Expanding the genetics of microcephalic primordial dwarfism

Jennie Elaine Murray

Thesis submitted for the degree of Doctor of Philosophy
The University of Edinburgh
October 2014

This thesis is composed of original research undertaken by myself, and where the work of others is included their contributions have been duly acknowledged. This work has not been submitted for any other degree or professional qualification.

Jennie Murray

Acknowledgements

Firstly, I would like to say a big thank you to my supervisor, Professor Andrew Jackson, for his fantastic guidance and support over the last 3 years. Your integrity and dedication has been inspiring and, although I still have a long way to go, your input has been invaluable in helping me grow and develop as a clinical scientist. I am also in a huge debt of gratitude to Louise Bicknell for her superb day to day supervision, teaching, mentoring, encouragement, general life coaching and eternal optimism. You always kept me going when things got tough. I am also incredibly grateful for the rapid proof reading of this thesis enabling me to achieve my rather tight deadline.

I also want to thank Professors David Fitzpatrick and Mary Porteous for their additional supervision and career guidance which helped me to obtain a clinical fellowship and complete this thesis.

Next, I would like to thank Alison Meynert for the huge bioinformatic support she provided and always promptly answering my endless stream of programming questions which made such a large part of this thesis possible; Mike Halachev also for all things bioinformatic and particularly for help with the CNV analysis. I would also like to thank the members of the IGMM computing services; especially Graeme Grimes for getting me started with SQL and John Ireland, Ewan McDowall and Semir Kazazic for always quickly coming to my rescue on the many occasions my minimal computer skills failed me.

I am also indebted to Carol-Anne Martin for inducting me into the field of cell biology and imparting some of her immense knowledge base to me as well as patiently teaching me many of the experimental techniques used in this thesis including cell culture, immunoblotting, immunofluorescence and microscopy. A big thank you also to Morad Ansari for all his technical help with performing array-CGH, generous donation of reagents and slides as well as support with AmpliSeq™.

Additionally I would like to thank all the other members of my lab, past and present for all their helpful input and discussions during lab meetings and general lab

support, in particular Anna Klingseissen for introducing me to the world of zebrafish, Maggie Harley for help with various techniques including cell culture and immunoblotting, Paula Carroll for being ever willing to help at all times, Martin Reijns and James Ding for help with cloning, Rachel Rigby for all things RNA related, Andrea Robertson for being such a lovely desk buddy, providing much needed light relief as well as providing invaluable help in the lab and Karen MacKenzie for lifts to work and regularly letting me vent my frustrations. I would also like to thank Anne Seawright for all her assistance in and around the lab including ensuring there was always milk and coffee in the office!

I would also like to express my gratitude to Stephen Brown, Agnes Gallacher and Stewart McKay in technical services for all their support in anything sequencing related; Lizzie Freyer for assistance with flow cytometry; Andrew Wood for insightful discussions and generous donation of antibodies; Craig Nicol for teaching me all things graphical for which I believe this thesis has greatly benefited and Matt Pearson for additional microscopy support.

A huge thank you also to all the patients and their families whose selfless contributions to this research in the hope of providing help to future generations has been incredibly humbling and has been a big motivation for me over the last three years. Additionally I'd also like to thank Eilidh Guild and Angela Sandilands for their administrative support, keeping me organised with patient data, samples and reports and also our clinical counterparts in the USA, Mike Bober and Angela Duker for their ongoing collaboration, providing access to patient samples and expert clinical opinion.

Finally, I would like to thank my family and friends for their eternal support, always believing in me and always attempting to be interested in what I was doing. And last but certainly not least, my incredible husband Andrew, for all your love, support and encouragement. You have been my absolute rock throughout as you always are in everything I do and therefore this thesis is dedicated to you along with our new beautiful baby daughter, Nina.

Abstract

Body mass varies considerably between different mammals and this variation is largely accounted for by a difference in total cell number rather than individual cell size. Insights into mechanisms regulating growth can therefore be gained by understanding what governs total cell number at any one point. In addition, control of cell proliferation and programmed cell death is fundamental to other areas of research such as cancer and stem cell research. Microcephalic Primordial Dwarfism (MPD) is a group of rare Mendelian human disorders in which there is an extreme global failure of growth with affected individuals often only reaching a height of around one metre in adulthood. To date, all identified disease genes follow an autosomal recessive mode of inheritance and encode key regulators of the cell cycle, where mutations impact on overall cell number and result in a substantially reduced body size. MPD therefore provides a valuable model for examining genetic and cellular mechanisms that impact on growth. The overall aims of this thesis were to identify novel disease causing genes, as well as provide further characterisation of known disease causing genes, through the analysis of whole exome sequencing (WES) within a large cohort of MPD patients. Following the design and implementation of an analytical bioinformatics pipeline, deleterious mutations were identified in multiple disease genes including *LIG4* and *XRCC4*. Both genes encode components of the non-homologous end joining machinery, a DNA repair mechanism not previously implicated in MPD. Additionally, the pathogenicity of novel mutations in subunits of a protein complex involved in chromosome segregation was assessed using patient-derived cells. These findings demonstrate WES can be successfully implemented to identify known and novel disease causing genes within a large heterogeneous cohort of patients, expanding the phenotype of known disorders and improving diagnosis as well as providing novel insights into intrinsic cellular mechanisms critical to growth.

Table of Contents

List of Figures

List of Tables

Abbreviations

AD	=	autosomal dominant
AR	=	autosomal recessive
bp	=	base pairs
BRCT	=	BRCA 1 C-terminal
BW	=	birth weight
Chr	=	chromosome
cDNA	=	complimentary DNA
CNV	=	copy number variation
Ct	=	control
dbSNP	=	database of single nucleotide polymorphisms
DDR	=	DNA damage response
DGV	=	database of genomic variation
dH_2O	=	distilled water
DMSO	=	dimethyl sulfoxide
DNA	=	deoxyribonucleic acid
DNAse	=	deoxyribonuclease
dNTP	=	deoxyribonucleotide triphosphate
DSB	=	DNA double strand break
DTT	=	dithiothreitol
EDTA	=	ethylenediaminetetraacetic acid
EtOH	=	ethanol
EVS	=	Exome Variants Server
FCS	=	fetal calf serum
Gest^n	=	gestation
Gb	=	gigabase, 1,000,000,000 base pairs
Het	=	heterozygous
Hgt	=	height
Hom	=	homozygous
IGV	=	Integrative Genomics Viewer
IR	=	ionizing radiation
Ix	=	investigation

1KG	=	1000genomes project
kb	=	kilobase, 1000 base pairs
kDa	=	kilodalton
LBC	=	lymphoblastoid cells
Lgt	=	length
MAF	=	minor allele frequency
M	=	molar
m	=	month
Mb	=	megabase, 1,000,000 base pairs
MCPH	=	microcephaly
1° MCPH	=	autosomal recessive primary microcephaly
MGS	=	Meier Gorlin Syndrome
MOPS	=	3-(N-morpholino)propanesulfonic acid
mRNA	=	messenger ribonucleic acid
MRI	=	magnetic resonance imaging
MPD	=	microcephalic primordial dwarfism
n	=	number
N/A	=	not available
NGS	=	next generation sequencing
NHEJ	=	Non-homologous end joining
NMD	=	nonsense mediated decay
ns	=	not significant
OFC	=	occipital-frontal circumference (head)
PBS	=	phosphate buffered saline
PCR	=	polymerase chain reaction
Pt	=	patient
pter	=	terminus of the short arm of a chromosome
qter	=	terminus of the long arm of a chromosome
RNA	=	ribonucleic acid
RNase	=	ribonuclease
RT	=	room temperature
s.d./SD	=	standard deviation

SDS	=	sodium dodecyl sulphate
siRNA	=	small interfering RNA
SNP	=	single nucleotide polymorphism
SNVs	=	single nucleotide variants
SSB	=	DNA single strand break
TBE	=	tris/borate/EDTA
T_m	=	melting temperature
Tris	=	tris(hydroxymethyl)aminomethane
UV	=	ultraviolet
V	=	volts
v:v	=	volume:volume
WCE	=	whole cell extract
WES	=	whole exome sequencing
WGS	=	whole genome sequencing
Wgt	=	weight
w:v	=	weight:volume
XR	=	X-linked recessive
y	=	year

Chapter 1: Introduction

1.1 Overview of microcephalic primordial dwarfism (MPD)

1.1.1 Definition

Primordial dwarfism refers to a group of rare Mendelian disorders characterised by an extreme failure of growth of prenatal onset with affected individuals often only reaching a height of around one metre in adulthood . Characteristically growth failure is global and so although height is severely reduced body proportions generally remain unchanged with a uniform reduction in body size (McKusick, 1955) (Figure 1.1). This distinguishes primordial dwarfism from other, more common causes of dwarfism, such as skeletal dysplasias (disorders of bone and cartilage growth) which typically result in an isolated reduction of limb length with relative preservation of head and torso size (Trotter *et al.*, 2005). Additionally in primordial dwarfism, head size is often severely reduced as a consequence of reduced brain growth (microcephalic primordial dwarfism, MPD) (Seckel, 1960). This can be in proportion to body size or disproportionately smaller (Klingseisen *et al.*, 2011). The presence of microcephaly distinguishes MPD from other short stature syndromes with relative macrocephaly but with a relative proportionate reduction in limb length and body size such as Russell Silver syndrome, SHORT syndrome and 3M syndrome (Wakeling *et al.*, 2010, Huber *et al.*, 2011, Dyment *et al.*, 2013).

Another key feature of MPD is the presence of intrauterine growth restriction (IUGR) demonstrating growth failure is prenatal in onset (Seckel, 1960). This differentiates MPD from growth failure in which onset only occurs postnatally. Affected patients are therefore much smaller at birth than expected for their gestational age. The exact timing of the onset of growth failure in MPD is unknown and likely to be variable but may be evident on antenatal ultrasound scans from as early as the first trimester (Mirzaa *et al.*, 2014). Thus MPD can be defined as a severe and global failure in growth occurring prior to birth resulting in a reduction in both head and body size.

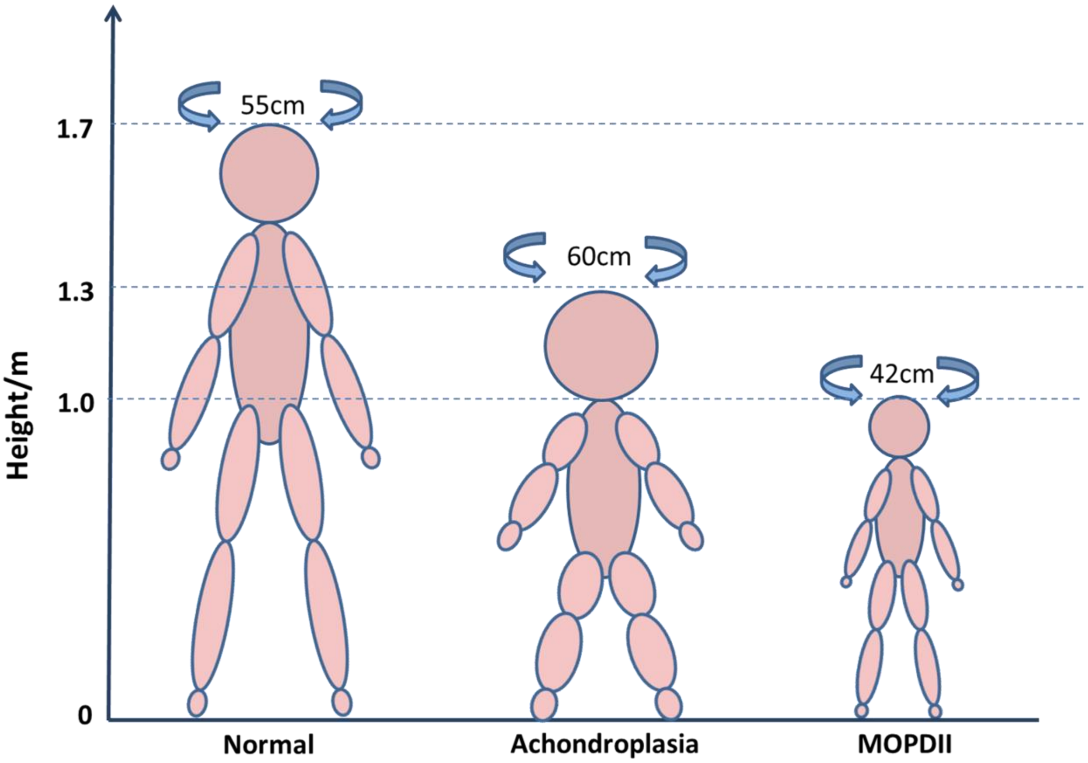


Figure 1.1. Comparison of body proportions in different disorders of dwarfism In MPD there is a global failure in growth resulting in an overall smaller body size with normal proportions (proportionate dwarfism). In achondroplasia (a common skeletal dysplasia), limb length is reduced disproportionately compared to head and torso (disproportionate dwarfism). Average height and OFC measurements are shown for an unaffected adult male, an adult male with achondroplasia (Horton *et al.*, 1978) and an adult with microcephalic osteodysplastic primordial dwarfism type II (MOPDII, the most common form of MPD, see Section 1.1.4.3) (Bober *et al.*, 2012).

1.1.2 Diagnostic criteria

Weight, height and head circumference are typical parameters used in the clinical setting to determine the general status of growth during childhood (Rogol *et al.*, 2014). However, a quantitative definition of MPD is lacking from older studies (McKusick, 1955, Seckel, 1960). Across a healthy population growth parameters display a normal distribution with 99.73% of individuals falling within 3 s.d. of the population mean (Cole, 2012). Given that MPD is a disorder of extreme growth failure, we have used postnatal growth parameters (OFC and height) of at least 4 s.d. below the population mean to classify patients with MPD (Klingseisen *et al.*, 2011). This is keeping with the human phenotype ontology database (Kohler *et al.*, 2014) which uses -4 s.d. to define severe short stature with less than 0.1% of the population falling into this group.

1.1.3 Aetiology

Many factors, both environmental and genetic, impact on growth either *in utero* or postnatally. These include nutritional deficiencies, metabolic diseases, infection, endocrine disorders, emotional neglect, chronic disease and drugs including both illegal and prescribed (Rogol *et al.*, 2014). In fact any disease process can hinder growth in childhood by diverting energy to another cause (Hochberg *et al.*, 2008). However, the rapidly increasing identification of disease causing genes in patients with MPD (O'Driscoll *et al.*, 2003, Rauch *et al.*, 2008, Al-Dosari *et al.*, 2010, Bicknell *et al.*, 2011a, Bicknell *et al.*, 2011b, Edery *et al.*, 2011, Guernsey *et al.*, 2011, He *et al.*, 2011, Kalay *et al.*, 2011, Qvist *et al.*, 2011, Ogi *et al.*, 2012) indicates growth impairment to such an extreme degree is primarily of genetic origin, most commonly a single gene disorder of autosomal recessive inheritance.

1.1.4 Spectrum of phenotypes

Although a diagnosis of MPD can be made based on an individual's growth parameters, a large degree of clinical heterogeneity is still apparent between individuals that fall within this group (Figure 1.2) reflecting large underlying genetic heterogeneity. Several distinct disease entities have now been described.

1.1.4.1 Seckel syndrome (MIM 210600)

In 1960, Seckel described a group of 15 patients with MPD (Seckel, 1960), which he then termed as 'bird headed dwarfism' or 'nanocephalic dwarfism' due to the significantly sloping forehead, large beaked nose and micrognathia (small chin) observed in these patients (Figure 1.2A). Growth retardation was prenatal in onset presenting with reduced growth parameters at birth. Postnatal growth continued to be poor in affected cases with significant developmental delay. Seckel's paper led to MPD being initially designated as 'Seckel syndrome' which is often still used interchangeably with the term 'MPD' particularly where a molecular diagnosis has not yet been ascertained (Sarici *et al.*, 2012). Large clinical heterogeneity therefore exists between many cases originally designated as Seckel syndrome however with improving knowledge of different clinical phenotypes and corresponding molecular defects several cases have since been reclassified (Willems *et al.*, 2010). Classical ‘Seckel syndrome’ is now often restricted to cases with severe pre- and post-natal microcephaly which is disproportionately more severe than short stature, typical facial features as described above and significant cognitive impairment (Faivre *et al.*, 2002).

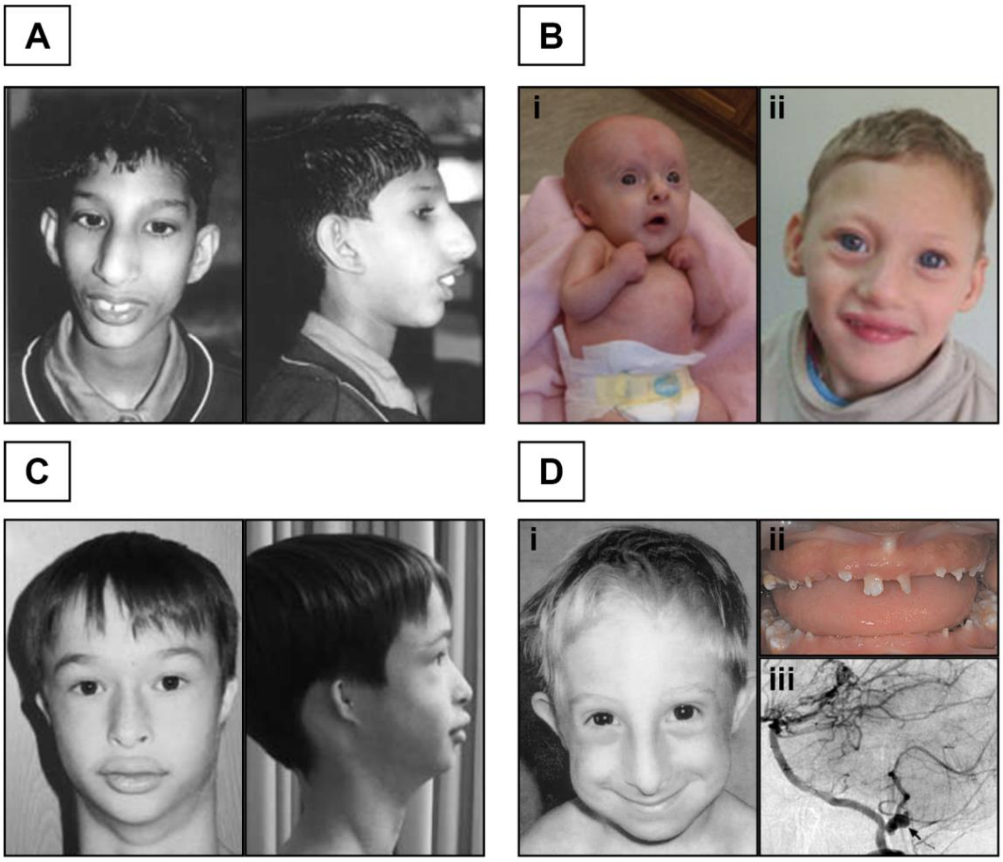


Figure 1.2. Clinical heterogeneity in MPD

Clinical photographs of patients with MPD with differing distinctive features. A) A patient with Seckel syndrome demonstrating microcephaly with characteristic sloping forehead and large beaked nose (Goodship *et al.*, 2000). B) Two unrelated patients with MOPDI at i) 6 months (Nagy *et al.*, 2012) and ii) 3 years (Abdel-Salam *et al.*, 2011) with sparse hair, prominent eyes, cerebral malformations, severe developmental delay and joint deformities. C) A patient with Meier-Gorlin syndrome characterised by microtia and patella hypoplasia (de Munnik *et al.*, 2012). D) i) A patient with MOPDII and typical facial features including long nose with overhanging columella, hypoplastic alae and micrognathia (Hall *et al.*, 2004), ii) Typically, patients with MOPDII also have severe microdontia and abnormally shaped teeth (Kantaputra *et al.*, 2011). iii) Angiogram demonstrating aneurysm of the posterior inferior cerebellar artery (black artery) in a patient with MOPDII (Bober *et al.*, 2010). MOPDII patients have an increased risk of cerebral vascular malformations.

1.1.4.2 Microcephalic Osteodysplastic Primordial Dwarfism Type I (MOPDI, MIM 210710)

MOPDI was first recognised as a distinct phenotype in 1967 (Taybi H & Linder D, 1967) and initially designated as Taybi-Linder syndrome. In an attempt to further delineate the phenotypes which were falling under the umbrella of 'Seckel syndrome', Majewski characterised MPD into three subgroups which he termed as Majewski osteodysplastic primordial dwarfism types I, II and III (Majewski *et al.*, 1982a, Majewski *et al.*, 1982b, Majewski *et al.*, 1982c). It soon became clear that Taybi-Linder syndrome and Majewski's MOPD type I and type III referred to the same disorder (Sigaudy *et al.*, 1998). For the purposes of this thesis this distinct subgroup of MPD will be referred to as MOPD type I.

MOPDI is characteristically associated with significant neurodisability and many do not survive beyond one year of age (Nagy *et al.*, 2012). Patients exhibit severe growth failure at birth with profound microcephaly (average weight and OFC at birth is -5 s.d. and -7 s.d. respectively). Structural brain abnormalities are common, especially neuronal migration defects such as agenesis of the corpus callosum, gyral abnormalities and heterotopias (Meinecke *et al.*, 1991, Sigaudy *et al.*, 1998, Klinge *et al.*, 2002, Pierce *et al.*, 2012). Consequently, severe developmental delay is common with accompanying seizures. Additionally, hypoplastic optic discs and Mondini malformations have also been reported resulting in vision and hearing impairment (Pierce *et al.*, 2012). Typical facial features include prominent eyes and an elongated, prominent nose (Figure 1.2B) (Edery *et al.*, 2011, Nagy *et al.*, 2012). Dry, aged-appearing skin and sparse hair are also characteristic along with skeletal abnormalities including joint dislocations, contractures, horizontal acetabulum, abnormal vertebrae, bowed long bones, elongated clavicles and delayed bone age (Sigaudy *et al.*, 1998). Recent identification of the underlying genetic aetiology in MOPDI has expanded the clinical spectrum of this disorder (see Section 1.3.4) and patients with milder growth and neurodevelopmental phenotypes have also been described (Abdel-Salam *et al.*, 2012) (Figure 1.2B).

1.1.4.3 Microcephalic Osteodysplastic Primordial Dwarfism Type II (MOPDII, MIM 210720)

Majewski's MOPD type II represents the most common and well characterised subgroup of MPD patients. MOPDII is characterised by a proportionate reduction in both head circumference and height to at least 5 s.d. below the population mean (average adult height -10 s.d. and OFC -8.5 s.d.) (Bober *et al.*, 2012). There is a characteristic facial appearance and body habitus with long nose, prominent nasal tip, hypoplastic alae, small mandible and mild truncal obesity (Figure 1.2C) (Rauch *et al.*, 2008). Skeletal abnormalities are typical and include gracile long bones, carpal fusion, brachydactyly, delayed bone age and hip deformities such as coxa vara and slipped epiphysis (Willems *et al.*, 2010, Karatas *et al.*, 2014). Intelligence is predominantly normal with only mild learning difficulties (Hall *et al.*, 2004). There is often severe microdontia with malformed teeth that can lack roots (Kantaputra *et al.*, 2011). There is also an increased incidence of early onset insulin resistance (Huang-Doran *et al.*, 2011).

Of most significance to affected families is the increased incidence of neurovascular complications predominantly cerebral aneurysms and vascular stenosis resulting in moya moya disease (30-50% of cases) (Brancati *et al.*, 2005, Bober *et al.*, 2010, Perry *et al.*, 2013), both of which impact on life expectancy. Awareness of this has led to the recommendation that screening with Magnetic Resonance Angiography (MRA) should be undertaken every 1-2 years as early intervention is likely to improve function and survival although long term follow up studies are not yet available (Perry *et al.*, 2013).

1.1.4.4 Meier-Gorlin Syndrome

Meier-Gorlin syndrome (MGS) is characterised by a triad of growth failure, microtia and hypoplastic or absent patella (Gorlin *et al.*, 1975). Other malformations have also been described including cortical hypoplasia and lobar congenital emphysema as well as musculoskeletal abnormalities (clindactyly, hypermobility, slender long bones, dislocated joints, contractures, delayed bone age) and urogenital tract anomalies (micropenis, hypospadias, cryptorchidism, hypoplastic labia, clitoromegaly) (Bongers *et al.*, 2001, de Munnik *et al.*, 2012). Additionally, failure of development of secondary sexual characteristics, particularly breast development, is often a feature. Patients also have similar facial characteristics which include microstomia and full lips with a prominent narrow convex nose becoming more apparent with age (Figure 1.2C) and intellect is usually normal. Although MGS is considered a subgroup of MPD, less severe growth failure has been reported in affected patients and growth parameters may even be within normal range (Bicknell *et al.*, 2011a).

1.1.4.5 Other disorders which overlap MPD

1.1.4.5.1 Syndromic conditions affecting global growth

A proportionate reduction in growth of prenatal onset is a common observation in many genetic disorders (Online Mendelian Inheritance in Man, OMIM$^{®}$) but few are associated with consistent growth failure that falls with the defined range of MPD (Figure 1.3). Therefore many syndromes are predominantly characterised on the basis of other, more distinguishing features such as the presence of malformations or distinctive dysmorphic features and the degree of growth failure is often poorly quantified. However, patients presenting with microcephaly and short stature within the defined MPD range (Section 1.1.2) have been described in several disorders including Bloom's syndrome, Wolf-Hirschhorn syndrome and Cohen syndrome (Keller *et al.*, 1999, Hennies *et al.*, 2004, Antonius *et al.*, 2008).

1.1.4.5.2 Autosomal recessive primary microcephaly (1° MCPH)

In 1° MCPH there is an isolated reduction in brain volume of prenatal onset to around a third of its normal size comparable to that observed in early humans (Woods *et al.*, 2005). Although the brain is small, often with a simplified gyral pattern, its structure remains otherwise normal (Mahmood *et al.*, 2011). Patients have a variable degree of intellectual disability but usually no neurological deficit. This is very similar to many patients with MPD however height is either within normal range or mildly reduced. The discovery of overlapping genetic aetiologies in MPD and 1° MCPH (see Section 1.3) has led to the hypothesis that these two disorders represent opposing ends of the same clinical spectrum (Klingseisen *et al.*, 2011).

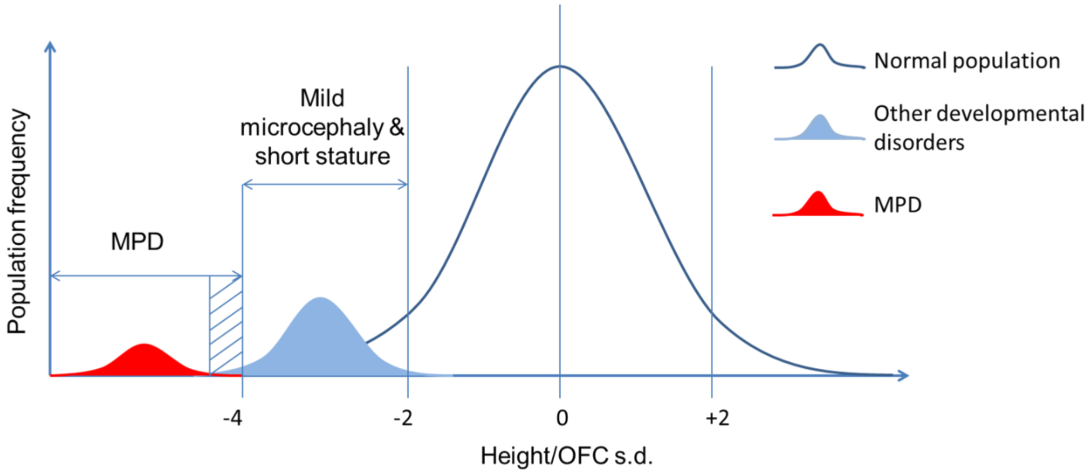


Figure 1.3. Graphical demonstration of overlapping growth distributions in MPD and other developmental disorders

Growth within a population is normally distributed. MPD is defined as height and OFC below -4 s.d. and therefore growth distribution in patients is below this threshold (red). However, the growth distribution for many other developmental disorders also lies below that of the normal population (blue) but is not reduced to the degree seen in MPD. Although mean height and OFC in many such disorders does not fall below -4 s.d., patients at the extreme ends of the corresponding growth spectrum can fall into the defined range for MPD (hashed region).

1.2 Mechanisms governing body size

For many years, there has been much interest in how organism size is determined although little is still known as to how this is so accurately regulated. For example, how do different organisms attain reproducible sizes whilst ensuring body proportions are consistently maintained? Body size is a reflection of total cell mass which in turn is determined by the number of cells and their individual size (Conlon *et al.*, 1999). A high degree of regulation of both cell growth and division must therefore exist in order to achieve a finite and consistent growth of different organs as well as the organism as a whole. Perturbation of any cellular mechanisms regulating cell proliferation or cell size can therefore impact on final organ/organism size.

Variation in body size between different mammals appears to be predominantly determined by cell number rather than cell size. This is illustrated by the 2800 fold difference in mass between a 70 kg human and a 25 g mouse which is largely accounted for by a similar fold difference in cell number with little variation in cell size (Conlon *et al.,* 1999). Cell size and cell number however, are intrinsically linked as proliferating cells are required to grow in mass prior to division in order to duplicate their content (Coelho *et al.,* 2000). In yeast, specific checkpoints through the cell cycle ensure that a minimum threshold in size is reached before the cell divides (Jorgensen *et al.*, 2004) and a similar mechanism may be in place in metazoans (Tzur *et al.*, 2009). Cells such as muscle and nerve cells can also continue to grow after they have ceased proliferating (Conlon *et al.*, 1999) but size regulation clearly exists at both organ and global levels ensuring growth does not exceed a set point. A balance must therefore occur between those factors which stimulate growth and promote cell survival and those which restrict final cell number and size.

The number of cells at any one point is a product of the rate at which cells are dividing minus the rate of cell death (Conlon *et al.*, 1999). Rate of cell division (or cell proliferation) is governed extrinsically by mitogens but also requires the cells intrinsic ability to undergo mitotic division and respond to both internal and external regulatory factors appropriately. A similar degree of complexity is also apparent in the control of cells undergoing programmed cell death (apoptosis) and those which become terminally differentiated having exited the cell cycle (Fuchs *et al.*, 2011).

1.2.1 Signalling pathways promoting cell growth and proliferation

1.2.1.1 Mitogen activated protein kinase (MAPK) pathways

Mitogens, which include growth factors, hormones, cytokines and environmental stresses, promote cell growth and proliferation by signalling through mitogen activated protein kinase (MAPK) pathways (Krishna *et al.,* 2008). Together these pathways coordinate the responses to a wide range of external and internal stimuli and can impact on many cellular processes including gene expression, cell metabolism, motility, apoptosis and differentiation. The mammalian MAPK extracellular regulated kinase 1 and 2 pathway (MAPK-ERK1/2) (Figure 1.4) has been specifically identified as a central regulator of cell proliferation by promoting the progression of the cell cycle from G1 phase to S phase through several mechanisms (Meloche *et al.*, 2007). Over 150 downstream phosphorylation targets of ERK have been identified (Krishna *et al.*, 2008). These include D-type cyclins which drives G1/S transition (Liu *et al.*, 1995, Sherr, 1995, Winston *et al.*, 1996) and the transcription factor c-myc which up regulates the expression of several genes such as translation initiation factors, ribosomal proteins and cyclins (Bouchard *et al.*, 1999, Hermeking *et al.*, 2000, Sears *et al.*, 2000, Adhikary *et al.*, 2005) promoting cell growth and proliferation. ERK activation also decreases expression of antiproliferative genes (Yamamoto *et al.*, 2006) and possibly the cyclin dependent kinase inhibitor p27 (Meloche *et al.*, 2007) further contributing to cell cycle progression.

1.2.1.2 Insulin-like growth factor (IGF) signalling

The IGF system is critical in growth both *in utero* and throughout childhood. Two insulin-like growth factors exist, IGF-I and IGF-II, which are secreted in an autocrine or paracrine fashion from most tissues during fetal life (Han *et al.*, 1988) in response to nutritional factors (Hietakangas *et al.*, 2009). In the postnatal period, production is predominantly via the liver (Yakar *et al.*, 1999) and largely under the control of growth hormone (GH) (Roberts *et al.*, 1987). Both IGF-I and -II activate Phosphotidyl-inosine-3 kinase (PI3K) signalling via the type 1 IGF receptor (IGF1R) present on all cell types (Siddle, 2011). Downstream targets include the master kinase target of rapamycin (TOR) which also integrates with other signalling pathways to modulate cell growth (Thomas *et al.*, 1997). Consequences of TOR activation include increased protein synthesis by initiating translation as well as preventing protein degradation and promoting ribosome biogenesis.

Disruption of the IGF system therefore results in a significant reduction in growth. Deficiency of growth hormone is characterised by a postnatal failure in growth whereas deficiency of IGF1 or IGF-1R results in both pre and post-natal growth failure in mice (Liu *et al.*, 1993) and humans (Walenkamp *et al.*, 2013) similar to

MPD. Investigation of IGF function in MPD patients has been advocated as growth hormone replacement therapy can improve growth in this patient group (Woods *et al.*, 2000). Interestingly, IGF-2 deficiency also results in growth failure of pre-natal onset but with relative preservation of head size (Russell-Silver syndrome) (Netchine *et al.*, 2007) possibly reflecting a tissue specific role during fetal development (Han *et al.*, 1988).

Although described separately a large degree of cross talk occurs between these signalling pathways (Mendoza *et al.*, 2011). IGF1R signalling can also activate MAPK pathways (King *et al.*, 1997) whereas ERK1/2 can activate TOR through phosphorylation (Carriere *et al.*, 2008).

1.2.2 Signalling pathways negatively regulating proliferation

In mammals, the Hippo pathway is the least well characterised of the three major signalling pathways but plays an important role in the control of organ size through the modulation of cell proliferation, migration, differentiation and apoptosis (Pan, 2007). Dysregulation of the Hippo pathway results in abnormal growth with organ overgrowth and tumorigenesis (Dong *et al.*, 2007). Core components of the Hippo pathway include the STE20 protein family kinases, Mst-1 and -2 (Harvey *et al.*, 2003) which activate the nuclear Dbf2-related (NDR) family protein kinases, Lats -1 and -2 (Justice *et al.*, 1995, Xu *et al.*, 1995, Wu *et al.*, 2003). The major downstream effectors of the Hippo pathway are the transcriptional coactivators YAP (Yes-associated protein) and its paralog TAZ (Lei *et al.*, 2008) which are inhibited through their phosphorylation by Lats kinase (Dong *et al.*, 2007). This prevents their accumulation in the nucleus where they associate with the TEAD family of DNA binding proteins inducing gene expression and promoting cell proliferation and survival (Ota *et al.*, 2008, Zhao *et al.*, 2008). In *Drosophila*, two proteins whose levels are decreased by Hippo signalling include cyclin E and DIAP1 (Tapon *et al.*, 2002) which negatively regulate cell cycle exit (Knoblich *et al.*, 1994) and apoptosis (Wang *et al.*, 1999) respectively. Reduced expression of bantam mRNA, another regulator of proliferation and apoptosis, has also been shown in *Drosophila melanogaster* (Peng *et al.*, 2009). Thus the net effect of the Hippo pathway following Mst and Lats kinase activation is suppression of growth by inhibiting proliferation and promoting apoptosis and differentiation.

Upstream regulators of the Hippo pathway so far identified include the cell surface protocadherins, Fat and Dachsous (Bennett *et al.*, 2006) and other apical membrane proteins including Crumbs (Chen *et al.*, 2010), Merlin, Expanded and Kibra (Genevet *et al.*, 2011) indicating that cell to cell interactions and cell polarity are important in governing this pathway. YAP/TAZ activity is also influenced directly by changes in mechanical tension through the F-actin cytoskeleton (Aragona *et al.*, 2013) and recent evidence has also shown Lats activity can either be activated or suppressed in response to different extracellular signals through G-protein coupled receptors (Yu *et al.*, 2012). In addition, the Hippo pathway can impact on the regulation of other key signalling pathways controlling tissue morphogenesis including Wnt, TGF-β, Sonic Hedgehog (Shh) and Notch pathways (Alarcon *et al.*, 2009, Varelas *et al.*, 2010, Zhao *et al.*, 2010, Heallen *et al.*, 2011). Changes in levels of morphogens which regulate these pathways governing the patterning of tissue during development can have large effects on organ size demonstrating such signalling is also important in controlling growth (Leevers *et al.*, 2005).

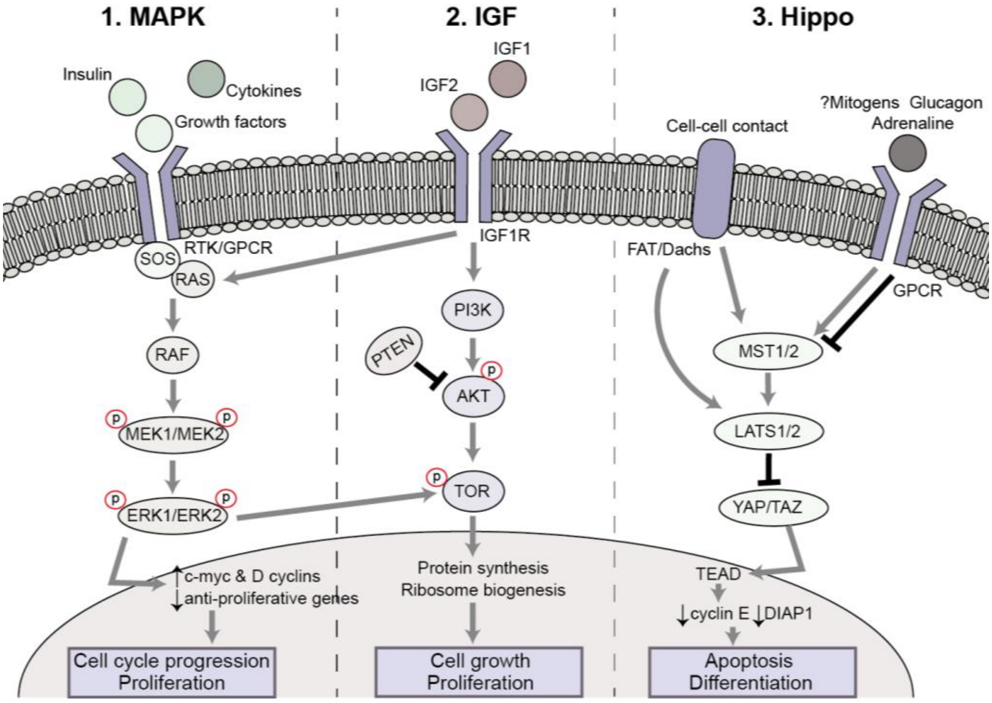


Figure 1.4. Signalling pathways regulating growth

1) Mitogens stimulate proliferation through the MAPK-ERK1/2 pathway. Activation of specific cell surface receptors (Goldsmith *et al.*, 2007, McKay *et al.*, 2007) recruit the nucleotide exchange factor, SOS (son of sevenless), activating the GTPase, Ras (Thomas *et al.*, 1992, Margarit *et al.*, 2003). The subsequent Raf activation phosphorylates MEK1/2 (Kyriakis *et al.*, 1992) which in turn phosphorylates ERK1/2 (Robinson *et al.*, 1996) promoting cell cycle progression through a variety of substrates including c-myc and D-cyclins. 2) IGF signalling activates PI3K which results in the sequential phosphorylation of AKT and TOR (Burgering *et al.*, 1995, Scott *et al.*, 1998). TOR activation has multiple downstream effects culminating in increased cell growth and proliferation. PTEN, whose dysfunction is associated with overgrowth, inhibits AKT activation thus suppressing growth (Mester *et al.*, 2013). 3) Activation of the Hippo pathway suppresses growth through inhibition of the transcriptional regulator YAP/TAZ (see Section 1.2.2 for details). As well as cell to cell contact, MST1/2 activity can also be modified by signalling induced by various stimuli through G protein coupled receptors. *Abbreviations: RTK=receptor tyrosine kinase, GPCR=G protein coupled receptor, Dachs=Dachsous.*

1.2.3 ‘Hypocellular’ hypothesis of MPD

It has been hypothesised that MPD is a condition of hypocellularity in which there is an overall reduction in cell number resulting in a smaller, but still proportionate person (Delaval *et al.*, 2008, Rauch *et al.*, 2008, Klingseisen *et al.*, 2011). In support of this hypothesis, the majority of disease genes so far identified in MPD have clear roles in cell cycle regulation (Section 1.3). Typically in affected patients neither increased nutritional support nor growth hormone administration have any impact on final body mass (Rauch, 2011) suggesting that cell growth is unable to be pushed beyond a particular threshold. Small changes in the rate of cell division can have a profound impact on the total cell number, for example, a cell which undergoes five rather than seven rounds of division will result in a four fold difference in final cell number (Klingseisen *et al.*, 2011) (Figure 1.5). This difference is then magnified with increasing rounds of division. The impact of altered proliferation and apoptosis on size has been demonstrated experimentally in embryonic mouse brains where inhibiting cell death or increasing neural progenitor proliferation both resulted in a dramatic increase in brain size (Kuida *et al.*, 1998, Chenn *et al.*, 2002). 1° MCPH is similarly thought to be due to reduced cell number arising from reduced expansion of the neural progenitor pool, either through premature neuronal differentiation or increased cell death, resulting in an isolated reduction in brain size (Mochida *et al.*, 2001).

1.2.4 Summary

Growth is controlled by a complex interplay of numerous pathways and mechanisms. Abnormalities in growth regulation are critical to many disease processes, most notably malignancy, therefore understanding which biological mechanisms govern body size and how is important to many areas of research. In MPD both pre- and post-natal growth is globally reduced as the result of a single gene defect. Identifying the molecular cause and the resulting impact on cellular processes in affected cases will give further insights into mechanisms critical in ensuring normal growth from early embryogenesis. Thus MPD provides us with a useful human model for investigating growth. The next Section details the disease causing genes so far identified in MPD and their associated function.

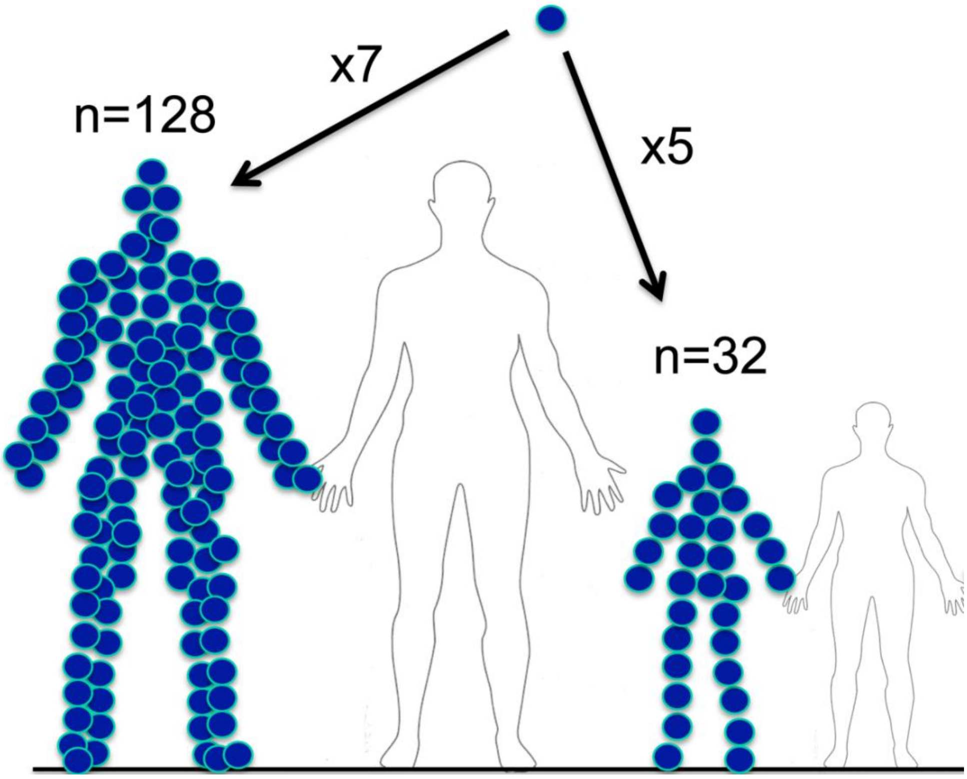


Figure 1.5. Impact on cell number by reduction in rate of cell division

In mammals, difference in body size is largely accounted for by a difference in cell number rather than cell size. Small differences in the number of cell divisions can have large impacts on final cell number and consequently body size. Reproduced from Klingseisen *et al.*, (2011).

1.3 Cellular mechanisms identified in MPD

Several genes have now been identified in MPD which act downstream of the pathways discussed in Section 1.2 and encode proteins integral to the cell cycle machinery. This includes centrosomal proteins with roles in the faithful segregation of chromosomes during mitosis (*CEP152, CENPJ* and *PCNT*) as well as those involved in DNA damage response and repair (*ATR* and *RBBP8*) which regulate cell cycle progression in order to maintain genomic integrity. Others are required for DNA replication (*ORC1, ORC4, ORC6, CDC6* and *CDT1*) and mRNA splicing (*RNU4ATAC*) although it is still currently unclear exactly how the latter impacts on cell cycle dynamics.

1.3.1 Centrosomal proteins

1.3.1.1 Centrosome structure and replication

The centrosome is a key organiser of the microtubule cytoskeleton integral to cell motility, adhesion, polarity and organelle transport as well as mitotic spindle poles (Badano *et al.*, 2005). Centrosomes are composed of two centrioles surrounded by a matrix of pericentriolar material (PCM) (Figure 1.6A) (Paintrand *et al.*, 1992). The two centrioles each consist of a barrel of nine microtubule triplets and are linked by fibres with one daughter (younger) centriole sitting in perpendicular orientation to the mother centriole. The older, mother centriole has subdistal and distal appendages to which microtubules are anchored (Piel *et al.*, 2000). During each cell cycle, the centrosome is replicated through the duplication of the centrioles, a process that only occurs once and is intricately linked to cell cycle progression (Figure 1.6B) (Robbins *et al.*, 1968). The process of centriole duplication involves four stages (Azimzadeh *et al.*, 2010). First the two centrioles disengage at the end of mitosis followed by nucleation of new daughter centrioles proximal to each mother centriole in G1-S phase. The daughter centrioles then elongate and mature in S-G2 phase before the two new centriole pairs separate. Each pair therefore contains an older ‘mother’ and younger ‘daughter’ centriole. The youngest set of centrioles then moves to the basal aspect of the cell where it gathers pericentriolar material forming a second centrosome (Rebollo *et al.*, 2007).

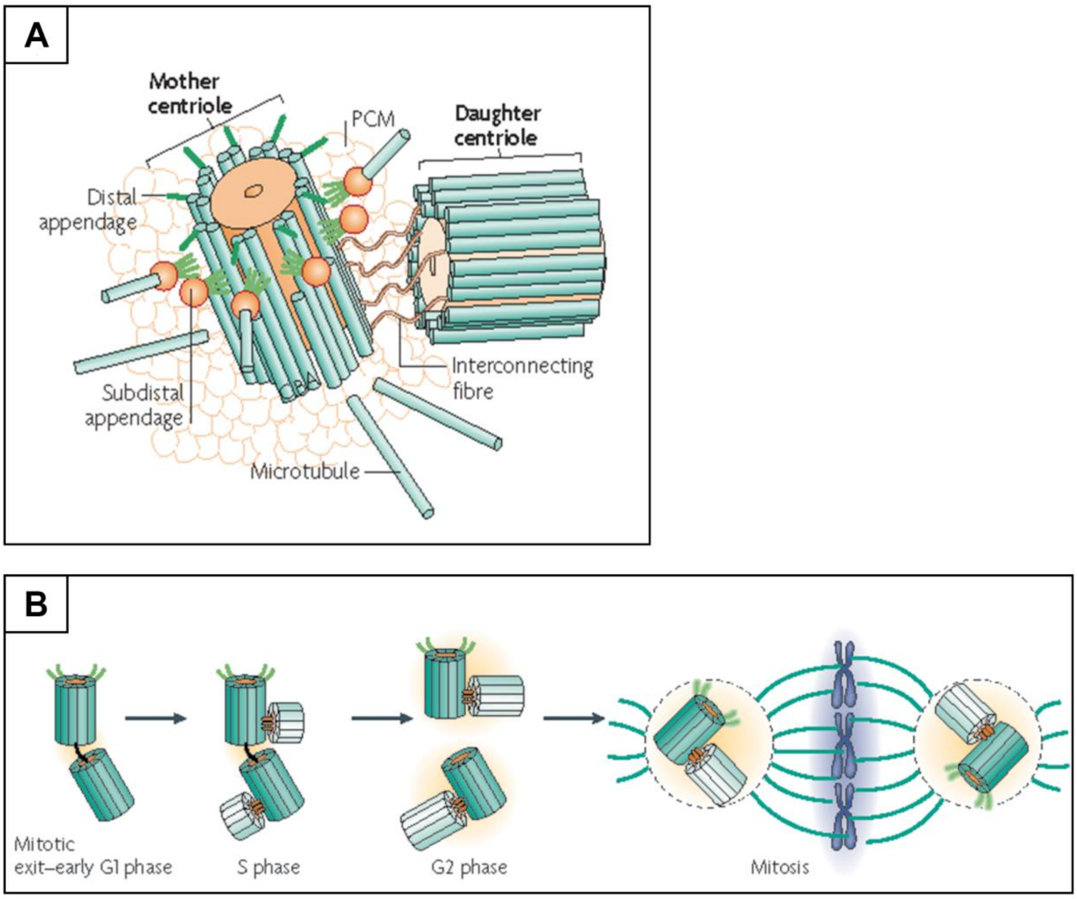


Figure 1.6. Centrosome structure and replication

A) Each centrosome consists of two centrioles composed of nine microtubule repeats. Subdistal and distal appendages exist on the mother centriole through which microtubules are attached. The centrioles are attached through interlinking fibres and surrounded by the pericentriolar matrix (PCM). B) Each cell contains one centrosome which is replicated once each cell cycle. Initially the two centrioles (dark green) disengage on mitotic exit, followed by nucleation of two daughter centrioles (light green) in G1 and S phase. The new procentrioles are then elongated in S and G2 phase followed by separation of the two pairs to opposite ends of the cell and formation of the mitotic spindle apparatus. Reproduced from Bettencourt-Dias *et al.*, (2007).

1.3.1.2 Centrosome function in the cell cycle

During interphase, proteins in the pericentriolar material nucleate a large number of surrounding microtubules which contribute to the formation of the mitotic spindle apparatus (Luders *et al.*, 2007, O'Connell *et al.*, 2007). This apparatus consists of a system of microtubules attached to the centrosome at one end and the kinetochore, a protein structure located at each centromere of centrally aligned chromosomes, at the other. This enables the separation of sister chromatids by forces generated through these microtubules which pull the chromosomes to opposite ends of the cell prior to cytokinesis. Centrosomes therefore contribute to the formation, organisation and subsequent orientation of the mitotic spindle apparatus (Rebollo *et al.*, 2007, Rusan *et al.*, 2007). The latter determines the plane of cleavage through which the cell divides (cytokinesis) which contributes to cell fate determination in certain stem cell populations (Yamashita *et al.*, 2007). Although microtubule arrays can form in the absence of centrosomes and cell division can still be completed (Khodjakov *et al.*, 2001), disrupting centrosome replication has been shown to result in aberrant cell divisions. In *Drosophila*, failure in centriole duplication results in slower mitotic spindle assembly and prolonged mitosis as well as abnormal cytokinesis (Basto *et al.*, 2006) whereas centrosome amplification (more than two centrosomes at mitosis) results in the formation of multipolar spindles and genome instability (Ko *et al.*, 2005, Loncarek *et al.*, 2008, Li *et al.*, 2014a).

Experiments in which the centrosome has been excised from the cell altogether results in G1 arrest (Hinchcliffe *et al.*, 2001, Khodjakov *et al.*, 2001) indicating centrosomes also play a role in cell cycle progression. In fact, centrosomes act as a platform for over 100 regulatory proteins (Doxsey *et al.*, 2005) including cyclin E which is required for G1 to S phase transition (Matsumoto *et al.*, 2004) and possibly CHK1, a cell cycle checkpoint kinase which inhibits G2 to M transition in response to DNA damage signalling (Kramer *et al.*, 2004, Loffler *et al.*, 2007). Thus centrosomes are a multifunctional unit ensuring mitotic fidelity.

1.3.1.3 PCNT in MOPDII

Biallelic truncating mutations in Pericentrin (*PCNT*) have so far been identified as the sole cause of MOPDII with large numbers of patients now reported (Griffith *et al.*, 2008, Rauch *et al.*, 2008, Willems *et al.*, 2010). PCNT is a well conserved large, coiled coil protein and a core component of the PCM (Flory *et al.*, 2000). It acts as a multifunctional scaffold protein anchoring other important centrosomal proteins including the microtubule nucleating protein γ-tubulin (Doxsey *et al.*, 1994, Dictenberg *et al.*, 1998, Zimmerman *et al.*, 2004). Loss of PCNT therefore results in the depletion of such proteins from the centrosome and consequently defects in the mitotic spindle apparatus. In keeping with this role, patient cells lacking PCNT show reduced γ-tubulin at spindle poles, disorganised mitotic spindles, chromosome misalignment and abnormal cytokinesis (Figure 1.7) (Griffith *et al.*, 2008, Rauch *et al.*, 2008). Additionally, premature sister chromatid separation was observed suggesting PCNT also contributes to the spindle assembly checkpoint (SAC) which ensures chromosomes are correctly attached to the mitotic microtubules in metaphase prior to segregation. Such mitotic failure is likely to result in reduced cell proliferation and/or increased cell death leading to a reduction in overall cell number (Delaval *et al.*, 2010).

PCNT, in combination with microcephalin (MCPH1), is also thought to recruit the cell cycle checkpoint kinase, CHK1 to the centrosome (Tibelius *et al.*, 2009). RNAi depletion of PCNT in cells results in loss of CHK1 from the centrosome, premature mitotic entry, delayed mitosis and increased cell death. In keeping with a role in CHK1-mediated cell cycle regulation, PCNT-deficient patient cells also show defective G2/M checkpoint arrest in response to UV-induced DNA damage and therefore ATR-dependent signalling may also contribute to growth failure (Griffith *et al.*, 2008, Tibelius *et al.*, 2009).

Although aberrant mitosis and impaired DNA damage signalling may well be responsible for the global growth failure seen in MOPDII patients, mechanisms underlying the other associated features in this syndrome (insulin resistance, cerebrovascular malformations) are less clear and may indicate additional cellular roles for PCNT (Jurczyk *et al.*, 2010).

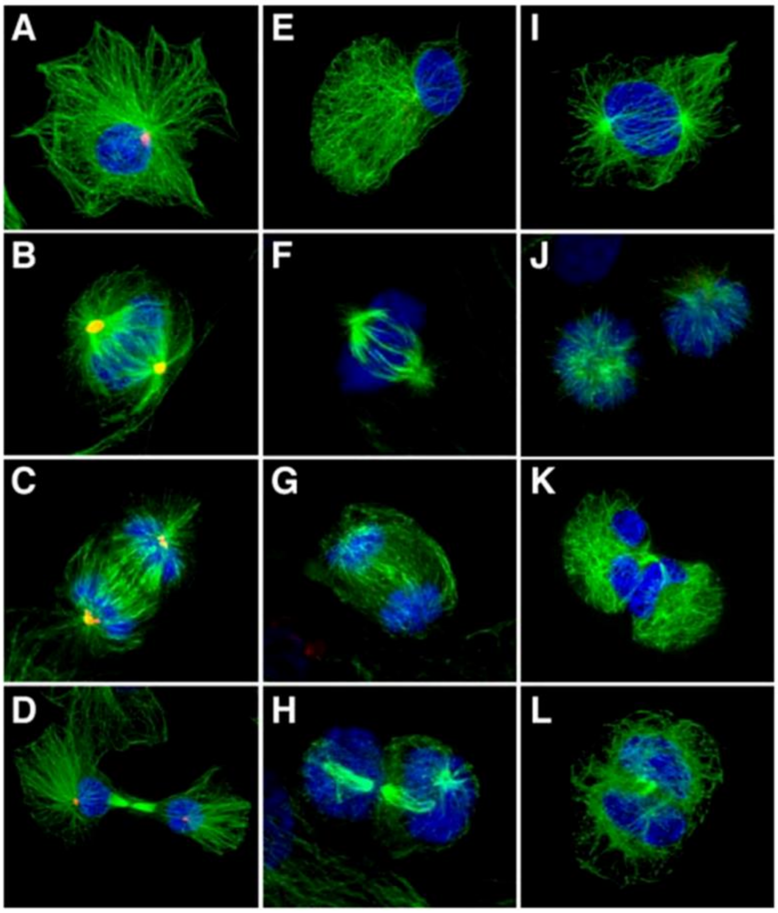


Figure 1.7. Mitotic abnormalities in patient fibroblasts lacking PCNT

Immunofluorescence images of mitotic cells. PCNT stained red, microtubules green and chromosomes blue. Mitotic cell from a healthy individual shown in A) interphase, B) metaphase, C) anaphase and D) during cytokinesis. In comparison, E-L represents images of MOPDII patient cells in which there is no detectable PCNT at the centrosomes. Patient cells displayed disorganised mitotic microtubules (I, F, J and G), abnormal alignment in metaphase (J) and disorganised cytokinesis (H, K and L). Reproduced from Rauch *et al.*, (2008).

1.3.1.4 Centriole duplication proteins in Seckel syndrome

Mutations in two genes encoding proteins essential to centriole duplication have been described in Seckel syndrome patients; Centrosomal protein of 152kDa (*CEP152)* and centromere protein J (*CENPJ)* (Al-Dosari *et al.*, 2010, Kalay *et al.*, 2011). Mutations in both genes have also been identified in patients with 1° MCPH (Bond *et al.*, 2005, Guernsey *et al.*, 2010). To date few mutations have been identified in

CEP152 in MPD including a founder intronic variant (c.261+1G>C) within several families of Turkish ancestry which disrupts transcript splicing (Kalay *et al.,* 2011). Patient cells deficient in CEP152 displayed delayed cell cycle progression and numerous mitotic abnormalities including multiple and fragmented centrosomes, incorrectly aligned chromosomes, premature separation of sister chromatids and monopolar spindles (Kalay *et al.*, 2011) . Only one mutation in *CENPJ* has been reported to date in association with Seckel syndrome, c.3302-1G>C (Al-Dosari *et al.*, 2010). This was identified in multiple affected individuals from a Middle-Eastern family with anthropometric measurement of -7 s.d. or less in all cases. The mutation was shown to alter mRNA splicing although no further characterisation of patient cells was performed.

CEP152 acts as a centrosomal scaffold for key regulators of centriole biogenesis including the polo-like kinase 4 protein (PLK4) and CENPJ (Kim *et al.*, 2013). Recently, mutations in *PLK4* have been identified in patients with MPD and retinopathy (Martin *et al.*, 2014). The authors also observed a similar phenotype in several patients with mutations in the tubulin, gamma complex associated protein 6 (*TUBGCP6*) which has previously been associated with microcephaly and retinopathy (Puffenberger *et al.*, 2012). TUBGCP6 is phosphorylated by PLK4 and is required for assembly of the γ-tubulin ring complex which nucleates microtubules in centriole biogenesis and formation of the mitotic spindle (Bahtz *et al.*, 2012).

1.3.1.5 Other mitotic spindle proteins implicated in MPD

Biallelic nonsynonymous mutations have recently been reported in *CENPE* (centromere-associated protein E) in a patient with MPD also exhibiting severe developmental delay, neurological abnormalities and congenital restrictive cardiomyopathy (OFC -9 s.d., height -7s.d.) (Mirzaa *et al.*, 2014). A sibling was similarly affected however height was less severely reduced (-3 s.d.). *CENPE* encodes a core component of the kinetochore, stabilizing the attachment of chromosomes to the mitotic microtubules (Yao *et al.,* 2000). Failure in attachment of the kinetochore to microtubules results in activation of the SAC preventing mitotic progression until all kinetochores are attached (Cleveland *et al.*, 2003). Patient cells showed mitotic spindle defects and impaired mitotic progression similar to those observed in PCNT-negative cells (Mirzaa *et al.*, 2014). Defects in other kinetochore proteins have also previously been associated with growth failure including the SAC kinase BUBR1 resulting in Mosaic Variegated Aneuploidy (MVA) characterised by growth retardation and cancer predisposition (Hanks *et al.*, 2004). Mutations in the cancer susceptibility candidate 5, *CASC5,* encoding another member of the kinetochore complex have also been identified in 1° MCPH patients (Genin *et al.*, 2012).

1.3.1.6 Centrosomal proteins identified in 1° MCPH and stem cell fate

In total nine genetic loci encoding centrosomal proteins have now been identified in 1° MCPH (Table 1.1); Microcephalin (*MCPH1*)*, ASPM, CDK5RAP2, CENPJ, STIL, CEP152, WDR62, CEP63* and *CEP135* (Bond *et al.*, 2002, Jackson *et al.*, 2002, Bond *et al.*, 2005, Kumar *et al.*, 2009, Bilguvar *et al.*, 2010, Guernsey *et al.*, 2010, Yu *et al.*, 2010, Sir *et al.*, 2011, Hussain *et al.*, 2012). All are expressed in the neuroepithelium during embryonic neurogenesis and play key roles in neurogenic mitosis (Megraw *et al.*, 2011) either through ensuring correct cell cycle check point signalling or through correct mitotic spindle organisation and orientation which determines cell fate following neural stem cell division. In early embryogenesis neural progenitors undergo symmetric cell division with the plane of division exactly perpendicular to the cell surface creating an adequate pool of stem cells from which to populate the developing cortex (Noctor *et al.*, 2004). Later in development, oblique cell cleavage of progenitor cells results in asymmetric cell division producing only one cell which continues to proliferate and another which is committed to become terminally differentiated as a neuron. In 1°MCPH it is postulated that mitotic spindle defects leads to an increase in early asymmetric divisions resulting in a reduction in the number of neural progenitors from which to populate the developing cortex. Disrupting asymmetric cell divisions has been shown to result in a depletion of progenitors in the proliferating region of the brain (ventricular zone, VZ) (Wang *et al.*, 2009).

As several genes are allelic for both MPD and 1° MCPH it has been hypothesised that MPD may arise from a similar cellular mechanism in which multiple stem cell pools are affected resulting in a more global growth failure (Delaval *et al.*, 2010). Interestingly, spindle misorientation has also been described in the neural precursors of PCNT deficient mice (Chen *et al.*, 2014). Compound heterozygous nonsynonymous mutations in *NIN* were reported in two siblings with MPD (height and OFC <-6 s.d.) (Dauber *et al.*, 2012). Ninein is another centrosomal protein which functions in the anchoring of the centrosome to microtubules (Shinohara *et al.*, 2013). Removal of ninein disrupts asymmetric cell division and leads to a reduction in neural progenitors in the VZ (Wang *et al.*, 2009). However, impaired ninein function in MPD patients as a direct consequence of *NIN* mutations has not been shown.

Table 1.1 provides a summary of all the centrosomal genes identified in MPD and 1° MCPH to date and their associated function (reviewed in Megraw *et al.*, (2011)). It is unclear, however, why some mutations result in MPD as opposed to 1° MCPH. The presence of other genetic modifiers in MPD patients or that some mutations only affect brain specific transcripts have been suggested (Al-Dosari *et al.*, 2010).

Table 1.1. Summary of genes encoding centrosomal proteins in MPD and 1° MCPH

Function	*Gene*	*Protein*	*Phenotype*	*OMIM*
PCM scaffold, nucleating MT, ATR-dependent cell cycle checkpoint activation	*PCNT*	Pericentrin	MOPDII	210720
Cell cycle checkpoint activation	*MCPH1*	Microcephalin	1° MCPH	251200
Centriole biogenesis	*CEP152*	Centrosomal protein of 152 kDa	1° MCPH & MPD	614852
	CENPJ	Centromeric protein J	1° MCPH & MPD	609393
	PLK4	Polo-like kinase 4	MPD and choriorentinopathy	605031
	TUBGCP6	Tubulin, gamma complex associated protein 6	MCPH/MPD and choriorentinopathy	251270
	STIL	SCL/TAL1 interrupting locus	1° MCPH	612703
	CEP63	Centrosomal protein of 63 kDa	1° MCPH	614728
	CEP135	Centrosomal protein of 135 kDa	1° MCPH	614673
Nucleating MT, spindle organisation & orientation	*ASPM*	Abnormal spindle-like microcephaly associated	1° MCPH	608716
	CDK5RAP2	Cyclin dependent kinase 5 regulatory subunit-associated protein 2	1° MCPH	604804
Mitotic spindle formation and orientation through JNK1 signalling	*WDR62*	WD-repeat containing protein 62	1° MCPH	604317

Abbreviations: PCM=pericentriolar matrix, MT=microtubules, JNK= c-Jun N-terminal kinase.

1.3.2 DNA damage response and repair proteins

DNA is under constant attack by both endogenous and exogenous reactive molecules with approximately 10^5 DNA lesions occurring per mammalian genome per day (Hoeijmakers, 2009). As well as promoting cell death and threatening organism viability, incorrect or ineffective repair can introduce genomic instability and predispose to cancer development with the introduction of mutations or larger structural rearrangements (Bender *et al.*, 1974, Frank *et al.*, 2000). Maintaining genomic integrity through effective DNA repair mechanisms is therefore integral to cell cycle control and survival.

1.3.2.1 Mechanisms of DNA damage

DNA damage can occur endogenously during physiological processes such as incorrectly incorporated nucleotides during DNA replication (DNA mismatch) as well as following abortive topoisomerase I and II activity resulting in single or double DNA strand breaks (Wang, 2002). Reactive oxygen compounds (ROS) can also induce DNA base lesions and are a by-product of several cellular processes including oxidative respiration and immune activation in response to inflammation and infection (Finkel *et al.*, 2000). In addition, exogenous agents capable of inducing DNA damage include ionising radiation, ultraviolet radiation, chemicals found in tobacco products and medication used in the treatment of cancer (Doll *et al.*, 1981, Ward, 1988, Espinosa *et al.*, 2003). Such extrinsic damage can block or impair DNA replication and transcription, resulting in nucleotide loss or induce single- or double-strand DNA breaks (SSB, DSB).

Of all possible DNA damage lesions that can occur, DSBs are the most deleterious and it is estimated that 10-20 DSBs occur in each cell every day (Martin *et al.*, 1985). The majority are pathologic occurring during replicative stress, mechanical stress, exposure to ionising radiation, ROS or the inadvertent action of nuclear enzymes (McClintock, 1941, Lea, 1946, Karanjawala *et al.*, 2002, Adachi *et al.*, 2003, Mahowald *et al.*, 2008, Ozeri-Galai *et al.*, 2011). Unrepaired SSBs can also become DSBs when encountered during replication (Kuzminov, 2001). Repair is important as DSBs promote cell death and impede proliferation (Frank *et al.*, 2000) whilst defective repair can lead to genomic rearrangements (Bender *et al.*, 1974) and telomere instability (Bailey *et al.*, 1999) which can promote the development of malignancy and contribute to ageing (Chang *et al.*, 2001, Schuler *et al.*, 2013).

Although DSBs can clearly be deleterious to a somatic cell they also have an important role in ensuring genetic diversity. Firstly, in the adaptive immune system by enabling variable (diversity) joining (V(D)J) recombination (Tonegawa, 1983, Taccioli *et al.*, 1993) and class switch recombination (CSR) (Casellas *et al.*, 1998, Pan-Hammarstrom *et al.*, 2005) in B and T lymphocyte maturation. Secondly, in germ cells DSBs are required to generate cross overs between homologous chromosomes during meiosis (Zickler *et al.*, 1999).

1.3.2.2 Mechanisms of DNA damage repair

Different repair mechanisms are present in the cell to deal with the diverse range of possible lesions and cells defective in certain repair mechanisms show increased sensitivity towards specific DNA damaging agents (Table 1.2).

1.3.2.2.1 Repair of aberrant nucleotides

Mismatch repair (MMR) machinery recognises and repairs incorrectly incorporated nucleotides during replication as well as DNA loops which have occurred during slippage of the replication machinery at repeat sequence motifs know as microsatellites (Strand *et al.*, 1993). The process involves endonucleolytic degradation of the error-containing strand to the site of the mismatch with removal of the incorrect base(s) followed by DNA resynthesis and ligation of the remaining nick (Jiricny, 2006). Damaged bases such as those induced by ROS can be corrected by base-excision repair (BER) in which the defect is recognised by a DNA glycosylase enzyme, followed by removal, infilling and ligation (David *et al.*, 2007). Base-lesions such as UV-induced pyrimidine photodimers which distort the DNA helix are removed by the nucleotide excision repair (NER) machinery which can also target lesions specifically blocking transcription (transcription-coupled NER). NER results in the excision of a 22-30 bp oligonucleotide removing the obstructive lesion and generating single-stranded DNA which is then recognised and repaired accordingly (Melis *et al.*, 2013).

1.3.2.2.2 Single-strand break (SSB) repair

SSBs occur as a direct consequence of damage by agents such as endogenous ROS and UV radiation as well as indirectly following nucleotide resection by MMR, BER or NER (Caldecott, 2008). SSBs also arise from abortive TOPI activity and decoupling of the replication machinery in stalled replication forks (Wang, 2002). SSBs are initially detected by poly (ADP-ribose) polymerase 1, PARP1 (D'Amours *et al.*, 1999). This is followed by the recruitment and activation of various end processing factors depending on the damage incurred. Such factors include X-ray repair cross-complementing protein 1 (XRCC1), polynucleotide kinase 3’-phosphatase (PNKP), tyrosyl-DNA phosphodiesterase 1 (TDP1) and Aprataxin (APTX) to resect and repair the damaged ends as well as polymerases to infill any remaining gap (Jilani *et al.*, 1999, Pouliot *et al.*, 1999, Whitehouse *et al.*, 2001, Liu *et al.*, 2005, Ahel *et al.*, 2006). This ensures the two ends are then compatible for final ligation by DNA ligases (LIG1 or LIG3) (Timson *et al.*, 2000).

1.3.2.2.3 Double-strand break (DSB) repair

In eukaryotic cells, there are two main DSB repair pathways; non-homologous end joining (NHEJ) and homologous recombination (HR). In NHEJ the broken ends of the DSB are ligated back together regardless of whether the DNA sequence remains intact (Roth *et al.*, 1985). This can be beneficial in promoting immune diversification in lymphocytes (Gerstein *et al.*, 1993) but can also introduce deleterious mutations and lead to genome instability with the generation of chromosomal rearrangements (Bender *et al.*, 1974). HR ensures the genomic sequence is preserved by using the alternative sister chromatid as a template for repair (Howard-Flanders, 1975). Due to their different mechanisms NHEJ can be active at any point in the cell cycle whereas HR is restricted to the late S and G2 phases of the cell cycle when there is a sister chromatid within close proximity (Rothkamm *et al.*, 2003). HR is therefore largely responsible for the repair of DSBs occurring during replication as well as those formed in meiosis whereas NHEJ has more impact on the repair of DSBs occurring as a consequence of DNA damage in G1-phase and non-cycling cells.

Although individual repair mechanisms have been described distinctly, marked overlap exists between these different processes. For example, HR and NER involves the generation of single-stranded DNA which then recruits factors required for SSB repair (Melis *et al.*, 2013). SSBs are also generated following the removal and correction of base lesions in BER and MMR and require components of the SSB repair machinery (Jiricny, 2006, David *et al.*, 2007). Some proteins also have roles in more than one mechanism, for example PNKP and APTX are often required for SSB repair but can also be utilised for DNA end-processing in NHEJ (Clements *et al.*, 2004, Koch *et al.*, 2004).

1.3.2.2.4 DNA repair mechanisms and human disease

Defects in many of these mechanisms are associated with human disease including cancer predisposition, neurodegeneration and developmental disorders in which growth is often affected (Table 1.2). The latter is illustrated by disorders such as Cockayne syndrome (impaired transcription-coupled NER) and Bloom syndrome (impaired HR), both associated with severe growth impairment which can fall into the defined range for MPD (Nance *et al.*, 1992, Keller *et al.*, 1999). However, the prominence of other features such as premature aging, immunodeficiency and cancer predisposition often distinguishes these syndromes from other MPD disorders.

Table 1.2. DNA damage repair mechanisms and associated disease

DNA repair mechanism	*Type of lesion repaired*	*Associated disorder*	*OMIM*	*Genes mutated*	*Associated phenotype*	*Cellular phenotype*
1 Mismatch repair (MMR)	Replication errors: DNA mismatches & slippage of replication machinery	Hereditary non-polyposis colon cancer	120435	*MSH2*, *MLH1*, *MSH6*, *PMS2*, *PMS1*, *TGFBR2*, *MLH3*	Cancer predisposition (Gastro-intestinal and reproductive systems)	Micro-satellite instability
2 Base excision repair (BER)	Abnormal DNA bases arising from damage due to cellular metabolism	Autosomal recessive adenomatous polyposis	132600	*MUTYH*	Colorectal polyps, cancer predisposition	Impaired 8-hydroxyguanine removal, increased mutagenesis
3 Nucleotide excision repair: Global genome (GG-NER) & transcription coupled (TC-NER)	Base lesions distorting DNA double helix interfering with transcription or replication	Cockayne Syndrome/COFS syndrome	216400 214150	*ERCC6*, *ERCC8* (TC-NER)	**MPCH, growth delay**, cataracts, arthrogryposis, photosensitivity, hearing loss, severe developmental delay, premature ageing	↑ sensitivity to UV
		Xeroderma pigmentosa	278760 278780	*ERCC4* *ERCC5*	Acute sun sensitivity, cancer predisposition (skin)	
		Trichothiodystrophy	601675	*ERCC2* *ERCC3*	Brittle hair, **growth delay**, congenital icthyosis, photosensitivity	
4 Single-strand break (SSB) repair	SSB: oxidative damage, DNA topoisomerase failure plus BER, MMR & NER/TCR induced SSB	Microcephaly with early onset, intractable seizures and developmental delay	613402	*PNKP*	**MCPH**, seizures, developmental delay	↑ sensitivity to camptothecin, UV & hydrogen peroxide
		Spinocerebellar ataxia with axonal neuropathy 1	607250	*TDP1*	Cerebellar atrophy, peripheral neuropathy	
		Ataxia oculomotor apraxia 1	208920	*APTX*	Cerebellar atrophy, ataxia, peripheral neuropathy, oculomotor apraxia	

5	Non-homologous end joining (NHEJ)	DSB	Ligase IV syndrome	606593	*LIG4*	SCID, **MCPH, growth delay**, cancer predisposition, pancytopenia	Extreme sensitivity to IR
			Radiosensitive-SCID	611291 615559 602450	*NHEJ1* *PRKCD* *DCLRE1C*	SCID +/- **growth delay**	
6	Homologous recombination (HR)	DSB	Nijmegen breakage syndrome (NBS)	251260	*NBN*	**MCPH, growth delay**, immune-deficiency, cancer predisposition	↑ sensitivity to IR, ↑spontaneous sister chromatid exchange
			NBS-like syndrome	613078	*RAD50, MRE11A*	**MCPH, growth delay**	
			Bloom syndrome	210900	*RECQL3*	**MCPH, growth delay**, photosensitivity, immunodeficiency, cancer predisposition	
			Werner syndrome	277700	*WRN*	Premature aging, cataracts, **growth delay**, skin abnormalities, cancer predisposition	
			Rothmund-Thomson, RAPADILINO & Baller-Gerold syndromes	268400 266280 218600	*RECQL4*	Poikiloderma, **growth delay**, sparse hair, radial ray defects, cancer predisposition, cataracts.	
			Familial breast cancer	604370 612555	*BRCA1, BRCA2,*	Cancer predisposition	
7	Fanconi anaemia (FA) pathway	Inter-strand DNA crosslinks (ICL)	Fanconi Anaemia	227650	*FANCA, FANCC, FANCG* most common (15 genes in total)	Progressive bone marrow failure, skeletal abnormalities, congenital malformations, **growth delay**, cancer predisposition	↑ sensitivity to ICL inducing agents (cisplatin, mitomycin C)

Mechanisms reviewed in the following articles; 1) Jiricny, (2006), 2) David *et al.*, (2007), 3) Melis *et al.*, (2013), 4) Caldecott, (2008), 5) Lieber, (2008) , 6) San Filippo *et al.*, (2008), 7) Su *et al.*, (2011). Disease associations reviewed in the following articles: O'Driscoll, (2012), Suhasini *et al.*, (2013). *Abbreviations: COFS=cerebroocularfacioskeletal, SCID=severe combined immunodeficiency.*

1.3.2.3 Cellular responses to DNA damage

To ensure repair occurs effectively and efficiently, signalling pathways are present to respond to DNA damage. Two main DNA damage response (DDR) pathways exist in the cell, one initiated by ATR (Ataxia Telangiectasia and Rad3-related) and the other by ATM (Ataxia Telangiectasia Mutated), both members of the phosphatidylinositol-3-kinase-like kinase (PIKKs) family (Marechal *et al.*, 2013). Following initial sensing of DNA damage, ATM/ATR activation results in the phosphorylation of hundreds of downstream effector proteins, many of which are common to both pathways (Shiloh, 2006). As well as recruiting relevant repair proteins to lesions, DDR pathways activate cell cycle checkpoints. Additionally DDR pathways regulate cell fate by inducing apoptosis or senescence to prevent the propagation of DNA damage and maintain genomic integrity. Checkpoint arrest of the cell cycle can occur at G1-S phase, intra-S-phase and G2-M phase postponing cell cycle progression until the lesion(s) is repaired (Lukas *et al.*, 2004).

1.3.2.3.1 ATR-mediated DNA damage signalling

ATR is recruited and activated by replication protein A (RPA) which coats single stranded DNA through obligatory binding to the ATR-interacting protein, ATRIP (Zou *et al.*, 2003) (Figure 1.8). Important downstream targets of ATR activation include CHK1 (Liu *et al.*, 2000) and the tumour suppressor protein p53 which can induce apoptosis (Tibbetts *et al.*, 1999). ATR is essential for development and complete absence leads to early embryonic lethality in mice (Brown *et al.*, 2000). ATR also plays an important role in the regulation of DNA replication in response to DNA damage by altering the distribution of active replication origins (Ge *et al.*, 2010).

1.3.2.3.1 ATM-mediated DNA damage signalling

ATM is activated in response to DSBs (Figure 1.8) interacting with NBS1 of the MRN (MRE11-RAD50-NBS1) complex at the site of a DSB (You *et al.*, 2005). Downstream targets of ATM include the checkpoint kinase CHK2, p53 and BRCA1 mediating cell cycle arrest, apoptosis and HR respectively (Matsuoka *et al.*, 1998, Fernandes *et al.*, 2005, Cheng *et al.*, 2010). This differs from DSB repair by NHEJ in which DSBs are detected by the KU70-80 complex (Mimori *et al.*, 1986, Walker *et al.*, 2001) and repair is initiated by the recruitment and activation of another PIKK, DNA-dependent protein kinase catalytic subunit (DNA-PKcs) (Gottlieb *et al.*, 1993). In contrast to ATR, ATM is non-essential for development. Mutations in *ATM* cause ataxia-telangiectasia characterised by neurodegeneration, immunodeficiency and cancer predisposition (Savitsky *et al.*, 1995).

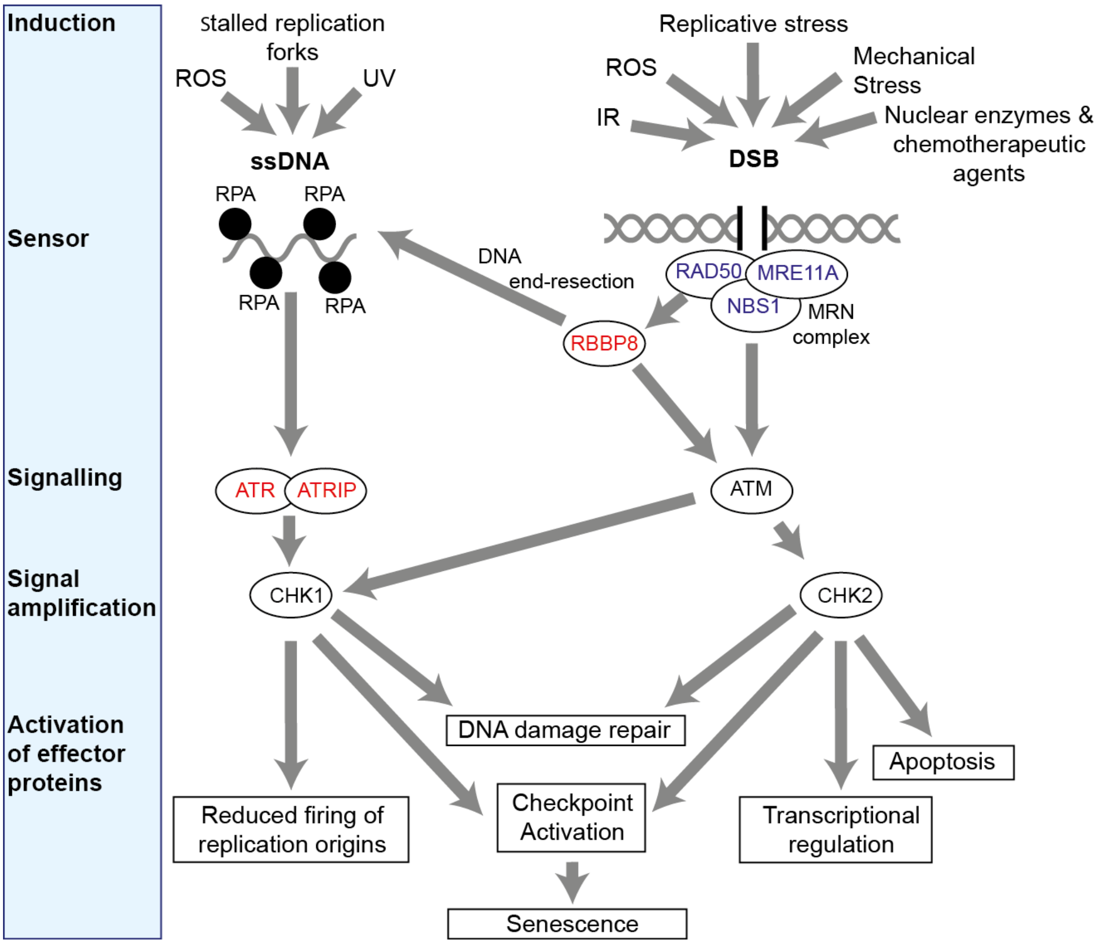


Figure 1.8. DNA damage response (DDR) pathways

ATR and ATM are both protein kinases central to the two main DDR signalling pathways. ATR is recruited to RPA-coated single-stranded DNA in association with ATRIP whereas ATM is activated by the MRN complex at sites of DSBs. Both activate downstream checkpoint kinases (CHK1 and CHK2 respectively) which phosphorylate a large number of downstream targets with a range of consequences all with the purpose of maintaining genomic integrity. Proteins encoded by genes mutated in MPD are highlighted in red and those which cause other disorders associated with impaired growth are shown in dark blue. *Abbreviations: ROS=reactive oxygen species, ssDNA=single-stranded DNA, DSB=DNA double strand break.*

1.3.2.4 ATR and ATRIP in Seckel syndrome

ATR was the first MPD disease gene to be identified. A homozygous synonymous mutation (c.2101A>G) that altered splicing resulting in partial loss of ATR function was identified in two related consanguineous families with Seckel syndrome through linkage studies (O'Driscoll *et al.*, 2003). Patient cells showed increased sensitivity to UV radiation and mitomycin C with impaired phosphorylation of downstream ATR DDR targets including p53. No impairment was seen in response to ionising radiation reflecting intact ATM-dependent signalling and repair. The patients were diagnosed with Seckel syndrome on the basis of microcephaly, receding forehead, prominent nose, micrognathia and developmental delay. In one patient severe growth retardation of prenatal onset was documented (OFC -12 s.d., height -4.8 at 9 years) (Goodship *et al.*, 2000). Three further cases have since been reported carrying compound heterozygous mutations in *ATR*. Two apparently unrelated patients have the same variants, a nonsense mutation in combination with a non-synonymous mutation (Ogi *et al.*, 2012). Both patients showed severe pre- and post-natal growth failure with severe microcephaly (average OFC -10 s.d., height -8 s.d.). The most recent patient reported carried a nonsynonymous coding variant which disrupted splicing in combination with a 540 kb deletion encompassing four genes including *ATR* (Mokrani-Benhelli *et al.*, 2013). This patient exhibited a less severe growth phenotype (OFC -5 s.d., height -1 s.d.) along with severe learning disability, epilepsy, a profound episode of bone marrow hypoplasia shortly after birth and mild immune abnormalities.

One patient has also been described with a heterozygous nonsense mutation (c.2278C>T, p.Arg760*) in *ATRIP* in combination with altered splicing of the other allele (Ogi *et al.*, 2012). ATRIP binding confers stability to ATR and assists in the localisation of ATR to single stranded DNA (Cortez *et al.*, 2001). This patient also had severe pre- and post-natal global growth failure (OFC -10 s.d., height -6.5 s.d.) as well as microtia, micrognathia, and dental crowding. Similar features were also described in the ATR patients, interestingly displaying some overlap with Meier-Gorlin Syndrome (Ogi *et al.*, 2012). Markedly reduced expression of both *ATRIP* and *ATR* was observed in both ATR and ATRIP patient cells along with reduced activation of downstream targets including CHK1 and FANCD2, reduced formation of 53BP1 foci and impaired G2/M checkpoint arrest in response to UV radiation consistent with impaired ATR DDR signalling. As so few cases have been reported to date, despite identification of ATR as a cause of MPD over a decade ago, this suggests *ATR/ATRIP* mutations are a relatively rare cause of MPD. An autosomal dominant nonsynonymous coding variant in *ATR* affecting p53 activation has also been reported in a large family with an oropharyngeal cancer syndrome (Tanaka *et al.*, 2012b) however additional families have yet to be reported. Furthermore, cancer predisposition has not yet been reported in association with MPD and autosomal recessive *ATR* mutations or in any of the parents carrying heterozygous mutations.

1.3.2.5 RBBP8 in Seckel syndrome

Two independent families with multiple affected individuals homozygous for nonsense mutations in the retinoblastoma binding protein 8 gene, *RBBP8* (also known as *CTIP*), have been reported (Qvist *et al.*, 2011). RBBP8 acts in combination with the MRE11-RAD50-NBS1 (MRN) complex in the end processing of DSBs during homologous recombination (Sartori *et al.*, 2007). The phosphorylation of RBBP8, a 5'-3' exonuclease, is required for the initiation of DNA end resection generating single-stranded DNA which consequently activates ATR signalling (Sartori *et al.*, 2007). RBBP8 has also been shown to be required for sustained ATR signalling and checkpoint activation ensuring complete repair occurs (Kousholt *et al.*, 2012).

In one family, aberrant splicing as the consequence of a new intronic splice site resulted in a truncated but stabilised protein, while in the second a homozygous 2 bp deletion in exon 11 was detected. In the former family, a diagnosis of Seckel syndrome had been made due to low birth weight and proportionate growth failure (height -3.5 to -5.5 s.d., OFC -4.7 s.d. to -5 s.d.) whereas affected cases in the second family showed a similar degree of microcephaly but with normal height (OFC -5 s.d. to -7 s.d.). Affected cases from both families had a characteristic ‘Seckel syndrome’ appearance with receding forehead and prominent nose, developmental delay, café au lait spots and digital abnormalities including phalangeal joint swelling, clinodactyly, polydactyly, syndactyly and absent nails. RBBP8 patient cells had increased sensitivity to ionising radiation demonstrating impaired DSB repair along with reduced CHK1 activation (Qvist *et al.*, 2011). This also reflects how different DNA damage response and repair mechanisms function cooperatively in the repair of certain lesions. Of note, mutations in all three genes encoding the MRN complex have also been associated with microcephaly and reduced growth (Matsuura *et al.*, 1998, Waltes *et al.*, 2009, Matsumoto *et al.*, 2011) (Table 1.2).

1.3.3 DNA replication proteins in Meier-Gorlin Syndrome

The discovery of mutations within components of the origin of replication complex (*ORC1*, *ORC4*, *ORC6*, *CDT1* and *CDC6*) in patients with Meier-Gorlin syndrome (Bicknell *et al.*, 2011a, Bicknell *et al.*, 2011b, Guernsey *et al.*, 2011) extended insights into cell cycle processes impacting on growth.

During S phase of the cell cycle, DNA replication is initiated at specific genomic sites referred to as origins of replication. Licensing of such origins at the end of M and during G1-phase requires the loading of a hexameric complex consisting of ORC1-ORC6 onto chromatin (Bell *et al.*, 1992, Gavin *et al.*, 1995) (Figure 1.9). This is followed by the recruitment of CDC6 (cell division cycle 6), CDT1 (chromatin licensing and DNA replication factor 1) and MCM (mini-chromosome maintenance) 2-7 helicase complex (Mendez *et al.*, 2003) forming the pre-replication complex (pre-RC). DNA replication is then initiated at the onset of S-phase by the phosphorylation of the pre-RC (Krude *et al.*, 1997, Lei *et al.*, 1997) resulting in association of replisomal proteins and activation of the DNA helicase complex (MCM2-7) necessary for DNA unwinding to form the replication fork (Moyer *et al.*, 2006, Ilves *et al.*, 2010). DNA replication proceeds bidirectionally from each origin until the entire genome is duplicated (Marheineke *et al.*, 2005).

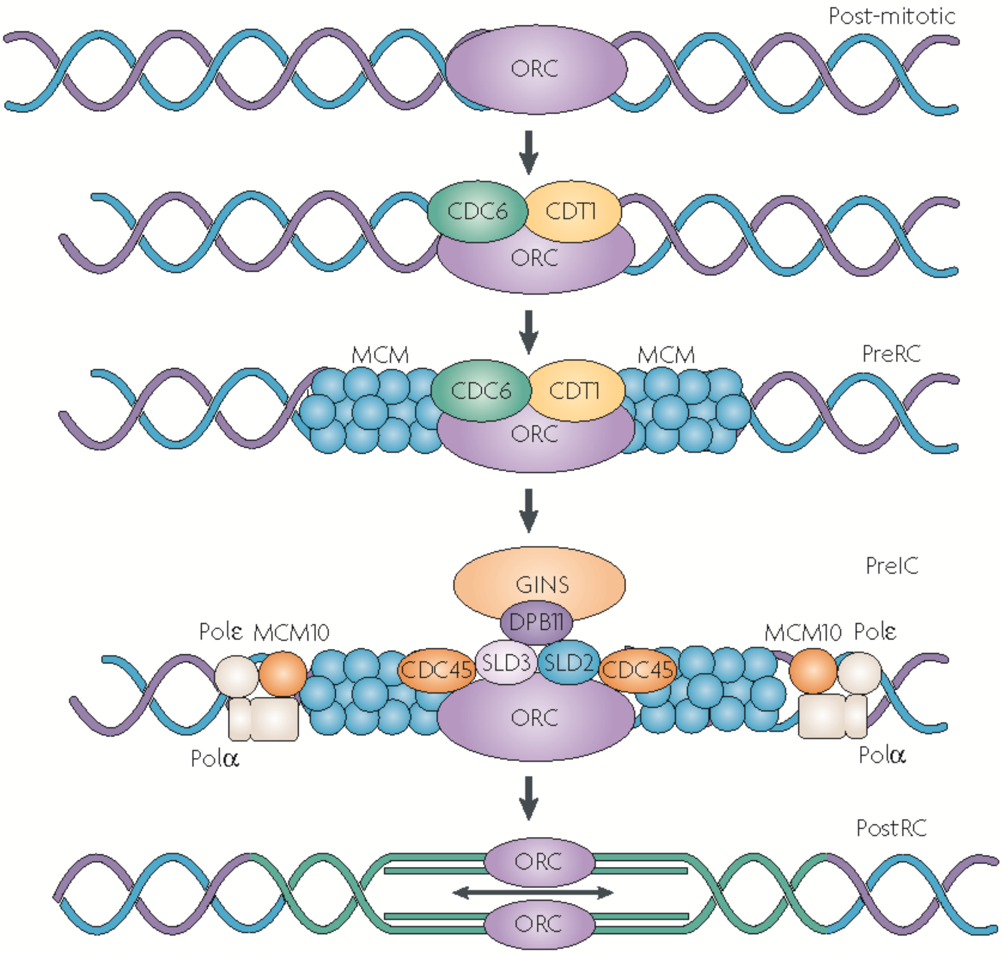


Figure 1.9. Pre-replication complex assembly and initiation of DNA replication in eukaryotes

During M/G1-phase, the ORC complex consisting of six subunits (ORC1-6) initially binds to chromatin, a process predominantly regulated by ORC1. Recruitment and binding of CDC6 and CDT1 follow with subsequent loading of at least 2 MCM helicase (MCM2-7) complexes (Ilves *et al.*, 2010) to complete the pre-replication complex (preRC) and licencing of the replication origin. At the onset of S-phase, the pre-RC complex is activated by cyclin-dependent kinases, recruiting several other proteins to form the pre-initiation complex (preIC), including the cell division cycle protein 45 (CDC45), the GINS complex, SLD2 and MCM10 (Zegerman *et al.*, 2007, Ilves *et al.*, 2010, van Deursen *et al.*, 2012). DNA topoisomerase 2-binding protein 1 (TOPBP1) and treslin (homologues of yeast DPB11 and SLD3 respectively) are also recruited to the preIC (Mueller *et al.*, 2011). The final recruitment of DNA polymerase machinery (Polε and Polα) initiates replication in both directions from the origin preceded by the MCM2-7 helicase which travels ahead of the polymerase to unwind the double stranded DNA. Reproduced from Aladjem, (2007).

To ensure the genome is replicated once and only once during each cell cycle, licencing of origins is strictly confined to M/G1 phase (Arias *et al.*, 2007). Under unstressed conditions only approximately 10% of origins licenced for replication are activated (Cayrou *et al.*, 2011). However, as new origins can not be licensed following entry into S phase (Wohlschlegel *et al.*, 2000), dormant origins can be utilised in the event of stalled replication forks or failed initiation of origins ensuring complete replication still occurs (Woodward *et al.*, 2006). To make certain that an adequate number of origins are licenced during G1 prior to progression into S-phase, a p53-dependent licencing checkpoint has been proposed to be activated preventing G1/S transition until a certain threshold is reached (Nevis *et al.*, 2009).

Reduction in assembly of the pre-RC reduces capacity to activate replication origins may result in slower progression through the cell cycle due to prolonged G1-phase and delayed S-phase entry (Noguchi *et al.*, 2006, Bicknell *et al.*, 2011b). Although a large degree of redundancy exists in the number of licenced origins it has been hypothesised that reduced licencing capacity may become rate limiting in certain cell populations with higher rates of proliferation (Bicknell *et al.*, 2011b, Klingseisen *et al.*, 2011), such as neuronal stem cells in which G1-phase is reduced (Takahashi *et al.*, 1995). The failure in growth seen in MGS patients may therefore be the result of a slower proliferation rate in specific cell populations critical to ensuring normal cellularity during development. Notably, patients with mutations in *ORC1* have a greater reduction in growth compared to mutations in other pre-RC genes (de Munnik *et al.*, 2012). The majority of *ORC1* mutations in MGS patients lie within the BAH chromatin binding domain (Bicknell *et al.*, 2011a). ORC1 is the only ORC protein with a BAH domain and appears to play an important role in the regulation of pre-RC loading onto chromatin (Noguchi *et al.*, 2006). The BAH domain has recently been shown to bind to chromatin bound H4K20me2, a modified histone enriched at replication origins and disruption of this interaction was found to impair normal growth in zebrafish (Kuo *et al.*, 2012).

Mice with mutations in the MCM subunits also have impaired growth along with genome instability and cancer predisposition (Kunnev *et al.*, 2010). In humans, a founder mutation in the *MCM4* helicase has been described in the Irish traveller community in association with short stature, adrenal insufficiency and natural killer cell deficiency, although OFC was normal or only mildly reduced and additionally microtia and patella abnormalities were not noted (Gineau *et al.*, 2012, Hughes *et al.*, 2012). Patient cells also exhibited increased chromosomal breakage in response to genotoxic agents (Gineau *et al.*, 2012). In contrast, genome instability and cancer predisposition have not been described in patients with MGS.

1.3.4 Splicing in MOPDI

Mutations within *RNU4ATAC*, which encodes a member of the minor spliceosome targeting splicing of U12 introns (Edery *et al.*, 2011, He *et al.*, 2011) have recently been identified including a founder mutation g.51G>A in several affected cases in the Ohio Amish population. The majority of mutations appear to lie in the structurally important 5' stem loop and are predicted to disrupt RNA secondary structure (He *et al.*, 2011). However, the exact mechanism by which *RNU4ATAC* mutations causes MOPDI is not yet understood. U12 introns are present in approximately 700 genes with diverse functions including cell cycle regulation (Sheth *et al.*, 2006, Alioto, 2007). Severe growth failure may therefore result from a cumulative effect of the abnormal splicing of several different RNA transcripts or even from the effect on just one specific gene transcript.

1.3.6 Common cellular pathways for MPD genes

A common feature of all the above genes so far identified in MPD is a disruption in cell cycle kinetics either through impaired DNA replication, impaired mitosis or failure in adequate DNA damage response. It is therefore conceivable that such defects can lead to a reduction in global cellularity either through reduced proliferation or increased cell death. Notably, there is also significant overlap between these different cellular processes. For example ATR activation alters replication origin distribution in response to stalled replication forks (Ge *et al.*, 2010) and many components of the ATR signalling pathway localise to the centrosome including CHK1 (Kramer *et al.*, 2004). Furthermore, both PCNT and CEP152 appear to contribute to the DNA damage response. CEP152 has been shown to bind to CINP, a genome maintenance protein which interacts with ATRIP (Lovejoy *et al.*, 2009). Thus, it is unlikely that each gene described above impacts on growth through the disruption of a single pathway.

