;10(11):R125.
2. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, et al. Large-scale copy number polymorphism in the human genome. Science. 2004;305(5683):525-8.
3. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, et al. Detection of large-scale variation in the human genome. Nat Genet. 2004;36(9):949-51.
4. Park H, Kim JI, Ju YS, Gokcumen O, Mills RE, Kim S, et al. Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. Nat Genet. 2010;42(5):400-5.
5. Cooper GM, Nickerson DA, Eichler EE. Mutational and selective effects on copy-number variants in the human genome. Nat Genet. 2007;39(7 Suppl):S22-9.
6. Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467(7311):52-8.
7. Hastings PJ, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. Nat Rev Genet. 2009;10(8):551-64.
8. Chia NL, Bryce M, Hickman PE, Potter JM, Glasgow N, Koerbin G, et al. High-resolution SNP microarray investigation of copy number variations on chromosome 18 in a control cohort. Cytogenet Genome Res. 2013; 141:16-25.
9. Hastings PJ, Ira G, Lupski JR. A microhomology-mediated break-induced replication model for the origin of human copy number variation. PLoS Genet. 2009;5(1):e1000327.
10. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, et al. Origins and functional impact of copy number variation in the human genome. Nature. 2010;464(7289):704-12.
11. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number in the human genome. Nature. 2006;444(7118):444-54.
12. Dellinger AE, Saw SM, Goh LK, Seielstad M, Young TL, Li YJ. Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. Nucleic Acids Res. 2010;38(9):e105.
13. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. 2007;17(11):1665-74.
14. Pinto D, Darvishi K, Shi X, Rajan D, Rigler D, Fitzgerald T, et al. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nat Biotechnol. 2011;29(6):512-20.
15. Pinto D, Marshall C, Feuk L, Scherer SW. Copy-number variation in control population cohorts. Hum Mol Genet. 2007;16 Spec No. 2:R168-73.
16. Girirajan S, Rosenfeld JA, Coe BP, Parikh S, Friedman N, Goldstein A, et al. Phenotypic heterogeneity of genomic disorders and rare copy-number variants. N Engl J Med. 2012;367(14):1321-31.
17. Bruno DL, Ganesamoorthy D, Schoumans J, Bankier A, Coman D, Delatycki M, et al. Detection of cryptic pathogenic copy number variations and constitutional loss of heterozygosity using high resolution SNP microarray analysis in 117 patients referred for cytogenetic analysis and impact on clinical practice. J Med Genet. 2009;46(2):123-31.
18. Miller DT, Adam MP, Aradhya S, Biesecker LG, Brothman AR, Carter NP, et al. Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. Am J Hum Genet. 2010;86(5):749-64.

19. Kaminsky EB, Kaul V, Paschall J, Church DM, Bunke B, Kunig D, et al. An evidence-based approach to establish the functional and clinical significance of copy number variants in intellectual and developmental disabilities. Genet Med. 2011;13(9):777-84.
20. Sanders SJ, Ercan-Sencicek AG, Hus V, Luo R, Murtha MT, Moreno-De-Luca D, et al. Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. Neuron. 2011;70(5):863-85.
21. Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, et al. A copy number variation morbidity map of developmental delay. Nat Genet. 2011;43(9):838-46.
22. Girirajan S, Eichler EE. Phenotypic variability and genetic susceptibility to genomic disorders. Hum Mol Genet. 2010;19(R2):R176-87.
23. Itsara A, Cooper GM, Baker C, Girirajan S, Li J, Absher D, et al. Population analysis of large copy number variants and hotspots of human genetic disease. Am J Hum Genet. 2009;84(2):148-61.
24. Liu X, Cheng R, Ye X, Verbitsky M, Kisselev S, Mejia-Santana H, et al. Increased Rate of Sporadic and Recurrent Rare Genic Copy Number Variants in Parkinson's Disease Among Ashkenazi Jews. Mol Genet Genomic Med. 2013;1(3):142-54.
25. Brothman AR, Persons DL, Shaffer LG. Nomenclature evolution: Changes in the ISCN from the 2005 to the 2009 edition. Cytogenet Genome Res. 2009;127(1):1-4.
26. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002;12(6):996-1006.
27. Knight SJ, Flint J. Perfect endings: a review of subtelomeric probes and their use in clinical diagnosis. J Med Genet. 2000 37(6):401-9.
28. Mi H, Muruganujan A, Casagrande JT, Thomas PD. Large-scale gene function analysis with the PANTHER classification system. Nat Protoc. 2013;8(8):1551-66.
29. Shaikh TH, Gai X, Perin JC, Glessner JT, Xie H, Murphy K, et al. High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications. Genome Res. 2009;19(9):1682-90.
30. Scherer SW, Lee C, Birney E, Altshuler DM, Eichler EE, Carter NP, et al. Challenges and standards in integrating surveys of structural variation. Nat Genet. 2007;39(7 Suppl):S7-15.
31. Pang AW, MacDonald JR, Pinto D, Wei J, Rafiq MA, Conrad DF, et al. Towards a comprehensive structural variation map of an individual human genome. Genome Biol. 2010 11(5):R52.
32. Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, et al. Copy number variation: new insights in genome diversity. Genome Res. 2006;16(8):949-61.
33. Knight SJ, Lese CM, Precht KS, Kuc J, Ning Y, Lucas S, et al. An optimized set of human telomere clones for studying telomere integrity and architecture. Am J Hum Genet. 2000;67(2):320-32.
34. Zhang L, Lu HH, Chung WY, Yang J, Li WH. Patterns of segmental duplication in the human genome. Mol Biol Evol. 2005;22(1):135-41.
35. Conrad DF, Hurles ME. The population genetics of structural variation. Nat Genet. 2007;39(7 Suppl):S30-6.
36. de Smith AJ, Walters RG, Froguel P, Blakemore AI. Human genes involved in copy number variation: mechanisms of origin, functional effects and implications for disease. Cytogenet Genome Res. 2008;123(1-4):17-26.
37. Gu W, Zhang F, Lupski JR. Mechanisms for human genomic rearrangements. Pathogenetics. 2008;1(1):4.
38. Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, et al. Segmental duplications and copy-number variation in the human genome. Am J Hum Genet. 2005;77(1):78-88.

39. Sharp AJ, Mefford HC, Li K, Baker C, Skinner C, Stevenson RE, et al. A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. Nat Genet. 2008;40(3):322-8.
40. Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK. A high-resolution survey of deletion polymorphism in the human genome. Nat Genet. 2006;38(1):75-81.
41. Bacolla A, Cooper DN, Vasquez KM. DNA structure matters. Genome Med.2013 5(6):51.
42. Riggs ER, Jackson L, Miller DT, Van Vooren S. Phenotypic information in genomic variant databases enhances clinical care and research: the International Standards for Cytogenomic Arrays Consortium experience. Hum Mutat. 2012;33(5):787-96.
43. Riggs ER, Church DM, Hanson K, Horner VL, Kaminsky EB, Kuhn RM, et al. Towards an evidence-based process for the clinical interpretation of copy number variation. Clin Genet. 2012;81(5):403-12.
44. Kearney HM, Thorland EC, Brown KK, Quintero-Rivera F, South ST. American College of Medical Genetics standards and guidelines for interpretation and reporting of postnatal constitutional copy number variants. Genet Med. 2011;13(7):680-5.
45. Feuk L, Marshall CR, Wintle RF, Scherer SW. Structural variants: changing the landscape of chromosomes and design of disease studies. Hum Mol Genet. 2006;15 Spec No 1:R57-66.
46. Van Den Bossche MJ, Strazisar M, De Bruyne S, Bervoets C, Lenaerts AS, De Zutter S, et al. Identification of a CACNA2D4 deletion in late onset bipolar disorder patients and implications for the involvement of voltage-dependent calcium channels in psychiatric disorders. Am J Med Genet B Neuropsychiatr Genet. 2012;159B(4):465-75.
47. Rooryck C, Stef M, Burgelin I, Simon D, Souakri N, Thambo JB, et al. 2.3 Mb terminal deletion in 12p13.33 associated with oculoauriculovertebral spectrum and evaluation of WNT5B as a candidate gene. Eur J Med Genet. 2009;52(6):446-9.
48. Abdelmoity AT, Hall JJ, Bittel DC, Yu S. 1.39 Mb inherited interstitial deletion in 12p13.33 associated with developmental delay. Eur J Med Genet. 2011;54(2):198-203.
49. Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, Regan R, et al. Functional impact of global rare copy number variation in autism spectrum disorders. Nature. 2010;466(7304):368-72.
50. Alkuraya FS. Autozygome decoded. Genet Med. 2010;12(12):765-71.
51. Hildebrandt F, Otto E, Rensing C, Nothwang HG, Vollmer M, Adolphs J, et al. A novel gene encoding an SH3 domain protein is mutated in nephronophthisis type 1. Nat Genet. 1997;17(2):149-53.
52. Zhou F, Xing Y, Xu X, Yang Y, Zhang J, Ma Z, et al. NBPF is a potential DNA-binding transcription factor that is directly regulated by NF-kappaB. Int J Biochem Cell Biol. 2013;45(11):2479-90.
53. Diskin SJ, Hou C, Glessner JT, Attiyeh EF, Laudenslager M, Bosse K, et al. Copy number variation at 1q21.1 associated with neuroblastoma. Nature. 2009;459(7249):987-91.
54. Dumas LJ, O'Bleness MS, Davis JM, Dickens CM, Anderson N, Keeney JG, et al. DUF1220-domain copy number implicated in human brain-size pathology and evolution. Am J Hum Genet. 2012;91(3):444-54.

# CHAPTER 6

# A Study of Long Stretches of Homozygosity in a Cohort of Healthy Australian Women

## 6.1 Abstract

Contiguous stretches of homozygosity are not uncommon in the healthy population representing common haplotypes and evidence of shared ancestry. Previous reports have described long contiguous stretched of homozygosity (LCSH) events from transformed cell lines and few studies detail the properties in healthy general populations.

In the study here, DNA was extracted from the whole blood of 64 randomly selected healthy females, which was analysed using Illumina Omni1-Quad to investigate the properties and incidence of LCSH in a Western European descendent population. A total of 765 events of stretches of homozygosity >1Mb were recorded with an average of 11.8 events of LCSH per person. All individuals and all autosomes are represented with regions of homozygosity measuring 1 Mb to 1.5 Mb. A non-random distribution of LCSH is observed with a high frequency of events in some regions. Comparison with published population studies reveals that 26% of LCSH in this cohort may be attributed to population specific LCSH and remnants of ancestral history. Two individuals with extensive homozygosity and LCSH properties consistent with close parental relationship are identified in the cohort, thus reflecting the diversity of the Western European descendent population. This suggests that consanguinity may be observed as an incidental finding in clinical referrals to diagnostic laboratories.

This study identifies common LCSH and demonstrates thresholds and analytical criteria that can be applied when interpreting events of LCSH in clinical applications.

## 6.2 Introduction

LCSH events are ubiquitous in the human genome. Population studies have shown that these regions are a genetic reminder of an individual's ancestral history (1-8), and may be indicative of geographic isolation, historical population bottlenecks (5, 9, 10) or consanguinity from religious or cultural choices (11, 12).

Homozygosity for a region that is "identical by descent" occurs when an individual inherits the same allelic segment resulting in homozygosity for that segment (1, 3, 4, 11, 13). Family members frequently have regions of the same haplotypes. Consistent with this is the observation that the segment size and overall contribution of LCSH to an individual's genome is indicative of the closeness of parental relatedness. More of the genome is shared between parents of close ancestral relationship compared to the outbred population. Large segments of shared haplotypes are reduced in size by meiotic recombination and are smaller in more distantly related individuals (11).

The clinical significance of regions of homozygosity has been well described with early studies demonstrating the increased risk of inheritance of autosomal recessive gene mutation (5, 9, 11, 14). More recently studies of complex diseases such as hypertension, cardiovascular disease and diabetes have been associated with LCSH (4, 7, 15-17). Extrapolation of this is the use of homozygosity mapping to identify candidate genes associated with a clinical phenotype or recessive Mendelian diseases (5).

Some segments of LCSH may occur by chance alone, representing "identity by state" (11). These regions represent the presence of haplotypes with high population frequency and may involve SNPs of low heterozygosity scores (level of heterozygosity in the population) (11), and low mutation or recombination rates (2, 3). This culminates in the presence of long stretches of homozygosity and may account for a proportion of the LCSH in outbred populations (2, 3, 11).

Tracts of extended homozygosity observed in population studies are reported to be co-localised in chromosomal distribution and occur non-randomly (2, 4-7, 18). Nothnagel et al. 2010 reviewed SNP genotyping data of European subpopulations and demonstrated LCSH events occurring at high frequency. These common LCSH events were shown to be independent of linkage disequilibrium (LD) and reflect patterns of meiotic recombination that are genome-wide (6). Kirin et al. 2010 reviewed LCSH length and concluded that the length of LCSH and prevalence varies among populations and correlates with migration patterns (1).

The implementation of SNP microarray technology to routine diagnostic laboratories has had a major impact on the detection of clinically significant copy number change at a resolution not detectable by conventional karyotyping methods. Consistent with this is the availability of homozygosity software programs and subsequent detection of LCSH from the genotyping data. This has culminated in the detection of regions of LCSH in clinical referrals to diagnostic laboratories. The interpretation of LCSH in this setting can be challenging due to the paucity of current knowledge of the prevalence and properties of LCSH in the healthy Western European descendent population.

This study describes the prevalence and properties of LCSH in a cohort of female Western European descendents. A systematic analysis of LCSH is performed to explore the contribution of LCSH to the genome in this cohort. Firstly the properties of LCSH are investigated for length, contribution to the genome and chromosomal distribution. Next the data is analysed to determine the thresholds applicable to analysis in diagnostic laboratories and thirdly, the incidental findings of LCSH consistent with consanguinity are reviewed.

## 6.3 Materials and Methods

## 6. 3.1 Sample collection

The female cohort presented here represents a Western European descendent population and was recruited from the "Aussie Normal" Collection in Canberra, Australia. This is a community based study of Australians recruited from the Australian Electoral Roll. Written consent was provided by the participants and ethics approval was initiated through the Australian National University Human Research Ethics committee. All participants are screened for complex disease association and status of general health recorded. Information regarding family history was not documented at the time of consultation.

## 6.3.2 Microarray investigation

DNA was extracted using standard protocols from blood samples collected from 64 females who are enrolled in the Aussie Normal Collection. Genome-wide analysis was performed using Illumina Human Omni1-Quad. This platform has a median probe interspacing of 1.2kb and overall resolution of 5 kb (Illumina, Inc.). There are 1,140,419 markers, comprised of 1,048,713 SNP markers and 91,706 intensity only "CNV markers". The standard Illumina clusterfile was used for determination of genotypes.

Regions of homozygosity were detected using the proprietary software, CNV Partition v2.3.4 (Illumina, Inc.). This algorithm utilizes genotype and allele frequencies. It predicts an LCSH event based on SNP allele frequencies and calculates the expectation of homozygosity for a number of consecutive SNPs. The genotypes along each chromosome are scanned and "windows" with loss of heterozygosity are identified. The likelihood of an LCSH event is predicted using a log odds ratio of the region being homozygous compared to not being homozygous (Technical Note, Illumina, Inc. 2010).

## 6.3.3 Definition of LCSH

Stringent quality control (QC) criteria were applied to all calls. For the purpose of this study an LCSH is defined as a region of extended homozygosity of consecutive SNP markers, at an average density of 1 SNP for every 5kb. For consistency this criteria was selected to be in accordance with previous reports of Gibson et al.

2006 (3). To ascertain the SNP density in an LCSH, a calculation was applied to each LCSH.

$$\text{SNP density} = \frac{\text{Number of SNP markers}}{(\text{LCSH Length (kb)} \times 0.001)}$$

Next, although the number of SNP markers within an LCSH may meet the numerical criteria, failure of these SNP markers for genotyping may result in false positive calls. To overcome this each LCSH was evaluated for the number of failed SNP markers that appear as "no call". In addition a visual inspection of plots was performed and LCSH events with "scattering" of markers were excluded from further analysis. Together these QC metrics ensure robustness of LCSH calls.

## 6.3.4 Contribution of LCSH in the cohort

The contribution of autozygosity to an individual's genome is defined as the cumulative length of the individual's autozygosity as a proportion of the overall autosome length. This is calculated by:

$$\%\ \text{Genome} = (\Sigma\ \text{LCSH events} / \Sigma\ \text{Autosome Length}) \times 100$$

where Σ LCSH represents the sum of the individual's LCSH > 1 Mb on chromosomes 1-22 and Σ autosome represents the total genome autosomal length covered by SNP markers (7, 13, 19). Here the Σ autosome is 2,699,116,387 bps representing the sum of the autosome covered by SNP markers by the Illumina Omni 1-Quad platform (13).

### 6.3.5 Website investigations

The physical position of LCSH events are mapped according to NCBI36/hg18 and the UCSC genome browser (http://genome.ucsc.edu/ hg18) with custom tracks applied was used for demonstration of overlap. Statistical analysis was done using Graphpad Quick Calcs (www.graphpad.com).

## 6.4 Results

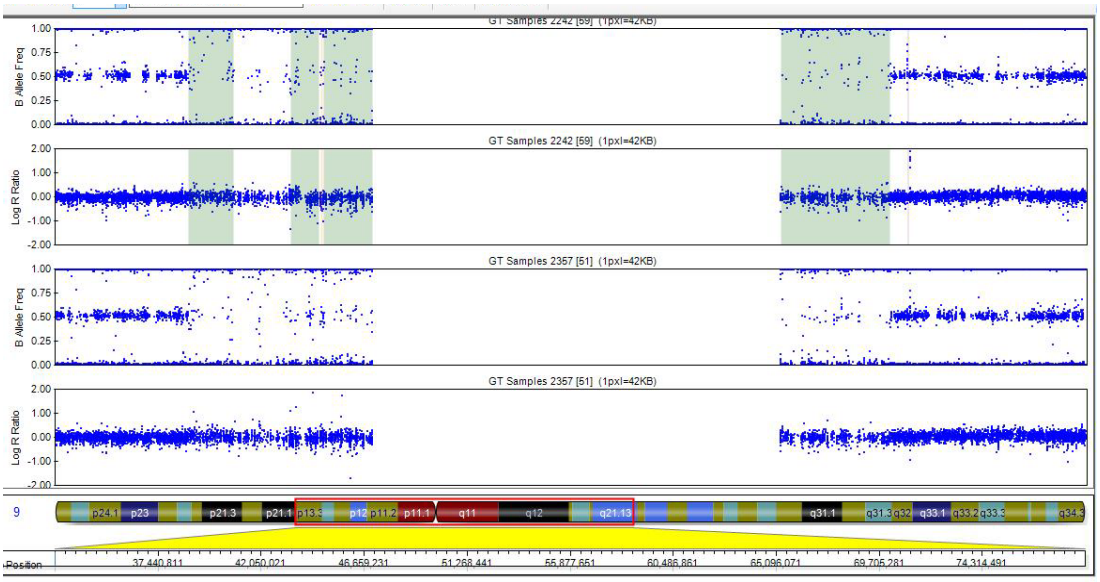
LCSH is observed as long stretches of consecutive SNP markers that record a homozygous genotype. For the purpose of this study the lower detection limit of 1Mb is applied. This is consistent with previous studies (3) and minimizes detection of false positive events due to low SNP frequency. Stringent QC metrics are applied to the LCSH events to ensure call confidence. The SNP density is calculated for each LCSH event that is detected by the proprietary software. A SNP density of >0.2 which equates to 1 SNP that achieved a genotype result for every 5 kb are included for analysis. Initially there are 1329 events of autosomal LCSH and after QC parameters are met a total of 765 events of LCSH >1Mb are included in the study. This amounts to an average of 11.8 per person in the cohort of females of Western European descent.

### 6.4.1 QC Metrics and putative false calls

The events that failed QC parameters (n=564) ranged in length up to 8Mb and 84.4% (476/564) are less than 2.5 Mb. These events were recorded on all chromosomes with the exception of chromosomes 12, 17, 19, 20, and 21. Common non-random events (occurrence in 2 or more individuals) predominated, accounting for 97% (547/564) of putative false positive events. An example of this is chromosome 9 where LCSH in the region flanking the centromere was called for all 64 individuals (Figure 1).

The association of these events with chromosomal landscape was investigated by mapping the number and physical position of LCSH events in consecutive 5Mb windows for the entire chromosome length. LCSH that failed QC metrics are found to be located predominantly in pericentromeric regions (Figure 4).

a)




b)

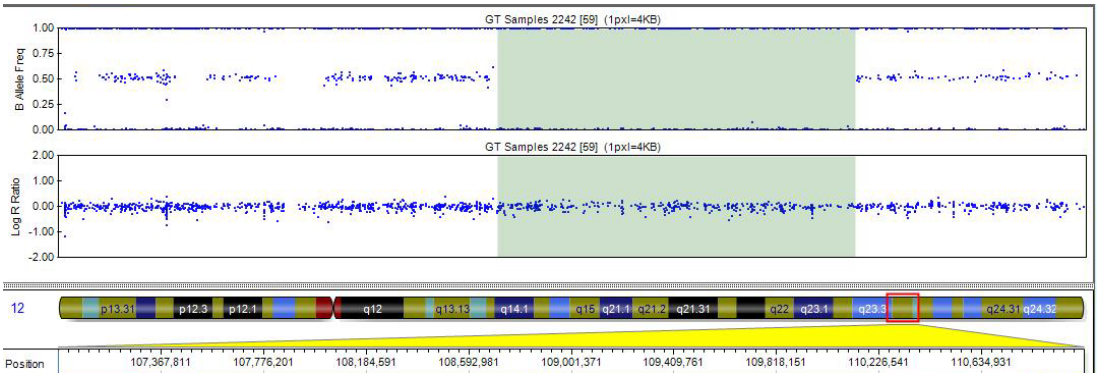



**Figure 1.** Plot of LCSH event is displayed a) the peri-centric region of chromosome 9. The average SNP density for this region is 0.05 and the region visually appeared as “scattered”, consisting of failed (“No call”) SNP markers, b) a 1.4Mb LCSH event on chromosome 12 with SNP density 0.277.

## 6.4.2 Incidence of LCSH

The average contribution of LCSH to the genome of individuals in the cohort presented here is calculated to be 0.54% and the mean cumulative total length per individual is 14.7 Mb ± 6 Mb. There is an average of 11.8 events of LCSH per individual with the mean LCSH length of 1.55 Mb ± 1.48 Mb.

There are 2 samples, AN2327 and AN2328 that recorded a contribution of 7% and 4% respectively. This finding is suggestive of an incidental finding reflecting a level of parental relatedness.

## 6.4.3 Properties of LCSH in cohort

The length properties of LCSH events were explored to characterise the contribution of LCSH in this cohort. It was observed that the number of events decreased with increasing length of LCSH and 94% of events measure 1-2.5Mb in length (Table 1).

**Table 1**. The stratification of LCSH according to length

| LCSH Length (Mb) | |
|---|---|
| **1.0-1.5** | 570 |
| **1.5-2.5** | 149 |
| **2.5-5.0** | 32 |
| **5.0-10.0** | 6 |
| **10.0-15.0** | 5 |
| **>15.0** | 3 |
| **Total** | **765** |

The number and length of LCSH events were then scored for each individual to determine the contribution to individual genomes. All 64 females recorded LCSH 1-2Mb in length and 20 individuals recorded LCSH 2-3Mb (Figure 2). The incidence declines with increasing LCSH size and 4/62 individuals (excluding AN2327 and AN2328), scored single events of LCSH 3-10Mb. LCSH >10Mb were only recorded in the two individuals that had genomic contribution consistent with consanguinity.

The number of individuals with LCSH events was statistically compared across a range of lengths to investigate if a lower limit threshold for LCSH length can be determined in this cohort. The results show a significant difference between the number of individuals with events less than 3Mb when compared to LCSH larger than 3Mb ($p<0.0001$; Fisher's Exact test) (Table 3). Supporting this is the finding of a single event in each of 3 individuals and 2 events in one individual in the length category 3-10Mb.

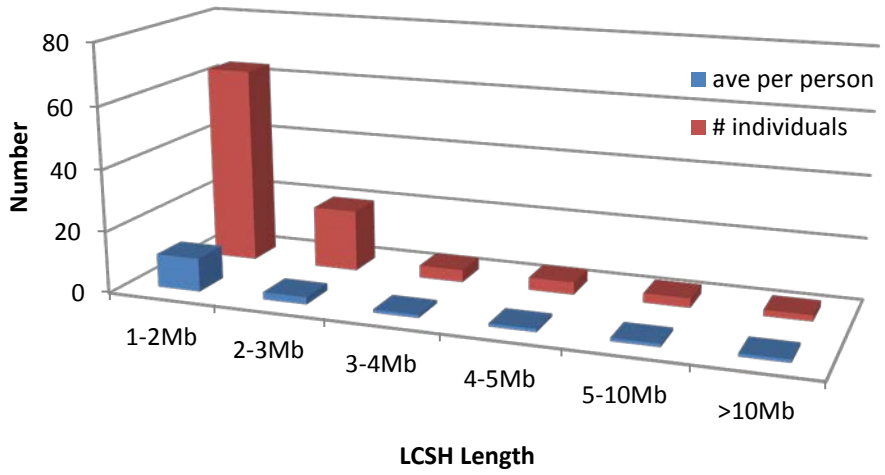



**Figure 2.** Stratification of the length of LCSH per person. The number of LCSH events per person is shown for each size range and can be compared to the number of individuals with these events. The number of LCSH events and the number of individuals with events decrease with increasing LCSH length.

## 6.4.4 Chromosomal landscape

The number and length of LCSH events is recorded for each chromosome (Figure 3). Firstly, with respect to the number of LCSH events, inter-chromosomal variation not consistent with the genomic length of each chromosome was observed. Chromosomes 18 and 21 had the least, which is in accordance with the copy number data presented previously. Secondly the number of events within size categories varied among the chromosomes. Although all chromosomes recorded LCSH events 1-1.5 Mb, the contribution of LCSH > 1.5 Mb varied among chromosomes.

The number and physical position of LCSH is mapped along each chromosome in adjacent 5Mb regions and all chromosomes showed LCSH distributed for the entire length (Figure 4). The intra-chromosomal variation for LCSH distribution is apparent. There is no evidence of enrichment of LCSH events with chromosome morphological structures such as centromeres and telomeres, for LCSH events that fulfill QC metrics.

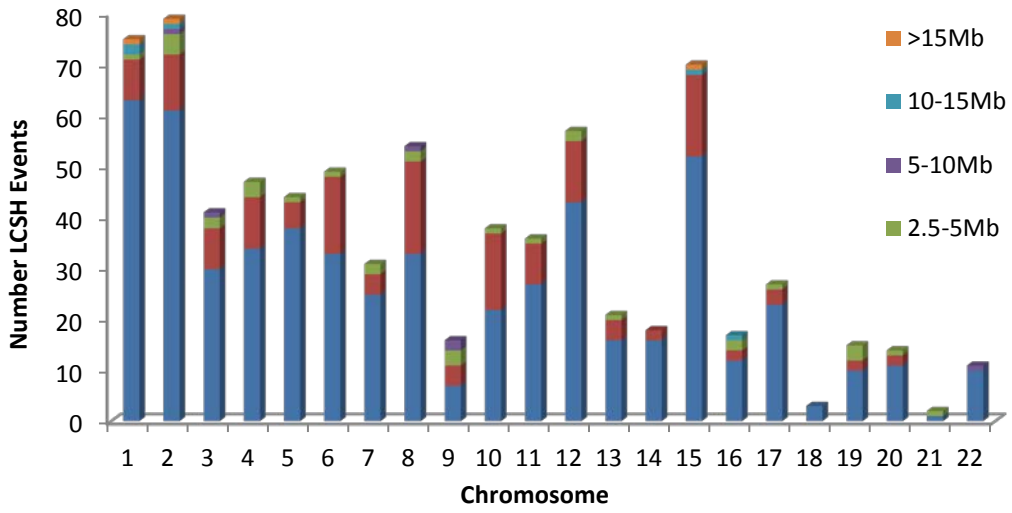



**Figure 3**. The number and lengths of autosomal LCSH events.

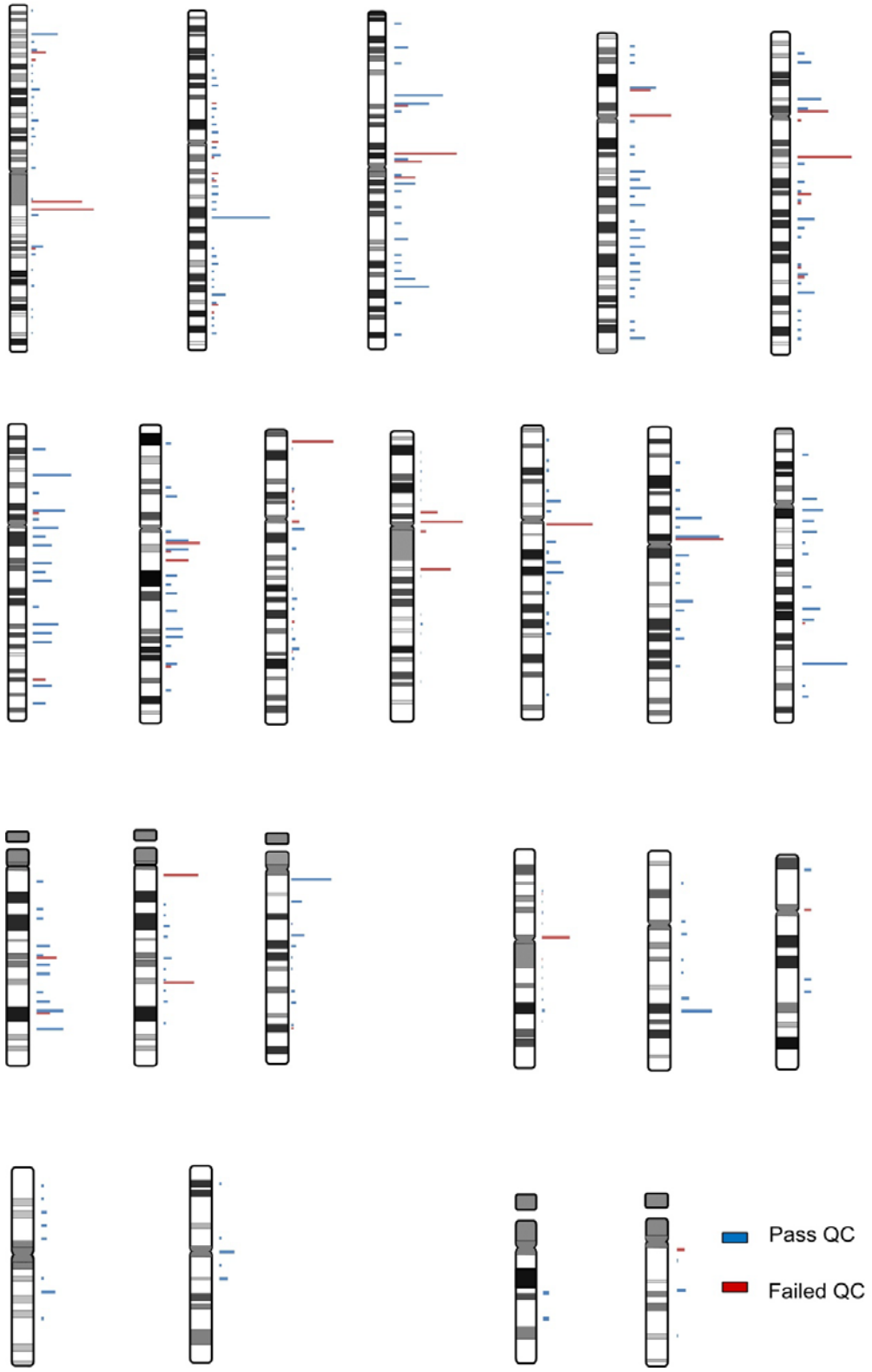



**Figure 4**. Chromosome landscape for LCSH in this cohort. The size of the bar correlates with the number of individuals with LCSH in the region.

## 6.4.5 Non-random distribution of LCSH events

To investigate the overlap of LCSH events the physical position for each LCSH according to NCBI36/hg18 is plotted. The genomic position is plotted on the Y axis and individuals with the corresponding LCSH event on the X axis (Figure 5). Common LCSH events are observed for all chromosomes with the exception of chromosome 18 and 21. All overlapping events are less than 2 Mb in length.

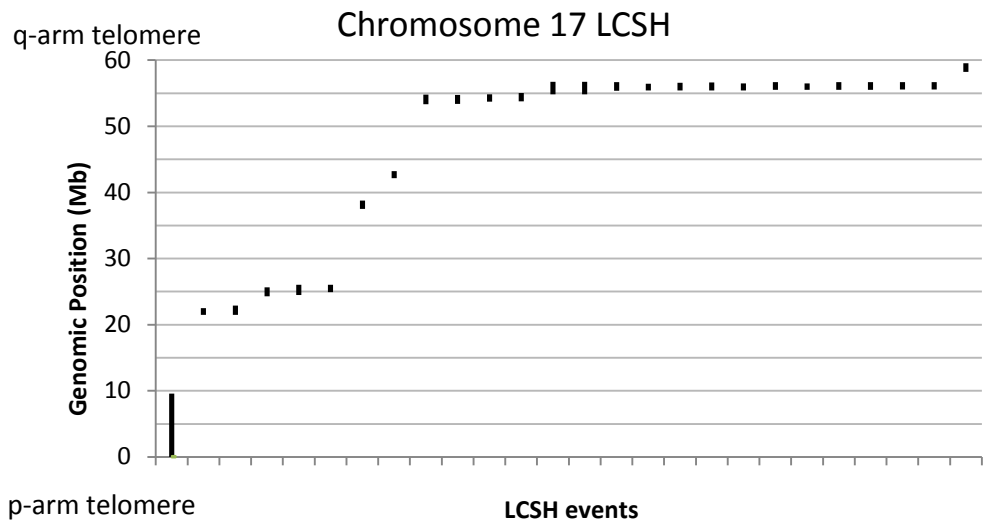



**Figure 5.** LCSH events are displayed for chromosome 17. The Y axis represents the genomic position and LCSH events in 24 individuals are displayed on the X axis. The size of the bar correlates with the LCSH length. There are 26 LCSH events detected on chromosome 17 in 24 individuals (X axis) and a common LCSH event is recorded in 13 individuals at 54.8-56.7Mb.

## 6.4.6 Comparison with population studies

Recent population studies have demonstrated non-random distribution of LCSH within populations and a correlation of the location and size of LCSH based on ancestral geographical backgrounds (1, 5, 6, 8). To investigate this in the cohort here, the LCSH events were compared to common LCSH events reported by Li et al. 2006 and Nothnagel et al. 2010 (4, 6). In the study by Li et al. 2006 the authors investigated a large cohort of the Han Chinese and Taiwanese aborigines, and compared these LCSH events to a cohort of Caucasian individuals registered on the Affymetrix database (4). The European population reported by Nothnagel et al. 2010, included 23 subpopulations (6).

Comparison of the LCSH events in the current study identified 46 LCSH regions that overlapped those reported by Li et al. 2006 and Nothnagel et al. 2010 (Table 2)(4, 6). There is greater concordance with the common LCSH events reported in the European population by Nothnagel et al. 2010 accounting for 41/46 (89%). Furthermore a higher frequency of shared calls is observed for chromosomes 1, 2, 3 and 4, an observation also reported by Nothnagel et al. 2010. The shared LCSH events are present at a frequency of 1-26 individuals and represent 204/765 (26.6%) of LCSH events in the Western European descendent population reported here.

