About the impact of altered RAS-MAPK and PI3K-AKT signalling in human developmental disorders

Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium

(Dr. rer. nat.)

genehmigt durch die Fakultät für Naturwissenschaften

der Otto-von-Guericke-Universität Magdeburg

von M.Sc., Sangamitra Boppudi

geb.am 16.Juli 1986 Visakhapatnam, Indien

Gutachter: Prof. Dr. med. Martin Zenker

Prof. Dr. rer. nat. Frank Kaiser

eingereichte am: 21-12-2017

verteidigt am: 25-09-2018

Table of Contents

List of Figures

List of Tables

Zusammenfassung

Der RAS/MAPK und der PI3K/AKT/mTOR Signalweg stellen komplexe und miteinander vernetzte zelluläre Signalwege dar, die viele biologische Prozesse regulieren. Mutationen, welche die Regulierung dieser Signalwege stören, sind verantwortlich für verschiedene Erkrankungen, die von Tumoren, entstanden durch erworbene somatische Mutationen, bis hin zu einem breiten Spektrum von seltenen Entwicklungsstörungen mit angeborenen Veränderungen reichen. RASopathien ist der Überbegriff für eine Gruppe von Erkrankungen, welche durch Mutationen in Genen, die für Komponenten oder Modulatoren des RAS/MAPK Signalweges kodieren, hervorgerufen werden. Betroffene Personen zeigen ein typisches Muster an physischen Anomalien sowie kognitive Defizite unterschiedlichen Ausmaßes. Ferner sind Mutationen in Komponenten des PI3K/AKT/mTOR Signalweges insbesondere mit angeborenen Erkrankungen assoziiert, bei welchen Wachstum von betroffenen Körperregionen oder Organen, sowie die Entwicklung des Gehirns beeinträchtigt sein können

Die massive Parallelsequenzierung (auch bekannt unter dem Begriff „Next Generation Sequencing“, NGS) wurde in den vergangenen zehn Jahren rapide weiterentwickelt und bietet neue Möglichkeiten für die Identifikation von genetischen Veränderungen, die Krankheiten verursachen können. Der hohe Durchsatz und die Sequenziertiefe dieser neuen Methoden machen die Untersuchung mehrerer Gene (bis hin zu Exomen und Genomen) in einem Experiment sowie die Detektion von sehr geringgradigen Mosaiken (wie in Tumoren) möglich.

Das Ziel dieser Arbeit war es, die Auswirkungen von genetischen Veränderungen, die zur Dysregulation von RAS/MAPK und PI3K/AKT/mTOR Signalwegen führen und menschliche Entwicklungsstörungen verursachen können, mit Hilfe neuer NGS Technologien weiter zu untersuchen.

Ein gemeinsames Merkmal der RASopathien ist die Beeinträchtigung der kognitiven Fähigkeiten. Unter anderem ist der RAS/MAPK Signalweg in neuronalen Zellen in die Regulation synaptischer Plastizität involviert. Ein Ziel dieser Arbeit war es, die Bedeutung genetischer Veränderungen, die direkt oder indirekt das Signalnetzwerk des RAS/MAPK Signalwegs betreffen, bei Patienten mit Intelligenzminderung (ID) zu untersuchen, welche nicht die äußerlich erkennbaren Merkmale der RASopathien zeigten. Zur Identifikation direkter und indirekter Interaktionspartner von Molekülen aus dem RAS/MAPK Signalweg wurde in silico ein Interaktionsnetzwerk um die RAS Moleküle konstruiert, anhand festgelegter Kriterien weiter differenziert, und somit eine Liste von 329 Kandidatengenen erstellt. In diesem Teilprojekt wurden mittels gezielter Re-Sequenzierung des ausgewählten Multigen-Panels eine Kohorte von 166 Patienten mit ungeklärter oder unspezifischer ID sowie eine Kontrollgruppe von 120 Patienten ohne ID per NGS untersucht. Die so gefundenen Sequenzvarianten wurden anhand ihrer Art, ihres Vorkommens in Datenbanken und der in silico-Prädiktion ihrer Auswirkungen auf das entstehende Genprodukt gefiltert.

Durchschnittlich wurden vier sehr seltene oder unbekannte Varianten pro Individuum (0-12 Varianten) in dem untersuchten Multigen-Panel gefunden. Die besten Kandidaten wurden mit konventioneller Sanger-Sequenzierung bestätigt und die Segregation in den Familien überprüft. In zwei der 166 Studienpatienten wurden wahrscheinlich krankheitsverursachende Veränderungen in Genen gefunden, welche nun als kausale Gene für ID etabliert sind. Die Ergebnisse zeigen, dass bei einem so heterogenen Krankheitsbild wie unspezifische/nicht-syndromale ID monogene Defekte mit Bezug auf den RAS/MAPK Signalweg einen kleinen aber durchaus relevanten Teil ausmachen. Darüber hinaus wurden in 47 weiteren Fällen potenziell pathogene Varianten in bereits etablierten Genen für ID sowie in neuen Genen mit Bezug auf den RAS/MAPK Signalweg gefunden, deren Pathogenität aber nicht definitiv bewiesen werden konnte. Um diese Ergebnisse zu verifizieren und ist eine Bestätigung in größeren Kohorten erforderlich.

Im zweiten Teilprojekt lag der Fokus auf einer Krankheitsgruppe, die durch somatische Mutationen in Genen des RAS/MAPK und des funktionell verbundenen PI3K/AKT/mTOR Signalwegs verursacht werden. In dieser Studie wurden spezifische KRAS Mutationen im Mosaik als krankheitsverursachende Veränderungen beim okuloektodermalen Syndrom (OES) und bei der enzephalo-kranio-kutanen Lipomatose (ECCL) gefunden. Dieser Zusammenhang bestätigte die kurz zuvor anhand von zwei Fällen postulierte Zugehörigkeit der Krankheitsgruppe zu den sogenannten “Mosaik-RASopathien”. Bei Patienten mit phänotypischen Merkmalen des „congenital lipomatous overgrowth, vascular malformations, and epidermal nevi“ (CLOVES) wurden PIK3CA-Hotspot-Mutationen in DNA aus verschiedenen Geweben untersucht. Mittels Sanger-Sequenzierung waren die Detektionsrate und die Quantifizierung der mutierten Allele auf minimal 10-15% begrenzt. Deswegen wurden zusätzlich Fragment-Analysen und NGS verwendet, mit welchen die Sensitivität auf einen Mosaikanteil von minimal <1-3% gesteigert werden konnte. Mittels NGS lag die Detektionsrate mutierter Allele sogar in Blutproben bei 1% oder darunter. Unsere Daten bestätigen, dass Material aus dem betroffenen Gewebe für die Detektion der zugrunde liegenden Mutation in PROS (PIK3CA-related overgrowth spectrum) notwendig ist, während Blutproben in den meisten Fällen eine ungeeignete Quelle darstellen. Verbesserte Detektionsmethoden werden ebenfalls für andere Gewebe mit niedriggradigem Mosaikanteil benötigt.

Zusammenfassend liefert diese Studie einen zusätzlichen Beitrag zum Verständnis der Bedeutuung von Mutationen in den RAS/MAPK und PI3K/AKT/mTOR Signalwegen bei Störungen der körperlichen und geistigen Entwicklung des Menschen. Sie belegt den Nutzen NGS-basierter Verfahren bei der Hochdurchsatz-Sequenzierung mehrerer Gene in größeren Patientenkohorten und bei der sensitiven Detektion von Mosaiken in DNA aus verschiedenen Geweben mittels sehr tiefer Sequenzierung.

1. Abstract

The RAS/MAPK and PI3K/AKT/mTOR pathways represent complex interconnected cellular signalling pathways that regulate many biological processes. Mutations leading to disturbance of these signalling pathways have been shown to be involved in various human diseases, ranging from tumors with acquired somatic mutations to a broad spectrum of rare developmental disorders with congenital alterations. RASopathies is the new umbrella term for a disease group caused by mutations in genes encoding various components and modulators of the RAS/MAPK signalling pathway leading to dysregulation of signal flow. Affected individuals display a common pattern of physical anomalies along with cognitive deficits of variable severity. Mutations in components of the PI3K/AKT/mTOR pathway, on the other hand, are particularly associated with congenital disorders showing altered growth of body parts or organs, again including the brain and affecting its function. During the last 10 years the technology of massive parallel sequencing (also known as next-generation sequencing, NGS) has rapidly evolved and provided new opportunities for identification of genetic alterations underlying human disorders. This thesis aimed at further investigating the impact of genetic changes leading to altered RAS/MAPK and PI3K/AKT/mTOR signalling in human developmental disorders with the help of novel NGS technologies.

Considering the impairment of cognitive functions as a common feature of RASopathies and the known importance of the RAS/MAPK signalling pathway is the regulation of synaptic plasticity, the aim of one subproject of this thesis was to further investigate the importance of genetic alterations directly or indirectly affecting the signalling network around RAS/MAPK for intellectual disability (ID) in humans, outside the syndromic context of RASopathies. For this purpose, an interaction network around the RAS molecules was constructed in silico and further prioritized according to various criteria, generating a list of 329 candidate genes for the investigations in humans. In the current project, a targeted resequencing approach was used and examined by means of NGS in a cohort of 166 patients with unexplained / unspecific ID as well as a control group of 120 subjects without ID. Thus identified sequence variants were filtered by occurrence in databases and the silico predicted effects on the gene product. On average, in the examined RAS pathway genes, ~4 very rare or unknown variants were identified per individual (Range: 0-12). The best candidates were validated by conventional sequencing and segregation studies in the family. In the 166 study patients, the most likely causative genetic change was found in two individuals in genes, which are now established as ID genes. The results show that even in a disease as heterogeneous as ID, monogenic defects of molecules with respect to the RAS signalling pathway make a small but relevant part. In order to identify further mutations, targeted examinations in larger cohorts are to be carried out for the best new candidate genes and/or co-operation with other groups that have operated exome sequencing in ID cohorts.

A second subproject was focused on a group of diseases caused by somatic mosaic mutations in genes of the RAS signalling pathway and the functionally linked PI3K/AKT/mTOR signalling pathway. The current study identified specific KRAS mutations as the cause of the oculoectodermal syndrome (OES) and the encephalo-cranio-cutaneous lipomatosis (ECCL), thus confirming the association of these related diseases to the group of "mosaic RASopathies". In patients presenting with PIK3CA-related overgrowth spectrum (PROS), hot spot mutations of PIK3CA were identified by studying DNA from various tissues. By Sanger sequencing, the detection levels and quantification of mutant alleles were limited to 10-15%. So, fragment analysis and NGS methods were further applied which increased the mutant allele detection to <1-3%. With NGS method, mutant allele ratios for blood samples could also be detected and was 1% or less. Our data confirm that material from affected tissue is essential for detecting the underlying mutation in PROS whereas blood DNA would be a secondary source in most cases. Improved detection methods may also be required for other tissues with low level somatic mosaicism.

2. Introduction

2.1 RAS signalling pathway

The RAS proteins or RAS GTPases form a superfamily of small GTP binding proteins (monomeric G proteins of molecular mass 20-40 kDa) which participate in signal pathways crucial for a wide variety of biological functions (Van Aelst L & D'Souza-Schorey C, 1997). The RAS superfamily of small GTPases are grouped into at least five major sub families based on their structure, sequence and function: RAS (RAS sarcoma), Rho (RAS homologous), Rab (RAS-like proteins in brain), Ran (The RAS-like nuclear) and Sar1/Arf (ADP ribosylation factor) (Wennerberg K, Rossman KL, Der CJ, 2005). The classical RAS proteins of the RAS subfamily include HRAS, KRAS, NRAS, RRAS, MRAS, RIT1, and RIT2. Other RAS family proteins, including Rap's (RAS-related proteins), Ral (RAS-like) and Rheb (RAS homolog enriched in brain) proteins also regulate signalling networks (Rojas et al., 2012).

Central to their activity is the ratio of their GTP/GDP bound forms subjected to complex regulation, illustrated in figure 2.1. The main known regulators of this ratio are

- Guanine nucleotide exchange factors (GEFs, ex: SOS, RASGRF's and RASGRP's) which promote formation of the active, GTP-bound form,
- GTPase activating proteins (GAPs, ex: p120 and neurofibromin), which accelerates the intrinsic GTPase activity to promote formation of the inactive GDP-bound form,
- Guanine nucleotide dissociation inhibitors (GDIs) - The Rho and Rab GTPases are regulated by these proteins which mask the prenyl modification and promote cytosolic sequestration of these GTPases

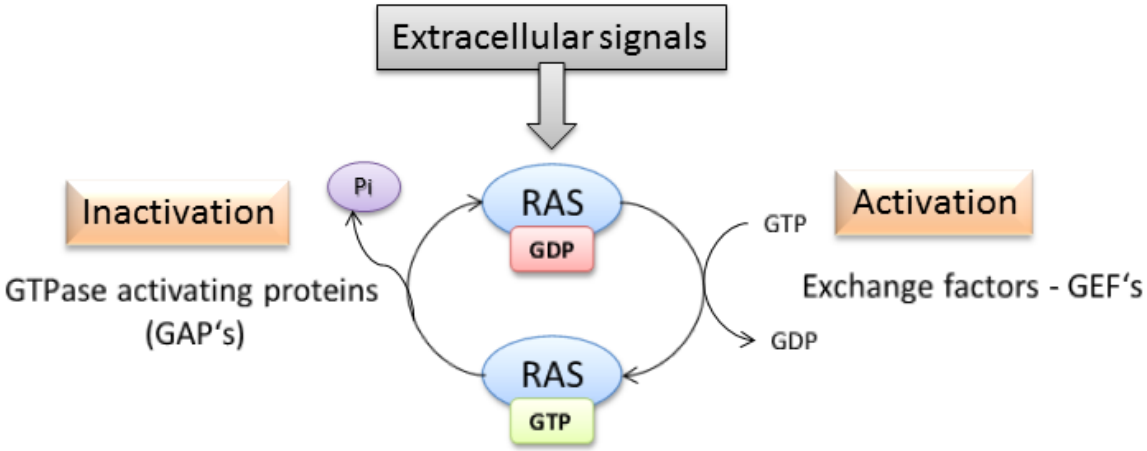


Figure 2.1: Regulation of RAS proteins. RAS acts as signal switch between active and inactive states, by converting GDP into GTP. In the active state, GDP is exchanged by GTP which is facilitated by GEF (Guanine Nucleotide Exchange Factor). In the inactive state, GTP is intrinsically converted to GDP by GAP. RAS can also be activated by the inhibition of the GAPs.

Active GTP-bound RAS interacts with a wide range of targets (effectors) including RAF kinases (Rapidly Accelerated Fibrosarcoma), phosphatidylinositol 3-kinase (PI-3 kinase), RalGDS, phospholipase C epsilon, p120GAP, and the Nore-MST1 complex and stimulates downstream signalling pathways (Figure 2.2). RAS-GTP induces a wide variety of cellular processes, such as transcription, translation, cell-cycle progression, apoptosis or cell survival, through direct interaction with various effectors. GAP proteins also interact with RAS-GTP and might also act as effectors (Aoki Y et al., 2008).

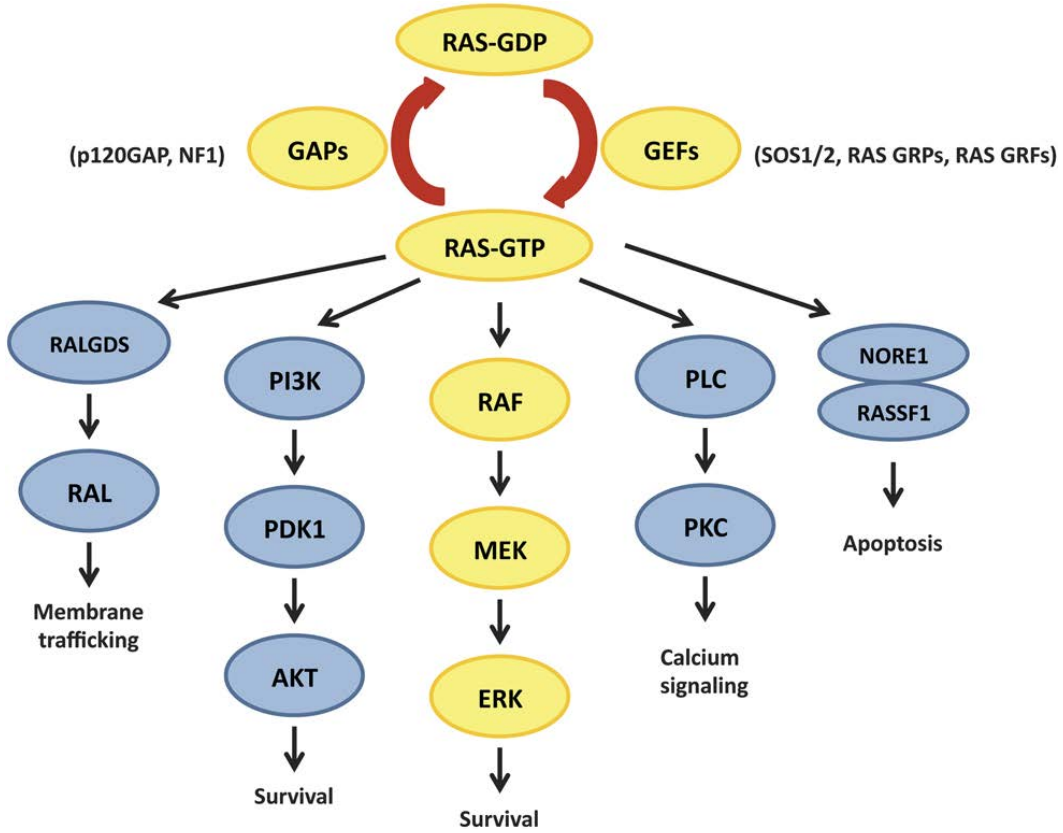


Figure 2.2: Overview of known RAS effectors and their corresponding biological responses. [Reprinted from Lourenco SV et al. ''Head and Neck Mucosal Melanoma: A Review.'' Am J Dermatopathol. 2014;36:578–587. Copyright(C) 2014 by Lippincott Williams, permission from Wolters Kluwer Health, Inc.]

The two main cellular pathways of the RAS proteins operated or studied are the MAPK (mitogen-activated protein kinases) and phosphoinositide-3 kinase (PI3K) pathways. Signal transmission via these cascades may be initiated by the activation of cell surface receptors by growth factors, hormones, and stress. RAF kinase is the first known RAS effector in the MAPK cascade which subsequently activates MEK and ERK (extracellular signal-regulated kinases). The early and late developmental processes are controlled by the RAF-MEK-ERK signalling cascade which includes determination of morphology, organogenesis, synaptic plasticity and growth (Tartaglia M, Gelb BD, Zenker M. 2011). Dysregulation of these cellular processes or functions is also involved in cancer, a major hallmark of RAS/MAPK signalling pathway.

2.2 PI3K/AKT/mTOR signalling pathway

Phosphoinositide 3-kinases (PI3Ks) belongs to the family of lipid protein kinases which regulate various cellular functions like cell growth, size, survival, proliferation, motility, and adhesion, and also fat metabolism/ blood vessel growth (Engelman and Cantley, 2006). The PI3K/AKT/mTOR is an intracellular signalling pathway prominently involved in cell cycle regulation. Upon activation by a wide range of factors like hormones, growth factors or extracellular matrix components, the PI3K adds a phosphate to phosphatidylinositol-4,5-bisphosphate (PIP2) generating an active form, PIP3. This reaction is negatively regulated by PTEN (Phosphatase and tensin homolog), by removal of phosphate, inactivating the PIP3 and slowing down the process. Membrane-associated PIP3 phosphorylates pyruvate dehydrogenase kinase isozyme 1 (PDK1) which inturn activates AKT (protein kinase B (PKB)). AKT, a serine/threonine kinase is translocated to the membrane by PI3K activity. Interaction with PIP3 results in conformational changes exposing the phosphorylation sites of AKT. Partial activation of AKT is done by allowing PDK1 to phosphorylate at Thr308 of AKT. Full activation of AKT is achieved by further phosphorylation at Ser473 by the PDK2 complex including the mammalian target of rapamycin complex 2 (mTORC2). Subsequently, AKT also inhibits the formation of the TSC1-TSC2 complex (Figure 2.3).

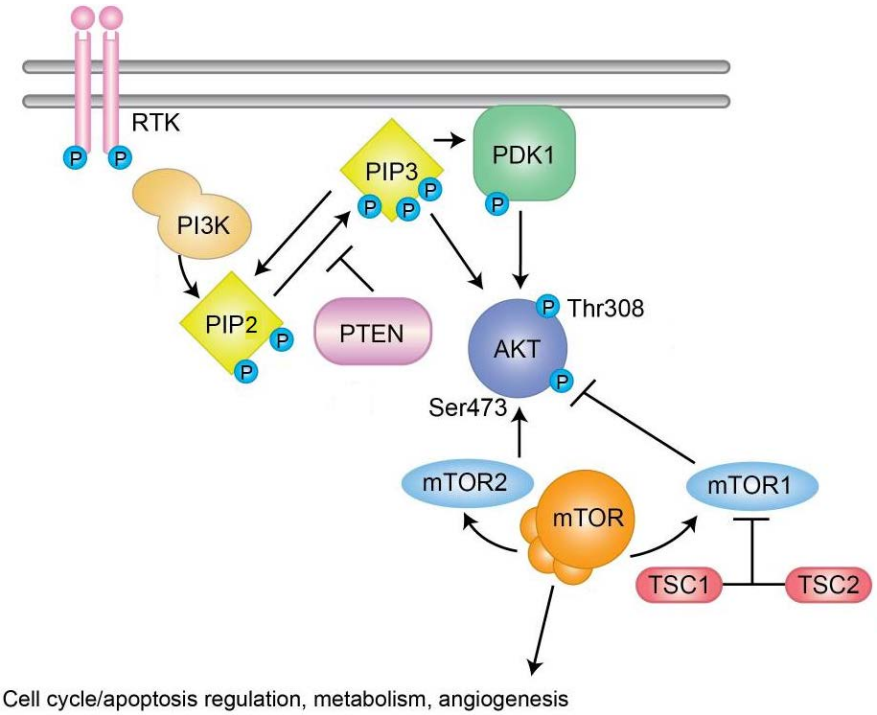


Figure 2.3: Overview of PI3K-AKT Pathway. [Adapted from Keppler-Noreuil, et al., 2015. Copyright 2014, with permission from American Journal of Medical Genetics Part A published by Wiley Periodicals, Inc]

Increased intracellular AKT promotes cell survival, differentiation, motility, proliferation, growth signalling and intracellular trafficking by phosphorylating a range of intracellular proteins.

2.3 Intellectual disability

Intellectual disability (ID), also called learning disability or cognitive deficit (formerly mental retardation) is a disability characterized by significant limitations both in intellectual functioning and adaptive behaviour as expressed in conceptual, social, and practical adaptive skills, which are apparent prior to the age of 18 (Definition from AAIDD, 11 ed., Schalock et al., 2010). ID can be seen as a symptom in certain groups of neurodevelopmental disorders or rare genetic diseases where various cognitive processes are differentially affected. The degree of severity of ID is usually defined by IQ scores - mild (50 < IQ< 70) to moderate/severe (IQ < 50). ID is present in about 1 to 3 percent of the general population in which 75-85% of these reported cases have mild ID and also majority of the cases receive no molecular diagnosis. ID is categorized into two major subclasses- syndromic ID and non-syndromic ID. Syndromic ID is the presence of intellectual deficits as one of the phenotypic feature along with other clinical and behavioural symptoms, in a more global clinical syndrome. Non-syndromic ID is the condition in which intellectual deficits is the only manifestation with no other abnormalities.

The causes for ID are heterogeneous and include both genetic and/or environmental factors which influence the development and function of the central nervous system (CNS) during the pre and postnatal period. The prenatal factors include syndrome disorders, developmental disorders (involving brain), chromosomal disorders, inborn errors of metabolism and environmental factors. Unfortunately, in ~30-50% of cases, the etiology is not identified even after thorough diagnostic evaluation. Environmental factors such as foetal teratogen exposure, malnutrition, premature birth, ischemia, head injury or infectious diseases can cause ID. Other perinatal and postnatal factors include majorly infections and traumas during one's life period.

2.3.1 Genetics of intellectual disability

About 30-50% cases account for ID are caused due to genetic factors; however ID is mostly sporadic with only around 5% of cases with hereditary factors (Daily, Ardinger, & Holmes, 2000). Genetic causes of ID include chromosomal abnormalities (Downs syndrome), microdeletions/duplications (Prader-Willi, Angelman and Williams syndromes) and monogenic diseases (Fragile X syndrome, Noonan syndrome). Unravelling the genetic causes of ID is one of the greater challenges and the study of individual ID-related genes is hindered by the rarity of large enough kindred for linkage analysis, a high rate of de novo mutations and extreme heterogeneity (Winnepenninckx et al., 2003). Many studies have also shown shared interactions between different molecular pathways for ID and various

neurodevelopmental disorders (Hoischen et al., 2014; Vissers, Gilissen, & Veltman, 2016). The best example would be ID and Autism Spectrum Disorder (ASD) in which 17% overlap of genes with de novo loss of function mutations were reported (Ronemus et al., 2014).

Until now, nearly 700 genes have been successfully linked to either syndromic or non-syndromic ID (Vissers, Gilissen, & Veltman, 2016). Around 10-12% of ID cases account for X-linked ID and in figure 2.4 it can be seen clearly that the X-linked genes for ID has reached a maximum plateau i.e. maximum number of genes are identified by now. A major step still impending is the identification of the many number of autosomal ID genes. An estimate of more than 2500 genes has been suggested as autosomal ID genes in which majority of them are recessive ones (Harripaul et al., 2017). Though variants in autosomal dominant genes with de novo occurrence contribute to a large proportion in sporadic cases, autosomal recessive gene variants serve a significant role in ID as they are endured in the population as heterozygous state (Hamdan FF et al., 2014). Recessive mutations causing ID occur mostly in populations with high levels of consanguinity and in normal populations (outbred population) accounts for 13-24% of total ID cases (Musante & Ropers, 2014).

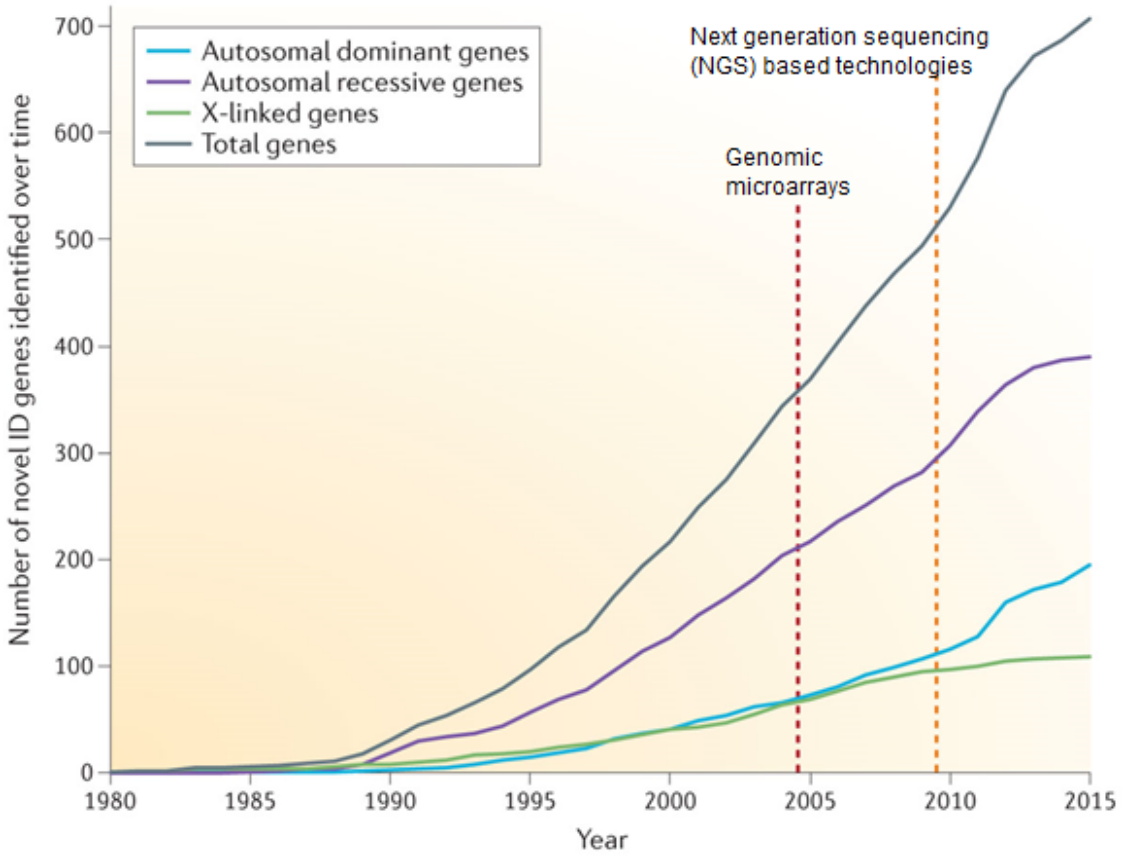


Figure 2.4: Graphical overview of the increase in gene discovery for isolated intellectual disability (ID) and ID-associated disorders over time, specified by the type of inheritance. Vertical dashed lines represent the introduction of novel technologies for the detection of new ID genes. [Adapted from Vissers, Gilissen, Veltman (2016). Genetic studies in intellectual disability and related disorders. Nat Rev Genet 17: 9–18. Copyright(C) by Nature Publishing Group, with permission from Nature Reviews Genetics.]

2.4 RAS signalling pathway and Intellectual disability

RAS is a ubiquitous eukaryotic protein and it is highly expressed in brain. Synaptic plasticity is crucial for neuronal networks development and regarded as the fundamental mechanism for learning and memory. Studies show the involvement of RAS/MAPK pathway playing a key role in regulation of synaptic plasticity - induction of LTP (long-term potentiation) and LTD (long-term depression) (Philips et al., 2013; Pagani et al., 2009; Mainberger et al., 2016). Any impairment in these processes or dysregulation of the RAS/MAPK cascade by germline or mosaic mutations tends to be a common molecular basis for various developmental disorders. An increase in significant knowledge of the role of RAS/MAPK pathway at different developmental time points shows to what degree a single pathway can source multiple anomalies with no distinct connections to each other.

2.4.1 RASopathies

RASopathies or Neuro-cardio-facio-cutaneous syndromes (NCFCS) are a group of developmental disorders with overlapping clinical features caused by mutations in genes that encode components or regulators of the RAS/MAPK pathway (Rauen K.A, 2013; Zenker M, 2011). RASopathies comprises neurofibromatosis type 1 (NF1), noonan syndrome (NS) and related disorders such as cardiofaciocutaneous (CFC), LEOPARD and Costello syndromes (Figure 2.5). RASopathies are pronounced as the largest known or most common group of developmental disorders with an incidence affecting 1 in 1,000 live births (San Martin and Pagani 2014). For RASopathies, strikingly a high level of both locus and allelic heterogeneity is observed. Individual entities of the RASopathies may be caused by the mutations in various genes of the RAS/MAPK pathway, and contrarily, some of these genes can be responsible for different individual syndromes. Recently, mutations in genes which do not belong to RAS /MAPK pathway (RIT1, RRAS) but transduce RAS signalling have been identified causing NS (Aoki et al., 2013; Flex et al., 2014). Although overlap of clinical features is present between the syndromes, each exhibits a distinct phenotype depending on the position of the variant in the RAS/MAPK pathway. Short stature, cardiovascular malformations, ectodermal and lymphatic abnormalities, a characteristic craniofacial phenotype, cancer predisposition are the major features of this group of disorders (Table 2.1). And a variety of neurological, cognitive, behavioural and/or motor coordination problems can be observed. A distinguishable feature also observed in these disorders is the varying degree of intellectual disability ranging from null to severe impairment. Due to the clinical and genetic heterogeneity in RASopathies, it is important to find out correlation between genotype-phenotype associations.

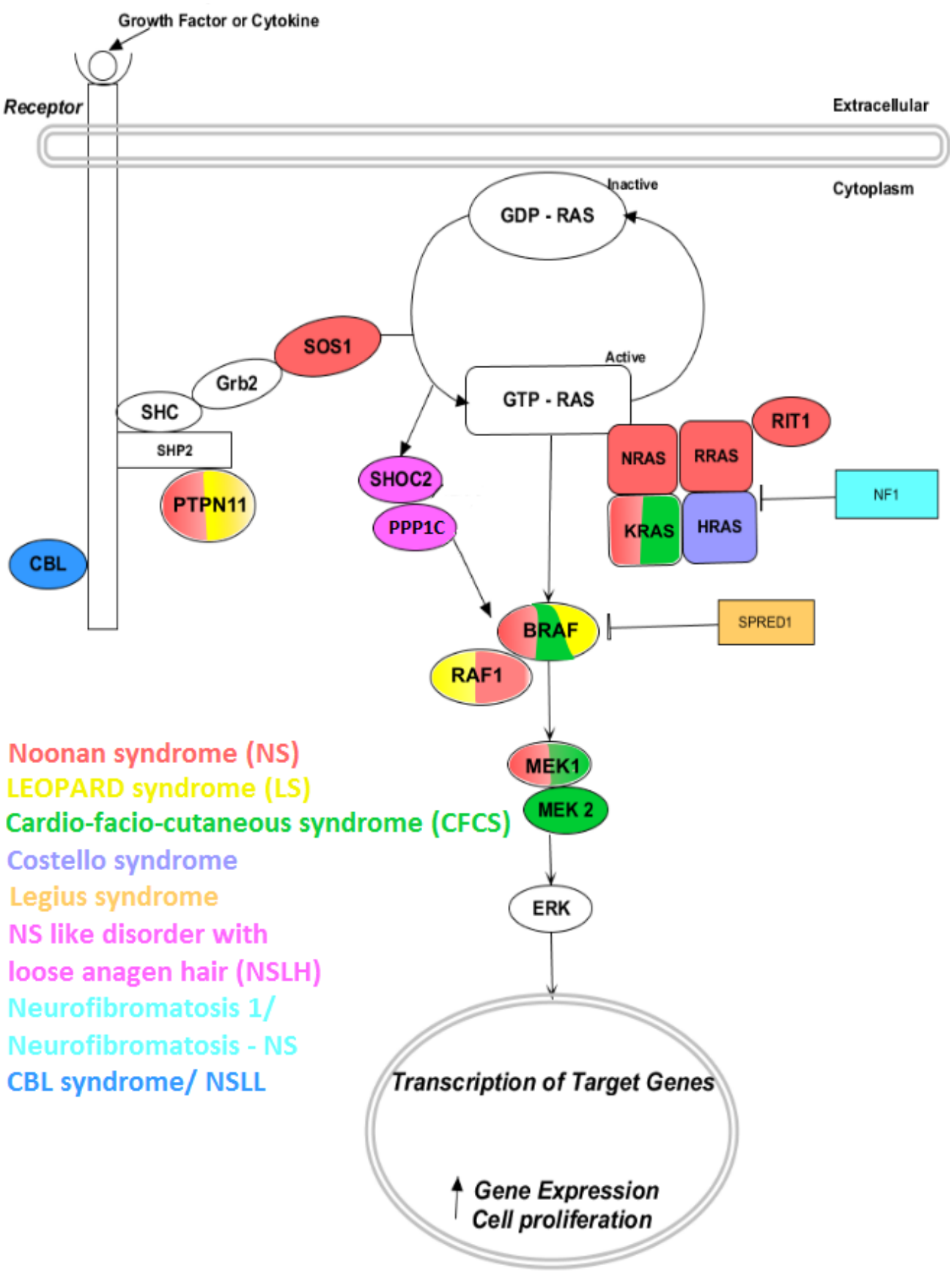


Figure 2.5: RAS/MAPK signalling pathway and disorders with germline mutations of related genes. [Prepared by using PathVisio software, version 3.2.4, www.pathvisio.org, Kutmon et al., 2015.]

Table 2.1: RASopathies and description of their clinical features.

Disorder	**Gene (s)**	**Clinical Features**	**ID[1]**	**OMIM #**
Noonan syndrome (NS)	PTPN11, SOS1, RAF1, KRAS, NRAS, BRAF, RRAS, RIT1	Typical craniofacial dysmorphic features; congenital heart defects; short stature; undescended testicles; ophthalmologic abnormalities; bleeding disorders; predisposition to cancer	0-+	163950
Neurofibromatosis type 1 (NF1)	NF1	Cafe-au-lait spots; Lisch nodules in eye; neurofibromas and plexiform neurofibromas; short in 13%; large head circumference in 24%	0-+	162200
Neurofibromatosis – NS (NFNS)	NF1	Features of both conditions	0-+	601321
Cardio-facio-cutaneous syndrome (CFC)	BRAF, MAP2K1, MAP2K2, KRAS	Distinctive facial appearance; heart defects; failure to thrive; short stature; ophthalmologic abnormalities; multiple skin manifestations, including progressive formation of nevi	+++	115150
Costello syndrome (CS)	HRAS	Coarse facies; distinctive hand posture and appearance; feeding difficulty; failure to thrive; congenital heart defects; short stature; ophthalmologic abnormalities; multiple skin manifestations; predisposition to cancer	++	218040
Legius syndrome	SPRED1	Café-au-lait maculae; intertriginous freckling; macrocephaly	0-+	611431
LEOPARD syndrome (LS) / NS with multiple lentigines	PTPN11, RAF1, BRAF	Noonan-like facial dysmorphism; multiple lentigines; congenital heart defects; short stature; sensorineural deafness	0-+	151100
NS-like disorder with or without juvenile myelomonocytic leukemia (NSLL) or CBL syndrome	CBL	Variable. NS-like facial appearance; microcephaly; predisposition to leukemia	+	613563
NS like disorder with loose anagen hair (NSLH)	SHOC2, PPP1CB	Macrocephaly; short stature with growth hormone deficiency; fine, sparse and easily pluckable hair; characteristic hair phenotype; diffuse skin hyperpigmentation.	+	607721

1. The severity of the characteristic is indicated by the number of + symbols and 0 for null or no ID.

2.4.2 RAS signalling pathway in the nervous system

The initial study by English & Sweatt, 1996 demonstrated the role of RAS/MAPK signalling in cognition in which they showed that MAPK is activated after LTP induction. Since then several studies have indicated that during development and for the normal functioning of the CNS, the RAS-mediated neuronal activities play an important role (Sweatt, 2001; Thomas and Huganir 2004; Ye and Carew, 2010). Further supporting the role of RAS/MAPK pathway in synaptic plasticity and cognitive function many studies have been done using genetically modified mutant mice (Satoh et al., 2007; Jindal et al., 2015; Hernandez-Porras and Guerra, 2017). The first line of evidence suggesting RAS signalling contribution to synaptic plasticity was provided by Heumann et al., 2000 by generating SynRas mice (transgenic mice overexpressing *Hras*G12V in neurons under the control of synapsin promoter) in which neuronal RAS was constitutively active with pronounced neuronal hypertrophy and they also showed an increased size of pyramidal neurons and increased size and complexity of dendritic spines, suggesting the role of such mutations in altering dendritic structures. RAS proteins have also been shown to down regulate the phosphorylation of NMDA receptor which regulates activity-dependent synaptic plasticity and learning and memory (Manabe et al., 2000). Mouse models having partial expression of ERK/MEK showed deficits in long term memory but with an intact short term memory, impaired spatial learning and deficits in long term fear memory suggesting the role of RAS/MAPK pathway in memory consolidation (Brambilla et al., 1997; Satoh et al., 2007; Kelleher et al., 2004). In post synaptic neurons with an increase in intracellular calcium levels, through NMDA receptors or voltage gated calcium channels in response to glutamate or membrane depolarization, also activates the RAS/MAPK pathway (Rosen et al., 1994; Fivaz and Meyer, 2005) and it has been reported that active RAS regulates morphological differentiation of neurons (Biou et al., 2008; Woolfrey et al., 2009). The NMDA receptor shows bidirectional synaptic plasticity depending on the type of subunit activated, causing either activation or inhibition of the RAS/MAPK pathway (Thomas and Huganir 2004). In mature neurons, the surface delivery/recycling of internalized AMPARs is impaired due to inhibition of RAS-ERK pathway by NR2B (an NMDA receptor subunit which drives surface delivery of GluR1) thereby weakening the synaptic transmission (Kim et al., 2005). Many other evidences suggested that signalling of RAS family proteins, either activation or inhibition is critical for memory formation and neuronal morphogenesis (Ye and Carew, 2010; Pierpont, Tworog-Dube, & Roberts, 2013; Lee et al., 2014).

The various genes encoding components or modulators of the RAS/MAPK signalling pathway have been involved in ID and/or ASD. One example, the autosomal gene SYNGAP1

encoding the RASGAP SynGAP (Synaptic GTPase activating protein) has been found to be involved in ID (Hamdan, Gauthier et al., 2009; Hamdan, Daoud et al., 2011; Clement et al., 2012). SYNGAP1 is localized mainly in the excitatory synapses of the neuron interacting with the PSD complex and by activating the glutamate receptors, it suppress RAS signalling activation (Figure 2.6). *Syngap1* knockout mice showed significant deficits in adult hippocampal LTP and also several behavioural deficits like in working memory, auditory fear conditioning, social interaction, contextual discrimination including spatial memory deficits (Clement et al., 2012). In a recent study by Araki et al., 2015, it has been shown that by phosphorylation of SynGAP in hippocampal neurons in response to LTP, synaptic dispersion/scattering of SynGAP was observed rapidly in spines, allowing synaptic incorporation of AMPA receptors through activation of RAS/MAPK signalling thereby increasing synaptic potentiation and spine enlargement. During early postnatal developmental stages, SYNGAP1 has been shown to be involved in negatively regulating synaptic AMPAR trafficking (Rumbaugh et al., 2006). No impairment in cognition or neurotransmission was observed in GABAergic inhibitory neurons when there was a reduction in SYNGAP1 expression (Ozkan et al., 2014). Along with these findings and many more demonstrates the important role of SYNGAP1 in neuronal development.

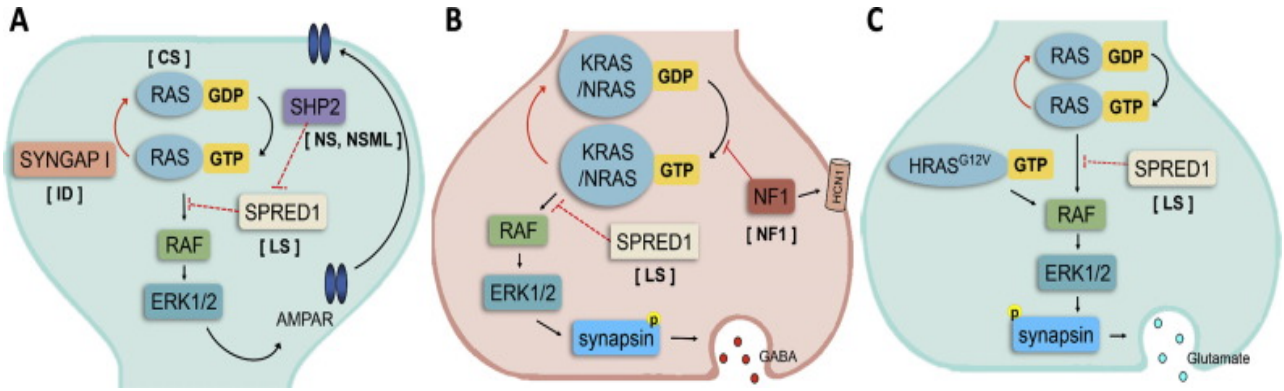


Figure 2.6: Cell type-specific regulation of the RAS/MAPK pathway by distinct regulators. (A) Postsynaptic neuron at excitatory synapses showing multiple positive and negative regulators. (B) Presynaptic neuron at inhibitory synapses showing neurofibromin (NF1) interactions. (C) Presynaptic neuron at excitatory synapses showing $HRAS^{G12V}$ interactions. Protein interactions missing conclusive supporting evidence are indicated with dashed lines. Black and red arrows represent positive and negative regulation, respectively. [Reprinted from "Cell type-specific roles of RAS-MAPK signalling in learning and memory: Implications in neurodevelopmental disorders," by Hyun-Hee Ryu, Yong-Seok Lee, 2016, Neurobiology of Learning and Memory, 135, 13–21. Copyright 2016, with permission from Elsevier]

Similarly many other known genes for RASopathies like PTPN11, HRAS, BRAF and NF1 have been shown at neuronal synapses altering the general mechanisms and affecting neuronal development in many aspects (Ryu & Lee, 2016). Mouse models having mutations in PTPN11 mimicking noonan syndrome showed an abnormal hyperactivation of RAS/MAPK signalling post synaptically by facilitating AMPAR trafficking, thereby causing

deficits in LTP and impairing learning (Lee et al., 2014). In the post-synapse, HRAS has been shown to be involved in phosphorylation of NMDAR, AMPAR trafficking and increased hippocampal LTP (Stornetta and Zhu, 2011; Zhu et al., 2002). In the pre-synapse, HRAS phosphorylates synapsin1 facilitating glutamate release and also enhanced LTP (Kushner et al., 2005). In mouse models, mutations in BRAF showed impaired spatial learning and hippocampal LTP as well as learning deficits in contextual fear conditioning (Chen et al., 2006; Moriya et al., 2015). In the pre-synapse of inhibitory neurons, NF1 is shown as negatively regulating MAPK pathway and its inhibition phosphorylates synapsin1 abnormally causing expedite release of GABA transmitter, and impaired LTP (Omrani et al., 2015, Shilyansky et al., 2010). Gene products of many other genes associated with ID are found in both pre and post synapses having specific synapse functions and aiding in synapse formation and development, including RSK2, CASK, RALGDS, PTEN, TSC1/2 and many more. All these studies strongly implicate a crucial involvement of RAS in neuronal plasticity which in turn may regulate memory formation at behavioural level.

2.5 PI3K/AKT/mTOR signalling pathway in the nervous system

The PI3K/AKT/mTOR signalling pathway has been shown to involve in normal brain development and its dysfunction is linked to many neurological diseases. In CNS, PI3K/AKT signalling pathway is important in development of the neocortex and neuronal survival regulation (Chan et al., 2011) and it has also been shown to play a role in various neuroprotective effects (Leinninger et al., 2004; Tapodi et al., 2005). Mouse models with conventional and conditional ablation of key components of the PI3K/AKT/mTOR pathway resulted in hyper activation downstream of the pathway exhibiting multiple roles in brain development and maintenance (Fraser et al. 2004; Roy et al., 2015). Studies through single cell sequencing identified a mutation burden in both non-neuronal and neuronal cells, indicating the occurrence of mutations in neural progenitor cells (NPCs) (Evrony et al. 2012; Poduri et al. 2013).

Various studies showed that gain of function mutations in the PI3K/AKT/mTOR pathway components resulted in various neurodevelopmental and neuropsychiatric diseases, with distinct clinical phenotypes (Jansen et al., 2015; Mirzaa GM et al., 2012; Rivière et al., 2012). The PTEN hamartoma tumour syndrome and tuberous sclerosis complex (TSC) caused by mutations in PTEN, TSC1, and TSC2 have been extensively studied in humans (Henske et al., 2016; Lachlan et al., 2007) and modelled in mice (Sperow M et al., 2012; Bateup et al., 2013). An increasing number of developmental brain malformations has recently been associated with novel mutations in genes encoding components of the PI3K/AKT/mTOR pathway like

megalencephaly-capillary malformation (MCAP) syndrome, megalencephaly-polymicrogyria-polydactyly-hydrocephalus (MPPH) syndrome, megalencephaly (MEG), focal cortical dysplasia (FCD) and also been identified in brain tissue resected from hemimegalencephaly (HMEG) individuals (Jansen et al., 2015; Mirzaa GM et al., 2012; Rivière et al., 2012). Variable degree of intellectual disability has been reported in these brain disorders ranging from mild learning disability to severe disability. Subsets of patients also have seizures, cortical dysplasia, hydrocephalus, gross motor delays, limb asymmetry or overgrowth, hypotonia, autism and connective tissue dysplasia (Roy et al., 2015). Recently, for the megalencephaly-related syndromes both germline and somatic point mutations in AKT3, PIK3R2, and PIK3CA have been identified (Rivière et al., 2012; Nakamura et al. 2014) and in HMEG, a severe form of megalencephaly, somatic gain of function mutations in AKT3, PIK3CA, and mTOR have been identified (Poduri et al. 2012; Lee JH et al. 2012).

AKT being the central node is a positive regulator for many cellular functions downstream the pathway. AKT3 is highly expressed in the brain and is the predominant isoform than AKT1/2 which are expressed at lower levels in the brain. Through localization of the phospho-Akt (all isoforms) in the developing cortex, it has been shown that AKT has primary role in brain development by enhancing the NPCs in the ventricular zone (Poduri et al. 2012). Somatic gain of function mutations in AKT1 causes Proteus syndrome (Lindhurst et al., 2011) and activating mutations in AKT2 are linked to overgrowth and hypoglycemia (Hussain et al., 2011). Germline and/or somatic mutations of AKT3 have been shown linked to megalencephaly-related syndromes, HMEG and malformations of cortical development (MCD) (Poduri et al. 2012; Alcantara et al., 2017; Wang L et al., 2017).

Together with all the evidences and studies, an activation of the PI3K/AKT/mTOR signalling pathway shows increased proliferation of NPCs, neuronal hypertrophy and increased dendritic branching and causes localized and restricted abnormalities depending on the type of mutation and cell type specific mutations (Evrony et al. 2012; Poduri et al. 2012; Wang L et al., 2017) suggesting the role of the pathway in CNS development.

2.6 Mosaic disorders

Genetic developmental disorders are mostly caused by germline mutations that may be either inherited or have occurred de novo in a parental germ cell. However, disease-causing mutations may also arise postzygotically at early embryonic stages. A **mosaic** or **mosaicism** denotes the presence of two or more populations of cells with different genotypes in one individual who has developed from a single fertilized egg. Specific genetic changes may even be seen predominantly or exclusively in a mosaic status. Happle

first postulated the concept of mosaicism in which a lethal mutation is survived in certain monogenic disorders when present in close proximity to normal cells or postzygotic de novo mutations in early embryonic stage (Happle, 1987). During the past few years, an increasing number of mosaic disorders involving RAS/MAPK and PI3K/AKT pathway components have been delineated clinically and molecularly, starting with Proteus syndrome caused by mosaic mutations of the AKT1 gene (Lindhurst. et al., 2011), and followed by several other disorders.

2.6.1 RAS pathway and mosaicism

Mosaic variants of disorders that are usually seen with germline mutations can occasionally be observed in neurofibromatosis type 1, and two cases of Costello syndrome with mosaic HRAS mutations have been described (Gripp, et al., 2006; Sol-Church, et al., 2009). Similarly, mosaic cases may also exist for other RASopathies like Legius syndrome (SPRED1), Rhomdoid nevus syndrome (capillary malformation–arteriovenous malformation; RASA1), and LEOPARD syndrome (PTPN11). Mosaic mutations affecting the RAF/RAS/MAPK signalling pathway have recently been described in an increasing number of (neuro) cutaneous disorders and congenital nevi including Schimmelpenning syndrome (HRAS, KRAS) (Groesser et al., 2012), Keratinocytic nevus (HRAS, KRAS, NRAS) (Hafner et al., 2012), Nevus sebaceous (HRAS, KRAS) (Sun et al., 2013), Neurocutaneous melanosis (NRAS, BRAF) (Charbel et al., 2014; Salgado et al., 2015), and Nevus spilus-type congenital melanocytic nevi (NRAS) (Kinsler et al., 2014; Krengel et al., 2016). The term “mosaic RASopathies” has been introduced and is now mainly used for disorders where typically the oncogenic type of mutations can be found in affected tissues but not in the blood or unaffected tissues (Luo and Tsao, 2014). These observations are in line with the hypothesis that these mutations are only tolerated, if they do not affect all cells of an organism. Clinically, mosaic RASopathies appear to have little in common with the germline RASopathies, and Noonan syndrome-like features are usually not recognizable.

Oculoectodermal syndrome (OES) and encephalocraniocutaneous lipomatosis (ECCL) are rare disorders that share many common features such as epibulbar dermoids, aplasia cutis congenita / focal alopecia, pigmentary changes following Blaschko lines, bony tumor-like lesions, and others. Neurodevelopmental symptoms like developmental delay, epilepsy, seizures, learning difficulties, and behavioural abnormalities have also been reported (Ardinger, et al., 2007; Moog, 2009). A distinct hairless fatty tissue nevus of the scalp (naevus psiloliparus) is regarded as the dermatological hallmark of ECCL (Happle and Kuster, 1998). Subcutaneous fatty masses in the frontotemporal or zygomatic region are common in ECCL

but have occasionally been reported also in children diagnosed with OES. In addition, giant cell granulomas of jaws and non-ossifying fibromas of long bones have also been reported in ECCL (Moog, 2009). About 20 cases with OES and more than 50 patients with ECCL have been reported in the literature. In both, OES and ECCL, exclusively sporadic occurrence has been observed. Together with the obvious mosaic pattern of skin involvement, this was considered suggestive of a genetic mosaicism with mutations that would confer embryonic lethality when occurring in the germline (Moog, 2009).

Recently, Peacock et al., identified mutations in the KRAS (V-Ki-RAS2 Kirsten rat sarcoma viral oncogene homolog) gene, namely c.38G>A (p.Gly13Asp) and c.57G>C (p.Leu19Phe), in affected tissues from two patients with OES, thus suggesting that OES is a mosaic RASopathy (Peacock, et al., 2015). Here in this current study, we present three further patients with OES and one with ECCL in all of which specific mosaic mutations in the KRAS gene could be demonstrated in lesional tissue.

2.6.2 PIK3CA-related overgrowth spectrum (PROS)

Correspondingly, in the group of disorders that is now known under the term “PIK3CA (Phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit alpha isoform)-related overgrowth spectrum (PROS)”, identical mosaic mutations may account for various phenotypic expressions depending solely on tissue distribution of the mutation (Keppler-Noreuil, et al., 2015). Activating PIK3CA somatic mutations have been shown in various regional overgrowth conditions like CLOVES syndrome (congenital lipomatous overgrowth, vascular malformations, epidermal nevi, and skeletal abnormalities; MIM 612918), megalencephaly-capillary malformation syndrome (MCAP; MIM 602501), dysplatic megalencephaly (DMEG), fibroadipose overgrowth (FAO), hemihyperplasia-multiple lipomatosis (HHML), isolated Macrodactyly, and few cases of Klippel-Trenaunay syndrome (KTS; MIM 149000) (Kurek et al., 2012; Rios et al., 2013; Mirzaa, Riviere, and Dobyns, 2013; Keppler-Noreuil et al., 2015; Vahidnezhad et al., 2016). Each disorder has distinct clinical features however frequent overlap with other PROS exists like vascular malformations, mosaic skin lesions like epidermal nevi and regional segmental overgrowths like macrodactyly.

The PIK3CA gene encodes a catalytic subunit p110a of the phosphoinositide-3-kinase heterodimer (Figure 2.7). The Adaptor binding domain (ABD, p85) and the RAS binding domain (RBD) domain interact with the PI3K/PI4K kinase domain. Major somatic mutations for PIK3CA gene have been identified in C2, helical and kinase domain. Very few mutations

in ABD and in the RBD have been identified until now related to developmental disorders (Mirzaa G. et al., 2016).

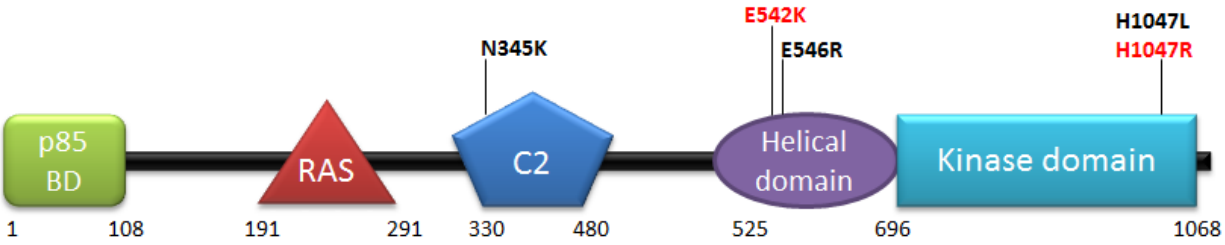


Figure 2.7: Schematic of PIK3CA gene structure and its key functional domains. High frequency of mutations is observed in the helical and kinase domain of p110a (Activating mutations E542K and H1047R are highlighted in red). (BD- binding domain)

Gain-of-function mutations in PIK3CA on chromosome 3q26 have been identified in affected tissues from patients affected by CLOVES syndrome, demonstrating somatic mosaicism of varying degrees (Kurek et al., 2012). CLOVES syndrome is a sporadically occurring, regional overgrowth disorder characterized by asymmetric somatic hypertrophy and anomalies in multiple organs. It is caused by somatic mosaicism for mutations affecting components of the PI3K/AKT/mTOR signalling pathway. CLOVES syndrome is differentiated from other regional overgrowth syndromes by the presence of truncal overgrowth and characteristic patterned macrodactyly at birth (Figure 2.8).

Clinical features for CLOVES syndrome:

- **Fatty Truncal Mass** - a soft fatty mass of variable size in one or both sides of the back and abdominal wall with extending into gluteal or groin regions
- **Vascular Anomalies** - capillary malformations, abnormal lymphatic and venous channels, spinal arteriovenous malformation
- **Abnormal extremities** (arms and legs) and **Scoliosis** (curving of the spine) - large wide hands or feet, large fingers or toes, wide space between digits and uneven size of extremities
- **Skin abnormalities** - birthmarks, prominent veins, lymphatic vesicles, moles and epidermal nevus (light brownish slightly raised skin in the upper chest, neck or face)
- **Neurological abnormalities** - Hemimegalencephaly, syringomyelia, agenesis of corpus callosum, seizures
- **Other abnormalities** include small or absent kidney, abnormal patella (knee cap), abnormal knee and hip joints

Mouse models expressing the common activating PIK3CA mutations (H1047R and E545K) showed resembling human clinical features, including brain enlargement, cortical malformations, hydrocephalus and epilepsy and further treatment with PI3K inhibitors and suppression of PI3K signalling ameliorated seizures in these animals (Roy et al., 2015).

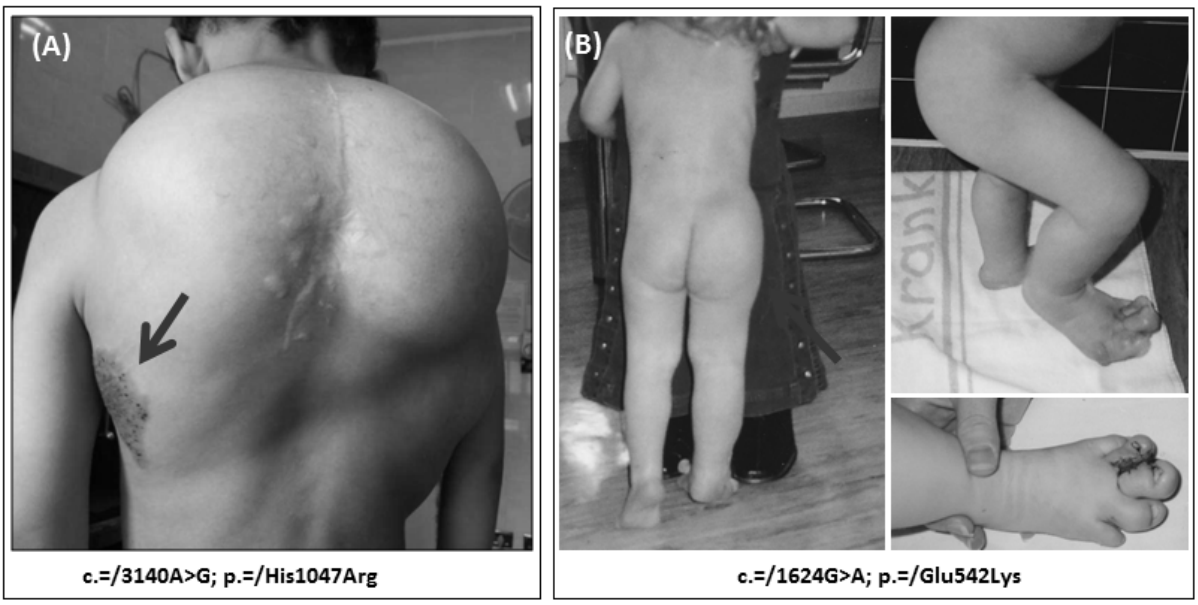


Figure 2.8: Clinical pictures of CLOVES patients having recurrent activating mutations His1047Arg and Glu542Lys. (A) A boy with mutation at c.3140A>G in PIK3CA showing vascular malformation with extensive lipomatosis and the mutant allele proportion detected in the lipoma is around 30-40% (picture courtesy by Dr. Eman Ragab, Tanta Faculty of Medicine and University Hospitals, Tanta, Egypt and the picture is printed with permission from parents) (B) A three year old girl with mutation at c.1624G>A in PIK3CA showing disproportionate growth of the toes II to IV occurred to the right (macrodactyly of the toes II-IV) and lipomatosis of the right gluteal region which extends into the proximal thigh. An increase of the soft tissue in the area of the right sole (lipomatosis) is also present and the mutant allele proportion detected in the lipoma is around 30% [Picture of the patient reprinted with permission from, Eva Schneckenhaus (2009). Mutationsanalyse des PTEN-Gens bei Proteus-Und Proteus-Like-Syndrom (Dissertation Thesis). , Medizinischen Fakultät, Otto-von-Guericke University Magdeburg, Germany.]

The current study focuses on identification of somatic mutations of PIK3CA in patients presenting with CLOVES syndrome. Here, we also delved into the liability of using other detection methods like amplicon deep sequencing and fragment analysis, next to Sanger sequencing, to detect somatic mutations in DNA from different tissue samples.

2.7 Next generation sequencing technology

Several novel approaches were explored to replace Sanger as the dominant provider of sequencing technologies as it was a leading method for over 30 years. In 2005, the 454 systems (Genome Analyzer) based on pyrosequencing method becomes the first next generation sequencing (NGS) technology which was developed by Roche Company. Since 2006, massive evolution of technologies, instruments and methods has emerged revolutionizing the world of genomics. During the last years, NGS technologies which employ massively parallel approaches to produce millions of sequence reads in a single run have made it possible to sequence genetic regions and complete genomes in a time-efficient manner with a low per-base cost (Schuster SC 2008). Next-generation sequencing has been applied mainly to de novo sequencing of bacterial and plant genomes, resequencing of entire

human genomes, exome sequencing and targeted resequencing of (entire) known susceptibility genes or loci of interest. NGS applications have widespread over areas from epigenetics to transcriptome sequencing and also with increasing use in single cell analysis and metagenomics.

2.7.1 Evolution of NGS

Since the introduction of 454 systems in 2005, rapid and important advancements in sequencing chemistries or methods have been achieved till date. Over the past decade, the sequencing technologies have also evolved continuously, increasing the capacity of the data by a factor of 100-1000 implementing revolutionary methods (Kircher & Kelso, 2010). Figure 2.9 represents a brief overview of the progress of the technology development over the years and one can view an excellent pace achieved through NGS until the ability to sequence an entire human genome under routine analysis with the cost decreasing rapidly and the data output increasing massively.

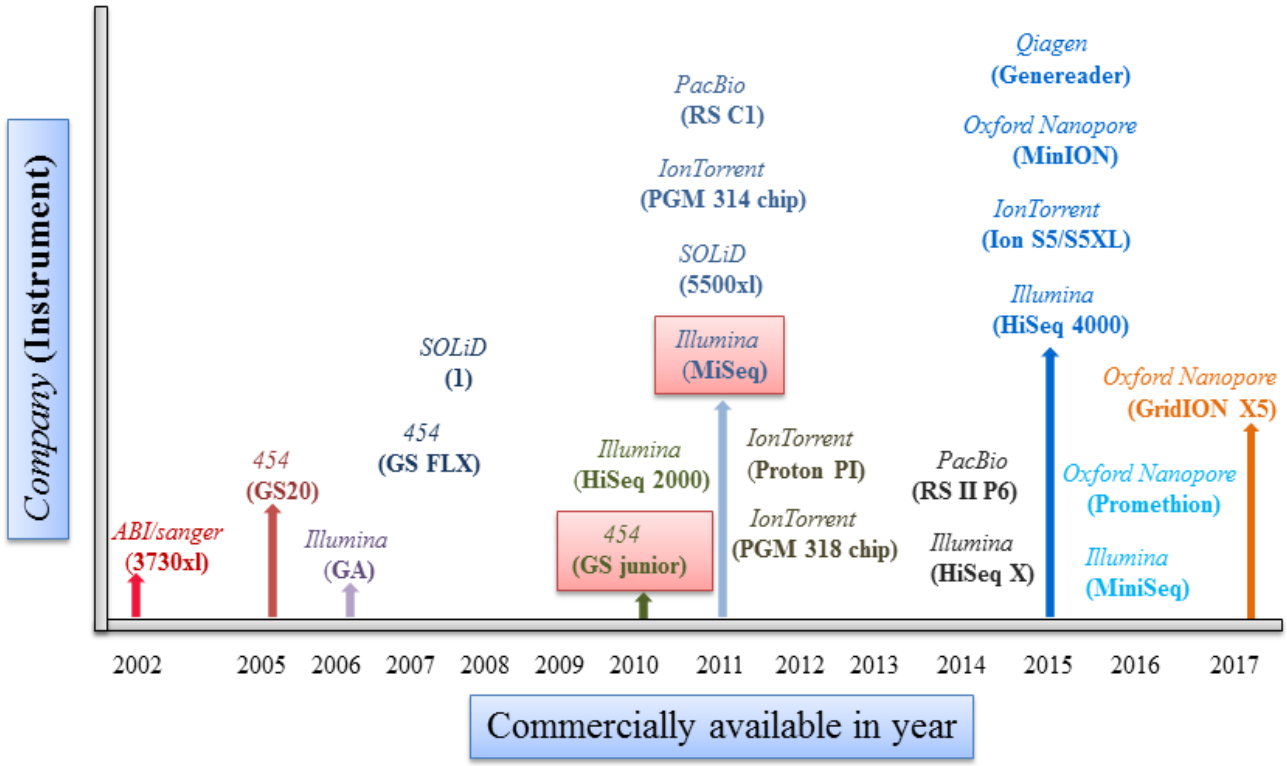


Figure 2.9: Commercially available sequencing platforms over the years. The sequencing instruments used in this current study are highlighted in red background.

From the year 2009, NGS is being used for different studies related to human gene mutations and in 2011 the FDA has approved the use of NGS in clinical diagnostics application. With the introduction of Illumina's Hiseq X in 2014, the entire human genomes could be sequenced in less than 3 days producing a data of 1.8Tb (Figure 2.10) with costs dropping to nearly $1000 per genome (Figure 2.11). The cost per genome has drastically dropped in the last two years making breakthroughs in genomics by generating more quality data.

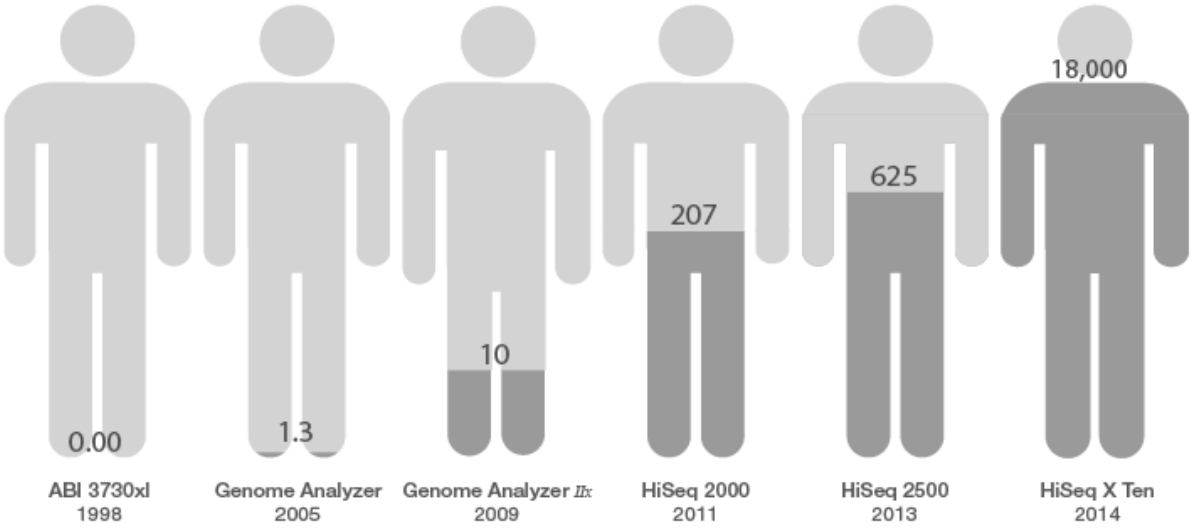


Figure 2.10: Human genomes sequenced annually over the years. The capacity to sequence the entire human genome (at 30x coverage) has increased massively over the years. [Figure is adapted with the permission from Illumina and remains their copyright, Courtesy of Illumina, Inc.]

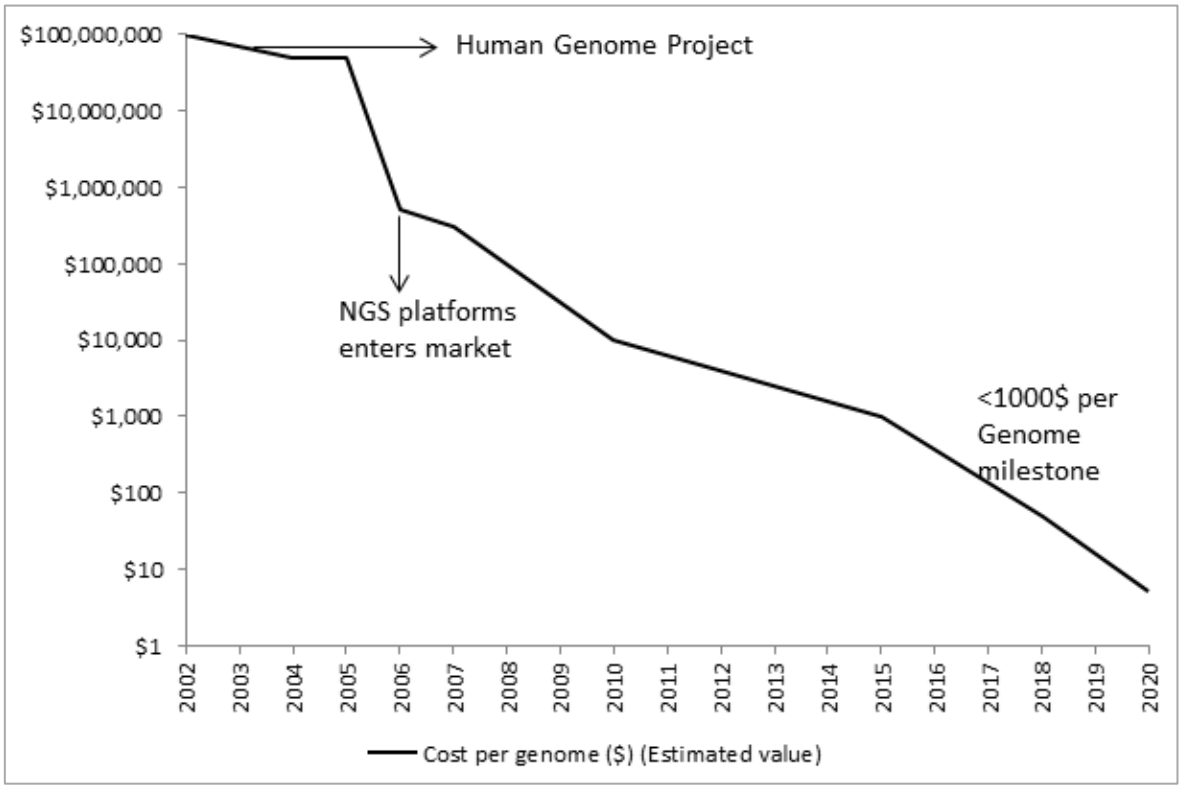


Figure 2.11: Graph illustrating the rapid decrease in the cost of genome over the years.