At least four genes (*CENPJ, CEP152, TUBGCP6, RBBP8*) are associated with a variable phenotype which ranges from microcephaly and normal stature to MPD suggesting the two conditions may represent a single disease spectrum with similar underlying pathogenesis (Bond *et al.*, 2005, Al-Dosari *et al.*, 2010, Guernsey *et al.*, 2010, Kalay *et al.*, 2011, Qvist *et al.*, 2011, Puffenberger *et al.*, 2012, Martin *et al.*, 2014). Why some mutations result in primary 1° MCPH and others in MPD is not yet fully understood but this suggests other genes implicated in 1° MCPH could also cause MPD.

1.4 Methods of Gene discovery

The previous Section highlighted that MPD is a highly heterogeneous disorder. Recent developments in sequencing technology may therefore be beneficial to further gene discovery in this group of conditions, particularly given the rarity of MPD and that patients are predominantly single cases from non-consanguineous families. This combined with genetic heterogeneity in an extremely rare disorder makes gene discovery difficult by traditional methods.

Prior to the advent of next generation sequencing technology, Sanger sequencing (Sanger *et al.*, 1977) was the most efficient method available by which to sequence DNA and identify causative mutations. In this method, a DNA polymerase replicates DNA from a primer bound to the single stranded template under investigation. Chain-terminating dideoxyribonucleotides (ddNTPs), along with unmodified dNTPs, are added to the sequencing reaction. ddNTPs lack the 3’-OH group required to form a phosphodiester bond thus terminating further synthesis once incorporated into the newly synthesised strand. The sequence can then be deduced using capillary electrophoresis to determine the different fragment lengths produced following addition of each of the four ddNTPs. In modern methods, the ddNTPs are fluorescently labelled to allow automated detection (dye-terminator sequencing) (Smith *et al.*, 1986).

Advantages of Sanger sequencing include a low error rate and the ability to produce reasonably long sequencing reads (up to 900 bp) in a single reaction (Liu *et al.*, 2012). Although automated capillary sequencing machines can analyse multiple sequencing reactions at one time (up to 384), only one DNA template can be sequenced per reaction. Therefore sequencing large numbers of genes by this method is not feasible for most researchers and diagnostic services. Genetic mapping techniques such as linkage analysis including homozygosity mapping (Lander *et al.*, 1987, Kerem *et al.*, 1989), karyotyping to detect translocations (Kurotaki *et al.*, 2002) and detection of duplications and deletions (Vissers *et al.*, 2005) can assist in limiting the genomic region in which the causative mutation might reside. Candidate genes within this region can then be prioritised and sequenced although the region may still be too large to sequence every gene. Methods utilising genetic linkage are also often only successful in homogeneous disorders with multiple affected individuals or where DNA from several family members is available to inform on commonly inherited regions.

1.4.1 Next Generation Sequencing (NGS) Technology

Mass paralleled sequencing technologies have revolutionised DNA sequencing. Sequencing the entire genome is now possible in under a day and costs are rapidly reducing as technology continues to advance. This has resulted in the rapid identification of a large number of new disease causing genes in recent years and provided valuable information on the degree of normal variation within the human genome (Genomes Project *et al.*, 2012). However, this technology is not without disadvantages with higher error rates and often limited read length compared to traditional Sanger sequencing methods (Metzker, 2010).

In NGS technologies numerous unique template fragments can be spatially separated on a single stage enabling vast numbers of sequencing reactions to be performed simultaneously. Each template is clonally amplified via emulsion PCR (Dressman *et al.*, 2003) or solid-phase amplification (Fedurco *et al.*, 2006) to enable detection of the fluorescent signal during DNA synthesis (Figure 1.10). Methods using single molecule templates are also employed by some systems (Harris *et al.*, 2008) and have the advantage of requiring lower input DNA, generally more straightforward preparation protocols and the potential to produce longer read lengths (Metzker, 2010). Ligating a short unique identifying sequence (barcode) onto each template fragment also allows multiple samples to be sequenced in one reaction (Meyer *et al.*, 2007, Parameswaran *et al.*, 2007).

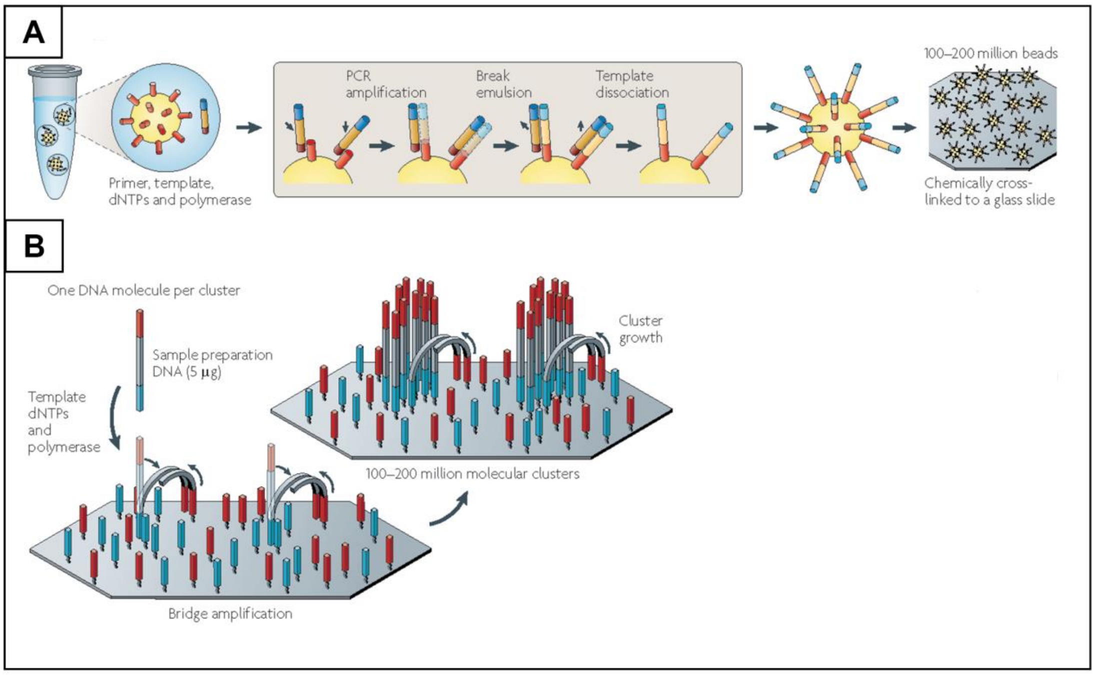

A
Primer, template, dNTPs and polymerase
PCR amplification
Break emulsion
Template dissociation
100–200 million beads
Chemically cross-linked to a glass slide
B
One DNA molecule per cluster
Sample preparation DNA (5 μg)
Template dNTPs and polymerase
Bridge amplification
100–200 million molecular clusters
Cluster growth

Figure 1.10. Methods of template immobilisation and amplification in Next Generation Sequencing Technologies

In NGS technology millions of template fragments can be separated and sequenced in a single reaction. The most commonly used technologies require clonal amplification of each fragment prior to sequencing in order for the fluorescent signal to be detected. Prior to this stage adapters containing universal primer sequences (blue and red) are ligated onto the ends of each fragment (Orange in A, grey in B). A) In emulsion PCR (e.g. used by Roche) one DNA fragment is hybridised to a single bead (yellow) in solution. The fragment is then repeatedly amplified so each bead holds several thousand copies of the same fragment. The beads can then be evenly distributed and immobilised on a solid platform prior to sequencing. B) In solid-phase amplification (e.g. used by Illumina) each fragment is hybridised to a solid platform containing millions of universal forward and reverse primers prior to amplification. Initial extension of each primer allows the fragment to form a bridge with neighbouring primers which is then repeatedly amplified. Reproduced from Metzker *et al.*, 2010.

Currently five different sequencing methods are used by commercially available NGS systems; pyrosequencing, ion semi-conductor sequencing, sequencing by ligation (SBL), cyclic reversible termination (CRT) and real-time sequencing (Figure 1.11). In CRT, one terminating fluorescent nucleotide is added at a time followed by imaging which detects where they have been incorporated (Ju *et al.*, 2006). The terminating group and dye is then cleaved allowing further synthesis on addition of the next nucleotide. SBL involves the use of DNA ligase to join cyclically introduced dye-labelled probes (Landegren *et al.*, 1988). The two base pair probe hybridises to the complimentary sequence on the template and is then ligated to the adjacent probe/primer by DNA ligase. Unligated probes are washed away prior to imaging to identify only those that have been incorporated. Pyrosequencing and ion semi-conductor sequencing differs from the other two methods in that they do not involve the cyclical termination of DNA synthesis. Pyrosequencing measures the release of inorganic pyrophosphate on incorporation of a complimentary nucleotide by converting it to a light signal. The intensity (proportional to the amount of nucleotide incorporated) and order of the light signal following sequential addition of each dNTP is then used to determine the sequence. Ion semi-conductor technology is similar except the voltage produced from emitted hydrogen ions following incorporation of nucleotides is measured (Rothberg *et al.*, 2011). However, runs of identical nucleotides in the sequence (homopolymer runs) are more prone to error in these latter two methods due to difficulties in accurately relating the intensity of the emitted signal to the number of bases incorporated (Harris *et al.*, 2008, Liu *et al.*, 2012).

Finally, and most recently available, is the real time imaging of incorporated dye-labelled nucleotides during uninterrupted template-directed synthesis using a zero-mode waveguide detector (real-time sequencing) (Eid *et al.*, 2009). This allows the sequencing of longer reads compared to other methods and improved sequencing of GC rich regions (Shin *et al.*, 2013). As single molecule templates are used, this method avoids errors which can potentially be introduced during clonal amplification (Metzker, 2010). Advantages and disadvantages of each method are highlighted in Table 1.3 which shows the most popular sequencing platforms commercially available and the associated technology.

Table 1.3. Comparison of different NGS sequencing platforms

Sequencing platform (company)	*Template prep^n*	*Sequencing method*	*Read length /bp*	*Run time /days*	*Out-put /Gb*	*Pros*	*Cons*
HiSeq 2000 (Illumina)	Solid-phase	CRT	100	3-10	600	Widely available, high output for cost	Short read length
454 GS FLX (Roche)	emPCR	Pyro-sequencing	700	1	0.7	Long read length	High error rate in homo-polymer runs
Ion Proton (Life Technologies)	emPCR	Semi-conductor	200	2-4 hours	10	High speed, low cost	High error rate in homo-polymer runs
SOLiDv4 (Applied biosystems)	emPCR	SBL	100	7-14	120	High accuracy	Short read length
PacBio RS (Pacific Biosciences)	Single molecule	Real-time	1300	<1	*N/A*	Long read length, short DNA prep^n time	Limited availability

All data obtained from Liu *et al.*, (2012) except Ion proton specifications obtained from company website (www.lifetechnologies.com, accessed 07.07.14). *Abbreviations: EmPCR=emulsion PCR, CRT=cyclic reversible termination, SBL=sequencing by ligation, prep^n=preparation.*

1.4.1 Whole Exome Sequencing (WES)

Sequencing a human genome requires the analysis of 6 Gb of DNA which can contain over three million variants (Wheeler *et al.*, 2008). Not only is this a large volume of data to analyse and store but identifying one or two causal pathogenic variants from such a large dataset is very challenging. As approximately 85% of pathogenic mutations lie in protein coding regions of the human genome (Botstein *et al.*, 2003), limiting sequencing to these regions drastically reduces the volume of data produced (300 fold reduction in variant number; Wheeler *et al.*, 2008) whilst still capturing the majority of disease causing mutations. Also the 20-fold reduction in sequencing required (Ng *et al.*, 2009) allows more samples to be analysed within the confines of available resources such as additional patients or family members.

Including more samples in the experimental design could also potentially increase the chances of identifying causative variants.

In WES, the desired genomic regions are captured by the hybridisation of sheared DNA to probes complimentary to the sequence of interest, the exons. Sequence probes may be fixed to a solid platform such as a microarray (Hodges *et al.*, 2007) or filter (Herman *et al.*, 2009) (solid-phase hybridization) or in suspension and biotinylated (liquid-phase hybridization) allowing them to bind to magnetic streptavidin beads (Gnirke *et al.,* 2009). Following hybridisation and capture of target regions, fragments containing non-coding sequences can be washed away and the remaining library enriched prior to sequencing. A range of different exome capture kits are now commercially available. Those based on liquid-phase hybridisation (Agilent and NimbleGen) have the advantage of being more amenable to automation and thus potentially have a higher throughput capacity.

Initially, proof of principle experiments successfully identified known disease causing mutations by using targeted capture methods and NGS in several patient groups (Ng *et al.*, 2009, Chou *et al.*, 2010, Hoischen *et al.*, 2010, Raca *et al.*, 2010). WES was then successfully employed to identify novel disease causing genes in 2010 where patients with Miller syndrome were found to share deleterious, autosomal recessive variants in the *DHODH* gene (Ng *et al.*, 2010b). This was shortly followed by the identification of *de novo* variants in *MLL2* in patients with Kabuki syndrome (Ng *et al.*, 2010a). Both discoveries, and many more since, have established WES as a useful and powerful diagnostic tool for isolated, unrelated cases of rare Mendelian disorders.

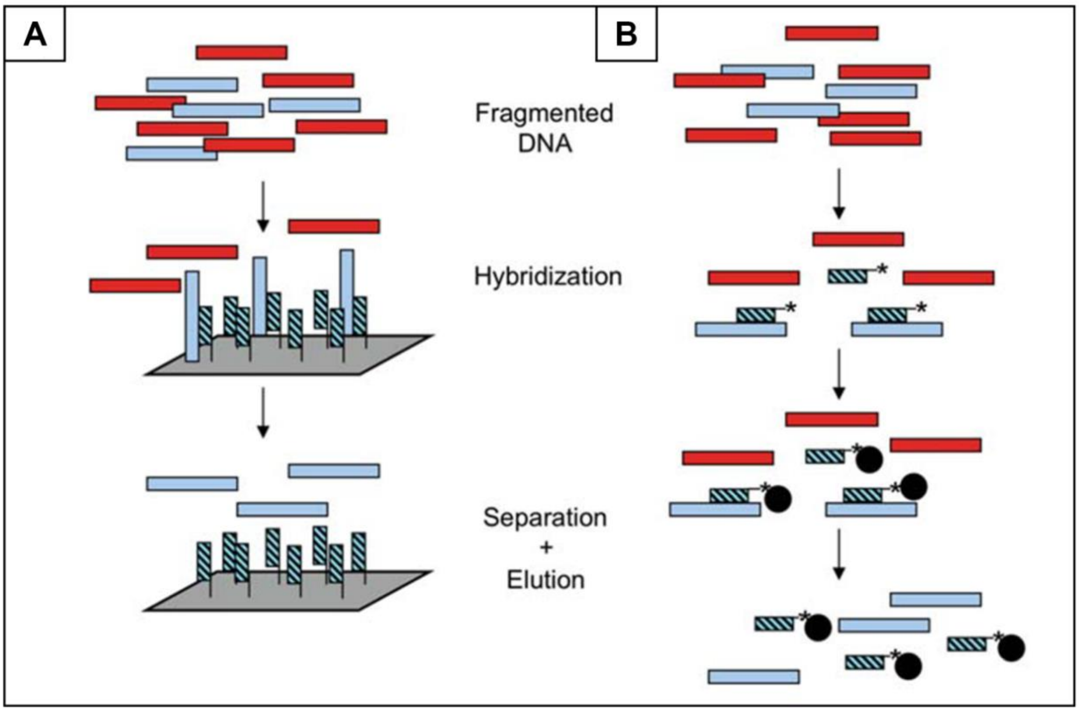


Figure 1.11. Different methods of exome capture

In both methods illustrated DNA fragments containing the sequence of interest (light blue) are captured by hybridisation to a set of complimentary probes (dark blue/black hashed). In A) Solid-phase hybridisation probes are fixed to a microarray whereas in B) Liquid-phase hybridisation probes are biotinylated (black) and in suspension which can then bind to streptavidin beads. Following capture, unwanted DNA fragments (red) are removed by washing and the desired fragments eluted from the probes. Reproduced from Teer *et al.*, (2010).

1.4.3 Summary

WES is now a commonly used research technique in gene discovery in a wide variety of both common and rare disease (Majewski *et al.*, 2011) and with decreasing costs the possible applications of NGS technologies has expanded (for example, detection of copy number variation, homozygosity mapping and RNA sequencing) (Medvedev *et al.*, 2009, Ku *et al.*, 2012, Seelow *et al.*, 2012). Given the large number of mechanisms that regulate growth (Conlon *et al.*, 1999, Klingseisen *et al.*, 2011), the number of potential candidate genes for MPD is extensive. It is therefore not feasible to sequence large numbers of genes in every patient by conventional capillary based sequencing methods. This is not only due to time and cost but also the large quantity of DNA required may not be realistic to obtain in young patients with a small body mass. Also many of these patients are single cases in non-consanguineous families and therefore are not suitable for gene discovery by genetic mapping methods. NGS affords the ability to perform massive parallel sequencing providing an attractive, and now cost effective, method for screening a large number of genes. WES has already led to the identification of novel disease causing genes in MPD including the centrosomal genes, *CEP152* (Kalay *et al.*, 2011) and *PLK4* (Martin *et al.*, 2014). As well as identifying disease genes in known pathways and mechanisms affecting growth, WES will also facilitate the discovery of unanticipated disease causing genes. These may be functionally different to those previously discovered giving new insights into growth regulation.

1.5 Thesis aims and objectives

1.5.1 Hypothesis: MPD is a heterogeneous genetic disorder resulting from defects in a number of cellular pathways and mechanisms that reduce global cellularity

Recent advances in DNA sequencing has led to a rapid increase in the discovery of novel disease causing genes in Mendelian disorders. Despite this, many MPD patients still remain without a molecular diagnosis. Within this patient group, marked clinical heterogeneity has been described suggesting a large degree of underlying genetic heterogeneity. This is supported by the multiple disease causing genes identified in MPD to date (Section 1.3). In addition, reduced growth is a common finding in many different developmental disorders, several of which may overlap with MPD further adding to the genetic diversity of this group of conditions. It is likely that genetic heterogeneity combined with the rarity of MPD has hindered gene discovery to date. For many disease causing genes in MPD, mutations have only been reported in a small number of families (often only one) preventing detailed characterisation of distinct phenotypes (Section 1.3). Identifying other possible cases with mutations in the same gene has therefore been difficult on a phenotypic basis.

To date, almost all disease causing genes identified in MPD encode proteins that operate in different but overlapping mechanisms which alter cell cycle kinetics and survival, including DNA replication, mitosis and DNA damage response and repair. Such processes involve many different proteins, regulated by complex signalling pathways. Therefore, any gene encoding proteins involved in such processes and pathways could be considered a possible candidate for MPD. The diversity of potentially affected processes is further illustrated by genome wide association studies in which a large number of genetic loci have been link to both height and head circumference (Weedon *et al.*, 2008, Ikram *et al.*, 2012b, Taal *et al.*, 2012a). In addition, growth retardation is a commonly reported finding in many knockout mouse strains (Reed *et al.*, 2008).

1.5.2 Thesis aims

The objective of this thesis was to define the molecular basis for profound growth impairment in patients for whom a recognisable syndrome is not apparent and a molecular diagnosis has not yet been determined in order to identify novel disease genes. Secondly, I aimed to characterise associated phenotypes, with a view to improving clinical diagnosis, management and outcomes. Finally, identifying novel genetic causes of MPD may provide new insights into physiological growth regulation.

I plan to address this with the following aims:-

1. Design and implement an analytical pipeline for the analysis of WES in a large cohort of clinically diverse MPD patients.

2. Identify novel candidate disease causing genes in MPD through the analysis of WES.

3. Establish the impact of mutations in novel disease genes on protein function and investigate the mechanism by which protein dysfunction impairs growth.

4. Identify structural aberrations altering gene dosage that results in the MPD phenotype.

Chapter 2: Materials & Methods

2.1 Patient recruitment

Patients were recruited to research studies at the MRC Human Genetics Unit in Edinburgh, UK, the Institute of Human Genetics at the University of Cologne, Germany and the Nemours Foundation, Delaware, USA by their local physician.

2.1.1 Ethical consent

This study was approved by the multi-centre research ethics committee for Scotland (04:MRE00/19). Collaborative studies, from which results in this thesis are also obtained, were approved by the Cologne hospitals ethics board or the Nemours Office of Human Subject Protection (NOHSP) and Institutional Review Board. Informed consent was obtained from all families.

2.1.1 Inclusion criteria

All patients selected into the study had a clinical diagnosis of MPD based on head (occipital-frontal) circumference (OFC) and height being more than 4 standard deviations (s.d.) below the population mean at the time of recruitment. Standard deviations for height, weight, and OFC normalized for age and sex were calculated using Cole's LMS method using UK 1990 cohort data (Freeman *et al.*, 1995). Age at examination was corrected for prematurity where birth was before 37 weeks gestation and age <2 years. Age was rounded to the nearest month prior to s.d. calculation.

2.1.3 Clinical data collection

Medical history, anthropometric data, examination findings and clinical laboratory results were obtained by questionnaire from the referring physician.

2.2 Chemical reagents and buffers

2.2.1 Sources of reagents

All chemicals were purchased from Sigma-Aldrich, Fisher Scientific or Amersham Biosciences (GE Healthcare Life Sciences). Enzymes were from New England Biolabs, Promega and Roche and cell culture materials from Gibco (Life Technologies).

2.2.2 Buffer solutions

All buffers listed were prepared with dH_2O (Elix® essential 5 water purification system, EMD Millipore).

10 X TBE: 0.89 M Tris base, 0.89 M boric acid, 20 mM EDTA

TE: 10 mM Tris-HCl and 1 mM EDTA (autoclaved at 121^{o}C for 15 minutes to sterilise)

Low TE: 10 mM Tris-HCl and 0.1 mM EDTA (pH8, autoclaved at 121^{o}C for 15 minutes to sterilise)

WCE Buffer: 50 mM Tris-HCl pH8, 280 mM NaCl, 0.5% NP40, 0.2 mM EDTA, 0.2 mM EGTA, 10% glycerol and EDTA-free protease inhibitor tablet (Roche)

1 X Western transfer buffer: 25 mM Tris base, 192 mM glycine, 0.1% (w/v) SDS, 20% (v/v) methanol

1 X TBST: 0.05 M Tris base, 0.15 M NaCl, 2% (v/v) Tween-20 **(**pH 7.5)

PHEM: 25 mM Hepes-NaOH pH 6.8, 100 mM EGTA, 60 mM PIPES, 2 mM $MgCl_2$

Hypotonic Buffer: 0.56% (w/v) KCl, 1% (w/v) $Na_3C_6H_5O_7$

2.3 Cell culture methods

2.3.1 Preparation and growth of human cell lines

2.3.1.1 Lymphoblastoid cells (LBCs)

Lymphocytes were isolated from fresh blood samples and transformed with Epstein Barr virus by Sean O'Neil (MRC HGU, IGMM, Edinburgh). LBCs were cultured in suspension in Roswell Park Memorial Institute (RPMI) 1640 medium (Gibco) supplemented with 10% fetal calf serum (FCS, HyClone™), 100 U/ml penicillin and 100 µg/ml streptomycin. Cells were maintained at 37^{o}C with 5% CO_2 and at a density of 2-10 x 10^{5}cells/ml.

2.3.1.2 Fibroblast cells

Fibroblasts were cultured from a 3x3 mm punch skin biopsy taken from an affected MPD patient and two healthy control individuals. Skin biopsies were performed by a local physician under aseptic technique (iodine and mercurochrome based antiseptics were avoided). The biopsy was transported in PBS or culture media. On arrival the sample was dissected into smaller segments using a scalpel and placed in a sterile 16 mm laten tube (Nunc™ cell culture tube, Thermo Scientific) under a glass coverslip (Nunc™ Thermanox™ coverslip, Thermo Scientific). These were cultured in pre-warmed AmnioMax™ C-100 basal media plus AmnioMax™ C-100 supplement (Gibco) at 37^{o}C with 3% O_2 and 5% CO_2. Initial skin biopsy samples were cultured for a minimum of 2 weeks prior to first passage. When confluent, cells were passaged by the removal of media and washed twice with pre-warmed PBS. Cells were detached from the vessel surface by the addition of trypsin:versene (1:1, v:v) and incubated at 37^{o}C for 5 minutes. Cells were re-suspended in fresh media and split 1:2-1:6 of their original density. Cells were maintained in the same culture medium at 37^{o}C with 3% O_2 and 5% CO_2.

2.3.1.3 Cell preservation

Initially cells were stored at -80^{o}C in 2ml cryostat tubes re-suspended in 1ml FCS and 10% DMSO prior to transfer to liquid nitrogen for longer term storage. Cell pellets for nucleic acid and protein extraction were obtained by centrifugation of re-suspended cells at 1200 rpm (Allegra™ X-22 centrifuge, Beckman Coulter) for 4 minutes followed by removal of supernatant and stored until required at -80^{o}C. Prior to use, cell pellets were thawed and washed in PBS by centrifugation at 6,000 rpm for 10 minutes (Heraeus Megafuge 1.0R centrifuge, Thermo Scientific).

2.3.2 Transfection with short interfering RNA (siRNA)

NCAPD3 and Luciferase control siRNA oligonucleotides were designed using the Whitehead Institute siRNA design tool (http://jura.wi.mit.edu/bioc/siRNA) (Yuan *et al.*, 2004) by Carol-Anne Martin (MRC HGU, IGMM, Edinburgh).

NCAPD3 siRNA: 5´CAU GGA UCU AUG GAG AGU AUU-3´
Luciferase control siRNA: 5´-CUU ACG CUG AGU ACU UCG AUU-3´

siRNA oligonucleotides were transfected into monolayer primary fibroblasts at 40-50% confluency using DharmaFect transfection reagent (Thermo Scientific) according to manufacturers' instructions. Each well of a 6-well plate was transfected with 25 nM siRNA. Two solutions were prepared, one containing 50 nmol siRNA in 200 µl of Opti-MEM I Medium (Life Technologies) and the second contained 5 µl of DharmaFect in 195 µl of Opti-MEM. Both solutions were incubated at RT for 5 minutes prior to combining and then incubated for a further 20 minutes before adding to the cells. Following addition of the transfection solution, a further 800 µl Opti-MEM was added and cells were incubated for 8 hours at 37^{o}C with 3% O_2 and 5% CO_2. The transfection media was then removed by aspiration and replaced with AmnioMax™ C-100 media plus supplement.

2.4 Nucleic acid methods

2.4.1 Nucleic acid extraction from human cells

2.4.1.1 Genomic DNA

Genomic DNA was extracted from whole blood using the Genomic DNA extraction kit (Illustra) or saliva samples using ORAgene® collection kits according to manufacturer's instructions. This was performed by Sean O'Neil (MRC HGU, IGMM, Edinburgh). Genomic DNA was extracted from ~1 x 10^6 primary fibroblasts using the DNeasy blood and tissue kit (Qiagen) following the manufacturers' instructions and eluted in 100 µl of elution buffer.

2.4.1.2 RNA and generation of cDNA

RNA was extracted from human LBCs and primary fibroblasts using the RNeasy Mini Kit (Qiagen) as per manufacturers' instructions. Approximately 1 x 10^6 cells were homogenised using a Qiashredder column (Qiagen) by centrifugation at 16000 g for 2 minutes. An on-column treatment was then performed with 30 U RNase-free DNase I (Qiagen) for 15 minutes to remove genomic DNA. RNA was eluted in 30 µl of RNase-free H_2O (Qiagen) and stored until required at -80°C.

Complimentary DNA (cDNA) was then generated from template RNA by reverse transcription. Each reaction contained 400 ng to 1 µg of RNA, 100 pmol random primers (Promega), 5 mM Dithiothreitol (DTT) and 40 U Protector RHase Inhibitor (Roche) made up to a total volume of 14µl with RNase-free H_2O. The reaction was initially heated to 70°C for 5 minutes followed by incubation for 5 minutes on ice to denature RNA secondary structures. The reverse transcriptase (RT) enzyme (20 U AMV RT, Roche) and buffer (1 X AMV RT buffer, Roche) were then added along with 1 µM dNTPs in 6 µl of RNase-free H_2O. The 20 µl reaction was incubated at 42°C for 60 minutes followed by incubation at 75°C for 8 minutes to inactivate the RT enzyme. For each RNA sample the reaction was also performed in the absence of RT (substituted with RNase-free H_2O) to ensure contamination with genomic DNA had not occurred during the generation of cDNA. Samples of cDNA were stored at -20°C.

2.4.2 Polymerase chain reaction (PCR)

2.4.2.1 Primer design

Primers were designed to amplify specific exons and intron-exon boundaries using the ExonPrimer tool (http://ihg.gsf.de/ihg/ExonPrimer.html) which utilises the Primer3 tool (v.0.4.0) (Untergasser *et al.*, 2012). Primers ranged from 17 to 27 bp in length with a melting temperature (T_m) of 57°C to 63°C. This script utilises a nearest neighbour formula to calculate the T_m of individual primers which did not differ by more than 4°C between forward and reverse primer pairs (SantaLucia, 1998). Primers designed and used in this thesis are listed in Appendix I.

2.4.2.2 Amplification of genomic DNA and cDNA

Regions of interest were amplified by polymerase chain reaction (PCR) using the relevant primers pairs designed as described in Section 2.4.2.1. PCR reactions were performed using Taq polymerase (Thermo Scientific). Each reaction contained 10 ng of genomic DNA, 1 X Reddymix PCR mastermix (Thermo Scientific) and 0.5 µM of both forward and reverse primers in a final reaction volume of 10 µl made up with dH_2O.

Prior to amplification of cDNA, each sample generated in Section 2.4.1.2 was diluted 5 fold and 1 µl used in each 10 µl reaction volume. For regions with a high GC content, 5% (v/v) DMSO was added to the PCR reaction to inhibit secondary structure formation.

A touchdown PCR programme was used to permit efficient amplification of different primer sets with a range of annealing temperatures to allow for variation in T_m between different primer pairs. This allowed multiple PCR targets to be routinely amplified on the same PCR plate, increasing throughput of samples. PCRs were performed using a DNA Engine Tetrad 2 thermal cycler (MJ research/Bio-Rad) using the following programme;

1. Denaturation: 94°C for 4 minutes

2. Cycle 1 (x3);
Denaturation: 94°C for 30 sec
Annealing: 65°C for 30 sec
Extension: 72°C for 45 sec

3. Cycle 2 (x3);
Denaturation: 94°C for 30 sec
Annealing: 62°C for 30 sec
Extension: 72°C for 45 sec

4. Cycle 3 (x3);
Denaturation: 94°C for 30 sec
Annealing: 59°C for 30 sec

Extension: 72^{o}C for 45 sec

5. Cycle 4 (x35);
Denaturation: 94^{o}C for 30 sec
Annealing: 56^{o}C for 30 sec
Extension: 72^{o}C for 45 sec

6. Extension: 72^{o}C for 10 minutes

7. Hold: 10^{o}C for 15 minutes

Amplification of cDNA with primers complimentary to *CENPJ* was performed using the following PCR programme;

1. Denaturation: 94^{o}C for 4 minutes

2. Cycle 1 (x35);
Denaturation: 94^{o}C for 30 sec
Annealing: 63^{o}C for 30 sec
Extension: 72^{o}C for 90 sec

3. Extension: 72^{o}C for 10 minutes

4. Hold: 10^{o}C for 15 minutes

2.4.3 Agarose gel electrophoresis

Gels were cast by dissolving 1% (w/v) agarose (Hi-Pure Low EEO agarose, Biogene) in 1 X TBE buffer (Section 2.2.2) by heating the solution to boiling in a conventional microwave for 1-2 minutes. To visualise nucleic acids, ethidium bromide (0.5 µg/ml), SYBR® gold (1 in 10,000 dilution) (Life Technologies) or SYBR® Safe (1 in 10,000 dilution) (Life Technologies) was added to the liquid gel (v/v) during cooling. If required, samples were mixed with 1x blue/orange loading dye (Promega) prior to loading. Samples were then loaded into the gel alongside a reference DNA marker containing fragments of known sizes (1 kb plus DNA ladder, Life Technologies and 100 bp ladder, Promega). A constant voltage of 60-120 volts (depending on gel volume) was applied to resolve DNA fragments of different sizes. Nucleic acids were then visualised using a UV (U:Genius3) or LED transilluminator (Syngene) to determine the presence and size of nucleic acid fragments.

2.4.4 PCR purification from agarose gel

Where multiple PCR products were resolved by electrophoresis, each band was excised from the gel using a scalpel. The PCR product was then purified from the gel using the QiaQuick Gel Extraction Kit (Qiagen) according to the manufacturers' instructions. DNA was finally eluted in 30µl of elution buffer.

2.4.5 Ligation of DNA into plasmids and transformation of chemically-competent cells

20 ng of gel-extracted PCR product was ligated into pGEM®-T Easy Vector (Promega) at a 3:1 insert:vector molar ratio as per manufacturers' instructions. The ligation reaction was incubated for 6 hours at RT prior to transformation of chemically competent cells.

Chemically competent DH5α *E. coli* was prepared by Martin Reijns (MRC HGU, IGMM, Edinburgh) (Reijns, 2006) and stored in 200 µl aliquots at -80^{o}C until required for transformation. Cells were thawed on ice and transformed by the addition of 1 µl of ligation reaction to 50 µl of DH5α cells and incubated on ice for 30 minutes. Cells were then heat-shocked at 42^{o}C for exactly 45 seconds and placed on ice for 2 minutes to recover. Cells were re-suspended in 500 µl Luria-Bertani (LB) broth medium and incubated for 60 minutes with shaking. 500 µl of cells were then evenly distributed over a pre-warmed LB-agar plate with ampicillin (0.1%, v/v) and incubated overnight at 37^{o}C following which discrete colonies were visible. Individual colonies were then harvested and grown for a further 24 hours in 5 ml LB broth with 50µg/ml ampicillin at 37^{o}C with constant agitation. Cells were pelleted by centrifugation 4000 rpm for 4 minutes (Allegra X-12R centrifuge, Beckman Coulter) and the residual broth discarded prior to DNA extraction (Section 2.4.6).

2.4.6 Purification of plasmid DNA from *E.coli* cells

DNA was extracted from transformed DH5α *E.coli* using the QIAprep Spin Miniprep Kit (Qiagen) following manufacturers' guidelines and eluted in 50 µl elution buffer.

2.4.7 Quantification and quality assessment of nucleic acids

Nucleic acid concentration was determined in 1.5µl of sample using a NanoDrop 1000 UV-Vis Spectrophotometer (Thermo Scientific) by measuring the sample absorbance at 260 nM. Purity was also assessed by measuring the absorbance at 230 nM and 280 nM. A 260/280 ratio and 230/260 ratio of greater than 1.8 indicates minimal protein, carbohydrate and lipid contamination.

Genomic DNA was additionally quantified by IGMM technical services using the QuantiFluor® dsDNA system (Promega). Double-stranded DNA in each sample was stained with QuantiFluor® dye (Promega) ($504nm_{Ex}/531nm_{Em}$) and the emitted wavelength measured using a 1420 Victor2 microplate reader (Perkin Elmer). DNA concentration was calculated based on a standard curve (0 - 100 ng/ml) plotted from measurements of Lambda DNA at a known concentration after serial dilutions. As many of the DNA samples had been stored for several years and obtained from multiple samples worldwide, each sample was also assessed for evidence of degradation by agarose gel electrophoresis. Over time DNA can fragment following exposure to various environmental insults including UV radiation and variations in temperature which will occur through repeated freeze-thaw of samples (Dean *et al.*, 2001) (Figure 2.1).

Samples which appeared severely degraded or those of low concentration (less than 30 ng/µl) were amplified using GenomiPhi V2 DNA amplification kit (Illustra™) as per manufacturers' instructions. The amplification reaction was incubated at 30^{o}C for 17 hours using a DNA Engine Tetrad 2 thermal cycler. The enzyme was then inactivated by heating to 65^{o}C for 10 minutes. Samples were then purified by ethanol precipitation. Absolute ethanol was added to each sample at a 2:1 ratio (EtOH:DNA, v/v) with 0.1 M NaOAc (pH 5.2) and incubated on ice for 40 minutes. DNA was collected by centrifugation at 2800 g for 15 minutes and then washed in 70% ethanol (EtOH:dH_2O, v/v). The DNA pellet was left to air dry for 5 minutes before resuspension in 20 µl low TE buffer (Section 2.2.2). The quantity and quality of DNA were then reassessed as prior to amplification. DNA yield ranged from 2 ng/µl to 7.7 µg/µl. Following the two methods of quantification, the concentration most consistent with the intensity of DNA visualised following gel electrophoresis was subsequently used.

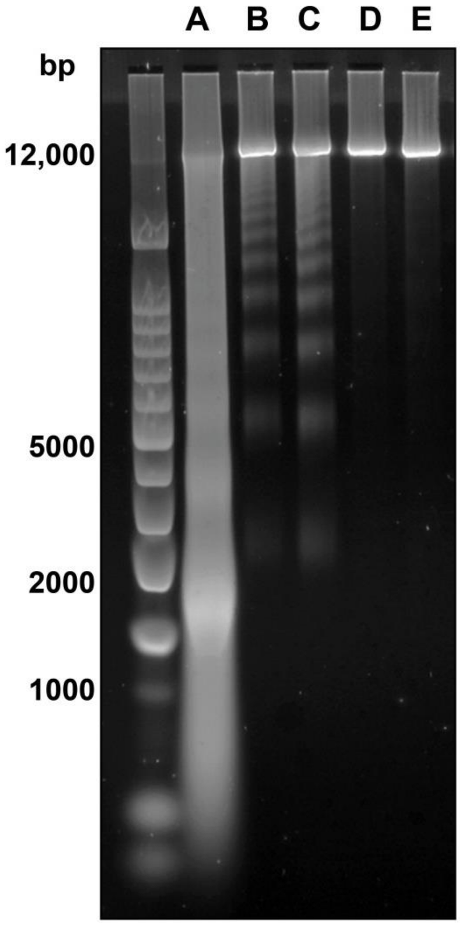


Figure 2.1. Agarose gel electrophoresis of genomic DNA

Example of resolution of different DNA samples following gel electrophoresis. A) Poor quality sample in which the DNA has degraded into multiple fragments of various sizes appearing as a large smear. B and C) Two samples show some evidence of degradation although the majority of DNA remains intact at 12 kb. D and E) Good quality samples with minimal degradation demonstrated by only fragments with a high molecular weight.

2.4.6 DNA methods

2.4.6.1 Whole exome capture

Samples were selected on the availability of DNA of adequate quantity and quality and also prioritised on whether parental DNA samples were available. To maximise the chances of identifying pathogenic variants, as many affected cases as possible were selected whilst also including samples from unaffected and affected parents and siblings where possible. In families with a strong degree of consanguinity (relationship between parents at least third cousins), only the affected case was sequenced as it was anticipated that the pathogenic variant would most likely be homozygous in these cases and thus sequencing parental DNA would be less informative compared to non-consanguineous families. In total, DNA samples from 166 individuals were prepared and sequenced, 154 of which were performed elsewhere (45 were prepared at the Wellcome Trust Sanger Institute (WTSI, Cambridge, UK) and 109 by Oxford Gene Technology (OGT, Oxfordshire, UK)).

Preparation of whole exome libraries from 12 samples was performed by myself using the SureSelect^{XT} target enrichment system for Illumina paired-end sequencing library (Agilent Technologies) as per manufacturers' instructions (Protocol version 1.3). All DNA quantification steps and assessment of library fragment size were performed using a 2100 Bioanalyzer with high sensitivity DNA kit (Agilent Technologies) unless otherwise stated. All purification steps were performed using Solid Phase Reversible Immobilization (SPRI) (Agencourt AMPure XP, Beckman Coulter) as per manufacturers' instructions, in DNA LoBind 1.5ml tubes (Eppendorf) using a Dynal DynaMag™-2 magnetic stand (Life Technologies). Beads were washed in freshly prepared 70% ethanol (EtOH:nuclease-free H_20, v/v) and DNA fragments eluted in nuclease-free H_20 (Life Technologies).

For each sample, 3 µg of genomic DNA was fragmented to approximately 200 bp (range 50 to 500 bp) by sonication using a Bioruptor® Plus (Diagenode) as per manufacturers' instructions. Genomic DNA was diluted to 0.01 µg/µl in TE buffer (Section 2.2.2) and sonication was performed in 30 second cycles for 2 to 3.5 hours at 4^{o}C. Size of fragments was initially assessed by agarose gel electrophoresis prior to SPRI bead purification at a 1:5 ratio (DNA:beads) to select 100-200 bp fragments. End repair and adapter ligation were performed using the Agilent SureSelect^{XT} reagent kit for Illumina HiSeq as per manufacturers' instructions. Libraries were amplified prior to exome capture by high fidelity PCR. Each reaction contained 15 µl of adapter ligated DNA library, 1 X KAPA HiFi™ HotStart ReadyMix (Kapa

BioSystems), 2.5 µl of PE 1.0 primer and indexing reverse primer (SureSelect^{XT} reagent kit) made up to 50 µl volume with nuclease-free H_20. Libraries were amplified using the following PCR programme:

1. Denaturation: 98^{o}C for 2 minutes

2. Cycle 1 (x5);
Denaturation: 98^{o}C for 30 sec
Annealing: 65^{o}C for 30 sec
Extension: 72^{o}C for 60 sec

3. Extension: 72^{o}C for 10 minutes

4. Hold: 10^{o}C for 10 minutes

500 ng of each amplified library was concentrated by dehydrating samples in a centrifugal vacuum concentrator (DNA 120 SpeedVac®, Thermo Scientific) at RT for 30 minutes at 300 g and re-suspended in 3.4 µl nuclease-free H_2O. Exome capture was performed using the Agilent SureSelect^{XT} Human All Exon V4 kit. This exome capture kit was used in the library preparation of all 166 samples. Block, hybridisation and bait solutions (SureSelect^{XT} reagent kit) were prepared as per manufacturers' instructions. Libraries were denatured after addition of 5.6 µl of blocking solution by heating to 95^{o}C for 5 minutes using a DNA Engine Tetrad 2 thermal cycler followed by incubation at 65^{o}C for 5 minutes. Immediately after, 13 µl of hybridisation solution was combined with 7 µl of bait solution and added to each library maintained at 65^{o}C followed by incubation at the same temperature for 24 hours. Hybridised fragments were isolated using Streptavidin-coated magnetic beads (Dynabeads® MyOne™ Streptavidin T1, Life Technologies) prepared as per manufacturers' instructions. Finally, the captured library was amplified indexing each library with a unique index sequence (SureSelect^{XT} reagent kit) to allow sequencing of multiple libraries in a single run. Each reaction contained 15 µl of captured DNA library, 1 X KAPA HiFi™ HotStart ReadyMix (Kapa BioSystems), 1µl each of post-capture forward and reverse primer with index (SureSelect^{XT} reagent kit) made up to 50 µl volume with nuclease-free H_20. Post-captured libraries were amplified using the following PCR programme:

1. Denaturation: 98^{o}C for 2 minutes

2. Cycle 1 (x12);

Denaturation: 98^{o}C for 30 sec
Annealing: 57^{o}C for 30 sec
Extension: 72^{o}C for 60 sec

3. Extension: 72^{o}C for 10 minutes

4. Hold: 10^{o}C for 10 minutes

Following final quantification, each library was dehydrated by centrifugation at 300 g in a vacuum concentrator at 43^{o}C for 30 minutes followed by reconstitution in low TE buffer to a concentration of 10 nM. Libraries were then combined in equimolar amounts into two pools. Libraries with similar average fragment length were pooled together and a 25% representation of A and C or T and G at each site in the unique index sequence were ensured in each pool.

2.4.6.2 Preparation of custom designed AmpliSeq™ libraries

An Ion AmpliSeq™ primer panel was custom designed using the online web tool (https://www.ampliseq.com, Life Technologies). All coding exons plus 25 bp into flanking introns of 20 genes were specified. The designed panel, based on the hg19 reference sequence, contained 686 amplicons in three pools with 98.5% coverage of the desired sequence. Genes (relevant to this thesis) included in the panel are listed in Table 2.1.

Table 2.1. Genes and their associated coverage in the custom designed AmpliSeq™ primer panel

Gene	*Coverage %*	*Gene*	*Coverage %*
SMC2	99.21	***NCAPD3***	99.78
SMC4	95.97	***XRCC4***	100
NCAPH	99.94	***PLK1***	100
NCAPH2	100	***AURKB***	100
NCAPG	96.24	***KIF4A***	100
NCAPG2	99.98	***USP2***	97.58
NCAPD2	99.38	***FAT1***	99.77

Libraries were prepared using the Ion AmpliSeq™ library kit 2.0 (Life Technologies) as per manufacturers' instructions (Ion AmpliSeq™ library preparation, v5.0). For each DNA sample, three individual PCR reactions were performed with each primer pool. Each PCR reaction consisted of 5 ng genomic DNA with 1 X primer pool and

1 X HiFi master mix in a 3.4 µl volume performed over 19 cycles. To prevent evaporation each reaction was performed in a 384-well plate (Roche) heat sealed with a foil lid (ALPS™ 35 manual heat sealer, Thermo Scientific). The three amplified primer pools were then combined prior to digestion of primer sequences. Reagent volumes in subsequent steps were scaled down according to the initial PCR reaction volume (1/6th of recommended reaction volumes used). Primer digest was performed at the following incubations; 50°C for 10 minutes, 55°C for 10 minutes and 65°C for 20 minutes. Unique Ion Xpress™ barcode adapters (Life Technologies), numbers 1-96, were ligated onto individual libraries to allow multiplexing of samples during sequencing. SPRI bead purifications were performed as described in Section 2.4.6.1 with the exception that bead immobilization was performed in a U-bottom 96-well plate (Greiner Bio-One) using a corresponding magnetic stand (DynaMag™-96 side magnet, Life technologies) and DNA eluted in low TE buffer. Following adapter ligation and purification, libraries were amplified as per manufacturers' instructions using the following PCR Programme;

1. Denaturation: 95°C for 5 minutes

2. Cycle 1 (x8);
Denaturation: 95°C for 15 sec
Annealing: 58°C for 15 sec
Extension: 70°C for 60 sec

3. Hold: 10°C for 10 minutes

Amplified libraries were then purified as per manufacturers' instructions and quantified as per Section 2.4.6.1. Serial dilutions of individual libraries was performed in low TE buffer (Section 2.2.2) to achieve a final concentration of 100 pM.

2.4.6.3 Array comparative genomic hybridisation (array-CGH)

Genomic DNA from six unaffected individuals (three males and three females) was combined in equi-molar amounts to create two reference pools of DNA, one for each sex. 500 ng of test sample and sex-matched reference genomic DNA was labelled with Cyanine-3 (Cy3) and Cyanine-5 (Cy5) dyes respectively using the NimbleGen Dual-Colour DNA labelling kit (Roche) as per manufacturers' instructions

(NimbleGen arrays user's guide: CGH and CNV arrays v8.1). The labelling was performed at half reaction volumes and samples were protected from the light where possible. Following denaturation, the labelling reaction was incubated overnight at 37°C before addition of the stop solution. Precipitated DNA was washed with freshly prepared 80% ethanol (EtOH:dH2O, v/v) at 4°C and dried by centrifugation at 300 g in a vacuum concentrator (DNA 120 SpeedVac®, Thermo Scientific) at 43°C for 5 minutes. DNA was rehydrated in 25 µl dH_2O for 20 minutes at RT prior to quantification using a NanoDrop 1000 UV-Vis Spectrophotometer (Thermo Scientific, Section 2.4.7). 20 µg of the labelled test sample was then combined with an equal amount of labelled (sex-matched) reference DNA and dehydrated as above for 15-20 minutes prior to re-suspending in 3.3 µl of a unique NimbleGen sample tracking control (Roche). Hybridisation solutions (NimbleGen hybridisation kit) were prepared and added to the sample as per manufacturers' instructions prior to loading onto a 12 x 135 K array slide (v3.1, Roche). Hybridisation was then carried out by incubation at 42°C for 72 hours using a NimbleGen hybridisation system 4 (Roche). Array slides were washed (NimbleGen wash buffer kit) and dried as per manufacturers' instructions and then immediately scanned using laser emission wavelengths of 532 nm and 635 nm (NimbleGen MS200 scanner). NimbleScan software (v1.2) was used to calculate the relative intensity at each probe. The CGH-segMNT module of NimbleScan was used for the analysis with a minimum segment length of 5 probes and averaging window of 130 kb. Regions of copy number variation were identified using SignalMap software (v1.9, NimbleGen) and compared to the Database of Genomic Variants (http://projects.tcag.ca/variations) to exclude polymorphic CNVs. SignalMap assigns the probe position to the hg18 reference genome. For the purposes of this thesis, genomic coordinates have been translated to the hg19 reference to maintain consistency in results.

2.5 Sequencing methods and variant detection

2.5.1 ABI dye-terminator sequencing

ABI sequencing was undertaken by the Institute of Genetics and Molecular Medicine (IGMM) sequencing service. Purification of PCR products was performed using

1.2X SPRI beads (Agencourt AMPure XP, Beckman Coulter) as per manufacturers' instructions. Beads were washed as described in Section 2.4.6.1 and PCR product eluted in 15 µl dH_2O. Dye termination cycle sequencing reactions were performed using BigDye® Terminator v3.1 cycle sequencing kit (Life Technologies) as per manufacturers' instructions. A standard reaction contained 0.0625 X BigDye® ready reaction premix, 1 X BigDye® sequencing buffer, 0.165 µM of both forward and reverse primers and 2 µl of purified PCR product made up to 10 µl volume with dH_2O. For sequencing reactions, either gene specific or N13 primers were used (Appendix I). For sequencing plasmids, 150 ng of plasmid plus 0.34 µM of both forward and reverse M13 primers (Appendix I) were heated to 96^{o}C for 2 minutes in 5 µl dH_2O followed by addition of 0.125 X BigDye® ready reaction premix, 0.65 X BigDye® sequencing buffer made up to a final volume of 10 µl dH_2O. Cycle sequencing reactions of both PCR products and plasmids were then performed under the following conditions: 25 cycles of 96^{o}C for 30 seconds, 50^{o}C for 15 seconds and 60^{o}C for 4 minutes.

Samples were precipitated in 24 µl 95% ethanol (EtOH:dH_2O, v/v) with 0.23 M NaOAc (pH 4.8) and washed twice with 100 µl of 70% ethanol (EtOH:dH_2O, v/v). DNA was precipitated by centrifugation at 3000 rpm for 30 minutes (Allegra™ X-12 centrifuge, Beckman Coulter) and the pelleted DNA left to air dry for 5 minutes prior to resuspension in 15 µl Hi-Di formamide (Life Technologies). Samples were then denatured by heating to 95^{o}C for 5 minutes and immediately cooled on ice.

Products were capillary-sequenced using an ABI 3730 (48 capillary) or 3130xl (16 capillary) Genetic Analyzer. Alignment of sequencing reads and detection of variants was performed using Mutation Surveyor® software (v3.30) (Softgenetics) or Sequencher 4.10.1 (Gene Codes Corp) with default settings. Reference sequences for individual genes are listed in Appendix II.

2.5.2 Illumina

2.5.2.1 Sequencing

100 bp paired-end sequencing of pooled whole exome libraries (166 in total) was performed using the Illumina® HiSeq 2000 sequencing system. 7-8 pooled libraries

were sequenced on an individual lane. This was conducted at three different centres: OGT (n = 109), WTSI (n = 45) and the National High-throughput DNA sequencing centre in Denmark (University of Copenhagen) (n = 12).

2.5.2.2 Alignment and variant calling

FASTQ files containing sequencing reads for 158 libraries were aligned and recalibrated using a pipeline designed by Alison Meynert, James Prendergast and Martin Taylor (MRC HGU, IGMM, Edinburgh). This is based on a previously established method (DePristo *et al.*, 2011), an overview of which is shown in Figure 2.2. I performed alignment and recalibration of the remaining eight samples (all sequenced in Denmark) using the same tools to ensure consistency in analysis (script supplied in Appendix III). Sequencing reads were initially aligned to the hg19 reference genome (ftp://hgdownload.cse.ucsc.edu, 11.04.12) using the Burrows-Wheeler Alignment tool (BWA, v6.2) (Li *et al.*, 2009), a high speed, short read aligner algorithm. Bases with a quality (PHRED) score of less than 10 were trimmed from the 3' end of each read prior to alignment. Unmapped reads were then further aligned with the Stampy alignment tool (v1.0.21) (Lunter *et al.*, 2011), a hybrid mapping algorithm with higher sensitivity. Any reads mapping to identical positions in the genome (duplicates) were marked using Picard tools (v1.79) (http://picard.sourceforge.net/) and ignored in downstream variant detection to reduce errors from PCR amplification (Meynert *et al.*, 2013). Coverage of target exons was determined using the GATK DepthOfCoverage tool (Genome Analysis Toolkit v2.4.9) (https://www.broadinstitute.org/gatk/) and a BED file format listing the genomic coordinates of designed baits supplied by the manufacturer. Realignment around known and suspected indels was performed using the GATK RealignerTargetCreator and IndelRealigner using the following reference data sets for known indels: Single Nucleotide Polymorphism Database (dbSNP) version 137 (Smigielski *et al.*, 2000), Phase I release of indels from the 1000 Genomes Project (1KG) (Clarke *et al.*, 2012), and a high confidence combined set of indels from Mills *et al.*, (2011a) and the 1KG project (ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/2.5/hg19/). Recalibration of base quality scores was performed using the GATK BaseRecalibrator and PrintReads tools.

Variant calling and recalibration was then performed on all 166 BAM (Binary Alignment Map) files by Alison Meynert using the GATK Unified Genotyper, VariantAnnotator and VariantRecalibrator tools (Figure 2.2). Recalibration of called variants was performed using reference datasets of validated variants obtained from the 1KG (1000G_omni2.5) and HapMap (HapMap_3.3) sequencing projects (Frazer *et al.*, 2007).

All variants were initially annotated with predicted protein consequence using SnpEff (v3.2a) with the Ensembl 70 human annotations (Cingolani *et al.*, 2012) and all variants not occurring in a gene were removed (Alison Meynert). SnpEff also annotates variants with the corresponding 'rs#' identifier (if present) in dbSNP. All remaining variants were then compiled into an SQL database and non-synonymous coding variants were additionally annotated using dbNSFP v2.0 (Alison Meynert) (Liu et al, 2013) which details the predictions from a further seven different programmes (see Chapter 3, Table 3.3). All variants were additionally annotated using the Alamut-HT (Interactive Biosoftware, Rouen, France) programme (performed by Louise Bicknell, MRC HGU, IGMM, Edinburgh) which allows batch processing of large datasets through five splice prediction programmes (further details provided in Section 3.3.2). Variants were then annotated with corresponding minor allele frequencies in the 1KG (Clarke *et al.*, 2012) and NHLBI GO Exome Sequencing Project (ESP) databases (Seattle, WA: http://evs.gs.washington.edu/EVS/) (Alison Meynert) and finally, with the full gene name and any associated disease (OMIM morbid) acquired from the Ensembl database (http://www.ensembl.org/biomart/martview/). I then performed all downstream filtering of variants using SQL scripts (Appendix IV) and the SQLyog MySQL graphical user interface tool (Webyog, Santa Clara, CA, USA).

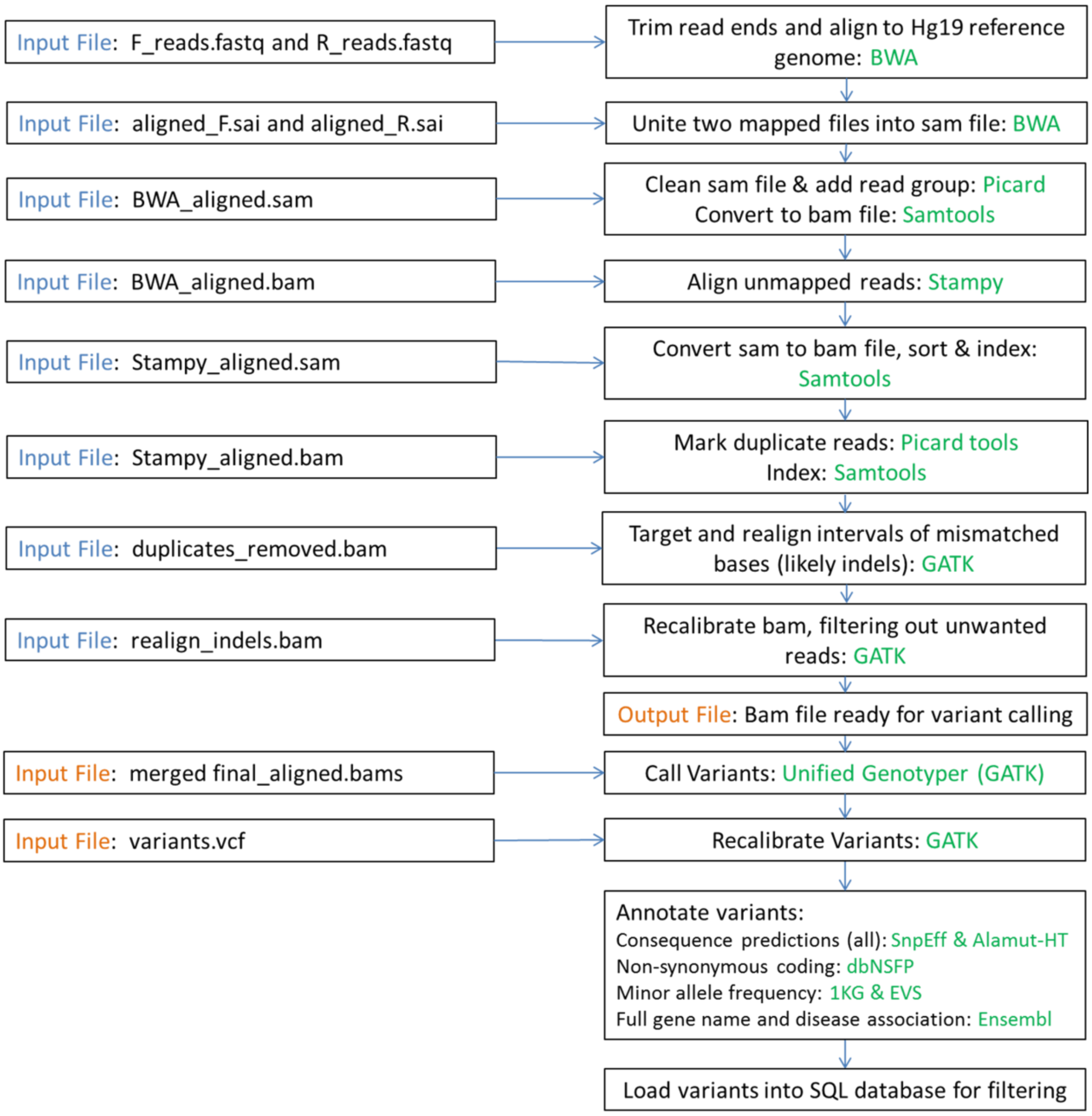

Input File: F_reads.fastq and R_reads.fastq
Trim read ends and align to Hg19 reference genome: BWA
Input File: aligned_F.sai and aligned_R.sai
Unite two mapped files into sam file: BWA
Input File: BWA_aligned.sam
Clean sam file & add read group: Picard
Convert to bam file: Samtools
Input File: BWA_aligned.bam
Align unmapped reads: Stampy
Input File: Stampy_aligned.sam
Convert sam to bam file, sort & index: Samtools
Input File: Stampy_aligned.bam
Mark duplicate reads: Picard tools
Index: Samtools
Input File: duplicates_removed.bam
Target and realign intervals of mismatched bases (likely indels): GATK
Input File: realign_indels.bam
Recalibrate bam, filtering out unwanted reads: GATK
Output File: Bam file ready for variant calling
Input File: merged final_aligned.bams
Call Variants: Unified Genotyper (GATK)
Input File: variants.vcf
Recalibrate Variants: GATK
Annotate variants:
Consequence predictions (all): SnpEff & Alamut-HT
Non-synonymous coding: dbNSFP
Minor allele frequency: 1KG & EVS
Full gene name and disease association: Ensembl
Load variants into SQL database for filtering

Figure 2.2. In-house alignment and variant calling pipeline for Illumina data. Samtools (v1.16) (http://samtools.sourceforge.net/).

2.5.2.3 Copy number variation (CNV) analysis

CNV analysis was performed in-house on all exome datasets of affected individuals by Mihail Halachev (MRC HGU, IGMM, Edinburgh) using ExomeCNV (Sathirapongsasuti *et al.*, 2011). ExomeCNV creates a reference read depth across each exon (as set by the bed file of designed capture baits obtained from the manufacturer) by either using a single reference sample or by pooling multiple samples that have been prepared and sequenced in a similar manner to the test sample. The presence of a CNV may then be identified by detecting a deviation in the depth of coverage in the test sample from the paired reference. This is initially performed on a per exon basis and then across widening chromosome segments with sequential merging using the circular binary segmentation algorithm (Olshen *et al.*, 2004). The latter improves power in CNV detection by combining depth of coverage over consecutive exons. Minimum specificity and sensitivity were set at 0.999 in exon-wise calling and 0.99 in segment-wise calling with an admixture rate of 0. The smallest region that has sufficient coverage to create enough power to detect a CNV is called and thus single exon CNVs will not be rejected if surrounded by copy neutral exons.

Two reference datasets were created by pooling the depth of coverage obtained at each exon in five parental samples prepared and sequenced at WTSI and 55 parental samples prepared and sequenced at OGT. The median read depth for each exon was then used as the reference to compare with each affected sample generating a log2 ratio for each exon. All singleton samples were matched according to where they were sequenced except for those samples prepared in Edinburgh and sequenced in Denmark which were matched to the OGT reference dataset. For all samples in which parents were additionally sequenced, the analysis was performed twice comparing the read depth at each exon to that of each parent. I then compiled all deleted regions into an SQL database and performed downstream filtering using SQL scripts (Appendix IV).

2.5.3 Ion Proton™ and Ion Torrent™

Two pools containing 96 uniquely barcoded libraries (Ion Xpress™ barcode adapters 1-96, Life Technologies) in equal concentrations were prepared (192 individual samples in total). Libraries were then combined in equal volumes and quantified using a 2100 Bioanalyzer. Each pool was sequenced at the Wellcome Trust Clinical Research Facility (WTCRF, Edinburgh) using an Ion P1™ chip kit (v2) and an Ion Proton™ semiconductor sequencing system (Life Technologies). Sequencing reads were aligned to the hg19 reference genome and variants called using the Ion Torrent™ Suite Software (v4.0.2) (also performed at the WTCRF). Variants were then annotated as per Illumina data by Mihail Halachev (MRC HGU, IGMM, Edinburgh). Equal volumes of the remaining 6 Ampliseq libraries were pooled and sequenced using an Ion 316™ chip kit (v2) and Ion Torrent™ semiconductor sequencing system (Life Technologies) performed by IGMM technical services (University of Edinburgh). FASTQ files generated from the Ion Torrent™ sequencer were aligned and analysed by myself using NextGENe (v2.2.0) (Softgenetics) with the following alignment settings: matching requirements ≤ 12 bp and ≥ 65%, filter out variants with mutation percentage ≤ 20, SNP allele ≤ 3 counts and total coverage ≤ 5 X except for homozygous variants, balance ratios ≤ 0.1 and frequency ≤ 80%. Reference sequences for individual genes are listed in Appendix II.

2.6 Protein Methods

2.6.1 Whole cell protein extraction

Protein was extracted from cultured primary fibroblasts by detergent mediated cell lysis. Harvested cell pellets were re-suspended in whole cell extract (WCE) buffer (Section 2.2.2) with 0.025 U benzinase nuclease (Novogen) and incubated on ice for 30 minutes. WCE buffer volume was matched to the number of harvested cells, for example, 3 x 10^6 fibroblasts were re-suspended in 200 µl of buffer. Cellular debris was pelleted by centrifugation at 16000 g for 10 minutes at 4^oC and the supernatant collected and stored until required at -80^oC.

2.6.2 Protein quantification

A Bradford assay was performed to determine the protein concentration in whole cell extracts using the Bio-Rad protein assay kit (Bio-Rad laboratories Ltd). 2 µl of sample was added to 1 ml of 1 X dye reagent and incubated for 5 minutes at RT. The A_{595} was measured using a WPA* UV1101 Biotech Spectrophotometer (Colonial Scientific) and the protein concentration calculated from the absorbance reading using a standard curve (prepared by Margaret Harley, MRC HGU, IGMM, Edinburgh) generated from BSA standards (concentration range 0.2-1.5 mg/ml) provided by the manufacturer. Three measurements were taken for each sample and the average used for quantification.

2.6.3 SDS-PAGE and Western immunoblotting

Sodium dodecyl sulphate polyacrylamide electrophoresis (SDS-PAGE) was performed to separate proteins according to size using the NuPage®Novex gel system (Life Technologies) and pre-cast 4-12% Bis-tris 1 mm gels (Life Technologies). 30 µg of protein from each sample was denatured in 1 X sample loading buffer (Life Technologies) with 0.05 M DTT by heating to a minimum temperature of 70°C for 20 minutes. Denatured samples were then loaded onto the gel (maximum 40 µl per well) alongside the precision plus protein standard (Life Technologies). Electrophoresis was performed in 1 X MOPs SDS running buffer (Life Technologies) with 0.25% (v/v) NuPAGE antioxidant at 200 volts for 60 minutes.

Following separation, electrophoretic transfer of proteins to a nitrocellulose blotting membrane (Amersham) was performed using a Mini-PROTEAN® 3 Cell system (Bio-Rad). The transfer was performed in 1 X western transfer buffer (Section 2.2.2) at 100 V for 60 minutes. The membrane was then blocked in 5% Marvel (Premier foods) in 1 X TBST for 1 hour at RT with constant agitation. Primary antibodies (Table 2.2) were diluted in fresh blocking solution at the relevant concentration. Following overnight incubation at 4°C with the primary antibody, the membrane was washed three times in 1 X TBST for 5 minutes. A second incubation was then performed with the appropriate HRP-labelled secondary antibody (Table 2.2) also diluted in fresh blocking solution for 1 hour at RT with constant agitation.

Table 2.2. Details of antibodies used in western immunoblotting

Antibody	*Dilution*	*Species*	*Catalogue Nº*	*Company*
Primary				
NCAPD3	1:250	Rabbit	A300-604A	Bethyl labs
NCAPG2	1:500	Rabbit	A300-605A	Bethyl labs
NCAPH2	1:500	Rabbit	A302-276A	Bethyl labs
NCAPH	1:500	Rabbit	A300-603A	Bethyl labs
Actin	1:1000	Rabbit	A2066	Sigma-Aldrich
Secondary				
HRP-linked anti-rabbit	1:5000	Goat	7074	Cell signalling

HRP was detected using an ECL detection kit (Amersham Biosciences) according to the manufacturers' instructions. ECL solutions A and B were combined in equal volumes (2 ml total volume for a 20 cm^2 membrane) and mixed vigorously prior to adding to the membrane ensuring complete coverage of the membrane surface. The membrane was incubated in the ECL detection solution for 1 minute at RT. Excess solution was then removed and the membrane placed between two acetate sheets in a Hypercassette™ (Amersham). In the absence of light, the membrane was exposed to photographic film (Kodak Biomax XAR film) and developed using a Konika SRX-101A Developer. The developed film was scanned and the image uploaded into Adobe Photoshop CS6 for presentation.

2.7 Microscopy Methods

2.7.1 Fixation of cells

To visualise mitosis, primary fibroblasts were cultured on 16 mm glass coverslips (thickness 0.13 - 0.17 mm, Thermo Scientific) and fixed at 60-80% confluency. 0.5% Triton™ X-100 (Sigma-Aldrich) in PHEM buffer (Section 2.2.2) was pre-warmed to 37^{o}C prior to the addition of 4% paraformaldehyde (Thermo Scientific). The fixative solution was then immediately added to the cells and incubated for 10 minutes at RT followed by removal of the fixative and 2 x 5 minute washes in PBS. Fixed cells were stored until required at 4^{o}C.

To examine chromosome morphology, primary fibroblasts were cultured to 80% confluency in a 25cc flask (CELLSTAR®, Sigma-Aldrich). Prior to fixation cells were incubated at 37^{o}C at 3% O_2 and 5% CO_2 for 30 minutes with 70 ng/ml colcemid (Sigma-Aldrich) in culture media to disrupt microtubules. Cells were washed in PBS and detached by trypsinisation followed by resuspension in culture media and centrifugation at 1200 rpm (Allegra™ X-22 centrifuge, Beckman Coulter) for 5 minutes to pellet the cells. Cells were gradually re-suspended in 5mls of hypotonic buffer (Section 2.2.2) added in a drop-wise manner with gentle flicking of the tube and incubated for 10 minutes at RT. Centrifugation was repeated as previously and supernatant removed. Cells were then fixed by the addition of 5 ml of freshly prepared methanol:acetic acid, 3:1 (v/v), at 4^{o}C in a drop-wise manner.

Centrifugation and re-suspension in fixative was repeated a further two times reducing the volume of methanol fixative each time to achieve a final suspension volume of 1 ml.

2.7.2 Immunostaining

PFA fixed cells on coverslips were blocked by incubation in 1% (v/v) bovine serum albumin (Sigma-Aldrich) in PBS at 37^{o}C in a humidity chamber for 30 minutes. Incubations with primary and secondary antibodies at the following dilutions were performed under the same conditions in fresh blocking solution;

Primary antibody: α-tubulin (B512) antibody (Cat. Number T6074, Sigma-Aldrich) diluted to 1:1000.

Secondary antibodies: Alexa Fluor 488-linked anti-mouse antibody at 1:500 dilution (Cat. Number A11029, Life Technologies) and 4´6-diamidino-2-phenylindole (DAPI) (Fisher Scientific) at 1 µg/ml.

Stained cells were washed three times in PBS for 5 minutes after incubation with primary and secondary antibodies. Coverslips were mounted in Vectorshield® mounting medium (Vector Laboratories). Methanol fixed cells were dropped onto a microscopy slide from a minimum height of 0.25 m. Following evaporation of methanol cells were mounted in Vectorshield® mounting medium with DAPI. All stained slides were stored at 4^{o}C protected from light prior to microscopy.

2.7.3 Microscopy

Imaging of fixed cells was performed using a Zeiss Axioplan 2 widefield fluorescence microscope with an objective lens mounted PIFOC collar. Images were obtained in multiple Z planes using a 63 X 1.4 plan apochromat objective (Zeiss) and fluorescence filters for DAPI (359 nm excitation, 461 nm emission) and FITC (489 nm excitation and 508 nm emission). Image capture and acquisition was performed with an ORCA-ER camera (Hamamatsu) using Volocity software (PerkinElmer). To improve resolution, images were deconvolved using constrained iterative restoration with Volocity software. Metaphase chromosomes were imaged in a single Z plane using a 100 X 1.4 plan apochromat objective (Zeiss) and DAPI filter. Image capture and acquisition was obtained using a CoolSnap HQ2 camera (Roper Scientific) and iVision software. Scale of images in iVision software (IPLab) was determined using a script designed by Matt Pearson (MRC HGU, IGMM, Edinburgh) in Volocity software.

2.7.4 Flow cytometry

3 x 10^6 cells were harvested by trypsinisation and then washed by the addition of 10 mls PBS added in a drop-wise manner. Cells were pelleted by centrifugation at 1200 rpm (Allegra™ X-22 centrifuge, Beckman Coulter) for 10 minutes and supernatant gently removed followed by fixation of cells in 900 µl 70% EtOH/dH_2O at 4^oC added in a drop-wise manner. Fixed cells were stored at -20^oC. At least 60 minutes prior to flow cytometry, cells were thawed and washed twice in PBS before being re-suspended in 100µg/ml RNase A and 50 µg/ml propidium iodide. Stained cells were incubated on ice and protected from light. Flow cytometry was performed by Elizabeth Freyer (MRC HGU, IGMM, Edinburgh) using a FACScalibur (BD Biosciences). Subsequent data analysis was performed using FlowJo software (v7.6.5, Tree Star).

2.8 Data Analysis

Graphs were generated and statistical analysis performed using GraphPad Prism v6 (GraphPad Software). Statistical comparisons of normally distributed continuous data were performed using the two tailed Students unpaired t-test and comparisons of categorical data performed using the Fisher’s exact test or Chi squared test with Yates correction. Statistical comparison of three or more groups of non-normally distributed data was performed using the non-parametric Kruskal-Wallis test followed by Dunn’s multiple comparison test.

Chapter 3: Diagnosis and gene discovery in MPD by whole exome sequencing (WES)

3.1 Introduction

In this Chapter, the development of a filtering workflow for the identification of pathological mutations in this cohort of MPD patients is described. The pipeline integrates many of the filtering strategies previously employed by other sequencing projects as well as incorporating new methods to remove sequencing errors and prioritise candidate disease genes. This led to an increase in diagnosis within this diverse patient group as well as the identification of novel disease genes.

3.2 Description of samples and sequencing data

3.2.1 Samples selected for WES

Analysis of WES was performed in 95 out of a possible 114 families within the MPD cohort who did not yet have a molecular diagnosis. 166 individuals were sequenced of which 102 were affected (Table 3.1). In 57 families only the affected case was sequenced due to the presence of consanguinity or lack of adequate DNA available from other family members. Families were then categorised as singletons, pairs, trios or quads depending on which additional family members' were also sequenced (Table 3.1). Data from eight families in which the causative variant(s) had already been identified were included as positive controls for validating the filtering pipeline.

Table 3.1. Family relationships of WES samples

	Number of samples
Individuals sequenced	166
Affected individuals sequenced	102
Singletons without any additional family samples	57
Pairs: Singleton with affected sibling (pairs)	6
Trios: Singleton with both parental samples (trios)	30
Trios(m): Singleton with sample from mother & unaffected sibling	1
Quads: Affected sibling pairs with parental samples (quads)	1
Affected cases from consanguineous families*	36
Number of cases with known diagnosis	8

* Relationship between parents 3rd cousin or closer.

3.2.2 Quality analysis of sequencing data

Overall, a mean of 82,370,661 sequencing reads (range 34,961,608 to 203,909,410) was generated per sample, of which an average of 92% uniquely mapped to the reference genome (range 74-98%). On-target coverage with a read depth of at least 15X ranged from 79% to 96% (mean=89%), with average on-target read depth ranging from 37 to 119 (mean=72) per sample.

Comparison between samples prepared and sequenced in the three different locations (Table 3.2) identified significant differences in the number of sequencing reads, mean read depth and on-target coverage between each group indicating technical variation in the efficiency in both library capture and sequencing (Table 3.2). WTSI samples were the earliest libraries prepared and showed the lowest capture efficiency with less on-target coverage indicating a less efficient protocol was used at this time. To compensate for early inadequacies in capture methods, these libraries were sequenced to a much higher read depth. In contrast, libraries prepared and sequenced a year later at OGT showed significantly improved on-target coverage and a higher proportion of useable reads. This likely reflects a lower number of duplicate reads arising from bias in PCR amplification and an increase in the number of mappable reads indicating a higher quality of sequencing was achieved. Improved uniformity in flowcell loading as well as sequencing software is likely to have contributed to the latter. Notably, those libraries prepared manually by myself in Edinburgh followed by sequencing in Denmark were comparable to OGT samples in terms of capture efficiency although read depth was higher in this group as one less sample was present in each sequencing run.

Table 3.2. Quality metrics on sequencing output

Per sample	***WTSI*** ***Mean*** *(+/- s.d.)*	***OGT*** ***Mean*** *(+/- s.d.)*	***Edinburgh/Denmark*** ***Mean*** *(+/- s.d.)*	***Statistical significance*** *P value*
Date Performed	July-Sept, 2011	July-Nov, 2012	June-Nov, 2012	
Total reads	**158,813,572** (31,682,839)	**53,130,285** (24,375882)	**61,309,819** (9,129,108)	<0.05
% reads uniquely mapped to hg19	**87** (4.5)	**94** (2.9)	**92** (2.8)	<0.001 (WTSI vs OGT)
Mean read depth	**98** (13.7)	**61** (10.0)	**74** (11.5)	<0.05
On-target coverage %	**85** (1.1)	**90** (3.4)	**93** (3.4)	<0.05

On-target coverage indicates percentage of targets sequenced to a depth of at least 15X. Each subgroup was compared to all others using a non-parametric Kruskal-Wallis test followed by a Dunn's multiple comparison test. P value indicates results for all three comparisons performed for that parameter except where stated. *Abbreviations: WTSI=Wellcome Trust Sanger Institute, OGT=Oxford Gene Technology.*

3.3 Description of filtering pipeline

In total, 11,540,535 variants were identified in the 102 affected individuals sequenced following the variant calling process described in Chapter 2. Therefore to identify pathogenic variants, I devised and implemented a filtering pipeline using SQL scripts (Appendix IV). An overview of this pipeline is shown in Figure 3.1 and each step is described in further detail in the next section.

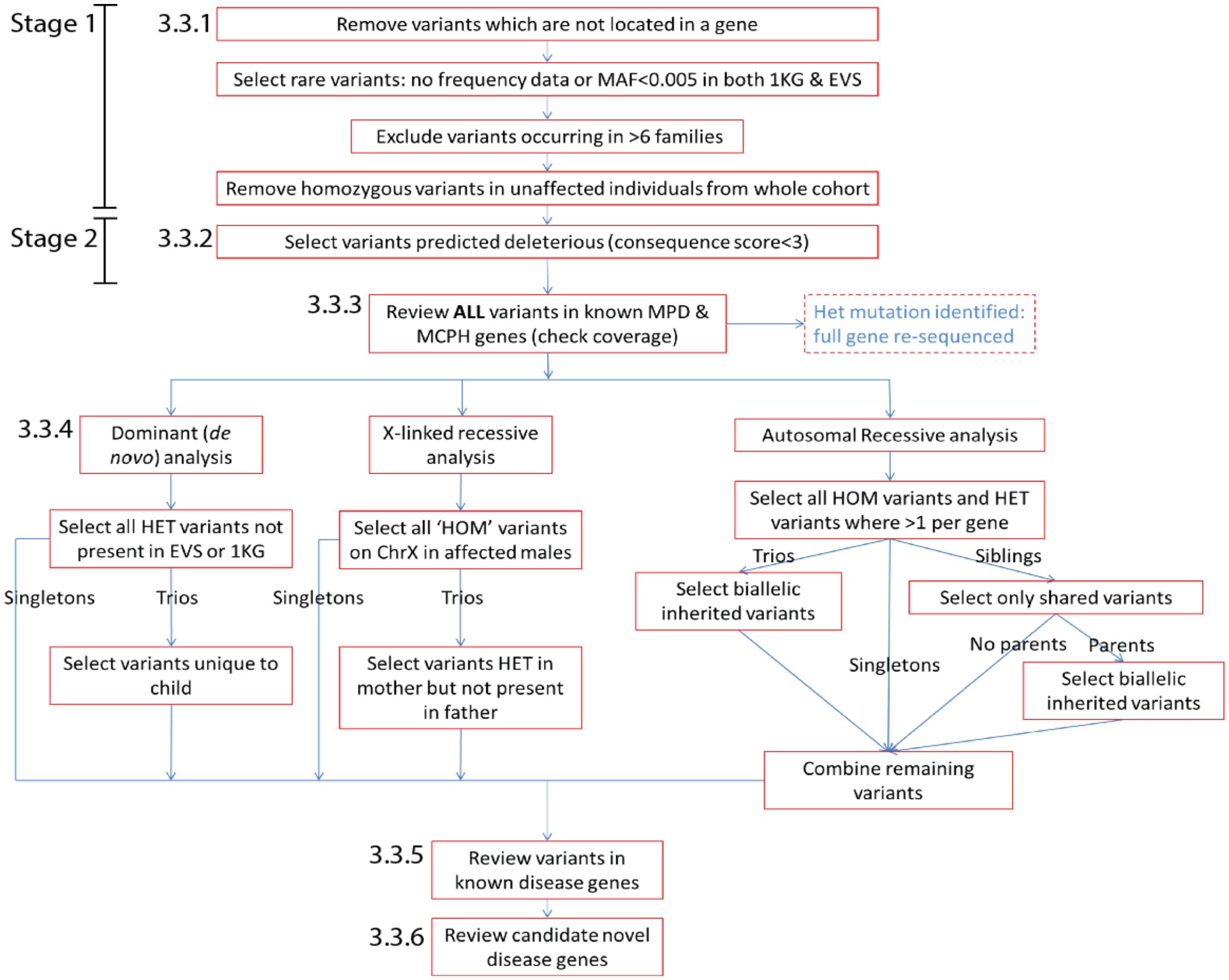

Stage 1
3.3.1
Remove variants which are not located in a gene
Select rare variants: no frequency data or MAF<0.005 in both 1KG & EVS
Exclude variants occurring in >6 families
Remove homozygous variants in unaffected individuals from whole cohort
Stage 2
3.3.2
Select variants predicted deleterious (consequence score<3)
3.3.3
Review ALL variants in known MPD & MCPH genes (check coverage)
Het mutation identified: full gene re-sequenced
3.3.4
Dominant (de novo) analysis
X-linked recessive analysis
Autosomal Recessive analysis
Select all HET variants not present in EVS or 1KG
Select all 'HOM' variants on ChrX in affected males
Select all HOM variants and HET variants where >1 per gene
Trios
Siblings
Select biallelic inherited variants
Select only shared variants
Singletons
Trios
Singletons
Trios
No parents
Parents
Singletons
Select variants unique to child
Select variants HET in mother but not present in father
Select biallelic inherited variants
Combine remaining variants
3.3.5
Review variants in known disease genes
3.3.6
Review candidate novel disease genes

Figure 3.1. Overview of variant filtering pipeline.

List of known MPD & 1º MCPH genes provided in Appendix V.

3.3.1 Stage 1 Filter: Excluding common variants and likely sequencing errors

Prior to filtering, all variants not occurring within a gene were removed from the SQL database (Alison Meynert, MRC HGU, IGMM, Edinburgh). This included all variants annotated as `intergenic` by SnpEff or those which were not annotated with a corresponding gene name removing many off target reads which are likely to have low read depth and therefore a higher false positive variant rate adding to variant 'noise' (Hoischen *et al.*, 2010, Meynert *et al.*, 2013). This resulted in an initial starting list of over 600,000 variants in more than 30,000 candidate genes across all samples corresponding to approximately 100,000-150,000 variants per family. I then performed all downstream filtering of variants using SQL scripts (Appendix IV).

To select for rare variants, those annotated with a minor allele frequency of 0.005 or less in either the 1KG or the EVS databases were filtered out. MPD is a very rare disorder likely to occur in around 1 in every million people therefore any pathogenic variant, assuming Hardy–Weinberg equilibrium (Stern, 1943) and a recessive model, would be expected at an allele frequency of 0.001 or less. However, a less stringent threshold was used in order not to miss a more common causative variant occurring in combination with another exceedingly rare pathogenic variant as the alternate allele. An extreme example of this is TAR syndrome (Albers *et al.*, 2012) which results from a deletion on one allele in combination with a SNP on the other allele. Additionally, allele frequency for rare variants have wide confidence intervals given current sample size in the 1KG and EVS cohorts leading to inaccuracies in the reported minor allele frequency.

As all samples were sequenced on the same Illumina HiSeq platform it was anticipated that false positive variants arising through sequencing errors, such as homopolymer runs, will likely occur at similar sites across multiple samples. To remove such technical artefacts, filtering was performed to exclude variants that occurred recurrently in those sequenced. It was reasoned that as the MPD cohort is clinically heterogeneous and from a wide range of ethnic backgrounds it was unlikely that the majority of families would share the exact same pathogenic variant.

Mutations in *PCNT* (MOPDII) are the most common cause of MPD occurring in approximately 20% of patients in this cohort. This is followed by mutations in *RNU4ATAC* (MOPDI) occurring in less than 3% of cases. Furthermore, recurrent mutations are rare in both disorders. It was therefore anticipated that further disease causing variants in novel genes were likely to be present at low frequency in the patients sequenced, therefore variants occurring in 6 families (5%) or more were excluded. Assuming a model of full penetrance, variants also highly unlikely to be pathogenic were those occurring in the homozygous state in unaffected individuals and therefore such variants were also removed on filtering.

This first stage of filtering removed 68% of variants (approximately 400,000 in total) but those remaining still covered 27,620 possible candidate genes, only 10% less than the list prior to filtering.

3.3.2 Stage 2 Filter: Selecting variants predicted to deleteriously affect protein function

To prioritise only those variants likely to have a deleterious effect on the protein, I graded consequence predictions generated by SnpEff, dbNSFP and Alamut HT on a scale of 1-7 (Table 3.3), the lowest score reflecting those variants predicted to have the most deleterious consequence. Non-synonymous coding variants were graded with a score of 2 if any one of the seven programmes used by dbNSFP predicted it to be deleterious. Alamut-HT software uses 5 splice prediction programmes (MaxEntScan, NNSPLICE, Human Splicing Finder, SpliceSiteFinder and GeneSplicer) to generate a value corresponding to the strength of predicted effect of the variant on either the nearest natural splice site or splicing in the local vicinity. Those variants predicted either to create a splice site or alter an existing splice site (parameters set: effect on nearest splice site >0.5 or <-0.5 or local effect='new' or 'strongly activated') were given a score of 2.5. Only variants with a consequence score of 2.5 or less were taken forward for further filtering excluding a further 24% of variants (92% in total) reducing the number of candidate genes in the cohort to nearly 15,000 (50% of the initial candidate gene list) with an average of 576 variants per family (range = 321-1,313).

Table 3.3. Classification and scoring of variants according to predicted protein consequence

Score	***Consequence prediction annotation with SnpEff***	***Prediction programmes used by dbNSFP***	***Non-synonymous prediction annotation***
1	STOP_GAINED FRAME_SHIFT SPLICE_SITE_DONOR SPLICE_SITE_ACCEPTOR START_LOST		
2	NON_SYNONYMOUS_START NON_SYNONYMOUS_CODING CODON_CHANGE_PLUS_CODON_INSERTION CODON_INSERTION STOP_LOST CODON_DELETION CODON_CHANGE_PLUS_CODON_DELETION UTR_5_DELETED	SIFT_score Polyphen2_HDIV Polyphen2_HVAR LRT Prediction Mutation Taster Mutation Assessor FATHMM_score	D D or P D or P D A or D High, medium D (or not annotated)
2.5	INTRON SYNONYMOUS_CODING	*If predicted to alter splicing by Alamut-HT*	
3	NON_SYNONYMOUS_START NON_SYNONYMOUS_CODING	SIFT_score Polyphen2_HDIV Polyphen2_HVAR LRT Prediction Mutation Taster Mutation Assessor FATHMM_score	T B B N or U N or P Low, neutral T
4	INTRON STOP_LOST		
5	SYNONYMOUS_CODING START_GAINED		
6	UTR_5_PRIME UTR_3_PRIME		
7	UPSTREAM DOWNSTREAM SYNONYMOUS_STOP EXON(in non-coding exon)		

Abbreviations used in annotation by various non-synonymous prediction programmes: SIFT and FATHMM_score; D=damaging, T=tolerated. Polyphen2; D=probably damaging, P=possibly damaging, B=benign, LRT prediction; D=deleterious, N=neutral, U=unknown, Mutation taster; A=disease causing automatic, D=disease causing, N=polymorphism, P=polymorphism automatic.

3.3.3 Variants in known MPD and 1° MCPH genes

Following the above two filtering steps, all variants in genes in which mutations are known to cause MPD and 1° MCPH (list in Appendix V) were manually reviewed in IGV and Alamut software (see Section 3.3.7). Additionally, if a deleterious heterozygous mutation (consequence score =1) was identified in an affected individual in one of these genes then the whole gene was resequenced by capillary sequencing to ensure that a second variant had not been missed by WES. This allowed patients with mutations in known disease genes to be readily identified early in the filtering process and excluded from further analysis. In total, 169 variants in 80 families were present in known MPD and 1° MCPH genes and in nine patients causality was attributed to these genes (Section 3.5).

3.3.4 Analysis by mode of inheritance

Previously identified mutations in MPD and 1° MCPH genes have been inherited in an autosomal recessive manner and so this was the default model for analysis. However, for multiplex datasets (more than one family member sequenced) autosomal dominant and X-linked recessive inheritance patterns were also examined (SQL scripts in Appendix IV).

3.3.4.1 Autosomal recessive (AR)

To identify all potential variants which had been inherited in an autosomal recessive manner, all homozygous and heterozygous variants in affected cases were selected. Heterozygous variants were then grouped per individual and gene to identify those genes where more than one variant was present per individual. All heterozygous variants occurring within these genes were combined with all homozygous variants identified. This removed a further 6.4% of variants (98.8% in total) reducing the number of candidate genes to 7.2% of the initial number (2189 candidate genes). In families where unaffected parents were sequenced (trios and quads), further filtering was performed to extract only those variants which were biallelic in the affected child. Therefore homozygous variants were selected only if the variant was identified in both parents in a heterozygous state and only those heterozygous variants were selected where at least two variants per gene were identified, one in each parent, consistent with autosomal recessive inheritance. Where an affected sibling was also sequenced (pairs and quads), only those variants in common between the two siblings were selected prior to the selection of recessive variants. This enabled variant number, and thus candidate gene number to be reduced much further with biallelic variants only occurring in 183 genes (0.6%).

3.3.4.2 Autosomal dominant (*de novo*)

As none of the parents were reported to be affected analysis was performed under the assumption that any pathogenic dominant variants had occurred *de novo* in the affected child. All heterozygous variants which did not occur in either the 1KG or EVS database (minor allele frequency=0.0000 or null) were selected. Variants occurring in dbSNP (i.e. annotated with an ‘rs’ number) were not removed as pathogenic dominant mutations are present in this database (Bhagwat, 2010). In all trios, only those variants not occurring in either parent were selected. Where an affected sibling was present (pairs and quads), a *de novo* event in both siblings was unlikely in the presence of unaffected parents and analysis was therefore not performed. Thus the possibility of gonadal mosaicism or incomplete penetrance in these families remains.

In total, *de novo* variants were identified in 612 genes of which only 91 contained a variant with a consequence score of 1. In singletons, heterozygous variants with a consequence score of 1 occurring in two or less families were initially prioritised for review due to the large number of variants in this group (>15,000 variants reduced to 789 variants in 734 genes).

3.3.4.3 X-linked recessive (XR)

All hemizygous variants in affected males (36 families) on the X chromosome were selected. In trios, only those variants also occurring in the mother in the heterozygous state but not present in the father were selected. In families designated ‘trio(m)’ and ‘quad’, analysis was not performed as affected offspring were female. Also only one sibling pair consisted of two affected males and so was analysed in this manner. This analysis identified 129 candidate genes in total.

3.3.5 Reduction in variant number and candidate genes through the filtering pipeline

Despite a large initial number of variants (mean 114,141 variants per family) the filtering pipeline efficiently reduced this number by 99.9% (Table 3.4). Removing common variants (stage 1 filter) had a large impact on variant number removing 68% but had little impact on the number of candidate genes with over 90% remaining (Figure 3.2). Selecting deleterious variants (stage 2 filter) appeared more effective in reducing the number of both variants and candidate genes (7.6% and 49.6% of total remaining respectively) although remaining numbers were still too large to review in each family individually (mean 576 variants per family, 15,382 candidate genes in total remaining) (Table 3.4).

Selecting only variants that have potentially been inherited in an autosomal recessive manner regardless of whether additional family members were sequenced dramatically reduced variant number to 1.2% and candidate gene number to 7.2% although this still equated to over 2,000 individual genes (Table 3.4). However filtering was greatly enhanced where parental DNA samples had also been sequenced (trios and quads) enabling biallelic variants to be identified reducing the candidate gene list to a manageable 183 genes. Examining the number of variants per family demonstrates the power of additionally sequencing parental samples. In trios and quads selecting only biallelic variants resulting in a five fold reduction in number of variants and candidate genes (mean of seven candidate genes per family in the trio group compared to 52 per family in the singleton group) (Table 3.5). Selecting variants in common between affected siblings (pairs) was also similarly effective in reducing variant number and therefore provides an effective but less costly strategy than sequencing both parents where multiple affecteds are present. Notably, in two trio sets, no variants remained following analysis for biallelic inheritance. This may reflect variation in exon capture and consequently coverage between samples although other possibilities include laboratory errors with DNA samples or non-paternity.

Selecting potential dominant variants (all rare, deleterious heterozygous variants) did not reduce variant number as much as selecting all potential recessive variants (over 11,000 candidate genes remaining) (Table 3.4, Figure 3.2). However, filtering was greatly enhanced when parental samples were used to identify only *de novo* variants. Variant number in trios was reduced to almost 10% of that in singletons (Table 3.5) with only 2% of candidate genes remaining in total (Table 3.4, Figure 3.2). Even in the family where only the mother and unaffected sibling were sequenced (trio(m)), 95% of possible dominant variants could be excluded in the *de novo* analysis although it was not possible to identify biallelic variants using this family structure. However, this does demonstrate the power of sequencing additional family members, even if DNA from both parents is unavailable. In contrast, little impact was made on variant number in the analysis of X-linked recessive variants where sequencing data was available from the mother compared to those without. This suggests that in disorders of clear X-linked inheritance sequencing parental DNA may be of little additional benefit.

Table 3.4. Reduction in variants and candidate gene number through filtering pipeline

	Stage of common filtering pipeline	***Total variant number***	***% of total variants***	***Mean variant N° per family*** *(range)*	***Total N° of candidate genes***	***% of total candidate genes***
1	All variants within a gene	601,908	**100**	**114,141** (97,868-149,104)	30,551	**100**
2	Post stage 1 (rare variants)	189,936	**32**	**2,454** (1,299-5,927)	26,620	**91.3**
3	Post stage 2 (deleterious variants)	45,944	**7.6**	**576** (321-1,313)	15,382	**49.7**
4	AR analysis	7,449	**1.2**	**85** (15-272)	2,189	**7.2**
5	Biallelic analysis (in multiplex families)	395	**0.07**	**25** (0-64)	183	**0.6**
6	AD analysis	21,142	**3.5**	**267** (113-629)	11,080	**36**
7	*De novo* analysis (trios)	737	**0.1**	**27** (7-144)	612	**2.0**
8	X-linked recessive analysis	154	**0.03**	**5** (1-9)	129	**0.4**

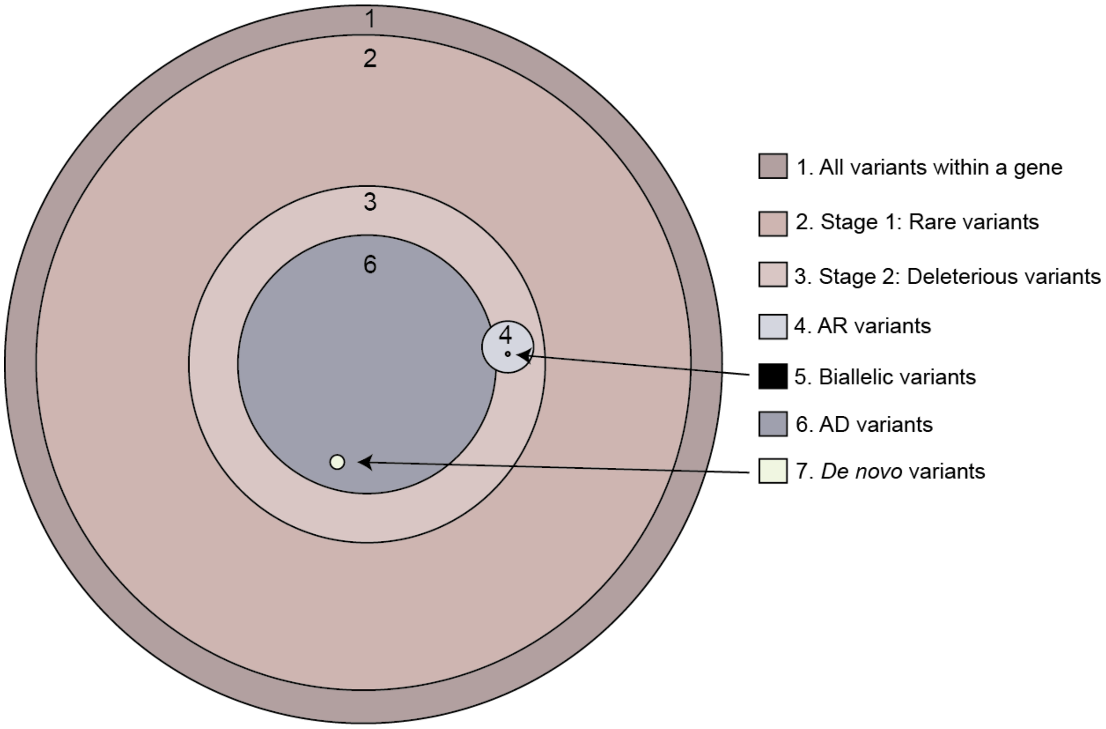

1
2
3
6
4
1. All variants within a gene
2. Stage 1: Rare variants
3. Stage 2: Deleterious variants
4. AR variants
5. Biallelic variants
6. AD variants
7. *De novo* variants

Figure 3.2. Venn diagram demonstrating reduction in candidate gene number through filtering pipeline

Removing all commonly occurring variants (stage 1 filter) had little impact on the number of candidate genes whereas further filtering for deleterious variants had a more pronounced impact on gene number. Autosomal recessive analysis reduced gene number further than selecting all potential dominant variants however using parental samples to select either biallelic and *de novo* variants dramatically reduced candidate genes to a manageable number.

Table 3.5. Variant and candidate gene number per family following analysis by different inheritance models

		Mean variant N° per family (range)				
Number of families after known MPD genes identified		***Singleton(51)***	***Trio(28)***	***Trio(m)(1)***	***Quad(1)***	***Sib pair(5)***
AR analysis	All potential AR variants	86 (15-225)	64 (26-200)	21	272	180 (117-240)
	Biallelic variants	N/A	13 (0-56)	N/A	16	38 (12-64)*
AD analysis	All Het variants	295 (131-629)	233 (113-436)	135		
	De novo variants	N/A	28 (8-144)	7		
X-linked recessive analysis	Hom variants in males on Chr	5 (1-9)	5 (1-7)			19
	Variants only inherited from mother	N/A	4(1-7)			6
		Mean candidate gene N° per family (range)				
Number of families after known MPD genes identified		***Singleton(51)***	***Trio(28)***	***Trio(m)(1)***	***Quad(1)***	***Sib pair(5)***
AR analysis		52 (13-113)	7 (0-29)	12	6	22 (6-38)
AD analysis		293 (133-609)	25 (7-139)	7		
X-linked recessive analysis		5 (1-9)	4 (1-7)			6

Once diagnosis identified in known MPD gene, families removed from further analysis. *as parents not sequenced, number indicates variants remaining after filtering for those shared between siblings. Light grey boxes indicate where analysis was not performed. X-linked recessive analysis was performed in 36 families with affected males including 10 trios and 1 sib pair.