**Table 2.** LCSH events that are shared with published population studies

| Chromosome | Start | End | Length | Incidence n=64 | Reference | Population |
|---|---|---|---|---|---|---|
| 1 | 35151706 | 36522512 | 1370806 | 18 | Nothnagel et al.2010 | Eur |
| 1 | 48839966 | 50433998 | 1594032 | 13 | Nothnagel et al.2010 | Eur |
| 1 | 72733119 | 73935797 | 1202678 | 4 | Nothnagel et al.2010 | Eur |
| 1 | 102478412 | 103555825 | 1077413 | 3 | Nothnagel et al.2010 | Eur |
| 1 | 105827431 | 109374806 | 3547375 | 1 | Li et al.2006 | Taiwan / Caucasian |
| 2 | 82341978 | 83464352 | 1122374 | 3 | Nothnagel et al.2010 | Eur |
| 2 | 135475565 | 136757591 | 1282026 | 26 | Nothnagel et al.2010 | Eur |
| 2 | 180943821 | 182009209 | 1065388 | 1 | Li et al. 2006 | Han Chinese |
| 2 | 188848297 | 190414337 | 1566040 | 4 | Nothnagel et al.2010 | Eur |
| 2 | 194436899 | 195676295 | 1239396 | 4 | Nothnagel et al.2010 | Eur |
| 2 | 195681856 | 197202742 | 1520886 | 1 | Nothnagel et al.2010 | Eur |
| 2 | 209845733 | 211178738 | 1333005 | 1 | Nothnagel et al.2010 | Eur |
| 3 | 46957741 | 48394901 | 1437160 | 3 | Nothnagel et al.2010 | Eur |
| 3 | 50497067 | 52806956 | 2309889 | 6 | Nothnagel et al.2010 | Eur |
| 3 | 80147137 | 81263958 | 1116821 | 3 | Nothnagel et al.2010 | Eur |
| 3 | 128553826 | 131246501 | 2692675 | 1 | Nothnagel et al.2010 | Eur |
| 4 | 32986285 | 34831774 | 1845489 | 9 | Nothnagel et al.2010 | Eur |
| 4 | 81328112 | 82566682 | 1238570 | 3 | Nothnagel et al.2010 | Eur |
| 4 | 151380952 | 152392430 | 1011478 | 1 | Nothnagel et al.2010 | Eur |
| 4 | 171230346 | 172610975 | 1380629 | 1 | Nothnagel et al.2010 | Eur |
| 5 | 41474996 | 42543090 | 1068094 | 1 | Nothnagel et al.2010 | Eur |
| 5 | 130360237 | 132565129 | 2204892 | 7 | Nothnagel et al.2010 | Eur |
| 6 | 34233259 | 35376511 | 1143252 | 1 | Nothnagel et al.2010 | Eur |
| 7 | 68551022 | 69568087 | 1017065 | 1 | Nothnagel et al.2010 | Eur |
| 7 | 117427893 | 118589202 | 1161309 | 2 | Nothnagel et al.2010 | Eur |
| 8 | 60491687 | 61705133 | 1213446 | 3 | Nothnagel et al.2010 | Eur |
| 8 | 118640383 | 120176806 | 1536423 | 1 | Li et al. 2006 | Han Chinese/ Caucasian |
| 8 | 120187202 | 126984731 | 6797529 | 1 | Li et al. 2006 | Han Chinese/ Caucasian |
| 10 | 21608953 | 22637180 | 1028227 | 1 | Nothnagel et al.2010 | Eur |
| 10 | 36925752 | 38913783 | 1988031 | 6 | Nothnagel et al.2010 | Eur |
| 10 | 57390788 | 58475353 | 1084565 | 1 | Nothnagel et al.2010 | Eur |
| 10 | 68277771 | 69420603 | 1142832 | 4 | Nothnagel et al.2010 | Eur |
| 10 | 73498754 | 75363938 | 1865184 | 7 | Nothnagel et al.2010 | Eur |
| 10 | 81194136 | 82209046 | 1014910 | 1 | Li et al. 2006 | Han Chinese |
| 11 | 37478062 | 39222671 | 1744609 | 5 | Nothnagel et al.2010 | Eur |
| 11 | 104480036 | 105760553 | 1280517 | 2 | Nothnagel et al.2010 | Eur |
| 12 | 36733128 | 37917629 | 1184501 | 5 | Nothnagel et al.2010 | Eur |
| 12 | 84664036 | 85729851 | 1065815 | 2 | Nothnagel et al.2010 | Eur |

| | | | | | | |
|---|---|---|---|---|---|---|
| 12 | 86870720 | 87890894 | 1020174 | 1 | Nothnagel et al.2010 | Eur |
| 12 | 109622941 | 111747901 | 2124960 | 10 | Nothnagel et al.2010 | Eur |
| 15 | 40227236 | 42134792 | 1907556 | 10 | Nothnagel et al.2010 | Eur |
| 15 | 69708893 | 70794003 | 1085110 | 1 | Nothnagel et al.2010 | Eur |
| 16 | 45474095 | 46789566 | 1315471 | 2 | Nothnagel et al.2010 | Eur |
| 16 | 65413148 | 66871656 | 1458508 | 4 | Nothnagel et al.2010 | Eur |
| 17 | 54835580 | 56729649 | 1894069 | 13 | Nothnagel et al.2010 | Eur |
| 22 | 29863796 | 30899263 | 1035467 | 5 | Nothnagel et al.2010 | Eur |

## 6.4.7 Incidental finding

Calculation of the contribution of LCSH to individual genomes revealed 2 samples that are outliers in comparison to the cohort. The LCSH events in these individuals are increased across all LCSH length categories. AN2327 has 46 LCSH events >1mb, amounting to a genomic contribution of 7%. LCSH 1-2Mb in length accounted for 20/46 (43.4%) of events and there are 7 events >10Mb with the largest measuring 19.29Mb. AN2328 with a genomic contribution of 4%, consisted of 41 LCSH events >1Mb. Of these 1-2Mb were recorded for 21/41 (51.2%) and a single event >10Mb, measuring 10.6Mb. Notably in this individual there are 9 events measuring 2-3Mb compared to the cohort average of 2.5 events observed in 18/64 individuals in the cohort (Table 3).

Common LCSH, events that are observed in multiple individuals in the cohort account for some LCSH in these individuals, however private LCSH events predominate. Chromosomes 1, 2, 15 and 16 are represented in LCSH > 10 Mb.

**Table 3.** Comparison of LCSH events in AN2327 and AN2328 with the cohort

| | LCSH Events | 1-2Mb | 2-3Mb | 3-4Mb | 4-5Mb | 5-10Mb | >10Mb |
|---|---|---|---|---|---|---|---|
| **AN2327** | 46 | 20 | 7 | 4 | 4 | 4 | 7 |
| **AN2328** | 41 | 21 | 9 | 4 | 4 | 2 | 1 |
| **Ave. Number Events** | | 10.8 | 2.5 | 1 | 1 | 1 | 0 |
| **# Individuals*** | | 62 | 18 | 2 | 2 | 1 | 0 |

*excluding 2327 and 2328

The allele plots were analysed for evidence of mosaicism and telomeric association, a hallmark of acquired loss of heterozygosity. All LCSH events are interstitial and no evidence of mosaicism detected. This finding infers inheritance of the LCSH events by identity by descent (Table 4). Likewise the possibility of uniparenatal disomy (UPD), where both chromosome homologues are inherited from the same parent, was excluded in these individuals as events > 5 Mb are located on numerous chromosomes. Due to the pattern of LCSH length distribution it is likely that these individuals are offspring of consanguineous relationships.

**Table 4.** Mechanisms of derivation of LCSH is shown

| **Mechanism** | | **Centromere** | **Chromosome arms** |
|---|---|---|---|
| LCSH/Autozygosity | IBD | Homozygous/ Heterozygous | Runs of homozygosity /Heterozygosity |
| Monosomy Rescue | UPD | Homozygous | Homozygous |
| Trisomy Rescue MI non disjunction | UPD | Heterodisomy | Runs of homozygosity /Heterozygosity |
| Trisomy Rescue – MII non disjunction | UPD | Homozygous | Runs of homozygosity /Heterozygosity |
| Segmental UPD somatic | Malignancy | Heterodisomy | Mosaic homozygosity +/- telomeric region |

## 6.5 Discussion

Recent studies have investigated the prevalence of LCSH in the context of population history and disease association using data derived from HapMap (3, 5, 7), volunteer population recruitment (4, 6) and clinical cohorts (2, 7, 13). The prevalence, genomic contribution and properties of LCSH in the outbred general population are not fully characterised for all populations and variation among studies exists. This may be explained by difference in design and investigation purposes of the studies. Of note are the differences in reports of the average number per individual and contribution of common non-random LCSH events. However, studies are in agreement on the pattern and proportion of the length of LCSH events (1, 3-7). LCSH 0.5-1.5Mb is documented in all populations but the prevalence and length of LCSH > 1.5 Mb varies and reflects individual and population variation with respect to parental relatedness and ancestral history (1, 3-7, 9, 10).

It is believed that the prevalence and properties of LCSH in a randomly selected sample of healthy Western European descendent females has not been previously described. Whole genomic DNA is used in the study to eliminate risk of cell line artefact and somatic LCSH events. Furthermore a stringent analytical approach is applied to ensure confidence and exclude putative false positive LCSH calls. The results show that the overall genomic contribution is consistent with previous reports of outbred populations however the average number of LCSH events per individual is lower. Comparison with previous studies shows that population specific LCSH accounts for some but not all of the LCSH in the cohort here.

### 6.5.1 The importance of QC parameters

There are notable differences among previously published studies in the method of analysis of LCSH. This also applies to the QC metrics employed to evaluate the robustness of LCSH calls. Some studies defined a true call by the number of SNP markers in a pre-determined length of LCSH (2, 3, 6, 7). Other studies accept the call confidence provided by proprietary homozygosity detection software (4, 13).

The QC metrics applied to the LCSH events here is stringent but consistent with that detailed by Gibson et al. 2006. The SNP density is calculated and threshold of > 0.2 (1 SNP per 5 kb), is applied to ensure adequate coverage of SNP markers. As a result 564 LCSH events were excluded from further analysis. It is of interest that LCSH events within the pericentromeric region predominated. These calls scored < 0.2 for SNP density, demonstrated a "scattered" effect of SNP markers on visual inspection of plots, and a coinciding "no-call" in the genotyping data. The interpretation of pericentromeric events of LCSH varies amongst studies. These are noted to be excluded from analysis in some studies (1, 3, 7), whilst Curtis et al. 2007 selectively included these LCSH and reported on the high population frequency of these events (2). The results shown here indicate that LCSH located in the pericentromeric region may be due to overcalling by the homozygosity software (11) and reflect low SNP density or integrity of SNP markers (3). Alternative methods are required to confidently call LCSH in regions of low SNP density or poor SNP marker integrity.

Notwithstanding the stringent approach applied here may result in under-estimation of the total number of LCSH events. The calculation of SNP density is an

algorithm based on the number of SNPs per given length. The homozygosity detection algorithm incorporated in CNV Partition (Illumina Inc.) may overestimate the length of an LCSH event by assigning an end "breakpoint" at a distant SNP marker in a region of low SNP density. As a result the algorithm for SNP density will under-estimate the total number of SNP markers for the length of the event and indicate a low SNP density leading to an incorrect assignment of false positive. Visual inspection of plots and evaluation of genotyping data for individual SNP markers within the region can assist in the determination of end points and call confidence.

## 6.5.2 Incidence of LCSH is consistent with an outbred population

The contribution of LCSH to the autosomal genome is calculated for each individual and an average cumulative length of 14.7 Mb representing 0.7% of the genome is presented. Review of the literature revealed variation in the properties of LCSH. McQuillan et al. 2008 illustrated the difference in contribution to the genome due to the technical thresholds applied. In that study the authors compared the cumulative length calculated where the lower detection limit is set at 0.5 Mb, 1.5 Mb and 5.0 Mb (7). In the current study the lower limit is defined as 1Mb consistent with that reported by Gibson et al. 2006 (3). Taking this into account the estimations of contribution to the genome in European and Scottish populations is 0.2% measuring 6Mb (> 1.5 Mb) and 3% measuring 84 Mb (> 0.5 Mb) (7). In a

similar study comparing inbreeding coefficients across populations, Kirin et al. 2010 reported between 0.21% (> 5 Mb) and 4.6% (> 0.5 Mb) for a European population (1). Given that 570/765 (74.5%) of events in the current study are between 1 Mb and 1.5 Mb it is determined that the value of 0.7% measuring 14.7 Mb is comparable to previous reports of LCSH in European populations.

The number of LCSH events per person was observed to vary among populations. Gibson et al. 2006 in a study of HapMap individuals reported levels of 4.4 (YRI), 5.5 (CHB), 8.3 (CEU) and 8.4 (JPT) LCSH events >1Mb per person. The CEU population represents Northern and Western European ancestry (20) and is comparable to the 11.8 events per person in the Western European descendent population. In contrast is the report of 35.9 events per person by Curtis et al. 2008. This study however includes LCSH events in centromeric regions and may reflect an over-estimation. A similar estimated mean of 38.74 events per individual with a median length of 1.3Mb is presented by Nothnagel et al. 2010 for 23 European subpopulations (6).

## 6.5.3 Common LCSH events reflect ancestral history

LCSH has been recently reported as indicative of ancestral population haplotypes (4, 5, 11) and reflect population migration patterns (5-7, 10). The number and length of LCSH events increases with migration patterns "out of Africa" consistent with a reduction in haplotype diversity (5). In addition, common co-localised events are suggestive of a level of specificity for the population (4-7). These non-

random LCSH events may be indicative of regions of the genome where there is reduced recombination or selection pressures allowing long segments of homozygosity to remain in the general population through generations (3-5).

To evaluate this hypothesis, the LCSH events detected in the cohort here were cross-referenced with previous reports. It was ascertained that population specific LCSH may explain up to 26% of all LCSH in the cohort. However this comparative analysis is not exhaustive and is hindered in part by the absence of a database similar to the Database of Genomic Variants (http://dgv.tcag.ca/dgv). Such a database would assist in the registration of population based LCSH and enable refinement of homozygosity mapping for recessive disease association.

## 6.5.4 Incidental finding of consanguinity in the cohort

Endogamous populations representing approximately 10% of the human population practise consanguinity for cultural or religious reasons (11) and have a higher prevalence of autozygosity and larger segments than outbred populations (3, 4, 6, 11). The pattern of distribution of LCSH number and length is consistent among studies and correlates with the degree of relatedness (1, 13, 19). Consanguinity is practised in middle Eastern, African and Indian populations (11) and demonstrated in south and Western Asians (1) with some populations recording levels of up to 60% of relationships (13). However, European populations have among the lowest levels of consanguinity and LCSH levels are consistent with shared ancestry (1).

Based on previous studies, the finding of 3% (2/64) of this cohort of randomly selected females with levels of homozygosity indicative of parental relatedness is not unexpected (1, 11, 19) and may reflect the diversity of cultural backgrounds of the cohort. Considering the percentage of homozygosity it is proposed that these individuals are the offspring of fourth (4%; AN2328) and third degree (7%; AN2327) relatives (13). Analysis of LCSH events revealed an increased number and length of LCSH events in these individuals when compared to other individuals in the cohort and notably the only representation of LCSH >10 Mb. This trend pertains for all length categories and the largest single event is 19.2 Mb (AN2327). The cumulative LCSH length is elevated, 199 Mb (AN2327) and 110 Mb (AN2328) compared to the cohort mean of 14.7 Mb.

It is of interest that some LCSH events in these individuals overlap common regions observed in the cohort and those reported in the European population (6) but are considerably larger. An example of this for AN2327 is an LSCH on chromosome 1 position 72.5-85.7Mb measuring 13.17 Mb compared to the cohort average 1.2 Mb and 1.43 Mb reported by Nothnagel et al. 2010. These long regions of homozygosity may represent regions that undergo recombination in successive generations resulting in shorter segments (11).

### 6.5.5 Recommendations for SNP microarray in diagnostic laboratories

SNP microarray technology, recently implemented to diagnostic laboratories, provides measurement of DNA copy number and genotype information. Variation of the proportion of lengths of LCSH is apparent and is indicative of an individual's ancestral history and closeness of parental relationships (5-7, 11). Although LCSH is not clinically significant, its presence may unmask autosomal recessive gene mutations or imprinted genes (11, 19, 21). The challenge for diagnostic laboratories is the distinction between background levels of LCSH seen in outbred populations and potentially clinically significant events such as a large number of rare and uncommon LCSH, as observed in consanguinity, that may harbour autosomal recessive disease genes or where LCSH on a single chromosome is suggestive of segmental or whole chromosome uniparental disomy (UPD) (4, 19).

To determine thresholds for LCSH analysis, the number of events, the number of individuals with these events and LCSH length was correlated. The results show a significant difference of the number of individuals with events < 3 Mb compared to events > 3Mb. Furthermore LCSH events > 3 Mb is observed as single events in only 4/64 (6%) individuals in this cohort. It is proposed here that 3 Mb is a lower detection limit threshold that can be applied when filtering data for molecular karyotype analysis.

The lower reporting level of the contribution of LCSH to the genome was considered. In this cohort of healthy Western European descendent females the

average autosomal contribution is 0.7%, 11.8 events per person and consisted predominantly of LCSH < 3 Mb. With this in mind the properties of LCSH in this cohort can be differentiated from that of consanguineous relationships of fourth degree relatives at 3%. However the level of reporting of LCSH in clinical referrals must be determined by the laboratory and consideration given to the background levels of the population.

## 6.6 Conclusion

The objective of this study is to investigate and document the properties of LCSH in a small cohort of randomly selected females representing a Western European descendent population. The study identified a background level of LCSH in all individuals and shared events providing further evidence for ancestral haplotypes. Whilst the average contribution to the genome is consistent with an outbred population two individuals are reported with autozygosity suggestive of consanguineous parental relationships.

The appreciation of properties of LCSH in the outbred population is necessary particularly with the recent implementation of molecular karyotyping by SNP microarray to routine diagnostic laboratories. In this setting the knowledge of the prevalence of LCSH, length and genomic location of common LCSH events will assist in differentiating ancestral autozygosity from close parental relatedness and UPD. Utilization of this information enables targeted analysis for candidate genes by homozgosity mapping. In addition the study here has ascertained thresholds for

diagnostic application and demonstrated the need for QC metrics and careful assessment of genotype data.

## 6.7 Reference

1. Kirin M, McQuillan R, Franklin CS, Campbell H, McKeigue PM, Wilson JF. Genomic runs of homozygosity record population history and consanguinity. PLoS One. 2010 5(11):e13996.
2. Curtis D, Vine AE, Knight J. Study of regions of extended homozygosity provides a powerful method to explore haplotype structure of human populations. Ann Hum Genet. 2008;72(Pt 2):261-78.
3. Gibson J, Morton NE, Collins A. Extended tracts of homozygosity in outbred human populations. Hum Mol Genet. 2006;15(5):789-95.
4. Li LH, Ho SF, Chen CH, Wei CY, Wong WC, Li LY, et al. Long contiguous stretches of homozygosity in the human genome. Hum Mutat. 2006;27(11):1115-21.
5. Pemberton TJ, Absher D, Feldman MW, Myers RM, Rosenberg NA, Li JZ. Genomic patterns of homozygosity in worldwide human populations. Am J Hum Genet. 2012;91(2):275-92.
6. Nothnagel M, Lu TT, Kayser M, Krawczak M. Genomic and geographic distribution of SNP-defined runs of homozygosity in Europeans. Hum Mol Genet. 2010;19(15):2927-35.
7. McQuillan R, Leutenegger AL, Abdel-Rahman R, Franklin CS, Pericic M, Barac-Lauc L, et al. Runs of homozygosity in European populations. Am J Hum Genet. 2008;83(3):359-72.
8. Khrunin AV, Khokhrin DV, Filippova IN, Esko T, Nelis M, Bebyakova NA, et al. A genome-wide analysis of populations from European Russia reveals a new pole of genetic diversity in northern Europe. PLoS One. 2013; 8(3):e58552.
9. Chong JX, Ouwenga R, Anderson RL, Waggoner DJ, Ober C. A population-based study of autosomal-recessive disease-causing mutations in a founder population. Am J Hum Genet. 2010;91(4):608-20.
10. Wang SR, Agarwala V, Flannick J, Chiang CW, Altshuler D, Hirschhorn JN. Simulation of Finnish population history, guided by empirical genetic data, to assess power of rare-variant tests in Finland. Am J Hum Genet. 2014;94(5):710-20.
11. Alkuraya FS. Autozygome decoded. Genet Med. 2010;12(12):765-71.
12. Rudan I, Smolej-Narancic N, Campbell H, Carothers A, Wright A, Janicijevic B, et al. Inbreeding and the genetic complexity of human hypertension. Genetics. 2003;163(3):1011-21.
13. Sund KL, Zimmerman SL, Thomas C, Mitchell AL, Prada CE, Grote L, et al. Regions of homozygosity identified by SNP microarray analysis aid in the diagnosis of autosomal recessive disease and incidentally detect parental blood relationships. Genet Med 2013;15(1):70-8.
14. Henrichsen CN, Chaignat E, Reymond A. Copy number variants, diseases and gene expression. Hum Mol Genet. 2009;18(R1):R1-8.
15. Lencz T, Lambert C, DeRosse P, Burdick KE, Morgan TV, Kane JM, et al. Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. Proc Natl Acad Sci U S A. 2007;104(50):19942-7.
16. Jaber L, Shohat T, Rotter JI, Shohat M. Consanguinity and common adult diseases in Israeli Arab communities. Am J Med Genet. 1997;70(4):346-8.

17. Campbell H, Carothers AD, Rudan I, Hayward C, Biloglav Z, Barac L, et al. Effects of genome-wide heterozygosity on a range of biomedically relevant human quantitative traits. Hum Mol Genet. 2007;16(2):233-41.
18. Wang S, Haynes C, Barany F, Ott J. Genome-wide autozygosity mapping in human populations. Genet Epidemiol. 2009;33(2):172-80.
19. Kearney HM, Kearney JB, Conlin LK. Diagnostic implications of excessive homozygosity detected by SNP-based microarrays: consanguinity, uniparental disomy, and recessive single-gene mutations. Clin Lab Med. 2011;31(4):595-613, ix.
20. Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467(7311):52-8.
21. Hildebrandt F, Heeringa SF, Ruschendorf F, Attanasio M, Nurnberg G, Becker C, et al. A systematic approach to mapping recessive disease genes in individuals from outbred populations. PLoS Genet. 2009;5(1):e1000353.

**Technical Note: Illumina Systems and Software**

DNA copy number and loss of heterozygosity analysis algorithms. Illumina, Inc. 2010

# CHAPTER 7

# Clinical Application of the “Aussie Normals” Cohort

## 7.1 Abstract

Common complex disease traits have a major demand on global public health resources. Through extensive investigation by genome-wide association studies (GWAS) and linkage studies, genetic pathways and monogenic causes have been identified for some common and late on-set diseases while the genetic causes of others such as hypertension remains elusive. The investigation of genetic causes of common disease traits is challenging due to phenotypic heterogeneity and poorly defined study cohorts.

The "Aussie Normals" is an *a priori* community based study of volunteers who represent the Australian population. This denotes a well defined phenotypic cohort where a comprehensive medical history was recorded and a large range of clinical chemistry analytes tested for each individual. The "Aussie Normals" collection provides a rare opportunity to elucidate predisposition to common disease traits in the Australian population.

In the study reported here, a sample cohort of Australian women respondents to the "Aussie Normals" Collection are investigated for genetic variants detected by high resolution SNP microarray. To demonstrate a clinical application of the "Aussie Normals" collection, individuals with hypertensive disease trait were selected for investigation. Review of the results of nearly 100 clinical chemistry analytes and medical history identified a subset of individuals with indicative markers. The molecular karyotypes were investigated firstly to establish correlation of copy number change or homozygosity with previously characterised quantitative trait loci (QTL) and secondly to identify genetic variants common to the subset of individuals. This process illustrates the value of general population

cohorts that have well defined phenotype to the search of the genetic basis of common complex and late onset disease traits.

## 7.2 Introduction

Copy number variation and long contiguous stretches of homozygosity (LCSH) are genetic variants now well described in clinical cohorts and since 2004 as genetic variation in normal populations (1-7). The development of high throughput technologies such as microarray and next generation sequencing has provided vast amounts of data that are continually mined for disease associations. This revolution is facilitated by expansive studies such as the International HapMap Consortium Phase 1, 2 and 3, Wellcome Trust Case Control Consortium and 1000 Genomes Project (8-11).

The focus of the International SNP Map Working Party and the International HapMap Project (8, 11) was to identify and validate genetic variants such as single nucleotide polymorphism (SNP) (11, 12). This is a genome-wide form of genetic variation that is estimated to occur once every 1000 base pairs (bp) (11). A single base pair variation in a DNA sequence generates two alleles. This is differentiated from a point mutation where the frequency in the population exceeds 1% and a common variant is noted where the minor allele frequency is >5% (11). The finding that SNPs were not independent and existed in linkage disequilibrium enabled more efficient use of this form of genetic variation in Genome-Wide Association Studies (GWAS) (8, 9, 11, 13).

The efficiency of GWAS in detecting risk alleles is limited by the investigation of common alleles (allele frequency >5%) that have small effect and do not

encompass the full extent of allelic frequencies anticipated in common complex disease (11). Despite this limitation SNP data has been mined in hundreds of genome-wide studies for association with complex disease and common disease trait and risk alleles in numerous conditions such as Crohn's disease, type 1 and type 2 diabetes, asthma, obesity and Coeliac disease (9, 11, 13). Furthermore multiple rare and common variants including both SNP and other forms of genetic variation such as copy number variation (CNV) have been reported in disease such as type 1 diabetes and Crohn's disease (13).

Despite the rapid progression of discovery of genetic variants and association with disease, many risk loci and genetic pathways are yet to be revealed. This is particularly noted in conditions where there is phenotypic heterogeneity and complex interactions of both environmental and genetic factors (14-16). Two models of inheritance were proposed to explain the search for candidate loci. The common variant/ common disease hypothesis infers that susceptibility alleles are common in the population (frequency > 1%). This model formed the basis of many genome-wide association studies (11, 14, 32). Alternatively the rare variant hypothesis implies the presence of rare variants abundantly in the genome (11, 14). The contribution of rare variants is yet to be characterised and will be the subject of future studies (11). Preliminary evidence suggests that this model may account for some complex diseases (16) and interaction of both common and rare variants may contribute to disease pathogenesis (11, 16).

Hypertension is an example of a common complex disease trait. This condition is a risk factor for cardiovascular disease at a prevalence that places significant burden on global public health resources (17-19). The causes of hypertension are believed

to be multi-factorial and the phenotype influenced by the interaction of environmental and genetic factors (30). Several hundreds of genes in numerous biological pathways are believed to be involved (14-17, 19). Candidate loci have been identified in association with familial hypertension (16, 20, 21) and conditions that lead to progression to hypertension (16, 20). However the aetiology of essential hypertension is relatively unknown. The initial GWAS studies failed to identify quantitative trait loci for hypertension which may be due to the study design and the biological complexity of the condition (9, 11, 22).

Several studies have demonstrated an association of LCSH load with a proportional increase in blood pressure as evidenced by a higher rate of hypertension in populations who practice endogamy and in the offspring of consanguineous couples (11, 15, 19). It is noted that few studies have investigated CNV and LCSH in general populations for genetic variants associated with hypertension (17, 22). Furthermore the genetic pathways for essential hypertension are yet to be defined (15, 16, 19, 32). This is due in part to the paucity of normal cohorts with well described clinical history and pathological evidence of “health” as well as the clinical heterogeneity of the disease resulting in challenges in defining a representative clinical cohort (15, 16).

In the study here CNV and LCSH events are reviewed to determine correlation with previously characterised genetic markers for essential hypertension and monogenic causes of familial hypertension. The identification of putative candidate genes is beyond the scope of this study however, the data is reviewed to identify potential candidate regions common to the clinical subset that may be considered for further investigation.

## 7.3 Materials and Methods

### 7.3.1 The “Aussie Normals” Cohort

The cohort represents 64 females between the ages of 23 and 84 years residing in Canberra, Australian Capital Territory, Australia. Volunteers for the community based study, the “Aussie Normals”, were recruited from the local electoral roll. Respondents to a letter of invitation undertook a telephone interview as an initial screening process. To ensure a robust cohort, individuals that provided the medical history of diabetes, malignancy, treatment dependent asthma and current pregnancy were not accepted to the study cohort. In the second phase of screening, phenotypic data including age, height, blood pressure, body mass index (BMI) and medications were recorded (23). Peripheral blood was collected and tested for 91 common clinical chemistry analytes (23), conventional cytogenetic analysis and DNA extraction. All test results were reviewed by medical professionals and specialist pathologists. Written consent was provided by the participants and ethics approval was initiated through the Australian National University Human Research Ethics committee.

### 7.3.2 Data analysis

Lymphocyte cultures were established for cytogenetic analysis and the metaphase spreads analysed for structural rearrangements. DNA was manually extracted by organic separation and ethanol precipitation. High resolution SNP microarray investigation was performed using Illumina Omni1-Quad for the detection of copy

number change and long stretches of homozygosity. Each of these methods is described in detail in chapters 3, 4, 5 and 6.

Hypertension was selected as a feasibility study for the investigation of common disease predisposition in the "Aussie Normals" collection. Hypertension is defined as blood pressure values of >140MmHg (systolic) or >90MmHg (diastolic) (16, 23) or individuals who reported a medical history of hypertension and are taking anti-hypertensive medications. This formed a clinical subset for the investigation of genetic variants and comparison with those without evidence of hypertension.

Copy number variants (CNV) and long contiguous stretches of homozygosity (LCSH) were extracted for the individuals in the clinical subset and demonstrates the clinical application of this general population cohort. Analysis was performed in two phases. Firstly by identifying CNV and LCSH events that overlapped previously characterised genetic variants that have a known association with essential hypertension (unknown aetiology), familial hypertension and cardio-vascular disease. Secondly CNV and LCSH were screened for events that occurred at a higher frequency in the clinical cohort than individuals without evidence of hypertension.

## 7.4 Results

### 7.4.1 Hypertension in the "Aussie Normals"

The medical history, blood pressure, body mass index (BMI) and age of 64 females in the study were reviewed. There are 18/64 individuals with diastolic blood

pressure > 140MmHg or systolic blood pressure >90MmHg or are noted to be on anti-hypertensive medication. These women are aged between 53 to 83 years (median 79.5) compared to 23- 83 years in the entire cohort (median 65.06). The body mass index (BMI) ranged from 20.4 to 42.04 and 2/18 individuals recorded an elevated BMI >35. The median BMI of the clinical subset is 26.02 compared to 24.03 in the entire cohort. There are 46/64 individuals with normal blood pressure values.

## 7.4.2 Review of previously characterised risk variants

The CNV and LCSH load for each individual was reviewed to evaluate the hypothesis that increased CNV and LCSH load correlate with elevated blood pressure values or associated clinical trait (15, 19, 24, 25) (Table 1). The average number of CNV >1kb per person was elevated but did not reach significance. The clinical subset recorded an average of 173 per person compared to 166 in the entire cohort. The number of LCSH events > 1Mb per person is consistent for both cohorts, reporting 11 events per individual. Similarly, the contribution of homozygosity to the individual genome accounting for 0.56% in the clinical subset compared to 0.54% in the entire cohort.

**Table 1.** CNV and LCSH properties for the individuals with hypertension

| Sample id | Age | CNV > 1kb | > 100kb | LSCH | % Genome |
|---|---|---|---|---|---|
| **2242** | 79 | 183 | 8 | 11 | 0.51 |
| **2243** | 82 | 193 | 8 | 17 | 0.88 |
| **2310** | 71 | 182 | 12 | 12 | 0.62 |
| **2313** | 83 | 216 | 6 | 13 | 0.63 |
| **2317** | 76 | 165 | 6 | 9 | 0.41 |
| **2320** | 53 | 151 | 4 | 12 | 0.6 |
| **2324** | 81 | 145 | 5 | 11 | 0.48 |
| **2326** | 79 | 154 | 5 | 8 | 0.37 |
| **2331** | 66 | 157 | 9 | 9 | 0.47 |
| **2337** | 80 | 155 | 5 | 6 | 0.25 |
| **2339** | 67 | 183 | 7 | 13 | 0.65 |
| **2346** | 78 | 180 | 10 | 13 | 0.65 |
| **2347** | 77 | 165 | 4 | 13 | 0.62 |
| **2350** | 68 | 199 | 7 | 16 | 0.75 |
| **2352** | 80 | 170 | 6 | 12 | 0.55 |
| **2357** | 81 | 158 | 6 | 9 | 0.4 |
| **2378** | 79 | 197 | 10 | 14 | 0.76 |
| **Ave** | | **173.7** | | **11.64** | **0.56** |

The molecular karyotypes were reviewed for evidence of CNV or LCSH that correlated with previously characterised risk alleles (Table 2). No evidence of CNV was detected at or near the 29 risk alleles reported by The International Consortium for Blood Pressure genome-wide association study (16) for the entire cohort. Events of LCSH correlated with the SNP alleles in 5/18 individuals involving 4/29 loci that harbour the *SLC39A8, FGF5, SH2B3* and *CYP1A1-ULK3* genes. It is noted that a higher incidence of these LCSH events occurred in individuals with blood pressure in the healthy range. Homozygosity for the region harbouring the *SH2B3* gene was detected in 8/46 females without evidence of hypertension. Additional loci were also detected in the healthy cohort involving

5/46 females across 4/29 loci and included *C10orf107, CYP17A1, ATP2B1* and *GOSR2* (Table 2).

Left ventricular hypertrophy is indicative of cardiac damage subsequent to chronic hypertension. BoonPeng et al. 2013 reviewed the contribution of CNV on genes that are reported to be involved in the molecular and biochemical pathways that lead to progression of left ventricular hypertrophy (LVH)(17). The authors identified candidate genes involved in hypertensive LVH pathogenesis, some of which are reported to overlap CNV registered in the Database of Genomic Variants (DGV; http://dgv.tcag.ca) (26). To test this hypothesis CNV and LCSH regions in the "Aussie Normals" cohort were analysed for autosomal regions that overlapped the candidate genes (Table 3). The results illustrate that there are no CNV in the clinical subset and one CNV that harbours the *CAMK2B* gene was detected in 3/46 individuals who do not present with hypertension. Similarly one LCSH event involving the *NFATC3* gene was detected in 2/18 individuals, but this is also observed in the healthy population at a frequency of 5/46.

**Table 2.** SNP alleles associated with variation in blood pressure

| Gene | Identifier SNP | Chr | Position NCBI36/Hg18 | CNV n=18 | CNV n=46 | LCSH n=18 | LCSH n=46 |
|---|---|---|---|---|---|---|---|
| **Ehret et al. 2012** | | | | | | | |
| *MOV10* | rs2932538 | 1 | 113,018,066 | 0 | 0 | 0 | 0 |
| *MTHFR-NPPB* | rs17367504 | 1 | 11,785,365 | 0 | 0 | 0 | 0 |
| *SLC4A7* | rs13082711 | 3 | 27,512,913 | 0 | 0 | 0 | 0 |
| *MECOM* | rs419076 | 3 | 170,583,580 | 0 | 0 | 0 | 0 |
| *ULK4* | rs3774372 | 3 | 41,852,418 | 0 | 0 | 0 | 0 |
| *SLC39A8* | rs13107325 | 4 | 103,407,732 | 0 | 0 | 1 | 1 |
| *GUCY1A3-GUCY1B3* | rs13139571 | 4 | 156,864,963 | 0 | 0 | 0 | 0 |
| *FGF5* | rs1458038 | 4 | 81,383,747 | 0 | 0 | 1 | 2 |
| *NPR3-C5orf23* | rs1173771 | 5 | 32,850,785 | 0 | 0 | 0 | 0 |
| *EBF1* | rs11953630 | 5 | 157,777,980 | 0 | 0 | 0 | 0 |
| *HFE* | rs1799945 | 6 | 26,199,158 | 0 | 0 | 0 | 0 |
| *BAT2-BAT5 (HLA)* | rs805303 | 6 | 31,724,345 | 0 | 0 | 0 | 0 |
| *CACNB2(5')* | rs4373814 | 10 | 18,459,978 | 0 | 0 | 0 | 0 |
| *PLCE1* | rs932764 | 10 | 95,885,930 | 0 | 0 | 0 | 0 |
| *CACNB2(3')* | rs1813353 | 10 | 18,747,454 | 0 | 0 | 0 | 0 |
| *C10orf107* | rs4590817 | 10 | 63,137,559 | 0 | 0 | 0 | 1 |
| *CYP17A1-NT5C2* | rs11191548 | 10 | 104,836,168 | 0 | 0 | 0 | 2 |
| *ADM* | rs7129220 | 11 | 10,307,114 | 0 | 0 | 0 | 0 |
| *FLJ32810-TMEM133* | rs633185 | 11 | 100,098,748 | 0 | 0 | 0 | 0 |
| *PLEKHA7* | rs381815 | 11 | 16,858,844 | 0 | 0 | 0 | 0 |
| *ATP2B1* | rs17249754 | 12 | 88,584,717 | 0 | 0 | 0 | 1 |
| *SH2B3* | rs3184504 | 12 | 110,368,991 | 0 | 0 | 2 | 8 |
| *TBX5-TBX3* | rs10850411 | 12 | 113,872,179 | 0 | 0 | 0 | 0 |
| *FURIN-FES* | rs2521501 | 15 | 89,238,392 | 0 | 0 | 0 | 0 |
| *CYP1A1-ULK3* | rs1378942 | 15 | 72,864,420 | 0 | 0 | 1 | 2 |
| *GOSR2* | rs17608766 | 17 | 42,368,270 | 0 | 0 | 0 | 1 |
| *ZNF652* | rs12940887 | 17 | 44,757,806 | 0 | 0 | 0 | 0 |
| *JAG1* | rs1327235 | 20 | 10,917,030 | 0 | 0 | 0 | 0 |
| *GNAS-EDN3* | rs6015450 | 20 | 57,184,512 | 0 | 0 | 0 | 0 |

**Table 3.** Candidate genes with roles in left ventricular hypertrophy and hypertension

| Gene | Chr | Position NCBI36/Hg18 | CNV n=18 | CNV n=46 | LCSH n=18 | LCSH n=46 |
|---|---|---|---|---|---|---|
| **Boon Peng et al. 2013** | | | | | | |
| *SLC8A1* | 2p22.1 | 40,192,790-40,510,948 | 0 | 0 | 0 | 1 |
| *SLC9A2* | 2q12.1 | 102,602,580-102,694,241 | 0 | 0 | 0 | 0 |
| *CTNNB1* | 3p22.1 | 41,215,946-41,256,943 | 0 | 0 | 0 | 0 |
| *PPP3CA* | 4q24 | 102,163,610-102,487,651 | 0 | 0 | 0 | 1 |
| *EGF* | 4q25 | 111,053,489-111,153,567 | 0 | 0 | 0 | 0 |
| *CAMK2B* | 7p14.3 | 44,223,274-44,331,755 | 0 | 3 | 0 | 0 |
| *EGFR* | 7p11.2 | 55,054,219-55,242,525 | 0 | 0 | 0 | 0 |
| *GATA4* | 8p23.1 | 11,599,126-11,654,918 | 0 | 0 | 0 | 0 |
| *KCNB2* | 8q13.3 | 73,612,180-74,013,138 | 0 | 0 | 0 | 1 |
| *JAK2* | 9p24.1 | 4,975,245-5,117,995 | 0 | 0 | 0 | 0 |
| *ADAM12* | 10q26.2 | 127,690,944-128,067,117 | 0 | 0 | 0 | 0 |
| *CALM1* | 14q32.1 | 89,933,080-89,944,372 | 0 | 0 | 0 | 0 |
| *MEF2A* | 15q26.3 | 97,956,185-98,071,524 | 0 | 0 | 0 | 0 |
| *MAPK3* | 16p11.2 | 30,032,927-30,042,131 | 0 | 0 | 0 | 0 |
| *NFATC3* | 16q22.1 | 66,676,876-66,818,338 | 0 | 0 | 2 | 5 |
| *ERBB2* | 17q21.1 | 35,097,863-35,138,441 | 0 | 0 | 0 | 0 |
| *ILF3* | 19p13.2 | 10,625,988-10,664,095 | 0 | 0 | 0 | 0 |
| *CALM3* | 19q13.2 | 51,796,352-51,805,879 | 0 | 0 | 0 | 0 |
| *MCIP-1 (RCAN1)* | 21q22.12 | 34,810,654-34,909,252 | 0 | 0 | 0 | 0 |
| *ITGB2* | 21q22.3 | 45,130,296-45,165,393 | 0 | 0 | 0 | 0 |

**Table 4.** Common CNV reported to be associated with hypertension

| CNV | Chr | Position NCBI36/Hg18 | CNV n=18 | CNV n=46 |
|---|---|---|---|---|
| **Margues et al. 2014** | | | | |
| rs2932538/ esv27061 | 1 | 112,494,152-113,047,786 | 0 | 9 |
| esv2757747 | 1 | 112,958,658-116,542,895 | 0 | 0 |
| rs7129220/ nsv483076 | 11 | 10,149,870-10,309,473 | 0 | 0 |
| rs17608766/ dgv976e1 | 17 | 41,439,751-42,632,332 | 2 | 13 |
| rs1327235/ dgv1306e1 | 20 | 10,892,138-11,116,725 | 0 | 0 |

A similar study by Marques et al. 2014 reviewed the molecular karyotypes of a cohort of participants in the Victorian Family Heart Study, Australia, for CNV that correlated with known risk alleles (22). The authors reported an overlap of known risk SNP alleles with common CNV that are registered in the DGV (http://dgv.tcag.ca) (26) on chromosome 1, 11, 17 and 20. Within their clinical cohort they reported a prevalence of a deletion at chromosome 1: 112Mb-113Mb (esv27061; NCBI36/Hg18) and chromosome 1: 112Mb-116Mb (esv2757747; NCBI36/Hg18) encompassing *MOV10* and several microRNA (22). To evaluate the contribution of these CNV to hypertension in the "Aussie Normals" cohort the molecular karyotypes were analysed (Table 4). There was no evidence of esv 27061 or esv 2757747 in the clinical subset, although a deletion was detected in 9/46 individuals in the cohort without hyptertension. A similar observation was reported for chromosome 17 (dgv976e1) where a greater prevalence of the deletion is identified in the individuals without hypertensive trait.