Among the different available platforms, there exist similarities and disparities between them due to their sequencing chemistries which yield to a wide range of capabilities and/or specifications in different applications. Although a number of different parameters are used for comparing the performance of the platforms, majorly two main criterions are considered, the number of reads produced by the instrument and their corresponding read lengths (Figure 2.12). Other parameters include cost per run/base, sample preparation time/cost, instrument run time/cost, percentage frequency of sequencing errors and overall efficiency.

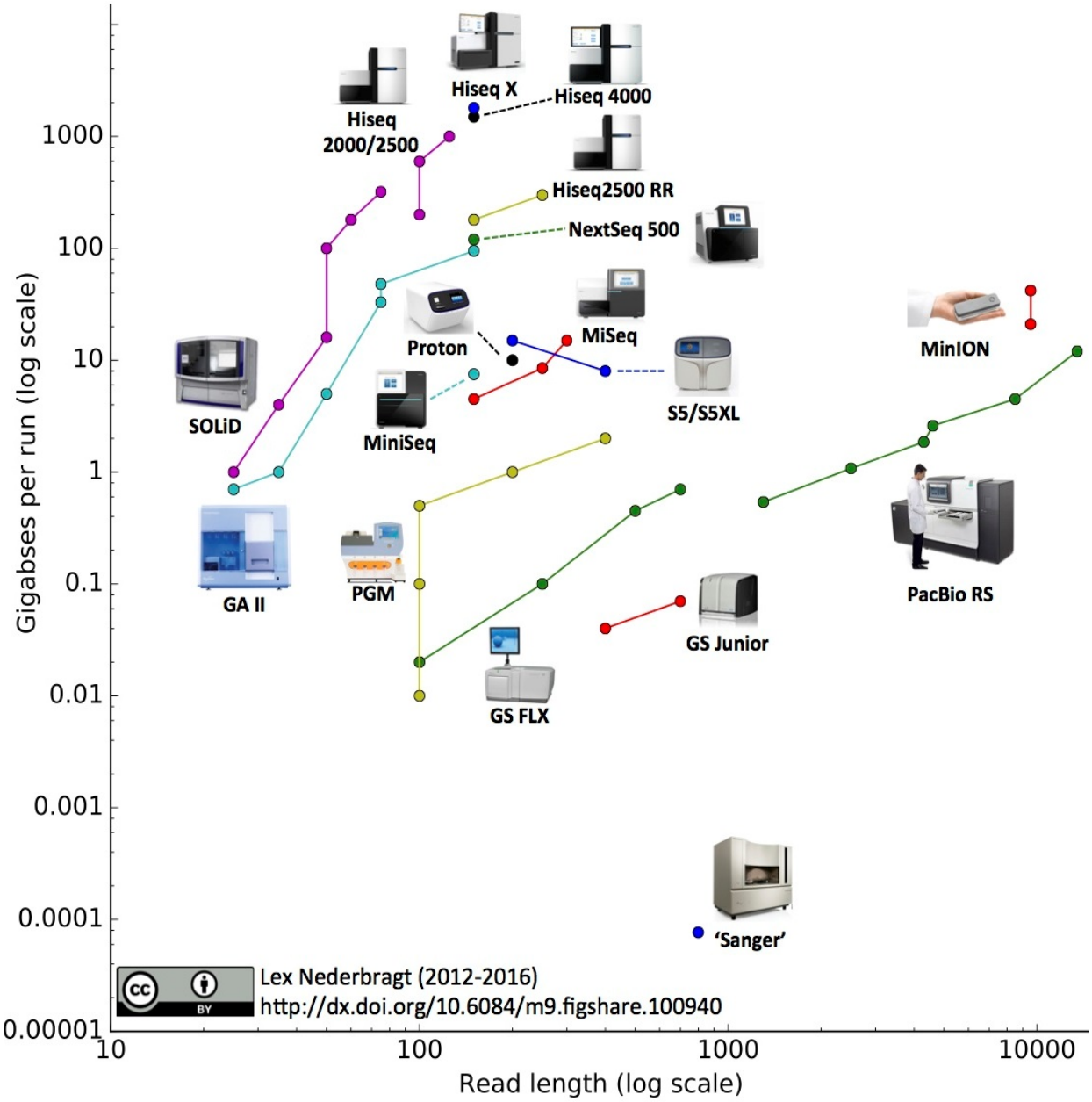


Figure 2.12: Graph representing the developments in high throughput sequencing. The data is based on the throughput metrics for the different platforms since their first instrument version came out. The figures visualize the results by plotting throughput in raw bases versus read length. [Nederbragt, Lex (2016): developments in NGS. figshare. https://doi.org/10.6084/m9.figshare.100940.v9 Retrieved: May 18, 2017 GMT). https://flxlexblog.wordpress.com/2016/07/08/developments-in-high-throughput-sequencing-july-2016-edition/. Under the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0)]

In the coming years a wide variety of new generation sequencing instruments with spectacular chemistries are about to come offering much higher read data in efficient time, easier usability, and low cost. Some of them include Quantum Biosystems company using nanogate technology, Base4 company using pyrophosphorolysis method, GenapSys company introducing GENIUS sequencer using sequencing by synthesis method and solid state detection, Qiagen Genereader using sequencing by synthesis method, Illumina Firefly using one-channel CMOS (complementary metal-oxide semiconductor) technology and Roche Genia using nanopore technology (Goodwin, McPherson, and McCombie, 2016).

2.7.2 Applications of NGS

The applications for NGS are very high and broad including all possible ways like de novo genome sequencing, epigenetics, metagenomics and microbiomes, and transcriptomics.

- Whole-genome sequencing (WGS) is one of the most widely used application in which an extensive genomic information and associated biological significance could be obtained. This method serves as a bridge in identifying many differences between samples by comparative analysis of multiple whole genomes. Exome sequencing (WES) or exomics is also an invaluable method in which only the coding exons are sequenced for a particular gene helping identification of mutations either for rare disorder like ID or more common disorders like cancer. In contrast to exomics, WGS can assess alterations in the coding genes and the regulatory and noncoding regions, especially multiallelic copy number variations (Handsaker et al., 2015). Majorly in humans, both WGS and exome analysis are aimed to detect and catalogue SNPs, de novo mutations, and sequence variants such as copy number, indels, and structural variations (Rabbini, Tekin & Mahdieh, 2014). Although the cost of DNA sequencing has come down a lot in recent years and still would be much decreased in near future, there still remains a question of cost management when sequencing large number of samples by exome or whole genome sequencing.

- Targeted resequencing by gene panels is a high sensitivity method used for ultra-deep sequencing of the PCR products of particular genes of interest. Targeted resequencing is done by constructing gene panels with definite number of potential genes (limiting the size of the genome) for a particular disorder, thereby generating high quality reads per run increasing the depth of the genomic study and reducing costs (Griffith M. et al., 2015). The data set generated is also smaller and manageable when compared to WES or WGS making analysis easier and efficient. Another important advantage of targeted resequencing is the requirement of low amount input DNA than WES or WGS. This method is ideally suited for clinical applications like genotyping, rare variant detection, disease associated gene sequencing, and Genome Wide Association Studies (GWAS).

- Another important application is detection of somatic mutations either by WGS, WES or targeted sequencing which serves as an important diagnostic tool. The high depth coverage of NGS helps in detecting even low grade mosaicism in mutant alleles from wild type which is often considered as background noise in Sanger sequencing.

- RNA-seq or the transcriptome analysis is a very important method in which all the RNA transcript sets expressed by the genome in cells, tissues, and organs at different stages of an organism's life cycle are sequenced. With this sequencing method knowledge about the biological intricacies of genome function could be obtained in detail which is limiting in the genomics sequencing (Mele M et al., 2015).

- NGS is also involved in Epigenomics (study of heritable gene regulation) and Methylomics (genome wide analysis of DNA methylations) providing insight into the regulatory mechanisms of the genome (Soon WW et al., 2013).

- Other important applications of NGS are Single-cell and metagenome sequencing. Through Single-cell sequencing individual cells are sequenced to gain information on cell based interactions and variations. Metagenomics helps in study of the microbial community thus helping gain important information on various parameters like knowing the ecosystem, in epidemiological studies, and identification of new species (Gilbert & Dupont, 2011; Treutlein et al., 2014).

NGS is currently used in detection of mutations linked to rare Mendelian disorders or more genetically heterogeneous complex disorders such as ID. Since the introduction of NGS in 2007, a rapid rise in disease gene identification for rare diseases is induced thus increasing the rate of diagnosis. Still many diseases await genetic cause and many mutations identified by NGS need to be catalogued. However careful scrutiny of the variants detected must be done since significant levels of false-positives and false-negatives might be generated due to sequencing errors or amplification biases NGS (Rieber et al., 2013).

In the current study, targeted resequencing approach with gene panels is implemented for identification of rare and novel variants linked to intellectual disability and mosaic disorders. This method is chosen rather than for genome or exome sequencing, as it provides an in-depth analysis of the mutations linked to specific disorders, costs and analysis time. Gene panel sequencing approach also reduces the chance of incidental findings and also generates low false-negatives than WGS or WES. The major limiting factor of this method is that the prediction of the disease-causing gene to be included in the panel. Two different sequencing platforms are used in this study. For somatic mutation detection, amplicon based resequencing method is opted which is performed on the Roche 454 GS Junior system. For identification of rare and novel variants linked to ID, targeted enrichment capture method is used and the sequencing is performed on the Illumina Miseq® system.

3. Objectives

3.1 Multigene panel sequencing in patients with ID and Short stature

Activating germline mutations in various genes encoding components or modulators of the RAS/MAPK signalling pathway have been found to cause a group of clinically overlapping syndromes (including Noonan, CFC) (Schubbert et al., 2006; Zenker M, 2011). Cognitive deficits of variable expression are part of all these diseases and are considered to reflect dysregulated RAS/MAPK signalling in the nervous system. We hypothesized that mutations in other modulators of the pathway that are preferentially expressed in neuronal cells may be responsible for non-syndromic ID without the typical physical symptoms. It has also been implicated that short stature is a common feature noticeable in these syndromes. Apart from the affected signal proteins of the RAS/MAPK pathway, SHP2, encoded by PTPN11, is known to be implicated in growth hormone (GH) signalling related to short stature. Besides SHP2, other interconnections exist between the RAS/MAPK and GH pathways that remain to be elucidated. Apart from activating MAP-kinase pathway, the activated RAS also binds to PI3K thereby activating the PI3K/AKT/mTOR pathway. The PI3K/AKT/mTOR pathway is also one of the major pathways involved in many neurological disorders comprising ID and ASD (Alomari AI. 2009a; Gucev ZS et al., 2008). The JAK-STAT cascade is also one of the major signalling pathways stimulated by cytokines and growth factors. We therefore intended to design a gene panel which consists of subset of genes from all the four pathways - RAS/RAF/MEK, GH, JAK-STAT and PI3K/AKT/mTOR pathway (Figure 3.1) which may be disease relevant and identify rare or novel variants in these selected genes.

The primary goal of this subproject is to evaluate the significance of mutation in genes related to the RAS/MAPK pathway in non-specific / non-syndromic types of ID and to evaluate the usefulness of NGS multigene panel sequencing to reach this goal. To this end, the study aimed at identification of RAS/MAPK-related genes that may play a role in the nervous system by a systematic data search and in silico evaluation. The secondary goal is to identify mutations in growth-related genes in a short stature cohort. The below following objectives were implemented in this current project for the RAS-GH custom-designed gene panel

1. Target gene selection related to neuronal components/ modulators of the RAS/MAPK pathway using different criteria and predicted protein-protein interactions.
2. Selection of target genes which are related to short stature, GH pathway along with PI3K and JAK-STAT pathways, under a common pathway termed as GH pathway in this current study.

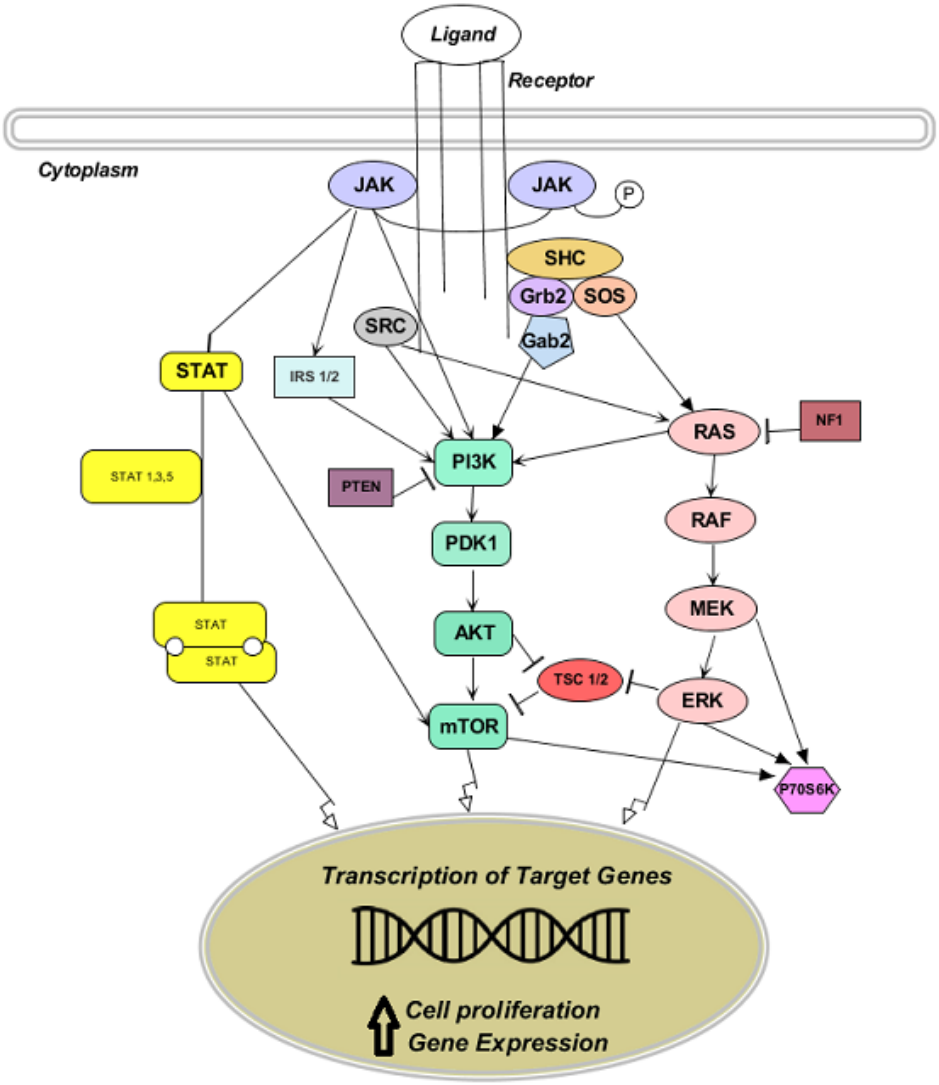


Figure 3.1: Schematic representation of four different signalling pathways: the RAS/ERK/MEK pathway, the JAK-STAT pathway, the PI3K/AKT/mTOR pathway and the GH pathway. [Prepared by using PathVisio software, version 3.2.4, www.pathvisio.org, Kutmon et al., 2015.]

3. Collection of all available data for the selected genes till date which is published online by different resources and prioritizing the selected genes of both RAS and GH pathway through scoring system based on different parameters.
4. Screen the genes of RAS/MAPK pathway in a cohort of patients with non-syndromic or non-specific ID and for GH pathway in a cohort of patients with GH insensitivity and short stature using next generation sequencing technologies. The Short stature cohort will be served as a control for comparison of results.
5. Evaluate the frequency of mutations in members of RAS/MAPK and GH signalling pathway and gain insight on gene-phenotype association. Also studying the variation prospect of the minor allele load on a gene-by-gene basis in both the cohorts.
6. Also evaluate the effects or activations in both the pathways and their inter relation. Collect relevant statistical data between the two cohort groups and two pathways in all different functional and basic categories.

7. All sequence variants identified will be filtered against various public online databases to eliminate already known, apparently non-pathogenic changes. In addition, the OMIM (Online mendelian inheritance in man) catalogue, ClinVar and the Human gene mutation database (HGMD) will also be used as a filter to identify all previously described changes and known human phenotypes associated with gene.
8. Novel sequence variants detected by this approach will be further validated by traditional sequencing methods (Sanger sequencing) and their pathogenicity will be checked by the various online prediction programmes.
9. For further confirmation of the novel mutations to be disease causing, segregation analysis will be performed on the parent's DNA samples through traditional sanger sequencing and by analyzing new DNA samples from patients not yet under investigation or by analyzing on the RNA level rather than in genomic level, so the effects on the m-RNA and protein level also needs to be shown. This can be investigated further by cloning and its expression can be studied. When possible, additional family members will also be recruited for follow-up cosegregation analysis.
10. We will also focus on the genes that are already implicated in studies related to ID (CTNNB1; SHANK's; KCNQ3 etc..,) and provide insight in further characterization of already-known disease-associated genes.

3.2 Mosaic disorders

The primary goal of this sub project is to identify mutations underlying OES and PROS in the specific phenotype cohorts and to check the distribution of mutations in various tissues and evaluate sensitivity of different detection methods. Patients with mosaic mutations can occasionally be observed in neurofibromatosis type 1, and in two cases of Costello syndrome mosaic HRAS mutations have been described (Gripp KW et al., 2006; Sol-Church K et al., 2009). Similarly, mosaic cases may also exist for other RASopathies. Recently OES has been termed as a mosaic RASopathy (Peacock et al., 2015). OES and encephalocraniocutaneous lipomatosis (ECCL) are rare disorders that share many common features. The current study focuses to corroborate the evidence of OES being a mosaic RASopathy and confirm the common etiology of OES and ECCL. Also interactions between the PI3K/AKT/mTOR and RAS/RAF/MEK pathways have been identified in various malignancies. So the current study also focuses on identification of somatic mutations of PIK3CA in patients presenting with PROS. Here, we also delved into the liability of using other detection methods like amplicon deep sequencing and fragment analysis, next to Sanger sequencing, to detect somatic mutations in DNA derived from different tissue samples.

4. Materials and Methods

4.1 Study subjects

All genetic studies were done on genomic DNA extracted from blood or tissue samples. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Analyses were done in diagnostic context and informed consent was given by the parents or a legal guardian according the directives of the respective countries. For all study subjects, signed consent for scientific evaluation and publication of clinical data was given.

A total of 166 ID patients and 120 Short stature patients were screened for mutation detection using the custom-designed panel (RAS-GH panel) in this current project.

ID Cohort – This cohort consists of individuals with non-specific ID presenting for genetic evaluation at the University Hospital Magdeburg, Germany during the recent 6 years (2010-2016). All individuals had had a routine cytogenetic analysis with a normal karyotype, molecular genetic testing for Fra(X) mental retardation with normal results, and microarray analysis using an Affymetrix genome-wide human SNP 6.0 array (Affymetrix, Inc., Santa Clara, CA) for the exclusion of pathogenic copy-number variation (CNV) in the genome. Experimental procedures had been performed according to the manufacturer's instructions. Image data of the Affymetrix array had been analysed with the Affymetrix Genotyping console 4.1 and the Chromosome Analysis Suite v1.2 and v3.0.
The cohort was recruited with ID and normal results for the aforementioned genetic tests as only inclusion criteria. Most of the patients had non-specific and/ or non-syndromic ID and no selection was made based on severity of cognitive impairment. The age range was 7-9 years and contained more males than females. Most of the patients had mild ID with an IQ within the range 50-70 and patients with severe ID contribute only 10-12% of the whole cohort. Priority selection of the patients within the cohort was done based on availability of parent DNA samples. When possible, additional family members were also recruited for segregation analysis.

Short stature Cohort - This cohort consists of individuals who were treated with recombined growth hormone because of idiopathic growth hormone deficiency (IGHD) or intrauterine growth retardation (IUGR) / small for gestational age (SGA) presenting at the University Hospital Homburg, Germany (This cohort will be termed as 'GH cohort' in this current study). Diagnosis of IGHD was based on two different growth hormone stimulation tests

showing no appropriate increase of the hormone. The SGA group includes children with birth weight and length below the 3rd percentile and insufficient catch-up growth during the first years of life. Children with known syndromes like Noonan syndrome, Leri-Weill syndrome (SHOX deficiency), and Prader-Willi syndrome were excluded in this cohort. Also children with brain malformations, radiation, brain cancer, all other forms of cancer and other diseases that have to be treated were excluded. Specifically, children that were pre-pubertal at the beginning of the therapy and during the first year were only included. The age of the children was around 5-7 years. This study had been approved by the ethical committee of the medical association of Saarland (No.58/06) and conducted in accordance with these guidelines and the Declaration of Helsinki principles.

Specific phenotype cohorts: For the mosaic disorder studies like OES/ECCL or PIK3CA-related overgrowth syndrome (PROS), patients were samples were collected from national and international partners according to the predefined clinical phenotype.

For the PIK3CA study, a total of 18 patients were selected with various tissue and blood samples of each patient comprising a total of 48 analysed samples. Out of the 18 patients analysed, 14 patients were classified as CLOVES, 2 patients as hemihyperplasia-multiple lipomatosis (HHML) and 2 patients as macrodactyly. The entire patient list and tissue samples presenting with PROS are shown in supplementary table 1. Informed consent was obtained for the collection of all blood and tissue samples for genetic testing.

For the OES/ECCL study, three patients with OES (patient 1, 3, and 4) and one with ECCL (patient 2) were selected. For DNA extraction and genetic analysis, blood and different tissue samples were available from all the patients'. Informed consent was obtained for the collection of all blood and tissue samples for genetic testing. Parental written permission was obtained to publish the patients' photographs.

Patient 1: The propositus is the first child of healthy nonconsanguineous parents of German origin. He was born at term after an uneventful pregnancy via normal spontaneous delivery. His birth weight was 3100 g (−0.6 SD), his length was 51 cm (−0.2 SD), and his head circumference 34 cm (−0.4 SD). He did not experience perinatal adaptation problems. At birth, small skin tags of the right upper eyelid and the right eyebrow as well as an epibulbar dermoid on the right ocular surface, involving conjunctiva as well as cornea, were noted (Figure 4.1a,b). At clinical genetic evaluation he presented as an active, friendly boy with normal psychomotor development. He had a triangular face with prominent zygomatic arches and mild frontal bossing, a frontal upsweep, as well as deeply set eyes with a relatively narrow intercanthal distance. He had a hairless patch with immovable, yellowish, leathery

skin on his right frontoparietal scalp, which measured approximately 8 × 4 cm (Figure 4.1c). Histological examination of a biopsy of this lesion revealed a paucity of hair follicles without scarring and loss of elastic fibers in the trichrome-masson staining. It was noticeable that the arrector pili muscles were arranged in the midcorium in a line parallel to the surface (Figure 4.1d). Because of the presentation of the three cardinal features: fibroma of the eyelids/eyebrows, epibulbar dermoid and the typical scalp lesion in a boy with normal psychomotor development, a clinical diagnosis of OES was made.

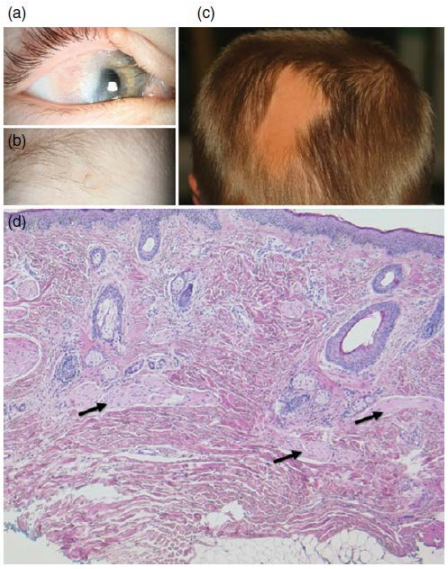


Figure 4.1: Patient 1 with clinical diagnosis of OES. (**a**, **b**) Right epibulbar dermoid, eyelid lesion and small skin tag of the eyebrow. (**c**) Parietal region of alopecia. (**d**) Histology from this region showing a linear arrangement of the arrector pili muscles in the midcorium (black arrows, HE-staining).

Patient 2: This male patient was the first child of non-consanguineous parents of Danish descent born at term with normal birth parameters and a relatively large head circumference of 37 cm (+1.5 SD). He presented with bilateral epibulbar dermoids, everted and hypoplastic upper eyelids with skin tags, and four well-circumscribed alopecic lesions on the upper right parietal scalp (Figure 4.2). Magnetic resonance imaging (MRI) of the brain performed at age 9 months showed hemiatrophy of left cerebral hemisphere and ventriculomegaly, but no definite evidence of intracranial lipomatosis. The boy had normal cognitive development, while his motor development was mildly delayed, probably because of visual impairment. He walked at 1.5–2 years and spoke his first words at 11 months. He has had no seizures or other neurological symptoms. A skeletal survey showed several benign cystic lesions in the long bones of all four extremities and clavicles (Fig. 4.2f). At the last clinical evaluation at the age 5 years, he showed mild growth failure (height: 100.7 cm, −2.4 SD; weight: 16.0 kg, −1.5 SD) with relative macrocephaly (head circumference: 53.5 cm, +1.5 SD). He had focal alopecia, streaks of pigmentary changes on the neck and both arms, and atopic dermatitis in several locations on the extremities and cheeks. No limb asymmetry was noted. Karyotype was normal, and array comparative genomic hybridisation (CGH) at 244 k resolution showed no

copy number variation. Based on these multiple abnormalities, the clinical diagnosis of ECCL was made according to published diagnostic criteria (Moog, 2009).

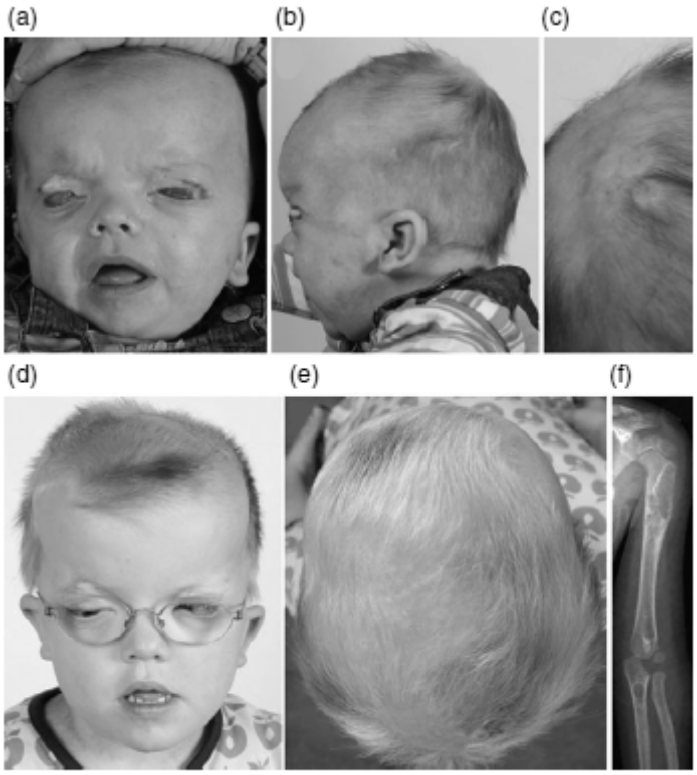


Figure 4.2: Patient 2 with clinical diagnosis of ECCL. (**a–c**) Clinical photos at the age of 10 months showing bilateral ocular dermoids, skin tags, hypoplasia of upper eyelids, and scalp lesions. (**d**, **e**) The same patient at age 5 years after ocular surgery. Note focal alopecia of the scalp. (**f**) X-ray image of the left upper limb with cystic lesions in the humerus and proximal ulna.

Patient 3: He is the first child of healthy non-consanguineous Turkish parents. His characteristic features supporting the diagnosis of OES included hairless areas in the parieto-occipital region of the scalp consistent with aplasia cutis congenital (ACC), a left-sided epibulbar dermoid, multiple areas of pigmentary changes following the lines of Blaschko mainly on the neck and trunk, a tumor of the lower jaw histologically classified as giant cell granuloma, and multiple cystic lesions of the right humerus and clavicle suggesting non-ossifying fibromas. Body length, weight and head circumference were in the normal ranges.

Patient 4: The clinical abnormalities of this 6 year old male patient included bilateral epibulbar dermoids, multiple areas of scalp alopecia, streaky areas of hypo and hyperpigmentation following the lines of Blaschko, intellectual disability, and attention deficit hyperactivity disorder. Brain MRI showed enlarged cisterna magna, an enlarged fluid space in the quadrigeminal cistern suggesting an arachnoid cyst, and several subcutaneous masses within the scalp. His anthropometric measurements were normal. Formally, he would also fulfill the criteria for a probable diagnosis of ECCL according to Moog, 2009 but classification as OES was preferred, as there was no intracranial or intraspinal lipoma.

4.2 DNA isolation and quantification

Genomic DNA was extracted from different affected tissues and blood using the QIAamp DNA mini kit (Qiagen, Hilden, Germany), according to the manufacturer's protocol. The

initial concentration and quality of the DNA was assessed using a NanoDrop ND-2000 spectrophotometer (NanoDrop technologies, Wilmington, USA) and 1% or 2% agarose gel electrophoresis respectively.

PIK3CA-related overgrowth study samples: For DNA extraction and genetic analysis, blood samples were available from all the patients'. Fresh or frozen affected tissue was available from the affected individuals who had undergone surgery or resection of the lipomatous overgrowth. Analysed tissues included epidermal nevi, skin, fatty tissue, bone and connective tissue. Formalin-fixed paraffin-embedded (FFPE) tissue blocks were also available for few patients. From this material thin serial sections each of 10 µm were generated. DNA was recovered from these sections using the BIOstic FFPE tissue DNA isolation kit (MO BIO laboratories, CA). Genomic DNA used in the experiments was diluted to a final concentration of 50 ng/µl.

OES/ECCL study samples: Blood samples were available from all the patients'. Different patient's tissue samples were used for DNA extraction and genetic analysis which included native tissue samples from the resected epibulbar dermoid (patient 1), native skin biopsy from the scalp lesion (patient 2), fibroblasts from a skin biopsy and FFPE tumor tissue from a giant cell granuloma of the mandible (patient 3), biopsies from a scalp lesion and a hyperpigmented area of the skin (patient 4). For patients 1 and 3, remaining tissue samples from surgical removal of epibulbar dermoids, and a mandibular giant cell granuloma, respectively, were used for the analysis. Patients 2, 3 and 4 had a 3 mm punch biopsy of the scalp lesion and from a hyperpigmented area of the skin, respectively, to obtain material for genetic testing. In addition, a cell culture was developed from skin fibroblasts of patient 2. For patient 3, only the fibroblast line but no native tissue was available from the skin biopsy. Genomic DNA was isolated from peripheral leukocytes, tissue samples, and cell cultures according to standard procedures.

RAS-GH panel study samples: All the patients tested for this panel had genomic DNA (gDNA) isolated from peripheral leukocytes. In NGS methodologies for generating quality data; a good quality genomic DNA template is crucial. Therefore, assessment of the gDNA integrity is important in the first step. To ensure appropriate DNA quality, all the samples for both ID and Short stature cohorts were quality checked by three different individual methods as described in figure 4.3 and only the DNA samples which passed through all the three methods were selected for inclusion in the study. For DNA quality check in all the experimental procedures, automated electrophoresis method was done on the Agilent TapeStation 2200 instrument using the genomic DNA ScreenTape assay and the procedure

was performed as instructed by the manufacturer and the data was analysed using the Agilent 2200 TapeStation software (Agilent Technologies, CA, USA). For accurate quantification of the DNA sample for use in NGS, two different fluorometric methods were used. Fluorometric methods were chosen for quantification as these assays are highly selective for double-stranded DNA (dsDNA) over RNA or common contaminants (such as salts, free nucleotides, solvents, detergents, or protein).

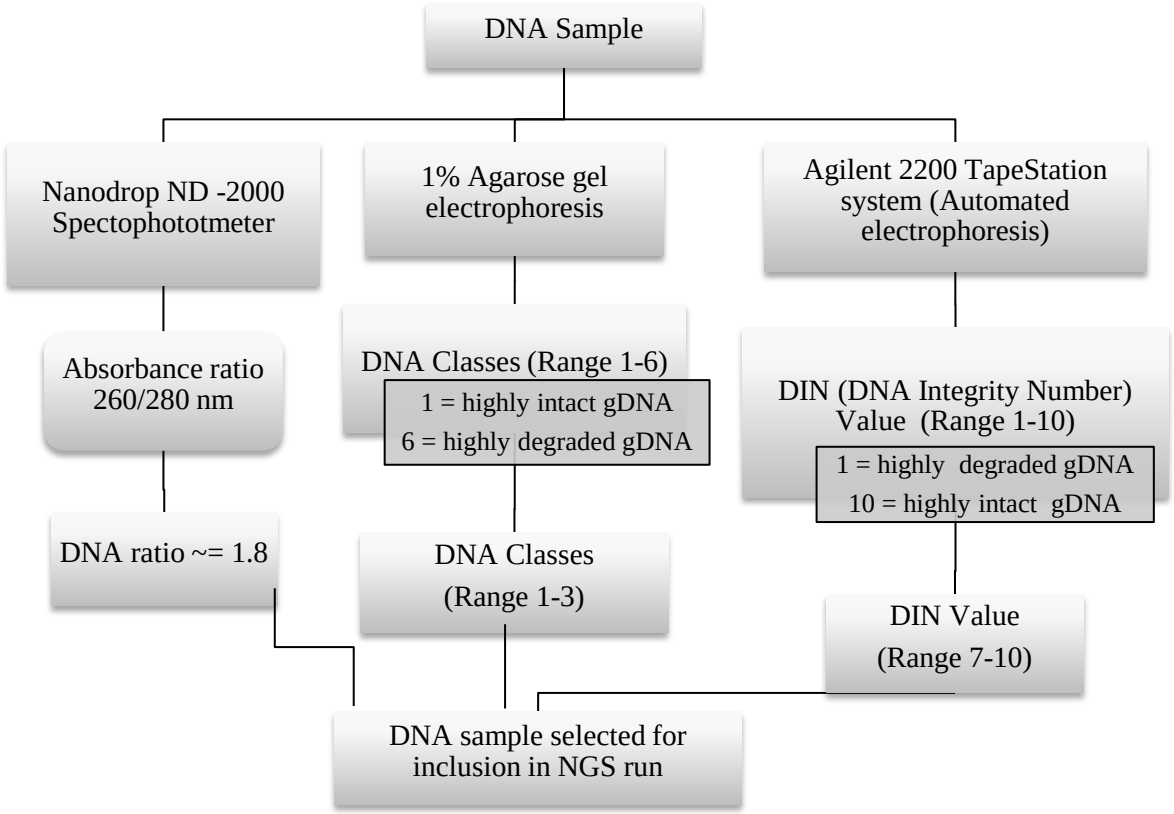


Figure 4.3: Different methods used in the current study for assessment of the quality of DNA

For 96 samples of the ID cohort, the DNA was quantified using the QuantiFluor® dsDNA dye on the QuantiFluor®-ST Fluorometer (Promega, USA). For the entire Short stature cohort and remaining 72 samples from ID cohort the Qubit® 3.0 Fluorometer (Invitrogen, Carlsbad, CA, USA) was used with the Qubit® dsDNA BR (Broad-Range) assay kit. All the samples were measured by taking triplicate readings and the procedures were performed according to the manufacturer's instructions. Detailed description of the fluorometric methods used in the current study is described in appendix I and II for reference. Genomic DNA used in the final experiments was diluted to a final concentration of 5 ng/µl (50 ng).

4.3 454 GS junior sequencing

For PIK3CA-related overgrowth study, next-generation pyrosequencing method was performed on a bench-top 454 GS Junior platform (Roche Diagnostics, Mannheim, Germany). The sequencing procedure was performed using the GS Junior titanium

sequencing Kit and the method described in the sequencing method manual, GS Junior titanium Series[1]. The pyrosequencing chemistry and a complete workflow of the 454 sequencing method are shown in supplementary figure 1.

Amplicon library preparation: 5 amplicons were generated with different sizes to cover mutation hotspots of the PIK3CA gene (Table 4.1). Two NGS runs were performed, due to insufficient data from the first run. From positive Sanger sequencing results, a total of 47 samples in Run I and 35 samples in Run II were sequenced by amplicon deep sequencing of the PIK3CA gene.

Table 4.1: Gene specific primer sequences (without Universal tag) - 5 Amplicons for the PIK3CA gene.

PIK3CA Exons	Gene specific forward primer	Gene specific reverse primer	PCR product size (bp)	Number of samples
Exon 2	ACTTTAGAATGCCTCCGTG	CGAAGGTATTGGTTTAGACAGAAA	349	8
Exon 5	CCTTTGCAGATTAATATGTAGTCATAA	CGGAGATTTGGATGTTCTCC	469	1
Exon 8	CCTTTTGGGGAAGAAAAGTG	GAGAGAAGGTTTGACTGCCATAA	284	4
Exon 10	TTGGTTCTTTCCTGTCTCTGAA	TTCCACAAATATCAATTTACAACCA	487	9
Exon 21	TGACATTTGAGCAAAGACCTG	ATGCTGTTCATGGATTGTGC	456	25

The samples were amplified by Universal Tailed Amplicon Sequencing (UTAS) method (Supplementary figure 1) using primers for the five coding exons (Table 4.1) with intron boundaries of the PIK3CA gene. The UTAS method is a two-step PCR strategy to attach sequencing adaptors and barcodes to the target-specific amplicons. In the first round PCR, the template specific sequences (defining the boundaries of the amplicons) were targeted by the gene specific primers fused to a universal sequence which would be the target for the second round primers. In order to maintain sequencing directionality, different universal tails were designed for the primer forward and reverse ends of the amplicons. The secondary PCR was performed in a volume of 25 µl according to manufacturer's instructions (“454 MID kit”, Multiplicom N.V., Niel, Belgium) with fusion primers targeting universal sequences containing the specific MID sequences and adaptors.

[1] GS Junior Titanium Series Amplicon Library Preparation Method Manual, Sequence emPCR Amplification Method Manual-Lib-A, and Sequencing Method Manual. Roche Diagnostics, Roche Applied Science, 68298 Mannheim, Germany.454 Sequencing 2012.

Amplicon purification and quantification: For Run I, the secondary PCR products obtained were purified twice using the Biomek® NXP laboratory automation workstation with 0.8 fold ratio of magnetic beads (Agencourt AMPure XP, Beckman Coulter, Krefeld, Germany) to the PCR quantity for removal of any remaining primers and adapters. For Run II, the secondary PCR products were purified manually or size selected with Solid Phase Reverse Immobilization (SPRI) magnetic bead method (SPRIselect reagent kit, Beckman Coulter, Krefeld, Germany) for efficient removal of primer dimers before emulsion. This amplicon purification was exclusively size selective limiting the presence of short or unspecific sequences in the run. For SPRI method, left side selection with ratio of 0.7x magnetic beads were used for a target selection of above 250 bp. Amplicons were then quantified and pooled by qPCR method on LightCycler® 480 (Roche Diagnostics, Mannheim, Germany) using Kapa library quantification kit (Kapa Biosystems, Wilmington, USA) as described by the manufacturer. On the basis of these measurements an equimolar pool was made with all the samples making one library. The library was further diluted to 10^5 molecules/µl by the same qPCR method mentioned. One more quality and quantity check of the library was made using the 2200 TapeStation instrument (Agilent Technologies, CA, USA). From the above dilution, 25 µl of DNA library having a concentration of 1 x 10^5 molecules/µl was used in emulsion PCR (emPCR) which gives an input of approximately 0.5 molecules of library DNA per Capture Bead (Cpb). The emPCR, corresponding to clonal amplification of the purified amplicon pool, was carried out using the 454 GS Junior Titanium Series Lib-A emPCR kit and was performed as instructed in the manual provided by Roche - emPCR amplification method manual-Lib-A[1].

Data analyses: The GS Amplicon Variant Analyzer (AVA) software (version 2.7; Roche Diagnostics, Mannheim, Germany). All nucleotide variations were identified by AVA software with read counts and frequencies. All variant calls made by the software were examined and confirmed (true or artifact) by reviewing individual flowgrams[2].

4.4 Sanger sequencing

All coding exons with flanking introns of PIK3CA gene (NM_006218.2), Exon 10 of PIK3R1 (NM_181523.2) and PIK3R2 (NM_005027.3), and exon 4 of AKT1 (NM_005163.2) were sequenced for the PIK3CA-related overgrowth study, and all the exons of KRAS gene (NM_004985) for OES/ECCL study. All amplicons were amplified by PCR in a 20 µl reaction volume from the isolated genomic DNA from various tissue sources and blood.

[2] A read from 454 sequencer is natively represented as a flowgram, which is a sequence of pairs of a nucleotide and its (fractional) intensity. The intensity is, in principle, proportional to the number of bases incorporated.

Direct bidirectional Sanger sequencing was performed using ABI 3500XL genetic analyser with Big Dye Terminator Cycle Sequencing Kit (PE Applied Biosystems, Foster City, Calif., USA). Sequences were aligned using the Seqpilot analysis software (JSI medical systems, Kippenheim, Germany) and compared with the reference sequences (PIK3CA; ENST00000263967, PIK3R1; ENST00000521381, PIK3R2; ENST00000222254, AKT1; ENST00000402615, KRASB; ENST00000311936) for genomic DNA and mRNA. Mosaic level for mutations were estimated by comparing the area under the curve of electropherograms for the wildtype and mutant peaks in the forward and reverse sequencing directions using the Seqpilot software.

The filtered potentially pathogenic variants identified in genes of the RAS-GH custom panel were also confirmed by conventional Sanger sequencing. The single exons of respective genes were amplified by PCR in a 20 µl reaction volume using primer pairs designed from a primer design software tool, Primer3 (http://bioinfo.ut.ee/primer3-0.4.0/). Further, segregation analysis was also performed for the particular candidate variations depending on the availability of parent/relative DNA samples.

4.5 Fragment analysis

A fragment analysis was developed and performed further to check for the amplification of each amplicon and presence of any short fragments for samples used in GS junior sequencing. Fragment analysis was also used to detect the presence of deletions in the samples tested for PIK3CA-related overgrowth study. Each sample was checked for presence of short fragments and if present then an additional purification was performed as mentioned before each run. This step was performed as a quality measure to remove short fragments. The secondary PCR products from amplicon library preparation (with the fusion primers) were used in this analysis. The secondary PCR products were further labelled with carboxylfluorescein (FAM) at the 5′-terminal end of the upper primer provided in the 454 MID kit (Multiplicom N.V., Niel, Belgium) targeting the adaptors. After the singleplex labelling PCR, the size standard mix (0.3 µl GS600 in 10 µl Hi Di Formamide) was added to 2 µl of PCR product and denatured at 95°C for 3 min, followed by immediate chilling. The products were electrophoresed on the ABI 3500XL genetic analyzer and later analysed with the GeneMapper® software v4.1 (Applied Biosystems, Foster City, Calif., USA).

4.6 Target selection: Custom-designed gene panel

In the current project instead of sequencing entire exomes, targeted resequencing approach was used for comparative analysis of candidate genes or regions and for obtaining a high level of accuracy to identify low frequency SNPs and structural variants. In the selection process,

already known RASopathy genes (PTPN11, RAF1, SOS1, KRAS, BRAF, MAP2K1, HRAS, NF1, SPRED1 etc..,), genes linked to ID and in relation to RAS pathway (SYNGAP1, RHEB, RIT1, GRIN2B, YWHAE etc..,) and Short stature genes (GHR, GRB2, IGF1, IGF1R, INSR, IRS2, JAK2, SHOX, SOX2 etc..,) were included in the panel. Also in the current panel, STAT family of transcription factors and PI3K pathway related genes were also included.

Genes for targeted resequencing which might be the possible candidates for monogenic forms of non-syndromic ID/Short stature were selected based on the following criteria.

1. Linkage to RAS pathway or growth hormone (GH) pathway - The known RASopathy, ID and Short stature genes served as basis to find other linked genes, such that known genes were connected by the other genes either as direct interactors or interactors that require one linker.
2. Expression data- mainly in brain for RAS pathway genes and bone/cartilage for Short stature pathway genes.
3. Evidence of known phenotypes/syndromes/disorders in Humans, monogenic or complex forms inferred from other studies related to ID and Short stature.
4. Implicated in ID and linked to short stature or either of one for both RAS and Short stature genes.
5. Animal model with Behaviour/neurological phenotype, shown in learning and memory, growth/size/body, mortality/aging, nervous/skeletal system development/ abnormalities and abnormal craniofacial features.
6. Shown functional aspects in vitro
    a. RAS pathway- affecting synaptic plasticity and/or neurotransmission, synaptic regulation, synaptic structure and/or number and NMDA/AMPA/GABA receptor trafficking or surface expression.
    b. Short stature pathway- affecting skeletal system development/morphogenesis, bone/bone marrow development, cartilage development/condensation, and regulation of cell growth.

From the above mentioned criteria a complete list of targeted genes was prepared which are mainly linked to both RAS and GH pathway. A total of 329 genes were selected, out of which 221 genes belong to RAS pathway/RAS related pathway and 108 genes belong to short stature related pathway with a total size of the target sequence spanning 777119 bp (~777.12 kb). The full regions of the targeted genes were considered for the study. Supplementary table 2 includes the total list of 329 genes studied in this project. A table with all the details of the selected genes like name, locus, family, evidence supporting various parameters, expression

data etc., has been listed in supplementary table 3 and supplementary table 4 for RAS and Short stature related pathways, respectively. All information/data regarding each mentioned criteria on which the genes were selected was collected using various online resources outlined in supplementary table 5. A priority list is made with the selected genes by giving scores to the following mentioned criteria in Table 4.2.

Table 4.2: Scores given to each category for prioritizing of the selected genes

Category	Score
Involved in ID or linked to short stature	3
Directly linked to RAS/GH signalling	2
Involved in learning and memory and/or phenotype in animal model	2
Shown functional aspects *in vitro*	2
Indirectly linked to RAS/GH signalling	1
Predominant brain or bone/cartilage expression	2

The priority list scoring given to the above mentioned categories is explained in detail in Table 4.3. A maximum of score 11 could be obtained if all the above mentioned features are displayed and a minimum of score null if none of the featured criteria are present. The RAS pathway genes all had a score of between 11-3 and for Short stature pathway genes between 11-0.

A network of genes was built in Cytoscape software (Shannon P et al., 2003) with all the selected genes and their interaction partners using data obtained from the STRING database version 10 (Szklarczyk et al, 2015) and IntAct Molecular Interaction Database (Orchard S et al., 2013). Gene symbols/names were used as input to let the database search tool determine mutual interactions among the input genes. This query was performed using human as organism with confidence score of 0.4 (medium confidence) between the target and its interaction partners and otherwise default settings. The database output files were reformatted for input and better visualization in Cytoscape version 3.3. The complete interaction network for all the selected genes for both the RAS and Short stature pathways are shown in Figure 4.4 and 4.5 respectively. For easy identification and representation of the different categories of each gene in the network, the following parameters were taken into account.

1. Colour coding was used to differentiate between genes which are already implicated in ID and short stature and
2. Different shapes are assigned to predominant brain/bone or various tissue expression
3. Genes which could not be mapped in the pathway are mentioned in separate special columns in the figures.

Table 4.3: Explanation for scoring each category mentioned in the priority list.

		RAS pathway genes	**Short stature pathway genes**
1	Involved in ID or in short stature	Score 3 is given, if the gene is directly involved in human ID (Known mutations, human ID syndromes etc.). Score 1 is given if the gene is shown involved in chromosome deletion regions with ID (genes within range of del region), or GWAS, Array studies etc., or other behavior anomalies (ADHD, depression, Autism, Schizophrenia, Alzheimer's etc.) in human ID	Score 3 is given if the gene is directly involved in human short stature (Known SNP's, deletions, GH deficiency, pituitary hormone deficiency, syndromes etc.). Score 2 is given to genes linked to short stature in animals (mostly mice), Score 1 to genes not directly affected in (human /organism model) short stature and score null with no relation to short stature.
2	Linkage to pathway directly	Score 2 is given to genes directly linked to HRAS (First neighbours) and genes linked to at least the first neighbours of HRAS (Second neighbours)	Score 2 is given to genes directly linked to GHR; JAK; IGF's; IRS's; SHC1 (First neighbours) and genes linked to HRAS along with its first neighbours.
3	Functional aspects *in vitro* (Score 2 is given if any of the mentioned functional aspect is shown)	affecting synaptic plasticity, and/or neurotransmission, regulation, synaptic structure and/or number and NMDA/AMPA/GABA receptor trafficking or surface expression	affecting skeletal system development/morphogenesis, bone/ bone marrow development, cartilage development/ condensation, and regulation of cell growth.
4	Animal Model (Score 2 is given if more than 2 features are present and Score 1 if less than 2)	Behaviour/neurological phenotype, shown in learning and memory, growth/size/body, mortality/aging, nervous system development/ abnormalities.	Skeletal system development/ abnormalities, growth/size/body, abnormal craniofacial features, nervous system abnormalities
5	Linkage to pathway indirectly	Score 1 is given to genes which are linked to at least the Second neighbours of HRAS.	Score 1 is given to genes which are linked to at least the Second neighbours of above.
6	Expression	Score 2 is given for genes showing predominant brain expression. Score 1 is given if genes are expressed in brain either moderately or low. Null score for no expression of genes in nervous system.	Score 2 is given for genes showing predominant bone/cartilage expression. Score 1 is given if genes are expressed in bone/ cartilage either moderately or low. Null score for no expression of genes in skeletal system.

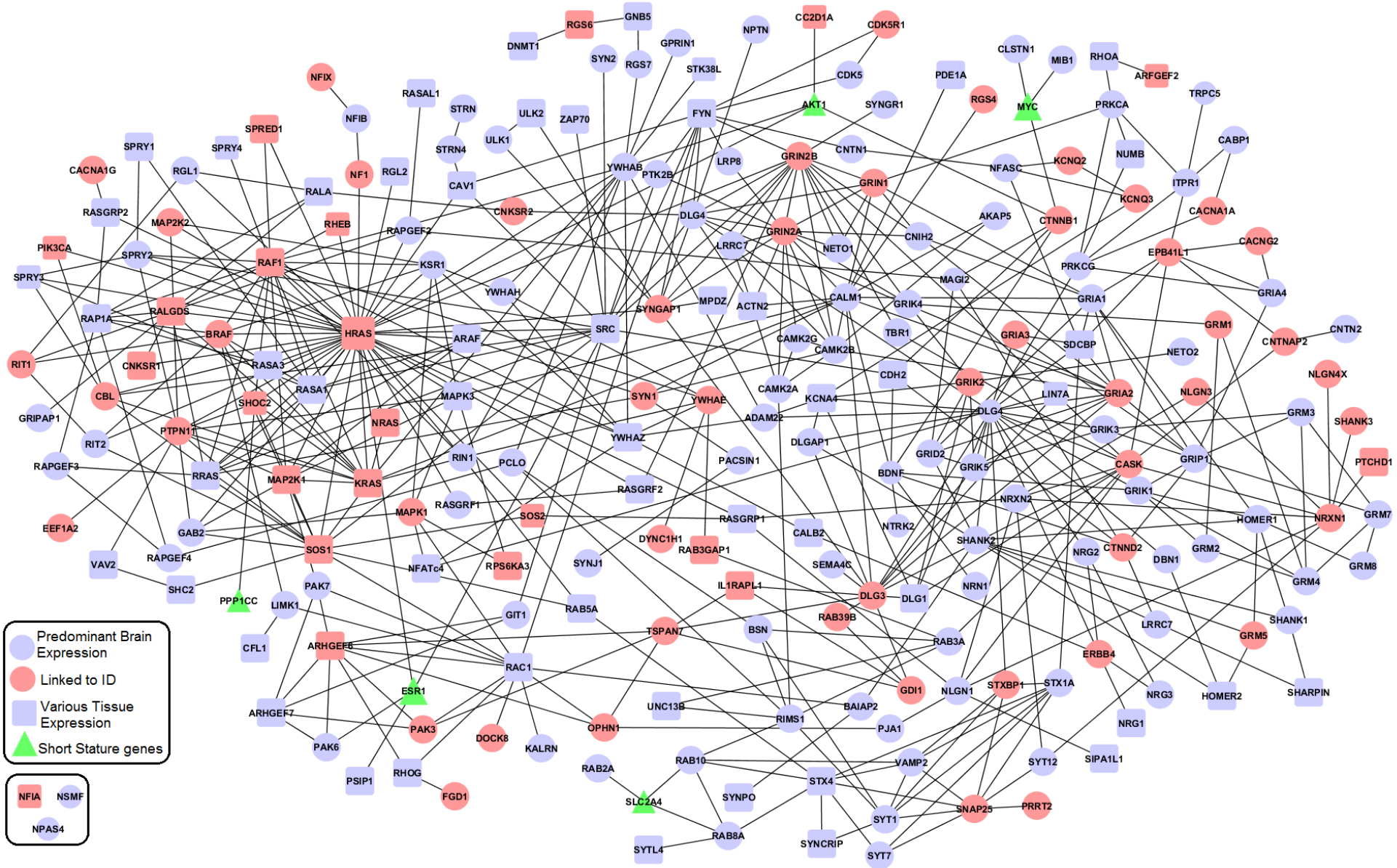


Figure 4.4: Network of genes related to RAS/MAPK pathway and their interaction partners to be studied in the project (prepared by using Cytoscape software, version 3.3) based on STRING interactions (Version 10). Lines indicate at least moderate evidence (STRING score > 0.3) for protein-protein interactions in humans. A total of 221 genes were selected which belong to RAS/RAS related pathway.

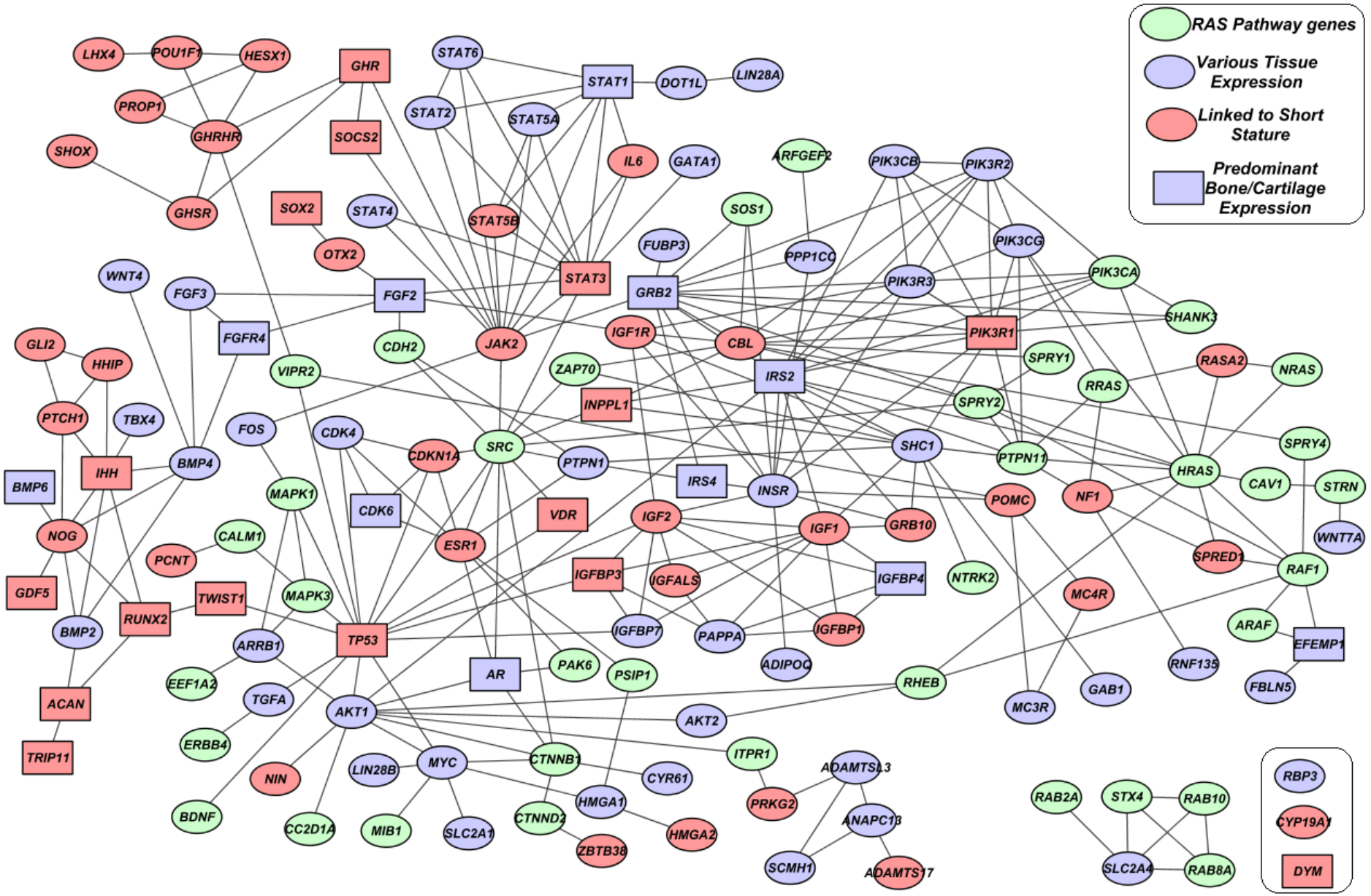


Figure 4.5: Network of genes related to GH-PI3K-JAK-STAT pathway and their interaction partners to be studied in the project (prepared by using Cytoscape software, version 3.3) based on STRING interactions (Version 10). Lines indicate at least moderate evidence (STRING score > 0.3) for protein-protein interactions in humans. A total of 108 genes were selected which belong to GH-PI3K-JAK-STAT related pathway.

4.7 Target enrichment sequencing

The fore mentioned work done on the GS junior provided an insight into the methodology of next generation sequencing. From this knowledge, in the current gene panel project more efficient NGS technology was used because of the increased size of target sequence, cost effectiveness and for high sensitive detection of molecular mutations. For deep sequencing of target genomic regions of interest and for high throughput of the Illumina System, commercially available solution-based hybridization enrichment approach was opted for targeting genomic sequence in this study. In the current project, Nextera® rapid capture custom enrichment assay was used and the cluster generation/sequencing was performed on the Illumina Miseq® desktop sequencer (Illumina, San Diego, CA, USA).

4.7.1 Probe design for selected genes

The hybridization probes for the selected genes from the above mentioned criteria were designed using the DesignStudio® software (Illumina, San Diego, CA, USA) which is a freely available web based tool (designstudio.illumina.com). In this software the target genes were added either by entering the HGNC names or defined target region coordinates. Depending on the probe interval selected (standard/dense) and targeting option (Exons/Full regions/CDS), the software determined the number of probes required to cover the target based on spacing, specificity, and GC Content and also checks for potential cross-binding of probes and duplicate assays. The 329 selected genes with dense probe spacing (120 bp center to center spacing for adjacent 80-mer probes) and covering full regions were given as input for the DesignStudio® software resulting in total captured area of ~778 kb at 100% probe coverage. The overall design panel resulted in 6879 target probes with 3% overlap and average region size of 141 bp (Supplementary figure 2). After the completion of the design panel with target regions, a detailed report of the panel is available in the DesignStudio® software showing the uniformity and specificity of the panel. Once the custom-design panel was completed, the manifest file was downloaded from the DesignStudio® software from the project dashboard. For defining regions of interest like the genomic coordinates (for a given target in the custom design) by the analysis software, a manifest file is required for the enrichment workflow. The manifest file contains the target and probes sequences and is used during the alignment of the sequenced reads.

4.7.2 Nextera® library preparation and quantification

In the current study, Nextera® Rapid Capture Enrichment (NRCE) method was used for the library preparation. This method uses an enzymatic DNA fragmentation step hence it is more

sensitive to the DNA input amount and also the enrichment strongly depends on accurately quantified starting material amount. All gDNA quantifications were done using the fluorometric methods and the procedure was performed as instructed by the manufacturer. After initial quantifications, the DNA was diluted further to 10 ng/µl in Tris-HCl (10 mM, pH 8.5) in 20 µl volume. Samples were then re-quantified using the similar fluorometric-based methods. Samples were measured in triplicates and an average was taken. Based on the quantification, gDNA samples were further diluted in Tris-HCl 10 mM, pH 8.5 to a final volume of 10 µl at 5 ng/µl (50 ng total). A total of 50 ng genomic DNA was used as input material for NRCE method with two-step gDNA normalization.

The enrichment procedure was performed using the Nextera® Rapid Capture custom enrichment Kit and the method described in the reference guide[3] as instructed by the manufacturer (Illumina Inc, CA, USA).

- In the first step, the genomic DNA was enzymatically fragmented by Nextera® transposome. The transposome cuts the gDNA randomly and simultaneously adds adapter sequences at the ends, known as "tagmentation", producing amplifiable library molecules. Post "tagmentation" a clean-up procedure was done for removal of excessive/unbound Nextera® transposome as it can bind to the ends of the DNA tightly and can interfere with subsequent downstream process. The adapters (for cluster generation and sequencing) and indexes (for sequencing) were added to the fragmented DNA by PCR amplification followed by PCR clean-up with magnetic beads to remove short fragments and unwanted products.
- As a quality measure to check the size distribution of the DNA fragments, each DNA sample was checked on the Agilent TapeStation 2200 using the D1000 ScreenTape system and the procedure was performed as instructed by the manufacturer (Agilent Technologies, CA, USA).
- In the second step, pooling of the libraries with different indexes was done resulting in sub-libraries. In the current study for each run of 24 samples, two enrichment reactions were done each with 12-plex pool complexity and with a total DNA library mass of 6000 ng for each enrichment reaction (500 ng of each DNA library). The libraries were quantified by fluorometric methods and the procedure was performed as instructed by the manufacturer. These two sub-libraries were then hybridized with custom-designed biotinylated capture probes to target regions of interest. The

[3] Nextera Rapid Capture Enrichment Reference Guide (January 2016). Document # 15037436 v01. Illumina, San Diego, California 92122 U.S.A.

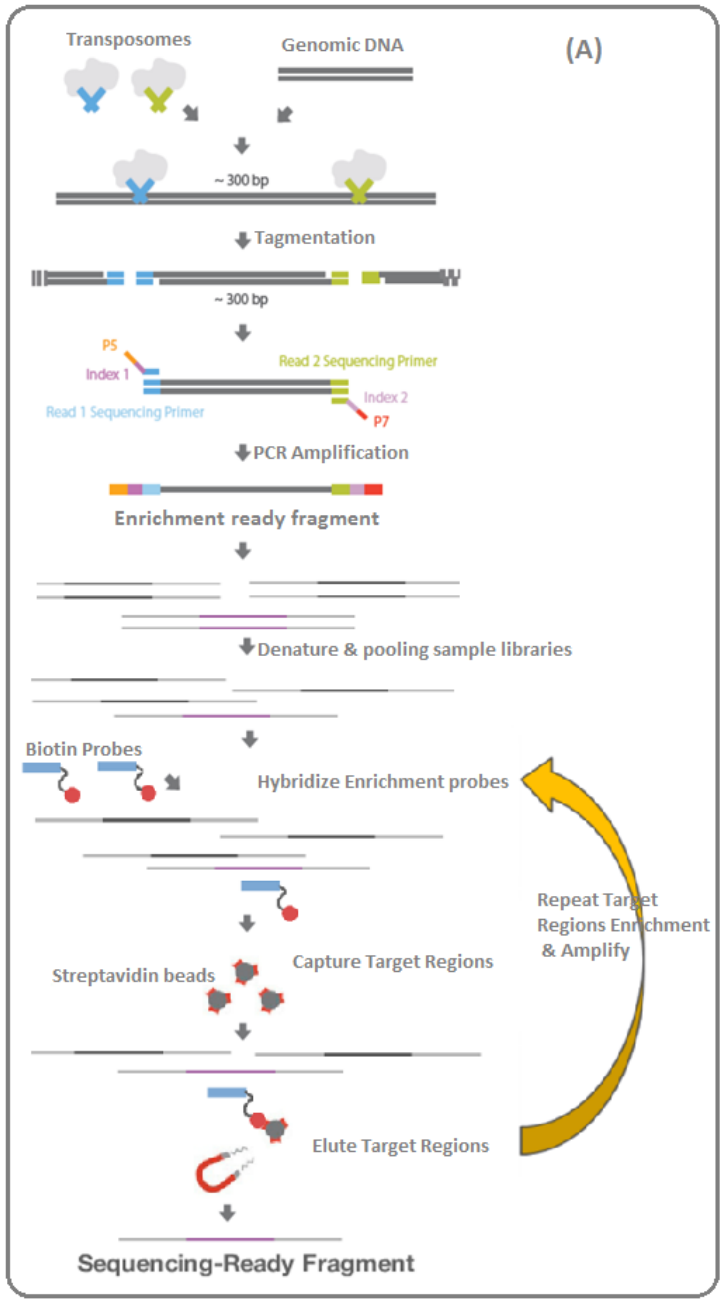
hybridized probes were then captured using streptavidin magnetic beads followed by elution of the enriched library from the beads.

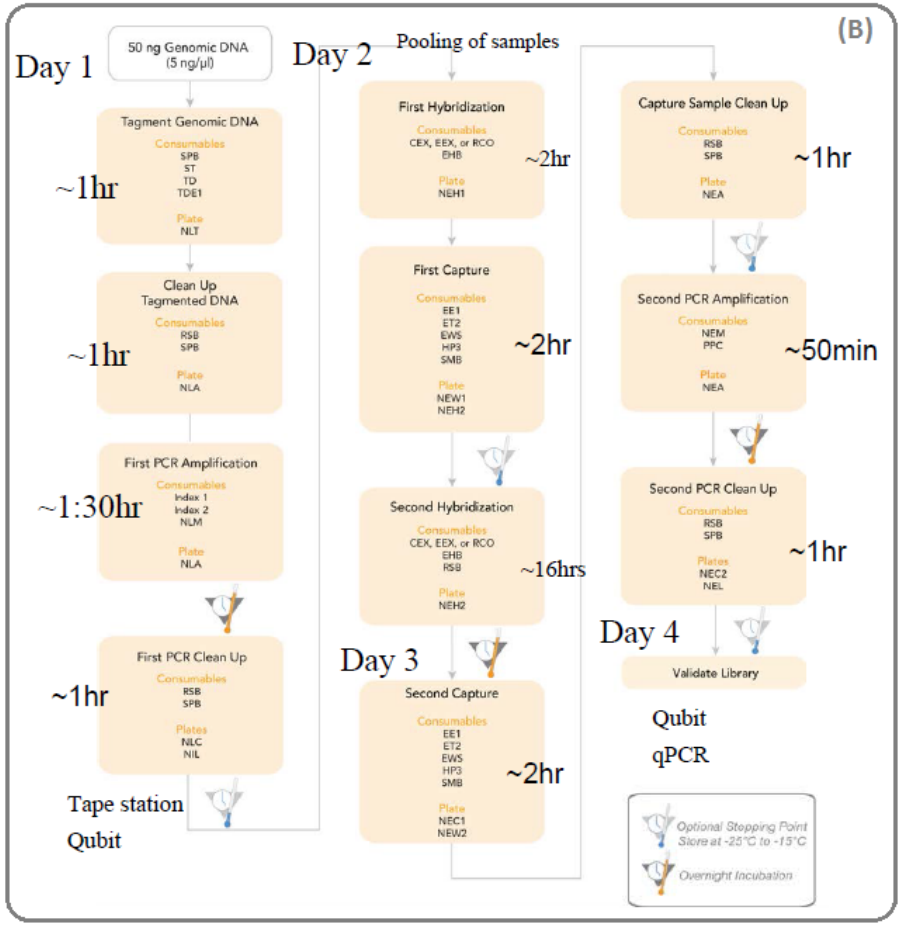
- In the third step, a second round of hybridization, capture and elution were performed on the eluted DNA to ensure high specificity of the captured regions. A capture sample clean-up was performed to remove unwanted products. A second PCR amplification of the enriched libraries was done followed by a final clean-up procedure. At the end of this protocol, enriched indexed DNA library ready for sequencing was generated.

The complete library preparation workflow of the Rapid capture method is shown in Figure 4.6 and the index sequences used in the current study is shown in appendix table III.

Libraries validation: For obtaining highest quality data, optimum cluster densities need to be present across each lane in the flow cell which is achieved through accurate quantification of the post-enriched library. In the current study, the DNA libraries were quantified by qPCR method on LightCycler480 (Roche Diagnostics, Mannheim, Germany) using Kapa library quantification kit for Illumina sequencing platforms (Kapa Biosystems, Wilmington, USA) as described by the manufacturer. Additionally, the libraries were also quantified by two fluorometric methods (Promega QuantiFluor® and Qubit® dsDNA HS (High sensitivity) assay) and the procedure was performed as instructed by the manufacturer. The quality control analysis to check the size of the libraries for distribution of DNA fragments was done on the TapeStation using the D1000 ScreenTape system and the procedure was performed as instructed by the manufacturer (Agilent Technologies, CA, USA).

Reduced volume Nextera® library preparation: In the current study, a total of 12 runs were performed with 24 samples per run, out of which 8 runs were performed with the standard enrichment protocol (4 each of ID and Short stature cohort) as instructed by the manufacturer. In order to increase cost efficiency and the number of samples analysed with this panel, the remaining 4 runs (3 ID and 1 Short stature cohort) were performed with half reagent protocol, built in-house with a few modifications to the standard protocol. All the modifications were tested initially on five different DNA samples changing various parameters like DNA concentration/ volume; incubation time/temperature and modifications in clean up procedures of the normal protocol as shown in table 4.4. Modifications were done only until first hybridization step and further steps involving capture, binding and elution was followed as mentioned in the normal protocol without any changes. All essential quality checks were done to prior implementing them in the actual runs. The final modifications applied to the runs are shown in Figure 4.7 in comparison to the standard protocol.





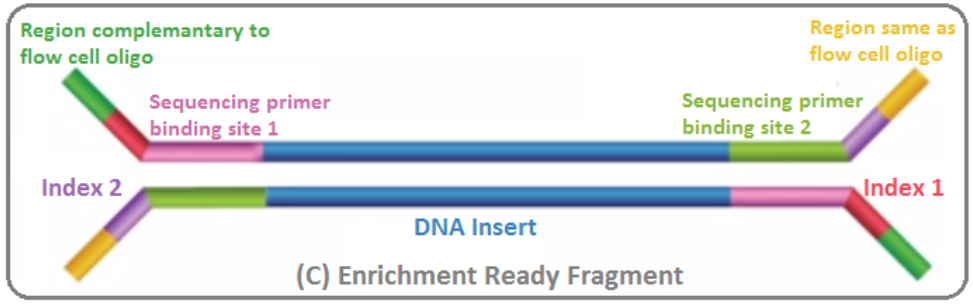


Figure 4.6: Brief overview of the Nextera® Rapid Capture enrichment method.

(A) Enrichment assay – sample preparation
(B) Library preparation workflow
(C) The sequencing ready fragment with primers and indexes (Figures are reproduced with the permission from Illumina and remain their copyright, Courtesy of Illumina, Inc.)

Table 4.4: Table showing the different modifications implemented to the Nextera® workflow in the current study.