3.3.6 Review of variants in other known disease genes affecting growth

OMIM morbid annotation was used to highlight variants occurring in genes already known to be associated with a developmental disorder. Many of these genes also impact on growth but may not have previously been associated with such an extreme failure in growth as seen in MPD patients. As well as diagnostic benefit to the patient, early identification allowed these cases to be excluded from further analysis of novel gene candidates and help prioritise genes for closer review.

3.3.7 Strategies to prioritise remaining candidate genes

Following the removal of families in whom a likely diagnosis had been identified, variants in candidate genes unlikely to be disease causing were filtered out. This included genes whose function was highly unlikely to be related to growth such as HLA genes, mucin genes, olfactory receptor genes and hypothetical genes (those which have not been experimentally validated and of unknown function). Also genes enriched for ‘rare’ variants, as previously identified by the NHLBI GO Exome Sequencing Project, were excluded (Tennessen *et al.*, 2012). This includes genes often poorly mapped, for example pseudogenes, duplicated genes and genes from large families sharing similar sequences, and those under weak selective constraint. Remaining genes with variants occurring in more than one family and genes with a highly deleterious variant (consequence score of 1) were then prioritised for further examination.

The sequencing reads corresponding to each variant were examined by visualising alignment files using the Integrative Genomics Viewer (IGV) (Robinson *et al.*, 2011) to assess the quality of reads, alignment and read depth at that site. A variant was deemed likely to be real if reads were accurately aligned, a sufficient read depth was present (>3 reads for homozygous variants and >10 reads for heterozygous variants) and at least 50% of reads had a mapping quality of >0. The protein consequence was then reviewed using Alamut software. Assuming *de novo* variants in MPD would be haploinsufficient, each candidate gene from the *de novo* analysis was also reviewed in the Database of Genomic Variants (DGV) (MacDonald *et al.*, 2014) and disregarded if a microdeletion incorporating that gene had been observed in the control population. After closer inspection, if the variant(s) within that gene still appeared potentially deleterious it was then validated by capillary sequencing. If DNA was available from any family members these were also re-sequenced to investigate whether the variants segregate correctly within the family.

Finally, a literature search was performed for each candidate gene. Genes were excluded if function had been well characterised and unlikely to have an impact on growth, for example *PRSS1* which encodes a pancreatic enzyme required for protein digestion. Also genes previously established to cause a disease totally unrelated to growth such as isolated non-syndromic deafness or cataracts were not investigated further. Existing animal models for that gene were then researched in the following databases: Mouse Genome Informatics (http://www.informatics.jax.org/, accessed 06.2013) (Blake *et al.*, 2014), Zebrafish Model Organism Database (ZFIN) (http://zfin.org/, accessed 06.2013) (Bradford *et al.*, 2011) and FlyBase (http://flybase.org/, accessed 06.2013) (St Pierre *et al.*, 2014). Genes where animal knockout studies had been performed and demonstrated no observable impact on growth were given lower priority. Gene expression patterns were also reviewed by examining microarray expression data using UCSC human gene sorter as genes previously identified in MPD are all widely if not ubiquitously expressed (https://genome.ucsc.edu/cgi-bin/hgNear, accessed 06.2013) (Kent *et al.*, 2005). Therefore genes with very limited and specific expression patterns, for example those only expressed in ovaries and testis, were deemed unlikely to cause a global growth phenotype.

3.4 Variation in variant number between samples

Following variant calling I noted that there was a large range in the number of variants called per sample with some samples having over 25% more variants than others (mean number of variants per sample = 111,381, range 97,868 to 132,821 prior to stage 1). The cause of such a large disparity between samples was therefore investigated by comparing samples prepared and sequenced at different locations and then by comparing those of different ethnic origin. Variant number between samples prepared and sequenced at either the Wellcome Trust Sanger Institute (WTSI), Oxford Gene technology (OGT) or Edinburgh-Denmark locations showed some variation in variant number prior to filtering with a significant difference occurring between the OGT and Edinburgh-Denmark samples and the OGT and WTSI samples ($p<0.01$, Figure 3.3). Those sequenced at WTSI had the highest number of variants (mean=120,508) whilst those at OGT had the lowest number (mean=107,134). A significant difference in variant number was still present between these two sample groups following both the stage 1 ($p<0.005$) and stage 2 filters ($p<0.01$) (Figure 3.3).

Samples were then grouped by European, Asian, African or Middle Eastern origin. Country of origin was unknown in 14 samples, 10 of which were in the WTSI group, and so were also excluded. Seven samples were from more than one ethnic background and so also not included. Prior to filtering, African samples had a significantly higher number of variants compared to both European and Asian groups ($p=0.01$) (Figure 3.4). Following filtering, variation in variant number became more evident between samples of different ethnicity with samples of African origin having the highest number of apparently rare deleterious variants followed by Middle Eastern and then Asian families. The number of variants in African, Middle-Eastern and Asian groups were all significantly increased compared to the European group ($p=0.0001$) but not compared to each other following both stage 1 and 2 filtering. This likely reflects the under representation of these ethnic groups in the control datasets used to filter out common variants as well as the increased variation known to be present in Africans (Tennessen *et al.*, 2012). Notably, the Edinburgh-Denmark and WTSI groups did not include any samples of African origin which may have contributed to the lower number of variants in these groups.

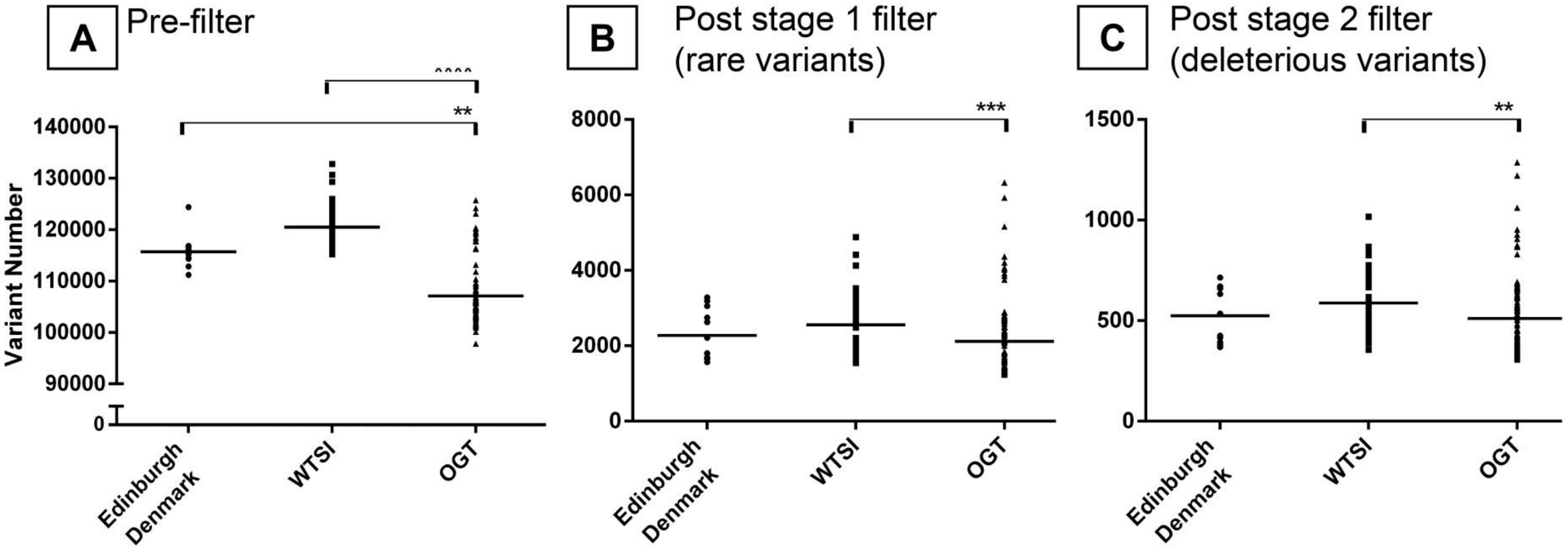


Figure 3.3. Comparison of variant number in samples prepared and sequenced in different locations

Range in variant numbers for each sample prepared and sequenced at the three different locations A) prior to filtering (mean variant number indicated by horizontal bar: Edinburgh/Denmark=115,736 +/-3196 s.d., WTSI=120,508 +/-3816 s.d., OGT=107,134 +/-5862 s.d.), B) post stage 1 filter (mean variant number: Edinburgh/Denmark=2,276 +/-666 s.d., WTSI=2,557 +/-806 s.d., OGT=2119 +/-964 s.d.) and (C) post stage 2 filter (mean variant number: Edinburgh/Denmark=524 +/-136 s.d., WTSI=588 +/-157 s.d., OGT=511 +/-193 s.d.). Each group was compared to all others using a Kruskal-Wallis test followed by a Dunn's multiple comparison test. A significant difference was consistently seen between the OGT samples compared to the WTSI samples. Comparisons in which a significant difference was present are indicated by overlying horizontal bars. ** = p<0.01, *** = p<0.005, ****p<0.0001. No significant difference was seen on other comparisons. *Abbreviations: WTSI=Wellcome Trust Sanger Institute, OGT=Oxford Gene Technology.*

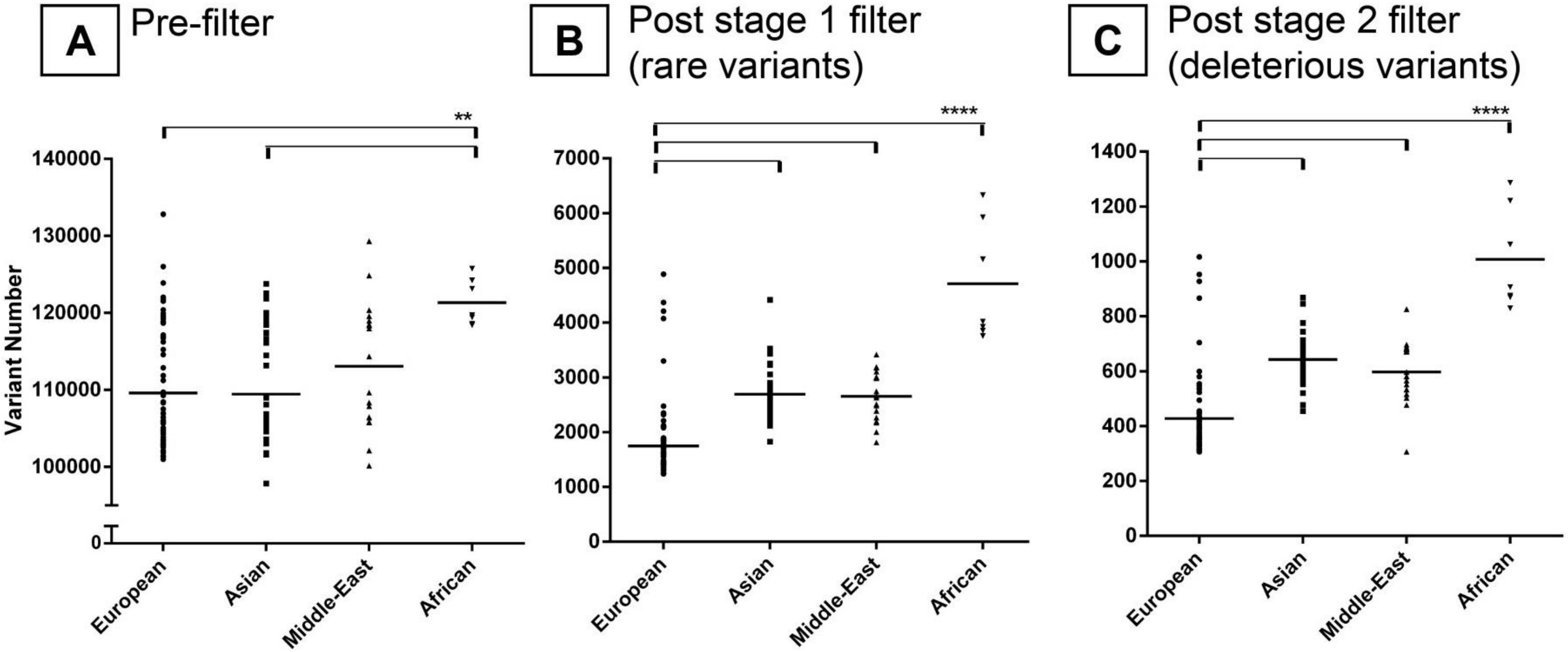


Figure 3.4. Comparison of variant number between samples of different ethnicity

A) Prior to filtering a significant increase in variant number was seen in African samples compared to European and Asian groups (mean variant number: European =109,582 +/-7484 s.d., Asian=109,435 +/-7210 s.d., Middle East=113,064 +/-8252 s.d., African=121,334 +/-2992 s.d.). Following B) the stage 1 filter (mean variant number: European=1,747 +/-678 s.d., Asian=2,693 +/-480 s.d., Middle East= 2,655 +/-457 s.d., African= 4,711 +/-1083 s.d.) and C) the stage 2 filter (mean variant number: European =428 +/-135 s.d., Asian=643 +/-88s.d., Middle East=596 +/-112s.d., African=1008+/-185s.d.) samples of African origin continued to show the highest number of variants. Samples of African, Middle-Eastern and Asian origin all showed a significantly higher number of variants compared to European samples. Each group was compared to all others using a Kruskal-Wallis test followed by a Dunn's multiple comparison test. No significant difference was seen on other group comparisons. **p<0.01, ****= p<0.0001.

3.5 Variants identified in known disease genes

3.5.1 Variants identified in known MPD and 1° MCPH genes

In total 169 variants in 80 families were identified. The quality and consequence of individual variants were assessed using IGV and Alamut (Section 3.3.6). This identified nine families with recessive deleterious variants within these genes. In five families, these variants were previously known, confirming the sensitivity of the pipeline to identify causative variants. These included mutations in *ASPM* (MIM 608716), *CDK5RAP2* (MIM 604804), *ATR* (MIM 210600), *PCNT* (MIM 210720) and *PLK4* (Martin *et al.*, 2014). In the other four families, novel variants were identified in *PCNT* (MOPDII), *ORC1* (Meier Gorlin syndrome) and *CENPJ* (Seckel syndrome) (Table 3.6). All occurred in the homozygous state except for one case with compound heterozygous mutations in *PCNT*. Variants were validated by capillary sequencing and found to segregate appropriately with the phenotype in the family.

The variants identified in *PCNT* were nonsense mutations and therefore transcrips are predicted to undergo nonsense mediated decay similar to those previously described in MOPDII patients (Rauch *et al.*, 2008). Both patients showed proportionate and severe growth failure similar to other MOPDII patients (Bober *et al.*, 2012) (Table 3.6). Patient 1 was also noted to have widespread confetti like hypopigmented patches with joint contractures and epilepsy whilst patient 2 was reported as having moderate developmental delay. These features are not characteristic of MOPDII and had precluded prior consideration of this diagnosis. However, on review of clinical phenotypes both patients were noted to also have typical facial features in keeping with this diagnosis. These cases therefore expand the clinical spectrum of MOPDII and demonstrate the power of WES in identifying diagnoses in atypical cases.

In another patient, a homozygous non-synonymous coding variant was identified in *ORC1* affecting a highly conserved residue within the BAH domain and predicted to be highly deleterious (Figure 3.5). Mutations in *ORC1* cause Meier-Gorlin syndrome characterised by microtia, hypoplastic or absent patella and short stature (Bicknell *et al.*, 2011b), although this triad of features is not universally present in every patient with MGS (de Munnik *et al.*, 2012). This patient was noted to have severe growth failure (Table 3.6) in keeping with mutations in *ORC1* although microtia was not present and no abnormalities were reported concerning the patella. The patient did, however, have some facial features consistent with the diagnosis including microstomia. A mutation altering the same amino acid (c.314G>A, p.Arg105Gln) has previously been identified in several MGS patients (de Munnik *et al.*, 2012) and occurs at a site conserved in vertebrates (Figure 3.5). Therefore the likelihood of this variant also being pathogenic is high and in conjunction with the clinical phenotype of this patient, was attributed as causal.

Table 3.6. Mutations identified in known MPD genes

Gene	*cDNA change*		*Protein change*	*Parents*		*OFC /s.d.*	*Height /s.d.*
				Mother	*Father*		
***PCNT*:** MOPDII	1	c.8761G>T & c.6773delAC	p.Glu2921* & p.Thr2258Ilefs*40	c.8761 G>T	c.6773 delAC	-14.4	-10.6
	2	c.9319delAG	p.Ser3107Phefs*39	Het	Het	-11.2	-10.6
***ORC1*:** MGS	3	c.313C>T	p.Arg105Trp	Het	Het	-11.1	-6.2
***CENPJ*:** Seckel syndrome	4	c.3302-15A>G		Het	Het	-11.1	-4

Abbreviations: MGS=Meier Gorlin Syndrome.

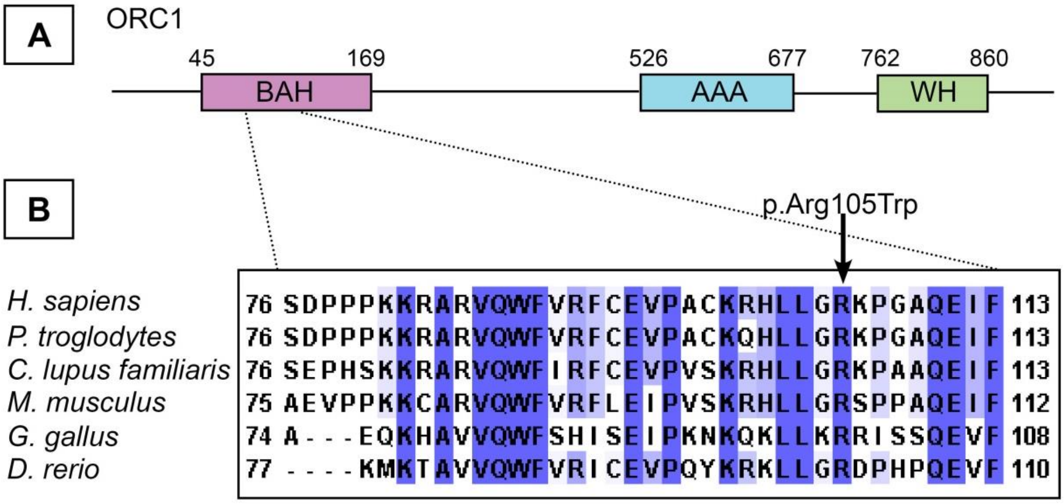

A
ORC1
45
169
BAH
526
677
AAA
762
860
WH
B
p.Arg105Trp
H. sapiens
P. troglodytes
C. lupus familiaris
M. musculus
G. gallus
D. rerio
76 SDPPPKKRARVQWFVRFCEVPACKRHLLGRKPGAQEIF 113
76 SDPPPKKRARVQWFVRFCEVPACKQHLLGRKPGAQEIF 113
76 SEPHSKKRARVQWFIRFCEVPVSKRHLLGRKPAAQEIF 113
75 AEVPPKKCARVQWFVRFLEIPVSKRHLLGRSPPAQEIF 112
74 A---EQKHAVVQWFSHISEIPKNKQKLLKRRISSQEVF 108
77 ----KMKTAVVQWFVRICEVPQYKRKLLGRDPHPQEVF 110

Figure 3.5. Conservation of amino acid altered by the c.313C>T variant identified in *ORC1*

A) Schematic of ORC1 protein demonstrating identifiable domains including the BAH (bromo-adjacent homology) domain, AAA+ ATPase domain (AAA) and a winged helix DNA-binding domain (WH). B) Alignment of sequences encoding the BAH domain in *ORC1* in different species using ClustalW2 (http://www.ebi.ac.uk/Tools/msa/clustalw2/) demonstrates the affected amino acid is conserved in vertebrates (sequence reference in Appendix II). A similar mutation, affecting the same amino acid, c.314G>A (p.Arg105Gln) has been previously reported in several MGS patients (de Munnik *et al.*, 2012).

Mutations in *CENPJ* have only previously been reported in one family with MPD (Al-Dosari *et al.*, 2010). A homozygous mutation was identified in this reported family affecting the splice acceptor site of exon 11 (c.3302-1G>C) resulting in a reduction in the wild type transcript in patients and the production of three alternate transcripts. These transcripts were found to lack exon 12, exons 11 and 12 and exons 11 to 13 respectively. In this study, a homozygous intronic mutation, 15 bases upstream of the same exon-intron boundary (c.3302-15A>C), was identified in one patient and was predicted to create an alternative splice acceptor site at the position of the variant (Figure 3.6A). This would be predicted to result in an out-of-frame transcript with the inclusion of 14 additional bases prior to exon 11. RT-PCR studies in patient lymphoblastoid cells did not identify this transcript possibly because it has undergone nonsense mediated decay. However, a reduction in full length transcript was apparent along with the production of three alternate out-of-frame transcripts.

Sequencing PCR products extracted from the four separate bands determined that they corresponded to the full length transcript and three smaller transcripts which were identical to those identified in the previously reported patient (Figure 3.6B). The patient in this study had severe microcephaly (-11 s.d.) with short stature (-4 s.d.), moderate developmental delay and a small right kidney with grade II vesicoureteric reflux. All growth parameters were reported to be below -7 s.d. in all five affected family members described in Al-Dosari *et al*, 2010 and displayed similar facial features to the patient identified here including high, prominent nasal bridge, hooked nose and receding chin.

In two families, deleterious heterozygous variants were identified in *ASPM* (c.1138C>T, p.Gln380*) and *PCNT* (c.2494_2497dupGACG, p.Ala833Glyfs*45) respectively (consequence score=1). Both variants were called with high confidence following WES however capillary sequencing did not validate either variant and re-sequencing of the full gene did not reveal any other likely pathogenic variants. Therefore neither case was excluded from further analysis.

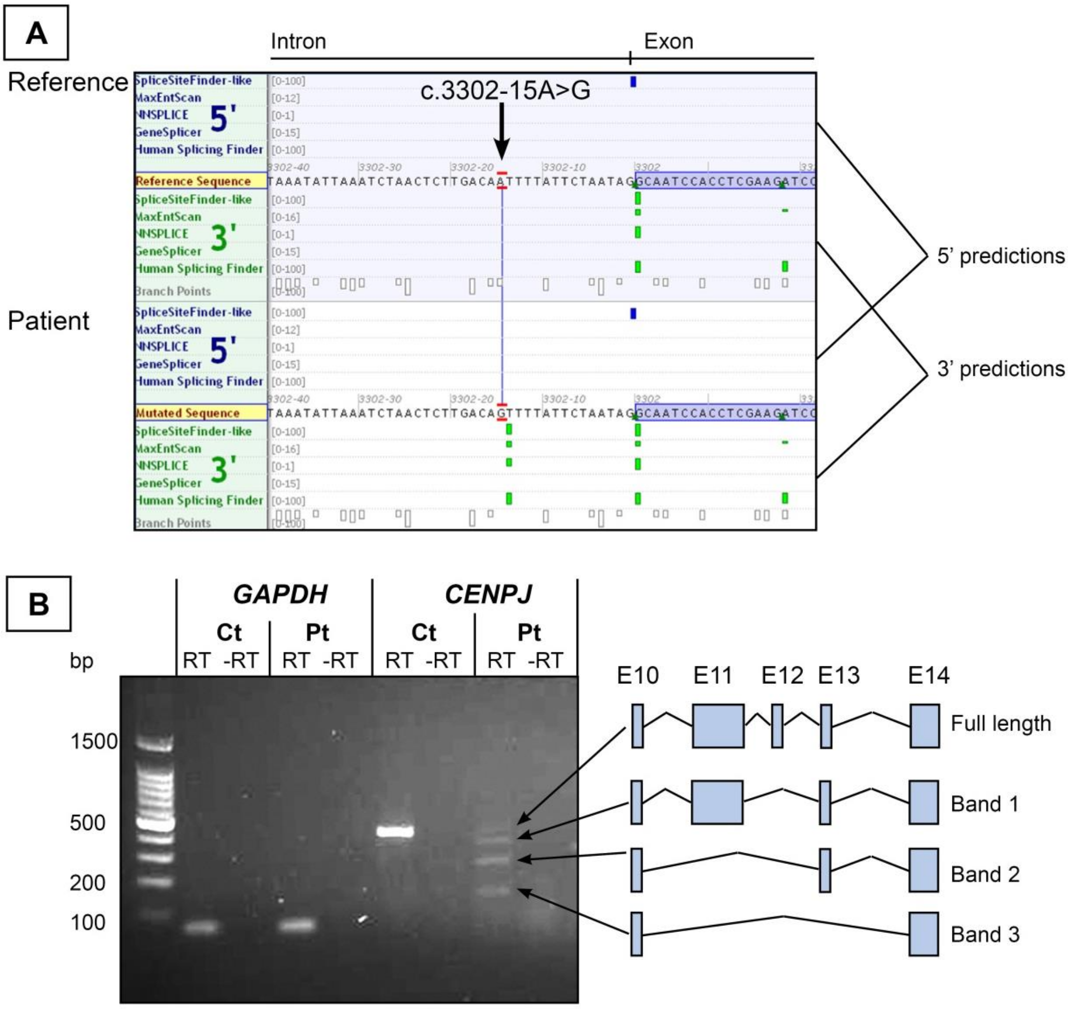


Figure 3.6. A homozygous *CENPJ* mutation, c.3302-15A>G, disrupting splicing

A) Alamut graphical display showing that four of the five splice site prediction programmes employed predict the presence of a splice acceptor site (green boxes) at the intron-exon boundary of the reference sequence (blue shading). The same programmes also predict the presence of an alternative splice acceptor site at the position of the intronic single nucleotide change, c.3302-15A>G identified in the patient (additional green boxes in white shading). B) RT-PCR studies showed a reduction in full length transcript in patient lymphoblastoid cells compared to control cells and the presence of three bands representing alternative transcripts. Sequencing of DNA extracted from these bands confirmed that exons 12 to 13 have been removed similar to that caused by the mutation c.3302-1G>C identified in Al-Dosari *et al*., 2010. No difference in the housekeeping gene, GAPDH, transcript levels was observed between patient and control cells. PCR following RT reaction without reverse transcriptase (-RT) demonstrates no sample contamination occurred.

3.5.2 Variants identified in other known disease genes affecting growth

Review of known disease genes using the OMIM morbid annotation (Section 2.5.2.2) identified deleterious variants in 20 families, 17 of which had not been previously identified (Table 3.7). Variants in *SMARCAL1* (Schimke immuno-osseous dysplasia, MIM 242900), *CREBBP* (Rubinstein Taybi syndrome, MIM 180849) and *NSUN2* (Dubowitz syndrome, MIM 223370) had been previously identified therefore provided validation of the designed pipeline. In combination with known MPD and 1° MCPH genes, 100% of known variants were detected. Ligase IV syndrome (MIM 606593) is a disorder characterised by microcephaly, immunodeficiency and cancer predisposition that has not been previously associated with growth failure to the degree seen in MPD and so it was surprising to identify mutations in *LIG4* in three families (Table 3.7). The contribution of *LIG4* mutations to MPD was therefore further investigated (Chapter 4).

A nonsense mutation was also identified in the gene *RBM10* (RNA binding motif protein 10) on chromosome X. *RBM10* mutations cause TARP syndrome (Talipes equinovarus, Atrial septal defect, Robin sequence and Persistent left superior vena cava, MIM 311900) (Johnston *et al.*, 2010). Few patients with this disorder are reported in the literature with very limited information on growth. Affected males are typically unwell from birth and die in early life often from unexplained causes (Kurpinski *et al.*, 2003). The patient identified here had features consistent with this diagnosis including early lethality and severe micrognathia, an aspect of Robin sequence (Table 3.7). The severe growth failure documented in this patient suggests this may be a previously unrecognised aspect of this syndrome.

Mutations were also identified in the disease genes *SRCAP* (Floating Harbor syndrome, MIM 611421), *BLM* (Bloom syndrome, MIM 210900), *ERCC6* (Cockayne syndrome, MIM 216400), *VPS13B* (Cohen syndrome, MIM 216550) and *TUBGCP6* (microcephaly with chorioretinopathy, MIM 251270) (Table 3.7). All were predicted to either truncate the protein or disrupt an essential splice site (consequence score =1) thus highly likely to be deleterious to protein function.

Growth failure, comparable to that observed in these patients, has been previously reported in other patients with similar mutations (Nance *et al.*, 1992, Keller *et al.*, 1999, Hennies *et al.*, 2004, Puffenberger *et al.*, 2012, Nikkel *et al.*, 2013). Additionally, the patients identified here were also reported to have other clinical features consistent with the associated syndrome and thus the mutations were likely to be responsible for the phenotype observed in these patients. Likely causative mutations were also identified in *ESCO2* (Roberts syndrome, MIM268300) and *MRE11A* (Ataxia-telangiectasia-like/Nijmegen-like breakage syndrome, MIM604391) in two further families (Table 3.7). In both cases intronic mutations were identified which were predicted to strongly impact on splicing. Similarly, the phenotype reported in the patients correlated well to those previously described (Schule *et al.*, 2005, Matsumoto *et al.*, 2011) however further RNA studies would be necessary to confirm the effect on splicing.

A 22 base pair deletion (c.817-2_837delAGGCCAACGAGGCACGCCCATA) disrupting the splice acceptor site of the last exon was also identified in *PNKP* (polynucleotide kinase 3'-phosphatase) in association with a nonsense variant on the alternate allele. Mutations in *PNKP* have previously been described in association with severe microcephaly, developmental delay and seizures which were both present in the two affected siblings (Shen *et al.*, 2010). One sibling also had severely reduced stature although height was within normal range in the other sibling suggesting short stature may be a variable feature of mutations in this gene. Additional RNA studies would also be valuable to confirm the effect on splicing.

Finally, a *de novo* heterozygous nonsense mutation was identified in a female patient in *MED12* which resides on the X chromosome. Non-synonymous coding mutations in males in *MED12* have been described in association with three different syndromes; Lujan Fryns syndrome (MIM309520), Opitz-Kaveggia syndrome (MIM305450) and Ohdo syndrome, the Maat–Kievit–Brunner type (OSMKB, MIM249620) (Risheg *et al.*, 2007, Schwartz *et al.*, 2007, Vulto-van Silfhout *et al.*, 2013). This female patient presented with features of growth failure with mild developmental delay and moderate deafness which are not features typical of *MED12* related disorders (Graham *et al.*, 2013). However, she did have facial features similar to those described in males with Lujan Fryns syndrome including maxillary hypoplasia, receding chin, thin upper lip, prominent forehead and low-set retroverted ears (Van Buggenhout *et al.*, 2006). Affected females with heterozygous *MED12* mutations have only been previously described in one family harbouring a frameshift mutation which impacts on splicing resulting in an alternate transcript with a 75bp in-frame deletion which does not undergo nonsense mediated decay (Lesca *et al.*, 2013). Both female carriers and affected males were reported to have significant cognitive impairment but without any growth abnormalities. The facial features and intellectual disability present in this patient could therefore be attributed to the *MED12* mutation however this does not explain the failure in growth. It is possible this patient also harbours a second disease causing mutation in another gene which is responsible for the MPD phenotype. However, it is interesting to note that Lujan Fryns syndrome is also associated with tall stature and macrocephaly (Van Buggenhout *et al.*, 2006) indicating that aberrant MED12 function can impact on growth. It is therefore conceivable that gain of function mutations in *MED12* may result in Lujan Fryns syndrome whereas severe loss of function mutations results in growth restriction. Notably, heterozygous mutations resulting in premature truncation of *MED12* have not been reported in control populations (EVS).

Table 3.7. Probable pathogenic variants identified in atypical disease genes

Gene & Disorder	***Variant details***			***Mode of inheritance***	***OFC /s.d.***	***Height /s.d.***	***Clinical features***
		cDNA	***Protein***				
LIG4: Ligase IV Syndrome	5	2386dupATTG & 2440C>T	Ala797fs*2 Arg814*	AR	-10.8	-5.9	*Further details in Chapter 4*
	6	2094C>G & 2440C>T	Tyr698* Arg814*		-12.3	-5.0	
	7	1277delAAGAG 2440C>T	Glu426Glyfs*17 Arg814*		-9.4	-6.0	
SRCAP: Floating Harbor Syndrome	8	7330C>T	Arg2444*	*De novo*	-4.0	-4.0	Duplex kidney
	9	7330C>T	Arg2444*		-5.1	-5.1	Stenosis of descending colon
	10	7236delTC	Pro2413Cysfs*29		-2.9	-3.1	
MRE11A: Nijmegen-like syndrome	11	1447C>T & 154-36A>G	Arg483*	AR	-10.5	-8.2	Severe micrognathia
BLM: Bloom Syndrome	12	1985delAA	Lys662Ilefs*5	AR	-8.3	-4.8	Crowded teeth
	13	479_480delTT	Phe160*		-5.3	-3.9	Hypopigmented & café au lait lesions
ERCC6: Cockayne Syndrome	14	4063-1G>C		AR	-5.4	-4.5	
	15	2170-2A>G			-6	-7	
MED12: Lujan-Fryns syndrome	16	6448C>T	Gln2150*	*De novo*	-3.4	-4.1	Dysmorphic features, bilateral hearing loss
RBM10: TARP syndrome	17	820C>T	Gln274*	XR	-6.0	-5.3	Severe micrognathia, unexplained death at 9months
ESCO2: Robert Syndrome	18	862-12A>G		AR	-8.1	-7.5	Bilateral mesomelia, absent thumbs, club hands and feet.
VPS13B: Cohen Syndrome	19	C10888T & 11936delC	Gln3630* & Phe3954Leufs*33	AR	-4.9	-2.3	Severe developmental delay, neutropenia
PNKP: Microcephaly & seizures	20	1517G>A & 817-2_837del	Trp506*	AR	-7.1 -6.0	-6.9 -1.3	Epilepsy, severe developmental delay
TUBGCP6: Microcephaly & Chorioretinopathy	21	4334insT	His1445Glnfs*24	AR	-11.1	-3.3	Retinal dystrophy, severe developmental delay

Deleterious variants were identified in genes causing recognisable disorders associated with growth restriction following WES within our cohort.

3.6 Novel candidate disease genes identified in MPD

61 families remained after 29 families in whom pathogenic variants were identified in known disease genes and five families in whom copy number variation abnormalities were identified through ExomeCNV analysis (Chapter 6) were removed from further analysis. However, for these 61 families, almost 2,000 candidate genes still remained for closer review (Table 3.8). Therefore further filtering steps were implemented to attempt to reduce the candidate gene list to a tractable size. Removing poorly mapped genes had little impact on this number, as did excluding those genes which had a high likelihood of being functionally unrelated to growth. However, prioritising genes with variants in more than one family and those variants likely to have the greatest functional impact on the protein (consequence score=1) was successful in generating a manageable number of candidate genes for further review.

Table 3.8. Strategies used to prioritise candidate genes for further investigation

	Autosomal Recessive (including biallelic)	*Autosomal Dominant (including de novo)*
Gene number following exclusion of families with diagnosis	**1594**	**458**
Gene number after excluding poorly mapped genes	**1567** (98%)	**447** (98%)
Gene number after excluding functionally unrelated genes	**1429** (90%)	**414** (90%)
Number of genes with variants in more than one family	**222** (14%)	**22** (5%)
Number of genes with at least one variant with a consequence score of 1.	**141** (9%)	**52** (11%)

Percentage of starting gene number at each stage of filtering shown in brackets.

Following manual inspection of individual variants and corresponding genes as described in Section 3.3.7, six genes were identified as strong candidates for further investigation (Table 3.9), all of which contained autosomal recessive variants. Most genes were excluded as likely false positives following review of variants with IGV due to low read depth or poorly mapped reads. Review of splice predictions (variants annotated with a consequence score of 2.5) also showed that many of these variants were unlikely to impact on splicing as predicted changes were either very small or reduced the strength of a predicted splice site deep within an intron. Many candidate genes following autosomal dominant and *de novo* analyses were also excluded following review of overlying structural variation in control datasets (DGV) which indicated that these genes were unlikely to be disease causing as a result of haploinsufficiency.

Table 3.9. Novel candidate disease genes identified in MPD following filtering of WES

Gene: *full name*	***Function***	***Variant details***		
		Pt	***cDNA***	***Protein***
XRCC4: X-ray repair complementing defective repair in Chinese hamster cells 4	Non-homologous end joining in DNA double strand break repair.	22	c.127T>C	p.Trp43Arg
NCAPD3: Non-SMC condensin II complex, subunit D3	Chromosome condensation and segregation in mitosis	23	c.[382+14A>G] +[1783delG]	p.Ser74Alafs*3p. Val595Serfs*34
NCAPD2: Non-SMC condensin I complex, subunit D2	Chromosome condensation and segregation in mitosis	24	c.4120+2T>C	[splice]
FAT1: FAT atypical cadherin 1	Modulates planar cell polarity, Hippo pathway	25	c.5611G>A	p.Asp1871Asn
		26	c.[11333C>T]+ [c.3446T>C]	p.Ala3778Val p.Met1149Thr
		27	c.[7130C>T]+ [483C>A]	p.Thr2377Met p.Asn161Lys
USP2: Ubiquitin specific peptidase 2	Deubiquitinates MDM2 & cyclin D1	28	c.216_221delTGCCCinsGTGAGCT	p.Gly3ValfsX33
DONSON: Downstream neighbour of *SON*	Unknown	29	c.[786-33A>G] +[1282C>T]	p.Gln428*
		30	c.[1047-9A>G] +[876C>G]	p.Phe292Leu

Of the six genes identified, there was strong functional evidence for five having a connection to growth. XRCC4 binds to LIG4 as part of the non-homologous end joining machinery (NHEJ) which repairs DNA double-strand breaks (Drouet *et al.*, 2005). Mutations in *LIG4*, causing Ligase IV syndrome, were also identified in MPD patients in this cohort indicating mutations in *XRCC4* may also have a similar impact on growth. Further investigation of mutations in *LIG4* and *XRCC4* in MPD patients is performed in Chapter 4.

NCAPD2 and *NCAPD3* both encode subunits of the condensin complexes which play an important role in the compaction of DNA into chromosomes and chromosome segregation during mitosis (Hagstrom *et al.*, 2002, Ono *et al.*, 2004). As perturbed mitosis has previously been associated with MPD and impaired cell cycle impacts on cell proliferation (Section 1.3.1.3), these genes represented strong candidates in growth failure and further investigation of condensin genes in MPD patients is performed in Chapter 5.

FAT1 and *USP2* were also both selected for further investigation based on their function. One patient was identified as homozygous for a complex indel variant in the first exon of *USP2*, a nonsense mutation which was predicted to result in a premature termination codon in the transcript (Table 3.8). *USP2* encodes an ubiquitin-specific protease which has been shown to stabilize cyclin D1, a cell cycle regulator which drives G1-S phase transition, in human cancer cells (Shan *et al.*, 2009). USP2 has also been shown to stabilize the ubiquitin ligase MDM2 which targets the tumour suppressor p53 for degradation promoting cell proliferation (Kim *et al.*, 2012). *FAT1* encodes a cadherin protein which may activate the hippo pathway (Bennett *et al.*, 2006) important in growth regulation (Section 1.2.2). Non-synonymous variants with a consequence score of 2 were identified in three families in this gene. Both genes were therefore included in the custom designed Ion AmpliSeq™ panel and sequenced in the remainder of the MPD cohort. No further patients were identified with mutations in *USP2* and hence this gene was not prioritised for further investigation in this thesis. In contrast, 31 patients were identified with at least one rare deleterious non-synonymous variant (consequence score of 2) in *FAT1* out of 199 patients sequenced. High numbers of similar variants are also present in the EVS control dataset at low frequencies suggesting this gene may be enriched for rare variation in the general population and thus less likely to be disease causing.

To investigate this further, statistical comparison of the frequency of rare alleles between the two cohorts was performed. Initially EVS variants were annotated with consequence scores using SNPEff and dbNSFP as described in section 3.3.2. A cumulative frequency was obtained for all minor allele variants with a consequence score of 2 in each cohort (MPD cohort maf = 0.1, EVS cohort maf = 0.02). Chi square test with Yates comparison indicated that *FAT1* is significantly enriched for rare deleterious non-synonymous coding variants in the MPD cohort compared to EVS (p<0.0001). However, these populations differ in ethnicity with EVS containing European-American and African-American populations whereas the the MPD cohort additionally contains many families from the Middle-Easte and sub-Indian continent. SNV calling and validation pipelines are also likely to differ in stringency between the two cohorts and performing the analysis on a cumulative frequency does not take into account that some variants may be part of the same allele. Although not performed in this thesis, further statistical analysis using the site frequency spectrum of variation in a matched control population (MacArthur *et al*., 2014) as well as validation of all deleterious missense variants in the MPD cohort in *FAT1* could be more informative.

Mutations in *DONSON* were identified in two families. Affected cases harboured a strongly deleterious mutation on one allele in combination with an intronic variant which was predicted to have a very mild impact on splicing. Both cases had severe microcephaly but more mildly reduced stature. Sequencing of patients with a similar short stature-microcephaly phenotype was performed by Louise Bicknell (HGU MRC, Edinburgh) which has identified several more patients with similar mutations. Furthermore, many of the additional patients identified also harbour the c.786-33A>G variant which has been found to reside within a rare haplotype shared by all patients with this variant (Louise Bicknell). *DONSON* may therefore represent a novel disease causing gene in patients with a phenotype that overlaps with MPD. Very little is known regarding the function of DONSON however the drosophila homologue, humpty dumpty, has been shown to play a role in DNA replication (Bandura *et al.*, 2005). Further investigation of this gene is ongoing by Louise Bicknell and Paula Carroll (HGU MRC, Edinburgh).

3.7 Conclusions

It is usual for more than 50,000 sequence variants to be identified per exome depending on the capture method and sequencing platform used (Gilissen *et al.*, 2012). Therefore filtering strategies are required to help identify potentially pathogenic variants from such large numbers. However, this needs to be tailored to the disease under investigation, in particular the prevalence of the disease and the anticipated mode of inheritance. Filtering variants also requires a fine balance between reducing the variant list to a manageable scale whilst still retaining those variants which are potentially damaging. Previous studies which have been most successful in implementing WES to identify pathogenic variants investigated a clinically homogeneous group of patients who all harboured mutations in the same disease gene (Ng *et al.*, 2010a, Ng *et al.*, 2010b). In many, DNA from affected siblings and/or parents was also sequenced and the mode of inheritance correctly anticipated. In this setting the number of candidate variants could be reduced to single figures allowing the disease gene to be readily identified. Earlier studies also have the advantage of uncovering those disease genes which represent a common cause of a particular disorder (Ng *et al.*, 2010a, Hood *et al.*, 2012, Schmidts *et al.*, 2013) whereas rare disease genes, or conditions with large locus heterogeneity such as MPD, pose more difficulty due to the low number of cases that will be present in the cohort under investigation.

In this large, diverse cohort where there is established locus heterogeneity many cases are likely to be unique in their underlying disease gene. Similarly family samples are not always available. Following standard filtering methods the final variant list for an isolated case following WES may still be as large as 500 (Gilissen *et al.*, 2012) and therefore prioritising variants further requires other strategies. This would also be similar to the situation in clinical diagnostics where often isolated cases are encountered rather than multi-affected families often prioritised in research studies. Large scale sequencing projects of multiple family members are also costly and so unlikely to be feasible in government funded healthcare.

This Chapter describes the development of a comprehensive filtering strategy for the analysis of WES in a Mendelian disorder with substantial locus heterogeneity.

Despite a high number of initial variants per sample (>100,000), filtering successfully reduced the number of variants to 0.01% whilst retaining established pathogenic mutations in the dataset as well as achieving a molecular diagnosis in a further 22% of those families sequenced. Strategies employed by previous studies included removing variants outside coding regions and synonymous variants which are deemed less likely to have a functional impact on the protein (Zhang *et al.*, 2012, Schmidts *et al.*, 2013). However, non-coding, intronic and synonymous variants can be disease causing by impacting on either splicing, gene expression or protein function (Richards *et al.*, 2012, Hunt *et al.*, 2014). Here we additionally analysed splice site effects of intronic and synonymous variants identifying mutations in four families which would have otherwise been missed. Furthermore, analysis using multiple models of inheritance and specifically examining known disease genes identified pathogenic mutations in 30% of families. However, despite extensive analysis disease variants have not yet been identified in 59% of cases. This may reflect the limitations of array-based capture methods and NGS technology although other factors including experimental design and stringency of variant filtering are also important considerations.

3.7.1 Experimental design

Although the cost of NGS technology is falling, undertaking WES in large numbers of samples is costly and consideration of experimental design, especially deciding who to sequence, is a key issue in using available research funding to maximise gene discovery. In this MPD cohort sequencing the parents and patient as a trio in non-consanguineous families enabled a large reduction in variant number by filtering according to a particular mode of inheritance (Table 3.5). This resulted in the identification of a probable diagnosis in 25% of these families. Conversely, in consanguineous families, only the affected offspring were sequenced as autosomal recessive inheritance and homozygosity-by-descent were anticipated in these cases. Similarly a diagnosis was obtained in 25% of cases in both trio and singleton groups suggesting this strategy was a favourable use of resources in a disorder where recessive inheritance was most likely. Furthermore, identifying regions of homozygosity in consanguineous cases could help narrow down candidate disease genes (Becker *et al.*, 2011) and tools are now available to perform this using NGS generated data (Seelow *et al.*, 2009). WES of the proband could also be combined with genetic linkage analysis to further limit the region of interest although this requires DNA to be available from other family members and is most powerful in multiplex families (Yamaguchi *et al.*, 2011).

This study also demonstrates the importance of considering different modes of inheritance in analysis, particularly if no candidate genes are found under the favoured model. Here, examining variants using a range of inheritance models uncovered unexpected diagnoses, for example TARP syndrome (X-linked recessive) and Floating Harbor syndrome (*de novo*), expanding the phenotype of these disorders. However, this approach is limited to trios and therefore unless consanguinity is likely, sequencing trios over singletons will improve the chances of identifying a diagnosis.

It is also important to consider that causative variants may lie outwith the exome either in a non-coding region which can affect splicing or gene regulation or within a non-protein coding gene which is not included in the exome capture design. For example, mutations in *RNU4ATAC* cause MOPDI (Edery *et al.*, 2011, He *et al.*, 2011), a non-coding gene which is not covered in most exome capture kits. Whole genome sequencing (WGS) may therefore be a preferential alternative to maximise the chances of discovering disease causing mutations. The benefits of WGS, including improved uniformity in coverage and absence of capture based bias, need to be weighed against problems of increased cost ($1000 per genome), sequencing fewer samples and increased data handling capacity. Furthermore, vastly increased variant numbers in non-coding sequence will be identified and predicting the functional consequence of variants outwith the exome will be very challenging although resources such as the ENCODE (ENCyclopedia Of DNA Elements) project (Consortium, 2004) and the development of prediction algorithms for genome wide variants (Kircher *et al.*, 2014) can now assist with interpretation. An alternative approach by which to identify abnormalities in gene expression and splicing as well as coding variation is to sequence the transcriptome using NGS technology (Cirulli *et al.*, 2010). Not only does this method negate the need for costly exome-enrichment but also entails significantly less sequencing compared to the whole genome. However, highly expressed genes will be over-represented and also nonsense mediated decay may deplete transcripts containing nonsense and splice mutations. As yet, causative variants have yet to be identified in Mendelian disorders through a solely RNA-based approach.

3.7.2 Library preparation

In this study, there was a high degree of variation in the number of sequencing reads produced for each sample (Table 3.2) which in part is explained by the depth of sequencing which affects target coverage and consequently the numbers of variants identified (Hoischen *et al.*, 2010, Meynert *et al.*, 2013). Although discrepancies in variant number occur with different commercially available capture kits (Chilamakuri *et al.*, 2014), other factors also account for this variation including library quantification and DNA quality.

3.7.2.1 Sample to sample variability

The quality of input DNA (encompassing parameters such as degradation and purity) affects the efficiency of library capture with higher levels of dropout occurring in samples of poorer quality (Hasmats *et al.*, 2014). In addition, pooling of sample libraries prior to NGS sequencing of several barcoded samples in one lane will determine the consistency of sequencing depth between samples with unequal pooling altering the representation of each sample in the sequencing run (Quail *et al.*, 2008). Furthermore, quantification of libraries prepared in Edinburgh was performed using the Agilent 2100 Bioanalyzer, the accuracy of which is influenced by DNA concentration (Panaro *et al.*, 2000). For example, the quantification of libraries with lower concentrations may be underestimated leading to an overrepresentation of such samples following pooling.

3.7.2.2 Batch variability

Significant variation in sequencing metrics was also seen between batches prepared and sequenced at different locations despite using the same capture method and sequencing platform (Table 3.2). This study was performed over a 16 month period with WTSI samples being prepared and sequenced over a year before the Edinburgh-

Denmark and OGT samples. Improvements in experimental protocols and software will have occurred during this period and likely contributed to the variability between different batches. Different centres also use different DNA quantification methods which often do not produce consistent results (Buehler *et al.*, 2010). Under- or over-representation of quantification can impact on cluster density and consequently the total number of sequencing reads generated per run (Quail *et al.*, 2008). Even with highly accurate quantification and equal pooling it is still then difficult to achieve complete and reproducible uniformity when manually loading a small volume of sample onto the flowcell. The degree to which the library preparation process is automated may also vary between the different centres and may impact on the efficacy of library capture.

3.7.3 False positive variants

One common difficulty encountered with NGS technologies is how to identify real variation from false positive variants. The three main sources of false positive variants include errors in sequencing, incorrect alignment of reads to the reference genome (mapping errors) and incorrect variant calling in regions of low coverage. Since the advent of NGS technology, great efforts have been taken to improve alignment and variant calling software. Despite this, false positive variants still present a significant issue with poor concordance between different alignment programmes and variant callers (Liu *et al.*, 2013b, Shang *et al.*, 2014). Strategies previously employed include filtering variants based on quality scores (Zhang *et al.*, 2012) however reads containing a potentially damaging insertion or deletion are more likely to be poorly mapped and consequently less reliably called (Chou *et al.*, 2010). Such variants are therefore more likely to be excluded when using stringent quality controls and therefore no quality filters were used in this pipeline.

3.7.3.1 Sequencing errors

Sequencing errors in NGS technology can result from inaccuracies in base calling which particularly occur in homopolymer runs and repetitive regions (Boland *et al.*, 2013). The degree and type of error varies depending on the method of exome capture, template preparation and sequencing technology used, for example, semi-conductor sequencing has greater inaccuracy in homopolymer runs compared to other methods (Harris *et al.*, 2008, Liu *et al.*, 2012). False positive variants resulting from such errors are therefore likely to occur at similar sites within different samples, particularly if the same methodologies were used. As a large degree of genetic heterogeneity was anticipated in this cohort, it was assumed repeatedly occurring variants were more likely to represent sequencing errors than common pathogenic variants and were excluded. However, in previous studies where one or a very limited number of genes were likely to cause the phenotype, removing variants in common between patients would have likely excluded pathogenic mutations (Smith *et al.*, 2014). An alternative and perhaps preferable approach is to use large publically available control datasets (matched in methodology) to identify commonly occurring errors although this will not remove false positive variants which are specific to a batch of samples prepared and sequenced together.

3.7.3.2 Mapping errors

Approximately 4% of the genome will map poorly to the reference (DePristo *et al.*, 2011) which provides an additional source of false positive calls. Genes more liable to incorrect alignment include pseudogenes, those from large gene families with very closely related sequences or duplicated genes (Tennessen *et al.*, 2012). Equally genes with multiple repeat regions will also be difficult to align accurately and may appear enriched for novel variants (Treangen *et al.*, 2012). Here, BWA plus Stampy was used for read alignment in combination with GATK for variant calling. This approach gives higher sensitivity in variant detection compared to other programmes but at the expense of specificity (Altmann *et al.*, 2012). Using this combination of software, Altmann *et al* found the total number of variants following whole exome capture and Illumina sequencing to be in the region of 250,000 which reduced eight fold when limiting variants to target regions (exons or 50 bases into flanking introns). The high numbers of starting variants reported here compared to other studies (Gilissen *et al.*, 2012) may reflect the inclusion of more off-target variants prior to filtering as only those outwith a gene locus (annotated by SnpEff as ‘intergenic’) were removed. This, of course, is weighed against the possibility that potentially causative variants in other regions (deep intronic, 5´UTR and 3´UTR) inadvertently uncovered by off-target capture will be missed. However, as previously mentioned in Section 3.7.1, interpreting the functional consequences of variants which lie outwith coding exons requires additional annotation.

3.7.3.3 Regions of low coverage

Regions covered by low read depth will contain more errors in variant calling and potentially pathogenic variants are more likely to be missed (Hoischen *et al.*, 2010, Meynert *et al.*, 2013). Low read depth can result from low capture efficiency which is a particular problem in regions of large repeats or extremes of GC content (Sathirapongsasuti *et al.*, 2011). Such regions are also more liable to errors during sequencing (Hodges *et al.*, 2007, van Dijk *et al.*, 2014). Attempting to validate all candidate variants in such areas is time consuming and inefficient as many are likely to be false positives. Filtering variants based on read depth can be a useful first line approach to help reduce noise however homozygous variants can still be confidently detected at low read depths (<5X) (Meynert *et al.*, 2013) and thus true-positive variants will also be excluded. If such regions include genes of high interest to the study it may be more worthwhile performing a focused capture using alternative methods with improved performance in such regions combined with increased sequencing to achieve a greater read depth. For example, (Oyola *et al.*, 2012) showed that improved coverage could be achieved in AT rich regions by adjusting PCR conditions. As well as capture by hybridisation methods using customised bait panels (e.g. SureSelect™ and HaloPlex™, Agilent), alternative methods which can be tailored to target specific genes of interest for high throughput sequencing include molecular inversion probes (Boyle *et al.*, 2014) and multiplex PCR primer panels (AmpliSeq™) (Millat *et al.*, 2014).

3.7.4 Identifying rare and deleterious variants

Pathogenic variants may also be missed when excluding variants previously identified in control populations such as those recorded in dbSNP (Sherry *et al.*, 2001), the 1000 Genomes Project (1KG) (Clarke *et al.*, 2012) or the NHLBI GO Exome Sequencing Project (ESP, Seattle, WA) (http://evs.gs.washington.edu/EVS/). The large numbers of individuals now sequenced in these projects has resulted in the ascertainment of rare heterozygous variants in such databases which may be disease causing in the homozygous state (Knowles *et al.*, 2013). Utilising control population datasets is common practice when filtering for rare variants however an appropriate minor allele frequency (maf) threshold needs to be set which takes into account the incidence of the disorder in the general population (Bamshad *et al.*, 2011).

Following the advent of NGS, large sequencing projects has revealed the extent of genetic variation between different ethnic groups (Genomes Project *et al.*, 2012). Such variation was evident in this study with an enrichment of apparently rare variants occurring in non-European samples. Therefore matching cases with controls of the same ethnicity would be ideal to reduce the candidate gene list further in such families. However, many of the families in the MPD cohort are from the Middle-East or sub-Indian continent, groups for which limited sequencing data is in publically available datasets (EVS, 1KG). Minor allele frequencies for some populations are available in both 1KG and EVS datasets and this additional data could be utilised by setting the threshold on the basis of the population with the highest occurring maf. An additional consideration is that disease prevalence may differ between population groups, particularly with a higher incidence of many recessive conditions occurring in consanguineous populations (Al-Owain *et al.*, 2012). Therefore a higher minor allele frequency threshold is necessary when filtering variants in a subpopulation with a relatively high incidence of the disorder.

Extensive annotation with a range of consequence prediction programmes assisted in the identification of potentially deleterious variants. Many prediction programmes are based on sequence conservation and related functional domains which is limited in less well-characterised genes (Wu *et al.*, 2013). For example a likely causative variant in Miller syndrome was initially excluded as it was not predicted to be deleterious in any programmes employed (Ng *et al.*, 2010b). Notably, synonymous variants and predicted benign non-synonymous variants (consequence score=3) were filtered out in this pipeline. Therefore causative variants may have been inadvertently removed during the filtering process and further refinement and optimisation of parameters on a study-by-study basis may be necessary. However, currently in the MPD cohort, the large number of synonymous and predicted benign non-synonymous variants precludes a realistic approach to their analysis.

3.7.5 Interpreting 'noisy' genes

During the examination of exome data in this study it was clear that certain genes were more enriched for rare deleterious variants compared to others. These may represent either false-positive, as described in the previous Section, or true-positive variants. Interpreting the numerous variants in these genes can be difficult but understanding why certain genes are noisy can help prioritise those more likely to have phenotypic consequences. For example, some genes, such as those encoding histones, also have a higher degree of non-disease causing genetic variation than those under greater selective constraint (Tennessen *et al.*, 2012). Such related genes have a high degree of functional redundancy and therefore numerous deleterious variants may be present but have no phenotypic consequence (Zhang, 2012). Similarly large genes, for example *TTN* (Titin), are often more tolerant of genetic variation although disease causing mutations can still be present in such genes (Chauveau *et al.*, 2014). Statistical comparison of the frequency of deleterious variants between case and control populations can help identify genes in which deleterious variants are enriched in the patient cohort and therefore more likely to be disease causing (MacArthur *et al.*, 2014). This was performed in this study identifying *FAT1* as a strong candidate disease causing gene in MPD (Section 3.6).

3.7.5 Why WES does not always provide the answer

Given all the above limitations of whole exome capture and sequencing, the probability of detecting a causative mutation in a gene has been estimated at 86% (Ng *et al.*, 2009). This was based on a study in which only four similarly affected patients were sequenced to identify dominant disease causing mutations in a known gene. Therefore the probability of identifying pathogenic variants in disorders where the disease gene and mode of inheritance are more uncertain is likely to be even lower. There are several reasons why WES may fail to identify a disease causing mutation. Firstly, target capture efficiency is not uniform across the genome and variants are less likely to be identified in regions of low coverage. In 35 samples in this study at least 15% of the exome was covered by less than 15 reads. To obtain a sensitivity of 95% in the detection of a heterozygous SNV a read depth of at least 13X is required compared to just 3X in the detection of homozygous SNVs (Meynert *et al.*, 2013). Sensitivity to detect such variants will therefore be reduced in genes which have been inadequately covered especially when identifying compound heterozygous mutations. This is particularly relevant when analysing parent-child segregation of variants, for example, during autosomal recessive and X-linked analysis where such analysis assumes all samples have equal coverage and that all variants have been called and genotyped correctly in each family member. Given the variability in sequencing metrics observed in this study, this is unlikely to always be the case. It is also worth noting that non-paternity or laboratory based errors can occur with samples giving misleading results when filtering using parental data. Performing a SNP array on samples prior to WES can be a useful way to confirm the fidelity of sample handling during preparation especially if this has been out-sourced. This has previously been successfully utilised to identify a sample handling error in the MPD exome cohort (L Bicknell, personal communication).

Secondly, reads containing large insertions, deletions or complex indels are less likely to align to the reference sequence. Such variants have a higher probability of impacting on protein function by causing a frameshift in the coding sequencing and hence are usually prioritised for review. However, indel detection from NGS has much lower sensitivity and specificity compared to SNVs, with higher false positive rates (Grimm *et al.*, 2013) and it is therefore harder to discriminate between error and true variation. One causal deletion, initially missed during variant calling, was identified in neurofibromatosis type 1 syndrome by examining unmapped reads and performing a *de novo* assembly (Chou *et al.*, 2010). This requires extended bioinformatic skill and in this instance the authors had the advantage of prior knowledge of the variant.

Finally, alternative possibilities for the causality of MPD not explored in this Chapter, but also potentially identifiable using WES, include digenic disease, inherited variants with incomplete penetrance, uniparental disomy (UPD), errors in imprinting, copy number variation (CNV) and mosaicism. Alternative filtering strategies could be employed to identify possible causative variants in more than one gene within the same pathway (Yoshimura *et al.*, 2014) and contiguous variants within a chromosomal region which do not appear to conform to Mendelian inheritance can indicate regions of UPD (King *et al.*, 2014). Increasing read depth or changing parameters used for variant calling may assist in identifying mosaic variants (Pritchard *et al.*, 2013) and several programmes now exist to identify CNVs in WES (Tan *et al.*, 2014). The use of one of such programme in CNV detection in WES is further explored in Chapter 6.

In summary, WES provides a method by which to screen a large numbers of genes for pathogenic mutations which has not been previously feasible by other methods. However, interpreting the large number of variants even with efficient filtering strategies remains challenging. Knowledge of limitations can improve the likelihood of success by tailoring experimental design and filtering strategies to the disorder in question. Although, significant progress has been made in the efficiency of target capture, sequencing technologies, accuracy of alignment and variant calling this method is still error prone and in many cases pathogenic variants are unable to be identified. In Chapter 7 alternative approaches for investigating the remaining patients in whom a molecular diagnosis has not yet been identified following WES are discussed.

Chapter 4: Mutations in components of the non-homologous end joining machinery cause MPD

4.1 Introduction

In Chapter 3, mutations were identified in three families in *LIG4* and in one family in *XRCC4* following exome sequencing. Both genes encode proteins with a critical role in non-homologous end joining (NHEJ), the predominant repair mechanism of DNA double-strand breaks within the cell.

4.1.1 Non-homologous end joining (NHEJ)

In NHEJ (Figure 4.1), the damaged termini of DNA at the site of a DSB promote rapid binding of the KU70-KU80 heterodimer to each free end independent of its sequence (Mimori *et al.*, 1986, Walker *et al.*, 2001). This recruits and activates the endonuclease DNA-dependent protein kinase catalytic subunit (DNA-PKcs) (Gottlieb *et al.*, 1993) along with the exonuclease Artemis which also acquires endonucleolytic activity on binding to DNA-PKcs (Ma *et al.*, 2002). The combined activity of Artemis and DNA-PKcs ensures damaged DNA overhangs can be cleaved at a variety of positions (Yannone *et al.*, 2008) allowing the removal of damaged or non-ligatable groups. Infilling of nucleotides is performed by Polymerases mu and/or lambda, as required, through their binding to KU complexes (Ma *et al.*, 2004). Polymerase mu does not require a template to synthesise DNA which particularly lends itself to processing breaks with no sequence homology (Ramadan *et al.*, 2004). The final ligation step is performed by the XRCC4-XLF-LIG4 complex (Ahnesorg *et al.*, 2006), recruited by DNA-PKcs (Drouet *et al.*, 2005). The X-ray repair cross-complementing protein 4 (XRCC4) and Cernunnos (XLF) have similar structures and it has been proposed they alternate to form a filament that wraps around the DNA ends, bridging the break (Hammel *et al.*, 2011, Ropars *et al.*, 2011, Wu *et al.*, 2011, Andres *et al.*, 2012). This may act to stabilise the free ends allowing Ligase IV (LIG4) to complete ligation (Roy *et al.*, 2012, Mahaney *et al.*, 2013).

One important aspect about NHEJ machinery is its flexibility. The KU-DNA end complex is able to recruit the different multifunctional components in any order (Ma *et al.*, 2004) and these different components can also involve a variety of other end processing factors as required. This increases the range of end-processing abilities allowing the repair of various DNA end configurations which may occur from different types of damage but also results in variable outcomes from the processing of identical DSBs (joining heterogeneity) (Ma *et al.*, 2004). Polynucleotide Kinase PNKP (PNKP) is phosphorylated by DNA-PKcs and can interact with XRCC4 (Zolner *et al.*, 2011). It can function to remove terminal blocking groups which prevent DNA ligation (Weinfeld *et al.*, 2011). Aprataxin (APTX) (Clements *et al.*, 2004) and Aprataxin and PNK like factor (APLF) (Iles *et al.*, 2007, Kanno *et al.*, 2007, Macrae *et al.*, 2008) have both also been shown to interact with XRCC4. APLF is both an endo- and exo-nuclease whereas APTX possesses DNA deadenylation activity able to convert 5′AMP groups resulting from unsuccessful ligation reactions. NHEJ components are therefore likely to perform iterative rounds of DNA processing until the free ends are compatible for ligation (Ma *et al.*, 2004). Tyrosyl-DNA phosphodiesterase 1 (TDP1) (Inamdar *et al.*, 2002) and Apurinic/apyrimidinic endonuclease (APE1) (Suh *et al.*, 1997) are also able to process free DNA ends in DSBs however currently no *in vivo* evidence exists demonstrating their direct role in NHEJ.

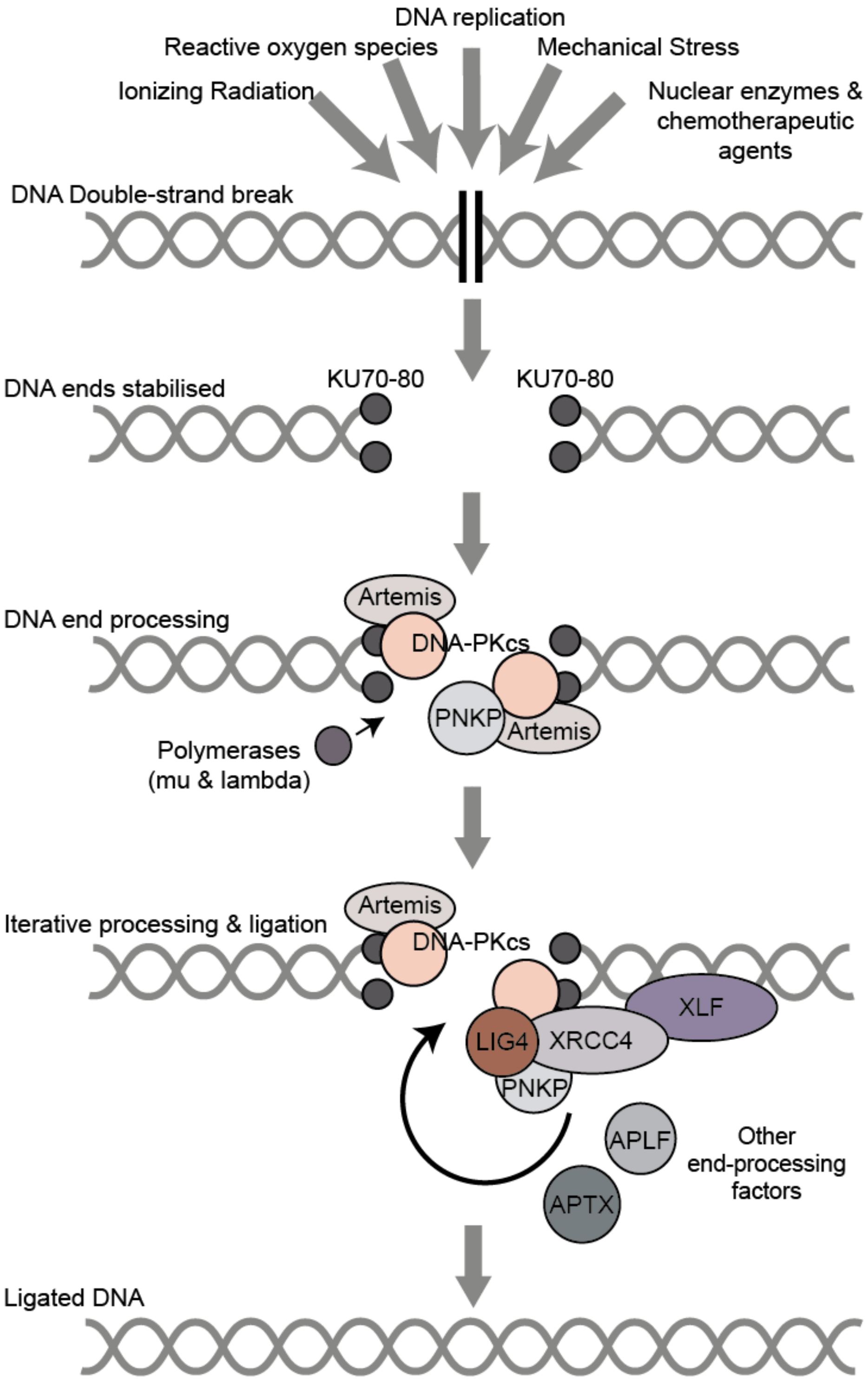


Figure 4.1. Overview of DNA double-strand break repair by NHEJ

The KU complex initially coats the damaged DNA ends recruiting the end processing factors DNA-PKcs and Artemis as well as polymerases for nucleotide infilling. Other enzymes such as PNKP may also be activated by DNA-PKcs to facilitate end processing. Final end joining is performed by the LIG4-XRCC4-XLF complex although iterative rounds of end processing involving the recruitment of further enzymes by XRCC4 may be required until the DNA ends are compatible for ligation by LIG4.

4.1.2 Mutations in NHEJ components cause radiosensitive severe combined immunodeficiency (RS-SCID)

In humans, mutations have been identified in *DCLRE1C* encoding Artemis (Moshous *et al.*, 2001), *PRKDC* encoding the catalytic subunit of DNA-PKcs (DNA dependent protein kinase) (van der Burg *et al.*, 2009), *NHEJ1* encoding Cernunnos-XLF (Buck *et al.*, 2006a) and *LIG4* encoding Ligase IV (Table 4.1). All are associated with increased cellular radiosensitivity and severe combined immunodeficiency due to defective V(D)J recombination leading to early arrest of both B and T cell maturation. Patients with mutations in *DCLRE1C* exhibit RS-SCID without any additional phenotypes whereas mutations in *PRKDC (Woodbine et al., 2013)*, *NHEJ1* (Buck *et al.*, 2006a, Dutrannoy *et al.*, 2010) and *LIG4* (Table 4.1) have been associated with other developmental abnormalities. Most notably these include microcephaly and short stature although growth failure to the degree seen in MPD has not been previously described.

Unlike mutations in other NHEJ genes, immune compromise has not always been evident in cases with *LIG4* mutations (Table 4.1). The first case to be reported was a ‘developmentally normal’ 14 year old boy (with presumably normal growth) who exhibited severe radiosensitivity during treatment for leukaemia without prior immunodeficiency (Riballo *et al.*, 1999, Riballo *et al.*, 2001). As well as in patients with RS-SCID (Buck *et al.*, 2006b, Enders *et al.*, 2006, van der Burg *et al.*, 2006), *LIG4* mutations have been identified in individuals with milder degrees of immunodeficiency, microcephaly, developmental delay and pancytopenia (O'Driscoll *et al.*, 2001, O'Driscoll *et al.*, 2004). Ligase IV syndrome has therefore often been considered as a rare differential to Nijmegen breakage syndrome (caused by mutations in Nibrin, *NBN*) characterised by microcephaly, mild immune dysfunction and malignancy predisposition as well as a cause of RS-SCID.

Following the identification of three MPD cases with *LIG4* mutations in Chapter 3, sequencing of *LIG4* was undertaken in the remainder of the cohort. Given the rarity of ligase IV syndrome, it was somewhat surprising that this identified a further seven families with truncating mutations in *LIG4* all initially presenting with severe growth failure. In this Chapter, more extensive characterisation of the Ligase IV phenotype is performed as well as the description of a genotype-phenotype correlation. As XRCC4 is a direct binding partner of LIG4 and another important component of NHEJ, resequencing of *XRCC4* was also performed in the remainder of the MPD cohort identifying a further five families, all with nonsense mutations in *XRCC4.* This substantiates *XRCC4* as a new disease causing gene in MPD and allows a common phenotype to be described which overlaps *LIG4* patients. These findings place mutations in *LIG4/XRCC4* as the second most common cause of MPD in this cohort behind mutations in *PCNT* (MOPD II).

Table 4.1. **Description of previously published cases with mutations in *LIG4***

Reference	*Mutations*	*Place of Origin*	*Sex*	*IUGR*	*Postnatal growth failure*	*Develop-ment*	*Immunodeficiency*	*Clinical Course*
Ben-Omran *et al.*, (2005)	p.Arg814* p.Arg814*	Europe	M	Yes	OFC -7.76sd Normal stature	Moderate delay	None evident	T-cell acute lymphoblastoid leukaemia, died at 4.5yrs from sepsis after chemotherapy induction
O'Driscoll *et al.*, (2001)	p.Arg814* p.Arg580*	- *Siblings*	-	-	'microcephaly and growth retardation'	-	Chronic respiratory infections	Photosensitivity, psoriasis, hypothyroid, amenorrhea
			-	-	'microcephaly and growth retardation'	-	Chronic skin conditions & sinusitis	Myelodysplasia, photosensitivity, telangiectasia, hypothyroid, type II diabetes, hypogonadism
O'Driscoll *et al.*, (2001), Gruhn *et al.*, (2007)	p.Arg814* p.Gln469Glu	Europe	F	Yes	OFC -8.24sd Height -2.83sd	Delayed	Recurrent respiratory infections	Myelodysplasia, bone marrow failure from 5yrs, Multiple psoriasiform erythrodermic squamous skin patches, atypical bone maturation, Had successful BMT
O'Driscoll *et al.*, (2001)	p.Arg278His p.Arg278His	-	-	Microcephaly	None	Global delay	Extensive plantar warts	Cytopenia
Plowman *et al.*, (1990), Riballo *et al.*, (1999)	p.Arg278His p.Arg278His	Europe	M	-	-	-	-	T cell lymphoblastic leukaemia at 14yrs
Buck *et al.*, (2006b)	p.Gln280Arg p.Lys424fs*20	Morocco *Siblings*	F	No	OFC -5.04sd 'growth & weight retardation'	-	Severe	Repeated infections since 3m. BMT at 19mths, developed EBV associated lymphoproliferative syndrome and died shortly after
			F	No	Microcephaly	-	Severe	BMT at 2mths but died shortly after following veno-occlusive disease

van der Burg *et al.*, (2006)	p.433delGln p.433delGln	Europe	F	-	No	Normal	RS-SCID	SCID diagnosed at 2yrs. Died shortly after conditioning for BMT
Enders *et al.*, (2006)	p.His282Leu p.Lys424fs*20	*Siblings*	F	Microcephaly	OFC<3rd centile Height at 10th centile	Normal	RS-SCID	SCID diagnosed at 2yrs. EBV encoded B cell non-Hodgkin's lymphoma
			F	Microcephaly	OFC<3rd centile Height at 25th centile	Significant delay	RS-SCID	Underwent successful BMT
Yue *et al.*, (2013)	p.Arg814* p.Leu205fs*17	Europe	F	No	OFC -6.62sd Height -3.86sd	Normal	Croup & bronchitis, poor social circumstances	Anal squamous cell carcinoma at 34yrs, died after severe reaction to local radiotherapy for malignancy. Preceding cytopenia.
Yamada *et al.*, (2001), Toita *et al.*, (2007)	p.Met249Val p.Lys424fs*20	Japan	F	No	OFC -5.52sd Height -3.64sd	Normal	Combined immunodeficiency	Episode of thrombocytopenia at3yrs, Pre-axial polydactyly, EBV associated large B cell lymphoma. Died during chemotherapy treatment.
Grunebaum *et al.*, (2008)	p.Ala845Thr p.His282Leu p.Arg581fs*	-	F	Microcephaly	Microcephaly Short stature	-	-	Protracted universal scaly erythroderma, hepatosplenomegaly, generalised lymphadenopathy and diarrhoea. Underwent successful BMT
Unal *et al.*, (2009)	p.588delLys p.588delLys	Europe *Siblings*	F	Yes	OFC & height <3rd centile	Normal	Recurrent respiratory tract infections	Progressive bone marrow failure from 8 yrs. Underwent successful BMT.
			F	No	OFC & height <3rd centile	Normal	Recurrent respiratory tract infections	Progressive bone marrow failure from 3.5 yrs.

Growth parameters given as standard deviations where measurements were reported otherwise description stated as reported in article. Dash indicates where no information was provided. *Abbreviations: RS-SCID=radiosensitive severe combined immunodeficiency, BMT=bone marrow failure, EBV=Epstein-Barr virus*

4.2 Results

4.2.1 Resequencing of *LIG4* in MPD patients

In Chapter 3, analysis of WES data using an autosomal recessive model of inheritance identified three cases with two heterozygous truncating mutations in *LIG4* (consequence score=1) (Figure 4.2A). All mutations were then validated in the affected by capillary sequencing (Figure 4.2B). The coding exon of *LIG4* was then sequenced in 138 patients with MPD in whom exome sequencing had not been previously performed. This identified a further five cases, all with two heterozygous truncating mutations in *LIG4* (in one family sequencing was performed collaboratively in The Netherlands, (Ijspeert *et al.*, 2013)). Two additional cases were then identified through WES by a collaborating group in Cologne, Germany, both of which also had a clinical diagnosis of MPD. Parental DNA was available in eight families, and all were found to be heterozygous carriers following capillary sequencing, confirming the mutations to be biallelic in the offspring (Table 4.2). All parents were reported to be healthy. In one family, there was also a similarly affected sibling (F1.2) and sequencing confirmed the same *LIG4* mutations were present in both affected offspring.

In total, 11 patients from 10 families were identified with biallelic truncating mutations in *LIG4* (Murray *et al.*, 2014). One further patient with MPD was found to carry the heterozygous truncating mutation, c.128delT, p.L43*, but a second likely deleterious variant was not observed in the remaining coding sequence. Cellular testing for sensitivity to ionising radiation, however, has been recommended as a second pathogenic non-coding variant may be present. *LIG4* was also sequenced in 43 patients with primary microcephaly and in 56 with microcephaly and milder short stature (OFC<4sd and height -2 to -4sd from population mean) but no further mutations were found. In total, eight different truncating mutations were identified, six of which have not been previously reported. One mutation, c.2440C>T (p.Arg814*), was identified recurrently in nine families. This mutation was also present in control datasets at a low frequency in the heterozygous state (dbSNP: rs104894419, EVS: maf=0.00015, 1000genomes: maf=0.0005). Comparing the frequency of the c.2440C>T allele reported in EVS with the frequency identified in this cohort (0.015) showed a significant enrichment of c.2440C>T in MPD patients (two-tailed p=0.0001 using Chi squared test with Yates correction).

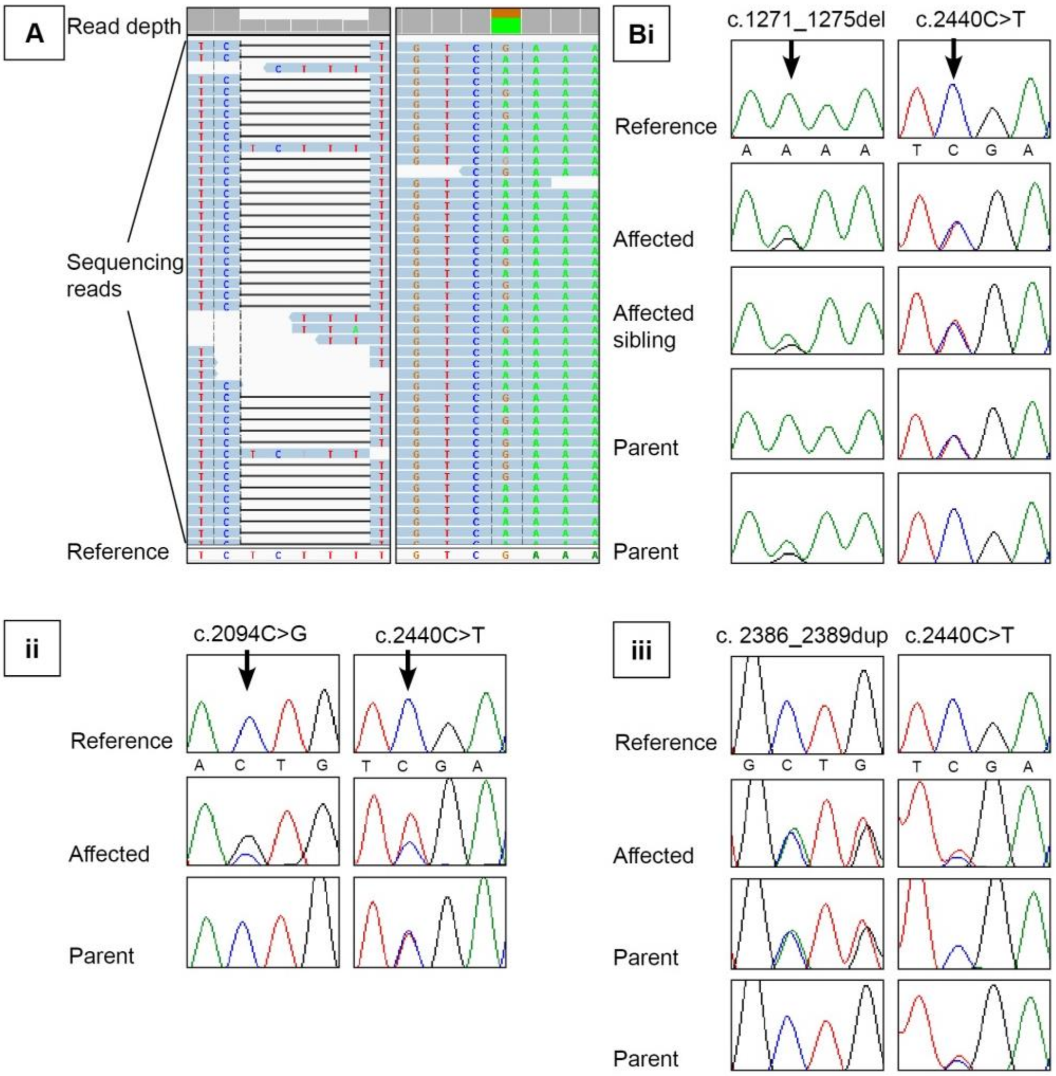


Figure 4.2. Mutations identified in three patients in LIG4 following WES

A) Sequencing reads following WES aligned to the *reverse compliment* of the *LIG4* reference sequence using the Integrative Genomics Viewer (IGV) demonstrating compound heterozygous truncating mutations identified in F1.1; c.1271_1275delAAAGA (horizontal line identified in 160 out of 335 reads) and c.2440C>T (green/brown identified in 76 out of 132 reads). B) Chromatograms following capillary sequencing in each child and parents/siblings (where DNA available) confirmed mutations segregated appropriately with the phenotype in i) F1.1 ii) F2 and iii) F3.

Table 4.2. ***LIG4* mutations identified in 10 families with a clinical diagnosis of MPD**

Patient	*Nucleotide change*	*Protein change*	*Country of Origin*	*Mother*	*Father*	*Size of p.R814* associated haplotype*

F1.1	c.[2440C>T]
	+[1271_1275delAAAGA]
		p.(Arg814*)
		p.(Lys424Argfs*20)
			Canada	p.(Arg814*)	p.(Lys424Argfs*20)	1.49Mb

F1.2	c.[2440C>T]
	+[1271_1275delAAAGA]
		p.(Arg814*)
		p.(Lys424Argfs*20)
			Canada	p.(Arg814*)	p.(Lys424Argfs*20)	1.49Mb

F2	c.[2440C>T]
	+[2094C>G]
		p.(Arg814*)
		p.(Tyr698*)
			USA	p.(Arg814*)	N/A	1.29Mb

F3	c.[2440C>T]
	+[2386_2389dupATTG]
		p.(Arg814*)
		p.(Ala797Aspfs*3)
			Australia	p.(Arg814*)	p.(Ala797Aspfs*3)	None

F4	c.[2440C>T]
	+[1271_1275delAAAGA]
		p.(Arg814*)
		p.(Lys424Argfs*20)
			UK	p.(Lys424Argfs*20)	p.(Arg814*)	>2Mb

F5	c.[2440C>T]
	+[1271_1275delAAAGA]
		p.(Arg814*)
		p.(Lys424Argfs*20)
			USA	p.(Arg814*)	p.(Lys424Argfs*20)	None

F6	c.[2440C>T]
	+[1271_1275delAAAGA]
		p.(Arg814*)
		p.(Lys424Argfs*20)
			Germany	p.(Arg814*)	p.(Lys424Argfs*20)	1.19Mb

F7	c.[2440C>T]
	+[1512_1513delTC]
		p.(Arg814*)
		p.(Arg505Cysfs*12)
			USA	N/A	N/A	N/A

F8	c.[2440C>T]
	+[1246_1250dupGATGC]
		p.(Arg814*)
		p.(Leu418Metfs*3)
			UK	N/A	N/A	>2Mb

F9	c.[2440C>T]
	+[1271_1275delAAAGA]
		p.(Arg814*)
		p.(Lys424Argfs*20)
			Turkey	p.(Arg814*)	p.(Lys424Argfs*20)	None

F10*	c.[613delT]
	+[1904delA]
		p.(Ser205Leufs*29)
		p.(Lys635Argfs*10)
			The Netherlands	p.(Ser205Leufs*29)	p.(Lys635Argfs*10)	N/A

Reproduced from Murray *et al*, 2014. F1.1 and F1.2 represent individual affected siblings. All mutations are predicted to cause premature truncation of the protein.

*Sequencing and identification of mutations occurred in The Netherlands (Ijspeert *et al.*, 2013). See Appendix II for details of reference sequence. N/A=not available.

4.2.2 Haplotype analysis in families sharing the recurrent p.R814* mutation

To determine whether c.2440C>T (p.Arg814*) represented a founder mutation in these families, polymorphic SNPs in a 2 Mb region surrounding *LIG4* were genotyped in the nine families (Figure 4.3). This identified a common haplotype of at least 2 Mb in two families (F4 and F8) and three families (F1, F2 and F6) shared 6-12 SNPs of the same haplotype immediately adjacent to the gene (Murray *et al.*, 2014). This established that c.2440C>T represents a founder mutation in at least two of the families and potentially five families in total. However in the remaining three families no common haplotype was seen indicating a mutation at this nucleotide has occurred on multiple occasions. This provides clear evidence that c.2440C>T is the causative mutation in these families rather than another co-segregating variant. Notably this nucleotide is part of a CpG motif which, when methylated, is prone to spontaneously deaminate (Scarano *et al.*, 1967). CpG motifs are therefore frequent sites of C>T mutations.

4.2.2 Characterisation of phenotype associated with truncating *LIG4* mutations

4.2.2.1 Growth impairment

Intrauterine growth retardation was observed in all cases (mean birth weight -3.0 s.d., birth OFC -3.6 s.d. and length -3.8s.d.) (Figure 4.4, Table 4.3). Growth then continued to be delayed postnatally with a severe reduction in weight, OFC and height. Again this was present in all cases (mean weight -6.8 s.d., OFC -10.1 s.d. and height -5.1 s.d.). Disproportionate microcephaly was also universally observed with OFC substantially more reduced than height (p=0.0001). Growth parameters of previously described cases (where available) are shown in Table 4.1 and Figure 4.4 (grey) (mean birth weight -1.3 s.d., birth OFC -1.5 s.d., birth length -1.5 s.d., postnatal weight -2.4 s.d., OFC -5.6 s.d. and height -2.4 s.d.). Notably, none of these cases had growth parameters which fell into the defined range for MPD (see Section 1.1.2). The growth failure observed in the patients identified in this study was found to be significantly greater than those cases previously reported (Figure 4.4) (Murray *et al.*, 2014).

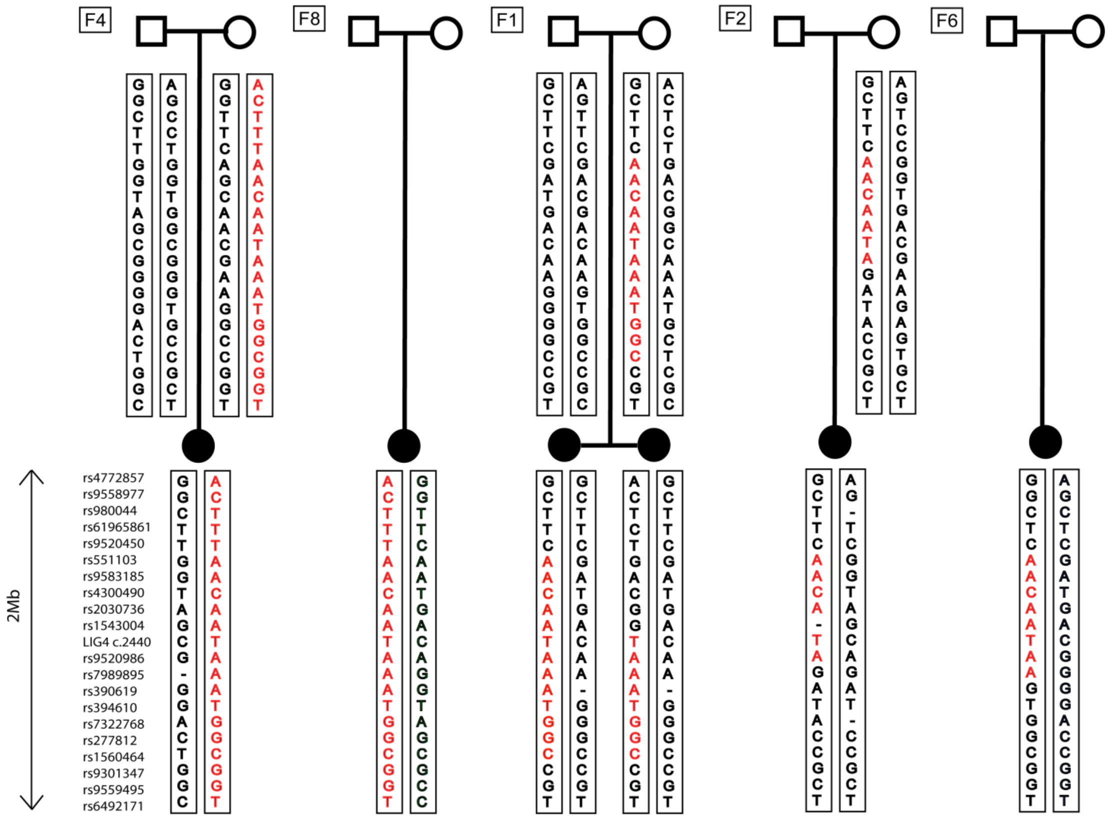


Figure 4.3. Haplotype analysis in 5 families sharing the recurrent c.2440C>T (p.R814*) mutation

Reproduced from Murray *et al*, 2014. A common 2Mb haplotype is present in two families, F4 & F8. F1, F2 and F6 also shared a smaller region of identical SNPs adjacent to *LIG4*, suggesting that the c.2440C>T mutation may have the same ancestral origin in these families. Dash indicates where the nucleotide could not be confidently identified following sequencing.

4.2.2.2 Facial characteristics, malformations and intellectual development

Clinical photographs were available for review in nine cases (Figure 4.5). All were noted to share similar facial characteristics particularly in early childhood. Common features included fine sparse hair, epicanthic folds, wide depressed nasal bridge, broad nasal tip and prominent chin. No dental problems were reported. Malformations were observed in some cases but were only present at a low frequency (Table 4.3). The most commonly reported malformation was congenital hip dysplasia occurring in three individuals. One case was also reported to have severe vomiting episodes occurring from birth and this had resulted in significant feeding difficulties throughout childhood. The cause for this has not been established. Most of the cases had not yet reached pubertal age but of the two oldest patients (F1.1 and F1.2) primary ovarian failure was observed in both.

In three cases development was reported to be normal whereas six showed evidence of mild developmental delay and one exhibited moderate delay. Where delay was present, expressive language skills were predominantly affected. All school-aged children were in mainstream school although some patients required additional support. Development could not be assessed in one case (F10) who died at 6 months of age from overwhelming sepsis.

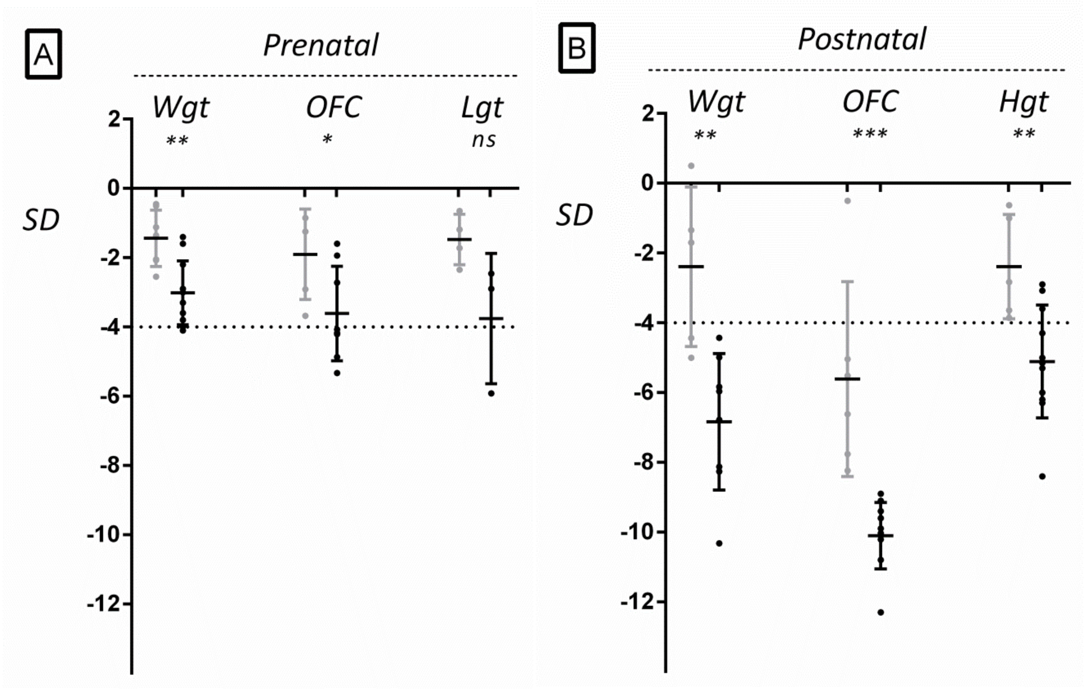


Figure 4.4. Growth impairment in patients with *LIG4* truncating mutations

Adapted from Murray *et al*, 2014. Measurements plotted as Z-scores (standard deviation of measurement from population mean for age and sex) for previously reported cases (grey) alongside the patients reported here (black) for each growth parameter. A) Measurements at birth of MPD cases (mean weight -3.0 +/-0.93 s.d. (n=11), OFC -3.6 +/-1.37 s.d. (n=8) and length -3.8 +/-1.88 s.d. (n=3)) compared to previously reported cases (mean weight -1.3 +/-0.8 s.d. (n=6), OFC -1.5 +/-1 s.d. (n=4) and length -1.5 +/-0.7 s.d. (n=4)). B) Most recent postnatal measurements of MPD cases (weight -6.8 +/-1.96 s.d., OFC -10.1 +/-0.95 s.d. and height -5.1 +/-1.62 s.d.) compared to previously reported cases (weight -2.4 +/-2.3 s.d. (n=5), OFC -5.6 +/-2.8 s.d. (n=6) and height -2.4 +/-1.5 s.d. (n=5)). Mean and standard deviation of patient group indicated by horizontal and vertical bars respectively. * p<0.05; ** p<0.01; *** p<0.001 (unpaired t-test).

Table 4.3. Anthropometric data and clinical findings of patients with truncating *LIG4* mutations

			A. Anthropometric Data					***B. Clinical Data***	
Pt	***Sex***	***Gest^n /weeks***	***BW /s.d. (kg)***	***Birth OFC /s.d. (cm)***	***Age at exam***	***OFC /s.d. (cm)***	***Height /s.d. (cm)***	***Develop -mental Delay***	***Malformations & additional features***
F1.1	F	38	**-3.6** (1.59)	**-4.87** (27.75)	17y6m	**-9.4** (42.5)	**-6** (127)	Mild	Small cerebral aneurysm, primary ovarian failure
F1.2	F	40	**-2.2** (1.96)	N/A	11y9m	**-9.9** (41.5)	**-4.3** (117.5)	Mild-moderate	Atrial-ventricular septal defect, atrophic kidney, rib hypoplasia, fusion of carpal bones, copper beaten skull, platybasia, abnormal C1 vertebrae and primary ovarian failure
F2	F	33	**-3.3** (1.01)	**-4.07** (25)	7y10m	**-12.3** (37.9)	**-5** (99)	Mild	Anal atresia with rectovaginal fistula, esotropia
F3	F	37	**-3.3** (1.58)	**-4.17** (28)	2y2m	**-10.8** (36.3)	**-6.3** (68.5)	Mild	Unilateral congenital hip dysplasia, cutis marmorata
F4	F	40	**-3.8** (1.84)	N/A	2y6m	**-8.9** (39)	**-3.6** (78.4)	None	Psoriasis
F5	M	38	**-1.6** (2.44)	**-2.72** (30.5)	2y	**-9.1** (38.2)	**-5.3** (70.4)	None	Unilateral congenital hip dysplasia
F6	F	32	**-2.95** (0.95)	**-1.94** (27)	2y	**-10.21** (36.5)	**-3.08** (75)	Mild	Congenital hip dysplasia, 2/3 toe syndactyly, excessive vomiting (gastrostomy in situ)
F7	F	37	**-4.1** (1.29)	**-4.21** (27.5)	3y8m	**-10.03** (38.8)	**-5.15** (79)	None	None
F8	F	34	**-2.93** (1.23)	N/A	1y9m	**-9.6** (37)	**-6.2** (65)	Mild	None
F9	M	37+	**-1.4** (2.25)	**-1.6** (30.8)	5y6m	**-10.8** (39)	**-2.9** (98)	Mild	Hypopigmentation, hypermobile knees, single palmar crease, 2/3 toe syndactyly, sandal gap
F10	M	37	**-4.0** (1.3)	**-5.33** (26.5)	3m	**-10.1** (29)	**-8.4** (43)	N/A*	Pre-axial polydactyly 2/5 toe syndactyly, dysplastic kidney and dysgenesis of corpus callosum.

Reproduced from Murray *et al*, 2014. Anthropometric data stated as Z scores (standard deviation from population mean for age and sex), actual measurements in brackets.