Collectively these findings do not provide support for the association of these common CNV or CNV encompassing the risk alleles with hypertension in the "Aussie Normals." To ensure the robustness of CNV detection in the study here the CNV locations listed in table 4 were analysed for marker distribution. There is greater than 20 SNP and CNV polymorphic markers in all regions indicating adequate coverage for CNV detection.

Next the molecular karyotypes of the clinical subset were reviewed for CNV or LCSH events that encompass genes known to be associated with familial forms of hypertension (Table 5). There was no evidence of CNV or LCSH within these regions, with the exception of 2/46 individuals in the healthy cohort with LCSH

events involving *CYP17A1*. This gene is a risk allele in adrenal hyperplasia and associated hypertension (16).

**Table 5.** Genes involved in familial forms of hypertension

| Gene | CNV | | LCSH | | Disease association | Reference |
|---|---|---|---|---|---|---|
| | n=18 | n=46 | n=18 | n=46 | | |
| *ATP1A1* | 0 | 0 | 0 | 0 | Primary Aldosteronism | Nishimoto et al. 2015 |
| *CACND1* | 0 | 0 | 0 | 0 | Primary Aldosteronism | Nishimoto et al. 2015 |
| *CYP11B1* | 0 | 0 | 0 | 0 | Familial Hyperaldosteronism | Funder et al. 2012 |
| *CYP11B2* | 0 | 0 | 0 | 0 | Familial Hyperaldosteronism | Funder et al. 2012 |
| *CYP17A1* | 0 | 0 | 0 | 2 | Adrenal Hyperplasia | Ehret et al. 2012 |
| *KCNJ5* | 0 | 0 | 0 | 0 | Hyperaldosteronism, Type III | Funder et al. 2012 |

## 7.4.3 Review of candidate regions in the "Aussie Normals"

The molecular karyotypes of the 18 individuals with hypertension were analysed for CNV or LCSH calls that are unique to the clinical subset. The main objective of this analysis is to identify chromosomal regions that may be considered candidates for further investigation. There is no evidence of either CNV or LCSH calls that are present in all 18 individuals. Consideration was then given to the events that were detected in greater than 50% (9/18) of the clinical cohort. This accounted for 60 CNV regions, of which 32/60 (53%) did not harbour genes. It was observed that these variants are not unique to the clinical subset and are detected at high frequency in the entire cohort (Table 6).

**Table 6.** List of CNV observed in more than 50% of the individuals with hypertension

| Chr | Start | End | n=18 | n=46 | Gene |
|---|---|---|---|---|---|
| 1 | 72537704 | 72584280 | 13 | 19 | Nil |
| 1 | 150822709 | 150853563 | 11 | 18 | LCE3C |
| 1 | 157122178 | 157136357 | 11 | 28 | Nil |
| 1 | 167412568 | 167507933 | 9 | 17 | NME7 |
| 1 | 192718169 | 192719701 | 11 | 20 | Nil |
| 2 | 1504536 | 1519601 | 18 | 46 | TPO |
| 2 | 88922426 | 89238301 | 13 | 41 | Nil |
| 2 | 126159751 | 126168232 | 10 | 25 | Nil |
| 2 | 127391209 | 127393549 | 13 | 30 | Nil |
| 2 | 159667833 | 159669697 | 11 | 21 | TANC1 |
| 2 | 176976251 | 176979982 | 11 | 26 | Nil |
| 2 | 194363988 | 194423573 | 18 | 39 | Nil |
| 2 | 241511624 | 241521395 | 18 | 46 | Nil |
| 3 | 32077059 | 32082421 | 14 | 37 | Nil |
| 3 | 100378327 | 100384895 | 11 | 31 | Nil |
| 3 | 163973719 | 164108624 | 9 | 35 | Nil |
| 3 | 164245767 | 164251701 | 12 | 35 | Nil |
| 4 | 9077736 | 9089333 | 14 | 36 | Nil |
| 4 | 9820366 | 9843358 | 10 | 25 | Nil |
| 4 | 10001819 | 10011587 | 10 | 18 | Nil |
| 4 | 146655600 | 146659357 | 10 | 22 | *SMAD1* |
| 4 | 166222921 | 166224107 | 15 | 40 | *TMEM192* |
| 4 | 173663817 | 173668254 | 15 | 33 | *GALNTL6* |
| 4 | 186678926 | 186681104 | 14 | 30 | *PDLIM3* |
| 5 | 57359369 | 57370582 | 11 | 31 | Nil |
| 5 | 90536386 | 90537807 | 13 | 32 | Nil |
| 6 | 32561042 | 32613985 | 18 | 46 | *HLA* |
| 6 | 35734478 | 35737647 | 9 | 12 | *FKBP5* |
| 6 | 54037736 | 54041833 | 18 | 22 | *MLIP* |
| 7 | 125835808 | 125842423 | 11 | 21 | Nil |
| 7 | 147700415 | 147707199 | 14 | 30 | *CNTNAP2* |
| 7 | 154831298 | 154835743 | 12 | 22 | Nil |
| 8 | 583707 | 589201 | 15 | 29 | *ERICH1* |
| 8 | 25117724 | 25126553 | 12 | 30 | *DOCK5* |
| 8 | 32799203 | 32810651 | 18 | 46 | Nil |
| 8 | 39352501 | 39506336 | 14 | 37 | *ADAM3A* |
| 8 | 42307707 | 42312756 | 13 | 33 | *IKBKB* |
| 8 | 47521923 | 47526841 | 12 | 24 | Nil |
| 9 | 17894496 | 17901627 | 12 | 30 | Nil |
| 9 | 22486651 | 22498382 | 11 | 35 | Nil |

| | | | | | |
|---|---|---|---|---|---|
| 9 | 70923423 | 70932920 | 11 | 33 | *TJP2* |
| 9 | 130452370 | 130453706 | 11 | 32 | *WDR34* |
| 10 | 4280129 | 4284071 | 13 | 36 | Nil |
| 10 | 4698627 | 4700298 | 10 | 30 | *LINC00704* |
| 10 | 77916098 | 77930393 | 13 | 29 | *C10orf11* |
| 10 | 108020351 | 108022353 | 12 | 19 | Nil |
| 10 | 122216937 | 122221235 | 11 | 24 | *PAPPDC1A* |
| 11 | 1003987 | 1008456 | 16 | 37 | *MUC6* |
| 11 | 55122337 | 55209410 | 10 | 14 | *OR4C11,OR4P4, OR4S2, OR4C6* |
| 13 | 49967095 | 49972694 | 13 | 32 | *DLEU1* |
| 14 | 36838520 | 36840751 | 18 | 46 | *MIPOL1* |
| 16 | 22953373 | 22956947 | 12 | 26 | Nil |
| 16 | 25246775 | 25258041 | 13 | 44 | Nil |
| 18 | 37118901 | 37127830 | 12 | 27 | Nil |
| 18 | 53090436 | 53099515 | 11 | 24 | Nil |
| 18 | 75411150 | 75413066 | 12 | 23 | Nil |
| 19 | 12555342 | 12558389 | 10 | 26 | *ZNF490* |
| 20 | 1517540 | 1532878 | 12 | 32 | *SIRPB1* |
| 22 | 16438001 | 16439664 | 13 | 39 | *SLC25A18* |
| 22 | 22604775 | 22619747 | 14 | 11 | Nil |

Review of LCSH identified 4 events that are present in greater than 50% of the clinical subset (Table 7). These are represented on 1q (141mb-144Mb and 146Mb-148Mb), 2q21.3(135Mb-136Mb) and 16p11.2 (31.4Mb -34.1Mb). It should be noted that all of these, with the exception of 2q21.3, represent LCSH calls that failed quality control parameters due to the insufficient coverage of SNP markers. The quality control metrics applied in this study are stringent. However for the purpose of this assessment these calls will be considered here. The incidence is elevated in the clinical cohort for chromosomes 1q21.1 (146Mb -148Mb), 2q21.3 and 16p11.2, although the LCSH events are not unique and are present in the entire cohort.

When specific CNV and LCSH calls are considered together, overlapping chromosome regions are observed, altering the frequency in the clinical subset (Table 7). For example a CNV at 1q21.1 (147.1-147.7) in one individual in the clinical subset is within the region of LCSH in 13 other individuals for a total of 14/18. Similarly the LCSH at 16p11.2 encompassing 31.4-34.1Mb was detected in 11 individuals and CNV located at the genomic position 32.04Mb-33.78Mb was detected in 6 individuals accounting for 94% (17/18) of the individuals with hypertension.

**Table 7.** List of LCSH and coinciding CNV in the clinical subset

| Chr | Position NCBI36/Hg18 | | CNV n=18 | CNV n=46 | LCSH n=18 | LCSH n=46 | Genes[#] |
|---|---|---|---|---|---|---|---|
| 1q21.1 | 141477151 | 144431772 | 0 | 0 | 9 | 26 | *NBPF8-NBPF20, HFE2, NOTCH2NL, PDE4DIP, PEX11B* |
| 1q21.1 | 146290400 | 148081703 | 1 | 6 | 13 | 30 | *NBPF25P,FCGR1C, FAM231D* |
| 2q21.3 | 135525827 | 136547005 | 0 | 0 | 10 | 16 | *RAB3GAP1, UBXN4, LCT, MCM6,DARS* |
| 16p11.2 | 31476343 | 34135123* | 6 | 11 | 11 | 36 | *TP53TG3, ENPP7P13* |

#minimal region of overlap, * combined region

## 7.5 Discussion

The investigation of general population cohorts provides a resource that can be investigated to determine benign variants and background levels of genetic variation, “what is normal”, and can be utilised as a control cohort in comparative studies for the investigation of pathogenicity. The “Aussie Normals” collection is an example of a general population cohort. The volunteers residing in Canberra, ACT recruited for the “Aussie Normals” collection according to the Australian Bureau of Statistics are representative of the Australian population with respect to age, gender and ethnicity (23) and indicative of a Western European descendent population. The cohort is well defined with respect to medical history and current pathology status (23). General population cohorts such as this provide an invaluable resource for numerous applications in medical research.

In the study presented here the “Aussie Normals” was investigated for CNV and LCSH events in a small cohort of females with clinical indicators of hypertension to exemplify an application of a clinically well defined general population cohort. The prevalence of high blood pressure in the general population in 2005 is 22% and expected to increase to 30% by 2025 (18). It therefore represents a clinical variation within a normal population. The study findings do not demonstrate a correlation of CNV load, common CNV or LCSH involving known risk alleles with hypertension. However this feasibility study investigated a small female cohort. The “Aussie Normals” has recruited 1856 volunteers (893M and 963F) (23) and expansion of this study is achievable. Unlike other common disease traits or monogenic Mendelian disease there is considerable phenotypic heterogeneity

which makes investigation challenging with respect to defining the clinical and non-clinical control cohort (32).

### 7.5.1 The role of control cohorts

Few published studies have well defined clinical cohorts supported by medical data or conversely a well defined control cohort to provide comparison. This is exemplified in the study by Marques et al. 2014 where they identified a prevalence of two common CNV (esv2757747 and esv27061) in a clinical cohort of Australians and proposed an association of these CNV in hypertensive disease. The authors did not validate their findings against a control cohort indicating that the genes encompassed by the two CNV were previously identified as risk alleles in essential hypertension by SNP and linkage disequilibrium studies (16, 22). In addition the authors were unable to corroborate their own findings when compared to a broader clinical cohort and postulated that the different prevalence was due to population specific variation. It is of interest that the finding of these common CNV is not replicated in the clinical subset reported here in the “Aussie Normals”. There was no evidence of esv2757747 or esv27061 in the clinical subset and in fact there was no evidence of esv2757747 in the entire cohort. In contrast to the findings of Marques et al. 2014 a higher prevalence of esv27061 was detected in individuals that did not present with biological indicators of hypertension. This exemplifies the need and potential application of clinically well defined control cohorts when investigating disease associations of genetic variants such as CNV.

## 7.5.2 The role of CNV in disease pathogenesis

Given the recent knowledge of the contribution of CNV and LCSH to human genetic variation the investigation for common and rare CNV and LCSH for common complex disease association is warranted. Some monogenic causes for familial hypertension and hyperaldesteronism (a cause of hypertension) have been described (16, 21, 27). However the genetic pathways for essential hypertension are less understood. The seminal genome-wide association study by the Wellcome Trust Case Control Consortium failed to identify risk alleles for blood pressure variation or hypertension (9). This study investigated common SNP loci that are present in the population at a frequency >5% missing any variants with potential disease association at population frequencies below this level or rare variants (9, 11, 22). The genome-wide study by the International Consortium for Blood Pressure reported on a total of 29 risk variants across 28 loci from a collaborative study involving 200,000 individuals of European descent (16).

Although it is recognised that the causes of hypertension are multi-factorial, polygenic and potentially involve numerous genetic mechanisms, few studies have investigated the role of CNV in disease pathogenesis (17, 22). There are a number of pathways that CNV may contribute to predisposition to disease (Figure 1). A deletion CNV may unmask a risk allele enabling expression of an autosomal recessive gene or disrupt long range gene regulation mechanisms (25, 28). Alternatively phenotypic variability with the deletion of contiguous genes is apparent in which the larger deletion may culminate in a more severe phenotype or result in a syndrome with phenotypic diversity. Although the effect of imprinting on genes is well described, there is the potential of yet undiscovered

imprinted genes of small effect that may contribute to common disease association. Finally the accumulative effect of CNV or other genetic variants in a two hit model is a way in which CNV may contribute to disease predisposition (25). These models have been proposed to explain the phenotypic heterogeneity and variable expressivity that is observed in microdeletion and microduplication of complex disorders such as autism and developmental delay (25). These models could equally well be applied to common complex diseases such as hypertension and cardiovascular disease predisposition.

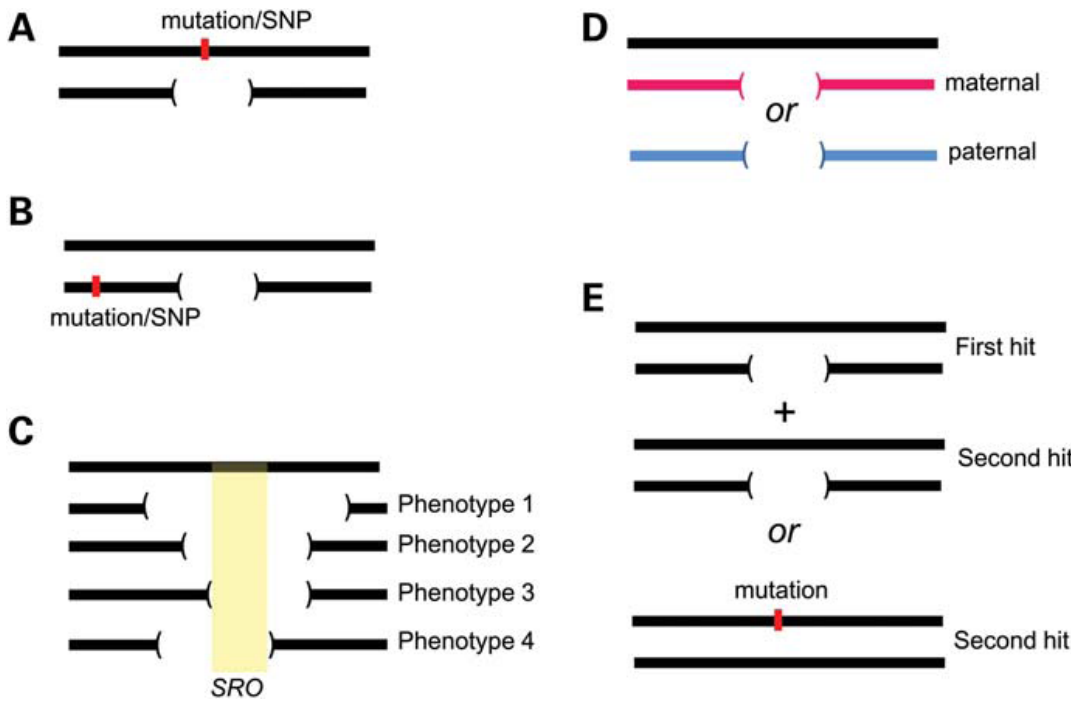



**Figure 1**. Models for the potential role of CNV deletion in disease pathogenesis and causes of phenotypic variability are shown; Unmasking a autosomal recessive allele (a), long range disruption of a gene (b), contiguous gene deletion effects (c), gene imprinting (d) and CNV load or compound effect (e). Figure from Girirajan, S. and Eichler, E. 2010 (25).

### 7.5.3 The role of long contiguous stretches of homozygosity

Several studies have demonstrated an association of LCSH load with an increase in blood pressure as evidenced by a higher rate of hypertension in populations who practise endogamy and in the offspring of consanguineous couples (15, 19). The overall contribution of LCSH to an individual's genome increases proportional to the degree of parental relationship. The number and length of tracts of shared haplotypes also increases compared to the outbred population (15, 19), and as such encompasses more genes. This is associated with an increased risk for the inheritance of autosomal recessive genes (29). This evidence was initially thought to support the common variant common disease model, it could also be hypothesised that the rare variant model applies (14, 32). The presence of homozygosity in an individual genome will increase the likelihood of disruption of genetic pathways due to the chance involvement of a larger number of homozygous risk alleles (15, 19, 29, 30) or alternatively represent expression of a large number of risk alleles with small clinical effect. As reported in this cohort the LCSH events observed in the presence of parental relatedness include events that are both common in the general population and rare. The two individuals in the current study with the incidental finding of homozygosity at 7% and 4% of their genomes, both had normal blood pressure 128/78 and 111/71 respectively. These individuals are 65 and 33 years of age and have normal BMI (25 and 24). This finding, although with significant sample size limitations, suggests that it is not overall load that contributes to disease development but rather the involvement of specific genes and allelic variants. Distinction of LCSH regions in the outbred

population with that for offspring of consanguineous couples has not been previously investigated for hypertensive disease. This comparison, together with that of in-bred individuals who do not have hypertensive indicators, may yield candidate regions for further investigation. A large collaborative study isolating shared regions of interest and comparison with medically characterised individuals with normal blood pressure will assist in the discovery of rare variants or identify the cumulative effect of large numbers of variants with small effect.

A point of interest identified in the study here is that genetic variants should not be considered in isolation. By considering specific CNV and LCSH events together the prevalence of genetic variation in the candidate region increased. This may be applicable when considering the disease association of rare or common risk SNP alleles and gene mutations.

## 7.5.4 Other genetic variants

It is likely that SNP allelic variants do not account for all of genetic causes of common complex disease and a combination of genetic variants, common or rare, will lead to disease predisposition (11). Multiple rare variants have been identified in type 1 diabetes, schizophrenia and elevated high density lipoprotein cholesterol (11, 15). Many of the CNV detected in this cohort, that is shared by the individuals with clinical indications of hypertension do not harbour RefSeq genes. The possibility of small non-coding microRNA (miRNA) and long non-coding RNA (LncRNA), some of which are yet to be discovered, must also be considered. Preliminary studies have shown expression of non-coding RNA in tissues (22,32).

The putative role of LncRNAs in epigenetic regulation and miRNA in inactivation of mRNA impacting protein synthesis has been reported making these potential candidates for common complex disease pathogenesis (22,32) .

The impact of CNV gene regulation mechanisms must also be considered. Disruption of transcription factors by deletion or duplication of DNA sequences either involving or near enhancers and promoters located downstream of candidate genes has been shown to alter gene expression (25, 28, 31). Although the CNV and LCSH events reported here did not overlap previously characterised risk variants a more indepth analysis of the genomic architecture surrounding the shared variants to identify potential gene regulation mechanisms is required. This study does not exclude the possible role of the known risk variants in disease pathogenesis in these individuals but the findings do not support a direct role of CNV and LCSH in the altered expression of these alleles.

## 7.6 Conclusion

Presented here is an example of the application of a clinically well defined general population. The result of this size limited study does not support common CNV as a risk factor for hypertension in this cohort. However the role of CNV in alternative mechanisms such as rare variants, the cumulative effect of numerous low incidence alleles or the role of long range gene regulation in CNV where no RefSeq genes were observed cannot be excluded. The possibility of non-coding RNA in regions poorly covered by SNP microarray markers, but where there is a high prevalence of CNV and LCSH detection, must also be considered. This study is

indicative of the challenges of detecting genetic risk factors for common disease trait. The 1000 Genome Project with a combination of SNP, copy number and exome sequencing will provide additional targets for disease correlations. This is a population based project using sequencing technology to identify a range of genetic variations and their patterns of linkage disequilibrium. However as demonstrated here genetic variants should not be considered in isolation, as this may result in omission of candidate regions.

## 7.7 References

1. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, et al. Large-scale copy number polymorphism in the human genome. Science. 2004;305(5683):525-8.
2. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number in the human genome. Nature. 2006;444(7118):444-54.
3. Pinto D, Marshall C, Feuk L, Scherer SW. Copy-number variation in control population cohorts. Hum Mol Genet. 2007;16 Spec No. 2:R168-73.
4. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, et al. Origins and functional impact of copy number variation in the human genome. Nature. 2010;464(7289):704-12.
5. Gibson J, Morton NE, Collins A. Extended tracts of homozygosity in outbred human populations. Hum Mol Genet. 2006;15(5):789-95.
6. McQuillan R, Leutenegger AL, Abdel-Rahman R, Franklin CS, Pericic M, Barac-Lauc L, et al. Runs of homozygosity in European populations. Am J Hum Genet. 2008;83(3):359-72.
7. Nothnagel M, Lu TT, Kayser M, Krawczak M. Genomic and geographic distribution of SNP-defined runs of homozygosity in Europeans. Hum Mol Genet. 2010;19(15):2927-35.
8. Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010 ;467(7311):52-8.
9. Craddock N, Hurles ME, Cardin N, Pearson RD, Plagnol V, Robson S, et al. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature. 2010;464(7289):713-20.
10. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56-65.
11. Ku CS, Loy EY, Salim A, Pawitan Y, Chia KS. The discovery of human genetic variations and their use as disease markers: past, present and future. J Hum Genet. 2010;55(7):403-15.
12. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature. 2001;409(6822):928-33.
13. Estivill X, Armengol L. Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. PLoS Genet. 2007;3(10):1787-99.
14. Wright A, Charlesworth B, Rudan I, Carothers A, Campbell H. A polygenic basis for late-onset disease. Trends Genet. 2003;19(2):97-106.
15. Campbell H, Carothers AD, Rudan I, Hayward C, Biloglav Z, Barac L, et al. Effects of genome-wide heterozygosity on a range of biomedically relevant human quantitative traits. Hum Mol Genet. 2007;16(2):233-41.
16. Ehret GB, Munroe PB, Rice KM, Bochud M, Johnson AD, Chasman DI, et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature. 2012;478(7367):103-9.
17. Boonpeng H, Yusoff K. The utility of copy number variation (CNV) in studies of hypertension-related left ventricular hypertrophy (LVH): rationale, potential and challenges. Mol Cytogenet. 2013; 6(1):8.
18. Kearney PM, Whelton M, Reynolds K, Muntner P, Whelton PK, He J. Global burden of hypertension: analysis of worldwide data. Lancet. 2005;365(9455):217-23.

19. Rudan I, Smolej-Narancic N, Campbell H, Carothers A, Wright A, Janicijevic B, et al. Inbreeding and the genetic complexity of human hypertension. Genetics. 2003;163(3):1011-21.
20. Nishimoto K, Tomlins SA, Kuick R, Cani AK, Giordano TJ, Hovelson DH, et al. Aldosterone-stimulating somatic gene mutations are common in normal adrenal glands. Proc Natl Acad Sci U S A. 2015;112(33):E4591-9.
21. Funder JW. The genetic basis of primary aldosteronism. Curr Hypertens Rep. 2012;14(2):120-4.
22. Marques FZ, Prestes PR, Pinheiro LB, Scurrah K, Emslie KR, Tomaszewski M, et al. Measurement of absolute copy number variation reveals association with essential hypertension. BMC Med Genomics. 2014; 7:44.
23. Koerbin G, Cavanaugh JA, Potter JM, Abhayaratna WP, West NP, Glasgow N, et al. 'Aussie normals': an a priori study to develop clinical chemistry reference intervals in a healthy Australian population. Pathology. 2015;47(2):138-44.
24. Glessner JT, Bick AG, Ito K, Homsy JG, Rodriguez-Murillo L, Fromer M, et al. Increased frequency of de novo copy number variants in congenital heart disease by integrative analysis of single nucleotide polymorphism array and exome sequence data. Circ Res. 2014;115(10):884-96.
25. Girirajan S, Eichler EE. Phenotypic variability and genetic susceptibility to genomic disorders. Hum Mol Genet. 2010;19(R2):R176-87.
26. MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW. The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res. 2014;42(Database issue):D986-92.
27. Nusbaum C, Zody MC, Borowsky ML, Kamal M, Kodira CD, Taylor TD, et al. DNA sequence and analysis of human chromosome 18. Nature. 2005;437(7058):551-5.
28. Kleinjan DA, van Heyningen V. Long-range control of gene expression: emerging mechanisms and disruption in disease. Am J Hum Genet. 2005;76(1):8-32.
29. Jaber L, Shohat T, Rotter JI, Shohat M. Consanguinity and common adult diseases in Israeli Arab communities. Am J Med Genet. 1997;70(4):346-8.
30. Ben Halim N, Ben Alaya Bouafif N, Romdhane L, Kefi Ben Atig R, Chouchane I, Bouyacoub Y, et al. Consanguinity, endogamy, and genetic disorders in Tunisia. J Community Genet. 2013;4(2):273-84.
31. Feuk L, Marshall CR, Wintle RF, Scherer SW. Structural variants: changing the landscape of chromosomes and design of disease studies. Hum Mol Genet. 2006;15 Spec No 1:R57-66.
32. Charchar FJ, Zimmerli LU, Tomaszewski M. The pressure of finding human hypertension genes: new tools, old dilemmas. J Hum Hypertension. 2008; 22: 821-28.

# CHAPTER 8

# Comparison of CNV Outputs from Two CNV Detection Software Programs

Part of this chapter was published in

Discordance between CNV detection software tools:

The challenges for the diagnostic setting.

Nicole L. Chia, Howard R. Slater, Julia M. Potter

Current Topics in Genetics, Vol 6, 2016

## 8.1 Abstract

Copy number variation (CNV) detection software tools applied to data generated from microarray platforms have been shown to yield varied results. However the variation of CNV calling and limitations of CNV detection programs are not widely appreciated in the diagnostic setting. In this study two CNV detection software programs were applied to generate CNV calls from the same raw data derived from a healthy female population cohort (n=64). The CNV outputs from the software programs were compared for the genomic coordinates, CNV length and copy number type for each CNV. Comparative analysis showed that the total numbers of calls differed for the software applications. CNV detected by one program accounted for 39.1% of calls > 1 kb and these CNV measured up to 350 kb in size. In CNV >100kb call fragmentation accounted for 12% and discordance of CNV length was observed in 19.9% of calls and recorded up to 500kb in length. The sensitivity and specificity of selected CNV regions (CNVR) were assessed by confirmation using alternative experimental methods. The current study demonstrates differences between the detection software programs that may impact on diagnostic interpretation. The need for awareness of the limitations of the CNV detection algorithms used in microarray investigations is emphasised and recommend careful assessment of CNV calls in clinical diagnosis. These findings, in conjunction with previous reports support the need for inclusion of software programs in laboratory validation and report protocols.

## 8.2 Introduction

CNV detection software programs are applied to the data generated from microarray platforms to estimate copy number state. There are a number of commercially available microarray platforms achieving a range of levels of resolution. There is an equally varied number and type of programs for CNV detection, ranging from commercially available platform specific programs such as CNV Partition (Illumina, Inc.) and Chromosome Copy Number Analysis Tool (CNAT, Affymetrix) to free access program such as PennCNV and QuantiSNP (1, 2). Within these software suites are bio-informatic algorithms that are employed to predict copy number change (1-5). These software programs were created for research purposes with specific goals such as prediction of disease association in genome wide association studies (4, 5). This technology has been rapidly adopted in diagnostic genetics laboratories where interpretation is in the clinical context for individual persons (6, 7).

A small number of studies have compared the CNV outputs attained from different platforms whilst others have compared the CNV detection software and/or algorithms for the accuracy of CNV calls (2-5, 7, 8). Discordance of CNV calling including CNV length, copy number state and presence of CNV event was reported by all studies. Modifications to genome coverage have provided increased resolution, particularly in coding regions, with subsequent calling of smaller CNV. However awareness of the limitations and performance characteristics of CNV detection software programs is not widely appreciated by those using the technology for diagnostic applications.

The purpose of this study is to

1) Compare CNV calling from two frequently used CNV detection software programs.
2) Identify differences in the CNV outputs from the programs
3) Investigate the causes of discordance of CNV outputs
4) Consider the significance for diagnostic laboratories
5) Explore ways to identify false calls in the diagnostic laboratory setting
6) Make recommendations to minimise the risk for CNV detection

To do this the study evaluates several parameters to determine the causes of CNV discordance and explores characteristics to assist in CNV interpretation and recognition of false calls. The significance of the discordance for clinical laboratories is highlighted supporting recommendations for comprehensive validation protocols. This study differs to others in the mode of analysis of discordant CNV calls and it is understood this is the first study to stratify the discordance of CNV calls and analyse CNV from a general population.

## 8.3. Materials and Methods

### 8.3.1 Comparison of CNV detection software outputs

A genome-wide investigation of 64 healthy female individuals representing a cohort of females from a Western European descendent population was performed using the Illumina Omni1-Quad (Illumina, Inc.) (9). This platform has a median probe interspacing of 1.2kb and overall resolution of 5 kb (Illumina, Inc.). Two CNV

calling software programs were applied to the same raw copy number and genotyping data. The CNV output from CNV Partition v2.3.4, a proprietary platform-specific software for Illumina and PennCNV (1) an academic freeware program designed for use on Illumina platforms are exported to an excel file (Table 1). The standard Illumina clusterfile was used for the determination of genotypes. To provide consistency of CNV detection the default parameters recommended by the manufacturers were applied. The format of the data extracted includes

1) Sample identification number
2) Chromosome number
3) CNV start position
4) CNV end position
5) CNV length
6) Copy number state
7) Number of markers (PennCNV)

**Table 1.** The format of the excel file output for CNVs a) CNV Partition and b) PennCNV.

a)

| Sample ID | Chr | Start | End | Size | Value |
|---|---|---|---|---|---|
| 2307 [6] | 9 | 105609407 | 106307173 | 697766 | 1 |

b)

| | | | | |
|---|---|---|---|---|
| chr9:105611023-106307173 | Num snp=223 | length=696,151 | state2,cn=1 | samplesplit6.txt |

The CNV calls from chromosome 1-22 were included in the study. The CNV outputs are compared by visual inspection of calls in excel for sample number, chromosome, CNV length and start and end positions. Program specific calls are stratified according to CNV length, copy number status, chromosome location and association with genome architecture. Discordant CNV calls > 100kb are further investigated for CNV length, copy number type, LogR, marker density and association with genomic architecture to ascertain criteria that can be used by diagnostic laboratories to evaluate the confidence of a CNV call.

## 8.3.2 Determining CNV concordance

The CNV outputs from CNV Partition and PennCNV were inspected to identify shared calls and establish a merged list of CNV that is a robust representation of CNV in the study cohort. Two levels of assessment were applied to the data.

1) Sample identification
    a. To identify CNV calls detected by only one algorithm " program specific"
2) Call concordance
    a. To identify the level of variation of breakpoint estimation and copy number state

In step one a program specific call is recorded where a CNV is reported in a sample in only one of the programs. In step two (Figure 1a and 1b) a CNV call is scored as concordant between the CNV outputs where there is

a) consistency in start or end position,

b) the CNV is within a larger CNV and/or

c) greater than 50% overlap.

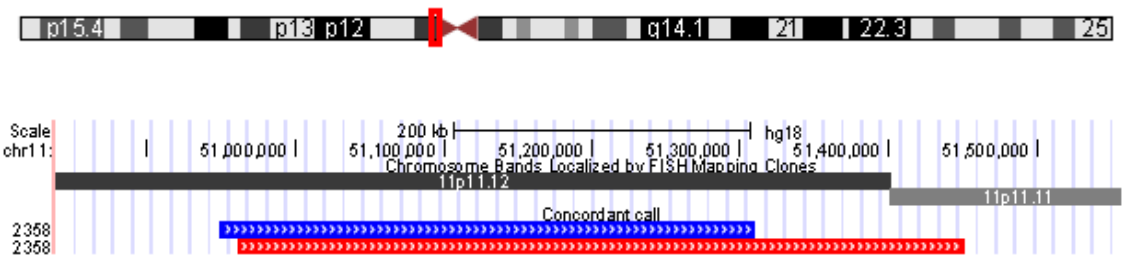



**Figure 1a.** A 359kb CNV gain recorded as concordant between CNV Partition (blue) and PennCNV (red) (http://genome.ucsc.edu). The CNV calls overlap by > 50%.