Reagents	Standard Protocol	Trial A	Trial B	Trial C
DNA tagmentation step:				
DNA Conc (ng)	50	50	25	25
DNA – Sample volume (µl)	10	10	5	5
Tagment DNA Buffer (µl)	25	12.5	12.5	12.5
Tagment DNA enzyme (µl)	15	7.5	7.5	7.5
Stop Tagment Buffer (µl)	15	7.5	7.5	7.5
Incubation Temp/Time	58°C/10min	58°C/20min	58°C/10min	58°C/10min
Clean up Tagmented DNA:				
Magnetic Beads (µl)	65	65	32.5	65
Alcohol (µl)	200	200	100	200
Resuspension buffer (µl)	22.5	22.5	11.5	22.5
First PCR amplification step:				
Index adapters (µl)	5	2.5	2.5	2.5
Library amplification mix (µl)	20	10	10	10
Tagmented DNA volume (µl)	20	10	10	10
PCR cycles (x)	10	12	14	14

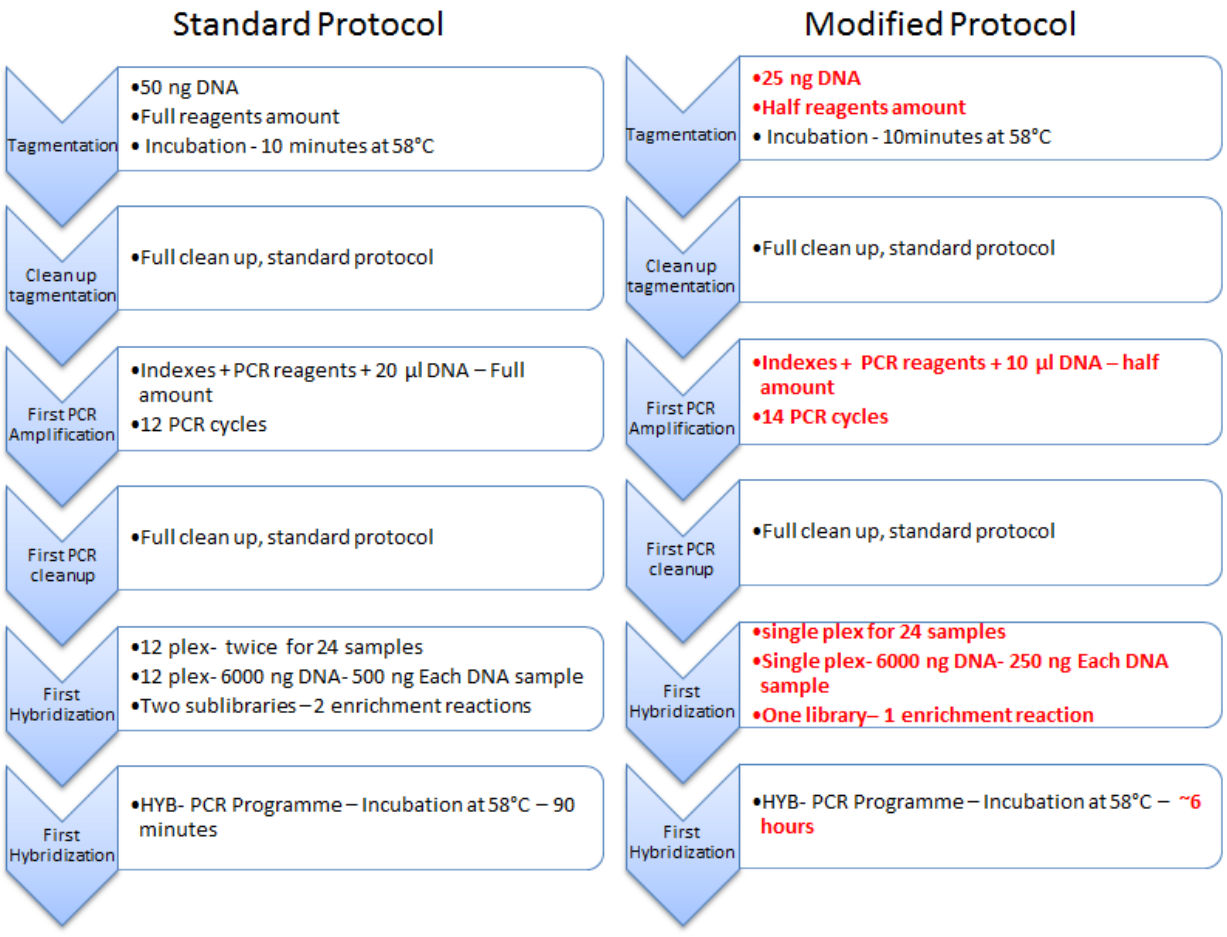


Figure 4.7: Comparison and modifications done between the standard protocol and in-house built modified protocol of the Nextera® amplification enrichment protocol. The text in red colour indicates the modifications done to the normal protocol.

4.7.3 Cluster generation/sequencing

The libraries were further prepared for subsequent cluster generation and DNA sequencing. The flowchart showing the quantitation and preparation of libraries for sequencing are mentioned in-detail in Figure 4.8.

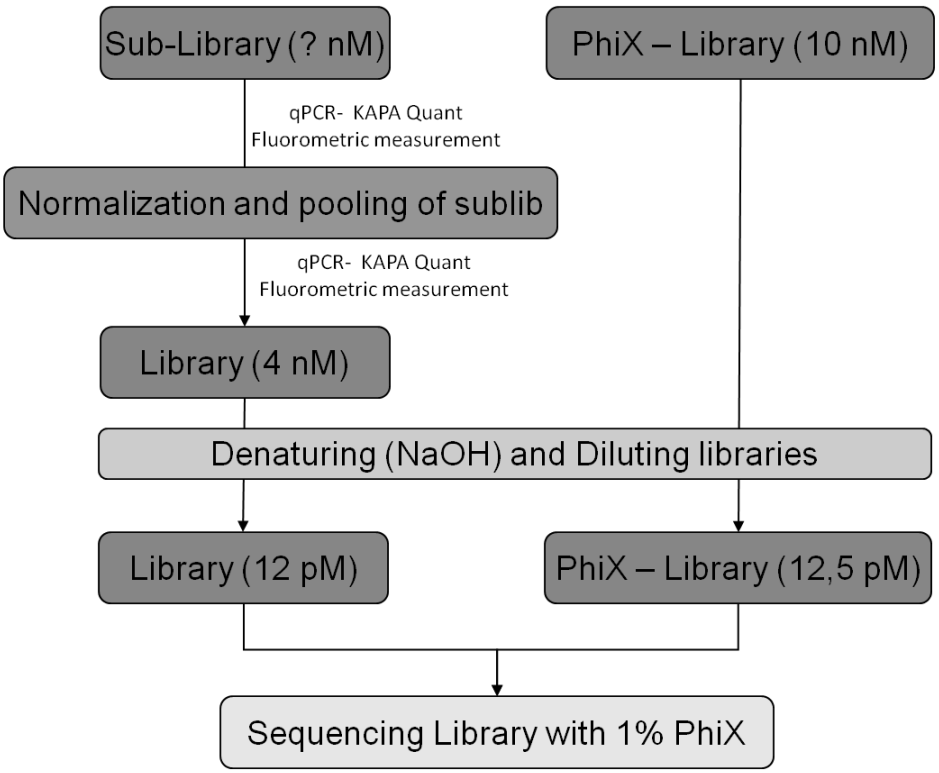


Figure 4.8: Flowchart showing the preparation of libraries for sequencing on the Miseq® instrument

In this current study, all the sequencing libraries were denatured and diluted to final amount of 12 pM except for one run where 15 pM was used for cluster generation. An increased concentration for one run was used since the final library showed low library concertation and low peak distribution size than recommended. The final sequencing library was spiked-in with 1% PhiX control (12.5 pM) for all the runs. During cluster generation, the single DNA molecules were bound to the flow cell surface and thereafter bridge amplification occurs to form clusters (Figure 4.9A).

The sample sheet with indexes was created using the Illumina experiment manger (IEM) software on the Miseq® instrument. For the current study, targeted resequencing category with enrichment application was selected in the IEM software. The final library mix was then loaded into the reagent cartridge in the provided reservoir given by the manufacturer. The sequencing run was set up using the Miseq® controller software (MCS) on the Miseq® instrument. Prior to loading the flow cell into the instrument, it was cleaned thoroughly with laboratory grade water and 70% alcohol wipes to remove excess salts, pat dried and visually

inspected that the flow cell was clear from any kind of obstructions. The flow cell was then loaded on to the Miseq® instrument followed by loading of the reagent cartridge and reagent buffer bottle for sequencing on the Miseq® instrument. The reagent buffer bottle, flow cell and reagent cartridge were provided with an RFID (Radio Frequency Identification), which is read by the MCS automatically thereby configuring their barcodes.

In the current study, 2 × 150 paired-end reads with dual indexing were generated using the Miseq® reagent kits v2, 300cycles (pre-filled, ready-to-use reagent cartridges). Paired end sequencing was performed through this technique, which means sequencing of the DNA fragments in a library done on both ends and the forward and reverse reads are aligned as read pairs. With the read pair alignment, a more accurate read alignment was achieved with a higher ability to detect indels and SNV calls. For enabling rapid and accurate sequencing, clonal amplification and SBS (Sequencing by synthesis) chemistry was done. The cluster amplification/generation and sequencing process is explained in detail in figure 4.9.

Since a single library template molecule would not generate enough signals for the system to detect, an amplification of the molecules is done on a solid surface, called the “flow cell”. After library preparation, the template fragments are flooded across the flow cell, which allows them to bind to the surface with the help of the custom adapters. The unbound fragments are then washed away. The bound fragments are clonally amplified on the surface of the flow cell by “bridge amplification” PCR process (cluster generation) generating millions of copies of each template. These tight physical clusters on the flow cell surface emit strong signals which are then detected by the system for image analysis. In SBS technology throughout each sequencing cycle, a single modified fluorescently labelled dNTP is added to the nucleic acid chain. The dNTP label serves as “reversible terminators” blocking further polymerization. The four dNTPs are added to the templates in every step followed by fluorescence recording, dNTPs are then enzymatically cleaved enabling the incorporation of the next nucleotide.

Lasers are passed over the flowcell to activate the fluorescent label on the nucleotide base. This fluorescence is detected by a camera and recorded. Each of the terminator bases (A, C, G and T) give off a different colour. Since only a single base is added each time, incorporation bias is minimized and the final result is base-by-base sequencing. Through this a high accuracy is achieved minimizing raw error rates, eliminating sequence context errors and enabling robust base calling within repetitive sequence regions and homopolymers.

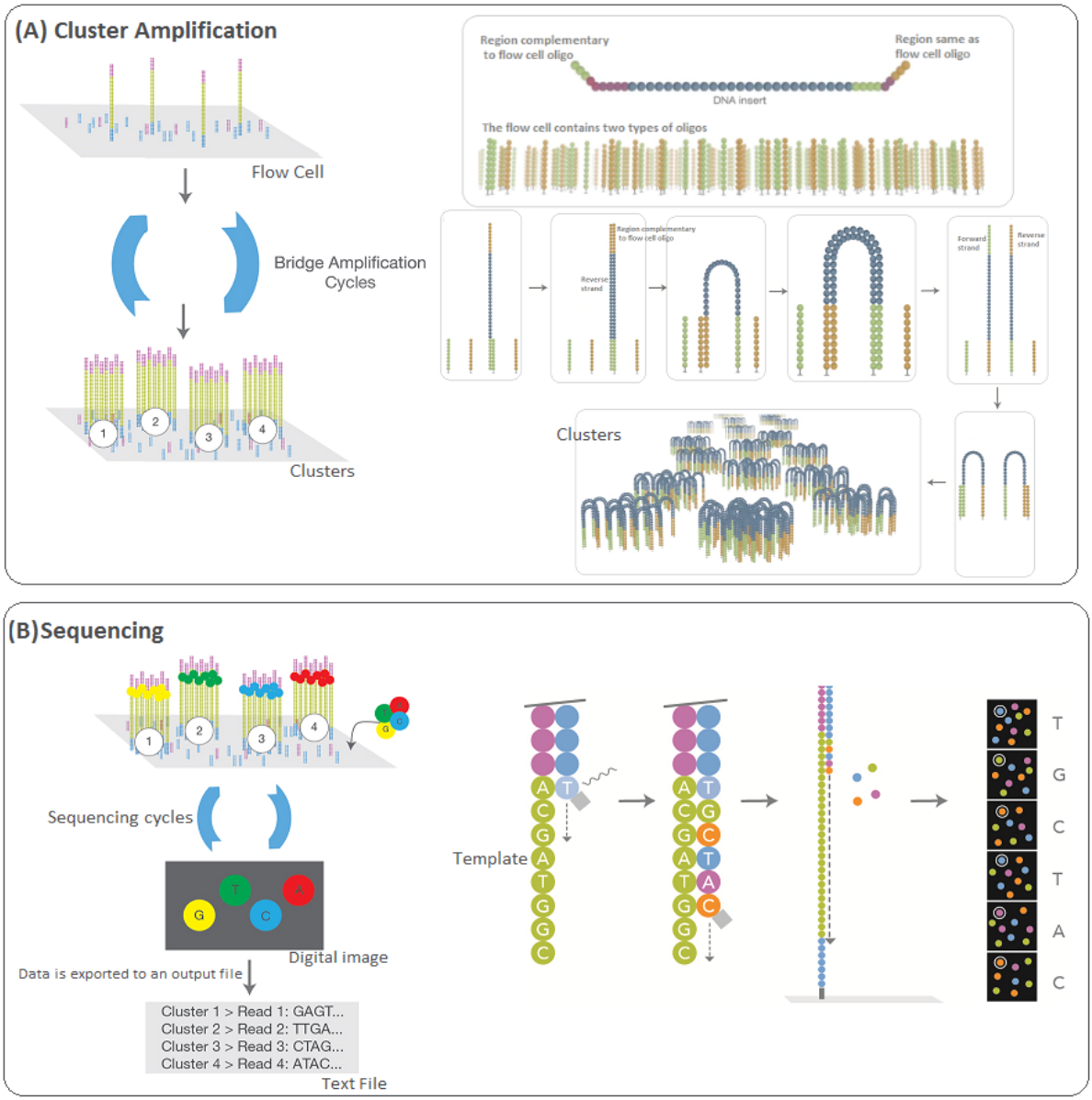


Figure 4.9: SBS (Sequencing by synthesis) chemistry overview. (A) Cluster Amplification: Libraries are loaded onto the flow cell which consists of the oligo's on the surface and the fragments are hybridized to the flow cell surface. By bridge amplification cycles, each fragment forms a clonal cluster on the flow cell. (B) Sequencing reaction: Along with sequencing reagents, fluorescently labelled nucleotides are added which results in incorporation of the first base. The flow cell is imaged and recordings of the emission from each cluster are done. To identify a base, the emission wavelength and intensity are used. This cycle is repeated 'n' times to create a read length of 'n' bases. [Figures for SBS chemistry are reproduced with the permission from Illumina and remain their copyright, Courtesy of Illumina, Inc.]

4.7.4 Data analysis and review

A brief overview of the data analysis is shown in figure 4.10 with the software's used and the outputs generated. Real-time analysis (RTA) uses early cycles of the run for template generation[4]. After the template of cluster positions is generated, images produced over every

[4] Template generation is the process by which cluster positions over the entire flow cell surface are defined according to X and Y coordinates position

subsequent cycle of imaging are aligned against the template. Individual cluster intensities in all 4 nucleotide colour channels are extracted and base calls are produced from the normalized cluster intensities. In the primary analysis, MCS controls the flow cell stage, fluidics system, and flow cell temperatures, and captures images of clusters on the flow cell. MCS also provides an overview of quality statistics, which can be monitored during the run. The MCS screen was view-only and no changes could be done during the run.

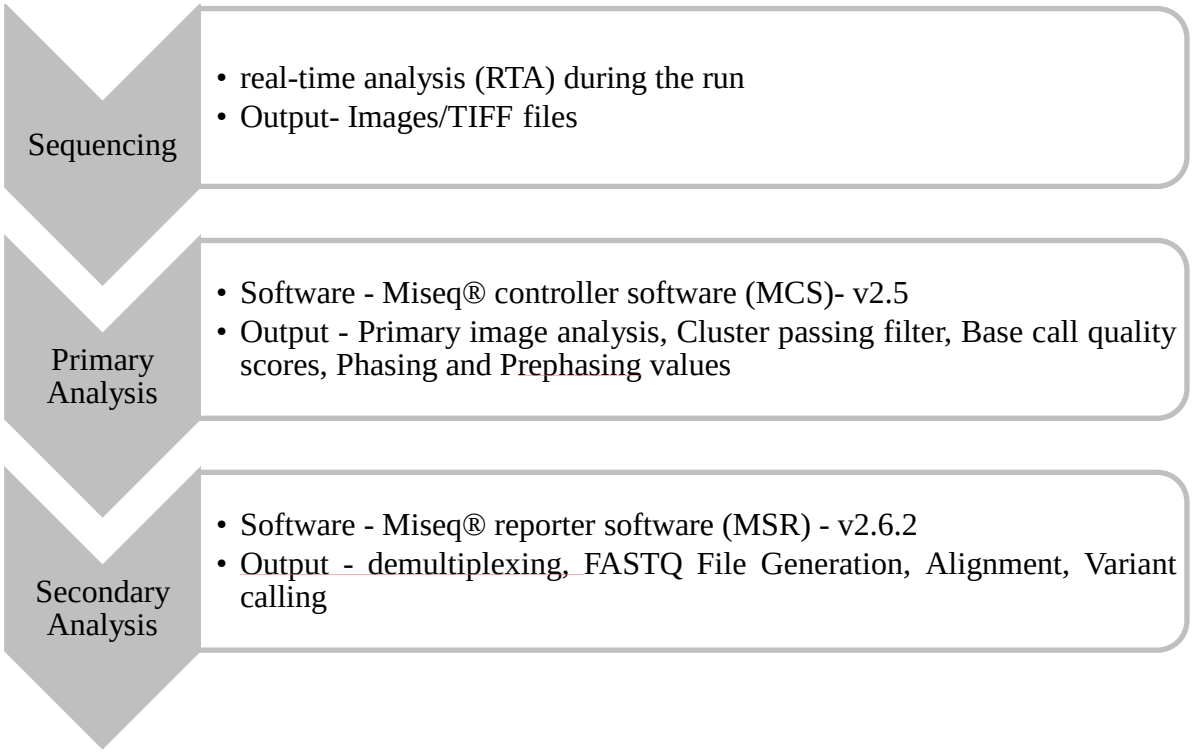


Figure 4.10: Flowchart showing the data analysis procedure in the Illumina Miseq® Instrument, with the respective outputs in each analysis.

For monitoring a run in more detail the Sequencing analysis viewer (SAV) can be used during and after the run. SAV allows reviewing metrics during a run as the metrics were generated, and also after a run has completed. After template generation, SAV provides metrics generated by RTA and organizes the metrics into plots, graphs, and tables (supplementary figure 4). In the secondary analysis, MSR processes base calls generated on-instrument during the sequencing run by RTA software. Upon completion of RTA, MSR produces information about alignment, structural variants, and contig assemblies for each genome and sample. MSR also displays the metrics into plots, graphs, and tables (supplementary figure 5).

In the current study, Image analysis and base calling were performed using Illumina real time analysis and high quality reads were aligned and mapped to the human assembly hg19 (GRCh37) using Burrows–Wheeler Aligner (BWA) (Version 0.6.1-r104-tpx) against the reference sequences. Following local realignment, single-nucleotide variants and small indels were called with GATK´s unified genotyper (version 1.6-22).

4.8 Variant detection and filtering

Identification of novel or very rare variants was the prime goal of this current study. To achieve this we have applied few filtering steps to determine promising variants for each sample which may be responsible for the phenotype. A brief overview of the filtering strategy is explained in figure 4.11 and figure 4.12. Careful scrutiny was done for each sample and the filters we applied were chosen not too stringent in order to keep the sensitivity for potential causative mutations high. In the current study, genomic variant analysis, filtering and interpretation was done using two different softwares – Illumina's VariantStudio v2.2 and Alamut software suite (Interactive Biosoftware, Rouen, France). The VCF (Variant call format) files generated during the analysis step of the sequencing data were used as input formats for variant detection.

The following filtering strategies were applied for each sample after variant calling

- Variants with all predicted consequences on mRNA/protein were considered including synonymous, missense, deletions/insertions, frameshift, intronic upstream and downstream from splice site affecting sites
- All variants having mean depth coverage and read counts ≥10x were only considered.
- All variants only with variant allele frequency >30% for heterozygous mutations and >90% for hemi or homozygous mutations were considered.
- For rare variants identification, minor allele frequency cutoff of 0.001, which is commonly used in rare variant association studies of available population data, was considered in this study.
- For artifact and cross sample verification, Integrative Genomics Viewer (IGV v2.3, www.broadinstitute.org/igv) was used. BAM files (.bam) which are the binary versions of a SAM file (Sequence Alignment Map) were used as input format into it and the coverage depth of reads for each variant called was also viewed on the reference.
- For novel variant identification, the filtered variants were then cross checked from a variety of public databases to exclude the rare variants with already described phenotype or disease association. Various databases used were NCBI SNP database (dbSNP), the 1000 Genomes Project (TGP), Exome Aggregation Consortium (ExAC), and for disease variants identification ClinVar and the Human Genome Mutation Database (HGMD) were used.
- For identification of any reported somatic mutation or oncogenic effect, the potential variants were also cross checked in COSMIC database (cancer.sanger.ac.uk/cosmic).

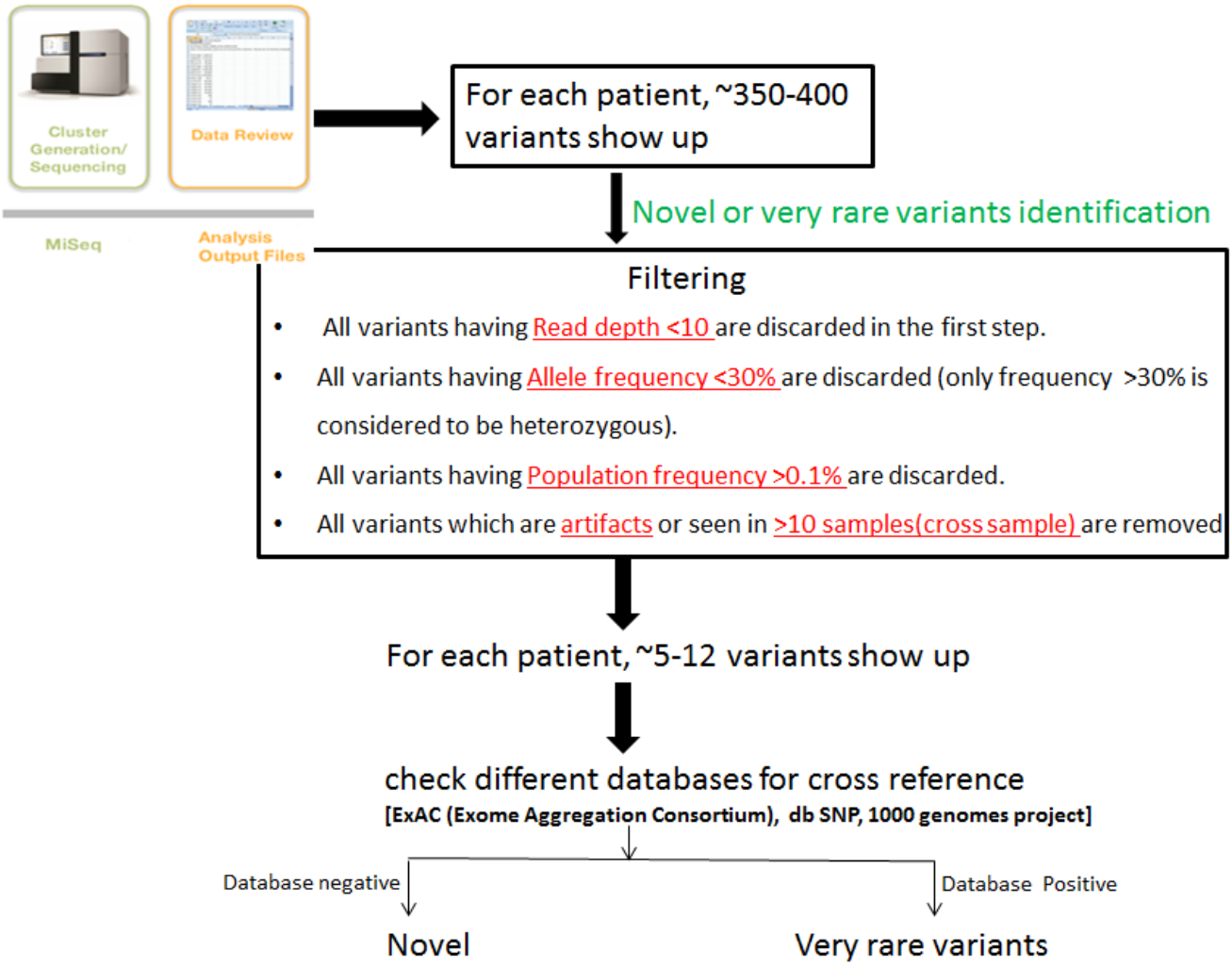


Figure 4.11: Flowchart showing the different strategical steps used for filtering of the rare variants identified in this current study.

Once the rare and novel variants have been filtered for each patient, a further second filtering pipeline was used for narrowing down to the potential disease causing variant matching the phenotype (Figure 4.12). The following filtering strategies were applied for each potential variant

- All the genes included in this study were divided into three groups based on the evidence of known phenotypes/syndromes/disorders in Humans, monogenic or complex forms inferred from other studies related to ID.
    - ID panel (75 genes) - genes with known monogenic disorders for ID or related disorders (Table 4.5)
    - Second panel (106 genes) - genes with any known monogenic disorders (listed in Appendix IV)
    - Third panel (147 genes) - genes with no known monogenic disorders description (listed in Appendix IV).

Table 4.5: The complete list of ID panel genes separated according to mode of inheritance.

Intellectual disability - ID Panel genes - 75			
Dominant (D)		**Carrier (C)**	**Diverse**
BRAF	NF1	ARFGEF2	AKT1
CACNA1A	NFIX	ARHGEF6	BDNF
CACNA1G	NRAS	CC2D1A	CTNND2
CACNG2	NRXN1	CDK6	NLGN3
CASK	PIK3R2	CNTNAP2	NLGN4X
CBL	PRRT2	DLG3	PIK3CA
CTNNB1	PTPN11	FGD1	RASA2
DOCK8	RAF1	GRIA3	RNF135
DYNC1H1	RHEB	GRIK2	SHANK1
EEF1A2	RIT1	GRIP1	SHANK2
EPB41L1	RPS6KA3	GRM1	SYN1
ERBB4	SHANK3	IGF1	
GDI1	SHOC2	IL1RAPL1	
GRIN1	SLC2A1	OPHN1	
GRIN2A	SNAP25	PAK3	
GRIN2B	SOS1	PTCHD1	
HRAS	SOS2	RAB39B	
ITPR1	SOX2	RAB3GAP1	
KCNQ2	SPRED1	TSPAN7	
KCNQ3	STXBP1		
KRAS	SYNGAP1		
MAP2K1	TWIST1		
MAP2K2			

- The selected variants were then filtered based on the type of inheritance. For variants identified in genes with only recessive phenotypes associated to date, a possible second mutation was screened thoroughly in other exons of the gene and the read coverage for each exon was also carefully checked for any gaps or low coverage. This was done by using the Integrative Genomics Viewer (IGV, v2.3).
- Only variants shown to have pathogenic or likely pathogenic effect according to the evidence categories outlined in the American College of Medical Genetics and Genomics (ACMG) recommendations (Richards, et al. 2015) were considered further. In this study to predict the possible impacts of a mutation to be pathogenic or benign and on the protein structure/function, different web-based prediction tools with different algorithms were used which include Mutation Taster; Polyphen-2; SIFT; UMD-Predictor; Align GVGD; MutPred; Meta-SNP which integrates four different existing methods: PANTHER, PhD-SNP, Meta-SNP, SNAP; and Multivariate analysis of protein polymorphism (MAPP). Variants which pass in more than 50% of the prediction programmes as pathogenic were included further.

- To predict the possible effects on splicing and effects on mRNA various online tools were used including the Human Splicing Finder (HSF 3.0); NetGene2Server; splice site prediction by neural network (NNSPLICE); MaxEntScan; and Genesplicer
- For identification of evolutionarily conserved elements in a multiple alignment, three different conservation scores were also used - PhyloP (basewise conservation score), PhastCons (region wise conservation score) and Grantham score (conservation of amino-acid residues).
- For rare variants identification with probable pathogenic effect, minor allele frequency cutoff was increased to 0.0001 (0.01%) of available ExAC data.
- For the detected potential pathogenic variant, further evaluation was performed by familial cosegregation analysis and phenotype match (figure 4.12). When sufficient data was collected supporting the overlapping of all clinical features described/published and inheritance mode for a particular variant, the variant was considered pathogenic and verification in a diagnostic lab was recommended.
- Further factors considered were like linkage to RAS or GH pathway either directly/indirectly or animal model with behaviour/neurological phenotype or shown in functional aspects in vitro related to brain/synaptic function involvement.

4.9 Statistical analysis

To explain the occurrence of a multifactorial concept, comparison tests were done between the two cohorts for the custom gene panel- ID (case group) and Short stature (control group) cohorts. The core/main genes in both the pathways were selected and analysed for particular enrichment of variants in certain genes in both the cohorts. Comparison tests between the cohorts were also done for various parameters like total number of variants identified, number of variants identified for each selected gene, frequency of heterozygous rare variants identified and frequency of the variants identified in known ID genes. A two sample t-Test and chi-square test was done to determine whether there was a significant difference in the number of variants between the controls versus case groups. For comparision purposes various parameters were tested between the standard and modified protocol runs. The different parmeters tested were mean region coverage depth (X), 1x coverage, 20x coverage, Q30 scores, number of gaps ≤ 20x coverage, uniformity of coverage (%), % duplicate paired reads, and percent aligned reads. A two sample t-Test was done to determine whether there was a significant difference for different parameters tested between the standard versus modified protocol.

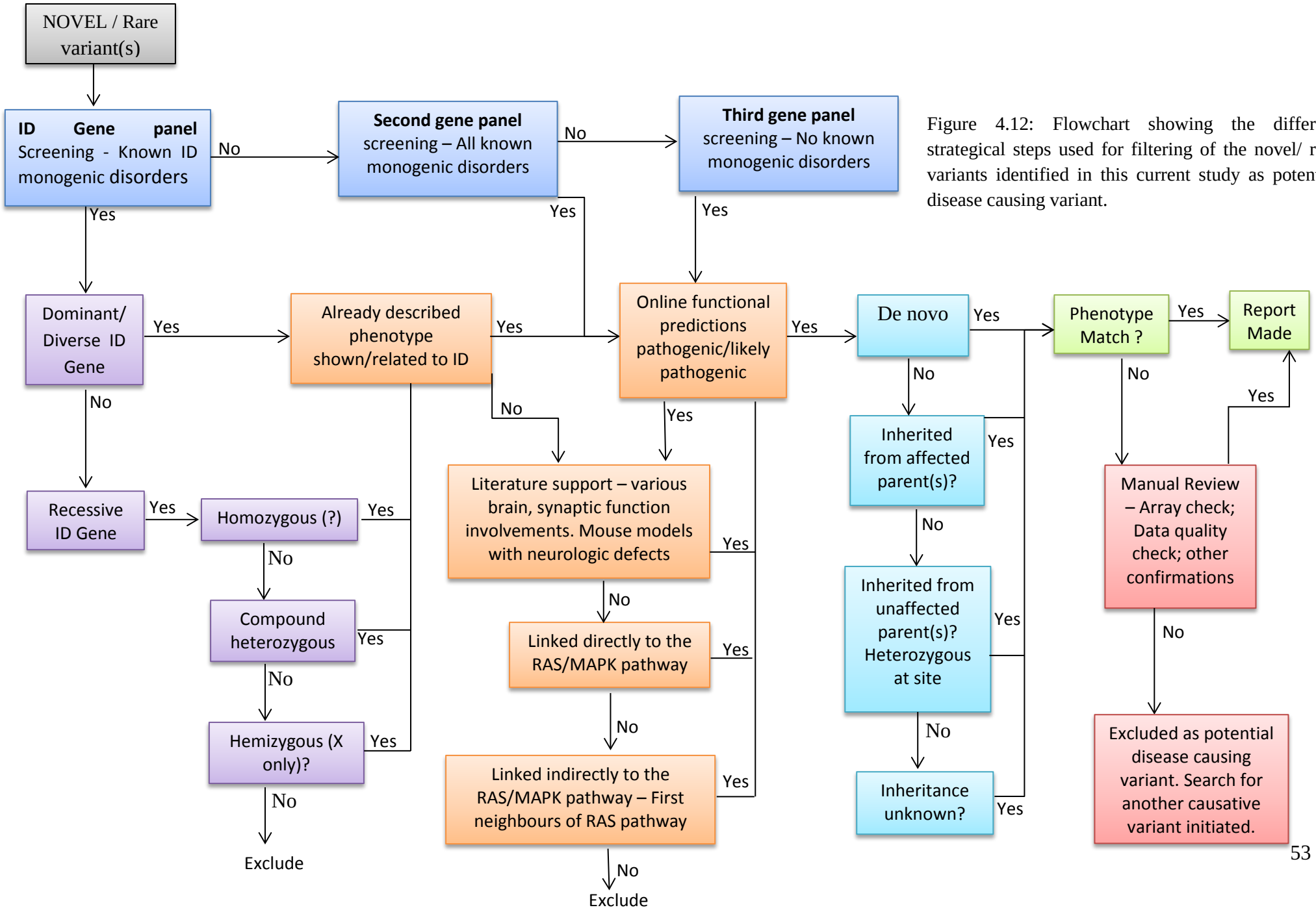


Figure 4.12: Flowchart showing the different strategical steps used for filtering of the novel/ rare variants identified in this current study as potential disease causing variant.

5. Results

5.1 Mosaic disorders

Two individual studies were done related to mosaic disorders, one study includes the identification of hot spot mutations of PIK3CA in patients presenting with disorders of the PROS (PIK3CA-related overgrowth spectrum) and second study was to corroborate the evidence of OES being a mosaic RASopathy. Analysed lesional tissues included epibulbar dermoids, epidermal nevi, skin, fatty tissues, and connective tissue. Blood was also available from the patients. The challenges involved in finding the causative somatic mutations are:

- Not present in every cell of the body
- Not present even in every cell of an affected tissue
- A wild-type allele is also present

Overcoming these challenges and identifying causative mutations was done by using different sequencing methods. Because detection levels and quantification of mutant alleles were limited to 10-15% by Sanger sequencing, other methods like fragment analysis and amplicon-deep sequencing (on a Roche GS Junior platform) were applied. This improved the mutant allele detection to mosaic levels of <1%.

5.1.1 PIK3CA mutation spectrum of patients with PROS

Molecular analysis was primarily performed for all patients by traditional Sanger sequencing[5]. All coding exons with flanking introns of PIK3CA were analysed. Mutant allele ratios of 30-50% were observed in scrapings from epidermal nevi or affected fatty tissue samples. In none of the blood samples PIK3CA mutations were detected by Sanger sequencing. In order to rule out other disorders, hot spot regions including Exon 10 of PIK3R1 and PIK3R2 (megalencephaly-related syndromes), and Exon 4 of AKT1 (Proteus syndrome) were also analysed. All of them were negative and no overlap between the disorders was identified.

A total of 9 missense mutations and 3 small in frame deletions were identified in PIK3CA gene by Sanger sequencing which are described in Table 5.1. Three novel mutations (1 missense and 2 deletions) were identified which have not been described related to PROS until now. All the mutations identified in this study were termed disease causing by different web-based prediction tools. In this study, majority of the mutations lie in the hot spot region

[5] The Sanger sequencing data was provided by apl. Prof. Dr. Ilse Wieland, Institute of Human Genetics, University Hospital Magdeburg, Germany.

of PIK3CA gene i.e. exon 21 and the common mutations already described. Different tissue samples yield different allele ratios for the same mutation in different patients.

Table 5.1: A summary of the identified hot spot mutations in PIK3CA for different tissue samples/blood by Sanger sequencing

Exon	cDNA	Predicted protein	COSMIC / db SNP	Samples (n)
2	c.317G>T	p.Gly106Val	COSM748 / -	1
2	c.317_328del	p.(Gly106_Glu109del)	- / -	4
2	c.328_330del	p.(Glu110del)	COSM4971083 / -	4
5	c.1035T>A	p.Asn345Lys	COSM754 / rs121913284	1
8	c.1338G>C	p.Trp446Cys	- / -	2
8	c.1340_1366del	p.(Pro447_Leu455del)	COSM5944102 / -	2
10	c.1624G>A	p.Glu542Lys	COSM760 / rs121913273	6
10	c.1633G>A	p.Glu545Lys	COSM763 / rs104886003	1
10	c.1637A>G	p.Gln546Arg	COSM12459/ rs397517201	2
21	c.3130A>T	p.Asn1044Tyr	COSM36288 / -	2
21	c.3140A>T	p.His1047Leu	COSM776 / rs121913279	4
21	c.3140A>G	p.His1047Arg	COSM775 / rs121913279	19

5.1.1. (A) Confirmation of the causal variants by NGS

For somatic mutation detection in all samples and tissues, the Sanger sequencing results were not sensitive enough. Improved detection methods were required for other tissues with low level of somatic mosaicism. So, a total of 47 samples in Run I and 35 samples in Run II were sequenced by amplicon deep sequencing of the PIK3CA gene on the GS Junior Instrument. All the mutations in Table 5.1 were confirmed by NGS except for Exon 8 c.1338G>C which was negative in the NGS run. Thus this Exon 8 mutation turned out to probably be a sequencing artifact in Sanger sequencing. From the patients from which multiple tissues were available for testing, one example is shown in figure 5.3 of a patient carrying a mutation at c.3140A>G in PIK3CA showing the different percentage of mutant allele frequencies in different tissues detected by various sequencing methods.

NGS Run I: The first run on the GS junior was performed with all the 47 samples. A shotgun processing was implemented since the position of the mutations was already known and also to get the maximum number of reads from both directions. Though necessary qualitative and quantitative measurements were performed, the run was a failure with only few results. There were more number of short and dot reads than the desired library which spoiled the run. Of the total reads, 91,557 reads (31.94%) could be assembled to the amplicon reference sequences (passed filter wells). There were 50,075 short quality reads around 100-150 bp region which are probably the primer dimers after secondary PCR or unspecific sequences.

Figure 5.1 shows the distribution of the failed and passed filtered reads for all the 47 patients included in the run. In summary, 68.06% did not pass the quality filter systems provided in the GS Junior software for various reasons: short read length (17.47%) or incomplete extension and mixed reads (50.26%). All these reads were excluded in further analysis as failed reads reducing the final data set from 286,616 reads to 91,557 passed filtered reads (31.94%). Figure 5.1c shows the distribution of amplicon read lengths for all the 47 patients included in the run.

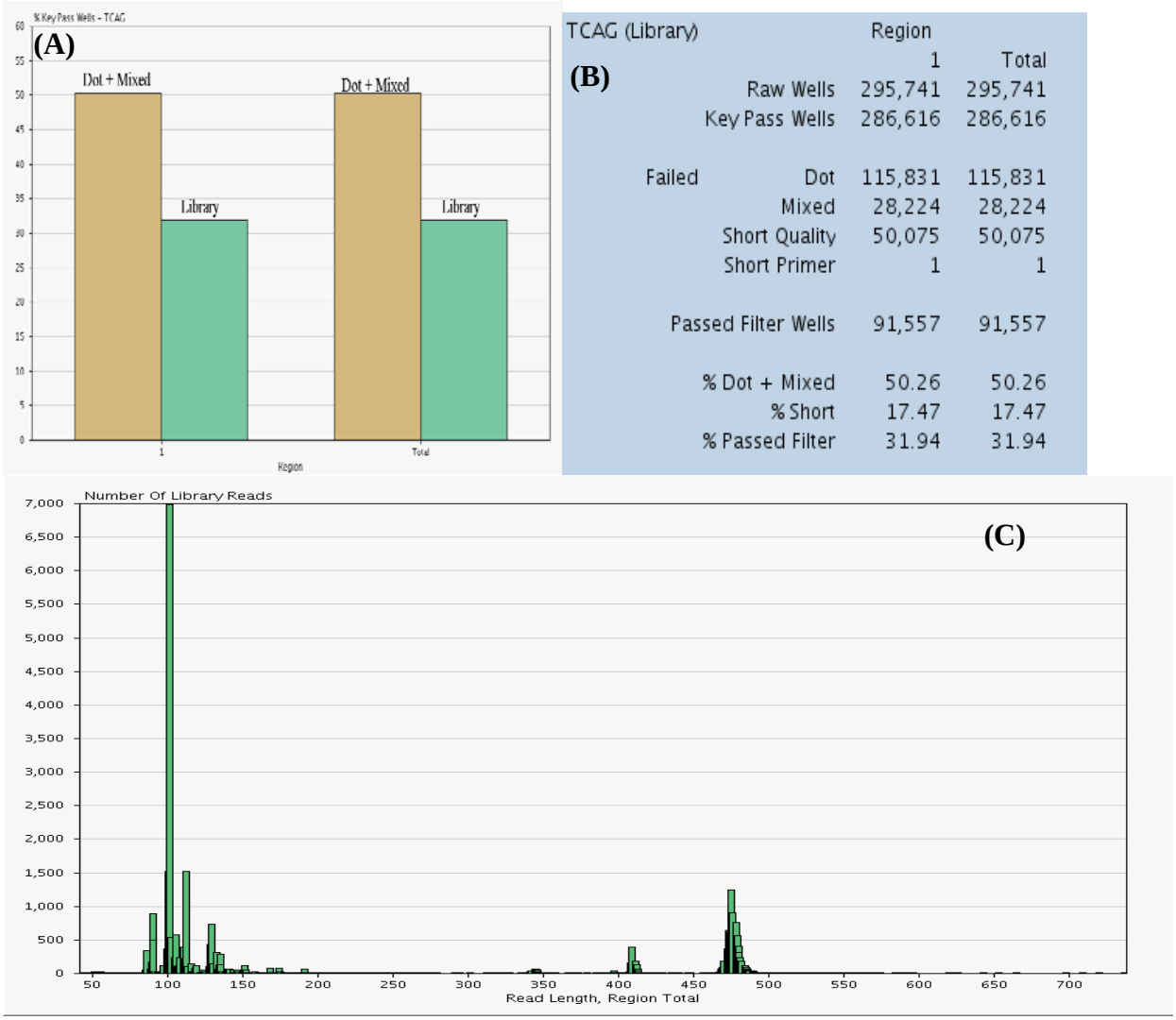
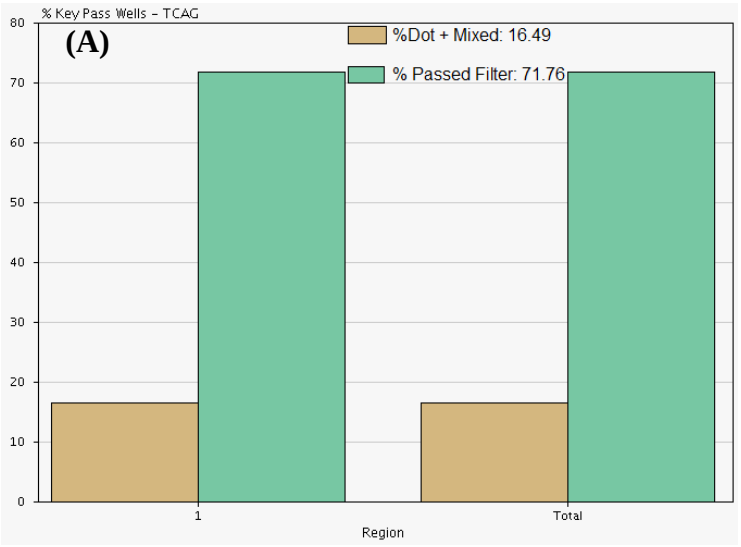


Figure 5.1: GS Junior Run I summary. (A) The distribution of key pass wells for 47 patients of the run and (B) the summary chart showing the statistics of the failed and passed filtered reads (C) The distribution of amplicon read lengths for 47 patients of the Run I. The frequency of read lengths (bp) is plotted from all the 91,557 reads generated in one run.

NGS Run II: The samples with deletions and few FFPE samples which showed higher background in Run I were not included in this run reducing the sample size to 35. The run was successful with the minimum read count of 2000 reads per sample. The dot and the short quality reads drastically reduced than the first run improving the quality of the run. Of the

total reads, 198157 reads (71.76%) could be assembled to the amplicon reference sequences (passed filter wells). There were 31,728 short quality reads around 100-150 bp region which are probably the primer dimers after secondary PCR or unspecific sequences. Figure 5.2 shows the distribution of the failed and passed filtered reads for all the 35 patients included in the run. In summary, only 27.98% did not pass the quality filter systems provided in the GS Junior software for various reasons: short read length (11.49%) or incomplete extension and mixed reads (16.49%). All these reads were excluded in further analysis as failed reads reducing the final data set from 276,128 reads to 198,157 passed filtered reads (71.76%). Figure 5.2c shows the distribution of amplicon read lengths for all the 35 patients included in the run. Table 5.2 shows the variants frequency for both sequencing methods- Sanger and NGS Run II representing all identified PIK3CA mutations.



(B)

TCAG (Library)		Region 1	Total
	Raw Wells	285,238	285,238
	Key Pass Wells	276,128	276,128
Failed	Dot	17,411	17,411
	Mixed	28,113	28,113
	Short Quality	31,728	31,728
	Short Primer	1	1
	Passed Filter Wells	198,157	198,157
	% Dot + Mixed	16.49	16.49
	% Short	11.49	11.49
	% Passed Filter	71.76	71.76

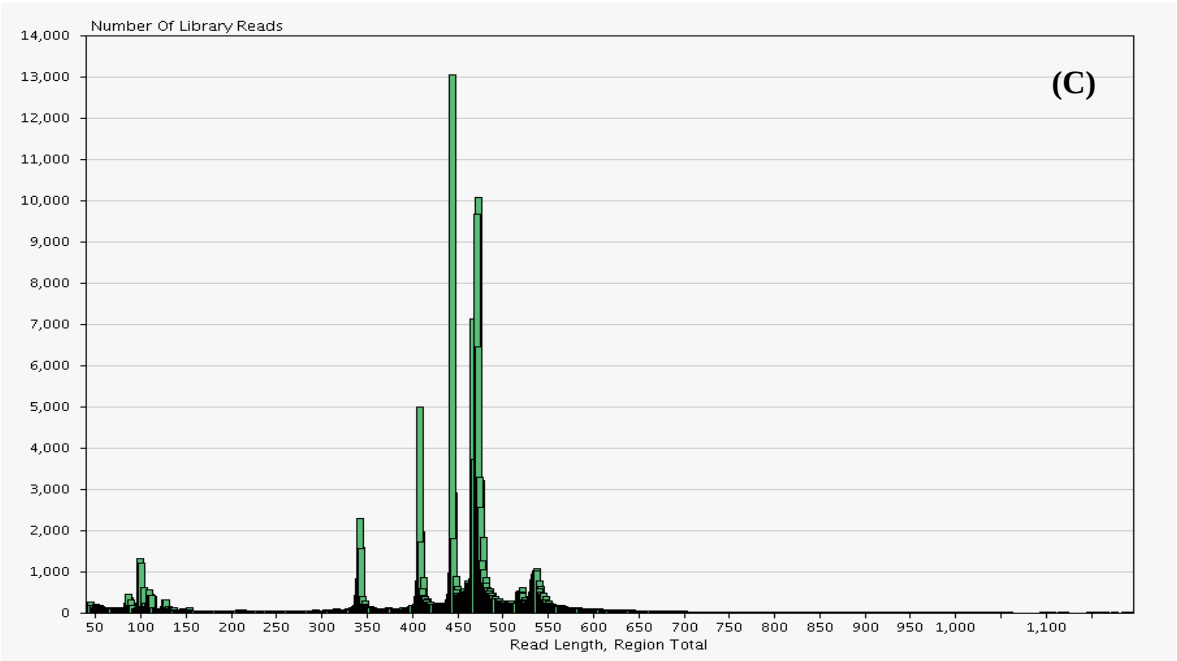


Figure 5.2: GS Junior Run II summary. (A) The distribution of key pass wells for 35 patients of the run and (B) the summary chart showing the statistics of the failed and passed filtered reads. (C) The distribution of amplicon read lengths for 35 patients of the Run II with median reads length of 467bp. The frequency of read lengths (bp) is plotted from all the 198,157 reads generated in one run.

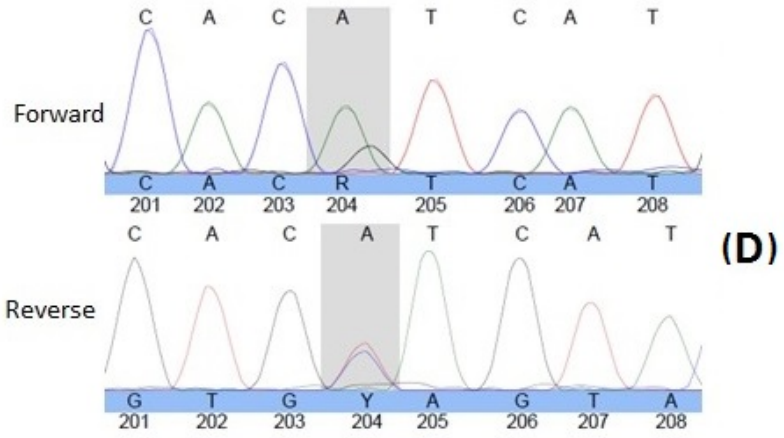
Table 5.2: Variants frequency table for Sanger sequencing and NGS Run II representing all identified PIK3CA mutations. The total number of samples with the total number of reads per sample and percentages of variants showing the mutations are shown below. The patient ID and sample number are in reference to Supplementary table 1. (CLOVES - congenital lipomatous overgrowth, vascular malformations, epidermal nevi, and skeletal abnormalities; HHML- hemihyperplasia-multiple lipomatosis; seq- sequencing)

Exon	**Patient ID**	**Classification of Phenotype**	**Sample number**	**Tissue**	**Sanger seq allele %**	**Number of Reads**		**Variant reads**		**% of Variant**	
						Total no of reads	**Variant reads**	**Forward**	**Reverse**	**Variant**	**Wild type**
Exon 2 **p.(Glu110del)**	P3 P2	CLOVES CLOVES	8 6	Blood Blood	**<10** **<10**	3244 5824	---- ----	---- ----	---- ----	**----** **----**	100 100
Exon 2 **p.Gly106Val**	P4	CLOVES	9	Blood	**<10**	9408	117	76	41	**1.24**	98.76
Exon 5 **p.Asn345Lys**	P7	Macrodactyly	14	Bone	**30-40**	4271	1567	829	738	**36.6**	63.4
Exon 10 **p.Glu542Lys**	P8	CLOVES	15	Fibroblasts	**20-30**	3818	1382	755	627	**36.1**	63.9
	P12	CLOVES	21	FFPE	**<10**	4363	75	46	29	**1.72**	98.28
	P10	CLOVES	18	Fat tissue	**15-25**	4769	1140	569	571	**23.9**	76.1
Exon 10 **p.Glu545Lys**	P11	CLOVES	20	Cartilage	**30-50**	-----	**----**	-----	**----**	-----	-----
Exon 10 **p.Gln546Arg**	P9	CLOVES	16	Skin	**30-40**	2577	435	36	399	**16.8**	83.2
Exon 21 **p.Asn1044Tyr**	P15	HHML	33 34	Fat Blood	**30-40** **<10**	5810 2737	1652 5	842 4	810 1	**28.4** **0.18**	71.6 99.82

Exon 21 **p.His1047Leu**	P18	CLOVES	45 46 47 48	FFPE- tumor FFPE- Bone Skin Blood	**38** **35** **30** **<10**	2373 2541 2486 2470	900 769 583 ---	471 379 285 ---	429 390 298 ---	**37.9** **30.2** **23.4** **---**	62.1 69.8 76.6 100
Exon 21 **p.His1047Arg**	P13	CLOVES	24 25 26 27 28 29	Blood Cartilage Tendon Skin Connective tissue Epiphyses tissue	**<10** **25-35** **30-40** **13-20** **30-50** **20-30**	3047 4420 4097 3301 3975 3483	3 1192 1459 533 1424 774	3 535 688 268 723 398	--- 657 771 265 701 376	**0.09** **26.9** **35.6** **16.1** **35.8** **22.2**	99.91 73.1 64.4 83.9 64.2 77.8
	P16	CLOVES	35 36 37 38 39 40 41	Blood Skin Adipose neck Adipose cervical Connective Keratinocytes Nevus	**<10** **40** **---** **15-30** **15-20** **34-50** **30-40**	2448 4116 2680 3523 5724 7614 3966	3 1278 374 602 837 3625 1419	--- 825 192 317 426 1903 736	3 453 182 285 411 1722 683	**0.12** **31** **13.9** **17** **14.6** **47.6** **35.7**	99.88 69 86.1 83 85.4 52.4 64.3
	P17	Macrodactyly	42 43 44	Subcutaneous tissue Deep fat skin	**<10** **20-30** **<10**	2811 2063 2664	117 420 203	57 224 112	60 196 91	**4.1** **20.3** **7.6**	95.9 79.7 92.4
	P14	HHML	32 30 31	Fat Bone Blood	**20-30** **7-13** **<10**	3632 3439 4816	736 283 7	401 142 5	335 141 2	**20.2** **8.2** **0.1**	79.8 91.8 99.9

(Classification of the phenotype for the patients and most of the patient samples were provided by Prof. Dr. Sigrid Tinschert, Institute of Clinical Genetics, Technical University of Dresden, Dresden, Germany)

Overall, we identified 11 PIK3CA mutations in 18 individuals which consist of 3 deletions and 8 missense mutations. All the identified missense mutations were previously known pathogenic variants in PROS except for one, p.Asn1044Tyr. From the patient P11 presenting with CLOVES, only cartilage sample was available and by Sanger sequencing one known missense change p.Glu545Lys was identified. The sample could not be included in NGS runs due to presence of high short fragments and degraded DNA quality and only Sanger data is available for this variant.



(F)

Tissue	Sample number	% of mutant alleles	
		Sanger	NGS Run II
Blood	24	<10	<2
Cartilage	25	25-35	27
Tendon	26	30-40	36
Skin	27	13-20	16
Connective tissue	28	30-50	36
Epiphyses	29	20-30	22

Figure 5.3: (A) A 4-year old boy (P13) with postaxial polydactyly with bilateral extreme overgrowth of feet and lower limbs. The arrow marks indicated are epidermal naevi on the neck and abdomen; (B) Local overgrowth of both extremities with polydactyly; (C) Lipomatosis of soles of feet; (D) Mutation at c.3140A>G in *PIK3CA* in connective tissue by Sanger sequencing in both forward and reverse directions; (E) GS Junior Flowgram tab for the read displaying the c.3140A>G mutation in *PIK3CA* in the same tissue; (F) Table showing the different percentage of mutant alleles in different tissues by comparing both sequencing methods for c.3140A>G in *PIK3CA*. [Picture of the patient reprinted with permission from, Eva Schneckenhaus (2009). Mutationsanalyse des PTEN-Gens bei Proteus-und Proteus-like-Syndrom (Dissertation Thesis). Medizinischen Fakultät, Otto-von-Guericke University, Magdeburg, Germany.]

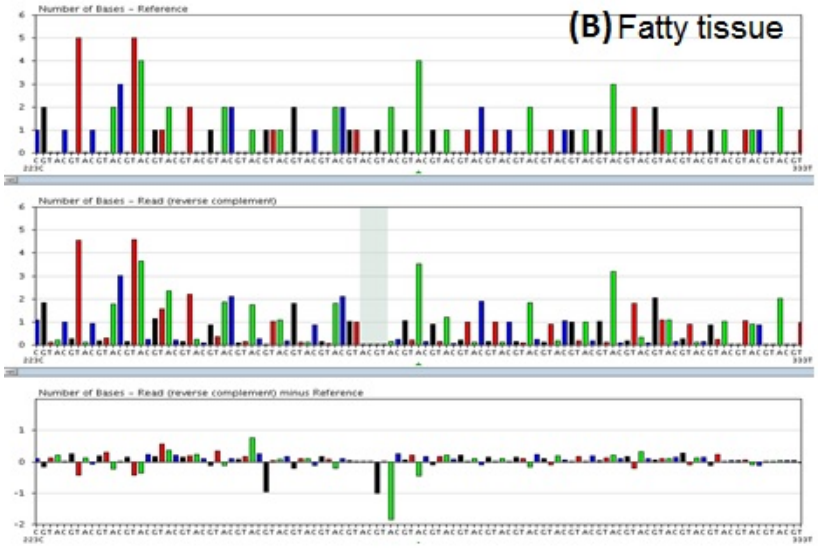
5.1.1. (B) Fragment Analysis

Fragment analysis was also used to detect the presence of deletions in the samples tested for PIK3CA-related overgrowth study. The Genemapper software provided the area under the peak from which the mutant allele ratios were calculated for the deletions. The deletion ratio (%) was calculated using the below formula

$$\% = \frac{\text{Area of mutant peak}}{\text{Combined area (Mutant peak + Wild type peak)}}$$

Two fragment analysis runs were performed for the deletion patients for all samples from which the mean was taken and compared to the NGS run I. Though Run I had poor read count the data was used for comparison purposes between the different sequencing methods (Table 5.3). In Run II the deletion patients were not included except for blood samples. Different percentages of mutant alleles in different tissues are shown by comparing all the three sequencing methods in Figure 5.4 and Figure 5.5 of exon 2 deletions of PIK3CA gene for two patients.

Tissue (A)	Sample number	% of mutant alleles		
		Sanger	Fragment Analysis	NGS Run I
Blood	6	<10	---	---
Fatty tissue	5	50	36	27



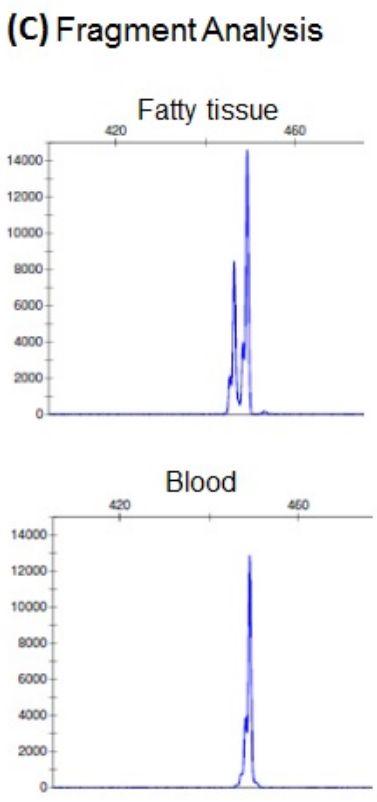


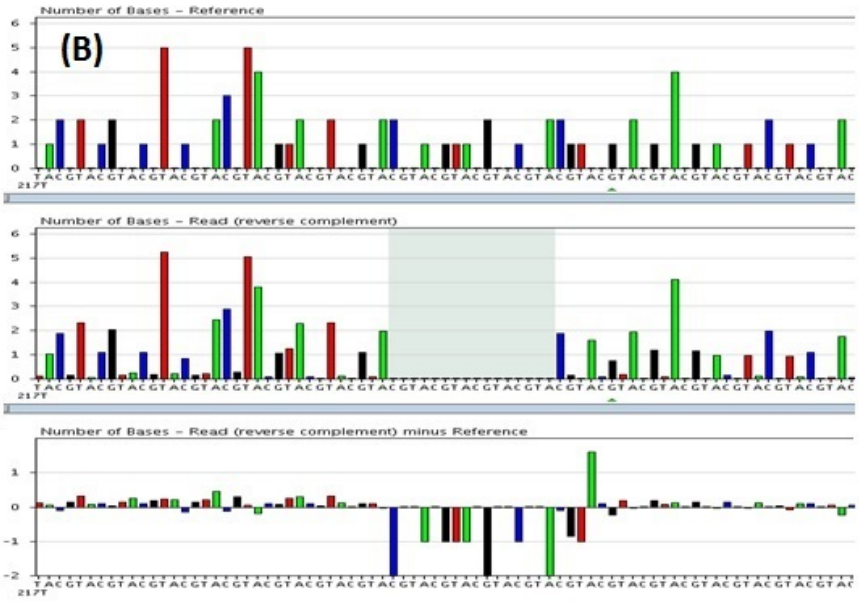
Figure 5.4: Example showing the deletion **c.328_330del; p. (Glu110del)** in exon 2 of PIK3CA for a patient (P2) in all samples by different sequencing methods. (A) Table showing the different percentage of mutant alleles in different tissues by comparing all the three sequencing methods. (B) GS Junior Flowgram tab for the read displaying the deletion in Fatty tissue. (C) Fragment analysis showing the wild type and mutated peaks for Fatty tissue and only wild type peak for the blood sample.

Table 5.3: Deletion ratios table for the samples analysed by Sanger sequencing, Fragment analysis and results of NGS Run I for the deletion samples.

Exon	**Patient ID**	**Classification of Phenotype**	**Sample number**	**Tissue**	**Sanger seq DEL %**	**Fragment Analysis**			**NGS Run I results**			
						Run I DEL %	Run II DEL %	Mean	Total reads	Mutant reads	DEL %	WT %
Exon 2 p.(Glu110del)	P2	CLOVES	5	Fat tissue	**50**	36	36	**36**	157	42	**27**	73
			6	Blood	**<10**	---	---	**---**	98	--	**--**	100
	P3	CLOVES	7	Fat tissue	**20-30**	26	25	**25.5**	129	27	**21**	79
			8	Blood	**<10**	---	---	**---**	---	--	**--**	100
Exon 2 p.(Gly106_Glu109del)	P1	CLOVES	1	Subcutaneous Fat tissue	**<10**	3.4	1.8	**2.6**	635	12	**1.9**	98.1
			2	Skin	**<10**	10.5	7	**8.8**	810	57	**7**	93
			3	Subcutaneous tissue	**30**	23	21	**22**	155	35	**22.5**	77.5
			4	Fat tissue	**30**	23	21	**22**	397	105	**26.4**	73.6
Exon 8 p.(Pro447_Leu455del)	P5	CLOVES	11	Bone	**20-40**	25	21	**23**	268	73	**27**	73
			10	Fat tissue	**40**	33	30	**31.5**	74	25	**34**	66

Note: Though the read count was very low in the NGS Run I, the results here were just used for comparison purposes between the different sequencing methods. (DEL- deletion; WT- Wild type allele)

(A) Tissue	Sample number	% of mutant alleles		
		Sanger	Fragment Analysis	NGS Run I
Subcutaneous fat tissue	1	<10	3	<2
Skin	2	<10	9	7
Subcutaneous tissue	3	30	22	23
Fatty tissue	4	30	22	26



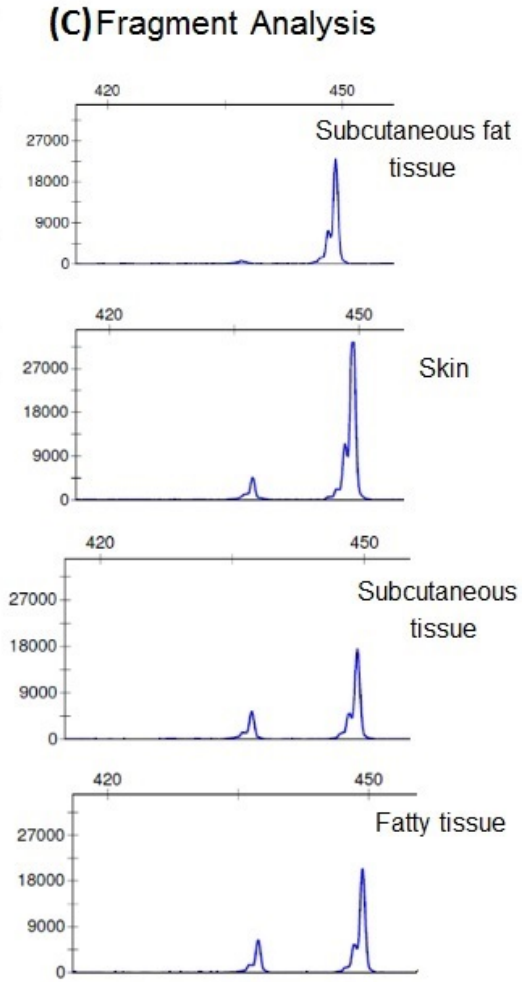


Figure 5.5: Example showing the deletion **c.317_328del; p. (Gly106_Glu109del)** in exon 2 of *PIK3CA* for a patient (P1) in all samples by different sequencing methods. (A) Table showing the different percentage of mutant alleles in different tissues by comparing all the three sequencing methods. (B) GS Junior Flowgram tab for the read displaying the deletion in one of the tissue. (C) Fragment analysis showing the wild type and mutated peaks for different tissues for the deletion.

5.1.2 Mosaic KRAS mutations in OES/ECCL patients

In lesional tissues from all four patients with a clinical diagnosis of OES and ECCL, respectively, we could identify KRAS mutations at various levels of mosaicism, while the mutations were not detected in leukocyte-derived DNA at the detection threshold of Sanger sequencing. The observed mutations in all four patients affected exon 4, codon 146 (Table 5.4). Two patients were found to carry a mosaic for the sequence variant c.437C>T (p.Ala146Val), and in the other two, the mutation c.436G>A (p.Ala146Thr) was identified.

Table 5.4: Results of KRAS genotyping in various tissues.

Patient #	KRAS mutation type	Tissue sample(s): Type (proportion of mutant allele[a])	Leukocyte DNA
1	c.437C>T (p.Ala146Val)	Epibulbar dermoid (17%)	Negative[b]
2	c.436G>A (p.Ala146Thr)	Skin biopsy of scalp lesion: Superficial (epidermal) fraction (43%); Lower (dermal) fraction (35%) Cultivated fibroblasts (50%)	Negative[b]
3	c.437C>T (p.Ala146Val)	Fibroblasts from skin biopsy (40%) Jaw giant cell tumor (<10%)	Negative[b]
4	c.436G>A (p.Ala146Thr)	Skin biopsy of scalp lesion: Superficial (epidermal) fraction (24%); Lower (dermal) fraction (38%) Skin biopsy of hyperpigmented skin: Superficial (epidermal) fraction (22%); Lower (dermal) fraction (11%)	Negative[b]

a A proportion of the mutant allele of 50% represents a non-mosaic pattern for a heterozygous mutation. b Below detection threshold of Sanger sequencing.

In **patient 1**, DNA extracted from the epibulbar dermoid sample showed the sequence variant c.437C>T (p.Ala146Val) in a calculated proportion of ~17% for the mutant allele (Figure 5.6A). In **patient 2**, the mutation c.436G>A (p.Ala146Thr) was identified at a high level of mosaicism in DNA from the skin biopsy of the scalp lesion. In the sample containing the epidermal fraction, the mutated allele was present in a slightly higher proportion of ~43% than in the lower (dermal) fraction (~35%). The calculated proportion of the mutated allele ranged from 11% to nearly 50% in the dermal fibroblast cultures from patient 2, indicating a non-mosaic heterozygous pattern in these cells (Figure 5.6B). The same mutation as in patient 1, c.437C>T (p.Ala146Val), was discovered in the fibroblast cultures from a skin biopsy of **patient 3** at a proportion of ~40% for the mutated allele (Figure 5.6C). We were unable to confirm the mutation in the DNA extracted from a formalin-fixed paraffin-embedded tissue sample derived from the giant cell lesion of the mandible of patient 3, but this could be due to a very low amount and quality of DNA that prevented reliable sequencing results (data not shown). The same mutation as in patient 2, c.436G>A (p.Ala146Thr), was discovered from a scalp lesion and a hyperpigmented area of the skin of **patient 4** (Figure 5.6D). In the sample

(skin biopsy of the scalp lesion) containing the dermal fraction, the mutated allele was present in a slightly higher proportion of ~38% than in the upper (epidermal) fraction (~24%). In the sample (skin biopsy of hyperpigmented skin) containing the epidermal fraction, the mutated allele was present in a slightly higher proportion of ~22% than in the lower (dermal) fraction (~11%). Both the observed alterations at KRAS codon 146 are known cancer-associated mutations and listed in the COSMIC database (mutation IDs: COSM19404, COSM19900). They are consistently predicted to be pathogenic by different web-based prediction tools (Mutation Taster, Polyphen, SIFT, SNAP, Meta-SNAP, UMD-Predictor and PhD-SNP).

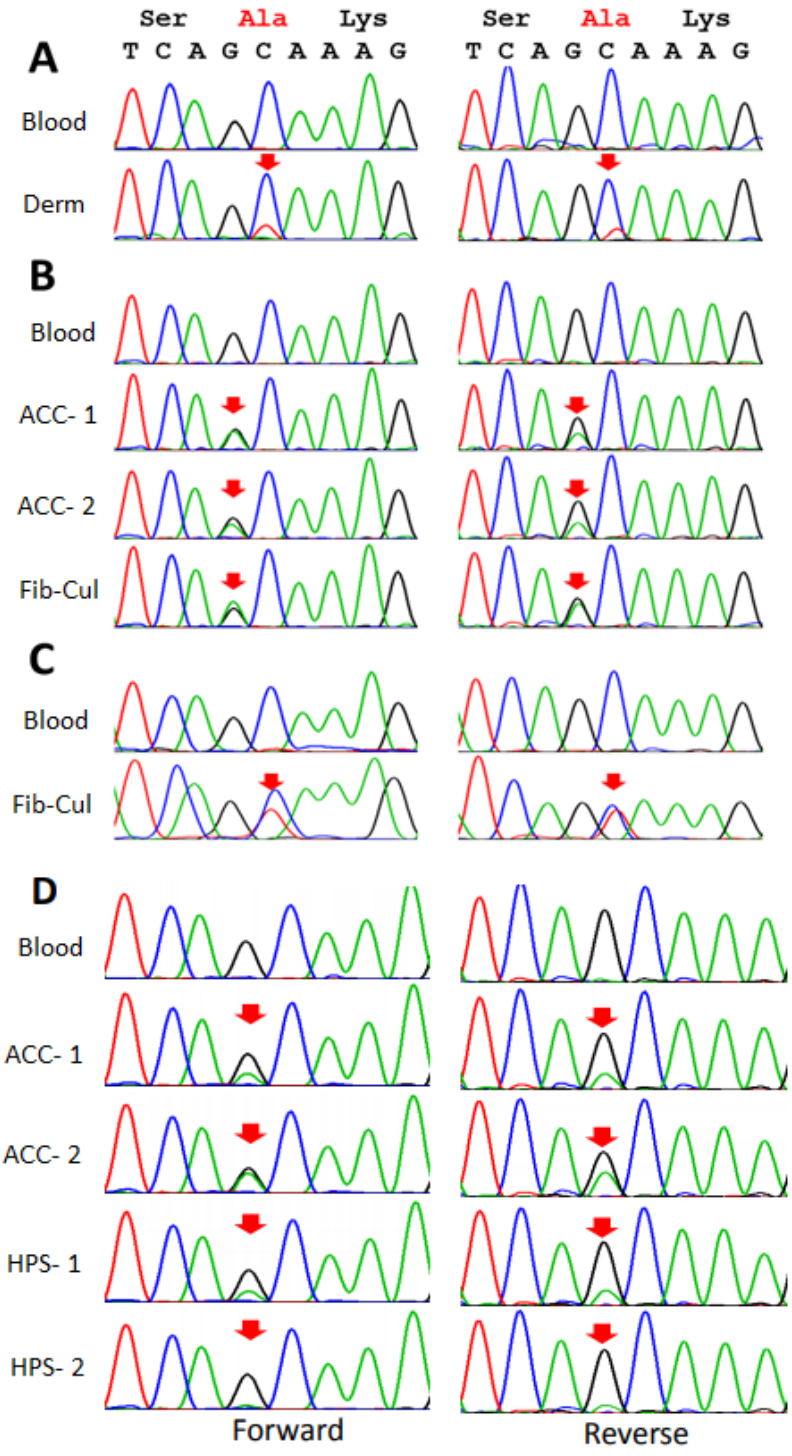


Figure 5.6: Results of bidirectional sequencing of KRAS exon 4 in DNA samples from patient 1 (A), patient 2 (B), patient 3 (C) and patient 4 (D). Electropherograms demonstrate mosaic mutations in lesional tissue and absence of the mutation in DNA extracted from blood samples. Sequences are displayed in forward and reverse sequencing direction. Arrows mark the position of the mutations.

Derm, epibulbar dermoid; ACC-1, aplasia cutis congenital, superficial (epidermal) layer; ACC-2, aplasia cutis congenital, dermal fraction; Fib-cul, fibroblast culture from lesional skin biopsy; HPS- 1, hyperpigmented skin superficial (epidermal) layer; HPS- 2, hyperpigmented skin lower (dermal) layer.

5.2 Multigene panel sequencing in patients with ID and Short stature

This project was aimed on identifying monogenic causes for ID and short stature. Considering the important role of RAS/MAPK signalling in the nervous system (reference to introduction 2.4.2) as well as for growth, this approach was mainly focussed on various genes encoding components or direct and indirect modulators of the RAS/MAPK signalling pathway. In addition, the panel of target genes comprised genes encoding components of the growth hormone signalling pathway as well as other genes involved in short stature. Two independent patient cohorts with ID and short stature, respectively, were investigated. In the current project, a targeted resequencing method is approached using a custom made enrichment based assay - Nextera® Rapid Capture Custom Enrichment Kit from Illumina.