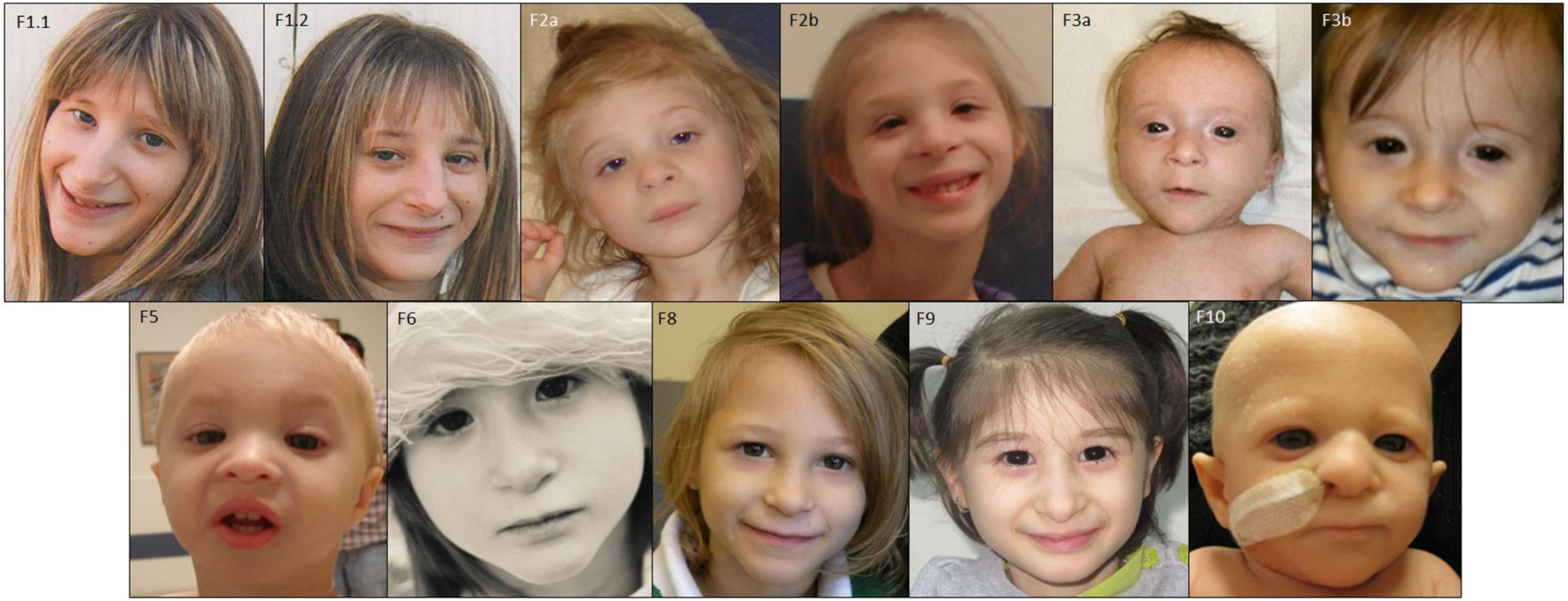


Figure 4.5. Facial features of nine *LIG4* cases with microcephalic primordial dwarfism

Reproduced from Murray *et al*, 2014. Age of patient at time of photos; F1.1 at 20 years, F1.2 at 14 years, F2 at (a) 3 years and (b) 6 years 10months, F3 at (a) 3 months and (b) 10months, F5 at 20 months, F6 at 4 years, F8 at 6 years, F9 at 6yrs 2 months and F10 at 5 months.

4.2.2.3 Haematological abnormalities

At time of recruitment to the study, three patients were reported to have experienced episodes of pancytopenia with no apparent trigger. This was subsequently reported in a further six patients, the first episode occurring after recruitment to the study (Table 4.4). The earliest reported episode occurred at two years of age (F3 and F4) and the latest at 15 years (F1.1) (Murray *et al.*, 2014).

Platelets were always the most significantly affected cell type (calculated as percent of normal; 11.8%, p=0.0001) followed by leukocytes (33.7%, p=0.0001) with only a mild reduction in haemoglobin levels (79.3%, p=0.002) (Figure 4.6A, Table 4.4). Seven patients followed a progressive course of bone marrow failure with an increase in transfusion frequency with age (Figure 4.6B). One patient (F1.1) experienced pancytopenia over a six month period requiring regular platelet transfusions but this spontaneously improved and whilst she is no longer transfusion dependent indices still remain subnormal on recent investigation (Table 4.4). Only one patient has had no documented evidence of cytopenia at the age of three years (F5). Finally, in patient F10 cytopenia occurred in association with acute sepsis towards the end of life at five months of age (Ijspeert *et al.*, 2013). No evidence of malignancy of any type was reported in any of the 11 cases with the oldest patient being 21 years at time of writing. Bone marrow aspiration was performed in four individuals and demonstrated a hypocellular marrow but no morphological abnormalities were seen.

4.2.2.3 Immune dysfunction

On recruitment to the study, immunodeficiency had not been clinically suspected in 10 of the 11 cases. However, on retrospective questioning, several children (F1.2, F3, F6 and F9) have had recurrent common childhood illnesses (respiratory, skin and gastrointestinal) and one patient (F1.1) had a severe influenza illness requiring hospitalisation despite vaccination (Murray *et al.*, 2014). Severe immunodeficiency had only become clinically apparent in one patient (F10) at an early age with recurrent and opportunistic infections from birth. The patient unfortunately died at six months of age due to an acute gastric bleed in association with a clinical sepsis syndrome. Immune function was noted to be profoundly impaired on investigation in this patient (Ijspeert *et al.*, 2013) (Table 4.5).

Subsequent to diagnosis with Ligase IV deficiency, detailed immunological investigations have been performed in many of the remaining patients. Eight of the nine patients tested demonstrated hypogammaglobulinemia, particularly low IgG, and in the eight patients for which T/B cell subset data is available, B cell counts were found to be severely depleted with a variable reduction in T cells (Table 4.5). Vaccination response was also found to be generally impaired where assessed. This has impacted on the clinical management of these patients and all are now under the care of a specialist paediatric immunologist. Currently bone marrow transplantation is underway in patient F3, and is also being considered in several others (F1.2, F6 and F7).

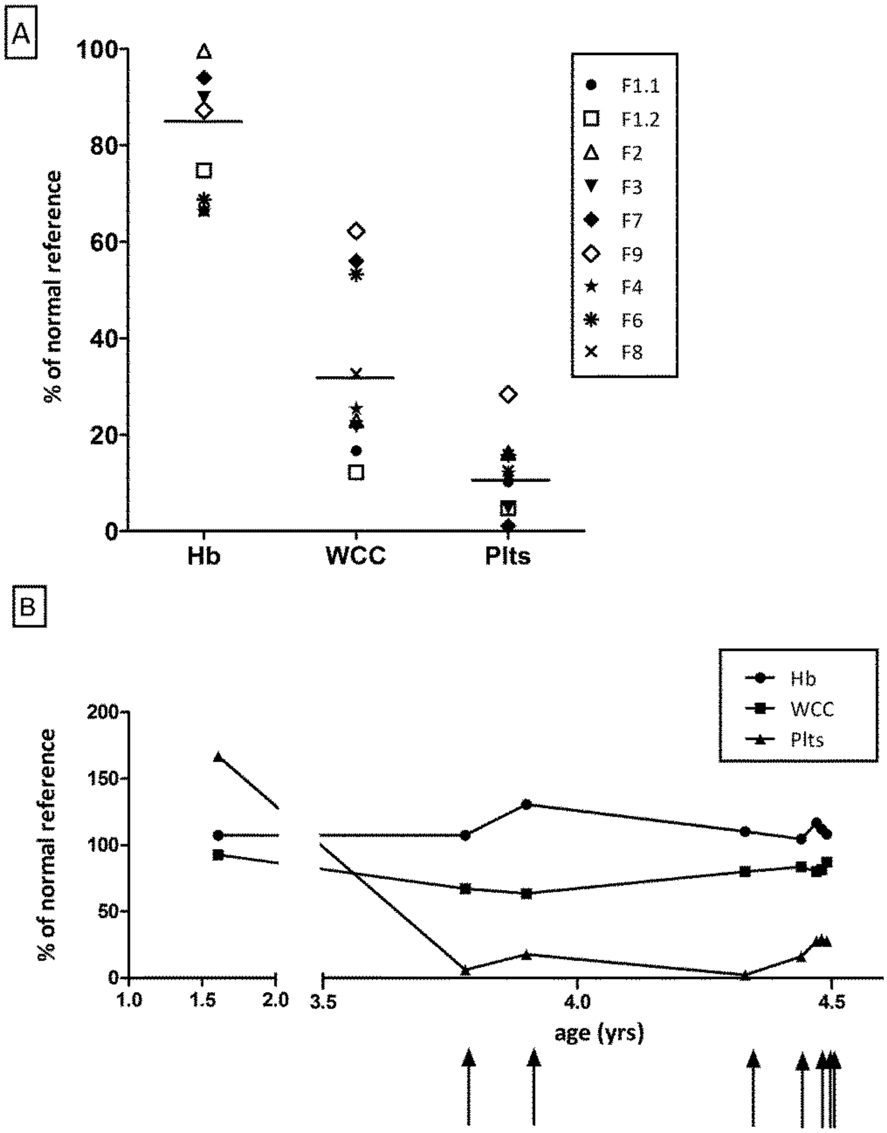


Figure 4.6. Haematological parameters in *LIG4* patients during episodes of cytopenia

Reproduced from Murray *et al*, 2014. A) Documented blood counts from nine patients during the most recent cytopenic episode. Values are shown as percentage of average normal count: Haemoglobin 79.3%+/-13 s.d. (p=0.002), leucocytes 33.7%+/-19 s.d. (p=0.0001) and platelets 11.8%+/-8 s.d. (p=0.0001). Platelets were the most significantly affected cell type. B) Serial blood counts in one patient (F8) over time demonstrating increasing transfusion dependency with age. Arrows indicate when platelet transfusion occurred. Statistical comparisons performed using unpaired t-test.

Table 4.4. Clinical course of bone marrow failure in *LIG4* patients and details of haematological investigations

A. Features at referral			*B. Current Haematological Test Results*						
Pt	***Age of referral to study***	***Pancyto-penia***	***Onset of cyto-penia***	***Age at I^x***	***Hb***	***WCC***			***Plts***
					g/dl	***Total $10^3/\mu l$***	***Neut %***	***Lymph %***	***$10^3/\mu l$***
F1.1	17y	6month period, self-resolved	15y	20y	**8.2**	**1.5**	53	40	**30**
F1.2	14y	Persistent	8y	14y	**10.1**	**1.1**	36	64	**14**
F2	3y	One episode, self-resolved	3y	3y	12.6	**1.8**	55	24	**43**
F3	4m	None	2y	2y	11.1	**2**	30	40	**18**
F4	2y	None	2y	6y	**8.2**	**2.34**	N/A	N/A	**32**
F5	1y	None	None at 3y	N/A	N/A	N/A	N/A	N/A	N/A
F6	Neonate	None	3y 10m	3y 10m	**8.5**	**4.9**	55	32	**42**
F7	3y8m	None	3y9m	4y4m	11.9	**4.9**	54.3	26	**3**
F8	4y	None	6y	6y2m	**8.2**	**3**	N/A	N/A	**33**
F9	2y6m	None	6y	6y	11.7	5.6	60.8	27.2	**78.1**
F10	4m	None	5m	6m	**4.4**	35	90	**7**	**58**

Adapted from Murray *et al*, 2014. Absolute full blood count values are the most recent recorded in which cytopenia is demonstrated. Normal range of absolute counts varies with age, sex and laboratory methods but approximate reference ranges are as follows: Hb 10.8-13.9g/dl, Total WCC 5.5-12.9 x$10^3/\mu l$, Neut 30-60%, Lymph 20-50%, Plts 130-400x$10^3/\mu l$. Levels considered below normal range (according to local laboratory) are highlighted in bold. *Abbreviations: Hb = Haemoglobin, WCC = White Cell Count, Plts = Platelet count, Neut = Neutrophil count, Lymph = Lymphocyte count***.** **N/A = not available or not performed.*

Table 4.5. Clinical course of bone marrow failure in *LIG4* patients and details of haematological investigations

A. Features at referral			*C. Current Immunological Test Results*								
Pt	*Age of referral to study*	*Increase in Infection*	*Age at I^x*	*Immunoglobulins g/dl*			*T and B Lymphocyte subset Cells/µl*				*Antibody response to Vaccination*
				IgG	*IgA*	*IgM*	*CD3*	*CD8*	*CD4*	*CD19*	
F1.1	17y	Mild	15y	**4.31**	1.06	0.14	680	380	**240**	**0**	N/A
F1.2	14y	Mild	14y	**5.59**	1.29	0.76	660	370	**190**	**10**	N/A
F2	3y	None	Awaiting investigations								
F3	4m	Mild	2y	**3.7**	0.58	0.74	**710**	**170**	500	**very low**	Poor response to pneumococcal vaccine
F4	2y	None	6y	**Low**	N/A	N/A	N/A	N/A	**low**	**very low**	Poor response to pneumococcal vaccine
F5	1y	None	2y	**3.66**	0.8	0.62	**467**	**204**	**164**	**6**	N/A
F6	Neonate	Mild	4y	**5.76**	4.23	0.27	983	292	583	**11**	Tetanus and Diphtheria normal, HiB and pneumococci low
F7	3y8m	None	Awaiting investigations								
F8	4y	None	6y	**4.8**	2.4	**<0.3**	N/A				N/A
F9	2y6m	Mild	5y6m	9.7	3.1	1.0	normal	normal	normal	**very low**	N/A
F10	4m	Severe	4m	**0.76**	**0.22**	0.85	5730	2410	3380	**60**	N/A

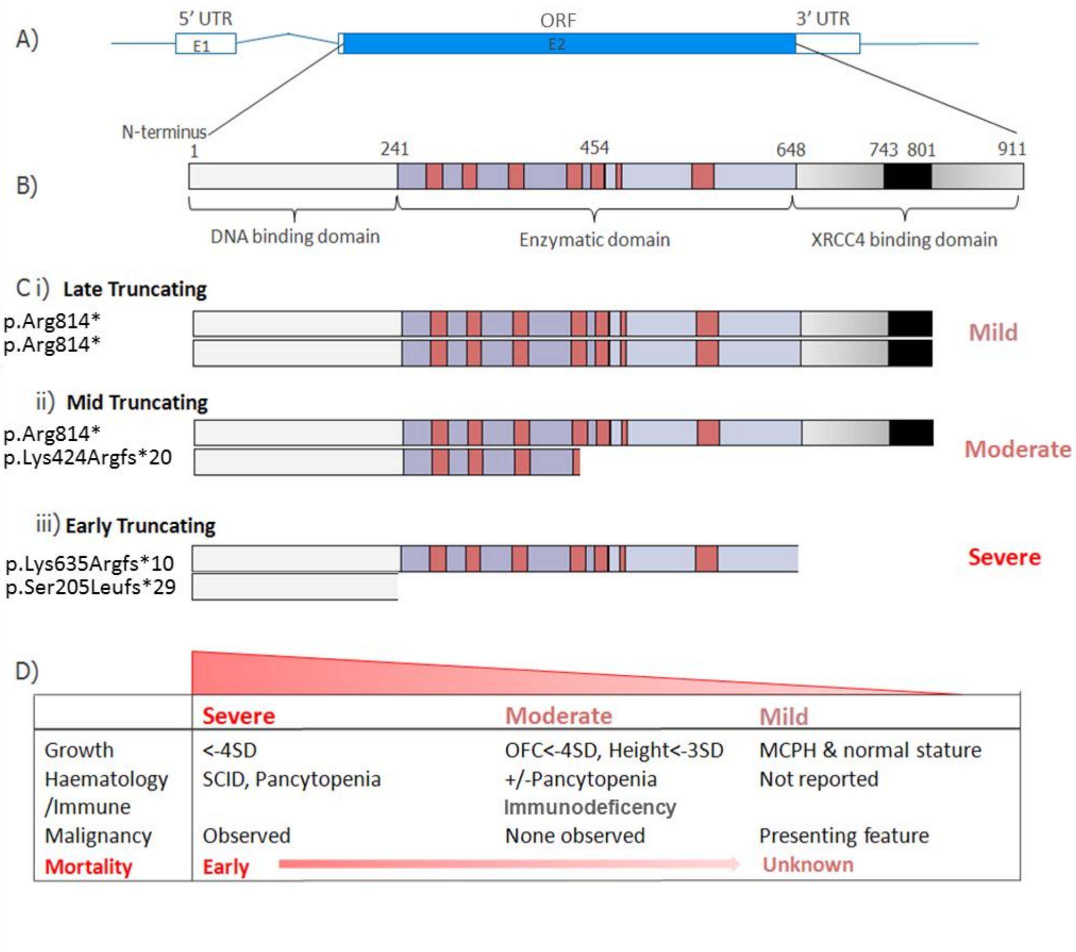
Adapted from Murray *et al*, 2014. Increase in infections: 'Mild' defined as an increased frequency of common pathogens (upper & lower respiratory, skin and gastrointestinal) or increased severity of a single illness above that deemed normal for childhood by their clinician. Normal range of absolute counts varies with age, sex and laboratory methods but approximate reference ranges are as follows: IgG 5.65-17.65g/dl, IgA 0.85-3.85g/dl, IgM 0.55-3.75g/dl, Total T lymphocytes (CD3+) 870-2080/µl, T helper subset (CD4+) 530-1290/µl, T cytotoxic subset (CD8+) 220-960/µl, B lymphocytes (CD19+) 100-400/µl. Levels considered below normal range (according to local laboratory) are highlighted in bold. In F1.1 and F1.2, naïve CD4 and CD8 cells were reported to be very low. In F5, total lymphocyte count was also reported to be reduced.

4.2.5 Sensitivity to ionising radiation in patient cells

Assessment of radiosensitivity of lymphoblastoid cells was performed in three patients (F2, F5 and F7) by S. Haghayegh (UCLA Department of Pathology & Laboratory Medicine, Los Angeles, USA) using a colony survival assay previously established for the testing of patients with ATM mutations and ataxic telangiectasia (Sun *et al.*, 2002). Increased sensitivity to DNA damage was observed in patient cells, with reduced survival of cells (3 to 5% cell survival) following exposure to 1 Gray of ionising radiation (lab reference range for wild-type cells: 50.1% +/-13.5, n=24). Notably, this was greater than the increase in sensitivity seen in ataxic telangiectasia patient cells (13.1% +/-7.2, n=104) and that observed in previous studies of *LIG4* patients, <30% cell survival at 1 Gray (O'Driscoll *et al.*, 2001). Increased cellular sensitivity to ionising radiation was also present in patient F10, <5% cell survival (Ijspeert *et al.*, 2013).

4.2.6 Delineation of a phenotype-genotype correlation in *LIG4* patients

The difference in growth phenotype between the cases identified in this study and those previously reported led me to examine if any genotype-phenotype correlation was evident. As the *LIG4* gene contains two exons (Figure 4.7A), with only exon 2 encoding the 911aa protein, transcripts should not be targeted for nonsense mediate decay and thus truncating mutations would not be anticipated to be null. The protein is comprised of an enzymatic domain containing highly conserved sequence motifs (Shuman *et al.*, 1995, Marchetti *et al.*, 2006), and a C-terminal XRCC4 binding domain comprised of two BRCT domains with a linker region that contains the minimal XRCC4 binding motif (Sibanda *et al.*, 2001) (Figure 4.7B). Mutations were defined as late-truncating if occurring after the XRCC4 binding motif, mid-truncating if located between the start of the enzymatic domain and the end of the XRCC4 binding motif and early truncating if occurring before the start of the enzymatic domain. The recurrent c.2440C>T, p.Arg814* mutation, represents the most distal truncating mutation reported and only affects the last BRCT domain of the XRCC4 binding domain (Figure 4.7C). Only one patient homozygous for this mutation has been reported in the literature to date. They exhibited microcephaly with normal stature (Ben-Omran *et al.*, 2005) and developed leukaemia at four years of age. No prior immune dysfunction or cytopenic episodes had been noted in this patient. The majority of cases identified in this study (F1.1-F9) also carry the c.2440C>T mutation but combined with a mid-truncating mutation on the alternate allele. These patients displayed a more severe growth phenotype with chronic or progressive cytopenia and evidence of immune dysfunction on investigation. The most severely affected case in this study presented with SCID early in life and severe growth failure (F10). They carried a 'mid-truncating' mutation which completely removed the XRCC4 binding site in combination with an 'early' truncating mutation that removed the entire enzymatic domain. Severity of the phenotype therefore appears to correspond to the degree of protein truncation with severe growth retardation only occurring when the critical XRCC4 binding region (between the two BRCT domains) has been disrupted in at least one allele. Immune compromise and bone marrow failure were also observed in these patients but severe immunodeficiency (SCID) was only evident in the patient where the XRCC4 binding region had been totally lost from both alleles. As XRCC4 binding confers stability to LIG4 (Bryans *et al.*, 1999), variation in the strength of this interaction may in part explain the differing degrees of phenotype severity seen with different mutations.



	Severe	Moderate	Mild
Growth	<-4SD	OFC<-4SD, Height<-3SD	MCPH & normal stature
Haematology /Immune	SCID, Pancytopenia	+/-Pancytopenia Immunodeficency	Not reported
Malignancy	Observed	None observed	Presenting feature
Mortality	Early		Unknown

Figure 4.7. Degree of protein truncation correlates with severity of growth failure and immune deficiency

Reproduced from Murray *et al*, 2014. **A)** Schematic of the *LIG4* gene. **B)** LIG4 protein domain structure: enzymatic domain (purple) with highly conserved sequence motifs (pink), XRCC4 binding domain (grey) with critical binding motif (black). **C, D)** Position of truncating mutations correlating with phenotype. **C)** Biallelic mutations representative for each group are depicted: **(i)** Late Truncating mutations are associated with the mildest phenotype **(ii)** Late & mid truncating were observed in patients with severe growth delay, cytopenia and subclinical immunodeficiency **(iii)** Mid and early truncating mutations were associated with the most severe immune phenotype as well as severe growth delay. *Abbreviations: ORF = open reading frame, UTR = untranslated region.*

4.2.7 Resequencing of *XRCC4* in MPD patients

In Chapter 3, a non-synonymous coding variant (consequence score=2) was identified in the homozygous state in a singleton case, c.127T>C (p.Trp43Arg) in *XRCC4* following WES (Figure 4.8A). The mutation was validated by capillary sequencing and both parents were found to be heterozygous carriers (Figure 4.8B). The predicted amino acid change occurred at a residue conserved in vertebrates (Figure 4.8C) and was predicted to have a functional impact on the protein by all prediction programmes used in analysis. The same mutation was subsequently reported in another case with MPD from the same country (Shaheen *et al.*, 2014).

XRCC4 was then re-sequenced in 199 patients with MPD, primary microcephaly or microcephaly and short stature using a custom designed Ion AmpliSeq panel to amplify desired coding regions and sequencing using either the Ion Proton™ or Ion Torrent™ platforms. This sequencing technology (semi-conductor method) has a markedly lower sensitivity for identifying indels compared to SNVs (see Section 1.4.1) therefore if a single deleterious heterozygous variant was identified, all coding regions were re-sequenced by capillary sequencing to ensure a second pathogenic variant was not present. This approach identified an additional five cases with recessive mutations in *XRCC4* (Table 4.6). In three families, where parental DNA was available, the mutations were confirmed to be biallelic in the affected offspring. Three patients of European origin all carried the same deletion on one allele, c.25delC, potentially representing a founder mutation in these families. Two of these patients also carried the same nonsense mutation on the alternate allele, c.823C>T.

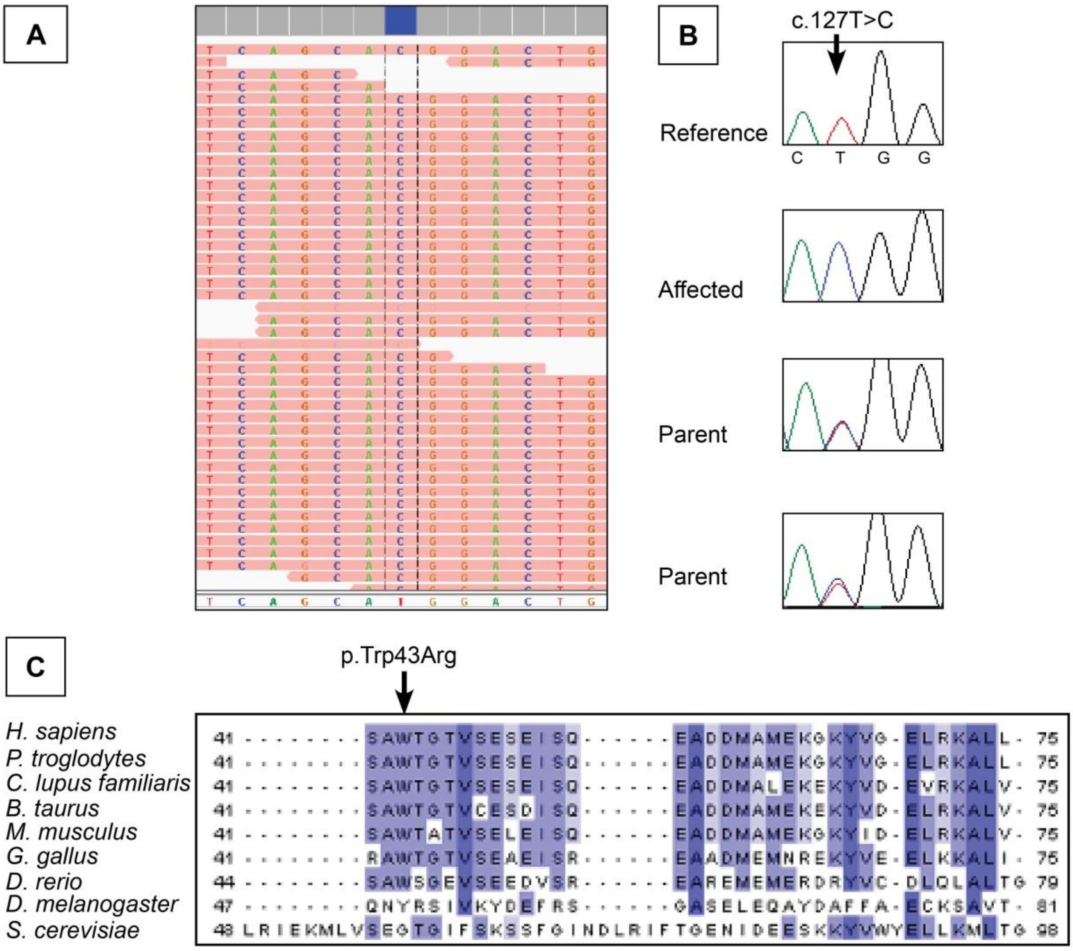

A
B
c.127T>C
Reference
C T G G
Affected
Parent
Parent
T C A G C A T G G A C T G
C
p.Trp43Arg
H. sapiens
P. troglodytes
C. lupus familiaris
B. taurus
M. musculus
G. gallus
D. rerio
D. melanogaster
S. cerevisiae
41 ---------SAWTGTVSESEISQ------EADDMAMEKGKYVG-ELRKALL- 75
41 ---------SAWTGTVSESEISQ------EADDMAMEKGKYVG-ELRKALL- 75
41 ---------SAWTGTVSESEISQ------EADDMALEKEKYVD-EVRKALV- 75
41 ---------SAWTGTVCESDISQ------EADDMAMEKEKYVD-ELRKALV- 75
41 ---------SAWTATVSELEISQ------EADDMAMEKGKYID-ELRKALV- 75
41 ---------RAWTGTVSEAEISR------EAADMEMNREKYVE-ELKKALI- 75
44 ---------SAWSGEVSEEDVSR------EAREMEMERDRYVC-DLQLALTG 79
47 ---------QNYRSIVKYDEFRS------GASELEQAYDAFFA-ECKSAVT- 81
48 LRIEKMLVSEGTGIFSKSSFGINDLRIFTGENIDEESKKYVWYELLKMLTG 98

Figure 4.8. Non-synonymous mutation identified in *XRCC4* following WES

A) Sequencing reads following WES aligned to *XRCC4* reference sequence using IGV identifies a homozygous variant, c.127T>C (blue) present in all reads (60 X) in the affected child. B) Capillary sequencing chromatograms confirming the mutation in the child and identified both parents as heterozygous carriers. C) Alignment of sequences from different species demonstrating the nucleotide change affects an amino acid which is conserved in vertebrates. Alignment performed using ClustalW2 (http://www.ebi.ac.uk/Tools/msa/clustalw2/, reference sequences used provided in Appendix II).

Table 4.6. Mutations in *XRCC4* were identified in six patients with MPD

Pt	*Nucleotide change*	*Protein change*	*Sex*	*Country of Origin*	*Mother*	*Father*
F11	c.[127T>C]	p.(Trp43Arg)	M	Saudi Arabia	p.(Trp43Arg)	p.(Trp43Arg)

F12	c.[481C>T]
	+[673C>T]
		p.(Arg161*)
		p.(Arg225*)
			M	Morocco	p.(Arg225*)	p.(Arg161*)

F13	c.[25delC]
	+[823C>T]
		p.(His9Thrfs*8)
		p.(Arg275*)
			M	Italy	N/A	

F14	c.[25delC]
	+[823C>T]
		p.(His9Thrfs*8)
		p.(Arg275*)
			M	France	N/A	

F15	c.[492dupA]	p.(Cys165Metfs*5)	F	Portugal	N/A	
F16	c.[25delC] +[-1-10G>T]	p.(His9Thrfs*8)	M	UK	p.(His9Thrfs*8)	c.[-1-10G>T]

See Appendix II for details of reference sequence.

4.2.8 Predicted Impact of *XRCC4* mutations on protein function

Nonsense mutations were predominantly identified in *XRCC4* and were distributed throughout the gene (Figure 4.9). As *XRCC4* is composed of eight exons and all mutations were upstream of the last exon (encodes amino acids 297-366), it is anticipated that transcripts would undergo nonsense mediated decay. As examination of other predicted transcripts did not identify an alternative transcript which would be unaffected by these nonsense mutations, it is likely that no functional protein would be produced by any of these alleles. One variant identified resulted in a single nucleotide change immediately adjacent to the first base of exon two, c.-10-1G>T. This was predicted to completely abolish the splice acceptor site of intron 1 (Figure 4.10), although the precise effect on splicing is yet to be determined by RNA studies.

XRCC4 is composed of an N-terminal head domain which forms a globular structure followed by a coiled-coil domain forming an α-helical stalk and terminates in an unstructured C terminal region (Junop *et al.*, 2000). The linker region between the two BRCT domains of LIG4 has been shown to bind directly to XRCC4 through an area in the coiled-coil stalk region (Sibanda *et al.*, 2001). It has been proposed that XRCC4 plays a structural role in NHEJ and may act with XLF to stabilise the DNA break prior to ligation (Roy *et al.*, 2012). Two XRCC4 proteins interact through the head domains to form homodimers in a similar fashion to XLF (Hammel *et al.*, 2011, Ropars *et al.*, 2011, Wu *et al.*, 2011, Andres *et al.*, 2012). XRCC4 and XLF homodimers then alternate through interactions of the head domains to form long protein filaments which may bridge the DNA gap promoting ligation. The non-synonymous coding variant identified in F11, c.127T>c (p.Trp43Arg), lies within the head domain of the encoded protein and may exert a deleterious effect on protein function by disrupting XRCC4-XLF interactions.

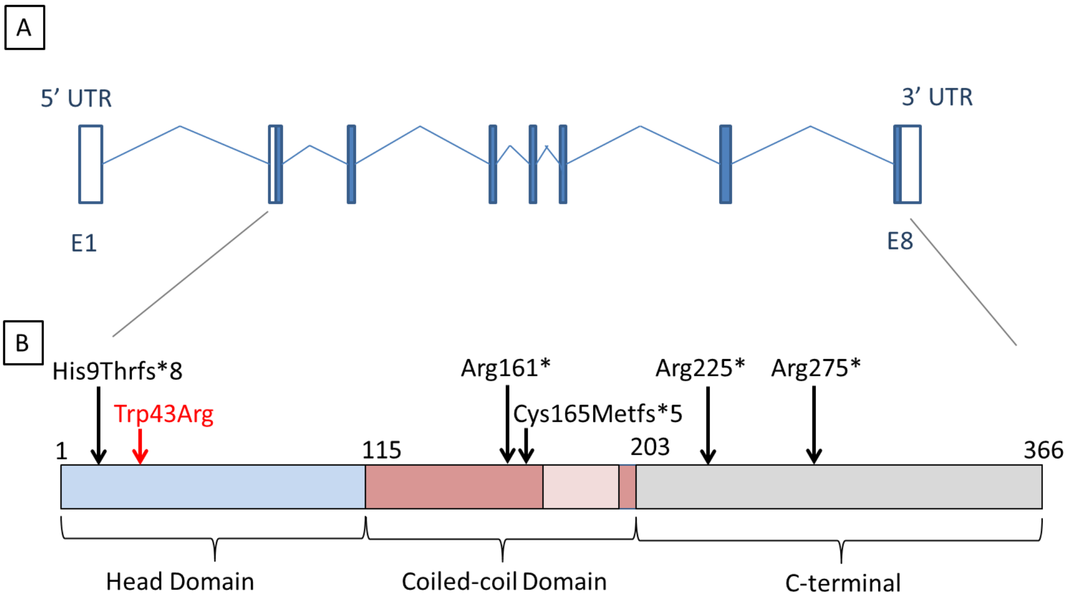


Figure 4.9. Mutations in *XRCC4* gene and location in the encoded protein

A) *XRCC4* is composed of eight exons. B) The 366aa protein is encoded by exons 2-8. The non-synonymous mutation, p.Trp43Arg (red), lies in the head domain of the protein which binds to XLF. LIG4 binds to XRCC4 via a region in the coiled-coil domain (light pink). The remaining coding mutations were predicted to produce truncated transcripts which would undergo nonsense mediated decay and thus would not produce any functional protein.

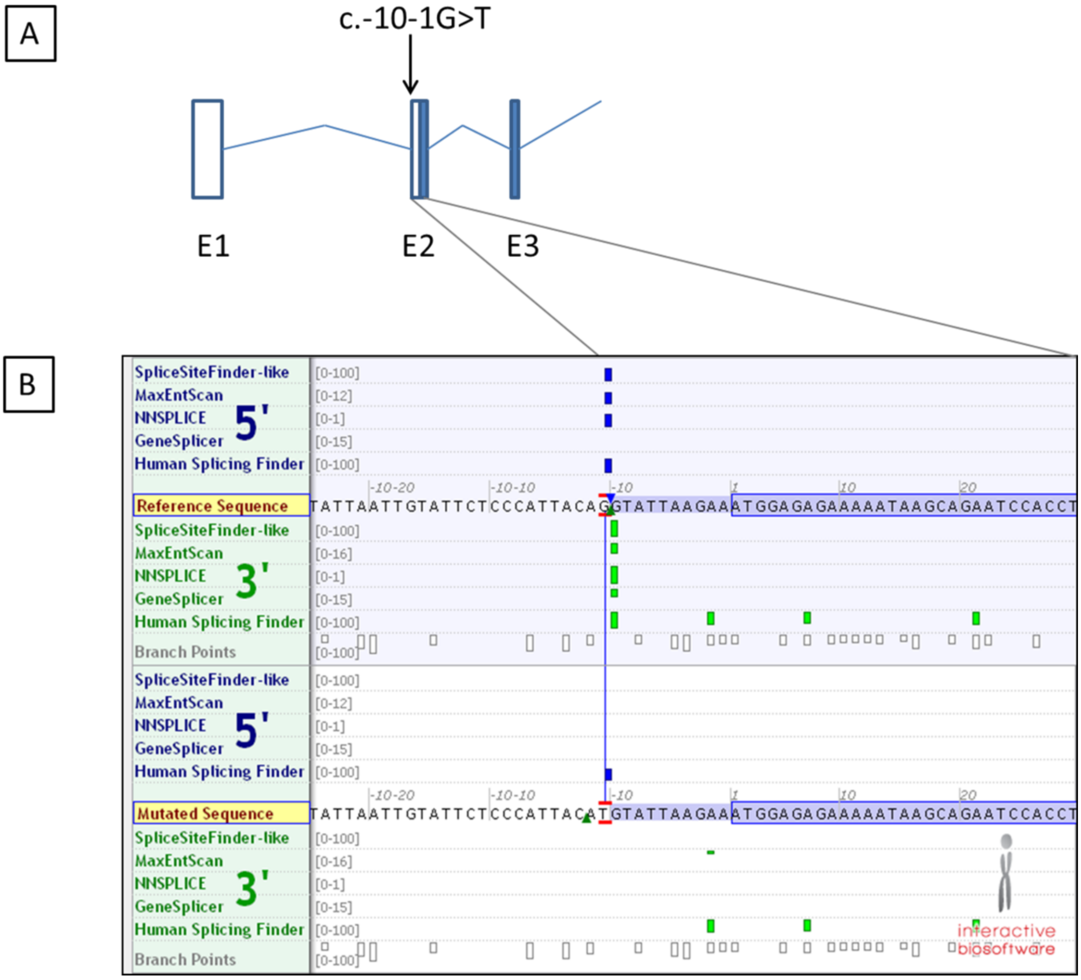


Figure 4.10. Predicted effect of a splice site variant identified in *XRCC4*

A) Schematic of exons 1-3 of *XRCC4* with position of splice variant indicated within the 5' UTR of the gene (white box). B) The 5'UTR of *XRCC4* spans the first two exons (white box). The c.-10-1G>A variant occurs immediately adjacent to the start of exon 2 and is predicted to completely abolish the splice acceptor site of intron 1 in all five prediction programmes used by Alamut software.

4.2.8 Characterising the phenotype associated with *XRCC4* mutations

4.2.8.1 Growth Impairment

Growth parameters were available in five of the six *XRCC4* patients (Table 4.7). IUGR was observed in four patients (mean birth weight -2.9 s.d., birth OFC -3.2 s.d. and length -3.9 s.d.). On comparison with *LIG4* cases there was no significant difference between growth parameters at birth (p>0.6) (Figure 4.11). Postnatal microcephaly was seen in all five cases with *XRCC4* mutations and this was disproportionate to the reduction in height (OFC -8.12 s.d., height -4.8 s.d.) similar to the *LIG4* patients. This is in contrast to patients with *PCNT* mutations (MOPD II) in which OFC is more proportionately reduced in comparison to height (mean postnatal OFC -10.3 s.d. and height -8.5 s.d.) (Bober *et al.*, 2012). Postnatal head circumference was slightly less reduced compared to *LIG4* patients (p=0.04) but no significant difference in stature (p=0.3) or weight (mean weight -5.0 s.d., p=0.2) was seen between the two groups (Figure 4.11).

4.2.8.2 Facial characteristics, malformations and intellectual development

Patients show various degrees of developmental delay from none at all to severe delay (Table 4.7). No malformations are as yet reported. One patient is described as having feeding difficulties and has a gastrostomy in situ. The cause of this is unclear although it is noted that there is also severe developmental delay in this case. Facial photographs were available in patient F16 only (Figure 4.12). Similar features to the *LIG4* patients were evident with fine sparse hair, epicanthic folds, small chin and broad nasal tip.

Table 4.7. Anthropometric data and clinical features of patients with *XRCC4* mutations

Pt	*Gest^n /weeks*	*Birth*			*Postnatal*			*Develop-mental delay*	*Immuno-globulins g/dl*	*Lymphocyte subset Cells/µl*				*Additional clinical features*
		Weight /sd (kg)	*OFC /sd (cm)*	*Length /sd (cm)*	*Age at Exam*	*OFC /sd (cm)*	*Height /sd (cm)*			*CD3*	*CD8*	*CD4*	*CD19*	
F11	40	-3.87 (1.8)	-4.87 (29)	-2.52 (46)	3y1m	-8.3 (39.8)	-4.7 (78.7)	Moderate	*N/A*	**110** (160-270)	**47** (60-100)	**47** (100-170)	**16** (40-80)	Nil reported
F12	32	-3.27 (0.903)	-4.57 (24)	*N/A* (34)	2y9m	-10.7 (36.4)	-5.7 (73.4)	Severe	*N/A*	*N/A*				Poor feeding Gastrostomy
F13	36	-3 (1.55)	-2.9 (29)	-4.49 (39)	8y4m	↓	-3.9 (108.2)	None	*N/A*	*N/A*				Nil reported
F14	*N/A*	*N/A*			*N/A*			*N/A*	*N/A*	*N/A*				*N/A*
F15	37	-0.6 (2.55)	-0.63 (32)	-2.08 (44)	10y5m	-4.6 (48)	-2.7 (123)	Mild	Total levels normal	**127** (147-230)	**392** (553-871)	492 (340-440)	**22** (27-43)	Nil reported
F16	36	-3.83 (1.25)	-2.9 (29)	-6.56 (35)	5m	-8.9 (33)	-7.2 (49.7)	Unknown	*N/A*	*N/A*				Eczema

Anthropometric data stated as Z scores (standard deviation from population mean for age and sex), actual measurements in brackets. F13 reported as having microcephaly although measurements were not available. Local reference ranges for lymphocyte subsets are shown in brackets. Low values are highlighted in bold.

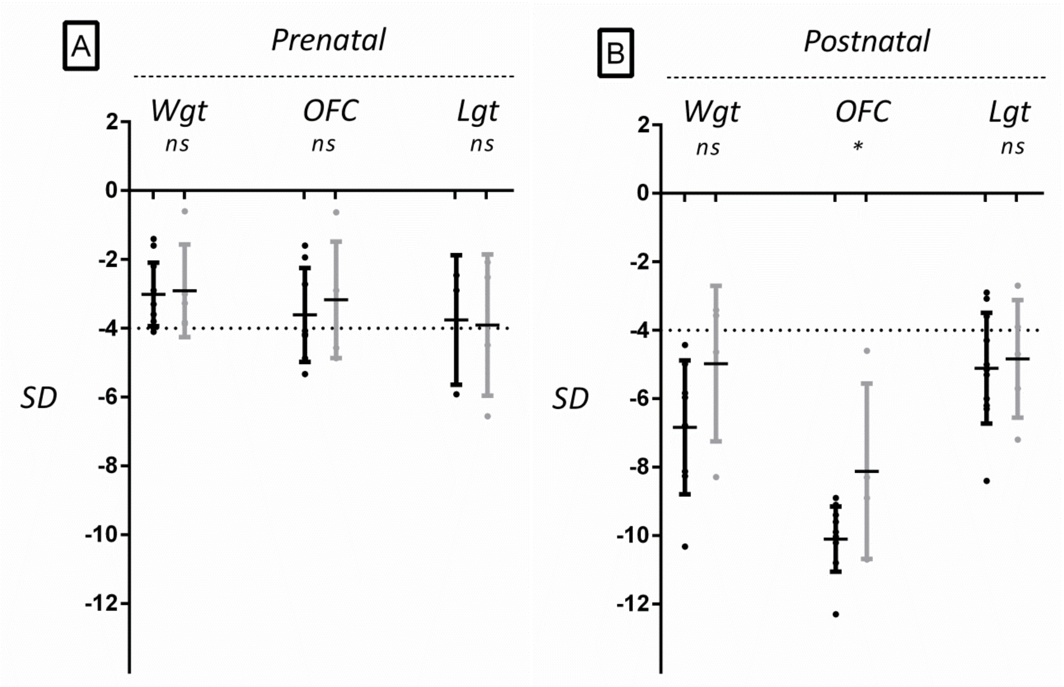


Figure 4.11. Growth impairment in cases with *XRCC4* mutations compared to *LIG4* patients with MPD

Measurements plotted as Z-scores (standard deviation of measurement from population mean for age and sex) for *LIG4* cases identified in this study (black) alongside those with mutations in *XRCC4* (grey) for each growth parameter. A) Measurements at birth of *XRCC4* patients (mean weight -2.9 +/-1.35 s.d. (n=5), OFC -3.2 +/-1.69 s.d. (n=5) and length -3.9 +/-2.05 s.d. (n=4)) compared to *LIG4* patients. B) Most recent postnatal measurements of *XRCC4* patients (mean weight -5.0 +/-2.27 s.d. (n=4), OFC -8.12 +/-2.56 s.d. (n=4) and height -4.8 +/-1.72 s.d. (n=5)) compared to *LIG4* patients. Mean and standard deviation of patient group indicated by horizontal and vertical bars respectively. * *p<0.05.*

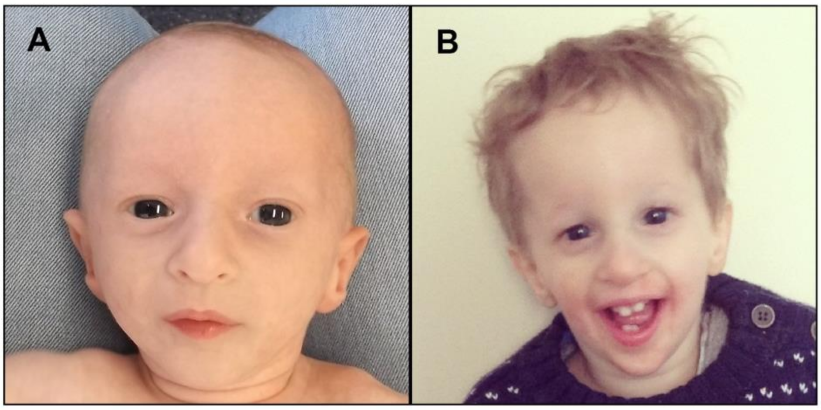


Figure 4.12. Clinical photographs of patient with mutations in *XRCC4*

Patient F16 at A) five months and B) two years of age shows similar features to the *LIG4* patients shown in Figure 4.5.

4.2.8.3 Haematological and immune dysfunction

No cytopenic episodes or clinical evidence of immunodeficiency had been reported in any of the *XRCC4* cases. Following identification of mutations, immune and haematological investigations were recommended to all clinicians. Results from two patients are currently available (Table 4.7). In F11 there was leucopenia (WBC=3.03x10^3/µl, approximate normal range 5.5-12.9) with a reduction in both T and B cell numbers. Total immunoglobulin levels have not yet been performed. In F15 total immunoglobulin levels were normal, total lymphocyte count was slightly reduced (1658/mm3, normal range 2100-3076) with some mild abnormalities in T and B cell counts.

4.2.8.4 Malignancy

Malignancy was not reported in any of the probands however one mother (confirmed heterozygous carrier of a nonsense mutation) is suffering from metastatic breast cancer at the age of 34 years. There was no other family history of malignancy known. All other parents were reported to be healthy except for the father of a separate *XRCC4* patient who had died many years earlier of a myocardial infarction.

4.3 Discussion

4.3.1 Mutations in *LIG4* and *XRCC4* are a common cause of MPD

Mutations were identified in two components of the NHEJ machinery, *LIG4* and *XRCC4*, in 16 families with MPD. This accounts for 4% of cases in our cohort and is the second most common cause of MPD behind *PCNT* mutations (approximately 20% of the cohort). Although severe microcephaly has been reported in patients with mutations in NHEJ genes, including *LIG4*, a global failure in growth to the degree observed in this group of patients has not been previously described (Table 4.1, Section 4.1.2). Importantly, the majority of patients identified with *LIG4* mutations in this study presented with severe growth failure whereas other features indicative of Ligase IV syndrome, such as immunodeficiency, pancytopenia or malignancy, were not readily apparent delaying the diagnosis in this patient group. During this study mutations were also identified in *XRCC4* in six independent cases. Since this discovery, mutations in *XRCC4* have only been reported in one additional patient who also presented with severe growth failure (OFC -8.3, height -7.1 at 4 years) (Shaheen *et al.*, 2014). Identification of this number of cases allows more detailed characterisation of this novel disorder.

4.3.2 Clinical implications: diagnosis and management

This is the first time a comprehensive and consistent clinical phenotype has been described in a group of *LIG4* patients and in contrast to many previous reports immunodeficiency was not suspected clinically prior to molecular diagnosis (Buck *et al.*, 2006b, Enders *et al.*, 2006, van der Burg *et al.*, 2006, Grunebaum *et al.*, 2008). However, further investigation has revealed significant humoral and cellular immunodeficiency is present in the majority of patients. This highlights the need for early recognition and subsequent immunological investigation even in asymptomatic patients. Bone marrow failure, primarily presenting with episodic thrombocytopenia, appears to be a strongly discriminative diagnostic feature of Ligase IV syndrome, occurring in 72% of all reported cases to date. However, cytopenia may not be evident at initial presentation, particularly in very young children. Increasing awareness in clinicians that Ligase IV syndrome commonly presents with severe growth failure without any other suggestive features will hopefully lead to improved diagnosis.

Early identification of *LIG4* patients is important as there are significant implications for clinical management. In view of increased cellular radiosensitivity in patient cells (O'Driscoll *et al.*, 2001, Murray *et al.*, 2014), it would be important to minimise exposure to ionising radiation where possible such as through radiographic imaging. Even though malignancy was not reported in any of the patients identified in this study, this has been a feature of previous reports (Plowman *et al.*, 1990, Ben-Omran *et al.*, 2005, Buck *et al.*, 2006b, Enders *et al.*, 2006, Toita *et al.*, 2007, Yue *et al.*, 2013) with a 25% incidence of malignancy in all cases reported in the literature to date. Therefore all *LIG4* patients should be considered at increased risk. Furthermore, bone marrow transplantation is being considered in several patients in this cohort. This has previously been associated with poor outcomes in *LIG4* patients and it is now apparent that modifications to conditioning regimes are required to improve survival due to increased sensitivity to certain immunosuppressants (O'Driscoll *et al.*, 2008).

Given the functional overlap between LIG4 and XRCC4 it is unsurprising that phenotypic overlap between the two patient groups was observed. Most notably, a comparable degree of growth failure was present in both *XRCC4* and *LIG4* patients as well as similarities in facial features. However, as yet, no evidence of bone marrow failure has been reported in any of the seven *XRCC4* patients identified to date with the oldest being 10 years of age (Shaheen *et al.*, 2014) (range of onset in *LIG4* patients was 2-15 years). Additionally, in the two patients tested in this study, relatively mild abnormalities were seen in T and B cell levels compared to *LIG4* patients in which B cells were severely depleted in all patients investigated. This is surprising given that significant immune dysfunction is common to all disorders of NHEJ repair (Moshous *et al.*, 2001, Buck *et al.*, 2006a, Buck *et al.*, 2006b, van der Burg *et al.*, 2009) and V(D)J recombination is impaired in Xrcc4 deficient mice (Guirouilh-Barbat *et al.*, 2007). Despite minimal evidence of either immune or haematological compromise in *XRCC4* patients identified to date it would therefore still be prudent to monitor these parameters until further data from additional patients is available. XRCC4 patient cells also exhibit increased cellular sensitivity to ionising radiation (Grant Stewart, personal communication, July 2014) which reflects defective DSB repair (Sun *et al.*, 2002). Additionally, radiosensitivity is observed in XRCC4 null mouse embryonic stem cells (Gao *et al.*, 1998b) along with marked genomic instability in XRCC4 deficient mouse embryonic fibroblasts (Gao *et al.*, 2000). Thus, similar to LIG4 patients, *XRCC4* patients should also be considered at increased risk of malignancy although longer follow up is required to establish the degree of risk. Of note, in the 16 families identified in this study malignancy was reported in one parent, a heterozygous carrier of a nonsense mutation in *XRCC4*. Modification of cancer risk in heterozygous carriers of *LIG4* SNPs has been suggested (Roddam *et al.*, 2002) however, longer follow up and the identification of more cases are required to ascertain whether this case is related to the *XRCC4* mutation or coincidental given the high rate of breast cancer in the general population (Ly *et al.*, 2013).

4.3.3 Genotype-phenotype correlations suggest LIG4-XRCC4 binding is critical to growth

As such a consistent phenotype was observed in the majority of *LIG4* patients in this cohort, it was hypothesised that this may be related to the type of mutation identified. All the patients identified in this cohort have biallelic truncating mutations in *LIG4* in combination with MPD whereas previously reported cases have predominantly non-synonymous coding mutations and generally less severe growth impairment (Table 4.1). Indeed, a genotype-phenotype correlation was observed in which the degree of protein truncation was related to the severity of the growth phenotype, in particular, the extent to which the XRCC4 binding domain was disrupted.

Furthermore, all *LIG4* patients identified to date (in which growth parameters are available) who harbour at least one truncating mutation have significant growth impairment ranging from severe microcephaly (OFC < -5 s.d.) with normal stature to MPD (O'Driscoll *et al.*, 2001, Ben-Omran *et al.*, 2005, Buck *et al.*, 2006b, Gruhn *et al.*, 2007, Toita *et al.*, 2007, Yue *et al.*, 2013, Murray *et al.*, 2014). Interestingly, patients harbouring biallelic non-synonymous coding or in-frame deletions in the enzymatic domain were the least growth impaired but exhibited severe immune or haematological dysfunction (Plowman *et al.*, 1990, Riballo *et al.*, 1999, O'Driscoll *et al.*, 2001, van der Burg *et al.*, 2006, Unal *et al.*, 2009). In these cases the XRCC4 binding domain was preserved in both alleles. In addition, MPD is evident in all patients identified with biallelic *XRCC4* mutations. This suggests the LIG4-XRCC4 interaction is the critical determinant in the severity of the growth phenotype.

XRCC4 binding is required for LIG4 stability (Bryans *et al.*, 1999), localisation to the nucleus (Girard *et al.*, 2004) and chromatin binding (Liu *et al.*, 2013a). Additionally, the LIG4-XRCC4 complex has been shown to prevent the degradation of terminal nucleotides, a function which appears independent of ligation (Smith *et al.*, 2003). The recurrently occurring c.2440C>T (p. Arg814*) mutation results in impaired LIG4 cellular activity, with reduced LIG4 protein levels (O'Driscoll *et al.*, 2001) and also impairs ligase activity (Girard *et al.*, 2004). Notably, the c.2440C>T mutation impairs enzymatic function less than earlier truncating mutations (Girard *et al.*, 2004). Arg814 lies at the start of the second BRCT domain of the XRCC4 binding site (Critchlow *et al.*, 1997). Loss of this BRCT domain impairs XRCC4 binding (Girard *et al.*, 2004), though the major determinant of XRCC4 binding, the ‘intervening linker sequence’, is retained (Grawunder *et al.*, 1998). Earlier (close to the N-terminus) protein truncation leads to a more complete loss of function, with truncated protein resulting from the c.2094C>G (p.Arg580*) mutation not interacting with XRCC4 or localising to the nucleus (Girard *et al.*, 2004).

Surprisingly, nine of the ten families identified with *LIG4* mutations harboured the g.2440C>T (p.Arg814*) mutation which is found only in European populations with an allele frequency of 0.0005 and 0.001, in EVS and 1KG databases respectively. Therefore the frequency of g.2440C>T homozygotes would be expected to be 1 in 1,000,000 to 1 in 4,000,000. Such patients may only be mildly affected with non-syndromic microcephaly and normal development (Ben-Omran *et al.*, 2005). This may therefore represent an under-diagnosed group of *LIG4* patients with increased malignancy risk. For those cases that reach the genetics clinic, the routine adoption of high-throughput sequencing in molecular diagnostics is likely to improve ascertainment; however, appropriate setting of allele frequency filters will be crucial.

4.3.3 Role of NHEJ in growth and neurogenesis

The identification of *LIG4* and *XRCC4* mutations as a relatively common cause of profound impairment in growth highlights the critical importance of NHEJ in early development. Despite mutations in NHEJ genes not previously being associated with such severe global growth failure in humans, animal models have shown otherwise. Ku80 and Ku70 mutant mice are viable and exhibit a 40-60% reduction in size (Nussenzweig *et al.*, 1996, Gu *et al.*, 1997). Null *Lig4* mice are early embryonic lethal (Barnes *et al.*, 1998, Frank *et al.*, 1998) but two viable hypomorphic mouse models show an approximately 40% reduction in size (Nijnik *et al.*, 2007, Rucci *et al.*, 2010). Additionally, *Xrcc4* null mice show late embryonic lethality with an approximately 30% reduction in size at E15.5 (Gao *et al.*, 1998b). Furthermore, mouse embryonic fibroblasts from null *Xrcc4* and *Lig4* embryos exhibited slower growth, failure to repair DSBs and growth arrest following exposure to ionising radiation (Gao *et al.*, 1998b). Interestingly, DNA-PKcs and Artemis deficient mice display RS-SCID without any apparent growth defects similar to that observed in humans (Gao *et al.*, 1998a, Rooney *et al.*, 2002) indicating LIG4 and XRCC4 are more critical to development compared to other NHEJ components. Artemis and DNA-PKcs are required for the opening of RAG-generated hairpins arising in V(D)J recombination (Ma *et al.*, 2002) and, in contrast to LIG4 and XRCC4, are therefore perhaps less critical in the repair of DSBs occurring from other sources which do not have the same processing requirements (Adachi *et al.*, 2004).

Notably, five of the six patients with *XRCC4* mutations identified in this study harboured biallelic nonsense mutations and therefore are anticipated to be null in contrast to *LIG4* truncating mutations which are predicted to be hypomorphic. It is unclear why *XRCC4* null mutations are embryonic lethal in mice but apparently viable in humans. Possible explanations include the presence of a human-specific alternate transcript which partly rescues function similar to that observed in MPD patients with mutations in *PLK4* (Martin *et al.*, 2014). However, this seems unlikely given that nonsense mutations were distributed throughout the gene. Alternatively, a greater degree of redundancy in the multiple components which can be utilised in NHEJ may exist in humans compared to mice. For example, LIG1 and LIG3 are both able to perform NHEJ in the absence of LIG4 in DT40 chicken cells, albeit, less efficiently (Paul *et al.,* 2013). Compensatory mechanisms unique to humans might therefore occur in the presence of XRCC4 deficiency.

One interesting observation in all *LIG4* and *XRCC4* patients described to date is the presence of severe and disproportionate microcephaly in comparison to height. A similar pattern is also observed in patients with mutations in genes encoding the MRN complex indicating that neuronal cells are especially sensitive to defective DSB repair (Weemaes, 2000, Waltes *et al.*, 2009, Matsumoto *et al.*, 2011).

Neuronal stem cells lie in the ventricular and sub-ventricular zones (VZ/SVZ) of the developing brain and undergo rapid symmetrical cell division in early development. Slightly later in embryogenesis there is a switch to asymmetric cell division generating post-mitotic neurons populating the intermediate zone (IZ) and distal cortical plate (CP) (Pontious *et al.*, 2008). High levels of endogenous DNA breakage are seen in the VZ/SVZ layers during embryogenesis as well as a marked increase in sensitivity to DNA damage with activation of apoptosis at much lower doses of IR compared to the adult brain (Gatz *et al.*, 2011). Additionally, neuronal stem cells are not fully sensitive to G2/M checkpoint arrest (Gatz *et al.*, 2011), perhaps to ensure rapid cell division occurs unperturbed by DNA damage in such cells. Any unrepaired DSBs will therefore persist in daughter cells including post-mitotic cells within the IZ following asymmetric division. Proliferating cells are able to utilise homologous recombination in DSB repair due to the availability of a sister chromatid whereas non-proliferating post-mitotic neurons are reliant on NHEJ for repair (Orii *et al.*, 2006). In both *Xrcc4*- and *Lig4*-null murine embryos, lethality results from severe p53-dependent neuronal cell death occurring predominantly in the IZ (Gao *et al.*, 1998b, Orii *et al.*, 2006). Gatz *et al.*, (2011) therefore postulated that p53-dependent apoptosis is triggered with increased frequency by the presence of persisting DSBs within IZ cells. Failure to repair such damage in LIG4/XRCC4 deficiency would therefore lead to substantial neuronal cell death and consequently fewer cells and a smaller brain.

Stem cell depletion along with reduced capacity for renewal has also been shown in the haematopoietic system in *Lig4* hypomorphic mice (Nijnik *et al.*, 2007). The depletion of other stem cell pools through defective DSB repair may therefore explain the global failure in growth in *LIG4* and *XRCC4* patients (Woodbine *et al.*, 2014). Bone marrow failure is not a feature of other DSB repair disorders such as Bloom syndrome (Kaneko *et al.*, 2004) or Nijmegen breakage syndrome (Weemaes, 2000) suggesting LIG4 specifically, whether through its role in NHEJ or while undertaking a different function, is critical to the development and maintenance of haematopoietic stem cell pools. Additionally, the increased IZ apoptosis observed in *Lig4*- and *Xrcc4*-deficient mice was found to be ATM-dependent (Sekiguchi *et al.*, 2001) however, patients with ATM mutations only exhibit a subtle postnatal microcephaly (Nissenkorn *et al.*, 2011). Finally, this does not account for why some *LIG4* patients with different mutations have no reported growth defect despite the fact that LIG4 enzymatic capacity and thus NHEJ is greatly reduced. These observations suggest the possibility that the role of LIG4-XRCC4 may extend beyond NHEJ in stem cells.

4.3.4 Conclusions

This Chapter describes the identification of mutations in *LIG4* and *XRCC4* as a common cause of MPD along with further characterisation of associated features which include immunodeficiency, bone marrow failure and malignancy predisposition. This will hopefully improve diagnosis in this patient group as well as impact on the clinical management of those affected. Identification of these mutations, along with the observation of a genotype-phenotype correlation, indicates that the LIG4-XRCC4 interaction specifically is critical in ensuring normal growth. Increased p53-mediated cell death in stem cell pools during embryogenesis may partly explain the hypocellularity in LIG4/XRCC4 deficient patients but does not account for phenotypic variability observed in LIG4 and ATM deficient patients suggesting that the LIG4-XRCC4 complex may have an additional, as yet unidentified, role in development. In the next Chapter the pathogenicity of mutations in two further novel candidate disease genes identified following WES (*NCAPD2* and *NCAPD3*) are investigated.

CHAPTER 5: Mutations in the Non-SMC subunits of Condensin I & II complexes in MPD

5.1 Introduction

In mitosis, replicated genetic material within each cell is required to be equally transmitted to two daughter cells. To ensure genomic integrity is maintained during segregation, DNA is compacted into chromosomes in a process referred to as chromosome condensation (Koshland *et al.*, 1996). Critical to the mechanism and maintenance of DNA compaction is the highly conserved protein complex, condensin (Hagstrom *et al.*, 2002). In vertebrates there are two types of condensin complexes, I and II, each composed of five subunits (Ono *et al.*, 2003). In Chapter 3 the identification of six candidate novel disease genes in MPD patients through WES was described. Two of which, *NCAPD2* and *NCAPD3,* encode subunits of the condensin I and condensin II complexes respectively.

Given their important role in mitosis, it is conceivable that abnormal condensin function impacts on the rate of cell division and consequently body mass. Genes encoding subunits of the condensin complexes are therefore plausible candidates that could affect growth. In this Chapter, further studies are performed to investigate whether the variants identified in these genes in MPD patients are deleterious to protein function as well as ascertain whether the disruption of these proteins is responsible for the observed growth phenotype.

5.1.1 Structure of condensin complexes

Both the condensin I and II complexes, share the same overall structural arrangement with two SMC subunits, SMC2 and SMC4 (members of the structural maintenance of chromosomes ATPase family), but differ in their three associated non-SMC subunits (Figure 5.1) (Ono *et al.*, 2003). The SMC subunits are composed of a distal central hinge domain flanked by two long coiled coils which fold back on each other and interact forming an antiparallel coil (Saitoh *et al.*, 1994). Consequently, the two outer ends of the coils are proximally located and form ATPase head domains. The SMC2 and SMC4 subunits together create a V-shaped heterodimer through interactions at their hinge domains (Hirano *et al.*, 2001). Bridging the gap between the two ATPase domains at the ends of the SMC subunits is a kleisin protein, encoded by *NCAPH* in the condensin I complex and *NCAPH2* in condensin II (Schleiffer *et al.*, 2003). The two remaining proteins, composed of multiple HEAT repeats, complete the complex through interactions with the kleisin subunit and, more weakly, each other (Onn *et al.*, 2007). These two proteins are encoded by *NCAPD2* and *NCAPG* in condensin I and *NCAPD3* and *NCAPG2* in condensin II (Figure 5.1).

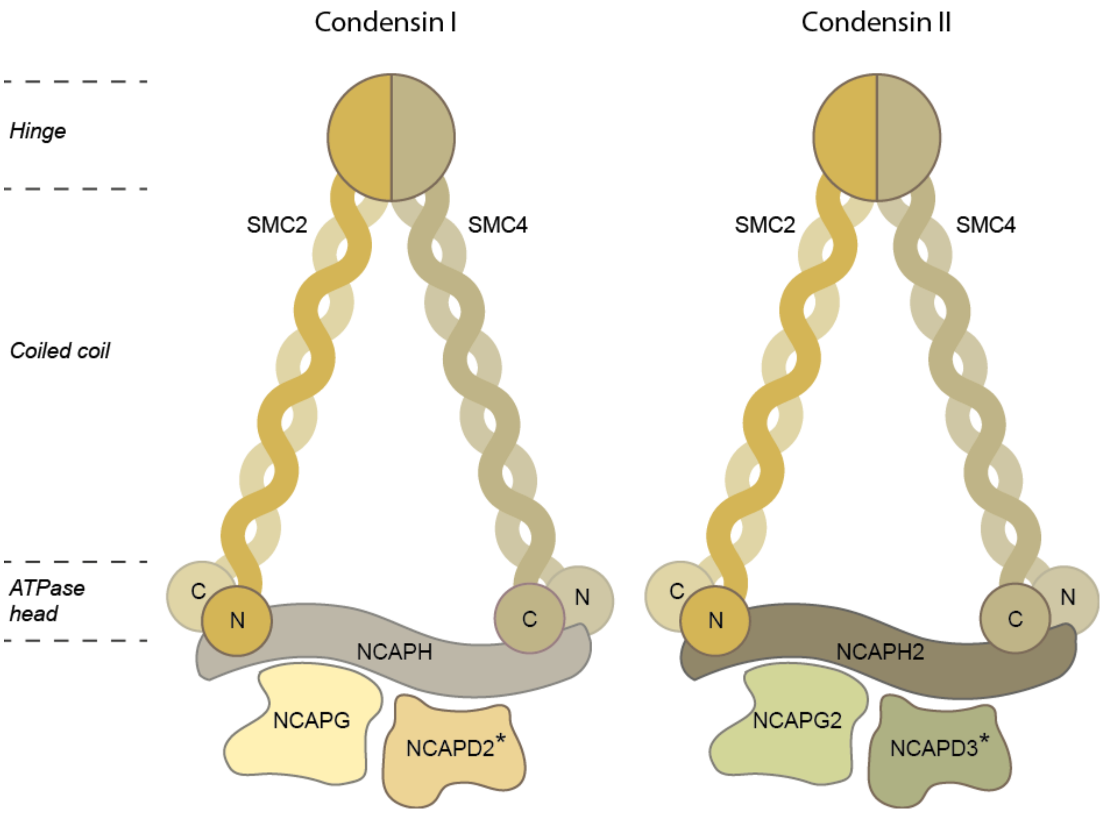


Figure 5.1. Structure of the two condensin complexes

Both complexes are composed of the same two SMC subunits which dimerize at the hinge domain (SMC2 and SMC4). The two coiled coil domains extending from each hinge domain form antiparallel coils ending in an ATPase head domain. The two complexes then differ in the three non-SMC subunits: A kleisin protein (NCAPH in condensin I and NCAPH2 in condensin II) which bridges the ATPase head domains of the two SMC subunits and two HEAT repeat proteins which associate with the complex predominantly through interactions with the kleisin subunit (NCAPG, NCAPD2 in condensin I and NCAPG2, NCAPD3 in condensin II) (figure adapted from (Piazza *et al.*, 2013). *Indicates subunits in which mutations were identified in MPD patients through WES.

5.1.2 Function of condensin complexes

The two complexes associate with chromosomes independently of each other and at distinct chromosomal regions with different dynamics suggesting they each have discreet and independent functions (Ono *et al.*, 2003).

Condensin II localises to the nucleus during interphase and is bound to chromosomes from early prophase until the end of mitosis (Yeong *et al.*, 2003). There is evidence supporting a role of condensin II in the early axial shortening of chromosomes in prophase. Depletion of condensin II in several organisms delays the onset of chromosome condensation until just prior to nuclear envelope breakdown and results in the appearance of elongated, 'curly' chromosomes (Ono *et al.*, 2003, Green *et al.*, 2012).

In contrast, condensin I is excluded from the nucleus only gaining access to chromosomes after breakdown of the nuclear envelope in prometaphase and then remains bound to chromosomes until late anaphase (Ono *et al.*, 2004). Condensin I appears to then be required for the lateral compaction of chromosomes (Ono *et al.*, 2003, Green *et al.*, 2012) as well as ensuring chromosome stability (Gerlich *et al.*, 2006) and aiding the separation of chromosome arms through the removal of cohesin (Hirota *et al.*, 2004). Depletion of condensin I results in a delay in progression through prometaphase and metaphase as well as the appearance of wider, shorter chromosomes (Ono *et al.*, 2003, Hirota *et al.*, 2004, Green *et al.*, 2012).

Despite extensive work on the condensin complexes, still relatively little is known regarding the mechanisms controlling condensin function. Now an active area of investigation, several factors have been implicated in the recruitment of condensins to chromosomes as well as the regulation of their activity. This also differs between the two condensin complexes in keeping with their different functions during mitosis. For example, A-kinase anchor protein 8 (AKAP8) has been implicated in condensin I binding to chromosomes but not condensin II (Eide *et al.*, 2002) whereas retinoblastoma 1 (RB1) assists in the chromatin association of condensin II but not condensin I (Longworth *et al.*, 2008). The regulation of condensins also varies between the two complexes and appears to involve a dynamic and intricate system of phosphorylation throughout mitosis. Regulatory proteins so far identified include cyclin-dependent kinase 1 (CDK1) (Kimura *et al.*, 1998), polo-like kinase 1 (PLK1) (Abe *et al.*, 2011), aurora B kinase (AURKB) (Lipp *et al.*, 2007, Takemoto *et al.*, 2007), casein kinase 2 (CK2) (Takemoto *et al.*, 2006) and protein phosphatase 2 (PP2A) (Takemoto *et al.*, 2009). Intriguingly, as well as a critical role in chromosome condensation and segregation, further evidence now also suggests a role for condensins in the DNA damage response (Blank *et al.*, 2006, Heale *et al.*, 2006, Tanaka *et al.*, 2012a) suggesting condensins may have a wider range of functions than previously thought.

5.2 Mutations identified in the non-SMC subunits of condensin complexes

5.2.1 *NCAPD2*

In Chapter 3, a splice site variant was identified in *NCAPD2*, c.4120+2T>C (consequence score=1) in a single case following WES (Figure 5.2A) which was subsequently confirmed to be homozygous by capillary sequencing (Figure 5.2B). DNA was also available from three additional unaffected family members; the mother, father and sibling. All were found to be heterozygous carriers confirming that the variant segregated correctly with the phenotype in these family members. The variant is located two nucleotides into intron 31 and is predicted to abolish the splice donor site of exon 31 in four out of five prediction programmes used by Alamut software (Figure 5.2C). It was also not present in any of the control datasets examined (EVS, 1KG or dbSNP). The predicted inclusion of intron 31 following disruption of this splice site would result in the introduction of a premature termination codon prior to the last exon (exon 32).

5.2.2 *NCAPD3*

Biallelic variants were also identified in *NCAPD3* following WES in a trio set as described in Chapter 3 (Figure 5.3A). Capillary sequencing confirmed the affected child to be heterozygous for the variants, c.382+14A>G (consequence score=2.5) and c.1783delG (p.Val595Serfs*34) (consequence score =1). Resequencing also confirmed the intronic variant was inherited from the mother and the nonsense mutation from the father (Figure 5.3B). The intronic variant was predicted to create an alternative splice site 13 nucleotides into intron 3 by four out of five prediction programmes used by the Alamut software (Figure 5.3C). The nonsense mutation is predicted to introduce a premature termination codon prior to the last exon and thus it would be anticipated that no functional protein would be produced by this allele (termed a ‘null’ allele) as the transcript would undergo nonsense mediated decay.

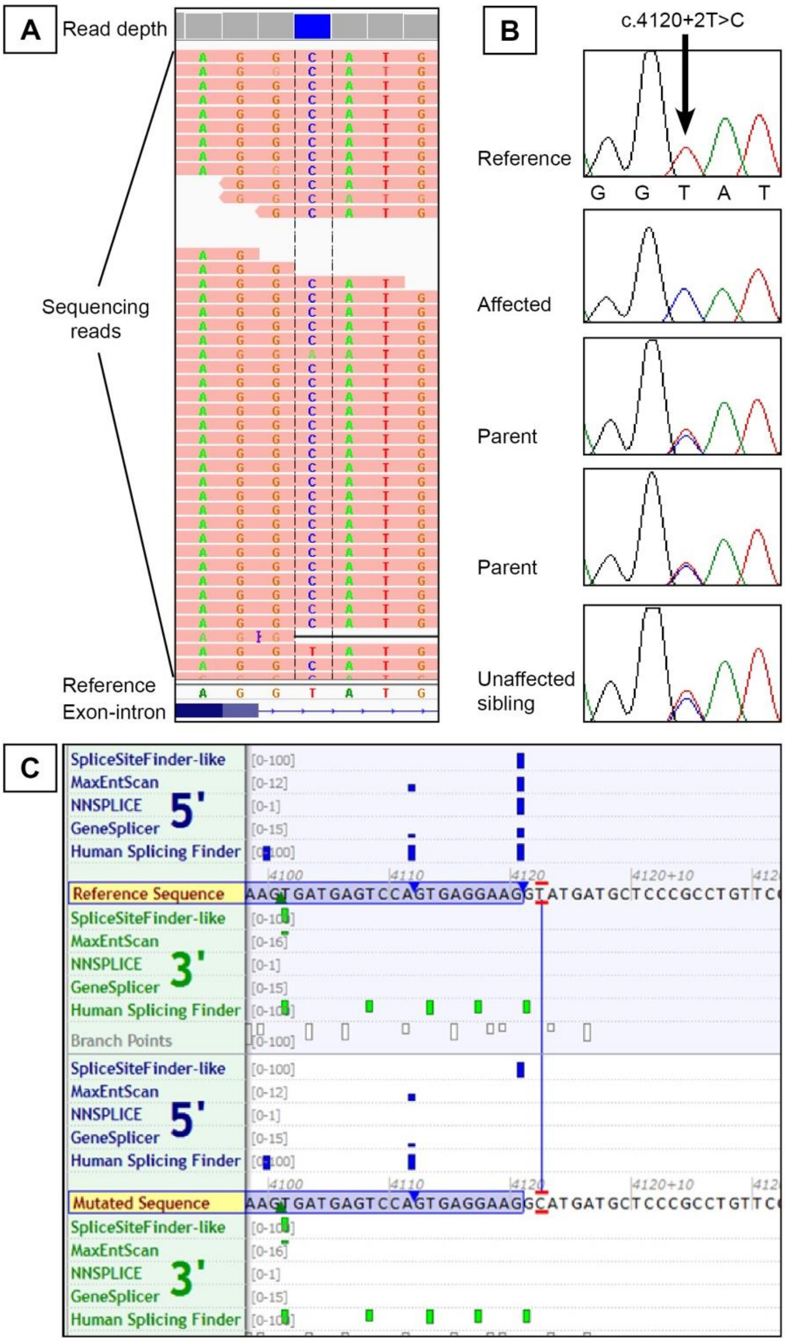


Figure 5.2. Splice site variant identified in *NCAPD2*.

A) Sequencing reads following WES aligned to *NCAPD2* reference sequence using the Integrative Genomics Viewer (IGV) identified the variant c.4120+2T>C (blue) in 50 out of 52 reads in the affected child. B) Chromatograms following capillary sequencing of all family members where DNA available confirmed the affected child to be homozygous for the variant and this segregated correctly with the observed phenotype in the parents. C) The variant was predicted to completely abolish the splice donor site of exon 31 as demonstrated by the graphic display produced by Alamut software.

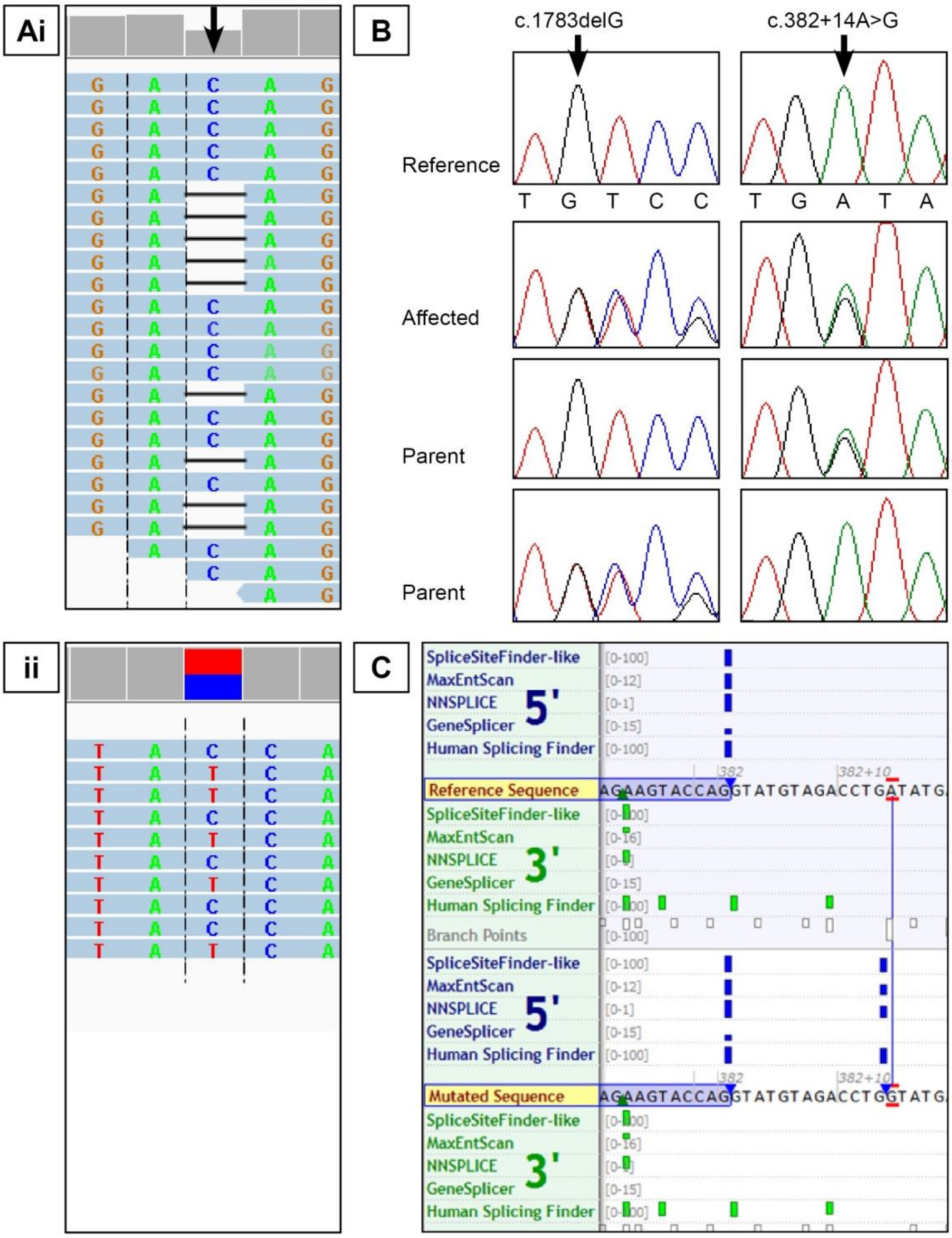


Figure 5.3. Compound heterozygous variants identified in *NCAPD3*.

A) Sequencing reads following WES aligned to the *reverse compliment* of the *NCAPD3* reference sequence using IGV identifies (i) a single nucleotide deletion, c.1783delG (arrow), in 14 out of 24 reads and (ii) an intronic variant, c.382+14A>G (red/blue) in 5 out of 10 reads. B) Chromatograms following capillary sequencing of the child and parents confirmed the affected child to be compound heterozygous for both variants. C) The intronic variant, c.382+14A>G, was predicted to create an alternative splice site in intron 3, 13 nucleotides upstream of the exon 3 splice donor site (graphic display using Alamut software).

5.2.3 *NCAPH*

To identify whether other MPD patients had mutations in genes affecting the same complex, all eight genes encoding the condensin subunits and two genes encoding the key condensin regulators *AURKB* and *PLK1* were sequenced in the remainder of the cohort (199 samples, all affected cases from unique families). This included 222 exons amplified using the custom designed Ampliseq primer set described in Section 2.4.6.2 followed by sequencing using the Ion Proton™ (192 samples) and Ion Torrent™ (6 samples) platforms. This included patients with primary microcephaly and microcephaly with short stature (height -2 s.d. to -4 s.d.). Identified variants in samples sequenced using the Ion Proton™ sequencer were annotated as previously described using SnpEff, dbNSFP and Alamut-HT (Section 2.5.2.1). Initially 463 variants were identified in total across the 10 genes. These were filtered in an SQL database using the same pipeline as described in Section 3.3 (Table 5.1). All remaining candidate variants (30 in total) were reviewed in IGV and, if likely true positives (see Section 3.3.7), resequenced along with any additional family members to check whether the variant segregates correctly with the phenotype. No potentially pathogenic recessive variants were identified in the 6 samples sequenced using the Ion Torrent™ sequencer which were analysed independently using NextGENe software (described in Section 2.5.3).

Any exons with poor coverage (read depth less than 10) were resequenced by capillary sequencing. In 39 cases (including four sequenced on the Ion Torrent™ sequencer), a single heterozygous, but potentially deleterious variant was identified in at least one gene. As the sequencing technology (semi-conductor method) used in this experiment has a markedly lower sensitivity for identifying indels compared to SNVs (see Section 1.4.1), repeated resequencing of the whole gene was performed by capillary sequencing to ensure a second, possibly pathogenic variant was not present. This, however, did not identify any further candidate variants. The possibility of causative *de novo* variants was also considered, however none were identified following capillary sequencing of the affected in combination with parental samples.

Table 5.1. Filtering of identified variants following targeted resequencing of condensin genes using Ion Proton™ sequencer.

Filtering Step	*Variant Number*
Initial number of unique variants prior to filtering	**463**
Select Rare Variants: 1KG, EVS (maf<0.005 or no frequency data)	**159**
Exclude variants occurring in more than 6 families	**143**
Variants with a consequence score<3	**129**
Variants following autosomal recessive analysis	**30**
Number of variants validated with correct segregation confirmed in parents	**1**

This analysis identified only one potentially deleterious variant which confirmed on resequencing and segregated correctly in the family. The non-synonymous coding variant c.728C>T (p.Pro243Leu) was identified in *NCAPH*, another subunit of the condensin I complex (Figure 5.4A), in a patient with primary microcephaly. The patient was confirmed to be homozygous by capillary sequencing and both parents were confirmed to be heterozygous carriers (Figure 5.4B). This variant was predicted to be deleterious by all prediction programmes used (Table 3.3) and alters an amino acid which is highly conserved across all species including *Schizosaccharomyces pombe* (Figure 5.5A).

Splicing programmes in Alamut software predicted that the variant would have minimal effect on neighbouring splice sites (Figure 5.5B). An impact on splicing is therefore unlikely to underlie the pathogenicity of this variant. Although this variant was not identified in any control datasets examined (1KG, EVS or dbSNP), another variant change has been reported in Europeans affecting the preceding nucleotide and altering the same amino acid, c.727C>T (p.Pro243Ser). It is present at a minor allele frequency of 0.0005 in the European population sequenced in EVS and was not observed in the homozygous state. Similarly it is predicted to be deleterious to protein function and so it is possible this may also be a disease causing variant when inherited in a biallelic manner.

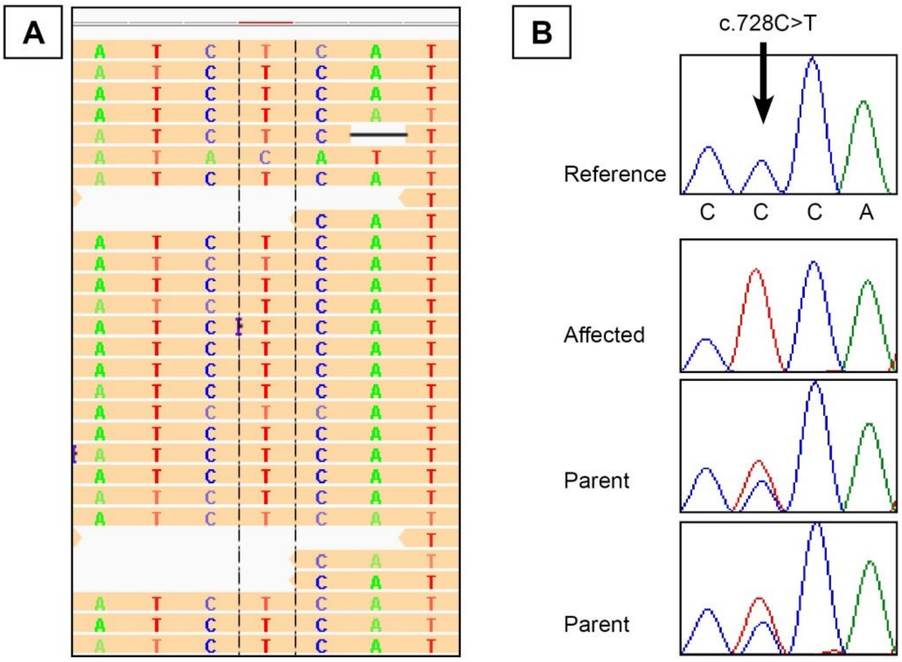


Figure 5.4. Non-synonymous coding variant identified in *NCAPH*.

A) Sequencing reads following Ampliseq multiplex PCR and next generation sequencing (Ion Proton) aligned to the *NCAPH* reference sequence using the Integrative Genomics Viewer (IGV). This identified a variant, c.728C>T (red) in 123 of 129 reads, in one affected case. B) Chromatograms following capillary sequencing of the affected patient and unaffected parents confirmed both parents to be heterozygous carriers and the patient to be homozygous.

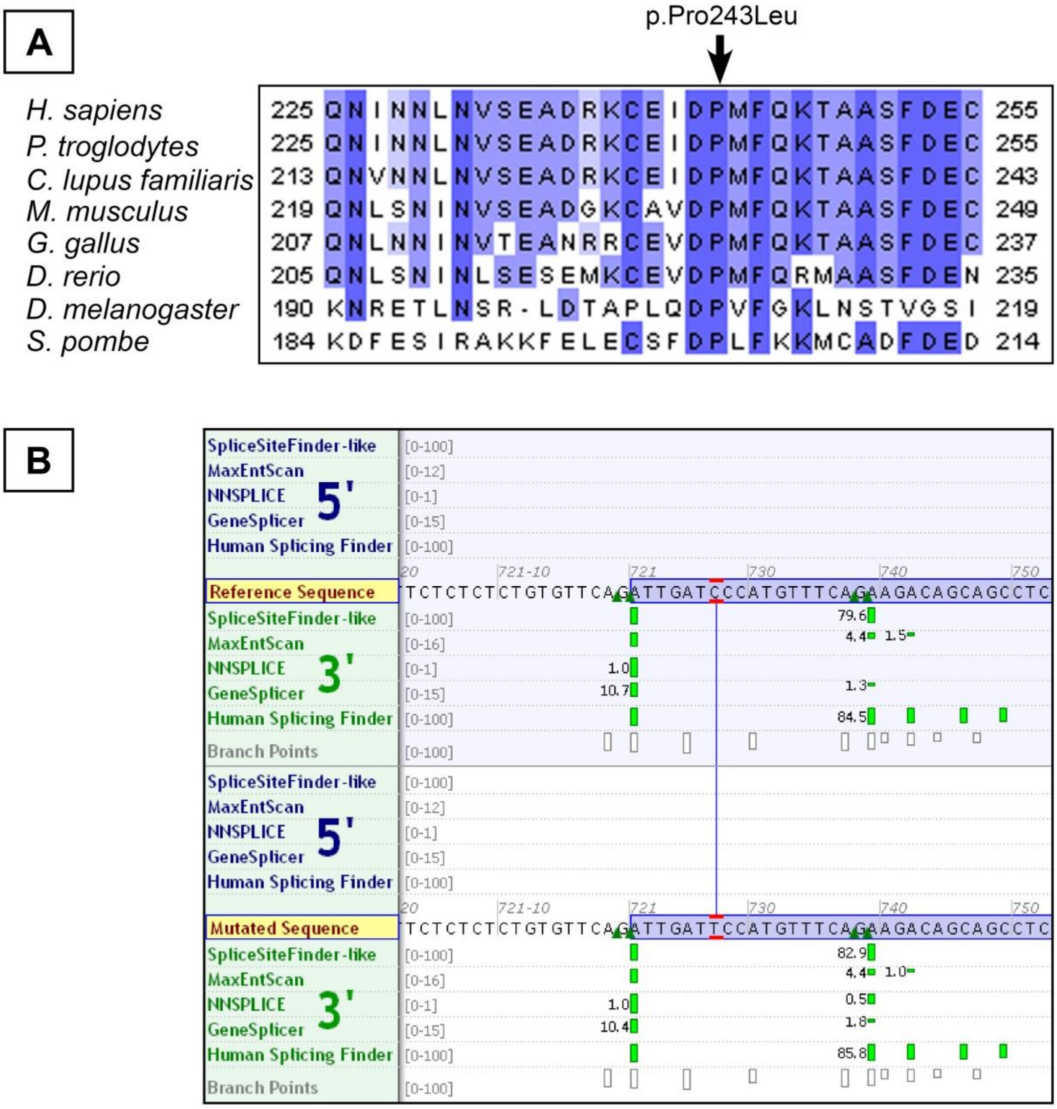

A
p.Pro243Leu
H. sapiens 225 QNINNLNVSEADRKCEIDPMFQKTAASFDEC 255
P. troglodytes 225 QNINNLNVSEADRKCEIDPMFQKTAASFDEC 255
C. lupus familiaris 213 QNVNNLNVSEADRKCEIDPMFQKTAASFDEC 243
M. musculus 219 QNLSNINVSEADGKCAVDPMFQKTAASFDEC 249
G. gallus 207 QNLNNINVTEANRRCEVDPMFQKTAASFDEC 237
D. rerio 205 QNLSNINLSESEMKCEVDPMFQRMAASFDEN 235
D. melanogaster 190 KNRETLNSR-LDTAPLQDPVFGKLNSTVGSI 219
S. pombe 184 KDFESIRAKKFELECSFDPLFKKMCADFDED 214
B
SpliceSiteFinder-like [0-100]
MaxEntScan [0-12]
NNSPLICE 5' [0-1]
GeneSplicer [0-15]
Human Splicing Finder [0-100]
20 721-10 721 730 740 750
Reference Sequence TCTCTCTCTGTGTTCAGATTGATCCCATGTTTCAGAAGACAGCAGCCTC
SpliceSiteFinder-like [0-100] 79.6
MaxEntScan [0-16] 4.4 1.5
NNSPLICE 3' [0-1] 1.0
GeneSplicer [0-15] 10.7 1.3
Human Splicing Finder [0-100] 84.5
Branch Points [0-100]
SpliceSiteFinder-like [0-100]
MaxEntScan [0-12]
NNSPLICE 5' [0-1]
GeneSplicer [0-15]
Human Splicing Finder [0-100]
20 721-10 721 730 740 750
Mutated Sequence TCTCTCTCTGTGTTCAGATTGATTCCATGTTTCAGAAGACAGCAGCCTC
SpliceSiteFinder-like [0-100] 82.9
MaxEntScan [0-16] 4.4 1.0
NNSPLICE 3' [0-1] 1.0 0.5
GeneSplicer [0-15] 10.4 1.8
Human Splicing Finder [0-100] 85.8
Branch Points [0-100]

Figure 5.5. Conservation of amino acid affected by NCAPH variant and predicted impact on splicing

A) Alignment of *NCAPH* sequences in different species illustrating the affected amino acid is highly conserved across both eukaryotes and prokaryotes. The rare variant, c.727C>T, reported in the EVS database also affects the same amino acid. Alignments performed using ClustalW2 (http://www.ebi.ac.uk/Tools/msa/clustalw2/, reference for sequences used given in Appendix II), B) Splice programmes in Alamut software predicted the variant to have a minimal impact on neighbouring splice sites. Percentage values indicate confidence of prediction with 100% being highly confident. Only altered values are shown.

5.3 Clinical description of patients with mutations in condensin genes

A summary of the clinical features of each patient along with corresponding mutation details is provided in Table 5.2.

5.3.1 *NCAPD2*

The patient is of Indian origin and was referred to the study at nine months of age for further investigation of severe microcephaly. He was born at 37weeks gestation to non-consanguineous parents after an uneventful pregnancy. No antenatal scans were performed but IUGR was evident at birth, weight -4.88 s.d. (1kg at 37weeks gestation). No medical support or intervention was required in the neonatal period. Hypospadias was noted on early examination along with reticular hyperpigmented patches on the right knee. Growth continued to be impaired postnatally. At 9months: OFC -12.76 s.d. (31cm), Height -3.8 s.d. (62cm), weight -5.87 s.d. (4.7kg). At 3years: OFC -11.9 s.d. (34.8cm), Height -5.8 s.d. (74cm), weight -7.25 s.d. (7kg). Development has been significantly delayed with social smile only occurring at 7months (normally present from six weeks) and at nine months only cooing sounds were apparent (babbling with a variety of sounds are usually present at this age). Abnormal movements were also noted at the nine month exam with gross hyperactivity and jerky movements of limbs being described. At three years he was able to walk although no speech was apparent at this stage. He was also noted to display little interaction with his parents and exhibited stereotypic repetitive movements similar to behaviours often described in autism. Facial features include large prominent eyes with upslanting palpebral fissures, prominence of the mid-forehead, a wide, flat nasal tip and small chin (Figure 5.6A).

5.3.2 *NCAPD3*

The patient was born to Caucasian, non-consanguineous parents at 37 weeks gestation. Birth weight was at the lower end of normal, weight -1.85 s.d. (2.15kg), whilst length was more substantially reduced, -4 s.d. (40.6cm) (birth OFC not available). At 6 years and 5 months considerable failure in growth was apparent;

OFC -5.4 s.d. (45cm), height -5.7 s.d. (90cm) and weight -9.47 s.d. (9.07kg). On examination he had a discrepancy in leg length and an asymmetry in chest wall shape. Facial features include mild hypertelorism, long philtrum, wide nasal tip and small, low set, posteriorly rotated ears (Figure 5.6). Further skeletal assessment showed bone age to be delayed, long gracile bones, coning of epiphysis in the hands and metaphyseal striations of uncertain significance. Development was reported to be entirely normal. Hyperpigmented patches were noted on the lower face and back along with slight webbing of the neck. Both of these features, along with short stature, are common to neuro-cardio-facial-cutaneous (NCFC) syndromes, a group of autosomal dominant disorders associated with mutations in the Ras-induced MAPK pathway (see Section 1.2.1.1) (Bentires-Alj *et al.*, 2006). Sequencing of genes associated with one of these conditions, Noonan syndrome, was undertaken by the local diagnostic laboratory who identified the patient to be homozygous for a non-synonymous coding variant in *SOS1* (c.1964C>T, p.Pro655Leu) (Roberts *et al.*, 2007). The variant was found to be relatively common in the European population with a minor allele frequency of 0.01 in EVS. This exceeds what would be expected for a disease causing variant in MPD given the low incidence of this disorder in the general population. The variant was predicted to be benign by all but one prediction programme used in Alamut and a leucine residue also occurs at the corresponding site in *Drosophila* indicating this change is tolerated in other species. Only dominant, gain of function mutations have so far been described in *SOS1* in association with Noonan syndrome (Roberts *et al.*, 2007). Notably, both parents were found to be healthy heterozygous carriers of the *SOS1* variant. The patient also did not exhibit any other features consistent with this diagnosis, such as cardiac defects, pectus excavatum/carinatum or coagulation defects (Tartaglia *et al.*, 2007). Short stature is also predominantly postnatal in onset and severe microcephaly is not a feature previously reported in this disorder (van der Burgt, 2007). Therefore this variant was unlikely to be responsible for the phenotype observed and further investigation with WES was subsequently performed.

At the age of 8 he was found to have a malignant brain tumour which was highly aggressive and poorly responsive to treatment. This unfortunately resulted in his death at 11 years of age. The malignancy was classified as an anaplastic medulloblastoma.

5.3.4 *NCAPH*

This patient originates from Portugal and was described as small at birth (growth parameters at birth are unavailable). At 42 years he was microcephalic (OFC= 51.1cm, -3.4 s.d.) but stature was within normal range (height= 168.5cm, -1.4 s.d.). A moderate degree of mental retardation was reported but no other associated medical issues. The patient had an elongated face with a long narrow nose and philtrum (Figure 5.6).

Table 5.2. Molecular and clinical details of three patients with mutations in the non-SMC condensin subunits.