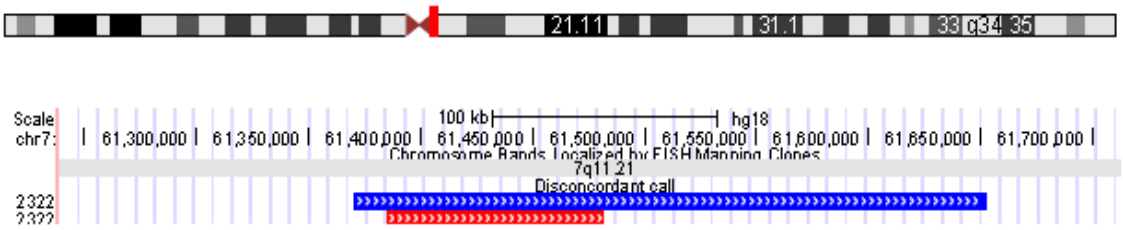



**Figure 1b.** A 282kb loss in CNV Partition (blue) is reported as a 96kb loss by PennCNV (red). This is recorded as length discordance with < 50% overlap.

### 8.3.3 QC metrics are applied for CNV detection

Assessment of the run QC showed that all samples achieved a genotype call rate of 99.8% and the sample standard deviation of the LogR is <0.28. The assay achieved an average LogRdev of 0.13 and BAFdev 0.029. The following quality metrics were applied to assess the quality of CNV calls. A minimum of 4 consecutive markers with LogR <-0.3 (loss) or >0.15 (gain) was required to detect a genuine call. This threshold was determined by calculating the mean LogR value of 40 samples with copy number 2 in 4 regions and compared with the LogR values of a total of 30 regions that recorded a copy number of 0, 1, 3, 4.

### 8.3.4 Confirmation of CNV

The sensitivity and specificity for selected CNVR were evaluated by determining the true positive and true negative CNV using alternative methods. To identify false calls and evaluate accuracy in breakpoint estimation, CNV were selected for confirmation according to the following criteria:

1) The CNV is present in multiple samples
2) The CNV is called by one and both algorithms in multiple samples
3) The start or end positions are the same OR
4) The start and end positions differ between samples

### 8.3.5 Confirmation of CNV breakpoints

Two CNVR < 5kb were experimentally confirmed by PCR and sequencing. The PCR method, previously described in Chapter 3 and 4 is designed to differentiate wild-type, homozygous and heterozygous deletions. Breakpoints were confirmed by capillary sequencing giving an accurate assessment of CNV length (9).

A selection of CNV >100kb were confirmed by fluorescence in-situ hybridisation (FISH) investigation using RP11 BAC clones. FISH probes were selected within the CNV region using custom tracks applied to the UCSC browser (http://genome.ucsc.edu). For CNV verification, FISH signals of samples with the CNV were compared to a sample without the CNV (copy number 2) as the negative control. Estimation of CNV length could not be made by FISH methods.

## 8.4 Results

### 8.4.1 Comparison of outputs from CNV detection software programs

The current study compares the detection of CNV from two CNV detection software programs, CNV Partition and PennCNV using the same raw data from Illumina Omni1-Quad SNP microarray. CNV events called by PennCNV that were also called by CNV Partition accounted for 61% of CNV calls > 1kb and increased to 365/415 (88%) in CNV >100kb. CNV Partition performed better with shared events >1kb accounting for 88.3% and 456/463 (98%) of calls >100kb.

## 8.4.2 Types of discrepancy between the CNV outputs

Differences were identified between the two outputs. These ranged from CNV detected by a single CNV detection program, discrepancy in breakpoint estimation resulting in variation of estimated length, and multiple CNV in one program correlating with a single CNV in the other (Figure 2). The discordant CNV were investigated further to determine a causative factor and if there is any significance for CNV detection in diagnostic laboratories.

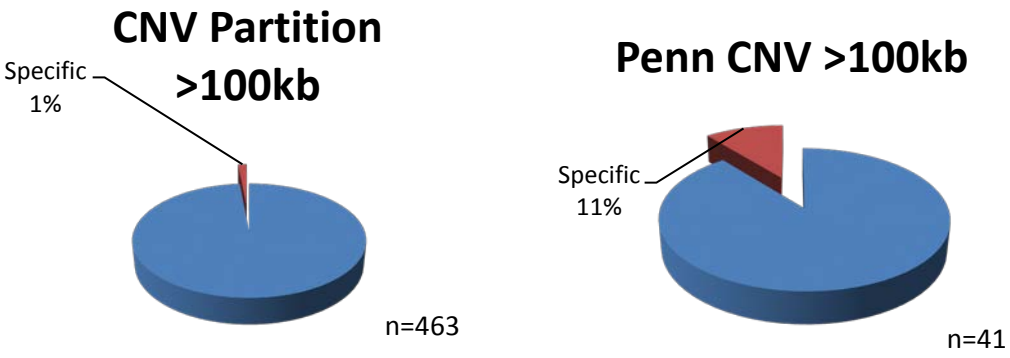



**Figure 2a.** The proportion of program specific CNV calls for CNV Partition and PennCNV.

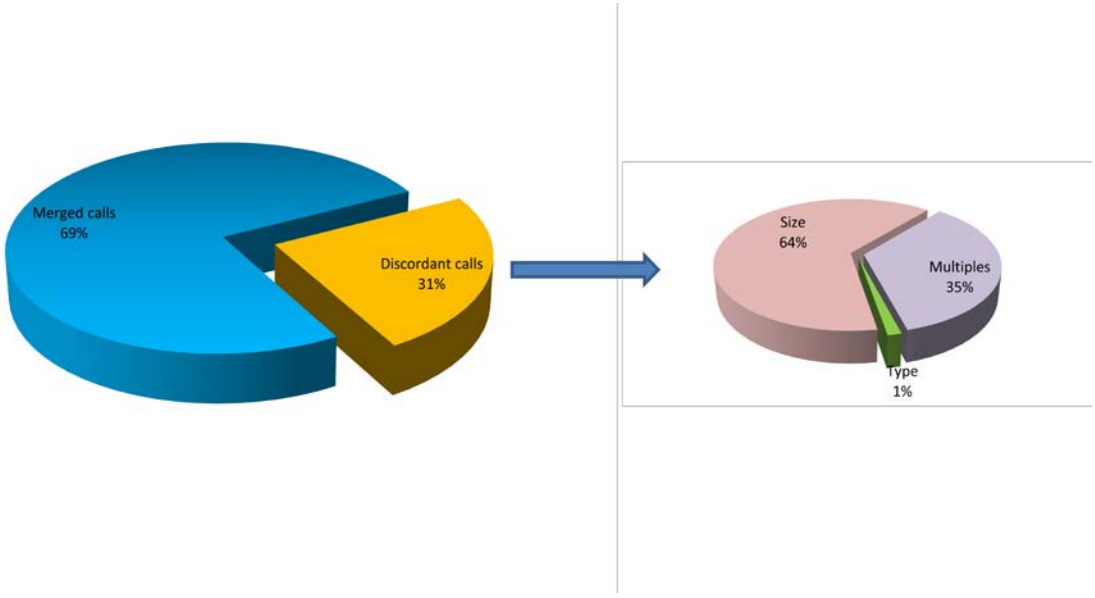



**Figure 2b.** The proportion of classes of discordance of CNV >100 kb. 31% show discordance of CNV length, one algorithm were called as multiple smaller CNVs by another, and some of these had altered copy number state.

## 8.4.2.1 Investigation of program specific CNV

Overall PennCNV called more CNV >1kb on chromosomes 1-22 than CNV Partition (17,292 and 13,422 respectively). The CNV outputs were compared for sample level concordance of CNV calls. CNV detected in a sample by a single algorithm represent either a false positive by the CNV detection program that made the call or false negative by the program that did not detect the CNV. PennCNV detected 6765 CNV > 1kb that are not detected by CNV Partition and 1359 CNV > 1kb are reported only in CNV Partition. The program specific calls were stratified according to length, chromosome distribution and copy number state.

### 8.4.2.1.1 Length distribution of program specific CNV

Sample level concordance of CNV outputs improved with increasing length of CNV (Figure 3). For CNV Partition 98% of program specific CNV are <100kb and 99% are recorded for Penn CNV. There was no evidence of program specific discordance >350kb.

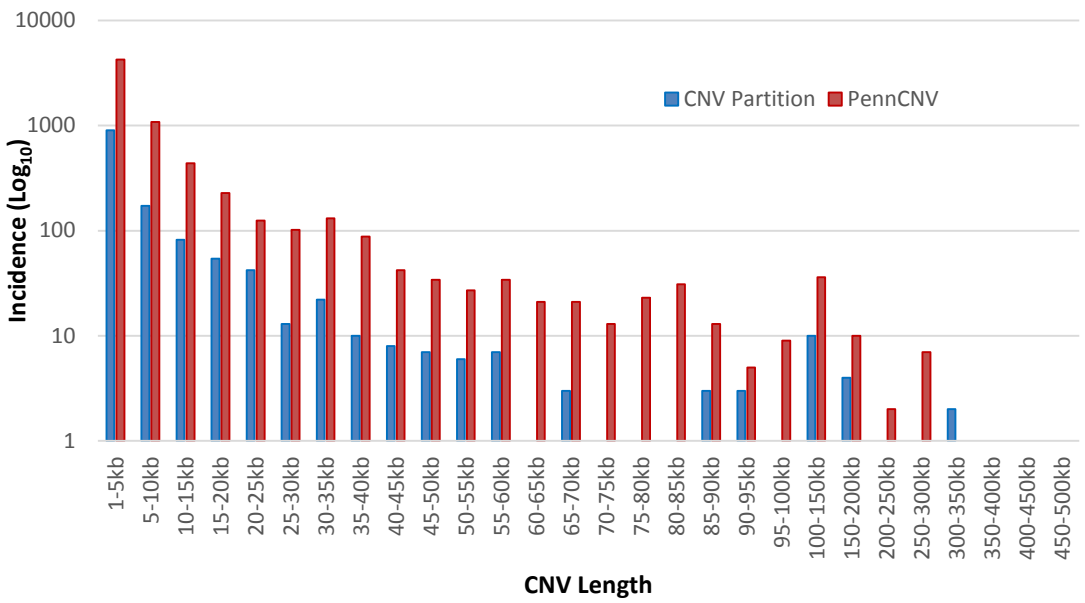



**Figure 3**. The incidence and length of CNV calls made only by CNV Partition (blue) or Penn CNV (red). The incidence of program specific CNV calls decreased with increasing size of the CNV and is apparent up to 350kb.

### 8.4.2.1.2 Chromosome distribution of program specific CNV

Inter-chromosomal variation of CNV incidence was observed between the software CNV outputs (Figure 4). Chromosome 1, 3, 5, 8, 13, 19 and 21 show up to 98% concordance, while the largest difference in the incidence of CNV between the CNV detection programs is observed for chromosomes 2, 4, 7, 9, 15, 16 and 17.

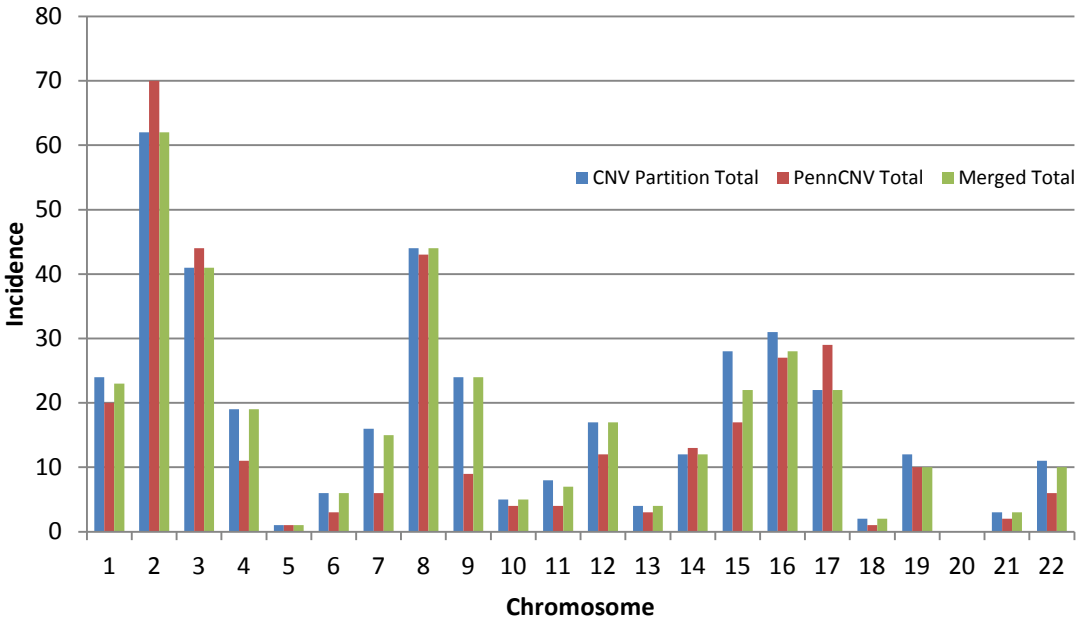



**Figure 4.** Comparison of the total number of CNV > 100kb called for CNV Partition and PennCNV for each chromosome. The merged total represents the sample level concordance of CNV events between the two algorithms. Call fragmentation and different CNV length estimates results in a lack of correspondence of the number of concordant calls with the CNV outputs.

## 8.4.2.1.3 Copy number state of program specific CNV

The program specific CNV were stratified according to copy number type (Figure 5). Analysis of the CNV>100kb made by either PennCNV or CNV Partition identified a predominance of duplication accounting for 64%.

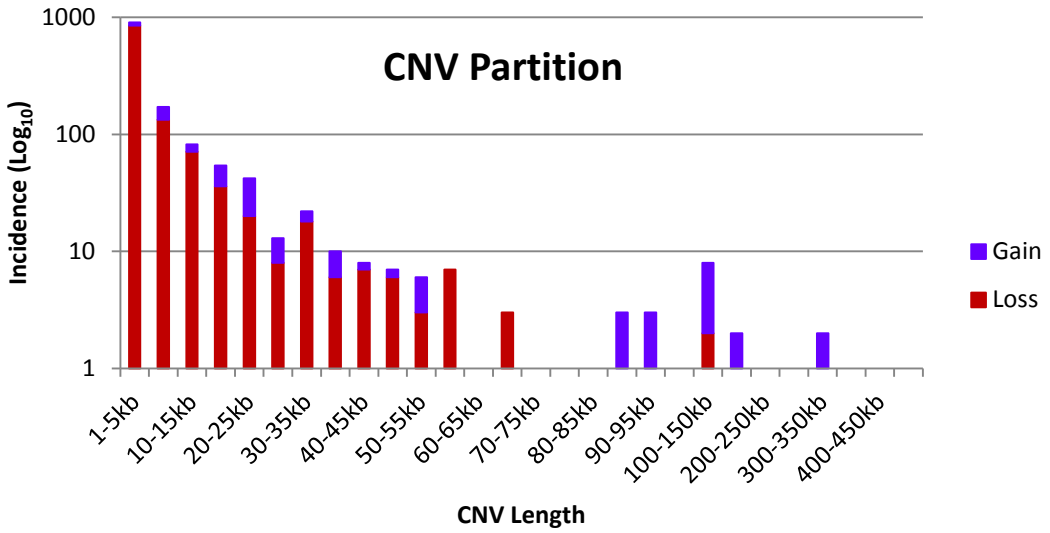



**Figure 5a.** Stratification of length of CNV and copy number made in CNV Partition and not by PennCNV. The 'X' axis represents the CNV length, the Y axis represent the $\log_{10}$ of the number of CNV. Calls exclusive to CNV Partition are <350 kb. CNV gain is predominant in CNV >85kb.

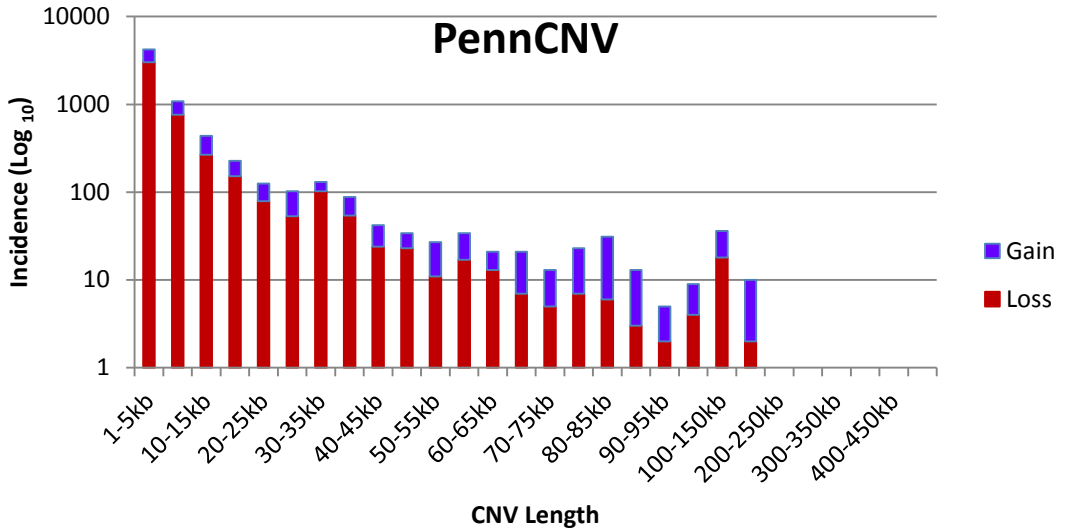



**Figure 5b.** Stratification of length of CNV and copy number called in PennCNV and not in CNV Partition. The X axis represents the CNV length, the Y axis represents the $\log_{10}$ of the number of CNV. Calls exclusive to PennCNV are < 250 kb.

## 8.4.2.2 Association of program specific CNV with genomic architecture

Common CNV accounted for 39/50 of the calls made only by PennCNV. These calls are shown to be sample specific as the CNVR was detected by both software applications for other samples. Program specific calls by CNV Partition represented singletons in 4/7 (57%). Review of the genomic architecture using UCSC browser with segmental duplication (SD) and repeat masker tracks applied, identified a correlation of these CNV with segmental duplication with 49/57 (86%) falling within segmental duplication regions, 2 CNV overlapped both segmental

duplication and a region of repetitive elements (RE) and the remainder are within RE regions (Figure 7).

However there was less association with chromosomal morphology where CNV in pericentromeric regions accounted for 29/57 (50.8%) of the discordant calls (Figure 6). Of these, 19/29 (65.5%) are common CNV and occur in segmental duplication regions, the remaining are singletons or low frequency variants within regions of repetitive elements or segmental duplication.

### 8.4.2.3 Investigation of QC metrics for program specific CNV

The LogR of program specific calls was analysed to determine if it differed from the expected value or from samples where the CNV was called by both programs. The average LogR for CNV >1kb detected only by PennCNV on chromosome 18 is -0.308 for loss and 0.228 for gain (n=120). This is consistent with the expected LogR of < -0.30 for loss or >0.15 for gain (Table 2).

The number of probe markers within the CNV was recorded to determine if this may be an indicator of the false calls. An average of 40 markers is recorded in CNV >100kb for samples where the CNV is called by a single program. This is compared to an average of 61 markers in samples where both programs called the CNV. For the CNV>100kb detected only by PennCNV (n=50), 46/50 recorded >20 markers and an average of 1 marker per 4.6kb, compared to an average of 1 marker per 3.26kb for CNV >100kb called by both programs.

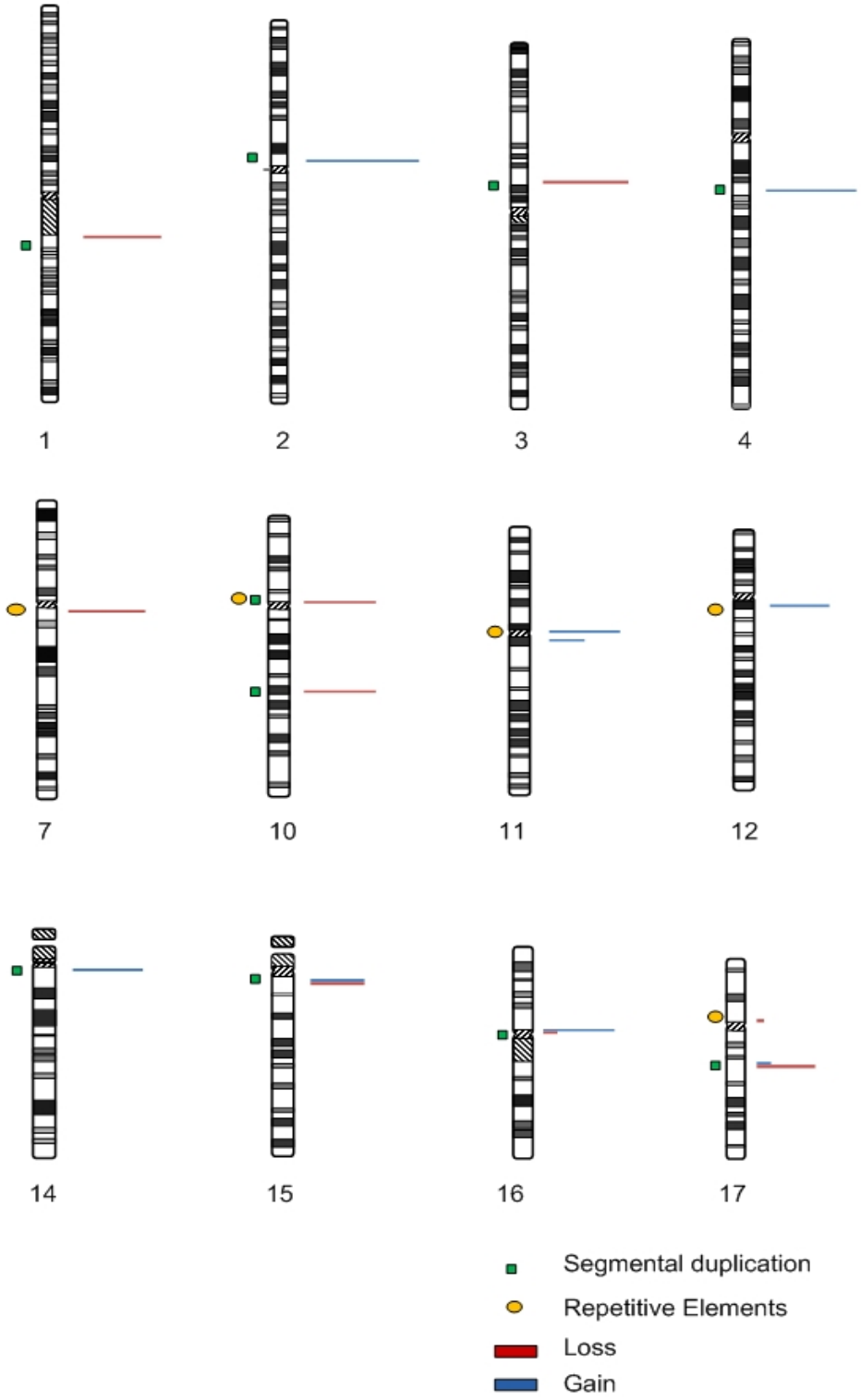



**Figure 6.** Location of program specific CNV > 100 kb and association with genomic architecture. The size of the bars represents the incidence of the CNV.

**Table 2.** Comparison of LogR values for sample level discordance. There is no significant difference of LogR values in program specific CNV compared to merged CNV.

| CNV | chr2:88925413-89090893 | chr16:34336806-34464860 | chr17:42003374-42132984 |
|---|---|---|---|
| Program specific CNV | 0.301<br>n=9 | 0.21<br>n=6 | -0.305<br>n=8 |
| CNV called in both software | 0.303<br>n=22 | 0.196<br>n=1 | -0.324<br>n=8 |
| Length<br>Copy Number | 168kb<br>gain | 270kb<br>gain | 128kb<br>loss |

The performance of the SNP markers within the CNV was reviewed. The number of SNP markers recording a genotype or "no call" (NC) was recorded. SNP marker "no call" was observed for the genotypes of >60% of SNP markers in high incidence CNVR and CNVR with a high rate of sample level discordance including false calls, breakpoint estimation and fragmentation of a CNV call. In contrast private and low frequency variants showed <3% of SNP markers with "no call" (Table 3).

**Table 3.** Evaluation of SNP marker performance for common and low frequency CNV and CNV with evidence of sample level discordance

| CNV | Both Programs | CNV Partition | PennCNV | Total Markers | SNP Markers | Failed SNP Markers | % Failed SNP Markers |
|---|---|---|---|---|---|---|---|
| **Common CNV** | | | | | | | |
| chr2:88930087-89090893 | 54 | 54 | 64 | 60 | 22 | 19 | 86.36 |
| chr3:163992511-164108624 | 34 | 34 | 37 | 27 | 24 | 23 | 95.83 |
| chr8:39352501-39506336 | 41 | 41 | 41 | 45 | 24 | 24 | 100.00 |
| **Sample level discordance** | | | | | | | |
| chr3:75502426-75621454 | 3 | 3 | 7 | 27 | 25 | 14 | 56.00 |
| chr4:69064675-69173198 | 10 | 10 | 12 | 28 | 24 | 16 | 66.67 |
| chr16:33298106-33537523 | 12 | 12 | 12 | 56 | 43 | 35 | 81.40 |
| chr16:34343935-34614585 | 1 | 1 | 6 | 35 | 27 | 2 | 7.41 |
| chr16:33384384-33537523 | 12 | 12 | 13 | 56 | 43 | 35 | 81.40 |
| chr17:42003374-42132984 | 5 | 5 | 16 | 29 | 11 | 8 | 72.73 |
| **Low Frequency CNV** | | | | | | | |
| chr1:202165699-202384983 | 1 | 1 | 1 | 113 | 113 | 0 | 0.00 |
| chr1:202457035-202838694 | 1 | 1 | 1 | 188 | 170 | 2 | 1.18 |
| chr2:86149308-86363012 | 1 | 1 | 1 | 105 | 105 | 2 | 1.90 |
| chr7:121033877-121359954 | 1 | 1 | 1 | 81 | 74 | 2 | 2.70 |
| chr12:1727243-1888251 | 1 | 1 | 1 | 87 | 86 | 2 | 2.33 |
| chr12:45519568-45772649 | 1 | 1 | 1 | 72 | 72 | 0 | 0.00 |

## 8.4.3 Discordance of breakpoint estimation

The start and end positions and estimated CNV length of shared CNV calls were compared. Discordance of CNV length was observed between the two outputs for some CNV. CNV length difference of > 50% was found in 92/463 calls (19.9%). The

inconsistency in reported CNV length was apparent up to 500kb. Likewise the discordance of CNV lengths reported by PennCNV increased with increasing size of the CNV (Figure 7).

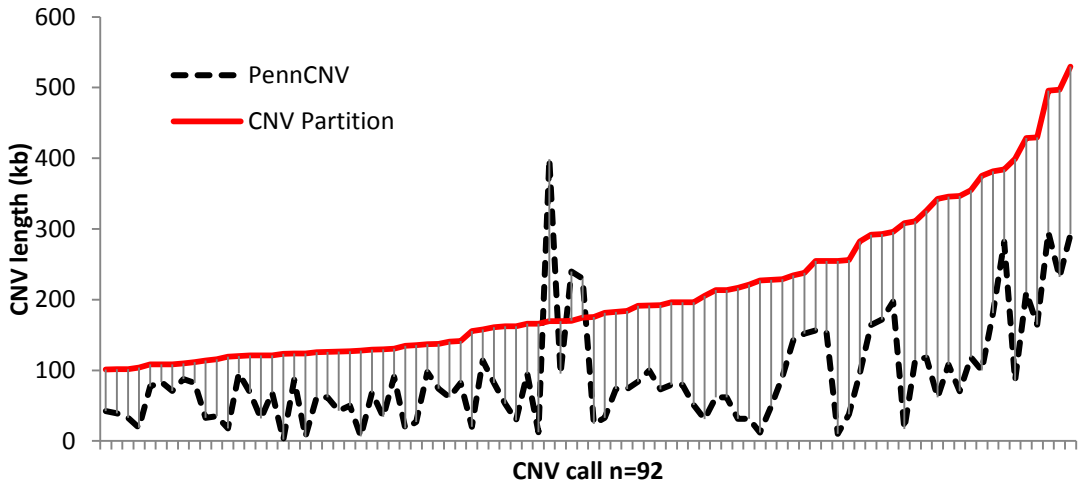



**Figure 7**. Comparison of CNV Partition (solid line) and PennCNV (dashed line) for CNV length of 92 CNV >100kb. The discordance in CNV size is demonstrated by the high low lines. There is an increase in discordance with increasing size of the calls in CNV Partition up to 500kb. There is no apparent CNV length discordance recorded in this cohort in CNV calls >500kb.

## 8.4.3.1 Experimental investigation of putative breakpoints

To evaluate the accuracy of breakpoint estimation of the software programs, the breakpoints of a selected CNV deletion at chromosome 18q12.3 (38,308,102-38,311,523) were confirmed by "walking" PCR and sequencing. According to the output by PennCNV the start and end positions and CNV length varied between individuals. The PCR product and sequencing results were consistent with the

same breakpoint for all samples assessed. This confirmed that deletions in 16 individuals were the same in length and breakpoint position. The breakpoints determined by sequencing are consistent with the start and end position estimated by CNV Partition.

The LogR for markers flanking the putative breakpoints and within the CNV were investigated and a fluctuation in the LogR values for individual SNP markers (Figure 8) was identified. This instability in LogR correlates with the breakpoints called by PennCNV, whereas the start and end position indicated by CNV Partition are consistent with the sequence position.

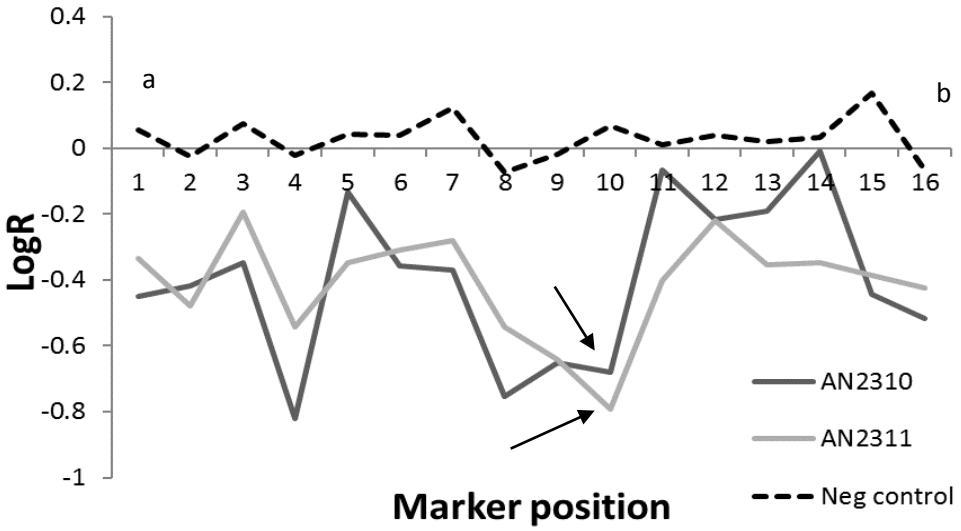



**Figure 8.** Variation of LogR ratio (y-axis) correlates with inconsistency of breakpoint calling. A 3kb deletion on chromosome 18q12.3 (38,308,078–38,311,652), (Chia et al. 2013) is mapped for two samples and one negative control (dashed line). The CNV confirmed by sequencing spans the entire length of the plot

(a to b). The same start position was called by both programs. The arrows indicate the end position called by PennCNV.

To investigate the involvement of GC content on the fluctuation of LogR and breakpoint estimation, the GC content of the sequence between the markers was recorded. Using UCSC genome browser and repeat masker tools the GC content is estimated and correlated against the LogR. The start and end positions of a deletion in sample (AN 2311) by CNV Partition is chr18:38308102-38311523 and Penn CNV is chr18:38308102-38309912 (Table 4). The end position for PennCNV (chr18: 38309912) correlates with a LogR -0.64. There is an increase in the LogR of the next consecutive markers to -0.11 and -0.15. These markers coincide with > 40% GC and are within regions of ALU sequences and simple repeats (http://genome.ucsc.edu). In contrast CNV partition encompassed all markers to chr 18: 38311523.

**Table 4**. GC content between consecutive markers in 18q12.3 (38,308,102-38,311,523)

| Locus | Position | LogR AN 2311 | GC % between consecutive loci | Repeat % | Type |
|---|---|---|---|---|---|
| rs346231 | 38308102 | -0.3047009 | | | |
| cnvi0072434 | 38308266 | -0.3571239 | 35.15 | | |
| cnvi0072435 | 38308484 | -0.2388388 | 43.84 | 52 | LTR |
| cnvi0072436 | 38308709 | -0.6967938 | 32.3 | 69 | LTR |
| cnvi0072437 | 38308891 | -0.2487170 | 38 | 0 | |
| rs346232 | 38309077 | -0.5158722 | 34.22 | 12.83 | LowC |
| cnvi0072438 | 38309347 | -0.4761522 | 38.45 | 0 | |
| cnvi0072439 | 38309559 | -0.7647809 | 34.74 | 0 | |
| rs10468964 | 38309785 | -0.4093818 | 24.6 | 0 | |
| cnvi0072441 | 38309912 | -0.6408494 | 31.62 | 0 | |
| cnvi0072442 | 38310331 | -0.1137405 | 40.24 | 50.24 | Alu |
| cnvi0072443 | 38310534 | -0.1560823 | 47.55 | 25 | Simple |

| | | | | | |
|---|---|---|---|---|---|
| cnvi0072444 | 38310742 | -0.1731966 | 43 | 0 | |
| cnvi0072445 | 38310988 | -0.1803744 | 39 | 0 | |
| rs16976167 | 38311238 | -0.2594561 | 38 | 0 | |
| cnvi0072446 | 38311523 | -0.6982208 | 47 | 12.59 | Simple GT(n) |

## 8.4.4 Evidence of CNV fragmentation

Fragmentation of calls, where a single CNV in one of the outputs is recorded as multiple smaller CNV in the other, was observed for 50/415 (12%) PennCNV calls (Figure 9 and 10). The effect of fragmentation was greater for PennCNV with no evidence of fragmentation recorded in CNV Partition. This finding was evaluated up to 1.4Mb which is the largest CNV called in this healthy cohort and no evidence of call fragmentation was observed in CNV >500kb. This anomaly may result in an underestimation of the length of the CNV and contributes to the overall higher number of calls made by PennCNV. The significance of this is highlighted in the CNV demonstrated in Figure 10a and 10b, where a 464kb loss at 15q13.3 encompassing 8 RefSeq genes in CNV Partition is called by PennCNV as two CNV measuring 120kb and 25 kb and involves no genes. The size of the CNV called by CNV Partition would be evaluated in a clinical setting, whereas the CNVs called by Penn CNV for the same sample is below general size thresholds and would not be investigated.

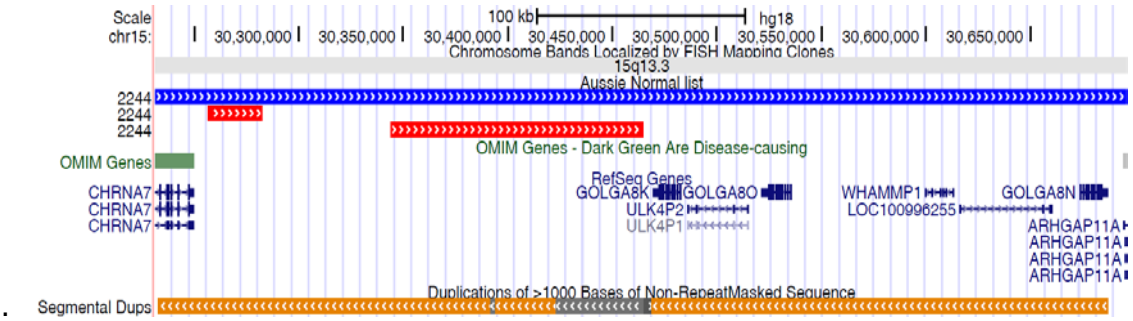



**Figure 9**. Image from UCSC Genome Browser (http://genome.ucsc.edu)(10) showing a CNV call from CNV Partition (blue) and PennCNV (red). Call fragmentation is observed as a 25kb and 120kb deletion in PennCNV corresponding to a single event of 467kb in CNV Partition for the same individual (AN2244). This CNV corresponds with a region of segmental duplication.

## 8.4.5 Copy number state discordance

Copy number state was also evaluated between the outputs and inconsistencies demonstrated. This was observed in association with call fragmentation by PennCNV with a copy number 0 and 1, consistent with homozygous and heterozygous deletions respectively, for a single copy number 1, heterozygous deletion in CNV Partition. This anomaly was observed in 2 calls >100kb (0.4%).