5.2.1 Phenotyping of the study cohorts

A total of 166 patients with ID and 120 growth-deficient patients with normal cognitive development were included. The ID cohort is the main case study group and Short stature cohort served as the control study group in the current project for comparison purposes. The male to female ratio among the probands in both the cohorts were 60:40. The median age at the time of the clinical assessment was 7.0 years for the ID cohort and 5.5 years for the Short stature cohort (SD 4.0, range 3–15). For the ID cohort all the patients had a previous karyotype analysis, microarray and Fragile X syndrome testing, which presented normal results. In the ID cohort, 97% of children had ID of varying degree, 42% had language impairment, 32% had a history of seizures, 10% had a learning disability, and 8% with autism. In the ID cohort, less than 5% of the diverse phenotype terms described symptoms like low set ears, microcephaly, muscular hypotonia, growth retardation, congenital heart defects etc. were present. In the Short stature cohort, all the patients included were devoid of any neurological abnormalities.

5.2.2 Classification of custom-designed gene panel

Genes which might be the possible candidates for monogenic forms of non-syndromic ID/short stature were selected for targeted re-sequencing based on certain criteria as explained in methods section. In the current study, a total of 329 genes were selected which were directly or indirectly related to the RAS pathway (221 genes) and genes related to growth or short stature by various mechanisms (108 genes) with cumulative target length of ~777.12 kb. The full regions of the targeted genes are considered for the study with an average region size of 141bp. The 221 selected genes related to the RAS pathway were broadly divided into eight protein classes according to their function - Synaptic Vesicles / Protein Transport (17 genes);

MAGUKs / Adaptors / Scaffolders (37 genes); G Proteins and modulators (49 genes); Kinases (30 genes); Signalling molecules and Enzymes (13 genes); Cell Adhesion and Cytoskeleton (24 genes); Channels and receptor(s) (31 genes); and regulatory / transcription factors (20 genes). Figure 5.7 here shows an example of two protein classes, Channels and receptor(s) and MAGUKs / Adaptors / Scaffolders with their corresponding genes. The remaining protein classes are shown in Supplementary figure 7.

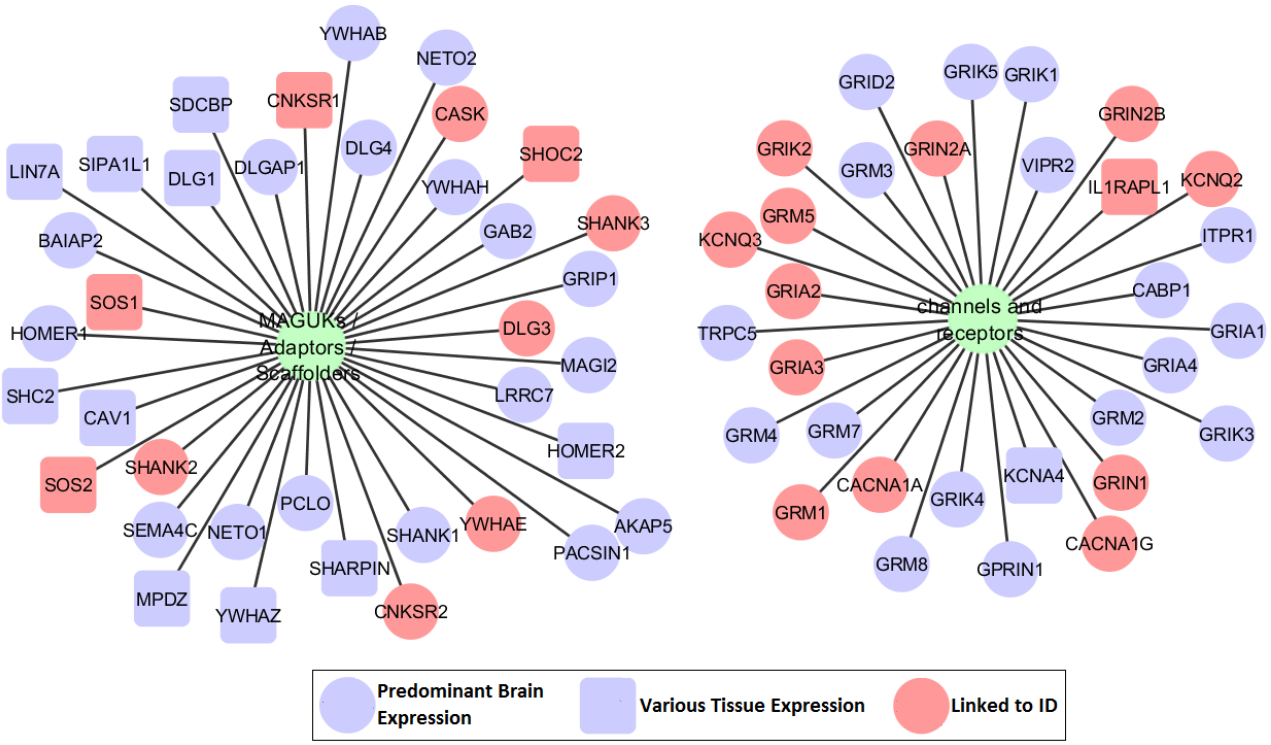


Figure 5.7: Example of two protein classes, MAGUKs / Adaptors / Scaffolders (left) and Channels and receptor(s) (right) with their corresponding genes, related to RAS/MAPK pathway, classified according to their function in the current project (prepared by using Cytoscape software, version 3.3).

The distributions of all 221genes which belong to RAS/RAS extended pathway are grouped according to their protein expression in brain either predominant or low and various tissue expressions are shown in figure 5.8. The figure 5.8 also shows the distribution of the 221 genes with relation to linkage to RAS pathway (direct or indirect interactions) and previously shown linked to ID or not. Out of the 221 genes, 146 genes are predominantly expressed in brain with 49 of them previously linked to ID with a strong monogenic cause. 53 genes had moderate or low brain expression with 15 of them previously linked to ID with a strong monogenic cause and 10 of them linked to mild ID. Only 22 genes had various tissue expressions with 13 of them not linked to ID. Out of the 221 genes, 60 genes are directly linked to RAS/MAPK pathway which are the first interactors of the RAS/MAPK pathway (genes directly linked to HRAS (First neighbours) and genes linked to at least the first neighbours of HRAS (Second neighbours)) and 79 genes are indirectly linked to the RAS/MAPK pathway (genes which are linked to at least the second neighbours of HRAS).

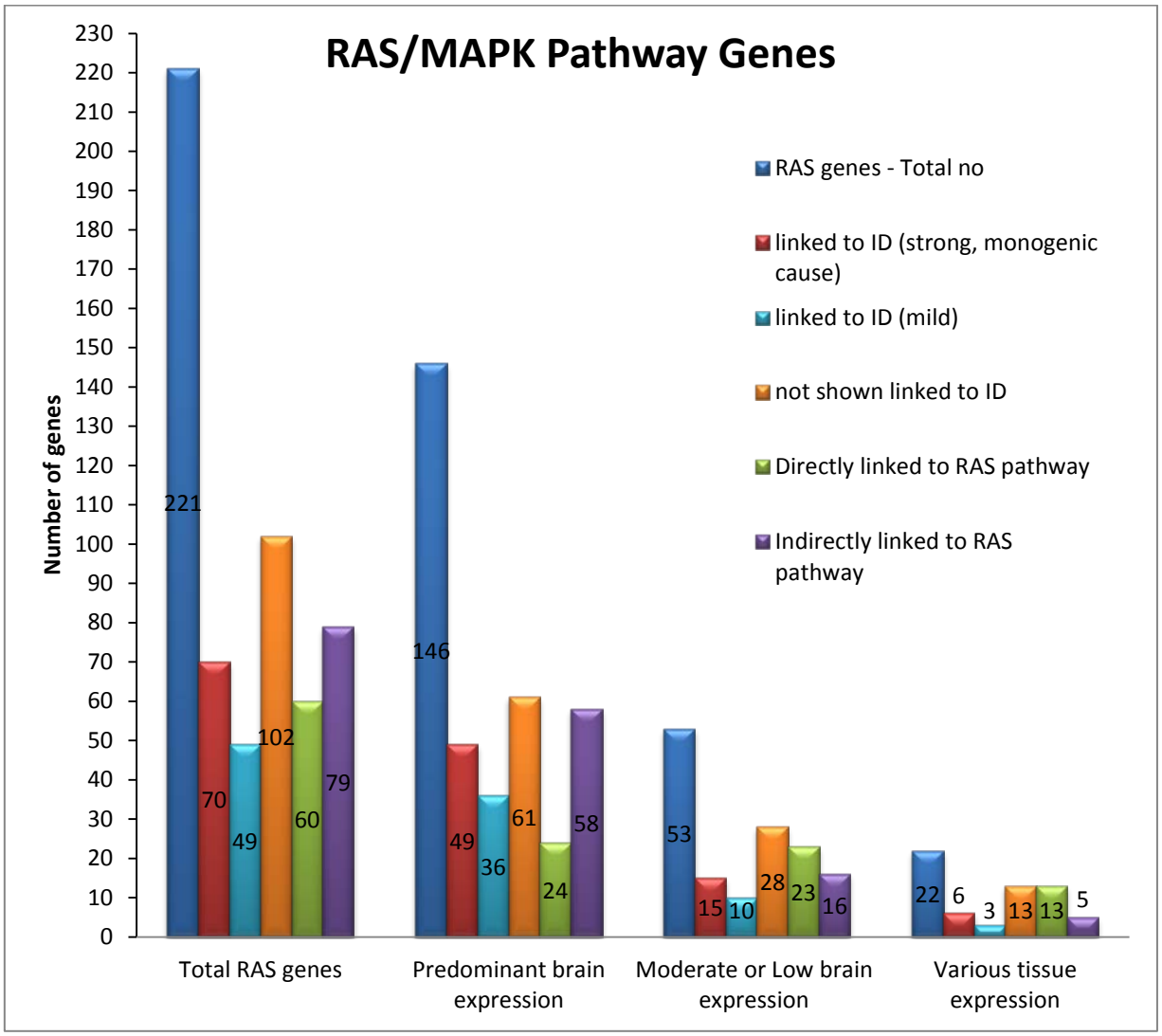


Figure 5.8: Graph showing the distribution of all 221genes which belong to RAS/RAS extended pathway according to their protein expression in brain with relation to linkage to RAS pathway and ID.

5.2.3 Evaluation of the gDNA quality and quantity

In NGS methods for generating high quality data, the quantity and quality of the starting genomic material is of utmost importance. To achieve consistent tagmentation and reproducible library size distributions, accurate DNA input amount is necessary and the quality of the enrichment strongly depends on accurately quantified starting material amount. In the current study to meet this requirement, four different methods for quantity check were used and an exemplary result for subset of samples (single NGS run- 24 samples) are displayed in Figure 5.9. For quantification of samples in further NGS runs, out of the three different methods used, fluorometric measurements were chosen as these assays were highly selective for double-stranded DNA (dsDNA) over RNA or common contaminants than spectrophotometric measurements (NanoDrop) and TapeStation analysis. Two-step gDNA normalization was done for accurate quantification of the starting material used by fluorometric method, using a final volume of 10µl at 5ng/µl (50ng total) in the NGS runs.

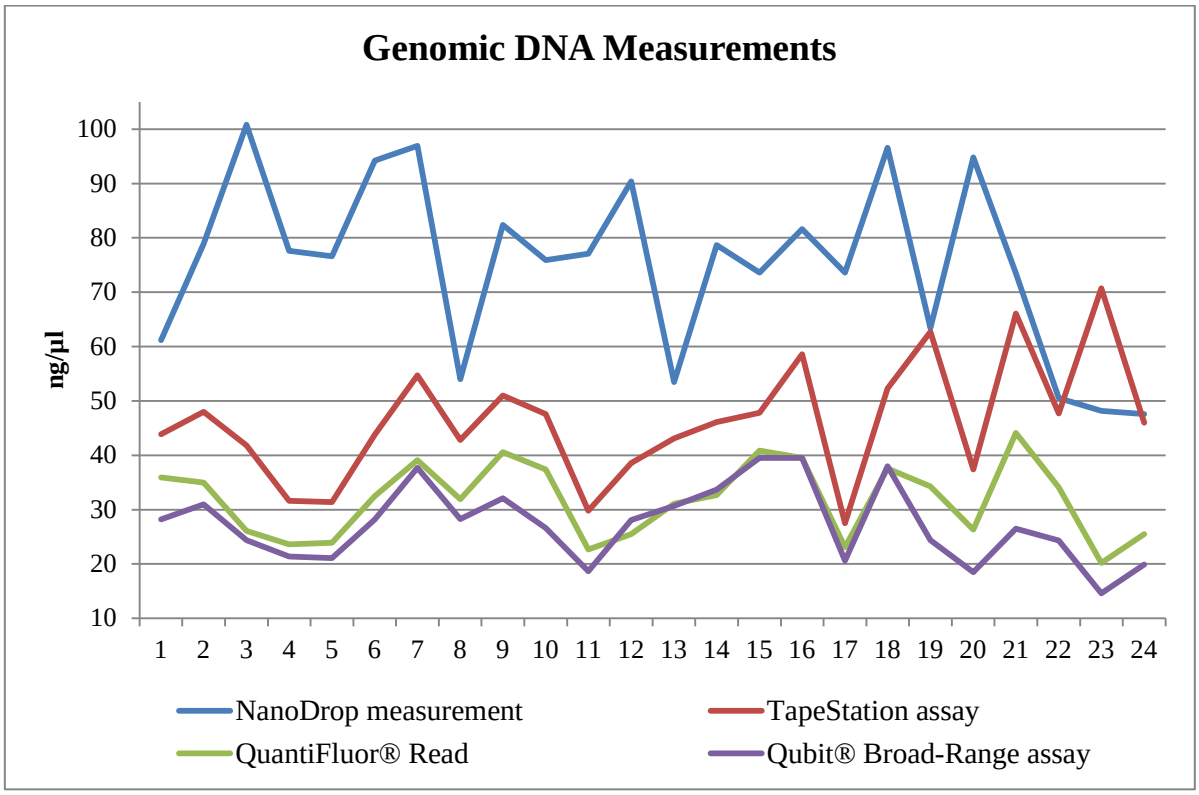


Figure 5.9: Assessment of quantity of gDNA using different methods in the current study for a single run (24 samples). The graph shows that different gDNA concentrations resulted for the same sample by using different methods representing sensitivity of each method.

In NGS methodologies for generating quality data an intact genomic DNA template is crucial. Therefore, assessment of the gDNA integrity is important in the first step itself. Two different gel electrophoretic methods were applied for all samples in the NGS runs as qualitative measures. The figure 5.10 is an example result for a subset of samples, showing highly intact gDNA and no degradation, through sharp bands on the agarose gel and with high DIN values >7 (0-10) (A high DIN indicates highly intact gDNA, and a low DIN a strongly degraded gDNA sample) from the TapeStation assay.

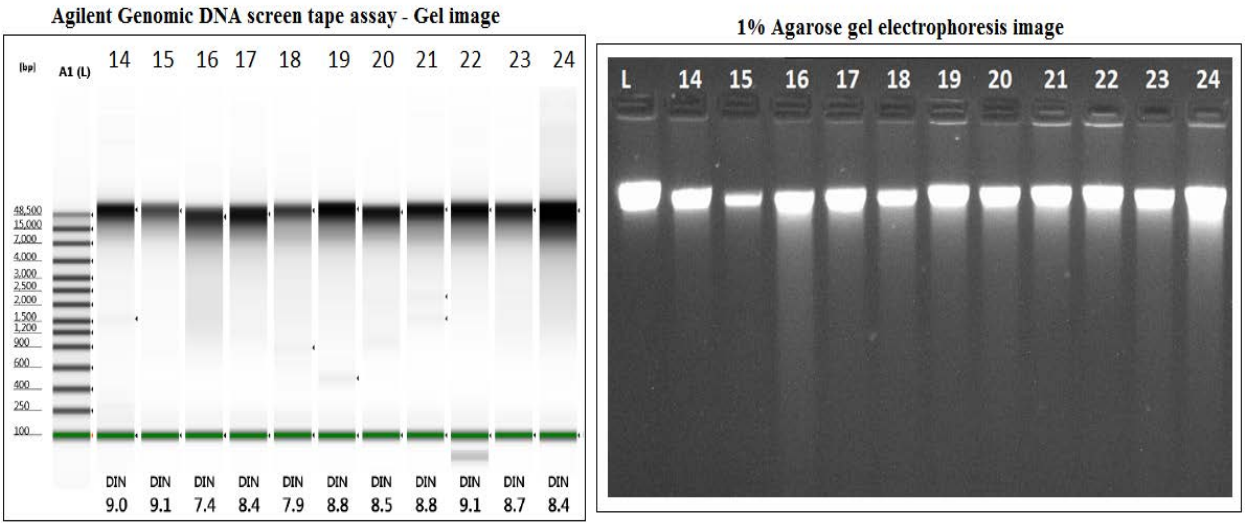


Figure 5.10: The gel images of two different electrophoretic methods used in the current study for assessment of the quality of DNA. Lanes (from left to right): DNA ladder (L), DNA samples 14-24.

5.2.4 Library preparation and quantification

The entire Nextera® Rapid Capture workflow was completed in 3 days generating enriched indexed DNA library ready for sequencing. The first step done was a single step gDNA enzymatic fragmentation followed by ligation of sequence adapters and sample indices. In the current project two sample indexes were added for each sample generating unique id for identification and the possibility to pool 24 samples per run. All the pre-enriched libraries generated in the current study resulted with a peak distribution size ranging from approximately 150 bp –1 kb with the sample peak around 300 bp as described in the protocol (Figure 5.11). This served as an important quality measure, as a larger peak distribution (>350 bp) would be an indication as use of more input DNA which could result in lower on-target reads and smaller peak distribution (<225 bp) would be an indication as use of less input DNA or low quality gDNA which could result in reduced library diversity or elevated duplicates.

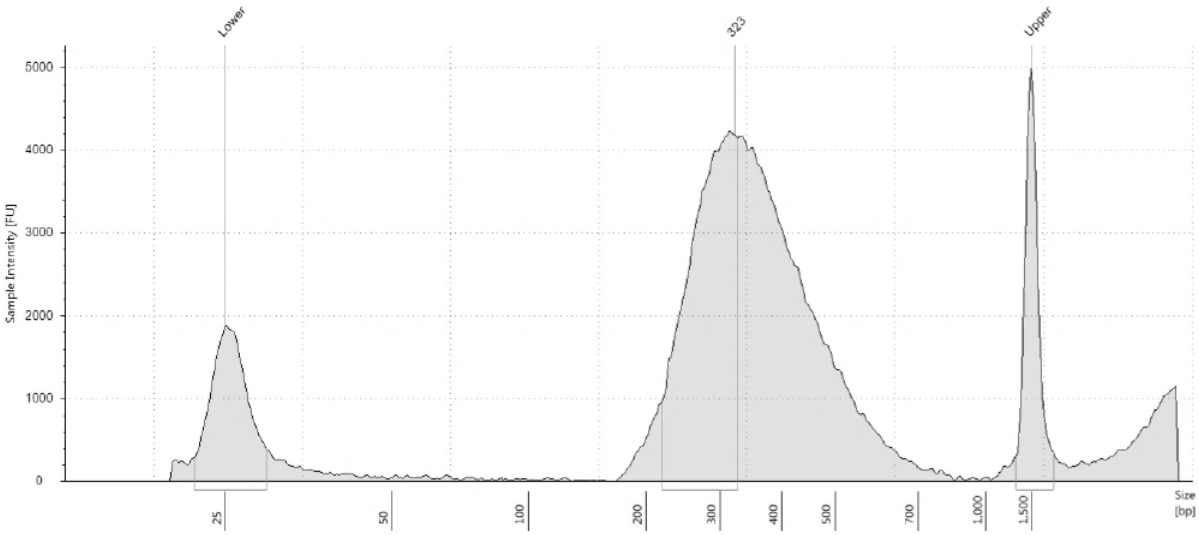


Figure 5.11: Electropherogram showing the Nextera® Rapid Capture Enrichment (NRCE) Post-PCR, Pre-Enriched Library Distribution of a single DNA sample. The peak distribution size ranging from approximately 150 bp –1 kb with the sample peak around 300 bp indicating accurate quantification of samples. The graph represents peak distribution with peak size (bp) on the X-axis and sample inensity (FU, fluorescence units) on the Y-axis, showing also the lower and upper markers peaks at 25 bp and 1500 bp respectively.

The pre-enriched libraries with different indexes were then pooled and at the end of the protocol, enriched indexed DNA library ready for sequencing was generated. Quality control checks for the enriched final library (after pooling of sub-libraries, 24 samples in one pool) generated in the current study resulted with a peak distribution size (insert size) ranging from approximately 200 bp –1 kb with the sample peak around 350 bp as recommended by the manufacturer and there was no contaminating adapter-dimers (Figure 5.12).

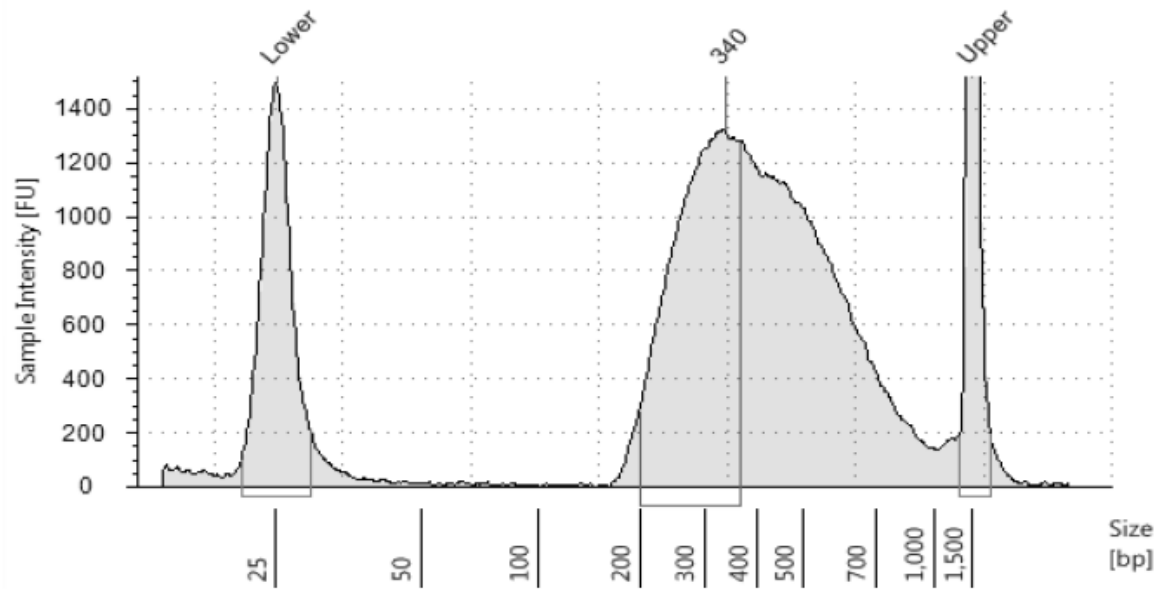


Figure 5.12: Electropherogram showing the NRCE Post-Enrichment (24-plex Enrichment) Library Distribution. The peak distribution size ranging from approximately 200 bp –1 kb with the sample peak around 350 bp indicating accurate quantification of samples. The graph represent peak distribution with peak size (bp) on the X-axis and sample inensity (FU, fluorescence units) on the Y-axis, showing also the lower and upper markers peaks at 25 bp and 1500 bp respectively.

5.2.5 Data analysis

Miseq® control software (MCS) does the primary image analysis and produces base calls and quality scores in the background. The MCS sequencing screen was 'view only' which showed the run progress, intensities, and quality scores (supplementary figure 3). After template generation, Sequencing Analysis Viewer (SAV) provides metrics generated by RTA and organizes the metrics into plots, graphs, and tables (supplementary figure 4). The indexing tab of the SAV lists provides information of the even distribution of the 24 samples (equal amount) in the enriched final library pool assessing the quantification steps done (data not shown).

5.2.6 Run statistics I – Overview of the quality of runs

A total of 12 runs were performed in this study with 24 samples in each run comprising of 286 patients' altogether. For the ID cohort (166 patients) and Short stature cohort (120 patients), a total of seven and five runs were performed respectively (Table 5.5). An average of 1.7 ± 0.5 million total passed filter reads per run was acquired, with approximately 96.7% mapping to the reference genome. A read enrichment (Target aligned reads/Total aligned reads) of about 75% ± 2% and about 85% ± 4% of Padded Read Enrichment (Padded target aligned reads/Total aligned reads) was acquired in the current study. Summary statistics of all the runs performed in the current study displaying the cluster passing filters, Q-scores, and cluster densities are displayed in figure 5.13.

Table 5.5: Run statistics of all the runs performed in the current study. The runs done using the modified protocol are highlighted in red text background. (ID: Magdeburg ID cohort; GH: Short stature cohort; Conc: concentration)

Run name	Final library Conc	Cluster density (K/mm²)	Cluster passing filter	Q Score >=Q30	Yield total (GB)	Error rate	% of gaps from total length of targeted reference
ID_Run1	12 pM	1657	84.7%	90.7%	7.6	0.57%	1.07%
ID_Run2	12 pM	1642	84.3%	90.8%	7.4	0.54%	1.16%
ID_Run3	12 pM	1291	91.7%	93.7%	7.1	0.51%	1.30%
ID_Run4	12 pM	1329	91.2%	93.1%	7.3	0.50%	1.14%
GH_Run1	12 pM	820	96.5%	96.4%	4.8	0.47%	1.73%
GH_Run2	12 pM	776	96.24%	96.2%	4.5	0.47%	1.69%
GH_Run3	12 pM	1020	93.2%	94.4%	5.7	0.48%	1.40%
GH_Run4	12 pM	1149	93.1 %	93.0%	6.4	0.51%	1.30%
ID_Run5	12 pM	1339	88.6%	83.1%	7.0	0.85%	1.21%
ID_Run6	12 pM	1463	86.3%	72.2%	7.4	2.59%	1.52%
GH_Run5	12 pM	1527	86.1%	80.6%	7.7	1.10%	1.38%
ID_Run7	15 pM	1251	90.7%	81.2%	6.8	1.05%	1.75%

- Final library concentration - The amount of final library pool (24 samples) used for sequencing runs.
- Cluster Density (K/mm²) - Shows the number of clusters per square millimeter for the run. (The density of clusters detected by image analysis, +/- one standard deviation).
- Clusters Passing Filter (%) - Shows the percentage of clusters passing filter based on the Illumina chastity filter, which measures quality.
- %Q>=30 - The percentage of bases with a quality score of 30 or higher, respectively. Higher scores indicate higher confidence in the variant and lower probability of errors. For a quality score of Q, the estimated probability of an error is $10^{-(Q/10)}$. For example, the set of Q30 calls has a 0.1% error rate.
- Yield total - The number of bases sequenced
- Error Rate - The calculated error rate, as determined by the PhiX alignment. The alignment of a PhiX spike-in as an external control to measure the percentage of reads with 0–4 mismatches, providing a direct measurement of the intrinsic error rate.
- Total length of targeted reference - Total length of sequenced bases in the target reference.

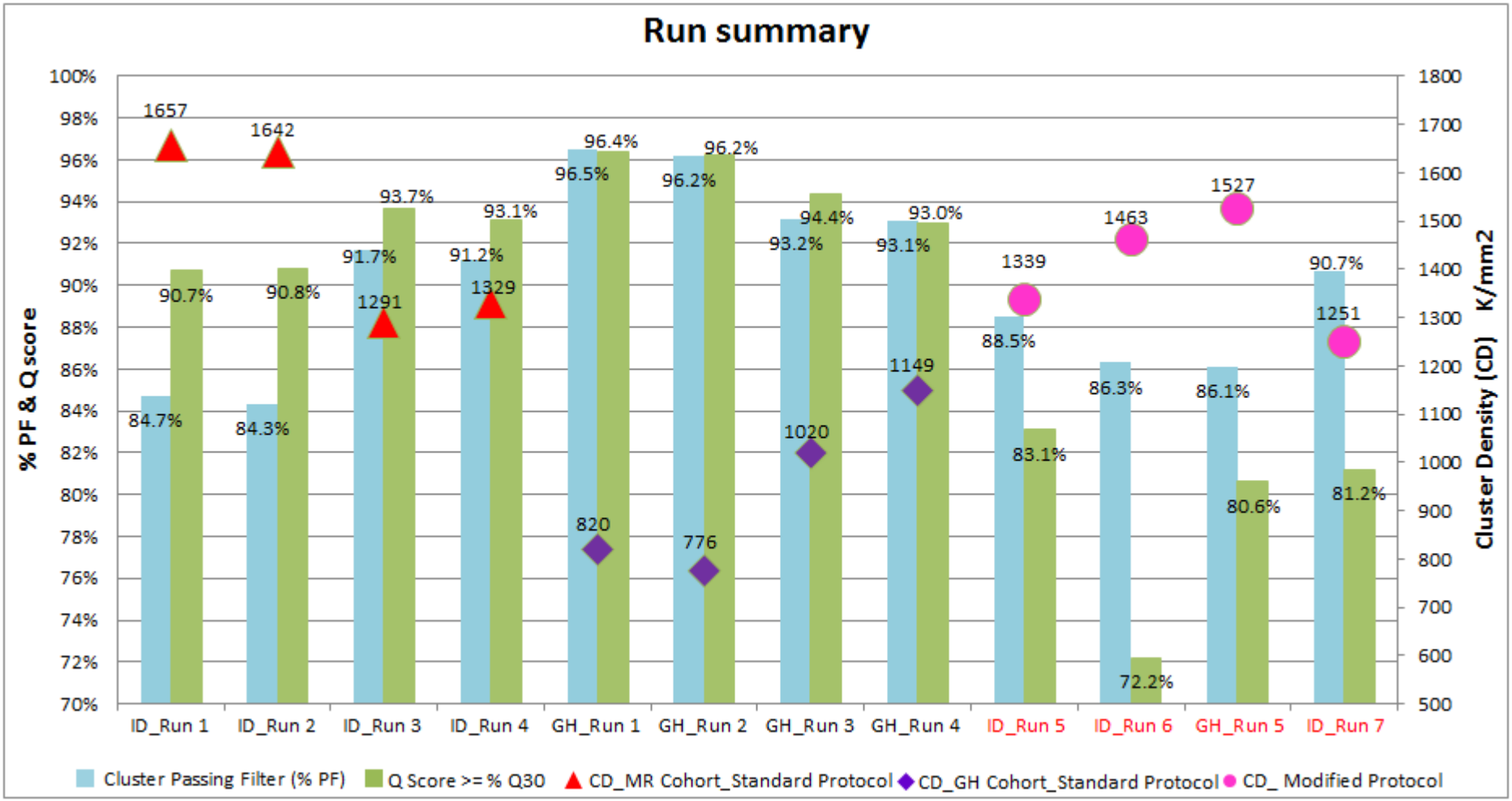


Figure 5.13: Summary statistics of all the runs performed in the current study displaying the cluster passing filters, Q-scores, and cluster densities (CD). The cluster densities between normal/modified protocols as well as between ID/ Short stature (GH) cohorts are indicated in detail by colour coding.

In the current study about >1200 k/mm2 clusters (770–1660 k/mm2) were generated and no major difference in the cluster passing filters was observed between the runs. All the standard protocol runs had a Phred Q score >= 30 values, a mean of more than 90% of sequenced bases, reflecting the high quality of the runs and increased probability of correct base calling whereas reduced Qscores for modified protocol runs was observed.

The average mean depth for the targeted regions was 151.6 ± 51.6 with approximately 90.2% of Percent Q30 (Figure 5.14) for all the runs performed in this study. The uniformity of coverage (Pct > 0.2*mean) (the percentage of targeted base positions in which the read depth is greater than 0.2 times the mean region target coverage depth) was 92.5 ± 1.1% (Figure 5.14). The run statistics per run is shown completely in Supplementary Table S7.

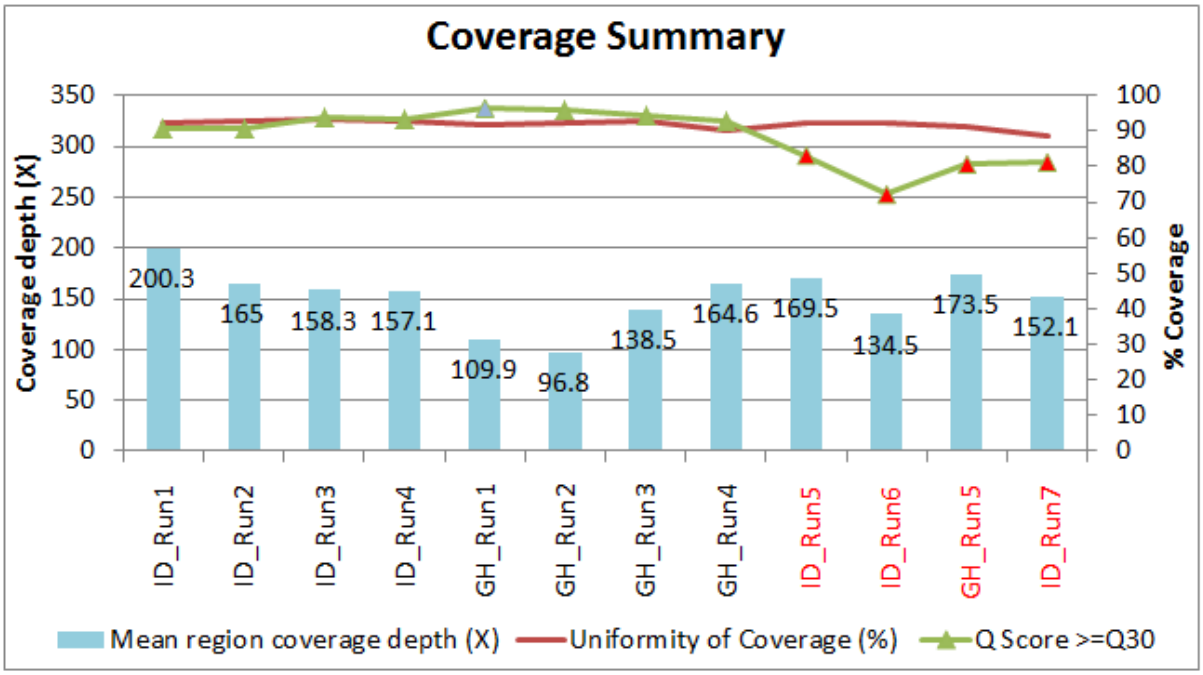


Figure 5.14: Coverage summary of all the runs in the current study. The coverage depth (number of reads per region) with percentage of uniformity of coverage and Q score values are displayed. The runs done using the modified protocol are highlighted in red text background. (ID: Magdeburg ID cohort; GH: Short stature cohort).

On average the target coverage at 1x, 10x, 20x, 50x was 98.9 ± 0.3%, 96.2 ± 1.1%, 93.6 ± 2.1% and 83.8 ± 5.9%, respectively. Target coverage graphs displaying percentage targets with coverage greater than 1x and 20x for all the runs in the current study are shown in Figures 5.15 and 5.16 respectively. Out of all 286 samples, few samples showed low coverages for both 1x and 20x than others probably due to poor sample quality. In particular, 3 samples (2 from ID cohort and one from GH cohort) failed having very low average mean depth (72.3, 65.3 and 23.7) and also very low 1x (<98%) and 20x (<85%) coverages. These 3 samples also had high number of gaps[6] less than 20x percentage coverage with less uniformity of coverage. The failure of these samples is due to poor sample quality, inaccurate quantification of the samples and due to low sample input, the library peak distribution sizes was less than 200 bp resulting in more elevated gaps. The two samples which failed in ID cohort were repeated in another run and achieved promising results.

[6] Given a depth threshold, a gap is defined as a consecutive run of bases in which all bases have coverage less than the threshold. It is in these regions that variants are filtered due to low depth.

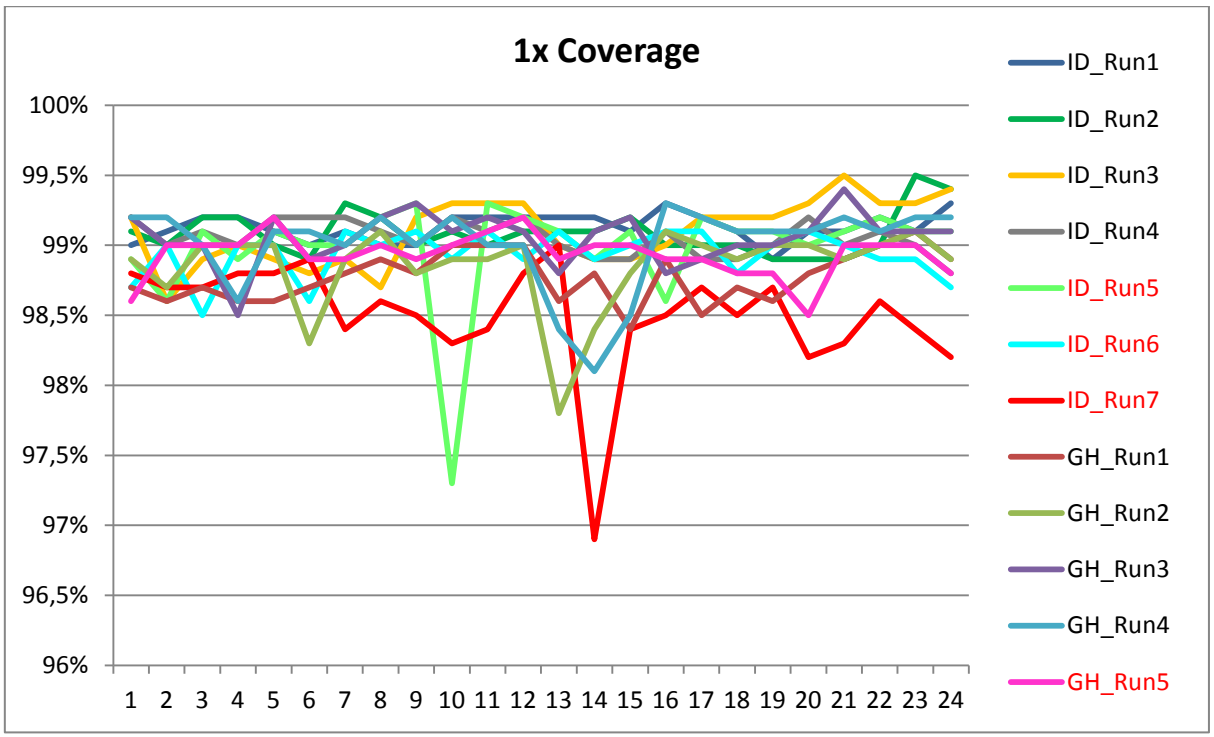


Figure 5.15: Target coverage graph displaying percentage targets with coverage greater than 1x for all the runs in the current study, along for each corresponding sample in individual run. On average 98.9 % of 1x coverage was observed for each sample in the runs except for few outliers. The runs done using the modified protocol are highlighted in red text background. (ID: Magdeburg ID cohort; GH: Short stature cohort).

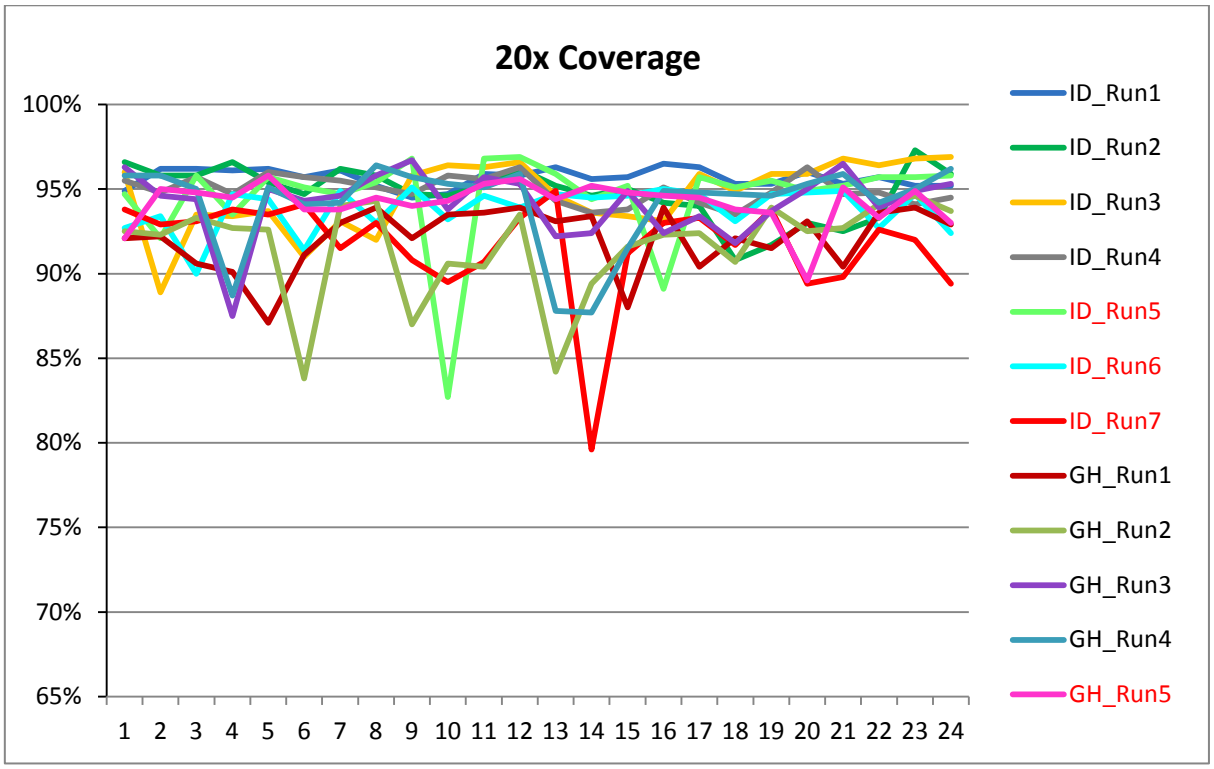


Figure 5.16: Target coverage graph displaying percentage targets with coverage greater than 20x for all the runs in the current study, along for each corresponding sample in individual run. On average 93.8 % of 20x coverage was observed for each sample in the runs except for few outliers. The runs done using the modified protocol are highlighted in red text background. (ID: Magdeburg ID cohort; GH: Short stature cohort).

There were quite a number of gaps with ≤ 20x coverage and the low coverage was due to that the gaps were either regions of high (>60%) or low (<30%) GC content. Target coverage graphs displaying number of gaps less than 20x percentage coverage for all the runs in the current study are shown in Figure 5.17.

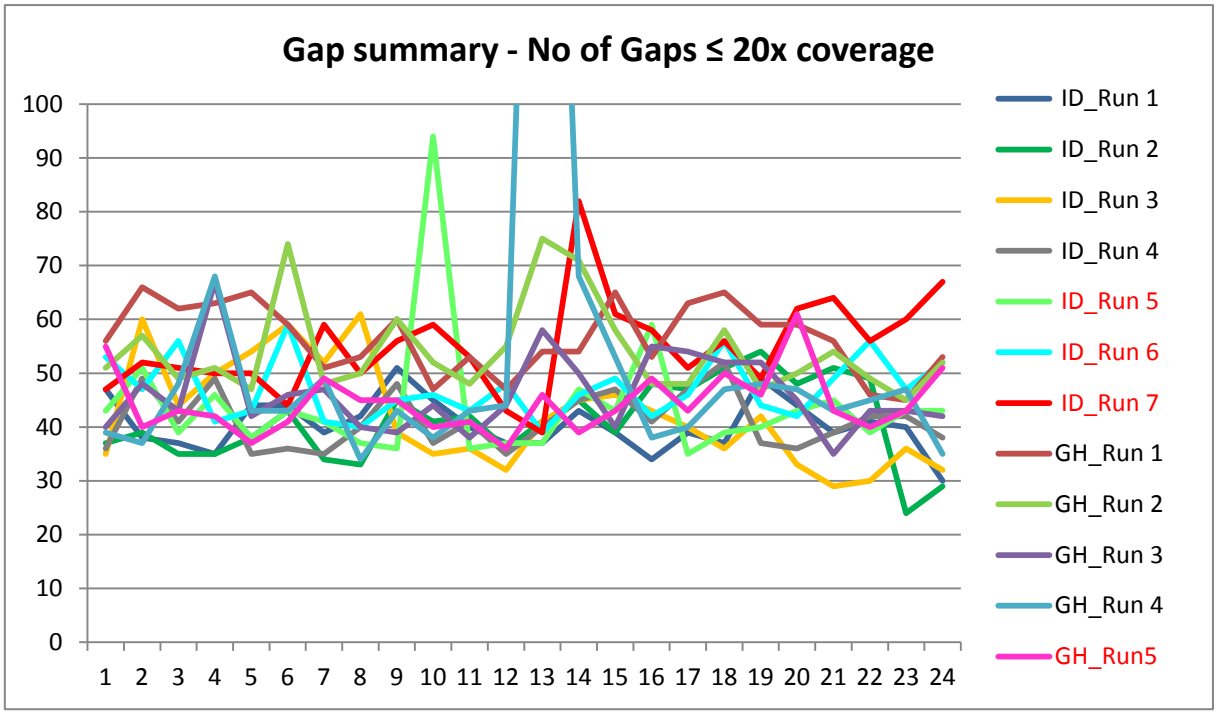


Figure 5.17: Gap summary graph displaying the number of gaps less than 20x percentage coverage for all the runs in the current study, along for each corresponding sample in individual run. On average 42 gaps per sample was present in each run with no significant differences between the modified and standard runs except for few outliers. The runs done using the modified protocol are highlighted in red text background. (ID: Magdeburg ID cohort; GH: Short stature cohort).

An example summary statistics graph showing comparison between standard protocol and modified protocol, for two individual runs performed in the current study, displaying the Q-scores, number of gaps and gap length in bp along for each corresponding sample in individual run is shown in detail in Supplementary figure 6.

5.2.7 Performance of modified protocol

In the current study, a total of 12 runs were performed with 24 samples per run, out of which 8 runs were performed with the standard enrichment protocol (4 each of ID and Short stature cohort) as instructed by the manufacturer. The remaining 4 runs (3 ID and 1 Short stature cohort) were performed with half reagent protocol, built in-house with a few modifications to the standard protocol. All the modifications were tested initially on five different DNA samples changing various parameters like DNA concentration/ volume; incubation time/temperature and modifications in clean up procedures of the normal protocol. All the modifications done to the protocol is shown completely in Table 4.4. All test volumes

produced libraries with a similarly broad peak ranging from 200-1000 bp but varying in the quantities of the peak, a decrease in the sample intensity (Figure 5.18).

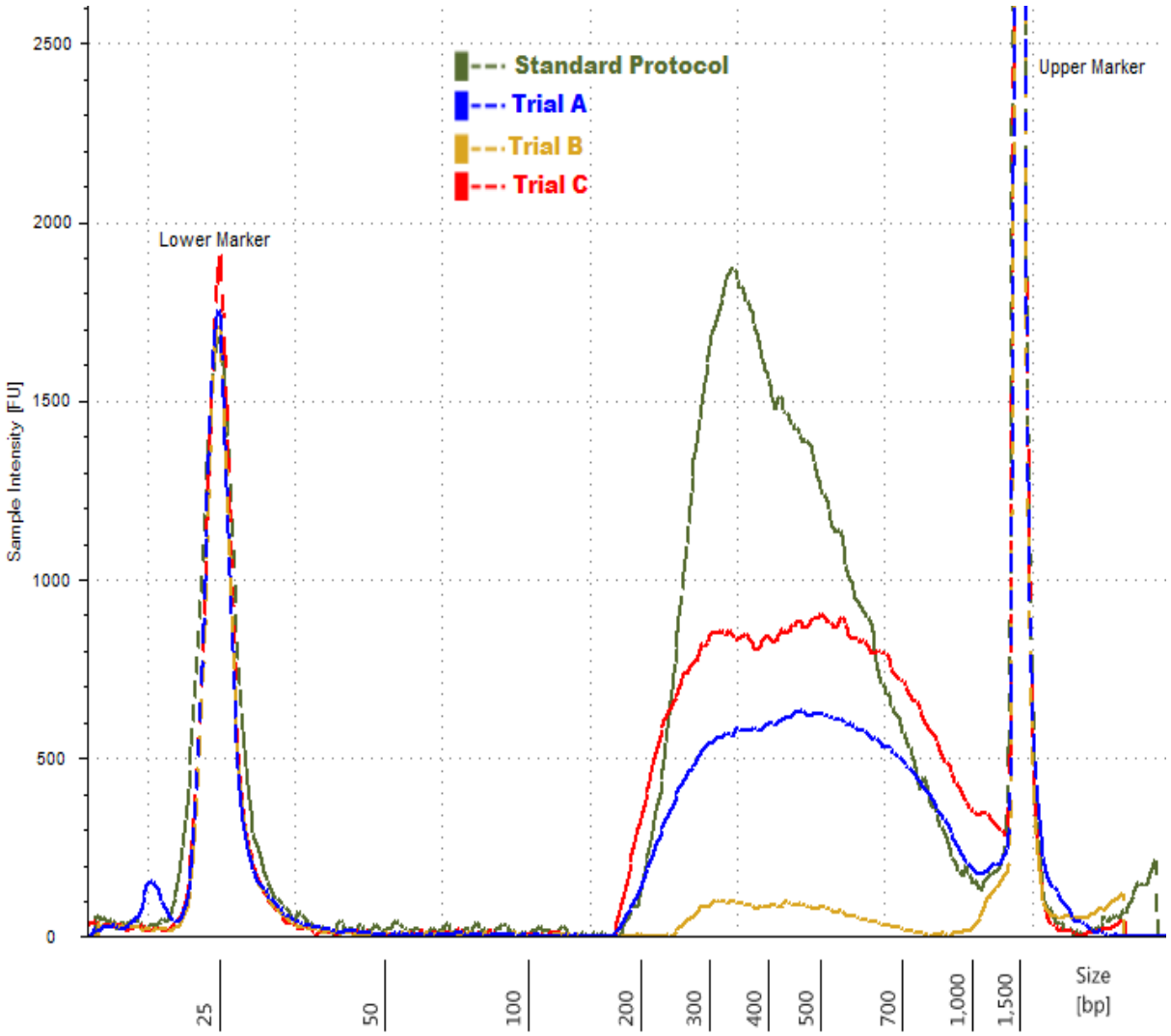


Figure 5.18: Electropherograms showing the NRCE Post-PCR, Pre-Enriched library distribution in comparsion of the standard protocol to the three different trials conducted in the current study. Sample peak intensities constructed using varying volume of Nextera reactions: standard (green), half-volume (orange, Trial B), half-volume with extended incubation time (blue, Trial A) and half-volume with increased PCR cycles (red, Trial C). The graphs represent peak distribution with peak size (bp) on the X-axis and sample inensity (FU, fluorescence units) on the Y-axis, showing also the lower and upper markers.

The Trail C (modified) protocol was used in subsequent runs and it was a success with no major differences or changes in various parameters with the normal protocol. In comparison to the example provided by the manufacturer, the sample peak distribution remains unchanged for both the protocols with the peak intensity at 350bp after post PCR amplification as it is important in determining the eventual distance between the mated pair reads in pair-end sequencing. The only change observed with the modified protocol was the decrease in the sample intensity. Figure 5.19 shows an equal distribution of NRCE Post-enrichment Library (Final library – 24 plex) in both the protocols with no difference.

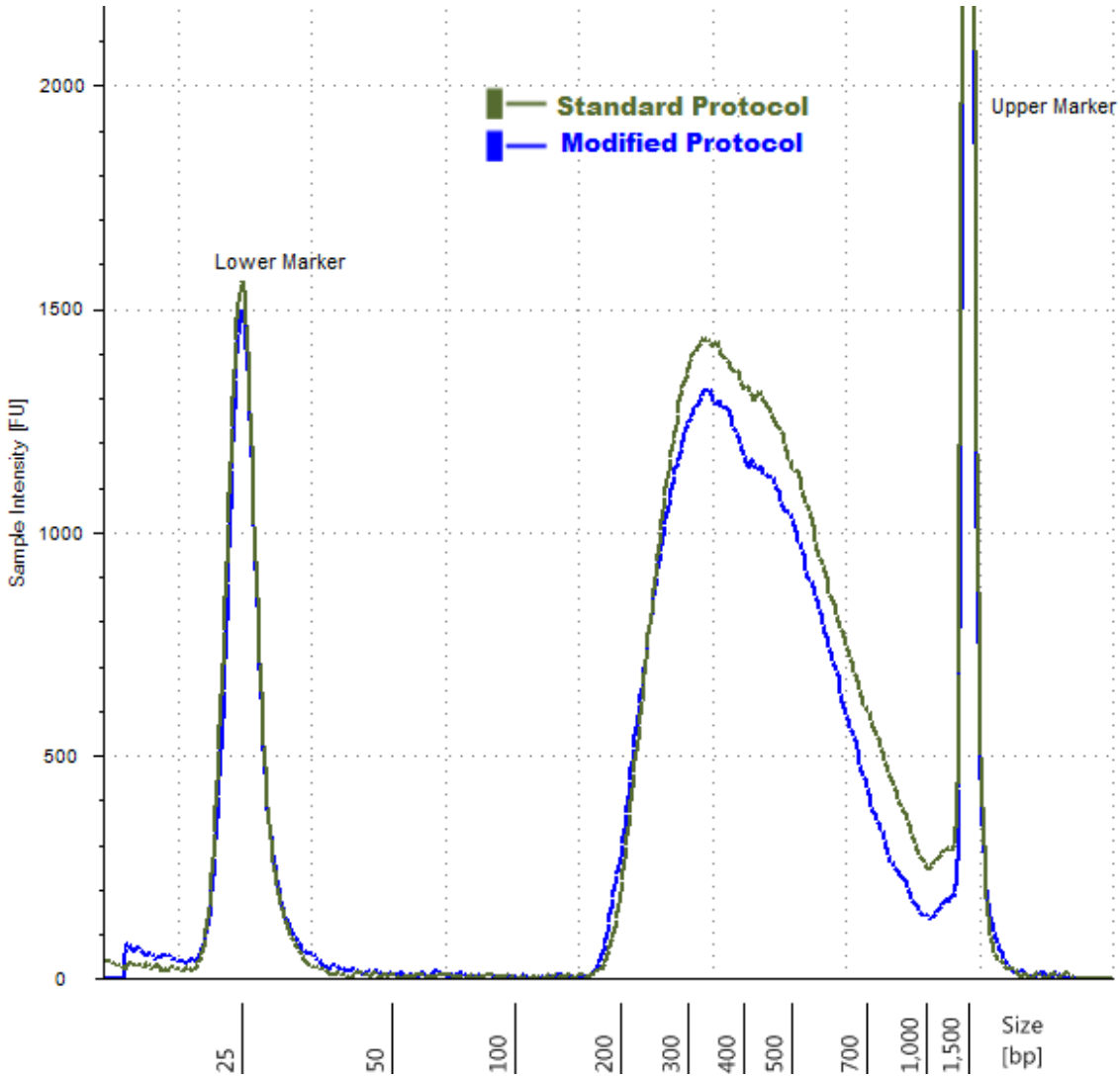


Figure 5.19: Electropherograms showing the NRCE Post-enrichment library (Final library – 24 plex) distribution approximately 200 bp -1 kb in both standard and modified protocol. The graphs represent peak distribution with peak size (bp) on the X-axis and sample inensity (FU, fluorescence units) on the Y-axis, showing also the lower and upper markers.

In the modified protocol runs, there was no complete sample drop for any sample. Only 2 samples (out of 96) had very low average mean depth with very low 1x and 20x coverages. For comparision purposes, various parameters were tested between the standard and modified protocol runs. For obtaining normalised results only four runs (first three runs from ID cohort and fourth run from Short stature cohort) from standard protocol were compared against the four runs of modified protocol. The different parmeters tested were mean region coverage depth (X), 1x coverage, 20x coverage, Q30 scores, number of gaps ≤ 20x coverage, uniformity of coverage (%), % duplicate paired reads, and percent aligned reads. Out of all parameters tested, only Q30 scores and percent aligned reads has shown difference between the two methods which is of statistical significance (Figure 5.20A, 5.20B). Though % duplicate paired reads did not differ much in both the methods but showed a statistical trend towards significance (Figure 5.20C).

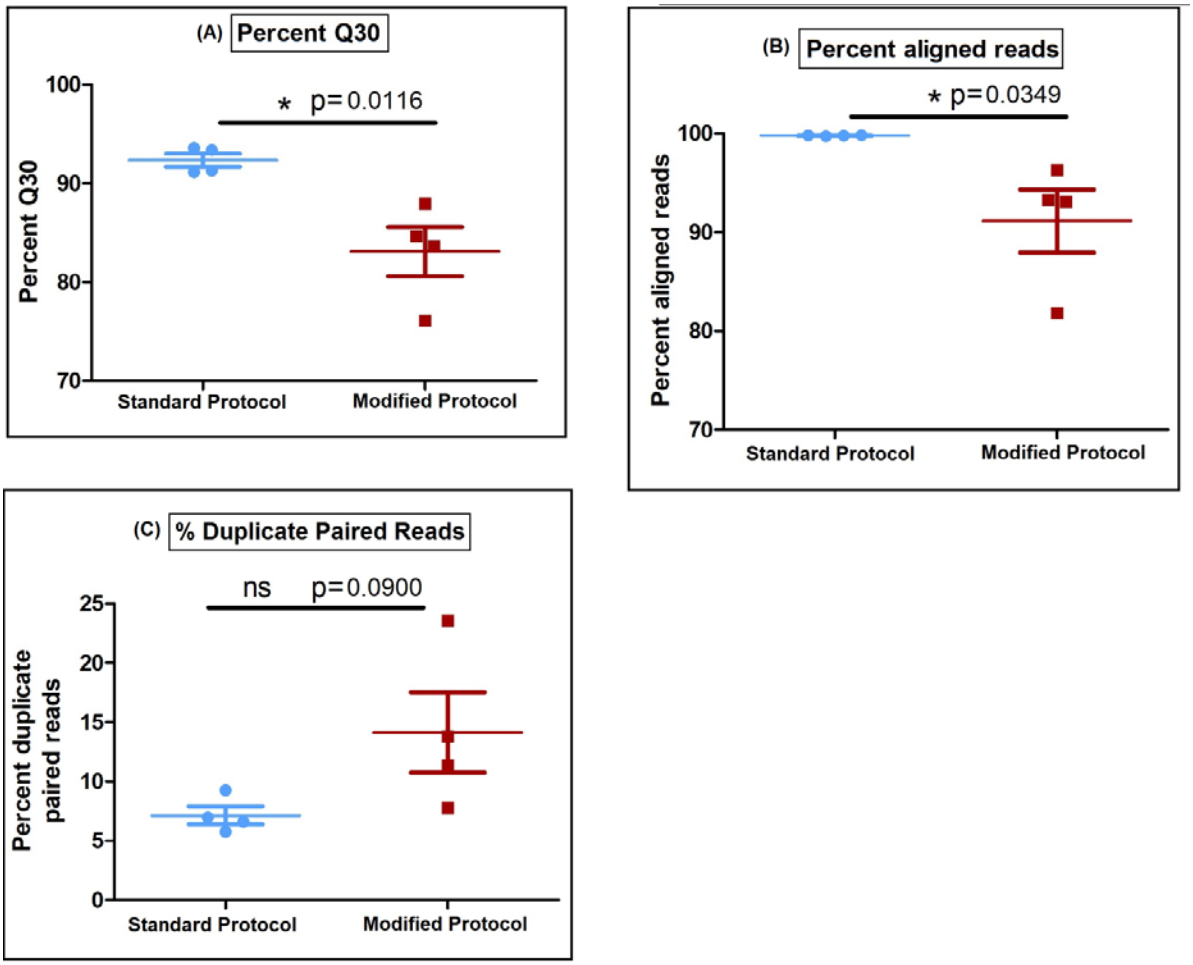


Figure 5.20: Comparison graph showing the (A) Percent Q30 scores (B) Percent aligned reads and (C) percentage duplicate paired reads between the two protocols in the current study. The data points representing the corresponding values obtained in the four runs each of normal protocol and modified protocol. The percent Q30 scores and percent aligned reads scores were lower in the modified protocol runs than the standard runs showing statistical significance. The percentage of duplicate paired reads though did not differ much between the protocols but showed a slight trend towards significance. All values are mean ± SEM, *$P \leq 0.05$, ns= not significant, Student's t-test.

5.2.8 Run statistics II – Variant identification and classification

After variant calling, a total of 72465 genomic variants for the ID Cohort and 49959 genomic variants for the Short stature cohort were identified in the current study with an average around 350-440 variants per sample. For the ~380 variants called per sample, about 68% were synonymous variants and 27% were non-synonymous variants with 82% SNVs already identified in various public databases. The remaining 4% include Indels and loss-of-function variants and 1% of SNVs occurring in intron and UTR regions. By applying all the filtering strategies for identification of very rare and novel variants, a total of 1198 and 780 rare variants in ID and Short stature cohorts were identified respectively (Table 5.6). Of these, 270 and 197 variants were found only once in the ID and Short stature cohort, respectively and were not previously reported in any database. In the current study for all statistical tests and comparisons between the cohorts, we focused mainly on a subset of genes, the ID gene panel comprising of 75 genes.

Table 5.6: Summary statistics of the total number of novel/rare variants in current study.

Cohort name (Number of patients)	Total no of Rare/Novel Variants			Total no of Novel variants	Total no of Rare/Novel variants in ID panel			Total no of ID panel Novel variants	
	All	Non-Syn	Syn		All	Non-Syn	Syn	Non-Syn	Syn
ID (166)	1198	730	468	270	298	167	131	56	15
Short stature (120)	780	471	309	197	184	107	77	47	16

Syn: Synonymous, Non-Syn: Non- Synonymous (missense, nonsense, indel and splice-site mutations)

In order to find possible accumulation of rare variants in certain genes of ID which might explain or provide a hint into the direction of a multifactorial concept, various comparison tests were done between the two cohorts - ID (case group) and Short stature (control group). The various parameters tested were total number of rare/novel variants, pathogenic/tolerated variants, synonymous/non-synonymous variants and other variants frequency (loss-of-function, indels, and splice mutations). A number of different factors were tested like

- Particular gene enrichment between the cohorts – The possibility of statistical enrichment of variants in a certain gene belonging to the RAS pathway was tested between the cohorts. A total of 44 genes belonging to known monogenic ID cause (BRAF, CBL, HRAS, KRAS, MAP2K1/ MAP2K2, SOS1, SPRED1, PTPN11, SHOC2, RIT1, NF1 etc.) were tested individually in both the cohorts. All the genes were statistically tested for various parameters. For particular gene enrichment, no parameter showed significance between the cohorts.
- Total number of variants identified between the cohorts – All the rare variants called after filtering were statistically tested by student's t-test method for various parameters. There was no significant difference observed between the cohorts in any of the parameter tested.
- Total number of variants identified in ID panel - All the variants of ID panel called after filtering were statistically tested by student's t-test method for various parameters. There was no significant difference observed between the cohorts in any of the parameter tested, likely because the study is underpowered to detect such an effect.
- Total number of rare/ novel variants identified in ID panel - The rare/novel variants of ID panel were again tested separately by student's t-test method depending on the mode of inheritance of the ID genes (Figure 5.21). Due to unequal sample numbers in each cohort, they were normalised prior to testing statistical significance. There was no significant difference observed between the cohorts in any of the parameter tested

except for one. An enrichment of non- synonymous variants in recessive ID genes in the ID cohort was seen with statistical significance.

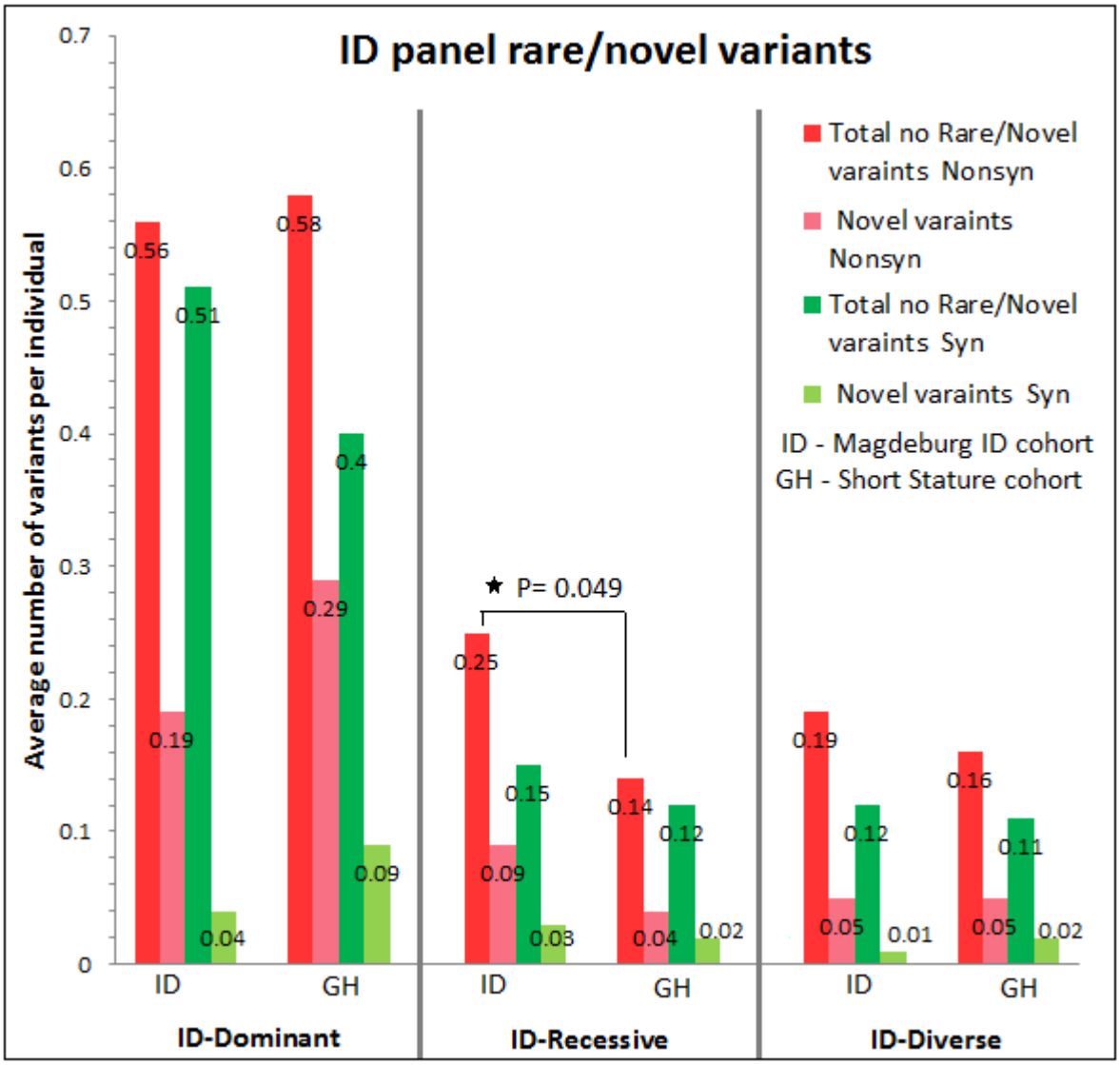


Figure 5.21: Classification of the identified rare/novel variants in the ID panel genes according to their mode of inheritance in both the cohorts. All values are mean ± SEM, *P≤ 0.05, Student's t-test. ID: Magdeburg ID cohort, GH: Short stature cohort, Syn: Synonymous, NonSyn: Non- Synonymous (missense, nonsense, indel and splice-site mutations).

Within the ID cohort on average, in the examined RAS pathway related genes, ~ 4 very rare or unknown variants were identified per individual (Range: 0-12). All potential mutations and variants of unknown significance classified as pathogenic by different filtering criteria were further confirmed by Sanger sequencing. By using the inheritance information derived from parental data, potential variants were refined to a median of 1-2 per sample (range 0–4).

5.2.9 Monogenic disorders - Mutations identified in known ID genes

After variant filtering, the entire 298 candidate variants of the ID panel were manually reviewed in the 166 patients of ID cohort. Several potential variants were observed, and in 2 patients of these, a single, most likely causative genetic change in known ID genes could be

identified. Both the variants identified were classified as pathogenic according to the ACMG standards and guidelines (Richards et al. 2015; Supplementary table 8). The mutations were further confirmed by Sanger sequencing and segregation analysis was also performed depending on the availability of the samples. For the patients, signed consent for scientific evaluation and publication of clinical data was given but permission to publish the patients' photographs could not be obtained.

Patient number- ID-001

We report here a novel nonsense mutation in the gene CTNNB1 (Catenin beta-1) with the sequence variant c.788T>A (p.Leu263*). The variant identified was de novo in the patient but without confirmation of paternity and maternity. The ACMG criteria (PVS1; PM6; PM2; PP3) defined the variant as pathogenic. The CTNNB1 gene is a known ID gene, associated phenotype listed as mental retardation, autosomal dominant 19 (OMIM: 615075).

Phenotype of the patient: The index patient (male) was born after 42 weeks of gestation (birth weight 3310 g, length 50 cm, OFC 33 cm). During pregnancy, amniocentesis was performed due to presence of plexus cysts and an intracardiac white spot, but revealed normal male karyotype. Postnatally he was noted to have pronounced muscular hypotonia. He had a significant delay of motor and speech development, walking at age 4yrs, deficits in fine motoric skills, intellectual disability, sleep disturbance, and autistic features. The MRI showed unilateral focal heterotopias in the periventricular white matter, and he also has dysgenesis of right pupil, ataxic gait with recurrent upper airway infections. At the age of 5y 5m he had measurements of length 112 cm (25-50th centile), weight 19 kg (25-50th centile), and OFC 49.5 cm (3-10th centile). He had some minor abnormal facial features like high forehead, strabismus, preauricular tag at right ear, narrow nasal bridge, pointed nasal tip, flat philtrum, thin upper lip, and talipes equinovarus. He was the second child of healthy unrelated German parents with a healthy sister. The mother of the index patient had a maternal half-sister who had a stillbirth of a boy with holoprosencephaly and cleft lip and palate.

Patient number- ID-002

We report here a missense mutation in the gene BRAF (v-Raf murine sarcoma viral oncogene homolog B) with the sequence variant c.1399T>G (p.Ser467Ala). The ACMG criteria (PS1; PS3; PM1; PM2; PP3; PP4) defined the variant as pathogenic. The BRAF gene is a known ID gene implicated in the phenotype, cardio-facio-cutaneous (CFC) syndrome (OMIM: 115150) and the identified variant has already been implied to be disease causing mutation in CFC syndrome (Rodriguez-Viciana et al., 2006; 2008; Yoon et al., 2007).

Phenotype of the patient: The index patient (male) was born after 40 weeks of gestation (weight 3630 g, length 50 cm, OFC 36 cm) after uneventful pregnancy. He had small atrial

septal defect (ASD) but operation was not necessary. He presented with global developmental delay at the age of 18 months, walking at age 18 months, delay of speech development, deficit of fine motor skills, autistic features, and anxiety. He had seizures at the age of 3yrs and was under medication and became free of seizures. He also had strabismus. At the age of 6y 9mo he had measurements of length 114.5 cm (3-10th centile), weight 25.3 kg (50-75th centile), and OFC 53 cm (50-75th centile). He had stereotypic movements, narrow forehead, lateral thinning of eyebrows, small eyes, hypotelorism, up-slanting palpebral fissures, narrow nasal bridge, anteverted nares, long and deep philtrum, hypotonic mouth, full cheeks, high palate, thin lower lip, large ears, up-turned earlobes, short neck, broad chest, short fingers, syndactyly of toes II/III, curly hair, and 2x0,5 cm CALF right side of back. He was the second child of healthy unrelated German couple. The patient had been seen by a clinical geneticist familiar with RASopathies but no specific suspicion of CFC syndrome was raised at the time of visit. After the diagnosis was established molecularly, clinical photos were reviewed again and the patient was confirmed to lack the typical facial gestalt, but some features (curly hair, bitemporal narrowing, broad chest) were recognized as compatible with a diagnosis of CFC syndrome. Publication of facial photographs was unfortunately not permitted by the parents.