	Condensin II subunit	*Condensin I subunits*	
Gene	*NCAPD3*	*NCAPD2*	*NCAPH*
Mutation details			
c.DNA	c.[382+14A>G] +[1783_1784delG]	c.4120+2T>C	c.728C>T
Protein	p.Ser74Alafs*3 p.Val595Serfs*34		p.Pro243Leu
Pre-natal growth parameters			
Gest^n /weeks	37	37	N/A
Weight /s.d.(kg)	-1.85 (2.15)	-4.88 (1)	N/A 'small at birth'
OFC /s.d.(cm)	N/A	N/A	N/A
Length /s.d.(cm)	-4 (40.6)	N/A	N/A
Post-natal growth parameters and clinical features			
Age of exam	6y5m	3y	42y
OFC /s.d.(cm)	-5.4 (45)	-11.9 (34.8)	-3.4 (51.1)
Height /s.d.(cm)	-5.7 (90)	-5.8 (74)	-1.4 (168.5)
Weight /s.d.(kg)	-9.47 (9.07)	-7.25 (7)	N/A
Development	Normal	Moderate delay	Moderate intellectual disability
Additional features	Medulloblastoma Hyperpigmented patches	Autistic behaviours Abnormal movements	Nil

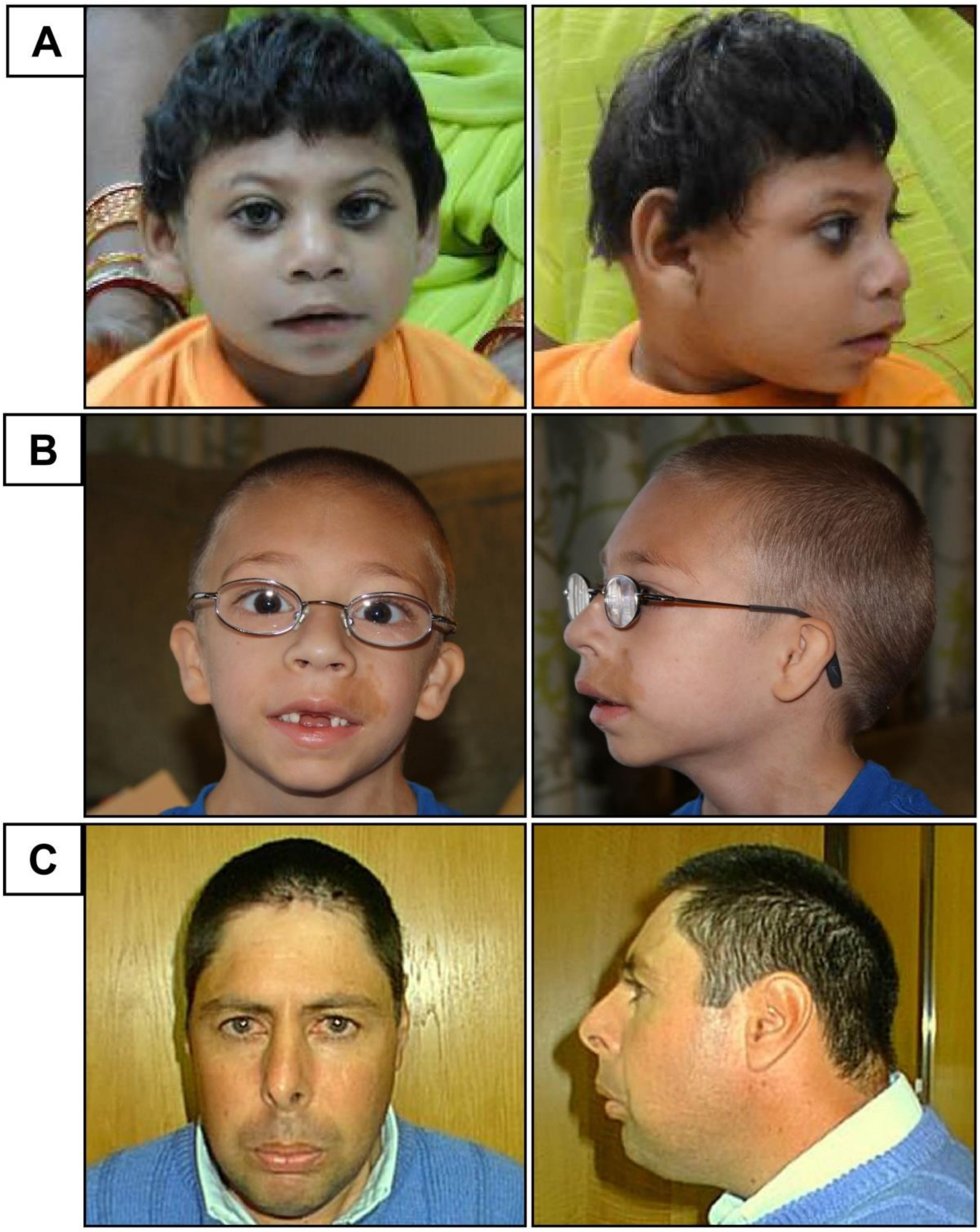


Figure 5.6. Clinical photographs of patients with mutations in the non-SMC condensin subunits

A) Patient at 9 months of age, homozygous for a splice site mutation in *NCAPD2*. B) Patient at 6 years of age, compound heterozygous for mutations in *NCAPD3*. C) Patient at 42 years of age, homozygous for a non-synonymous mutation in *NCAPH*.

5.4 Functional impact of *NCAPD3* mutations on protein

The mutations identified were predicted to significantly impair but not totally abrogate protein function. To establish whether the variants identified disrupted condensin function, cell lines were requested from referring clinicians. Only cells from the patient with compound heterozygous mutations in *NCAPD3* (c.[382+14A>G]+[1783_1784delG]) were available during this thesis. These were therefore compared to fibroblasts from two independent healthy controls referred to as control 1 and control 2. Early passage (≤6) fibroblasts were used with control and patient cells matched to within two passages for each experiment.

5.4.1. Impact of intronic mutation on splicing

The intronic mutation identified in *NCAPD3* was predicted to create an alternative splice donor site in intron 3 (Figure 5.3). To investigate whether this variant disrupted normal splicing, RT-PCR studies were performed. Primers were used to amplify the coding region between exon 2 and 4 (Figure 5.7, red), exon 2 and 6 (Figure 5.7, green) and exon 2 and 10 (Figure 5.7, orange) and demonstrated a marked reduction in the level of full length transcript present in patient cells compared to control 1 cells. Notably there was also a marked increase in an alternate smaller transcript which was detected with all three primer pairs. The difference in band size between the full length and alternate transcripts corresponded to the size of exon 3 (163 bp). To confirm this, the DNA present in each band was extracted and sequenced which confirmed the large PCR fragment corresponded to the full length reference transcript and the smaller fragment corresponded to a transcript in which exon 3 had been omitted. A third, even smaller transcript was also identified in both patient and control cell lines following amplification between exon 2 and 6 and between exon 2 and 10. Extraction and sequencing of the DNA in this band revealed the presence of a transcript which lacks exon 3 and 4. This transcript appeared similarly expressed in both patient and control cells as judged by the intensity of the band on the gel following electrophoresis (Figure 5.7B). Absence of exon 3 results in a frameshift of the coding sequence whereas no shift in frame occurs when exon 3 and 4 are omitted.

These findings are different from the predicted impact on splicing shown in Figure 5.3 which suggests an alternative transcript incorporating an additional 13 bp of intron 3 may occur. However, this fragment was not detected following extraction and sequencing of the large PCR fragment obtained from the patient cells which corresponded to the full length transcript. The difference between predicted and observed transcripts in patient cells may be explained by the presence of a co-segregating variant which was not identified on WES. Alternatively, this may reflect the limitations of splice prediction programmes in encapsulating the full complexity of mRNA splicing.

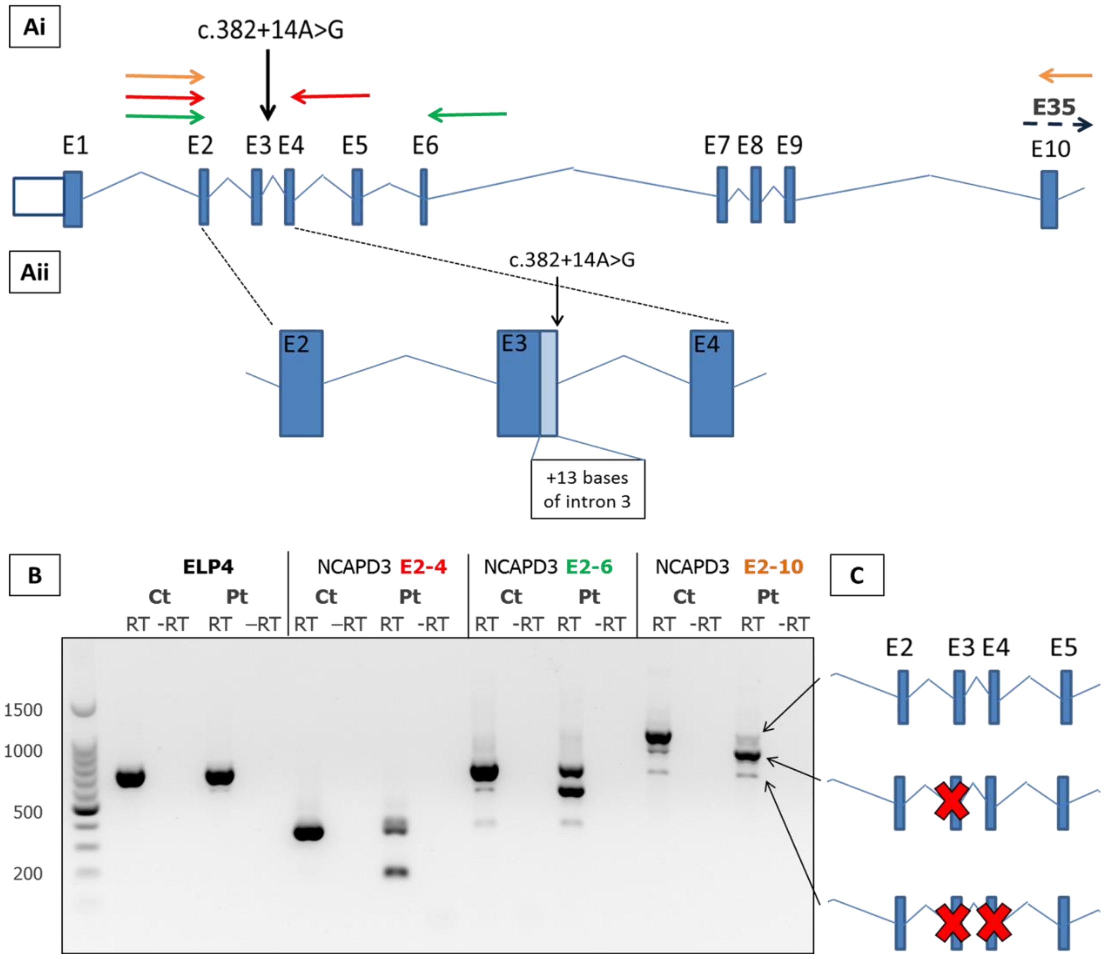


Figure 5.7. Impact of intronic mutation in *NCAPD3* on splicing

Ai) *NCAPD3* is composed of 35 exons, the first 10 exons (E1-10) of which are depicted to show the position of the intronic mutation identified and the position of three primer pairs used in the amplification of cDNA (orange , red and green). The second nonsense mutation (c.1783_1784delG) identified in this patient resides in exon 15. Aii) The intronic mutation is predicted to create an alternative splice site which would be expected to result in a transcript with an additional 13 bases of intron 3. B) RT-PCR of RNA extracted from patient fibroblasts (Pt) and control 1 fibroblasts (WT) showed alternate splicing of *NCAPD3* was present in patient cells but there was no change in splicing of a control transcript (*ELP4)*. Notably, a band corresponding to the predicted transcript shown in Aii was not present in patient cells. C) Sequencing of gel-extracted DNA from the different bands amplified by primers at exon 2 and 10 in both patient and control established that the largest band which appeared reduced in patient cells, corresponded to the full length transcript while the second, shorter PCR fragment, which appeared increased in the patient cells, corresponded to a transcript in which exon 3 had been skipped. The smallest band identified following amplification with primers between exon 2 to 6 (green) and 2 to 10 (orange) was composed of a transcript in which exon 3 and 4 had been removed. Levels of this transcript were similar between patients and controls.

Abbreviations: RT=reverse transcription, -RT=reverse transcription performed in the absence of reverse transcriptase enzyme.

5.4.2 Impact of mutations on protein levels

To determine if NCAPD3 protein levels were reduced in patient cells, immunoblotting was performed on whole cell extracts from patient-derived fibroblasts. NCAPD3 protein levels were markedly reduced in patient cells compared to control 1 and 2 (Figure 5.8). Depletion of NCAPD3 with RNAi in a control fibroblast line showed reduction in a protein band of similar molecular weight confirming specificity of the NCAPD3 antibody. No change in NCAPD3 levels was seen following transfection with RNAi targeting luciferase in the same control cell line.

Protein immunoblotting was also performed to determine whether cellular levels of other condensin subunits were affected by the reduction in NCAPD3. No change in protein levels were observed in the other two non-SMC condensin II subunits, NCAPH2 and NCAPG2, or in the condensin I subunit, NCAPH in patient cells compared to control cells. RNAi depletion of NCAPD3 in control 1 cells also had no effect on the protein levels of these subunits.

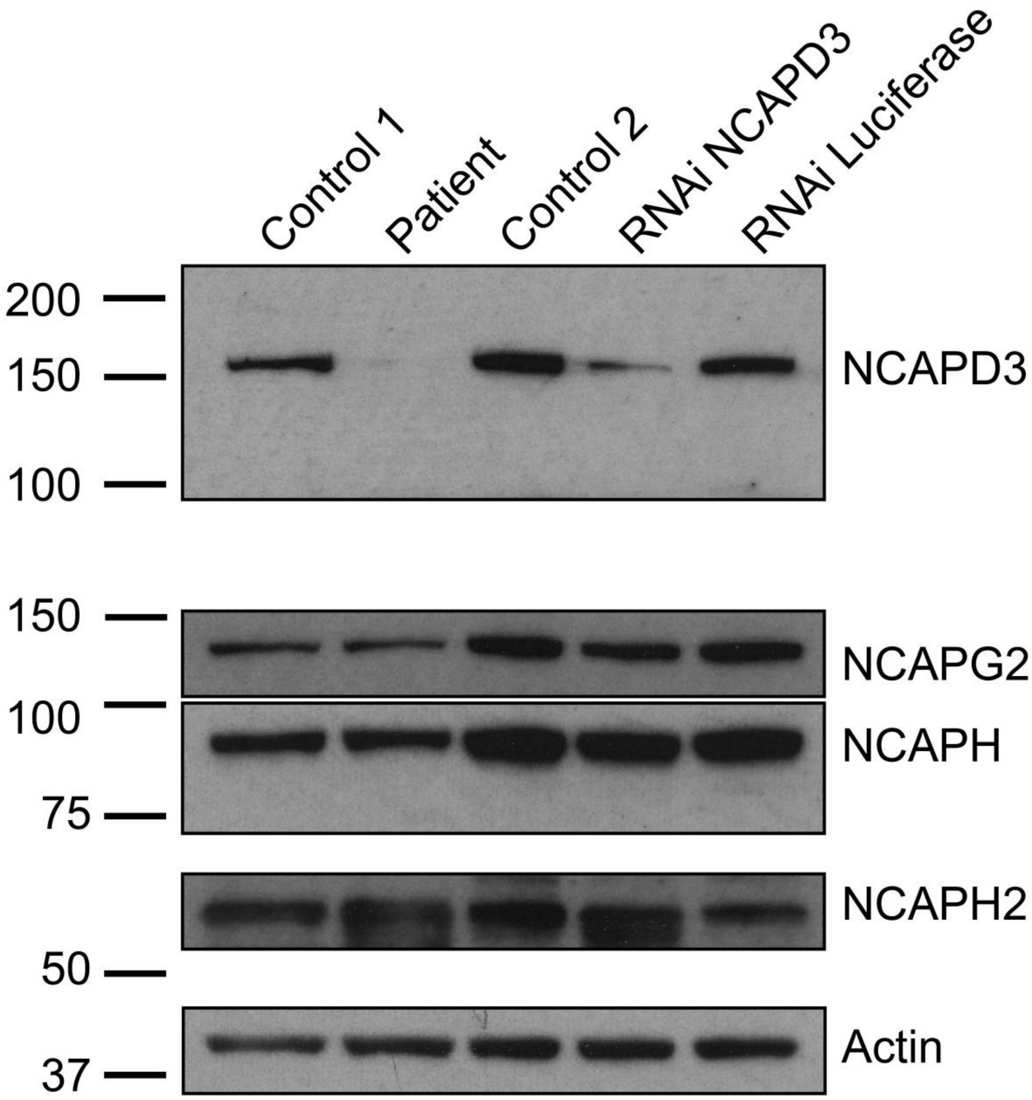


Figure 5.8. Impact of *NCAPD3* mutations on protein levels of condensin subunits

Immunoblotting showed reduced NCAPD3 protein levels in patient fibroblasts compared to two control fibroblast cell lines. Actin used as loading control. No consistent difference was seen in the protein levels of other condensin I and II subunits. NCAPD3 protein levels were also reduced in control cells transfected with RNAi targeted to NCAPD3. Again, other condensin subunits were unaffected in these cells. Transfection with RNAi targeted to luciferase had no impact on NCAPD3 or other condensin subunits.

5.5 Abnormalities in *NCAPD3* patient fibroblasts during mitosis

5.5.1 Abnormal chromosome morphology

Previous studies have shown that disruption of *NCAPD3* expression results in abnormally shaped chromosomes in mitosis (Green *et al.*, 2012). To determine whether the patient cells carrying *NCAPD3* mutations display a similar phenotype, metaphase spreads were prepared from the patient-derived cultured fibroblasts. Strikingly, condensed chromosomes in the patient cells also appeared 'thin and curly' compared to both control fibroblast lines (Figure 5.9 A-C). 100% of metaphase cells examined in the patient (n = 20) exhibited the same chromosome morphology. Metaphase spreads were also examined following RNAi knockdown of NCAPD3 in one of the fibroblast control lines, control 1. Condensed chromosomes displayed a similar, although less severe phenotype to the NCAPD3 patient cells (Figure 5.9 D), possibly because NCAPD3 protein levels were not depleted to the same extent as in patient cells (Figure 5.8). Transfection with RNAi targeting luciferase in the same cell line (Figure 5.9 E) had no impact on morphology.

5.5.2 Chromosome segregation defects

Previous studies have also shown that NCAPD3 and the condensin II complex are required for faithful chromosome segregation (Green *et al.*, 2012). To determine whether patient cells show a similar phenotype, fibroblasts were fixed on coverslips and both DNA and microtubules were visualised with immunofluorescent markers. Chromosome segregation abnormalities were observed in patient cells including anaphase bridges and lagging chromosomes (Figure 5.10A).

Chromosome segregation defects were found to be significantly increased in the patient fibroblasts (14%) compared to both controls (control 1=3%, control 2=2%, t-test; $p<0.006$) (Figure 5.10B). The most frequently observed abnormality in patient cells was lagging chromosomes (8.5% +/-4.1 s.d.) followed by anaphase bridges (5.8% +/-1.6 s.d.) although this difference was not statistically significant.

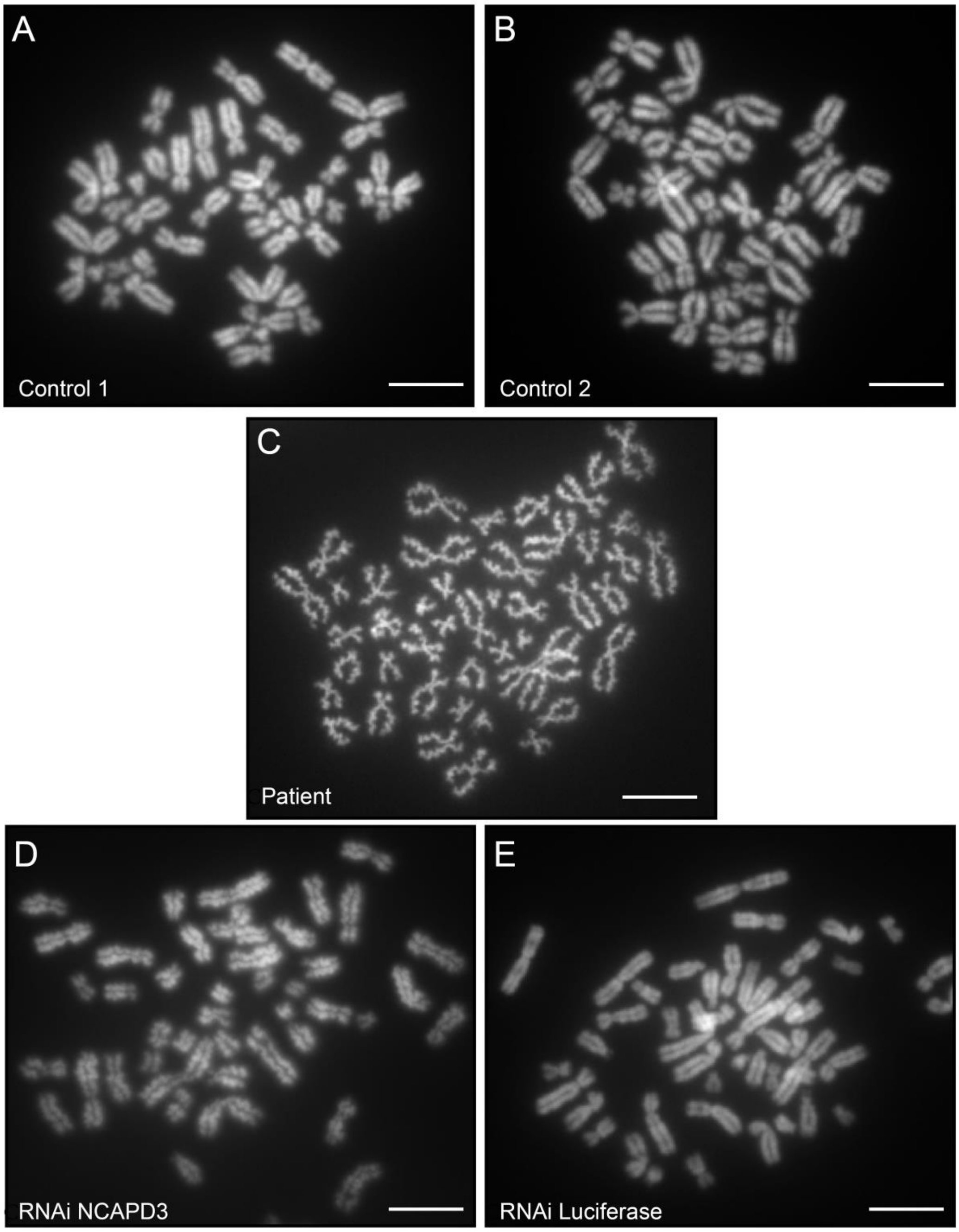


Figure 5.9. Metaphase spreads in *NCAPD3* patient fibroblasts

Chromosomes appear normally condensed in two independent control cell lines, control 1 (A) and control 2 (B). In contrast, chromosomes in the *NCAPD3* patient cell line are abnormally shaped appearing thin and curly (C). A similar but less extreme phenotype was observed in control cells transfected with RNAi targeted to NCAPD3 (D) whereas transfection with RNAi targeted to luciferase in the same control line had no impact on chromosome morphology (E). Scale bar = 7µm.

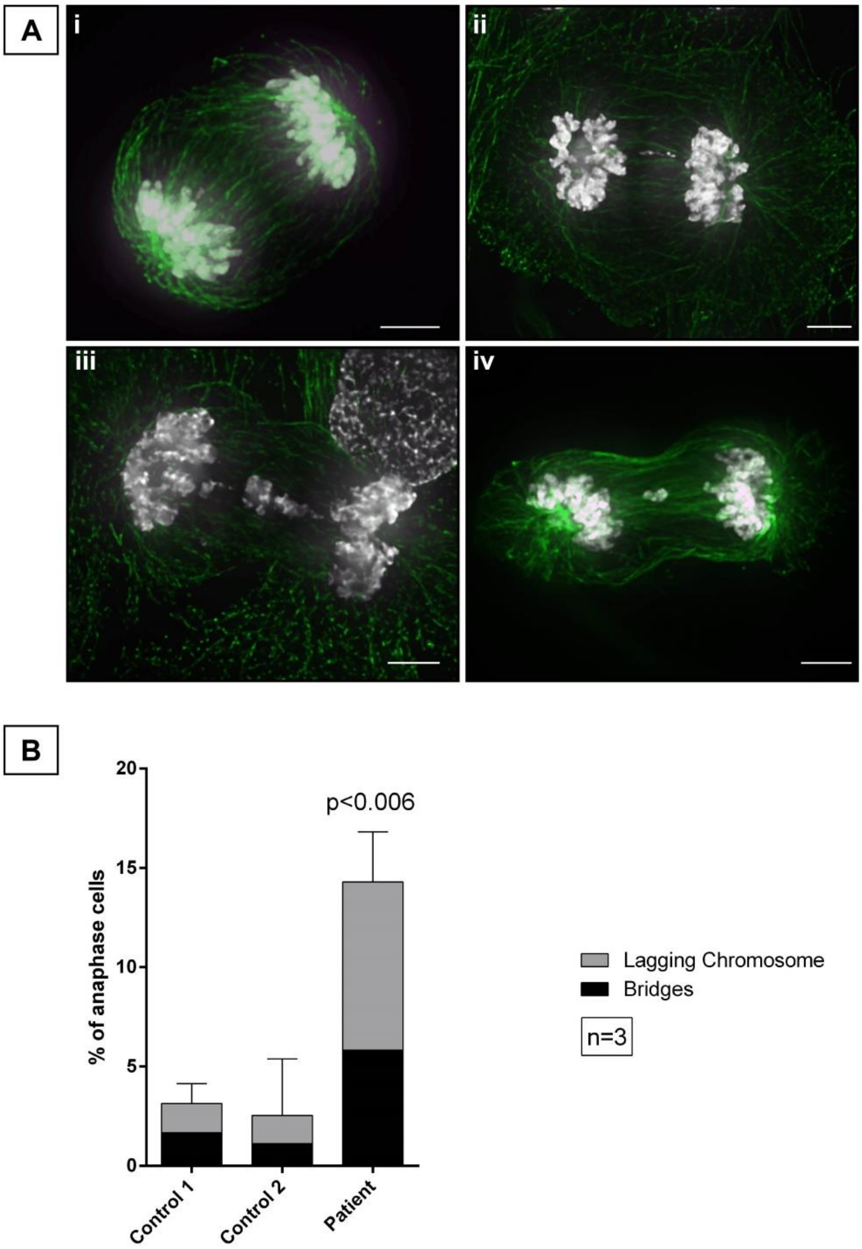


Figure 5.10. Chromosome segregation defects in *NCAPD3* patient fibroblasts

A) Images of patient-derived fibroblasts in anaphase. DNA stained with DAPI (white) and microtubules stained with α-tubulin antibody (green). i) Normal chromosome segregation. ii and iii) Anaphase bridges are shown in which chromatin can be visualised spanning the distance between segregating chromosomes. This suggests a failure in separation of homologous chromosomes. (iv) a lagging chromosome is present suggesting a failure in the spindle assembly checkpoint (SAC) has occurred. B) 80-100 anaphase cells were scored in each cell line in three independent experiments. There was a significant increase (unpaired t-test) in the total number of cells with chromosome segregation defects in the patient (mean=14% +/-2.5 s.d.) compared to two independent controls (control 1; mean=3% +/-1 s.d., control 2; mean=2.5% +/-2.8 s.d.). Scale bar =5.4µm.

5.5.3 Prevalence of aneuploidy

As *NCAPD3* patient cells showed an increase in chromosome segregation defects this may result in an increase in the proportion of cells with an abnormal DNA content. To determine if low NCAPD3 levels results in an increase in the proportion of aneuploid (DNA content>4N) or tetraploid cells (DNA content =8N), flow cytometric analysis of DNA content following staining with propidium iodide was performed in patient cells compared to both control lines (Figure 5.11). Over three independent experiments, no significant difference was seen in the proportion of cells with a DNA content of more than 4N between any of the cell lines (patient; mean=0.39% +/-0.06 s.d., control 1; mean=0.37% +/-0.27 s.d., control 2; mean=0.31% +/-0.4 s.d., unpaired t-test p > 0.7 for all comparisons).

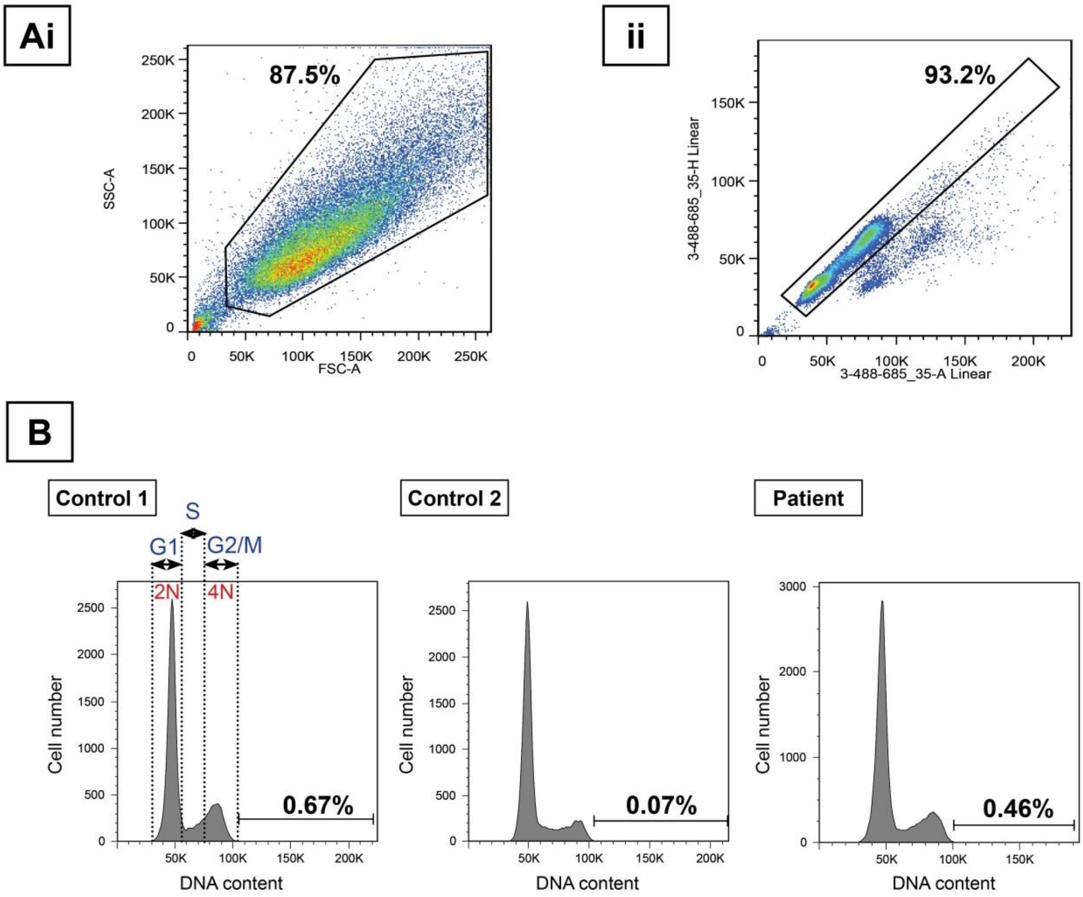


Figure 5.11. Measurement of DNA content (>4N) in *NCAPD3* patient fibroblasts by flow cytometry.

A) Following flow cytometry, cells were gated using scatter plots of FSC (a measure of size) vs SSC (a measure of granularity) to exclude debris (i). Cell aggregates were then excluded from further analysis based on area vs height (ii). Percentage indicates % of cells within gated region. B) Histograms of a representative flow cytometric experiment showing cells in G1 (2N), S phase (between 2N and 4N peaks) and G2/M (4N) in two independent control cell lines and NCAPD3 patient cells. Very few cells were seen in the area indicating a DNA content > 4N (horizontal bar) in any of the cell lines. Overlying number indicates percentage of live cells in this area.

5.7 Discussion

5.7.1 Deleterious mutations in condensin complex genes identified in MPD patients

In this Chapter, mutations in *NCAPD2* and *NCAPH*, encoding two of the non-SMC subunits of condensin I and mutations in *NCAPD3,* encoding a non-SMC subunit of condensin II, are described in three separate cases presenting with varying degrees of growth failure. This is the first time mutations in genes encoding subunits of the condensin complexes have been described in association with human disease. The advent of NGS has led to a large increase in the number of genes implicated as disease-causing in MPD following the identification of apparently deleterious variants (Alazami *et al.*, 2012, Dauber *et al.*, 2012, Shaheen *et al.*, 2014). However, as each genome harbours approximately 100 loss of function variants (MacArthur *et al.*, 2012) assigning causality without supporting evidence can lead to variants being incorrectly documented as disease causing (Bell *et al.*, 2011, Norton *et al.*, 2012) which is misleading for clinicians and potentially harmful to patients and their families (MacArthur *et al.*, 2014). Therefore, further cellular studies to verify the pathogenicity of the variants identified in these novel candidate disease genes were performed.

Experiments in primary fibroblasts derived from the patient with mutations in *NCAPD3* demonstrated a marked reduction in total NCAPD3 protein levels as a consequence of altered splicing of one allele in combination with a null allele confirming the mutations have a deleterious impact on the protein. Total cellular levels of the other condensin II subunits appeared unaffected although isolating and examining chromatin bound protein would be useful in the future to determine whether the localisation or stability of the condensin complex is adversely affected by loss of NCAPD3. Additionally, chromosome morphology was strikingly abnormal appearing elongated and ‘curly’ similar to observations following depletion of condensin II in various other cell types including human HeLa cells (Ono *et al.*, 2003, Green *et al.*, 2012). Although not performed in this study, measuring axial length and width of chromosomes would allow quantification of the defect and an indication of severity (Hudson *et al.*, 2003). Furthermore, an increase in chromosome segregation defects were also observed in keeping with findings from other studies (Hudson *et al.*, 2003, Green *et al.*, 2012) providing evidence that condensin II function is impaired in this patient. Similar experiments examining the effect of the variants discovered in *NCAPD2* and *NCAPH* would also be useful to establish the impact of these mutations on protein function but unfortunately patient cells were not available in these cases.

Notably, the phenotypes of the three cases varied although microcephaly was present in all patients. This is perhaps either a reflection of the differing functions of the two condensin complexes (Section 5.1.2) or of the severity of different mutations identified. The phenotypes included isolated microcephaly with intellectual disability (*NCAPH*), severe microcephaly with short stature and severe developmental delay (*NCAPD2*) and lastly, MPD with normal development (*NCAPD3*). Strikingly, the patient with mutations in *NCAPD3* developed malignancy in childhood. The unusual and aggressive nature of the malignancy, an anaplastic medulloblastoma, in combination with a rare growth disorder, raises the possibility the two are causally related. However, confidently determining associated clinical phenotypes is difficult at this point in time with only mutations identified in single patients for each gene. Identifying additional mutations in other similarly affected patients will increase confidence that mutations in condensin genes cause growth failure and will help to clarify commonly associated phenotypes. Longer term follow up will also be required to establish whether cancer predisposition is common to condensin related disorders or a chance association in this one case. However, as only three patients were identified from a total of 301 patients sequenced (1%), mutations in condensin genes likely represent a rare cause of growth failure and identifying further patients may be difficult. This emphasises the importance of open access databases and sharing of data between researchers to create larger more informative patient cohorts. Accessing data from large sequencing projects such as Deciphering Developmental Disorders (Firth *et al.*, 2011) may also help identify additional patients.

5.7.2 Mechanism of growth failure in condensin dysfunction

Condensin complexes are critical in the regulation and coordination of chromosome condensation as well as ensuring resolution of sister chromatids during mitosis (Hirano, 2005). It has now been demonstrated in several organisms that disruption of either condensin I or II results in aberrant mitosis with increase in chromosome segregation defects and mitotic delay (Bhalla *et al.*, 2002, Hudson *et al.*, 2003, Hirota *et al.*, 2004, Green *et al.*, 2012). It is therefore conceivable that mutations in condensin complex genes can impact on global cellularity by preventing efficient and effective cell division. Further supporting evidence for a role of condensin complexes in ensuring normal growth is seen in animal studies with microcephaly being reported in *ncapg*, *ncapd2* and *ncaph* morphant zebrafish (Seipold *et al.*, 2009). Additionally, decreased head size and length were reported in *smc4* mutant zebrafish (Amsterdam *et al.*, 2004) (ZFIN historical data, 2006). Early embryonic lethality is observed in *Ncapg2* null mice due to failed expansion of the inner cell mass (Smith *et al.*, 2004) and embryonic lethality has also been documented in *Ncaph* and *Ncaph2* null mice (http://www.sanger.ac.uk/mouseportal/). Interestingly, a homozygous hypomorphic mutation in *Ncaph2* in mice affecting two of three predicted transcripts results in impaired T cell development with reduced thymus size and circulating numbers of T-lymphocytes (Gosling *et al.*, 2007). No other abnormalities in development or growth however were reported suggesting a lineage specific role of the affected transcript or a separation of function effect in which chromosome segregation is not impaired. In humans, aberrant chromosome condensation is also seen in microcephalin (MCPH1) deficient cells, mutations in which cause autosomal recessive primary microcephaly (Trimborn *et al.*, 2004, Trimborn *et al.*, 2006). MCPH1 appears to prevent premature mitotic entry through early activation of chromosome condensation providing evidence of a link between neurogenesis and the misregulation of chromosome condensation.

5.7.3 Possible mechanisms of cancer predisposition in patients with condensin mutations

The occurrence of a severe and unusual malignancy during childhood in one of the patients suggests cancer predisposition may be a feature of condensin mutations. Cancer predisposition due to abnormal condensin function is plausible given the segregation defects observed that may lead to structural rearrangements or abnormal DNA content in daughter cells. However, the mechanism by which condensin dysfunction results in chromosome segregation defects has not yet been fully elucidated. In simplistic terms, it is widely considered that failure in chromosome compaction leads to entanglement between sister chromatids which persist during segregation (Hirano, 2005, Jeppsson *et al.*, 2014). Additionally, condensin I complex assists in the removal of cohesin just prior to anaphase (Hirota *et al.*, 2004) and also appears to have a specific function in the resolution of repetitive DNAs in *Saccharomyces cerevisiae* (D'Amours et al., 2004). Thus segregation defects may be a direct consequence of loss of condensin at chromatin but may also in part be due to indirect effects on other proteins including type II topoisomerase which assists in the detanglement of sister chromatids and whose localisation to chromosomes is disrupted by depletion of condensin in other organisms (Coelho *et al.*, 2003, Hudson *et al.*, 2003). Furthermore, condensin depletion impairs kinetochore orientation impacting on kinetochore-microtubule attachments but without activating the spindle assembly checkpoint (Cimini *et al.*, 2001, Ono *et al.*, 2004). As well as essential roles in chromosome condensation and segregation, accumulating evidence also supports a role for condensin in the regulation of gene transcription (Bhalla *et al.*, 2002, Rawlings *et al.*, 2011). It is hypothesised that intramolecular linking and compaction of DNA by condensins reduces the accessibility of transcriptional activators to chromatin (Jeppsson *et al.*, 2014). Failure in gene silencing during mitosis is therefore another factor which could possibly contribute towards the development of malignancy. Finally, condensin complexes have been directly implicated in DNA repair during interphase presenting another possible source of genome instability in condensin dysfunction (Heale *et al.*, 2006).

Despite a marked increase in chromosome segregation defects, no increase was seen in aneuploidy or tetraploidy in patient cells compared to controls following flow cytometry in this study. However, flow cytometry may not be the ideal experiment to address this question for two reasons. Firstly, a delay in mitosis may cause patient cells to grow much slower in culture compared to control cells and this may make small increases in aneuploidy difficult to detect as fewer patient cells will be actively dividing at any one point. Secondly, in this study the DNA content of every cell was measured regardless of its stage in the cell cycle. As the chromosome number changes as a cell progresses through the cell cycle, a post mitotic tetraploid cell in G1 phase will be indistinguishable in terms of DNA content from a diploid cell which has just replicated its DNA in preparation for mitosis. Alternatively, changes in chromosome number as well as structural rearrangements can be detected in interphase cells using a range of techniques such as G-banding, spectral karyotyping and fluorescence in situ hybridisation (FISH) analysis either using a small panel of probes or whole chromosome painting (Ried *et al.*, 1998). Large numbers of cells would need to be reviewed to identify subtle increases in aneuploid cells which would be possible with automated analysis (Wang *et al.*, 2012). This would enable more sensitive detection of subtle alterations in chromosome number independent of variation in DNA content throughout the cell cycle.

5.7.3 Conclusions

In this Chapter, variants identified in genes encoding subunits of the condensin complexes were characterised further by examining their effect on RNA, protein and mitosis providing supporting evidence for the discovery of mutations in condensin genes as a novel but rare cause of growth failure. Aberrant and delayed mitosis resulting from impaired condensin function could be responsible for the growth phenotype observed in these patients and future experimental plans to investigate this are discussed in Chapter 7.

CHAPTER 6: Contribution of changes in chromosomal copy number variation (CNV) to the aetiology of MPD

6.1 Introduction

The preceding three Chapters describe the identification of single nucleotide variants (SNVs) or small insertions/deletions (less than 20 base pairs) in disease causing genes in MPD patients. However, more than 50% of families within this cohort still remain without a diagnosis despite performing WES in the majority of cases. One possibility is that some patients have larger alterations in chromosome structure resulting in the gain or loss of genetic material which are unable to be detected by standard variant calling methods. As each human somatic cell contains two copies of its DNA, any variation in DNA content will alter the copy number of the affected region. Variation in copy number could involve a whole chromosome, which most commonly results from a failure in the segregation of chromosome pairs during meiosis (non-disjunction) (Sankaranarayanan, 1979), or just part of a chromosome. The latter can occur following the breakage and re-joining of chromosome segments (Currall *et al.*, 2013) in which genetic material is either lost (microdeletion) or gained (microduplication) resulting in a reduction or increase in copy number respectively. The size of the region affected can vary greatly from 1 kb to several megabases covering a large number of genes (Iafrate *et al.*, 2004, Redon *et al.*, 2006).

CNVs, defined as a DNA segment that is present at a variable copy number in comparison with a reference genome, are a common occurrence in the general population affecting both coding and noncoding regions (Redon *et al.,* 2006). On average each individual harbours approximately 12 large scale CNVs (greater than 1 Mb) in their genome (Iafrate *et al.*, 2004). The majority of these occur in regions in which variation in copy number is frequently observed in healthy individuals (polymorphic regions) and therefore are unlikely to be of clinical significance. However, if the CNV involves one or several genes whose function is dosage dependent then this may have a negative impact on cellular processes resulting in phenotypic abnormalities. Loss or gain of a specific genomic region has now been identified as a common cause of intellectual disability (Shaw-Smith *et al.*, 2004) and many malformation syndromes have been attributed to contiguous gene deletions (Theisen *et al.*, 2010). Some of the more commonly recognisable syndromes resulting from chromosomal microdeletions include Williams syndrome (7q11.23 deletion) (Martens *et al.*, 2008), DiGeorge syndrome (22q11.2 deletion) (Monteiro *et al.*, 2013) and Miller-Dieker syndrome (17p13.3 deletion) (Stratton *et al.*, 1984). Abnormal growth has also been reported in association with variation in copy number. For example, deletion of the 5q35 region encompassing the gene *NSD1* (nuclear receptor binding SET domain protein 1) causes Soto syndrome characterised by overgrowth and macrocephaly (Kurotaki *et al.*, 2002). Additionally, duplications of *NSD1* have been identified in patients with growth restriction (Franco *et al.*, 2010). Reciprocal growth phenotypes have also been seen in association with CNVs at other loci including 16p11.2 (Jacquemont *et al.*, 2011) and 1q21.1 (Brunetti-Pierri *et al.*, 2008). It is therefore important to explore the possibility of CNVs as a cause of MPD. This may lead to the discovery of novel disease genes by identifying those in the affected regions which are critical to the growth phenotype.

Submicroscopic CNVs can be detected by comparative genomic hybridisation using chromosomal microarrays (array-CGH) and currently this is the first line clinical diagnostic test for the detection of CNVs in patients with intellectual disability and developmental abnormalities (Miller *et al.*, 2010). Although array-CGH is the current diagnostic gold standard method for identifying CNVs, the size of CNVs detectable is limited by the probe density of the microarray. The increasing implementation of large scale sequencing projects has led researchers to design strategies for detecting CNVs in NGS data (Chiang *et al.*, 2009). Current NGS methods result in the production of multiple short reads (less than 200 base pairs) across each defined genomic region within a DNA sample. A reduction in copy number within the sample will result in a lower number of sequencing reads covering the affected region (reduced depth of coverage). Conversely, duplicated regions will lead to an over-representation of reads corresponding to a higher read depth. This provides a potential method for the detection of CNVs from NGS data with the possibility of high resolution given the short read length.

Many programmes have now been designed to identify CNVs from NGS data (Zhao *et al.*, 2013) and have been successfully implemented to identify clinically relevant CNVs with patient cohorts (de Ligt *et al.*, 2013, Gilissen *et al.*, 2014). This Chapter investigates the utility of WES in identifying CNVs using ExomeCNV (Sathirapongsasuti *et al.*, 2011) (Section 2.5.2.3), a programme specifically designed for this purpose and potentially able to identify single exon deletions (120 bp). ExomeCNV analysis was performed in 105 MPD patients resulting in the identification of likely causative chromosomal microdeletions in six families.

6.2 Identification of pathogenic microdeletions from WES following ExomeCNV analysis

Although ExomeCNV was performed in all 102 affected individuals from 95 families, 18 samples failed analysis (9 singletons and 9 samples with parents) as a consistent baseline indicating neutral copy number was not achieved in at least 50% of the exome, therefore these samples were too noisy to interpret. Only microdeletions were prioritised for further analysis as these are a more common cause of disease compared to microduplications (Vissers *et al.*, 2012) with analysis of gain in copy number planned for the future.

In total, 20,347 microdeletions were called by ExomeCNV ranging from a single exon (118 bp) to 248 Mb in size in 84 affected cases (mean 242 per patient). As it was not feasible to validate every region called following ExomeCNV analysis, I designed and implemented a filtering process (similar to that described in Section 3.3) to prioritise possible microdeletions for validation. Figure 6.1 shows an overview of the designed filtering pipeline.

Array-CGH data was available from two patients prior to WES in which CNVs of uncertain significance had already been identified. Both CNVs were identified on ExomeCNV analysis and this was used to define a log2 ratio (log2R) threshold for prioritising deleted regions (minimum log2R of known deletion = -0.79). Therefore all regions with a log2R of -0.70 or below were selected for further review. This identified a total of 744 regions in 84 affected patients (mean deletions per patient=9, range 1 to 50) (Table 6.1) with size of deletion ranging from 188 bp to nearly 18 Mb. These regions were then uploaded into an SQL database for further filtering.

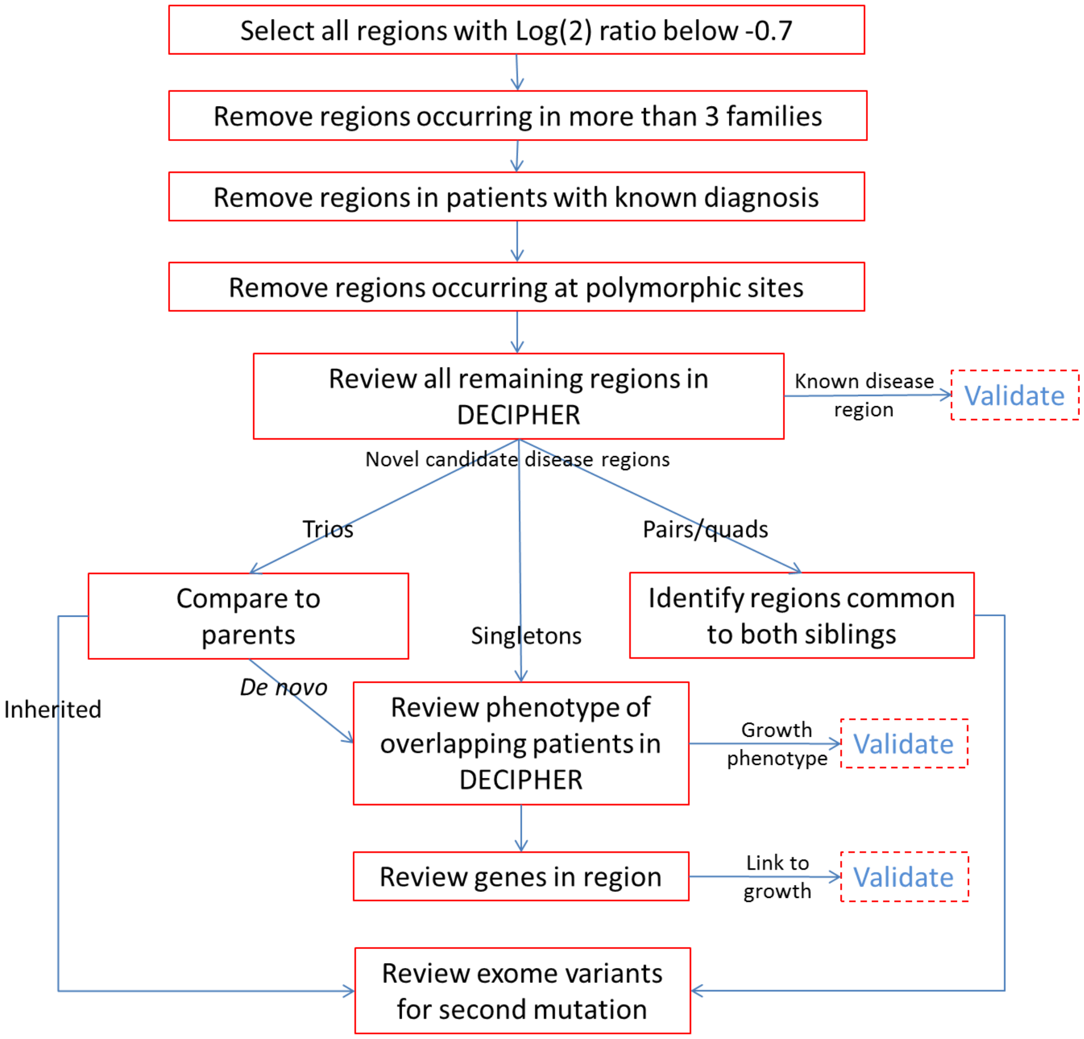

Select all regions with Log(2) ratio below -0.7
Remove regions occurring in more than 3 families
Remove regions in patients with known diagnosis
Remove regions occurring at polymorphic sites
Review all remaining regions in DECIPHER
Known disease region
Validate
Novel candidate disease regions
Trios
Pairs/quads
Compare to parents
Singletons
Identify regions common to both siblings
Inherited
De novo
Review phenotype of overlapping patients in DECIPHER
Growth phenotype
Validate
Review genes in region
Link to growth
Validate
Review exome variants for second mutation

Figure 6.1. Analysis pipeline of ExomeCNV data for the detection of pathogenic microdeletions

Abbreviations; DGV=database of genomic variation, DECIPHER= Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources.

Of the deleted regions identified with a Log2R of less than -0.7, 90% were under 100 kb in size and only 2.4% were over 1 Mb (Figure 6.2A). The most frequently occurring deletion size was between 1 and 50 kb (68%). Only 16 single exon deletions of less than 500 bp were called in the analysis (2.3%). There was no significant difference in the Log2R for each group (Figure 6.2B) indicating that regions with a log2R of less than -0.7 were called with similar confidence independent of size.

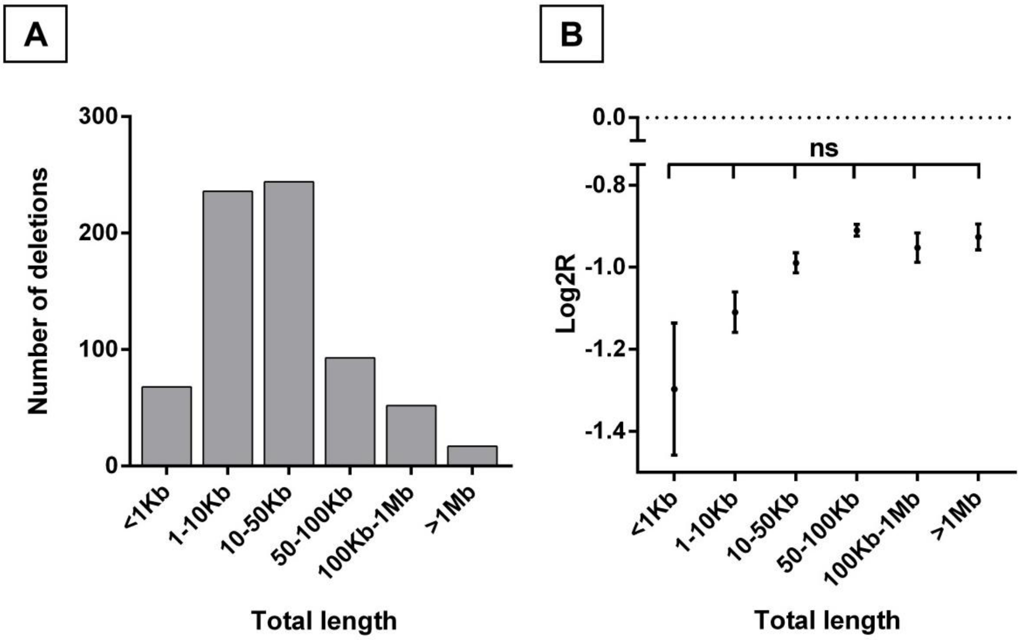


Figure 6.2. Comparison of frequency and Log2 ratios of different sized deletions

A) Number of deletions with log2R less than -0.7 within each size range demonstrating the majority of deletions called were less than 50 kb in size. B) Corresponding mean Log2R of deletions in each size range. Vertical lines indicate standard error of the mean. Each group was compared to all others using a Kruskal-Wallis test followed by a Dunn’s multiple comparison test. No significant difference was seen in Log2R between each group. Only regions in autosomes were included in this analysis.

As it was anticipated that large genetic heterogeneity is present within the cohort, apparent deletions incorporating exactly the same region in multiple cases were deemed more likely to represent sites of common polymorphism or result from inconsistencies in exon capture between case and control samples leading to a consistent source of error in CNV calling. Therefore, as a first line approach, only deletions called in three or less patients (3.5% of affected cases sequenced) were prioritised for review. This excluded 49.5% of regions (376 remaining, Table 6.1). To further prioritise likely causative regions, those occurring in patients in whom a molecular diagnosis had been identified, as described in Chapter 3, were also removed excluding a further 20% of regions (228 remaining in 49 patients, Table 6.1).

Deleted regions were then uploaded into the UCSC genome browser (http://genome.ucsc.edu/, accessed 06.2013) and compared to CNVs identified in control cohorts in the Database of Genomic Variants (DGV; http://dgv.tcag.ca/; accessed 13.06.13) (MacDonald *et al.*, 2014). Any region overlapping with commonly occurring CNVs (polymorphic regions) in DGV were removed resulting in 97 possible deletions in 22 affected cases (Table 6.1). These remaining regions were then compared to CNVs in DECIPHER (Databas*e* of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources) (Bragin *et al.*, 2014) (https://decipher.sanger.ac.uk/, accessed 06.2013) which contains the results of array-CGH analysis performed in patients with developmental disorders. This allowed identification of any patients with possible deletions overlapping CNV regions previously associated with abnormalities in growth. These regions were then prioritized for validation (three regions in three cases, Table 6.1).

Table 6.1. Reduction in number of candidate CNVs through filtering pipeline

Filtering Step	***Number of remaining regions***	***Number of patients***
Log2 ratio below -0.7	744	84
Regions occurring in more than 3 patients removed	376	83
Regions occurring in patients with single gene diagnosis removed	228	49
Microdeletions occurring in polymorphic regions removed	97	22
Microdeletions identified in known disease regions	3	3
Microdeletions identified with evidence to support possible impact on growth (Phenotypes in DECIPHER or gene function)	2	2

In trios, remaining regions which were identified in the child but not in either parent (*de novo* deletions) were then examined in closer detail along with all possible remaining regions in singleton samples (parents not sequenced). Using DECIPHER, the phenotypes of all patients with CNVs overlapping these regions were reviewed. If growth abnormalities were frequently reported, then these regions were also prioritised for validation. This identified two possible causative deletions encompassing a similar region of (3q27) in two independent patients. If no overlapping cases were reported in either DGV or DECIPHER, then all genes within the region were reviewed to determine if they provided a possible link to growth either through the function of the encoded protein or through animal models as previously described (Section 3.3.6).

Finally, the possibility of autosomal recessive inheritance was explored by reviewing genes lying within potentially inherited regions. Such regions included those in common between affected siblings or those also found to be present in the parent of an affected child. Any gene with a potential functional link to growth was then reviewed for a second possible pathogenic variant (consequence score < 3) in the WES analysis described in Chapter 3. No candidate disease genes were identified from this autosomal recessive analysis.

In total, five potentially pathogenic microdeletions were identified with a minimum size of 3Mb. All five were subsequently validated by array-CGH.

6.3 CNVs identified encompassing known disease regions

6.3.1 Wolf-Hirschhorn syndrome

Two microdeletions located at the terminus of the short arm of chromosome 4 (4pter) were identified in two independent patients of 18.5Mb (4p16.3-p15.31 in patient 1) and 3.1Mb (4p16.3 in patient 2) (Figure 6.3A). Array-CGH performed on DNA from the affected patient and parents confirmed the CNVs to be *de novo* in both cases (Figure 6.3B).

The deletions both overlapped a region associated with Wolf-Hirschhorn syndrome (WHS) (Figure 6.4). WHS is a well characterised developmental disorder in which affected patients typically show distinctive facial features with a wide forehead, ocular hypertelorism and prominent glabella which is often described as a 'Greek helmet' appearance (Wilson *et al.*, 1981). Other features of this syndrome include pre- and post-natal growth retardation, developmental delay, congenital hypotonia, seizures, malformations (cardiac, brain and genito-urinary) and common variable immunodeficiency (Battaglia *et al.*, 2008). Growth failure is considered one of the cardinal features of WHS occurring in at least 75% of patients and growth parameters to the degree seen in MPD have been previously described (mean OFC and height measurements being -3.5 s.d. and -4 s.d. respectively) (Antonius *et al.*, 2008). Following identification of the microdeletion, on clinical review both patients were found to have subtle but characteristic facial appearance along with severe growth retardation in keeping with a diagnosis of WHS but without significant congenital malformations (Table 6.2).

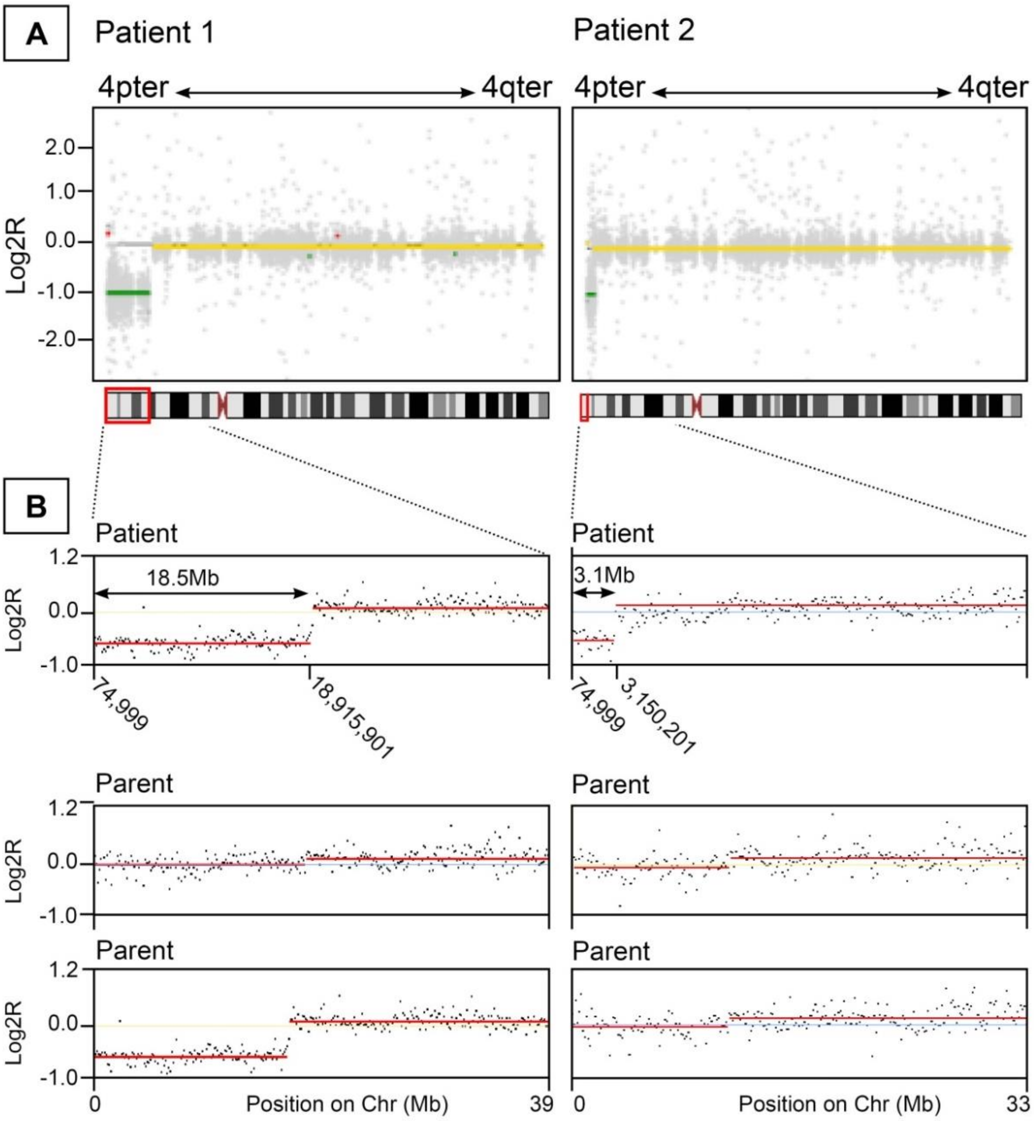


Figure 6.3. 4p terminal deletions identified in two patients following ExomeCNV analysis

A) Log2 ratios at each exon (Log2R) generated following ExomeCNV analysis (yellow line = copy neutral, red = amplification, green = deletion). Grey dots indicate log2R at exons with insufficient coverage to provide adequate power to call a CNV at that exon by itself. A two fold reduction in copy number (Log2R = -1, green line) was identified at the terminus of the short arm of chromosome 4 (4pter) in two independent patients indicating a heterozygous microdeletion of this region. Underlying schematic of chromosome 4 indicates region of microdeletion (red). B) Array-CGH confirmed the deletions to be *de novo* in the affected cases. Black dots indicate log2R at each probe. Red line indicates log2R across each segment (minimum of 5 consecutive probes). Genomic position of deletion as indicated corresponds to the hg19 reference genome.

Table 6.2. Clinical features of two patients identified with 4p terminal deletions

Pt	*CNV Size /Mb*	*Sex*	*Gest^n*	*Birth*		*Age at exam*	*OFC /s.d. (kg)*	*Height /s.d. (cm)*	*Reported clinical features*
				Weight /s.d. (kg)	*OFC /s.d. (cm)*				
1	18.5	M	37	-3.95 (1.34)	-4.94 (27)	16m	-7.5 (39.5)	-5.1 (66)	Hypospadias, undescended testis, prominent glabella, hypertelorism
2	3.1	F	31	-3.05 (0.81)	-2.72 (25.3)	17m	-4.6 (42.3)	-6.9 (60.6)	Febrile seizures, developmental delay, prominent eyes

Genomic coordinates are according to hg19 reference genome.

Two critical regions (WHSCR1 and WHSCR2) have been defined in WHS (Altherr *et al.*, 1997, Wright *et al.*, 1997, Zollino *et al.*, 2003) with two respective candidate genes proposed to be primarily responsible for the cardinal features of this syndrome including severe growth failure (Figure 6.4) (Stec *et al.*, 1998, Wright *et al.*, 1999). The two genes are Wolf-Hirschhorn syndrome candidate 1 (*WHSC1*) and negative elongation factor complex member A (*NELFA*). Both deletions identified in this Chapter encompassed both these genes. *WHSC1* encodes a protein ubiquitously expressed in early development (Stec *et al.*, 1998) and has been implicated in the DNA damage response following replicative stress (Hajdu *et al.*, 2011). Haploinsufficient mice also show growth retardation (Nimura *et al.*, 2009). *NELFA* (also referred to as *WHSC2*) shows a similar expression pattern to *WHSC1* and haploinsufficiency results in slow progression from S to M phase, reduced DNA replication and altered higher ordered chromatin assembly (Kerzendorfer *et al.*, 2012). Thus both genes have potential functions that may impact on the cell cycle and consequently affect growth. Reviewing heterozygous variants with a consequence score of 1 in *WHSC1* and *NELFA* in WES data from this cohort of MPD patients did not identify any likely pathogenic *de novo* mutations in these two genes.

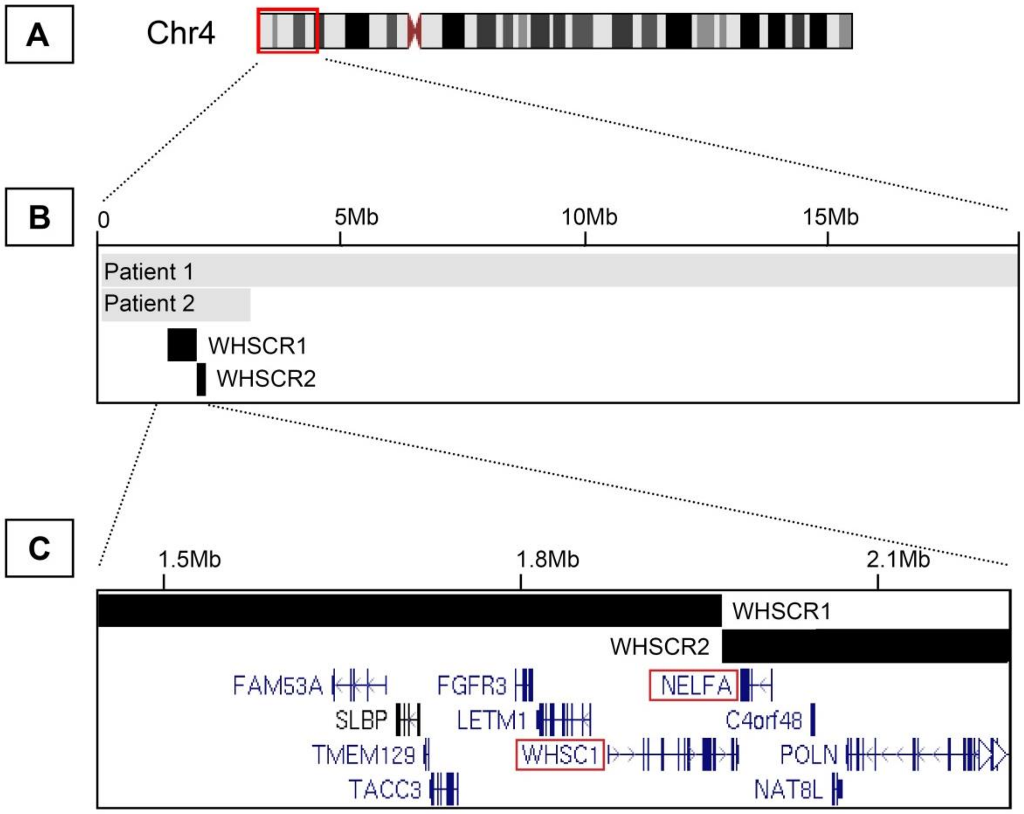


Figure 6.4. The identified 4p microdeletions overlap a critical region for growth in Wolf-Hirschhorn syndrome

A) Schematic of chromosome 4 indicating 18.1 Mb region deleted in patient 1. B) USCS genome browser graphical display showing the two regions deleted in MPD patients (grey) and the two previously defined critical regions (black) responsible for the cardinal features of WHS. C) This has led to the identification of two candidate genes, *WHSC1* and *NELFA* (also referred to as *WHSC2*) thought to play a critical role in the phenotype of WHS including growth failure. *Abbreviations: WHSCR=Wolf-Hirschhorn syndrome critical region.*

6.3.2 Chr15 deletion encompassing *IGF1R* identified in two affected siblings

One patient was identified with a 2.99 Mb heterozygous microdeletion in the long arm of chromosome 15 (15q26.1 to 15qter) which was validated by array-CGH (Figure 6.5). Deletion of the same region has been previously reported in patients with growth retardation, microcephaly and developmental delay (Rudaks *et al.*, 2011) whereas duplication of the same region is associated with tall stature (Kant *et al.*, 2007). The deletion includes 19 coding genes including *IGF1R* which encodes the insulin-like growth factor 1 receptor. IGF1R is a receptor tyrosine kinase which binds insulin-like growth factor leading to a range of cellular responses that promote cell proliferation and differentiation (Liu *et al.*, 1993, Worrall *et al.*, 2013). Deleterious heterozygous and compound heterozygous mutations in this gene have previously been identified in patients with MPD (Abuzzahab *et al.*, 2003, Walenkamp *et al.*, 2013) and therefore haploinsufficiency of this gene has been thought to be primarily responsible for the growth phenotype in this microdeletion syndrome. A literature review of the other genes in the deleted region did not identify any other biologically plausible candidates.

The patient identified here was male and reportedly showed growth restriction at birth (birth weight unavailable, birth length -3 s.d.) although birth OFC was normal (-0.6 s.d.). Growth failure was also evident post-natally (height and OFC both -3s.d. at 11 years of age) similar to patients previously described with microdeletions of this region (Rudaks *et al.*, 2011). Growth hormone deficiency was suspected following endocrine investigation of IGF1 and growth hormone levels and he was commenced on replacement therapy. Over four years of treatment his height increased by 1 s.d. (-4 s.d. prior to treatment). The patient also had a similarly affected sister with IUGR reported at birth (birth measurements unavailable) and postnatal growth failure (OFC -4 s.d. and height -6 s.d. at approximately four years of age). Additionally, this sibling had an atrial septal defect.

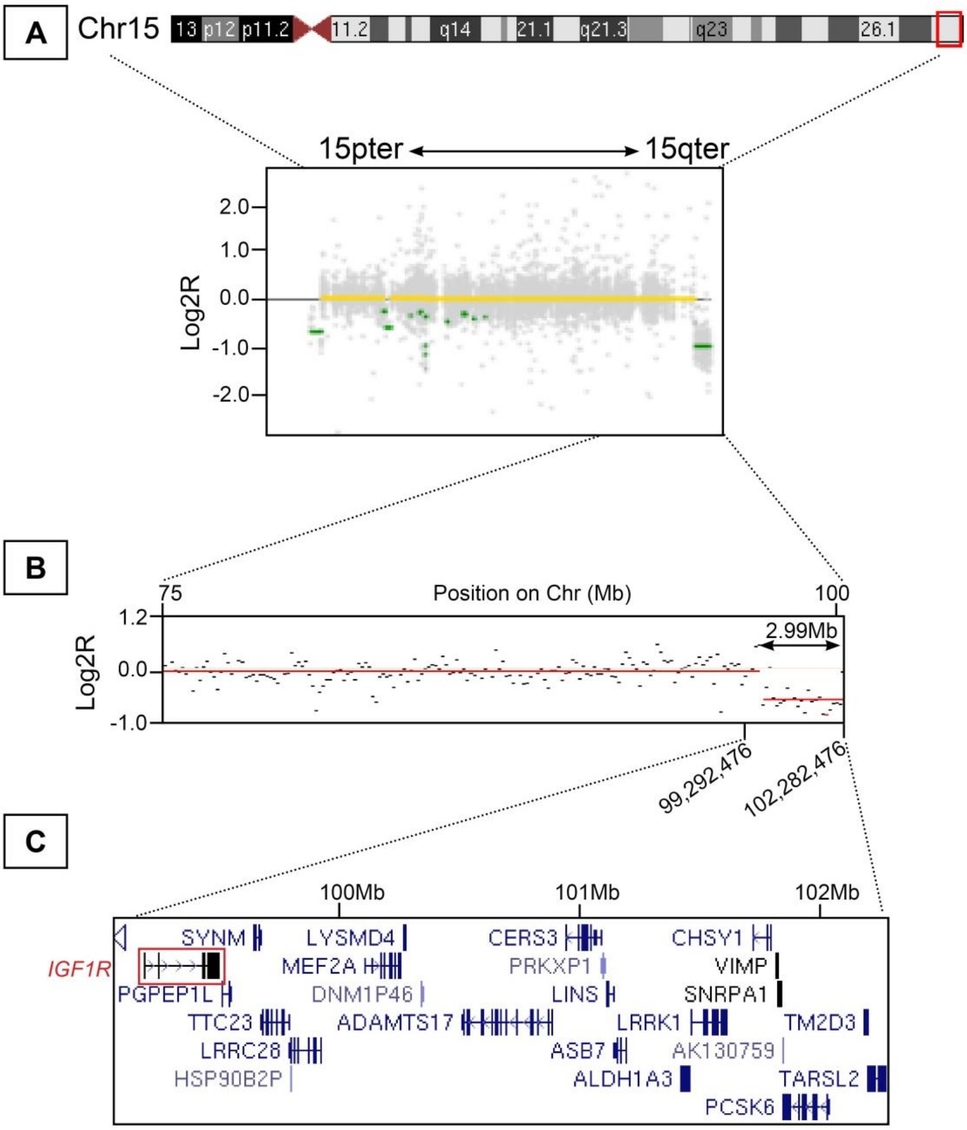


Figure 6.5. 15q microdeletion identified in one case encompassing *IGF1R* following ExomeCNV analysis

A) Schematic of chromosome 15 with deleted region highlighted in red. A deletion at the terminus of the long arm of chromosome 15 (15qter) was identified in one patient following ExomeCNV analysis as indicated by a Log2R of -1 at the end of chr15 (green line). Other possible deletions indicated in green were excluded during filtering. B) Array-CGH confirmed a 2.99 Mb deletion at chromosome 15 (q26.3). Genomic coordinates correspond to the hg19 reference genome. C) The deleted region contained 20 coding genes including *IGF1R* (red), haploinsufficiency of which has previously been associated with severe growth failure.

The deletion initially appeared to be *de novo* in the child as it was not identified in either parent (Figure 6.6 A, B). It was subsequently found to be present in both the affected siblings as well as a third sibling who was reportedly unaffected (Figure 6.6 C,D), suggesting instead the presence of gonadal mosaicism or more likely, a balanced structural rearrangement in one of the parents. Notably, growth parameters for the unaffected sibling were unavailable. Repeat examination of the third unaffected sibling with careful growth measurements would be prudent as they may have a mild phenotype which has previously been overlooked. Although uncommon, it is also important to be aware that sample mix up can occur and it would therefore be appropriate to reconfirm the results in a fresh DNA sample from the unaffected sibling and parents.

It was also considered whether further modifiers are present in the two affected siblings additionally impacting on growth. Capillary sequencing of *IGF1R* coding regions was undertaken in all five family members to ascertain whether a second mutation was present on the alternate allele. However, no likely pathogenic variant was identified. It is still possible that a variant in a noncoding region of the gene or a regulatory region may be impacting on gene expression which could be further investigated by cellular studies examining *IGF1R* transcript and protein levels. Alternatively, a variant in another disease gene could be compounding the haploinsufficiency of *IGF1R* in the two affected siblings although no other candidate variants were identified on WES in the proband.

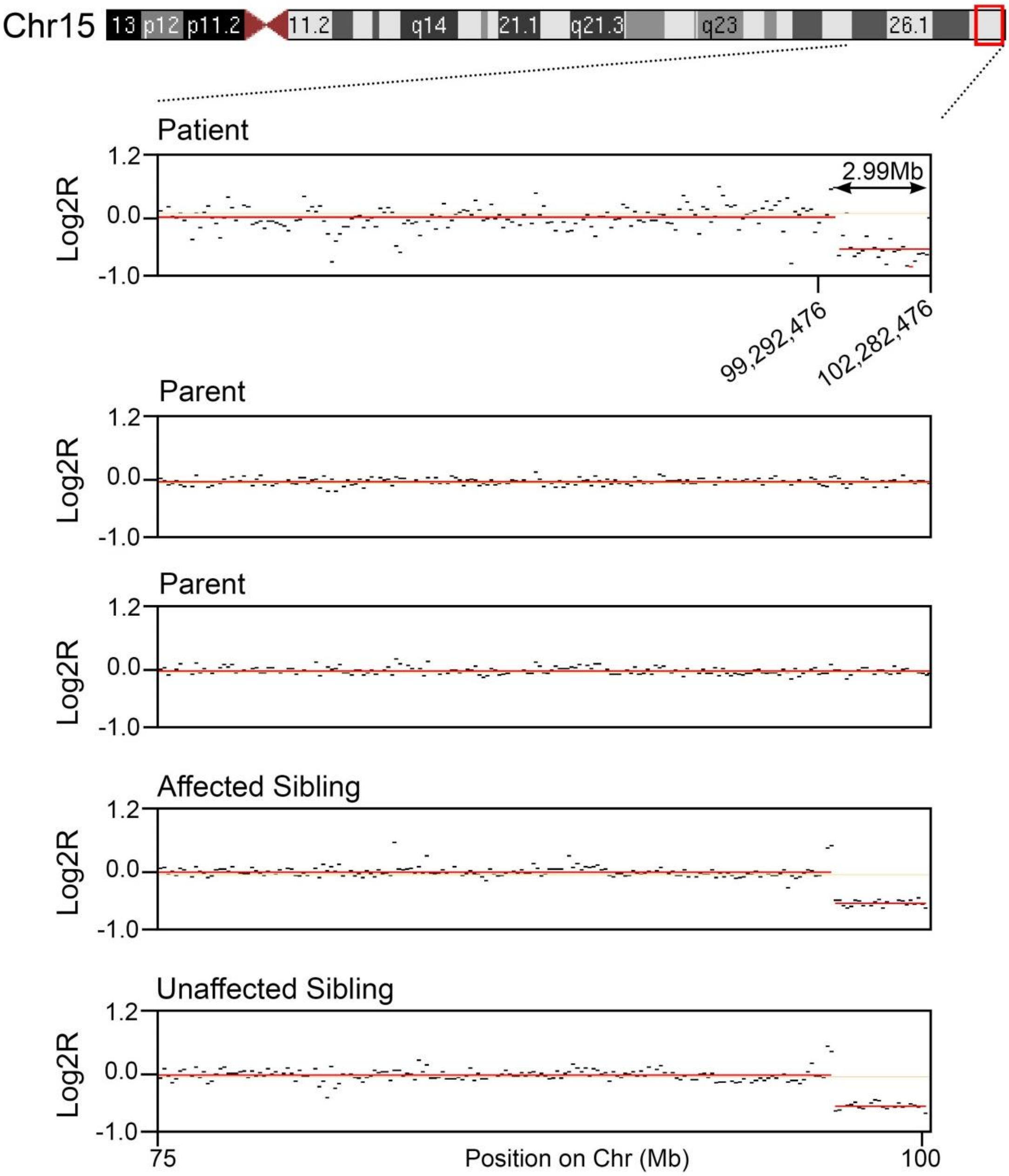


Figure 6.6. Array-CGH analysis of the 15q microdeletion in multiple family members

The 15qter deletion was not apparent in either parent following array-CGH but was present in both the affected and unaffected sibling. Genomic coordinates correspond to the hg19 reference genome.

6.4 Identification of a novel CNV as a cause of MPD

6.4.1 3q27 microdeletion

Two patients were identified with microdeletions encompassing a region on the long arm of chromosome 3 (3q27-3q28). Both deletions were validated by array-CGH (Figure 6.7). Parental DNA was only available in parents of one of the patients and this confirmed the deletion to be *de novo* in their affected child (patient 3). Both cases showed similar growth parameters with prenatal and postnatal growth failure although postnatally OFC was disproportionately reduced in comparison to height (Pt3 and 4, Table 6.3). Developmental delay was reported in both cases but otherwise there were no malformations or other associated features. Given this microdeletion was present in two of the 85 patients where ExomeCNV analysis was able to be performed, it could potentially represent a relatively common cause of MPD. Therefore array-CGH was performed in five further patients with similar phenotypes. This identified another patient with a similar, overlapping deletion at 3q27 (Pt 5) and array-CGH analysis of parental DNA confirmed this to be *de novo* in the child.

Table 6.3. Clinical features of three patients identified with 3q microdeletions

Pt	*Size /Mb*	*Sex*	*Gest^n*	*Birth*		*Age at exam*	*OFC /s.d. (kg)*	*Height /s.d. (cm)*	*Clinical Features*
				Weight /s.d. (kg)	*OFC /s.d. (cm)*				
3	3.77	F	41	-3.28 (2.17)	-3.34 (31)	4y5m	-7.42 (42.5)	-4.12 (87.3)	Small teeth, high pitched voice, moderate developmental delay
4	4.03	F	36	-4.28 (1.1)	-3.33 (28)	2y10m	-8.8 (39.5)	-4.6 (76.5)	Low set posteriorly rotated ears. Mild developmental delay
5	3.51	F	N/A	-2.3* (1.75)	N/A	9y5m	-8.1 (43.5)	-5.0 (105)	Small ears, high arched palate, triangular face, shallow orbits, delayed bone age, mild developmental delay

*assumed term gestation.

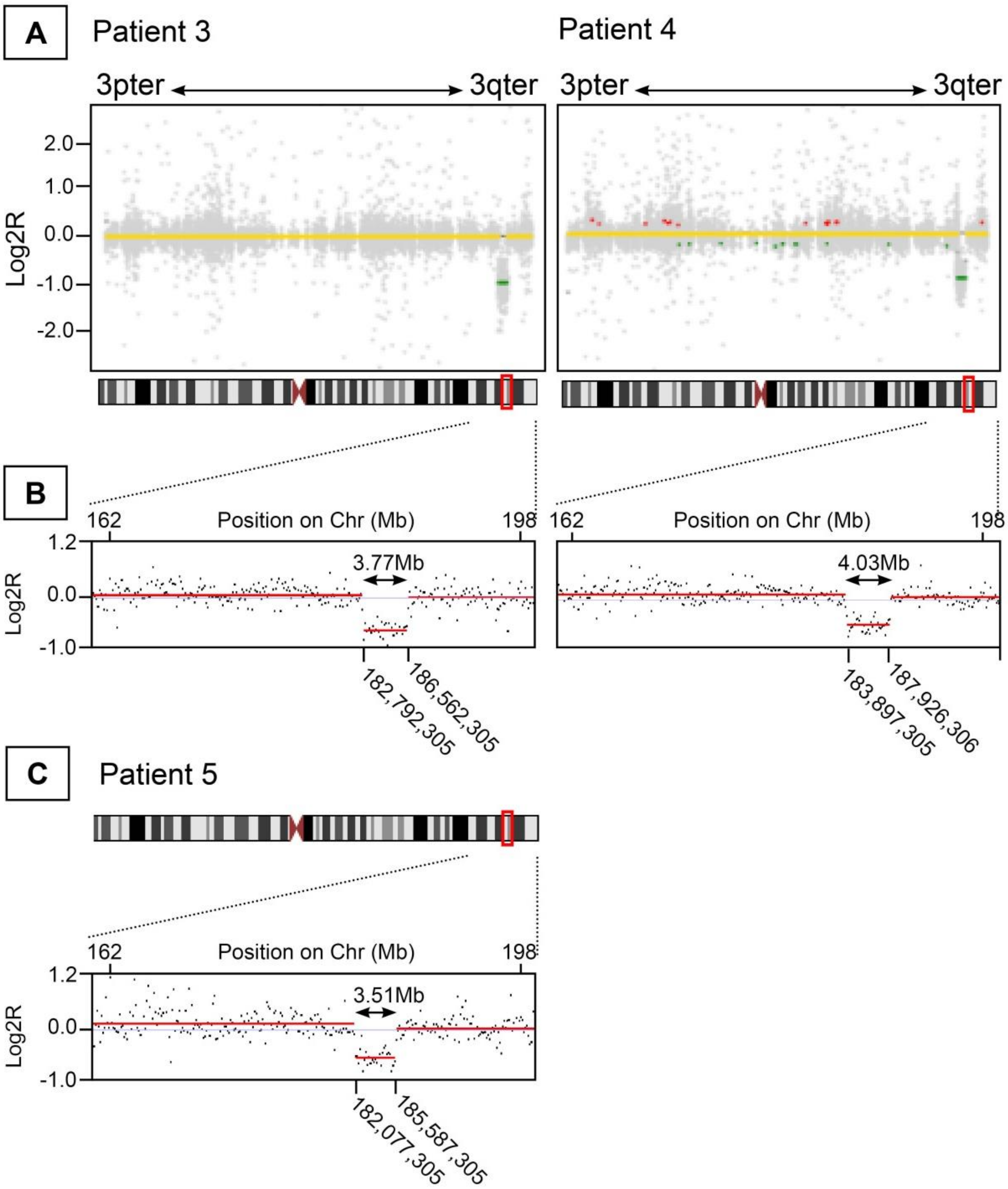


Figure 6.7. 3q27 microdeletions identified in three MPD patients

A) Log(2) ratios generated following ExomeCNV analysis identifies two patients with a two fold reduction in copy number (green line) within the long arm of chromosome 3(q27.1-q27.3) indicating a heterozygous microdeletion of this region. Underlying schematic of chromosome 3 indicates region of microdeletion (red). B) Validation by array-CGH confirmed both deletions. Genomic position of deletion corresponds to the hg19 reference genome. C) Array-CGH in patients with a similar phenotype identified one further case with a deletion overlapping the same region.

6.4.2 *IGF2BP2* is a novel candidate gene for growth restriction at the 3q27 locus

All three deletions shared a common region 1.7 Mb in size (chr3:183,897,305-185,587,305, hg19) encompassing 24 genes (Figure 6.8). Of all the candidate genes in this region (Table 6.4), *IGF2BP2* had the strongest supporting evidence for a functional role in regulating growth along with a low HI score. A brief summary of these genes is shown in Table 6.4 along with corresponding haploinsufficiency (HI) scores obtained from the DECIPHER database. HI scores are predictive, based on a classification model trained on the characteristics of genes known to be either haploinsufficient or sufficient from previous documentation of structural variation and associated disease (Huang *et al.*, 2010). A low percentage score (<10%) indicates the gene is more likely to be haploinsufficient.

IGF2BP2 encodes one of three IGF2 mRNA binding proteins, otherwise known as *IMP2,* which facilitates nuclear export, stability and translation of IGF2 mRNA. All three IGF2 mRNA binding proteins are ubiquitously expressed during embryogenesis (Nielsen *et al.*, 1999, Christiansen *et al.*, 2009). In humans, down regulation of *IGF2* through hypomethylation of a differentially methylated region results in Russell Silver syndrome, a short stature syndrome in which head size is typically normal. Conversely, up regulation of *IGF2* occurs in Beckwith-Wiedemann syndrome which is characterised by overgrowth (Nativio *et al.*, 2011). In addition, IGF2-deficient mice are noted to be small with microcephaly (Liu *et al.*, 1993). Therefore haploinsufficiency of *IGF2BP2* may impact on growth by destabilising IGF2 mRNA resulting in reduced protein levels. IGF2BP2 is phosphorylated by mammalian target of rapamycin (mTOR) (Dai *et al.*, 2011) and its expression is regulated by *HMGA2* (Brants *et al.*, 2004), a gene known to be strongly and reproducibly associated with height (Weedon *et al.*, 2007, Weedon *et al.*, 2008, Lango Allen *et al.*, 2010) and head circumference in genome wide association studies (Ikram *et al.*, 2012a, Taal *et al.*, 2012b). Additionally, *Hmga2* null mice show a severe reduction in body size described as a pygmy phenotype (Zhou *et al.*,

1995) and in humans haploinsufficiency of *HMGA2* is also associated with growth restriction as part of the 12q14 microdeletion (Mari *et al.*, 2009).

Therefore capillary sequencing of *IGF2BP2* was performed in 239 patients with MPD, short stature and microcephaly (OFC>-4 s.d., height -2 to -4 s.d.) or primary microcephaly (normal height). However, this did not identify any potentially pathogenic mutations in this cohort. Additionally, no likely pathogenic variants were detected in the remaining copy of *IGF2BP2* in the three patients with a 3q27 microdeletion suggesting that growth failure is not the result of biallelic abnormalities in *IGF2BP2.* Considering at least two of the deletions are *de novo* it is therefore most likely that gene haploinsufficiency is responsible for the observed growth phenotype.

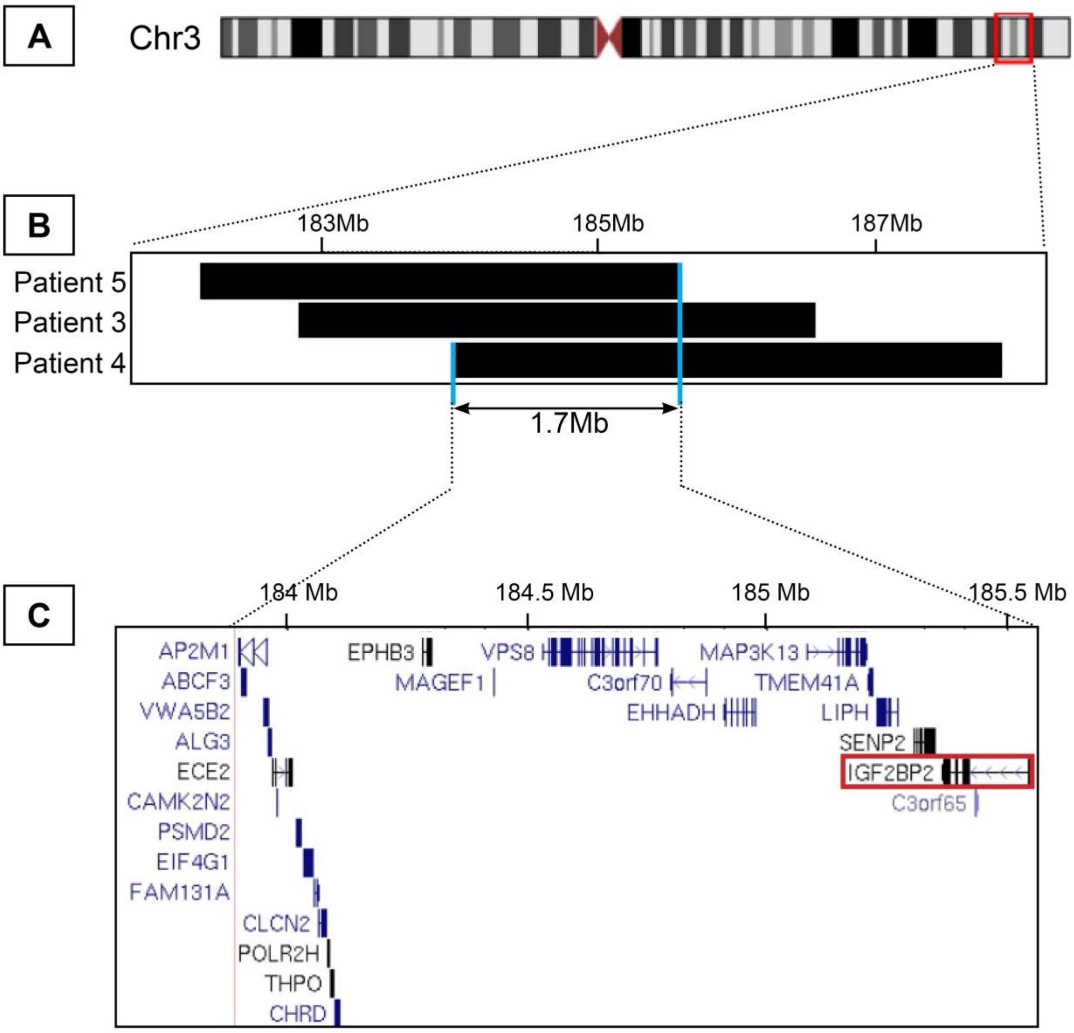


Figure 6.8. 3q27 microdeletions identified in three MPD patients

A) Schematic of chromosome 3 with region displayed in B highlighted in red. B) The three patients identified in this study with 3q27.3 microdeletions shared a common region 1.7 Mb in size (chr3:183,897,305-185,587,305, hg19). C) This region encompassed 24 genes including *IGF2BP2* which has a strong functional link to growth and a low haploinsufficiency (HI) score highlighting it as a strong candidate gene for the growth phenotype in these patients.

Table 6.4. Genes within 3q27 region deleted in three MPD patients

Gene	*Full Name*	*HI Score %*	*Function*
SENP2	SUMO1/sentrin/SMT3 specific peptidase 2	21	Processes SUMO1, a ubiquitin-like protein involved in nuclear transport, transcriptional regulation, apoptosis, and protein stability
IGF2BP2	Insulin-like growth factor 2 mRNA binding protein 2	7	Regulates IGF2 translation
C3orf65	Chromosome 3 open reading frame 65	40	Unknown
LIPH	Lipase, member H	88	Catalyses the production of the lipid mediator LPA. Recessive mutations cause hypotrichosis (Woolly hair syndrome).
TMEM41A	Transmembrane protein 41A	60	Unknown
C3orf70	Chromosome 3 open reading frame 70	N/A	Unknown
MAP3K13	Mitogen-activated protein kinase kinase kinase 13	37.6	Possible role in the JNK signalling pathway mediating neuronal apoptosis
EHHADH	Enoyl-CoA, hydratase/3-hydroxyacyl CoA dehydrogenase	92.2	Peroxisomal beta-oxidation pathway
VPS8	Vacuolar protein sorting 8 homolog (S. cerevisiae)	68.9	Endosomal trafficking in yeast
EPHB3	EPH receptor B3	22.5	Ephrin receptor
MAGEF1	Melanoma antigen family F, 1	93.7	Member of MAGE superfamily, function unknown
CHRD	Chordin	38.1	Dorsalizes early vertebrate embryonic tissues
THPO	Thrombopoietin	52.1	Platelet production. Mutations cause thrombocytopenia 1,
POLR2H	Polymerase (RNA) II (DNA directed) polypeptide H	78.5	RNA II polymerase subunit
CLCN2	Chloride channel, voltage-sensitive 2	51.7	Voltage gated cholride channel
FAM131A	Family with sequence similarity 131, member A	26.9	Unknown
EIF4G1	Eukaryotic translation initiation factor 4 gamma, 1	4.1	Initiation of protein synthesis by recruiting mRNA to the ribosome
PSMD2	Proteasome 26S subunit, non-ATPase, 2	2	Processes class I MHC peptides, TNF signalling pathway

Gene	***Full Name***	***HI Score %***	***Function***
CAMK2N2	Calcium/calmodulin-dependent protein kinase II inhibitor 2	31.5	Regulates neuronal synaptic plasticity
ECE2	endothelin converting enzyme 2	56.4	Endothelin 1 production and processing of neuroendocrine peptides
ALG3	ALG3, alpha-1,3-mannosyltransferase	44.7	N-glycosylation
VWA5B2	von Willebrand factor A domain containing 5B2	N/A	Unknown
ABCF3	ATP-binding cassette, sub-family F (GCN20), member 3	14.2	ABC transporter superfamily, function unknown
AP2M1	adaptor-related protein complex 2, mu 1 subunit	44.9	Acidification of endosomes and lysosomes

6.4.3 Variable growth in other reported patients with 3q27 deletions argues against *IGF2BP2* haploinsufficiency as a cause of MPD

Clinical information concerning other patients with overlapping *de novo* microdeletions to further refine the region responsible for reduced growth were obtained from DECIPHER (http://decipher.sanger.ac.uk, accessed 24/03/12) and by literature review. In total, 16 patients with overlapping *de novo* microdeletions were identified, 12 in DECIPHER and an additional four from the literature (Figure 6.9B). Growth parameters were available in nine cases which, for DECIPHER patients, were obtained directly from the listed clinician if not previously published. Six patients had microdeletions which spanned the entire 1.7 Mb critical region defined by the three patients identified in this study (Figure 6.9B, purple and black respectively). Although none of these patients had growth failure which fell into the defined range of MPD, microcephaly was reported in all six patients and two had additional short stature (height -3 s.d. to -4 s.d.). Ten patients were reported with deletions which only partially overlapped the region and, where growth parameters were available, this was associated with a significantly milder reduction in OFC (mean OFC -2.6 s.d. +/- 1.5 s.d., p=0.007, n=5) in comparison to those with deletions encompassing the entire 1.7 Mb region (mean OFC -6.3 s.d. +/- 2 s.d., n=7). No significant difference was observed in height between the two groups (mean height of patients with 1.7 Mb deletion; -3.32 s.d. +/- 1.7 s.d., n=7, mean height of patients with partial deletion; -1.68 s.d. +/-1 s.d., n=5, p=0.08). This suggest that complete deletion of this region is associated with growth failure which ranges from isolated microcephaly to MPD and that the three patients identified in this study may lie at the extreme end of this spectrum.

Of the nine patients with deletions encompassing *IGF2BP2*, six had microcephaly (OFC < -3 s.d.) with or without short stature, however two patients showed only a very mild reduction in growth parameters (OFC -2 s.d., height 0 to -2 s.d.) and one patient was described as disproportionately tall (patient 277546). Additionally, no growth phenotype was reported in three DECIPHER patients with partial *IGF2BP2* deletions despite other phenotype data being recorded and therefore it is probable that growth is unremarkable in these cases. These latter patients argue against haploinsufficiency of *IGF2BP2,* in isolation, as critical for normal growth and suggest that other genetic modifiers, possibly *in cis* with *IGF2BP2,* are contributing to the growth phenotype.

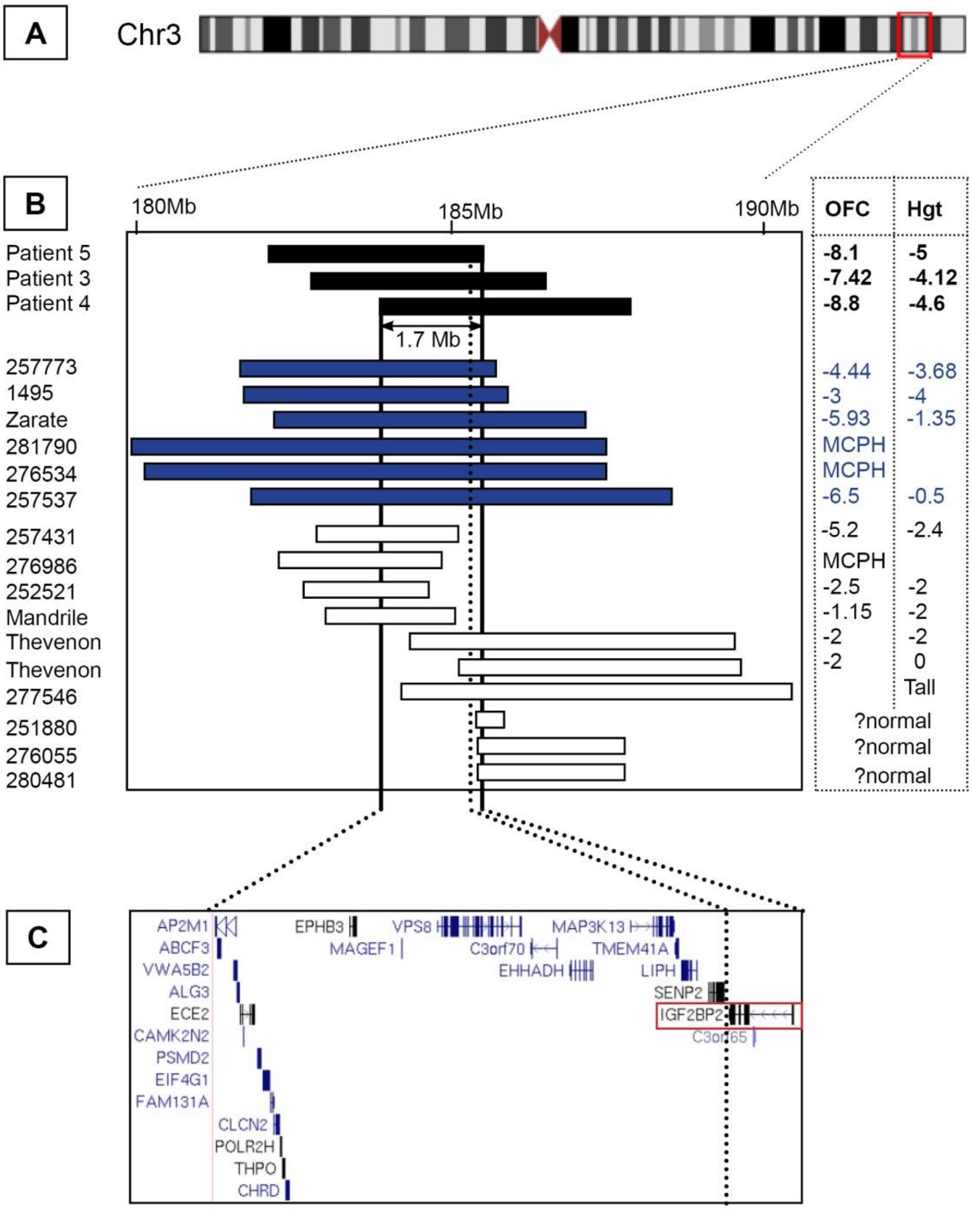


Figure 6.9. Additional patients reported with 3q27 microdeletions and associated growth phenotype

A) Schematic of chromosome 3 indicating region shown in B (red). B) Reported patients with microdeletions overlapping the entire 1.7 Mb region (purple) shared by the three MPD patients identified in this study (black) and reported patients with partial deletions of this region (white). Left hand annotation indicates DECIPHER identification number or reference (Mandrile *et al.*, 2013, Zarate *et al.*, 2013, Thevenon *et al.*, 2014). Available growth data of each patient shown on the right. Patients with the full 1.7 Mb deletion showed a significant reduction in OFC compared to patients with partial deletions (p=0.007, unpaired t-test). Growth was within normal limits in at least three of the twelve additional patients with deletions encompassing *IGF2BP2* (C, red box) suggesting haploinsufficiency of this gene is not primarily responsible for the growth failure observed in patients 3-5.