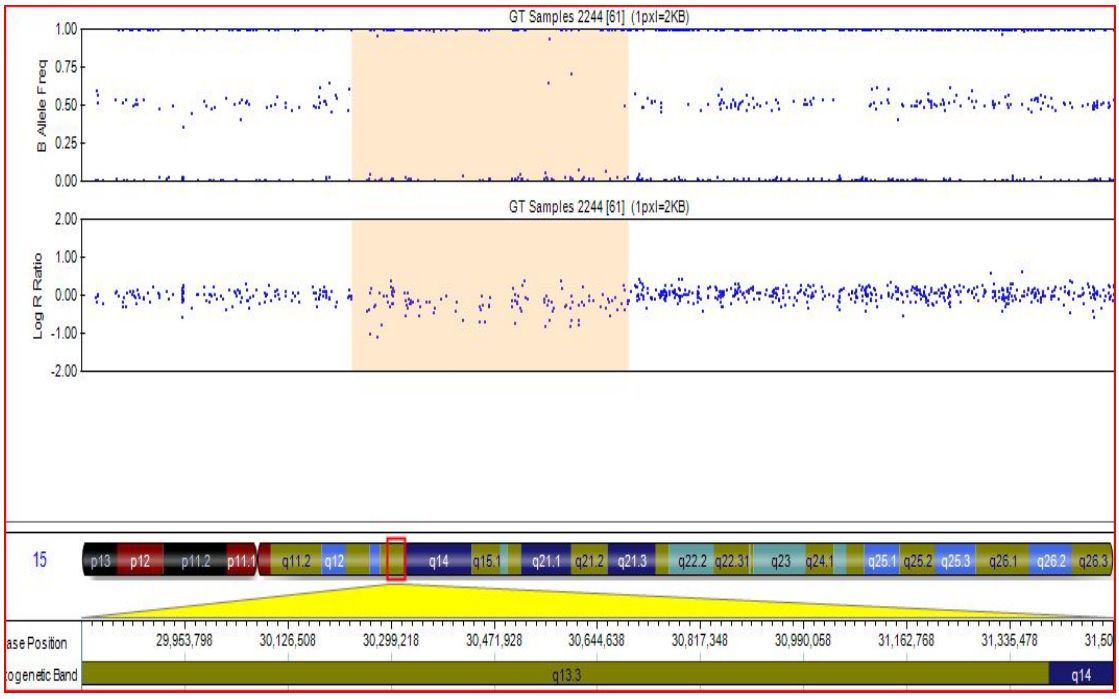



**Figure 10.** Screen shot of the CNV plot in Genome Studio (Illumina, Inc). The 467kb CNV loss called by CNV Partition is highlighted by the pink zone. Call fragmentation by PennCNV is demonstrated in Figure 9.

## 8.4.6 Confirmation of CNV

### 8.4.6.1 CNV <5kb confirmed by molecular methods

To investigate the sensitivity and specificity, two common CNV <5kb on chromosome 18 in 64 subjects were selected for experimental investigation by PCR. One common deletion at 18q12.3 (38,308,078–38,311,652) was called in 16/64 individuals by PennCNV and 8/64 by CNV Partition. All 16 calls (100%) made by PennCNV were confirmed by PCR and 4/4 were verified by DNA sequence analysis. The sensitivity and specificity for PennCNV at this CNV is 100%

(www.Medcalc.org). All 8 calls (100%) made by CNV Partition were confirmed by PCR. However, CNV Partition failed to detect deletions in 8/16 (50%) individuals that were made by PennCNV and confirmed by PCR. This suggests a false negative call by CNV Partition in 8/64 samples. The sensitivity for this CNV detected by CNV Partition is 50% and specificity of 100% and negative predictive value of 85.7% (Table 5).

A second common deletion at 18q21.2 (49,390,404-49,391,772) recorded an incidence of 8/64 by PennCNV. All of the calls made by PennCNV were confirmed by PCR and 2/2 by sequencing, representing a sensitivity and specificity 100%. CNV Partition detected 6/64 and failed to detect 2/8 of the CNV called by Penn CNV (Figure 11). This represents a sensitivity of 75%, specificity of 100% and negative predictive value of 96.55% (www.medcalc.org).

**Table 5.** Sensitivity and specificity of CNV Partition and PennCNV for selected CNV.

| | | | | | CNV Partition | | | PennCNV | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Chr.band | Copy number | Length | LogR | Confirmation | Incidence | Sensitivity | Specificity | Incidence | Sensitivity | Specificity |
| 18q12.3 | Loss | 3.2kb | -0.35 | PCR/SEQ | 8/64 | 50 | 100 | 16/64 | 100 | 100 |
| 18q21.2 | Loss | 1.2kb | -0.39 | PCR/SEQ | 6/64 | 75 | 100 | 8/64 | 100 | 100 |
| 16p11.1 | Gain | 277kb | 0.19-0.2 | FISH | 1/64 | 64 | 100 | 6/64 | 100 | 100 |

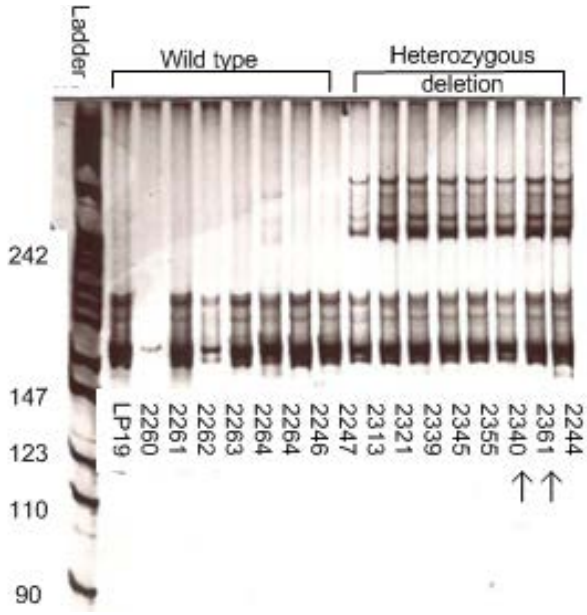



**Figure 11.** Confirmation of CNV18027 at chr 18:49,390,404-49,391,772 by PCR and visualisation on polyacrylamide gel. PCR products are demonstrated for wild type (159 bp only) and heterozygous deletions (159/240bp). PennCNV called a heterozygous deletion in 8 samples and CNV Partition called the deletion in only 6 of these samples. The two samples where a CNV was not called by CNV Partition (2340 and 2361) are confirmed as heterozygous deletions by PCR (arrowed).

## 8.4.6.2 CNV >100kb

CNV >100kb were verified by FISH. A 270kb common copy number gain on chromosome 16p11.1 was called only by PennCNV in 5/64 subjects and 1/64 was called by both CNV Partition and PennCNV (Figure 12a). Dual colour FISH analysis using clones RP11-488I20 and RP11-80F22 verified the copy number gain in all 6/6 events representing sensitivity and specificity of 100% for PennCNV (Figure 12b). A 130kb deletion at 17q21.32 detected by PennCNV in 16/64 and both

programs in 7/64 individuals was not confirmed by FISH. This may be due to lack of specificity of the probe location or a false positive call by PennCNV and CNV Partition. Novel copy number gains detected by PennCNV and CNV Partition in single subjects on 1q32.1, 2p11.2, 12p13.33 and 7q31.32 were verified by FISH.

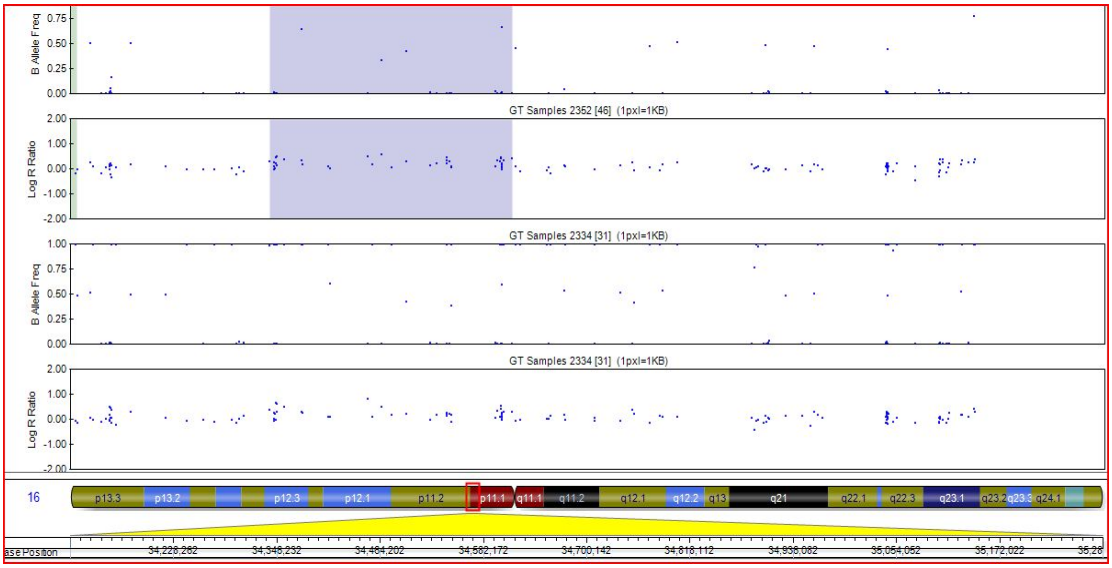



**Figure 12a**. A 277kb gain at 16p11.1. The duplication was called by both programs in AN 2352 (blue shaded box) and not by CNV Partition in AN 2334.

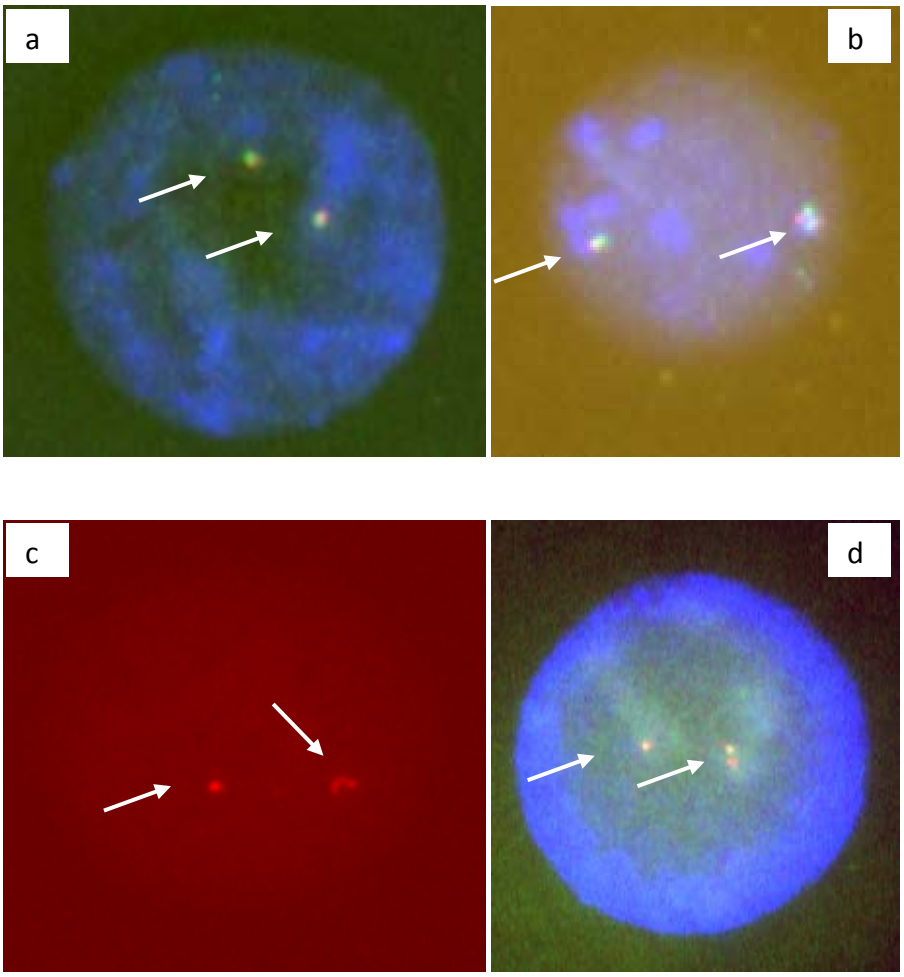



**Figure 12b**. FISH confirmation of the CNV gain at 16p11.1 using red and green fluorescent labelled RCPI human BAC library 11 (RP11-488I20/RP11-80F22) clones; a) The CNV negative control (wild type control) shows two fusions and b) a single fusion and double fusion is observed for AN 2352 confirming the CNV gain recorded by both algorithms. FISH investigation of AN 2334 shows a single and double spectrum red signal (c) and a single fusion and a double fusion signal (d) consistent with a CNV gain as detected by PennCNV but not by CNV Partition.

## 8.5 Discussion

The reporting of CNV in the diagnostic population relies on the efficiency of CNV detection software programs. This study investigates the differences of CNV outputs from two CNV detection software programs using the same raw data. It is believed that this is the first comparative study using raw data from a cohort of a Western European descendent population, thereby more closely simulating the routine diagnostic setting than previously reported comparative studies. This study differs to other comparative studies as it provides a sample level comparison of CNV outputs and is the first to provide stratification of discordant CNV. Differences in CNV length and breakpoint estimation occurring up to 500kb in size are reported. Likewise, CNV calls made by one of the algorithms representing false positive or negative calls were apparent to 350kb in size. These findings overlap the reportable CNV length by diagnostic laboratories, highlighting the need for further evaluation.

A limited number of studies have previously evaluated microarray platforms and CNV detection software programs for reproducibility, breakpoint accuracy and efficiency of CNV detection (2-5, 7, 11, 12) (Table 6). The common element between these studies is the level of inconsistency of CNV calls between platforms and CNV software detection tools. A review of these studies revealed that not one program recorded total CNV concordance. One approach used by these studies is to compare the CNV outputs from the raw data generated from one or more HapMap individuals (2, 5, 7). In a comprehensive study of microarray platforms and CNV detection software programs Pinto et al. 2011 reported < 70%

reproducibility of CNV calls in replicate experiments and < 50% concordance where the same raw data were analysed by CNV detection programs including CNV Partition, PennCNV, iPattern and QuantSNP (7). Winchester et al. 2009 reviewed 7 programs including PennCNV and CNV Partition against published HapMap data and demonstrated <50% overlap of CNV events (2, 5). In the study by Zhang et al. 2011 the authors found a poor correlation of CNV detection against the sequencing studies of Kidd et al. 2005 and CGH data of Conrad et al. 2010 (5, 13, 14). They report that small and common CNV have low recovery rates and deletions have a higher recovery rate than duplications.

**Table 6.** Review of comparison studies of CNV detection software programs and outcomes reported.

| Literature | Platform | Algorithm | Cohort | Findings |
|---|---|---|---|---|
| Current | Illumina Omni1-Quad | CNV Partition Penn CNV | 64 females "Aussie Normals collection" | 60.9% > 1kb 88% >100kb |
| Pinto et al. 2011 | CGH, SNP Platforms: Affymetrix and Illumina 660W, 1M, | CNV Partition Penn CNV, Partek, QuantiSNP , others | 6 HapMap control samples | < 70% reproducibility < 50% concordance |
| Zhang et al. 2011 | Affymetrix SNP | PennCNV, Birdsuite, Partek,HelixTree | HapMap (published from Kidd et al. ) and Bipolar study | < 11% concordance >100kb (cohort = 8 HapMap samples) |
| Winchester et al. 2009 | Affymetrix and Illumina SNP arrays | 12 programs including PennCNV, CNV Partition | NA12156 | Concordance of 20% of calls detected by all programs, 27% of calls detected by more than one program |
| Wineinger et al. 2012 | Computational analysis | HMM models such as Penn CNV, Birdseye | Not stated | 4-5 probe high false neg rate , 7 probes 90% recovery, 10 probes nearly all detected |

## 8.5.1 Discordance of CNV outputs between the algorithms

In the current study the performance of CNV Partition and PennCNV were compared for the detection of CNV in a Western European healthy population cohort. Discordance of CNV calls between the CNV outputs was detected. The causes of discordance, characteristic features of discordant calls and potential significance in a diagnostic setting were investigated. To do this the discordant CNV are stratified for length and copy number state and investigated for parameters that are applied in routine laboratory investigations including QC metrics and correlation with genomic architecture.

This investigation revealed that discordant CNV calls can be divided into three groups based on the type of discordance. Group 1 are CNV called in only one algorithm, group 2 is CNV fragmentation and CNV with discordance in breakpoint estimation (CNV length) are represented in group 3 (Table 7).

**Table 7.** The CNV outputs for the same raw data for CNV Partition and PennCNV were compared. The total represents the incidence of CNV called by each program. The total incidence of discordant calls >100kb stratified into groups 1-3.

| | | | Group 1 | Group 2 | Group 3 |
|---|---|---|---|---|---|
| | >1kb Total | >100kb Total | Single algorithm | Call fragmentation | CNV Length |
| CNV Partition | 13,422 | 463 | 7 | 0 | 4 |
| Penn CNV | 17,292 | 415 | 50 | 50 | 88 |

CNV calls made by a single algorithm represent either a false positive by the CNV program that made the call or false negative by the program that did not make the call. Program specific CNV are shown here to occur up to 350kb in size, most of which are less than 100kb in size however still pertained to >100kb for 2% (CNV Partition) and 12% (PennCNV). This discordance overlaps the enrichment of CNV >250kb as reported in a study of children with neurological disorders (15). Further investigation of the program specific CNV showed that they occurred more frequently in common, high frequency CNV and are enriched in regions of segmental duplication.

Call fragmentation is recorded where a single event in CNV Partition is detected as multiple calls by PennCNV. In this study call fragmentation in CNV >100kb was observed in 12% of calls. This discordance was assessed in the current study for CNV up to 1Mb and no evidence of call fragmentation was observed in CNV >500kb. This anomaly may result in an underestimation of the length of the CNV and explains the overall higher number of calls made by PennCNV.

In the study reported here discordance of breakpoint estimates resulting in a difference in predicted CNV length was recorded in 20% of CNV > 100kb and in CNV up to 500kb. Breakpoint estimates as a form of discordance between CNV detection software programs has rarely been investigated in published comparative studies. Winchester et al. 2009 observed a “difference in the size of predicted events between algorithms” but failed to investigate this further (2). The findings reported here are consistent with the study reported by Pinto et al. 2011. The authors considered the reproducibility of breakpoint estimates within a program by performing triplicate experiments and they reported within program

variation. To assess the precision of breakpoint estimation between programs they used datasets from sequencing projects and compared the distance of the known breakpoint with the putative start and end positions. In doing this they identified variation between the programs (7).

## 8.5.2 The role of the algorithms

To understand how CNV outputs are generated from the CNV programs the algorithms in each of the programs were investigated. CNV call outputs from software programs are based on background bio-information algorithms.

### 8.5.2.1 CNV Partition

CNV Partition is a modification of the circular binary segmentation method described by Ohlsen and Venkatraman (2004) and uses two outputs, LogR ratio (LRR) and B allele frequency (BAF) to predict changes in copy number state (Technical note, Illumina Inc. 2010). The LRR is the logged ratio of normalised signal intensity of observed to expected ($Log2(R_o/R_e)$) (2) and zero is expected for a normal sample. Evidence of copy number change is determined by deviations from zero. The BAF is a representation of the B allele frequency carried by the sample. The expected BAF in a normal sample is 0.0, 0.5 and 1.0 for each locus and this represents AA, AB and BB genotypes respectively. Discordance from these values is suggestive of a copy number change (2).

Copy number estimation (X) is made by comparing the observed LRR and BAF values for each locus to the expected values for 14 genotypes using bivariate Gaussian distribution (Technical note, Illumina Inc. 2010) (2, 16, 17). The likelihood of observing an estimated LRR and BAF is calculated for each of the genotypes and summarised. The copy number estimate (X) is the average of the five modelled copy numbers (L0, L1, L2, L3 and L4) divided by their individual likelihoods. This value provides the input for estimation of breakpoints (Technical note, Illumina Inc. 2010)

Putative breakpoints are identified across the whole genome (16, 17). The values determined for X are ordered along the whole chromosome according to position. Regions within the genome that achieve a value of X that is consistently higher or lower than the expected value of X for a diploid sample are identified. The values of X at loci are compared to identify regions where the copy number estimate between these loci is maximally different from adjacent regions. A sliding window algorithm is applied to identify regions where there is maximum variation. In circular binary segmentation models the segment is not defined and the analysis of contiguous segments is continuous (16). The modification applied here provides a more efficient identification of change point by setting a defined segment size (window). The calculation is repeated for defined window sizes 4, 8, 16, 32 etc expanding in both directions, until the optimal size window is identified that represents the region of maximum difference (Technical note, Illumina Inc. 2010).

Copy number is then assigned to the partitioned regions between putative breakpoints across the genome. The likelihood of copy number for each copy number state 0, 1, 2, 3, 4, is estimated and summed for the region. The copy

number with the highest sum is then assigned to the region (Technical note, Illumina Inc.) (Figure 13).

### 8.5.2.2 PennCNV

PennCNV uses the Hidden Markov Model (HMM) for CNV detection. This incorporates the LRR, BAF , population frequency of the B allele and distance between consecutive SNPs to estimate copy number change (1). This algorithm estimates the probability of moving from one copy number state to another between SNPs and assumes that there are dependence structures between nearby SNPs (1, 2). To predict a CNV the hidden Markov model algorithm assumes that the hidden copy number state at a SNP is related to the copy number state of the adjacent SNP. A transition matrix is used to determine the probability of moving from one state to another across adjacent SNPs. Hidden Markov models are reported to be associated with a high false positive rate (11), however in this study we show a higher call rate for PennCNV but confirmation of selected CNV by alternative methods suggest that this does not correlate with false positive CNV in PennCNV.

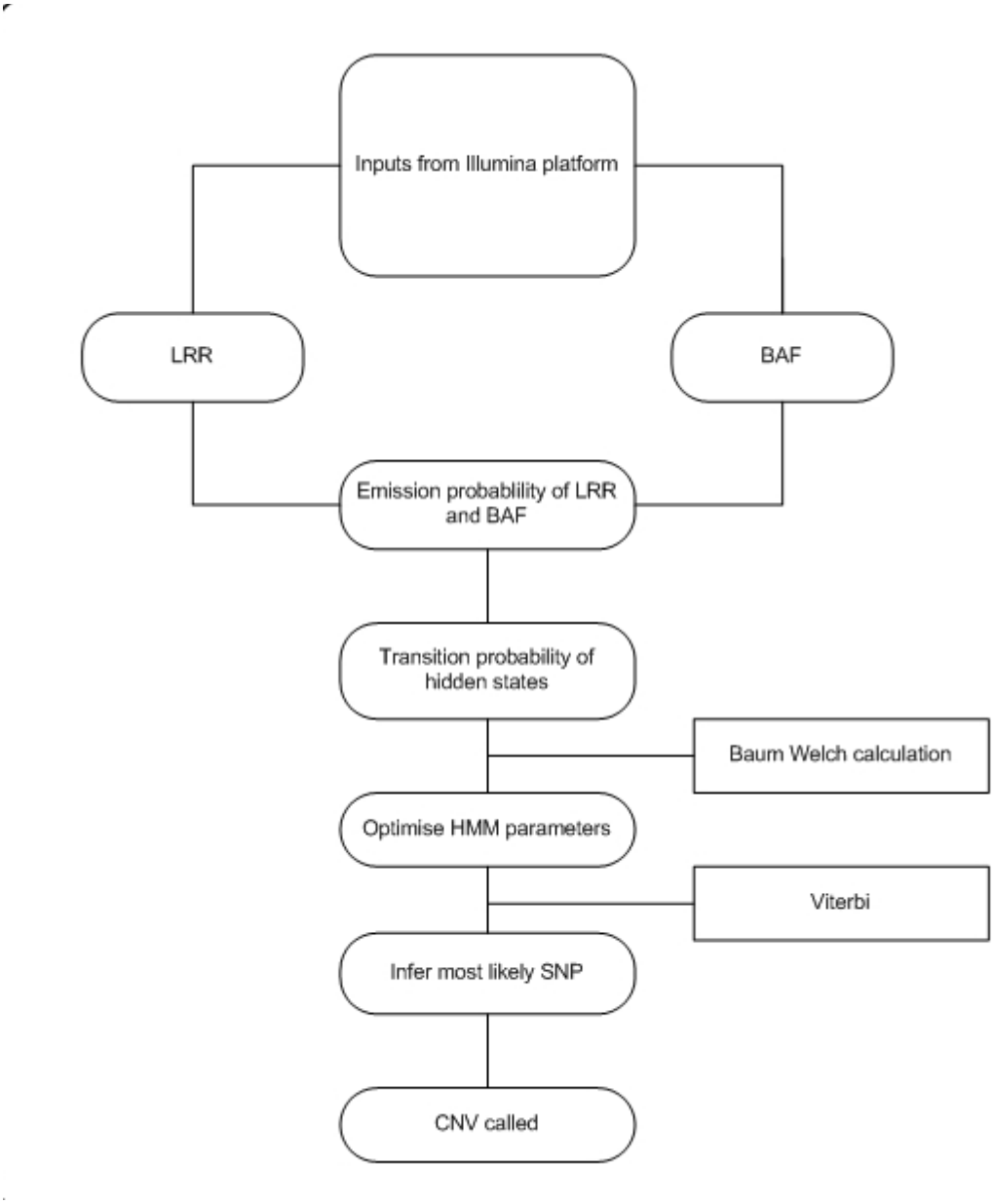



**Figure 13a**. Workflow for CNV output for PennCNV (1)

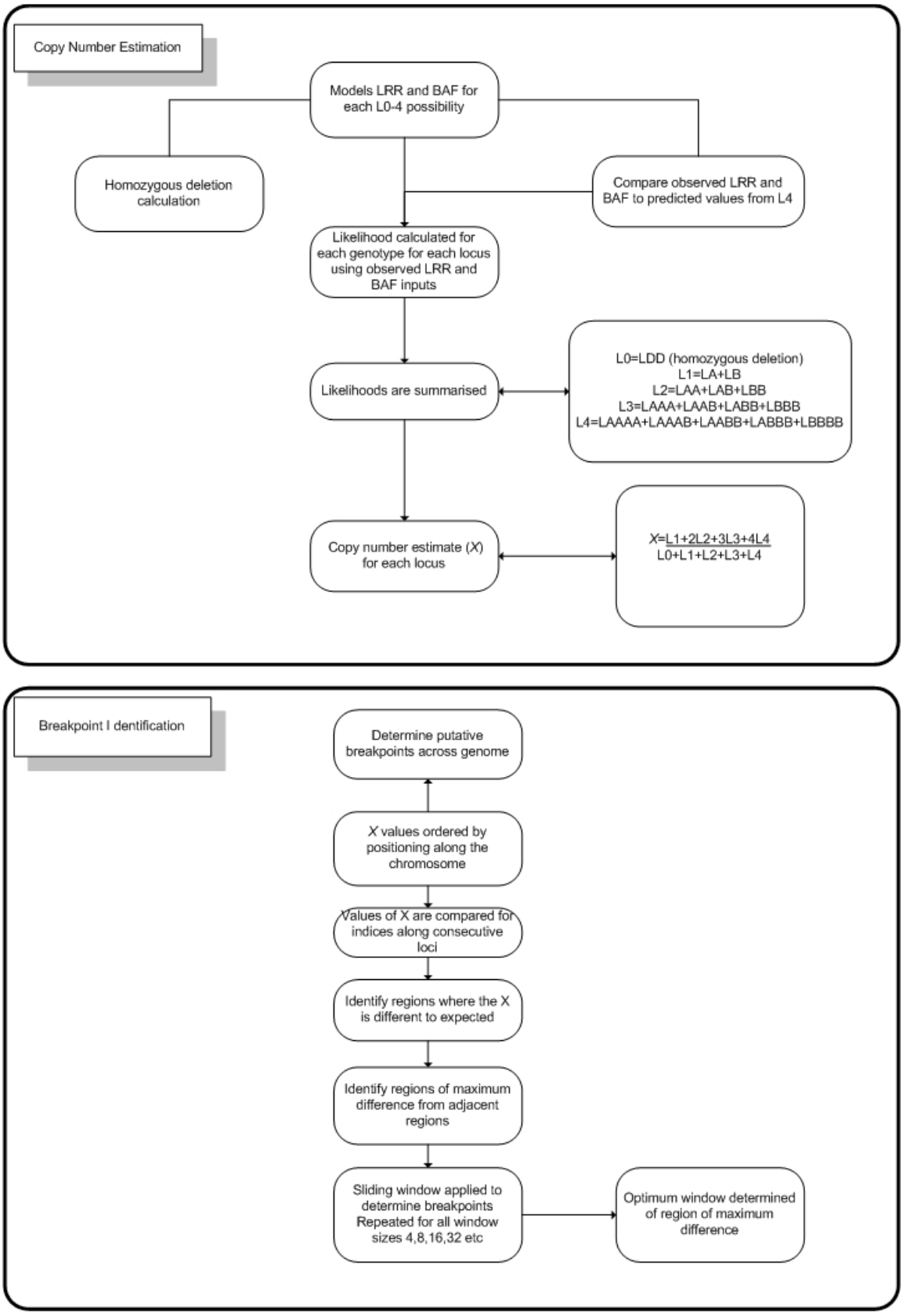



**Figure 13b.** Stage 1 and 2 of CNV Partition CNV detection algorithm

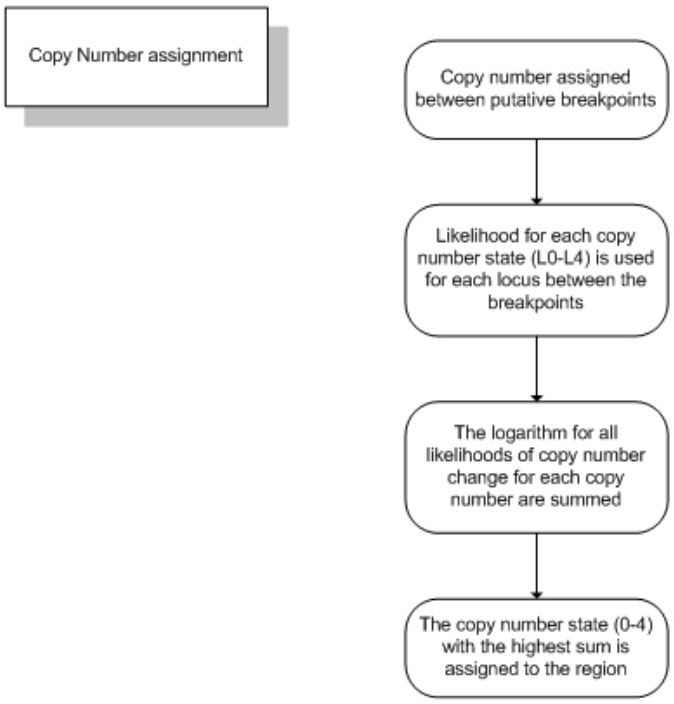



**Figure 13b**. Work flow chart for the algorithm of CNV detection in the CNV Partition software program (Illumina, Inc). There are three stages to CNV detection using the two outputs LogR ratio (LRR) and B allele frequency (BAF). The copy number estimate for loci is followed by breakpoint estimation and finally copy number assignment of determined regions of copy number change (adapted from Technical Note; Illumina, Inc.).

### 8.5.3 Potential causes for discordance between CNV detection programs

The occurrence of CNV call discordance may be explained in part by the probe design for SNP based platforms which is influenced by regions of genome complexity. A review of the SNP markers in CNVR where there are discordant CNV showed a higher incidence of "no call" genotypes compared to singletons and low frequency variants that were called in both algorithms. The "GenCall" category is an indication of the confidence of the SNP call and can be used at the sample level or to identify poor quality CNV calls (12). In this study a "no call" was observed for the genotypes of >60% of SNP markers in high incidence CNVR and CNVR with a high rate of sample level discordance including false calls, breakpoint estimation and fragmentation of a CNV call. This may be representative of poorly performing SNP markers however the calls due to low frequency alleles cannot be excluded.

Likewise genome complexity may account for the higher rate of program specific CNV in segmental duplication regions. The potential for cross hybridization in regions of segmental duplication results in platform designs with low probe density in these regions (7). This leads to reduced detection rates (5, 7) and reduced robustness of probe markers leading to the potential for false CNV calls. The importance of CNV detection in these regions is indicated, as segmental duplication regions are reported to be hotspots for CNV formation (7, 15, 18, 19) with pathogenic and common CNV enriched in these regions (15).

Analysis of the putative breakpoints for selected CNV compared to accurate breakpoints determined by sequencing allows the evaluation of the causes for differences in breakpoint estimation. Inspection of the LogR values of markers within the CNV showed fluctuation towards zero that correlated with the estimated breakpoints reported by PennCNV. Further investigation of the GC content between markers within the CNV suggested a correlation of the variation of the LogR with a change in the GC content and the different breakpoint estimation by the programs. However this finding represents a limited number of CNV investigated in this cohort.

GC content has been reported previously in association with variation of LogR values. Marioni et al. 2007 demonstrates a strong correlation of peaks and troughs of LogR values with the GC content of the clones, and a subsequent study by Diskin et al.2008 (20), investigated genomic and technical factors that may influence the quality of signal intensity data generated by SNP array platforms (20, 21). They demonstrated that GC content is a significant factor in the appearance of fluctuation in LogR values (20).

CNV discordance may also be explained by the different bio-informatic algorithms employed to detect copy number change. The mechanism of CNV calling differed between the two programs. CNV Partition uses a “sliding window” principle to identify breakpoints, whereas the prediction of breakpoints between SNP markers is probability based in hidden Markov model algorithms. As such regions of poor SNP marker performance or low density of SNP markers could impact breakpoint estimation. This may be reflected in call fragmentation and variation of breakpoint

estimates demonstrated as discordance of CNV length between CNV software outputs.

### 8.5.4 How to recognise a false call

I sought to determine if the QC parameters applied in CNV analysis are indicators of the presence of a false call. The LogR and marker density were assessed for the current study. There appeared to be no significant difference in either parameters compared to expected values in confirmed CNV and CNV called by both algorithms. Given the confirmation of CNV by alternative methods for selected CNV and indication that these are false negative by one program rather than a false positive call, it is suggested that LogR and marker density can be used as indicators when assessing the sensitivity of a CNV call. However only a limited number of program specific CNV were confirmed by alternative methods.

There is sufficient evidence to indicate that careful consideration of LogR and marker density should be applied where diagnostic laboratories report CNV below length thresholds. Variation in breakpoint estimates correlates in the two CNV sequenced with fluctuations in LogR which corresponded with altered GC content. Although this finding is limited to the small CNV analysed it suggests that careful interpretation of dot plots flanking the estimated breakpoint is required. LogR remains a useful indicator of breakpoint estimation.

A similar finding was report by Weininger et al. 2012. In an effort to evaluate the confidence of CNV in a GWAS study the authors reviewed two commonly occurring errors of CNV calling in Hidden Markov model CNV detection algorithms. They

demonstrated that the log relative ratio (LRR) is a reliable indicator of copy number change and breakpoint estimation and can be applied for small CNV and regions of variable copy number state.

### 8.5.5 Limitations of the study

The Illumina Omni1-Quad platform is a higher resolution platform with approximately 1 million markers and median interspacing of 1.5kb, than platforms currently used by diagnostic laboratories and as such may not be representative of calls identified by lower resolution platforms that are currently utilized by diagnostic laboratories (7, 9). However with decreasing cost and the need for more information increasingly higher resolution platforms are being introduced.

A further limitation of this study was encountered when assessing the level of true concordance. This was complicated by CNV fragmentation and variation in CNV length. For example a CNV measuring larger than 100kb in one program may correspond to CNV as small as 1kb in the other program. For this purpose strict criteria was applied to define concordance so that a baseline of CNV calls could be achieved.

The default parameters for each program were applied in this study. Optimisation was not performed due to the small sample cohort and the aim to replicate conditions in routine diagnostic laboratories. Evidence suggests that optimising parameters of the CNV detection software with a reference set may reduce noise and improve detection rates and breakpoint accuracy (3, 8) .

## 8.5.6 Recommendations

From the results of this study it is recommended that commercial developers of CNV detection software programs offer more than one algorithm as a selection within the software tools. While the level of discordance identified in the few studies reported to date suggests that the use of two CNV detection programs and selection of only shared calls would increase the false negative rate, the application of an additional algorithm when required would increase the confidence of calls. It would allow a laboratory scientist to confirm a call that does not meet all QC requirements or one that is visible in the CNV plot but not called by a single algorithm. In addition alternative methods of confirmation can be initiated where there is insufficient evidence for confirmation of the call. To apply alternative methods of confirmation e.g FISH or an alternative platform can be costly and time consuming. The rate of CNV requiring alternative confirmation can be reduced with confirmation using an inbuilt second CNV detection algorithm.

The inconsistencies between CNV detection programs are offset by the continual improvement of both platform design and CNV detection programs, some CNV will not be detected by all software applications. Platform validation is well recognized by clinical laboratories but the validation of the software programs calling the CNV are frequently overlooked or inadequately performed. Laboratories should be aware of the limitations of the CNV detection software employed and method

validation should include confirmation of CNV and breakpoint estimation by alternatives such as FISH, an alternative CNV detection program or inter-laboratory comparison of the same raw data. Confirmation of 'low confidence' CNV calls is recommended for CNV < 500kb in clinically significant regions where there is complex architecture, poor SNP integrity and low marker density. In the absence of SNP data, as with CGH arrays, the need for confirmation of CNV is more pertinent.