5.2.10 Unclassified novel/rare variants identified in known dominant ID genes

For patients in the ID cohort, a total of 33 novel and rare deleterious/damaging mutations have been identified in genes, which have been already implicated in ID with autosomal dominant inheritance pattern after the second filtering pipeline was used (as described in methods section 4.8). Although protein truncating variants [PTVs (Rivas et al. 2015)] are generally more conclusive than missense mutations, the 17 novel/rare mutations described in the table 5.7 are of particular interest depending on the known nature or function of the protein or the predicted effect of the amino acid substitution on protein function. All the mutations described in table 5.7 were classified as according to the ACMG standards and guidelines (supplementary table 8). A total of eight novel mutations were identified in different patients consisting of seven missense changes and one splice site variant. The mutations were further confirmed by Sanger sequencing and segregation analysis was also performed depending on the availability of the samples. Though classified as pathogenic/ likely pathogenic, they were not rated as clearly causative mutations, as either the clinical features of the patients are not completely compatible to the phenotypes related to the respective gene as described in literature or because the variants were inherited from an apparently healthy parent. Nevertheless, without further functional evidence of the variant to support or confirm pathogenicity, narrowing down as causative mutation is not possible.

Table 5.7: Novel/rare variants identified in known dominant genes for ID or related disorders. The list provides information of patients only from ID cohort.

Patient #	Sex	Gene	Chr Nr	cDNA	Protein	Inheritance	ExAC Frequency; ExAC Allele count	Additional comments	ACMG - criteria	ACMG classification	Phenotype (MIM number)
ID-003	F	DOCK8	9	c.4626+1G>A	p.?	NT	------	------	PM2; PP3; PVS1; PM1	Pathogenic	Mental Retardation, AD 2 (614113)
ID-004	F	DOCK8	9	c.1286G>A	p.Gly429Glu	Ma; sibling	Freq-8.242e-06; Allele count- 1 / 121334	Splice region variant, New Acceptor site. rs776197556	PP3; PM2	VUS	
ID-005	M	KCNQ3	8	c.1091G>A	p.Arg364His	Ma	------	BECTS (Fusco et al., 2015)	PP3; PM2; PM1; PS1	Likely Pathogenic	Seizures, benign neonatal, type 2 (BFNS2) (121201)
ID-006	M	EPB41L1	20	c.1049T>C	p.Ile350Thr	Pa	------	------	PP3; PM2; PM1	VUS	Mental Retardation, AD 11 (614257)
ID-007	M	STXBP1	9	c.1672C>A	p.Gln558Lys	NT	------	------	PP3; PM2; PM1	VUS	Epileptic encephalopathy, early infantile, 4 (612164)
ID-008	F	SOS1	2	c.800T>C	p.Val267Ala	NT in Father	Freq- 8.241e-06; Allele count-1 / 121348	------	PP3; PM2; PM1	VUS	Noonan syndrome 4 (610733)
ID-009	M	DYNC1H1	14	c.7769C>T	p.Pro2590Leu	Pa	------	------	PP3; PM2; PM1	VUS	Mental Retardation, AD 13 with Neuronal Migration Defects (614563)
ID-010	F	DYNC1H1	14	c.12047C>G	p.Ser4016Cys	NT	Freq- 2e-05; Allele count- 2 / 121190	------	PP3; PM2	VUS	
ID-011	M	DYNC1H1	14	c.12811C>T	p.Arg4271Cys	NT in Father	------	rs573730260	PP3; PM2	VUS	
ID-012	F	CACNG2	22	c.638G>T	p.Arg213Leu	Sibling; NT in Parents	Freq-8.292e-06; Allele count- 1 / 120598	rs768099799	PP3; PM2	VUS	Mental Retardation, AD 10 (614256)
ID-013	M	CACNA1G	17	c.3569G>C	p.Arg1190Pro	NT	Freq-5.219e-05; Allele count- 6 / 114954	rs199761120	PP3; PM1	VUS	Spinocerebellar ataxia 42 (616795)
ID-014	M	CACNA1G	17	c.5510C>T	p.Thr1837Met	NT	Freq- 4.979e-05; Allele count- 6 / 120512	rs202107134	PP3; PM1	VUS	

ID-015	F	BRAF	7	c.2131C>T	p.Leu711Phe	NT	------	------	PP3; PM2; PM1	VUS	Cardiofaciocutaneous syndrome (115150); Noonan Syndrome 7 (613706); LEOPARD Syndrome 3 (613707)
ID-016	M	SHANK3	22	c.845C>T	p.Ser282Leu	NT	------	------	PP3; PM2	VUS	Phelan-McDermid syndrome (606232)
ID-017	F	SHANK3	22	c.4208C>T	p.Ser1403Leu	Pa	Freq- 4e-05; Allele count-2/ 49416	------	PP3	VUS	
ID-018	F	NTRK2	9	c.1075T>C	p.Tyr359His	Ma	------	------	PP3; PM2; PM1	VUS	Obesity, hyperphagia, and developmental delay (613886)
ID-004	F	NTRK2	9	c.2375G>A	p.Arg792His	Pa	------	------	PP3; PM2; PM1; PP2	Likely Pathogenic	

- number, Chr Nr- chromosome number, Magdeburg ID cohort (ID), F- Female, M- Male, Ma- Maternal, Pa – Paternal, NT – Not tested, Autosomal dominant (AD), VUS- Variant with uncertain significance, BECTS- benign childhood epilepsy with centrotemporal spikes.

A total of eight novel mutations were identified in known ID genes consisting of 7 missense changes and 1 splice site variant. With the use of available online prediction tools, the DOCK8 splice site variant, c.4626+1G>A showed prediction of most probably affecting splicing by altering the WT donor site and could not be proved as de novo due to the unavailability of parental DNA samples. Another rare DOCK8 variant c.1286G>A is classified as missense mutation but due to its position (splice region variant) and according to prediction tools, it may also affect splicing The KCNQ3 variant classified as likely pathogenic has been described by Fusco et al., 2015 in one patient suffering from BECTS (benign childhood epilepsy with centrotemporal spikes) with no further functional evidences and the variant was neither found in ExAC nor 1000G. The rare variants described here project as potential causative variants due to their low population frequency and allele count. All the described variants were predicted to have pathogenic effect obtained through multiple computational evidence support, but according to the ACMG standards and guidelines they are classified as uncertain significance.

5.2.11 Unclassified novel/rare variants identified in known autosomal recessive/ X-linked ID genes

In the current study, we found 18 rare/novel variants in already known autosomal recessive (AR) / X-linked ID genes in few patients of the ID cohort after second filtering step. A total of 7 mutations were identified affecting evolutionarily conserved amino acid residues and are of potential pathogenic significance predicted to affect protein function by widely used prediction algorithms in the current study (Table 5.8). Four mutations were found related to known AR genes which consist of one stop variant and three missense variants in heterozygous state. Three missense mutations were found related to known X-linked ID genes, out of which two were novel hemizygous missense variants. All the mutations described in table 5.8 were classified as according to the ACMG standards and guidelines (supplementary table 8). The mutations were further confirmed by Sanger sequencing and segregation analysis was also performed depending on the availability of the samples. For the identified heterozygous mutations having recessive pattern, a possible second mutation could not be identified in the patients for the respective genes, which places the variants in questionable position of whether the identified mutation really contributed to the phenotype.

5.2.12 Unclassified novel/rare variants identified in genes that only have association findings with ID

In the current study, we also found nine rare/novel variants in genes which have been implicated in ID through associated studies, position in microdeletions, data from animal models etc., without confirmation of their role by disease-causing point mutations. A total of 7 missense mutations were identified of potential pathogenic significance, out of which three were in SHANK 1 & 2, two each in CTNND2 and AKT1 (Table 5.9). The mutations were further confirmed by Sanger sequencing and segregation analysis was also performed depending on the availability of the samples. Previously mutations in SHANK genes were shown to be detected in a whole spectrum of autism with a gradient of severity in cognitive impairment (Grozeva et al., 2015; Berkel et al., 2010). The clinical pertinence of SHANK1 and SHANK2 genes still remains to be ascertained due to the rare frequency of mutations identified in these genes when compared to SHANK3. With the advent of new technologies, gene variants in CTNND2 were linked to many complex human disorders like Alzheimer's disease, other neurological disorders like bipolar disorder, schizophrenia, autism with ID and it is has been postulated that CTNND2 haploinsufficiency may play a role for ID in Cri-du-chat syndrome (Lu, Q et al., 2016; Hofmeister et al., 2015). AKT1 gene variants have been shown to be associated with schizophrenia along with mood disorders and in Cowden syndrome with mild to moderate ID.

Table 5.8: Novel/rare variants identified in known recessive/ X-linked genes for ID or related disorders. The list provides information of patients only from ID cohort.

Patient #	Sex	Gene	Chr Nr	cDNA	Protein	Inheritance	ExAC Frequency; ExAC Allele count	Additional comments	ACMG - criteria	ACMG classification	Phenotype (MIM number)
ID-019	F	CC2D1A	19	c.1516C>T	p.Gln506*	Pa	------	------	PM2; PP3; PVS1	Pathogenic	Mental Retardation, AR 3 (608443)
ID-020	F	CC2D1A	19	c.2302C>T	p.Arg768Trp	NT	Freq- 1.658e-05; Allele count- 2 / 120656	ExAC allele count in Asian population	PP3; PM2; PM1	VUS	
ID-021	F	ARFGEF2	20	c.3010C>G	p.Leu1004Val	NT	------	Splicing- New donor site is created at 3' end	PP3; PM2	VUS	Periventricular heterotopia with microcephaly (608097)
ID-022	F	ARFGEF2	20	c.2191G>A	p.Glu731Lys	NT	Freq- 7e-05; Allele count-8/ 121382	Splicing- Cryptic acceptor strongly activated	PP3; PM2; PM1	VUS	
ID-023	M	GRIA3	X	c.343T>C	p.Ser115Pro	Ma	------	Hemizygous variant	PP3; PM2; PM1	VUS	Mental retardation, X-linked 94, recessive (300699)
ID-024	M	ARHGEF6	X	c.1999T>A	p.Cys667Ser	NT	------	Hemizygous variant	PP3; PM2; PM1	VUS	Mental retardation, X-linked 46, recessive (300436)
ID-025	F	ARHGEF6	X	c.1763C>T	p.Pro588Leu	NT	Freq-2.28e-05; Allele count- 2 / 87715	------	PP3; PM2; PM1	VUS	

- number, Chr Nr- chromosome number, Magdeburg ID cohort (ID), F- Female, M- Male, Ma- Maternal, Pa – Paternal, NT – Not tested, Autosomal recessive (AR), VUS- Variant with uncertain significance, *Termination- codon (stop codon)

The majority of variants identified in the genes were missense changes, with only one nonsense mutation. The nonsense mutation in CC2D1A c.1516C>T, was identified in heterozygous state inherited from father and a further second mutation could not be identified. The GRIA3 mutation c.343T>C was inherited from the mother, so X-chromosome inactivation (XCI) test need to be done further to confirm pathogenicity. While for the others segregation could not be tested due to unavailability of parental DNA samples.

Table 5.9: List of novel/rare variants identified in genes associated with ID or related disorders. The list provides information of variants of patients only from ID cohort.

Patient #	Sex	Gene	Chr Nr	cDNA	Protein	Inheritance	ExAC Frequency; ExAC Allele count	Additional comments	ACMG - criteria	ACMG classification	Phenotype (MIM number)
ID-023	M	SHANK2	11	c.1211G>A	p.Arg404His	Pa	------	predisposition to ASD or ID (Grozeva et al., 2015; Berkel et al., 2010)	PP3; PM2; PM1	VUS	{Autism susceptibility 17} (613436)
ID-026	F	SHANK2	11	c.5519T>C	p.Ile1840Thr	Ma; siblings	------		PP3; PM2	VUS	
ID-027	M	SHANK1	19	c.5531C>T	p.Pro1844Leu	NT	Freq-2.259e-05; Allele count- 2 / 88540	rs768792411. ExAC allele count in South Asian population	PP3; PM1	VUS	No OMIM phenotype
ID-028	M	CTNND2	5	c.1960C>T	p.Arg654Trp	Ma	------	Autism (Turner et al., 2015), ID (Hofmeister et al., 2015)	PP3; PM2; PM1	VUS	No OMIM phenotype
ID-029	M	CTNND2	5	c.1997C>T	p.Ser666Leu	NT	------		PP3; PM2; PM1	VUS	
ID-030	M	AKT1	14	c.1112C>G	p.Thr371Arg	NT	------	Splicing- New acceptor site	PP3; PM2; PM1; PP2	Likely Pathogenic	Cowden syndrome 6 (615109); {Schizophrenia, susceptibility to} (181500)
ID-031; 032	F, F	AKT1	14	c.533T>C	p.Met178Thr	NT	Freq-8.26e-06; Allele count- 1 / 121062	rs769619023	PP3; PM2; PM1; PP2	Likely Pathogenic	

- number, Chr Nr- chromosome number, Magdeburg ID cohort (ID), F- Female, M- Male, Ma- Maternal, Pa – Paternal, NT – Not tested, VUS- Variant with uncertain significance

All the variants identified in the genes were missense changes. The SHANK2 variant, p.Arg404His was inherited from a healthy father. The SHANK2 variant, p.Ile1840Thr was inherited from an intellectually disabled mother and also present in other siblings of the patient who had developmental delay. One CTNND2 variant was inherited from a healthy mother while the other could not be tested. Segregation analysis for AKT1 variants could not be tested due to unavailability of parental samples.

5.2.13 Potentially disease-causing variants in genes not previously linked to ID

Among the several mutations found in the current study, few of variants in specific genes were particularly interesting based on their nature of their function specific to neuron or brain and also some of these genes were functionally linked to known ID genes formulating them as candidate genes for ID (Table 5.10). They enact as probable pathogenic variants as most of the proteins encoded by these genes interact with either products of known ID genes or shown to have some neurological dysfunction in animal models. All the mutations described were classified according to the ACMG standards and guidelines (supplementary table 8).

Table 5.10: List of novel variants identified in genes not previously linked to ID.

Patient #	Sex	Gene	Chr Nr	cDNA	Protein	Inheritance	ACMG - criteria	ACMG classification	Phenotype (MIM number) (Inheritance)
ID-033	F	MAPK3	16	c.999C>A	p.Tyr333*	NT	PM2;PP3; PVS1; PM1	P	------
ID-034	M	NPTN	15	c.475C>T	p.Arg159*	Ma	PM2;PP3; PVS1	P	------
ID-024	M	TBR1	2	c.1275C>A	p.Tyr425*	NT	PM2; PP3; PVS1	P	-----
ID-035	M	TBR1	2	c.739A>G	p.Thr247Ala	NT	PP3; PM2; PM1	VUS	------
ID-036	M	TBR1	2	c.418G>C	p.Gly140Arg	NT	PP3; PM2	VUS	------
ID-037	M	SYT12	11	c.1063A>T	p.Ile355 Phe	Ma	PP3; PM2; PM1	VUS	------
ID-038	M	BSN	3	c.6325G>C	p.Asp2109His	NT	PP3; PM2; PM1	VUS	-------
ID-039	M	BSN	3	c.11166G>C	p.Lys3722Asn	NT	BP4; PM2; PM1	VUS	-------
ID-040	M	LRRC7	1	c.3388A>T	p.Arg1130Trp	Pa	PP3; PM2	VUS	-------
ID-041	M	LRRC7	1	c.3014A>G	p.Asp1005Gly	NT	PP3; PM2	VUS	-------
ID-042	M	NRXN2	11	c.2032G>T	p.Gly678Trp	father NT	PP3; PM2; PM1	VUS	---------
ID-043	M	LIMK1	7	c.655G>T	p.Val219Phe	NT	PP3; PM2; PM1	VUS	---------
ID-044	M	LIMK1	7	c.1028A>T	p.Glu343Val	NT	PP3; PM2; PM1	VUS	---------
ID-045	M	NRG1	8	c.1475G>A	p.Arg492Gln	NT	PP3; PM2; PM1	VUS	{?Schizophrenia, susceptibility to} (603013)
ID-013	M	BDNF	11	c.478C>T	p.Arg160Trp	NT	PP3; PM2; PM1	VUS	{Memory impairment, susceptibility to}
ID-046	F	PCLO	7	c.15098G>A	p.Cys5033Tyr	NT	PP3; PM2; PM1	VUS	?Pontocerebellar hypoplasia, type 3 (608027) (AR)
ID-047	M	PCLO	7	c.460C>T	p.Pro154Ser	NT	PP3; PM2; PM1	VUS	
ID-014	M	PCLO	7	c.12209A>G	p.Glu4070Gly	Ma	PP3; PM2; PM1	VUS	

ID-009	M	CNTN1	12	c.1139A>T	p.Asp380Val	Ma	PP3; PM2; PM1	VUS	?Myopathy, congenital, Compton-North (612540) (AR)
ID-003	F	CNTN2	1	c.1247C>A	p.Ala416Asp	NT	PP3; PM2; PM1; BP1	VUS	?Epilepsy, myoclonic, familial adult, 5 (615400) (AR)

Chr Nr- chromosome number, Freq- frequency, ID- Magdeburg ID cohort; F- Female, M- Male; *Termination-codon (stop codon), Ma- Maternal, Pa – Paternal, NT – Not tested, VUS- Variant with uncertain significance, P-Pathogenic, AR – Autosomal Recessive

The majority of variants identified in potential new candidate genes were missense changes, while 3 variants were predicting loss of function due to premature translation stop codons. However, one of them was inherited from a healthy parent while for the other two segregation could not be tested. For few of the genes in the list, a possible human phenotype was postulated on the basis of finding of mutations in a single family. All the missense variants were predicted to have pathogenic effect obtained through multiple computational evidence support, but according to the ACMG standards and guidelines they are classified as uncertain significance.

5.2.14 Case studies: Patients with Multiple Variants

For some patients after filtering, around 2-4 putative deleterious/damaging mutations have been identified which fulfilled all classification criteria used in this study and also segregation in the families (for available samples) was also observed. Nevertheless without functional evidence to support pathogenicity, narrowing down to single potential candidate mutation is not possible.

Case Study 01 (CS01): ID-026

The index patient (I) is a 9-year-old girl born after 39 weeks of gestation after uneventful pregnancy. At the time of birth, weight was 2820 g with a height 54 cm. At age 18-24 months' developmental delay became obvious along with speech delay. Her academic skills remained impaired and she received permanent educational support at special school (GB school; Geistig Behinderte). She is a friendly and socially well integrated girl with no malformations, no seizures and otherwise healthy. At the time of clinical visit at the age of 6½ years her weight was 20.5 kg (~3. centile), height 111.8 cm (3.-10. centile) and 50 cm OFC (10.-25. centile). She displayed mild muscular hypotonia, but otherwise no neurologic deficit. The family of the patient is remarkable with a healthy father, intellectually disabled mother and maternal grandmother, and three brothers with developmental delay and behavioural anomalies. All three boys are otherwise healthy and have no malformations. The half-sister of the patient's mother has learning problems, too.

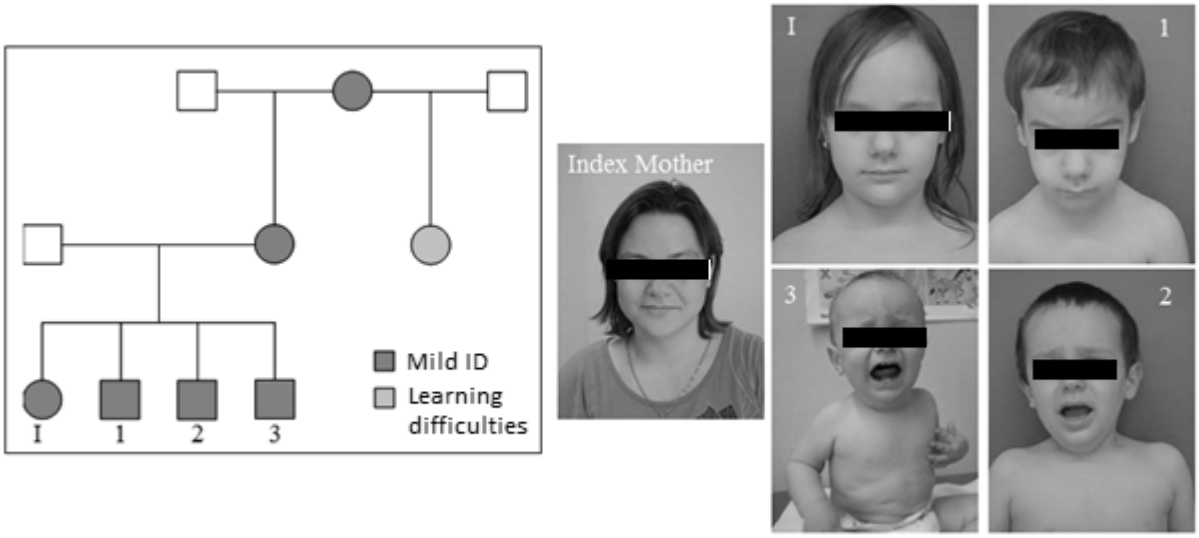


Figure 5.22: Pedigree and Clinical photos of the family. (A) Pedigree of the family (B) Clinical photos of (I) Patient ID-025 at the age of 6 years 6 months (1) Sibling 1 of index at age 4 years 2 months (2) Sibling 2 of index at age 3 years 6 months (3) Sibling 3 of index at age 15 months and their mother.

Table 5.11: Molecular findings in Patient ID-026 in the current study with NGS panel.

Gene	Chr no	cDNA	Protein	ExAC All Freq	PP	Inheritance	ACMG - criteria	ACMG classification
SHANK2	11	c.5519T>C	p.Ile1840Thr	-----	D	Ma; S1; S3	PM2; PP3	VUS
JAK2	9	c.975_976 insGTCA	p.Ile328Glnfs*3	-----	D	Ma; S1; S3	PP3; PM2	VUS

Pathogenicity Prediction (PP) - D=deleterious; T=Tolerated; Chr Nr- chromosome number; Freq- frequency; Ma- Maternal; Pa- Paternal; S – Sibling; VUS - Variant with uncertain significance; *Termination- codon (stop codon).

In CS01, two heterozygous novel mutations were detected, one missense in SHANK2 (SH3 and multiple ankyrin repeat domains 2) gene and one frameshift mutation in JAK2 (Janus kinase 2) gene. SHANK2 (also known as ProSAP1) is a scaffolding protein present at PSD which interacts with various proteins at the synapse with high expression levels in brain. SHANK2 mutations have been shown associated with ID, schizophrenia and ASD (Berkel et al., 2010) and knockout mouse models for Shank2 showed deficits in learning, social behaviours and synaptic plasticity defects (Lim et al., 2017). In the mouse model described by Schmeisser et al., 2012, the hyperactive Shank2-mutant mice had reduced spine density and body weight with abnormal social behaviour and also upregulation of ionotropic glutamate receptors was seen at the synapse. JAK2 is involved in cytokine receptor signalling which is predominantly expressed in the brain (present at PSD) and has been shown to have involvement in hippocampal synaptic plasticity and in modulation of learning and memory (Donzis & Tronson, 2014; Nicolas et al., 2012). From the gene-gene networks it is evident that JAK2 interacts with SHANK2 with intermediate partners all involved in growth hormone signalling pathway and MAPK pathway. Both SHANK2 and JAK2 had low probability of

loss-of-function intolerance suggesting an increased evidence of haploinsufficiency effect. Further testing in additional family members need to be done to confirm pathogenicity and for better segregation analysis.

Case Study 02 (CS02): ID-023

The index patient (male) was born after uneventful pregnancy. At the time of birth, weight was 3040 g with a height 50 cm and OFC 33 cm. He had developmental delay and started walking at age of 13 months. His first words started at age of 18 months and only used single words for a very long time. He showed behavioural anomalies, autistic behaviour, and temper tantrums with aggressive outbursts, attention deficit, hyperactivity, and sleeping disorder but otherwise healthy with no major malformations. He attends special school and at the time of clinical presentation he was 7 years 9 months of age with measurements: height 127 cm (25.-50. centile), weight 36.1 kg (90.-97. centile) and 50 cm OFC (~3. centile). The parents are healthy and mother originates from Russia. He has a healthy half-sister from her mother.

Figure 5.23: Clinical photos of (I) Patient ID-023 at the age of 7 years.

Table 5.12: Molecular findings in Patient ID-023 in the current study with NGS panel.

Gene	Chr no	cDNA	Protein	ExAC All Freq	PP	inheritance	ACMG - criteria	ACMG classification
SHANK2	11	c.1211G>A	p.Arg404His	-----	D	Pa	PP3; PM2	VUS
GRIA3	X	c.343T>C	p.Ser115 Pro	-----	D	Ma	PP3; PM2; PM1	VUS

Pathogenicity Prediction (PP) - D=deleterious; T=Tolerated; Chr Nr- chromosome number; Freq- frequency; Ma- Maternal; Pa- Paternal; NT – Not tested in further family members; VUS - Variant with uncertain significance.

CS02 presented two heterozygous novel missense mutations in SHANK2 and GRIA3 genes. GRIA3 (Glutamate receptor, ionotropic, AMPA 3; GLUR3) is also implicated in the phenotype, Mental retardation, X-linked 94, recessive (OMIM 300699). As previously described, SHANK2 mutations have been shown associated with ID, schizophrenia and ASD and the phenotype of the patient is in line with the described mouse model related to SHANK2 (Schmeisser et al., 2012). The phenotype of the patient also correlates with the described clinical features for MRX94 like hyperactivity, aggression and autistic behaviour with ID. The GRIA3 mutation is located in a functionally well-established domain, extracellular ligand-binding receptor (Periplasmic binding protein-like I). The GRIA3

mutation is inherited from the healthy mother, which is formally compatible with X-linked recessive inheritance. An X-chromosome inactivation (XCI) test needs to be done further to confirm a skewed XCI in the mother. The SHANK2 mutation is inherited from the father and further testing in other family members need to be done. According to the mouse model described above and also according to the gene-gene network analysis, a strong connection between the two genes exists and alterations in either of the genes or both acting together could potentially result in ID.

Case Study 03 (CS03): ID-038

This 11-year-old boy was the first child of healthy, unrelated German parents and has a healthy younger brother. He was born after 42 weeks of gestation after uneventful pregnancy. At the time of birth, weight was 2780 g with a height 50 cm and 32.5 cm OFC. He sat at the age of 6 months, crawled at age of 9 months and started to walk at age of 12 months. At age 5-6 years developmental delay became obvious. He showed attention deficits, learning difficulties and deficits of fine motor skills but otherwise healthy with no major malformations. He has recurrent infections in the upper airways. He showed minor craniofacial features like hypertelorism, broad nasal root, broad philtrum, small and low set ears, simple ears, and mild pterygium colli.

Figure 5.24: Clinical photos of (I) Patient ID-038 at the age of 9 years.

Table 5.13: Molecular findings in Patient ID-038 in the current study with NGS panel.

Gene	Chr no	cDNA	Protein	ExAC All Freq	PP	inheritance	ACMG - criteria	ACMG classification
BSN	3	c.6325G>C	p.Asp2109His	-----	D	NT	PP3; PM2; PM1	VUS
GRM1	6	c.1322C>T	p.Ala441Val	-----	D	NT	PP3; PM2; PM1	VUS
CNTNAP2	7	c.2606T>C	p.Ile869Thr	rs121908445; ExAC Freq- 0.0003. Known disease mutation at this position (HGMD CM080154) (Risk factor, ASD).	D	NT	PM2; BP1	VUS

Pathogenicity Prediction (PP) - D=deleterious; T=Tolerated; Chr No- chromosome number; Freq- frequency; M-Maternal; P- Paternal; NT – Not tested in further family members; VUS - Variant with uncertain significance

In CS03, two heterozygous novel missense mutations were detected in BSN (Bassoon) gene and in GRM1 (Glutamate receptor, Metabotropic, 1; MGLUR1) gene. GRM1 is implicated in

the phenotype, Spinocerebellar ataxia, autosomal recessive 13 (OMIM 614831) and Spinocerebellar ataxia, autosomal dominant 44 (OMIM 617691). GRM1 was also shown to associate with autosomal-recessive ID in a single consanguineous family (Davarniya et al., 2015). GRM1 has been shown to play an important role in synaptic plasticity and cerebellar development (Conquet et al., 1994). The reported GRM1 variant is located in the extracellular ligand-binding region and located near regions responsible for the regulation of mGluR1 activation in response to glutamate signalling (Watson et al., 2017). Recent studies have been shown linking BSN as a strong candidate gene for Landau-Kleffner syndrome (LKS, OMIM 245570) (Conroy et al., 2014) and also chromosome 3p21.31 microdeletions with characteristic clinical features of developmental delay and distinctive facial features encompass BSN with other two genes as potential genes for ID (Eto et al., 2013). The major clinical features of the patient in the current study share the clinical findings with already described previous reports. Studies with mutant mouse models of BSN gene have shown the role of Bassoon in network maturation, which affects the specific forms of learning and memory. In one study (Cheyne et al., 2011) it was shown during sequential synapse formation occurring onto newborn neurons and integrating them into the existing neuronal circuit, the presynaptic protein BSN was shown to present in high density. Along with BSN, higher levels of mGluR1/5 receptors were also present suggesting important roles in early neuronal outgrowth and neuronal integration and in synaptic plasticity. The CNTNAP2 missense variant (c.2606T>C) has been shown as may be a susceptibility allele for ASD (Bakkologlu et al., 2008) and functional studies indicate that it may alter protein expression, leads to CNTNAP2 protein mislocalization (Falivelli et al., 2012). This variant is present in population databases (ExAC 0.05%). This variant has been reported in 4 individuals affected with autism spectrum disorder from 3 families, and in all cases was inherited from an unaffected parent (Bakkologlu et al., 2008). Therefore, the available evidence is currently insufficient to prove it as disease causing conclusively and remains a question of whether as a pathogenic variant or a rare benign variant. Further testing of these mutations in additional family members need to be done to confirm pathogenicity and for better segregation analysis. From the gene-gene networks it is evident that all the genes interact with intermediate partners involved in neurotransmitter secretion and neuropeptide signalling pathway.

Case Study 04 (CS04): ID-005

The index patient (male) was the first child of healthy, unrelated German parents and had a healthy younger sister. Family history was unremarkable. He was born after 40 weeks of gestation after uneventful pregnancy but treated for bradycardia. He sat at the age of 8 months and started to walk at the age of 13 months. He had global developmental delay with

regression in speech development. His first words were around 18 months, 2-word sentences with 24 months, after 3 years used only some single words mostly non-verbal communication. At age 14 his IQ score was 20. He showed behavioural anomalies, stereotypic behaviour, unprovoked laughing, sleep disturbance, and absent feeling of satiety. He had his first seizure at age 13 years and before had abnormal EEG for many years. He is otherwise healthy with no major malformations. All preliminary genetic tests were resulted normal, which included GTG-banding, FISH of SHANK3-region (Phelan-McDermid-syndrome), fragile X, Angelman-syndrome, and array results. The MRI scan was also normal. At the time of clinical visit his age was 15 years with weight 56 kg (BMI 20.6), height 165 cm (-0.6 SD) and OFC 54.8cm (-0.33 SD) and no striking dysmorphologic features. He used only single words with very limited active speech and also had impaired passive langu

Figure 5.25: Clinical photos of (I) Patient ID-005 at the age of 15 years

Table 5.14: Molecular findings in Patient ID-005 in the current study with NGS panel.

Gene	Chr no	cDNA	Protein	ExAC All Freq	PP	inheritance	ACMG - criteria	ACMG classification
KCNQ3	8	c.1091G>A	p.Arg364His	-----	D	Ma	PP3; PM2; PM1; PS1	LP
PCLO	7	c.11261C>T	p.Thr3754Ile	-----	D	NT	PP3; PM2	VUS

Pathogenicity Prediction (PP) - D=deleterious; T=Tolerated; Chr Nr- chromosome number; Freq- frequency; Ma- Maternal; NT – Not tested in further family members; LP – Likely pathogenic, VUS - Variant with uncertain significance.

In CS04, two heterozygous novel missense mutations were detected in KCNQ3 (Potassium channel, voltage-gated, kqt-like subfamily, member 3) gene and in PCLO (Piccolo) gene. The same KCNQ3 mutation was previously described by Fusco et al., 2015 in a patient with benign childhood epilepsy with centrotemporal spikes (BECTS) as a disease causing pathogenic mutation and no functional studies were done. However recent reports have shown that KCNQ3 mutations have been in patients with more severe phenotypes of neurocognitive disorders including ID in 15% of patients (Miceli et al., 2015). The phenotype of the patient is in line with previously described reports with seizures and severe intellectual disability (Soldovieri et al., 2014; Bosch et al., 2016; Rauch et al., 2012). Therefore, the phenotypic spectrum of KCNQ3 variants appeared to be broader than benign epilepsy only. The KCNQ3 variant is located in a functionally well-established domain, Ion transport domain. The variant is maternally inherited and further testing in other maternal family members need to be done

to confirm pathogenicity and for better segregation analysis. Another novel missense mutation was identified in exon 7 of the PCLO gene, which corresponds to the coiled-coil region 3 (CC3) domain which promotes interactions between BSN-PCLO-ELKS/CAST and/or Munc13. These interaction partners are required for scaffolding and assembly of CAZ complex and for synaptic vesicle priming and any mutations on PCLO gene may result in altered interactions with their respective partners resulting in functional imbalance of the synapse thereby affecting synaptic plasticity, which is linked to several neurodevelopmental disorders.

Case Study 05 (CS05): ID-035

This 5-year-old boy was the first and only child of healthy, non-consanguineous German parents with unremarkable family history. He was born after 39 weeks of gestation after pregnancy showing intra-uterine growth retardation (IUGR). At the time of birth, weight was 2125 g with a height 42 cm and 31 cm OFC. The cranial ultrasound showed periventricular cysts. The boy has hip dysplasia and severe global developmental delay. As an initial classification the phenotype was considered as primary microcephalic dwarfism. At age 27 months all the measurements were <3rd centile with height of -2.1 SD and OFC of -4.6 SD and brachycephaly was also evident. At age 39 months microcephaly and microsomia became even more pronounced with height of -4.8 SD and OFC of -5.3 SD. There was a delay of bone age, no active speech and he was only able to walk with support.

Figure 5.26: Clinical photos of (I) Patient ID-035 at the age of 5 years.

Table 5.15: Molecular findings in Patient ID-035 in the current study with NGS panel.

Gene	Chr no	cDNA	Protein	Exac All Freq	PP	Inheritance	ACMG - criteria	ACMG classification
NFIB	9	c.998C>T	p.Ser333 Phe	-----	D	NT	PM1; PM2; PP3	US
TBR1	2	c.739A>G	p.Thr247 Ala	-----	D	NT	PM2; PP3; PM1	US
RASGRF1	15	c.3685C>T	p.Arg1229Cys	Exac Freq- 8.236e-05; Exac Allele count- 10 / 121412	D	NT	PM1; PP3	US

Pathogenicity Prediction (PP) - D=deleterious; T=Tolerated; Chr Nr- chromosome number; Freq- frequency; M-Maternal; P- Paternal; NT – Not tested in further family members; VUS - Variant with uncertain significance.

In CS05, two heterozygous novel missense mutations were detected in TBR1 (T-box, brain 1) and NFIB (Nuclear factor I/B) genes. Transcription factor, NFIB plays a fundamental role in various biologic processes, particularly in developmental regulation of cell differentiation in a number of organ systems (Becker-Santos et al. 2017). Haploinsufficiency of NFIB is laid as an underlying pathogenic mechanism, introducing NFIB as a novel causative ID gene (manuscript in preparation). Mouse models have pointed out the important role of the Nfi family for brain development (Chaudhry AZ et al., 1997). Nfib knockout mice show severe glial defects at the midline, as well as agenesis of the corpus callosum and die at birth due to defective lung maturation (Piper et al., 2009). Recently, NFIB has also been shown to regulate hippocampal neural stem cell fate (Rolando et al., 2016). The reported missense change is within the DNA-binding and dimerization domain, which is consistently rated to be likely pathogenic by various online prediction tools. TBR1 is specifically expressed in the brain and has been proposed to control neuronal migration and axonal projection of the cerebral cortex and amygdala. In mice, Tbr1 haploinsufficiency results in defective axonal projections and impairments of social interactions, ultrasonic vocalization, associative memory, and cognitive flexibility (Huang TN et al., 2014). TBR1 is a locus that is associated with high-confidence risk factor for ASD and ID and microdeletions region harbouring only TBR1 (Palumbo et al., 2014) and missense variants (Hamdan, F. F et al., 2014) have been described in patients with ID along with the variable presence of ASD or growth retardation. The TBR1 variant is located in a functionally well-established domain, DNA binding domain. A very rare RASGRF1 (Ras-guanine nucleotide-releasing factor 1) missense variant was also identified which is located in the functionally well-established and highly conserved domain, Ras guanine-nucleotide exchange factors catalytic domain. RASGRF1 is highly expressed in the CNS and mediates neuroplasticity via the RAS/MAPK signalling by directly binding to the NR2B subunit of NMDA receptors (Krapivinsky et al., 2003). It is also shown to be involved in synaptic plasticity and neurite outgrowth by activating the Rac signalling pathway, which is also important for the dynamics of the cytoskeleton (Baldassa et al., 2007). Further testing of these mutations in additional family members need to be done to confirm pathogenicity and for better segregation analysis.

6. Discussion

The RAS/MAPK and PI3K/AKT/mTOR pathways are interrelated signalling modules with multiple possible modifiers and effectors. Both pathways are very complex and only part of their functions and interactions is understood to date. Active GTP-bound RAS interacts with a wide range of targets (effectors) and stimulates downstream signalling pathways. Apart from activating MAP-kinase pathway, the activated RAS also binds to PI3K thereby activating the PI3K/AKT/mTOR pathway (Figure 6.1). Both signalling pathways are involved in many processes like regulating cell growth, metabolism and survival.

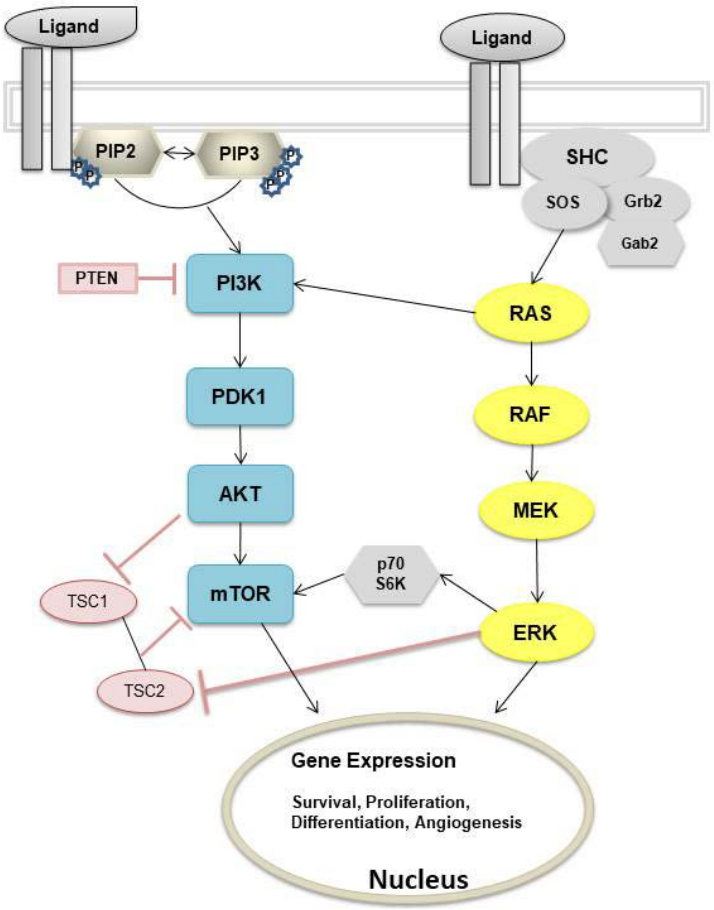


Figure 6.1: Interactions between the PI3K/AKT/mTOR and RAS/RAF/MEK pathways.

The interactions between these two pathways have been extensively described having impact in many cancers and both pathways are involved in many neurological disorders comprising ID and ASD (Sarah C et al., 2017). Known genetic defects in components of these pathways point out their crucial role for growth and neurodevelopment. Examples are RASopathies and the group of disorders related to disturbed PI3K/AKT/mTOR signalling (which includes PROS, Proteus syndrome, PTEN hamartoma tumor syndrome (PHTS), conditions with macrocephaly and ID, Tuberous Sclerosis complex). RASopathies constitute a group of neurodevelopmental disorders with overlapping clinical features caused by mutations in genes that encode components or regulators of the Ras/MAPK pathway (Rauen K.A, 2013). Studies

have shown that activation of both the pathways has a major role in the pathophysiology of RASopathies and depending on the tissue specificity and time, certain organ manifestations might occur not only due to a single pathway but a specific interaction of the two (Janku et al., 2011; Goodwin et al., 2012).

The advent of NGS technology has revolutionized molecular genetic testing in clinical diagnosis and research. One of the major advantages of this technology is the feasibility of testing of multiple genes (multigene panels, exomes, and genomes) of an individual in one experiment. This has been the precondition for the identification of many causes of genetic disorders that could not be solved by traditional approaches (positional cloning, functional candidate approaches), for example the large group of sporadically occurring disorders that are due to de novo mutations. Multigene panel and exome analysis has meanwhile also been introduced in clinical genetic testing with big success. The second advantage of NGS is the possibility of deep sequencing allowing the detection of very low levels of sequence alterations, which has been an important achievement in cancer genetics and mosaic disorders.

In this current study NGS technologies were applied for both, mosaic detection and identification of disease-causing mutations in a large multigene panel with a specific focus on genes related to RAS/MAPK and PI3K/AKT/mTOR pathways. The projects described here used the technology platforms available at the Institute of Human Genetics of the University Hospital Magdeburg and applied them for research projects. Thereby, the projects and approaches described here reflect also the rapid evolution of NGS technologies during the last few years represented by the two NGS platforms used in this thesis work (Roche GS Junior and Illumina MiSeq system).

6.1 Next generation sequencing technology

Since the introduction of NGS, the rate of gene identification for rare diseases is in rapid succession due to the increased evidence and biological understanding of the number of genes linking to the diseases thereby increasing the diagnostic yield (Fernández, Gouveia & Couce, 2017; Danielsson et al., 2014). NGS applications are not only limited to clinical genetics but are also very beneficial for other branches like microbiology, pathology, plant-biology and haematology. In addition, it is providing possibilities for carrier screening and prenatal screening, as well as in developing new targets for treatment and preventing ineffective drug treatments (Yohe & Thyagarajan, 2017). With these advantages, a rapid evolution of NGS methods and increase of novel technologies is seen over the past ten years since the first next

generation sequencing technologies. In recent years, many factors like improvements in NGS chemistries, more feasible, affordable and compact bench-top sequencers, decreases in experimental cost and computing power costs, and invention of several target capture technologies have resulted in widespread applications of these technologies to address questions that could not previously be solved (Singh et al., 2016). The yield and costs limits of the NGS methods were pushed with the advent of Illumina's Hiseq X and ONT PromethION platforms. Some technologies over the years could not keep up their pace in the race for evolution which resulted in their closure like the 454 pyrosequencing from Roche, Helicos Genetic Analysis System and the Revolocity system from Complete Genomics. In the next few years, more efficient and powerful third-generation technologies with novel sequencing solutions such as Oxford Nanopore, DNA transistor technologies from IBM and electron microscopy-based techniques will standardize the field (Goodwin, McPherson, and McCombie, 2016). These existing and forthcoming technologies would revolutionize sequencing world, making routine large scale genomic sequencing more feasible and extending towards direct sequencing of proteins and RNA or even real time precision medicine by drastically changing clinical genomics scenario.

The current work was also an exploration of the use of new NGS technologies available at the institute for research projects. In a first project for the application of the NGS technology, the GS Junior (GSJ) sequencing platform was used to validate the mutations identified in patients presenting with PROS. This provided the necessary insights into the limitations and pitfalls of the method and in the design. The GSJ platform is outdated but still available at our institute, was more efficient for smaller research projects like amplicon design panels for few numbers of genes. For the second project, a more evolved NGS instrument, Illumina MiSeq system, in combination with a custom-designed targeted enrichment was used for identification of rare variants in patients presenting with ID in a large multigene panel. Since in the current work two individual projects were done, a comparison study between these two different technologies has not been done. But few studies comparing the performances of these systems in different works/fields have been done, demonstrating that both technologies are accurate in variant calling and the Roche GS Junior is more prone to noise than the MiSeq, especially in areas containing homopolymers (Raymond et al., 2017; Alame et al., 2016; Loman NJ et al., 2012).

Regardless of these advantages, NGS applications and analysis still remain a challenge. The huge diversity of the human genome often limits the idea of simple analysis and the huge amount of data generated has an increasing risk of generating false-positive and false-negative

results (Lelieveld et al., 2015). Furthermore, NGS technologies are still limited in giving accurate results for repetitive regions and technical errors may arise in any step like library preparation, sequencing or data analysis further impeding the final outcome (van Dijk et al., 2014). In clinical scenarios, such errors giving inaccurate analysis may lead to misdiagnosis. Future NGS technologies promise the use of sequencing single molecules without the need for amplification or enrichment steps which introduce bias (Goodwin, McPherson, and McCombie, 2016).

In the current study working with different NGS technologies which are technically different from each other, suggest an evolution of the work with the available small scale Roche GS Junior towards a more medium scaled MiSeq, also dramatically increasing the sequencing capacity. These projects were intended to use the facilities available at our institute and at the time of start of these projects, exome sequencing was still a matter of cost.

6.2 Mosaic disorders

A characteristic of mosaic disorders is their extreme variability of phenotypic expression. The developmental timing when a mutation occurs and hence it's distribution over the body are major determinants for this variability (Campbell et al., 2015). For example, the same recurrent HRAS mutation p.Gly13Arg predominates in different clinical manifestations, obviously depending just on its tissue distribution: nevus sebaceous and Schimmelpenning syndrome, non-organoid keratinocytic epidermal nevus, nevus spilus, and phakomatosis pigmentokeratotica. Correspondingly, in the group of disorders that is now known under the term PROS identical mosaic mutations may account for various phenotypic expressions just depending on tissue distribution (Keppler-Noureuil et al., 2015). The second important determining factor for the phenotype of mosaic disorders appears to be the genotype, as certain disorders are related predominantly or exclusively to specific genes. For example, congenital giant melanocytic nevi and neurocutaneous melanosis are associated with NRAS and BRAF mutations, while sebaceous nevi and Schimmelpenning syndrome are caused by HRAS and KRAS mutations.

6.2.1 Mosaic KRAS mutations in OES/ECCL patients

Mosaic KRAS mutations that have previously been observed in some cases of nevus sebaceous, Schimmelpenning syndrome, and in non-organoid keratinocytic nevi are p.Gly12Asp and p.Gly12Val (Boppudi et al., 2016). In contrast, OES has previously been associated with the mutations p.Leu19Phe and p.Gly13Asp in two affected individuals (Peacock et al., 2015). In the present study recurrent mutations at codon 146, p.Ala146Val

and p.Ala146Thr, were first identified in OES/ECCL. Although the numbers of cases are still quite low, it appears unlikely that these non-overlapping mutation spectra in different types of KRAS mosaic diseases differ just by chance. The findings presented here corroborate that OES consistently results from a mosaic status for specific KRAS mutations and confirm for the first time the common genetic etiology of ECCL, which was the clinical diagnosis in patient 2. Both OES and ECCL, thus, belong to the mosaic RASopathies, a rapidly growing group of neurocutaneous disorders.

Clinically, mosaic RASopathies appear to have little in common with the germline RASopathies, and Noonan syndrome-like features are usually not recognizable. This may be related to the restricted tissue distribution of the mosaic mutations as well as to the more severe effects in mutated cells and tissues. Notably however, in OES and ECCL, we noticed that among the less commonly reported abnormalities, there are also some typical symptoms that occur in germline RASopathies, such as congenital heart defects (atrial septal defects, aortic coarctation), foetal hydrothorax, lymphedema, macrocephaly, large birth measurements, and relative or absolute short stature (Ardinger et al., 2007; Moog et al., 2009). Moreover, specific tumor types have repeatedly been reported in both, germline RASopathies as well as OES/ECCL, such as giant cell tumors and embryonal rhabdomyosarcoma (Kratz et al., 2011; Lees et al., 2000; Neumann et al., 2009). Oncogenic KRAS mutations, now established as the genetic basis of OES/ECCL, provide a reasonable explanation for such tumor associations, and consequently regular clinical follow-up should be recommended for early diagnosis of oncologic complications.

It has been shown for germline as well as for somatic cancer-associated mutations that different KRAS alterations may have unique biochemical effects (Gremer et al., 2011; Hunter et al., 2015). Activating effects of RAS mutations caused by either accelerated (GEF-independent) GDP/GTP exchange or defective GTPase activity (GAP resistance). The latter is the major activating mechanism of the most common cancer-associated KRAS mutations (such as Gly12Asp, Gly12Val and others) (Hunter et al., 2015). Alterations of Gly-13, Leu-19 and Ala-146 are instead predicted to rather affect GDP/GTP exchange kinetics because of their specific spatial relationship to the nucleotide binding cleft (Boppudi et al., 2016). Indeed, RAS Ala146Val and Gly13Asp mutants were both shown to have a strongly increased nucleotide exchange rate with no and only minor impairment of GTPase activity, respectively (Hunter et al., 2015, Feig LA & Cooper GM, 1988). KRAS Leu19Phe has been shown to accumulate in the GTP-bound state and to have oncogenic activity (Akagi et al., 2007). The precise activating mechanism for this mutant has not been resolved, but the position of Leu-19

relative to Ala-146 would be compatible with the assumption that functional effects might be similar (Boppudi et al., 2016). Notably, in NIH3T3 cells expressing mutant KRAS isoforms, transcription-profiling experiments and hierarchical clustering analysis revealed the presence of two major clusters for differential gene expression, one of them containing Gly13Asp, Ala146Thr, and Leu19Phe, and one containing the Gly12Val, Gly12Cys and Gly12Asp mutants (Smith et al., 2010). Although it is too preliminary to speculate about the exact pathophysiological mechanisms, these findings suggest that all hitherto known OES/ECCL-associated mutations generate KRAS proteins with similarly altered biological properties probably related to their ability to over-activate through GEF-independent nucleotide exchange. These specific functional consequences may determine the phenotype because of distinct and tissue-dependent effects on cell fate decisions. This hypothesis is further supported by the observation that specific oncogenic RAS mutations also display differential coupling to specific cancers (Prior et al., 2012). KRAS p.Gly13Asp as well as codon 146 mutations are commonly found in colorectal cancer but are infrequent in pancreatic or lung cancer (Smith et al., 2010; Edkins et al., 2006). Moreover, mutation-specific phenotype associations have also been observed in other mosaic RASopathies, as shown for NRAS mutations in congenital melanocytic nevi (CMN) vs. the nevus spilus type of CMN (Kinsler et al., 2014, Krengel et al., 2016, Boppudi et al., 2016). Nevertheless, despite significant point mutation biases, the borders between mosaic RASopathy phenotypes and their association with specific mutations may not be absolute.

Recently, two recurrent mosaic activating mutations in FGFR1 have been shown to cause ECCL (Bennett JT et al., 2016). The reported mutations in FGFR1, c.1638C>A (p.Asn546Lys) and c.1966A>G (p.Lys656Glu) were both known oncogenic mutations and are the two most commonly mutated residues among FGFR1 mutation-containing tumors (Lew ED et al., 2009). Along with this evidence, ECCL thus represents the first known example of a developmental disorder in the FGFR family with an increased risk for cancer, specifically low-grade gliomas (Bieser et al., 2015; Kocak et al., 2015). Specific differences in RAS/MAPK pathway activation due to mutations in KRAS versus FGFR1 might also play a role in the striking phenotypic overlap between OES and ECCL. Essentially, it could make sense that “self-activating” KRAS mutants have the same biological consequences as an activating mutant of the upstream receptor like FGFR1.

In the current study, for detection of mosaic mutations in KRAS gene conventional Sanger sequencing was used. Mosaic mutations were detected only in lesional tissues and there was absence of the mutations in DNA extracted from blood samples and one FFPE tissue sample

from jaw giant cell tumor (Table 5.4, Results section 5.1.2). There is a chance that low-frequency mosaic mutations in blood could be missed using Sanger sequencing as the typical Sanger detection limit is up to 10% (Riviere et al., 2012). With these results, it is suggestive that for blood samples or low quality DNA samples the detection might be low or absent and that sequencing of biopsy derived DNA will provide a higher diagnostic yield. Moreover use of highly sensitive NGS technology is more advantageous as it can detect alleles present as low as 0.1% (Keppler-Noreuil et al., 2015).

In conclusion, we have established OES and ECCL as mosaic RASopathies and confirm the common etiology of OES and ECCL. KRAS codon 146 mutations, as well as the previously reported OES-associated alterations, are known oncogenic KRAS mutations with distinct functional consequences. We define codon 146 of KRAS as a hotspot for mutations associated with these related disorders. Despite some overlaps between the various mosaic RASopathies both phenotypically and genotypically, there is growing evidence for mutation-specific phenotype associations, the pathophysiological basis of which needs to be addressed by future research.

6.2.2 PIK3CA-related overgrowth syndrome (PROS)

PROS are non-hereditary regional overgrowth disorders with somatic mutations in PIK3CA. Until now no instance of confirmed recurrence in siblings or vertical transmission in PROS has been reported. Activating PIK3CA somatic mutations have also been shown in further regional overgrowth conditions like megalencephaly (MEG), megalencephaly-capillary malformation syndrome (MCAP), hemihyperplasia-multiple lipomatosis (HHML), isolated macrodactyly etc. All these conditions have been subsumed under the term PROS. (Keppler-Noreuil et al., 2015; Rios et al., 2013; Mirzaa, Riviere, and Dobyns, 2013). A wide number of possible explanations can be given on how similar somatic mutation in one gene can cause different clinical conditions. The phenotype may depend on (1) the time point of somatic mutation arises during embryogenesis and (2) the specific cell type that contains the mutation. Some individuals have suggestive phenotype of the disorder but do not meet the clinical diagnostic criteria and also detection of no or low-level mosaicism for PIK3CA mutations in one or the other tissue cannot exclude the possibility of PIK3CA-associated segmental overgrowth disorders (Mirzaa et al., 2016; Martinez-Lopez et al., 2017).

In the current study, hot spot PIK3CA mutations, a total of 8 missense mutations and 3 deletions were identified in 18 patients presenting with PROS. Three novel mutations (p.Gly106_Glu109del; p.Pro447_Leu455del and p.Asn1044Tyr) were identified which have

not been described as being related to PROS before. Somatic mutation detection has been done in this study by three different detection methods: Sanger sequencing, Fragment analysis and NGS and all the mutations identified were confirmed by the three methods used. The identified mutations in this study are distributed in the adaptor binding domain (ABD), helical and kinase domain of the PIK3CA gene which play important roles in regulatory function. All the mutations identified were termed disease causing by different web-based prediction tools.

Two NGS runs were performed in this study and all candidate PIK3CA variants identified by our first screening were confirmed by a second, independent amplification reaction and sequencing experiment in the primary sample tested, confirming the high specificity of our assay. NGS-based ultra-deep sequencing of PIK3CA achieved a mean coverage of almost 4000 fold at coding bases and splice junctions. The NGS run 1 was a considerable failure in which only 31.94% of the total reads could be assembled to the amplicon reference sequences (passed filter wells). A number of different reasons could be speculated for this run failure like short read lengths (primer dimers, unspecific sequences), mixed reads, type of tissue sample used, number of samples used, unequal sequencing efficiencies and poor quantification or purification of amplicons. The two main reasons for run failure we present through this study is the use of FFPE samples which resulted in higher number of non-specifically mapped reads and short read lengths. As described in earlier studies FFPE samples are tough to process and are very sensitive to NGS technologies due to highly fragmented DNA which can reduce library fragment size and uniformity (Hedegaard et al., 2014). Moreover for the FFPE DNA, the PCR success rate is strongly correlated with the size of the amplicons (Do H & Dobrovic A, 2015). In recent NGS studies, very short amplicon lengths (120 bp or less) are being used for generation of excellent data from FFPE samples along with the use of advanced or specialized kits (like the Illumina amplicon kits). Based on these data, one reasonable explanation for the failure of the FFPE samples in this study is due to the use of either standard PCR methods for amplification or due to the varied large size of the amplicons (280 – 490 bp). Such problems can be overcome which we have experienced with this self-designed amplicon approach with the available new enrichment protocols for NGS (like capture-based sequencing approaches) and data analyses adjusted to interpret FFPE data reducing the occurrence of artifacts/false positives (Kotoula et al., 2015; Kamps R et al., 2017). The FFPE samples in this study also showed higher background (short fragments) in fragment analysis than compared to other samples derived from fresh tissues or blood (Supplementary figure 8). So considering the above reasons the second run was performed with limited number of samples excluding the FFPE samples with high background and the samples containing deletion mutations. The NGS run 2 was successful in which 71.76% of the

total reads could be assembled to the amplicon reference sequences (passed filter wells). The amount of short reads was drastically reduced when compared to run 1 providing better quality specific reads. Besides reducing the sample number the other important cause for the successful run 2 is the use of SPRI beads for removal of primer dimers before emulsion. This amplicon purification was exclusively size selective limiting the presence of short or unspecific sequences in the run.

Due to tissue specific somatic mutations, DNA samples derived from different tissues along with blood samples were included in the study. By using Sanger sequencing method, calculation of allele frequencies with absolute values is difficult to achieve since detection levels and quantification of mutant alleles were calculated from electropherograms (area under the curve) and the threshold is limited to 10-15%. As explained in above section due to low level detection limit by Sanger sequencing (Riviere et al., 2012), in none of the blood samples PIK3CA mutations could be detected. We therefore overcame this issue by deep-targeted NGS sequencing in both tissue and blood samples utilizing amplicon deep sequencing method, which has been particularly demonstrated to be effective for detecting low-frequency variants (O'Roak BJ et al., 2012). We used a multiplex bar-coded amplicon sequencing approach for PIK3CA in patients suffering with PROS. A total of 8 blood samples were present in the current study and mutation levels as low as 0.1% was detected in 5 blood samples using NGS. In three blood samples, mosaic mutations could not be detected even though having high read coverage of more than 2500 reads but alternatively could be detected in tissue samples. In other studies it has been already described that active AKT in the PI3K/AKT pathway has a deleterious effect to hematopoiesis (Kharas et al., 2010). This is confirmed through our study and previous studies (Keppler-Noreuil et al., 2014; Mirzaa et al., 2016; Hucthagowder et al., 2017) that blood source is not preferable for detection of somatic mutations. Although the mutations causing PROS are usually not detected in blood by Sanger sequencing, ultra-deep sequencing with its high sensitivity may be able to detect the causative mutation in leukocytes in more than half of the cases.

Variable levels of allele frequency were detected in different tissues by different methods used in this study. Mutant allele ratios of 30-50% were observed in scrapings from epidermal nevi (50%) or affected fatty tissue (30%) samples by Sanger sequencing, whereas by NGS method, the mutant allele ratios were 35% for epidermal nevi or 20% for affected fatty tissue. It has been already shown in several studies that mutations are not equally detectable from available "surrogate" tissues samples (Mirzaa et al., 2016; Kuentz et al., 2017). This is in line with the current findings showing a strong variability of mutation distribution based on the

tissue sample with mutant allele ratios ranging from 20-50%. For example, in one of the patient harbouring the mutation c.3140A>G in PIK3CA as mosaic allele frequencies ranging from 15-50% were found in different tissues. The presence of same mutant allele in mosaic in different tissues proves that the mutation was a post-zygotic event and occurred during early embryonic development affecting the cell lineages.

In the current study, assessing the mosaic allele frequency for the deletions identified by Sanger sequencing also remained challenging because the calculations were again based on the electropherograms. So for attaining better accuracy, fragment analysis method was also performed in this study. This method is fast, simple and accurate which adds information to the data interpretation over Sanger sequencing data providing more accurate information about the deletions. The fragment analysis method used in this study was also helpful as a great quality measure for checking the presence of short fragments (Supplementary figure 8). Post hoc tests for determining the deletion ratios using different methods (using the Bonferroni correction) revealed that the fragment analysis calculations showed a slight higher allele frequency from NGS, which was not statistically significant. However, Sanger sequencing calculations were highly over estimated which was statistically significantly different to fragment analysis ($p = .0004$) and NGS ($p = .002$) frequencies. For the missense variant allele frequency, Sanger results were gain overestimated which was statistically significantly different to NGS ($p < .0005$) frequencies. Therefore from our results and previous studies (Rohlin et al., 2009; Arsenic et al., 2015), we could say that Sanger sequencing for mosaic disorders results in systematic over estimation of the allele frequencies which could prone into more possibilities of artifacts. Therefore, we can conclude that the use of other sensitive methods is recommended for somatic mutations detection and NGS is probably the most precise and sensitive method of determining the mosaic level.

Overall, we identified 11 different mutations in 18 mutation-positive individuals having recurrent mutations (i.e., seen in more than one affected individual). Three mutations had not been previously identified in patients with PROS (p.Gly106_Glu109del; p.Pro447_Leu455del and p.Asn1044Tyr). Recurrent known pathogenic variants were found at established hotspots in exon 10 (p.Glu542Lys) and exon 21 (p.His1047Arg) which are also known as hotspots for somatic mutations in different cancers. Although most mutations were missense changes, we identified three in frame deletions encoding p.Gly106_Glu109del; p.Pro447_Leu455del and p.Glu110del in four patients presenting with CLOVES. The three in frame deletions identified in the present study are supposed to lead to expression of a mutant protein with gain of function, as it has been shown for several of the PROS-associated missense mutations (Rios JJ

et al., 2013; Mirzaa et al., 2016). The identified novel p.Gly106_Glu109del variant lies in a linker sequence between the adaptor-binding and the Ras-binding domains with no reports in cancer samples (COSMIC database). The novel p.Pro447_Leu455del variant is in the C2 domain (facilitates recruitment to plasma membrane) and reported in only one cancer sample (COSMIC database; COSM5944102; breast cancer). In a patient presenting with HHML, a novel missense mutation p.Asn1044Tyr was identified which is located in the kinase domain and reported in only four cancer samples (COSMIC database; COSM36288). In a recent study by Kuentz et al., 2017, in a patient presenting with CLOVES, a different amino acid change (p.Asn1044Lys) was reported with a total of 6 samples in COSMIC and described the variant demonstrating of having weak oncogenic activity. According to the ACMG guidelines, the novel missense mutation identified in this study was classified as likely pathogenic (PM1, PM2, PM5, PP2, PP3) establishing it as a pathogenic variant for PROS. Functional Studies showing activity of PIK3CA mutations demonstrate a gain of function mechanism for at least 3 of 8 missense mutations identified (p.Asn345Lys, p.Glu545Lys, p.His1047Leu). For two other mutations (p.Gln546Arg, p.His1047Arg) identified in this study, the gain of function mechanism has been shown for different missense mutations at the same codon (Gymnopoulos et al., 2007).

Activating mutations in PIK3CA are well described in various cancers and mutations have been demonstrated showing high oncogenic potential (Gymnopoulos et al., 2007). Various studies showing the functional impact of PIK3CA mutations were done through in vitro and in vivo cell growth and oncogenic transformation experiments (Dogruluk T et al., 2015; Meyer DS et al., 2013) and the oncogenic potency was classified from weak to strong (Gymnopoulos et al., 2007; Keppler-Noreuil et al., 2015). Most of the identified mutations in the current study were also present in the COSMIC database at varying frequencies. The majority of the mutations (6/11) identified in this current study had strong oncogenic activity with the high frequencies of somatic mutations identified at p.Glu542 and p.His1047 positions, which are the described hotspot mutation sites. Previous studies have shown that these mutations lead to a higher increase in kinase activity (Mandelker et al., 2009) and also occur frequently in cancer (Gymnopoulos, Elsliger, and Vogt, 2007). The binding of the p85 regulatory subunit at the N-SH2 domain of PIK3CA is disrupted due to the mutation at position p.Glu542, while mutations at position p.His1047 were shown to lead to RAS-independent activation by causing a conformational change (Zhao and Vogt, 2008). The two variants (p.Glu110del and p.Gly106Val) found in this study were recently assigned to PROS described having unknown oncogenic activity (Mirzaa et al., 2016; Kuentz et al., 2017).

In the current study, a novel exon 8 c.1338G>C variant was detected by Sanger sequencing in one patient in two different samples (skin abrasion and FFPE) but was not so clear due to the threshold limit (<10%) for Sanger detection and high background noise. When NGS was performed, the variant was not detected even though a total of 430 reads were generated with 100% WT allele in the two samples. Higher depth coverage could not be achieved for these samples due to the low quality DNA derived from these samples. But still puts this variant in question suggesting that Sanger sequencing might have produced a very likely artifact which could not be fully excluded by NGS since the read count was limited.

Until now treatment for disorders with regional overgrowth is done either by surgical intervention (removal of excess tissue/fat) or orthopedic procedures. Targeted therapy by applying PI3K/AKT/mTOR inhibitors is being used in various cancer types presenting the same mutations as above. Even though this may not cure the disease like PROS it may probably prevent progression or recurrence. In vitro studies have demonstrated that the use of targeted inhibitors resulted in significant reduction of the proliferation rate of mutant cells and also to suppression of the PI3K/AKT/mTOR signalling pathway (Lindhurst et al., 2015; Loconte et al., 2015). In a recent study by Suzuki Y et al., 2017, four different compounds were tested on fibroblasts cell lines from a patient presenting with CLOVES, harbouring the mutation c.3140A>G. The effects of two direct (rapamycin and NVP-BEZ235) and two indirect (aspirin and metformin) inhibitors of PI3K/AKT/mTOR signalling pathway were analysed in this study. They have shown that all four compounds suppressed S6 phosphorylation and inhibited cell growth of the patient-derived fibroblast cell lines but metformin was the only compound which showed mild inhibition on the control cells too, thus the authors suggest metformin as a candidate drug for treating PROS. Inhibitor treatment may therefore be helpful for in future therapies for patients with PROS or other overgrowth syndromes.

Variable degree of intellectual disability has been reported in the brain disorders of PROS belonging to the PI3K/AKT/mTOR pathway which ranges from mild learning disability to severe disability (Jansen et al., 2015). Subsets of patients also have seizures, cortical dysplasia, hydrocephalus, gross motor delays, limb asymmetry or overgrowth, hypotonia, autism and connective tissue dysplasia (Roy et al., 2015). Individuals with CLOVES syndrome have also been shown to have variable degrees of intellectual disability along with their brain malformations showing an overlap with those in MCAP syndrome (Sapp et al., 2007; Alomari AI, 2009). In one of the studies, CNS malformations and seizures have been reported in CLOVES syndrome patient who had neuronal migration defects and hemimegalencephaly with agenesis of the corpus callosum (Gucev ZS et al., 2008). This data

suggests that CNS manifestations could be an important feature for CLOVES syndrome which makes it distinct from Proteus syndrome. In another study, PTPN11 germline variant was identified in an MCAP patient along with PIK3CA mosaic variant (Döcker et al., 2015). These studies prove the interaction between the PI3K/AKT/mTOR and RAS/RAF/MEK pathways giving an insight into the second hit hypothesis. In principle PIK3CA mutations might be restricted to the brain tissue and somatic mosaicism may even exist for them to explain further causes for ID. Further studies including more patients and animal models need to be done to show the interactions between overgrowth syndromes and RASopathies, along with analysis of a second tissue sample for somatic mutation detections.

In conclusion, our data confirm that cells from affected tissue are more preferable for mutation analysis in PROS whereas blood would be a secondary source. With the help of ultra-deep sequencing, mosaic levels less than 1% could be detected even in blood samples. This suggests that in the absence of lesional tissue from the patients, blood could also be tested by very sensitive NGS methods which can be helpful for diagnosis. The limitations for use of different materials can be partially overcome with the help of new technologies in sequencing and use of advanced kits. Finally, somatic mutation detection levels by three different detection methods used in this study are: Sanger Sequencing – 10-15%, Fragment Analysis – 5%, NGS - <1%. More improved detection methods like NGS is necessary for identification of low mutant allelic frequency.

6.3 Multigene panel sequencing in patients with ID and Short stature

Intellectual disability causes are heterogeneous and the number of pathogenic or causative genes identified to date is still only a few hundred and a genetic diagnosis is still lacking in most cases (Musante & Ropers, 2014). With the advent of new technologies and methods, in the recent years, only a fraction of ID disease genes have been identified and described, and for numerous candidate mutations the pathogenic significance still needs to be verified (Vissers, Gilissen, & Veltman, 2016). To expedite the establishment of new candidate genes for ID and its related pathways, screening of large heterogeneous study groups still need to be done.

Since exome sequencing at the time of the project start was still expensive, we searched for an approach that could be realized on the instrument that was available and that worked with sequencing costs of about 200 € per sample. Herein, we used an approach focusing particularly on a panel of genes related directly or indirectly to the RAS/MAPK signalling pathway with one of the currently available and most frequently used platform in clinical

diagnostics scenario due to its reliability and versatility, the Illumina Miseq® System, for mutation detection and for disease gene discovery. A moderate size gene panel was constructed with 329 genes related to RAS/MAPK and PI3K/AKT/mTOR pathways and analysis of these genes was done by using custom-designed targeted enrichment capture method. This method was chosen as it was more economical, less laborious, possibility of multiplexing (sequencing many samples simultaneously) than sager sequencing and requires less infrastructure and also superior in its ability to generate a large amount of sequence data in one instrument run to achieve hundreds fold coverage at ease and the required amount of input DNA was also low. Target capture hybridization generates reads beyond the region of interest in immediate adjacent regions called off-target or near-target reads providing more data (Bodi et al., 2013; Lelieveld et al., 2015).

The current study included a cohort of 166 patients with non-specific ID and 120 growth-deficient patients. The study focused specifically on detection of rare and novel variants linked to RAS/MAPK pathway. The ID cohort is the main case study group and Short stature cohort served as the control study group in the current project for comparison purposes. The application of this technology allowed us to identify numerous novel and rare variations with definite pathogenic mutations in genes (two patients), which have been previously described for ID. Considering the large heterogeneity of non-specific ID, it has to be assumed – and empirically confirmed in some cases – that the contribution of individual monogenic disorders is usually less than 1% of the whole.

6.3.1 Target selection: Custom-designed gene panel

In the current study, a custom-designed targeted gene panel was designed containing selected set of genes belonging to RAS/MAPK and Short stature related pathways. A targeted gene panel approach was selected due to its many advantages like (1) having improved sensitivity than WES or WGS and also ensuring high specificity (2) to achieve high read depth coverages for selective targets allowing identification of very rare and novel variants (3) this approach allowed us to analyse 329 genes of interest at costs of about 200 € per sample. (4) Time efficient for both library generation and data analysis (5) more accurate reducing the probability of incidental sequence findings (low false positives) (Grozeva et al., 2015; Fernández, Gouveia & Couce, 2017). The overall design panel resulted in 6879 target probes for the 329 selected genes with 3% overlap with no gap length. Care was also taken in aspect of balancing i.e., equal efficiency was present between the designed capture probes and the capture targeted regions. Even though utmost care was taken during panel design to achieve full 100% coverage with no gaps, there were some gaps at the end result in the target probes

due to various factors like complex genomic regions, low specificity of the probe, high GC content, or other factors intrinsic for probe based enrichment (poor capture, amplification) (Rehm 2013; Lelieveld et al., 2015). Despite those limitations that are typical for enrichment methods, we ended up with 94% of target coverage of >20x. The performance of the gene panel was a bit below than other studies where it has been shown that for target coverage of >20x is around 96-99% (García-García G et al., 2016; Sun Y et al., 2015) but adequate coverage could be achieved. The designed panel overall produced high stringency and specificity still leaving place for improvements for gap coverages related to extreme GC conditions or repetitive elements.

There were certain limitations regarding the selection of genes for the panel like (1) only genes for which a connection to RAS/MAPK pathway of their gene products is known to date was included. Some genes prone to linage to RAS pathway through only computational evidence were also included in the panel with no or little evidence for biological relevance yet. But there may be important modifiers of this pathway, for which the connection is not known. Examples are SHOC2 or LZTR1 which had not been related to the RAS/MAPK pathway until mutations in these genes were found in RASopathies (Cordeddu V et al., 2009; Aoki Y et al., 2016) (2) Another limitation is that we had to cut the list of candidates at a certain threshold score according to our classification rules. Therefore the list cannot be complete. And also our classification rules are somewhat arbitrary, because they are based on arguments that are not equal for each gene. For example, a gene that is long known and well-studied is more likely to have any findings in the literature and databases that link it to the RAS/MAPK pathway in contrast to a gene for which only little functional data is available.