6.6 Discussion

6.6.1 Microdeletions are a relatively frequent cause of MPD

In this Chapter, microdeletions were identified in 5% of families in whom WES was performed in comparison to single gene mutations which were identified in 33% (Chapter 3). Extending this to all MPD patients in the cohort (including those in which a diagnosis was made prior to WES), alteration in gene copy number due to microdeletions is responsible for approximately 2.5% of all cases and is therefore as common as mutations in NHEJ genes (Chapter 4). Pathogenic microdeletions are frequently associated with microcephaly and/or short stature (Nevado *et al.*, 2014) and thus several microdeletion syndromes could potentially be associated with a growth phenotype that falls into the range of MPD (Section 1.1.4.5.1). This analysis enabled the detection of microdeletions covering known disease regions which have previously been associated with severe growth failure, Wolf-Hirschhorn syndrome and the *IGF1R* locus (Antonius *et al.*, 2008, Walenkamp *et al.*, 2013). However, this analysis also identified a previously uncharacterised microdeletion on chromosome 3q27 in three independent MPD patients.

6.6.2 3q27 is a novel microdeletion syndrome associated with MPD

The 3q27 deletion detected in these three cases allowed a novel locus for the growth phenotype to be defined which contained a strong candidate disease gene, *IGF2BP2*. However, no cases were identified with point mutations in this gene on resequencing over 200 patients with growth deficiency. Furthermore, other cases out with this study have since been reported with microdeletions encompassing *IGF2BP2* and relatively normal growth (Thevenon *et al.*, 2014). This suggests that haploinsufficiency of another gene in this region may be responsible for the phenotype. Alternatively, the growth phenotype is not exclusively caused by the loss of a single gene and rather results from the loss of several contiguous genes and/or regulatory regions. This is supported by the compilation of growth data from other reported cases with overlapping deletions (Figure 6.9) which showed a significantly greater reduction in OFC in those with a complete deletion of the 1.7 Mb region compared to those with only a partial loss of this region. Other deleted genes which may contribute to the growth failure include the Mitogen-activated protein kinase kinase kinase, *MAP3K13*, which may function in c-Jun N-terminal kinase (JNK) signalling, a pathway regulating neuronal apoptosis (Ikeda *et al.*, 2001, Xu *et al.*, 2001) as well as the SUMO-specific protease, SENP2, which deconjugates sumolated proteins including the p53 regulator, MDM2 (Jiang *et al.*, 2011). SENP2 deficient mice are embryonically lethal with impaired cell cycle progression in trophoblastic stem cells and defective placental development (Chiu *et al.*, 2008). SENP2 depletion in human fibroblasts has also been shown to induce p-53 dependent premature cell senescence (Yates *et al.*, 2008) and thus conceivably impacts on global cellularity in development. Additionally, the functions of seven of the deleted genes are largely unknown and so also remain potential growth modulating candidates (Table 6.4).

6.6.3 Difficulties in CNV detection from WES

In contrast to WGS, capture bias and discontinuity in coverage are inherent to WES and need to be taken into account during copy number analysis. At least14 tools now exist for CNV detection from WES (Tan *et al.*, 2014) and all follow a four step approach; mapping of reads to the reference, normalization to correct for biases in read depth resulting from variable capture efficiencies, estimation of copy number variation and segmentation to detect discordancy in copy number between adjacent regions (Zhao *et al.*, 2013). However, comparison of different CNV detection programmes highlights the difficulties with WES with poor concordance between different tools and large sample to sample variability occurring within each programme (Samarakoon *et al.*, 2014, Tan *et al.*, 2014).

ExomeCNV is one such tool primarily designed to compare a test sample to a single paired control matched in terms of capture method and sequencing platform. However the results are consequently highly dependent on the depth of coverage in the one reference sample as it assumes that the read density across each exon will be the same in both test and control samples in copy neutral regions. Given the large sample to sample variability in sequencing and depth of coverage identified in this study (Section 3.2.2) this method is likely to incur a high false positive rate. Similar difficulties were experienced by Samarakoon *et al.*, (2014) who found the number of microdeletions called per individual ranged from zero to over 2,000 following analysis of 30 different exomes using ExomeCNV. To limit errors occurring from sample variation both case and control should be prepared in the same batch and sequenced in the same run. The importance of this was noted in this study as all samples failed analysis where the affected was compared to a single parental sample prepared and sequenced at a different location.

The authors of ExomeCNV did report a greater precision in CNV detection following pooling independent samples to create a reference thus reducing the variance in read depth between different samples (Sathirapongsasuti *et al.*, 2011). Consequently this method was implemented for singleton cases in this study where parental sequencing data was not available. Combining samples to create a reference is a method used by many of the other programmes which have been designed specifically for the detection of CNVs in large cohorts (Fromer *et al.*, 2012, Krumm *et al.*, 2012) with improved consistency between samples compared to ExomeCNV (Samarakoon *et al.*, 2014). Improved accuracy in CNV detection may also be achieved by combining the results of multiple programmes (Zhao *et al.*, 2013, Samarakoon *et al.*, 2014).

Another difficulty with CNV detection in WES data is achieving high resolution with many tools performing poorly at the single exon level (de Ligt *et al.*, 2013). In this study at least 2/3^{rds} of deletions called in this analysis with a log2R below -0.7 covered less than 50 Kb and 43% covered less than 10 Kb. This is similar to previous observations in which structural variants of less than 10 Kb were found to be the most abundant in control populations (Mills *et al.*, 2011b). All deletions less than 1 Mb were either excluded during the filtering pipeline or considered unlikely to be pathogenic on manual review. In contrast 50% of deletions covering over 1Mb of sequence were identified as likely causative deletions and subsequently validated on array-CGH indicating reasonable specificity at this level as well as fewer polymorphisms occurring in this size range. Additionally, single exon deletions of less than 500 bp only represented 2.3% of deletions which is much lower than that anticipated from WGS studies (Mills *et al.*, 2011b) suggesting a lower sensitivity in CNV detection at this level. As so few of the called deletions were validated it is not possible to draw firm conclusions regarding the sensitivity and specificity of ExomeCNV for different sized deletions. However, other studies have noted improved sensitivity in detecting smaller CNVs (1-4 exons) with programmes using other strategies such as ExomeCopy (Samarakoon *et al.*, 2014) which utilises read count normalisation followed by a Hidden Markov Model based approach to identify CNVs (Love *et al.*, 2011).

Other studies reporting on the detection of CNVs from WES have either focused on the identification of previously validated CNVs (de Ligt *et al.*, 2013), identified CNVs in known disease genes (Bademci *et al.*, 2014, Lieber *et al.*, 2014) or estimation of CNV frequency within a patient cohort (Poultney *et al.*, 2013). As yet, the identification of novel causative deletions from WES has not been described and thus filtering strategies for handling such a large number of false positive calls have not been reported. In this study a pragmatic approach was taken to filtering out noise by excluding commonly occurring regions with identical genomic coordinates. However, this would not be an advisable approach in more homogenous disorders as the likelihood of removing pathogenic deletions would be much greater. An alternative approach would be to use larger independent datasets captured and sequenced using the same method although, as discussed above, different tools are likely to be more suitable in such an analysis.

6.6.4 Alternative approaches to CNV detection

The discontinuity in coverage arising from selected capture precludes the accurate identification of CNV breakpoints. As these could occur anywhere in the region between adjacent exons the distance between the CNV boundary identified from WES analysis and the actual breakpoint could range from as little as 125 bp to over 22 Mb in size (Sathirapongsasuti *et al.*, 2011). CNV analysis from WGS circumvents many of the difficulties arising from WES by providing more even coverage and continuity in read distribution and thus is able to achieve higher resolution with lower false positive rates along with accurate breakpoint detection (Duan *et al.*, 2013). However these advantages need to be weighed against the increased sequencing costs incurred by WGS.

Array-CGH represents the gold standard for diagnostic CNV detection and high resolution arrays can detect deletions as small as 200 bp (Urban *et al.*, 2006) although such genome wide arrays are costly. Alternative strategies such as MAPH (Mulitplex Amplifiable Probe Hybridisation) (Armour *et al.*, 2000), MLPA (Mulitplex Ligation-dependent Probe Amplification) (Schouten *et al.*, 2002), qPCR (Babashah *et al.*, 2009) and digital droplet PCR (Mazaika *et al.*, 2014) could present a more cost-effective option to validate small deletions (<100 Kb) in specific genes. These latter methods could also be used to screen a large number of patients for intragenic deletions in a small number of candidate genes which have been identified through the detection of SNVs/indels or reside within a larger CNV region such as *IGF2BP2* or *WHSC1*.

6.6.5 Conclusions

Despite the limitations of CNV detection in WES, it was still possible to detect large CNVs (greater than 1Mb) in this cohort of MPD patients using ExomeCNV, establishing a diagnosis in a further five families. Given the frequency of large CNVs in the MPD cohort, performing diagnostic array-CGH prior to embarking on large scale NGS in the future should be considered. Notably, this analysis does not include the analysis of microduplication in the WES dataset and will not detect balanced structural rearrangements which could also be disease causing. With decreasing costs, WGS may provide a more cost-effective approach to WES in the future by combining mutation discovery in both coding and non-coding regions as well as greater sensitivity in CNV detection (Gilissen *et al.*, 2014) with the added benefit of detecting breakpoints and chromosomal rearrangements.

This Chapter demonstrates that WES can be used to successfully identify causal *de novo* CNVs in developmental disorders although this method is highly error prone and lacks sensitivity in the detection of small intragenic deletions. Although a potential strategy for novel gene discovery, identification of possible pathogenic deletions was biased in this study by prioritising regions encompassing genes with a known functional link to growth. This was required due to the high level of noise in the dataset which rendered validation of all CNVs infeasible despite filtering. Given the relative frequency in which pathogenic CNVs have been identified despite suboptimal methodology, improved resolution in CNV detection either by improved WES analysis or WGS will likely increase the identification of pathogenic CNVs in this cohort.

Chapter 7: Summary of thesis findings and future plans

7.1 Improving molecular diagnosis in MPD patients

7.1.1 Outcome and observations of WES analysis

Performing WES in combination with a comprehensive filtering pipeline to detect both point mutations and copy number variation resulted in the identification of a molecular diagnosis in 39 families (41% of those sequenced). As expected, a large degree of genetic heterogeneity was observed with mutations being identified in 25 different genes (Chapter 3) as well as three distinct microdeletion syndromes (Chapter 6). In comparison a diagnostic yield of only 16% was achieved following WES of another large heterogenous cohort with severe intellectual disability (de Ligt *et al.*, 2012), although CNV analysis was not undertaken. The majority of molecular diagnoses made in this study were in known disease genes/regions demonstrating a large degree of overlap between MPD and other previously characterised developmental disorders expanding the phenotypic spectrum of these conditions. Additionally in this study, WES has led to the discovery of several novel disease causing genes including *XRCC4*, *NCAPD3*, *NCAPD2*, *NCAPH*, *DONSON* and possibly *FAT1* (Section 3.6). This demonstrates the power of WES in enabling the rapid identification of causal variants in a rare, heterogeneous disorder which were unlikely to be detected through previous gene finding methods (Section 1.4).

However, although WES has enabled sequencing of the majority of the exome it is by no means fully comprehensive and in 59% of families no candidate disease-causing variants were identified. Difficulties in WES experienced in this and other studies (Ng *et al.*, 2009, Mamanova *et al.*, 2010) include variation in capture efficiency between different samples which impacts on sequencing volume and consequently depth of coverage (Section 3.2.2) (Sathirapongsasuti *et al.*, 2011, Chilamakuri *et al.*, 2014). Consistency may be further optimised by ensuring the same methodology is used and that matched samples, such as affected and parental samples, are prepared and sequenced in the same batch. However, despite this marked variability can still occur which will impact on downstream filtering particularly when identifying *de novo* or biallelic variants in the analysis of trios and quads (Section 3.7.5). The second issue is dealing with the high number of false positives variants that result from sequencing and mapping errors (Section 3.7.3) (DePristo *et al.,* 2011). Although a pragmatic approach was taken in this study to filter out commonly occurring variants, this would not be desirable in a more homogeneous cohort where multiple patients may share a common pathogenic variant. Removing this step would considerably increase the number of false positive variants remaining following filtering. Additionally, although filtering out common polymorphisms and ‘ benign’ variants reduced the total number of variants by over 90% (Section 3.3.5), large variability in remaining variant number was still present between individual samples. This reflected the patients’ countries of origin (Section 3.4) as limited availability of large scale sequencing data in African, Middle-Eastern, sub-Indian continent and Asian populations prevented the use of ethnically matched controls in filtering (Section 3.7.4).

Finally, it is important to be aware of the limitations of NGS technology with the identification of indels being particularly problematic (Chilamakuri *et al.*, 2014). Notably, there is a significant difference in the performance of different technologies with semi-conductor sequencers having substantially reduced performance in indel detection in homopolymer runs (Harris *et al.*, 2008, Liu *et al.*, 2012). This was observed in Chapter 4 in which indel mutations in *XRCC4* were initially missed in three patients following sequencing using the Ion Proton/Torrent™ platforms and only subsequently identified on capillary resequencing (Section 4.2.7). Consideration of such limitations during experimental design is therefore crucial to maximise the mutation finding potential of WES.

7.1.2 Implications for the future application of WES in clinical practice

This study demonstrates the advantage of WES in uncovering unexpected diagnosis and the potential benefits of WES in the clinical diagnostic setting. Some disease genes, for example *VPS13B* which causes Cohen syndrome (Kolehmainen *et al.*, 2003), are very large and therefore expensive to sequence by capillary methods.

Furthermore the phenotype associated with mutations in this gene can be very variable (Seifert *et al.*, 2006) making a clinical diagnosis difficult. WES provides a cost effective way to sequence a large number of plausible disease genes regardless of size and detect causal mutations in genes which are not initially clinically apparent.

Despite the huge expansion in gene discovery following the advent of NGS technology, WES has yet to become a mainstay investigation in clinical diagnostic genetics in the UK and will be challenging to implement, not least because of the costly bioinformatics skill, processing power and storage capabilities needed to handle the large amount of data generated from NGS technologies. There are also ethical considerations such as the inadvertent uncovering of disease related variants which are not part of the condition being investigated, an issue highlighted in the Miller syndrome study by Ng *et al*, 2010a which also identified a mutation in *DNAH5* in one family establishing a second diagnosis of primary ciliary dyskinesia. Although a likely advantageous discovery for this patient, identifying variants which confer susceptibility to certain conditions such as malignancy (e.g. BRACA1/2 mutations) or predicate the onset of disease (e.g. Alzheimer’s or Parkinson’s disease) may not always be welcomed by that individual or family (van El *et al.*, 2013). Therefore, in the clinical setting, limiting the review of variants to only those occurring in known disease genes relevant to the condition being investigated may circumvent such issues. However, the American Society of Medical Genetics also recommends that incidental findings which potentially provide significant medical benefit for the patient should be reported, for example pathogenic mutations in genes known to cause hypertrophic obstructive cardiomyopathy or familial hypercholesterolemia (Green *et al.*, 2013).

Finally, there also lies the inherent challenge of assigning pathogenicity to previously unclassified variants especially relevant where prenatal diagnosis or pre-implantation genetic diagnosis is being sought by the family. Although several freely available databases detailing variants discovered within each gene exist, for example the human gene mutation database (Stenson *et al.*, 2014) or the Leiden open variation database (Fokkema *et al.*, 2011), one single comprehensive database would clearly be desirable. Additionally, to avoid misleading other clinicians it is critical that variants entered into such databases are correctly annotated, for example ‘implicated’ rather than ‘pathogenic’ is perhaps preferable terminology to classify novel variants lacking further genetic evidence and/or supportive functional studies (MacArthur *et al.*, 2014). Such databases of collated variants will be a powerful resource for determining criteria on which to sensitively classify previously unassigned variants as performed recently in a large cohort of cystic fibrosis patients and variants in the *CFTR* gene (Sosnay *et al.*, 2013).

7.1.3 Implications for clinical diagnostics and patient management

Identifying a molecular diagnosis is often of vital importance to the patient and their family. They may desire an explanation for why they or their child is affected in such a way and also whether this is likely to occur again in subsequent pregnancies. Uncovering the molecular cause can also assist clinical management and potentially improve outcomes. This has been the case in MOPDII where annual screening for cerebral vascular malformations in patients lacking PCNT has enabled early intervention (Perry *et al.*, 2013). Additionally, in this study, the discovery of *LIG4* mutations resulted in the detection of previously unrecognised immune dysfunction in patients which is now being monitored and managed appropriately (Section 4.2.2.3). In several cases the diagnosis has also expedited the consideration of bone marrow transplantation enabling early intervention and tailoring of conditioning regimes which is associated with improved outcomes in such patients (O'Driscoll *et al.*, 2008). These issues also need to be taken into consideration in the management of *XRCC4* patients who are likely to share similar characteristics. Continued follow up and data collection of both *LIG4* and *XRCC4* patients will further improve the characterisation of associated phenotypes as well as inform on prognosis and malignancy risk hopefully leading to increased diagnosis and improved management in these disorders.

As well as implications for clinical management following molecular diagnosis, findings from this study can also be used to inform on the first line investigation of patients presenting to the clinic with severe growth failure. Given the relatively high incidence of *LIG4/XRCC4* mutations in the MPD cohort, full blood count, immunoglobulin levels and T and B cell subset counts are simple clinical investigations in addition to baseline endocrinological testing in newly present primordial dwarfism patients. In addition, following the discovery of causative microdeletions in MPD patients in this study (Chapter 6), diagnostic array-CGH should also be considered even in the absence of significant malformations or intellectual disability. It would also be important to carefully consider DNA damage repair disorders during examination which were identified in 10% of those sequenced (Section 3.5.2). Features which may indicate the presence of such include microcephaly disproportionately reduced in comparison to height, abnormalities in skin pigmentation, sun-sensitivity, malignancy and immunodeficiency (Weemaes, 2000, Kaneko *et al.*, 2004, Matsumoto *et al.*, 2011, Murray *et al.*, 2014).

7.1.4 Identifying a molecular diagnosis in remaining patients

As the cost of sequencing continues to decline there is an expectation that WES will be superseded by WGS which has the advantages of improved uniformity in coverage by avoiding capture bias, improved CNV detection including determination of breakpoints and sequencing of non-coding genes and regulatory regions (Duan *et al.*, 2013, Gilissen *et al.*, 2014). Although interpreting non-coding variants is challenging, focusing on coding variants alone along with CNV analysis in the first instance may still result in improved diagnostic yield. This was recently demonstrated in a cohort of patients with severe intellectual disability in which WGS following WES and microarray based CNV analysis achieved a diagnosis in a further 42% of patients (Gilissen *et al.*, 2014). Furthermore, the generation of larger WGS datasets in control populations, for example, the 100,000 genomes project (http://www.genomicsengland.co.uk/the-100000-genomes-project/) will assist in the filtering of commonly occurring variants. Other large scale projects are also underway sequencing specific subpopulations such as the Born in Bradford study (http://www.borninbradford.nhs.uk/) and this will allow improved matching of case and control populations in the future. Interpreting the functional consequence of variants out with the exome is also likely to improve with the development of specific prediction algorithms such as Combined Annotation-Dependent Depletion (CADD) (Kircher *et al.*, 2014). Finally, whole genome genotyping of SNVs, either by WGS or SNP array, will enable regions of homozygosity to be more accurately defined which can assist in narrowing down candidate variants in consanguineous families (Kaasinen *et al.*, 2014).

Notably, a substantial reduction in variant number was achieved when sequencing parents in addition to the affected (Section 3.3.5). Although at significantly increased cost, the benefits of this approach are likely to be particularly important for WGS given that each individual genome harbours over 3 million SNVs, at least 100 times more than in the exome alone (Wheeler *et al.*, 2008, Gilissen *et al.*, 2014). Sequencing trios also has the advantage of allowing identification of *de novo* CNVs. This will be particularly beneficial in prioritising candidate disease regions in the smaller size range (under 10 Kb) in which the majority of CNVs occur (Mills *et al.*, 2011b). Another lower cost alternative to WGS is to perform transcriptome sequencing in the remaining patients without a diagnosis which could assist in the identification of deleterious variants previously removed on filtering. However variant detection is highly dependent on expression levels of individual genes (Cirulli *et al.*, 2010) and thus developmentally relevant genes expressed during embryogenesis may not be well represented in an analysis performed in blood or tissue obtained in late childhood or early adulthood.

Lastly, structural variation has also not been fully investigated in this cohort and this could be a significant source of pathogenicity in MPD following the identification of causal microdeletions in six families, 5% of those sequenced (Chapter 6). In particular, gains in copy number identified by ExomeCNV analysis have not been reviewed and the presence of balanced structural rearrangements has not been investigated. In the future it would be worthwhile repeating the CNV analysis using another programme(s) with improved performance in CNV detection in a large cohort such as ExomeCopy or ExCopyDepth (Love *et al.*, 2011, Samarakoon *et al.*, 2014). Cytogenetic techniques including spectral karyotyping and FISH analysis (Ried *et al.*, 1998) have previously been the mainstay in the detection of balanced structural rearrangements however such aberrations can also now be identified using an NGS-based approach by analysing discordantly mapped paired-end reads (Koboldt *et al.*, 2012). Subsequent *de novo* assembly of such reads can then be used to sensitively detect structural breakpoints (Korbel *et al.*, 2007). Moving from WES to WGS for investigating MPD patients in the future will therefore be preferential as it provides greater scope for the analysis of different types of genetic variation within one experiment.

7.1.5 Novel gene discovery in MPD

Due to the large numbers of candidate disease genes still remaining after filtering (Section 3.6), further strategies were used to prioritise novel candidates. These were largely based on researching gene function and identifying a functional link to growth (for example, *NCAPD3*, *NCAPD2* and *FAT1*). However this approach favours genes which have been well characterised and reduces the chance of discovering entirely novel disease genes whose function may provide the greatest insights into growth regulation. Additionally, plausible variants may be identified in multiple genes within the same patient and assigning causality to one particular variant/gene requires supporting evidence. Discovering additional mutations in other similarly affected cases substantially improves confidence that a gene is disease causing, for example *DONSON* and *FAT1*. Recently, several new disease genes have been identified on the basis of a single mutation identified following WES (Shaheen *et al.*, 2014), however, sequencing such as large number of genes substantially lowers the prior probability of a gene being disease causing. Therefore prior to assigning causality in such cases it is prudent to determine the functional impact in patient cells (or cells engineered to express the mutation) and also confirm the phenotypic consequence of each variant in model organisms (MacArthur *et al.*, 2014). The question then remains as to how, if at all, the affected gene is involved in the disease process. Addressing this may not be straightforward but is of particular importance in genes whose function has not been fully characterised as incorrectly assigning causality can have adverse consequences in clinical management.

7.2 Insights into cellular mechanisms in MPD

The majority of genes so far identified in MPD have well defined roles in cell cycle regulation leading to the hypothesis that MPD is a disorder of reduced global cellularity (Section 1.2.3) (Delaval *et al.*, 2008, Rauch *et al.*, 2008, Klingseisen *et al.*, 2011). Additionally, there is also clear genetic overlap with the phenotypically related disorder 1° MCPH in which there is an isolated reduction in brain size (Verloes *et al.*, 1993, Woods *et al.*, 2005). Accumulating evidence suggests that a key pathogenic mechanism reducing neuronal stem cell population in 1° MPCH is the disruption of mitotic spindle orientation in neural progenitors impacting on the plane of cell division and consequently stem cell fate (Section 1.3.1.6) (Wang *et al.*, 2009, Megraw *et al.*, 2011, Chen *et al.*, 2014). Mutations in three centrosomal genes have now been identified in patients with microcephaly and normal stature as well as patients with MPD (*CEP152*, *CENPJ* and *TUBGC6*) (Bond *et al.*, 2005, Al-Dosari *et al.*, 2010, Guernsey *et al.*, 2010, Kalay *et al.*, 2011, Puffenberger *et al.*, 2012, Martin *et al.*, 2014) and thus the two disorders appear to be part of the same clinical spectrum with similar underlying pathogenesis. Why some mutations in these genes result in MPD and others in 1° MCPH however is not yet known. Furthermore, in addition to centrosomal proteins and mitotic spindle formation, DNA damage response/repair pathways have also been implicated in MPD and 1° MCPH (O'Driscoll *et al.*, 2003, Alderton *et al.*, 2006, Qvist *et al.*, 2011, Ogi *et al.*, 2012) suggesting increased genome instability in embryogenesis can also result in reduced cellularity most likely through increased apoptosis. Thus global growth failure appears to result from a range of aetiologies which perhaps act through their impact on all stem cell pool populations, with neural progenitors more sensitive to certain cellular defects resulting in a greater impact on brain growth.

Surprisingly in this study, causal mutations were identified in MPD patients in several known disease genes associated with previously distinct disorders such as Ligase IV syndrome (Section 3.5.2). Identifying disorders which overlap with MPD can provide insights into other processes which impact on growth. Many of the known disease genes identified in Section 3.5.2 function in well characterised cellular mechanisms and detailed review of filtered WES variants in related genes resulted in the identification of the novel disease gene, *XRCC4* (Section 4.2.7) establishing the role of the LIG4-XRCC4 complex in ensuring normal growth. In the analysis of future exomes/genomes in MPD patients, particular attention should therefore be paid to variants in other genes involved in the same mechanisms, such as chromatin remodelling and transcriptional activation (*SRCAP*, *MED12*) (Wong *et al.*, 2007, Lai *et al.*, 2013), mRNA processing (*RBM10*, *NSUN2*) (Tuorto *et al.*, 2012, Inoue *et al.*, 2014) and other DNA damage repair mechanisms (*BLM, ERCC6, PNKP*, *SRCAP*) (Kaneko *et al.*, 2004, Fousteri *et al.*, 2008, Weinfeld *et al.*, 2011, Dong *et al.*, 2014) (Sections 1.3.2.2.4 and 3.5.2).

7.2.1 NHEJ

A novel finding of this study was the identification of mutations in two genes encoding key components of the NHEJ machinery, *LIG4* and *XRCC4* as a relatively common cause of MPD (4% of total MPD cohort) further emphasising the importance of mechanisms which maintain genomic integrity in global embryogenesis. In *LIG4*- and *XRCC4*-deficient mice severe p53-mediated apoptosis occurs in post-mitotic neurons (Gao *et al.*, 1998b, Orii *et al.*, 2006) suggesting that accumulation of DSBs and increased cell death is the primary mechanism of hypocellularity in these patients (Gatz *et al.*, 2011). Additionally, increased DSBs and enhanced apoptosis were observed in human induced pluripotent stem (iPS) cells derived from LIG4 patient fibroblasts (Tilgner *et al.*, 2013) and stem cell depletion was also observed in Lig4 hypomorphic mouse bone marrow (Nijnik *et al.*, 2007). Similar to other disorders of DSB repair such as Nijmegen breakage and Nijmegen-like breakage syndrome (Weemaes, 2000, Waltes *et al.*, 2009, Matsumoto *et al.*, 2011), head size was more severely affected than body size indicating neurogenesis is particularly sensitive to DSBs (Section 4.3.3). This raises the question of whether microcephaly in all disorders of DSB repair arises as a consequence of increased apoptosis. However, the extreme and global growth failure seen in *LIG4* and *XRCC4* patients reported here in combination with genotype-phenotype observations from this study (Section 4.3.3) suggests an additional mechanism specific to the LIG4-XRCC4 complex may also be contributing to the severe growth failure in these patients.

Recently, a reduction in telomere length was observed in lymphocytes from an additional patient with biallelic truncating mutations in *LIG4* (the sibling of which was reported in Yue *et al.*, (2013)) (Stewart *et al.*, 2014). Although a role of some NHEJ components in telomere maintenance has been implied in various organisms (Riha *et al.*, 2006), telomere shortening in vertebrates has only previously been observed in KU deficient cells and was not apparent in LIG4 or XRCC4 deficient mouse embryonic fibroblasts (d'Adda di Fagagna *et al.*, 2001). Notably in this study, predominant features observed in LIG4 deficient patients include pre- and post-natal growth failure, progressive bone marrow failure often from early childhood and immunodeficiency, all features characteristic of Hoyeraal-Hreidarsson syndrome, a severe form of dyskeratosis congenita caused by deficiency in the telomere maintenance helicase, *RTEL1* (Walne *et al.*, 2013). The specific role of NHEJ in telomere maintenance may therefore be pertinent to other cell populations, not just haematopoietic cells, and could be contributing to both bone marrow failure and the global failure in growth over time. Additional signs of a telomere maintenance disorder could be specifically reviewed in the further follow up of patients, such as premature hair loss/greying and abnormalities in nail growth and skin pigmentation (Savage *et al.*, 2010). Telomere length could also be measured in different patient cell lines obtained in this study by flow cytometry with fluorescent *in situ* hybridisation (Baerlocher *et al.*, 2006) however patient derived iPS cells may enable more relevant modelling for telemoric anomalies during development as well as enabling investigation of different cellular lineages along with the possibility of generating organoids (Fan *et al.*, 2011, Tilgner *et al.*, 2013, Lancaster *et al.*, 2014).

7.2.2 Condensin complexes

The identification of mutations in condensin genes in MPD patients provides insight into another cellular mechanism which can impact on growth, chromosome condensation and segregation. This is not the first time abnormalities in chromosome condensation have been described in relation to growth failure. Mutations in the 1° MCPH gene, *MCPH1* (microcephalin), result in a characteristic cellular phenotype of premature chromosome condensation and perturbed mitosis associated with failure in ATR-checkpoint signalling (Trimborn *et al.*, 2004, Alderton *et al.*, 2006).

As the condensin genes have not been previously implicated in human disease, it is important to establish whether the mutations identified are deleterious to condensin complex function. The availability of a patient derived fibroblast line in this study allowed demonstration of reduced protein levels and aberrant splicing in the presence of compound heterozygous mutations in *NCAPD3*. In addition, abnormal chromosome morphology and chromosome segregation defects were observed similar to findings in previous knockdown studies (Ono *et al.*, 2003, Green *et al.*, 2012). This work could be extended by examining chromosome morphology in further detail using super-resolution microscopy and quantifying the defect by measuring individual chromosomes (Green *et al.*, 2012). The stability of the chromosome structure can also be assessed using a TEEN assay which assesses the ability of chromosomes to resume their normal morphology after repeated disruption (Hudson *et al.*, 2003). Chromosomes in cells lacking the condensin subunit SMC2 fail to recover their normal appearance after exposure to the hypotonic TEEN buffer which removes Mg^{2+} perturbing chromatin higher order structure. Therefore patient cells would be expected to exhibit similar defects. Additionally, the stability of the condensin complexes could be determined by examining condensin localisation at chromosomes with immunofluorescent labelling of the various subunits or by co-immunopreciptation studies in patient cells. Mutations may also impact on other proteins such as TOPII or INCENP whose chromosomal localisation is disrupted in SMC2 deficient cells (Hudson *et al.*, 2003). Live imaging of patient cells could be performed to determine whether mitosis is delayed which may possibly explain the hypocellularity in these patients (Hirota *et al.*, 2004). As discussed in Section 5.7.3, the observed increase in chromosome segregation defects could result in an increase in chromosomal rearrangements and aneuploidy, providing a mechanism for increased cell death and consequent dwarfism. Additionally this might also result in cancer predisposition in affected patients. Various cytogenetic techniques could be used to assess the incidence of such aberrations in patient cells including FISH and spectral karyotyping (Ried *et al.*, 1998).

To determine whether condensin dysfunction impacts on global growth, detailed anthropometric measurements could be taken in existing animal models such as the nessy mutant mouse which harbours a homozygous non-synonymous coding mutation in the condensin II subunit, *NCAPH2* (Gosling *et al.*, 2007). Additionally, recently developed genome editing technologies such as transcription activator-like effector nucleases (TALEN) and clustered regularly interspaced short palindromic repeat (CRISPR)/CAS$_9$ RNA-guided nucleases allows patient mutations to be modelled in a variety of different systems including human iPS cells (Li *et al.*, 2014b), mice (Inui *et al.*, 2014) and zebrafish (Auer *et al.*, 2014). Establishing condensin mutations as a cause of growth failure could enable further exploration of condensin function, for example can mutations give further insights into the role of condensin at kinetochores or in transcriptional regulation as well as determine whether different subunits have specific roles within the protein complex?

7.3 Physiological relevance to growth

The overarching question posed by this thesis is can disease causing genes in MPD be used to inform on growth regulation? In MPD all body structures are reduced in size with relative preservation of body proportions consistent with a global effect on growth. Accumulating evidence indicates that MPD results from a reduced proliferative capacity present in all cell types commencing from early embryogenesis indicating that the cells intrinsic ability to divide and survive is critical in attaining a normal body size. A reduced capacity for growth determined by stem cell pools may explain why growth is unable to continue over a longer period of time to allow a normal size to be achieved in MPD patients. The recent generation of cerebral organoids from patient derived iPS cells has enabled human specific modelling of microcephaly demonstrating premature neuronal differentiation (Lancaster *et al.*, 2013) and thus provides a potential model system by which growth in MPD patients can be examined from embryogenesis.

Somewhat surprisingly, relatively few MPD genes have been identified as key components of growth regulatory pathways such as those described in Section 1.2. With the exception of *IGF1* and *IGF1R*, mutations in which phenocopy MPD (Walenkamp *et al.*, 2013), mutations in genes encoding other components of growth signalling pathways have more distinct developmental phenotypes. For example, mutations in components of the MAPK pathway are associated with neurocognitive impairment, pigmentary changes in the skin and cardiac abnormalities (Denayer *et al.*, 2008). Mechanisms controlling growth are clearly complex with cross talk occurring between different signalling pathways perhaps allowing for compensation to occur and thus preservation of body size when certain components are lacking or inadequate. Alternatively, different signalling pathways may only control growth of specific regions and thus global body size is determined by an orchestrated effect of the various different regulatory pathways which all ultimately impact on cell proliferation, survival and differentiation (Klingseisen *et al.*, 2011). Perhaps it is therefore unsurprising that MPD genes all appear to disrupt this common end point.

Interestingly in this study, rare deleterious mutations in *FAT1* appeared enriched in MPD patients compared to control populations (Section 3.6). In Drosophila, Fat functions as a cell surface receptor regulating cell proliferation through the hippo pathway (Silva *et al.*, 2006). Fat1 deficiency in mice results in perinatal lethality (Ciani *et al.*, 2003) and mice with reduced expression of fat1 show a distinct developmental phenotype similar to patients with facioscapulohumeral dystrophy with altered shoulder girdle and facial musculature, retinal vasculopathy, abnormal inner ear patterning and kidney defects (Caruso *et al.*, 2013). Examining the location of mutations across the gene in the MPD cohort compared to control populations may display a clustering of mutations in a certain region which may have a distinct function in growth. It still remains to be established whether the mutations identified in MPD patients have a deleterious effect on protein function however if this proves to be the case patient cells could provide valuable insights into the function of such atypical protocadherins in development.

To date MPD genes have provided relatively few new insights into growth regulation as most genes discovered so far function in cell cycle machinery or DNA repair processes impacting on cell proliferation and cell death thus explaining a reduction in global cellularity. However, the discovery of novel disease genes in MPD patients with relatively unknown function may provide significant breakthroughs in understanding growth regulation, for example, the possible identification of *FAT1*, a novel gene strongly implicated in a developmental growth signalling pathway.

Appendix I: Primer sequences

Genomic DNA primers

Gene	*Exon*	*Forward*	*Reverse*
LIG4	**1A**	AAACGAGAAGATTCATCACCG	TCCTTTCTGTAAACATCTTGGCT
	1B	TGTTGAAGCCAAGATGTTTACAG	TGCATTATGAATGAATGGGG
	1C	AACCAAGCTAGATGGTGAACG	TCCAATTCATCCATTAGTCCAC
	1D	CCAATTCCAGGTAGAATAGAAATAGTG	TTGTTCTAGGTCGTCCAGGG
	1E	TTGAACCTTGTAATTCTGTCATTG	GGCAAAATGTTCTTTGGTTG
	1F	CTGTGTAATTGCAGGGTCTGAG	TGCAACACGACTATGATCTTCC
	1G	ACACCGTTTATTTGGACTCG	CCTGCTGCAATGAGTCTGC
XRCC4	**2**	TCCTTGGTGTTTGTGTAGCTG	TGGAAAAGTATCCCTGAGGAC
	3	CGAGATGTGAGTCTAAATTAATGC	CAAGCTTTTGCTTCTGAGGTG
	4	TGCTTAAAACCAGGCTTCTC	TGTTTATAAATTTGCTGGTGCC
	5	TCTAAAGGCTTATTTGTATCAGTTTTC	AACAGCCTGAACATCCACATC
	6	AATCTGCTGCCTAGCAGGG	TTCTTGTCATTAGAATAAGAAGCCC
	7	CTGACTTGATTCAACAAATCTGC	TTGCTTACGTTCTCAGCATTTC
	8	GAAACAGGATTTAACTGTCATTTCAC	TAATAGCGGCTGCTGACTTG
PCNT	**1**	GTAGCGCGACGGCCAGTCGTGACAGTTGTCGCGGG	CAGGGCGCAGCGATGACGACCGCGTCCTCTCCTGG
	2	GTAGCGCGACGGCCAGTCCTCATGGTCCCCACCAGTC	CAGGGCGCAGCGATGACCCTCTTTCCTCACTTTTGAAAGGC
	3	GTAGCGCGACGGCCAGTGCCTGCAGATAGAGGAGCTGG	CAGGGCGCAGCGATGACCCCTGCAGAGATGGAAGGTCC
	4	GTAGCGCGACGGCCAGTAGGACGTGCGTCGTCAGTTC	CAGGGCGCAGCGATGACAAAGGAGATGGCAGCGCCC
	5	GTAGCGCGACGGCCAGTCTTCCACGGGATGTCTGCTG	CAGGGCGCAGCGATGACTCCCAACCCCAAGAGGGAAC
	6	GTAGCGCGACGGCCAGTCCTGGTCGTTTCCTGGGC	CAGGGCGCAGCGATGACGCCTCGACTGGCTTCTGTCTG
	7	GTAGCGCGACGGCCAGTTGGCCTGCTCAGAGGTTTTG	CAGGGCGCAGCGATGACCGCAATGACGCATGGTGAC
	8	GTAGCGCGACGGCCAGTGCTCTGGGTGCACCATTATTGTC	CAGGGCGCAGCGATGACGCAAATAACACAATACGAAATAGCCC
	9	GTAGCGCGACGGCCAGTTTAGGATCGCAGTGGCAGGG	CAGGGCGCAGCGATGACGCAAACCTCCACTTTCAAACCAG
	10	GTAGCGCGACGGCCAGTCCCACCACAGGTAACCAGGC	CAGGGCGCAGCGATGACGAAATCACCAACAAAACTACCCCTG
	11	GTAGCGCGACGGCCAGTCTGTGAGCAGTCGGTCCTGG	CAGGGCGCAGCGATGACGGAAGGACGGAAACACAGGC
	12	GTAGCGCGACGGCCAGTAGCAGGAAACACCTTTGAGGG	CAGGGCGCAGCGATGACTTCACGGAGGACTTGGATCG

Gene	***Exon***	***Forward***	***Reverse***
PCNT	**13**	GTAGCGCGACGGCCAGTAGGTTGCCGTTCTGCCTGTG	CAGGGCGCAGCGATGACCCCAGGATGCCCCTCCATAC
	14	GTAGCGCGACGGCCAGTTTCAGAGGTGGGTTTGGGTTG	CAGGGCGCAGCGATGACCACAGCGCCTCCTCCCAG
	15A	GTAGCGCGACGGCCAGTTCCCGAAACGATGACCTGAAC	CAGGGCGCAGCGATGACTCTAGGGCCTGCCTGTGCTC
	15B	GTAGCGCGACGGCCAGTCCTCCTTAGAGAGCAAGCAGGG	CAGGGCGCAGCGATGACCACGCAGACATGTGACACGC
	16	GTAGCGCGACGGCCAGTCTCCATGCATATCTGCTGAATG	CAGGGCGCAGCGATGACCCGCCACTGAATGTACAACAC
	17	GTAGCGCGACGGCCAGTCACCAAGCTGAAAAGTCCTAAGTCAG	CAGGGCGCAGCGATGACGAGCCGTGGGGACCAATG
	18	GTAGCGCGACGGCCAGTCGAGGTGTGCAAACTGGTGG	CAGGGCGCAGCGATGACCGAGGTGTGCAAACTGGTGG
	19	GTAGCGCGACGGCCAGTTGGGGCGACGTTCTGAGTTC	CAGGGCGCAGCGATGACGCACCGGCATCCACCAG
	20	GTAGCGCGACGGCCAGTGCCCTGCCTGGGTGGTG	CAGGGCGCAGCGATGACGCCTCCCATGTTGGCTTTG
	21	GTAGCGCGACGGCCAGTAGCTCTGCTGCTCTCAGGGG	CAGGGCGCAGCGATGACTCCTTGGGAGCTGGGGAAAC
	22	GTAGCGCGACGGCCAGTTCACCATCAGGAGATGCACG	CAGGGCGCAGCGATGACAGAAGCTTACGTAACGATCTGGAAAG
	23-24	GTAGCGCGACGGCCAGTCAGTGGAAGCCTGGGTGGAC	CAGGGCGCAGCGATGACGAAAAGCGAGTGGGTGGCAG
	25	GTAGCGCGACGGCCAGTTCTAGGGGAGGGCATAGGGC	CAGGGCGCAGCGATGACTTCTGGTGCAGACGTGGTGG
	26	GTAGCGCGACGGCCAGTTCCACCTGCTCTGCTTCAGG	CAGGGCGCAGCGATGACCAGTGGTCACAAGCCCATCG
	27	GTAGCGCGACGGCCAGTCAGGGTAGTAATTGCTTTAGGCATGTG	CAGGGCGCAGCGATGACGACCAGAGACTCGCCTCCCC
	28A	GTAGCGCGACGGCCAGTTGCATTCAGGATGTAACGTGCC	CAGGGCGCAGCGATGACGGCAATCGTCGCTTCCTTTG
	28B	GTAGCGCGACGGCCAGTCTGGCTGAGCTGGAGCG	CAGGGCGCAGCGATGACGCCCAGCGCAGAGAGAAGTC
	29	GTAGCGCGACGGCCAGTCTGGCTGCCGTACTGGTTCC	CAGGGCGCAGCGATGACGATTACAAGAATAAATCTGAGGC
	30A	GTAGCGCGACGGCCAGTTGCTCCGAAACTCCCTGAAATC	CAGGGCGCAGCGATGACACCAAGTACGCTGGTCGGTG
	30B	GTAGCGCGACGGCCAGTTGATGCCAATACAACCCCAGG	CAGGGCGCAGCGATGACGAGGTGAACAAACAATGGCAGC
	31	GTAGCGCGACGGCCAGTTGGGAAGATAAATTCAGGCCTTTG	CAGGGCGCAGCGATGACCCAGGCAAAGGATGCAGG
	32	GTAGCGCGACGGCCAGTGATCACCCCTGTCCTGCCTC	CAGGGCGCAGCGATGACCTGGCCCTCCAAGGCTCTC
	33	GTAGCGCGACGGCCAGTGCCCTTCACAGAGTCCTGGC	CAGGGCGCAGCGATGACCCACGTGGCCCTAAGGACG
	34	GTAGCGCGACGGCCAGTGCCCCATCTCCAAACGCAG	CAGGGCGCAGCGATGACCCGAAATTCACACACAAAACAGTC

Gene	*Exon*	*Forward*	*Reverse*
PCNT	**35**	GTAGCGCGACGGCCAGTGGCGGAGCTGGTTTTGAGG	CAGGGCGCAGCGATGACCACCACCCTCTGCTCCCAAG
	36	GTAGCGCGACGGCCAGTCCCCACGAGTCTGTCTCTAACCTG	CAGGGCGCAGCGATGACGGCCCACCAGAGGTTCTGTC
	37	GTAGCGCGACGGCCAGTCTCTGCCTCCCCTCCTGG	CAGGGCGCAGCGATGACTGCACTGCCATGAAGAACAGG
	38A	GTAGCGCGACGGCCAGTTTCAAAAGGTAGAATTGCTGGGC	CAGGGCGCAGCGATGACTGCTGAGAATGCAGGGCTTG
	38B	GTAGCGCGACGGCCAGTACAGGCCCATCACGCTCTG	CAGGGCGCAGCGATGACCCAGACATAAATCTCGCGTCCC
	39	GTAGCGCGACGGCCAGTGGCAGTTTTGTTTTTGGACACTG	CAGGGCGCAGCGATGACCAGGGGCCTCACACCAGACG
	40	GTAGCGCGACGGCCAGTTTGAGCCGTGAGCTCTTGTTTAG	CAGGGCGCAGCGATGACGGAAGAACAATGGGGAAGCG
	41	GTAGCGCGACGGCCAGTTCCACAAGCTTCTCATTGAACCC	CAGGGCGCAGCGATGACAGAAGGTGACGCCCACATCC
	42	GTAGCGCGACGGCCAGTGCGTCTCTAAGATGCTCTTGTTG	CAGGGCGCAGCGATGACGCAATTTGCAAACCCAGG
	43	GTAGCGCGACGGCCAGTTACTGTTGGAAGGCCGATGG	CAGGGCGCAGCGATGACGGCCTCAAATGCCATTCATC
	44	GTAGCGCGACGGCCAGTGCTTGTTTGGTCACAGTGGGG	CAGGGCGCAGCGATGACTTCACGGATTTTCTTACCGTGC
	45	GTAGCGCGACGGCCAGTTGTCTTAAAATTGCCATACAGGCTC	CAGGGCGCAGCGATGACGAGTTTCCTGCCTTGCCCTG
	46	GTAGCGCGACGGCCAGTCCAGAGTCACCCATCCCCAC	CAGGGCGCAGCGATGACTGACCTCAGCACGTATCACTGAG
	47	GTAGCGCGACGGCCAGTTCGACCAGGATATAATTTGTTCAGTG	CAGGGCGCAGCGATGACAATGCTCCAGCTGGCTTTGC
ASPM	**1**	GTAGCGCGACGGCCAGTAAGCGGTCAGCGTAAGTCC	CAGGGCGCAGCGATGACTCTCCAATCGTCAACCTTCC
	2	GTAGCGCGACGGCCAGTAATTAAGCAGATAGGGTAGGAGAAA	CAGGGCGCAGCGATGACTGCAAAAATAAGGAAAATATACCAA
	3A	GTAGCGCGACGGCCAGTGGAAATGCAGAAGAGCAGAAA	CAGGGCGCAGCGATGACTTGAAGAACAGTTGGGGGTAA
	3B	GTAGCGCGACGGCCAGTGTGCAACTTGCTTGCCACT	CAGGGCGCAGCGATGACAAAAATTGATTAGGGGATAAAATAGGA
	3C	GTAGCGCGACGGCCAGTAGAAAATTTTAAGTCCAGATTCTTTCA	CAGGGCGCAGCGATGACTTTTCATGTTCACCCACTGC
	3D	GTAGCGCGACGGCCAGTAGGCCACCTGTACCAGAGAA	CAGGGCGCAGCGATGACGCTAAGGAAATGTACCCAGCA
	4	GTAGCGCGACGGCCAGTGAATATGATTGTGAAGAACCCAAAC	CAGGGCGCAGCGATGACTTCTTCCAGGCTGTTATTCAAC
	5	GTAGCGCGACGGCCAGTTTGTTCAGTGTTTTAAAGATGGTATTG	CAGGGCGCAGCGATGACGCTAATGAACAGGGAATTATGC
	6	GTAGCGCGACGGCCAGTTGAAATTGCATTTTATTGCTGG	CAGGGCGCAGCGATGACTATGTCAATAAAGCCGGGG
	7	GTAGCGCGACGGCCAGTCATGCTTTAGCTTTGCTGCC	CAGGGCGCAGCGATGACATGGCATAGTCTATTTACCTAATAAGC

Gene	*Exon*	*Forward*	*Reverse*
ASPM	**8**	GTAGCGCGACGGCCAGTTCCTCAGTCACTTCCCTTTG	CAGGGCGCAGCGATGACAAACAGGAAGAATGACAATAAGCC
	9	GTAGCGCGACGGCCAGTTTTGTGCTTGCTACCCTACAC	CAGGGCGCAGCGATGACGCAAGCAAAAGTCGTAAATGG
	10	GTAGCGCGACGGCCAGTCAGAATGATTTGGAGGATTTG	CAGGGCGCAGCGATGACAAAAGTGTTTTCCAGAAAATGTTAGTC
	11-12	GTAGCGCGACGGCCAGTAACTGTTGGGATTTATGTGGG	CAGGGCGCAGCGATGACAAATGATGGTTGTTGTTGTTATTC
	13	GTAGCGCGACGGCCAGTTCCATTTCAGGCACTTTATTTTC	CAGGGCGCAGCGATGACTTTGAGGGAAAGTTTGCTTACAC
	13-15	GTAGCGCGACGGCCAGTAACTTGCCATTACTTGCCTTG	CAGGGCGCAGCGATGACGAAATGACAAATAGGTAATAACCACC
	16	GTAGCGCGACGGCCAGTTCTGTTTTTATCTTTTGTGGGTTTT	CAGGGCGCAGCGATGACACCTCCCCAACCCAAAATAC
	17	GTAGCGCGACGGCCAGTTGTAGGGGTGTTTTATTTCCAG	CAGGGCGCAGCGATGACAAACTTCATCACATTTTGCCTTC
	18A	GTAGCGCGACGGCCAGTGGATTTCTGAATTGGCTA	CAGGGCGCAGCGATGACTTCTTAAATGCCATTCTCTAAAAGC
	18B	GTAGCGCGACGGCCAGTTTCAGAAAATGGAAGCAACG	CAGGGCGCAGCGATGACGCAGAGCGTGTTTTCTGGTA
	18C	GTAGCGCGACGGCCAGTCAGGCACATGTAAGAAAACATCA	CAGGGCGCAGCGATGACCTTCCTCTGATTGACCTGTGC
	18D	GTAGCGCGACGGCCAGTTACAGTCTTATTTCAGAATGAGAAAGG	CAGGGCGCAGCGATGACTGTTGCCTTTGAAGCTGTCT
	18E	GTAGCGCGACGGCCAGTGACTGCAGGAAGGAAGCAAT	CAGGGCGCAGCGATGACTCACGCTGCATTTTACCTTG
	18F	GTAGCGCGACGGCCAGTAAGTTACCATACAATGAGAAAAGCAG	CAGGGCGCAGCGATGACGATGAAAGTAGCAGCCCTGTG
	18G	GTAGCGCGACGGCCAGTTTCCTTCAGGTACAAAATGCAG	CAGGGCGCAGCGATGACGAAGCATGTTTCCAAGTCTGAA
	18H	GTAGCGCGACGGCCAGTGAAAGGCAGCCATTACAATACA	CAGGGCGCAGCGATGACACTGCTTGGGTACGCACTG
	18I	GTAGCGCGACGGCCAGTGCATTATCTCCACCTTAGAGCAA	CAGGGCGCAGCGATGACCCGATACACAGCCATCTGAA
	18J	GTAGCGCGACGGCCAGTGGGCTGCAGTAACAATTCAAA	CAGGGCGCAGCGATGACGTGCCCTTTCCCTCTTTCA
	19	GTAGCGCGACGGCCAGTTTTTGTCTGCACTAACCTTATTTTAAC	CAGGGCGCAGCGATGACAAAGCAAGACAGTACGAGAGATG
	20	GTAGCGCGACGGCCAGTTGCGTGTGTAAATTCTGATTG	CAGGGCGCAGCGATGACTGTGTGAAATAAATGCATACTTAGGTC
	21	GTAGCGCGACGGCCAGTGAACTTCTAGTGGGTTGGAAATC	CAGGGCGCAGCGATGACTTCTTAACACTATTTAACATCAAGTGC
	22	GTAGCGCGACGGCCAGTCCTTCTCTAATAGGGCAGTCTAGTTG	CAGGGCGCAGCGATGACCTCAAGACTTTGCTGGCAGG
	23	GTAGCGCGACGGCCAGTGGGGAAGTAATGCCTCTGTG	CAGGGCGCAGCGATGACAAAGAGCTTAGCAATGAAATTATGG
	24-25	GTAGCGCGACGGCCAGTTTTGGTCGATAAATGCTGTCC	CAGGGCGCAGCGATGACCAAACTTAATTTGCAGGGGC

Gene	*Exon*	*Forward*	*Reverse*
ASPM	**26**	GTAGCGCGACGGCCAGTTTGGTTGGGTTGTTTGTAAATG	CAGGGCGCAGCGATGACCAGGTTTGAACACACATAAAACC
	27	GTAGCGCGACGGCCAGTGCGACAGAGCAAGAGAGACC	CAGGGCGCAGCGATGACTTTCTCCACTGAAAAGCACATC
	28	GTAGCGCGACGGCCAGTATGACCAAAAGGCAGTGGTC	CAGGGCGCAGCGATGACTGCCATTGATATGAATTTGTGAG
IGF1R	**1**	TGAGTTTGAGACTTGTTTCCTTTC	GGGTTTCGCAAACAGGG
	2A	AGGATTCCTGAAAACCAACTG	TCTCAATCCTGATGGCCC
	2B	CGGCTGGAAACTCTTCTACAAC	AGGTCAAGGAGGAGGAGAGG
	3	AGAGAAGGCGGTGCCTC	ACCTTTGTGTGCTAGGGTGG
	4	GGGGTGAGATACCATGTGAC	TAGTGGTGACTCAGGGACGG
	5	TCCAAGTATGTCACCCTTACACC	CAGCAACACAGTTTCATAAGCAC
	6	AGGCTAGAGGGGACTGTGG	AAGCTGCCATCACACATGG
	7	GCAAGACAGGTGCTTTTCAG	TCAAGGATCGAAAGACTGGC
	8	AGAAAGTGTGACATGCTGGG	ACACCTGGTACAAGCAAGGC
	9	GCAGTTTCCTGTTGGCTTG	GGCTCAGGCACATTACAACC
	10	GCTATCTTCTTGATTAAAGGTACTGAG	ACTACCTGGTGGGGAGAAAG
	11	TGTGCCCATTCTGACACTTG	GAGAGCGTGGGCTCTGTTC
	12	CTCAGGTCAGCCCAGTGTTG	GCATATGCTGTCAATGGATG
	13	AGTTTAGTTGGCAGGCCCC	GGGCTGTCCTGATCCTGG
	14	CAGGAATTCTTACTGTATGATGGG	TCACCTTCACTCACATTCAAATC
	15	TTCTGATACCGTGTGAGAGAGG	GCTCAAAATAATGCAAACCTCC
	16	AAAGCACGTTCTGTCTAAGGG	GCCAAGAACATACTGGGAGG
	17	CAGTTCCAGACAACACAGGC	TTTAAAGACACAGCATTTCCTTG
	18	CTCGAAAGAAATTGGCATGG	TGCCAACAAAGTCCTCAAAAC
	19	GTGTAGGGTCCTCTGCTGTG	ACTGAGCTGGTGGAAAGTGC
	20	GCATTGTTCAGTCCATCCC	ACAAGGACAAACCTCTTAGCC
	21	CCGCCGTACGCTTGTATG	TGAGGTACAGGAGGCTTGTG
ORC1	**4**	GAATTGCTCAGTAAAGTGTCCTTG	CATGAAGTTCAGGAGGCAGAC
CENPJ	**12-13**	AAAGGACAGCAGTTCACAGG	GCATCTTGTCCAAAGAGCCC
SRCAP	**34**	GACCTGGAAGCTGCTAATTTC	CGGAGCTGAAGCTGGAGTG
MRE11A	**4**	CCAGATTGAAAGTCCCTTTG	GGAAGGCAAAACAGTTGTGTG
	13	ACGTGTCCTGTACTCCTCCC	CAACCATATGCAAGACTCTGTTC
BLM	**8**	AACGTGTGCCAGTGATTCTG	GGAGGTTTAAGAGGTCCCTAAATG
	3	CAGCAGGACAGGAAACACAG	CACATCGCTGCTTAACCATTC
ERCC6	**21**	TCTTGTGACCCTTCACAGCC	AGTGCAGCCAACTTCCATTG
	11	TTAAGGAATGGAAGCAATTGAG	CACCAGACTCTCTCATCCGC
MED12	**44**	TGGGAAAGGAGGTTGAAGAAG	TTTCAGCTACTTTGGCCCAC
RBM10	**8**	GACCAGTCGTGGAGCCTTC	GTCCTGCTCACTGCCATTTC
ESCO2	**4**	CGCTTAGTGAATACAGGGGAAG	TAACCACAGGCACGCTACTG
VPS13B	**56B**	CAAAGAGTTGGAAGAATACAAGGAA	AAGCGTGTCATTTCTACCCG
	62	TTTTGTTCTGGAACTCTCGTAAG	CAATGCAGTGAGACCCTGTC

Gene	*Exon*	*Forward*	*Reverse*
PNKP	**9**	CCTGATGTGGGGCACAG	AACACACGGGACACCCC
	17	GGCTACAGGTACTGTTGGGG	TCCATGTAAGAAACTGGCAGG
TUB-GCP6	**19**	CATCTCCTGTGGATGGGG	GGAGCTGGAGTCAGGGC
FAT1	**2A**	GGACCAAAGGAGGAAATACAGC	GGACCAAAGGAGGAAATACAGC
	3	TTCTGAAGCTGGCAGTTTTG	AAATGCTAATTCTTACTTTTCCCC
	10A	TTTTGCTATTGATTCTAGCACTGG	GATGTTGCCTTCGGTGATG
	10B	TGTTTGGGAATCACAGCAAG	TGTCTATGTCTGAACTGTCTGCATC
	19	TTCCGTGACTGACATTGAGG	AAGGCCGTCTGAAACCATC
USP2	**1**	GTCAGCTGGTGCTCACTGC	ACTTTCTGAGCTCCAAACCG
DONSON	**5**	GCGAGTCCTCATAACCAATTTC	AAAAGTTTGCGGACTGCTG
	6-7	AACTCTACTGCCTTAATAGATCATTTG	AAAACTGCCAATGTTAAACTTCTC
	8B	TGGGAGAATAGGTGTGTTTTGAC	AGCGCAAGACTCCATCTCG
NCAPD3	**1**	GATTGGTCCACAGGAACG	ACGCATAGGACCCTCGC
	2	TCTCCTCTTGAACACACCTGG	CCTGTGAACTGTACACCAAGAAAG
	3	GGAGCCAGGCTACTCTATTCC	TCTCTGCCATGGGTTAGCAC
	4	CATGAAAAGCCTGTGTTTTGC	TGCCTATAATCATGATGCCTG
	5-6	CCATGGGACAAAGTGTTTGG	TGGTATTTCCATTTCTGCTGG
	7	CATTTAAACATGGCTTGTAGGAC	TATGGTTCATGTGCCTTCG
	8	TCTGAGACCCTTGTGTTCTCC	CTTAGTGCAGGGCCCAGAG
	9	CCACTCTTATAGCGTGAGGTAGAG	AACAATGCACACTCATTCCC
	10	AATATGTTTCAGCCATCCTGC	CCAAAGAACCCTCCTCATTC
	11	TGAGGAGGGTTCTTTGGTTC	AAAGAACGATTCCCTTGCAG
	12-13	TCTTGATTTTGATTCCTCCG	ACGGTTGTCTGAAATCCAGC
	14	TGGTGAAATGTTAGCTGCTG	ACAACAGGATCAGGGTCAAC
	15	CAGTTTATCTTTGTGAGTTGATGC	GCAGCAGGTAATTATTGTGTGG
	16	GGTCTATTTGCAGTTGAAGGTG	CATACACTAAGTAATCTGGAAAGCT
	17	CCTGGAGCAGCAGAAGTTG	TCACTTCCTTAAGTCTTGGGC
	18-19	TTGGCAGCAGTGTTTAACTTG	CAAAACACATACAGTAACATCCTAAGC
	20	AACATGGTGTTTCATGGGTAG	AGAACCTCACGGATAAAATGG
	21-22	GGGATAAAGGAAATACTCCATTGAC	CAGCCCCTGACACAGTCC
	23	AAATTTAGGAATGGAATTAGGATTTAG	TCCCTTAAACCAGAAGCAATG
	24	CTGGGAACAGTGAGGCTTG	CCAGCATACATCAGAAATGAGG
	25	TTGGTGTATTCAGAGACTGTTGG	AGCCTAACCTCCCCTGATTC
	26	CAGTGTTAAATCCCTTACACCC	CTGGGCCATTATCAACACAG
	27	AATTACACCTCCTCTCTGGGC	CTTAGGGGAGCAGCACACTC
	28	TGTGGTAGGGTGGCTCTACTC	AATCATTCACGGATTGCCTG
	29	TTGTTACACAGGCGAAGGTG	GATGACGGAGATCTTGAGAGG
	30	TGATTTGATCATTACGTTGTCC	ACAATGACTTTCCTGCCCAC

Gene	*Exon*	*Forward*	*Reverse*
NCAPD3	**31**	GTGCAGCACTGTCACCTCTG	CACCGCTTTGAAAGGAATAAG
	32	TGTGGTTGTTTTGCTCTTCG	ACACATCACGAATGCAGGAG
	33	CCCACGGACAACTAGAACAG	CATTCAGCTTTCCCAGAACC
	34-35	CCTTCTGGGCTGATAGATGC	GGACACGAGACTGCTTCCTC
NCAPD2	**2**	CAGGTCTTGAATATAGACCCTGAC	CCATGTGAGATGGTCCTTTG
	3-4	TTTTCTGCCATGGATAGAATTTAC	CATGTGGCTGACAAAATGTAGG
	5	TGCTGGGTGTTTTGAAGTTG	TGGATCATTGGATTCCCTTC
	6	TGGTACTAGATTGTTTGGGCAC	TCAACAGAGCCTCACTCTCAC
	7-8	TAGAATGAATGAGGCAGGGC	TCTTTATTCCCTCCCCAAGG
	9	CCTTCCCTTACTGGGAACTG	TTGTCAAAGCTACTTTCTTGTTCTG
	10	AGTTATGGCCAAGCACAAGC	CACTGGATTCCTCTTTACACTGG
	11-12	GAACAAAATGGATGATGTGGG	TTTCTGACTGCAAGGAGAAGC
	13	AGACAGGCTTCTCCTTGCAG	CACCTGGTGGCAGCATATC
	14	TTACAAACGGCTGCAGTGAC	AGGCCAATGATGGCATAAAG
	15	GTTGGATTCCCCTGTATCCC	GGAAGGCCCCATCTCTATTG
	16	TTCAGAAAGGTTCTGTCGTTAGG	GAAGTCCTGTGAGAAGCTGACC
	17	TCCAGATCCATAGGCAGTCC	GATCACGAGGTCAAGGGTTC
	18	GCAATATCATCTTCAGAAAAGCAG	CCTCTCAGAACCTTCCTTTGC
	19-20	CCCCGATGAGTCCAATCC	CAGAGGAGGAGGGAGAAGTTG
	21	CACAAAAGGTGAGCTCTCAGG	CCTGCAGGGAGAACAGAAG
	22	TGAATAGCACGTGAGCAAGG	AGCAGAGGAGCAGAGATTGC
	23-24	CATGCAAAACAAAGTATGGGC	AAGCCATTTCATCCCCATC
	25	TCTCAAGGAAGATGGGGATG	AGCCCTGGTGATTCTCACAC
	26-27	CCAGTGTTAGGGTGTAGCCC	GTGTCTGATTCCTAAGCCCG
	28-29	TCCAGGGTTGGTCTTAGTGG	AGATGTCTTAGCCAGCCCAG
	30-31	GTCTTCCATGGCCTTGTTTC	CCCTCTGCACCCTAGCTG
	32	GTAGATGCTCTGCCCCACTG	CTTGGTATCGCAAAGTGCTG
NCAPH	**1**	AGGCCCGAGGTGGTCTG	AGTACTGCTTCGCTAGGGG
	2-3	TCAGTGGACAGTGTGTGGTG	TAAAACCACCAGCACTAAGATG
	4	TGGGTATTCCTATGTAAAGACTTAGC	TCATGTATTTGCAGCATTCAAC
	5	TGAAGCTTGCTGTGTTTTACC	TCAAGTAACTACCTGGTATGCCC
	6	CGACAGGTTTTGAAGAACTTGAG	TTAGCCCTGCCAAGTCTCAG
	7	TCTTTCCCTTGGCTACATGC	TGATGGGCTTATTGTACCCC
	8	TGAAGGGAGTAAGGAATGCAC	GCAGGACCACTCTTGAGCC
	9	GTCCCTCCAGTGGTGATTTG	AATGCAGTGAGATGCCCG
	10	AATGCAGCTCATTTGCGG	ACACAACAGGGCTGTGACTG
	11	CCTTGATAAGGCTGTCAGCAG	GGAACAAGGAAGAATGAAACC
	12	AAAGTAAGATGTTTTGCCCACTG	GGCACTGCCTAACTGTCATAC

Gene	*Exon*	*Forward*	*Reverse*
NCAPH	**13**	AGGAGGAGATTGGGCATAGG	AATTGCCAACAGTCTCCCAC
	14	GTGTTTCTGAGCTCCTCTGC	GCTCAGATTGCGTGACTCTC
	15	AGCTCCTTGGTTGAAACTGG	AGACATCAACCACTCCCCTC
	16	TGAGGCACTGGTTCTCTTTGTAG	ACAGGTTGTGTGTGGAAGGC
	17	CATACTTGATGAACGAGCTTTTG	CTGTATTTAAGAAATCCTATGTTAGCC
	18	CACTACCTTGTGACTGTGATTGG	GTCCCGGCAACTGAGTAAGG
NCAPH2	**1**	CGCCTACGCATTTTCCTG	AGAGCGCCAGATAGCCATAG
	2	AGACCAAAGATGGGTGGGG	ACTGTGGTCCCACTGGC
	3-4	ACCGATGGATAGCTGGGAAG	GGAGCCCAGGCAAACAC
	5	TGAGTTTCTTTGGCATGTGG	CCCAGGGAAAGTGCTGAC
	6-7	ATGTGGGTCCTGTTCTCCAC	GTCCAACCAGAAGACCCTCC
	8	GGAGAGGCTTGTTGAACACC	CTCCCTGGTCCCCTCTTTG
	9	GGAAGTAGCCATCAGGGTCC	CCAGGGTTGCCAGTTCAC
	10	GTACCCAGAGCCTTGAGGG	GCAGGCAGCCAGTCCTC
	11-12	TACTCCTTTGGGAGGGACG	CTGATCCTCCTCCACTGGC
	13	GACAGAAGCGCAAGAGGAAG	GGGTGACACCAGGTGGG
	14-15	CAAAGCCTTGGGTGGACTG	AGAGTGAACAGCACGCAGG
	16-17	GGCACCTGCCGACTAGC	AGGAGCTACTGCCTCCCAAG
	18-19	GGGACAGGTGAGCGGTG	ACATCAGCAGGGTCTGGG
	20	GAGTAGCCTGGGATACGTGG	GCCAGCCAGGAGCAGAC
NCAPG	**1**	GGGCTGGCAGGCTGTAG	GGCTGGAGTCACGGTGC
	2	TCATGCAAATTGATTGCCC	AAAGGAGAACTGGTTTAAAGAGG
	3	TGTTGTAACTTCTTGCATTCCG	TTTTAAGCTGGGTGGTTGC
	4	TGCCCAGCCTATCTTTATCAC	AAGAAAATGGCTCCTTGTCC
	5	TTGGAGATTCTGGGTAATTTTC	CTGGCAGGCTTACTTGACC
	6	AAAATACAAGAGATGGAAGCTAGAGG	AATATAATCTTAAGCTATGGGGAGG
	7	GTGGTTTTAAAGCAAATCTTATTATTC	ACTTTGAGGAAGGGGACTGG
	8	AATATGACCTGAAAATGGCCC	TTTTATACCACTTTAGAGAGCTAGGG
	9	GAAAGAACTGAGGCAAATGTAATTAAG	GAGACGAACCTAAAACAAAACAGAC
	10	TCAATATATGGTGGATACTACTTGGAG	TTTCAGAACTGCACAGCAAG
	11	CCTGACTCATTAAATTCTGGGC	CAAGGAAGCAAATAAAGTCCTACC
	12	ACACAAAGCCCTGAGAAAGC	GTCCTTTACAGACATCTTATTTAATCC
	13	AAAATGACACCTTTGAGGTAATAGAC	TCGCCACAGAGTATTTCAGG

Gene	*Exon*	*Forward*	*Reverse*
NCAPG	**14**	TTATTGACTGTGAAGATTAACTTTGTC	CACACTGGGACCTATCAGGG
	15	AGGACAGGCTACATTGAGAGG	CATTTTCATCTAGGAAAGAGCC
	16	GCTCTTTCCTAGATGAAAATGCAC	TGCAGTTCCTCAGTATTTCAAGG
	17	GAAGGGGAGCTTATTTTGGG	ACCGCATTCCCTAATCTTTG
	18	TCAAAGATTAGGGAATGCGG	AAACATTTATCAGTAATGCTCTTTCC
	19	TGCTAGAAGGACTTTGTCATTTC	GGAAACAAAGATAAGAAAACTTGC
	20	CATGTGTTGAAAGGATTGTTGTG	GCTCTTGATGATGGAATTTTAGAC
	21	CACTGATAGAATGTGAGTCAGAAAC	TTTGTTCTGATTTTGACAAGGG
NCAPG2	**2**	TGATTGTGTTTTCCATTAGGG	TGGTCGCTAAGGATTCACTG
	3	GAATCCCTGTGCTTCTTCCC	TTCTTGACTTCTAACTTCAAAATCTG
	4	CCCAGGGTGAAGACAATATG	AGGCAACAAGAGCGAAACTC
	5	GATTTCTCTGAGACTTCCAGTTCC	TGTGGCAAGTGGCTACTACG
	6	TTGATTTGAAATATGACTTTACTGTTC	GTGCAGACAGGAACCATGTG
	7	TTTTAGCAATAGCAATGTTTAGCTTAC	CCAAATGGAATCTTTCAAAACC
	8	TTCACTATGGTTGGAGAGTAGCTG	TGGATCTTAAATAACCCTTCTTTG
	9	TCAGGTCGTCAGTAGGACACAC	CCCCTAACTGATGTCTGTGAAAG
	10	CTGTTGCTCTTTCCAAATGC	TTTGCTAAATTCAAAAGCACAAG
	11	TTTGGGAACACTGTACTTAAGCC	AGTACATGCCACAGCCTGG
	12	TGAAGCAGATGTGGATTTCG	TCTTTGTGGAGTAGGATTACAGG
	13	CTACCTGAACCTTTCATTGGC	TGTTTGATGAACACATGTCCC
	14	CTGTTTCCAGGGTTACAGGC	TGTAATTTGTATTATGCTTTTGGG
	15	CACCAACATAGGTAGGGAGTGC	CTGGGACAAGATGTGACTGG
	16	TTCTGAGATGCTGTTACCTTTTAG	TTGTTAAATTCTAAACAAACGTAACC
	17	GGAATTTACGTAGAGAGACTTTCAG	TCAGCACAACAATGGACAGG
	18-19	ACCTTTTGACAAAGTTGGGG	TGAAAGCTTCACACTTTGATTTG
	20-21	ATTGTTCTTGCTGTACGGGG	AATCCTGAAACATAACAAACTTGG
	22	GTGAGGCTGGATTTACTCCC	AGGGAGGGAAGGACTGGTTC
	23	GGCTTTTGGTATTGGCTCAC	GGCTGATGCATCTGTGTAGG
	24	GTGACTCCTTGAGTCCTGGG	TTGAGAAACGTAGAGAATGTACTGC
	25	TGTTGAAGAAAACCCAAAGC	GCTGAGCGCTTACAGCCAC
	26	GTGGTTGTGGTTGAACCTTG	CAGGAGTAGCCAAGTGCAGG
	27	ATAGGGAGGCCCAGCTTAGG	CCACCTGGGTGTAAGTAGCAG
	28	GGGGATTCTGACTCCATTTTAC	CACCTTCCCACATTCCCAC

Gene	*Exon*	*Forward*	*Reverse*
SMC2	**2**	TTTGTGGCCTGTTTGATTCC	GGGACGTCTCCAAAGTTCAG
	3	TGCTGCTTTGTACGGTAGGG	ACGTTTTGCAAAGGCATAGG
	4	TGACCATGTCATACTTGTTCCAC	CAGTGCAAGTAACTAAACCATATTTC
	5	TCAGTCTAGATAAACATGAATGGC	AATCATTCCCAAATAATGTCTCC
	6-7	AAATTCCAGTTGGACACATTG	TTTTGGATACATTGCTGACTTTG
	8	TGACTTCTGATGTTACGTAAAGTTTC	GAAGACAATGCCCAATGAATC
	9	CATACAATAAAATGACAGCGTGAAC	GCCCAAACAACTGTACTTAGCTTATAG
	10	GCTCACTCCAAATTGCTATGG	AAACAAACAAGTTCCTTGAGACTG
	11	AACCAATCTGAATCTTCGGC	AATGTTTCCTACTTGCTTGCG
	12	TGCTTTGGAAATAGTGGTTGC	AATTTTCACACTGCTTCCAGC
	13	CAAAAGAGCTCCTGATGAGTAAAG	GCACATGACTTTTCAAGCTAACC
	14	CTTGATTGGGATAAATGGCAC	GGTTCTGCATATGATTATGTCCC
	15	GAAAATTCAGAAAGAGTTTTAATGC	GAAGAAGTGGGTAGTTGGGC
	16	TCACGTAGGCAAGTCCAGTG	CCAGCAACCCTAGCCAAG
	17	AAGGTAAGAATCTGAAGTTTGAATG	GATACCTAAAGTAAGGTAACGGATATG
	18	CCTGTAATTTTCCAAACCATTTC	TCAACCAATGATGCTCTACCAC
	19	TGGTAACAAAGCTGTGGGTC	TGATGACGAGTGAAAAGACTGAG
	20	CTGCCTGTAATCACATTGCC	ATGCCAGGCTGTAATTTTGC
	21	TTAAGAAACTGGCTTATACATATTGG	TGCAAAAGGTTAACTCTAATAATTGC
	22	TGGCTGCTAAGGAGATAAGGG	AAATGGAATTGGGATTTTGTTG
	23	CAAAGTAGACCTGCTGTGTTGC	AAACACAAATACAGTTTTAAGGTAAGG
	24	AGGATCACATATTTGCTTCAGTG	ACAATGCTTTAGTATTTATCAACCC
	25	TGGACAGCTAGCATTCATTTAGG	CCACAATAATTGGACCATAATTAGAC
SMC4	**2**	GTGGACCGAGATTTCCCC	AACTCCGCGCTCAACAAC
	3	TGCACTAATGTAGGCTTTTCCC	AAATAGGGAGTGAATGAATTTTAGG
	4	TGCCTGTGATTAGCAATGAG	AAACCAATGCTTTCCGCTG
	5	TTTGAAGTAGGCCTATGACTTAATACC	TGTAAACCATGATGTGCTGC
	6	GGGAAATGAGTGTGTGTAAGTGG	CATCCTTTTCCTTTTCCACC
	7	GACTTTCATTGTAAATCAGAATGTTTC	ATGGGGTCTTGCTATGTTGC
	8	GGATTTTCCCCTCATAGCC	GGCCACCACTTGATATGCTG
	9	TGTTCCAGGTAGGAAATGTGG	TCCATGAAAACTAGGTATGATATTCAG

Gene	*Exon*	*Forward*	*Reverse*
SMC4	**10**	AAAAGAAGAACTAGAGGACCCTGAG	AGGTTGACTAGTTTTAGGAATGTTG
	11	TGGTTTCATTTCAGTGTAGGTTTG	AAATGCTGACCTCATTCTTTTG
	12	CATACAGGTGCTCAGTTTTGG	CTTGAAATATGGTTGTCAAGGC
	13	TGAATGAAATGGTCCTGTGC	ATGGTGTGAACCCGGGAG
	14	CATGAGATGTCTTCCGTTTAGG	TTGTACCTGGAAAACAGGGG
	15	GGGACAAATCATAGAACAGTCAGG	CCTAGTCATTTATAGAAAACGCCATC
	16	GGGCGACAGAAGTTTTAGTTTTC	CAGTTGAATCCCAATAGTTTCAG
	17	TTCTTGAAGATTTTAAGTGCCC	GCAATGACTTGCATATACTTAGCTTC
	18	CATCACGTCAAGTCATAGCAAC	CTTTCCAATGTGAAGGCAGG
	19	TCTAAAGTTGAATGATACAAGCAGC	ACCAATCTCTCCTCATGCCC
	20	TGTAATGTGAAAATGATGGCTC	AACAATCCTCCCCAAACCC
	21	CACAACTCCTGTTGTTTTCATAGC	AACTAAAGATCAGAATTTAGGCTAAGG
	22	GGCGTCAGGGCAAGACTAC	TGCCTTCACATCCTACCTCC
	23	TGTGAAAAGCAAGTTACAGATTCC	CATGTGTAATGTTTAGGATACCCC
	24	TTCCTGCCATCATTCCTTTAC	TGACAACTGCATAAAACAGAAACC
Sequencing primers for ASPM and PCNT oligonucleotides			
N13		GTAGCGCGACGGCCAGT	CAGGGCGCAGCGATGAC
Sequencing primers for plasmids			
M13		GTAAAACGACGGCCAGT	AACAGCTATGACCATG

LIG4 SNPs primers

dbSNP ID	*Forward*	*Forward*
rs277812	ATATTTCCCCAATGGGCTTC	GGGTAGACATGGTTGAACACAC
rs390619	GCTACATAACTCACCCAAACTGG	AGGCCTGATGGAAAGTCG
rs394610	CCTTCCTCAGAAAGAAATACGTG	TCTATACTTCATAAAGTGGTTGAGGG
rs551103	CTCCTCAGAAAGGCAGCATC	AGGACACTAGTGCTTTGTGGG
rs980044	AGTTTGCTGGATGAAGAGGC	CTCCTCTCTGTGTCTGTCCC
rs1543004	AAGGAAAATATTGACAGTTTACCCTC	TCCAGATGTGCAATTTTAAAGC
rs1560464	GACAGGTTTTGCCATGTTG	TTTCTAGTGGACACAGCCAAG
rs2030736	AAGCTGAAATCGCACCATTG	TGAGATGGGGTTTCTCTCTTG
rs4300490	TGCTACTGGAAGGGATGTAAAC	CTCTTGGAGTGCCCAGTGTC
rs4772857	TTCCCATAGACCCTCTTCTCC	AATTACACAAGCTGGCTCCG
rs6492171	CAGCAAATAAAAGCACATGGTC	CCTAAATATATTGTGGGGAGGAG
rs7322768	AAACAACAGCTAACCGGAGG	CAGCACTTATCACCTGACATCC
rs7989895	CCGCCTATCTGGCTAAAGTG	TGCTACTGAGCATCTTTTGGTATG
rs9301347	TTTCGCATTTATGACAAAACATC	CTGATTTGTGAATGGGACTCTG
rs9520450	TTGGCGTATGACAAAAGTGG	TTTTAATGGTCTTCACAGAGCAG
rs9520986	AGAAAGTTACAAATGAAATAGTGCAG	TCCACACCTCTGGTTTGTATTAG
rs9558977	GCTGCCTTGATTTCTCTTTTG	TTTGATCAGAAAGTTCAAGTTCATTC
rs9559495	GTGGTAGATGGTTGGCAAGG	AGAGGTCATGTCACTGTTTGATG
rs9583185	CACCAGGGTGAATCAGGAAC	TTTAATACGAGAGATAGGCAATTCTAC
rs61965861	CCATCAGTGTCGTTTTCTCTG	TTGGATGGGAAGTTAACCC

cDNA primers

Gene	*Exon*	*Forward*	*Exon*	*Reverse*
NCAPD3	**2**	GCCTTTGGATCCCAGCATAG	**4**	CGATTCAAGTTAGATTCCTGGG
	2	TGACACAGTGTGGGAACTGG	**6**	GGTTAAGAGCTCTGGCTTGTG
	2	TGAAAGCCTTTTACCCTTTGC	**10**	TTGTAAAGCCAGGCAATGAA
CENPJ	**9**	AAAGCTGAGAACGCATCTTTAG	**15**	TTGTACCATCTGGGAAAATGC

Appendix II: Reference sequences

Genomic DNA

Gene	*Reference*	*Gene*	*Reference*
LIG4	NM_002312.3	***ORC1***	NC_000001.9
XRCC4	NM_003401.3	***CENPJ***	NG_009165.1
PCNT	NG_008961.1	***SRCAP***	NG_032135.1
ASPM	NG_015867.1	***MRE11A***	NG_007261.1
IGF1R	NG_009492.1	***BLM***	NG_007272.1
FAT1	NC_000004.11	***ERCC6***	NG_009442.1
USP2	NC_018922.1	***MED12***	NG_012808.1
DONSON	NC_000021.8	***RBM10***	NG_012548.1
NCAPD3	NC_000011.9	***ESCO2***	NG_008117.1
NCAPD2	NC_000012.11	***VPS13B***	NG_007098.2
NCAPG2	NC_000007.13	***PNKP***	NG_027717.1
NCAPG	NC_000004.12	***TUBGCP6***	NG_032160.1
NCAPH	NC_000002.11	***SMC2***	NC_000009.11
NCAPH2	NG_021419.1	***SMC4***	NC_000003.12
IGF2BP2	NG_011602.1		

Nucleotides numbered from first base of initiation codon (ATG) in coding DNA sequence

mRNA

Gene	*Reference*
CENPJ	NM_018451.3
NCAPD3	NM_015261.2

Protein

Species	*Gene*	*Reference*
Homo sapiens	***ORC1***	NP_001177747.1
Pan troglodytes		XP_513408.2
Canis lupus familiaris		XP_005629058.1
Mus musculus		NP_035145.2
Gallus gallus		NP_001026457.1
Danio rerio		NP_956227.1
Homo sapiens	***NCAPH***	NP_056156.2
Pan troglodytes		XP_001148661.1
Canis lupus familiaris		XP_005630442.1
Mus musculus		NP_659067.2
Gallus gallus		NP_001244261.1
Danio rerio		NP_001073665.1
Drosophila melanogaster		NP_477106.1
Schizosaccharomyces pombe		NP_587811.1
Homo sapiens	***XRCC4***	NP_003392.1
Pan troglodytes		NP_001267327.1
Canis lupus familiaris		XP_005618209.1

Species	***Gene***	***Reference***
Bos taurus	***XRCC4***	NP_001075084.1
Mus musculus		NP_082288.1
Gallus gallus		XP_424905.4
Danio rerio		NP_957080.1
Drosophila melanogaster		NP_648316.1
Saccharomyces cerevisiae		NP_011425.1

Appendix III: Script to align Illumina sequencing reads

```
#!/bin/bash -f
#
#Align FASTQ file to reference genome using BWA, further align unmapped reads using Stampy, create sorted bam, remove duplicates, realign indels and create statistics
#Output is indexed bam ready for calling variants
#Fields need to be completed for individual files – add in patient identifier used in FASTQ and associated index #sequence
#
# -- SGE options :
#$ -S /bin/bash
#$ -cwd
#$ -q serial.q
#$ -m bea
#$ -M jennie.murray@igmm.ed.ac.uk

# Configure modules
. /etc/profile.d/modules.sh

# Load required modules
module load apps/gcc/BWA/0.6.2
module load apps/java/hdf_java/2.8
module load apps/gcc/samtools/0.1.16
module load apps/gcc/python/2.7.1
module load apps/gcc/gatk
module load apps/gcc/BEDTools/2.16.2
module load apps/gcc/vcftools/0.1.9

#Set variables
patient="insert patient id"
RG_Index="insert unique index"

#Set reference files **also need ref.stidx * ref.sthash files available for Stampy***
ref=/mnt/lustre2/jmurray/Index/hg19.fa
ref2=/mnt/lustre2/jmurray/Index/hg19_2.fa
GATK=/mnt/lustre2/jmurray/GenomeAnalysisTK-2.4-9-g532efad
indels_1=/mnt/lustre2/jmurray/References/1000G_phase1.indels.hg19.vcf
indel_2=/mnt/lustre2/jmurray/References/Mills_and_1000G_gold_standard.indels.hg19.vcf
snps=/mnt/lustre2/jmurray/References/dbsnp_137.hg19.sorted.vcf

#Use qlogin to login to node and run below commands interactively
#Merg FASTQ files for each direction and remove individual files
```

```
gunzip -c "$patient"_TAGCTT_L00*_R1_*.fastq.gz | gzip -c > "$patient"_1.fastq.gz
gunzip -c "$patient"_"$RG_Index"_L00*_R2_*.fastq.gz | gzip -c > "$patient"_2.fastq.gz
rm "$patient"_"$RG_Index"_L00*_R1_*.fastq.gz
rm "$patient"_"$RG_Index"_L00*_R2_*.fastq.gz

# Run below program using qsub
#
# Map reads to indexed reference
#-q10 removes reads at end with quality (PHRED) score less than 10.
bwa aln -q10 $ref "$patient"_1.fastq.gz > "$patient"_1.sai
bwa aln -q10 $ref "$patient"_2.fastq.gz > "$patient"_2.sai
#
#Unite first and second mapping in SAM format and remove FASTQ and indexes
bwa sampe $ref "$patient"_1.sai "$patient"_2.sai "$patient"_1.fastq.gz "$patient"_2.fastq.gz
> "$patient".bwa.sam

#Clean sam
java -Xmx4g -jar /mnt/lustre2/jmurray/picard-tools-1.79/CleanSam.jar
INPUT=$patient.bwa.sam OUTPUT=$patient.bwa.clean.sam

#Add read group
java -jar /mnt/lustre2/jmurray/picard-tools-1.79/AddOrReplaceReadGroups.jar
INPUT=$patient.bwa.clean.sam O=$patient.rg.bwa.sam RGID="1"
RGLB="DENMARK_Ex" RGPL="ILLUMINA" RGPU="insert unique index"
RGSM="insert patient id"

#Convert sam to bam
samtools view -hbS -o $patient.rg.bwa.bam $patient.rg.bwa.clean.sam

#Align unmapped reads with stampy
python ./stampy-1.0.21/stampy.py --solexa --readgroup="ID:1,SM:insert patient
id,PL:ILLUMINA,PU:insert unique index" --keepreforder -g hg19 -h hg19 --
bamkeepgoodreads -M $patient.rg.bwa.bam > $patient.st.sam

#Convert sam to bam
samtools view -hbS -o $patient.st.bam $patient.st.sam

#Sort bam file
samtools sort $patient.st.bam $patient.sorted

#Index bam
samtools index $patient.sorted.bam

#Mark duplicates with picard
java -Xmx4g -jar /mnt/lustre2/jmurray/picard-tools-1.79/MarkDuplicates.jar
INPUT=$patient.sorted.bam OUTPUT=$patient.rmdup.bam
METRICS_FILE=$patient.rmdup.metrics.txt ASSUME_SORTED=TRUE
VALIDATION_STRINGENCY=SILENT

#Index Bam
samtools index $patient.rmdup.bam

#Mapping statistics
```

```
#bamtools stat -in $patient.rmdup.bam > $patient.markdup.stats.txt
#java -jar /mnt/lustre2/jmurray/picard-tools-1.79/QualityScoreDistribution.jar
I=$patient.rmdup.bam O=$patient.rmdup.scores.txt CHART=$patient.rm.scores.pdf
#java -jar /mnt/lustre2/jmurray/picard-tools-1.79/CollectInsertSizeMetrics.jar
I=$patient.rmdup.bam O=$patient.rmdup.inserts.txt H=$patient.rm.inserts.pdf

#Mapping coverage statistics
bedtools coverage -abam $patient.rmdup.bam -b /mnt/lustre2/jmurray/v4.bed >
$patient.rmdup.targetstats.bed #bait coverage
bedtools coverage -abam $patient.rmdup.bam -b /mnt/lustre2/jmurray/USCS_all_exons.bed
> $patient.rmdup.exonstats.bed #Exon coverage

#Order bam file for gatk using correctly ordered reference *only need to do this if not used
correctly ordered reference in first place
java -jar /mnt/lustre2/jmurray/picard-tools-1.79/ReorderSam.jar INPUT=$patient.rmdup.bam
OUTPUT=$patient.rmdup.order.bam REFERENCE=$ref2

#Index
samtools index $patient.rmdup.order.bam

#Realign indels with gatk *make sure path to correctly ordered reference and indexes
java -Xmx16g -jar $GATK/GenomeAnalysisTK.jar -l INFO -T RealignerTargetCreator -R
$ref2 -known $indels_1 -known $indels_2 -o $patient.intervals -I $patient.rmdup.order.bam

#Determining regions to be realigned
java -Xmx16g -jar $GATK/GenomeAnalysisTK.jar -l INFO -T IndelRealigner -R $ref2 -
known $indels_1 -known $indels_2 -targetIntervals $patient.intervals -I
$patient.rmdup.order.bam -o $patient.realigned.bam

#Calculate recalibration statistics
java -Xmx16g -jar $GATK/GenomeAnalysisTK.jar -l INFO -T BaseRecalibrator -R $ref2 -
knownSites $snps -I $patient.realigned.bam -o $patient.recal.grp

#Recalibrate reads in bam
java -Xmx16g -jar $GATK/GenomeAnalysisTK.jar -l INFO -T PrintReads -R $ref2 -BQSR
$patient.recal.grp -I $patient.realigned.bam -o $patient.recal.bam

#Indicate run finished
echo "Aligning_Complete"
```

Appendix IV: SQL scripts

SNV and indel filtering

Prior to filtering, information obtained for every intragenic variant identified was collated and uploaded into the following tables in an SQL database named `mopd_06_13`. Each unique variant and patient was given a distinct identification number (‘site_id’ and ‘sample_id’ respectively) which was used to link the different tables.