## 8.6 Conclusion

The purpose of this study is to assess the variation that exists between detection programs and highlight the significance for diagnostic interpretation. Stratification of CNV in the current study identified discordance in breakpoint estimation, the presence of call fragmentation and CNV detected by a single algorithm, the last representing either false positive or false negative CNV. These anomalies are within the size range of CNV analysed and interpreted in diagnostic laboratories and overlap the CNV length reported in association with pathogenic significance (15).

The findings here suggest a correlation of discordant CNV calls with genomic architecture. The performance of algorithms appears to be compromised in complex high repeat regions and with fluctuation in GC content. The ability to estimate copy number state or breakpoints appears to be relative to the bio-informatic calculations applied, for example hidden markov models compared to

circular binary methods. There is some reassurance for diagnostic laboratories that the discordance due to SNP marker performance did not occur in low frequency CNV.

The comparison performed in this study and stratification of the differences in CNV outputs between algorithms provides insight to the limitations of CNV detection and highlights the need for careful assessment of CNV calls and verification of CNV detection software in clinical laboratories.

## 8.7 References

1. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. 2007;17(11):1665-74.
2. Winchester L, Yau C, Ragoussis J. Comparing CNV detection methods for SNP arrays. Brief Funct Genomic Proteomic. 2009;8(5):353-66.
3. Dellinger AE, Saw SM, Goh LK, Seielstad M, Young TL, Li YJ. Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. Nucleic Acids Res. 2010;38(9):e105.
4. Wineinger NE, Tiwari HK. The impact of errors in copy number variation detection algorithms on association results. PLoS One.2012; 7(4):e32396.
5. Zhang D, Qian Y, Akula N, Alliey-Rodriguez N, Tang J, Gershon ES, et al. Accuracy of CNV Detection from GWAS Data. PLoS One.2011; 6(1):e14511.
6. Bruno DL, Ganesamoorthy D, Schoumans J, Bankier A, Coman D, Delatycki M, et al. Detection of cryptic pathogenic copy number variations and constitutional loss of heterozygosity using high resolution SNP microarray analysis in 117 patients referred for cytogenetic analysis and impact on clinical practice. J Med Genet. 2009;46(2):123-31.
7. Pinto D, Darvishi K, Shi X, Rajan D, Rigler D, Fitzgerald T, et al. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nat Biotechnol. 2011;29(6):512-20.
8. Baross A, Delaney AD, Li HI, Nayar T, Flibotte S, Qian H, et al. Assessment of algorithms for high throughput detection of genomic copy number variation in oligonucleotide microarray data. BMC Bioinformatics. 2007;8:368.
9. Chia NL, Bryce M, Hickman PE, Potter JM, Glasgow N, Koerbin G, et al. High-resolution SNP microarray investigation of copy number variations on chromosome 18 in a control cohort. Cytogenet Genome Res. 2013; 141:16-25.
10. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002;12(6):996-1006.
11. Karimpour-Fard A, Dumas L, Phang T, Sikela JM, Hunter LE. A survey of analysis software for array-comparative genomic hybridisation studies to detect copy number variation. Hum Genomics. 2010;4(6):421-7.
12. Ritchie ME, Liu R, Carvalho BS, Irizarry RA. Comparing genotyping algorithms for Illumina's Infinium whole-genome SNP BeadChips. BMC Bioinformatics.2011; 12:68.
13. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, et al. Mapping and sequencing of structural variation from eight human genomes. Nature. 2008;453(7191):56-64.
14. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, et al. Origins and functional impact of copy number variation in the human genome. Nature. 2010 ;464(7289):704-12.
15. Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, et al. A copy number variation morbidity map of developmental delay. Nat Genet. 2011;43(9):838-46.

16. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics. 2004;5(4):557-72.
17. Venkatraman ES, Olshen AB. A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics. 2007; 23(6):657-63.
18. Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, et al. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet. 2009;41(10):1061-7.
19. Cooper GM, Nickerson DA, Eichler EE. Mutational and selective effects on copy-number variants in the human genome. Nat Genet. 2007 ;39(7 Suppl):S22-9.
20. Diskin SJ, Li M, Hou C, Yang S, Glessner J, Hakonarson H, et al. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. Nucleic Acids Res. 2008;36(19):e126.
21. Marioni JC, Thorne NP, Valsesia A, Fitzgerald T, Redon R, Fiegler H, et al. Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. Genome Biol. 2007;8(10):R228.

**Technical Note: Illumina Systems and Software**

DNA copy number and loss of heterozygosity analysis algorithms. Illumina, Inc. 2010

# CHAPTER 9

# Conclusion

This thesis reports on the findings of a comprehensive genome-wide investigation of genetic variants, CNV and LCSH, in a cohort of females recruited in the "Aussie Normals" Collection. This study differs from other studies of control cohorts and general populations in participants investigated, scope of investigation and diagnostic application. It began in early 2010 when microarray technology was still developing and studies of "normal/healthy" population cohorts was largely derived from The International HapMap Consortium. Coinciding with this was the rapid progression of this technology from research and development to implementation for clinical diagnosis which highlighted the need for the understanding of genetic variants in the "normal" population.

Stratification of CNV in the study presented here identified a predominance of CNV measuring less than 50kb in length. Deletion accounted for the majority of CNV < 50kb but the incidence of duplication and deletion nearly equalised in CNV >100kb. In the "Aussie Normals" cohort CNV > 100 kb was detected in all individuals and CNV measuring 500kb-1.4Mb are private variants in 8 individuals (12.5%). Inter-chromosomal variation was observed for CNV load and copy number state. The incidence of CNV is consistent with chromosomal length for the total number of CNV detected however inter-chromosomal variation is apparent when CNV > 100kb is considered.

CNV is ubiquitous in the genome but intra-chromosomal variation of CNV distribution is apparent. CNV are non-randomly distributed and in contrast with previous reports are enriched at morphological structures such as subtelomeric and pericentromeric regions in only a few chromosomes. Review of breakpoints and flanking regions revealed that all CNV > 100kb are associated with either

segmental duplication or repetitive elements. Notably, common, high incidence CNV are associated with regions of segmental duplication and low frequency and private variants correlated with repetitive elements. Confirmation of CNV at the base pair level provided the opportunity to investigate the characteristics of the breakpoint and intervening sequence. The sequence signature identified DNA sequence predisposing the region to genomic instability potentially leading to CNV formation.

Review of the chromosomal landscape of CNV on chromosome 18 revealed a bias away from deleterious genes and a pattern of fewer genes and larger segments of gene deserts are associated with the high incidence common CNV. Extension of this investigation to chromosomes 1-22 revealed that CNV harboured genes with roles in immunity, signal transduction and a large proportion of genes with undefined function. Gene function differed for deletion and duplications. Genes with roles in signal transduction were observed in deletions whereas genes involved in immunity are over represented in duplications.

Together these findings suggest that the chromosomal CNV landscape, inter and intra-chromosomal distribution and contribution of duplication and deletion are attributed to the structural, functional and molecular aspects of DNA. Demonstrated here and consistent with HapMap based population studies, enrichment of CNV < 50kb and paucity of CNV > 500kb is characteristic of genetic variation in the general population. The enrichment of common CNV in regions harbouring immunity, signal transduction or gene deserts and notable absence of CNV involving deleterious genes may contribute to patterns of CNV distribution. The molecular mechanisms that contribute to the formation of CNV may also

provide an explanation. This is illustrated in the investigation of chromosome 18. Chromosome 18 is shown in this study to be under-represented for CNV >100kb and in particular for duplications. In accordance with this, chromosome 18 has the lowest genomic contribution of segmental duplication compared to other autosomes.

The prevalence of CNV <50kb and enrichment of deletion may be attributed to exogenous and endogenous causes of genomic instability predisposing to DNA breakage and the repair process resulting in deletion. This is demonstrated for a deletion on chromosome 18. Characterisation of the intervening sequence revealed the presence of sequence signatures that have the potential to form structures that predispose the DNA to genomic instability. Numerous mechanisms of CNV formation can result in CNV formation and reflect characteristics of the surrounding sequence. The finding of microhomology at CNV breakpoints illustrates a correlation of CNV location and the potential role of repetitive elements in CNV chromosome landscape and CNV properties.

General population studies have a role in the documentation of rare and novel CNV associated with a benign outcome. The finding of 31 novel CNV and rare CNV that overlap categories of unknown or uncertain clinical significance highlights the utility of general population studies and the dynamic nature of CNV formation. Accurate cataloguing of these variants in publicly available databases provides evidence for interpretation in clinical diagnostics. Likewise the documentation of CNV in healthy general populations will contribute to the investigation of candidate genes by refining regions of interest in CNV associated with pathogenic significance.

The employment of high resolution SNP microarray technology in this investigation provided the opportunity to investigate the properties and distribution of long contiguous stretches on homozygosity in a cohort of Australian women. The findings demonstrate the background level and length of LCSH. Recommendations of thresholds characterised for this local outbred population are provided and can be applied in clinical diagnostics.

The clinical application of general population studies is further exemplified in the comparison with clinical cohorts. The investigation of CNV provided evidence for an increased level of duplications, and in particular CNV < 250kb in selected clinical cohorts. The investigation of specific CNV identified a candidate region present as a heterozygous deletion in the general population but observed at increased incidence as a homozygous deletion in the CNV of fetal demise. Comparison of both CNV and LCSH in common complex disease provided evidence of regions for further investigation. Previously published association of common CNV with the disease trait, hypertension, was not supported here and CNV involving high risk alleles was not observed as a mechanism in this cohort. Interestingly this study provided examples of regions of CNV and LCSH that encompassed the same genomic region in different individuals. This provides evidence that investigation of disease association for candidate genes should consider all types of genetic variants collectively rather than in isolation. In addition the “smaller” CNV may refine the candidate genes within the “larger” region of LCSH. It may also provide an explanation for a degree of phenotypic heterogeneity.

This study considers the technical challenges and limitations of the CNV detection algorithms applied to data generated from microarray technology. The lack of concordance of CNV detection among early studies created some confusion with regard to the extent, properties and chromosomal distribution of CNV. A comprehensive investigation in the study here of the discordance of CNV detection between two algorithms identified an overlap of genomic region and CNV length with that reported in clinical diagnostics. This is further applied in the investigation of LCSH with the potential for over-calling of LCSH events in regions of poor SNP marker integrity. These findings support the recommendation of appropriate guidelines and QC metrics for CNV and LCSH detection in clinical diagnostics. It should be noted however that recent improvements in platform and algorithm design have made progress in mitigating these issues.

The limitations of this study should be considered. The cohort, though representing a random recruitment of females in the Australian population and clinically defined, fulfilling the category of "healthy/normal", is small. Any correlations should be confirmed in larger cohort studies. The study began in 2010 prior to improvements in platform and algorithm design and may have contributed to the level of discordance between the algorithms potentially elevating the estimation of false calls. In addition the NCBI36/hg18 genome build was used for investigation. Genome build GRCh37/hg19 released shortly after and more recently GRCh38/hg38 has resulted in alteration of the location of genes and sequence characteristics. This provided some challenges in interpretation which are noted where applicable, and may impact on the evaluation of candidate genes or significance of breakpoints within some of the CNV and LCSH reported here.

## 9.1 Recommendations and future directions

Since beginning this study, Next Generation Sequencing (NGS) has demonstrated clinical utility in the investigation of single gene mutations. The 1000 Genomes Project employs sequencing technology with a goal to define single nucleotide polymorphism, short insertions and deletions and in particular, document low frequency variants in over 1000 individuals representing 14 populations.

The investigation presented in this thesis illustrates the ubiquity of genetic variation at a level of resolution afforded by molecular karyotyping in a sample of a general population. However the extent of genetic variation at the level of resolution between molecular karyotyping and NGS technologies is yet to be fully appreciated. A predominance of CNV measuring less than 50kb was detected in the 'Aussie Normals" study. This is at the lower limit of resolution for most microarray platforms and this category of variants remains undefined. Characterisation of the true contribution of CNV at this level is within the scope of NGS or other technologies yet to be developed. The sequencing of large numbers of individuals in studies such as the 1000 Genomes project and general population studies is required to reveal the full extent of genomic variation and contribution to genetic and phenotypic diversity.

A recommendation from this thesis is for future disease association studies to incorporate the interactions of all genomic variants collectively rather than investigating the contribution of variant types in isolation. This is exemplified in the investigation of genotype and phenotype correlations of common disease trait in this thesis, where both CNV and LCSH in different individuals encompassed the

same genomic region. Employment of a similar comprehensive investigative approach that includes all categories of variants such as but not limited to CNV, LCSH, single nucleotide polymorphisms and indel, will expand the representative study cohort and more accurately identify regions of the genome that harbour putative candidate genes.

Continual and progressive revision of the literature and curation of publicly available databases of variants is required. The continued updating of these resources, with all categories of variants detected using a range of technologies, will provide evidence of novel and rare variants associated with a benign outcome and refine candidate regions associated with pathogenic significance. It is noted that there is no equivalent resource for the registration of common regions of homozygosity. A recommendation from this thesis is the creation of a database of population specific and common LCSH that would assist in the targeted investigation for candidate genes by homozygosity mapping using technologies such as NGS and exome sequencing.

The challenges in determining the extent of CNV and LCSH in early studies was due in part to the difference in study design and poorly defined control cohorts. A standardised approach to study design and strategies for determining sensitivity and specificity is recommended for future investigations to ensure accurate cataloguing of variants. A further recommendation is collation of clinical evidence of "healthy/normal" for participants representing general population and control cohorts. As illustrated in the "Aussie Normals" study, this will permit informative comparison with pathogenic cohorts for the determination of similarities or differences of significance. A clinical evidence based approach ensures definition of

benign variants and refinement of genes in benign variants that overlap regions of clinical significance.

Microarray platforms detect copy number change which is influenced by platform resolution as determined by the number and genomic position of markers. CNV detection is equally influenced by the algorithm applied to predict copy number change. As illustrated in this thesis, different outcomes from the same raw data can be achieved with a variety of algorithms. This may also be a limitation of the bio-informatic approaches of NGS. It is recommended from the evidence reported here that a more stringent approach is applied to QC metrics, mandatory variant confirmation and study design to ensure a consistent and robust approach to the determination of human genetic variation using future technologies.

The investigation of CNV at the base pair level has illustrated the role of sequence analysis in the investigation of genetic variants. This is applicable to the confirmation of structural variants and extrapolation to other aspects such as the characterisation of the molecular mechanism of formation of these variants. Future research studies will utilize the current evidence as demonstrated in studies such as this, and with the wider application of high volume sequence analysis offered by NGS, will further expand our knowledge of the derivation of genetic variants. The availability of the sequence motif at CNV breakpoints will enable further investigation of the exogenous and endogenous factors that contribute to genomic instability. It is projected that future research with this evidence may lead to the mitigation of disease causing genomic instability.

It is anticipated that molecular karyotyping by microarray in the diagnostic and research setting will continue to complement high resolution technologies by

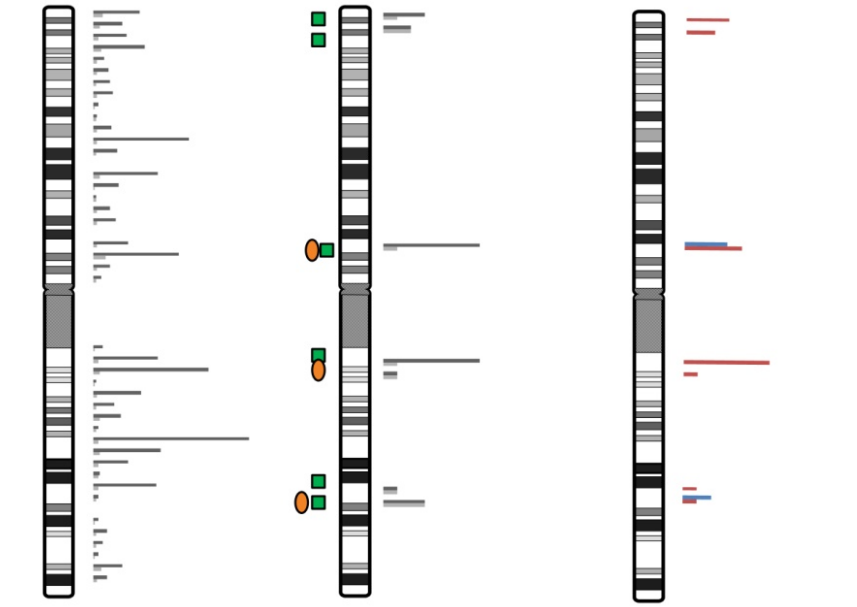
identifying regions of the genome that warrant further investigation by NGS and exome sequencing technologies. As NGS methodology expands to specialised diagnosis more clinically significant changes to DNA and RNA will be described and current knowledge of genetic variation refined. As such, the discoveries from general population studies such as the "Aussie Normals", genome-wide association, case studies, and data culminating from the 1000 Genomes Project will further expand the knowledge of phenotype/genotype correlations.

Documented here is a comprehensive investigation of the genetic variants, CNV and LCSH, of a small cohort of Australian women. The findings have contributed to the understanding of the magnitude, role and mechanisms of formation of these variants in the human genome. The CNV detected in this study on chromosome 18, including novel and rare variants, are published in the Database of Genomic Variants, contributing to the global aim to develop a map of human genetic variation.

# Appendix 1

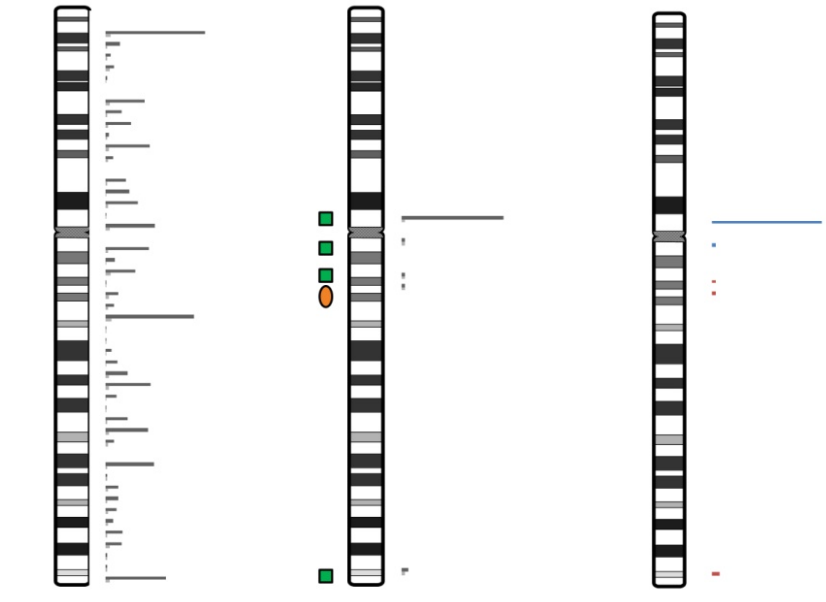
# Chromosome CNV Landscape

## Chromosome 1



Total incidence CNV >100kb Loss / gain >100kb

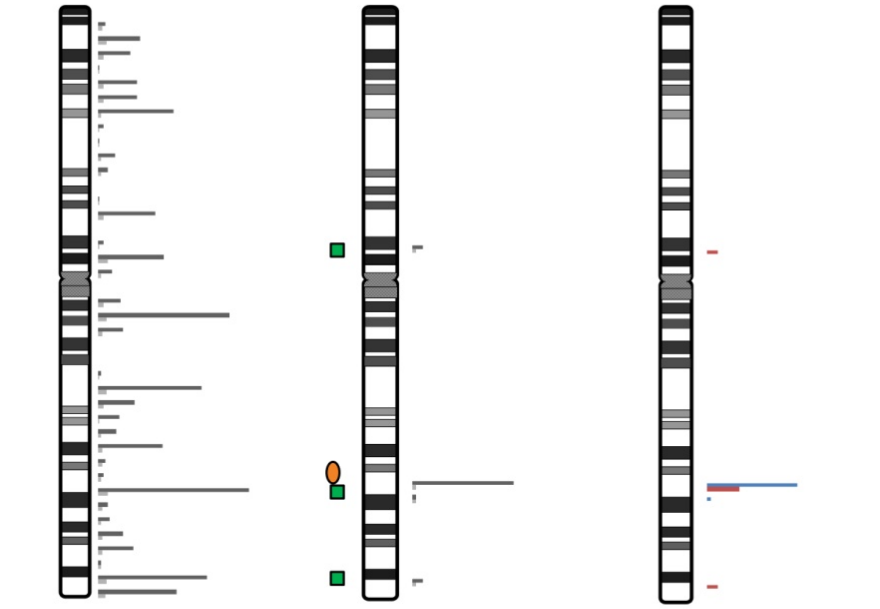
## Chromosome 2



Total incidence CNV >100kb Loss/ gain >100kb

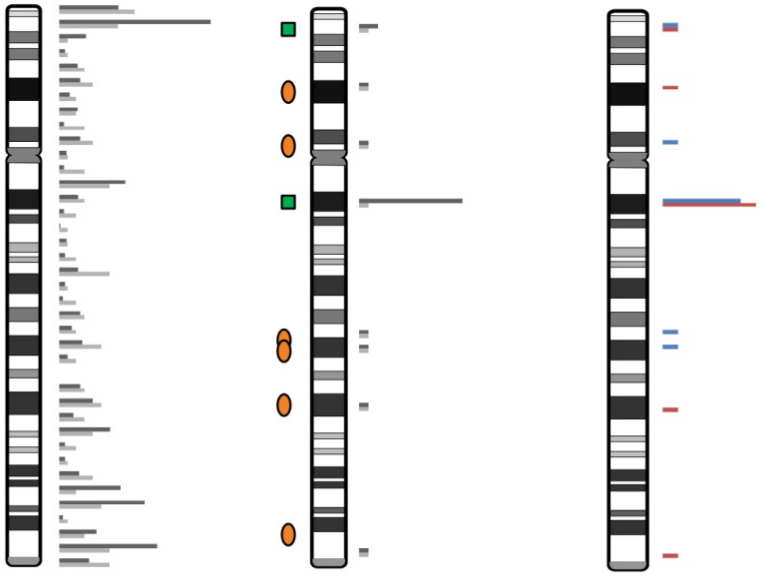
CNV
CNVR
Loss
Gain
Segmental Duplication
Repetitive Elements

**Chromosome 3**

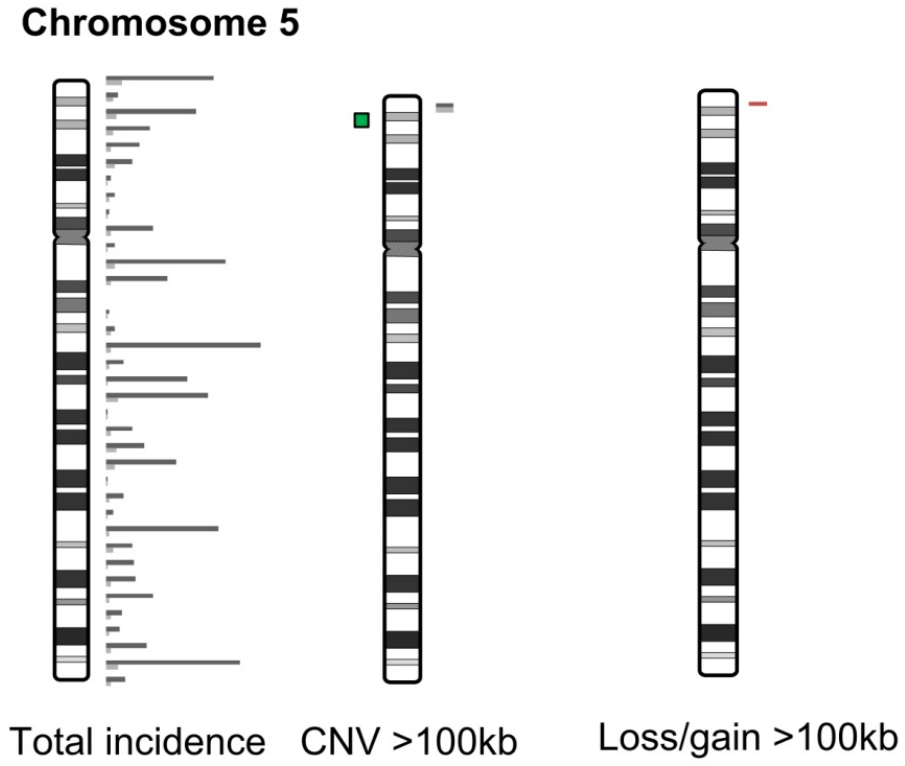


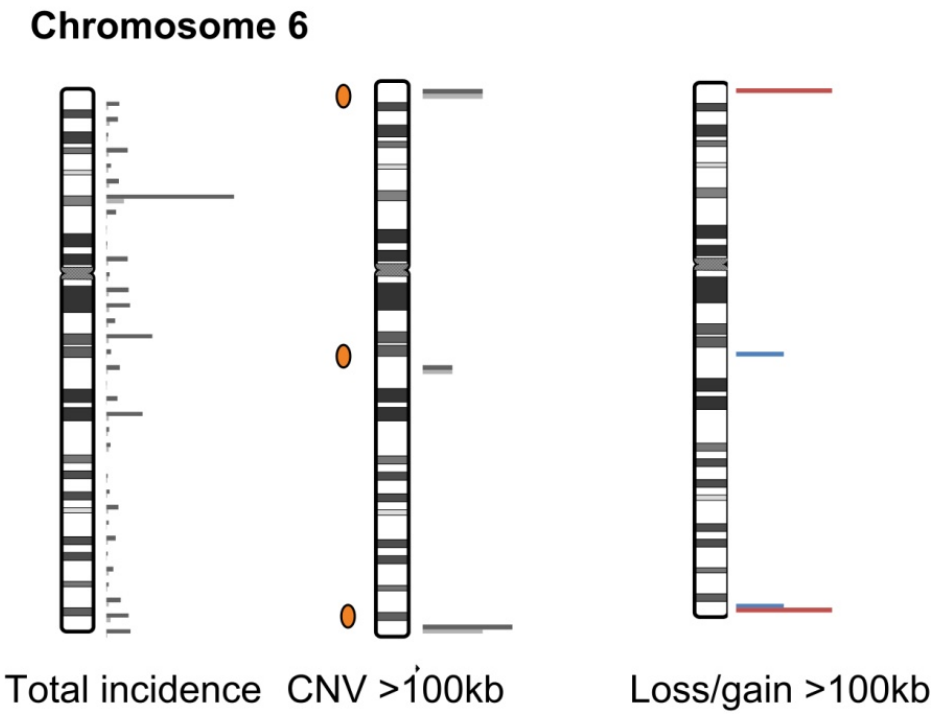
Total incidence CNV > 100kb Loss/gain >100kb

**Chromosome 4**

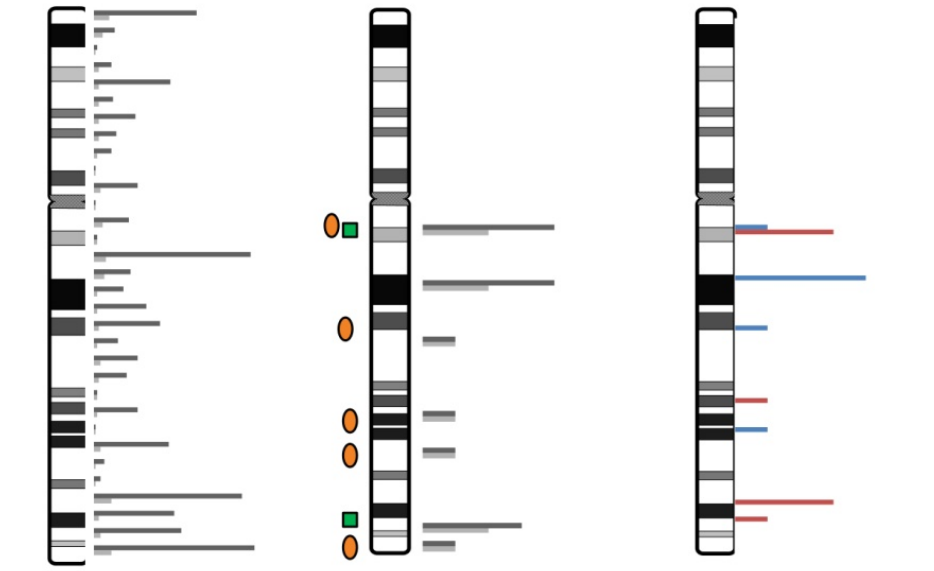


Total incidence CNV > 100kb Loss/gain >100kb

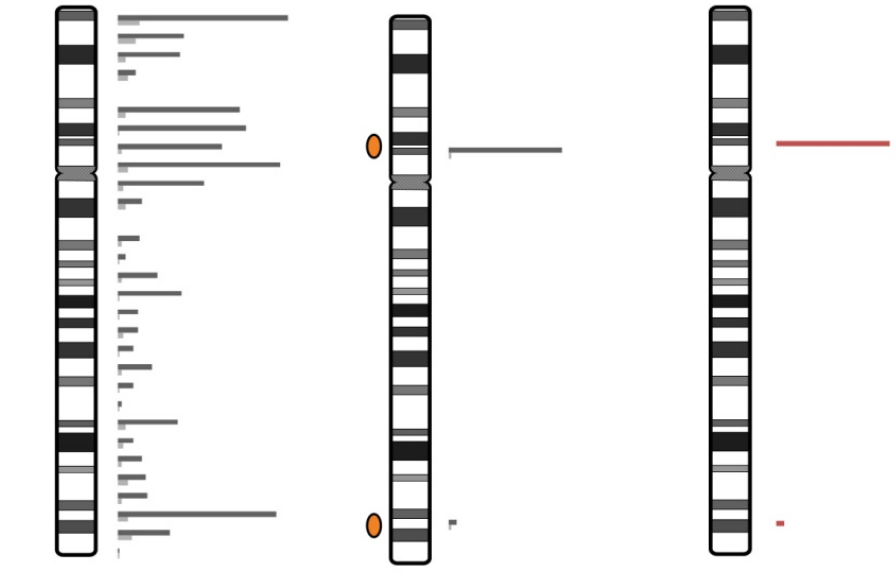

Chromosome 5
Total incidence
CNV >100kb
Loss/gain >100kb

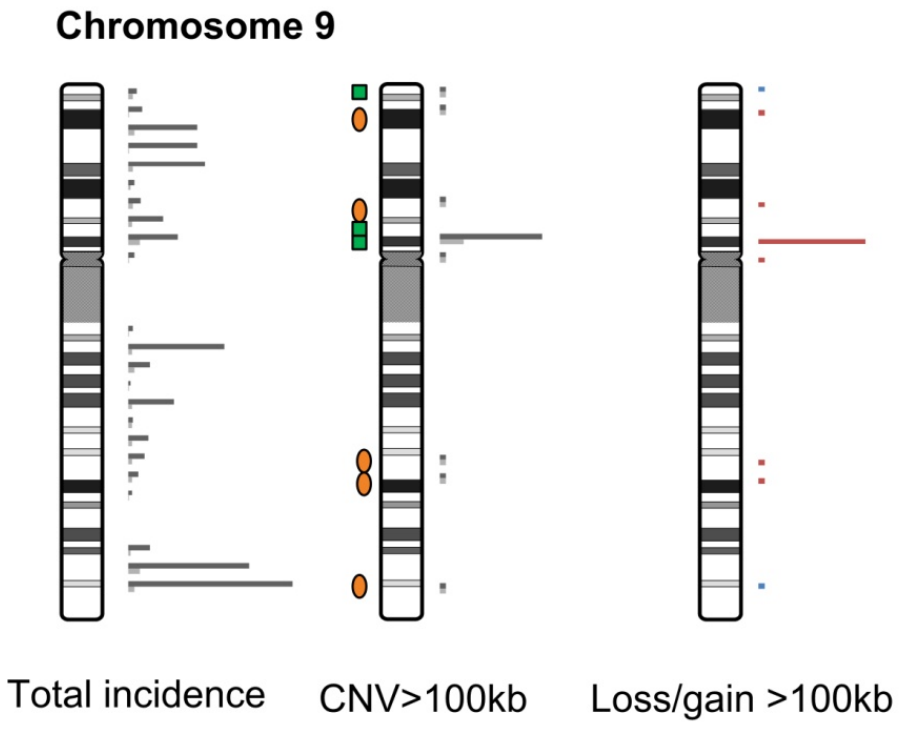

Chromosome 6
Total incidence
CNV >100kb
Loss/gain >100kb

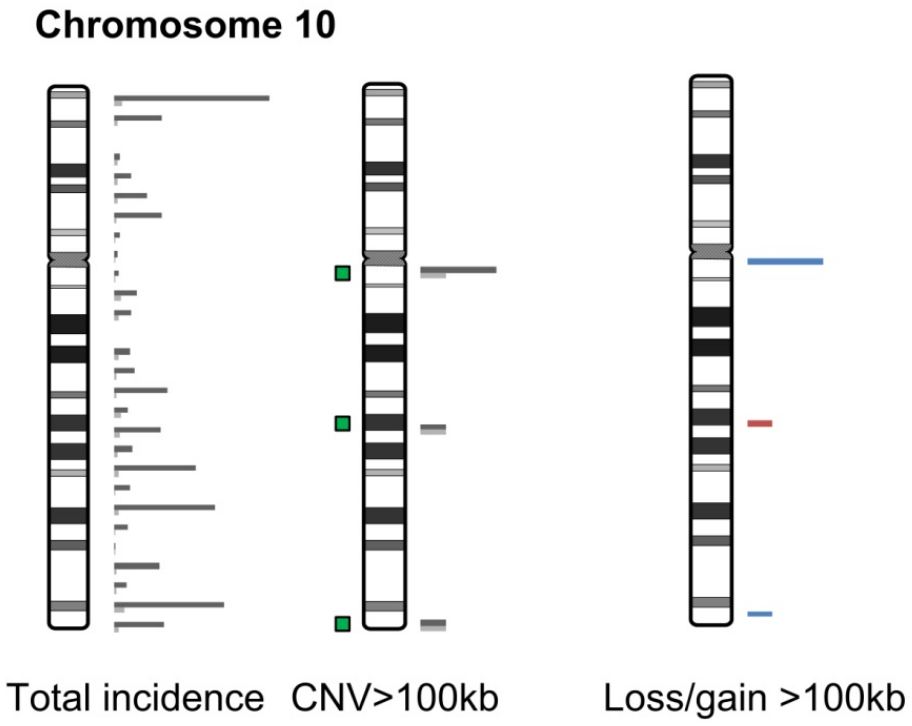
**Chromosome 7**



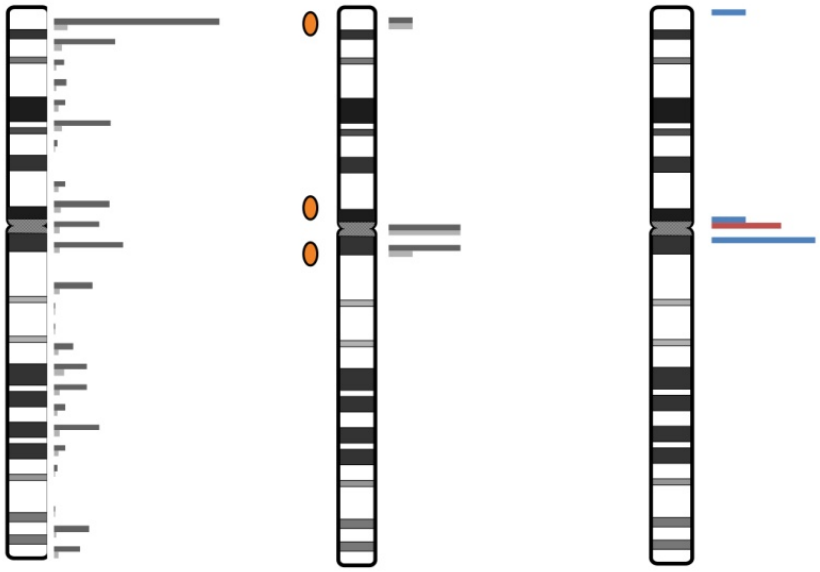

**Chromosome 8**




Chromosome 9
Total incidence
CNV>100kb
Loss/gain >100kb

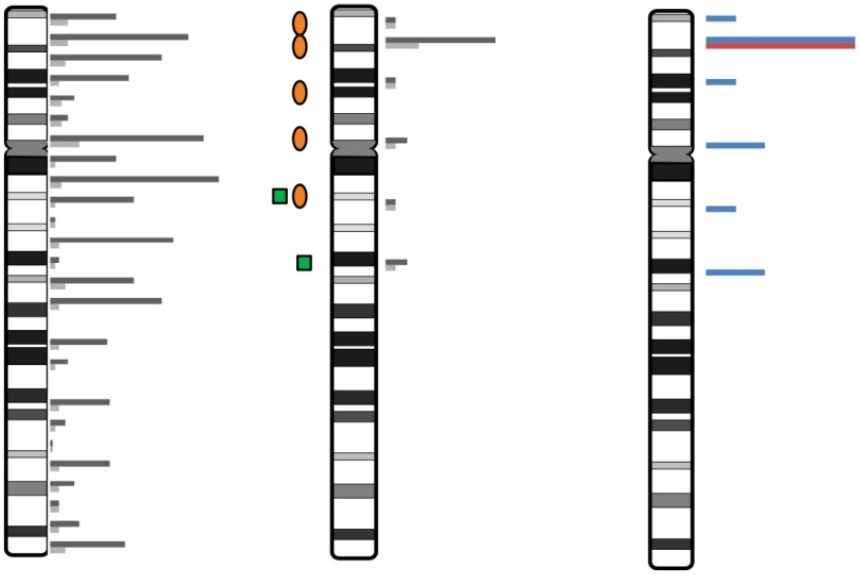

Chromosome 10
Total incidence
CNV>100kb
Loss/gain >100kb

**Chromosome 11**

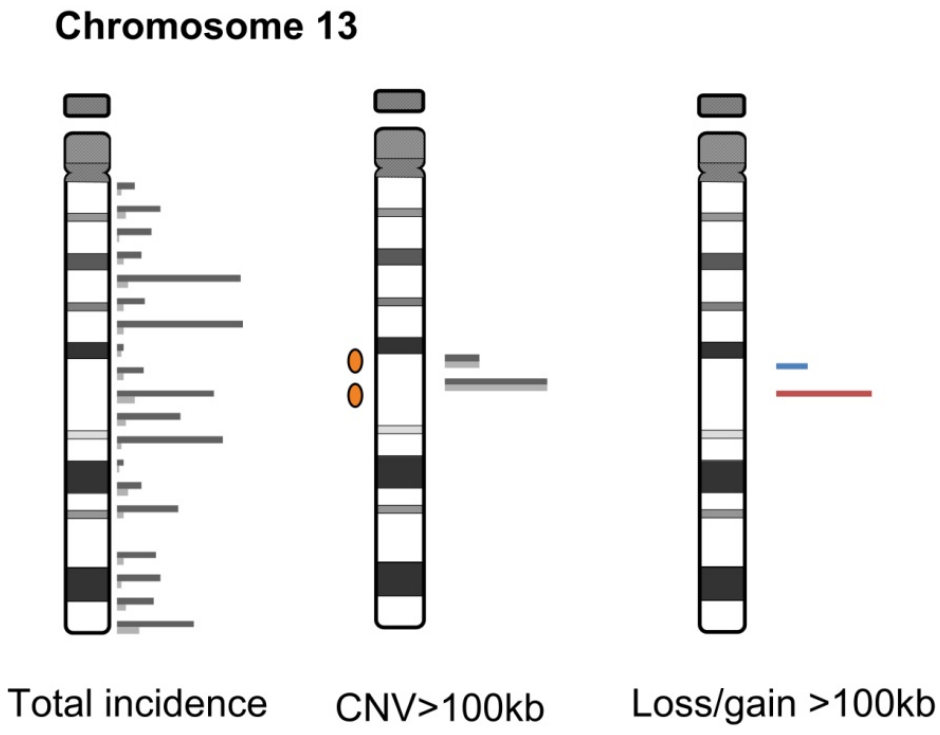


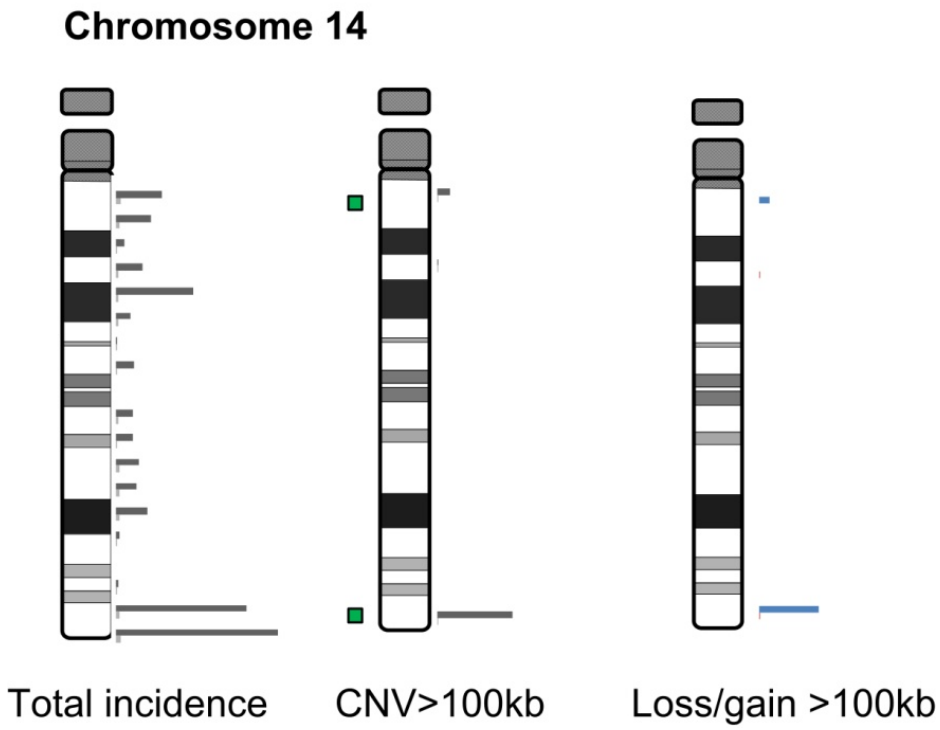
Total incidence CNV>100kb Loss/gain >100kb