6.3.2 Sample quality and quantity checks

One of the major limitations of the Nextera® method is the constraints it places on input samples. The enrichment method strongly depends on accurately quantified starting material and superior quality DNA for achieving consistent tagmentation, reproducible library size distributions and good sequencing results. The most common cause of over or under-clustering[7] is inaccurately quantifying the library. In order to fulfil these criteria, prior to library preparation, various quality and quantity checks were performed for the selected DNA samples. For quantification of samples four different methods were compared which were one spectrophotometric (NanoDrop), one automated electrophoresis (TapeStation) and two

[7] If you load too little DNA, you're likely to 'under-cluster' the flow cell. Under-clustering usually maintains data quality, but results in lower data output. If you load too much DNA, clusters will be too close together (over-clustering), resulting in poor image resolution and analysis problems. Over-clustered flow cells have lower Q30 scores and reduced data output. In each case (over/under clustering) the caveat is lower data output.

different fluorometric methods (QuantiFluor® and Qubit®) (Figure 5.9, section 5.2.3). For the quantification of samples, NanoDrop measurements results were overestimated which was statistically significantly different to other methods compared ($p < .0005$) and which is in line with the previously reported studies (Hussing C et al., 2015; Simbolo M et al., 2013). The study by Hussing et al., 2015 also showed that for quantifying double-stranded DNA (dsDNA), TapeStation and Qubit® instruments were more accurate and quantification of samples up to 20–40 pg/µl could be accurately measured by the Qubit® instrument. From the obtained results and previous studies, fluorometric method using the Qubit® instrument was chosen in the current study and recommended as an ideal assay as it is highly selective for dsDNA over RNA or common contaminants (such as salts, free nucleotides, solvents, detergents, or protein). In the current study, utmost care was taken to have a final volume of 10 µl at 5 ng/µl gDNA (50 ng total) in the NGS runs as prescribed by the manufacturer which is reflected by the consistent peak distribution size with the sample peak around 300bp, as described in the protocol for the pre-enriched libraries generated. Out of 286 samples, pre-enriched libraries for three samples had a mean peak less than 300 bp indicating a poor sample quality, inaccurate quantification of the samples or low sample input which resulted in low coverage depth of the particular samples than the rest (section 5.2.6). Two of the samples were again re-quantified by Qubit® instrument which were successfully amplified and achieved promising results in another NGS runs. Successful reproducibility of the results using different quality and quantity methods in the current study provides an insight and protocol into the necessary checks need to be done on the samples for NGS methodologies.

6.3.3 Performance of modified Nextera® Rapid Capture Protocol

In the current study, Nextera® rapid capture enrichment method was performed and the chosen method was successful in the current study in terms of assay performance for analytical sensitivity and specificity, assay's repeatability and reproducibility. Reproducibility of the standard assay has been tested by using the same sample in two different runs under the same assay conditions and produced identical results (Supplementary figure 9a).

In order to evaluate the robustness of the Nextera® kit with further reduction in costs and increase in sample number per run, an in-house developed modified protocol method using reduced volumes of reagents for enrichment was tested and successfully implemented for four runs. Initial tests were performed with using standard protocol, Trail A (half volume), Trail B and Trail C scaled tagmentation reactions with proportionately reduced input DNA amounts. Following amplification and PCR clean-up, the quantities and size distributions of the libraries were compared. All test volumes produced libraries with a similarly broad peak (size

of the library for a distribution of DNA fragments) ranging from 200-1000 bp but varying in the quantities of the peak (Figure 5.18, section 5.2.7). The modified protocol (Trail C) was successful with no major differences or changes in various parameters compared the normal protocol. There was also no sample drop for any of the samples in the modified protocol and the fragment sizes generated after post enrichment did not differ between the two protocols. This is important in determining the eventual distance between the mated pair reads in pair-end sequencing. But there was a very slight decrease in sample peak intensity maybe due to a drop in quality in certain parameters (Figure 5.19, section 5.2.7). Only two samples (out of 96) had very low average mean depth with very low 1x and 20x coverages in the modified protocol due to inaccurate quantification of input material. One sample was repeated again in the next run and achieved very good coverage and results (Supplementary figure 9b).

Various parameters when tested beteween the modified and standrad runs, the only parameters which showed statistical siginficance to the normal protocol were the Q30 scores (% of bases with a quality score of 30 or higher) and percent aligned reads (% of reads passing filter that aligned to the reference genome) suggesting poor base calling accuracy increasing false positives variant calls and reduced read quality towards the end of the reads. This may be due to the fact that there was a slight over clustering in the modified runs than what is recommended by the manufacturer which results in overloaded signal intensities and poor template generation causing a decrease in these parameters. This problem can be overcome with accurate quantifications by qPCR method or reducing the library concertation in the sequencing reaction. For testing assay's reproducibility, a sample undergone sequencing using the standard protocol was included in the modified runs which yielded identical results in terms of variant calling except for a slight decrease in Q30 scores and more artifacts. A major advantage of the modified protocol generated in the current study is that with low input DNA concentration (half of recommend by manufacturer); libraries of acceptable complexity could be produced. It is very much beneficial for patient cases having limiting starting material and also allows for additional validation analysis and follow-up studies for the samples. But the modified protocol is not ideal for study related to mosaic disorders since low level detection of allele frequencies with higher read depths and accurate base calling are necessary for such studies.

The error rate percentage between the two protocols varied very much. The modified protocol showed an error rate of 1.40% which was twice much more than the normal protocol (0.53%). This may be due to over clustering which can lead to poor run performance, lower Q30 scores, a possibility for introduction of sequencing artifacts, and—counterintuitively—lower

total data output. One kind of enrichment bias is PCR duplicates which can arise due to (1) low inputs of DNA (2) high number of PCR amplification cycles (3) during enzymatic fragmentation of gDNA as it has been already shown that transposon systems have an increased predisposition for insertions (van Dijk et al., 2014) and (4) increased variance in fragment size. In the current study, though percent duplicate paired reads did not differ much in both the methods but showed a statistical trend towards significance with the modified protocol generating slightly higher PCR duplicates than standard protocol (Figure 5.20, section 5.2.7). This may be due to decreased amount of input DNA or increased number of amplification cycles during first PCR amplification step in the modified protocol. Furthermore a wide distribution of the duplicate reads was observed suggesting the issue of PCR amplifications as more and the transposase preference for AT rich sequences but not dependent on the regions or samples. Statistically more number of variants come up at threshold of >30x with the modified protocol due to increased number of wrong calls. This can be overcome by increasing the threshold for detection of rare variants although some potential variants might be missed. The modified protocol could still be optimised with further tests by changing few parameters like the number of PCR cycles, the incubation time for hybridization etc. to achieve increased specificity and sensitivity like the standard protocol.

6.3.4 Run statistics – Overview of the quality of runs

From each library pool, high quality raw data with more than 5 GB of output for each run (gigabases [GB] per flow cell) was obtained representing quite a high yield for an Illumina Miseq® platform with v2 reagents. In the current study on average about 1200 k/mm2 clusters (770–1660 k/mm2) were generated per run with an average of 1.7 ± 0.5 million total passed filter reads per run. The cluster densities generated in this study was slightly more than the optimal raw cluster density recommended by the manufacturer (1000–1200 k/mm2). Usually an increase in cluster density results in a higher number of reads but contrarily an increased cluster density also results in a lower number of reads passing quality filter and an increase in mismatches (Mitra A et al., 2015). Approximately 96.7% of passed filter reads were mapping to the reference genome. All the runs had a mean of more than 90% of sequenced bases with a Phred Q score $>=30$ reflecting the high quality of the runs and increased probability of correct base calling. The Q scores achieved in the current study are comparable to the published studies (García-García G et al., 2016; van Dijk et al., 2014).

Sufficient sequencing coverage is important for sensitivity and assured variant calling as low coverage tends to increase uncertainty and limit sensitivity (Lelieveld et al., 2015). In the

current study, the read coverage assessed at 20x (accepted threshold for proper variant calling) was above 90% and for 50x (accepted threshold for Indel detection) were above 83% for all the runs and also has high mean of depth of coverage and evenness score. This is of critical importance as it allows for multiplexing of more samples in future and reliable variant detection. Greater depth of coverage is also required to overcome sequencing errors but it is limited by costs, library complexity and methods preferred. Even though the numbers achieved in our study are a bit low than average what is recommended by various studies, it is still commendable and usable for rare variants identification (García-García G et al., 2016). The uniformity of coverage which is independent of sequencing depth was above 90% for all the runs which reflects an evenness of the distribution of base coverage among the targeted regions relative to the mean coverage (meaning getting sequencing as even as possible). The evenness scores (essential for experiment efficiency) achieved in the current study is slightly higher to the published studies (Bodi K et al., 2013; Mokry et al., 2010), but this cannot be compared with other methods, as it dependent on the targeted regions. Coverage variability in enrichment-based approaches is observed in regions with a high GC content or in a homologous region due to the probe-binding affinities or of the probes-nonspecific binding, respectively (Shin & Park, 2016; Chilamakuri et al., 2014). There were quite a number of gaps with ≤ 20x coverage and the low coverage was due to that the gaps were either regions of high (>60%) or low (<30%) GC content. This may be due to PCR amplification bias or during hybridization step as these regions tend to have reduced capture effectiveness. But interestingly there is no significant difference in the gap coverage between the samples and also between the two different protocols in the current study (Supplementary figure 8).

6.3.5 Variant identification and classification in patients with ID

Identification of novel or very rare variants that represent possible or likely pathogenic mutations causing ID was the primary goal of this current study. Distinguishing true variants from sequencing errors and artifacts is the main challenge for variant identification. Accurate identification of genomic variants is an important step for recognising disease-associated mutations considering several criteria, such as sequencing quality at the area, sufficient sequencing coverage or depth, allelic frequency and appearance of the variant in the forward and reverse strands with no strand bias. Various factors interfere with variant identification like during sequencing on the Illumina Miseq® system, errors may arise due to (1) base miscalls because of low signal strength, (2) SNPs in homopolymers stretch, (3) errors in inverted repeats, (4) read-through problems and in specific motifs. And during analysis, errors may be caused due to bad mapping (the mapping algorithm can map a read to the wrong location in the reference), incorrect trimming of reads and adapters, inclusion of PCR

duplicates, error in reference sequence, sequence contamination like adapters, multi-mapping to repeat or paralogous regions, variant calling with insufficient read mapping, and misaligned indels (Shin & Park, 2016; Singh et al., 2016).

In the current study, a total of 166 patients diagnosed with ID were screened using targeted NGS analysis through a panel of 329 genes, which were related to RAS/MAPK and PI3K/AKT/mTOR pathways. According to various studies, the number of high confidence variants identified by exome sequencing in ID studies was an average of 20-25,000 genetic variants per patient (de Ligt et al., 2012; Rauch et al., 2012; Hamdan FF et al., 2014) whereas the average number of variants for targeted gene panels of ID was between 400-700 genetic variants per patient (Martínez F et al., 2017; Redin C et al., 2014). In the current study after variant calling, an average around 350-440 variants per sample/patient were reported which is in line with the published studies. By applying all the filtering strategies for identification of very rare and novel variants, a total of 1199 rare high quality variants in ID cohort were identified. Within the ID cohort an average, in the examined RAS pathway related genes, ~ 4 very rare or unknown variants were identified per individual (Range: 0-12). But when comparing with exome sequencing, the number of rare and novel variants identified in human exomes which have about the 100-fold number of genes would be higher with an average of ~ 35 variants with around 7 de novo mutations (range of 2-20) identified per patient. There was a lack of any statistical accumulation of variants in the RAS pathway-related genes in the ID cohort probably due to the under limited cohort size in the current study. The contribution of variants in the genes investigated in this study is too low to discover accumulation of variants in a cohort of this size and clinical heterogeneity.

For a number of patient cases no potential pathogenic novel or rare variants could be identified which can be due to a number of likely reasons like (1) the causative gene is not included in the custom-designed panel; whole exome sequencing may address this issue (2) poor coverage of certain regions in genes or gaps (3) deep intronic or intergenic causative mutations which cannot be detected by selective gene panel, maybe this can be achieved through whole genome sequencing (4) for some complex disorders like ID, some common variants have been described associated with pathogenicity (Clarke T. et al., 2015) and also which are more common in general population (have low penetrance) could plausibly cause the traits in significant number of ID cases due to familial aggregation by possible interactions (5) In some families, non-genetic factors could also play a role both prenatally and perinatal.

For the identified variants in this study, classification was done according to the ACMG recommendation (Richards, et al. 2015) which is a more conservative and safe method for

running into preliminary pathogenic variants. One limitation of using this approach is that the ACMG criteria are not useful for classifying variants in a gene that is not an established disease gene, because many arguments do not apply. A second limitation for example would be testing for de novo occurrence was not possible for a significant number of interesting variants due to the lack of parental DNA samples.

6.3.6 Monogenic disorders – Mutations identified in known ID genes

The primary aim of this study was to identify mutations in RAS pathway-related genes as a cause of non-specific types of ID and thereby to evaluate the significance of RAS pathway alteration in ID. In two individuals out of a total of 166 patients, a single, most likely causative genetic change in currently known ID genes could be identified. Through this study, high number of likely clear pathogenic variants was expected to be identified but could not be found due to the high locus heterogeneity of ID. For example, SYNGAP1 mutations have been reported in 2% of patients with non-specific ID and STXBP1 mutations in 3% (Hamdan FF, Gauthier et al., 2009; Hamdan FF, Piton et al., 2009) of ID cases. We have expected around 4-5% solved cases in our cohort according to previous studies. But we could only identify 1% pathogenic variants in our cohort. However, this mutation rate is probably an underestimation of the true contribution of each gene to the overall burden of ID. For some of the variants identified, the analysis was significantly limited due to lack of parental samples or further family members to distinguish from de novo to rare familial causes. Therefore it is likely that there are additional causal variants in this cohort that we were unable to confidently identify as such. The unclassified variants in known ID genes are discussed further in section 6.3.7.

The patient ID-001 had a novel heterozygous de novo nonsense mutation (c.788T>A; p.Leu263*) in the gene CTNNB1 (beta 1 catenin) which has been associated with phenotype listed as mental retardation, autosomal dominant 19. The beta catenin is a highly conserved protein which mediates cell adhesion and is also part of Wnt-signalling pathway. The first study linking ß-catenin to ID was in a trio-exome sequencing study by de Ligt et al., 2012, followed by many individual case reports. To date, a total of 31 cases with loss-of-function mutations of the CTNNB1 gene linking to ID have been reported (Kharbanda et al., 2017; Li N et al., 2017). Till date, ten nonsense mutations were reported in CTTNB1 gene with two nonsense mutations identified in exon 6 of the gene (Leu251* and Gln309*) (Li N et al., 2017). The clinical features of the patient in the current study are in line with the reported cases in the literature which includes ID, microcephaly, abnormal facial features, motor and speech delays, hypotonia, behavioural abnormalities, and visual defects (Kuechler et al.,

2015; Li N et al., 2017). The identified novel mutation lead to premature truncation of the protein which also agrees with the previous reported de novo mutations and deletions, causing loss of function of the gene. In various tumors, somatic gain-of-function mutations in CTNNB1 have already been identified. In the current patient and in the previously reported cases, no tumor manifestations have been reported and are unlikely to be expected due to the opposite mechanism of haploinsufficiency (Kuechler et al., 2015). Mouse models were also reported showing functional characterizations of the mutations identified with deficits in dendritic branching, long-term potentiation, and cognitive function (Tucci et al., 2014). Reports were also shown with consistent results displaying various behavioural traits like autism in mice (Dong et al., 2016). There is no direct link between the gene CTNNB1 and of RAS/MAPK signalling pathway, but it is linked directly to AKT1 gene. Some studies have been shown that coordinate activation of RAS and β-catenin drives AKT-dependent tumor cell proliferation, migration as well as tumor growth in various cancers (Polosukhina et al., 2017; Yi Y et al., 2015; Gosens et al., 2010). But still it remains unclear whether any alteration of RAS/MAPK signalling plays a role for the neuronal deficits caused by CTNNB1 haploinsufficiency.

The patient ID-002 was found to have a mutation (c.1399T>G, p.Ser467Ala) in BRAF gene which is associated with Cardio-facio-cutaneous (CFC) syndrome. The identified mutation has indeed already been reported as disease causing mutation associated with CFC syndrome (Rodriguez-Viciana et al., 2006; Yoon et al., 2007; Rodriguez-Viciana and Rauen, 2008). It is a very rare mutation in CFC syndrome with only x counts in the NSEuroNet database (URL). The missense change affects the protein kinase domain (glycine loop) of the BRAF protein. Functional studies done by Rodriguez-Viciana et al, have shown that the variant resulted in higher levels of ERK and MEK phosphorylation in transfected COS cells. Cheng et al., 2012 through simplified mathematical models (computational evidence) has demonstrated that the variant have a systemic impact which was quantified as a combination of protein stability change and pathway perturbation. The position of the variant is within the phosphate binding P-loop of the kinase domain which is highly conserved across species and is also a hotspot region for missense variants described in CFC syndrome (p.Gly464Arg/Val, p.Gly466Ser, p.Phe468Ser, p.Gly469Arg/Glu, p.Thr470Pro), which further supports the functional importance of this region. Missense variants are the common type of diseases-causing variants in this gene, whereas loss-of-function mutations for BRAF have not been associated with a human disease, so far. The identified variant is a very rare variant with only three reported cases until now. Since CFC syndrome is usually a clinical diagnosis and the study cohort has had a routine genetic workup before, we did not expect detecting typical RASopathy cases in

this cohort. In fact, the clinical features of this patient, particularly the facial phenotype, were quite atypical and CFC syndrome was not suspected clinically. However, with the knowledge of a disease-causing BRAF mutation one can recognize some features that fit with a diagnosis of CFC syndrome, including some facial anomalies, cardiac, musculoskeletal and ocular abnormalities, ID, seizures, and hypotonia. This observation illustrates that CFC syndrome is not always an easy clinical diagnosis, and some patients with this syndrome may not be diagnosed, unless broad non-targeted sequencing is performed.

Additionally, one novel missense mutation in BRAF gene with uncertain pathogenicity (p.Leu711Phe) was also identified in the current study in a female patient presenting with ID and failure to thrive. The mutation is located in a well-established functional domain, protein kinase domain which is highly conserved across species. A neighboring mutation, p.Gln709Arg is associated with CFC syndrome (Sarkozy et al., 2009). Additional functional studies and segregation analysis in family members need to be done to further confirm the pathogenicity of this mutation.

6.3.7 Unclassified novel/rare variants identified in known ID genes

In total, 730 high-quality rare non-synonymous variants (missense, nonsense, indel and splice-site mutations) with minor allele frequency <0.1% were observed in the ID cohort. Classifying a missense variant as pathogenic is more complex as a minimum amount of missense variants affect protein function compared to loss-of-function variants. This even increases when there is no clear inheritance pattern for the missense variant due to lack of segregation analysis. According to one study it has been shown that every individual has a potential to carry large number of rare variants in their genome (Xue et al., 2012). In the current study eight novel mutations in dominant ID genes, 2 novel mutations in recessive ID genes, 2 novel mutations in X-linked ID genes and 5 novel mutations in genes showing association findings with ID were described (Table 5.7; 5.8; 5.9). Their pathogenicity was classified according to the ACMG standards and guidelines. They were of particular interest based on the known nature or function of the protein or the likely effect of the amino acid substitution on protein function. The level of evidence for a causative role of these variants was variable.

For example, here we reported a patient presenting with developmental delay and language impairment having a novel splice donor variant (c.4626+1G>A) in DOCK8 gene which predicts the abrogation of the splice donor site of exon 36 (out of 48 exons of the longest isoform) and most probably leads to aberrant splicing. DOCK8 is highly expressed in the immune system, and DOCK8 deficiency primarily causes immune-related disorders (Ham H

et al., 2013). The gene DOCK8 (dedicator of cytokinesis 8) although has been implicated in autosomal dominant Mental Retardation 2 (MRD2; OMIM# 614113). Genetic alterations affecting this gene have been published in only two patient cases until now, one with a deletion and another one with a translocation leading to disruption of the longest isoform of DOCK8 (Griggs et al., 2008). DOCK8 has also been suggested in few studies as a promising candidate gene in deletion regions linked to ID and autism (Vinci G et al., 2007; Shi, L., 2013). In a study by Redin C et al., 2014, 106 patients with ID were analysed and they found a de novo truncating variant in DOCK8 (p.Glu1166*) in a single patient. But the proband also carried a causative splice site mutation in PHF8 gene (PHD finger protein 8), and the implication of DOCK8 in ID was stated as most probably substantial or innocuous due to the weak evidences of DOCK8 linked to autosomal dominant ID. Intragenic point mutations linked to ID have not been reported, so far. The mutation observed in the patient presented here could not be identified as de novo due to the lack of parental samples. Without functional evidence to support pathogenicity by checking the splicing effect on m-RNA and further segregation analysis in the family, narrowing down as potential candidate mutation is not possible with the current study.

Another rare heterozygous missense variation c.1286G>A (p.Gly429Glu) in DOCK8 gene was observed in our cohort which is of particular interest. This variant is present in ExAC with one heterozygous count in the European (Non-Finnish) population out of a total of 121334 alleles (ExAC.broadinstitute.org). The variant has not previously been reported in any publications. This mutation occurs in the early exonic positions in a highly conserved region and is a splice region variant. The variation is predicted to activation of an exonic cryptic acceptor site thereby may also causing potential alteration of splicing suggesting loss of one or two domains downstream of altered spice site (HSF programme). The female patient has severe ID and so does her sibling. Both parents of the patient have attended special school but appear to be less severely affected than their children. Segregation analysis revealed the DOCK8 variant in mother and the affected sibling. Taking together the published evidence and the observation in our study cohort the role of DOCK8 as a gene for autosomal dominant ID remains unclear.

Another instance, we report a novel heterozygous missense variation c.1672C>A[8], p.Gln558Lys in the gene STXBP1 (syntaxin-binding protein 1) which has been implicated in autosomal dominant early infantile epileptic encephalopathy, 4 (OMIM# 612164). STXBP1,

[8] At the same position, a single base de novo deletion (c.1672delC) in STXBP1 was reported by Stamberger et al., 2016 causing frameshift in a patient presenting with severe ID and early-onset epilepsy and encephalopathy.

also known as MUNC18-1, a membrane trafficking protein is predominantly expressed in the brain and plays an important role in the presynapse for synaptic vesicle docking and fusion (Suri M et al., 2017). It was pointed out that STXBP1- related encephalopathy rather should considered as a complex neurodevelopmental disorder with ID added as one of the main criteria for classification (Stamberger et al., 2016). Several studies have been reported associating heterozygous pathogenic variants in the STXBP1 gene to ID, with or without epilepsy (Hamdan FF et al., 2009, 2011; Stamberger et al., 2016; Suri M et al., 2017) along with other neurological and behavioural abnormalities (Neale BM et al., 2012; Carvill GL et al., 2014). Mouse model studies showed that functional STXBP1 is crucial to synaptic maintenance and function (Verhage M et al., 2000; Toonen RF et al., 2006). The variant observed in the patient form our cohort is located in the Sec1-like, domain-2 of the protein which is a hot spot region for many described pathogenic recurrent mutations (p.Arg551Cys/His, p.Cys552Arg, p.Gly561Arg) and the variant is novel with no reports in the literature or in public databases. The male patient presenting with the p.Gln558Lys mutation has developmental delay, microcephaly, short stature and hypotonia. Due to absence of segregation data (no proof of de novo occurrence), the classification of this missense variation remained uncertain.

We also observed two novel heterozygous missense variations (c.1075T>C, p.Tyr359His; c.2375G>A, p.Arg792His) in the gene NTRK2. The BDNF receptor, TrkB, is encoded by the gene NTRK2 (Neurotrophic tyrosine kinase, receptor, type 2; OMIM 600456). Mature BDNF binds to the TrkB (tropomyosin-related kinase B) receptor, which on activation promotes neuronal differentiation and synaptic potentiation. Upon ligand binding, the receptor gets activated by phosphorylating itself and other several downstream effectors thereby activating the RAS-PI3K-AKT signalling cascade. Animal model studies revealed functional interactions of both BDNF and NTRK2 and they regulate LTP and also play an important role in hippocampal synaptic plasticity and learning (Xu B et al., 2000; Minichiello et al., 2002). NTRK2 has been already implicated in a syndromic type of ID with obesity, hyperphagia, and developmental delay (OMIM# 613886) in two individual case reports by Yeo et al., 2004 and Miller et al., 2017. In vitro functional studies of the identified human NTRK2 mutations was done by Gray J et al., 2007, showing that TrkB mutation impaired activation of MAPK and AKT with a reduction in the ability of BDNF- response neurite outgrowth and cell survival impairing neurogenesis, thus suggesting a contributing role of the mutations in severe hyperphagia and obesity. Both of the patients presenting with the missense variations in this study cohort had non-specific and non-syndromic ID and the mutations were each inherited from one parent showing milder degrees of developmental delay. The p.Arg792His variant

was paternally inherited and p.Tyr359His was maternally inherited. The female patient carrying the mutation p.Arg792His has severe ID and so does her sibling. The p.Arg792His variant was negative in the sibling. The father of the patient has been reported with some learning difficulties but appear to be less severely affected than the children. The female patient carrying the mutation p.Tyr359His has severe ID and the mother of the patient has been reported with some learning difficulties. Segregation analysis of the variant in further family members need to be done since low intelligence was reported in other relatives. The pathogenic significance of both variations remained unclear, but since the transmitting parent in both cases also had some degree of cognitive impairment it remains possible that the observed NTRK2 changes are causally related to the familial ID phenotypes.

Here we also report novel/rare mutations in SHANK (SH3 and multiple ankyrin repeat domains) family proteins which are predominantly expressed in the brain. Mutations in SHANK genes (SHANK 1, 2 & 3) have been shown associated with ID, schizophrenia and ASD (Berkel et al., 2010) and knockout mouse models for these genes showed deficits in learning, social behaviours and synaptic plasticity defects (Lim et al., 2017). The clinical pertinence of SHANK1 and SHANK2 genes still remains to be ascertained due to the rare frequency of mutations identified in these genes when compared to SHANK3.

In SHANK3 gene, two novel variants (p.Ser282Leu; p.Ala473Thr) and four rare variants (p.Ala29Pro; p.Ala986Ser; p.Ser1403Leu; p.Pro1255Leu) were identified in patients presenting with developmental delay. SHANK3 is the major haploinsufficient gene implicated in Phelan-McDermid syndrome (OMIM# 606232) and Schizophrenia. Heterozygous point mutations in SHANK3 have also been described in patients with autism spectrum disorders and schizophrenia associated to moderate to severe ID and poor language (Gauthier J et al., 2010; 2009). The two rare variants (p.Ser1403Leu; p.Pro1255Leu) were inherited from a healthy father whereas no segregation data was available for the other variants identified. In SHANK2 gene, two novel variants missense variants (p.Arg404His; p.Ile1840Thr) were identified in the current study and was explained in more detail in their individual case studies in results section (section 5.2.14, CS01 & CS02). In SHANK1 gene, two novel missense variants (p.Gly164Asp; p.Pro880Ser) and one rare missense variant (p.Pro1844Leu) was identified. Both of the patients presenting with the novel missense variations in this study cohort had non-specific and non-syndromic ID with muscular hypotonia. The rare variant is present in ExAC with two heterozygous counts in the south Asian population out of a total of 88540 alleles (ExAC.broadinstitute.org) which has not previously been reported in the literature and the patient had developmental delay with few dysmorphisms. Unfortunately, segregation data was not available for the variants identified due to lack of parental samples.

The current study also detected heterozygous mutations/variations in genes previously established as recessive ID genes (Table 5.8). Two variations in CC2D1A gene (p.Gln506* and p.Arg768Trp) and two missense mutations in ARFGEF2 gene (p.Leu1004Val and p.Glu731Lys) were detected in the current study. An observed single heterozygous mutation might act in a recessive model together with an unidentified mutation on the second allele. The sequencing data of the particular gene was manually screened on the Integrative Genomics Viewer (IGV, v2.3) to check for such a mutation on the second allele. The genes was thoroughly checked for (1) in all exons, splice sites and already predicted pathogenic intronic variants (2) the read coverage for each exon was also carefully checked for the presence of any gaps or low coverage and (3) checked for possible deletions or insertions. A further second mutation could not be identified in any of the reported variants. The nonsense mutation in CC2D1A c.1516C>T, was identified in heterozygous state inherited from healthy father, while for the others segregation could not be tested due to unavailability of parental DNA samples. Heterozygous mutations for recessive disorders might also be associated with mild clinical manifestations (semi-dominant) or act as genetic predispositions in a multifactorial model. There is some evidence from other disorders like increased risk of cancer in heterozygous mutations carriers for Nijmegen breakage syndrome, ataxia telangiectasia etc. For autosomal recessive ID genes, studies showing associations of pleotropic functions of these genes have been shown (Musante & Ropers, 2014). For example, FASN (fatty acid synthase) a functional candidate gene for recessive ID predisposes to leiomyomatosis (Eggert et al., 2012) and FTO gene (fat mass- and obesity-associated gene) which had been linked to obesity has been shown to cause severe ID with malformations (Frayling et al., 2007). With the limited published evidence, the concept would have some plausibility.

The current study also detected two novel missense variants of uncertain pathogenic significance in established X-linked genes for ID. Hemizygous mutations in GRIA3 (p.Ser115Pro) and ARHGEF6 (p.Cys667Ser), respectively, were identified in two male patients presenting with non-specific developmental delay. Both the genes are implicated in recessive forms of X-linked ID and previous studies showed skewed X inactivation in female mutation carriers (Bonnet et al., 2012; Kutsche et al., 2000). The phenotype of the patient presenting with the GRIA3 variant (Case study 02, Section 5.2.14) correlates with the described clinical features for MRX94 (OMIM# 300699) like hyperactivity, aggression and autistic behaviour along with ID. The GRIA3 mutation is located in a functionally well-established domain, extracellular ligand-binding receptor (Periplasmic binding protein-like I). Though the GRIA3 variant was classified as VUS according to the ACMG guidelines, it is

still can be argued of having pathogenic role since missense mutations in the transmembrane and ligand binding domains of the GRIA3 gene were reported in patients with ID and behavioral disturbances (Yuan H et al., 2015; Philips AK et al., 2014) and in vitro functional studies of the GRIA3 missense mutations was shown to impair channel function of the Ionotropic glutamate receptor 3 which is found to be linked with moderate ID (Wu Y et al., 2007; Chiyonobu et al., 2007). The GRIA3 variant was maternally inherited and additional functional studies or X-inactivation studies need to be done to further confirm pathogenicity. No segregation analysis could be done on the ARHGEF6 variant due to the absence of parental samples. For X-chromosomal mutations observed in males it is a general problem that the presence of the mutation in a healthy mother does not exclude its pathogenic role, and only extensive segregation analysis until another affected or unaffected male carrier is found, has the chance to collect more evidence for or against a pathogenic role.

In total, 468 high-quality rare synonymous variants (minor allele frequency <0.1%) were observed in the ID cohort. The discussion mainly focused on missense mutations and no results were shown for silent mutations or intronic changes due to their predicted lack of any impact on protein sequence or expression. Even though synonymous changes are generally assumed to have no functional consequences, there are several examples where such mutations have been shown to alter protein structure and clinical phenotypes (Livingstone M et al., 2017; Orion J et al., 2015). So one must take care not to neglect silent mutations and they must be carefully studied with further interest especially mutations showing splicing effects. For example in the current study after the second filtering pipeline was used (as described in methods section 4.8), two silent variants have been identified which were classified as pathogenic by different in-silico prediction programmes. Two heterozygous mutations, one in CC2D1A (c.2028C>T, p.Gly676Gly) and other in NRXN1 (c.3975C>T, p. p.Gly1325Gly) were identified in patients presenting with developmental delay and language impairment. Both the genes are implicated in recessive forms of ID and both the identified variants were present in ExAC with one heterozygous count in the European (Non-Finnish) population out of a total of 121334 alleles (ExAC.broadinstitute.org). The patient harbouring the mutation p.Gly676Gly in CC2D1A has ID, language impairment with few malformations which is in line with the reported phenotype of Mental retardation, autosomal recessive 3 (OMIM # 608443). The patient harbouring the mutation p.Gly1325Gly in NRXN1 (neurexin 1) has ID, epilepsy and behavioural anomalies which is in line with the reported phenotype of Pitt-Hopkins-like syndrome 2 (OMIM # 614325). By using the different in-silico prediction programmes for checking of the splicing effect, both the mutations showed predictions of splicing effects on the m-RNA by creating a new donor site. No segregation analysis could be

done for both variants due to the absence of parental samples. Further testing for splicing effects needs to be done in-vivo for confirming the pathogenicity of these variants.

Besides the two clearly solved cases (Section 6.3.6), possible candidate variations in established genes for ID related to the RAS/MAPK pathway were observed in 25 additional families. If all of them were really causative, the rate of ID caused by changes in RAS/MAPK pathway-related genes would rise from about 1% (2/166) to up to 15%. However, the evidence for a pathogenic role of the variants discussed in this section mostly remained weak. The major obstacles for a more conclusive classification were insufficient evidence for the gene itself being an ID gene (as shown for DOCK8, NTRK2, and ARHGEF6), lack of segregation data, and segregation within families with complex familial clustering of cognitive impairment of varying degree. Particularly in families where ID and learning difficulties appeared in multiple family members of both the paternal as well as the maternal lines, even an extended segregation analysis is unlikely to fully confirm or exclude the pathogenic role of a candidate variant.

6.3.8 Potentially disease-causing variants in genes previously not linked to ID

The second aim of this study was the attempt to identify any novel potential ID-associated genes, within those genes where most of the proteins encoded by these genes interact with either products of known ID genes or shown to have some neurological dysfunction in animal models. It is difficult to prove causality for any mutation in a novel gene when there is only one affected individual. Replication in larger populations is generally needed. Designating a candidate gene as causative for the disease requires a high degree of evidence like functional studies, additional families, co-segregation essential to support the link and similarly large data sets from corresponding cohorts with ID to identify the remaining rarer novel disease causing genes for this highly heterogeneous disorder (MacArthur et al., 2014). It is shown by the international scientific community that a prerequisite for new disease gene identification is the detection of more than one mutation in a gene in more than one pedigree (Fernández, Gouveia & Couce, 2017). The variants identified through the present study in novel candidate genes for ID were classified according to the ACMG guidelines. Most of them had to be designated as variants of unknown significance which pose a challenge to determine their pathogenicity. These guidelines have been developed primarily for classification of novel variants in genes with established disease-associations. They have limitations regarding the classification of variants in genes, which have never been associated with disease or in a gene previously associated with a different phenotype as many criteria are not applicable in these scenarios. Therefore, variants in novel candidate genes for a human disease formally remain

at the “variant of uncertain significance (VUS)” level, if ACMG criteria are applied. This is also true for variations in novel candidate genes for ID identified in this study.

Among the several novel and very rare variations that were found in the current study in RAS/MAPK pathway-related genes with no previously assigned human phenotype, few of them were particularly interesting based on their nature of their function specific to neuron or brain and few of the genes were functionally linked to known ID genes formulating them as potential candidate genes for ID (Figure 6.2). In three genes truncating variants were identified: TBR1, MAPK3, and NPTN. Proteins encoded by these genes were shown to be involved in various neurological functions or in related animal models. Other candidate genes with observed novel or very rare missense variants were NRG1, BSN, PCLO, SYT12, LRRC7, CNTN1/2, BDNF, LIMK1 and NRXN2. For each of them variations were identified in more than one patient case. A protein-protein interaction network analysis was done between the genes and the candidate genes showed direct interactions with already described known ID genes or its products or with each other further supporting their role as novel candidates for ID (Figure 6.2). The potential candidate genes described in the current study were also related in different pathways, which were previously described as linked to ID.

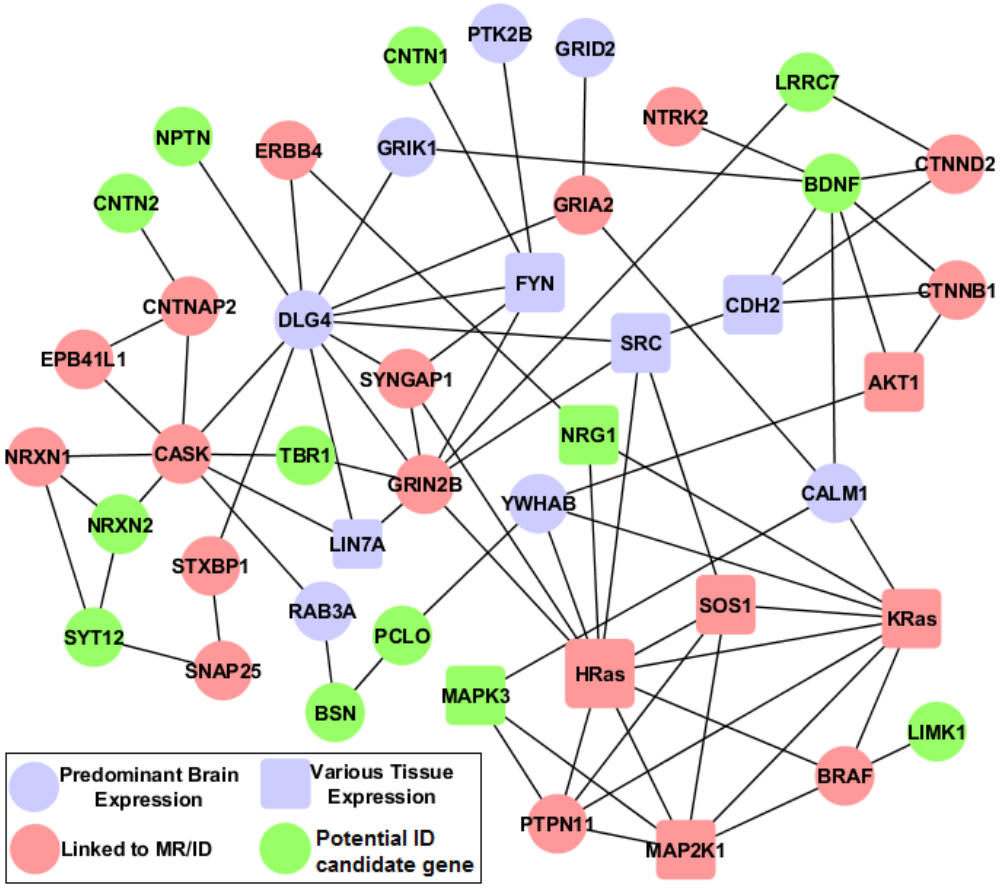


Figure 6.2: Network of genes showing the interactions between potential ID genes and already known ID genes in the project (prepared by using Cytoscape software, version 3.3) based on STRING interactions (Version 10). Lines indicate at least moderate evidence (STRING score > 0.3) for protein-protein interactions in humans.

6.3.8.1 T-box, brain 1 (TBR1)

We report a novel nonsense mutation (c.1275C>A, p.Tyr425*) and two novel heterozygous missense mutations (p.Gly140Arg, p.Thr247Ala) in three unrelated patients in the gene T-box, brain 1 (TBR1, OMIM 604616). TBR1 mutations are high-confidence risk factor for ASD and ID. Its evidence related to ID relies on the reference as candidate gene in a microdeletion region harbouring only TBR1 (Palumbo et al., 2014) and one reported observation of a de novo missense variant (Hamdan, F. F et al., 2014) in patients with ID along with the variable presence of ASD or growth retardation. Despite of these observations, TBR1 has not been classified as a human disease-associated gene in the OMIM catalogue due to insufficient additional reports and no robust supporting phenotype and functional data. The three male patients carrying TBR1 variations showed developmental delay with short stature, a phenotype which is consistent with described features in previous reports (Hamdan, F. F et al., 2014; Palumbo et al., 2014). The variants identified could not be confirmed as de novo due to lack of segregation analysis. The TBR1 gene product is specifically expressed in the brain. It is supposed to control neuronal migration and axonal projection of the cerebral cortex and amygdala. In mice, Tbr1 haploinsufficiency results in defective axonal projections and impairments of social interactions, ultrasonic vocalization, associative memory, and cognitive flexibility (Huang TN et al., 2014). TBR1 directly interacts with the genes CASK and GRIN2B which are already implicated in ID (Hsueh YP et al., 2000). It has been shown that the transcriptional activity of TBR1 increases with interaction with CASK thereby upregulating GRIN2B expression, resulting in synaptic stimulation and neurodevelopment (Huang TN., & Hsueh YP., 2017). All these functions suggest the potential role of TBR1 in learning and memory thereby making it a good candidate for ID in humans.

6.3.8.2 Mitogen-activated protein kinase 3 (MAPK3)

Here we report a novel truncating stop-gain mutation in MAPK3 gene (c.999 C>A, p.Tyr333*) in a patient presenting with non-specific developmental delay, language impairment and behavioural disorder. Mitogen-activated protein kinase 3 (MAPK3 or ERK1, OMIM 601795), a member of the MAPK family having important role in one of the prominent intracellular signalling pathway which controls various cellular responses like cell growth, differentiation, survival and also governing neural development and synaptic plasticity (Aoki Y et al., 2008; Ye and Carew, 2010) and is highly expressed in the brain. Mutations in various signalling molecules in the same pathway upstream from ERK are linked to RASopathies and other neurocognitive disorders (Rauen, 2013). Various studies through animal models support the role of ERK’s in brain development and maturation further

associating it with developmental disorders, including ID (Pucilowska et al., 2012). However, the suggested pathomechanism for genetic alterations in the RAS/MAPK pathway is overactivation causing to ERK hyper phosphorylation. The stop variant in MAPK3 instead very likely leads to loss of function of mutant allele suggesting haploinsufficiency as a possible pathomechanism, if this mutation was really disease-causing. The predicted likelihood of haploinsufficiency for this gene in ExAC is in medium range, suggesting probability of pathogenic effect for this particular variant. This nonsense mutation is of particular interest since the variant is located in the protein kinase-like domain of the protein which is a critical or well established functional domain and the domain is lost downstream to the altered splice site. The variant identified could not be confirmed as de novo due to lack of segregation analysis. But nevertheless the variant identified was an interesting one and with segregation analysis and more functional studies its pathogenicity could be confirmed.

6.3.8.3 Bassoon (BSN) & Piccolo (PCLO)

Bassoon (BSN, OMIM 604020) and Piccolo (PCLO, OMIM 604918) are two large presynaptic proteins at cytomatrix of active zone (CAZ), where neurotransmitters are released (tom Dieck et al., 1998). Their suggested functions include scaffolding and assembly of CAZ, organization of neurotransmitter release machinery, linkage of actin dynamics and endocytosis, maintenance of synapse integrity as well as integration of signalling pathways and synapto-nuclear signalling (Gundelfinger, Reissner & Garner, 2016; Fejtova et al., 2009; Hallermann et al., 2010). Defects in these proteins might therefore compromise the structure and function of synapses thus leading to “synaptopathies” that may manifest neurodevelopmental (ID, ASD, Fragile X syndrome), neurodegenerative (Alzheimer's, Parkinson's) and neuropsychiatric (bipolar, schizophrenia) disorders (Torres, Vallejo & Inestrosa, 2017).

Here we report novel missense mutations in PCLO gene which have been identified in the current study with pathogenicity of uncertain significance (p.Pro154Ser; p.Gln2956Glu; p.Glu4070Gly; p.Cys5033Tyr; p.Thr3754Ile; p.His3851Asn). Recent studies have suggested a possible linkage of PCLO mutations to autosomal recessive pontocerebellar hypoplasia type III (PCH3, OMIM 608027) (Ahmed et al. 2015), ASD and neuropsychiatric disorders (Minelli A et al. 2012). However, the confirmation of PCLO mutations as monogenic causes of human neurodevelopmental disorders is still pending. One of our patients who was found to carry a novel heterozygous missense mutation in the PCLO gene (c.12209A>G, p.Glu4070Gly) displayed major clinical features of severe developmental delay with short stature, hydrocephalus and hypoplasia of the corpus callosum. Some of these findings are overlapping with the described features in previous reports of PCH3. However, our patient did not have

any detectable mutation on the second allele. Therefore, the identified variant is insufficient to explain PCH3. In the current study, three (out of 6) novel missense mutations and five (out of 13) very rare missense variants were identified in exon 7 of the PCLO gene. This part of the gene encodes the coiled-coil region 3 (CC3) domain of the protein, which promotes interactions between Bassoon-Piccolo-ELKS/CAST and/or Munc13. These interactions are required for scaffolding and assembly of CAZ complex and for synaptic vesicle priming. Mutations affecting this part of the PCLO protein may result in altered interactions with their respective partners resulting in functional imbalance of the synapse thereby affecting synaptic plasticity, which is linked to several neurodevelopmental disorders (Gundelfinger, Reissner & Garner, 2016).

We also observed novel missense variations of uncertain significance in the BSN gene (p.Lys3722Asn; p.Ile2001Val; p.Asp2109His; p.Pro2901Ala; p.Arg3610Gln; p.Thr1585Ile). Recent studies have suggested BSN as a strong candidate gene for Landau-Kleffner syndrome (LKS, OMIM 245570) (Conroy J et al. 2014). Moreover chromosome 3p21.31 microdeletions encompassing BSN together with two other potential candidate genes for ID are associated with characteristic clinical features of developmental delay and distinctive facial features (Eto K et al., 2013). Bassoon mouse mutants (BsnΔEx4/5) display spontaneous epileptic seizures (Altrock et al., 2003) and also show enlargement of brain structures like cortex and hippocampus (Angenstein et al., 2007). Increased brain size in this model is correlated with unbalanced neurogenesis, reduced apoptosis and elevated BDNF levels in the hippocampus and other forebrain regions (Heyden et al., 2011). One of the patients presented here had a novel heterozygous missense mutation in BSN (c.6325G>C, p.Asp2109His). His major clinical features consisted of developmental delay, attention deficits, learning difficulties and facial dysmorphism (hypertelorism, broad nasal root, broad philtrum, small and low set ears, simple ears, and mild pterygium colli), sharing clinical findings with the described previous reports of LKS (Conroy J et al. 2014). Three novel mutations (out of 6) were identified in exon 5 of the BSN gene which corresponds to the coiled-coil region 2 (CC2), the largest domain. This exon points to a high evolutionary pressure frequently associated with a strong functional impact which involves in assembly of CAZ complex. The CC2 domain also promotes interactions between Piccolo-Bassoon (Gundelfinger, Reissner & Garner, 2016). In the current study, 18 very rare missense variants with population frequency <0.1% were also reported in both the genes. Particularly due to the lack of segregation data and since both are very large genes with a considerable genetic variation, and what we observed were only missense changes with no mutations indicating a clear loss of function, the essential data for further validation of PCLO and BSN as causative genes is still limiting. But with all the

functions inferred from mouse models suggest the potential role of PCLO and BSN in synapse maintenance and integrity, thereby making them as functionally good candidate genes for ID in human genetic studies.

6.3.8.4 Neuroplastin (NPTN)

Here we report a novel nonsense mutation in NPTN (c.475 C>T, p.Arg159*). The protein encoded by this gene, neuroplastin (NPTN, OMIM 612820), is a cell adhesion molecule which regulates synaptic plasticity i.e. long-term potentiation and formation/stabilization of excitatory synapses and balances the ratio of excitatory/inhibitory synapses (Herrera-Molina et al., 2017). NPTN has been proposed to have a potential role for regional synaptic dysfunctions in forms of intellectual deficits in an association study showing that polymorphisms in the regulatory region of the gene are associated with cortical thickness and intellectual abilities in adolescents (Desrivières et al., 2015). Knockout mouse models for the NPTN gene show a clear reduction in synaptic LTP and excitatory postsynaptic current amplitudes with several symptoms of autism, depression-like behaviour and retrograde amnesia for previously acquired associative memories (Bhattacharya et al., 2017). Recently, Herrera-Molina et al., 2017 have shown that deletion of NPTN in glutamatergic neurons impaired selective brain functions in mice and also calcium regulations, empowering the investigation of circuit-coded learning and memory and identification of causal mechanisms leading to cognitive deterioration. The nonsense mutation that was identified in our patient displaying developmental delay, speech delay with mild dysmorphisms including clinodactyly of 5th fingers, difficulties in balance and dosage of strength was found to be inherited from the healthy mother. The mutation likely leads to loss of function of the affected allele. However, the evidence for haploinsufficiency of NPTN is limited. The predicted likelihood of haploinsufficiency for this gene in ExAC is very high, suggesting the gene is extremely intolerant of loss-of-function variation. Our findings, although a single case reported thus raise the possibility of NPTN as a candidate ID gene, though further functional studies need to be done to confirm its pathogenicity.

6.3.8.5 Brain-derived Neurotrophic Factor (BDNF)

We report a novel heterozygous missense variation c.478 C>T, p.Arg160Trp in the gene encoding brain-derived neurotrophic factor (BDNF, OMIM 113505) in a patient presenting with non-specific developmental delay. Recently in an exome study done in 192 consanguineous families with non-syndromic ID, 26 novels genes not previously linked to recessive ID were reported (Harripaul R et al., 2017). Out of the 26 novel genes, nine genes had loss-of-function mutations and remaining missense mutations which included the first

reports of variants in BDNF or TET1 associated with ID. A homozygous missense mutation (p.Met122Thr) in BDNF gene was reported in a family with non-syndromic ID (Harripaul R et al., 2017). Several other studies have shown microdeletions encompassing BDNF in patients with various neurodevelopmental abnormalities, including ID (Ernst et al., 2012; Shinawi et al., 2011). BDNF, a member of the neurotrophin family is expressed predominantly in the brain and is involved in synapse formation and maturation, maintenance of synaptic plasticity and neuronal differentiation, growth and survival. Several studies in humans were done showing BDNF in memory impairment, linked to be associated affecting development, cognition, attention and behaviour (Park & Poo, 2013; Vilar & Mira, 2016) and also implied in various neuropsychiatric disorders (Autry & Monteggia 2012). Several animal models have been reported suggesting a role of BDNF in learning and memory (Cunha, Brambilla, & Thomas, 2010). Remarkably, association studies in humans with a single nucleotide polymorphism in the BDNF gene, c.196G>A (p.Val66Met; rs6265) which affects regulated release of BDNF, showed deficits in hippocampal and prefrontal cortical (PFC) plasticity and cognitive behaviours and is also known to influence both axonal and dendritic morphology and alteration of the initial dendritic outgrowth (Horch et al., 1999; Ninan I, 2014). BDNF is a known gene target of MECP2, the Rett syndrome gene and it has been suggested that haploinsufficiency for BDNF is associated with neurodevelopmental deficits in WAGR/11p13 deletion syndrome (Han JC et al., 2013). The variation identified in the patient of the present study cohort is of great interest and further segregation analyses need to be done to confirm its pathogenicity. Based on the human population studies showing various genetic associations and various experimental findings in animal models accumulate evidence that BDNF is a potential candidate gene for ID and further studies need to be performed.

Here we also report three novel missense variants in contactins, CNTN1 (c.1139A>T, p.Asp380Val) and CNTN2 (c.1247C>A, p.Ala416Asp; c.1856T>C, p.Ile619Thr). Contactins (CNTNs) are neuronal cell adhesion molecules which control major aspects of neurogenesis like synapse formation, axon growth and guidance, adhesion and migration of neuronal cells (Mohebiany et al., 2014). In the current study, ten very rare missense variants with population frequency <0.1% were also reported in both the genes. CNTN1 and CNTN2 are not yet listed as OMIM genes for ID or any other human disease except for a vague association with a myopathy and type of epilepsy, respectively in recessive forms. Mutations in CNTN1 gene though in humans cause congenital myopathy (OMIM # 612540), it has been shown otherwise in mice, where spontaneous mutations revealed neurodevelopmental phenotypes (Davisson et al., 2011). The *Cntn1* mice did not show any signs of myopathy but instead showed dysfunction in the nervous system with a severe ataxic phenotype and consistent

defects in the cerebellum especially in granule cell axon guidance and in dendritic projections (Davisson et al., 2011; Berglund EO et al., 1999). The *Cntn2* mice showed an increased LTP and improved spatial and object recognition memory with a possible role in promoting adult hippocampal neurogenesis (Puzzo et al., 2013). Moreover a very strong association between the genes CNTN2 and CNTNAP2 (contactin-associated protein 2; CASPR2) was shown in various studies (Lu Z et al., 2016; Buchner DA et al., 2012). In humans, mutations in the CNTNAP2 gene are associated with a variety of neurological disorders including ID, ASD and language delay (Lu Z et al., 2016). But with all the functions inferred from mouse models suggest the potential role CNTNs in involvement in improved synaptic function and memory, thereby making them as functionally good candidate genes for ID in human genetic studies.

Furthermore, in the current study other variants of possible functional relevance were also identified within genes like SYT12, LRRC7, LIMK1, NRXN2 and NRG1. These genes had no clear known function and are involved in diverse pathways. In the current study, although only few numbers of significant mutational changes were identified in genes not linked to ID and there is still scope that some candidate genes described might emerge as false positives. Despite the fact that some of the genes listed above appear to be really good candidate genes for ID, none of the changes observed in patients from this cohort confidently predict pathogenic significance at the current stage of the studies. One weakness is of course, as mentioned above is that segregation analysis for many variants is still missing, which would be the first step in further validation. Replication of the identified mutations and establishing genotype–phenotype correlations would be the second step to increase evidence followed by further functional studies. Functional studies, either in-vitro using patient-derived cells or in-vivo using animal models provides deeper knowledge in the function of the genes or mutations.

6.3.9 Case studies: Patients with Multiple Variants

ID is not always monogenic. Instead, a multifactorial pathogenesis has to be considered to also contribute. Here we presented some cases with more than one variant of possible pathogenicity (Section 5.2.14). We were considering that multifactorial / digenic or polygenic inheritance may contribute to ID, especially for cases with familial learning disabilities and cases with milder expression of ID. In the families we have described, there are different lines of evidence for the possible pathogenic role of the two or more observed variations. But one has to keep in mind that the study covers only a small fraction of all human genes. Therefore the picture cannot be complete. In some cases, interestingly, variants in genes for interacting proteins were found. This could be a hint that genetic impairment of a pathway or biological

functional complex by rare mutations leading to minor alterations at different sites might cumulatively lead to relevant changes that would not be caused by any of these changes alone. But the presented observations cannot be more than just a very first hint at such mechanisms. With the increase of genetic studies for ID and use of NGS methods, many new genes and common variants are being detected and studies showing associations or combinations between the genes/variants are signifying ID as a highly genetically heterogeneous polygenic disorder involving multiple different genetic loci/variant combinations (McCarroll & Hyman, 2013). Recent hypothesis suggest that multiple variants which are inter-related in some biological pathways act together presenting ID (Franić S et al., 2015). However, it remains to be proven that the combination of two mutations is indeed the cause of disease rather than the simple co-occurrence of two mutations by chance (Cooper et al., 2013). In the current study, multigenic studies were not able to be proven because of low sample size and unclear inheritance patterns or limited segregation analysis. The families (case studies) with more than one putative deleterious mutation described in this study might also have false positives. Thus the evidence we provided here of the variant-gene effect remains hypothetical and the data are much too preliminary to postulate multigenic concept only providing just a hint towards it.

6.4 Short stature study cohort

Short stature is a heterogeneous trait which can be caused due to multiple molecular defects in various intracellular signalling pathways (like RAS/MAPK; cyclic AMP (cAMP)-dependent or WNT5A/JNK signalling pathways), extracellular matrix components (genes that encode matrix collagens; proteoglycans; non-collagenous proteins; and their processing enzymes) and paracrine (FGFs; Indian hedgehog (IHH); bone morphogenetic factors (BMPs)) and endocrine (growth hormone (GH); STAT5B) regulation (Jee Y.H. et al., 2017). The longitudinal growth of the bones is mainly due to the chondrogenesis at the growth plate and a decrease in this configuration results in all forms of short stature (Jee Y.H. & Baron J, 2016). A key pathway which has been identified in regulation of growth plate chondrogenesis is RAS/MAPK signalling, in which alterations result in several disorders called RASopathies (Aoki Y et al., 2016). It has also been implicated that short stature is a common feature noticeable in these syndromes. Apart from the affected signal proteins of the RAS/MAPK pathway, SHP2, encoded by PTPN11, is known to be implicated in GH signalling related to short stature (Serra-Nedelec et al., 2012). Besides SHP2, other interconnections exist between the RAS/MAPK and GH pathways that remain to be elucidated. In the current study, a total of 120 patients diagnosed with short stature were screened using targeted NGS analysis through

a panel of 329 genes, which were related directly or indirectly to the RAS/MAPK signalling pathway and also either known short stature genes or candidate GH-associated genes. By applying all the filtering strategies for identification of very rare and novel variants, a total of 780 rare high quality variants in short stature cohort were generated. Within this cohort an average, in the examined short stature related pathway genes, ~ 3 very rare or unknown variants were identified per individual (Range: 0-12). A wide number of novel and rare variants were identified in this cohort but further analysis of these variants was not done in this study as it belongs to a collaborative project. In the current study, this cohort served as an internal control for comparing the enrichment of variants in certain genes and for total number of variants identified between the ID and Short stature cohorts.

This study was combined with a study on genetic basis in a short stature cohort, because of the shared RAS/MAPK pathway genes as candidates for both cohorts. In addition to the shared targets, each cohort had its specific set of target genes. By combination of these studies with an analysis using the same tool (panel) we could take advantage of the possibility to use the other cohort as control. Although the analysis of the sequence data from the short stature cohort was not a primary goal of this thesis work, we could contribute two cases to a study reporting ACAN (Aggrecan) mutations in non-syndromic short stature. From the current study and collaboration with other researchers, heterozygous mutations in the ACAN gene were associated with idiopathic short stature (Hauer et al., 2017). Aggrecan is the main proteoglycan of the extracellular matrix of the growth plate cartilage (Lauing et al., 2014) and mutations in this gene are associated with growth defects ranging from mild idiopathic short stature to severe skeletal dysplasias (Gibson & Briggs, 2016). From the current study, out of 120 patients analysed, one likely pathogenic missense variant (c.1702G>A; p.Asp568Asn) and another missense variant with uncertain pathogenic significance (c.1636G>A; p.Val546 Met) in ACAN gene was identified. The male patient presenting with the p.Asp568Asn had a height of −3.2 SDS, head circumference of −0.2 SDS with proportionate short stature and delayed bone aging. This variant is present in ExAC with seven heterozygous counts in the European (Non-Finnish) population out of a total of 120688 alleles (ExAC.broadinstitute.org). This variant affects a highly conserved amino acid, and was inherited from the mother with a height of −2.0 SDS. The p.Asp568Asn variant is located in the third link domain (residues 478–573), a hyaluronan-binding protein module and affects the interactions between the C-terminus and the N-terminus by weakening them which may cause destabilization of entire domain thereby affecting its ligand binding properties (Hauer et al., 2017). The patient received growth hormone therapy improving his height from −3.2 SDS to −1.7 SDS. So, identifying and understanding the genetic basis of short stature will have significant impact on

the care of children seeking medical attention for severe short stature and which also improves basic understanding of skeletal development and growth.

6.5 Conclusion

In conclusion, our study evaluated the importance and frequency of mutations in a panel of RAS/MAPK pathway-related genes in patients with non-specific and non-syndromic ID.
The methodology (multigene panel sequencing) proved to be useful. However, the yield of solved cases was quite low in this setting. So for addressing ID in all its genetic heterogeneity and complexity exome or genome sequencing has to be recommended – which has already become routine in many labs in the meantime. The sequencing runs overall quality in the present study was good, yielding potential pathogenic variants. We had some interesting novel variations in novel candidates, but for all of them the current study with its limitations could not provide sufficient evidence for pathogenicity. Two of 166 cases have been definitely solved. Another 45 cases have possible candidate variants in known ID genes or novel candidates, but with variable degrees of evidence supporting the pathogenic role of the respective variants. In order to identify further mutations, targeted examinations in larger cohorts are to be carried out for the best new candidate genes and/or co-operation with other groups that have operated exome sequencing in ID cohorts. Further analyses with functional studies have to be done for the candidate ID genes to conclude that these rare variants are associated with RASopathy or RAS/MAPK pathogenesis. For increasing the diagnostic yields and for identification of novel ID-associated genes, a trio analysis would be priority criteria since in the current study a proband-only approach was done where one or both parents were unavailable for testing in maximum cases. In the current study, we have also compared the performance of the two enrichment protocols and significant differences were reported.

The use of NGS provided enhanced efficiency and accuracy for high sensitive detection of molecular mutations in patient cohorts and complex phenotypes at less cost. The limitations of the studies presented here have to be considered from the perspective of rapidly evolving NGS technologies. When this study was started, big sequencers were only available in huge genome sequencing centres and everyone started with the low-capacity instruments and small panels. The same project designed today would probably use exome sequencing due to the decrease in cost and increase in efficiency of the NGS technologies.

7. Outlook

With the advent of new sequencing technologies, there is a rapid increase in the identification of novel genes that cause ID thus opening new interactions between different pathways and emerging evidence for certain biological pathways as principal contributors for normal cognitive development. Novel findings would further help in providing new insights and exploring the roles and functions of the associated genes in different biological pathways involved in human neurological development and cognitive functioning. Until now most of the studies or research related to ID is based on germline highly penetrant monogenic causes of ID which has been very successful but leaving a large section of unexplained milder forms of ID having no neurologic features or malformations. New research directions further explaining the causes need to be studied like somatic causes, non-coding causes and studying increasing inheritance complexity in ID. Learning more about the different functions and classes of genes/proteins affecting ID would help us improve our understanding of these mutations effecting at molecular level as well as the phenotypic consequences. For understanding the genetics of ID and improve diagnostic screening, various studies by different groups and in addition to our findings helps in for development of targeted gene sequencing panels thereby adding many more genes to the current catalogue of ID genes and which is also essential for whole exome/genome sequencing. To achieve these three aspects should be considered like replication of the findings, more functional studies for identified variants and consecutive new discoveries is very much essential. Through replication studies the confidence in validation of the discoveries increases which could be achieved by large data sets. For determining the functional role of identified mutations and for comparison purposes, developing model organisms would provide a deeper knowledge in the function related to human development. Moreover in near future, development of model organisms using new technologies like CRISPR-Cas9 system will allow rapid generation of models and studying the expression of all mutations associated with the syndrome or study specific mutations in different genetic backgrounds. Mimicking human cell nature with iPSCs derived from patient could also be used for accelerating functional studies and providing clues as to molecular processes that may underlie for ID. Studying the molecular causes of ID will be important, not only for the purposes of genetic counselling and screening, but also by knowing the molecular mechanisms and components would increase our knowledge on brain function and helps in further understanding of the full functioning of CNS. Furthermore, to achieve the complete story many more studies and novel discoveries have to be made thus pushing the challenge, a step closer, in finding therapeutic approaches for ID patients.

References

Ahmed, M.Y., Chioza, B.A., Rajab, A., et al. (2015). Loss of PCLO function underlies pontocerebellar hypoplasia type III. Neurology 84, 1745-1750.

Akagi, K., Uchibori, R., Yamaguchi, K., et al. (2007). Characterization of a novel oncogenic K-ras mutation in colon cancer. Biochemical and biophysical research communications 352, 728-732.

Alame, M., Lacourt, D., Zenagui, R., et al. (2016). Implementation of a Reliable Next-Generation Sequencing Strategy for Molecular Diagnosis of Dystrophinopathies. J Mol Diagn 18, 731-740.

Alcantara, D., Timms, A.E., Gripp, K., et al. (2017). Mutations of AKT3 are associated with a wide spectrum of developmental disorders including extreme megalencephaly. Brain 140, 2610-2622.

Alomari, A.I. (2009). Characterization of a distinct syndrome that associates complex truncal overgrowth, vascular, and acral anomalies: a descriptive study of 18 cases of CLOVES syndrome. Clin Dysmorphol 18, 1-7.

Altrock, W.D., tom Dieck, S., Sokolov, M., et al. (2003). Functional inactivation of a fraction of excitatory synapses in mice deficient for the active zone protein bassoon. Neuron 37, 787-800.

Ana, F.-M., Sofía, G., and María, L.C. (2017). NGS Technologies as a Turning Point in Rare Disease Research, Diagnosis, and Treatment. Current Medicinal Chemistry 24, 1-29.

Aoki, Y., Niihori, T., Banjo, T., et al. (2013). Gain-of-function mutations in RIT1 cause Noonan syndrome, a RAS/MAPK pathway syndrome. Am J Hum Genet 93, 173-180.

Aoki, Y., Niihori, T., Inoue, S., et al. (2016). Recent advances in RASopathies. J Hum Genet 61, 33-39.

Aoki, Y., Niihori, T., Narumi, Y., et al. (2008). The RAS/MAPK syndromes: novel roles of the RAS pathway in human genetic disorders. Hum Mutat 29, 992-1006.

Araki, Y., Zeng, M., Zhang, M., et al. (2015). Rapid dispersion of SynGAP from synaptic spines triggers AMPA receptor insertion and spine enlargement during LTP. Neuron 85, 173-189.

Ardinger, H.H., Horii, K.A., and Begleiter, M.L. (2007). Expanding the phenotype of oculoectodermal syndrome: possible relationship to encephalocraniocutaneous lipomatosis. Am J Med Genet A 143A, 2959-2962.

Arsenic, R., Treue, D., Lehmann, A., et al. (2015). Comparison of targeted next-generation sequencing and Sanger sequencing for the detection of PIK3CA mutations in breast cancer. BMC Clin Pathol 15, 20.

Autry, A.E., and Monteggia, L.M. (2012). Brain-derived neurotrophic factor and neuropsychiatric disorders. Pharmacol Rev 64, 238-258.

Bakkaloglu, B., O'Roak, B.J., Louvi, A., et al. (2008). Molecular cytogenetic analysis and resequencing of contactin associated protein-like 2 in autism spectrum disorders. Am J Hum Genet 82, 165-173.

Baldassa, S., Gnesutta, N., Fascio, U., et al. (2007). SCLIP, a microtubule-destabilizing factor, interacts with RasGRF1 and inhibits its ability to promote Rac activation and neurite outgrowth. The Journal of biological chemistry 282, 2333-2345.

Bateup, H.S., Johnson, C.A., Denefrio, C.L., et al. (2013). Excitatory/inhibitory synaptic imbalance leads to hippocampal hyperexcitability in mouse models of tuberous sclerosis. Neuron 78, 510-522.

Becker-Santos, D.D., Lonergan, K.M., Gronostajski, R.M., et al. (2017). Nuclear Factor I/B: A Master Regulator of Cell Differentiation with Paradoxical Roles in Cancer. EBioMedicine 22, 2-9.

Bennett, James T., Tan, Tiong Y., Alcantara, D., et al. Mosaic Activating Mutations in FGFR1 Cause Encephalocraniocutaneous Lipomatosis. The American Journal of Human Genetics 98, 579-587.

Berkel, S., Marshall, C.R., Weiss, B., et al. (2010). Mutations in the SHANK2 synaptic scaffolding gene in autism spectrum disorder and mental retardation. Nat Genet 42, 489-491.

Berglund EO, Murai KK, Fredette B, et al. (1999) Ataxia and abnormal cerebellar microorganization in mice with ablated contactin gene expression. Neuron. 24:739–750.

Bhattacharya, S., Herrera-Molina, R., Sabanov, V., et al. (2017). Genetically Induced Retrograde Amnesia of Associative Memories After Neuroplastin Ablation. Biological psychiatry 81, 124-135.

Bieser, S., Reis, M., Guzman, M., et al. (2015). Grade II pilocytic astrocytoma in a 3-month-old patient with encephalocraniocutaneous lipomatosis (ECCL): case report and literature review of low grade gliomas in ECCL. Am J Med Genet A 167A, 878-881.

Biou, V., Brinkhaus, H., Malenka, R.C., et al. (2008). Interactions between drebrin and Ras regulate dendritic spine plasticity. Eur J Neurosci 27, 2847-2859.

Bodi, K., Perera, A.G., Adams, P.S., et al. (2013). Comparison of commercially available target enrichment methods for next-generation sequencing. J Biomol Tech 24, 73-86.

Bonnet, C., Masurel-Paulet, A., Khan, A.A., et al. (2012). Retracted: Exploring the potential role of disease-causing mutation in a gene desert: Duplication of noncoding elements 5′ of GRIA3 is associated with GRIA3 silencing and X-linked intellectual disability. Human Mutation 33, 355-358.

Boppudi, S., Bogershausen, N., Hove, H.B., et al. (2016). Specific mosaic KRAS mutations affecting codon 146 cause oculoectodermal syndrome and encephalocraniocutaneous lipomatosis. Clin Genet 90, 334-342.

Borrie, S.C., Brems, H., Legius, E., et al. (2017). Cognitive Dysfunctions in Intellectual Disabilities: The Contributions of the Ras-MAPK and PI3K-AKT-mTOR Pathways. Annual Review of Genomics and Human Genetics 18, 115-142.

Bosch, D.G., Boonstra, F.N., de Leeuw, N., et al. (2016). Novel genetic causes for cerebral visual impairment. Eur J Hum Genet 24, 660-665.

Brambilla, R., Gnesutta, N., Minichiello, L., et al. (1997). A role for the Ras signalling pathway in synaptic transmission and long-term memory. Nature 390, 281-286.

Buchner DA, Geisinger JM, Glazebrook PA, et al. (2012). The juxtaparanodal proteins CNTNAP2 and TAG1 regulate diet-induced obesity. Mammalian genome : official journal of the International Mammalian Genome Society. 23(0):431-442.

Buske, O.J., Manickaraj, A., Mital, S., et al. (2013). Identification of deleterious synonymous variants in human genomes. Bioinformatics 29, 1843-1850.

Campbell, I.M., Shaw, C.A., Stankiewicz, P., et al. (2015). Somatic mosaicism: implications for disease and transmission genetics. Trends Genet 31, 382-392.

Carvill, G.L., Weckhuysen, S., McMahon, J.M., et al. (2014). GABRA1 and STXBP1: novel genetic causes of Dravet syndrome. Neurology 82, 1245-1253.

Chan, C.B., Liu, X., Pradoldej, S., et al. (2011). Phosphoinositide 3-kinase enhancer regulates neuronal dendritogenesis and survival in neocortex. The Journal of neuroscience : the official journal of the Society for Neuroscience 31, 8083-8092.

Charbel, C., Fontaine, R.H., Malouf, G.G., et al. (2014). NRAS mutation is the sole recurrent somatic mutation in large congenital melanocytic nevi. J Invest Dermatol 134, 1067-1074.

Chaudhry, A.Z., Lyons, G.E., and Gronostajski, R.M. (1997). Expression patterns of the four nuclear factor I genes during mouse embryogenesis indicate a potential role in development. Dev Dyn 208, 313-325.

Chen, A.P., Ohno, M., Giese, K.P., et al. (2006). Forebrain-specific knockout of B-raf kinase leads to deficits in hippocampal long-term potentiation, learning, and memory. J Neurosci Res 83, 28-38.

Cheng, T.M., Goehring, L., Jeffery, L., et al. (2012). A structural systems biology approach for quantifying the systemic consequences of missense mutations in proteins. PLoS Comput Biol 8, e1002738.

Cheyne, J.E., Grant, L., Butler-Munro, C., et al. (2011). Synaptic integration of newly generated neurons in rat dissociated hippocampal cultures. Molecular and cellular neurosciences 47, 203-214.

Chilamakuri, C.S., Lorenz, S., Madoui, M.A., et al. (2014). Performance comparison of four exome capture systems for deep sequencing. BMC Genomics 15, 449.

Chiyonobu, T., Hayashi, S., Kobayashi, K., et al. (2007). Partial tandem duplication of GRIA3 in a male with mental retardation. Am J Med Genet A 143A, 1448-1455.

Clarke, C. (2015). Autism Spectrum Disorder and Amplified Pain. Case Rep Psychiatry 2015, 930874.

Clarke, T.K., Lupton, M.K., Fernandez-Pujals, A.M., et al. (2016). Common polygenic risk for autism spectrum disorder (ASD) is associated with cognitive ability in the general population. Mol Psychiatry 21, 419-425.

Clement, J.P., Aceti, M., Creson, T.K., et al. (2012). Pathogenic SYNGAP1 mutations impair cognitive development by disrupting maturation of dendritic spine synapses. Cell 151, 709-723.

Conquet, F., Bashir, Z.I., Davies, C.H., et al. (1994). Motor deficit and impairment of synaptic plasticity in mice lacking mGluR1. Nature 372, 237-243.

Conroy, J., McGettigan, P.A., McCreary, D., et al. (2014). Towards the identification of a genetic basis for Landau-Kleffner syndrome. Epilepsia 55, 858-865.

Cooper, D.N., Krawczak, M., Polychronakos, C., et al. (2013). Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. Hum Genet 132, 1077-1130.