Name of Table	***Description***
site	Genomic position, reference and alternate alleles, allele count and frequency, quality scores and minor allele frequencies in 1KG and EVS databases
consequence	Affected gene name and transcript with corresponding cDNA and protein change for each unique variant and SnpEff consequence prediction
genotype	Genotype of each variant in each individual patient with genotype quality scores, total and allele read depths in each patient
dbnsfp	dbNSFP annotation of each unique non-synonymous coding variant
alamut_output_all	Output of AlamutHT analysis listing all variants predicted to impact on splicing
GeneDataBiomart	Gene name, Ensembl ID, gene description, MIM morbid description, gene status and chromosome number of all genes downloaded from Ensembl_67 database
Pedigree_all	All sample IDs, pedigree number, mothers ID, fathers ID, siblings ID, sex, status, diagnosis and number of family members sequenced in each pedigree (ped_count)
Teir1	List of known disease genes in MPD and 1° MCPH (Appendix V)
Problem_genes	List of noisy genes following NGS as identified by Tennessen *et al.*, (2012)

Stage 1 and 2 filtering

#Annotate consequence table with consequence score

```
#add in `Consequence_score` field (double, Nullable) to dbnsfp table using ‘Alter Table’
tool in SQLyog and populate with corresponding score
UPDATE `dbnsfp` SET `Consequence_score`=2
WHERE (sift_score<0.05 OR `polyphen2_hdiv_pred`='D' OR `polyphen2_hdiv_pred`='P'
OR `polyphen2_hvar_pred`='D' OR `polyphen2_hvar_pred`='P'
OR `lrt_pred`='D' OR `mutation_taster_pred`='A' OR `mutation_taster_pred`='D'
OR `mutation_assessor_pred`='high' OR `mutation_assessor_pred`='medium');

UPDATE `dbnsfp` SET `Consequence_score`=3 WHERE `Consequence_score` IS NULL;
```

```
# add in `type_score` field (double, Nullable) to table ‘consequence’
UPDATE `consequence` r, `dbnsfp` d
SET r.`Type_score`=d.`Consequence_score` WHERE r.site_id=d.`site_id`;
UPDATE `consequence` SET `Type_score`=1.0
WHERE (`type`='STOP_GAINED' OR `type`='FRAME_SHIFT' OR
`type`='SPLICE_SITE_ACCEPTOR' OR `type`='SPLICE_SITE_DONOR' OR
`type`='START_LOST');

UPDATE `consequence` SET `Type_score`=2.0
WHERE (`type`='CODON_CHANGE_PLUS_CODON_INSERTION'
OR `type`='CODON_INSERTION' OR `type`='CODON_DELETION' OR
`type`='CODON_CHANGE_PLUS_CODON_DELETION')

UPDATE `consequence` SET Type_score=4.0 WHERE ((`type`='INTRON' or
`type`='STOP_LOST') and type_score is null);

UPDATE `consequence` SET Type_score=5.0
WHERE ((`type`='SYNONYMOUS_CODING' OR `type`='START_GAINED') and
type_score is null));

UPDATE `consequence` SET Type_score=6.0
WHERE (`type`='UTR_5_PRIME' OR `type`='UTR_3_PRIME');

UPDATE `consequence` SET Type_score=7.0
WHERE (`type`='UPSTREAM' OR `type`='DOWNSTREAM' OR
`type`='SYNONYMOUS_STOP'
OR `type`='EXON');

#Add in field ‘type2_score’ (double, Nullable) to annotate intronic and synonymous variants
with splice prediction
UPDATE `consequence` c, `alamut_output_all` a SET c.`type2_score`=1
WHERE c.`site_id`=a.`id` AND c.`gene_name`=a.`gene_name`;

UPDATE `consequence` SET type_score=2.5 WHERE type2_score=1 AND (type_score=3
OR type_score=4 OR type_score=5);
```

#Filter for rare variants (Stage 1)

```
#Create table of all rare variants
CREATE TABLE `Variants_maf_005` SELECT `site_id`, `1kg_maf`, `evs_maf` FROM
`site` WHERE (`site`.`1kg_maf` <0.005 OR `site`.`1kg_maf` IS NULL) AND
(`site`.`evs_maf` <0.005 OR `site`.`evs_maf` IS NULL);

#Add in field for gene_name (varchar, 255) to allow counting of genes with rare variants
UPDATE `Variants_maf_005` v, `consequence` c
SET v.`gene_name`=c.`gene_name` WHERE v.`site`=c.`site_id`;
#change `site_id` to `site` in table `Variants_maf_005`
```

```
#Create table of all rare variants annotated with associated consequence and consequence
score
CREATE TABLE Variants_005_consequence SELECT
 `Variants_maf_005`.`site_id`
 , `Variants_maf_005`.`1kg_maf`
 , `Variants_maf_005`.`evs_maf`
 , `consequence`.`gene_name`
 , `consequence`.`transcript`
 , `consequence`.`type`
 , `consequence`.`type_score`
FROM `mopd_06_13`.`Variants_maf_005`
INNER JOIN `mopd_06_13`.`consequence`
ON (`Variants_maf_005`.`site_id` = `consequence`.`site_id`);
```

#Filter for variants occurring in less than 6 families (Stage 1)

```
#Create table with number of families site_id occurred in
CREATE TABLE `Variants_005_genotype` SELECT
`Variants_maf_005`.`site_id`
, `Variants_maf_005`.`gene_name`
, `genotype`.`genotype`
, `Pedigree_all`.`Pedigree_Name`
, `Pedigree_all`.`Individual_ID`
, `Pedigree_all`.`Affection_status`
FROM
`mopd_06_13`.`Variants_maf_005`
INNER JOIN `mopd_06_13`.`genotype`
ON (`genotype`.`site_id` = `Variants_maf_005`.`site_id`)
INNER JOIN `mopd_06_13`.`Pedigree_all`
ON (`genotype`.`sample_id` = `Pedigree_all`.`sample_id`);

#Count number of families in which each variant was identified
CREATE VIEW occurance_count AS (SELECT site_id, COUNT(DISTINCT
Pedigree_Name)
FROM Variants_005_genotype GROUP BY site_id);

#Select all rare variants occurring in less than 6 families
CREATE TABLE occurance_count_under6FN SELECT *
FROM `occurance_count` WHERE (`count(distinct Pedigree_Name)` <6);

#Add in new field `site_count` (varchar, 5) to `Variants_005_consequence` and
Variants_005_genotype tables and populate with occurrance scores
UPDATE `Variants_005_consequence` g, `occurance_count_under6FN` o
SET g.`site_count`=o.`COUNT(DISTINCT Pedigree_Name)` WHERE
g.`site_id`=o.`site_id`;

UPDATE `Variants_005_genotype` g, `occurance_count_under6FN` oSET
g.`site_count`=o.`COUNT(DISTINCT Pedigree_Name)` WHERE g.`site_id`=o.`site_id`;
```

```
#Select those with occurrance score of less than 6
CREATE TABLE `Variants_005_6FN_consequence`
SELECT * FROM `Variants_005_consequence` WHERE `site_count` IS NOT NULL;

CREATE TABLE `Variants_005_6FN_genotype`
SELECT * FROM `Variants_005_genotype` WHERE `site_count` IS NOT NULL;
# index `site_id` in both tables using ‘Alter Table’ tool in SQLyog

#Merg `Variants_005_6FN_genotype` and `Variants_005_6FN_consequence` tables to
create table with fields required for further filtering
CREATE TABLE `Variants_005_6FN` SELECT
`Variants_005_6FN_consequence`.`site_id`
, `Variants_005_6FN_consequence`.`1kg_maf`
, `Variants_005_6FN_consequence`.`evs_maf`
, `Variants_005_6FN_genotype`.`Pedigree_Name`
, `Variants_005_6FN_genotype`.`Individual_ID`
, `Variants_005_6FN_genotype`.`Affection_status`
, `Variants_005_6FN_genotype`.`site_count`
, `Variants_005_6FN_consequence`.`gene_name`
, `Variants_005_6FN_consequence`.`transcript`
, `Variants_005_6FN_consequence`.`type`
, `Variants_005_6FN_consequence`.`type_score`
, `Variants_005_6FN_genotype`.`genotype`
FROM
`mopd_06_13`.`Variants_005_6FN_consequence`
INNER JOIN `mopd_06_13`.`Variants_005_6FN_genotype`
ON (`Variants_005_6FN_consequence`.`site_id` = `Variants_005_6FN_genotype`.`site_id`);
```

#Identify variants homozygous in unaffected individuals (Stage 1)

```
#Create table of all homozygous variants occurring in unaffected samples
CREATE TABLE HOM_variants_unaffected
SELECT site_id FROM `Variants_005_6FN`
WHERE `Affection_status`=1 AND genotype='1/1';

#Add in `Hom_unaffected` field (double, Nullable) to `Variants_005_6FN`
UPDATE `Variants_005_6FN` h, `Hom_variants_unaffected` v
SET h.`Hom_unaffected`=2 WHERE h.`site_id`=v.`site_id`;

UPDATE `Variants_005_6FN` SET `Hom_unaffected`=1 WHERE `Hom_unaffected` IS
NULL
```

#Select deleterious variants (Stage 2)

```
CREATE TABLE `Variants_all_deleterious` SELECT * FROM `Variants_005_6FN`
WHERE `type_score`<3 and `Hom_unaffected`=1;
```

Identifying mutations in known MPD and 1° MCPH genes

```
#Create table of all rare, deleterious variants in known MPD and 1° MCPH genes in
#affected samples
CREATE TABLE Tier_1_variants SELECT
`Variants_all_deleterious`.`site_id`
, `Variants_all_deleterious`.`1kg_maf`
, `Variants_all_deleterious`.`evs_maf`
, `Variants_all_deleterious`.`Pedigree_Name`
, `Variants_all_deleterious`.`Individual_ID`
, `Variants_all_deleterious`.`Affection_status`
, `Variants_all_deleterious`.`site_count`
, `Variants_all_deleterious`.`gene_name`
, `Variants_all_deleterious`.`transcript`
, `Variants_all_deleterious`.`type_score`
, `Variants_all_deleterious`.`Hom_unaffected`
FROM
   `mopd_06_13`.`Variants_all_deleterious `
   INNER JOIN `mopd_06_13`.`Teir1`
      ON (`Variants_all_deleterious`.`gene_name` = `Teir1`.`Gene_name`) WHERE
`Variants_all_deleterious`.`Affection_status`=2;

#Manually review variants using IGV and Alamut and manually update diagnosis #field in
`Pedigree_all` table with gene_name in which causal variants identified #(also insert
diagnosis for all samples within the same pedigree)

#Add `diagnosis` field (varchar, 255) to `Variants_all_deleterious` table and update #with
diagnosis
UPDATE `Variants_all_deleterious` v, `Pedigree_all` p SET v.`diagnosis`=p.`diagnosis`
WHERE v.`pedigree_name`=p`Pedigree_name`;

#Remove those where diagnosis identified
CREATE TABLE `Variants_all_deleterious_unknown` SELECT * FROM
`Variants_all_deleterious` WHERE `diagnosis` IS NULL;
```

Mode of inheritance analysis

```
#Add field `ped_count` (double, Nullable) to `Variants_all_deleterious_unknown` #and
update to identify number of family members sequenced
UPDATE `Variants_all_deleterious_unknown` v, `Pedigree_all` p SET
v.`ped_count`=p.`ped_count` WHERE v.`pedigree_name`=p.`pedigree_name`;
```

#Autosomal recessive analysis

```
#Create individual tables of heterozygous and homozygous variants in all affected and
unaffected samples
CREATE TABLE `JM_Het_affected` SELECT * FROM
`Variants_all_deleterious_unknown`
WHERE genotype!='1/1' AND `Affection_status`=2 ;
```

```
CREATE TABLE `JM_Hom_affected` SELECT * FROM
`Variants_all_deleterious_unknown`
WHERE genotype='1/1' AND `Affection_status`=2;

CREATE TABLE `JM_Het_unaffected` SELECT * FROM
`Variants_all_deleterious_unknown`
WHERE `genotype`!='1/1' AND `Affection_status`=1 ;

CREATE TABLE `JM_Hom_unaffected` SELECT * FROM
`Variants_all_deleterious_unknown`
WHERE `genotype`='1/1' AND `Affection_status`=1;
#Index all fields in above four tables

#Count number of unique variants per gene per affected sample
CREATE VIEW `HET_allele_count` AS
(SELECT `Individual_ID`, gene_name, COUNT(DISTINCT site_id)
FROM `JM_Het_affected` GROUP BY `Individual_ID`, gene_name ) ;

#Select genes with more than one heterozygous variant present in affected individual
CREATE TABLE `COMP_HET_allele` SELECT * FROM `HET_allele_count` WHERE
`COUNT(DISTINCT site_id)`>1;

#Create table of recessive heterozygous variants in affecteds
CREATE TABLE `JM_COMP_HET_affected` SELECT
JM_Het_affected`.`site_id`
, `JM_Het_affected`.`1kg_maf`
, `JM_Het_affected`.`evs_maf`
, `JM_Het_affected`.`Pedigree_Name`
, `JM_Het_affected`.`Individual_ID`
, `JM_Het_affected`.`Affection_status`
, `JM_Het_affected`.`site_count`
, `JM_Het_affected`.`gene_name`
, `JM_Het_affected`.`transcript`
, `JM_Het_affected`.`type_score`
, `JM_Het_affected`.`Hom_unaffected`
, `JM_Het_affected`.`diagnosis`
, `JM_Het_affected`.`ped_count`
FROM `JM_Het_affected`
JOIN `COMP_HET_allele` ON `JM_Het_affected`.`individual_ID`=
`COMP_HET_allele`.`Individual_ID`
AND `JM_Het_affected`.`gene_name`= `COMP_HET_allele`.`gene_name`;

#Create table of all recessive variants in affecteds
CREATE TABLE `JM_recessive_affected`
(SELECT * FROM `JM_COMP_HET_affected`)
UNION ALL (SELECT * FROM `JM_Hom_affected`);

#Create table of all recessive variants in singletons
CREATE TABLE `JM_recessive_singletons_final` SELECT * FROM
`JM_recessive_affected` WHERE `ped_count`=1;
```

```
#Select only variants in common between sib pairs
#Create table of all homozygous variants in pairs
CREATE TABLE `JM_hom_sibs` SELECT * FROM `JM_hom_affected` WHERE
(`ped_count`=2 OR `ped_count`=4);

#Identify shared homozygous variants in pairs
CREATE VIEW `JM_hom_sib_count` AS (SELECT `site_id`, `pedigree_name`,
COUNT(DISTINCT `individual_ID`) FROM `JM_hom_sibs` GROUP BY `site_id`,
`pedigree_name`);

CREATE TABLE `JM_hom_sib_sites` SELECT * FROM `JM_hom_sib_count` WHERE
`COUNT(DISTINCT `individual_ID`)`=2;

#Create table of shared homozygous variants in pairs
CREATE TABLE `JM_hom_biallelc_pairs` SELECT
JM_hom_sibs`.`site_id`
,`JM_hom_sibs`.`1kg_maf`
, `JM_hom_sibs`.`evs_maf`
, `JM_hom_sibs`.`Pedigree_Name`
, `JM_hom_sibs`.`Individual_ID`
, `JM_hom_sibs`.`Affection_status`
, `JM_hom_sibs`.`site_count`
, `JM_hom_sibs`.`gene_name`
, `JM_hom_sibs`.`transcript`
, `JM_hom_sibs`.`type_score`
, `JM_hom_sibs`.`Hom_unaffected`
, `JM_hom_sibs`.`diagnosis`
, `JM_hom_sibs`.`ped_count`
FROM `JM_hom_sibs`
JOIN `JM_hom_sibs_sites` ON `JM_hom_sibs`.`site_id`= `JM_hom_sib_sites`.`site_id`
AND `JM_hom_sibs`.`pedigree_name`= `JM_hom_sib_sites`.`pedigree_name`;

#Create table of all heterozygous variants in pairs
CREATE TABLE `JM_het_sibs` SELECT * FROM `JM_recessive_affected` WHERE
(`ped_count`=2 OR `ped_count`=4) AND `genotype`!=’1/1’;

#Identify shared heterozygous variants in pairs
CREATE VIEW `JM_het_sib_count` AS (SELECT `site_id`, `pedigree_name`,
`gene_name`, COUNT(DISTINCT `individual_ID`) FROM `JM_het_sibs` GROUP BY
`site_id`, `pedigree_name`);

CREATE TABLE `JM_het_sib_sites` SELECT * FROM `JM_het_sib_count` WHERE
`COUNT(DISTINCT `individual_ID`)`=2;
```

```
#Create table of shared heterozygous variants in pairs
CREATE TABLE `JM_het_shared_pairs` SELECT
JM_het_sibs`.`site_id`
,`JM_het_sibs`.`1kg_maf`
, `JM_het_sibs`.`evs_maf`
, `JM_het_sibs`.`Pedigree_Name`
, `JM_het_sibs`.`Individual_ID`
, `JM_het_sibs`.`Affection_status`
, `JM_het_sibs`.`site_count`
, `JM_het_sibs`.`gene_name`
, `JM_het_sibs`.`transcript`
, `JM_het_sibs`.`type_score`
, `JM_het_sibs`.`Hom_unaffected`
, `JM_het_sibs`.`diagnosis`
, `JM_het_sibs`.`ped_count`
FROM `JM_het_sibs`
JOIN `JM_het_sibs_sites` ON `JM_het_sibs`.`site_id`= `JM_het_sib_sites`.`site_id`
AND `JM_het_sibs`.`pedigree_name`= `JM_het_sib_sites`.`pedigree_name`;

#Identify shared heterozygous variants in pairs where more than one variant per gene
CREATE VIEW `JM_het_sib_gene_count` AS (SELECT `pedigree_name`, `gene_name`,
COUNT(DISTINCT `site_id`) FROM `JM_het_sibs_sites` GROUP BY `pedigree_name`,
`gene_name`);

CREATE TABLE `JM_het_sib_common` SELECT * FROM `JM_het_sib_gene_count`
WHERE `COUNT(DISTINCT `site_id`)`>1;

#Create table of all shared recessive heterozygous variants in pairs
CREATE TABLE `JM_het_biallelc_pairs` SELECT
`JM_het_shared_pairs`.`site_id`
, `JM_het_shared_pairs`.`1kg_maf`
, `JM_het_shared_pairs`.`evs_maf`
, `JM_het_shared_pairs`.`Pedigree_Name`
, `JM_het_shared_pairs`.`Individual_ID`
, `JM_het_shared_pairs`.`Affection_status`
, `JM_het_shared_pairs`.`site_count`
, `JM_het_shared_pairs`.`gene_name`
, `JM_het_shared_pairs`.`transcript`
, `JM_het_shared_pairs`.`type_score`
, `JM_het_shared_pairs`.`Hom_unaffected`
, `JM_het_shared_pairs`.`diagnosis`
, `JM_het_shared_pairs`.`ped_count`
FROM `JM_het_shared_pairs`
JOIN JM_het_sib_common ON `JM_het_shared_pairs`.`pedigree_name`=
`JM_het_sib_common`.`pedigree_name`
AND `JM_het_shared_pairs`.`gene_name`= JM_het_sib_common`.`gene_name`;

#Create table of all shared recessive variants in sibling pairs
CREATE TABLE `JM_biallelic_pairs_final` (SELECT * FROM `JM_het_biallelc_pairs`
WHERE `ped_count`=2) UNION ALL (SELECT * FROM `JM_hom_biallelc_pairs`
WHERE `ped_count`=2);
```

```
#Identify biallelic homozygous variants in trios
#Create table of all homozygous variants in affected of trio and heterozygous variants in
parents
CREATE TABLE JM_hom_trio (SELECT * FROM `JM_Hom_affected` WHERE
`ped_count`=3) UNION ALL (SELECT * FROM `JM_Het_unaffected` WHERE
`ped_count`=3);

#Count number of family members in which each variant occurs and select those occurring
in all three members
CREATE VIEW `JM_hom_trio_count` AS (SELECT `site_id`, `pedigree_name`,
COUNT(DISTINCT `individual_ID`) FROM `JM_hom_trio` GROUP BY `site_id`,
`pedigree_name`);

CREATE TABLE `JM_hom_trio_sites` SELECT * FROM `JM_hom_trio_count` WHERE
`COUNT(DISTINCT `individual_ID`)`=3;

#Create table of all homozygous biallelic variants in affected
CREATE TABLE `JM_hom_biallelc_trios` SELECT
JM_hom_trio`.`site_id`
,`JM_hom_trio`.`1kg_maf`
, `JM_hom_trio`.`evs_maf`
, `JM_hom_trio`.`Pedigree_Name`
, `JM_hom_trio`.`Individual_ID`
, `JM_hom_trio`.`Affection_status`
, `JM_hom_trio`.`site_count`
, `JM_hom_trio`.`gene_name`
, `JM_hom_trio`.`transcript`
, `JM_hom_trio`.`type_score`
, `JM_hom_trio`.`Hom_unaffected`
, `JM_hom_trio`.`diagnosis`
, `JM_hom_trio`.`ped_count`
FROM `JM_hom_trio`
JOIN `JM_hom_trio_sites` ON `JM_hom_trio`.`site_id`= `JM_hom_trio_sites`.`site_id`
AND `JM_hom_trio`.`pedigree_name`= `JM_hom_trio_sites`.`pedigree_name` WHERE
`JM_hom_trio`.`Affection_status`=2;

#Identify biallelic heterozygous variants in trios
#Create table of all heterozygous recessive variants in affecteds of trios and heterozygous
variants in parents of trios
CREATE TABLE `JM_het_trio` (SELECT * FROM `JM_COMP_HET_affected` WHERE
`ped_count`=3) UNION ALL (SELECT * FROM `JM_Het_unaffected` WHERE
`ped_count`=3);

#Count number of family members in which each variant occurs and select those occurring
in only two members
CREATE VIEW `JM_het_trio_count` AS (SELECT `site_id`, `pedigree_name`,
COUNT(DISTINCT `individual_ID`) FROM `JM_het_trio` GROUP BY `site_id`,
`pedigree_name`);

CREATE TABLE `JM_het_trio_sites` SELECT * FROM `JM_het_trio_count` WHERE
`COUNT(DISTINCT `individual_ID`)`=2;
```

```
#Create table of all inherited heterozygous variants
CREATE TABLE `JM_het_interim_trio` SELECT
JM_het_trio`.`site_id`
,`JM_het_trio`.`1kg_maf`
, `JM_het_trio`.`evs_maf`
, `JM_het_trio`.`Pedigree_Name`
, `JM_het_trio`.`Individual_ID`
, `JM_het_trio`.`Affection_status`
, `JM_het_trio`.`site_count`
, `JM_het_trio`.`gene_name`
, `JM_het_trio`.`transcript`
, `JM_het_trio`.`type_score`
, `JM_het_trio`.`Hom_unaffected`
, `JM_het_trio`.`diagnosis`
, `JM_het_trio`.`ped_count`
FROM `JM_het_trio`
JOIN `JM_het_trio_sites` ON `JM_het_trio`.`site_id`= `JM_het_trio_sites`.`site_id`
AND `JM_het_trio`.`pedigree_name`= `JM_het_trio_sites`.`pedigree_name`;

#Create table listing genes in which inherited variants occur from both parents
CREATE VIEW `trio_biallelic_genes` AS (SELECT DISTINCT `pedigree_name`,
`gene_name`, COUNT(DISTINCT `individual_ID`)
FROM `JM_het_interim_trio` GROUP BY `pedigree_name`, `gene_name`);

CREATE TABLE `trio_biallelic_genes_t` SELECT * FROM `trio_biallelic_genes` WHERE
`COUNT(DISTINCT individual_ID)`=3;

#Create table of all heterozygous biallelic variants in affecteds in trios
CREATE TABLE `JM_het_biallelic_trios` SELECT
JM_het_interim_trio`.`site_id`
,`JM_het_interim_trio`.`1kg_maf`
, `JM_het_interim_trio`.`evs_maf`
, `JM_het_inerim_trio`.`Pedigree_Name`
, `JM_het_interim_trio`.`Individual_ID`
, `JM_het_interim_trio`.`Affection_status`
, `JM_het_interim_trio`.`site_count`
, `JM_het_interim_trio`.`gene_name`
, `JM_het_interim_trio`.`transcript`
, `JM_het_interim_trio`.`type_score`
, `JM_het_interim_trio`.`Hom_unaffected`
, `JM_het_interim_trio`.`diagnosis`
, `JM_het_interim_trio`.`ped_count`
FROM `JM_het_interim_trio`
JOIN `trio_biallelic_genes_t` ON
`JM_het_interim_trio`.`pedigree_name`=`trio_biallelic_genes_t`.`pedigree_name` AND
`JM_het_interim_trio`.`gene_name`=`trio_biallelic_genes_t`.`gene_name` WHERE
`JM_het_interim_trio`.`Affection_status`=2;

#Create table of all biallelic variants in trios
CREATE TABLE `JM_biallelic_trios_final` (SELECT * FROM `JM_het_biallelc_trios`)
UNION ALL (SELECT * FROM `JM_hom_biallelc_trios`);
```

```
#Identify biallelic homozygous variants in quads
#Create table of all shared homozygous variants in affected sibs of quad and heterozygous
variants in parents
CREATE TABLE JM_hom_quad (SELECT * FROM `JM_hom_biallelc_pairs` WHERE
`ped_count`=4) UNION ALL (SELECT * FROM `JM_Het_unaffected` WHERE
`ped_count`=4);

#Count number of family members in which each variant occurs and select those occurring
in all four members
CREATE VIEW `JM_hom_quad_count` AS (SELECT `site_id`, `pedigree_name`,
COUNT(DISTINCT `individual_ID`) FROM `JM_hom_quad` GROUP BY `site_id`,
`pedigree_name`);

CREATE TABLE `JM_hom_quad_sites` SELECT * FROM `JM_hom_quad_count`
WHERE `COUNT(DISTINCT `individual_ID`)`=4;

#Create table of all homozygous biallelic variants in affected
CREATE TABLE `JM_hom_biallelc_quad` SELECT
JM_hom_quad`.`site_id`
,`JM_hom_quad`.`1kg_maf`
, `JM_hom_quad.`evs_maf`
, `JM_hom_quad`.`Pedigree_Name`
, `JM_hom_quad`.`Individual_ID`
, `JM_hom_quad`.`Affection_status`
, `JM_hom_quad`.`site_count`
, `JM_hom_quad`.`gene_name`
, `JM_hom_quad`.`transcript`
, `JM_hom_quad`.`type_score`
, `JM_hom_quad`.`Hom_unaffected`
, `JM_hom_quad`.`diagnosis`
, `JM_hom_quad`.`ped_count`
FROM `JM_hom_quad`
JOIN `JM_hom_quad_sites` ON `JM_hom_qad`.`site_id`= `JM_hom_quad_sites`.`site_id`
AND `JM_hom_quad`.`pedigree_name`= `JM_hom_quad_sites`.`pedigree_name` where
`JM_hom_quad`.`Affection_status`=2;

#Identify biallelic heterozygous variants in quads
#Create table of all shared heterozygous recessive variants in affecteds of quads and
heterozygous variants in parents of quads
CREATE TABLE `JM_het_quad` (SELECT * FROM `JM_het_biallelc_pairs` WHERE
`ped_count`=4) UNION ALL (SELECT * FROM `JM_Het_unaffected` WHERE
`ped_count`=4);

#Count number of family members in which each variant occurs and select those occurring
in only three members
CREATE VIEW `JM_het_quad_count` AS (SELECT `site_id`, `pedigree_name`,
COUNT(DISTINCT `individual_ID`) FROM `JM_het_quad` GROUP BY `site_id`,
`pedigree_name`);

CREATE TABLE `JM_het_quad_sites` SELECT * FROM `JM_het_quad_count` WHERE
`COUNT(DISTINCT `individual_ID`)`=3;
```

```
#Create table of all inherited heterozygous variants
CREATE TABLE `JM_het_interim_quad` SELECT
JM_het_quad`.`site_id`
,`JM_het_quad`.`1kg_maf`
, `JM_het_quad`.`evs_maf`
, `JM_het_quad`.`Pedigree_Name`
, `JM_het_quad`.`Individual_ID`
, `JM_het_quad`.`Affection_status`
, `JM_het_quad`.`site_count`
, `JM_het_quad`.`gene_name`
, `JM_het_quad`.`transcript`
, `JM_het_quad`.`type_score`
, `JM_het_quad`.`Hom_unaffected`
, `JM_het_quad`.`diagnosis`
, `JM_het_quad`.`ped_count`
FROM `JM_het_quad`
JOIN `JM_het_quad_sites` ON `JM_het_quad`.`site_id`= `JM_het_quad_sites`.`site_id`
AND `JM_het_quad`.`pedigree_name`= `JM_het_quad_sites`.`pedigree_name`;

#Create table listing genes in which inherited variants occur from both parents
CREATE VIEW `quad_biallelic_genes` AS (SELECT DISTINCT `pedigree_name`,
`gene_name`, COUNT(DISTINCT `individual_ID`)
FROM `JM_het_interim_quad` GROUP BY `pedigree_name`, `gene_name`);

CREATE TABLE `quad_biallelic_genes_t` SELECT * FROM `quad_biallelic_genes`
WHERE `COUNT(DISTINCT individual_ID)`=4;

#Create table of all heterozygous biallelic variants in affecteds in quads
CREATE TABLE `JM_het_biallelic_quad` SELECT
JM_het_interim_quad`.`site_id`
,`JM_het_interim_quad`.`1kg_maf`
, `JM_het_interim_quad`.`evs_maf`
, `JM_het_inerim_quad`.`Pedigree_Name`
, `JM_het_interim_quad`.`Individual_ID`
, `JM_het_interim_quad`.`Affection_status`
, `JM_het_interim_quad`.`site_count`
, `JM_het_interim_quad`.`gene_name`
, `JM_het_interim_quad`.`transcript`
, `JM_het_interim_quad`.`type_score`
, `JM_het_interim_quad`.`Hom_unaffected`
, `JM_het_interim_quad`.`diagnosis`
, `JM_het_interim_quad`.`ped_count`
FROM `JM_het_interim_quad`
JOIN `quad_biallelic_genes_t` ON
`JM_het_interim_quad`.`pedigree_name`=`quad_biallelic_genes_t`.`pedigree_name` AND
`JM_het_interim_quad`.`gene_name`=`quad_biallelic_genes_t`.`gene_name` WHERE
`JM_het_interim_quad`.`Affection_status`=2;

#Create table of all shared biallelic variants in quads
CREATE TABLE `JM_biallelic_quad_final` (SELECT * FROM `JM_het_biallelc_quad`)
UNION ALL (SELECT * FROM `JM_hom_biallelc_quad`);
```

```
#Merg all biallelic variant tables
CREATE TABLE `JM_AR_final` SELECT * FROM
`JM_recessive_singletons_final`;

INSERT INTO `JM_AR_final` SELECT * FROM `JM_biallelic_pairs_final`;

INSERT INTO `JM_AR_final` SELECT * FROM
`JM_biallelic_trios_final`;

INSERT INTO `JM_AR_final` SELECT * FROM
`JM_biallelic_quads_final`;
```

#Autosomal dominant (*de novo*) analysis

```
#Create table of all possible autosomal dominant variants
CREATE TABLE JM_AD_all_affected SELECT * FROM `JM_Het_affected` WHERE
`1kg_maf` IS NULL and `evs_maf` IS NULL;

#Create table of all AD variants in singletons
CREATE TABLE `JM_AD_singletons` SELECT * FROM `JM_AD_all_affected` WHERE
`ped_count`=1;

#Create table of all AD variants in singletons where consequence score =1
CREATE TABLE `JM_AD_singletons_1` SELECT * FROM `JM_AD_singletons`
WHERE `type_score`=1;

#Create table of all AD variants in affecteds in trios
CREATE TABLE `JM_AD_trios` SELECT * FROM `JM_AD_all_affected` WHERE
`ped_count`=3;

#Create table of all novel variants in unaffecteds in trios
CREATE TABLE `Variants_novel_trio_parents` SELECT * FROM
`Variants_all_deleterious_unknown` WHERE (`1kg_maf` IS NULL AND `evs_maf` IS
NULL AND `Affection_status`=1 AND `ped_count`=3);

#Merg above two tables
CREATE TABLE `Variants_all_trio` (SELECT * FROM `Variants_novel_trio_parents`)
UNION ALL (SELECT * FROM `JM_AD_trios`);

#Count number of family members in which each variant occurs
CREATE VIEW `Trio_denovo_sites_count` AS
(SELECT `site_id`, `pedigree_name`, COUNT(DISTINCT `Individual_id`) FROM
`Variants_all_trio` GROUP BY `site_id`, `pedigree_name`);

#Create table of variants which only occur in one trio member
CREATE TABLE Trio_denovo_sites SELECT * FROM Trio_denovo_sites_count WHERE
`COUNT(DISTINCT Individual_id)`=1;
```

```
#Create table of de novo variants in affecteds of trios
CREATE TABLE JM_denovo SELECT
`Variants_all_trio`.`site_id`
, `Variants_all_trio`.`1kg_maf`
, `Variants_all_trio`.`evs_maf`
, `Variants_all_trio`.`Pedigree_Name`
, `Variants_all_trio`.`Individual_ID`
, `Variants_all_trio`.`Affection_status`
, `Variants_all_trio`.`site_count`
, `Variants_all_trio`.`gene_name`
, `Variants_all_trio`.`transcript`
, `Variants_all_trio`.`type_score`
, `Variants_all_trio`.`Hom_unaffected`
, `Variants_all_trio`.`diagnosis`
, `Variants_all_trio`.`ped_count`
FROM
`mopd_06_13`.`Variants_all_trio`
INNER JOIN `mopd_06_13`. `Trio_denovo_sites`
ON (`Variants_all_trio`.`site_id` = ` Trio_denovo_sites`.`site_id`) WHERE
`Variants_all_trio`.`Affection_status`=2;

#Create final table of all AD variants
CREATE TABLE `JM_AD_final`
(SELECT * FROM `JM_denovo`)
UNION ALL (SELECT * FROM `JM_AD_singletons`)
```

#X-linked recessive analysis

```
#Create table of all variants in X chromosome
CREATE TABLE `chrX_sites` SELECT `site_id` FROM `site` WHERE `chr`='chrX';

#Select all variants on ChrX present in affected in homozgous state
CREATE TABLE `JM_hom_affected_X` SELECT DISTINCT
`JM_Hom_affected`.`site_id`
, `JM_Hom_affected`.`1kg_maf`
, `JM_Hom_affected`.`evs_maf`
, `JM_Hom_affected`.`Pedigree_Name`
, `JM_Hom_affected`.`Individual_ID`
, `JM_Hom_affected`.`diagnosis`
, `JM_Hom_affected`.`Affection_status`
, `JM_Hom_affected`.`site_count`
, `JM_Hom_affected`.`gene_name`
, `JM_Hom_affected`.`type`
, `JM_Hom_affected`.`type_score`
, `JM_Hom_affected`.`genotype`
, `JM_Hom_affected`.`Hom_unaffected`
, `JM_Hom_affected`.`Ped_count`
FROM `mopd_06_13`.`JM_Hom_affected`
INNER JOIN `mopd_06_13`.`chrX_sites`
ON (`JM_Hom_affected`.`site_id` = `chrX_sites`.`site_id`);
```

```
#Add field `sex` (double, Nullable) to `JM_hom_affected_X` and populate with gender for
each sample
UPDATE `JM_hom_affected_X` j, `Pedigree_all` p SET j.`sex`=p.`sex` WHERE
j.individual_ID=p.individual_ID;

#Select all variants in affected males only (`sex`=1)
DELETE FROM `JM_hom_affected_X` WHERE `sex`=2;

#Select those in common between sib pair
CREATE VIEW `JMpair_homXcommon` as (SELECT * from `JM_hom_affected_X`
WHERE `ped_count`=2);

CREATE VIEW `JMpair_homXcommon_count` AS (SELECT `site_id`,
COUNT(DISTINCT individual_ID) FROM `JMpair_homXcommon` GROUP BY `site_id`);

CREATE TABLE `JM_pairXvariants` SELECT `site_id` FROM
`JMpair_homXcommon_count` WHERE `count(distinct individual_ID)`=2;

CREATE TABLE `JM_pairX_final` SELECT
`JM_hom_affected_X`.`site_id`
, `JM_hom_affected_X`.`evs_maf`
, `JM_hom_affected_X`.`1kg_maf`
, `JM_hom_affected_X`.`Pedigree_Name`
, `JM_hom_affected_X`.`Individual_ID`
, `JM_hom_affected_X`.`diagnosis`
, `JM_hom_affected_X`.`Affection_status`
, `JM_hom_affected_X`.`site_count`
, `JM_hom_affected_X`.`gene_name`
, `JM_hom_affected_X`.`type`
, `JM_hom_affected_X`.`type_score`
, `JM_hom_affected_X`.`genotype`
, `JM_hom_affected_X`.`Hom_unaffected`
, `JM_hom_affected_X`.`Ped_count`
, `JM_hom_affected_X`.`sex`
FROM `mopd_06_13`.`JM_hom_affected_X`
INNER JOIN `mopd_06_13`.`JM_pairXvariants`
ON (`JM_hom_affected_X`.`site_id` = `JM_pairXvariants`.`site_id`);

#Select those inherited from mother only in trios
#Create table of all variants in unaffecteds in trios
CREATE TABLE `Variants_all_trio_parents` SELECT * FROM
`Variants_all_deleterious_unknown` WHERE (`Affection_status`=1 AND `ped_count`=3);
```

```
#Change field names in `Variants_all_trio_parents` from `Individual_ID` to `
#Individual_IDp` and from `genotype` to `genotypep` using ‘Alter Table’ tool.
#Select all variants on X chromosome in affected where also present in a parent
CREATE TABLE `JM_Xtrios` SELECT
`JM_hom_affected_X`.`site_id`
, `JM_hom_affected_X`.`evs_maf`
, `JM_hom_affected_X`.`1kg_maf`
, `JM_hom_affected_X`.`Pedigree_Name`
, `JM_hom_affected_X`.`Individual_ID`
, `JM_hom_affected_X`.`diagnosis`
, `JM_hom_affected_X`.`Affection_status`
, `JM_hom_affected_X`.`site_count`
, `JM_hom_affected_X`.`gene_name`
, `JM_hom_affected_X`.`type`
, `JM_hom_affected_X`.`type_score`
, `JM_hom_affected_X`.`genotype`
, `JM_hom_affected_X`.`Hom_unaffected`
, `JM_hom_affected_X`.`Ped_count`
, `JM_hom_affected_X`.`sex`
, `Variants_all_trio_parents`.`Individual_IDp`
, `Variants_all_trio_parents`.`genotypep`
FROM `mopd_06_13`.`JM_hom_affected_X`
INNER JOIN `mopd_06_13`.`Variants_all_trio_parents`
ON (`JM_hom_affected_X`.`site_id` = `Variants_all_trio_parents`.`site_id`) AND
(`Variants_all_trio_parents`.`Pedigree_Name` = `JM_hom_affected_X`.`Pedigree_Name`);

#Add in field `sexp` (double, Nullable) to `JM_Xtrios` and populate
UPDATE `JM_Xtrios` j, `Pedigree_all` p SET j.`sexp`=p.`sex` WHERE
j.`individual_Idp`=p.`individual_id`;

#Remove all variants in affected present in father
DELETE FROM `JM_Xtrios` WHERE `sexp`=1;

#Remove all variants in affected where homozygous in the mother
DELETE FROM `JM_Xtrios` WHERE `sexp`=2 and `genotype`='1/1';

#Create table of candidate variants in singletons
CREATE TABLE `JM_Xfinal` SELECT * FROM `JM_hom_affected_X` WHERE
`ped_count`=1;

#Add results from pair and trio analysis to crate final table of possible X-linked recessive
variants
INSERT INTO `JM_Xfinal` SELECT * FROM `JM_pairX_final`;
INSERT INTO `JM_Xfinal` SELECT * FROM `JM_Xtrios`;
```

Identifying mutations in other known disease genes

```
#To tables `JM_Xfinal`, `JM_AD_final` and `JM_AR_final` add in fields
#`MIM_Morbid_Description` (varchar, 255) and `description` (varchar, 255)
#Update new fields in each table with gene data and disease association from #ensemble
database
UPDATE `JM_AR_final` f, GeneDataBiomart g SET
f.`MIM_Morbid_Description`=g.`MIM_Morbid_Description` AND
f. `description`=g.`description` WHERE f.`gene_name`=g.`gene_name`;
#Repeat for `JM_Xfinal` and `JM_AD_final`

#Manually review variants with relevant disease association using IGV and Alamut #and
manually update diagnosis field in `Pedigree_all` table with gene_name in #which causal
variants identified

#Update diagnosis field in final tables
UPDATE `JM_AR_final` f, `Pedigree_all` p SET f.diagnosis=p.diagnosis WHERE
f.`pedigree_name`=p.`pedigree_name`
#Repeat for `JM_Xfinal` and `JM_AD_final`

#Remove variants in families with diagnosis
DELETE FROM `JM_AR_final` WHERE `diagnosis` IS NOT NULL;
#Repeat for `JM_Xfinal` and `JM_AD_final`
```

Prioritising remaining candidate genes

#Remove variants in poorly mapped genes

```
CREATE TABLE JM_AR_minus_noisy_genes SELECT
` JM_AR_final`.`site_id`
, `JM_AR_final`.`1kg_maf`
, `JM_AR_final`.`evs_maf`
, `JM_AR_final`.`Pedigree_Name`
, `JM_AR_final`.`Individual_ID`
, `JM_AR_final`.`diagnosis`
, `JM_AR_final`.`Affection_status`
, `JM_AR_final`.`site_count`
, `JM_AR_final`.`gene_name`
, `JM_AR_final`.`type`
, `JM_AR_final`.`type_score`
, `JM_AR_final`.`genotype`
, `JM_AR_final`.`Hom_unaffected`
, `JM_AR_final`.`Ped_count`
, `JM_AR_final`.`sex`
, `Problem_genes`.`Gene`
FROM
`mopd_06_13`.`JM_AR_final`
INNER JOIN `mopd_06_13`.`Problem_genes`
ON (`JM_AR_final`.`gene_name` = `Problem_genes`.`Gene`) WHERE
`Problem_genes`.`Gene` IS NULL;
#Repeat for `JM_Xfinal` and `JM_AD_final`
```

#Manually remove those functionally unrelated to growth

#Select variants occurring in genes present in more than one family

```
#Count number of families with variants in each gene
CREATE VIEW JM_gene_pedigree_count AS (SELECT DISTINCT `gene_name,
COUNT(DISTINCT pedigree_name) FROM JM_AR_minus_noisy_genes GROUP BY
`gene_name`)
#Repeat for `JM_X_minus_noisy_genes` and `JM_AD_minus_noisy_genes`

CREATE TABLE JM_AR_novel_multiple_peds SELECT
`JM_AR_minus_noisy_genes`.`site_id`
, `JM_AR_minus_noisy_genes`.`1kg_maf`
, `JM_AR_minus_noisy_genes`.`evs_maf`
, `JM_AR_minus_noisy_genes`.`Pedigree_Name`
, `JM_AR_minus_noisy_genes`.`Individual_ID`
, `JM_AR_minus_noisy_genes`.`diagnosis`
, `JM_AR_minus_noisy_genes`.`Affection_status`
, `JM_AR_minus_noisy_genes`.`site_count`
, `JM_AR_minus_noisy_genes`.`gene_name`
, `JM_AR_minus_noisy_genes`.`type`
, `JM_AR_minus_noisy_genes`.`type_score`
, `JM_AR_minus_noisy_genes`.`genotype`
, `JM_AR_minus_noisy_genes`.`Hom_unaffected`
, `JM_AR_minus_noisy_genes`.`Ped_count`
, `JM_AR_minus_noisy_genes`.`sex`
FROM
`mopd_06_13`.`JM_AR_minus_noisy_genes`
INNER JOIN `mopd_06_13`.` JM_gene_pedigree_count` ON
(`JM_AR_minus_noisy_genes`.`gene_name` = `JM_gene_pedigree_count`.`gene_name`)
WHERE `JM_gene_pedigree_count`.`COUNT(DISTINCT pedigree_name)`>1;
#Repeat for `JM_X_minus_noisy_genes` and `JM_AD_minus_noisy_genes `
```

#Select genes where at least one variant has a consequence score=1.

```
#Create table of genes from final tables where one variant has consequence score=1
CREATE TABLE `Novel_AR_genes_1` SELECT DISTINCT `gene_name` FROM
`JM_AR_minus_noisy_genes` WHERE `type_score`=1;
#Repeat for `JM_X_minus_noisy_genes` and `JM_AD_minus_noisy_genes `
```

```
#Select variants in `Novel_AR_genes_1`
CREATE TABLE JM_AR_novel_1 SELECT
`JM_AR_minus_noisy_genes`.`site_id`
, `JM_AR_minus_noisy_genes`.`1kg_maf`
, `JM_AR_minus_noisy_genes`.`evs_maf`
, `JM_AR_minus_noisy_genes`.`Pedigree_Name`
, `JM_AR_minus_noisy_genes`.`Individual_ID`
, `JM_AR_minus_noisy_genes`.`diagnosis`
, `JM_AR_minus_noisy_genes`.`Affection_status`
, `JM_AR_minus_noisy_genes`.`site_count`
, `JM_AR_minus_noisy_genes`.`gene_name`
, `JM_AR_minus_noisy_genes`.`type`
, `JM_AR_minus_noisy_genes`.`type_score`
, `JM_AR_minus_noisy_genes`.`genotype`
, `JM_AR_minus_noisy_genes`.`Hom_unaffected`
, `JM_AR_minus_noisy_genes`.`Ped_count`
, `JM_AR_minus_noisy_genes`.`sex`
FROM
`mopd_06_13`.`JM_AR_minus_noisy_genes`
INNER JOIN `mopd_06_13`.`Novel_AR_genes_1` ON
(`JM_AR_minus_noisy_genes`.`gene_name` = `Novel_AR_genes_1`.`gene_name`);
#Repeat for `JM_X_minus_noisy_genes` and `JM_AD_minus_noisy_genes` joining on
`Novel_X_genes_1` and `Novel_AD_genes_1` respectively
```

CNV filtering

All deleted regions with log2R below -0.7 from ExomeCNV analysis compiled into SQL database in table `JM_CNV_all` with the following fields:

```
`chr` (varchar, 255)
`start` (double)
`end` (double)
`patient` (varchar, 255)
`parent` (varchar, 255)
`diagnosis` (varchar, 255)
#Index fields `patient`, `chr`, `start`, `end` in `JM_CNV_all`
```

#Remove regions occurring in more than 3 patients

```
CREATE VIEW `JM_CNV_count` AS (SELECT DISTINCT chr, `start`, `end`,
COUNT(DISTINCT patient) FROM `JM_CNV_all` GROUP BY chr, `start`, `end`);

CREATE TABLE `JM_CNV_count_less4` SELECT * FROM `JM_CNV_count` WHERE
`count(distinct patient)` <4;
#Index fields `chr`, `start`, `end` in `JM_CNV_count_less4`
```

```
CREATE TABLE JM_CNV_rare SELECT
`JM_CNV_all`.`chr`
, `JM_CNV_all`.`start`
, `JM_CNV_all`.`end`
, `JM_CNV_all`.`count`
, `JM_CNV_all`.`patient`
, `JM_CNV_all`.`parent`
, `JM_CNV_all`.`diagnosis`
FROM
`mopd_06_13`.`JM_CNV_count_less4`
INNER JOIN `mopd_06_13`.`JM_CNV_all`
ON (`JM_CNV_count_less4`.`chr` = `JM_CNV_all`.`chr`) AND
(`JM_CNV_count_less4`.`start` = `JM_CNV_all`.`start`) AND
(`JM_CNV_count_less4`.`end` = `JM_CNV_all`.`end`);
```

#Select regions in patients without diagnosis

```
CREATE TABLE `JM_CNV_rare_unknown` SELECT * FROM `JM_CNV_rare` WHERE
diagnosis IS NULL;
```

Appendix V: List of known MPD and 1° MCPH

Ensembl Gene ID	*Gene*	*Full name*	*MIM Morbid accession Nº*
ENSG00000066279	*ASPM*	asp (abnormal spindle) homolog, microcephaly associated (Drosophila)	608716
ENSG00000175054	*ATR*	ataxia telangiectasia and Rad3 related	210600
ENSG00000164053	*ATRIP*	ATR interacting protein	
ENSG00000147044	*CASK*	calcium/calmodulin-dependent serine protein kinase (MAGUK family)	300749
ENSG00000094804	*CDC6*	cell division cycle 6	613805
ENSG00000136861	*CDK5RAP2*	CDK5 regulatory subunit associated protein 2	604804
ENSG00000167513	*CDT1*	chromatin licensing and DNA replication factor 1	613804
ENSG00000151849	*CENPJ*	centromere protein J	613676/608393
ENSG00000174799	*CEP135*	centrosomal protein 135kDa	614673
ENSG00000103995	*CEP152*	centrosomal protein 152kDa	614852/613823
ENSG00000143702	*CEP170*	centrosomal protein 170kDa	
ENSG00000182923	*CEP63*	centrosomal protein 63kDa	614728
ENSG00000149948	*HMGA2*	high mobility group AT-hook 2	611547
ENSG00000017427	*IGF1*	insulin-like growth factor 1 (somatomedin C)	608747
ENSG00000140443	*IGF1R*	insulin-like growth factor 1 receptor	270450
ENSG00000167244	*IGF2*	insulin-like growth factor 2 (somatomedin A)	180860
ENSG00000073792	*IGF2BP2*	insulin-like growth factor 2 mRNA binding protein 2	-
ENSG00000073111	*MCM2*	minichromosome maintenance complex component 2	-

Ensembl Gene ID	***Gene***	***Full name***	***MIM Morbid accession Nº***
ENSG00000104738	*MCM4*	minichromosome maintenance complex component 4	609981
ENSG00000100297	*MCM5*	minichromosome maintenance complex component 5	-
ENSG00000076003	*MCM6*	minichromosome maintenance complex component 6	-
ENSG00000166508	*MCM7*	minichromosome maintenance complex component 7	-
ENSG00000147316	*MCPH1*	microcephalin 1	251200
ENSG00000072864	*NDE1*	nudE neurodevelopment protein 1	614019
ENSG00000085840	*ORC1*	origin recognition complex, subunit 1	224690
ENSG00000115947	*ORC4*	origin recognition complex, subunit 4	613800
ENSG00000091651	*ORC6*	origin recognition complex, subunit 6	613803
ENSG00000160299	*PCNT*	pericentrin	210720
ENSG00000142731	*PLK4*	polo-like kinase 4	-
ENSG00000101773	*RBBP8*	retinoblastoma binding protein 8	606744
ENSG00000264229	*RNU4ATAC*	RNA, U4atac small nuclear (U12-dependent splicing)	-
ENSG00000123473	*STIL*	SCL/TAL1 interrupting locus	612703
ENSG00000183763	*TRAIP*	TRAF interacting protein	-
ENSG00000075702	*WDR62*	WD repeat domain 62	604317

Appendix VI: Publications during this thesis

Murray, J. E., L. S. Bicknell, G. Yigit, A. L. Duker, M. van Kogelenberg, S. Haghayegh, D. Wieczorek, H. Kayserili, *et al.* (2014). Extreme growth failure is a common presentation of ligase IV deficiency. *Hum Mutat* 35(1): 76-85.

Bober, M. B., T. Niiler, A. L. Duker, J. E. Murray, T. Ketterer, M. E. Harley, S. Alvi, C. Flora, *et al.* (2012). Growth in individuals with Majewski osteodysplastic primordial dwarfism type II caused by pericentrin mutations. *Am J Med Genet A* 158A(11): 2719-2725.

Murray, J. E. and A. P. Jackson (2012). Exploring microcephaly and human brain evolution. *Dev Med Child Neurol* 54(7): 580-581.

Pagnamenta, A. T., J. E. Murray, G. Yoon, E. Sadighi Akha, V. Harrison, L. S. Bicknell, K. Ajilogba, H. Stewart, *et al.* (2012). A novel nonsense CDK5RAP2 mutation in a Somali child with primary microcephaly and sensorineural hearing loss. *Am J Med Genet A* 158A(10): 2577-2582.

Martin, C. A., I. Ahmad, A. Klingseisen, M. S. Hussain, L. S. Bicknell, A. Leitch, G. Nurnberg, M. R. Toliat, J. E. Murray, *et al.* 2014. Mutations in PLK4, encoding a master regulator of centriole biogenesis, cause microcephaly, growth failure and retinopathy. *Nat Genet* 46(12): 1283-1292.

References

Abdel-Salam, G. M., M. S. Abdel-Hamid, M. Issa, A. Magdy, A. El-Kotoury and K. Amr 2012. Expanding the phenotypic and mutational spectrum in microcephalic osteodysplastic primordial dwarfism type I. *Am J Med Genet A* 158A(6): 1455-1461.

Abdel-Salam, G. M., N. Miyake, M. M. Eid, M. S. Abdel-Hamid, N. A. Hassan, O. M. Eid, L. K. Effat, T. H. El-Badry*, et al.* 2011. A homozygous mutation in RNU4ATAC as a cause of microcephalic osteodysplastic primordial dwarfism type I (MOPD I) with associated pigmentary disorder. *Am J Med Genet A* 155A(11): 2885-2896.

Abe, S., K. Nagasaka, Y. Hirayama, H. Kozuka-Hata, M. Oyama, Y. Aoyagi, C. Obuse and T. Hirota 2011. The initial phase of chromosome condensation requires Cdk1-mediated phosphorylation of the CAP-D3 subunit of condensin II. *Genes Dev* 25(8): 863-874.

Abuzzahab, M. J., A. Schneider, A. Goddard, F. Grigorescu, C. Lautier, E. Keller, W. Kiess, J. Klammt*, et al.* 2003. IGF-I receptor mutations resulting in intrauterine and postnatal growth retardation. *N Engl J Med* 349(23): 2211-2222.

Adachi, N., S. Iiizumi, S. So and H. Koyama 2004. Genetic evidence for involvement of two distinct nonhomologous end-joining pathways in repair of topoisomerase II-mediated DNA damage. *Biochem Biophys Res Commun* 318(4): 856-861.

Adachi, N., H. Suzuki, S. Iiizumi and H. Koyama 2003. Hypersensitivity of nonhomologous DNA end-joining mutants to VP-16 and ICRF-193: implications for the repair of topoisomerase II-mediated DNA damage. *J Biol Chem* 278(38): 35897-35902.

Adhikary, S. and M. Eilers 2005. Transcriptional regulation and transformation by Myc proteins. *Nat Rev Mol Cell Biol* 6(8): 635-645.

Ahel, I., U. Rass, S. F. El-Khamisy, S. Katyal, P. M. Clements, P. J. McKinnon, K. W. Caldecott and S. C. West 2006. The neurodegenerative disease protein aprataxin resolves abortive DNA ligation intermediates. *Nature* 443(7112): 713-716.

Ahnesorg, P., P. Smith and S. P. Jackson 2006. XLF interacts with the XRCC4-DNA ligase IV complex to promote DNA nonhomologous end-joining. *Cell* 124(2): 301-313.

Al-Dosari, M. S., R. Shaheen, D. Colak and F. S. Alkuraya 2010. Novel CENPJ mutation causes Seckel syndrome. *J Med Genet* 47(6): 411-414.

Al-Owain, M., H. Al-Zaidan and Z. Al-Hassnan 2012. Map of autosomal recessive genetic disorders in Saudi Arabia: concepts and future directions. *Am J Med Genet A* 158A(10): 2629-2640.

Aladjem, M. I. 2007. Replication in context: dynamic regulation of DNA replication patterns in metazoans. *Nat Rev Genet* 8(8): 588-600.

Alarcon, C., A. I. Zaromytidou, Q. Xi, S. Gao, J. Yu, S. Fujisawa, A. Barlas, A. N. Miller*, et al.* 2009. Nuclear CDKs drive Smad transcriptional activation and turnover in BMP and TGF-beta pathways. *Cell* 139(4): 757-769.

Alazami, A. M., M. Al-Owain, F. Alzahrani, T. Shuaib, H. Al-Shamrani, Y. H. Al-Falki, S. M. Al-Qahtani, T. Alsheddi*, et al.* 2012. Loss of function mutation in LARP7, chaperone of 7SK ncRNA, causes a syndrome of facial dysmorphism, intellectual disability, and primordial dwarfism. *Hum Mutat* 33(10): 1429-1434.

Albers, C. A., D. S. Paul, H. Schulze, K. Freson, J. C. Stephens, P. A. Smethurst, J. D. Jolley, A. Cvejic, *et al.* 2012. Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in exon-junction complex subunit RBM8A causes TAR syndrome. *Nat Genet* 44(4): 435-439, S431-432.

Alderton, G. K., L. Galbiati, E. Griffith, K. H. Surinya, H. Neitzel, A. P. Jackson, P. A. Jeggo and M. O'Driscoll 2006. Regulation of mitotic entry by microcephalin and its overlap with ATR signalling. *Nat Cell Biol* 8(7): 725-733.

Alioto, T. S. 2007. U12DB: a database of orthologous U12-type spliceosomal introns. *Nucleic Acids Res* 35(Database issue): D110-115.

Altherr, M. R., T. J. Wright, K. Denison, A. V. Perez-Castro and V. P. Johnson 1997. Delimiting the Wolf-Hirschhorn syndrome critical region to 750 kilobase pairs. *Am J Med Genet* 71(1): 47-53.

Altmann, A., P. Weber, D. Bader, M. Preuss, E. B. Binder and B. Muller-Myhsok 2012. A beginners guide to SNP calling from high-throughput DNA-sequencing data. *Hum Genet* 131(10): 1541-1554.

Amsterdam, A., R. M. Nissen, Z. Sun, E. C. Swindell, S. Farrington and N. Hopkins 2004. Identification of 315 genes essential for early zebrafish development. *Proc Natl Acad Sci U S A* 101(35): 12792-12797.

Andres, S. N., A. Vergnes, D. Ristic, C. Wyman, M. Modesti and M. Junop 2012. A human XRCC4-XLF complex bridges DNA. *Nucleic Acids Res* 40(4): 1868-1878.

Antonius, T., J. Draaisma, E. Levtchenko, N. Knoers, W. Renier and C. van Ravenswaaij 2008. Growth charts for Wolf-Hirschhorn syndrome (0-4 years of age). *Eur J Pediatr* 167(7): 807-810.

Aragona, M., T. Panciera, A. Manfrin, S. Giulitti, F. Michielin, N. Elvassore, S. Dupont and S. Piccolo 2013. A mechanical checkpoint controls multicellular growth through YAP/TAZ regulation by actin-processing factors. *Cell* 154(5): 1047-1059.

Arias, E. E. and J. C. Walter 2007. Strength in numbers: preventing rereplication via multiple mechanisms in eukaryotic cells. *Genes Dev* 21(5): 497-518.

Armour, J. A., C. Sismani, P. C. Patsalis and G. Cross 2000. Measurement of locus copy number by hybridisation with amplifiable probes. *Nucleic Acids Res* 28(2): 605-609.

Auer, T. O. and F. Del Bene 2014. CRISPR/Cas9 and TALEN-mediated knock-in approaches in zebrafish. *Methods* 69(2): 142-150.

Azimzadeh, J. and W. F. Marshall 2010. Building the centriole. *Curr Biol* 20(18): R816-825.

Babashah, S., S. Jamali, R. Mahdian, M. H. Nosaeid, M. Karimipoor, R. Alimohammadi, M. Raeisi, F. Maryami, *et al.* 2009. Detection of unknown deletions in beta-globin gene cluster using relative quantitative PCR methods. *Eur J Haematol* 83(3): 261-269.

Badano, J. L., T. M. Teslovich and N. Katsanis 2005. The centrosome in human genetic disease. *Nat Rev Genet* 6(3): 194-205.

Bademci, G., O. Diaz-Horta, S. Guo, D. Duman, D. Van Booven, J. Foster Ii, F. B. Cengiz, S. Blanton and M. Tekin 2014. Identification of Copy Number Variants Through Whole-Exome Sequencing in Autosomal Recessive Nonsyndromic Hearing Loss. *Genet Test Mol Biomarkers*.

Baerlocher, G. M., I. Vulto, G. de Jong and P. M. Lansdorp 2006. Flow cytometry and FISH to measure the average length of telomeres (flow FISH). *Nat Protoc* 1(5): 2365-2376.

Bahtz, R., J. Seidler, M. Arnold, U. Haselmann-Weiss, C. Antony, W. D. Lehmann and I. Hoffmann 2012. GCP6 is a substrate of Plk4 and required for centriole duplication. *J Cell Sci* 125(Pt 2): 486-496.

Bailey, S. M., J. Meyne, D. J. Chen, A. Kurimasa, G. C. Li, B. E. Lehnert and E. H. Goodwin 1999. DNA double-strand break repair proteins are required to cap the ends of mammalian chromosomes. *Proc Natl Acad Sci U S A* 96(26): 14899-14904.

Bamshad, M. J., S. B. Ng, A. W. Bigham, H. K. Tabor, M. J. Emond, D. A. Nickerson and J. Shendure 2011. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* 12(11): 745-755.

Bandura, J. L., E. L. Beall, M. Bell, H. R. Silver, M. R. Botchan and B. R. Calvi 2005. humpty dumpty is required for developmental DNA amplification and cell proliferation in Drosophila. *Curr Biol* 15(8): 755-759.

Barnes, D. E., G. Stamp, I. Rosewell, A. Denzel and T. Lindahl 1998. Targeted disruption of the gene encoding DNA ligase IV leads to lethality in embryonic mice. *Curr Biol* 8(25): 1395-1398.

Basto, R., J. Lau, T. Vinogradova, A. Gardiol, C. G. Woods, A. Khodjakov and J. W. Raff 2006. Flies without centrioles. *Cell* 125(7): 1375-1386.

Battaglia, A., T. Filippi and J. C. Carey 2008. Update on the clinical features and natural history of Wolf-Hirschhorn (4p-) syndrome: experience with 87 patients and recommendations for routine health supervision. *Am J Med Genet C Semin Med Genet* 148C(4): 246-251.

Becker, J., O. Semler, C. Gilissen, Y. Li, H. J. Bolz, C. Giunta, C. Bergmann, M. Rohrbach, *et al.* 2011. Exome sequencing identifies truncating mutations in human SERPINF1 in autosomal-recessive osteogenesis imperfecta. *Am J Hum Genet* 88(3): 362-371.

Bell, C. J., D. L. Dinwiddie, N. A. Miller, S. L. Hateley, E. E. Ganusova, J. Mudge, R. J. Langley, L. Zhang, *et al.* 2011. Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Sci Transl Med* 3(65): 65ra64.

Bell, S. P. and B. Stillman 1992. ATP-dependent recognition of eukaryotic origins of DNA replication by a multiprotein complex. *Nature* 357(6374): 128-134.

Ben-Omran, T. I., K. Cerosaletti, P. Concannon, S. Weitzman and M. M. Nezarati 2005. A patient with mutations in DNA Ligase IV: clinical features and overlap with Nijmegen breakage syndrome. *Am J Med Genet A* 137A(3): 283-287.

Bender, M. A., H. G. Griggs and J. S. Bedford 1974. Mechanisms of chromosomal aberration production. 3. Chemicals and ionizing radiation. *Mutat Res* 23(2): 197-212.

Bennett, F. C. and K. F. Harvey 2006. Fat cadherin modulates organ size in Drosophila via the Salvador/Warts/Hippo signaling pathway. *Curr Biol* 16(21): 2101-2110.

Bentires-Alj, M., M. I. Kontaridis and B. G. Neel 2006. Stops along the RAS pathway in human genetic disease. *Nat Med* 12(3): 283-285.

Bettencourt-Dias, M. and D. M. Glover 2007. Centrosome biogenesis and function: centrosomics brings new understanding. *Nat Rev Mol Cell Biol* 8(6): 451-463.

Bhagwat, M. 2010. Searching NCBI's dbSNP database. *Curr Protoc Bioinformatics* Chapter 1: Unit 1.19.

Bhalla, N., S. Biggins and A. W. Murray 2002. Mutation of YCS4, a budding yeast condensin subunit, affects mitotic and nonmitotic chromosome behavior. *Mol Biol Cell* 13(2): 632-645.

Bicknell, L. S., E. M. Bongers, A. Leitch, S. Brown, J. Schoots, M. E. Harley, S. Aftimos, J. Y. Al-Aama, *et al.* 2011a. Mutations in the pre-replication complex cause Meier-Gorlin syndrome. *Nat Genet* 43(4): 356-359.

Bicknell, L. S., S. Walker, A. Klingseisen, T. Stiff, A. Leitch, C. Kerzendorfer, C. A. Martin, P. Yeyati, *et al.* 2011b. Mutations in ORC1, encoding the largest subunit of the origin recognition complex, cause microcephalic primordial dwarfism resembling Meier-Gorlin syndrome. *Nat Genet* 43(4): 350-355.

Bilguvar, K., A. K. Ozturk, A. Louvi, K. Y. Kwan, M. Choi, B. Tatli, D. Yalnizoglu, B. Tuysuz, *et al.* 2010. Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. *Nature* 467(7312): 207-210.

Blake, J. A., C. J. Bult, J. T. Eppig, J. A. Kadin, J. E. Richardson and M. G. D. Group 2014. The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. *Nucleic Acids Res* 42(Database issue): D810-817.

Blank, M., Y. Lerenthal, L. Mittelman and Y. Shiloh 2006. Condensin I recruitment and uneven chromatin condensation precede mitotic cell death in response to DNA damage. *J Cell Biol* 174(2): 195-206.

Bober, M. B., N. Khan, J. Kaplan, K. Lewis, J. A. Feinstein, C. I. Scott, Jr. and G. K. Steinberg 2010. Majewski osteodysplastic primordial dwarfism type II (MOPD II): expanding the vascular phenotype. *Am J Med Genet A* 152A(4): 960-965.

Bober, M. B., T. Niiler, A. L. Duker, J. E. Murray, T. Ketterer, M. E. Harley, S. Alvi, C. Flora, *et al.* 2012. Growth in individuals with Majewski osteodysplastic primordial dwarfism type II caused by pericentrin mutations. *Am J Med Genet A* 158A(11): 2719-2725.

Boland, J. F., C. C. Chung, D. Roberson, J. Mitchell, X. Zhang, K. M. Im, J. He, S. J. Chanock, *et al.* 2013. The new sequencer on the block: comparison of Life Technology's Proton sequencer to an Illumina HiSeq for whole-exome sequencing. *Hum Genet* 132(10): 1153-1163.

Bond, J., E. Roberts, G. H. Mochida, D. J. Hampshire, S. Scott, J. M. Askham, K. Springell, M. Mahadevan, *et al.* 2002. ASPM is a major determinant of cerebral cortical size. *Nat Genet* 32(2): 316-320.

Bond, J., E. Roberts, K. Springell, S. B. Lizarraga, S. Scott, J. Higgins, D. J. Hampshire, E. E. Morrison, *et al.* 2005. A centrosomal mechanism involving CDK5RAP2 and CENPJ controls brain size. *Nat Genet* 37(4): 353-355.

Bongers, E. M., J. M. Opitz, A. Fryer, P. Sarda, R. C. Hennekam, B. D. Hall, D. W. Superneau, M. Harbison, *et al.* 2001. Meier-Gorlin syndrome: report of eight additional cases and review. *Am J Med Genet* 102(2): 115-124.

Botstein, D. and N. Risch 2003. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet* 33 Suppl: 228-237.

Bouchard, C., K. Thieke, A. Maier, R. Saffrich, J. Hanley-Hyde, W. Ansorge, S. Reed, P. Sicinski, *et al.* 1999. Direct induction of cyclin D2 by Myc contributes to cell cycle progression and sequestration of p27. *EMBO J* 18(19): 5321-5333.

Boyle, E. A., B. J. O'Roak, B. K. Martin, A. Kumar and J. Shendure 2014. MIPgen: optimized modeling and design of molecular inversion probes for targeted resequencing. *Bioinformatics*.

Bradford, Y., T. Conlin, N. Dunn, D. Fashena, K. Frazer, D. G. Howe, J. Knight, P. Mani, *et al.* 2011. ZFIN: enhancements and updates to the Zebrafish Model Organism Database. *Nucleic Acids Res* 39(Database issue): D822-829.

Bragin, E., E. A. Chatzimichali, C. F. Wright, M. E. Hurles, H. V. Firth, A. P. Bevan and G. J. Swaminathan 2014. DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. *Nucleic Acids Res* 42(Database issue): D993-D1000.

Brancati, F., M. Castori, R. Mingarelli and B. Dallapiccola 2005. Majewski osteodysplastic primordial dwarfism type II (MOPD II) complicated by stroke: clinical report and review of cerebral vascular anomalies. *Am J Med Genet A* 139(3): 212-215.

Brants, J. R., T. A. Ayoubi, K. Chada, K. Marchal, W. J. Van de Ven and M. M. Petit 2004. Differential regulation of the insulin-like growth factor II mRNA-binding protein genes by architectural transcription factor HMGA2. *FEBS Lett* 569(1-3): 277-283.

Brown, E. J. and D. Baltimore 2000. ATR disruption leads to chromosomal fragmentation and early embryonic lethality. *Genes Dev* 14(4): 397-402.

Brunetti-Pierri, N., J. S. Berg, F. Scaglia, J. Belmont, C. A. Bacino, T. Sahoo, S. R. Lalani, B. Graham, *et al.* 2008. Recurrent reciprocal 1q21.1 deletions and duplications associated with microcephaly or macrocephaly and developmental and behavioral abnormalities. *Nat Genet* 40(12): 1466-1471.

Bryans, M., M. C. Valenzano and T. D. Stamato 1999. Absence of DNA ligase IV protein in XR-1 cells: evidence for stabilization by XRCC4. *Mutat Res* 433(1): 53-58.

Buck, D., L. Malivert, R. de Chasseval, A. Barraud, M. C. Fondaneche, O. Sanal, A. Plebani, J. L. Stephan, *et al.* 2006a. Cernunnos, a novel nonhomologous end-joining factor, is mutated in human immunodeficiency with microcephaly. *Cell* 124(2): 287-299.

Buck, D., D. Moshous, R. de Chasseval, Y. Ma, F. le Deist, M. Cavazzana-Calvo, A. Fischer, J. L. Casanova, *et al.* 2006b. Severe combined immunodeficiency and microcephaly in siblings with hypomorphic mutations in DNA ligase IV. *Eur J Immunol* 36(1): 224-235.

Buehler, B., H. H. Hogrefe, G. Scott, H. Ravi, C. Pabon-Pena, S. O'Brien, R. Formosa and S. Happe 2010. Rapid quantification of DNA libraries for next-generation sequencing. *Methods* 50(4): S15-18.

Burgering, B. M. and P. J. Coffer 1995. Protein kinase B (c-Akt) in phosphatidylinositol-3-OH kinase signal transduction. *Nature* 376(6541): 599-602.

Caldecott, K. W. 2008. Single-strand break repair and genetic disease. *Nat Rev Genet* 9(8): 619-631.

Carriere, A., M. Cargnello, L. A. Julien, H. Gao, E. Bonneil, P. Thibault and P. P. Roux 2008. Oncogenic MAPK signaling stimulates mTORC1 activity by promoting RSK-mediated raptor phosphorylation. *Curr Biol* 18(17): 1269-1277.

Caruso, N., B. Herberth, M. Bartoli, F. Puppo, J. Dumonceaux, A. Zimmermann, S. Denadai, M. Lebosse, *et al.* 2013. Deregulation of the protocadherin gene FAT1 alters muscle shapes: implications for the pathogenesis of facioscapulohumeral dystrophy. *PLoS Genet* 9(6): e1003550.

Casellas, R., A. Nussenzweig, R. Wuerffel, R. Pelanda, A. Reichlin, H. Suh, X. F. Qin, E. Besmer, *et al.* 1998. Ku80 is required for immunoglobulin isotype switching. *EMBO J* 17(8): 2404-2411.

Cayrou, C., P. Coulombe, A. Vigneron, S. Stanojcic, O. Ganier, I. Peiffer, E. Rivals, A. Puy, *et al.* 2011. Genome-scale analysis of metazoan replication origins reveals their organization in specific but flexible sites defined by conserved features. *Genome Res* 21(9): 1438-1449.

Chang, S., C. Khoo and R. A. DePinho 2001. Modeling chromosomal instability and epithelial carcinogenesis in the telomerase-deficient mouse. *Semin Cancer Biol* 11(3): 227-239.

Chauveau, C., J. Rowell and A. Ferreiro 2014. A Rising Titan: TTN Review and Mutation Update. *Hum Mutat*.

Chen, C. L., K. M. Gajewski, F. Hamaratoglu, W. Bossuyt, L. Sansores-Garcia, C. Tao and G. Halder 2010. The apical-basal cell polarity determinant Crumbs regulates Hippo signaling in Drosophila. *Proc Natl Acad Sci U S A* 107(36): 15810-15815.

Chen, C. T., H. Hehnly, Q. Yu, D. Farkas, G. Zheng, S. D. Redick, H. F. Hung, R. Samtani, *et al.* 2014. A Unique Set of Centrosome Proteins Requires Pericentrin for Spindle-Pole Localization and Spindle Orientation. *Curr Biol*.

Cheng, Q. and J. Chen 2010. Mechanism of p53 stabilization by ATM after DNA damage. *Cell Cycle* 9(3): 472-478.

Chenn, A. and C. A. Walsh 2002. Regulation of cerebral cortical size by control of cell cycle exit in neural precursors. *Science* 297(5580): 365-369.

Chiang, D. Y., G. Getz, D. B. Jaffe, M. J. O'Kelly, X. Zhao, S. L. Carter, C. Russ, C. Nusbaum, *et al.* 2009. High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Methods* 6(1): 99-103.

Chilamakuri, C. S., S. Lorenz, M. A. Madoui, D. Vodak, J. Sun, E. Hovig, O. Myklebost and L. A. Meza-Zepeda 2014. Performance comparison of four exome capture systems for deep sequencing. *BMC Genomics* 15: 449.

Chiu, S. Y., N. Asai, F. Costantini and W. Hsu 2008. SUMO-specific protease 2 is essential for modulating p53-Mdm2 in development of trophoblast stem cell niches and lineages. *PLoS Biol* 6(12): e310.

Chou, L. S., C. S. Liu, B. Boese, X. Zhang and R. Mao 2010. DNA sequence capture and enrichment by microarray followed by next-generation sequencing for targeted resequencing: neurofibromatosis type 1 gene as a model. *Clin Chem* 56(1): 62-72.

Christiansen, J., A. M. Kolte, T. O. Hansen and F. C. Nielsen 2009. IGF2 mRNA-binding protein 2: biological function and putative role in type 2 diabetes. *J Mol Endocrinol* 43(5): 187-195.

Ciani, L., A. Patel, N. D. Allen and C. ffrench-Constant 2003. Mice lacking the giant protocadherin mFAT1 exhibit renal slit junction abnormalities and a partially penetrant cyclopia and anophthalmia phenotype. *Mol Cell Biol* 23(10): 3575-3582.

Cimini, D., B. Howell, P. Maddox, A. Khodjakov, F. Degrassi and E. D. Salmon 2001. Merotelic kinetochore orientation is a major mechanism of aneuploidy in mitotic mammalian tissue cells. *J Cell Biol* 153(3): 517-527.

Cingolani, P., A. Platts, l. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu and D. M. Ruden 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* 6(2): 80-92.

Cirulli, E. T., A. Singh, K. V. Shianna, D. Ge, J. P. Smith, J. M. Maia, E. L. Heinzen, J. J. Goedert, *et al.* 2010. Screening the human exome: a comparison of whole genome and whole transcriptome sequencing. *Genome Biol* 11(5): R57.

Clarke, L., X. Zheng-Bradley, R. Smith, E. Kulesha, C. Xiao, I. Toneva, B. Vaughan, D. Preuss, *et al.* 2012. The 1000 Genomes Project: data management and community access. *Nat Methods* 9(5): 459-462.

Clements, P. M., C. Breslin, E. D. Deeks, P. J. Byrd, L. Ju, P. Bieganowski, C. Brenner, M. C. Moreira, *et al.* 2004. The ataxia-oculomotor apraxia 1 gene product has a role distinct from ATM and interacts with the DNA strand break repair proteins XRCC1 and XRCC4. *DNA Repair (Amst)* 3(11): 1493-1502.

Cleveland, D. W., Y. Mao and K. F. Sullivan 2003. Centromeres and kinetochores: from epigenetics to mitotic checkpoint signaling. *Cell* 112(4): 407-421.

Coelho, C. M. and S. J. Leevers 2000. Do growth and cell division rates determine cell size in multicellular organisms? *J Cell Sci* 113 (Pt 17): 2927-2934.

Coelho, P. A., J. Queiroz-Machado and C. E. Sunkel 2003. Condensin-dependent localisation of topoisomerase II to an axial chromosomal structure is required for sister chromatid resolution during mitosis. *J Cell Sci* 116(Pt 23): 4763-4776.

Cole, T. J. 2012. The development of growth references and growth charts. *Ann Hum Biol* 39(5): 382-394.

Conlon, I. and M. Raff 1999. Size control in animal development. *Cell* 96(2): 235-244.

Consortium, E. P. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 306(5696): 636-640.

Cortez, D., S. Guntuku, J. Qin and S. J. Elledge 2001. ATR and ATRIP: partners in checkpoint signaling. *Science* 294(5547): 1713-1716.

Critchlow, S. E., R. P. Bowater and S. P. Jackson 1997. Mammalian DNA double-strand break repair protein XRCC4 interacts with DNA ligase IV. *Curr Biol* 7(8): 588-598.

Currall, B. B., C. Chiang, M. E. Talkowski and C. C. Morton 2013. Mechanisms for Structural Variation in the Human Genome. *Curr Genet Med Rep* 1(2): 81-90.

d'Adda di Fagagna, F., M. P. Hande, W. M. Tong, D. Roth, P. M. Lansdorp, Z. Q. Wang and S. P. Jackson 2001. Effects of DNA nonhomologous end-joining factors on telomere length and chromosomal stability in mammalian cells. *Curr Biol* 11(15): 1192-1196.

D'Amours, D., S. Desnoyers, I. D'Silva and G. G. Poirier 1999. Poly(ADP-ribosyl)ation reactions in the regulation of nuclear functions. *Biochem J* 342 (Pt 2): 249-268.

D'Amours, D., F. Stegmeier and A. Amon 2004. Cdc14 and condensin control the dissolution of cohesin-independent chromosome linkages at repeated DNA. *Cell* 117(4): 455-469.

D, T. H. L. 1967. Congenital familial dwarfism with cephaloskeletal dysplasia. *Radiology* 89: 275-281.

Dai, N., J. Rapley, M. Angel, M. F. Yanik, M. D. Blower and J. Avruch 2011. mTOR phosphorylates IMP2 to promote IGF2 mRNA translation by internal ribosomal entry. *Genes Dev* 25(11): 1159-1172.

Dauber, A., S. H. Lafranchi, Z. Maliga, J. C. Lui, J. E. Moon, C. McDeed, K. Henke, J. Zonana, *et al.* 2012. Novel microcephalic primordial dwarfism disorder associated with variants in the centrosomal protein ninein. *J Clin Endocrinol Metab* 97(11): E2140-2151.

David, S. S., V. L. O'Shea and S. Kundu 2007. Base-excision repair of oxidative DNA damage. *Nature* 447(7147): 941-950.

de Ligt, J., P. M. Boone, R. Pfundt, L. E. Vissers, T. Richmond, J. Geoghegan, K. O'Moore, N. de Leeuw, *et al.* 2013. Detection of clinically relevant copy number variants with whole-exome sequencing. *Hum Mutat* 34(10): 1439-1448.

de Ligt, J., M. H. Willemsen, B. W. van Bon, T. Kleefstra, H. G. Yntema, T. Kroes, A. T. Vulto-van Silfhout, D. A. Koolen, *et al.* 2012. Diagnostic exome sequencing in persons with severe intellectual disability. *N Engl J Med* 367(20): 1921-1929.

de Munnik, S. A., L. S. Bicknell, S. Aftimos, J. Y. Al-Aama, Y. van Bever, M. B. Bober, J. Clayton-Smith, A. Y. Edrees, *et al.* 2012. Meier-Gorlin syndrome genotype-phenotype studies: 35 individuals with pre-replication complex gene mutations and 10 without molecular diagnosis. *Eur J Hum Genet* 20(6): 598-606.

Dean, M. D. and J. W. O. Ballard 2001. Factors affecting mitochondrial DNA quality from museum preserved Drosophila simulans. *Entomologia Experimentalis et Applicata* 98(3): 279-283.

Delaval, B. and S. Doxsey 2008. Genetics. Dwarfism, where pericentrin gains stature. *Science* 319(5864): 732-733.

Delaval, B. and S. J. Doxsey 2010. Pericentrin in cellular function and disease. *J Cell Biol* 188(2): 181-190.

Denayer, E., T. de Ravel and E. Legius 2008. Clinical and molecular aspects of RAS related disorders. *J Med Genet* 45(11): 695-703.

DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, *et al.* 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43(5): 491-498.

Dictenberg, J. B., W. Zimmerman, C. A. Sparks, A. Young, C. Vidair, Y. Zheng, W. Carrington, F. S. Fay and S. J. Doxsey 1998. Pericentrin and gamma-tubulin form a protein complex and are organized into a novel lattice at the centrosome. *J Cell Biol* 141(1): 163-174.

Doll, R. and R. Peto 1981. The causes of cancer: quantitative estimates of avoidable risks of cancer in the United States today. *J Natl Cancer Inst* 66(6): 1191-1308.

Dong, J., G. Feldmann, J. Huang, S. Wu, N. Zhang, S. A. Comerford, M. F. Gayyed, R. A. Anders, *et al.* 2007. Elucidation of a universal size-control mechanism in Drosophila and mammals. *Cell* 130(6): 1120-1133.

Dong, S., J. Han, H. Chen, T. Liu, M. S. Huen, Y. Yang, C. Guo and J. Huang 2014. The Human SRCAP Chromatin Remodeling Complex Promotes DNA-End Resection. *Curr Biol*.

Doxsey, S., D. McCollum and W. Theurkauf 2005. Centrosomes in cellular regulation. *Annu Rev Cell Dev Biol* 21: 411-434.

Doxsey, S. J., P. Stein, L. Evans, P. D. Calarco and M. Kirschner 1994. Pericentrin, a highly conserved centrosome protein involved in microtubule organization. *Cell* 76(4): 639-650.

Dressman, D., H. Yan, G. Traverso, K. W. Kinzler and B. Vogelstein 2003. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc Natl Acad Sci U S A* 100(15): 8817-8822.

Drouet, J., C. Delteil, J. Lefrancois, P. Concannon, B. Salles and P. Calsou 2005. DNA-dependent protein kinase and XRCC4-DNA ligase IV mobilization in the cell in response to DNA double strand breaks. *J Biol Chem* 280(8): 7060-7069.

Duan, J., J. G. Zhang, H. W. Deng and Y. P. Wang 2013. Comparative studies of copy number variation detection methods for next-generation sequencing technologies. *PLoS One* 8(3): e59128.

Dutrannoy, V., I. Demuth, U. Baumann, D. Schindler, K. Konrat, H. Neitzel, G. Gillessen-Kaesbach, J. Radszewski, *et al.* 2010. Clinical variability and novel mutations in the NHEJ1 gene in patients with a Nijmegen breakage syndrome-like phenotype. *Hum Mutat* 31(9): 1059-1068.

Dyment, D. A., A. C. Smith, D. Alcantara, J. A. Schwartzentruber, L. Basel-Vanagaite, C. J. Curry, I. K. Temple, W. Reardon, *et al.* 2013. Mutations in PIK3R1 cause SHORT syndrome. *Am J Hum Genet* 93(1): 158-166.

Edery, P., C. Marcaillou, M. Sahbatou, A. Labalme, J. Chastang, R. Touraine, E. Tubacher, F. Senni, *et al.* 2011. Association of TALS developmental disorder with defect in minor splicing component U4atac snRNA. *Science* 332(6026): 240-243.

Eid, J., A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, *et al.* 2009. Real-time DNA sequencing from single polymerase molecules. *Science* 323(5910): 133-138.

Eide, T., C. Carlson, K. A. Tasken, T. Hirano, K. Tasken and P. Collas 2002. Distinct but overlapping domains of AKAP95 are implicated in chromosome condensation and condensin targeting. *EMBO Rep* 3(5): 426-432.

Enders, A., P. Fisch, K. Schwarz, U. Duffner, U. Pannicke, E. Nikolopoulos, A. Peters, M. Orlowska-Volk, *et al.* 2006. A severe form of human combined immunodeficiency due to mutations in DNA ligase IV. *J Immunol* 176(8): 5060-5068.

Espinosa, E., P. Zamora, J. Feliu and M. Gonzalez Baron 2003. Classification of anticancer drugs--a new system based on therapeutic targets. *Cancer Treat Rev* 29(6): 515-523.

Faivre, L., M. Le Merrer, S. Lyonnet, H. Plauchu, N. Dagoneau, A. B. Campos-Xavier, J. Attia-Sobol, A. Verloes, *et al.* 2002. Clinical and genetic heterogeneity of Seckel syndrome. *Am J Med Genet* 112(4): 379-383.

Fan, J., C. Robert, Y. Y. Jang, H. Liu, S. Sharkis, S. B. Baylin and F. V. Rassool 2011. Human induced pluripotent cells resemble embryonic stem cells demonstrating enhanced levels of DNA repair and efficacy of nonhomologous end-joining. *Mutat Res* 713(1-2): 8-17.

Fedurco, M., A. Romieu, S. Williams, I. Lawrence and G. Turcatti 2006. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res* 34(3): e22.

Fernandes, N., Y. Sun, S. Chen, P. Paul, R. J. Shaw, L. C. Cantley and B. D. Price 2005. DNA damage-induced association of ATM with its target proteins requires a protein interaction domain in the N terminus of ATM. *J Biol Chem* 280(15): 15158-15164.

Finkel, T. and N. J. Holbrook 2000. Oxidants, oxidative stress and the biology of ageing. *Nature* 408(6809): 239-247.

Firth, H. V., C. F. Wright and D. D. D. Study 2011. The Deciphering Developmental Disorders (DDD) study. *Dev Med Child Neurol* 53(8): 702-703.

Flory, M. R., M. J. Moser, R. J. Monnat, Jr. and T. N. Davis 2000. Identification of a human centrosomal calmodulin-binding protein that shares homology with pericentrin. *Proc Natl Acad Sci U S A* 97(11): 5919-5923.

Fokkema, I. F., P. E. Taschner, G. C. Schaafsma, J. Celli, J. F. Laros and J. T. den Dunnen 2011. LOVD v.2.0: the next generation in gene variant databases. *Hum Mutat* 32(5): 557-563.

Fousteri, M. and L. H. Mullenders 2008. Transcription-coupled nucleotide excision repair in mammalian cells: molecular mechanisms and biological effects. *Cell Res* 18(1): 73-84.

Franco, L. M., T. de Ravel, B. H. Graham, S. M. Frenkel, J. Van Driessche, P. Stankiewicz, J. R. Lupski, J. R. Vermeesch and S. W. Cheung 2010. A syndrome of short stature, microcephaly and speech delay is associated with duplications reciprocal to the common Sotos syndrome deletion. *Eur J Hum Genet* 18(2): 258-261.

Frank, K. M., J. M. Sekiguchi, K. J. Seidl, W. Swat, G. A. Rathbun, H. L. Cheng, L. Davidson, L. Kangaloo and F. W. Alt 1998. Late embryonic lethality and impaired V(D)J recombination in mice lacking DNA ligase IV. *Nature* 396(6707): 173-177.

Frank, K. M., N. E. Sharpless, Y. Gao, J. M. Sekiguchi, D. O. Ferguson, C. Zhu, J. P. Manis, J. Horner, *et al.* 2000. DNA ligase IV deficiency in mice leads to defective neurogenesis and embryonic lethality via the p53 pathway. *Mol Cell* 5(6): 993-1002.

Frazer, K. A., D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, A. Boudreau, *et al.* 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449(7164): 851-861.

Freeman, J. V., T. J. Cole, S. Chinn, P. R. Jones, E. M. White and M. A. Preece 1995. Cross sectional stature and weight reference curves for the UK, 1990. *Arch Dis Child* 73(1): 17-24.

Fromer, M., J. L. Moran, K. Chambert, E. Banks, S. E. Bergen, D. M. Ruderfer, R. E. Handsaker, S. A. McCarroll, *et al.* 2012. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am J Hum Genet* 91(4): 597-607.

Fuchs, Y. and H. Steller 2011. Programmed cell death in animal development and disease. *Cell* 147(4): 742-758.

Gao, Y., J. Chaudhuri, C. Zhu, L. Davidson, D. T. Weaver and F. W. Alt 1998a. A targeted DNA-PKcs-null mutation reveals DNA-PK-independent functions for KU in V(D)J recombination. *Immunity* 9(3): 367-376.

Gao, Y., D. O. Ferguson, W. Xie, J. P. Manis, J. Sekiguchi, K. M. Frank, J. Chaudhuri, J. Horner, *et al.* 2000. Interplay of p53 and DNA-repair protein XRCC4 in tumorigenesis, genomic stability and development. *Nature* 404(6780): 897-900.

Gao, Y., Y. Sun, K. M. Frank, P. Dikkes, Y. Fujiwara, K. J. Seidl, J. M. Sekiguchi, G. A. Rathbun, *et al.* 1998b. A critical role for DNA end-joining proteins in both lymphogenesis and neurogenesis. *Cell* 95(7): 891-902.

Gatz, S. A., L. Ju, R. Gruber, E. Hoffmann, A. M. Carr, Z. Q. Wang, C. Liu and P. A. Jeggo 2011. Requirement for DNA ligase IV during embryonic neuronal development. *J Neurosci* 31(27): 10088-10100.

Gavin, K. A., M. Hidaka and B. Stillman 1995. Conserved initiator proteins in eukaryotes. *Science* 270(5242): 1667-1671.

Ge, X. Q. and J. J. Blow 2010. Chk1 inhibits replication factory activation but allows dormant origin firing in existing factories. *J Cell Biol* 191(7): 1285-1297.

Genevet, A. and N. Tapon 2011. The Hippo pathway and apico-basal cell polarity. *Biochem J* 436(2): 213-224.

Genin, A., J. Desir, N. Lambert, M. Biervliet, N. Van Der Aa, G. Pierquin, A. Killian, M. Tosi, *et al.* 2012. Kinetochore KMN network gene CASC5 mutated in primary microcephaly. *Hum Mol Genet* 21(24): 5306-5317.

Genomes Project, C., G. R. Abecasis, A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin, R. E. Handsaker, H. M. Kang, *et al.* 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491(7422): 56-65.

Gerlich, D., T. Hirota, B. Koch, J. M. Peters and J. Ellenberg 2006. Condensin I stabilizes chromosomes mechanically through a dynamic interaction in live cells. *Curr Biol* 16(4): 333-344.

Gerstein, R. M. and M. R. Lieber 1993. Extent to which homology can constrain coding exon junctional diversity in V(D)J recombination. *Nature* 363(6430): 625-627.

Gilissen, C., J. Y. Hehir-Kwa, D. T. Thung, M. van de Vorst, B. W. van Bon, M. H. Willemsen, M. Kwint, I. M. Janssen, *et al.* 2014. Genome sequencing identifies major causes of severe intellectual disability. *Nature* 511(7509): 344-347.

Gilissen, C., A. Hoischen, H. G. Brunner and J. A. Veltman 2012. Disease gene identification strategies for exome sequencing. *Eur J Hum Genet* 20(5): 490-497.

Gineau, L., C. Cognet, N. Kara, F. P. Lach, J. Dunne, U. Veturi, C. Picard, C. Trouillet, *et al.* 2012. Partial MCM4 deficiency in patients with growth retardation, adrenal insufficiency, and natural killer cell deficiency. *J Clin Invest* 122(3): 821-832.

Girard, P. M., B. Kysela, C. J. Harer, A. J. Doherty and P. A. Jeggo 2004. Analysis of DNA ligase IV mutations found in LIG4 syndrome patients: the impact of two linked polymorphisms. *Hum Mol Genet* 13(20): 2369-2376.

Gnirke, A., A. Melnikov, J. Maguire, P. Rogov, E. M. LeProust, W. Brockman, T. Fennell, G. Giannoukos, *et al.* 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* 27(2): 182-189.

Goldsmith, Z. G. and D. N. Dhanasekaran 2007. G protein regulation of MAPK networks. *Oncogene* 26(22): 3122-3142.

Goodship, J., H. Gill, J. Carter, A. Jackson, M. Splitt and M. Wright 2000. Autozygosity mapping of a seckel syndrome locus to chromosome 3q22. 1-q24. *Am J Hum Genet* 67(2): 498-503.

Gorlin, R. J., J. Cervenka, K. Moller, M. Horrobin and C. J. Witkop, Jr. 1975. Malformation syndromes. A selected miscellany. *Birth Defects Orig Artic Ser* 11(2): 39-50.

Gosling, K. M., L. E. Makaroff, A. Theodoratos, Y. H. Kim, B. Whittle, L. Rui, H. Wu, N. A. Hong, *et al.* 2007. A mutation in a chromosome condensin II subunit, kleisin beta, specifically disrupts T cell development. *Proc Natl Acad Sci U S A* 104(30): 12445-12450.

Gottlieb, T. M. and S. P. Jackson 1993. The DNA-dependent protein kinase: requirement for DNA ends and association with Ku antigen. *Cell* 72(1): 131-142.

Graham, J. M., Jr. and C. E. Schwartz 2013. MED12 related disorders. *Am J Med Genet A* 161A(11): 2734-2740.

Grawunder, U., D. Zimmer and M. R. Leiber 1998. DNA ligase IV binds to XRCC4 via a motif located between rather than within its BRCT domains. *Curr Biol* 8(15): 873-876.

Green, L. C., P. Kalitsis, T. M. Chang, M. Cipetic, J. H. Kim, O. Marshall, L. Turnbull, C. B. Whitchurch, *et al.* 2012. Contrasting roles of condensin I and condensin II in mitotic chromosome formation. *J Cell Sci* 125(Pt 6): 1591-1604.

Green, R. C., J. S. Berg, W. W. Grody, S. S. Kalia, B. R. Korf, C. L. Martin, A. L. McGuire, R. L. Nussbaum, *et al.* 2013. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet Med* 15(7): 565-574.

Griffith, E., S. Walker, C. A. Martin, P. Vagnarelli, T. Stiff, B. Vernay, N. Al Sanna, A. Saggar, *et al.* 2008. Mutations in pericentrin cause Seckel syndrome with defective ATR-dependent DNA damage signaling. *Nat Genet* 40(2): 232-236.

Grimm, D., J. Hagmann, D. Koenig, D. Weigel and K. Borgwardt 2013. Accurate indel prediction using paired-end short reads. *BMC Genomics* 14: 132.

Gruhn, B., J. Seidel, F. Zintl, R. Varon, H. Tonnies, H. Neitzel, A. Bechtold, H. Hoehn and D. Schindler 2007. Successful bone marrow transplantation in a patient with DNA ligase IV deficiency and bone marrow failure. *Orphanet J Rare Dis* 2: 5.

Grunebaum, E., A. Bates and C. M. Roifman 2008. Omenn syndrome is associated with mutations in DNA ligase IV. *J Allergy Clin Immunol* 122(6): 1219-1220.

Gu, Y., K. J. Seidl, G. A. Rathbun, C. Zhu, J. P. Manis, N. van der Stoep, L. Davidson, H. L. Cheng, *et al.* 1997. Growth retardation and leaky SCID phenotype of Ku70-deficient mice. *Immunity* 7(5): 653-665.

Guernsey, D. L., H. Jiang, J. Hussin, M. Arnold, K. Bouyakdan, S. Perry, T. Babineau-Sturk, J. Beis, *et al.* 2010. Mutations in centrosomal protein CEP152 in primary microcephaly families linked to MCPH4. *Am J Hum Genet* 87(1): 40-51.

Guernsey, D. L., M. Matsuoka, H. Jiang, S. Evans, C. Macgillivray, M. Nightingale, S. Perry, M. Ferguson, *et al.* 2011. Mutations in origin recognition complex gene ORC4 cause Meier-Gorlin syndrome. *Nat Genet* 43(4): 360-364.

Guirouilh-Barbat, J., E. Rass, I. Plo, P. Bertrand and B. S. Lopez 2007. Defects in XRCC4 and KU80 differentially affect the joining of distal nonhomologous ends. *Proc Natl Acad Sci U S A* 104(52): 20902-20907.

Hagstrom, K. A., V. F. Holmes, N. R. Cozzarelli and B. J. Meyer 2002. C. elegans condensin promotes mitotic chromosome architecture, centromere organization, and sister chromatid segregation during mitosis and meiosis. *Genes Dev* 16(6): 729-742.

Hajdu, I., A. Ciccia, S. M. Lewis and S. J. Elledge 2011. Wolf-Hirschhorn syndrome candidate 1 is involved in the cellular response to DNA damage. *Proc Natl Acad Sci U S A* 108(32): 13130-13134.

Hall, J. G., C. Flora, C. I. Scott, Jr., R. M. Pauli and K. I. Tanaka 2004. Majewski osteodysplastic primordial dwarfism type II (MOPD II): natural history and clinical findings. *Am J Med Genet A* 130A(1): 55-72.

Hammel, M., M. Rey, Y. Yu, R. S. Mani, S. Classen, M. Liu, M. E. Pique, S. Fang, *et al.* 2011. XRCC4 protein interactions with XRCC4-like factor (XLF) create an extended grooved scaffold for DNA ligation and double strand break repair. *J Biol Chem* 286(37): 32638-32650.

Han, V. K., P. K. Lund, D. C. Lee and A. J. D'Ercole 1988. Expression of somatomedin/insulin-like growth factor messenger ribonucleic acids in the human fetus: identification, characterization, and tissue distribution. *J Clin Endocrinol Metab* 66(2): 422-429.

Hanks, S., K. Coleman, S. Reid, A. Plaja, H. Firth, D. Fitzpatrick, A. Kidd, K. Mehes, *et al.* 2004. Constitutional aneuploidy and cancer predisposition caused by biallelic mutations in BUB1B. *Nat Genet* 36(11): 1159-1161.

Harris, T. D., P. R. Buzby, H. Babcock, E. Beer, J. Bowers, I. Braslavsky, M. Causey, J. Colonell, *et al.* 2008. Single-molecule DNA sequencing of a viral genome. *Science* 320(5872): 106-109.

Harvey, K. F., C. M. Pfleger and I. K. Hariharan 2003. The Drosophila Mst ortholog, hippo, restricts growth and cell proliferation and promotes apoptosis. *Cell* 114(4): 457-467.

Hasmats, J., H. Green, C. Orear, P. Validire, M. Huss, M. Kaller and J. Lundeberg 2014. Assessment of whole genome amplification for sequence capture and massively parallel sequencing. *PLoS One* 9(1): e84785.

He, H., S. Liyanarachchi, K. Akagi, R. Nagy, J. Li, R. C. Dietrich, W. Li, N. Sebastian*, et al.* 2011. Mutations in U4atac snRNA, a component of the minor spliceosome, in the developmental disorder MOPD I. *Science* 332(6026): 238-240.

Heale, J. T., A. R. Ball, Jr., J. A. Schmiesing, J. S. Kim, X. Kong, S. Zhou, D. F. Hudson, W. C. Earnshaw and K. Yokomori 2006. Condensin I interacts with the PARP-1-XRCC1 complex and functions in DNA single-strand break repair. *Mol Cell* 21(6): 837-848.

Heallen, T., M. Zhang, J. Wang, M. Bonilla-Claudio, E. Klysik, R. L. Johnson and J. F. Martin 2011. Hippo pathway inhibits Wnt signaling to restrain cardiomyocyte proliferation and heart size. *Science* 332(6028): 458-461.

Hennies, H. C., A. Rauch, W. Seifert, C. Schumi, E. Moser, E. Al-Taji, G. Tariverdian, K. H. Chrzanowska*, et al.* 2004. Allelic heterogeneity in the COH1 gene explains clinical variability in Cohen syndrome. *Am J Hum Genet* 75(1): 138-145.

Herman, D. S., G. K. Hovingh, O. Iartchouk, H. L. Rehm, R. Kucherlapati, J. G. Seidman and C. E. Seidman 2009. Filter-based hybridization capture of subgenomes enables resequencing and copy-number detection. *Nat Methods* 6(7): 507-510.

Hermeking, H., C. Rago, M. Schuhmacher, Q. Li, J. F. Barrett, A. J. Obaya, B. C. O'Connell, M. K. Mateyak*, et al.* 2000. Identification of CDK4 as a target of c-MYC. *Proc Natl Acad Sci U S A* 97(5): 2229-2234.

Hietakangas, V. and S. M. Cohen 2009. Regulation of tissue growth through nutrient sensing. *Annu Rev Genet* 43: 389-410.

Hinchcliffe, E. H., F. J. Miller, M. Cham, A. Khodjakov and G. Sluder 2001. Requirement of a centrosomal activity for cell cycle progression through G1 into S phase. *Science* 291(5508): 1547-1550.

Hirano, M., D. E. Anderson, H. P. Erickson and T. Hirano 2001. Bimodal activation of SMC ATPase by intra- and inter-molecular interactions. *EMBO J* 20(12): 3238-3250.

Hirano, T. 2005. Condensins: organizing and segregating the genome. *Curr Biol* 15(7): R265-275.

Hirota, T., D. Gerlich, B. Koch, J. Ellenberg and J. M. Peters 2004. Distinct functions of condensin I and II in mitotic chromosome assembly. *J Cell Sci* 117(Pt 26): 6435-6445.

Hochberg, Z. and K. Albertsson-Wikland 2008. Evo-devo of infantile and childhood growth. *Pediatr Res* 64(1): 2-7.

Hodges, E., Z. Xuan, V. Balija, M. Kramer, M. N. Molla, S. W. Smith, C. M. Middle, M. J. Rodesch*, et al.* 2007. Genome-wide in situ exon capture for selective resequencing. *Nat Genet* 39(12): 1522-1527.

Hoischen, A., C. Gilissen, P. Arts, N. Wieskamp, W. van der Vliet, S. Vermeer, M. Steehouwer, P. de Vries*, et al.* 2010. Massively parallel sequencing of ataxia genes after array-based enrichment. *Hum Mutat* 31(4): 494-499.

Hood, R. L., M. A. Lines, S. M. Nikkel, J. Schwartzentruber, C. Beaulieu, M. J. Nowaczyk, J. Allanson, C. A. Kim*, et al.* 2012. Mutations in SRCAP, encoding SNF2-related CREBBP activator protein, cause Floating-Harbor syndrome. *Am J Hum Genet* 90(2): 308-313.

Horton, W. A., J. I. Rotter, D. L. Rimoin, C. I. Scott and J. G. Hall 1978. Standard growth curves for achondroplasia. *J Pediatr* 93(3): 435-438.

Howard-Flanders, P. 1975. Repair by genetic recombination in bacteria: overview. *Basic Life Sci* 5A: 265-274.

Huang-Doran, I., L. S. Bicknell, F. M. Finucane, N. Rocha, K. M. Porter, Y. C. Tung, F. Szekeres, A. Krook, *et al.* 2011. Genetic defects in human pericentrin are associated with severe insulin resistance and diabetes. *Diabetes* 60(3): 925-935.

Huang, N., I. Lee, E. M. Marcotte and M. E. Hurles 2010. Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet* 6(10): e1001154.

Huber, C., A. Munnich and V. Cormier-Daire 2011. The 3M syndrome. *Best Pract Res Clin Endocrinol Metab* 25(1): 143-151.

Hudson, D. F., P. Vagnarelli, R. Gassmann and W. C. Earnshaw 2003. Condensin is required for nonhistone protein assembly and structural integrity of vertebrate mitotic chromosomes. *Dev Cell* 5(2): 323-336.

Hughes, C. R., L. Guasti, E. Meimaridou, C. H. Chuang, J. C. Schimenti, P. J. King, C. Costigan, A. J. Clark and L. A. Metherell 2012. MCM4 mutation causes adrenal failure, short stature, and natural killer cell deficiency in humans. *J Clin Invest* 122(3): 814-820.

Hunt, R. C., V. L. Simhadri, M. Iandoli, Z. E. Sauna and C. Kimchi-Sarfaty 2014. Exposing synonymous mutations. *Trends Genet*.

Hussain, M. S., S. M. Baig, S. Neumann, G. Nurnberg, M. Farooq, I. Ahmad, T. Alef, H. C. Hennies, *et al.* 2012. A Truncating Mutation of CEP135 Causes Primary Microcephaly and Disturbed Centrosomal Function. *Am J Hum Genet* 90(5): 871-878.

Iafrate, A. J., L. Feuk, M. N. Rivera, M. L. Listewnik, P. K. Donahoe, Y. Qi, S. W. Scherer and C. Lee 2004. Detection of large-scale variation in the human genome. *Nat Genet* 36(9): 949-951.