**Chromosome 12**

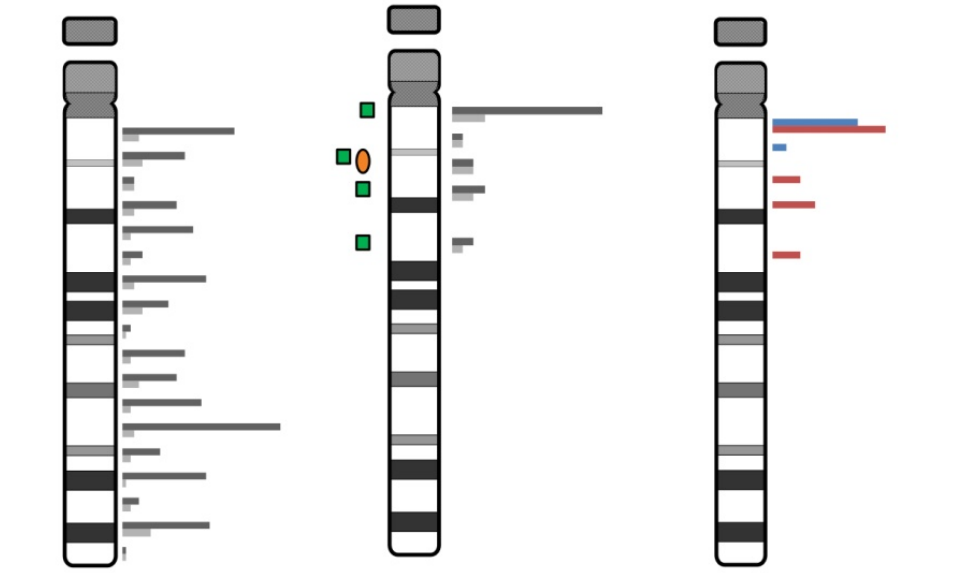


Total incidence CNV>100kb Loss/gain >100kb


Chromosome 13
Total incidence
CNV>100kb
Loss/gain >100kb

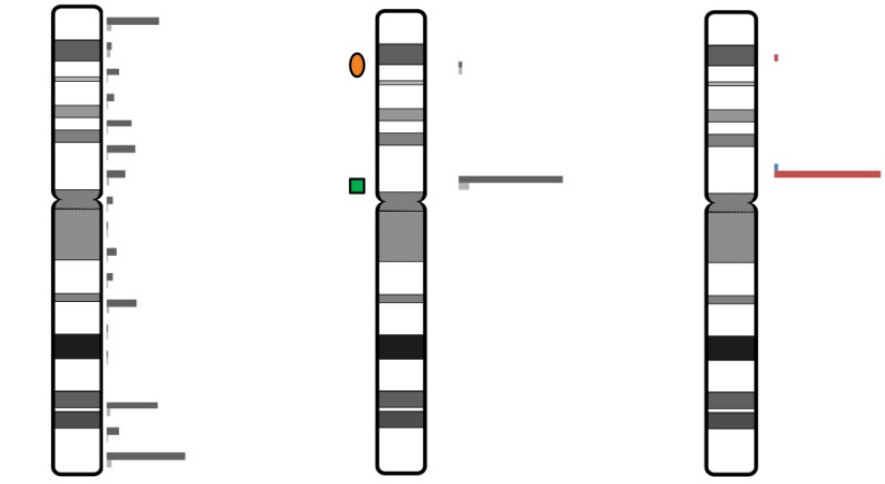

Chromosome 14
Total incidence
CNV>100kb
Loss/gain >100kb

## Chromosome 15

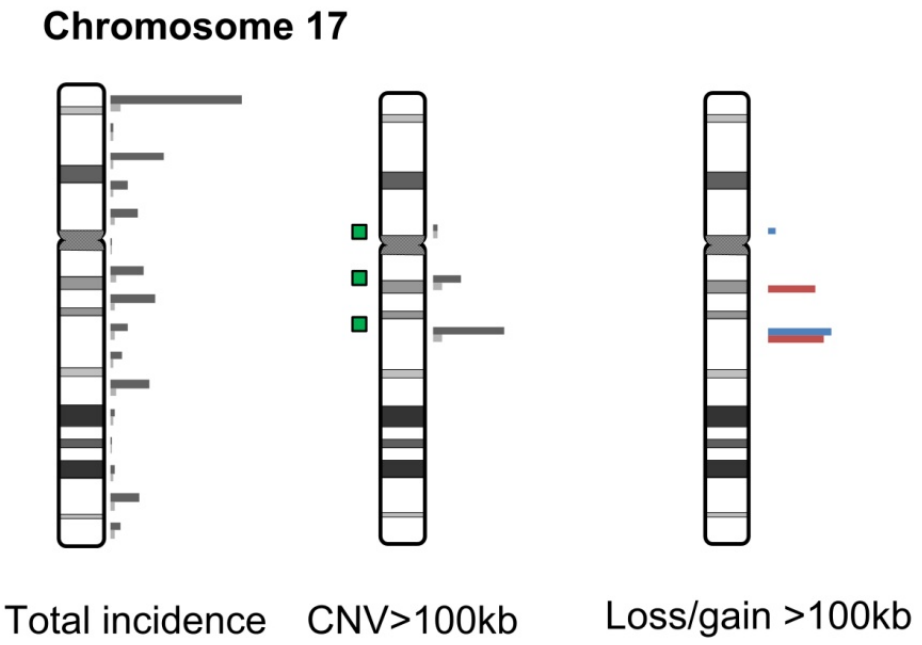


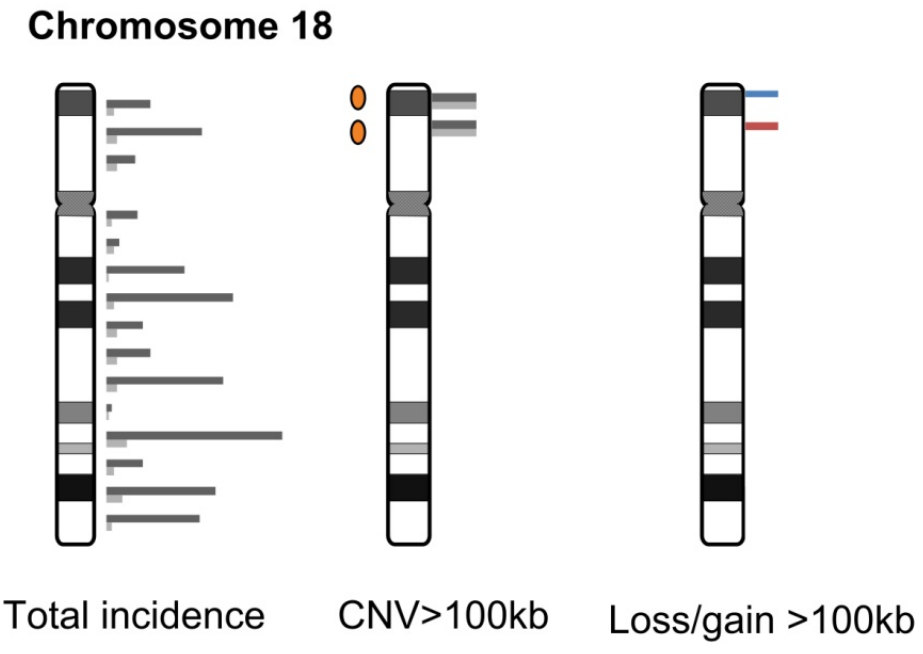
Total incidence CNV>100kb Loss/gain >100kb

## Chromosome 16

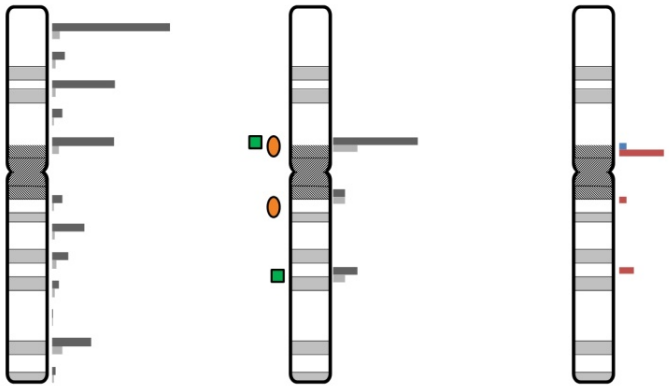


Total incidence CNV>100kb Loss/gain >100kb


Chromosome 17
Total incidence
CNV>100kb
Loss/gain >100kb

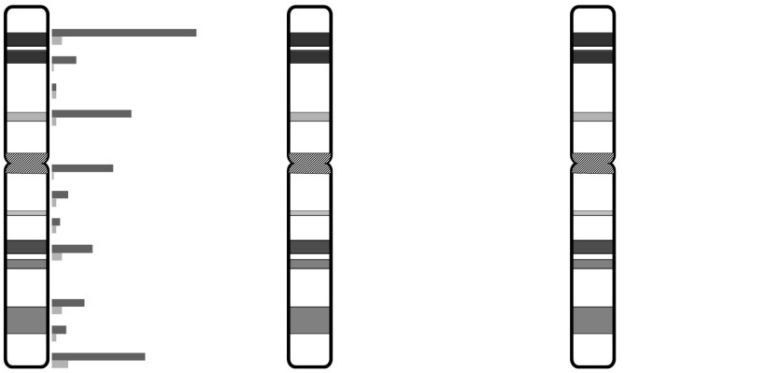

Chromosome 18
Total incidence
CNV>100kb
Loss/gain >100kb

**Chromosome 19**



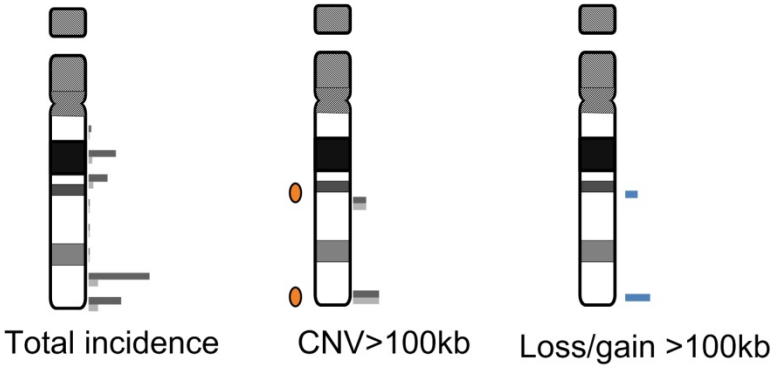
Total incidence CNV>100kb Loss/gain >100kb

**Chromosome 20**

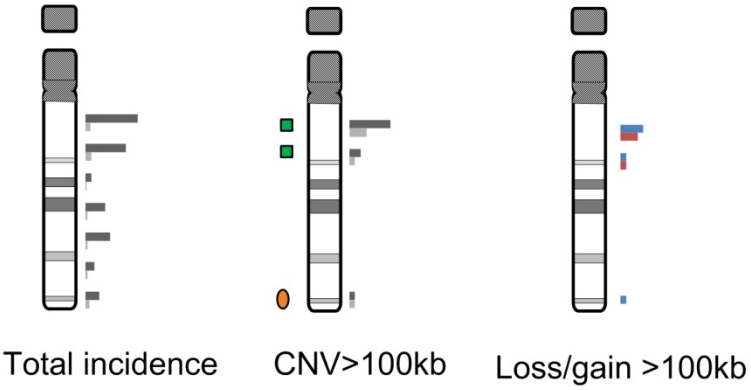


Total incidence CNV>100kb Loss/gain >100kb

**Chromosome 21**




**Chromosome 22**



# Appendix 2

# Method development: Fluorescence in-situ hybridisation techniques for confirmation of copy number change in the diagnostic laboratory

Part of this chapter was published in

Detection of segmental chromosome copy number gains using improved fluorescence in situ hybridisation techniques.

Nicole L. Chia, Howard R. Slater, Julia M. Potter

The Journal of the Association of Genetic Technologists 41 (1) 2015

## A2.1 Abstract

Fluorescence in-situ hybridisation (FISH) techniques are used for the targeted investigation of microduplication, microdeletion, and structural rearrangements. More recently FISH techniques using probes specific to the region of interest, have been applied to confirm genomic copy number variation (CNV). However, there are limitations in the assessment of FISH signal patterns. Tandem duplication of small CNVs appear as an increased signal size when standard FISH methods are applied. As such interpretation of signal patterns is subjective and further complicated in the presence of mosaicism. Here we describe a pretreatment that enhances the demonstration of tandem duplication. We assessed the sensitivity to CNVs of a minimum of 120kb in size and determined that the lower limit of detection of mosaicism is 10%. In contrast to some methods of chromatin extension and elongation, this technique is done using fixed cell preparations from routine cytogenetic harvesting, and can be applied to freshly harvested or stored fixed cell specimens. This modification to standard FISH preparations has the scope to be used as a screening tool for family and prenatal investigations.

## A2.2 Introduction

Copy number variation (CNV) is a recently identified major form of genetic variation in the human genome (1-5). Genome-wide microarray investigation for the detection of clinically significant copy number alteration has been introduced to an increasing number of routine diagnostic laboratories as the first tier

investigation for constitutional anomalies (6) and its application to prenatal diagnosis (7-10) and oncology is progressing rapidly (10-14).

Fluorescence in-situ hybridisation (FISH) is one method by which laboratories confirm the presence of copy number change (15-18). However, current methods of metaphase and interphase FISH investigation are limited in their effectiveness to detect copy number gains, particularly tandem duplications, due to limited resolution (15). The process of chromatin extension has been utilized to provide higher levels of resolution. These methods achieve a resolution 10 times greater than standard metaphase FISH achieving an overall resolution of 10-100kb compared to 1,000 to 3,000kb for standard metaphase FISH (19-22). This technique has been applied to the fine physical mapping of genes, characterising chromosome architecture, ordering of BAC clones and defining chromosomal rearrangement breakpoints (14, 20, 23, 24).

This investigation reports on a FISH method that can be applied to routine cytogenetic cell preparations. In contrast with other published fibre FISH and chromatin elongation methods, some of which are complex in their approach, require specific equipment or culturing treatments, the method described here can be incorporated easily into routine FISH investigations in any diagnostic laboratory. Performed in conjunction with routine methods this investigation algorithm enhances the effectiveness of FISH investigations and provides a comprehensive approach to the investigation of copy number change. It can be used as a cost effective screening for family studies.

## A2.3 Materials and Method

CNV data was obtained from the Aussie Normals Collection using Illumina Human Omni1-Quad and CNV Partition (Illumina, Inc) and PennCNV (25) detection software applied. Written consent was provided by the participants and ethics approval was provided by the Australian National University Human Research Ethics committee. In addition, CNV data was collected from diagnostic samples referred to The Canberra Hospital, Australia using Illumina CytoSNP12 and CNV Partition software (Illumina, Inc). The analysis of de-identified CNV data was undertaken with permission from the ACT Health Research and Ethics Committee.

Lymphocyte cultures from individuals enrolled in the "Aussie Normals" Collection and samples referred to a routine diagnostic laboratory were established for conventional karyotype and FISH investigations (5).

### A2.3.1 Cytogenetic cell preparation

Peripheral blood was inoculated into RPMI medium and HAMs F10 containing phytohaemagglutinin (PHA) (Life Technologies) for 72-96 hrs. The cells were harvested with Colcemid (40ug/ml, Sigma-Aldrich) for 9 min, 0.075M KCL for 13 min, 5% acetic acid wash and 4 changes of Carnoy fixative (3:1 methanol and acetic acid).

## A2.3.2 FISH lysis pretreatment

### A2.3.2.1 Freshly harvested suspensions

A FISH pretreatment was applied to enhance the demonstration of FISH signals for CNV gain >100kb in size. The cell suspension was dropped across the width of a Poly-L-Lysine coated slide (Fisher Scientific) using a glass pasteur pipette and air dried for 1 min. The slides are checked using phase microscopy and 10 x objective. An area on the slide with an optimum cell concentration of 30-50 individual nuclei per field of view with no evidence of overlapping or crowding is identified and marked with a diamond pen (Figure 1). The slides are placed in lysis buffer (0.5% SDS, EDTA 50mM, TRIS HCL 0.2M pH 8.0) for 5 minutes at room temperature (Table 1). Ethanol (94%) was gently added to the top of the buffer to form a layer to a depth of 1cm for a further 10 min.

**Table 1.** Reagent concentration of the lysis buffer

| Reagent | Volume |
|---|---|
| SDS 0.5% | 0.625ml |
| EDTA 50mM | 2.5 ml |
| TRIS 0.2M ph 8.0 | 5.0 ml |
| $H_2O$ | 16.875 ml |

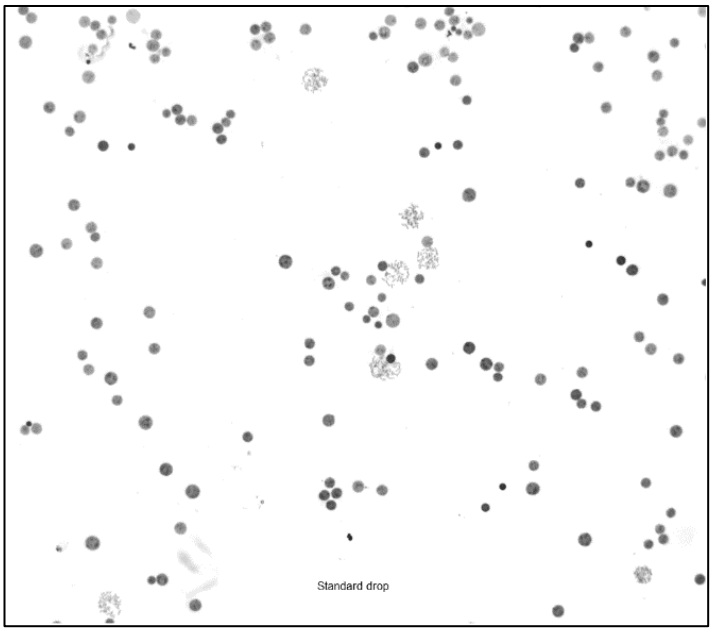



**Figure 1.** The optimal spreading cytogenetic cell suspension shows no overcrowding or overlapping of nuclei and 30-50 nuclei per field of view (10x objective captured)

## A2.3.2.2 Cytoplasmic Cell Preparations

The method was optimised for fixed cell suspensions greater than 6 months old or preparations with excessive cytoplasm (Figure 2). Cell suspensions that showed dense cytoplasm surrounding metaphase spreads and nuclei were tested with varying concentrations of SDS, TRIS or EDTA in the lysis buffer (Table 2). The effect of the temperature of the lysis buffer was also assessed. To do this the standard concentration of lysis buffer was preheated to 60°C and the slides stained in Leishman stain to determine the effect to the cytoplasmic cells.

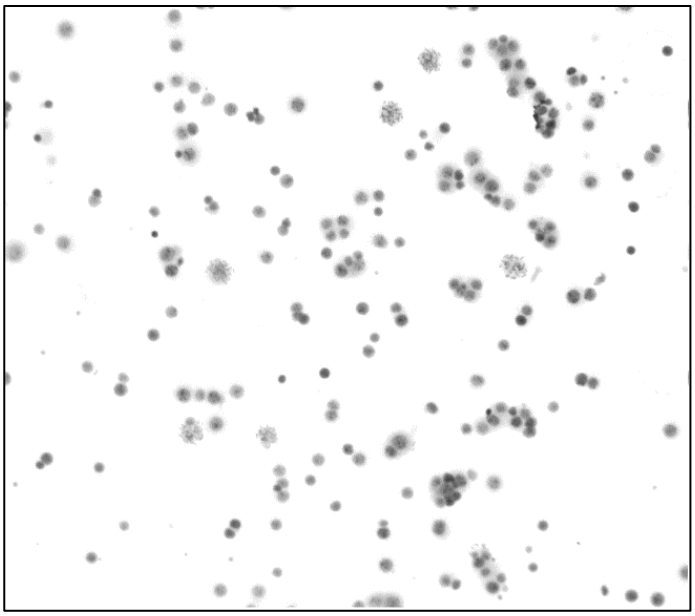


**Figure 2.** Cell preparations with thick cytoplasm surrounding the cells

**Table 2**. Reagent concentrations evaluated for the optimisation of the lysis buffer for cytoplasmic preparations.

| **Reagent** | **Volume** | | |
|---|---|---|---|
| | **Mix 1** | **Mix 2** | **Mix 3** |
| SDS 0.5% | 0.625ml | 0.625ml | 1.25ml |
| EDTA 50mM | 2.5 ml | 5.0ml | 2.5ml |
| TRIS 0.2M ph 8.0 | 5.0 ml | 10.0ml | 5.0ml |
| $H_2O$ | 16.875 ml | 9.875ml | 16.25ml |

## A2.3.3 Slide Preparation

The slides were removed slowly through the layer of ethanol at a 30° angle (Figure 3), then fixed in 70% ethanol for 30 min and dehydrated in 95% and 100% ethanol successively for 5 min each (14, 26, 27).

## A2.3.4 FISH Protocol

FISH processing is completed using standard laboratory procedures. 200µl of RNAse is dispensed onto each slide and incubated at $37^0$C for 30 mins. The slides are washed 2x5 mins in 2xSSC. Dispense 200µl of pepsin onto each slide and incubate at $37^0$C for 2 min. Wash slides 5 mins in PBS fix in 1% paraformaldehyde for 2 mins. Wash slides 5 mins in PBS. Dehydrate through 70, 90 and 100% ethanol and air dry slides. The probe is prepared (Table 3) and denatured in a $75^0$C waterbath for 5 mins.

**Table 3.** Probe mix for FISH hybridisation

| Reagent | per sample |
|---|---|
| **Probe A** | 0.5ul |
| **Probe B** | 0.5ul |
| **Cot 1** | 0.5ul |
| **Hybridisation Buffer** | 1.0ul |
| **Total** | 2.5ul |

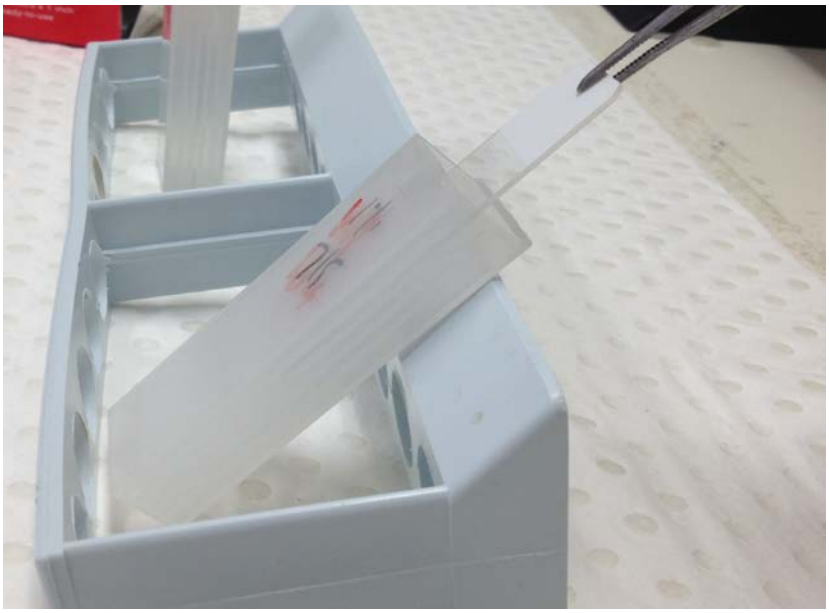


**Figure 3.** The slide is carefully removed at an angle from the pretreatment buffer to create tension on the cells and drag the fibres and nuclei along the slide

Slide denaturation is done with a 70% solution of formamide (7 parts formamide, 1 part 20xSSC, 2 parts distilled water) at 73-74$^0$C for 5 mins. The slides are dehydrated through cold 70, 90 and 100% ethanol for 3 mins per wash. After air drying 2.5 ul of probe mix is dispensed onto the marked area on the slide and hybridised overnight at 37$^0$C. The slides are washed in 0.4 SSC / 0.3% NP40 at 74$^0$C for 2 min and at room temperature in 2 SSC / 0.1% NP40 for 30sec. The slides are air dried and counter stained with DAPI.

## A2.3.4 Probe design

The ELN locus specific probe supplied by Vysis (Abbott Molecular, USA) was selected to demonstrate the 2.1Mb duplication located at 7q11.23.

For 9 CNV regions locus-specific, pre-labelled bacterial artificial chromosome clones (BAC) were selected using the UCSC Genome Browser (http://genome.ucsc.edu)(28) (Figure 4) and sourced from an external provider (SickKids, Toronto, Canada). To demonstrate gains, a single probe was selected from within the relevant CNV or where possible, two different RP11 clones were selected pre-labelled with spectrum orange and spectrum green. The location of the RP11 clones to the correct chromosome region was confirmed on metaphase spreads prepared from the suspensions of CNV negative controls in the "Aussie Normals" cohort.

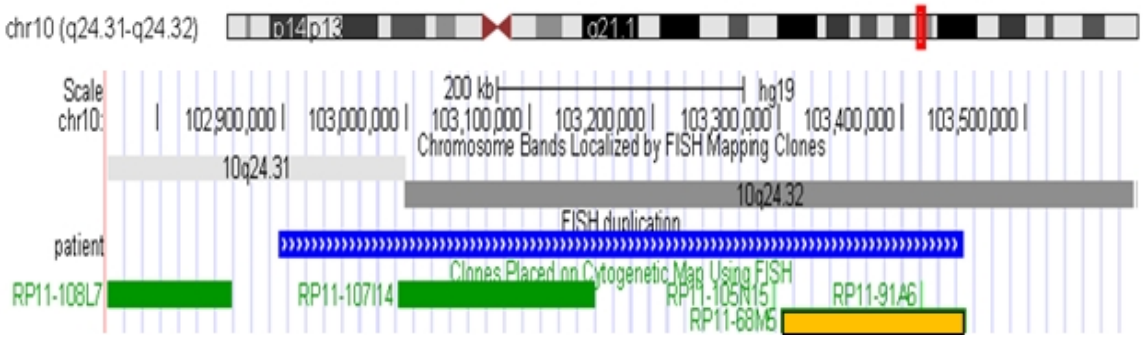



**Figure 4.** FISH probes were designed using the UCSC Genome Browser with custom track applied. Two FISH clones RP11-107I14 and RP11-68M5 were selected within a 500kb CNV gain at 10q24.31q24.32. To optimize demonstration of the CNV FISH, the two adjacent FISH clones were chosen pre-labelled in spectrum orange and spectrum green.

## A2.3.5 Method validation

The method was validated following the experimental design shown in Figure 5. To determine the effectiveness of the pretreatment buffer the slides were prepared with and without lysis buffer pretreatment and the FISH signals in interphase nuclei scored (Figure 6). The cell suspensions of samples without the copy number change (CNV negative control) and samples with the copy number change (CNV positive control) were analysed. The sensitivity of the method was determined by testing 10 CNV regions with duplications ranging in CNV length from 128kb to 2,100kb. The reproducibility was assessed with multiple samples for 2 CNV regions. To establish the background incidence of signal separation due to DNA replication a known CNV negative control was used for each of the regions.

## A2.3.6 Assessment of positional information

Metaphase spreads in the positive control prepared without pretreatment were analysed to exclude inter and intra-chromosomal insertions and unbalanced translocations (Figure 6c). Metaphase spreads were analysed and captured using a Zeiss Axioscope microscope and MetaSystems digital imaging software (MetaSystems, Germany).

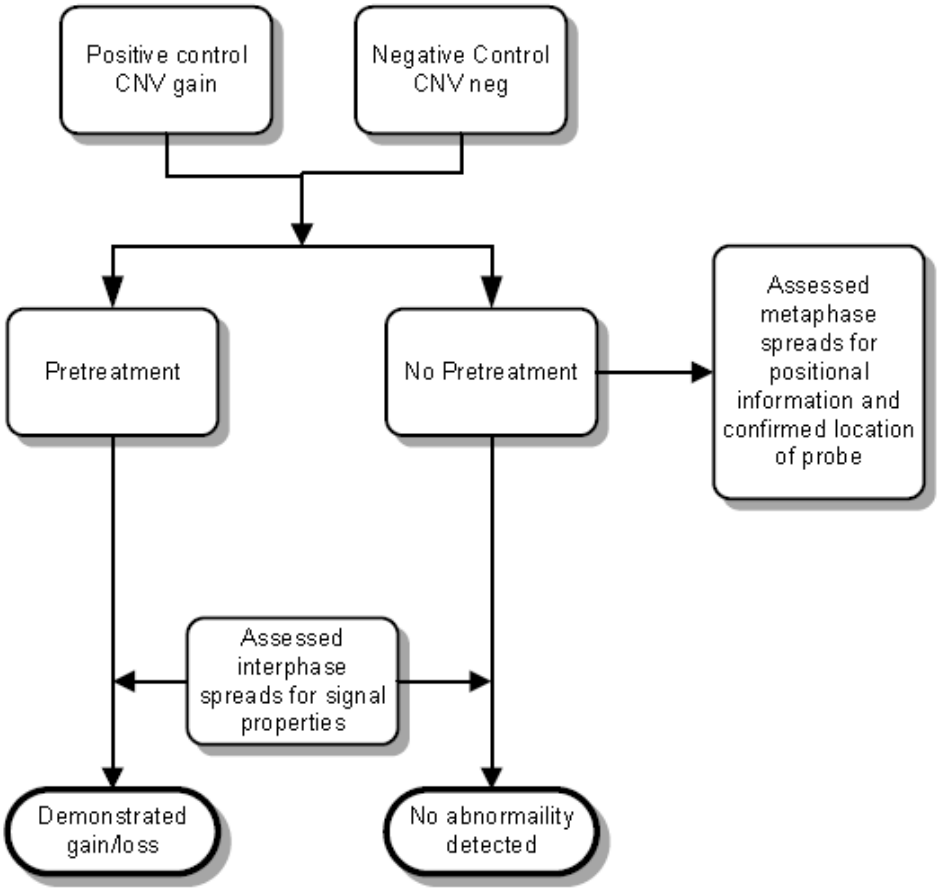



**Figure. 5**. The chart shows the workflow for the experiment design for verification of the CNV lysis pretreatment. Analysis of metaphase spreads for a CNV positive control excludes intra or interchromosomal insertions. Confirmation of correct band assignment of the FISH probe is made with metaphase spreads of a CNV negative control. The FISH signal properties for slides with and without pretreatment for the CNV positive and negative control are compared for signal properties.