Cordeddu, V., Di Schiavi, E., Pennacchio, L.A., et al. (2009). Mutation in SHOC2 promotes aberrant protein N-myristoylation and underlies Noonan-like syndrome with loose anagen hair. Nature genetics 41, 1022-1026.

Cunha, C., Brambilla, R., and Thomas, K.L. (2010). A simple role for BDNF in learning and memory? Front Mol Neurosci 3, 1.

Daily, D.K., Ardinger, H.H., and Holmes, G.E. (2000). Identification and evaluation of mental retardation. Am Fam Physician 61, 1059-1067, 1070.

Danielsson, K., Mun, L.J., Lordemann, A., et al. (2014). Next-generation sequencing applied to rare diseases genomics. Expert Rev Mol Diagn 14, 469-487.

Davarniya, B., Hu, H., Kahrizi, K., et al. (2015). The Role of a Novel TRMT1 Gene Mutation and Rare GRM1 Gene Defect in Intellectual Disability in Two Azeri Families. PloS one 10, e0129631.

Davisson, M.T., Bronson, R.T., Tadenev, A.L., et al. (2011). A spontaneous mutation in contactin 1 in the mouse. PloS one 6, e29538.

de Ligt, J., Willemsen, M.H., van Bon, B.W., et al. (2012). Diagnostic exome sequencing in persons with severe intellectual disability. N Engl J Med 367, 1921-1929.

De Rocca Serra-Nedelec, A., Edouard, T., Treguer, K., et al. (2012). Noonan syndrome-causing SHP2 mutants inhibit insulin-like growth factor 1 release via growth hormone-induced ERK hyperactivation, which contributes to short stature. Proceedings of the National Academy of Sciences of the United States of America 109, 4257-4262.

Desrivieres, S., Lourdusamy, A., Tao, C., et al. (2015). Single nucleotide polymorphism in the neuroplastin locus associates with cortical thickness and intellectual ability in adolescents. Mol Psychiatry 20, 263-274.

Do, H., and Dobrovic, A. (2015). Sequence artifacts in DNA from formalin-fixed tissues: causes and strategies for minimization. Clin Chem 61, 64-71.

Docker, D., Schubach, M., Menzel, M., et al. (2015). Germline PTPN11 and somatic PIK3CA variant in a boy with megalencephaly-capillary malformation syndrome (MCAP)--pure coincidence? Eur J Hum Genet 23, 409-412.

Dogruluk, T., Tsang, Y.H., Espitia, M., et al. (2015). Identification of Variant-Specific Functions of PIK3CA by Rapid Phenotyping of Rare Mutations. Cancer Res 75, 5341-5354.

Dong, F., Jiang, J., McSweeney, C., et al. (2016). Deletion of CTNNB1 in inhibitory circuitry contributes to autism-associated behavioral defects. Hum Mol Genet 25, 2738-2751.

Donzis, E.J., and Tronson, N.C. (2014). Modulation of learning and memory by cytokines: signaling mechanisms and long term consequences. Neurobiology of learning and memory 115, 68-77.

Edkins, S., O'Meara, S., Parker, A., et al. (2006). Recurrent KRAS codon 146 mutations in human colorectal cancer. Cancer Biol Ther 5, 928-932.

Eggert, S.L., Huyck, K.L., Somasundaram, P., et al. (2012). Genome-wide linkage and association analyses implicate FASN in predisposition to Uterine Leiomyomata. Am J Hum Genet 91, 621-628.

Engelman, J.A., Luo, J., and Cantley, L.C. (2006). The evolution of phosphatidylinositol 3-kinases as regulators of growth and metabolism. Nat Rev Genet 7, 606-619.

English, J.D., and Sweatt, J.D. (1997). A requirement for the mitogen-activated protein kinase cascade in hippocampal long term potentiation. The Journal of biological chemistry 272, 19103-19106.

Ernst, C., Marshall, C.R., Shen, Y., et al. (2012). Highly penetrant alterations of a critical region including BDNF in human psychopathology and obesity. Arch Gen Psychiatry 69, 1238-1246.

Eto, K., Sakai, N., Shimada, S., et al. (2013). Microdeletions of 3p21.31 characterized by developmental delay, distinctive features, elevated serum creatine kinase levels, and white matter involvement. Am J Med Genet A 161A, 3049-3056.

Evrony, G.D., Cai, X., Lee, E., et al. (2012). Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. Cell 151, 483-496.

Falivelli, G., De Jaco, A., Favaloro, F.L., et al. (2012). Inherited genetic variants in autism-related CNTNAP2 show perturbed trafficking and ATF6 activation. Human Molecular Genetics 21, 4761-4773.

Feig, L.A., and Cooper, G.M. (1988). Relationship among guanine nucleotide exchange, GTP hydrolysis, and transforming potential of mutated ras proteins. Molecular and Cellular Biology 8, 2472-2478.

Fejtova, A., Davydova, D., Bischof, F., et al. (2009). Dynein light chain regulates axonal trafficking and synaptic levels of Bassoon. The Journal of cell biology 185, 341-355.

Fivaz, M., and Meyer, T. (2005). Reversible intracellular translocation of KRas but not HRas in hippocampal neurons regulated by Ca2+/calmodulin. The Journal of cell biology 170, 429-441.

Flex, E., Jaiswal, M., Pantaleoni, F., et al. (2014). Activating mutations in RRAS underlie a phenotype within the RASopathy spectrum and contribute to leukaemogenesis. Hum Mol Genet 23, 4315-4327.

Fraser, M.M., Zhu, X., Kwon, C.H., et al. (2004). Pten loss causes hypertrophy and increased proliferation of astrocytes in vivo. Cancer Res 64, 7773-7779.

Frayling, T.M., Timpson, N.J., Weedon, M.N., et al. (2007). A Common Variant in the FTO Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity. Science 316, 889-894.

Franić, S., Groen-Blokhuis, M. M., Dolan, C. V., et al. (2015). Intelligence: shared genetic basis between Mendelian disorders and a polygenic trait. European Journal of Human Genetics, 23(10), 1378–1383.

Fusco, C., Frattini, D., and Bassi, M.T. (2015). A novel KCNQ3 gene mutation in a child with infantile convulsions and partial epilepsy with centrotemporal spikes. Eur J Paediatr Neurol 19, 102-103.

Garcia-Garcia, G., Baux, D., Faugere, V., et al. (2016). Assessment of the latest NGS enrichment capture methods in clinical context. Sci Rep 6, 20948.

Gauthier, J., Champagne, N., Lafreniere, R.G., et al. (2010). De novo mutations in the gene encoding the synaptic scaffolding protein SHANK3 in patients ascertained for schizophrenia. Proceedings of the National Academy of Sciences of the United States of America 107, 7863-7868.

Gauthier, J., Spiegelman, D., Piton, A., et al. (2009). Novel de novo SHANK3 mutation in autistic patients. Am J Med Genet B Neuropsychiatr Genet 150B, 421-424.

Gibson, B.G., and Briggs, M.D. (2016). The aggrecanopathies; an evolving phenotypic spectrum of human genetic skeletal diseases. Orphanet J Rare Dis 11, 86.

Gilbert, J.A., and Dupont, C.L. (2011). Microbial metagenomics: beyond the genome. Ann Rev Mar Sci 3, 347-371.

Giniger, E. (2012). Notch signaling and neural connectivity. Curr Opin Genet Dev 22, 339-346.

Goodwin, C.B., Yang, Z., Yin, F., et al. (2012). Genetic disruption of the PI3K regulatory subunits, p85alpha, p55alpha, and p50alpha, normalizes mutant PTPN11-induced hypersensitivity to GM-CSF. Haematologica 97, 1042-1047.

Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17, 333-351.

Gosens, R., Baarsma, H.A., Heijink, I.H., et al. (2010). De novo synthesis of {beta}-catenin via H-Ras and MEK regulates airway smooth muscle growth. FASEB J 24, 757-768.

Gray, J., Yeo, G., Hung, C., et al. (2007). Functional characterization of human NTRK2 mutations identified in patients with severe early-onset obesity. Int J Obes (Lond) 31, 359-364.

Gremer, L., Merbitz-Zahradnik, T., Dvorsky, R., et al. (2011). Germline KRAS mutations cause aberrant biochemical and physical properties leading to developmental disorders. Hum Mutat 32, 33-43.

Griffith, M., Miller, C.A., Griffith, O.L., et al. (2015). Optimizing cancer genome sequencing and analysis. Cell Syst 1, 210-223.

Gripp, K.W., Stabley, D.L., Nicholson, L., et al. (2006). Somatic mosaicism for an HRAS mutation causes Costello syndrome. Am J Med Genet A 140, 2163-2169.

Groesser, L., Herschberger, E., Ruetten, A., et al. (2012). Postzygotic HRAS and KRAS mutations cause nevus sebaceous and Schimmelpenning syndrome. Nat Genet 44, 783-787.

Grozeva, D., Carss, K., Spasic-Boskovic, O., et al. (2015). Targeted Next-Generation Sequencing Analysis of 1,000 Individuals with Intellectual Disability. Hum Mutat 36, 1197-1204.

Gucev, Z.S., Tasic, V., Jancevska, A., et al. (2008). Congenital lipomatous overgrowth, vascular malformations, and epidermal nevi (CLOVE) syndrome: CNS malformations and seizures may be a component of this disorder. Am J Med Genet A 146A, 2688-2690.

Gundelfinger, E.D., Reissner, C., and Garner, C.C. (2015). Role of Bassoon and Piccolo in Assembly and Molecular Organization of the Active Zone. Frontiers in synaptic neuroscience 7, 19.

Gymnopoulos, M., Elsliger, M.-A., and Vogt, P.K. (2007). Rare cancer-specific mutations in PIK3CA show gain of function. Proceedings of the National Academy of Sciences 104, 5569-5574.

Hafner, C., Toll, A., Gantner, S., et al. (2012). Keratinocytic epidermal nevi are associated with mosaic RAS mutations. J Med Genet 49, 249-253.

Hallermann, S., Fejtova, A., Schmidt, H., et al. Bassoon Speeds Vesicle Reloading at a Central Excitatory Synapse. Neuron 68, 710-723.

Ham, H., Guerrier, S., Kim, J., et al. (2013). Dedicator of cytokinesis 8 interacts with talin and Wiskott-Aldrich syndrome protein to regulate NK cell cytotoxicity. J Immunol 190, 3661-3669.

Hamdan, F.F., Daoud, H., Piton, A., et al. (2011). De novo SYNGAP1 mutations in nonsyndromic intellectual disability and autism. Biological psychiatry 69, 898-901.

Hamdan, F.F., Gauthier, J., Dobrzeniecka, S., et al. (2011). Intellectual disability without epilepsy associated with STXBP1 disruption. Eur J Hum Genet 19, 607-609.

Hamdan, F.F., Gauthier, J., Spiegelman, D., et al. (2009). Mutations in SYNGAP1 in autosomal nonsyndromic mental retardation. N Engl J Med 360, 599-605.
Hamdan, F.F., Piton, A., Gauthier, J., et al. (2009). De novo STXBP1 mutations in mental retardation and nonsyndromic epilepsy. Ann Neurol 65, 748-753.
Hamdan, F.F., Srour, M., Capo-Chichi, J.M., et al. (2014). De novo mutations in moderate or severe intellectual disability. PLoS Genet 10, e1004772.
Han, J.C., Thurm, A., Golden Williams, C., et al. (2013). Association of brain-derived neurotrophic factor (BDNF) haploinsufficiency with lower adaptive behaviour and reduced cognitive functioning in WAGR/11p13 deletion syndrome. Cortex 49, 2700-2710.
Handsaker, R.E., Van Doren, V., Berman, J.R., et al. (2015). Large multiallelic copy number variations in humans. Nat Genet 47, 296-303.
Happle, R. (1987). Lethal genes surviving by mosaicism: a possible explanation for sporadic birth defects involving the skin. J Am Acad Dermatol 16, 899-906.
Happle, R., and Kuster, W. (1998). Nevus psiloliparus: a distinct fatty tissue nevus. Dermatology 197, 6-10.
Harripaul, R., Vasli, N., Mikhailov, A., et al. (2017). Mapping autosomal recessive intellectual disability: combined microarray and exome sequencing identifies 26 novel candidate genes in 192 consanguineous families. Mol Psychiatry.
Hauer, N.N., Sticht, H., Boppudi, S., et al. (2017). Genetic screening confirms heterozygous mutations in ACAN as a major cause of idiopathic short stature. Sci Rep 7, 12225.
Hedegaard, J., Thorsen, K., Lund, M.K., et al. (2014). Next-generation sequencing of RNA and DNA isolated from paired fresh-frozen and formalin-fixed paraffin-embedded samples of human cancer and normal tissue. PloS one 9, e98187.
Henske, E.P., Jozwiak, S., Kingswood, J.C., et al. (2016). Tuberous sclerosis complex. Nat Rev Dis Primers 2, 16035.
Hernandez-Porras, I., and Guerra, C. (2017). Modeling RASopathies with Genetically Modified Mouse Models. Methods Mol Biol 1487, 379-408.
Herrera-Molina, R., Mlinac-Jerkovic, K., Ilic, K., et al. (2017). Neuroplastin deletion in glutamatergic neurons impairs selective brain functions and calcium regulation: implication for cognitive deterioration. Sci Rep 7, 7273.
Heumann, R., Goemans, C., Bartsch, D., et al. (2000). Transgenic activation of Ras in neurons promotes hypertrophy and protects from lesion-induced degeneration. The Journal of cell biology 151, 1537-1548.
Heyden, A., Ionescu, M.C., Romorini, S., et al. (2011). Hippocampal enlargement in Bassoon-mutant mice is associated with enhanced neurogenesis, reduced apoptosis, and abnormal BDNF levels. Cell Tissue Res 346, 11-26.
Hofmeister, W., Nilsson, D., Topa, A., et al. (2015). CTNND2-a candidate gene for reading problems and mild intellectual disability. J Med Genet 52, 111-122.
Hoischen, A., Krumm, N., and Eichler, E.E. (2014). Prioritization of neurodevelopmental disease genes by discovery of new mutations. Nature neuroscience 17, 764-772.
Horch, H.W., Kruttgen, A., Portbury, S.D., et al. (1999). Destabilization of cortical dendrites and spines by BDNF. Neuron 23, 353-364.
Hsueh, Y.P., Wang, T.F., Yang, F.C., et al. (2000). Nuclear translocation and transcription regulation by the membrane-associated guanylate kinase CASK/LIN-2. Nature 404, 298-302.
Huang, T.N., Chuang, H.C., Chou, W.H., et al. (2014). Tbr1 haploinsufficiency impairs amygdalar axonal projections and results in cognitive abnormality. Nature neuroscience 17, 240-247.
Huang, T.N., and Hsueh, Y.P. (2017). Calcium/calmodulin-dependent serine protein kinase (CASK), a protein implicated in mental retardation and autism-spectrum disorders, interacts with T-Brain-1 (TBR1) to control extinction of associative memory in male mice. J Psychiatry Neurosci 42, 37-47.
Hucthagowder, V., Shenoy, A., Corliss, M., et al. (2017). Utility of clinical high-depth next generation sequencing for somatic variant detection in the PIK3CA-related overgrowth spectrum. Clin Genet 91, 79-85.
Hunter, J.C., Manandhar, A., Carrasco, M.A., et al. (2015). Biochemical and Structural Analysis of Common Cancer-Associated KRAS Mutations. Mol Cancer Res 13, 1325-1335.
Hussain, K., Challis, B., Rocha, N., et al. (2011). An activating mutation of AKT2 and human hypoglycemia. Science 334, 474.
Hussing, C., Kampmann, M.L., Mogensen, H.S., et al. (2015). Comparison of techniques for quantification of next-generation sequencing libraries. Forensic Science International: Genetics Supplement Series 5, e276-e278.

Janku, F., Lee, J.J., Tsimberidou, A.M., et al. (2011). PIK3CA mutations frequently coexist with RAS and BRAF mutations in patients with advanced cancers. PloS one 6, e22769.

Jansen, L.A., Mirzaa, G.M., Ishak, G.E., et al. (2015). PI3K/AKT pathway mutations cause a spectrum of brain malformations from megalencephaly to focal cortical dysplasia. Brain 138, 1613-1628.

Jee, Y.H., Andrade, A.C., Baron, J., et al. (2017). Genetics of Short Stature. Endocrinol Metab Clin North Am 46, 259-281.

Jee, Y.H., and Baron, J. (2016). The Biology of Stature. J Pediatr 173, 32-38.

Jindal, G.A., Goyal, Y., Burdine, R.D., et al. (2015). RASopathies: unraveling mechanisms with animal models. Dis Model Mech 8, 1167.

Kamps, R., Brandao, R.D., Bosch, B.J., et al. (2017). Next-Generation Sequencing in Oncology: Genetic Diagnosis, Risk Prediction and Cancer Classification. Int J Mol Sci 18.

Kelleher, R.J., 3rd, Govindarajan, A., Jung, H.Y., et al. (2004). Translational control by MAPK signaling in long-term synaptic plasticity and memory. Cell 116, 467-479.

Keppler-Noreuil, K.M., Rios, J.J., Parker, V.E., et al. (2015). PIK3CA-related overgrowth spectrum (PROS): diagnostic and testing eligibility criteria, differential diagnosis, and evaluation. Am J Med Genet A 167A, 287-295.

Keppler-Noreuil, K.M., Sapp, J.C., Lindhurst, M.J., et al. (2014). Clinical delineation and natural history of the PIK3CA-related overgrowth spectrum. Am J Med Genet A 164A, 1713-1733.

Kharas, M.G., Okabe, R., Ganis, J.J., et al. (2010). Constitutively active AKT depletes hematopoietic stem cells and induces leukemia in mice. Blood 115, 1406-1415.

Kharbanda, M., Pilz, D.T., Tomkins, S., et al. (2017). Clinical features associated with CTNNB1 de novo loss of function mutations in ten individuals. Eur J Med Genet 60, 130-135.

Kim, M.J., Dunah, A.W., Wang, Y.T., et al. (2005). Differential roles of NR2A- and NR2B-containing NMDA receptors in Ras-ERK signaling and AMPA receptor trafficking. Neuron 46, 745-760.

Kinsler, V.A., Krengel, S., Riviere, J.B., et al. (2014). Next-generation sequencing of nevus spilus-type congenital melanocytic nevus: exquisite genotype-phenotype correlation in mosaic RASopathies. J Invest Dermatol 134, 2658-2660.

Kocak, O., Yarar, C., and Carman, K.B. (2016). Encephalocraniocutaneous lipomatosis, a rare neurocutaneous disorder: report of additional three cases. Childs Nerv Syst 32, 559-562.

Kotoula, V., Lyberopoulou, A., Papadopoulou, K., et al. (2015). Evaluation of two highly-multiplexed custom panels for massively parallel semiconductor sequencing on paraffin DNA. PloS one 10, e0128818.

Krapivinsky, G., Krapivinsky, L., Manasian, Y., et al. The NMDA Receptor Is Coupled to the ERK Pathway by a Direct Interaction between NR2B and RasGRF1. Neuron 40, 775-784.

Kratz, C.P., Rapisuwon, S., Reed, H., et al. (2011). Cancer in Noonan, Costello, cardiofaciocutaneous and LEOPARD syndromes. Am J Med Genet C Semin Med Genet 157C, 83-89.

Krencik, R., Hokanson, K.C., Narayan, A.R., et al. (2015). Dysregulation of astrocyte extracellular signaling in Costello syndrome. Sci Transl Med 7, 286ra266.

Krengel, S., Widmer, D.S., Kerl, K., et al. (2016). Naevus spilus-type congenital melanocytic naevus associated with a novel NRAS codon 61 mutation. Br J Dermatol 174, 642-644.

Kuechler, A., Willemsen, M.H., Albrecht, B., et al. (2015). De novo mutations in beta-catenin (CTNNB1) appear to be a frequent cause of intellectual disability: expanding the mutational and clinical spectrum. Hum Genet 134, 97-109.

Kuentz, P., St-Onge, J., Duffourd, Y., et al. (2017). Molecular diagnosis of PIK3CA-related overgrowth spectrum (PROS) in 162 patients and recommendations for genetic testing. Genet Med 19, 989-997.

Kurek, K.C., Luks, V.L., Ayturk, U.M., et al. (2012). Somatic mosaic activating mutations in PIK3CA cause CLOVES syndrome. Am J Hum Genet 90, 1108-1115.

Kushner, S.A., Elgersma, Y., Murphy, G.G., et al. (2005). Modulation of presynaptic plasticity and learning by the H-ras/extracellular signal-regulated kinase/synapsin I signaling pathway. The Journal of neuroscience : the official journal of the Society for Neuroscience 25, 9721-9734.

Kutmon, M., van Iersel, M.P., Bohler, A., et al. (2015). PathVisio 3: an extendable pathway analysis toolbox. PLoS Comput Biol 11, e1004085.

Kutsche, K., Yntema, H., Brandt, A., et al. (2000). Mutations in ARHGEF6, encoding a guanine nucleotide exchange factor for Rho GTPases, in patients with X-linked mental retardation. Nat Genet 26, 247-250.

Lachlan, K.L., Lucassen, A.M., Bunyan, D., et al. (2007). Cowden syndrome and Bannayan Riley Ruvalcaba syndrome represent one condition with variable expression and age-related penetrance: results of a clinical study of PTEN mutation carriers. J Med Genet 44, 579-585.

Lauing, K.L., Cortes, M., Domowicz, M.S., et al. (2014). Aggrecan is required for growth plate cytoarchitecture and differentiation. Dev Biol 396, 224-236.

Lee, J.H., Huynh, M., Silhavy, J.L., et al. (2012). De novo somatic mutations in components of the PI3K-AKT3-mTOR pathway cause hemimegalencephaly. Nat Genet 44, 941-945.

Lee, Y.S., Ehninger, D., Zhou, M., et al. (2014). Mechanism and treatment for learning and memory deficits in mouse models of Noonan syndrome. Nature neuroscience 17, 1736-1743.

Lees, M., Taylor, D., Atherton, D., et al. (2000). Oculo-ectodermal syndrome: report of two further cases. Am J Med Genet 91, 391-395.

Leinninger, G.M., Backus, C., Uhler, M.D., et al. (2004). Phosphatidylinositol 3-kinase and Akt effectors mediate insulin-like growth factor-I neuroprotection in dorsal root ganglia neurons. FASEB J 18, 1544-1546.

Lelieveld, S.H., Spielmann, M., Mundlos, S., et al. (2015). Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions. Hum Mutat 36, 815-822.

Lew, E.D., Furdui, C.M., Anderson, K.S., et al. (2009). The Precise Sequence of FGF Receptor Autophosphorylation Is Kinetically Driven and Is Disrupted by Oncogenic Mutations. Science Signaling 2, ra6-ra6.

Li, N., Xu, Y., Li, G., et al. (2017). Exome sequencing identifies a de novo mutation of CTNNB1 gene in a patient mainly presented with retinal detachment, lens and vitreous opacities, microcephaly, and developmental delay: Case report and literature review. Medicine (Baltimore) 96, e6914.

Lim, C.S., Kim, H., Yu, N.K., et al. (2017). Enhancing inhibitory synaptic function reverses spatial memory deficits in Shank2 mutant mice. Neuropharmacology 112, 104-112.

Lindhurst, M.J., Sapp, J.C., Teer, J.K., et al. (2011). A mosaic activating mutation in AKT1 associated with the Proteus syndrome. N Engl J Med 365, 611-619.

Lindhurst, M.J., Yourick, M.R., Yu, Y., et al. (2015). Repression of AKT signaling by ARQ 092 in cells and tissues from patients with Proteus syndrome. Sci Rep 5, 17162.

Livingstone, M., Folkman, L., Yang, Y., et al. (2017). Investigating DNA-, RNA-, and protein-based features as a means to discriminate pathogenic synonymous variants. Human Mutation 38, 1336-1347.

Loconte, D.C., Grossi, V., Bozzao, C., et al. (2015). Molecular and Functional Characterization of Three Different Postzygotic Mutations in PIK3CA-Related Overgrowth Spectrum (PROS) Patients: Effects on PI3K/AKT/mTOR Signaling and Sensitivity to PIK3 Inhibitors. PloS one 10, e0123092.

Loman, N.J., Misra, R.V., Dallman, T.J., et al. (2012). Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol 30, 434-439.

Lourenco, S.V., Fernandes, J.D., Hsieh, R., et al. (2014). Head and neck mucosal melanoma: a review. Am J Dermatopathol 36, 578-587.

Lu, Q., Aguilar, B.J., Li, M., et al. (2016). Genetic Alterations of δ-Catenin/NPRAP/Neurojungin (CTNND2): Functional Implications in Complex Human Diseases. Human Genetics, 135(10), 1107–1116

Lu Z, Reddy MVVVS, Liu J, et al. (2016). Molecular Architecture of Contactin-associated Protein-like 2 (CNTNAP2) and Its Interaction with Contactin 2 (CNTN2). The Journal of Biological Chemistry. 291(46):24133-24147.

Luo, S., and Tsao, H. (2014). Epidermal, sebaceous, and melanocytic nevoid proliferations are spectrums of mosaic RASopathies. J Invest Dermatol 134, 2493-2496.

MacArthur, D.G., Manolio, T.A., Dimmock, D.P., et al. (2014). Guidelines for investigating causality of sequence variants in human disease. Nature 508, 469-476.

Mainberger, F., Langer, S., Mall, V., et al. (2016). Impaired synaptic plasticity in RASopathies: a mini-review. J Neural Transm (Vienna) 123, 1133-1138.

Manabe, T., Aiba, A., Yamada, A., et al. (2000). Regulation of long-term potentiation by H-Ras through NMDA receptor phosphorylation. The Journal of neuroscience : the official journal of the Society for Neuroscience 20, 2504-2511.

Mandelker, D., Gabelli, S.B., Schmidt-Kittler, O., et al. (2009). A frequent kinase domain mutation that changes the interaction between PI3Kα and the membrane. Proceedings of the National Academy of Sciences 106, 16996-17001.

Marsh, S. (2007). Pyrosequencing applications. Methods Mol Biol 373, 15-24.

Martinez, F., Caro-Llopis, A., Rosello, M., et al. (2017). High diagnostic yield of syndromic intellectual disability by targeted next-generation sequencing. J Med Genet 54, 87-92.
Martinez-Lopez, A., Blasco-Morente, G., Perez-Lopez, I., et al. (2017). CLOVES syndrome: review of a PIK3CA-related overgrowth spectrum (PROS). Clin Genet 91, 14-21.
McCarroll, S.A., and Hyman, S.E. (2013). Progress in the genetics of polygenic brain disorders: significant new challenges for neurobiology. Neuron 80, 578-587.
Mele, M., Ferreira, P.G., Reverter, F., et al. (2015). Human genomics. The human transcriptome across tissues and individuals. Science 348, 660-665.
Mertz, J., Tan, H., Pagala, V., et al. (2015). Sequential Elution Interactome Analysis of the Mind Bomb 1 Ubiquitin Ligase Reveals a Novel Role in Dendritic Spine Outgrowth. Mol Cell Proteomics 14, 1898-1910.
Meyer, D.S., Koren, S., Leroy, C., et al. (2013). Expression of PIK3CA mutant E545K in the mammary gland induces heterogeneous tumors but is less potent than mutant H1047R. Oncogenesis 2, e74.
Miceli, F., Striano, P., Soldovieri, M.V., et al. (2015). A novel KCNQ3 mutation in familial epilepsy with focal seizures and intellectual disability. Epilepsia 56, e15-20.
Miller, K.A., Twigg, S.R., McGowan, S.J., et al. (2017). Diagnostic value of exome and whole genome sequencing in craniosynostosis. J Med Genet 54, 260-268.
Minelli, A., Scassellati, C., Cloninger, C.R., et al. (2012). PCLO gene: its role in vulnerability to major depressive disorder. J Affect Disord 139, 250-255.
Minichiello, L., Calella, A.M., Medina, D.L., et al. (2002). Mechanism of TrkB-mediated hippocampal long-term potentiation. Neuron 36, 121-137.
Mirzaa, G., Timms, A.E., Conti, V., et al. (2016). PIK3CA-associated developmental disorders exhibit distinct classes of mutations with variable expression and tissue distribution. JCI Insight 1.
Mirzaa, G.M., Conway, R.L., Gripp, K.W., et al. (2012). Megalencephaly-capillary malformation (MCAP) and megalencephaly-polydactyly-polymicrogyria-hydrocephalus (MPPH) syndromes: two closely related disorders of brain overgrowth and abnormal brain and body morphogenesis. Am J Med Genet A 158A, 269-291.
Mirzaa, G.M., Riviere, J.B., and Dobyns, W.B. (2013). Megalencephaly syndromes and activating mutations in the PI3K-AKT pathway: MPPH and MCAP. Am J Med Genet C Semin Med Genet 163C, 122-130.
Mitra, A., Skrzypczak, M., Ginalski, K., et al. (2015). Strategies for Achieving High Sequencing Accuracy for Low Diversity Samples and Avoiding Sample Bleeding Using Illumina Platform. PloS one 10, e0120520.
Mohebiany, A.N., Harroch, S., and Bouyain, S. (2014). New insights into the roles of the contactin cell adhesion molecules in neural development. Adv Neurobiol 8, 165-194.
Mokry, M., Feitsma, H., Nijman, I.J., et al. (2010). Accurate SNP and mutation detection by targeted custom microarray-based genomic enrichment of short-fragment sequencing libraries. Nucleic Acids Res 38, e116.
Moog, U. (2009). Encephalocraniocutaneous lipomatosis. J Med Genet 46, 721-729.
Moriya, M., Inoue, S., Miyagawa-Tomita, S., et al. (2015). Adult mice expressing a Braf Q241R mutation on an ICR/CD-1 background exhibit a cardio-facio-cutaneous syndrome phenotype. Hum Mol Genet 24, 7349-7360.
Musante, L., and Ropers, H.H. (2014). Genetics of recessive cognitive disorders. Trends Genet 30, 32-39.
Nakamura, K., Kato, M., Tohyama, J., et al. (2014). AKT3 and PIK3R2 mutations in two patients with megalencephaly-related syndromes: MCAP and MPPH. Clin Genet 85, 396-398.
Neale, B.M., Kou, Y., Liu, L., et al. (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature 485, 242-245.
Neumann, T.E., Allanson, J., Kavamura, I., et al. (2009). Multiple giant cell lesions in patients with Noonan syndrome and cardio-facio-cutaneous syndrome. Eur J Hum Genet 17, 420-425.
Nicholas, D.B., Attridge, M., Zwaigenbaum, L., et al. (2015). Vocational support approaches in autism spectrum disorder: a synthesis review of the literature. Autism 19, 235-245.
Nicolas, C.S., Peineau, S., Amici, M., et al. (2012). The Jak/STAT pathway is involved in synaptic plasticity. Neuron 73, 374-390.
Ninan, I. (2014). Synaptic regulation of affective behaviors; role of BDNF. Neuropharmacology 76 Pt C, 684-695.
Omrani, A., van der Vaart, T., Mientjes, E., et al. (2015). HCN channels are a novel therapeutic target for cognitive dysfunction in Neurofibromatosis type 1. Mol Psychiatry 20, 1311-1321.
O'Roak, B.J., Vives, L., Fu, W., et al. (2012). Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. Science 338, 1619-1622.

Ozkan, E.D., Creson, T.K., Kramar, E.A., et al. (2014). Reduced cognition in Syngap1 mutants is caused by isolated damage within developing forebrain excitatory neurons. Neuron 82, 1317-1333.

Pagani, M.R., Oishi, K., Gelb, B.D., et al. (2009). The phosphatase SHP2 regulates the spacing effect for long-term memory induction. Cell 139, 186-198.

Palumbo, O., Fichera, M., Palumbo, P., et al. (2014). TBR1 is the candidate gene for intellectual disability in patients with a 2q24.2 interstitial deletion. Am J Med Genet A 164A, 828-833.

Park, H., and Poo, M.M. (2013). Neurotrophin regulation of neural circuit development and function. Nature reviews Neuroscience 14, 7-23.

Peacock, J.D., Dykema, K.J., Toriello, H.V., et al. (2015). Oculoectodermal syndrome is a mosaic RASopathy associated with KRAS alterations. American Journal of Medical Genetics Part A 167, 1429-1435.

Philips, A.K., Sirén, A., Avela, K., et al. (2014). X-exome sequencing in Finnish families with Intellectual Disability - four novel mutations and two novel syndromic phenotypes. Orphanet Journal of Rare Diseases 9, 49.

Philips, G.T., Ye, X., Kopec, A.M., et al. (2013). MAPK establishes a molecular context that defines effective training patterns for long-term memory formation. The Journal of neuroscience : the official journal of the Society for Neuroscience 33, 7565-7573.

Pierpont, E.I., Tworog-Dube, E., and Roberts, A.E. (2013). Learning and memory in children with Noonan syndrome. Am J Med Genet A 161A, 2250-2257.

Piper, M., Moldrich, R.X., Lindwall, C., et al. (2009). Multiple non-cell-autonomous defects underlie neocortical callosal dysgenesis in Nfib-deficient mice. Neural Dev 4, 43.

Poduri, A., Evrony, G.D., Cai, X., et al. (2012). Somatic activation of AKT3 causes hemispheric developmental brain malformations. Neuron 74, 41-48.

Poduri, A., Evrony, G.D., Cai, X., et al. (2013). Somatic mutation, genomic variation, and neurological disease. Science 341, 1237758.

Polosukhina, D., Love, H.D., Correa, H., et al. (2017). Functional KRAS mutations and a potential role for PI3K/AKT activation in Wilms tumors. Mol Oncol 11, 405-421.

Prior, I.A., Lewis, P.D., and Mattos, C. (2012). A comprehensive survey of Ras mutations in cancer. Cancer Res 72, 2457-2467.

Pucilowska, J., Puzerey, P.A., Karlo, J.C., et al. (2012). Disrupted ERK signaling during cortical development leads to abnormal progenitor proliferation, neuronal and network excitability and behavior, modeling human neuro-cardio-facial-cutaneous and related syndromes. The Journal of neuroscience : the official journal of the Society for Neuroscience 32, 8663-8677.

Puzzo, D., Bizzoca, A., Privitera, L., et al. (2013). F3/Contactin promotes hippocampal neurogenesis, synaptic plasticity, and memory in adult mice. Hippocampus 23, 1367-1382.

Rabbani, B., Tekin, M., and Mahdieh, N. (2014). The promise of whole-exome sequencing in medical genetics. J Hum Genet 59, 5-15.

Rauch, A., Wieczorek, D., Graf, E., et al. (2012). Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. Lancet 380, 1674-1682.

Rauen, K.A. (2013). The RASopathies. Annu Rev Genomics Hum Genet 14, 355-369.

Raymond, S., Nicot, F., Jeanne, N., et al. (2017). Performance comparison of next-generation sequencing platforms for determining HIV-1 coreceptor use. Sci Rep 7, 42215.

Redin, C., Gérard, B., Lauer, J., et al. (2014). Efficient strategy for the molecular diagnosis of intellectual disability using targeted high-throughput sequencing. Journal of Medical Genetics 51, 724-736.

Rehm, H.L. (2013). Disease-targeted sequencing: a cornerstone in the clinic. Nat Rev Genet 14, 295-300.

Rehm, H.L., Bale, S.J., Bayrak-Toydemir, P., et al. (2013). ACMG clinical laboratory standards for next-generation sequencing. Genet Med 15, 733-747.

Richards, S., Aziz, N., Bale, S., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 17, 405-424.

Rieber, N., Zapatka, M., Lasitschka, B., et al. (2013). Coverage bias and sensitivity of variant calling for four whole-genome sequencing technologies. PloS one 8, e66621.

Rios, J.J., Paria, N., Burns, D.K., et al. (2013). Somatic gain-of-function mutations in PIK3CA in patients with macrodactyly. Hum Mol Genet 22, 444-451.

Rivas, M.A., Pirinen, M., Conrad, D.F., et al. (2015). Human genomics. Effect of predicted protein-truncating genetic variants on the human transcriptome. Science 348, 666-669.

Riviere, J.B., Mirzaa, G.M., O'Roak, B.J., et al. (2012). De novo germline and postzygotic mutations in AKT3, PIK3R2 and PIK3CA cause a spectrum of related megalencephaly syndromes. Nat Genet 44, 934-940.
Rodriguez-Viciana, P., and Rauen, K.A. (2008). Biochemical characterization of novel germline BRAF and MEK mutations in cardio-facio-cutaneous syndrome. Methods Enzymol 438, 277-289.
Rodriguez-Viciana, P., Tetsu, O., Tidyman, W.E., et al. (2006). Germline Mutations in Genes Within the MAPK Pathway Cause Cardio-facio-cutaneous Syndrome. Science 311, 1287-1290.
Rohlin, A., Wernersson, J., Engwall, Y., et al. (2009). Parallel sequencing used in detection of mosaic mutations: Comparison with four diagnostic DNA screening techniques. Human Mutation 30, 1012-1020.
Rojas, A.M., Fuentes, G., Rausell, A., et al. (2012). The Ras protein superfamily: evolutionary tree and role of conserved amino acids. The Journal of cell biology 196, 189-201.
Rolando, C., Erni, A., Grison, A., et al. Multipotency of Adult Hippocampal NSCs In Vivo Is Restricted by Drosha/NFIB. Cell Stem Cell 19, 653-662.
Ronemus, M., Iossifov, I., Levy, D., et al. (2014). The role of de novo mutations in the genetics of autism spectrum disorders. Nat Rev Genet 15, 133-141.
Rosen, L.B., Ginty, D.D., Weber, M.J., et al. (1994). Membrane depolarization and calcium influx stimulate MEK and MAP kinase via activation of Ras. Neuron 12, 1207-1221.
Roy, A., Skibo, J., Kalume, F., et al. (2015). Mouse models of human PIK3CA-related brain overgrowth have acutely treatable epilepsy. Elife 4.
Rumbaugh, G., Adams, J.P., Kim, J.H., et al. (2006). SynGAP regulates synaptic strength and mitogen-activated protein kinases in cultured neurons. Proceedings of the National Academy of Sciences of the United States of America 103, 4344-4351.
Ryu, H.H., and Lee, Y.S. (2016). Cell type-specific roles of RAS-MAPK signaling in learning and memory: Implications in neurodevelopmental disorders. Neurobiology of learning and memory 135, 13-21.
Salgado, C.M., Basu, D., Nikiforova, M., et al. (2015). BRAF mutations are also associated with neurocutaneous melanocytosis and large/giant congenital melanocytic nevi. Pediatr Dev Pathol 18, 1-9.
San Martin, A., and Pagani, M.R. (2014). Understanding intellectual disability through RASopathies. J Physiol Paris 108, 232-239.
Sapp, J.C., Turner, J.T., van de Kamp, J.M., et al. (2007). Newly delineated syndrome of congenital lipomatous overgrowth, vascular malformations, and epidermal nevi (CLOVE syndrome) in seven patients. Am J Med Genet A 143A, 2944-2958.
Sarkozy, A., Carta, C., Moretti, S., et al. (2009). Germline BRAF mutations in Noonan, LEOPARD, and cardiofaciocutaneous syndromes: molecular diversity and associated phenotypic spectrum. Hum Mutat 30, 695-702.
Satoh, Y., Endo, S., Ikeda, T., et al. (2007). Extracellular signal-regulated kinase 2 (ERK2) knockdown mice show deficits in long-term memory; ERK2 has a specific function in learning and memory. The Journal of neuroscience : the official journal of the Society for Neuroscience 27, 10765-10776.
Schmeisser, M.J., Ey, E., Wegener, S., et al. (2012). Autistic-like behaviours and hyperactivity in mice lacking ProSAP1/Shank2. Nature 486, 256-260.
Schuster, S.C. (2008). Next-generation sequencing transforms today's biology. Nat Methods 5, 16-18.
Shi, L. (2013). Dock protein family in brain development and neurological disease. Commun Integr Biol 6, e26839.
Shilyansky, C., Lee, Y.S., and Silva, A.J. (2010). Molecular and cellular mechanisms of learning disabilities: a focus on NF1. Annual review of neuroscience 33, 221-243.
Shin, S., and Park, J. (2016). Characterization of sequence-specific errors in various next-generation sequencing systems. Molecular BioSystems 12, 914-922.
Shinawi, M., Sahoo, T., Maranda, B., et al. (2011). 11p14.1 microdeletions associated with ADHD, autism, developmental delay, and obesity. Am J Med Genet A 155A, 1272-1280.
Simbolo, M., Gottardi, M., Corbo, V., et al. (2013). DNA qualification workflow for next generation sequencing of histopathological samples. PloS one 8, e62692.
Singh, R.R., Luthra, R., Routbort, M.J., et al. (2016). Implementation of next generation sequencing in clinical molecular diagnostic laboratories: advantages, challenges and potential. Expert Review of Precision Medicine and Drug Development 1, 109-120.
Sloan, S.A., and Barres, B.A. (2014). Mechanisms of astrocyte development and their contributions to neurodevelopmental disorders. Curr Opin Neurobiol 27, 75-81.
Smith, G., Bounds, R., Wolf, H., et al. (2010). Activating K-Ras mutations outwith 'hotspot' codons in sporadic colorectal tumours - implications for personalised cancer medicine. Br J Cancer 102, 693-703.

Sol-Church, K., Stabley, D.L., Demmer, L.A., et al. (2009). Male-to-male transmission of Costello syndrome: G12S HRAS germline mutation inherited from a father with somatic mosaicism. Am J Med Genet A 149A, 315-321.

Soldovieri, M.V., Boutry-Kryza, N., Milh, M., et al. (2014). Novel KCNQ2 and KCNQ3 mutations in a large cohort of families with benign neonatal epilepsy: first evidence for an altered channel regulation by syntaxin-1A. Hum Mutat 35, 356-367.

Soon, W.W., Hariharan, M., and Snyder, M.P. (2013). High-throughput sequencing for biology and medicine. Mol Syst Biol 9, 640.

Sperow, M., Berry, R.B., Bayazitov, I.T., et al. (2012). Phosphatase and tensin homologue (PTEN) regulates synaptic plasticity independently of its effect on neuronal morphology and migration. The Journal of physiology 590, 777-792.

Stamberger, H., Nikanorova, M., Willemsen, M.H., et al. (2016). STXBP1 encephalopathy: A neurodevelopmental disorder including epilepsy. Neurology 86, 954-962.

Stornetta, R.L., and Zhu, J.J. (2011). Ras and Rap signaling in synaptic plasticity and mental disorders. Neuroscientist 17, 54-78.

Sturgeon, M., Davis, D., Albers, A., et al. (2016). The Notch ligand E3 ligase, Mind Bomb1, regulates glutamate receptor localization in Drosophila. Molecular and cellular neurosciences 70, 11-21.

Sun, B.K., Saggini, A., Sarin, K.Y., et al. (2013). Mosaic activating RAS mutations in nevus sebaceus and nevus sebaceus syndrome. J Invest Dermatol 133, 824-827.

Sun, Y., Ruivenkamp, C.A., Hoffer, M.J., et al. (2015). Next-generation diagnostics: gene panel, exome, or whole genome? Hum Mutat 36, 648-655.

Sundaram, M.V. (2006). RTK/Ras/MAPK signaling. WormBook, 1-19.

Suri, M., Evers, J.M.G., Laskowski, R.A., et al. (2017). Protein structure and phenotypic analysis of pathogenic and population missense variants in STXBP1. Mol Genet Genomic Med 5, 495-507.

Suzuki, Y., Enokido, Y., Yamada, K., et al. (2017). The effect of rapamycin, NVP-BEZ235, aspirin, and metformin on PI3K/AKT/mTOR signaling pathway of PIK3CA-related overgrowth spectrum (PROS). Oncotarget 8, 45470-45483.

Sweatt, J.D. (2001). The neuronal MAP kinase cascade: a biochemical signal integration system subserving synaptic plasticity and memory. J Neurochem 76, 1-10.

Tapodi, A., Debreceni, B., Hanto, K., et al. (2005). Pivotal role of Akt activation in mitochondrial protection and cell survival by poly(ADP-ribose)polymerase-1 inhibition in oxidative stress. The Journal of biological chemistry 280, 35767-35775.

Tartaglia, M., Gelb, B.D., and Zenker, M. (2011). Noonan syndrome and clinically related disorders. Best Pract Res Clin Endocrinol Metab 25, 161-179.

Thomas, G.M., and Huganir, R.L. (2004). MAPK cascade signalling and synaptic plasticity. Nature reviews Neuroscience 5, 173-183.

tom Dieck, S., Sanmarti-Vila, L., Langnaese, K., et al. (1998). Bassoon, a novel zinc-finger CAG/glutamine-repeat protein selectively localized at the active zone of presynaptic nerve terminals. The Journal of cell biology 142, 499-509.

Toonen, R.F., Wierda, K., Sons, M.S., et al. (2006). Munc18-1 expression levels control synapse recovery by regulating readily releasable pool size. Proceedings of the National Academy of Sciences of the United States of America 103, 18332-18337.

Torres, V.I., Vallejo, D., and Inestrosa, N.C. (2017). Emerging Synaptic Molecules as Candidates in the Etiology of Neurological Disorders. Neural Plasticity 2017, 8081758.

Treutlein, B., Brownfield, D.G., Wu, A.R., et al. (2014). Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature 509, 371-375.

Tucci, V., Kleefstra, T., Hardy, A., et al. (2014). Dominant beta-catenin mutations cause intellectual disability with recognizable syndromic features. J Clin Invest 124, 1468-1482.

Turner, T.N., Sharma, K., Oh, E.C., et al. (2015). Loss of delta-catenin function in severe autism. Nature 520, 51-56.

Vahidnezhad, H., Youssefian, L., and Uitto, J. (2016). Klippel-Trenaunay syndrome belongs to the PIK3CA-related overgrowth spectrum (PROS). Exp Dermatol 25, 17-19.

Van Aelst, L., and D'Souza-Schorey, C. (1997). Rho GTPases and signaling networks. Genes Dev 11, 2295-2322.

van Dijk, E.L., Jaszczyszyn, Y., and Thermes, C. (2014). Library preparation methods for next-generation sequencing: tone down the bias. Exp Cell Res 322, 12-20.

Verhage, M., Maia, A.S., Plomp, J.J., et al. (2000). Synaptic assembly of the brain in the absence of neurotransmitter secretion. Science 287, 864-869.
Vilar, M., and Mira, H. (2016). Regulation of Neurogenesis by Neurotrophins during Adulthood: Expected and Unexpected Roles. Front Neurosci 10, 26.
Vinci, G., Chantot-Bastaraud, S., El Houate, B., et al. (2007). Association of deletion 9p, 46,XY gonadal dysgenesis and autistic spectrum disorder. Mol Hum Reprod 13, 685-689.
Vinet, E., Pineau, C.A., Clarke, A.E., et al. (2015). Increased Risk of Autism Spectrum Disorders in Children Born to Women With Systemic Lupus Erythematosus: Results From a Large Population-Based Cohort. Arthritis Rheumatol 67, 3201-3208.
Vissers, L.E.L.M., Gilissen, C., and Veltman, J.A. (2016). Genetic studies in intellectual disability and related disorders. Nat Rev Genet 17, 9-18.
Wang, L., Zhou, K., Fu, Z., et al. (2017). Brain Development and Akt Signaling: the Crossroads of Signaling Pathway and Neurodevelopmental Diseases. J Mol Neurosci 61, 379-384.
Watson, L.M., Bamber, E., Schnekenberg, R.P., et al. (2017). Dominant Mutations in GRM1 Cause Spinocerebellar Ataxia Type 44. Am J Hum Genet 101, 451-458.
Wennerberg, K., Rossman, K.L., and Der, C.J. (2005). The Ras superfamily at a glance. J Cell Sci 118, 843-846.
Winnepenninckx, B., Rooms, L., and Kooy, R.F. (2003). Mental retardation: a review of the genetic causes. The British Journal of Development Disabilities 49, 29-44.
Woolfrey, K.M., Srivastava, D.P., Photowala, H., et al. (2009). Epac2 induces synapse remodeling and depression and its disease-associated forms alter spines. Nature neuroscience 12, 1275-1284.
Wu, Y., Arai, A.C., Rumbaugh, G., et al. (2007). Mutations in ionotropic AMPA receptor 3 alter channel properties and are associated with moderate cognitive impairment in humans. Proceedings of the National Academy of Sciences of the United States of America 104, 18163-18168.
Xu, B., Gottschalk, W., Chow, A., et al. (2000). The role of brain-derived neurotrophic factor receptors in the mature hippocampus: modulation of long-term potentiation through a presynaptic mechanism involving TrkB. The Journal of neuroscience : the official journal of the Society for Neuroscience 20, 6888-6897.
Xue, Y., Chen, Y., Ayub, Q., et al. (2012). Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. Am J Hum Genet 91, 1022-1032.
Ye, X., and Carew, T.J. (2010). Small G protein signaling in neuronal plasticity and memory formation: the specific role of ras family proteins. Neuron 68, 340-361.
Yeo, G.S., Connie Hung, C.C., Rochford, J., et al. (2004). A de novo mutation affecting human TrkB associated with severe obesity and developmental delay. Nature neuroscience 7, 1187-1189.
Yi, Y., Polosukhina, D., Love, H.D., et al. A Murine Model of K-RAS and β-Catenin Induced Renal Tumors Expresses High Levels of E2F1 and Resembles Human Wilms Tumor. The Journal of Urology 194, 1762-1770.
Yohe, S., and Thyagarajan, B. (2017). Review of Clinical Next-Generation Sequencing. Archives of Pathology & Laboratory Medicine 141, 1544-1557.
Yoon, G., Rosenberg, J., Blaser, S., et al. (2007). Neurological complications of cardio-facio-cutaneous syndrome. Dev Med Child Neurol 49, 894-899.
Yoon, K.-J., Lee, H.-R., Jo, Y.S., et al. (2012). Mind bomb-1 is an essential modulator of long-term memory and synaptic plasticity via the Notch signaling pathway. Molecular Brain 5, 40-40.
Yuan, H., Low, C.M., Moody, O.A., et al. (2015). Ionotropic GABA and Glutamate Receptor Mutations and Human Neurologic Diseases. Mol Pharmacol 88, 203-217.
Zenker, M. (2011). Clinical manifestations of mutations in RAS and related intracellular signal transduction factors. Curr Opin Pediatr 23, 443-451.
Zhao, L., and Vogt, P.K. (2008). Helical domain and kinase domain mutations in p110alpha of phosphatidylinositol 3-kinase induce gain of function by different mechanisms. Proceedings of the National Academy of Sciences of the United States of America 105, 2652-2657.
Zhu, J.J., Qin, Y., Zhao, M., et al. (2002). Ras and Rap control AMPA receptor trafficking during synaptic plasticity. Cell 110, 443-455.
Zimerman, M., Wessel, M.J., Timmermann, J.E., et al. (2015). Impairment of Procedural Learning and Motor Intracortical Inhibition in Neurofibromatosis Type 1 Patients. EBioMedicine 2, 1430-1437.

List of Abbreviations

AA	amino acid
ABD	adaptor binding domain
ACMG	american college of medical genetics and genomics
AD	autosomal dominant
ADHD	attention-deficit hyperactivity disorder
ADP	adenosine diphosphate
AKT	protein kinase B (PKB), also known as Akt
AMPAR	α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid receptor
AR	autosomal recessive
ASD	autism spectrum disorder
ATP	adenosine triphosphate
BDNF	brain-derived neurotrophic factor
BRAF	v-Raf murine sarcoma viral oncogene homolog B
cDNA	complementary DNA
CDS	coding sequence
CFC	cardiofaciocutaneous syndrome
CHR	chromosome
CLOVES	congenital lipomatous overgrowth, vascular malformations, epidermal nevi, and skeletal abnormalities
CNS	central nervous system
CNV	copy number variation
CONTIG	contiguous
COS cells	cells being CV-1 (simian) in Origin, and carrying the SV40 genetic material
COSMIC	catalogue of somatic mutations in cancer
CTNNB1	catenin beta-1
dbSNP	single Nucleotide Polymorphism database
Del	Deletion
DIN	DNA Integrity Number
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
dNTP	deoxynucleotide
ds DNA	double stranded DNA
ECCL	encephalocraniocutaneous lipomatosis
ERK	extracellular signal–regulated kinase
et al.	et alii/aliae/alia (and others)
Ex	example
ExAC	exome aggregation consortium
FFPE	formalin-fixed, paraffin-embedded
G protein	GTPase; guanine nucleotide binding protein
GABA	gamma-Aminobutyric acid
GAP	GTPase-activating protein
GDI	guanosine nucleotide dissociation inhibitor
gDNA	genomic deoxyribonucleic acid
GDP	guanosine diphosphate
GEF	guanine nucleotide exchange factor
GH	growth hormone
GLUR1	glutamate receptor 1
GSJ	genome sequencer junior
GTP	guanosine triphosphate
GTPase	guanine nucleotide binding protein
GWAS	genome-wide association study
HCl	hydrochloric acid
HGMD	human gene mutation database
HRAS	Harvey Rat Sarcoma Viral Oncogene
ID	intellectual disability
IEM	Illumina experiment manager
IGV	integrative genomics viewer
INDELS	insertion or deletion of bases
IQ	intelligence quotient
IUGR	intrauterine growth restriction
JAK	janus kinase
KRAS	V-Ki-Ras2 Kirsten rat sarcoma viral oncogene homolog gene
LOF’s	loss of function
LP	likely pathogenic
LS	leopard syndrome
LTD	long-term depression
LTM	long-term memory
LTP	long-term potentiation
MAF	minor allele frequency
MAGUK	membrane-associated guanylate kinases
MAPK	mitogen-activated protein kinases

MCS	miseq controller software
MEK/ MAPKK	mitogen-activated protein kinase kinase
MR/MRT	mental retardation
ID cohort	magdeburg cohort
MRI	magnetic resonance imaging
mRNA	messenger RNA
mTOR	mammalian target of rapamycin
Munc-18	mammalian uncoordinated-18
NaOH	sodium hydroxide
NCFCS	neuro-cardio-facio-cutaneous syndromes
NF1	neurofibromin 1
NGS	next generation sequencing
NHLBI	national heart, lung, and blood institute
NMDA	N-Methyl-D-aspartate
NON-SYN	non-synonymous
NR2B/ GRIN2B	glutamate [NMDA] receptor subunit epsilon-2
NRCE	nextera rapid capture enrichment
NS	noonan syndrome
NT	not tested
OES	oculoectodermal syndrome
OFC	orbitofrontal cortex
OMIM	online mendelian inheritance in men
P	pathogenic
PCR	polymerase chain reaction
PDK1	pyruvate dehydrogenase kinase isozyme 1
PHIX	phi X 174 (or ΦX174) bacteriophage
PI3K	phosphoinositide 3-kinase
PIK3CA	phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha
PIP2	phosphatidylinositol-4,5-bisphosphate
PIP3	phosphatidylinositol (3,4,5)-trisphosphate
PSD	post synaptic density
PTEN	phosphatase and tensin homolog
PTV	protein truncating variants
qPCR	quantitative polymerase chain reaction
RAF	rapidly accelerated fibrosarcoma
RALGDS	Ral guanine nucleotide dissociation stimulator
RAS	rat sarcoma
RBD	ras binding domain
RIT1	Ras Like Without CAAX 1
RNA	ribonucleic acid
RTA	real time analysis
SAV	sequencing analysis viewer
SD	standard deviation
SHP2	Src homology region 2 (SH2)-containing protein tyrosine phosphatase 2
SIFT	sorting intolerant from tolerant
SNP	single-nucleotide polymorphism
SNV	single-nucleotide variant
SPRI	solid phase reversible immobilization
STAT	signal transducer and activator of transcription
STRING	search tool for recurring instances of neighbouring genes (database)
SYN	synonymous
SYNGAP1	synaptic Ras GTPase-activating protein 1
Tris-HCl	Tris (hydroxymethyl) aminomethane hydrochloride
US	uncertain significance
UTR	untranslated region
VCF	variant call format
WES	whole exome sequencing
WGS	whole genome sequencing
XD /XR	X-linked dominant/recessive

UNITS

°C	degree Celsius
∞	infinity
bp	base pair(s)
cm	centimetre(s)
cm²	square centimetre(s)
g	grams
kb	kilo base pair(s)
kDa	kilodalton(s)
kg	kilograms(s)
M	molar
mA	milliampere(s)
Mb	mega base pair(s)
mg	milligram(s)
min	minute(s)
ml	millilitre(s)
mM	millimolar
mm^2	square millimetre(s)
µg	microgram(s)
µl	microlitre(s)
µM	micrometre(s)
nm	nanometre(s)
ng	nanogram(s)
nt	nucleotide(s)
pH	negative of the logarithm to base 10 of the concentration
pM	picomole(s)
sec	second(s)
U	unit(s)
V	volt(s)
µ	micro
%	percentange
~	approximately equal to
Y/ Yrs	years
Mo	months
FU	fluorescence units
®	registered trade mark
v	version

AMINO ACIDS

A	Ala	Alanine
C	Cys	Cysteine
D	Asp	Aspartic acid
E	Glu	Glutamic acid
F	Phe	Phenylalanine
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
K	Lys	Lysine
L	Leu	Leucine
M	Met	Methionine
N	Asn	Asparagine
P	Pro	Proline
Q	Gln	Glutamine
R	Arg	Arginine
S	Ser	Serine
T	Thr	Threonine
V	Val	Valine
W	Trp	Tryptophan
Y	Tyr	Tyrosine
*	Ter	Termination (Stop)

NUCLEOBASES

A	Adenine/Adenosine
C	Cytosine/Cytidine
G	Guanine/Guanosine
T	Thymine/5-Methyluridine

Supplementary material

Supplementary Table 1: Patients and samples in PIK3CA-related overgrowth syndrome

Patients and samples involved in the CLOVES project. All the samples are included in NGS Run1 except for P4. The patients and samples highlighted in grey are included in RunII.

PIK3CA Exons	Patient ID	Sample number	Material	DNA conc (ng/µl)
Exon 2	P1	1	Fat tissue	10
		2	Skin	28
		3	Subcutaneous tissue	21
		4	Fat tissue	22
	P2	5	Fat tissue	7
		6	Blood	586
	P3	7	Fat tissue	48
		8	Blood	25
	P4	9	Blood	240
Exon 8	P5	10	Fat tissue	230
		11	Bone	237
	P6	12	Skin abrasia	41
		13	FFPE	56
Exon 5	P7	14	Bone	86
Exon 10	P8	15	Fibroblasts	120
	P9	16	Skin	8
		17	FFPE	88,3
	P10	18	Fat tissue	304
		19	Blood	399
	P11	20	Cartilage	119
	P12	21	FFPE	15,7
		22	Lipoma	64
		23	Blood	72
Exon 21	P13	24	Blood	182
		25	Cartilage	56
		26	Tendon	10
		27	Skin	68

PIK3CA Exons	Patient ID	Sample number	Material	DNA conc (ng/µl)
Exon 21	P13	28	Connective tissue	22
		29	Epiphyses tissue	38
	P 14	30	Bone	122
		31	Blood	63,2
		32	Fat	69
	P15	33	Fat	305
		34	Blood	34,8
	P16	35	Blood	240
		36	Skin	1390
		37	Adipose neck tissue	70
		38	Adipose cervical area	160
		39	Connective/adipose neck	330
		40	Keratinocytes	2980
		41	Nevüs	2400
	P17	42	Sub cutaneous tissue	101
		43	Deep Fat	118
		44	Skin	105
	P18	45	FFPE- Tumor	29,3
		46	FFPE- bone caspel	51,3
		47	Skin	13
		48	Blood	327

Supplementary Table 2: The complete list of targeted genes selected for this project. A total of 329 genes selected with 221 genes belonging to RAS related pathway and 108 genes belonging to GH pathway.

Short stature genes (108)			**RAS genes (221)**					
ACAN	GRB2	PIK3R2	ACTN2	DLG1	GRM7	NRG1	RAPGEF2	SPRY1
ADAMTS17	HESX1	PIK3R3	ADAM22	DLG3	GRM8	NRG2	RAPGEF3	SPRY2
ADAMTSL3	HHIP	POMC	AKAP5	DLG4	HOMER1	NRG3	RAPGEF4	SPRY3
ADIPOQ	HMGA1	POU1F1	ARAF	DLGAP1	HOMER2	NRN1	RASA1	SPRY4
AKT1	HMGA2	PPP1CC	ARFGEF2	DNMT1	HRAS	NRXN1	RASA3	SRC
AKT2	IGF1	PRKG2	ARHGEF6	DOCK8	IL1RAPL1	NRXN2	RASAL1	STK38L
ANAPC13	IGF1R	PROP1	ARHGEF7	DYNC1H1	ITPR1	NSMF	RASGRF1	STRN
AR	IGF2	PTCH1	BAIAP2	EEF1A2	KALRN	NTRK2	RASGRF2	STRN4
ARRB1	IGFALS	PTPN1	BDNF	EPB41L1	KCNA4	NUMB	RASGRP1	STX1A
BMP2	IGFBP1	RASA2	BRAF	ERBB4	KCNQ2	OPHN1	RASGRP2	STX4
BMP4	IGFBP3	RBP3	BSN	FGD1	KCNQ3	PACSIN1	RGL1	STXBP1
BMP6	IGFBP3_ promoter	RNF135	CABP1	FYN	KRAS	PAK3	RGL2	SYN1
CBL	IGFBP4	RUNX2	CACNA1A	GAB2	KSR1	PAK6	RGS4	SYN2
CDK4	IGFBP7	SCMH1	CACNA1G	GDI1	LIMK1	PAK7	RGS6	SYNCRIP
CDK6	IHH	SHC1	CACNG2	GIT1	LIN7A	PCLO	RGS7	SYNGAP1
CDKN1A	IL6	SHOX	CALB2	GNB5	LRP8	PDE1A	RHEB	SYNGR1
CYP19A1	INPPL1	SLC2A1	CALM1	GPRIN1	LRRC7	PIK3CA	RHOA	SYNJ1
CYR61	INSR	SLC2A4	CAMK2A	GRIA1	MAGI2	PJA1	RHOG	SYNPO
DOT1L	IRS2	SOCS2	CAMK2B	GRIA2	MAP2K1	PRKCA	RIMS1	SYT1
DYM	IRS4	SOX2	CAMK2G	GRIA3	MAP2K2	PRKCG	RIN1	SYT12
EFEMP1	JAK2	STAT1	CASK	GRIA4	MAPK1	PRRT2	RIT1	SYT7
ESR1	LHX4	STAT2	CAV1	GRID2	MAPK3	PSIP1	RIT2	SYTL4
FBLN5	LIN28A	STAT3	CC2D1A	GRIK1	MIB1	PTCHD1	RPS6KA3	TBR1
FGF2	LIN28B	STAT4	CDH2	GRIK2	MPDZ	PTK2B	RRAS	TRPC5
FGF3	MC3R	STAT5A	CDK5	GRIK3	NETO1	PTPN11	SDCBP	TSPAN7
FGFR4	MC4R	STAT5B	CDK5R1	GRIK4	NETO2	RAB10	SEMA4C	ULK1
FOS	MYC	STAT6	CFL1	GRIK5	NFASC	RAB2A	SHANK1	ULK2
FUBP3	NF1	TBX4	CLSTN1	GRIN1	NFATC4	RAB39B	SHANK2	UNC13B
GAB1	NIN	TGFA	CNIH2	GRIN2A	NFIA	RAB3A	SHANK3	VAMP2
GATA1	NOG	TP53	CNKSR1	GRIN2B	NFIB	RAB3GAP1	SHARPIN	VAV2
GDF5	OTX2	TRIP11	CNKSR2	GRIP1	NFIX	RAB5A	SHC2	VIPR2
GHR	PAPPA	TWIST1	CNTN1	GRIPAP1	NLGN1	RAB8A	SHOC2	YWHAB
GHRHR	PCNT	VDR	CNTN2	GRM1	NLGN3	RAC1	SIPA1L1	YWHAE
GHSR	PIK3CB	WNT4	CNTNAP2	GRM2	NLGN4X	RAF1	SNAP25	YWHAH
GLI2	PIK3CG	WNT7A	CTNNB1	GRM3	NPAS4	RALA	SOS1	YWHAZ
GRB10	PIK3R1	ZBTB38	CTNND2	GRM4	NPTN	RALGDS	SOS2	ZAP70
			DBN1	GRM5	NRAS	RAP1A	SPRED1	

Supplementary Table 3: RAS pathway/RAS related pathway gene list

The complete list of targeted genes for RAS pathway/RAS related pathway selected for this project. All supporting information collected through PubMed is represented with the PubMed ID numbers [unique identifier number used in PubMed] for easy identification for each gene. Predominant Brain expression is indicated by giving a score of 1 in the current list. HGNC- HUGO Gene Nomenclature Committee; ID – Intellectual disability

HGNC	Locus	Gene name; other names	Family	affects neurotransmission and/or synaptic plasticity, regulation	affects synapse structure and/or number	Neurophenotype models and/or affects learning/ memory	affects receptor trafficking or surface expression	predominant brain expression	linked to ID	Other evidence
ACTN2	1q43	F-actin cross-linking protein |alpha-actinin skeletal muscle	Cell Adhesion and Cytoskeletal	22427672		9454847	9009191			15072553
ADAM22	7q21.12	A disintegrin and metalloproteinase domain 22 (ADAM22), MDCL1	Cell Adhesion and Cytoskeletal	16990550		15876356	24227725	1	27066583 (single case)	20220021, 19692335, 20089912, 18206289
AKAP5	14q23.3	A kinase (PRKA) anchor protein 5, AKAP75, AKAP79, DAKAP1, H21, AKAP79/150, AKAP150	MAGUKs / Adaptors / Scaffolders	18711127, 23392692		18711127	19169250, 16510716	1	24665034 (chrm del)	20428246, 23462372, 23649627, 16504338
ARAF	Xp11.23	A-RAF | Proto-oncogene A-Raf, PKS2, RAFA1, ARAF1	Kinases			8805280				25097033
ARFGEF2	20q13.13	ADP-ribosylation factor guanine nucleotide-exchange factor 2 (brefeldin A-inhibited), BIG2	G proteins and modulators		24089482	22956851	15198677		14647276, 23812912, 19073947	19384555, 22676038, 21222180
ARHGEF6	Xq26.3	Rac/Cdc42 guanine nucleotide exchange factor (GEF) 6 , alpha-PIX, COOL2, COOL-2, PIXA	G proteins and modulators	21989057, 24715854	22554054	17105769, 21989057, 24987507			11017088, 12499396, 23871722	20861843, 23406282
ARHGEF7	13q34	Rho guanine nucleotide exchange factor (GEF) 7, BETA-PIX, COOL1	G proteins and modulators	12626503	22114281	10860822, 11266127			18203171 (chrm del)	22554054, 18184567, 25009260
BAIAP2/ IRSp53	17q25.3	brain-specific angiogenesis inhibitor 1-associated protein 2, Insulin Receptor Substrate P53, IRS-58	MAGUKs / Adaptors / Scaffolders	19208628		19193906, 24392092	15758177	1		24639075, 20888579, 19733838
BDNF	11p14.1	Brain-Derived Neurotrophic Factor, neurotrophin, NTF2, BULN2	NGF beta family	19293383, 24782711, 23602987	22161912, 19966931	7746324	24621058; 24391468	1	21567907, 20943059, 25714755	22435671, 22801293, 22922352, 25566435
BRAF	7q34	RAFB1 | B-Raf proto-oncogene serine/threonine-protein kinase (p94)	Kinases	16342120	24733831, 23505473	21383153		1	19206169, 23950000, 20859831	16474404, 8954940, 25389051, 25268071

BSN	3p21.31	Bassoon, ZNF231, KIAA0434, zinc finger protein 231, neuronal double zinc finger protein	Cell Adhesion and Cytoskeletal	12628169, 24442636, 24698275	23403927	20421286	10833299	1		21700703, 23516560
CABP1	12q24.31	calcium binding protein 1, Caldendrin, CALBRAIN	channels and receptors	11906216, 23027954		19224364, 24631676	18303947	1		12871994, 17055077, 23650371, 25058677
CACNA1A	19p13.2	calcium channel, voltage-dependent, P/Q type, alpha 1A subunit, **Brain calcium channel**	channels and receptors	10627593		9060410; 21228161, 24205277	20670620	1	19874387; 20097664, 25735478 , 23495138	24963350, 23495138
CACNA1G	17q21.33	voltage-gated calcium channel subunit alpha Cav3.1	channels and receptors	20371816		21296848		1	19455149; 22842074, 23949819, 21937992	11498049; 14526084, 22572848
CACNG2	22q12.3	Stargazin; TARPgamma2	Signalling molecules and Enzymes	15664178	17070505	20529126	11140673	1	21376300, 24581832, 25730879	
CALB2	16q22.2	calbindin 2, 29kDa (calretinin)	Signalling molecules and enzymes	9294225; 9758174		9294225				
CALM1	14q32.11	calmodulin 1	Kinases		10026200	25519244	9009191	1		
CAMK2A	5q32	ca2+/calmodulin-dependent protein kinase type II subunit alpha	Kinases	7781067		7781067, 25186740	19858198	1	23695276, 25790162	22592532
CAMK2B	7p13	CamKinase II beta subunit	Kinases	19503086		17913888		1		
CAMK2G	10q22.2	ca2+/calmodulin-dependent protein kinase type II subunit gamma	Kinases				7509863	1		15590149, 16679740
CASK	Xp11.4	calcium/calmodulin-dependent serine protein kinase (MAGUK family)	MAGUKs / Adaptors / Scaffolders		20623620	19165920, 25009461, 24062638	19620977	1	8786425, 21735175 , 24505460, 21735175	24278995, 20029458
CAV1	7q31.2	caveolin 1	MAGUKs / Adaptors / Scaffolders	21098662, 25611593	21203469	21203469	21799010			
CC2D1A/ Freud-1	19p13.12	coiled-coil and C2 domain containing 1A	fibrillar collagen family	21273312		22375002			16033914, 17394259, 25066123	23826796, 22023432
CDH2	18q12.1	cadherin 2, type 1, N-cadherin (neuronal)	Cell Adhesion and Cytoskeletal	16807326	15569714	17785185	16515543			24995881
CDK5	7q36.1	cyclin-dependent kinase 5	Kinases	21145377 , 17597494	15067135	17529984, 11978846	17529984, 19529798	1		10934255
CDK5R1	17q11.2	cyclin-dependent kinase 5, regulatory subunit 1 (p35)	Signalling molecules and Enzymes	15992381, 18053171		15992381		1	16425041, 14729829	
CFL1	11q13.1	Cofilin 1 (Non-Muscle)	Cell Adhesion and Cytoskeletal	20442266, 20835250	17875668	15649475, 23055942, 20407421	20407421, 23575840	0	22986108 (novel chrm del)	
CLSTN1	1p36.22	calsyntenin 1, Alcadeinalpha (Alcalpha)	Cell Adhesion and Cytoskeletal	12498782		17009726		1		17332754, 11161476
CNIH2	11q13.2	Cornichon Family AMPA Receptor Auxiliary Protein 2	cornichon	24853943		23522044	22815494, 21543622	1	22986108 (novel chrm del)	