Ijspeert, H., A. Warris, M. van der Flier, I. Reisli, S. Keles, S. Chishimba, J. J. van Dongen, D. C. van Gent and M. van der Burg 2013. Clinical Spectrum of LIG4 Deficiency is Broadened with Severe Dysmaturity, Primordial Dwarfism, and Neurological Abnormalities. *Hum Mutat*.

Ikeda, A., M. Masaki, Y. Kozutsumi, S. Oka and T. Kawasaki 2001. Identification and characterization of functional domains in a mixed lineage kinase LZK. *FEBS Lett* 488(3): 190-195.

Ikram, M. A., M. Fornage, A. V. Smith, S. Seshadri, R. Schmidt, S. Debette, H. A. Vrooman, S. Sigurdsson, *et al.* 2012a. Common variants at 6q22 and 17q21 are associated with intracranial volume. *Nat Genet* 44(5): 539-544.

Ikram, M. A., M. Fornage, A. V. Smith, S. Seshadri, R. Schmidt, S. Debette, H. A. Vrooman, S. Sigurdsson, *et al.* 2012b. Common variants at 6q22 and 17q21 are associated with intracranial volume. *Nat Genet* 44(5): 539-544.

Iles, N., S. Rulten, S. F. El-Khamisy and K. W. Caldecott 2007. APLF (C2orf13) is a novel human protein involved in the cellular response to chromosomal DNA strand breaks. *Mol Cell Biol* 27(10): 3793-3803.

Ilves, I., T. Petojevic, J. J. Pesavento and M. R. Botchan 2010. Activation of the MCM2-7 helicase by association with Cdc45 and GINS proteins. *Mol Cell* 37(2): 247-258.

Inamdar, K. V., J. J. Pouliot, T. Zhou, S. P. Lees-Miller, A. Rasouli-Nia and L. F. Povirk 2002. Conversion of phosphoglycolate to phosphate termini on 3' overhangs of DNA double strand breaks by the human tyrosyl-DNA phosphodiesterase hTdp1. *J Biol Chem* 277(30): 27162-27168.

Inoue, A., N. Yamamoto, M. Kimura, K. Nishio, H. Yamane and K. Nakajima 2014. RBM10 regulates alternative splicing. *FEBS Lett* 588(6): 942-947.

Inui, M., M. Miyado, M. Igarashi, M. Tamano, A. Kubo, S. Yamashita, H. Asahara, M. Fukami and S. Takada 2014. Rapid generation of mouse models with defined point mutations by the CRISPR/Cas9 system. *Sci Rep* 4: 5396.

Jackson, A. P., H. Eastwood, S. M. Bell, J. Adu, C. Toomes, I. M. Carr, E. Roberts, D. J. Hampshire*, et al.* 2002. Identification of microcephalin, a protein implicated in determining the size of the human brain. *Am J Hum Genet* 71(1): 136-142.

Jacquemont, S., A. Reymond, F. Zufferey, L. Harewood, R. G. Walters, Z. Kutalik, D. Martinet, Y. Shen*, et al.* 2011. Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. *Nature* 478(7367): 97-102.

Jeppsson, K., T. Kanno, K. Shirahige and C. Sjogren 2014. The maintenance of chromosome structure: positioning and functioning of SMC complexes. *Nat Rev Mol Cell Biol* 15(9): 601-614.

Jiang, M., S. Y. Chiu and W. Hsu 2011. SUMO-specific protease 2 in Mdm2-mediated regulation of p53. *Cell Death Differ* 18(6): 1005-1015.

Jilani, A., D. Ramotar, C. Slack, C. Ong, X. M. Yang, S. W. Scherer and D. D. Lasko 1999. Molecular cloning of the human gene, PNKP, encoding a polynucleotide kinase 3'-phosphatase and evidence for its role in repair of DNA strand breaks caused by oxidative damage. *J Biol Chem* 274(34): 24176-24186.

Jiricny, J. 2006. The multifaceted mismatch-repair system. *Nat Rev Mol Cell Biol* 7(5): 335-346.

Johnston, J. J., J. K. Teer, P. F. Cherukuri, N. F. Hansen, S. K. Loftus, K. Chong, J. C. Mullikin and L. G. Biesecker 2010. Massively parallel sequencing of exons on the X chromosome identifies RBM10 as the gene that causes a syndromic form of cleft palate. *Am J Hum Genet* 86(5): 743-748.

Jorgensen, P. and M. Tyers 2004. How cells coordinate growth and division. *Curr Biol* 14(23): R1014-1027.

Ju, J., D. H. Kim, L. Bi, Q. Meng, X. Bai, Z. Li, X. Li, M. S. Marma*, et al.* 2006. Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proc Natl Acad Sci U S A* 103(52): 19635-19640.

Junop, M. S., M. Modesti, A. Guarne, R. Ghirlando, M. Gellert and W. Yang 2000. Crystal structure of the Xrcc4 DNA repair protein and implications for end joining. *EMBO J* 19(22): 5962-5970.

Jurczyk, A., S. C. Pino, B. O'Sullivan-Murphy, M. Addorio, E. A. Lidstone, P. Diiorio, K. L. Lipson, C. Standley*, et al.* 2010. A novel role for the centrosomal protein, pericentrin, in regulation of insulin secretory vesicle docking in mouse pancreatic beta-cells. *PLoS One* 5(7): e11812.

Justice, R. W., O. Zilian, D. F. Woods, M. Noll and P. J. Bryant 1995. The Drosophila tumor suppressor gene warts encodes a homolog of human myotonic dystrophy kinase and is required for the control of cell shape and proliferation. *Genes Dev* 9(5): 534-546.

Kaasinen, E., E. Rahikkala, P. Koivunen, S. Miettinen, M. M. Wamelink, M. Aavikko, K. Palin, J. Myllyharju, *et al.* 2014. Clinical characterization, genetic mapping and whole-genome sequence analysis of a novel autosomal recessive intellectual disability syndrome. *Eur J Med Genet*.

Kalay, E., G. Yigit, Y. Aslan, K. E. Brown, E. Pohl, L. S. Bicknell, H. Kayserili, Y. Li, *et al.* 2011. CEP152 is a genome maintenance protein disrupted in Seckel syndrome. *Nat Genet* 43(1): 23-26.

Kaneko, H. and N. Kondo 2004. Clinical features of Bloom syndrome and function of the causative gene, BLM helicase. *Expert Rev Mol Diagn* 4(3): 393-401.

Kanno, S., H. Kuzuoka, S. Sasao, Z. Hong, L. Lan, S. Nakajima and A. Yasui 2007. A novel human AP endonuclease with conserved zinc-finger-like motifs involved in DNA strand break responses. *EMBO J* 26(8): 2094-2103.

Kant, S. G., M. Kriek, M. J. Walenkamp, K. B. Hansson, A. van Rhijn, J. Clayton-Smith, J. M. Wit and M. H. Breuning 2007. Tall stature and duplication of the insulin-like growth factor I receptor gene. *Eur J Med Genet* 50(1): 1-10.

Kantaputra, P., P. Tanpaiboon, T. Porntaveetus, A. Ohazama, P. Sharpe, A. Rauch, A. Hussadaloy and C. T. Thiel 2011. The smallest teeth in the world are caused by mutations in the PCNT gene. *Am J Med Genet A* 155A(6): 1398-1403.

Karanjawala, Z. E., N. Murphy, D. R. Hinton, C. L. Hsieh and M. R. Lieber 2002. Oxygen metabolism causes chromosome breaks and is associated with the neuronal apoptosis observed in DNA double-strand break repair mutants. *Curr Biol* 12(5): 397-402.

Karatas, A. F., M. B. Bober, K. Rogers, A. L. Duker, C. P. Ditro and W. G. Mackenzie 2014. Hip Pathology in Majewski Osteodysplastic Primordial Dwarfism Type II. *J Pediatr Orthop*.

Keller, C., K. R. Keller, S. B. Shew and S. E. Plon 1999. Growth deficiency and malnutrition in Bloom syndrome. *J Pediatr* 134(4): 472-479.

Kent, W. J., F. Hsu, D. Karolchik, R. M. Kuhn, H. Clawson, H. Trumbower and D. Haussler 2005. Exploring relationships and mining data with the UCSC Gene Sorter. *Genome Res* 15(5): 737-741.

Kerem, B., J. M. Rommens, J. A. Buchanan, D. Markiewicz, T. K. Cox, A. Chakravarti, M. Buchwald and L. C. Tsui 1989. Identification of the cystic fibrosis gene: genetic analysis. *Science* 245(4922): 1073-1080.

Kerzendorfer, C., F. Hannes, R. Colnaghi, I. Abramowicz, G. Carpenter, J. R. Vermeesch and M. O'Driscoll 2012. Characterizing the functional consequences of haploinsufficiency of NELF-A (WHSC2) and SLBP identifies novel cellular phenotypes in Wolf-Hirschhorn syndrome. *Hum Mol Genet* 21(10): 2181-2193.

Khodjakov, A. and C. L. Rieder 2001. Centrosomes enhance the fidelity of cytokinesis in vertebrates and are required for cell cycle progression. *J Cell Biol* 153(1): 237-242.

Kim, J., S. K. Keay, S. You, M. Loda and M. R. Freeman 2012. A synthetic form of frizzled 8-associated antiproliferative factor enhances p53 stability through USP2a and MDM2. *PLoS One* 7(12): e50392.

Kim, T. S., J. E. Park, A. Shukla, S. Choi, R. N. Murugan, J. H. Lee, M. Ahn, K. Rhee, *et al.* 2013. Hierarchical recruitment of Plk4 and regulation of centriole biogenesis by two centrosomal scaffolds, Cep192 and Cep152. *Proc Natl Acad Sci U S A* 110(50): E4849-4857.

Kimura, K., M. Hirano, R. Kobayashi and T. Hirano 1998. Phosphorylation and activation of 13S condensin by Cdc2 in vitro. *Science* 282(5388): 487-490.

King, D. A., T. W. Fitzgerald, R. Miller, N. Canham, J. Clayton-Smith, D. Johnson, S. Mansour, F. Stewart, *et al.* 2014. A novel method for detecting uniparental disomy from trio genotypes identifies a significant excess in children with developmental disorders. *Genome Res* 24(4): 673-687.

King, W. G., M. D. Mattaliano, T. O. Chan, P. N. Tsichlis and J. S. Brugge 1997. Phosphatidylinositol 3-kinase is required for integrin-stimulated AKT and Raf-1/mitogen-activated protein kinase pathway activation. *Mol Cell Biol* 17(8): 4406-4418.

Kircher, M., D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper and J. Shendure 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 46(3): 310-315.

Klinge, L., J. Schaper, D. Wieczorek and T. Voit 2002. Microlissencephaly in microcephalic osteodysplastic primordial dwarfism: a case report and review of the literature. *Neuropediatrics* 33(6): 309-313.

Klingseisen, A. and A. P. Jackson 2011. Mechanisms and pathways of growth failure in primordial dwarfism. *Genes Dev* 25(19): 2011-2024.

Knoblich, J. A., K. Sauer, L. Jones, H. Richardson, R. Saint and C. F. Lehner 1994. Cyclin E controls S phase progression and its down-regulation during Drosophila embryogenesis is required for the arrest of cell proliferation. *Cell* 77(1): 107-120.

Knowles, M. R., M. W. Leigh, L. E. Ostrowski, L. Huang, J. L. Carson, M. J. Hazucha, W. Yin, J. S. Berg, *et al.* 2013. Exome sequencing identifies mutations in CCDC114 as a cause of primary ciliary dyskinesia. *Am J Hum Genet* 92(1): 99-106.

Ko, M. A., C. O. Rosario, J. W. Hudson, S. Kulkarni, A. Pollett, J. W. Dennis and C. J. Swallow 2005. Plk4 haploinsufficiency causes mitotic infidelity and carcinogenesis. *Nat Genet* 37(8): 883-888.

Koboldt, D. C., D. E. Larson, K. Chen, L. Ding and R. K. Wilson 2012. Massively parallel sequencing approaches for characterization of structural variation. *Methods Mol Biol* 838: 369-384.

Koch, C. A., R. Agyei, S. Galicia, P. Metalnikov, P. O'Donnell, A. Starostine, M. Weinfeld and D. Durocher 2004. Xrcc4 physically links DNA end processing by polynucleotide kinase to DNA ligation by DNA ligase IV. *EMBO J* 23(19): 3874-3885.

Kohler, S., S. C. Doelken, C. J. Mungall, S. Bauer, H. V. Firth, I. Bailleul-Forestier, G. C. Black, D. L. Brown, *et al.* 2014. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res* 42(Database issue): D966-974.

Kolehmainen, J., G. C. Black, A. Saarinen, K. Chandler, J. Clayton-Smith, A. L. Traskelin, R. Perveen, S. Kivitie-Kallio, *et al.* 2003. Cohen syndrome is caused by mutations in a novel gene, COH1, encoding a transmembrane protein with a presumed role in vesicle-mediated sorting and intracellular protein transport. *Am J Hum Genet* 72(6): 1359-1369.

Korbel, J. O., A. E. Urban, J. P. Affourtit, B. Godwin, F. Grubert, J. F. Simons, P. M. Kim, D. Palejev, *et al.* 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science* 318(5849): 420-426.

Koshland, D. and A. Strunnikov 1996. Mitotic chromosome condensation. *Annu Rev Cell Dev Biol* 12: 305-333.

Kousholt, A. N., K. Fugger, S. Hoffmann, B. D. Larsen, T. Menzel, A. A. Sartori and C. S. Sorensen 2012. CtIP-dependent DNA resection is required for DNA damage checkpoint maintenance but not initiation. *J Cell Biol* 197(7): 869-876.

Kramer, A., N. Mailand, C. Lukas, R. G. Syljuasen, C. J. Wilkinson, E. A. Nigg, J. Bartek and J. Lukas 2004. Centrosome-associated Chk1 prevents premature activation of cyclin-B-Cdk1 kinase. *Nat Cell Biol* 6(9): 884-891.

Krishna, M. and H. Narang 2008. The complexity of mitogen-activated protein kinases (MAPKs) made simple. *Cell Mol Life Sci* 65(22): 3525-3544.

Krude, T., M. Jackman, J. Pines and R. A. Laskey 1997. Cyclin/Cdk-dependent initiation of DNA replication in a human cell-free system. *Cell* 88(1): 109-119.

Krumm, N., P. H. Sudmant, A. Ko, B. J. O'Roak, M. Malig, B. P. Coe, N. E. S. Project, A. R. Quinlan, *et al.* 2012. Copy number variation detection and genotyping from exome sequence data. *Genome Res* 22(8): 1525-1532.

Ku, C. S., M. Wu, D. N. Cooper, N. Naidoo, Y. Pawitan, B. Pang, B. Iacopetta and R. Soong 2012. Exome versus transcriptome sequencing in identifying coding region variants. *Expert Rev Mol Diagn* 12(3): 241-251.

Kuida, K., T. F. Haydar, C. Y. Kuan, Y. Gu, C. Taya, H. Karasuyama, M. S. Su, P. Rakic and R. A. Flavell 1998. Reduced apoptosis and cytochrome c-mediated caspase activation in mice lacking caspase 9. *Cell* 94(3): 325-337.

Kumar, A., S. C. Girimaji, M. R. Duvvari and S. H. Blanton 2009. Mutations in STIL, encoding a pericentriolar and centrosomal protein, cause primary microcephaly. *Am J Hum Genet* 84(2): 286-290.

Kunnev, D., M. E. Rusiniak, A. Kudla, A. Freeland, G. K. Cady and S. C. Pruitt 2010. DNA damage response and tumorigenesis in Mcm2-deficient mice. *Oncogene* 29(25): 3630-3638.

Kuo, A. J., J. Song, P. Cheung, S. Ishibe-Murakami, S. Yamazoe, J. K. Chen, D. J. Patel and O. Gozani 2012. The BAH domain of ORC1 links H4K20me2 to DNA replication licensing and Meier-Gorlin syndrome. *Nature* 484(7392): 115-119.

Kurotaki, N., K. Imaizumi, N. Harada, M. Masuno, T. Kondoh, T. Nagai, H. Ohashi, K. Naritomi, *et al.* 2002. Haploinsufficiency of NSD1 causes Sotos syndrome. *Nat Genet* 30(4): 365-366.

Kurpinski, K. T., P. A. Magyari, R. J. Gorlin, D. Ng and L. G. Biesecker 2003. Designation of the TARP syndrome and linkage to Xp11.23-q13.3 without samples from affected patients. *Am J Med Genet A* 120A(1): 1-4.

Kuzminov, A. 2001. Single-strand interruptions in replicating chromosomes cause double-strand breaks. *Proc Natl Acad Sci U S A* 98(15): 8241-8246.

Kyriakis, J. M., H. App, X. F. Zhang, P. Banerjee, D. L. Brautigan, U. R. Rapp and J. Avruch 1992. Raf-1 activates MAP kinase-kinase. *Nature* 358(6385): 417-421.

Lai, F., U. A. Orom, M. Cesaroni, M. Beringer, D. J. Taatjes, G. A. Blobel and R. Shiekhattar 2013. Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. *Nature* 494(7438): 497-501.

Lancaster, M. A. and J. A. Knoblich 2014. Organogenesis in a dish: modeling development and disease using organoid technologies. *Science* 345(6194): 1247125.

Lancaster, M. A., M. Renner, C. A. Martin, D. Wenzel, L. S. Bicknell, M. E. Hurles, T. Homfray, J. M. Penninger, *et al.* 2013. Cerebral organoids model human brain development and microcephaly. *Nature* 501(7467): 373-379.

Landegren, U., R. Kaiser, J. Sanders and L. Hood 1988. A ligase-mediated gene detection technique. *Science* 241(4869): 1077-1080.

Lander, E. S. and D. Botstein 1987. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science* 236(4808): 1567-1570.

Lango Allen, H., K. Estrada, G. Lettre, S. I. Berndt, M. N. Weedon, F. Rivadeneira, C. J. Willer, A. U. Jackson, *et al.* 2010. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 467(7317): 832-838.

Lea, D. E. (1946). Actions of Radiations on Living Cells. Cambridge, Cambridge University Press.

Leevers, S. J. and H. McNeill 2005. Controlling the size of organs and organisms. *Curr Opin Cell Biol* 17(6): 604-609.

Lei, M., Y. Kawasaki, M. R. Young, M. Kihara, A. Sugino and B. K. Tye 1997. Mcm2 is a target of regulation by Cdc7-Dbf4 during the initiation of DNA synthesis. *Genes Dev* 11(24): 3365-3374.

Lei, Q. Y., H. Zhang, B. Zhao, Z. Y. Zha, F. Bai, X. H. Pei, S. Zhao, Y. Xiong and K. L. Guan 2008. TAZ promotes cell proliferation and epithelial-mesenchymal transition and is inhibited by the hippo pathway. *Mol Cell Biol* 28(7): 2426-2436.

Lesca, G., M. P. Moizard, G. Bussy, D. Boggio, H. Hu, S. A. Haas, H. H. Ropers, V. M. Kalscheuer, *et al.* 2013. Clinical and neurocognitive characterization of a family with a novel MED12 gene frameshift mutation. *Am J Med Genet A* 161A(12): 3063-3071.

Li, H. and R. Durbin 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14): 1754-1760.

Li, J., J. W. Xuan, V. Khatamianfar, F. Valiyeva, M. Moussa, A. Sadek, B. B. Yang, B. J. Dong, *et al.* 2014a. SKA1 overexpression promotes centriole over-duplication, centrosome amplification and prostate tumorigenesis. *J Pathol*.

Li, M., K. Suzuki, N. Y. Kim, G. H. Liu and J. C. Izpisua Belmonte 2014b. A cut above the rest: targeted genome editing technologies in human pluripotent stem cells. *J Biol Chem* 289(8): 4594-4599.

Lieber, D. S., S. G. Hershman, N. G. Slate, S. E. Calvo, K. B. Sims, J. D. Schmahmann and V. K. Mootha 2014. Next generation sequencing with copy number variant detection expands the phenotypic spectrum of HSD17B4-deficiency. *BMC Med Genet* 15: 30.

Lieber, M. R. 2008. The mechanism of human nonhomologous DNA end joining. *J Biol Chem* 283(1): 1-5.

Lipp, J. J., T. Hirota, I. Poser and J. M. Peters 2007. Aurora B controls the association of condensin I but not condensin II with mitotic chromosomes. *J Cell Sci* 120(Pt 7): 1245-1255.

Liu, J. J., J. R. Chao, M. C. Jiang, S. Y. Ng, J. J. Yen and H. F. Yang-Yen 1995. Ras transformation results in an elevated level of cyclin D1 and acceleration of G1 progression in NIH 3T3 cells. *Mol Cell Biol* 15(7): 3654-3663.

Liu, J. P., J. Baker, A. S. Perkins, E. J. Robertson and A. Efstratiadis 1993. Mice carrying null mutations of the genes encoding insulin-like growth factor I (Igf-1) and type 1 IGF receptor (Igf1r). *Cell* 75(1): 59-72.

Liu, L., Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu and M. Law 2012. Comparison of next-generation sequencing systems. *J Biomed Biotechnol* 2012: 251364.

Liu, Q., S. Guntuku, X. S. Cui, S. Matsuoka, D. Cortez, K. Tamai, G. Luo, S. Carattini-Rivera*, et al.* 2000. Chk1 is an essential kinase that is regulated by Atr and required for the G(2)/M DNA damage checkpoint. *Genes Dev* 14(12): 1448-1459.

Liu, S., X. Liu, R. P. Kamdar, R. Wanotayan, M. K. Sharma, N. Adachi and Y. Matsumoto 2013a. C-Terminal region of DNA ligase IV drives XRCC4/DNA ligase IV complex to chromatin. *Biochem Biophys Res Commun* 439(2): 173-178.

Liu, X., S. Han, Z. Wang, J. Gelernter and B. Z. Yang 2013b. Variant callers for next-generation sequencing data: a comparison study. *PLoS One* 8(9): e75619.

Liu, Y., W. A. Beard, D. D. Shock, R. Prasad, E. W. Hou and S. H. Wilson 2005. DNA polymerase beta and flap endonuclease 1 enzymatic specificities sustain DNA synthesis for long patch base excision repair. *J Biol Chem* 280(5): 3665-3674.

Loffler, H., T. Bochtler, B. Fritz, B. Tews, A. D. Ho, J. Lukas, J. Bartek and A. Kramer 2007. DNA damage-induced accumulation of centrosomal Chk1 contributes to its checkpoint function. *Cell Cycle* 6(20): 2541-2548.

Loncarek, J., P. Hergert, V. Magidson and A. Khodjakov 2008. Control of daughter centriole formation by the pericentriolar material. *Nat Cell Biol* 10(3): 322-328.

Longworth, M. S., A. Herr, J. Y. Ji and N. J. Dyson 2008. RBF1 promotes chromatin condensation through a conserved interaction with the Condensin II protein dCAP-D3. *Genes Dev* 22(8): 1011-1024.

Love, M. I., A. Mysickova, R. Sun, V. Kalscheuer, M. Vingron and S. A. Haas 2011. Modeling read counts for CNV detection in exome sequencing data. *Stat Appl Genet Mol Biol* 10(1).

Lovejoy, C. A., X. Xu, C. E. Bansbach, G. G. Glick, R. Zhao, F. Ye, B. M. Sirbu, L. C. Titus*, et al.* 2009. Functional genomic screens identify CINP as a genome maintenance protein. *Proc Natl Acad Sci U S A* 106(46): 19304-19309.

Luders, J. and T. Stearns 2007. Microtubule-organizing centres: a re-evaluation. *Nat Rev Mol Cell Biol* 8(2): 161-167.

Lukas, J., C. Lukas and J. Bartek 2004. Mammalian cell cycle checkpoints: signalling pathways and their organization in space and time. *DNA Repair (Amst)* 3(8-9): 997-1007.

Lunter, G. and M. Goodson 2011. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res* 21(6): 936-939.

Ly, D., D. Forman, J. Ferlay, L. A. Brinton and M. B. Cook 2013. An international comparison of male and female breast cancer incidence rates. *Int J Cancer* 132(8): 1918-1926.

Ma, Y., H. Lu, B. Tippin, M. F. Goodman, N. Shimazaki, O. Koiwai, C. L. Hsieh, K. Schwarz and M. R. Lieber 2004. A biochemically defined system for mammalian nonhomologous DNA end joining. *Mol Cell* 16(5): 701-713.

Ma, Y., U. Pannicke, K. Schwarz and M. R. Lieber 2002. Hairpin opening and overhang processing by an Artemis/DNA-dependent protein kinase complex in nonhomologous end joining and V(D)J recombination. *Cell* 108(6): 781-794.

MacArthur, D. G., S. Balasubramanian, A. Frankish, N. Huang, J. Morris, K. Walter, L. Jostins, L. Habegger, *et al.* 2012. A systematic survey of loss-of-function variants in human protein-coding genes. *Science* 335(6070): 823-828.

MacArthur, D. G., T. A. Manolio, D. P. Dimmock, H. L. Rehm, J. Shendure, G. R. Abecasis, D. R. Adams, R. B. Altman, *et al.* 2014. Guidelines for investigating causality of sequence variants in human disease. *Nature* 508(7497): 469-476.

MacDonald, J. R., R. Ziman, R. K. Yuen, L. Feuk and S. W. Scherer 2014. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res* 42(Database issue): D986-992.

Macrae, C. J., R. D. McCulloch, J. Ylanko, D. Durocher and C. A. Koch 2008. APLF (C2orf13) facilitates nonhomologous end-joining and undergoes ATM-dependent hyperphosphorylation following ionizing radiation. *DNA Repair (Amst)* 7(2): 292-302.

Mahaney, B. L., M. Hammel, K. Meek, J. A. Tainer and S. P. Lees-Miller 2013. XRCC4 and XLF form long helical protein filaments suitable for DNA end protection and alignment to facilitate DNA double strand break repair. *Biochem Cell Biol* 91(1): 31-41.

Mahmood, S., W. Ahmad and M. J. Hassan 2011. Autosomal Recessive Primary Microcephaly (MCPH): clinical manifestations, genetic heterogeneity and mutation continuum. *Orphanet J Rare Dis* 6: 39.

Mahowald, G. K., J. M. Baron and B. P. Sleckman 2008. Collateral damage from antigen receptor gene diversification. *Cell* 135(6): 1009-1012.

Majewski, F. and T. Goecke 1982a. Studies of microcephalic primordial dwarfism I: approach to a delineation of the Seckel syndrome. *Am J Med Genet* 12(1): 7-21.

Majewski, F., M. Ranke and A. Schinzel 1982b. Studies of microcephalic primordial dwarfism II: the osteodysplastic type II of primordial dwarfism. *Am J Med Genet* 12(1): 23-35.

Majewski, F., M. Stoeckenius and H. Kemperdick 1982c. Studies of microcephalic primordial dwarfism III: an intrauterine dwarf with platyspondyly and anomalies of pelvis and clavicles--osteodysplastic primordial dwarfism type III. *Am J Med Genet* 12(1): 37-42.

Majewski, J., J. Schwartzentruber, E. Lalonde, A. Montpetit and N. Jabado 2011. What can exome sequencing do for you? *J Med Genet* 48(9): 580-589.

Mamanova, L., A. J. Coffey, C. E. Scott, I. Kozarewa, E. H. Turner, A. Kumar, E. Howard, J. Shendure and D. J. Turner 2010. Target-enrichment strategies for next-generation sequencing. *Nat Methods* 7(2): 111-118.

Mandrile, G., A. Dubois, J. D. Hoffman, V. Uliana, E. Di Maria, M. Malacarne, D. Coviello, F. Faravelli, *et al.* 2013. 3q26.33-3q27.2 microdeletion: a new microdeletion syndrome? *Eur J Med Genet* 56(4): 216-221.

Marchetti, C., S. A. Walker, F. Odreman, A. Vindigni, A. J. Doherty and P. Jeggo 2006. Identification of a novel motif in DNA ligases exemplified by DNA ligase IV. *DNA Repair (Amst)* 5(7): 788-798.

Marechal, A. and L. Zou 2013. DNA damage sensing by the ATM and ATR kinases. *Cold Spring Harb Perspect Biol* 5(9).

Margarit, S. M., H. Sondermann, B. E. Hall, B. Nagar, A. Hoelz, M. Pirruccello, D. Bar-Sagi and J. Kuriyan 2003. Structural evidence for feedback activation by Ras.GTP of the Ras-specific nucleotide exchange factor SOS. *Cell* 112(5): 685-695.

Marheineke, K., O. Hyrien and T. Krude 2005. Visualization of bidirectional initiation of chromosomal DNA replication in a human cell free system. *Nucleic Acids Res* 33(21): 6931-6941.

Mari, F., P. Hermanns, M. L. Giovannucci-Uzielli, F. Galluzzi, D. Scott, B. Lee, A. Renieri, S. Unger, *et al.* 2009. Refinement of the 12q14 microdeletion syndrome: primordial dwarfism and developmental delay with or without osteopoikilosis. *Eur J Hum Genet* 17(9): 1141-1147.

Martens, M. A., S. J. Wilson and D. C. Reutens 2008. Research Review: Williams syndrome: a critical review of the cognitive, behavioral, and neuroanatomical phenotype. *J Child Psychol Psychiatry* 49(6): 576-608.

Martin, C. A., I. Ahmad, A. Klingseisen, M. S. Hussain, L. S. Bicknell, A. Leitch, G. Nurnberg, M. R. Toliat, *et al.* 2014. Mutations in PLK4, encoding a master regulator of centriole biogenesis, cause microcephaly, growth failure and retinopathy. *Nat Genet* 46(12): 1283-1292.

Martin, G. M., A. C. Smith, D. J. Ketterer, C. E. Ogburn and C. M. Disteche 1985. Increased chromosomal aberrations in first metaphases of cells isolated from the kidneys of aged mice. *Isr J Med Sci* 21(3): 296-301.

Matsumoto, Y. and J. L. Maller 2004. A centrosomal localization signal in cyclin E required for Cdk2-independent S phase entry. *Science* 306(5697): 885-888.

Matsumoto, Y., T. Miyamoto, H. Sakamoto, H. Izumi, Y. Nakazawa, T. Ogi, H. Tahara, S. Oku, *et al.* 2011. Two unrelated patients with MRE11A mutations and Nijmegen breakage syndrome-like severe microcephaly. *DNA Repair (Amst)* 10(3): 314-321.

Matsuoka, S., M. Huang and S. J. Elledge 1998. Linkage of ATM to cell cycle regulation by the Chk2 protein kinase. *Science* 282(5395): 1893-1897.

Matsuura, S., H. Tauchi, A. Nakamura, N. Kondo, S. Sakamoto, S. Endo, D. Smeets, B. Solder, *et al.* 1998. Positional cloning of the gene for Nijmegen breakage syndrome. *Nat Genet* 19(2): 179-181.

Mazaika, E. and J. Homsy 2014. Digital Droplet PCR: CNV Analysis and Other Applications. *Curr Protoc Hum Genet* 82: 7 24 21-27 24 13.

McClintock, B. 1941. The Stability of Broken Ends of Chromosomes in Zea Mays. *Genetics* 26(2): 234-282.

McKay, M. M. and D. K. Morrison 2007. Integrating signals from RTKs to ERK/MAPK. *Oncogene* 26(22): 3113-3121.

McKusick, V. 1955. Primordial dwarfism and ectopia lentis. *Am J Hum Genet* 7(2): 189-198.

Medvedev, P., M. Stanciu and M. Brudno 2009. Computational methods for discovering structural variation with next-generation sequencing. *Nat Methods* 6(11 Suppl): S13-20.

Megraw, T. L., J. T. Sharkey and R. S. Nowakowski 2011. Cdk5rap2 exposes the centrosomal root of microcephaly syndromes. *Trends Cell Biol* 21(8): 470-480.

Meinecke, P. and E. Passarge 1991. Microcephalic osteodysplastic primordial dwarfism type I/III in sibs. *J Med Genet* 28(11): 795-800.

Melis, J. P., H. van Steeg and M. Luijten 2013. Oxidative DNA damage and nucleotide excision repair. *Antioxid Redox Signal* 18(18): 2409-2419.

Meloche, S. and J. Pouyssegur 2007. The ERK1/2 mitogen-activated protein kinase pathway as a master regulator of the G1- to S-phase transition. *Oncogene* 26(22): 3227-3239.

Mendez, J. and B. Stillman 2003. Perpetuating the double helix: molecular machines at eukaryotic DNA replication origins. *Bioessays* 25(12): 1158-1167.

Mendoza, M. C., E. E. Er and J. Blenis 2011. The Ras-ERK and PI3K-mTOR pathways: cross-talk and compensation. *Trends Biochem Sci* 36(6): 320-328.

Mester, J. and C. Eng 2013. When overgrowth bumps into cancer: the PTEN-opathies. *Am J Med Genet C Semin Med Genet* 163C(2): 114-121.

Metzker, M. L. 2010. Sequencing technologies - the next generation. *Nat Rev Genet* 11(1): 31-46.

Meyer, M., U. Stenzel, S. Myles, K. Prufer and M. Hofreiter 2007. Targeted high-throughput sequencing of tagged nucleic acid samples. *Nucleic Acids Res* 35(15): e97.

Meynert, A. M., L. S. Bicknell, M. E. Hurles, A. P. Jackson and M. S. Taylor 2013. Quantifying single nucleotide variant detection sensitivity in exome sequencing. *BMC Bioinformatics* 14: 195.

Millat, G., V. Chanavat and R. Rousson 2014. Evaluation of a New High-Throughput Next-Generation Sequencing Method Based on a Custom AmpliSeq Library and Ion Torrent PGM Sequencing for the Rapid Detection of Genetic Variations in Long QT Syndrome. *Mol Diagn Ther*.

Miller, D. T., M. P. Adam, S. Aradhya, L. G. Biesecker, A. R. Brothman, N. P. Carter, D. M. Church, J. A. Crolla, *et al.* 2010. Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *Am J Hum Genet* 86(5): 749-764.

Mills, R. E., W. S. Pittard, J. M. Mullaney, U. Farooq, T. H. Creasy, A. A. Mahurkar, D. M. Kemeza, D. S. Strassler, *et al.* 2011a. Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Res* 21(6): 830-839.

Mills, R. E., K. Walter, C. Stewart, R. E. Handsaker, K. Chen, C. Alkan, A. Abyzov, S. C. Yoon*, et al.* 2011b. Mapping copy number variation by population-scale genome sequencing. *Nature* 470(7332): 59-65.

Mimori, T. and J. A. Hardin 1986. Mechanism of interaction between Ku protein and DNA. *J Biol Chem* 261(22): 10375-10379.

Mirzaa, G. M., B. Vitre, G. Carpenter, I. Abramowicz, J. G. Gleeson, A. R. Paciorkowski, D. W. Cleveland, W. B. Dobyns and M. O'Driscoll 2014. Mutations in CENPE define a novel kinetochore-centromeric mechanism for microcephalic primordial dwarfism. *Hum Genet* 133(8): 1023-1039.

Mochida, G. H. and C. A. Walsh 2001. Molecular genetics of human microcephaly. *Curr Opin Neurol* 14(2): 151-156.

Mokrani-Benhelli, H., L. Gaillard, P. Biasutto, T. Le Guen, F. Touzot, N. Vasquez, J. Komatsu, E. Conseiller*, et al.* 2013. Primary microcephaly, impaired DNA replication, and genomic instability caused by compound heterozygous ATR mutations. *Hum Mutat* 34(2): 374-384.

Monteiro, F. P., T. P. Vieira, I. C. Sgardioli, M. C. Molck, A. P. Damiano, J. Souza, I. L. Monlleo, M. I. Fontes*, et al.* 2013. Defining new guidelines for screening the 22q11.2 deletion based on a clinical and dysmorphologic evaluation of 194 individuals and review of the literature. *Eur J Pediatr* 172(7): 927-945.

Moshous, D., I. Callebaut, R. de Chasseval, B. Corneo, M. Cavazzana-Calvo, F. Le Deist, I. Tezcan, O. Sanal*, et al.* 2001. Artemis, a novel DNA double-strand break repair/V(D)J recombination protein, is mutated in human severe combined immune deficiency. *Cell* 105(2): 177-186.

Moyer, S. E., P. W. Lewis and M. R. Botchan 2006. Isolation of the Cdc45/Mcm2-7/GINS (CMG) complex, a candidate for the eukaryotic DNA replication fork helicase. *Proc Natl Acad Sci U S A* 103(27): 10236-10241.

Mueller, A. C., M. A. Keaton and A. Dutta 2011. DNA replication: mammalian Treslin-TopBP1 interaction mirrors yeast Sld3-Dpb11. *Curr Biol* 21(16): R638-640.

Murray, J. E., L. S. Bicknell, G. Yigit, A. L. Duker, M. van Kogelenberg, S. Haghayegh, D. Wieczorek, H. Kayserili*, et al.* 2014. Extreme growth failure is a common presentation of ligase IV deficiency. *Hum Mutat* 35(1): 76-85.

Nagy, R., H. Wang, B. Albrecht, D. Wieczorek, G. Gillessen-Kaesbach, E. Haan, P. Meinecke, A. de la Chapelle and J. A. Westman 2012. Microcephalic osteodysplastic primordial dwarfism type I with biallelic mutations in the RNU4ATAC gene. *Clin Genet* 82(2): 140-146.

Nance, M. A. and S. A. Berry 1992. Cockayne syndrome: review of 140 cases. *Am J Med Genet* 42(1): 68-84.

Nativio, R., A. Sparago, Y. Ito, R. Weksberg, A. Riccio and A. Murrell 2011. Disruption of genomic neighbourhood at the imprinted IGF2-H19 locus in Beckwith-Wiedemann syndrome and Silver-Russell syndrome. *Hum Mol Genet* 20(7): 1363-1374.

Netchine, I., S. Rossignol, M. N. Dufourg, S. Azzi, A. Rousseau, L. Perin, M. Houang, V. Steunou*, et al.* 2007. 11p15 imprinting center region 1 loss of methylation is a common and specific cause of typical Russell-Silver syndrome: clinical scoring system and epigenetic-phenotypic correlations. *J Clin Endocrinol Metab* 92(8): 3148-3154.

Nevado, J., R. Mergener, M. Palomares-Bralo, K. R. Souza, E. Vallespin, R. Mena, V. Martinez-Glez, M. A. Mori, *et al.* 2014. New microdeletion and microduplication syndromes: A comprehensive review. *Genet Mol Biol* 37(1 Suppl): 210-219.

Nevis, K. R., M. Cordeiro-Stone and J. G. Cook 2009. Origin licensing and p53 status regulate Cdk2 activity during G(1). *Cell Cycle* 8(12): 1952-1963.

Ng, S. B., A. W. Bigham, K. J. Buckingham, M. C. Hannibal, M. J. McMillin, H. I. Gildersleeve, A. E. Beck, H. K. Tabor, *et al.* 2010a. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat Genet* 42(9): 790-793.

Ng, S. B., K. J. Buckingham, C. Lee, A. W. Bigham, H. K. Tabor, K. M. Dent, C. D. Huff, P. T. Shannon, *et al.* 2010b. Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* 42(1): 30-35.

Ng, S. B., E. H. Turner, P. D. Robertson, S. D. Flygare, A. W. Bigham, C. Lee, T. Shaffer, M. Wong, *et al.* 2009. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 461(7261): 272-276.

Nielsen, J., J. Christiansen, J. Lykke-Andersen, A. H. Johnsen, U. M. Wewer and F. C. Nielsen 1999. A family of insulin-like growth factor II mRNA-binding proteins represses translation in late development. *Mol Cell Biol* 19(2): 1262-1270.

Nijnik, A., L. Woodbine, C. Marchetti, S. Dawson, T. Lambe, C. Liu, N. P. Rodrigues, T. L. Crockford, *et al.* 2007. DNA repair is limiting for haematopoietic stem cells during ageing. *Nature* 447(7145): 686-690.

Nikkel, S. M., A. Dauber, S. de Munnik, M. Connolly, R. L. Hood, O. Caluseriu, J. Hurst, U. Kini, *et al.* 2013. The phenotype of floating-harbor syndrome: clinical characterization of 52 individuals with mutations in exon 34 of SRCAP. *Orphanet J Rare Dis* 8(1): 63.

Nimura, K., K. Ura, H. Shiratori, M. Ikawa, M. Okabe, R. J. Schwartz and Y. Kaneda 2009. A histone H3 lysine 36 trimethyltransferase links Nkx2-5 to Wolf-Hirschhorn syndrome. *Nature* 460(7252): 287-291.

Nissenkorn, A., Y. B. Levi, D. Vilozni, Y. Berkun, O. Efrati, M. Frydman, J. Yahav, D. Waldman, *et al.* 2011. Neurologic presentation in children with ataxia-telangiectasia: is small head circumference a hallmark of the disease? *J Pediatr* 159(3): 466-471 e461.

Noctor, S. C., V. Martinez-Cerdeno, L. Ivic and A. R. Kriegstein 2004. Cortical neurons arise in symmetric and asymmetric division zones and migrate through specific phases. *Nat Neurosci* 7(2): 136-144.

Noguchi, K., A. Vassilev, S. Ghosh, J. L. Yates and M. L. DePamphilis 2006. The BAH domain facilitates the ability of human Orc1 protein to activate replication origins in vivo. *EMBO J* 25(22): 5372-5382.

Norton, N., P. D. Robertson, M. J. Rieder, S. Zuchner, E. Rampersaud, E. Martin, D. Li, D. A. Nickerson, *et al.* 2012. Evaluating pathogenicity of rare variants from dilated cardiomyopathy in the exome era. *Circ Cardiovasc Genet* 5(2): 167-174.

Nussenzweig, A., C. Chen, V. da Costa Soares, M. Sanchez, K. Sokol, M. C. Nussenzweig and G. C. Li 1996. Requirement for Ku80 in growth and immunoglobulin V(D)J recombination. *Nature* 382(6591): 551-555.

O'Connell, C. B. and A. L. Khodjakov 2007. Cooperative mechanisms of mitotic spindle formation. *J Cell Sci* 120(Pt 10): 1717-1722.

O'Driscoll, M. 2012. Diseases associated with defective responses to DNA damage. *Cold Spring Harb Perspect Biol* 4(12).

O'Driscoll, M., K. M. Cerosaletti, P. M. Girard, Y. Dai, M. Stumm, B. Kysela, B. Hirsch, A. Gennery, *et al.* 2001. DNA ligase IV mutations identified in patients exhibiting developmental delay and immunodeficiency. *Mol Cell* 8(6): 1175-1185.

O'Driscoll, M., A. R. Gennery, J. Seidel, P. Concannon and P. A. Jeggo 2004. An overview of three new disorders associated with genetic instability: LIG4 syndrome, RS-SCID and ATR-Seckel syndrome. *DNA Repair (Amst)* 3(8-9): 1227-1235.

O'Driscoll, M. and P. A. Jeggo 2008. CsA can induce DNA double-strand breaks: implications for BMT regimens particularly for individuals with defective DNA repair. *Bone Marrow Transplant* 41(11): 983-989.

O'Driscoll, M., V. L. Ruiz-Perez, C. G. Woods, P. A. Jeggo and J. A. Goodship 2003. A splicing mutation affecting expression of ataxia-telangiectasia and Rad3-related protein (ATR) results in Seckel syndrome. *Nat Genet* 33(4): 497-501.

Ogi, T., S. Walker, T. Stiff, E. Hobson, S. Limsirichaikul, G. Carpenter, K. Prescott, M. Suri, *et al.* 2012. Identification of the first ATRIP-deficient patient and novel mutations in ATR define a clinical spectrum for ATR-ATRIP Seckel Syndrome. *PLoS Genet* 8(11): e1002945.

Olshen, A. B., E. S. Venkatraman, R. Lucito and M. Wigler 2004. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 5(4): 557-572.

Onn, I., N. Aono, M. Hirano and T. Hirano 2007. Reconstitution and subunit geometry of human condensin complexes. *EMBO J* 26(4): 1024-1034.

Ono, T., Y. Fang, D. L. Spector and T. Hirano 2004. Spatial and temporal regulation of Condensins I and II in mitotic chromosome assembly in human cells. *Mol Biol Cell* 15(7): 3296-3308.

Ono, T., A. Losada, M. Hirano, M. P. Myers, A. F. Neuwald and T. Hirano 2003. Differential contributions of condensin I and condensin II to mitotic chromosome architecture in vertebrate cells. *Cell* 115(1): 109-121.

Orii, K. E., Y. Lee, N. Kondo and P. J. McKinnon 2006. Selective utilization of nonhomologous end-joining and homologous recombination DNA repair pathways during nervous system development. *Proc Natl Acad Sci U S A* 103(26): 10017-10022.

Ota, M. and H. Sasaki 2008. Mammalian Tead proteins regulate cell proliferation and contact inhibition as transcriptional mediators of Hippo signaling. *Development* 135(24): 4059-4069.

Oyola, S. O., T. D. Otto, Y. Gu, G. Maslen, M. Manske, S. Campino, D. J. Turner, B. Macinnis, *et al.* 2012. Optimizing Illumina next-generation sequencing library preparation for extremely AT-biased genomes. *BMC Genomics* 13: 1.

Ozeri-Galai, E., R. Lebofsky, A. Rahat, A. C. Bester, A. Bensimon and B. Kerem 2011. Failure of origin activation in response to fork stalling leads to chromosomal instability at fragile sites. *Mol Cell* 43(1): 122-131.

Paintrand, M., M. Moudjou, H. Delacroix and M. Bornens 1992. Centrosome organization and centriole architecture: their sensitivity to divalent cations. *J Struct Biol* 108(2): 107-128.

Pan-Hammarstrom, Q., A. M. Jones, A. Lahdesmaki, W. Zhou, R. A. Gatti, L. Hammarstrom, A. R. Gennery and M. R. Ehrenstein 2005. Impact of DNA ligase IV on nonhomologous end joining pathways during class switch recombination in human cells. *J Exp Med* 201(2): 189-194.

Pan, D. 2007. Hippo signaling in organ size control. *Genes Dev* 21(8): 886-897.

Panaro, N. J., P. K. Yuen, T. Sakazume, P. Fortina, L. J. Kricka and P. Wilding 2000. Evaluation of DNA fragment sizing and quantification by the agilent 2100 bioanalyzer. *Clin Chem* 46(11): 1851-1853.

Parameswaran, P., R. Jalili, L. Tao, S. Shokralla, B. Gharizadeh, M. Ronaghi and A. Z. Fire 2007. A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. *Nucleic Acids Res* 35(19): e130.

Paul, K., M. Wang, E. Mladenov, A. Bencsik-Theilen, T. Bednar, W. Wu, H. Arakawa and G. Iliakis 2013. DNA ligases I and III cooperate in alternative non-homologous end-joining in vertebrates. *PLoS One* 8(3): e59505.

Peng, H. W., M. Slattery and R. S. Mann 2009. Transcription factor choice in the Hippo signaling pathway: homothorax and yorkie regulation of the microRNA bantam in the progenitor domain of the Drosophila eye imaginal disc. *Genes Dev* 23(19): 2307-2319.

Perry, L. D., F. Robertson and V. Ganesan 2013. Screening for cerebrovascular disease in microcephalic osteodysplastic primordial dwarfism type II (MOPD II): an evidence-based proposal. *Pediatr Neurol* 48(4): 294-298.

Piazza, I., C. H. Haering and A. Rutkowska 2013. Condensin: crafting the chromosome landscape. *Chromosoma* 122(3): 175-190.

Piel, M., P. Meyer, A. Khodjakov, C. L. Rieder and M. Bornens 2000. The respective contributions of the mother and daughter centrioles to centrosome activity and behavior in vertebrate cells. *J Cell Biol* 149(2): 317-330.

Pierce, M. J. and R. P. Morse 2012. The neurologic findings in Taybi-Linder syndrome (MOPD I/III): case report and review of the literature. *Am J Med Genet A* 158A(3): 606-610.

Plowman, P. N., B. A. Bridges, C. F. Arlett, A. Hinney and J. E. Kingston 1990. An instance of clinical radiation morbidity and cellular radiosensitivity, not associated with ataxia-telangiectasia. *Br J Radiol* 63(752): 624-628.

Pontious, A., T. Kowalczyk, C. Englund and R. F. Hevner 2008. Role of intermediate progenitor cells in cerebral cortex development. *Dev Neurosci* 30(1-3): 24-32.

Pouliot, J. J., K. C. Yao, C. A. Robertson and H. A. Nash 1999. Yeast gene for a Tyr-DNA phosphodiesterase that repairs topoisomerase I complexes. *Science* 286(5439): 552-555.

Poultney, C. S., A. P. Goldberg, E. Drapeau, Y. Kou, H. Harony-Nicolas, Y. Kajiwara, S. De Rubeis, S. Durand, *et al.* 2013. Identification of small exonic CNV from whole-exome sequence data and application to autism spectrum disorder. *Am J Hum Genet* 93(4): 607-619.

Pritchard, C. C., C. Smith, T. Marushchak, K. Koehler, H. Holmes, W. Raskind, T. Walsh and R. L. Bennett 2013. A mosaic PTEN mutation causing Cowden syndrome identified by deep sequencing. *Genet Med* 15(12): 1004-1007.

Puffenberger, E. G., R. N. Jinks, C. Sougnez, K. Cibulskis, R. A. Willert, N. P. Achilly, R. P. Cassidy, C. J. Fiorentini, *et al.* 2012. Genetic mapping and exome sequencing identify variants associated with five novel diseases. *PLoS One* 7(1): e28936.

Quail, M. A., I. Kozarewa, F. Smith, A. Scally, P. J. Stephens, R. Durbin, H. Swerdlow and D. J. Turner 2008. A large genome center's improvements to the Illumina sequencing system. *Nat Methods* 5(12): 1005-1010.

Qvist, P., P. Huertas, S. Jimeno, M. Nyegaard, M. J. Hassan, S. P. Jackson and A. D. Borglum 2011. CtIP Mutations Cause Seckel and Jawad Syndromes. *PLoS Genet* 7(10): e1002310.

Raca, G., C. Jackson, B. Warman, T. Bair and L. A. Schimmenti 2010. Next generation sequencing in research and diagnostics of ocular birth defects. *Mol Genet Metab* 100(2): 184-192.

Ramadan, K., I. V. Shevelev, G. Maga and U. Hubscher 2004. De novo DNA synthesis by human DNA polymerase lambda, DNA polymerase mu and terminal deoxyribonucleotidyl transferase. *J Mol Biol* 339(2): 395-404.

Rauch, A. 2011. The shortest of the short: pericentrin mutations and beyond. *Best Pract Res Clin Endocrinol Metab* 25(1): 125-130.

Rauch, A., C. T. Thiel, D. Schindler, U. Wick, Y. J. Crow, A. B. Ekici, A. J. van Essen, T. O. Goecke, *et al.* 2008. Mutations in the pericentrin (PCNT) gene cause primordial dwarfism. *Science* 319(5864): 816-819.

Rawlings, J. S., M. Gatzka, P. G. Thomas and J. N. Ihle 2011. Chromatin condensation via the condensin II complex is required for peripheral T-cell quiescence. *EMBO J* 30(2): 263-276.

Rebollo, E., P. Sampaio, J. Januschke, S. Llamazares, H. Varmark and C. Gonzalez 2007. Functionally unequal centrosomes drive spindle orientation in asymmetrically dividing Drosophila neural stem cells. *Dev Cell* 12(3): 467-474.

Redon, R., S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H. Shapero, *et al.* 2006. Global variation in copy number in the human genome. *Nature* 444(7118): 444-454.

Reed, D. R., M. P. Lawler and M. G. Tordoff 2008. Reduced body weight is a common effect of gene knockout in mice. *BMC Genet* 9: 4.

Reijns, M. (2006). An analysis of Lsm protein complexes. Doctorate of Philosophy, University of Edinburgh.

Riballo, E., S. E. Critchlow, S. H. Teo, A. J. Doherty, A. Priestley, B. Broughton, B. Kysela, H. Beamish, *et al.* 1999. Identification of a defect in DNA ligase IV in a radiosensitive leukaemia patient. *Curr Biol* 9(13): 699-702.

Riballo, E., A. J. Doherty, Y. Dai, T. Stiff, M. A. Oettinger, P. A. Jeggo and B. Kysela 2001. Cellular and biochemical impact of a mutation in DNA ligase IV conferring clinical radiosensitivity. *J Biol Chem* 276(33): 31124-31132.

Richards, A. J., A. McNinch, J. Whittaker, B. Treacy, K. Oakhill, A. Poulson and M. P. Snead 2012. Splicing analysis of unclassified variants in COL2A1 and COL11A1 identifies deep intronic pathogenic mutations. *Eur J Hum Genet* 20(5): 552-558.

Ried, T., E. Schrock, Y. Ning and J. Wienberg 1998. Chromosome painting: a useful art. *Hum Mol Genet* 7(10): 1619-1626.

Riha, K., M. L. Heacock and D. E. Shippen 2006. The role of the nonhomologous end-joining DNA double-strand break repair pathway in telomere biology. *Annu Rev Genet* 40: 237-277.

Risheg, H., J. M. Graham, Jr., R. D. Clark, R. C. Rogers, J. M. Opitz, J. B. Moeschler, A. P. Peiffer, M. May*, et al.* 2007. A recurrent mutation in MED12 leading to R961W causes Opitz-Kaveggia syndrome. *Nat Genet* 39(4): 451-453.

Robbins, E., G. Jentzsch and A. Micali 1968. The centriole cycle in synchronized HeLa cells. *J Cell Biol* 36(2): 329-339.

Roberts, A. E., T. Araki, K. D. Swanson, K. T. Montgomery, T. A. Schiripo, V. A. Joshi, L. Li, Y. Yassin*, et al.* 2007. Germline gain-of-function mutations in SOS1 cause Noonan syndrome. *Nat Genet* 39(1): 70-74.

Roberts, C. T., Jr., S. R. Lasky, W. L. Lowe, Jr., W. T. Seaman and D. LeRoith 1987. Molecular cloning of rat insulin-like growth factor I complementary deoxyribonucleic acids: differential messenger ribonucleic acid processing and regulation by growth hormone in extrahepatic tissues. *Mol Endocrinol* 1(3): 243-248.

Robinson, J. T., H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz and J. P. Mesirov 2011. Integrative genomics viewer. *Nat Biotechnol* 29(1): 24-26.

Robinson, M. J., M. Cheng, A. Khokhlatchev, D. Ebert, N. Ahn, K. L. Guan, B. Stein, E. Goldsmith and M. H. Cobb 1996. Contributions of the mitogen-activated protein (MAP) kinase backbone and phosphorylation loop to MEK specificity. *J Biol Chem* 271(47): 29734-29739.

Roddam, P. L., S. Rollinson, M. O'Driscoll, P. A. Jeggo, A. Jack and G. J. Morgan 2002. Genetic variants of NHEJ DNA ligase IV can affect the risk of developing multiple myeloma, a tumour characterised by aberrant class switch recombination. *J Med Genet* 39(12): 900-905.

Rogol, A. D. and G. F. Hayden 2014. Etiologies and early diagnosis of short stature and growth failure in children and adolescents. *J Pediatr* 164(5 Suppl): S1-14 e16.

Rooney, S., J. Sekiguchi, C. Zhu, H. L. Cheng, J. Manis, S. Whitlow, J. DeVido, D. Foy*, et al.* 2002. Leaky Scid phenotype associated with defective V(D)J coding end processing in Artemis-deficient mice. *Mol Cell* 10(6): 1379-1390.

Ropars, V., P. Drevet, P. Legrand, S. Baconnais, J. Amram, G. Faure, J. A. Marquez, O. Pietrement*, et al.* 2011. Structural characterization of filaments formed by human Xrcc4-Cernunnos/XLF complex involved in nonhomologous DNA end-joining. *Proc Natl Acad Sci U S A* 108(31): 12663-12668.

Roth, D. B., T. N. Porter and J. H. Wilson 1985. Mechanisms of nonhomologous recombination in mammalian cells. *Mol Cell Biol* 5(10): 2599-2607.

Rothberg, J. M., W. Hinz, T. M. Rearick, J. Schultz, W. Mileski, M. Davey, J. H. Leamon, K. Johnson*, et al.* 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 475(7356): 348-352.

Rothkamm, K., I. Kruger, L. H. Thompson and M. Lobrich 2003. Pathways of DNA double-strand break repair during the mammalian cell cycle. *Mol Cell Biol* 23(16): 5706-5715.

Roy, S., S. N. Andres, A. Vergnes, J. A. Neal, Y. Xu, Y. Yu, S. P. Lees-Miller, M. Junop, *et al.* 2012. XRCC4's interaction with XLF is required for coding (but not signal) end joining. *Nucleic Acids Res* 40(4): 1684-1694.

Rucci, F., L. D. Notarangelo, A. Fazeli, L. Patrizi, T. Hickernell, T. Paganini, K. M. Coakley, C. Detre, *et al.* 2010. Homozygous DNA ligase IV R278H mutation in mice leads to leaky SCID and represents a model for human LIG4 syndrome. *Proc Natl Acad Sci U S A* 107(7): 3024-3029.

Rudaks, L. I., J. K. Nicholl, D. Bratkovic and C. P. Barnett 2011. Short stature due to 15q26 microdeletion involving IGF1R: report of an additional case and review of the literature. *Am J Med Genet A* 155A(12): 3139-3143.

Rusan, N. M. and M. Peifer 2007. A role for a novel centrosome cycle in asymmetric cell division. *J Cell Biol* 177(1): 13-20.

Saitoh, N., I. G. Goldberg, E. R. Wood and W. C. Earnshaw 1994. ScII: an abundant chromosome scaffold protein is a member of a family of putative ATPases with an unusual predicted tertiary structure. *J Cell Biol* 127(2): 303-318.

Samarakoon, P. S., H. S. Sorte, B. E. Kristiansen, T. Skodje, Y. Sheng, G. E. Tjonnfjord, B. Stadheim, A. Stray-Pedersen, *et al.* 2014. Identification of copy number variants from exome sequence data. *BMC Genomics* 15: 661.

San Filippo, J., P. Sung and H. Klein 2008. Mechanism of eukaryotic homologous recombination. *Annu Rev Biochem* 77: 229-257.

Sanger, F., S. Nicklen and A. R. Coulson 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* 74(12): 5463-5467.

Sankaranarayanan, K. 1979. The role of non-disjunction in aneuploidy in man. An overview. *Mutat Res* 61(1): 1-28.

SantaLucia, J., Jr. 1998. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc Natl Acad Sci U S A* 95(4): 1460-1465.

Sarici, D., M. A. Akin, A. Kara, S. Doganay and S. Kurtoglu 2012. Seckel syndrome accompanied by semilobar holoprosencephaly and arthrogryposis. *Pediatr Neurol* 46(3): 189-191.

Sartori, A. A., C. Lukas, J. Coates, M. Mistrik, S. Fu, J. Bartek, R. Baer, J. Lukas and S. P. Jackson 2007. Human CtIP promotes DNA end resection. *Nature* 450(7169): 509-514.

Sathirapongsasuti, J. F., H. Lee, B. A. Horst, G. Brunner, A. J. Cochran, S. Binder, J. Quackenbush and S. F. Nelson 2011. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics* 27(19): 2648-2654.

Savage, S. A. and A. A. Bertuch 2010. The genetics and clinical manifestations of telomere biology disorders. *Genet Med* 12(12): 753-764.

Savitsky, K., A. Bar-Shira, S. Gilad, G. Rotman, Y. Ziv, L. Vanagaite, D. A. Tagle, S. Smith, *et al.* 1995. A single ataxia telangiectasia gene with a product similar to PI-3 kinase. *Science* 268(5218): 1749-1753.

Scarano, E., M. Iaccarino, P. Grippo and E. Parisi 1967. The heterogeneity of thymine methyl group origin in DNA pyrimidine isostichs of developing sea urchin embryos. *Proc Natl Acad Sci U S A* 57(5): 1394-1400.

Schleiffer, A., S. Kaitna, S. Maurer-Stroh, M. Glotzer, K. Nasmyth and F. Eisenhaber 2003. Kleisins: a superfamily of bacterial and eukaryotic SMC protein partners. *Mol Cell* 11(3): 571-575.

Schmidts, M., H. H. Arts, E. M. Bongers, Z. Yap, M. M. Oud, D. Antony, L. Duijkers, R. D. Emes, *et al.* 2013. Exome sequencing identifies DYNC2H1 mutations as a common cause of asphyxiating thoracic dystrophy (Jeune syndrome) without major polydactyly, renal or retinal involvement. *J Med Genet* 50(5): 309-323.

Schouten, J. P., C. J. McElgunn, R. Waaijer, D. Zwijnenburg, F. Diepvens and G. Pals 2002. Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res* 30(12): e57.

Schule, B., A. Oviedo, K. Johnston, S. Pai and U. Francke 2005. Inactivating mutations in ESCO2 cause SC phocomelia and Roberts syndrome: no phenotype-genotype correlation. *Am J Hum Genet* 77(6): 1117-1128.

Schuler, N. and C. E. Rube 2013. Accumulation of DNA damage-induced chromatin alterations in tissue-specific stem cells: the driving force of aging? *PLoS One* 8(5): e63932.

Schwartz, C. E., P. S. Tarpey, H. A. Lubs, A. Verloes, M. M. May, H. Risheg, M. J. Friez, P. A. Futreal, *et al.* 2007. The original Lujan syndrome family has a novel missense mutation (p.N1007S) in the MED12 gene. *J Med Genet* 44(7): 472-477.

Scott, P. H., G. J. Brunn, A. D. Kohn, R. A. Roth and J. C. Lawrence, Jr. 1998. Evidence of insulin-stimulated phosphorylation and activation of the mammalian target of rapamycin mediated by a protein kinase B signaling pathway. *Proc Natl Acad Sci U S A* 95(13): 7772-7777.

Sears, R., F. Nuckolls, E. Haura, Y. Taya, K. Tamai and J. R. Nevins 2000. Multiple Ras-dependent phosphorylation pathways regulate Myc protein stability. *Genes Dev* 14(19): 2501-2514.

Seckel, H. P. G. (1960). Bird-headed Dwarfs: Studies in Developmental Anthropology Including Human Proportions. Springfield, IL, Charles C Thomas.

Seelow, D. and M. Schuelke 2012. HomozygosityMapper2012--bridging the gap between homozygosity mapping and deep sequencing. *Nucleic Acids Res* 40(Web Server issue): W516-520.

Seelow, D., M. Schuelke, F. Hildebrandt and P. Nurnberg 2009. HomozygosityMapper--an interactive approach to homozygosity mapping. *Nucleic Acids Res* 37(Web Server issue): W593-599.

Seifert, W., M. Holder-Espinasse, S. Spranger, M. Hoeltzenbein, E. Rossier, H. Dollfus, D. Lacombe, A. Verloes, *et al.* 2006. Mutational spectrum of COH1 and clinical heterogeneity in Cohen syndrome. *J Med Genet* 43(5): e22.

Seipold, S., F. C. Priller, P. Goldsmith, W. A. Harris, H. Baier and S. Abdelilah-Seyfried 2009. Non-SMC condensin I complex proteins control chromosome segregation and survival of proliferating cells in the zebrafish neural retina. *BMC Dev Biol* 9: 40.

Sekiguchi, J., D. O. Ferguson, H. T. Chen, E. M. Yang, J. Earle, K. Frank, S. Whitlow, Y. Gu, *et al.* 2001. Genetic interactions between ATM and the nonhomologous end-joining factors in genomic stability and development. *Proc Natl Acad Sci U S A* 98(6): 3243-3248.

Shaheen, R., E. Faqeih, S. Ansari, G. Abdel-Salam, Z. N. Al-Hassnan, T. Al-Shidi, R. Alomar, S. Sogaty and F. S. Alkuraya 2014. Genomic analysis of primordial dwarfism reveals novel disease genes. *Genome Res* 24(2): 291-299.

Shan, J., W. Zhao and W. Gu 2009. Suppression of cancer cell growth by promoting cyclin D1 degradation. *Mol Cell* 36(3): 469-476.

Shang, J., F. Zhu, W. Vongsangnak, Y. Tang, W. Zhang and B. Shen 2014. Evaluation and comparison of multiple aligners for next-generation sequencing data analysis. *Biomed Res Int* 2014: 309650.

Shaw-Smith, C., R. Redon, L. Rickman, M. Rio, L. Willatt, H. Fiegler, H. Firth, D. Sanlaville*, et al.* 2004. Microarray based comparative genomic hybridisation (array-CGH) detects submicroscopic chromosomal deletions and duplications in patients with learning disability/mental retardation and dysmorphic features. *J Med Genet* 41(4): 241-248.

Shen, J., E. C. Gilmore, C. A. Marshall, M. Haddadin, J. J. Reynolds, W. Eyaid, A. Bodell, B. Barry*, et al.* 2010. Mutations in PNKP cause microcephaly, seizures and defects in DNA repair. *Nat Genet* 42(3): 245-249.

Sherr, C. J. 1995. D-type cyclins. *Trends Biochem Sci* 20(5): 187-190.

Sherry, S. T., M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski and K. Sirotkin 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 29(1): 308-311.

Sheth, N., X. Roca, M. L. Hastings, T. Roeder, A. R. Krainer and R. Sachidanandam 2006. Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res* 34(14): 3955-3967.

Shiloh, Y. 2006. The ATM-mediated DNA-damage response: taking shape. *Trends Biochem Sci* 31(7): 402-410.

Shin, S. C., H. Ahn do, S. J. Kim, H. Lee, T. J. Oh, J. E. Lee and H. Park 2013. Advantages of Single-Molecule Real-Time Sequencing in High-GC Content Genomes. *PLoS One* 8(7): e68824.

Shinohara, H., N. Sakayori, M. Takahashi and N. Osumi 2013. Ninein is essential for the maintenance of the cortical progenitor character by anchoring the centrosome to microtubules. *Biol Open* 2(7): 739-749.

Shuman, S. and B. Schwer 1995. RNA capping enzyme and DNA ligase: a superfamily of covalent nucleotidyl transferases. *Mol Microbiol* 17(3): 405-410.

Sibanda, B. L., S. E. Critchlow, J. Begun, X. Y. Pei, S. P. Jackson, T. L. Blundell and L. Pellegrini 2001. Crystal structure of an Xrcc4-DNA ligase IV complex. *Nat Struct Biol* 8(12): 1015-1019.

Siddle, K. 2011. Signalling by insulin and IGF receptors: supporting acts and new players. *J Mol Endocrinol* 47(1): R1-10.

Sigaudy, S., A. Toutain, A. Moncla, C. Fredouille, B. Bourliere, S. Ayme and N. Philip 1998. Microcephalic osteodysplastic primordial dwarfism Taybi-Linder type: report of four cases and review of the literature. *Am J Med Genet* 80(1): 16-24.

Silva, E., Y. Tsatskis, L. Gardano, N. Tapon and H. McNeill 2006. The tumor-suppressor gene fat controls tissue growth upstream of expanded in the hippo signaling pathway. *Curr Biol* 16(21): 2081-2089.

Sir, J. H., A. R. Barr, A. K. Nicholas, O. P. Carvalho, M. Khurshid, A. Sossick, S. Reichelt, C. D'Santos*, et al.* 2011. A primary microcephaly protein complex forms a ring around parental centrioles. *Nat Genet* 43(11): 1147-1153.

Smigielski, E. M., K. Sirotkin, M. Ward and S. T. Sherry 2000. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res* 28(1): 352-355.

Smith, E. D., Y. Xu, B. N. Tomson, C. G. Leung, Y. Fujiwara, S. H. Orkin and J. D. Crispino 2004. More than blood, a novel gene required for mammalian postimplantation development. *Mol Cell Biol* 24(3): 1168-1173.

Smith, J., E. Riballo, B. Kysela, C. Baldeyron, K. Manolis, C. Masson, M. R. Lieber, D. Papadopoulo and P. Jeggo 2003. Impact of DNA ligase IV on the fidelity of end joining in human cells. *Nucleic Acids Res* 31(8): 2157-2167.

Smith, J. D., A. V. Hing, C. M. Clarke, N. M. Johnson, F. A. Perez, S. S. Park, J. A. Horst, B. Mecham, *et al.* 2014. Exome Sequencing Identifies a Recurrent De Novo ZSWIM6 Mutation Associated with Acromelic Frontonasal Dysostosis. *Am J Hum Genet* 95(2): 235-240.

Smith, L. M., J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, C. R. Connell, C. Heiner, S. B. Kent and L. E. Hood 1986. Fluorescence detection in automated DNA sequence analysis. *Nature* 321(6071): 674-679.

Sosnay, P. R., K. R. Siklosi, F. Van Goor, K. Kaniecki, H. Yu, N. Sharma, A. S. Ramalho, M. D. Amaral, *et al.* 2013. Defining the disease liability of variants in the cystic fibrosis transmembrane conductance regulator gene. *Nat Genet* 45(10): 1160-1167.

St Pierre, S. E., L. Ponting, R. Stefancsik and P. McQuilton 2014. FlyBase 102--advanced approaches to interrogating FlyBase. *Nucleic Acids Res* 42(Database issue): D780-788.

Stec, I., T. J. Wright, G. J. van Ommen, P. A. de Boer, A. van Haeringen, A. F. Moorman, M. R. Altherr and J. T. den Dunnen 1998. WHSC1, a 90 kb SET domain-containing gene, expressed in early development and homologous to a Drosophila dysmorphy gene maps in the Wolf-Hirschhorn syndrome critical region and is fused to IgH in t(4;14) multiple myeloma. *Hum Mol Genet* 7(7): 1071-1082.

Stenson, P. D., M. Mort, E. V. Ball, K. Shaw, A. Phillips and D. N. Cooper 2014. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet* 133(1): 1-9.

Stern, C. 1943. The Hardy-Weinberg Law. *Science* 97(2510): 137-138.

Stewart, D. R., A. Pemov, J. J. Johnston, J. C. Sapp, M. Yeager, J. He, J. F. Boland, L. Burdett, *et al.* 2014. Dubowitz syndrome is a complex comprised of multiple, genetically distinct and phenotypically overlapping disorders. *PLoS One* 9(6): e98686.

Strand, M., T. A. Prolla, R. M. Liskay and T. D. Petes 1993. Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. *Nature* 365(6443): 274-276.

Stratton, R. F., W. B. Dobyns, S. D. Airhart and D. H. Ledbetter 1984. New chromosomal syndrome: Miller-Dieker syndrome and monosomy 17p13. *Hum Genet* 67(2): 193-200.

Su, X. and J. Huang 2011. The Fanconi anemia pathway and DNA interstrand cross-link repair. *Protein Cell* 2(9): 704-711.

Suh, D., D. M. Wilson, 3rd and L. F. Povirk 1997. 3'-phosphodiesterase activity of human apurinic/apyrimidinic endonuclease at DNA double-strand break ends. *Nucleic Acids Res* 25(12): 2495-2500.

Suhasini, A. N. and R. M. Brosh, Jr. 2013. DNA helicases associated with genetic instability, cancer, and aging. *Adv Exp Med Biol* 767: 123-144.

Sun, X., S. G. Becker-Catania, H. H. Chun, M. J. Hwang, Y. Huo, Z. Wang, M. Mitui, O. Sanal*, et al.* 2002. Early diagnosis of ataxia-telangiectasia using radiosensitivity testing. *J Pediatr* 140(6): 724-731.

Taal, H. R., B. St Pourcain, E. Thiering, S. Das, D. O. Mook-Kanamori, N. M. Warrington, M. Kaakinen, E. Kreiner-Moller*, et al.* 2012a. Common variants at 12q15 and 12q24 are associated with infant head circumference. *Nat Genet* 44(5): 532-538.

Taal, H. R., B. St Pourcain, E. Thiering, S. Das, D. O. Mook-Kanamori, N. M. Warrington, M. Kaakinen, E. Kreiner-Moller*, et al.* 2012b. Common variants at 12q15 and 12q24 are associated with infant head circumference. *Nat Genet* 44(5): 532-538.

Taccioli, G. E., G. Rathbun, E. Oltz, T. Stamato, P. A. Jeggo and F. W. Alt 1993. Impairment of V(D)J recombination in double-strand break repair mutants. *Science* 260(5105): 207-210.

Takahashi, T., R. S. Nowakowski and V. S. Caviness, Jr. 1995. The cell cycle of the pseudostratified ventricular epithelium of the embryonic murine cerebral wall. *J Neurosci* 15(9): 6046-6057.

Takemoto, A., K. Kimura, J. Yanagisawa, S. Yokoyama and F. Hanaoka 2006. Negative regulation of condensin I by CK2-mediated phosphorylation. *EMBO J* 25(22): 5339-5348.

Takemoto, A., K. Maeshima, T. Ikehara, K. Yamaguchi, A. Murayama, S. Imamura, N. Imamoto, S. Yokoyama*, et al.* 2009. The chromosomal association of condensin II is regulated by a noncatalytic function of PP2A. *Nat Struct Mol Biol* 16(12): 1302-1308.

Takemoto, A., A. Murayama, M. Katano, T. Urano, K. Furukawa, S. Yokoyama, J. Yanagisawa, F. Hanaoka and K. Kimura 2007. Analysis of the role of Aurora B on the chromosomal targeting of condensin I. *Nucleic Acids Res* 35(7): 2403-2412.

Tan, R., Y. Wang, S. E. Kleinstein, Y. Liu, X. Zhu, H. Guo, Q. Jiang, A. S. Allen and M. Zhu 2014. An evaluation of copy number variation detection tools from whole-exome sequencing data. *Hum Mutat* 35(7): 899-907.

Tanaka, A., H. Tanizawa, S. Sriswasdi, O. Iwasaki, A. G. Chatterjee, D. W. Speicher, H. L. Levin, E. Noguchi and K. Noma 2012a. Epigenetic regulation of condensin-mediated genome organization during the cell cycle and upon DNA damage through histone H3 lysine 56 acetylation. *Mol Cell* 48(4): 532-546.

Tanaka, A., S. Weinel, N. Nagy, M. O'Driscoll, J. E. Lai-Cheong, C. L. Kulp-Shorten, A. Knable, G. Carpenter*, et al.* 2012b. Germline mutation in ATR in autosomal- dominant oropharyngeal cancer syndrome. *Am J Hum Genet* 90(3): 511-517.

Tapon, N., K. F. Harvey, D. W. Bell, D. C. Wahrer, T. A. Schiripo, D. Haber and I. K. Hariharan 2002. salvador Promotes both cell cycle exit and apoptosis in Drosophila and is mutated in human cancer cell lines. *Cell* 110(4): 467-478.

Tartaglia, M., L. A. Pennacchio, C. Zhao, K. K. Yadav, V. Fodale, A. Sarkozy, B. Pandit, K. Oishi*, et al.* 2007. Gain-of-function SOS1 mutations cause a distinctive form of Noonan syndrome. *Nat Genet* 39(1): 75-79.

Teer, J. K. and J. C. Mullikin 2010. Exome sequencing: the sweet spot before whole genomes. *Hum Mol Genet* 19(R2): R145-151.

Tennessen, J. A., A. W. Bigham, T. D. O'Connor, W. Fu, E. E. Kenny, S. Gravel, S. McGee, R. Do*, et al.* 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337(6090): 64-69.

Theisen, A. and L. G. Shaffer 2010. Disorders caused by chromosome abnormalities. *Appl Clin Genet* 3: 159-174.

Thevenon, J., P. Callier, H. Poquet, I. Bache, B. Menten, V. Malan, M. L. Cavaliere, J. P. Girod*, et al.* 2014. 3q27.3 microdeletional syndrome: a recognisable clinical entity associating dysmorphic features, marfanoid habitus, intellectual disability and psychosis with mood disorder. *J Med Genet* 51(1): 21-27.

Thomas, G. and M. N. Hall 1997. TOR signalling and control of cell growth. *Curr Opin Cell Biol* 9(6): 782-787.

Thomas, S. M., M. DeMarco, G. D'Arcangelo, S. Halegoua and J. S. Brugge 1992. Ras is essential for nerve growth factor- and phorbol ester-induced tyrosine phosphorylation of MAP kinases. *Cell* 68(6): 1031-1040.

Tibbetts, R. S., K. M. Brumbaugh, J. M. Williams, J. N. Sarkaria, W. A. Cliby, S. Y. Shieh, Y. Taya, C. Prives and R. T. Abraham 1999. A role for ATR in the DNA damage-induced phosphorylation of p53. *Genes Dev* 13(2): 152-157.

Tibelius, A., J. Marhold, H. Zentgraf, C. E. Heilig, H. Neitzel, B. Ducommun, A. Rauch, A. D. Ho*, et al.* 2009. Microcephalin and pericentrin regulate mitotic entry via centrosome-associated Chk1. *J Cell Biol* 185(7): 1149-1157.

Tilgner, K., I. Neganova, I. Moreno-Gimeno, J. Y. Al-Aama, D. Burks, S. Yung, C. Singhapol, G. Saretzki*, et al.* 2013. A human iPSC model of Ligase IV deficiency reveals an important role for NHEJ-mediated-DSB repair in the survival and genomic stability of induced pluripotent stem cells and emerging haematopoietic progenitors. *Cell Death Differ* 20(8): 1089-1100.

Timson, D. J., M. R. Singleton and D. B. Wigley 2000. DNA ligases in the repair and replication of DNA. *Mutat Res* 460(3-4): 301-318.

Toita, N., N. Hatano, S. Ono, M. Yamada, R. Kobayashi, I. Kobayashi, N. Kawamura, M. Okano*, et al.* 2007. Epstein-Barr virus-associated B-cell lymphoma in a patient with DNA ligase IV (LIG4) syndrome. *Am J Med Genet A* 143(7): 742-745.

Tonegawa, S. 1983. Somatic generation of antibody diversity. *Nature* 302(5909): 575-581.

Treangen, T. J. and S. L. Salzberg 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* 13(1): 36-46.

Trimborn, M., S. M. Bell, C. Felix, Y. Rashid, H. Jafri, P. D. Griffiths, L. M. Neumann, A. Krebs*, et al.* 2004. Mutations in microcephalin cause aberrant regulation of chromosome condensation. *Am J Hum Genet* 75(2): 261-266.

Trimborn, M., D. Schindler, H. Neitzel and T. Hirano 2006. Misregulated chromosome condensation in MCPH1 primary microcephaly is mediated by condensin II. *Cell Cycle* 5(3): 322-326.

Trotter, T. L., J. G. Hall and G. American Academy of Pediatrics Committee on 2005. Health supervision for children with achondroplasia. *Pediatrics* 116(3): 771-783.

Tuorto, F., R. Liebers, T. Musch, M. Schaefer, S. Hofmann, S. Kellner, M. Frye, M. Helm, *et al.* 2012. RNA cytosine methylation by Dnmt2 and NSun2 promotes tRNA stability and protein synthesis. *Nat Struct Mol Biol* 19(9): 900-905.

Tzur, A., R. Kafri, V. S. LeBleu, G. Lahav and M. W. Kirschner 2009. Cell growth and size homeostasis in proliferating animal cells. *Science* 325(5937): 167-171.

Unal, S., K. Cerosaletti, D. Uckan-Cetinkaya, M. Cetin and F. Gumruk 2009. A novel mutation in a family with DNA ligase IV deficiency syndrome. *Pediatr Blood Cancer* 53(3): 482-484.

Untergasser, A., I. Cutcutache, T. Koressaar, J. Ye, B. C. Faircloth, M. Remm and S. G. Rozen 2012. Primer3--new capabilities and interfaces. *Nucleic Acids Res* 40(15): e115.

Urban, A. E., J. O. Korbel, R. Selzer, T. Richmond, A. Hacker, G. V. Popescu, J. F. Cubells, R. Green, *et al.* 2006. High-resolution mapping of DNA copy alterations in human chromosome 22 using high-density tiling oligonucleotide arrays. *Proc Natl Acad Sci U S A* 103(12): 4534-4539.

Van Buggenhout, G. and J. P. Fryns 2006. Lujan-Fryns syndrome (mental retardation, X-linked, marfanoid habitus). *Orphanet J Rare Dis* 1: 26.

van der Burg, M., H. Ijspeert, N. S. Verkaik, T. Turul, W. W. Wiegant, K. Morotomi-Yano, P. O. Mari, I. Tezcan, *et al.* 2009. A DNA-PKcs mutation in a radiosensitive T-B- SCID patient inhibits Artemis activation and nonhomologous end-joining. *J Clin Invest* 119(1): 91-98.

van der Burg, M., L. R. van Veelen, N. S. Verkaik, W. W. Wiegant, N. G. Hartwig, B. H. Barendregt, L. Brugmans, A. Raams, *et al.* 2006. A new type of radiosensitive T-B-NK+ severe combined immunodeficiency caused by a LIG4 mutation. *J Clin Invest* 116(1): 137-145.

van der Burgt, I. 2007. Noonan syndrome. *Orphanet J Rare Dis* 2: 4.

van Deursen, F., S. Sengupta, G. De Piccoli, A. Sanchez-Diaz and K. Labib 2012. Mcm10 associates with the loaded DNA helicase at replication origins and defines a novel step in its activation. *EMBO J* 31(9): 2195-2206.

van Dijk, E. L., Y. Jaszczyszyn and C. Thermes 2014. Library preparation methods for next-generation sequencing: tone down the bias. *Exp Cell Res* 322(1): 12-20.

van El, C. G., M. C. Cornel, P. Borry, R. J. Hastings, F. Fellmann, S. V. Hodgson, H. C. Howard, A. Cambon-Thomsen, *et al.* 2013. Whole-genome sequencing in health care: recommendations of the European Society of Human Genetics. *Eur J Hum Genet* 21(6): 580-584.

Varelas, X., B. W. Miller, R. Sopko, S. Song, A. Gregorieff, F. A. Fellouse, R. Sakuma, T. Pawson, *et al.* 2010. The Hippo pathway regulates Wnt/beta-catenin signaling. *Dev Cell* 18(4): 579-591.

Verloes, A., S. Drunat, P. Gressens and S. Passemard (1993). Primary Autosomal Recessive Microcephalies and Seckel Syndrome Spectrum Disorders. GeneReviews(R). R. A. Pagon, M. P. Adam, H. H. Ardinger et al. Seattle (WA).

Vissers, L. E. and P. Stankiewicz 2012. Microdeletion and microduplication syndromes. *Methods Mol Biol* 838: 29-75.

Vissers, L. E., J. A. Veltman, A. G. van Kessel and H. G. Brunner 2005. Identification of disease genes by whole genome CGH arrays. *Hum Mol Genet* 14 Spec No. 2: R215-223.

Vulto-van Silfhout, A. T., B. B. de Vries, B. W. van Bon, A. Hoischen, M. Ruiterkamp-Versteeg, C. Gilissen, F. Gao, M. van Zwam, *et al.* 2013. Mutations in MED12 cause X-linked Ohdo syndrome. *Am J Hum Genet* 92(3): 401-406.

Wakeling, E. L., S. A. Amero, M. Alders, J. Bliek, E. Forsythe, S. Kumar, D. H. Lim, F. MacDonald, *et al.* 2010. Epigenotype-phenotype correlations in Silver-Russell syndrome. *J Med Genet* 47(11): 760-768.

Walenkamp, M. J., M. Losekoot and J. M. Wit 2013. Molecular IGF-1 and IGF-1 receptor defects: from genetics to clinical management. *Endocr Dev* 24: 128-137.

Walker, J. R., R. A. Corpina and J. Goldberg 2001. Structure of the Ku heterodimer bound to DNA and its implications for double-strand break repair. *Nature* 412(6847): 607-614.

Walne, A. J., T. Vulliamy, M. Kirwan, V. Plagnol and I. Dokal 2013. Constitutional mutations in RTEL1 cause severe dyskeratosis congenita. *Am J Hum Genet* 92(3): 448-453.

Waltes, R., R. Kalb, M. Gatei, A. W. Kijas, M. Stumm, A. Sobeck, B. Wieland, R. Varon, *et al.* 2009. Human RAD50 deficiency in a Nijmegen breakage syndrome-like disorder. *Am J Hum Genet* 84(5): 605-616.

Wang, J. C. 2002. Cellular roles of DNA topoisomerases: a molecular perspective. *Nat Rev Mol Cell Biol* 3(6): 430-440.

Wang, S. L., C. J. Hawkins, S. J. Yoo, H. A. Muller and B. A. Hay 1999. The Drosophila caspase inhibitor DIAP1 is essential for cell survival and is negatively regulated by HID. *Cell* 98(4): 453-463.

Wang, X., X. Chen, Y. Li, H. Liu, S. Li, R. R. Zhang and B. Zheng 2012. Fluorescence in situ hybridization (FISH) signal analysis using automated generated projection images. *Anal Cell Pathol (Amst)* 35(5-6): 395-405.

Wang, X., J. W. Tsai, J. H. Imai, W. N. Lian, R. B. Vallee and S. H. Shi 2009. Asymmetric centrosome inheritance maintains neural progenitors in the neocortex. *Nature* 461(7266): 947-955.

Ward, J. F. 1988. DNA damage produced by ionizing radiation in mammalian cells: identities, mechanisms of formation, and reparability. *Prog Nucleic Acid Res Mol Biol* 35: 95-125.

Weedon, M. N., H. Lango, C. M. Lindgren, C. Wallace, D. M. Evans, M. Mangino, R. M. Freathy, J. R. Perry, *et al.* 2008. Genome-wide association analysis identifies 20 loci that influence adult height. *Nat Genet* 40(5): 575-583.

Weedon, M. N., G. Lettre, R. M. Freathy, C. M. Lindgren, B. F. Voight, J. R. Perry, K. S. Elliott, R. Hackett, *et al.* 2007. A common variant of HMGA2 is associated with adult and childhood height in the general population. *Nat Genet* 39(10): 1245-1250.

Weemaes, C. 2000. Nijmegen breakage syndrome. The International Nijmegen Breakage Syndrome Study Group. *Arch Dis Child* 82(5): 400-406.

Weinfeld, M., R. S. Mani, I. Abdou, R. D. Aceytuno and J. N. Glover 2011. Tidying up loose ends: the role of polynucleotide kinase/phosphatase in DNA strand break repair. *Trends Biochem Sci* 36(5): 262-271.

Wheeler, D. A., M. Srinivasan, M. Egholm, Y. Shen, L. Chen, A. McGuire, W. He, Y. J. Chen, *et al.* 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452(7189): 872-876.

Whitehouse, C. J., R. M. Taylor, A. Thistlethwaite, H. Zhang, F. Karimi-Busheri, D. D. Lasko, M. Weinfeld and K. W. Caldecott 2001. XRCC1 stimulates human polynucleotide kinase activity at damaged DNA termini and accelerates DNA single-strand break repair. *Cell* 104(1): 107-117.

Willems, M., D. Genevieve, G. Borck, C. Baumann, G. Baujat, E. Bieth, P. Edery, C. Farra, *et al.* 2010. Molecular analysis of pericentrin gene (PCNT) in a series of 24 Seckel/microcephalic osteodysplastic primordial dwarfism type II (MOPD II) families. *J Med Genet* 47(12): 797-802.

Wilson, M. G., J. W. Towner, G. S. Coffin, A. J. Ebbin, E. Siris and P. Brager 1981. Genetic and clinical studies in 13 patients with the Wolf-Hirschhorn syndrome [del(4p)]. *Hum Genet* 59(4): 297-307.

Winston, J. T., S. R. Coats, Y. Z. Wang and W. J. Pledger 1996. Regulation of the cell cycle machinery by oncogenic ras. *Oncogene* 12(1): 127-134.

Wohlschlegel, J. A., B. T. Dwyer, S. K. Dhar, C. Cvetic, J. C. Walter and A. Dutta 2000. Inhibition of eukaryotic DNA replication by geminin binding to Cdt1. *Science* 290(5500): 2309-2312.

Wong, M. M., L. K. Cox and J. C. Chrivia 2007. The chromatin remodeling protein, SRCAP, is critical for deposition of the histone variant H2A.Z at promoters. *J Biol Chem* 282(36): 26132-26139.

Woodbine, L., A. R. Gennery and P. A. Jeggo 2014. The clinical impact of deficiency in DNA non-homologous end-joining. *DNA Repair (Amst)* 16: 84-96.

Woodbine, L., J. A. Neal, N. K. Sasi, M. Shimada, K. Deem, H. Coleman, W. B. Dobyns, T. Ogi, *et al.* 2013. PRKDC mutations in a SCID patient with profound neurological abnormalities. *J Clin Invest* 123(7): 2969-2980.

Woods, C. G., J. Bond and W. Enard 2005. Autosomal recessive primary microcephaly (MCPH): a review of clinical, molecular, and evolutionary findings. *Am J Hum Genet* 76(5): 717-728.

Woods, K. A., C. Camacho-Hubner, R. N. Bergman, D. Barter, A. J. Clark and M. O. Savage 2000. Effects of insulin-like growth factor I (IGF-I) therapy on body composition and insulin resistance in IGF-I gene deletion. *J Clin Endocrinol Metab* 85(4): 1407-1411.

Woodward, A. M., T. Gohler, M. G. Luciani, M. Oehlmann, X. Ge, A. Gartner, D. A. Jackson and J. J. Blow 2006. Excess Mcm2-7 license dormant origins of replication that can be used under conditions of replicative stress. *J Cell Biol* 173(5): 673-683.

Worrall, C., D. Nedelcu, J. Serly, N. Suleymanova, I. Oprea, A. Girnita and L. Girnita 2013. Novel mechanisms of regulation of IGF-1R action: functional and therapeutic implications. *Pediatr Endocrinol Rev* 10(4): 473-484.

Wright, T. J., J. L. Costa, C. Naranjo, P. Francis-West and M. R. Altherr 1999. Comparative analysis of a novel gene from the Wolf-Hirschhorn/Pitt-Rogers-Danks syndrome critical region. *Genomics* 59(2): 203-212.

Wright, T. J., D. O. Ricke, K. Denison, S. Abmayr, P. D. Cotter, K. Hirschhorn, M. Keinanen, D. McDonald-McGinn, *et al.* 1997. A transcript map of the newly defined 165 kb Wolf-Hirschhorn syndrome critical region. *Hum Mol Genet* 6(2): 317-324.

Wu, J. and R. Jiang 2013. Prediction of deleterious nonsynonymous single-nucleotide polymorphism for human diseases. *ScientificWorldJournal* 2013: 675851.

Wu, Q., T. Ochi, D. Matak-Vinkovic, C. V. Robinson, D. Y. Chirgadze and T. L. Blundell 2011. Non-homologous end-joining partners in a helical dance: structural studies of XLF-XRCC4 interactions. *Biochem Soc Trans* 39(5): 1387-1392, suppl 1382 p following 1392.

Wu, S., J. Huang, J. Dong and D. Pan 2003. hippo encodes a Ste-20 family protein kinase that restricts cell proliferation and promotes apoptosis in conjunction with salvador and warts. *Cell* 114(4): 445-456.

Xu, T., W. Wang, S. Zhang, R. A. Stewart and W. Yu 1995. Identifying tumor suppressors in genetic mosaics: the Drosophila lats gene encodes a putative protein kinase. *Development* 121(4): 1053-1063.

Xu, Z., A. C. Maroney, P. Dobrzanski, N. V. Kukekov and L. A. Greene 2001. The MLK family mediates c-Jun N-terminal kinase activation in neuronal apoptosis. *Mol Cell Biol* 21(14): 4713-4724.

Yakar, S., J. L. Liu, B. Stannard, A. Butler, D. Accili, B. Sauer and D. LeRoith 1999. Normal growth and development in the absence of hepatic insulin-like growth factor I. *Proc Natl Acad Sci U S A* 96(13): 7324-7329.

Yamada, M., S. Matsuura, M. Tsukahara, K. Ebe, M. Ohtsu, H. Furuta, I. Kobayashi, N. Kawamura, *et al.* 2001. Combined immunodeficiency, chromosomal instability, and postnatal growth deficiency in a Japanese girl. *Am J Med Genet* 100(1): 9-12.

Yamaguchi, T., K. Hosomichi, A. Narita, T. Shirota, Y. Tomoyasu, K. Maki and I. Inoue 2011. Exome resequencing combined with linkage analysis identifies novel PTH1R variants in primary failure of tooth eruption in Japanese. *J Bone Miner Res* 26(7): 1655-1661.

Yamamoto, T., M. Ebisuya, F. Ashida, K. Okamoto, S. Yonehara and E. Nishida 2006. Continuous ERK activation downregulates antiproliferative genes throughout G1 phase to allow cell-cycle progression. *Curr Biol* 16(12): 1171-1182.

Yamashita, Y. M., A. P. Mahowald, J. R. Perlin and M. T. Fuller 2007. Asymmetric inheritance of mother versus daughter centrosome in stem cell division. *Science* 315(5811): 518-521.

Yannone, S. M., I. S. Khan, R. Z. Zhou, T. Zhou, K. Valerie and L. F. Povirk 2008. Coordinate 5' and 3' endonucleolytic trimming of terminally blocked blunt DNA double-strand break ends by Artemis nuclease and DNA-dependent protein kinase. *Nucleic Acids Res* 36(10): 3354-3365.

Yao, X., A. Abrieu, Y. Zheng, K. F. Sullivan and D. W. Cleveland 2000. CENP-E forms a link between attachment of spindle microtubules to kinetochores and the mitotic checkpoint. *Nat Cell Biol* 2(8): 484-491.

Yates, K. E., G. A. Korbel, M. Shtutman, I. B. Roninson and D. DiMaio 2008. Repression of the SUMO-specific protease Senp1 induces p53-dependent premature senescence in normal human fibroblasts. *Aging Cell* 7(5): 609-621.

Yeong, F. M., H. Hombauer, K. S. Wendt, T. Hirota, I. Mudrak, K. Mechtler, T. Loregger, A. Marchler-Bauer, *et al.* 2003. Identification of a subunit of a novel Kleisin-beta/SMC complex as a potential substrate of protein phosphatase 2A. *Curr Biol* 13(23): 2058-2064.

Yoshimura, H., S. Iwasaki, S. Y. Nishio, K. Kumakawa, T. Tono, Y. Kobayashi, H. Sato, K. Nagai, *et al.* 2014. Massively parallel DNA sequencing facilitates diagnosis of patients with Usher syndrome type 1. *PLoS One* 9(3): e90688.

You, Z., C. Chahwan, J. Bailis, T. Hunter and P. Russell 2005. ATM activation and its recruitment to damaged DNA require binding to the C terminus of Nbs1. *Mol Cell Biol* 25(13): 5363-5379.

Yu, F. X., B. Zhao, N. Panupinthu, J. L. Jewell, I. Lian, L. H. Wang, J. Zhao, H. Yuan, *et al.* 2012. Regulation of the Hippo-YAP pathway by G-protein-coupled receptor signaling. *Cell* 150(4): 780-791.

Yu, T. W., G. H. Mochida, D. J. Tischfield, S. K. Sgaier, L. Flores-Sarnat, C. M. Sergi, M. Topcu, M. T. McDonald, *et al.* 2010. Mutations in WDR62, encoding a centrosome-associated protein, cause microcephaly with simplified gyri and abnormal cortical architecture. *Nat Genet* 42(11): 1015-1020.

Yuan, B., R. Latek, M. Hossbach, T. Tuschl and F. Lewitter 2004. siRNA Selection Server: an automated siRNA oligonucleotide prediction server. *Nucleic Acids Res* 32(Web Server issue): W130-134.

Yue, J., H. Lu, S. Lan, J. Liu, M. N. Stein, B. G. Haffty and Z. Shen 2013. Identification of the DNA repair defects in a case of dubowitz syndrome. *PLoS One* 8(1): e54389.

Zarate, Y. A., C. Bell and B. Schaefer 2013. Description of another case of 3q26.33-3q27.2 microdeletion supports a recognizable phenotype. *Eur J Med Genet* 56(11): 624-625.

Zegerman, P. and J. F. Diffley 2007. Phosphorylation of Sld2 and Sld3 by cyclin-dependent kinases promotes DNA replication in budding yeast. *Nature* 445(7125): 281-285.

ZFIN Historical Data. Phenotype Annotation (1994-2006) (2006).

Zhang, J. 2012. Genetic redundancies and their evolutionary maintenance. *Adv Exp Med Biol* 751: 279-300.

Zhang, S. Q., T. Jiang, M. Li, X. Zhang, Y. Q. Ren, S. C. Wei, L. D. Sun, H. Cheng, *et al.* 2012. Exome sequencing identifies MVK mutations in disseminated superficial actinic porokeratosis. *Nat Genet* 44(10): 1156-1160.

Zhao, B., L. Li and K. L. Guan 2010. Hippo signaling at a glance. *J Cell Sci* 123(Pt 23): 4001-4006.

Zhao, B., X. Ye, J. Yu, L. Li, W. Li, S. Li, J. Yu, J. D. Lin, *et al.* 2008. TEAD mediates YAP-dependent gene induction and growth control. *Genes Dev* 22(14): 1962-1971.

Zhao, M., Q. Wang, Q. Wang, P. Jia and Z. Zhao 2013. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics* 14 Suppl 11: S1.

Zhou, X., K. F. Benson, H. R. Ashar and K. Chada 1995. Mutation responsible for the mouse pygmy phenotype in the developmentally regulated factor HMGI-C. *Nature* 376(6543): 771-774.

Zickler, D. and N. Kleckner 1999. Meiotic chromosomes: integrating structure and function. *Annu Rev Genet* 33: 603-754.

Zimmerman, W. C., J. Sillibourne, J. Rosa and S. J. Doxsey 2004. Mitosis-specific anchoring of gamma tubulin complexes by pericentrin controls spindle organization and mitotic entry. *Mol Biol Cell* 15(8): 3642-3657.

Zollino, M., R. Lecce, R. Fischetto, M. Murdolo, F. Faravelli, A. Selicorni, C. Butte, L. Memo, *et al.* 2003. Mapping the Wolf-Hirschhorn syndrome phenotype outside the currently accepted WHS critical region and defining a new critical region, WHSCR-2. *Am J Hum Genet* 72(3): 590-597.

Zolner, A. E., I. Abdou, R. Ye, R. S. Mani, M. Fanta, Y. Yu, P. Douglas, N. Tahbaz, *et al.* 2011. Phosphorylation of polynucleotide kinase/ phosphatase by DNA-dependent protein kinase and ataxia-telangiectasia mutated regulates its association with sites of DNA damage. *Nucleic Acids Res* 39(21): 9224-9237.

Zou, L. and S. J. Elledge 2003. Sensing DNA damage through ATRIP recognition of RPA-ssDNA complexes. *Science* 300(5625): 1542-1548.