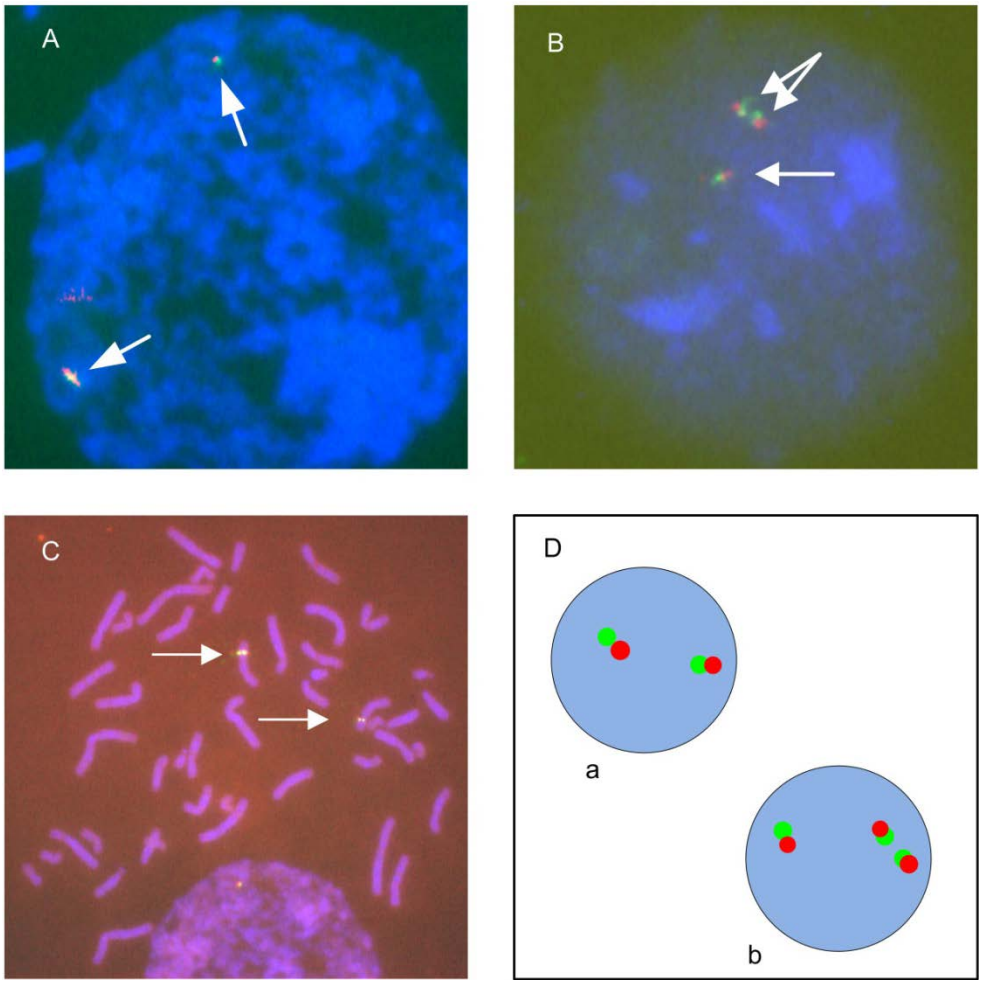



**Figure. 6** The FISH investigation of a 500kb CNV gain at 10q24.31q24.32 compared with and without pretreatment lysis buffer. (A) A single fusion signal (RP11-107I14 and RP11-68M5) on two homologues is observed with the standard FISH preparation without pretreatment, whereas (B) the same sample using the pretreatment method demonstrated two co-located fusion signals in addition to the homologue. Metaphase spreads on untreated slides were analysed for positional information and a translocation is excluded in this sample (C). The schematic representation of FISH signals are shown in (D) for a normal (a) and sequential duplication (b).

## A2.3.7 Assessment of mosaicism

The limit of detection of mosaicism was ascertained with a serial dilution of a non-mosaic constitutional 2.1Mb tandem duplication at 7q11.23 using the LSI ELN FISH probe (Vysis, Abbott Molecular, US). The suspension was mixed with a known normal control (Table 4). To ascertain that the concentrated suspensions for the normal and CNV positive were of similar cell density interphase nuclei were counted in 3 fields of view on phase microscopy (40x objective). An aliquot of each of the suspensions were mixed to produce suspensions that represented levels of mosaicism. The signal pattern was scored from 100 consecutive interphase nuclei for each dilution representing 10%, 20%, 30%, 50% and 80% mosaicism.

**Table 4.** Cell suspension dilution for levels of mosaicism

| | 0% | 10% | 20% | 30% | 50% | 80% | 100% |
|---|---|---|---|---|---|---|---|
| **7q11.23 gain** | 0 | 100ul | 200ul | 300ul | 500ul | 800ul | 1000ul |
| **7q11.23 diploid** | 1000ul | 900ul | 800ul | 700ul | 500ul | 200ul | 0 |

## A2.4 Results

## A2.4.1 Evaluation of the pretreatment buffer

Slides prepared from cell suspensions of freshly harvested lymphocyte cultures were treated with lysis buffer. DNA fibres were released from nuclei and stretched along the slide. Some nuclei remained intact (Figure 7).

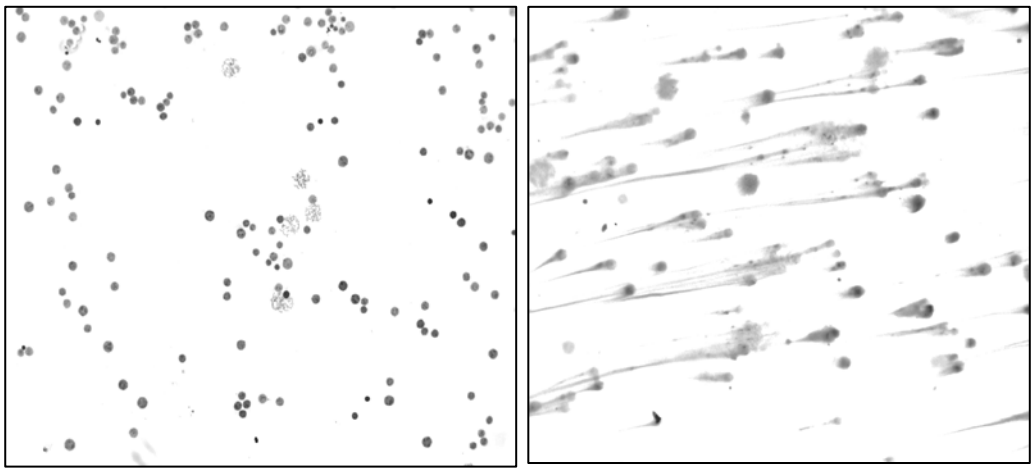


**Figure 7.** Slide preparation with the same suspension treated without (A) and with lysis pretreatment buffer (B). DNA fibres are released from the nuclear membrane after pretreatment with lysis buffer and intact nuclei appear larger than without the pretreatment (10x magnification).

## A2.4.2 Optimisation of cytoplasmic cell preparations

The pretreatment had minimal effect on nuclei that are surrounded by cytoplasm. Cytoplasm was apparent around cells and there was no release of DNA fibres. Variation to the concentration of SDS, EDTA and TRIS in the lysis buffer did not alter the appearance of cytoplasm surrounding the cells. The slides treated with lysis buffer at 60 C showed nuclei free of cytoplasm but few DNA fibres (Figure 8).

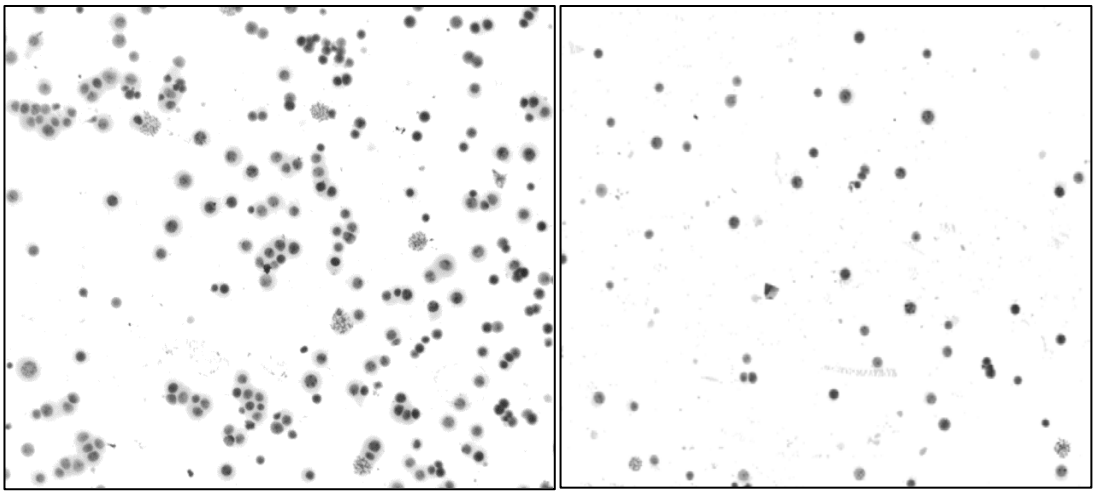


**Figure. 8**. Cytoplasmic cell preparations were assessed with different concentration of lysis buffer and heat denaturation. A) There was no change to the cytoplasm with 2x SDS. B) Pretreatment with lysis buffer at 60 C removed the cytoplasm.

## A2.4.3 Fibre-FISH evaluation

FISH signals were observed in DNA fibres as a series of green and red signals. Scoring of signal numbers within a nucleus was complicated by the uneven spreading of fibres. Likewise, overlapping of DNA fibres from multiple cells made interpretation of the number of FISH signals within a nucleus unreliable (Figure 9). The FISH signal pattern in intact interphase nuclei showed signal pattern and number that are representative of a single cell. As such assessment of signal numbers and interpretation of copy number gain and loss was valid.

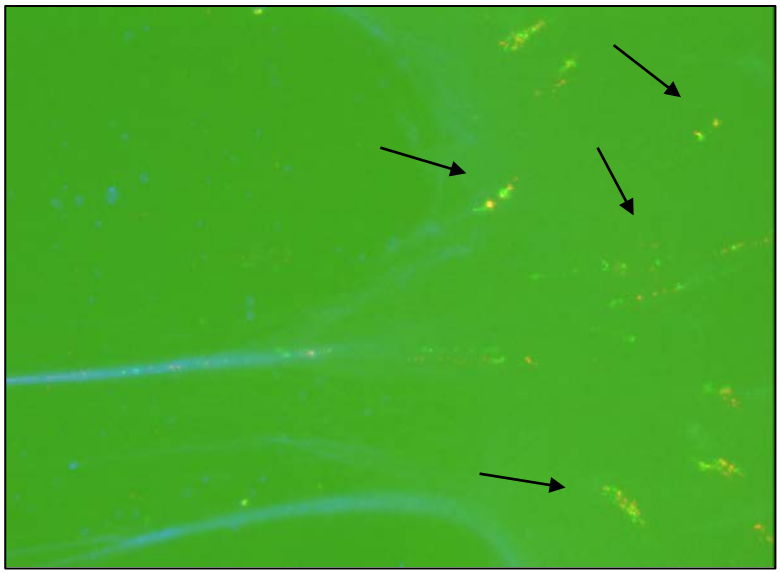


**Figure 9.** The FISH signals observed for the DNA fibres were unreliable for accurate assessment of CNV gain. Overlapping fibres makes interpretation challenging and unreliable.

## A2.4.4 Interpretation of FISH signals

Metaphase spreads from the negative control was analysed and the correct band assignment of the fluorescently labelled RP11 clone was confirmed. Cross hybridisation with several chromosome regions was observed with one RP11 clone and this was excluded from further investigation. Metaphase spreads for the CNV positive control were analysed and chromosomal rearrangements including inter-chromosomal insertion and unbalanced translocations were excluded in all CNV regions.

In total, 500 interphase cells were scored representing a minimum of 25 consecutive cells in 10 CNV regions for CNV positive and CNV negative controls

(Table 5). Co-location of signals measuring 1-2 signal widths apart in addition to a single signal was observed in the positive controls, whereas 2 distinct signals were observed in the negative control (Figure 6). Using these criteria signals consistent with tandem duplication were defined as clear separation of signals (1-2 signal width apart) (Figure 6d). To evaluate the sensitivity of the pretreatment a range of CNV lengths were analysed. The largest tandem duplication demonstrated was 2.1Mb within region 7q11.23 and the smallest tandem duplication was a 160kb gain within 12p13.33 (Table 6).

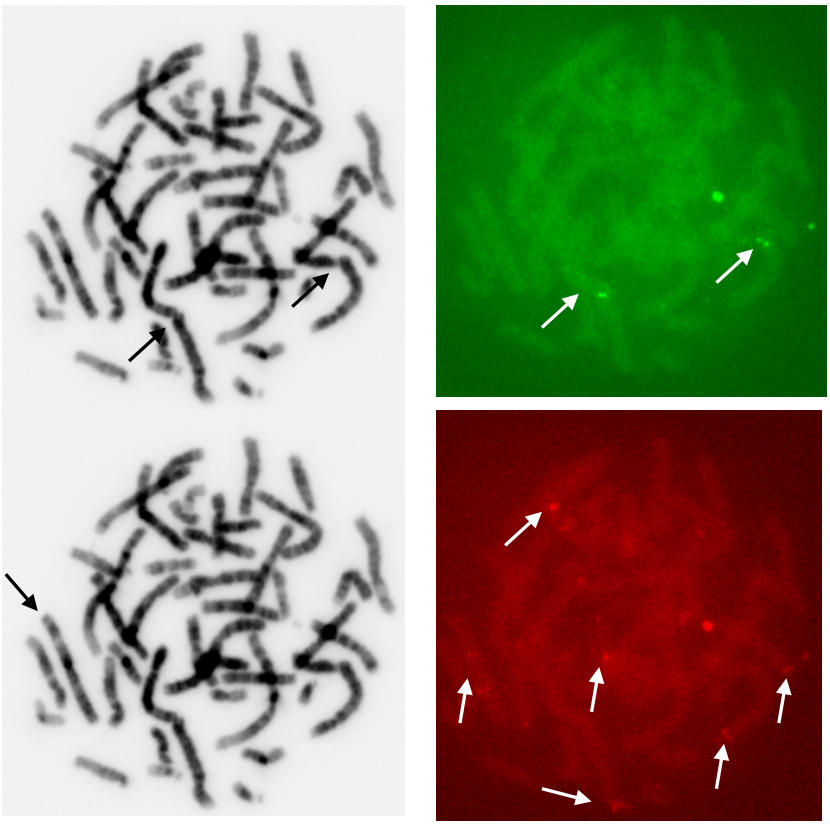


**Figure 10.** Cross hybridisation of FISH probe a) RP11-14H21 at 3p12.3 with 4q subtelomere and b) RP11- 634L22 (3p12.3) with many chromosomes

**Table 5**. CNV positive and negative controls were analysed for 10 CNV regions

| Chromosome Location | Sample id. | RP11 clone | CNV Negative | CNV Gain |
|---|---|---|---|---|
| 1q32.1 | AN-2311 | RP11-1104P23 | 2/30 | 25/30 |
| 1q32.1 | AN-2311 | RP11-284G5 | 2/30 | 25/30 |
| 2p11.2 | AN-2351 | RP11-301O19 | 1/30 | 20/25 |
| 7q31.32 | AN-2338 | RP11-1078A22 | 0/30 | 24/30 |
| 12p13.33 | AN-2306 | RP11-73O21 | 2/30 | 25/30 |
| 12q13.11 | AN-2344 | RP11-14O3 | 3/30 | 8/30 |
| 16p11.2 | AN-2322 | RP11-80F22/ RP11-488I20 | 0/25 | 22/25 |
| | AN-2345 | RP11-80F22/ RP11-488I20 | | 20/25 |
| | AN-2352 | RP11-80F22/ RP11-488I20 | | 20/25 |
| 17q21.31 | AN-2340 | RP11-244K17 | 3/25 | 21/25 |
| 10q24 | TCH-105 | RP11-68M5/ RP1-107II4 | 0/25 | 24/25 |
| 7q11.23 | TCH-1516 | ELN(Vysis) | 4/100 | 90/100 |

**Table 6**. The CNV regions tested with the pretreatment method show a range of CNV length and age of cell suspension.

| Chromosome Location | Start | End | Size | | Result | Suspension storage |
|---|---|---|---|---|---|---|
| 1q32.1 | 202165699 | 202389494 | 223795 | 1 | Tandem | 2 years |
| 1q32.1 | 202455140 | 202838694 | 383554 | 1 | Tandem | 2 years |
| 2p11.2 | 86149308 | 86363012 | 213704 | 1 | Tandem | 2 years |
| 7q31.32 | 121033877 | 121359954 | 326077 | 1 | Tandem | 2 years |
| 12p13.33 | 1727243 | 1888251 | 161008 | 1 | Tandem | 2 years |
| 12q13.11 | 45519568 | 45772649 | 253081 | 1 | Unconfrimed | 2 years |
| 16p11.2 | 34343935 | 34614585 | 270651 | 3 | Tandem | 2 years |
| 17q21.31 | 41519627 | 41713128 | 193501 | 1 | Tandem | 2 years |
| 7q11.23 | 72401192 | 74488253 | 2087061 | 1 | Tandem | < 6mths |
| 10q24 | 102896801 | 103448649 | 551848 | 4 | Tandem | 1 year |

## A2.4.5 Assessment of lower limit of detection for mosaicism

To assess the limit of detection of duplications using the FISH pretreatment method, the interphase signal pattern for a non-mosaic tandem duplication at 7q11.23 was analysed. Additional and co-located signals consistent with a tandem duplication were recorded in 90% (90/100) of interphase cells scored consecutively from pretreated slides. The CNV negative control yielded signal separation in only 4% of interphase nuclei scored. The lower limit of detection of mosaicism was investigated by producing a serial dilution for the non-mosaic tandem duplication at 7q11.23 that is representative of a range of levels of mosaicism (Figure 11). The lower limit of detection for the CNV assessed here was

determined to be 10% (p=0.0081; Fisher's Exact test) (Table 7). However due to the limitations of producing serial dilutions of fixed cell suspensions to reflect mosaicism, levels of mosaicism below 10% were not assessed.

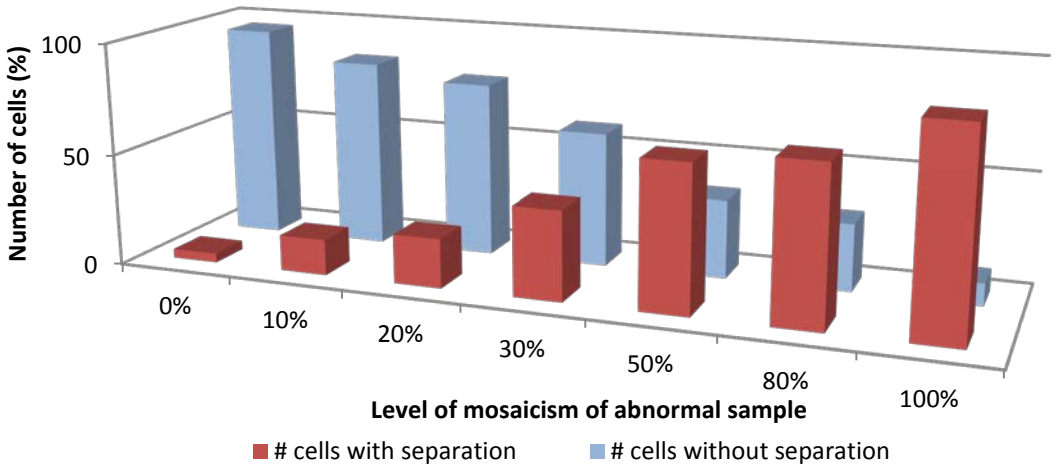



**Figure 11.** The lower limit of detection of mosaicism was determined by analysing the signal patterns in a serial dilution of a sequential duplication. The number of cells showing separation of signals is consistent with the level of mosaicism although 4% of split signals could be observed in the normal (0%) sample and 10% of cells in the abnormal sample did not effectively show the duplication. The lower limit of detection is 10% mosaicism with 16% of cells with a signal pattern demonstrating the sequential duplication.

**Table 7**. FISH signals in interphase cells scored for the serial dilution mosaic experiment.

| Relative mosaicism | 0% | 10% | 20% | 30% | 50% | 80% | 100% |
|---|---|---|---|---|---|---|---|
| # cells with separation | 4 | 16 | 22 | 40 | 65 | 70 | 90 |
| # cells without separation | 96 | 84 | 78 | 60 | 35 | 30 | 10 |
| Fisher's Exact test p= | | 0.0081 | 0.0002 | <0.0001 | <0.0001 | <0.0001 | <0.0001 |

## A2.5 Discussion

## A2.5.1 FISH verification of CNV in diagnostic laboratories

An inherent feature of microarray analysis of genomic copy number is the lack of positional information for genomic gains. Copy number gain may be the result of an intra- or inter-chromosome insertion, which carries with it an increased risk of transmission of an unbalanced gamete to a fetus in a future pregnancy (16). The gain may also represent a tandem duplication of DNA sequences. Demonstration of the chromosomal rearrangement and ascertainment of familial inheritance will assist in determination of clinical relevance (16, 29).

Confirmation of genomic copy number imbalance detected in the clinical setting may be required in regions of poor probe coverage and complex regions of genome architecture.[30] Recent studies have indicated that many factors interact with CNV calling. Technical factors such as marker specificity on platforms and the calling power of CNV detection algorithms have been shown previously to result in call variation (15, 31, 32). For instance, Park et al. 2010 demonstrated discordance in CNV calling with comparative genome hybridisation arrays (aCGH) in relation to

the reference genome (31). The authors demonstrated that a copy loss in the genome may be observed as a copy gain in the test sample when it is actually diploid (copy 2).

Verification of some CNVs has remained challenging in the routine laboratory. Methods such as MLPA (multiplex ligation dependent probe amplification) and QFPCR (quantitative fluorescent polymerase chain reaction) have recently been utilised by some laboratories for the demonstration of copy number change (15, 17). Although these methods are effective, their application may be outside the realm of some diagnostic cytogenetic laboratories. FISH investigation has traditionally been the preferred method albeit with limitation of sensitivity with respect to CNV size and demonstration of tandem duplications. Whole genome libraries of RPCI-11 male human BAC library are available for all chromosomes at an average resolution of ~1Mb (33-35). The resolving power of FISH methods is limited when closely juxtaposing BAC clones are hybridised to metaphase spreads (20). The resolution is improved by the application of less condensed DNA such as early pachytene stage or fibre FISH cell preparations (14, 33).

## A2.5.2 Comparison with previously described methods of chromosome elongation

Methods for the elongation of chromosomes and fibre-FISH have been described previously (14, 20, 22, 23, 26, 27, 36). Cytocentrifugation of a hypotonic cell suspension onto glass slides results in mechanically stretched chromosomes which

enhances the mapping resolution of chromosomes without losing the physical location of the region of interest (20, 24). Nuclei embedded in agarose gel melted onto Poly-L-Lysine slides release DNA fibres that can be used for high resolution physical mapping (20, 36). However there are few methods of chromatin extension that can be applied to stored fixed cell suspensions from cytogenetic cell preparations to enhance the resolution of FISH investigation. Techniques resulting in the spreading of DNA fibres whilst providing the highest level of resolution (19-21) are not optimal for the interpretation of FISH signals for copy number change due to the excess spreading of DNA fibres on the slide and tendency for fibre overlap (39). Other investigation protocols require specialist equipment and are applied to pre-harvested cell preparations (20) and as such are unsuitable for stored fixed cell suspensions.

## A2.5.3 Method development

Here we describe a simple and effective method of chromatin extension that can be used for the confirmation of CNV in diagnostic laboratories. This is based on the method described in Pole et al. 2006 and previously in Pulcini et al. 1998 and Mann et al. 1997 (14, 26, 27).

### A2.5.3.1 Preliminary evaluation of the lysis pretreatment

The method was evaluated for the demonstration of CNV >100kb. To determine the effectiveness of the lysis buffer pretreatment in providing elongation of chromatin the slides were prepared with and without the lysis buffer. This was done for samples with the copy number change and without the copy number change. Due to the spreading of DNA fibres released from nuclei and the need for accurate assessment of signal number, it was determined that the FISH signals in the DNA fibres could not be assessed. Accurate assessment of signal numbers and signal pattern was complicated by fibre spreading and overlapping of signals. As such analysis of copy number loss and gain was unreliable.

It was observed that intact nuclei showed signals that were confined within the nucleus and could be assessed for number. Due to apparently less compacted DNA signal pattern could also be assessed. For example duplication appeared as adjacent but separated signals.

### A2.5.3.2 Technical limitations of the pretreatment

It was observed that stored suspensions and nuclei surrounded by cytoplasm are resistant to the lysis pretreatment. The inclusion of heat to this step reduced this limitation and the cytoplasm was removed. Heat denatures cellular proteins (37) degrading the cytoplasmic and nuclear membranes resulting in optimum exposure of the chromatin to the FISH probes.

## A2.5.4 Determination of sensitivity and limit of detection

The sensitivity of the method has been assessed with respect to CNV size and lower limits of detection. Several CNV gains ranging in size from 0.16-2.1Mb were assessed using this method and with one exception, tandem duplication was demonstrated in all cases. A limitation of the method is the differential rate of swelling of cells which resulted in the lack of demonstration of tandem duplication in some cells. The limit of detection of mosaicism is shown to be 10%. However, due to variability of cell preparation assessment of mosaicism should be done in conjunction with the microarray result and copy number metric (log R value) (38).

## A2.5.5 Contribution to diagnostic testing

The slide pretreatment step with lysis buffer provides a fast, easy and cost effective modification to standard FISH procedures. The method, first described by Mann et al. 1997 (26) adapted for biological applications (27) and demonstration of translocation breakpoints (14) is applied here for the confirmation of CNV in a diagnostic laboratory for clinical interpretation. A key benefit is the use of fixed cell suspensions from routine cytogenetic cell preparations and standard FISH protocols. The pretreatment method can be applied to either fresh or stored suspensions and is effective with cell preparations where the cytoplasm surrounds the nuclei. When used in conjunction with the preparation of metaphase spreads allows the determination of tandem duplication and differentiation from inter or intra-chromosomal insertion.

## A2.6 Conclusion

The FISH method described here has direct application in genetics diagnostic laboratory for the confirmation of chromosome copy number gain. In the study here a practical approach to the confirmation of CNV in a routine diagnostic laboratory by enhancing the resolution of interphase FISH is described. When used in conjunction with FISH analysis of metaphase spreads and confirmed on the proband, this FISH method provides positional information of potential clinical significance not only for the proband but also for other family members through the determination of inheritance and ascertainment of familial risk factors.

# A2.6 References


1. Pinto D, Marshall C, Feuk L, Scherer SW. Copy-number variation in control population cohorts. Hum Mol Genet. 2007;16 Spec No. 2:R168-73.
2. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number in the human genome. Nature. 2006;444(7118):444-54.
3. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, et al. Detection of large-scale variation in the human genome. Nat Genet. 2004;36(9):949-51.
4. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, et al. Large-scale copy number polymorphism in the human genome. Science. 2004;305(5683):525-8.
5. Chia NL, Bryce M, Hickman PE, Potter JM, Glasgow N, Koerbin G, et al. High-resolution SNP microarray investigation of copy number variations on chromosome 18 in a control cohort. Cytogenet Genome Res.2013; 141:16-25.
6. Miller DT, Adam MP, Aradhya S, Biesecker LG, Brothman AR, Carter NP, et al. Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. Am J Hum Genet. 2010;86(5):749-64.
7. Lamb AN. Laboratory aspects of prenatal microarray analysis. Clin Lab Med. 2011;31(4):615-30, ix.
8. Hillman SC, Pretlove S, Coomarasamy A, McMullan DJ, Davison EV, Maher ER, et al. Additional information from array comparative genomic hybridization technology over conventional karyotyping in prenatal diagnosis: a systematic review and meta-analysis. Ultrasound Obstet Gynecol. 2010;37(1):6-14.
9. Leung TY, Vogel I, Lau TK, Chong W, Hyett JA, Petersen OB, et al. Identification of submicroscopic chromosomal aberrations in fetuses with increased nuchal translucency and apparently normal karyotype. Ultrasound Obstet Gynecol. 2011;38(3):314-9.
10. Schwartz S. Clinical utility of single nucleotide polymorphism arrays. Clin Lab Med. 2011;31(4):581-94, viii.
11. Tiu RV, Gondek LP, O'Keefe CL, Elson P, Huh J, Mohamedali A, et al. Prognostic impact of SNP array karyotyping in myelodysplastic syndromes and related myeloid malignancies. Blood. 2011;117(17):4552-60.
12. Talseth-Palmer BA, Holliday EG, Evans TJ, McEvoy M, Attia J, Grice DM, et al. Continuing difficulties in interpreting CNV data: lessons from a genome-wide CNV association study of Australian HNPCC/lynch syndrome patients. BMC Med Genomics. 2013; 6:10.
13. Heinrichs S, Kulkarni RV, Bueso-Ramos CE, Levine RL, Loh ML, Li C, et al. Accurate detection of uniparental disomy and microdeletions by SNP array analysis in myelodysplastic syndromes with normal cytogenetics. Leukemia. 2009;23(9):1605-13.
14. Pole JC, Courtay-Cahen C, Garcia MJ, Blood KA, Cooke SL, Alsop AE, et al. High-resolution analysis of chromosome rearrangements on 8p in breast, colon and pancreatic cancer reveals a complex pattern of loss, gain and translocation. Oncogene. 2006;25(41):5693-706.
15. Qiao Y, Liu X, Harvard C, Nolin SL, Brown WT, Koochek M, et al. Large-scale copy number variants (CNVs): distribution in normal subjects and FISH/real-time qPCR analysis. BMC Genomics. 2007;8:167.
16. Neill NJ, Ballif BC, Lamb AN, Parikh S, Ravnan JB, Schultz RA, et al. Recurrence, submicroscopic complexity, and potential clinical relevance of copy gains detected by array CGH that are shown to be unbalanced insertions by FISH. Genome Res. 2011;21(4):535-44.

17. Bruno DL, Ganesamoorthy D, Schoumans J, Bankier A, Coman D, Delatycki M, et al. Detection of cryptic pathogenic copy number variations and constitutional loss of heterozygosity using high resolution SNP microarray analysis in 117 patients referred for cytogenetic analysis and impact on clinical practice. J Med Genet. 2009;46(2):123-31.
18. South ST. Chromosomal structural rearrangements: detection and elucidation of mechanisms using cytogenomic technologies. Clin Lab Med. 2011;31(4):513-24, vii.
19. Haaf T, Ward DC. Structural analysis of alpha-satellite DNA and centromere proteins using extended chromatin and chromosomes. Hum Mol Genet. 1994;3(5):697-709.
20. Laan M, Isosomppi J, Klockars T, Peltonen L, Palotie A. Utilization of FISH in positional cloning: an example on 13q22. Genome Res. 1996;6(10):1002-12.
21. Haaf T, Ward DC. High resolution ordering of YAC contigs using extended chromatin and chromosomes. Hum Mol Genet. 1994;3(4):629-33.
22. Trask B, Pinkel D, van den Engh G. The proximity of DNA sequences in interphase cell nuclei is correlated to genomic distance and permits ordering of cosmids spanning 250 kilobase pairs. Genomics. 1989;5(4):710-7.
23. Weier HU. DNA fiber mapping techniques for the assembly of high-resolution physical maps. J Histochem Cytochem. 2001;49(8):939-48.
24. Molina O, Blanco J, Anton E, Vidal F, Volpi EV. High-resolution fish on DNA fibers for low-copy repeats genome architecture studies. Genomics. 2012;100(6):380-6.
25. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. 2007;17(11):1665-74.
26. Mann SM, Burkin DJ, Grin DK, Ferguson-Smith MA. A fast, novel approach for DNA fibre-fluorescence in situ hybridization analysis. Chromosome Res. 1997;5(2):145-7.
27. Pulcini F, Devignes MD. Fibre-fluorescence in situ hybridization directly performed from fresh biological samples: novel perspectives for genetic diagnosis. Chromosome Res. 1998;6(6):501-3.
28. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002;12(6):996-1006.
29. South ST, Brothman AR. Clinical laboratory implementation of cytogenomic microarrays. Cytogenet Genome Res. 2011;135(3-4):203-11.
30. Pinto D, Darvishi K, Shi X, Rajan D, Rigler D, Fitzgerald T, et al. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nat Biotechnol. 2011;29(6):512-20.
31. Park H, Kim JI, Ju YS, Gokcumen O, Mills RE, Kim S, et al. Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. Nat Genet. 2010;42(5):400-5.
32. Dellinger AE, Saw SM, Goh LK, Seielstad M, Young TL, Li YJ. Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. Nucleic Acids Res. 2010;38(9):e105.
33. Cheng Z, Buell CR, Wing RA, Jiang J. Resolution of fluorescence in-situ hybridization mapping on rice mitotic prometaphase chromosomes, meiotic pachytene chromosomes and extended DNA fibers. Chromosome Res. 2002;10(5):379-87.
34. Cheung VG, Dalrymple HL, Narasimhan S, Watts J, Schuler G, Raap AK, et al. A resource of mapped human bacterial artificial chromosome clones. Genome Res. 1999;9(10):989-93.
35. Morley M, Arcaro M, Burdick J, Yonescu R, Reid T, Kirsch IR, et al. GenMapDB: a database of mapped human BAC clones. Nucleic Acids Res. 2001;29(1):144-7.
36. Heiskanen M, Karhu R, Hellsten E, Peltonen L, Kallioniemi OP, Palotie A. High resolution mapping using fluorescence in situ hybridization to extended DNA fibers prepared from agarose-embedded cells. Biotechniques. 1994;17(5):928-9, 32-3.

37. Burgman PW, Konings AW. Heat induced protein denaturation in the particulate fraction of HeLa S3 cells: effect of thermotolerance. J Cell Physiol. 1992;153(1):88-94.
38. Conlin LK, Thiel BD, Bonnemann CG, Medne L, Ernst LM, Zackai EH, et al. Mechanisms of mosaicism, chimerism and uniparental disomy identified by single nucleotide polymorphism array analysis. Hum Mol Genet. 2010;19(7):1263-75.

# Appendix 3

# Glossary

**Autozygosity**: Homozygosity for alleles at loci that is identical by descent, inherited from a common ancestor (Li et al. 2006).

**BIR:** A recombination based mechanism that repairs breaks in a replication fork.

**Copy number variation (CNV):** A DNA segment with a change in copy number compared to a reference genome and measures >1kb.

**CNV activity:** A description of the incidence of CNV or CNVR in a genomic region.

**Copy number variant region (CNVR):** A description of CNV where CNV in different individuals has a common start and end position, or that shares significant overlap.

**Double strand break (DSB):** Both DNA strands are broken at the same site.

**Fork stalling and template switch (FoSTeS):** A replicative mechanism characterised by fork stalling resulting in the invasion of a single strand template into another fork.

**Genomic architecture:** The entireity of functional elements of DNA, such as genes, regulatory elements and sequence properties including repetitive elements and sequence motifs.

**Homozygosity:** Allele with identical haplotype.

**Identical by descent (IBD):** Homozygous allele where the haplotype is identical as a result of inheritance from a common ancestor.

**Identity by state (IBS):** Homozygous allele where the haplotype is identical by chance.

**LINE:** Long interspersed nuclear elements.

**Long contiguous stretches of homozygosity (LCSH**): where there are uninterrupted consecutive SNP markers with a homozygous genotype.

**LTR:** Long terminal repeats

**Microhomology mediated end joining (MMEJ):** A non homologous non replicative mechanism of CNV formation resulting in deletion of sequence between sequences of base homology.

**Microhomology:** Short sequence of base pair identity**.**

**Microhomology mediated break induced recombination (MMBIR):** Repair of single strand break during replication between sequences with base identity.

**Non-allelic homologous recombination (NAHR):** Homologous recombination involving sequences of extensive homology.

**Non homologous end joining (NHEJ):** Repair of double stranded break that is not mediated by homology.

**Novel variant:** A variant not previously documented for the genomic location.

**No call:** A SNP marker fails to record a genotype.

**SINE:** Short interspersed nuclear elements.

**Single strand break (ssDNA):** A break in a single strand of DNA.

# Appendix 4

# Publications and Presentations

# Publications

**Chia NL**, Bryce M, Hickman PE, Potter JM, Glasgow N, Koerbin G, Danoy P, Brown MA, Cavanaugh J: High-resolution SNP microarray investigation of copy number variations on chromosome 18 in a control cohort. Cytogenet Genome Res 141:16-25, 2013.

**Chia NL**, Slater HR, Potter JM: Modified fibre fluorescence in situ hybridisation techniques for the verification of copy number variation. The Journal of the Association of Genetic Technologists 41 (1), 2015.

**Nicole L Chia**, Howard R Slater, Julia M Potter: Discordance between CNV detection software tools: The challenges for the diagnostic setting. Current Topics in Genetics, 6: 2016.

# Website Publication

Database of Genomic Variants: http://dgv.tcag.ca/dgv Accession number estd198

# Published Abstracts

**Chia NL**, Slater H, Potter JM: CNVs provide the clue for a cryptic t(5;11) in acute myeloid leukemia. Cancer 206:145-216, 2013

## Oral Presentations

**Chia NL.** DNA copy number variation in the normal population: The Aussie Normals. Post Graduate Conference, Australian National University, 2010

**Chia NL**, Brown M, Danoy P, Cavanaugh J. Validation and population screening of copy number variation on chromosome 18 in the normal population. GeneMapper Scientific Meeting, Hobart, Tasmania, 2011

**Chia NL.** Chromosome 18: A model for investigation of Copy Number Variation. Post Graduate Conference, Australian National University, 2011

**Chia NL**. The Ups and Downs, Ins and Outs of Copy Number Variation. Australasian Genomics workshop, Sydney, NSW, 2012

**Chia NL** Slater H, Potter JM. CNVs provide the clue for a cryptic t(5;11) in acute myeloid leukemia. Cancer Cytogenomics Microarray Consortium and Cytogenomics Array Group, Chicago, USA, 2014

**Chia NL**. Consanguinity, LCSH and lessons learned from the Australian Population. Australasian Genomics workshop, Brisbane , QLD, 2014

## Poster Presentation

**Chia NL**, Slater HR, Potter JM. Confirmation of novel segmental copy number gain using improved fluorescence in-situ hybridisation techniques. RCPA Update, Melbourne, Vic. 2013