CNKSR1	1p36.11	connector enhancer of kinase suppressor of Ras 1, CNK1	MAGUKs / Adaptors / Scaffolders						21937992	15845549, 14749388
CNKSR2	Xp22.12	connector enhancer of kinase suppressor of Ras 2	MAGUKs / Adaptors / Scaffolders	21937992	24656827			1	22511892, 25223753, 25644381, 25754917	
CNTN1	12q12	contactin-1	Cell Adhesion and Cytoskeletal	1729438	1729438	12441289; 22242131		1		15691716; 16367788; 21969550;
CNTN2	1q32.1	contactin-2, axonin 1, axonal glycoprotein TAG-1	Cell Adhesion and Cytoskeletal	11178983		18760366		1		15813926; 21205796; 19372728, 10469653
CNTNAP2	7q35	contactin associated protein-like 2	Cell Adhesion and Cytoskeletal	19896112		12963709		1	21082657; 19896112, 22670139, 25045150	21962519; 20164332, 23714751
CTNNB1	3p22.1	Catenin (Cadherin-Associated Protein), Beta 1, 88kDa, Armadillo	Cell Adhesion and Cytoskeletal	21903109	23536073	22190459	17270735	1	23111993, 25326669, 24668549, 24614104	22007134, 11262227, 23377854,
CTNND2	5p15.2	Delta-catenin; NPRAP	Cell Adhesion and Cytoskeletal	19914181	19403811, 25724647	15380068	17687028	1	10673328, 25473103, 24677774; 25839933	
DBN1	5q35.3	Drebrin A	Cell Adhesion and Cytoskeletal	19174472	12878700	19837137	16635259	1		25329999
DENSIN/ LRRC7	1p31.1	LRRC7, leucine rich repeat containing 7, Densin-180, LAP1	MAGUKs / Adaptors / Scaffolders	10827168	15647492	22072671		1	22072671	18248607, 12390249
DLG1	3q29	discs, large homolog 1; SAP97	MAGUKs / Adaptors / Scaffolders	12805297		18842882	19357261, 19620977	1	24838842, 20830797, 15918153	21850710
DLG3	Xq13.1	discs, large homolog 3; SAP102	MAGUKs / Adaptors / Scaffolders	17344405, 22896795		17344405	19104036	1	19795139, 24721225, 25649377, 15185169	23329067
DLG4	17p13.1	post-synaptic density protein 95 | PSD95 | SAP-90	MAGUKs / Adaptors / Scaffolders	9853749		9853749	19169250, 19104036	1	25123844, 19617690	23921260, 21151988
DLGAP1	18p11.31	discs, large (Drosophila) homolog-associated protein 1; GKAP	MAGUKs / Adaptors / Scaffolders	15496675	15496675	9221768		1		22940546
DNMT1	19p13.2	DNA methyltransferase 1	C5-methyltransferase family	20228804		20228804	21492995			8898232, 22338191
DOCK8	9p24.3	dedicator of cytokinesis protein 8	G proteins and modulators					1	18060736, 25435912	25713392, 25760145, 21948691
DYNC1H1	14q32.31	dynein, cytoplasmic 1, heavy chain 1	Synaptic Vesicles / Protein Transport		21346813	21380844, 24755273		1	21076407, 22368300, 24307404	8227145; 18952079
EEF1A2	20q13.33	Eukaryotic Translation Elongation Factor 1 Alpha 2	GTP-binding elongation factor family			15835265		1	24697219 (novel)	19909265, 17965018, 22848658
EPB41L1	20q11.23	erythrocyte membrane protein band 4.1-like 1	Cell Adhesion and Cytoskeletal	19225127		19225127	11050113	1	21376300, 25572454	19503082; 17335044
ERBB4	2q34	v-erb-a erythroblastic leukemia viral oncogene homolog 4	Kinases	22378872	17521571	15219717	17521571	1	22378872, 23633123	22972991; 22115776

FGD1	Xp11.22	FYVE, RhoGEF and PH domain-containing protein 1	G proteins and modulators					1	11940089, 25258334	24446295, 22876573
FYN	6q21	FYN oncogene related to SRC, FGR, YES	Kinases	1361685	7962063	1361685	14722243		26052347 (chrm del)	21872217; 9987012
GAB2	11q14.1	GRB2-associated binding protein 2	MAGUKs / Adaptors / Scaffolders			16009726		1		24161894, 19118819
GDI1	Xq28	GDP dissociation inhibitor 1, Oligophrenin-2	Rab family	7543319; 11027356	18829665	9620768, 12354782, 22291894		1	18487148, 22002931, 21736009, 11050624, 9620768	11498055, 24715854
GIT1	17q11.2	G Protein-Coupled Receptor Kinase Interacting ArfGAP 1, CAT-1, ARF GAP GIT1	G proteins and modulators	12695502, 23889934, 12473661	15800193	21499268, 20043896	12629171, 25284783	1		25792865
GNB5	15q21.2	guanine nucleotide binding protein 5	G proteins and modulators	21766168		10749990, 21883221, 21804131	23804514	1	27523599 (novel)	9606987
GPRIN1	5q35.2	G protein regulated inducer of neurite outgrowth 1	channels and receptors	18729205	15585744	20345915	18940194	1		17625504, 18068825
GRIA1	5q33.2	glutamate receptor, ionotropic, AMPA 1; GluR1	channels and receptors	22197030	19470654	12628184		1		16007085, 19403825, 18484081
GRIA2	4q32.1	glutamate receptor, ionotropic, AMPA 2; GluR2	channels and receptors	12805550	12848940	16914668		1	22669415, 20358617	8938126, 15619119
GRIA3	Xq25	Glutamate receptor ionotropic, AMPA 3	channels and receptors	12848940	18316356	17989220		1	10644433, 22124977, 24721225, 23637084	19449417, 17568425, 17989220
GRIA4	11q22.3	glutamate receptor, ionotrophic, AMPA 4; GluR4	channels and receptors	16190873	20662939	17233759		1		17610819
GRID2	4q22.1	glutamate receptor, ionotropic, delta 2; GluR-delta-2	channels and receptors	7715804	18551174	11672610, 18583162	12833050	1	23611888	12752376, 11672610, 18583162
GRIK1	21q21.3	glutamate receptor, ionotropic, kainate 1; GluR5	channels and receptors	11069933		15673679	10516295, 14969737	1		17245443, 22730074
GRIK2	6q16.3	glutamate receptor, ionotropic, kainate 2; GluR6	channels and receptors	11182092		9580260	11182093	1	17847003, 25039795, 21557188	16219388, 11144357
GRIK3	1p34.3	glutamate receptor, ionotropic, kainate 3; GluR7	channels and receptors	17620617		17620617		1	24449200	19369569, 19921975, 16958029
GRIK4	11q23.3	glutamate receptor, ionotropic, kainate 4; KA1	channels and receptors	19778510		22203159		1	16819533, 18824690	19142228
GRIK5	19q13.2	glutamate receptor, ionotropic, kainate 5; KA2	channels and receptors	12533602		19778510		1		12954862, 19086053
GRIN1	9q34.3	glutamate receptor, ionotropic, N-methyl D-aspartate 1: NR1	channels and receptors	8980238		20345915	18701073, 19776282	1	21376300, 25167861	10729336, 19142228
GRIN2A	16p13.2	glutamate [NMDA] receptor subunit epsilon-1:NR2A	channels and receptors	7816096		7816096		1	20384727, 20890276, 25596524; 23033978	
GRIN2B	12p13.1	GluN2B | N-methyl D-aspartate receptor subtype 2B | NR2B	channels and receptors	19838302	19726645	10485705		1	20890276, 24272827, 23918416, 23718928	20696923, 25356899

GRIP1	12q14.3	glutamate receptor interacting protein 1	MAGUKs / Adaptors / Scaffolders	18509036	15965473	18509036	10985351	1	17220210	12597860, 9069286
GRIPAP1	Xp11.23	GRIP1 associated protein 1 | KIAA1167 | **GRASP1**	G Proteins and modulators				10896157, 21576009	1		12207967, 17761173
GRM1	6q24.3	glutamate receptor, metabotropic 1; mGluR1	channels and receptors	9247074	11818568	7954802, 23045678	11591458	1	22901947	7954803, 23045678, 19924463
GRM2	3p21.2	glutamate receptor, metabotropic 2; mGluR2	channels and receptors	8662555	11805343	15619115	21326193	1		15081787, 21078304
GRM3	7q21.11	glutamate receptor, metabotropic 3; mGluR3	channels and receptors	12213275	15635057	15619115	18720515	1	24585043, 18256595, 15310849, 25150943	21945652, 21832989
GRM4	6p21.31	glutamate receptor, metabotropic 4	channels and receptors	8815915	11906782	9676970	12223229	1		16039800, 22227508
GRM5	11q14.3	glutamate receptor, metabotropic 5; mGluR5	channels and receptors	9185557		19321764, 25581360	17881561, 22652057	1	21548960, 24332449, 21901840, 25581360	22539775, 22072671
GRM7	3p26.1	glutamate receptor, metabotropic 7; mGluR7	channels and receptors	15820696, 12805997		15313036	22287544	1	23201551, 24804643, 23295062	18255102, 20371242
GRM8	7q31.33	glutamate receptor, metabotropic 8; mGluR8	channels and receptors	16188284		12213278	1612930	1		12213279, 19740090
HOMER1	5q14.1	HOMER1	MAGUKs / Adaptors / Scaffolders	18184796	12867517	16011574		1		21945599, 20673876
HOMER2	15q25.2	homer homolog 2 (Drosophila)	MAGUKs / Adaptors / Scaffolders	9808459		10493740	14511114			19914345
HRAS	11p15.5	c-H-ras; v-Ha-ras Harvey rat sarcoma viral oncogene homolog	G proteins and modulators	15182302		19371735	15182302		19371735, 25677562, 22926243, 23335589	11238881
IL1RAPL1	Xp21.3	interleukin 1 receptor accessory protein-like 1, Oligophrenin-4	channels and receptors	21926414	18657618	10471494, 19811529,	25312502		21940441, 25305082, 23785489, 23613341	20096586, 20714405, 21933724
ITPR1	3p26.1	type 1 InsP3 receptor | inositol 1,4,5-triphosphate receptor, type 1	channels and receptors	15016804, 15872106		8115665		1	10932263	11356237
KALRN	3q21.1	kalirin, RhoGEF kinase	G proteins and modulators	19625617, 24334022		19020030	18031682	1	27421267 (novel??)	
KCNA4	11p14.1	voltage-gated potassium channel subunit Kv1.4	channels and receptors	16000151		9547391	11723117		27582084 (novel??)	11122345, 9334400
KCNQ2	20q13.33	potassium voltage-gated channel, KQT-like subfamily, member 2	channels and receptors			10854243		1	15249611, 22169383, 24318194 , 23621294	
KCNQ3	8q24.22	Potassium Voltage-Gated Channel, KQT-Like Subfamily, Member 3	channels and receptors	18483067		18483067, 19060215		1	25524373 (novel)	
KRAS	12p12.1	Ki-Ras;v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog	G proteins and modulators	11713472		11713472			17056636, 22488932, 21797849	16474404, 9294608
KSR1	17q11.1	kinase suppressor of ras 1	Kinases	21471251, 10891492		16731514		1		10891492
LIMK1	7q11.23	LIM Domain Kinase 1	Kinases	12123613, 25645926	23086941	12123600, 12123613	23575840	1		23401971, 15458846

LIN7A	12q21.31	Lin-7 Homolog A (C. Elegans), Mammalian Lin-Seven Protein 1	MAGUKs / Adaptors / Scaffolders	16186258, 12393911		11104771			24658322, 18286632	16504495
LRP8	1p32.3	low density lipoprotein receptor-related protein 8, apolipoprotein e receptor	Signalling molecules and Enzymes	16102539		16102539		1		
MAGI2	7q21.11	Membrane Associated Guanylate Kinase, WW And PDZ Domain Containing 2	MAGUKs / Adaptors / Scaffolders		22128856	17438139	22878254, 22593060	1	22196487, 18565486	17724123, 11526121
MAP2K1	15q22.31	MKK1 | MAPK/ERK kinase 1	Kinases	15016380	15056710	14994337			19156172, 25423878,	19835659, 17586814
MAP2K2	19p13.3	MEK2 | dual specificity mitogen-activated protein kinase kinase 2	Kinases		10704499	22998872			19156172, 25487361, 23610052	21615688, 18432190
MAPK1	22q11.22	Mitogen-activated protein kinase 2 | ERK2	Kinases	22232579	24113259	17913910	16672655	1	18596172	21849556, 15194111
MAPK3	16p11.2	mitogen-activated protein kinase 3 | ERK1	Kinases	12062026		12062026				11160759, 10196568
MIB1	18q11.2	mindbomb E3 ubiquitin protein ligase 1		23111145	17728463	16000382; 16061358		1		18043734, 18498734
MPDZ	9p23	MUPP1 | Multi-PDZ domain protein 1	MAGUKs / Adaptors / Scaffolders	15312654, 18417361			18417361, 15312654		28397838 (novel)	19483657, 22994563
NDR2/ STK38L	12p11.23	serine/threonine kinase 38 like, Nuclear Dbf2-related kinase 2	Kinases	15308672	22445341	24719112				
NSMF	9q34.3	NMDA receptor synaptonuclear signaling and neuronal migration factor		21364755		18303947, 19608740	18303947	1		15018815, 19014935
NETO1	18q22.3	neuropilin (NRP) and tolloid (TLL)-like 1	MAGUKs / Adaptors / Scaffolders	19243221, 24403160	19243221	21623363	19243221	1		
NETO2	16q12.1	neuropilin (NRP) and tolloid (TLL)-like 2	MAGUKs / Adaptors / Scaffolders	19217376, 24403160		21734292		1		
NFASC	1q32.1	Neurofascin; NF	Cell Adhesion and Cytoskeletal	22306302	21382554	20188654		1	28940097 (novel??)	18573915
NFATC4	14q12	Nuclear Factor Of Activated T-Cells, Cytoplasmic, Calcineurin-Dependent 4	Transcription factors	10537109	15537643	22586092	19955386, 18184313			22977251, 11514544
NFIA	1p31.3	Nuclear factor I/A	CTF/NF-I family		12514217	10518556	20516644		22031302, 24462883, 24657733	17530927
NFIB	9p24.1	nuclear factor I/B, CTF, CCAAT-box-binding transcription factor	CTF/NF-I family	19020026	17553984	15632069		1		19961580
NFIX	19p13.2	Nuclear factor I/X	CTF/NF-I family		23042739	17353270		1	22301465, 20673863, 23495138	21800304, 18477394, 24963350
NLGN1	3q26.31	Neuroligin1	Cell Adhesion and Cytoskeletal	18579781		18579781, 18650386	19098102, 19450252	1	22106001, 15551338	16298368
NLGN3	Xq13.1	neuroligin-3	Cell Adhesion and Cytoskeletal	17823315	10996085	19360662	21642956	1	12669065, 16648374, 19406211, 25167861	

NLGN4X	Xp22.31	neuroligin-4, X-linked | Neuroligin X	Cell Adhesion and Cytoskeletal	21278334		18227507, 22989184		1	19545860; 14963808, 19125102	18194880
NPAS4/ NXF	11q13.2	Neuronal PAS Domain Protein 4		18815592, 24024041	23172225	22194569, 23029555, 21887312		1	22030050, 22986108	18182044, 23637184
NPTN	15q24.1	neuroplastin, GP55, GP65, SDFR1, SDR1, np65, np55	Cell Adhesion and Cytoskeletal	10759566, 25040546, 25300133	24554721	21480899	16925595, 22389504	1	24514566	17123723
NRAS	1p13.2	v-ras neuroblastoma RAS viral oncogene homolog	G proteins and modulators			11238881			19966803, 22887781, 22855653	
NRG1	8p12	pro-neuregulin-1, membrane-bound isoform | neuregulin 1 type IV beta 3	Cell Adhesion and Cytoskeletal	11399426, 10839362		12145742, 23719163	22972991			
NRG2	5q31.2	neuregulin 2	Cell Adhesion and Cytoskeletal	15207852; 15048684		15340081		1	22711443; 21594995	10369162
NRG3	10q23.1	neuregulin 3	Cell Adhesion and Cytoskeletal	16478787, 24431462	20713722	25093331		1	20713722; 22831755	12648463; 20548296
NRN1	6p25.1	neuritin 1	Signalling molecules and enzymes	9122250		22190461	14664806	1	23183317	19569075
NRXN1	2p16.3	Neurexin I-beta	Cell Adhesion and Cytoskeletal			19822762	14983056	1	19896112, 25486015, 24832020, 22670139	18057082, 1621094
NRXN2	11q13.1	neurexin-2-beta	Cell Adhesion and Cytoskeletal	17347997		12827191	14983056	1, eye	21424692, 24700553	17035546; 16406382
NTRK2/ TRKB	9q21.33	Neurotrophic Tyrosine Kinase, Receptor, Type 2, BDNF/NT-3 Growth Factors Receptor	Kinases	10995833, 12367511, 24212565	22949667, 17442456	12676531, 8402890, 9728915	24757570, 15829735, 21848649	1	27884935	10985347
NUMB	14q24.3	numb homolog (Drosophila)	Signalling molecules and enzymes		16394100	14729486	16394100			14687546
OPHN1	Xq12	oligophrenin-1	G Proteins and modulators	19487570		17728457, 19401298	19487570	1	9582072, 24105372 , 25649377, 21796728	19401298, 24637888
PACSIN1	6p21.31	protein kinase C and casein kinase substrate in neurons 1; Syndapin1	MAGUKs / Adaptors / Scaffolders	16648848, 24751538	23420842	24509844	16617342, 23918399	1		
PAK3	Xq23	p21-activated kinase 3 | OPHN3	Kinases	17537723	17105769, 18481281	15574732, 16014725		1	9731525, 18523455, 24556213 , 22238087	17853471
PAK6	15q15.1	p21 protein (Cdc42/Rac)-activated kinase 6	Kinases			18675265, 23593460		1		
PAK7/ PAK5	20p12.2	p21 protein (Cdc42/Rac)-activated kinase 7	Kinases	22371566		18675265, 23593460		1		22732262, 12032833, 24474471
PCLO	7q21.11	piccolo (presynaptic cytomatrix protein), aczonin, KIAA0559, ACZ	MAGUKs / Adaptors / Scaffolders	11285225, 23474894	23403927	19766155, 23620758		1	23862039 (micro deletion)	21712437, 24167553, 12175852
PDE1A	2q32.1	calmodulin-dependent phosphodiesterase	Signalling molecules and Enzymes	19422887					20552675 (chrm del region)	15272012, 20552675
PIK3CA	3q26.32	phosphatidylinositol 3-kinase, catalytic, 110-kd, alpha	PI3/PI4-kinase family							

PJA1	Xq13.1	praja ring finger 1, E3 ubiquitin protein ligase, RNF70	E3 ubiquitin-protein ligase			11533224		1	17941886	12036302, 9393880, 23717400
PRKCA	17q24.2	PKC alpha	Kinases	15541307	16740968	14500987		1	22166941	
PRKCG	19q13.42	protein kinase C gamma type	Kinases	8269510		8269509	16571747, 15606904	1	21937992	15313841, 1456172, 16763984
PRRT2	16p11.2	proline-rich transmembrane protein 2		25194488, 23077019				1	21937992, 25060993, 25595153, 23126439	128200; 602066; 605000
PSIP1	9p22.3	PC4 and SFRS1 interacting protein 1, LEDGF, Transcriptional coactivator p75/p52	HDGF family		15642333	16980622				15464262, 18652779
PTCHD1	Xp22.11	patched domain-containing protein 1	patched family						20844286, 25131214, 21091464, 25782667	
PTK2B	8p21.2	protein tyrosine kinase 2 beta | FADK 2 , CAKB, FAK2, PYK2, RAFTK, PKB, CADTK, PTK, FRNK	Kinases	11239437, 10354603, 20071509	24718602	8910543	15814199, 11239437	1		15537634, 12946883
PTPN11	12q24.13	SH-PTP2 | protein-tyrosine phosphatase 1D | Shp2	Signalling molecules and enzymes	19047464	17679554	17442246		1	11704759, 22585553, 19864201	16573649, 20400923, 17546245
RAB10	2p23.3	ras-related GTP-binding protein RAB10	G Proteins and modulators	21856246, 24478353			17761527	1		
RAB2A	8q12.1	ras-related protein Rab-2A	G-proteins and Modulators		2111712			1	20308991	
RAB39B	Xq28	ras-related protein Rab-39B	Rab family	20159109		24357492	25784538	1	21076407, 25434005, 24700761, 25784538	20159109
RAB3A	19p13.11	RAS-associated protein RAB3A	Synaptic Vesicles / Protein Transport	9856469; 9194562	20338242; 17640821	15078563, 11598194		1		16436611, 9252190, 16584842
RAB3GAP 1	2q21.3	rab3 GTPase-activating protein catalytic subunit	Rab family	16782817		16782817	19356697		15696165, 25332050, 23124039, 20512159	18413245, 23176487, 23420520, 16532399
RAB5A	3p24.3	Ras-related protein 5A	G Proteins and modulators	16141272			15629704			8043272
RAB8A	19p13.12	RAB8A, member RAS oncogene family	G Proteins and modulators			18381201	15297461	1		18772192
RAC1	7p22.1	ras-related C3 botulinum toxin substrate 1 (rho family, small GTP binding protein Rac1), TC25, MIG5, TC-25, p21-Rac1	G proteins and modulators	19394428	15234347	19394428, 25613020	16291935		25818528, 21645877	18440141
RAF1	3p25.2	RAF | proto-oncogene c-RAF	Kinases			17396120			20052757, 23877478, 20683980	
RALA	7p14.1	ras-related protein Ral-A	G Proteins and modulators	11865051	16330713, 19383721	23433113	19823667			24284074
RALGDS	9q34.3	ral guanine nucleotide dissociation stimulator	G Proteins and modulators				15470141	1	21937992	11748241, 7809086

RAP1A	1p13.2	KREV-1 | ras-related protein Rap-1A	G Proteins and modulators	12824760, 18305243		18305243	12824760		26280580 (novel??)	24451631
RAPGEF2	4q32.1	CNrasGEF | PDZ domain-containing guanine nucleotide exchange factor 1	G proteins and modulators		25189171	19453629	11168587	1		23800469
RAPGEF3 /EPAC	12q13.11	Rap guanine nucleotide exchange factor (GEF) 3	G Proteins and modulators	18509114	20851749	20516079, 22365550		1		22260665, 19001039, 19734897
RAPGEF4	2q31.1	Rap guanine nucleotide exchange factor (GEF) 4; EPAC2	G proteins and modulators		19734897	22915127	19734897	1		
RASA1	5q14.3	RAS p21 protein activator (GTPase activating protein) 1	G Proteins and modulators			7477259			21626678, 22670137	
RASA3	13q34	ras GTPase-activating protein 3	G Proteins and modulators						23639964 (chrm del region)	18952607
RASAL1	12q24.13	RAS protein activator like 1 (GAP1 like)	G Proteins and modulators							16009725
RASGRF1	15q25.1	Ras protein-specific guanine nucleotide-releasing factor 1	G proteins and modulators	16467520, 23766509	16481401, 25644714, 22744634	11640934, 21251221, 22164138	14622581, 15029245	1	24808846	9384379, 10373510, 24470023
RASGRF2	5q14.1	Ras protein-specific guanine nucleotide-releasing factor 2	G proteins and modulators	16467520		9819557	9384379		23274185, 22678782	9707409, 11909944, 10733575
RASGRP1	15q14	CalDAG-GEFII | ras activator RasGRP	G Proteins and modulators	9789079	11738253				21866111, 24987507	18650386, 11955714, 18650386
RASGRP2	11q13.1	CALDAG-GEFI, RAS guanyl releasing protein 2	G Proteins and modulators							14988412, 10918068
RGL1	1q25.3	RalGDS-like 1	G Proteins and modulators			15037549		1	27431290 (novel??)	7935463
RGL2	6p21.32	RalGDS-like 2 | RAB2L	G Proteins and modulators							18540861
RGS4	1q23.3	regulator of G-protein signalling 4	G proteins and modulators	9437012, 18493969		15870291, 15661377	17101972	1	18834502, 21215802, 20414142, 19282471	23093381, 17301167, 17006672
RGS6	14q24.2	regulator of G-protein signalling 6	G proteins and modulators		12140291	22179605			25525169	14734556
RGS7	1q43	regulator of G-protein signalling 7	G proteins and modulators	19042037		20100282		1	15822126	agoraphobia, 11906535
RHEB	7q36.1	Ras Homolog Enriched In Brain	G Proteins and modulators	24889507; 22008911	23392671, 8206940	21238928; 23028662			23892008, 23485365 , 15562827	15150271, 24391850
RHOA	3p21.31	ras homolog gene family, member A	G proteins and modulators	22411227	16449195	21502507	15031678, 17438139			21451048, 22405202, 10455249, 19401298
RHOG	11p15.4	ras homolog gene family, member G (rho G) | ARHG	G proteins and modulators	21900250	22588079	19812248				22991824, 12393274
RIMS1	6q13	RIM 1 | Rab-3-interacting molecule 1 |	Rab family	11797009	21241895	17124501	15967982	1	25284784 (ASD)	19074017, 11797010
RIN1	11q13.2	Ras and Rab interactor 1	RIN (Ras interaction /interference) family	12574403		19830836		1		9144171, 11784866

RIT1	1q22	Ras-like without CAAX 1, RIBB, ROC1, Ras-like protein in tissues	G proteins and modulators	18388731	12668729	22815504			25124994, 25049390, 24939608, 23791108	18378158, 27699752,
RIT2	18q12.3	Ras-like without CAAX 2, RIBA, RIN, ROC2	G proteins and modulators	20489179	17460085	23805044, 19303663		1	21145994 (chrm del region)	8824319, 22451204
RPS6KA3/ RSK-2	Xp22.12	Ribosomal Protein S6 Kinase, 90kDa, Polypeptide 3	Kinases		24336713	23742761, 17033934	21116650, 16217014		17100996, 23495752, 21930553	21838783, 12393804, 24416220
RRAS	19q13.33	ras-related protein R-Ras	G proteins and modulators	25043327	15297673	22174156				15772154, 16237331
SDCBP	8q12.1	syndecan binding protein (syntenin) , ST1, SYCL, MDA-9, TACIP18	MAGUKs / Adaptors / Scaffolders	18434645	15797720, 21964490		14689485			15276154, 20233453
SEMA4C	2q11.2	SEMAPHORIN 4C	MAGUKs / Adaptors / Scaffolders	11134026	15978582	21122816		1		17498836
SHANK1	19q13.33	SH3 and multiple ankyrin repeat domains 1; SPANK-1	MAGUKs / Adaptors / Scaffolders	18272690	18272690, 25692235	18272690, 21695253, 20868654		1	22503632	20868654, 21695253, 24124131
SHANK2	11q13.4	SH3 and multiple ankyrin repeat domains 2; PROSAP1, CORTBP1, SPANK3, CTTNBP1	MAGUKs / Adaptors / Scaffolders	21994763		22699619	10433268, 22699620	1	20473310, 22699620, 23350639, 20880122	15632121, 25188300, 25560758
SHANK3	22q13.33	SH3 and multiple ankyrin repeat domains 3; PROSAP2	MAGUKs / Adaptors / Scaffolders	21558424, 23739967	15814786	21423165, 22573675	23739967, 21795692, 24652766, 21565394	1	21376300, 21167025, 25646853; 24089484	17173049, 21795692, 21378602, 22922660
SHARPIN	8q24.3	Shank-interacting protein-like 1	MAGUKs / Adaptors / Scaffolders		11178875	17538631				
SHC2	19p13.3	Src homology 2 domain-containing-transforming protein C2	MAGUKs / Adaptors / Scaffolders	9507002		11163269				8610109, 11791173
SHOC2	10q25.2	soc-2 suppressor of clear homolog (C. elegans)	MAGUKs / Adaptors / Scaffolders		25514808, 21732489	21732489			20882035, 23918763, 25123707	21396583, 21548061
SIPA1L1	14q24.2	signal-induced proliferation-associated 1-like 1 | E6TP1 | KIAA0440	MAGUKs / Adaptors / Scaffolders	21987493	11502259	19442707	21362453			17706945, 20547063
SNAP25	20p12.2	synaptosomal-associated protein, 25kD	Synaptic Vesicles / Protein Transport	16020741, 23732542	8332215	16870134, 21949876		1	25003006, 22762387, 25650683	19679075, 11753414, 9617921
SOS1	2p22.1	son of sevenless homolog 1 (Drosophila)	MAGUKs / Adaptors / Scaffolders	10517950		21041952			18651097, 22585553, 21387466, 20683980	11777939
SOS2	14q21.3	son of sevenless (Drosophilia) homolog 2	MAGUKs / Adaptors / Scaffolders	19429099		10938118	19429099		25795793 (novel)	21779500
SPRED1	15q14	sprouty-related, EVH1 domain containing	Sprouty	19118178	20047999	19118178			19366998, 24334617, 21089071	20047999, 22802525
SPRY1	4q28.1	Sprouty Homolog 1 (Drosophila)	Sprouty		21595564	17707653; 18423429	24980605		24980605 (Chrm del)	25049390, 21362415, 12402043

SPRY2	13q31.1	Sprouty Homolog 2 (Drosophila)	Sprouty	22383529, 21240919, 19683577, 17599098	25822989	15937482		1	19022413	21693512, 18335055, 18297599 , 19022413
SPRY3	Xq28	Sprouty Homolog 3 (Drosophila)	Sprouty		21062861	21062861				
SPRY4	5q31.3	Sprouty Homolog 4 (Drosophila)	Sprouty	22384148		17156747, 21693512			28539120 (novel??)	23219993, 16251209
SRC	20q11.23	v-src sarcoma (Schmidt-Ruppin A-2) viral oncogene homolog	Kinases	9478899	11867627, 25711940, 25451123	7958873	22993256			25128699
STRN	2p22.2	striatin, calmodulin binding protein	Signalling molecules and enzymes	9712839		10413453		1		16460920
STRN4	19q13.32	striatin, calmodulin binding protein 4, Zinedin	Signalling molecules and enzymes	23015759				1		18466332, 10748158
STX1A	7q11.23	Synatxin 1A	Synaptic Vesicles / Protein Transport	18703708	8973813	16723534		1	20662849, 15219469, 25445064	16722236, 20576034
STX4	16p11.2	syntaxin-4A (placental), STX4A, SYN-4, SYN4	Synaptic Vesicles / Protein Transport	20959521	20434989		20434989, 20814225			12648463
STXBP1	9q34.11	syntaxin binding protein 1	Synaptic Vesicles / Protein Transport	10657302		19557857		1	21364700, 19557857, 23020937, 24253858	22722545, 21762454, 24095819, 20876469
SYN1	Xp11.23	synapsin I ; BRAIN PROTEIN 4.1	Synaptic Vesicles / Protein Transport	17093089		18057210		1	14985377, 16118346, 20662849	21798362, 20530578
SYN2	3p25.2	synapsin II	Synaptic Vesicles / Protein Transport	8964517		22805168		1		11404405, 22384280
SYNCRIP	6q14.3	Synaptotagmin-binding, cytoplasmic RNA-interacting protein		18045242		15475564			27479843 (novel??)	10734137
SYNGAP1	6p21.3	synaptic Ras GTPase activating protein 1	G Proteins and modulators	16537406, 23785156	15470153, 23141534	12427827, 24945774	16537406	1	21237447, 23687080, 23161826, 20683986	19196676, 22050443, 20188038,
SYNGR1	22q13.1	Synaptogyrin 1	Synaptic Vesicles / Protein Transport	10595519		10595519		1		17239033
SYNJ1	21q22.11	Synaptojanin 1	Synaptic Vesicles / Protein Transport	11717343, 10595519		10535736		1	15261714, 11443522, 25302295	14622570, 10931870
SYNPO	5q33.1	Synaptopodin	Cell Adhesion and Cytoskeletal	12928494, 23630268	12928494, 25164660, 23884954	12928494				
SYT1	12q21.2	Synaptotagmin I	Synaptic Vesicles / Protein Transport	7954835	11078930	16729982		1	25705886 (novel??)	
SYT12	11q13.2	synaptotagmin XII	Synaptic Vesicles / Protein Transport	17190793, 23739973				1		11404423, 24164654

SYT7	11q12.2	Synaptotagmin VII	Synaptic Vesicles / Protein Transport	18308933		12925704		1		21576241, 14532108
SYTL4	Xq22.1	synaptotagmin-like protein 4	Synaptic Vesicles / Protein Transport			17761531			22091964	11773082
TBR1	2q24.2	T-box, brain, 1	T-box genes family			11239428	15584924	1	23444363, 25356899, 24458984, 23112752	9883721, 21285371, 10749215
TRPC5	Xq23	Transient Receptor Potential Cation Channel, Subfamily C, Member 5	channels and receptors	19450521	17626205, 16469785, 12858178	22135323, 19450521, 10493832	22699894, 22539836	1	24817631	15254149, 24756705, 17217053
TSPAN7	Xp11.4	Tetraspanin 7, MRX58, T4S2, TM4SF2	tetraspanin (TM4SF) family	22880149, 12207950			22445342	1	12070254, 12150222, 12376945, 22511893	20479760, 12376945
ULK1	12q24.33	unc-51-like kinase 1 (C. elegans)	Kinases	8878822	15014045			1	21457577 (chrm del region)	11146101, 21457577
ULK2	17p11.2	unc-51-like kinase 2 (C. elegans)	Kinases	15014045		21734278				17394779, 10624947, 24923441
UNC13B	9p13.3	unc-13 homolog B (C. elegans)	Synaptic Vesicles / Protein Transport	15294163, 11792327, 22966208	19700493, 19558601	12070347	10440376			11832228, 17267576, 22674279
VAMP2	17p13.1	vesicle-associated membrane protein 2 (synaptobrevin 2)	Synaptic Vesicles / Protein Transport	11691998	10958975	16793874		1	23966691, 25445064	15788763
VAV2	9q34.2	Vav 2 Guanine Nucleotide Exchange Factor	G proteins and modulators	21880903		15848800, 17202406, 20089829				
VIPR2	7q36.3	Vasoactive Intestinal Peptide Receptor 2, VPAC2, PACAP Type III Receptor	channels and receptors			16641377		1	21346763, 24095776, 24334122, 22813947	16202621
YWHAB	20q13.12	brain protein 14-3-3, beta isoform	MAGUKs / Adaptors / Scaffolders					1		
YWHAE	17p13.3	14-3-3 epsilon	MAGUKs / Adaptors / Scaffolders		12796778	20228804		1	18658164; 19584063; 19635726	14651979, 9581554
YWHAH	22q12.3	14-3-3 protein eta	MAGUKs / Adaptors / Scaffolders	15543142			18417361	1	19160447 (schizophrenia and bipolar disorder)	19483657
YWHAZ	8q22.3	(14-3-3 zeta); KCIP-1	MAGUKs / Adaptors / Scaffolders	24367683		11606630			23999528 (ASD)	
ZAP70	2q11.2	zeta-chain (TCR) associated protein kinase 70kDa, SRK, STD, TZK,	Kinases		21895656					21354221

Supplementary Table 4: Short stature pathway gene list

The complete list of targeted genes selected in this project for short stature related pathway. The evidence of known phenotypes/syndromes/disorders in Humans is given through OMIM with their mode of inheritance. In the current list, score 2 is given for predominant Bone/cartilage expression; Score 1 for moderately or low and Null score for no expression in skeletal system. HGNC- HUGO Gene Nomenclature Committee; AD- autosomal dominant, AR- autosomal recessive; XR – X-linked recessive.

HGNC	Locus	Family	Gene name; other names	predominant bone/ cartilage expression	Human Diseases	Mode of inheritance	OMIM
ACAN	15q26.1	aggrecan/ Versican proteoglycan family	CSPG1 | aggrecan core protein | cartilage-specific proteoglycan core protein | Chondroitin sulfate proteoglycan 1	2	(1) Achondroplasia; (2) osteochondritis dissecans, short stature, and early-onset osteoarthritis; od	AD AD	100800 165800
ADAMTS17	15q26.3	ADAMTS protein family	ADAMTS-17 |A disintegrin and metalloproteinase with thrombospondin motifs 17 | FLJ16363 | ADAM-TS17 | ADAM-TS 17 |	0	Weill-marchesani-like syndrome	AR	613195
ADAMTSL3	15q25.2	ADAMTS protein family	ADAMTS-like protein 3 | Punctin-2 | KIAA1233 | a disintegrin-like and metalloprotease domain with thrombospondin type I motifs-like 3	0	----------		
ADIPOQ	3q27.3	soluble defense collagens family	ADIPQTL1 | gelatin-binding protein 28 | ACRP30| Adipocyte, C1q and collagen domain-containing protein | adiponectin |ADPN | GBP28	0	Adiponectin deficiency		612556
AKT1	14q32.33	Ser/Thr protein kinase family	protein kinase B alpha | proto-oncogene c-Akt | v-akt murine thymoma viral oncogene homolog 1 | RAC-alpha serine/threonine-protein kinase	0	(1)Proteus syndrome (PROTEUSS) (2)Cowden Syndrome 6; CWS6	Somatic AD	176920 615109
AKT2	19q13.2	Ser/Thr protein kinase family	V-Akt Murine Thymoma Viral Oncogene Homolog 2, Protein Kinase Akt-2 , RAC-BETA	0	(1) Diabetes Mellitus, Noninsulin-Dependent; Diabetes Mellitus, Type II (2) Hypoinsulinemic hypoglycemia with hemihypertrophy	AD	125853 240900
ANAPC13	3q22.2	Anaphase Promoting complex family	Cyclosome subunit 13 | anaphase promoting complex subunit 13 | SWM1 | DKFZP566D193 |cyclosome subunit 13 | APC13	0	----------		
AR (in girls)	Xq12	nuclear hormone receptors family	Androgen Receptor, Dihydrotestosterone Receptor DHTR	2	(1)Spinal and Bulbar Muscular Atrophy, X-Linked 1; SMAX1 (2) Androgen insensitivity (3) Androgen insensitivity, partial, with or without breast cancer (4) Hypospadias 1, X-linked	XR	313200 300068 312300 300633
ARRB1	11q13.4	arrestin family	beta-arrestin-1 | arrestin 2 |ARB1 | ARR1	0	----------		
BMP2	20p12.3	bone morphogenetic protein family	bone morphogenetic protein 2A |BDA2 | BMP-2A	1	(1) Brachydactyly, type A2 (2) {HFE hemochromatosis, modifier of}	AD AR	112600 235200

BMP4	14q22.2	bone morphogenetic protein family	BMP2B1 | Bone morphogenetic protein 2B | bone morphogenetic protein 4 BMP-4 | OFC11 | MCOPS6 | DVR4	1	(1) Microphthalmia, syndromic 6 (2) Orofacial cleft 11	(1) AD	607932 600625
BMP6	6p24.3	bone morphogenetic protein family	VGR-1 |Vg1-related sequence | VGR | bone morphogenetic protein 6 | BMP-6 | vegetal related growth factor (TGFB-related)	2	----------		
CBL	11q23.3		E3 ubiquitin-protein ligase CBL | RING finger protein | casitas B-lineage lymphoma proto-oncogene | CBL2 | NSLL | FRA11B | RNF55	0	Noonan syndrome-like disorder with or without juvenile myelomonocytic leukemia		613563
CDK4	12q14.1	CMGC Ser/Thr protein kinase family	cell division protein kinase 4 | cyclin-dependent kinase 4 | PSK-J3 | CMM3	0	Melanoma, cutaneous malignant 3 (CMM3)	AD	609048
CDK6	7q21.2	CMGC Ser/Thr protein kinase family	cell division protein kinase 6 | serine/threonine-protein kinase PLSTIRE	2	Microcephaly 12, primary, autosomal recessive; mcph12	AR	616080
CDKN1A	6p21.2	CDK inhibitor family	DNA synthesis inhibitor | cyclin-dependent kinase inhibitor 1 | MDA-6 | CAP20 | p21 | WAF1 | PIC1 | MDA6	0	cip1/waf1 tumor-associated polymorphism 1		116899
CYP19A1	15q21.2	cytochrome P450 subfamily XIX	cytochrome P450 19A1 | microsomal monooxygenase | P-450AROM | cytochrome P450, family 19, subfamily A, polypeptide 1 | CPV1 | Estrogen synthase | flavoprotein-linked monooxygenase | CYP19 | ARO | CYAR | Aromatase | ARO1 |cytochrome P-450AROM	0	(1) Aromatase deficiency (2) Aromatase excess syndrome	(1)AR (2) Male limited AD vs AR or X-linked	613546 139300
CYR61	1p22.3	CCN (CYR61/CTGF/NOV) protein family of matricellular proteins	IBP-10 | CCN family member 1 | Insulin-like growth factor-binding protein 10 | Protein GIG1 | GIG1 | protein CYR61 | IGFBP10 | Cysteine-rich angiogenic inducer 61 |cysteine-rich heparin-binding protein 61	1	----------		
DOT1L	19p13.3	DOT1 family	DOT1-like protein, histone-lysine N-methyltransferase	0	----------		
DYM	18q21.1		DMC | FLJ20071 | FLJ90130 |dyggve-Melchior-Clausen syndrome protein | SMC | dymeclin | Dyggve-Melchior-Clausen syndrome protein	2	(1) Dyggve-Melchior-Clausen disease (2) Smith-McCort dysplasia	AR AR	223800 607326
EFEMP1	2p16.1	fibulin family	FBLN3 | EGF-containing fibulin-like extracellular matrix protein 1 | DRAD | extracellular protein S1-5 | MTLV | Fibulin-3 | FBNL	2	DOYNE HONEYCOMB RETINAL DYSTROPHY; DHRD	AD	126600
ESR1	6q25.1	nuclear hormone receptor family	estrogen receptor alpha delta | estrogen receptor alpha E1-E2-1-2 | ER | ESR | estrogen receptor alpha E1-N2-E2-1-2 | estradiol receptor | Estradiol receptor | NR3A1 | ER-alpha | estrogen nuclear receptor alpha | Nuclear receptor subfamily 3 group A member 1 | Era |ESRA	1	(1) Breast Cancer (2) Estrogen Resistance; ESTRR (3) Migraine with or without Aura, Susceptibility to, 1	AD AR AD	114480 615363 157300
FBLN5	14q32.12	fibulin family	FIBL-5 | fibulin-5 | developmental arteries and neural crest EGF-like protein | Developmental arteries and neural crest EGF-like protein | urine p50 protein | ARMD3 | UP50 | DANCE	1	(1) Cutis laxa, autosomal dominant 2 (2) Cutis laxa, autosomal recessive, type IA (3) Macular degeneration, age-related, 3	(1)AD (2) AR	614434 219100 608895
FGF2	4q27	heparin-binding fibroblast growth factor family	fibroblast growth factor 2 | basic fibroblast growth factor bFGF | prostatropin | HBGF-2 | bFGF | Heparin-binding growth factor 2	2	----------		
FGF3	11q13.3	heparin-binding fibroblast growth factor family	Heparin-binding growth factor 3 |INT-2 proto-oncogene protein | murine mammary tumor virus integration site 2, mouse | fibroblast growth factor 3 | oncogene INT2 HBGF-3	0	Deafness, congenital, with inner ear agenesis, microtia, and microdontia	AR	610706
FGFR4	5q35.2	Tyr protein kinase family	CD334 antigen | TKF | JTK2 | tyrosylprotein kinase | CD334 protein-tyrosine kinase | fibroblast growth factor receptor 4	2	----------		

FOS	14q24.3	bZIP family	FBJ Murine Osteosarcoma Viral Oncogene Homolog, Cellular Oncogene C-Fos, G0/G1 Switch Regulatory Protein 7, P55	0	----------		
FUBP3	9q34.11		FBP3 | FUSE-binding protein 3 |far upstream element-binding protein 3 | far upstream element (FUSE) binding protein 3	0	----------		
GAB1	4q31.21	GAB family	GRB2-associated binding protein 1 |growth factor receptor bound protein 2-associated protein 1	0	----------		
GATA1	Xp11.23	GATA zinc finger transcription factor family	NFE1 | ERYF1 | erythroid transcription factor 1 | GATA-binding protein 1 (globin transcription factor 1) XLANP | GF-1 | transcription factor GATA1 | XLTT | XLTDA	0	(1) Anemia, X-linked, with/without neutropenia and/or platelet abnormalities (2) Leukemia, megakaryoblastic, with or without Down syndrome, somatic (3) Thrombocytopenia with beta-thalassemia, X-linked (4) Thrombocytopenia, X-linked, with or without dyserythropoietic anemia	XR isolated cases XR XR	300835 190685 314050 300367
GDF5	20q11.22	bone morphogenetic protein (BMP) family	GDF-5 | Cartilage-derived morphogenetic protein 1 | OS5 | BMP14 | growth differentiation factor 5 | CDMP1 |LAP4 | Radotermin | CDMP-1 | SYNS2	2	(1) Acromesomelic dysplasia, Hunter-Thompson type (2) Brachydactyly, type A1, C (3) Brachydactyly, type A2 (4) Brachydactyly, type C (5) Chondrodysplasia, Grebe type (6) Du Pan syndrome (7) Multiple synostoses syndrome 2 (8) Symphalangism, proximal, 1B (9) {Osteoarthritis-5}	AR AR AD AD AR AR AD AD Multifactorial	201250 615072 112600 113100 200700 228900 610017 615298 612400
GHR	5p13.1	cytokine family of receptors	growth hormone receptor |GHBP | growth hormone binding protein | somatotropin receptor | serum binding protein | Somatotropin receptor | GH receptor	2	(1) Laron dwarfism (2) Short stature (3) {Hypercholesterolemia, familial, modifier of}	(1) AR (2)Dominant ,pseudoautosomal (3) AD	262500 604271 143890
GHRHR	7p14.3	G protein coupled receptor family	IGHD1B | GRFR | growth hormone releasing hormone receptor | GHRFR | GRF receptor | GHRH receptor	0	Isolated growth hormone deficiency, type 1b; IGHD1B		612781
GHSR	3q26.31	G protein coupled receptor family	growth hormone secretagogue receptor type 1 | GHS-R | Ghrelin receptor | GH-releasing peptide receptor |ghrelin receptor | growth hormone secretagogue receptor | GHRP	0	Short stature	Dominant, pseudo autosomal	604271
GLI2	2q14.2	GLI C2H2-type zinc-finger protein family	tax-responsive element-25-bp sequence binding protein | tax helper protein 1 | glioma-associated oncogene family zinc finger 2	0	(1) Culler-Jones syndrome (2) Holoprosencephaly-9	AD AD	615849 610829
GRB10	7p12.1	GRB7/10/14 family	GRB10 adaptor protein | GRBIR | KIAA0207 | maternally expressed gene 1 | insulin receptor-binding protein Grb-IR | RSS | MEG1	1	----------		
GRB2	17q25.1	GRB2/sem-5/DRK family	growth factor receptor-bound protein 2, Growth Factor Receptor-Bound Protein 3, SH2/SH3 Adapter GRB2, NCKAP2	2	----------		
HESX1	3p14.3	ANF homeobox family	HANF | Homeobox protein ANF | CPHD5 | RPX | homeobox expressed in ES cells 1 | Rathke pouch homeobox | ANF	0	SEPTOOPTIC DYSPLASIA	AD/AR	182230

HHIP	4q31.21	hedgehog family	HIP |FLJ20992 | hedgehog interacting protein	0	----------		
HMGA1	6p21.31	HMGA family	High Mobility Group AT-Hook 1	0	Diabetes Mellitus, Noninsulin-Dependent; Diabetes Mellitus, Type II	AD	125853
HMGA2	12q14.3	HMGA family	High Mobility Group AT-Hook 2	0	LEIOMYOMA, UTERINE; UL	Somatic	150699
IGF1	12q23.2	insulin family	somatomedin-C | MGF | insulin-like growth factor IB | Somatomedin-C | insulin-like growth factor IA | IGF-IB |IGF-IA	1	Growth retardation with deafness and mental retardation due to IGF1 deficiency	AR	608747
IGF1R	15q26.3	Tyr protein kinase family	CD221 | Insulin-like growth factor I receptor | soluble IGF1R variant 2 JTK13 | IGF-I receptor | IGFR	1	Insulin-like growth factor I, resistance to	AD/AR	270450
IGF2	11p15.5	insulin family	C11orf43 | insulin-like growth factor type 2 |PP9974 | Somatomedin-A | chromosome 11 open reading frame 43 | insulin-like growth factor 2 (somatomedin A) | FLJ44734 | IGF-II	0	(1) Beckwith-Wiedemann Syndrome; BWS (2) Silver-Russell Syndrome; SRS (3) Wilms Tumor 1; WT1	AD isolated cases AD/Somatic	130650 180860 194070
IGFALS	16p13.3		insulin-like growth factor binding protein complex acid labile chain	0	acid-labile subunit deficiency; aclsd		615961
IGFBP1	7p12.3	insulin-like growth factor binding protein family	binding protein-25/26/28 | AFBP | Placental protein 12 | insulin-like growth factor-binding protein 1 |alpha-pregnancy-associated endometrial globulin |amniotic fluid binding protein	0	----------		
IGFBP3	7p12.3	insulin-like growth factor binding protein family	insulin-like growth factor binding protein 3 | IBP-3 | BP-53 |acid stable subunit of the 140 K IGF complex | IGF-binding protein 3 | growth hormone-dependent binding protein	2	----------		
IGFBP4	17q21.2	insulin-like growth factor binding protein family	IBP-4 | IGF-binding protein 4 | IGFBP-4 | IBP4 | BP-4 | HT29-IGFBP | insulin-like growth factor-binding protein 4 | insulin-like growth factor binding protein 4 |BP-4	2	----------		
IGFBP7	4q12	insulin-like growth factor binding protein family	AGM | FSTL2 | TAF | IGFBPRP1 | IGFBP-7 |IGF-binding protein 7 | Prostacyclin-stimulating factor | MAC25 protein | RAMSVPS | insulin-like growth factor binding protein 7 | angiomodulin | PSF	0	Retinal arterial macroaneurysm with supravalvular pulmonic stenosis	AR	614224
IHH	2q35	hedgehog family	Indian hedgehog homolog | HHG2 | HHG-2 |BDA1	2	(1) Acrocapitofemoral dysplasia (2) Brachydactyly, type A1	AR AR	607778 112500
IL6	7p15.3	IL6 superfamily	interleukin 6 (interferon, beta 2), INF Beta, HGF; HSF; B-Cell Stimulatory Factor 2, CTL Differentiation Factor, IFNB2	1	(1) Crohn disease-associated growth failure} (2) Diabetes, susceptibility to (3) Intracranial hemorrhage in brain cerebrovascular malformations, susceptibility to (4) Kaposi sarcoma, susceptibility to (5) Rheumatoid arthritis, systemic juvenile}	Multifactorial AR AD AD --------	266600 222100 108010 148000 604302
INPPL1	11q13.4	inositol-1,4,5-trisphosphate 5-phosphatase family	SHIP2 | phosphatidylinositol-3,4,5-trisphosphate 5-phosphatase 2 | 51C protein| Inositol polyphosphate phosphatase-like protein 1 | SH2 domain-containing inositol 5'-phosphatase 2 | INPPL-1 | SHIP-2	2	Opsismodysplasia	AR	258480
INSR	19p13.2	Tyr protein kinase family	Insulin Receptor, CD220 Antigen, HHF5, IR	1	(1) Diabetes mellitus, insulin-resistant, with acanthosis nigricans (2) Hyperinsulinemic hypoglycemia,	——— AD AR	610549 609968 246200

					familial, 5 (3) Leprechaunism (4) Rabson-Mendenhall syndrome	AR	262190
IRS2	13q34	IRS family	IRS-2 | insulin receptor substrate 2 | IRS-2	2	Diabetes Mellitus, Noninsulin-Dependent; Diabetes Mellitus, Type II	AD	125853
IRS4	Xq22.3	IRS family	py160 |phosphoprotein of 160 kDa | pp160 insulin receptor substrate 4	2	----------		
JAK2	9p24.1	Tyr protein kinase family	Janus kinase 2 (a protein tyrosine kinase) | Janus kinase 2 | tyrosine-protein kinase JAK2 | JAK-2 | THCYT3 | JTK10 | Janus kinase 2 |EC 2.7.10.2	1	(1) Erythrocytosis, somatic (2) Leukemia, acute myelogenous (3) Myelofibrosis, somatic (4) Polycythemia vera (5) Thrombocythemia 3 (6) {Budd-Chiari syndrome}	AD AD AD Somatic AD/Somatic AR	133100 601626 254450 263300 614521 600880
LHX4	1q25.2	LIM-homeobox gene family	LIM homeobox protein 4 |LIM/homeobox protein Lhx4 | Gsh4 | CPHD4 | LIM homeobox 4	1	Pituitary hormone deficiency, combined, 4	AD	262700
LIN28A	1p36.11	lin-28 family	ZCCHC1 | LIN-28 | lin-28 homolog (C. elegans) | zinc finger CCHC domain-containing protein 1	1	----------		
LIN28B	6q16.3	lin-28 family	lin-28B | lin-28 homolog B (C. elegans) | FLJ16517	0	----------		
MC3R	20q13.2	G protein coupled receptor superfamily	obesity quantitative trait locus | MC3 | melanocortin 3 receptor | OQTL | melanocortin receptor 3 | BMIQ9 | MC3-R |OB20	0	(1) Mycobacterium tuberculosis, protection against} (2) Obesity, severe, susceptibility to, BMIQ9}	____ Broad locus on 20q	607948 602025
MC4R	18q21.32	G protein coupled receptor superfamily	melanocortin receptor 4 | melanocortin 4 receptor | MC4-R	0	Obesity, autosomal dominant	AD	601665
MYC	8q24.21		v-myc myelocytomatosis viral oncogene homolog (avian), Transcription Factor P64, MRTL,	0	Burkitt lymphoma	Isolated cases	113970
NF1	17q11.2		NFNS | Neurofibromatosis-related protein NF-1 | neurofibromin 1 | VRNF | WSS	1	(1) Leukemia, juvenile myelomonocytic (2) Neurofibromatosis, familial spinal (3) Neurofibromatosis, type 1 (4) Neurofibromatosis-Noonan syndrome (5) Watson syndrome		607785 162210 162200 601321 193520
NIN	14q22.1		ninein | ninein (GSK3B interacting protein) | KIAA1565 |SCKL7 | Glycogen synthase kinase 3 beta-interacting protein | ninein centrosomal protein	1	Seckel syndrome 7	AR	614851
NOG	17q22	noggin family	symphalangism 1 (proximal) | noggin | SYM1 | SYNS1 |synostoses (multiple) syndrome 1	1	(1) Brachydactyly, type B2 (2) Multiple synostoses syndrome 1 (3) Stapes ankylosis with broad thumb and toes (4) Symphalangism, proximal (5) Tarsal-carpal coalition syndrome	AD AD AD AD AD	611377 186500 184460 185800 186570

OTX2	14q22.3	paired homeobox family	MCOPS5 | orthodenticle homeobox 2 | orthodenticle homolog 2 (Drosophila) | homeobox protein OTX2 |CPHD6	0	(1) Microphthalmia, syndromic 5 (2) Pituitary hormone deficiency, combined, 6 (3) Retinal dystrophy, early-onset, and pituitary dysfunction	AD AD AD	610125 613986 610125
PAPPA	9q33.1		PAPPA1 | insulin-like growth factor-dependent IGF-binding protein 4 protease | pregnancy-associated plasma protein A, pappalysin 1	0	----------		
PCNT	21q22.3		PCNT2 | PCN | pericentrin 2 (kendrin) | KIAA0402 | pericentrin B |PCTN2 | MOPD2 SCKL4 | KEN | pericentrin-380 | PCNTB	1	Microcephalic osteodysplastic primordial dwarfism, type II	AR	210720
PIK3CB	3q22.3	PI3/PI4-kinase family	phosphoinositide-3-kinase, catalytic, beta polypeptide | PtdIns-3-kinase p110 | Phosphatidylinositol 4,5-bisphosphate 3-kinase 110 kDa catalytic subunit beta | PI3KBETA	1	----------		
PIK3CG	7q22.3	PI3/PI4-kinase family	PI3Kgamma | serine/threonine protein kinase PIK3CG | phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit gamma | PI3-kinase subunit gamma	1	----------		
PIK3R1	5q13.1	PI3K p85 subunit family	phosphoinositide-3-kinase, regulatory subunit 1 (alpha) | p85-ALPHA | phosphoinositide-3-kinase regulatory subunit	2	(1) Agammaglobulinemia 7, autosomal recessive (2) Immunodeficiency 36 (3) SHORT syndrome	AR AD AD	615214 616005 269880
PIK3R2	19p13.11	PI3K p85 subunit family	Phosphoinositide-3-Kinase, Regulatory Subunit 2 (Beta), P85, MPPH,	0	Megalencephaly-polymicrogyria-polydactyly-hydrocephalus syndrome 1	AD	603387
PIK3R3	1p34.1	PI3K p85 subunit family	p55PIK | phosphoinositide-3-kinase, regulatory subunit, polypeptide 3 (p55, gamma) |100% homology to SWISS-PROT Q92569	1	----------		
POMC	2p23.3	pomc family	NPP | corticotropin-like intermediary peptide | beta-LPH | adrenocorticotropin | ACTH	0	(1) Obesity, adrenal insufficiency, and red hair due to POMC deficiency (2) {Obesity, early-onset, susceptibility to}		609734 601665
POU1F1	3p11.2	POU transcription factor family	pituitary-specific transcription factor 1 | GHF-1 | PIT-1 | POU class 1 homeobox 1 | Pit-1 | CPHD1 | Growth hormone factor 1 | GHF1 | POU domain class 1, transcription factor 1 | POU1F1a	0	Pituitary hormone deficiency, combined, 1	AD/AR	613038
PPP1CC	12q24.11	PPP phosphatase family	protein phosphatase 1, catalytic subunit, gamma isozyme, Serine/Threonine-Protein Phosphatase PP1-Gamma Catalytic Subunit	1	----------		
PRKG2	4q21.21	ser/thr protein kinase family	cGKII | cGK2 | PRKGR2 | cGMP-dependent protein kinase II |cGMP-dependent protein kinase 2 | protein kinase, cGMP-dependent, type II | cGK 2 |EC 2.7.11.12	1	----------		
PROP1	5q35.3	paired homeobox protein family (non HOX)	Pituitary-specific homeodomain factor | PROP-1 | prophet of Pit1, paired-like homeodomain transcription factor | PROP paired-like homeobox 1 | CPHD2	0	PITUITARY HORMONE DEFICIENCY, COMBINED, 2; CPHD2	AR	262600
PTCH1	9q22.32	patched family	PTC | protein patched homolog 1	1	(1) Basal cell carcinoma, somatic (2) Basal cell nevus syndrome (3) Holoprosencephaly-7	Somatic AD AR	605462 109400 610828
PTPN1	20q13.13	protein-tyrosine phosphatase family	protein tyrosine phosphatase, placental | PTP1B | protein-tyrosine phosphatase 1B | tyrosine-protein phosphatase non-receptor type 1	0	{Insulin resistance, susceptibility to}	AD	125853

RASA2	3q23	GAP1 family of GTPase-activating proteins	GAP1m |GTPase-activating protein of RAS | GAP1M | RAS p21 protein activator 2 | ras GTPase-activating protein 2 | RASGAP	0	----------		
RBP3	10q11.22	peptidase S41A family	D10S64 | Interstitial retinol-binding protein | Interphotoreceptor retinoid-binding protein | IRBP | retinol-binding protein 3, interstitial | retinol-binding protein 3	0	Retinitis pigmentosa 66	AR	615233
RNF135	17q11.2	zinc finger family	RING finger protein 135 | REUL |E3 ubiquitin-protein ligase RNF135 | Riplet	1	Macrocephaly, macrosomia, facial dysmorphism syndrome		614192
RUNX2	6p21.1	RUNX family of transcription factor	Runt-Related Transcription Factor 2, SL3-3 Enhancer Factor 1 Alpha A Subunit, Acute Myeloid Leukemia 3 Protein, Oncogene AML-3, PEBP2A1	2	(1) Cleidocranial dysplasia (2) Metaphyseal dysplasia with maxillary hypoplasia with or without brachydactyly	AD AD	119600 156510
SCMH1	1p34.2	SCM family	polycomb protein SCMH1 | Sex comb on midleg homolog 1 | Scml3 | sex comb on midleg homolog 1 (Drosophila)	1	----------		
SHC1	1q21.3		SHC | SHCA | SHC-transforming protein 3 | SHC (Src homology 2 domain-containing) transforming protein 1 | SHC-transforming protein A | SH2 domain protein C1	0	----------		
SHOX	Xp22.33	paired homeobox family	SS | pseudoautosomal homeobox-containing osteogenic protein | Short stature homeobox-containing protein |growth control factor, X-linked | PHOG | GCFX | SHOXY	1	----------		
SLC2A1	1p34.2	sugar transporter family	hepG2 glucose transporter | Glucose transporter type 1, erythrocyte/brain | GLUT1DS | GLUT-1 | solute carrier family 2, facilitated glucose transporter member 1	0	----------		
SLC2A4	17p13.1	sugar transporter family	solute carrier family 2 (facilitated glucose transporter), member 4 | GLUT-4 | insulin-responsive glucose transporter type 4	1	----------		
SOCS2	12q22	suppressor of cytokine signaling (SOCS)	SSI-2 | Cish2 | STAT-induced STAT inhibitor-2 | STATI2 | CIS2 | SOCS-2 |CIS-2 | SSI2 | cytokine-inducible SH2 protein 2 | suppressor of cytokine signaling 2	0	----------		
SOX2	3q26.33	SRY-related HMG box family of transcription factors	transcription factor SOX-2 | transcription factor SOX2 | SRY-related HMG-box gene 2 | SRY (sex determining region Y)-box 2 | MCOPS3 |ANOP3	0	(1) Microphthalmia, syndromic 3 (2) Optic nerve hypoplasia and abnormalities of the central nervous system	AD AD	206900 206900
SPRED1	15q14	sprouty family	sprouty-related, EVH1 domain containing 1 | FLJ33903 | suppressor of Ras/MAPK activation | Spred-1 | hSpred1 | NFLS	1	Legius syndrome	AD	611431
STAT1	2q32.2	transcription factor STAT family	STAT91 |CANDF7 | ISGF-3 | transcription factor ISGF-3 components p91/p84 | signal transducer and activator of transcription 1-alpha/beta | Transcription factor ISGF-3 components p91/p84	0	(1) Candidiasis, familial, 7 (2) Immunodeficiency 31A, mycobacteriosis, autosomal dominant (3) Immunodeficiency 31B, mycobacterial and viral infections, autosomal recessive	AD AD AR	614162 614892 613796
STAT2	12q13.3	transcription factor STAT family	signal transducer and activator of transcription 2, 113kDa | interferon alpha induced transcriptional activator | p113 | STAT113 |ISGF-3	0	----------		
STAT3	17q21.2	transcription factor STAT family	HIES | signal transducer and activator of transcription 3 (acute-phase response factor) | APRF|DNA-binding protein APRF	1	(1) Autoimmune disease, multisystem, infantile-onset (2) Hyper-IgE recurrent infection syndrome	AD AD	615952 147060

STAT4	2q32.3	transcription factor STAT family	SLEB11 | signal transducer and activator of transcription 4 | SLEB11	1	Systemic lupus erythematosus, susceptitbility to, 11	AD	612253
STAT5A	17q21.2	transcription factor STAT family	MGF | STAT5 | signal transducer and activator of transcription 5A |MGF	0	----------		
STAT5B	17q21.2	transcription factor STAT family	signal transducer and activator of transcription 5B | transcription factor STAT5B |STAT5	1	Growth hormone insensitivity with immunodeficiency	------	245590
STAT6	12q13.3	transcription factor STAT family	IL-4-STAT | STAT6B |STAT, interleukin4-induced | transcription factor IL-4 STAT D12S1644 | STAT6C | signal transducer and activator of transcription 6, interleukin-4 induced | IL-4 Stat	0	----------		
TBX4	17q23.2	T box family	SPS | T-box protein 4 | T-box transcription factor TBX4 | T-box 4	1	Small patella syndrome	AD	147891
TGFA	2p13.3	epidermal growth factor family	transforming growth factor, alpha |TFGA | protransforming growth factor alpha | TGF-alpha	0	----------		
TP53	17p13.1	p53 family	tumor protein p53, Transformation-Related Protein 53, TRP53, Cellular Tumor Antigen P53	0	(1) Adrenal cortical carcinoma (2) Breast cancer (3) Choroid plexus papilloma (4) Colorectal cancer (5) Hepatocellular carcinoma (6) Li-Fraumeni syndrome (7) Nasopharyngeal carcinoma (8) Osteosarcoma (9) Pancreatic cancer (10) Basal cell carcinoma 7} (11) Glioma susceptibility 1}		202300 114480 260500 114500 114550 151623 607107 259500 260350 614740 137800
TRIP11	14q32.12		ACG1A | golgi-associated microtubule-binding protein 210 | thyroid receptor-interacting protein 11 GMAP-210 | CEV14 | TRIP-11	1	Achondrogenesis, type IA	AR	200600
TWIST1	7p21.1	BHLH family of transcription factors	Twist Basic Helix-Loop-Helix Transcription Factor 1, CRS1, Class A Basic Helix-Loop-Helix Protein 3, Acrocephalosyndactyly 3, Blepharophimosis, Epicanthus Inversus And Ptosis 3	2	(1) Craniosynostosis, type 1 (2) Robinow-Sorauf syndrome (3) Saethre-Chotzen syndrome	AD AD AD	123100 180750 101400
VDR	12q13.11	nuclear steroid/thyroid hormone receptor superfamily	vitamin D nuclear receptor variant 1 | 1,25-dihydroxyvitamin D3 receptor | Nuclear receptor subfamily 1 group I member 1 | nuclear receptor subfamily 1 group I member 1	0	(1) Osteoporosis, involutional (2) Rickets, vitamin D-resistant, type IIA	AD AR	166710 277440
WNT4	1p36.12	WNT family	protein Wnt-4 | SERKAL | wingless-type MMTV integration site family, member 4 | WNT-4	0	(1) Mullerian aplasia and hyperandrogenism (2) SERKAL syndrome	AD AR	158330 611812
WNT7A	3p25.1	WNT family	wingless-type MMTV integration site family, member 7A | proto-oncogene Wnt7a protein |protein Wnt-7a	0	(1) Fuhrmann syndrome (2) Ulna and fibula, absence of, with severe limb deficiency	AR AR	228930 276820
ZBTB38	3q23	krueppel C2H2-type zinc-finger protein family	CIBZ |zinc finger and BTB domain containing 38 | ZNF921	0	----------		

Supplementary Table 5: List of various online resources used in the current study for collecting all information/data regarding each mentioned criteria on which the genes were selected.

	Database/website name	***URL***
Basic information about the genes	GeneCards (Human gene database)	www.genecards.org
	Gene	www.ncbi.nlm.nih.gov/gene
	Ensembl genome browser	www.ensembl.org/index.html
	UCSC (University of California, Santa Cruz) genome browser	www.genome.ucsc.edu
Expression data	The Human Protein Atlas (HPA)	www.proteinatlas.org/
	UniProt	www.uniprot.org/
	Genevisible	genevisible.com/search
	GenAtlas	www.genatlas.org/
Evidence of known Syndromes/disorders in Humans	Human Phenotype Ontology	www.human-phenotype-ontology.org/
	OMIM (Online Mendelian Inheritance in Man)	www.omim.org
Animal model and/or Human/ animal disease connections	Mouse Genome Informatics (MGI)	www.informatics.jax.org
	Rat Genome Database (RGD)	rgd.mcw.edu
Known and Predicted Protein-Protein Interactions	STRING database	string-db.org
	KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway database	www.genome.jp/kegg/pathway.html
	Integrative Multi-species Prediction	http://imp.princeton.edu/
	IntAct Molecular Interaction Database	www.ebi.ac.uk/intact/
Supporting information and evidences	PubMed	www.ncbi.nlm.nih.gov/pubmed
	AmiGO2 (gene ontology browser)	amigo2.berkeleybop.org
	PathCards	pathcards.genecards.org

Supplementary Table 6: List of various in-silico/ web based prediction programs used in the current study for collecting all information/data regarding an identified variant and to predict the possible impacts of a mutation to be pathogenic or benign.

	Database/website name	***URL***
Variant filtration – Public databases	NCBI SNP database (dbSNP)	www.ncbi.nlm.nih.gov/SNP/index.html
	1000 Genomes Project (TGP)	browser.1000genomes.org/index.html
	Exome Aggregation Consortium (ExAC)	ExAC.broadinstitute.org
	ClinVar	www.ncbi.nlm.nih.gov/clinvar
	Human Genome Mutation Database (HGMD)	www.hgmd.cf.ac.uk/ac/index.php
Various in-silico/ web based prediction programs	Mutation Taster	www.mutationtaster.org
	Polyphen-2	genetics.bwh.harvard.edu/pph2
	SIFT (Sorting intolerant from tolerant)	sift.bii.a-star.edu.sg
	UMD-Predictor (Universal Mutation Database)	umd-predictor.eu/index.php
	Align GVGD (Grantham variation (GV) and Grantham deviation (GD))	agvgd.hci.utah.edu/index.php
	MutPred	mutpred.mutdb.org/about.html
	Meta-SNP	snps.biofold.org/meta-snp/index.html
	Multivariate Analysis of Protein Polymorphism (MAPP)	mendel.stanford.edu/SidowLab/downloads/MAPP/
web-based prediction tools for splicing effects	Human Splicing Finder (HSF 3.0)	www.umd.be/HSF3/
	NetGene2Server	www.cbs.dtu.dk/services/NetGene2
	splice site prediction by neural network (NNSPLICE)	www.fruitfly.org/seq_tools/splice.html
	MaxEntScan	genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html
	Genesplicer	ccb.jhu.edu/software/genesplicer

Supplementary Table 7: Complete enrichment summary report per run (24 samples in one run) of both ID and Short stature (GH) cohort

	ID_Run 1	ID_Run 2	ID_Run 3	ID_Run 4	GH_Run 1	GH_Run 2	GH_Run 3	GH_Run 4	ID_Run 5	ID_Run 6	GH_Run 5	ID_Run 7
Total PF reads:	2029182	1974176.4	1753240.6	1770438.6	1179340.76	1085909.76	1403013.6	1590537.76	1703697.92	1795732.08	1921499.24	1690550.6
Percent Q30:	91.12%	90.92%	93.46%	93.40%	96.64%	96.48%	95.13%	93.38%	84.25%	76.16%	84.67%	87.98%
Total aligned reads:	2025091.72	1969524.16	1749891.36	1766550.28	1171954.96	1079195.2	1391365.76	1573223.88	1644928.92	1472421.16	1793605.32	1574315.6
Percent aligned reads:	99.81%	99.74%	99.79%	99.78%	99.39%	99.38%	99.14%	98.84%	96.30%	81.85%	93.31%	93.14%
Targeted aligned reads:	1557079.28	1300564.56	1232568.84	1229531.96	840202.44	744081.04	1081047.12	1256682	1299271.56	1126656.92	1466802.44	1387850.04
Read enrichment:	76.63%	65.15%	70.08%	69.52%	71.92%	69.42%	77.57%	80.10%	78.48%	77.37%	82.44%	88.28%
Padded target aligned reads:	1729746.72	1484412.68	1451962.84	1436031.4	988081.52	895859	1210522.08	1381759.56	1462470.44	1257852.68	1545980.88	1420843.44
Padded read enrichment:	85.40%	74.42%	83.00%	81.27%	84.40%	83.16%	86.87%	87.88%	88.70%	86.32%	86.84%	90.28%
Total PF bases:	295083225	285199751	259673680	259880603	174537510.8	160663083	197543745	225552633	234882280	217333652	242087853	192583269
Total aligned bases:	292865667	282509496	257942633	257923261	173554906.3	159812758	195886382	222825258	227267191	193432710	226957950	178818322
Percent aligned bases:	99.25%	99.02%	99.30%	99.25%	99.44%	99.47%	99.12%	98.68%	96.54%	88.81%	93.69%	92.84%
Targeted aligned bases:	163916617	134680841	128630678	127772651	88829852.28	78238799.4	111962850	133066727	129314324	108714192	140239009	122963290
Base enrichment:	55.72%	47.04%	49.52%	49.50%	51.41%	49.41%	57.07%	60.71%	56.43%	56.93%	62.42%	69.19%
Padded target aligned bases:	241783677	204554421	202013069	198593374	138247188.3	123989365	163438027	189264189	193228671	159305698	192016852	159925493
Padded base enrichment:	82.44%	71.52%	78.18%	76.96%	79.82%	77.91%	83.32%	85.17%	84.72%	83.32%	85.36%	89.52%
Percent duplicate paired reads:	9.27%	5.76%	6.58%	6.98%	------	------	------	------	23.60%	7.81%	11.36%	13.81%
Mean region coverage depth:	202.824	166.628	159.152	158.088	109.9	96.804	138.532	164.64	160	134.52	173.528	152.136
Uniformity of coverage (Pct > 0.2*mean):	92.52%	92.90%	93.14%	92.64%	91.56%	92.20%	92.86%	90.44%	92.23%	92.14%	91.02%	88.76%
Target coverage at 1X:	99.13%	99.09%	99.10%	99.06%	98.78%	98.85%	99.06%	98.59%	98.94%	98.94%	98.93%	98.50%
Target coverage at 10X:	97.37%	96.94%	96.87%	96.94%	95.51%	95.42%	96.66%	95.02%	96.46%	96.34%	96.39%	95.04%
Target coverage at 20X:	95.76%	94.88%	94.69%	95.02%	92.06%	91.63%	94.18%	92.07%	94.12%	93.84%	94.18%	91.70%
Target coverage at 50X:	90.41%	86.85%	85.09%	87.31%	77.17%	73.99%	83.77%	82.94%	85.68%	84.50%	86.86%	81.04%
Fragment length median:	227.6	243.72	287.6	277.88	289.08	309.68	264.08	241.12	271.04	246.24	210.28	186.6
Fragment length min:	51.4	40.68	59.84	54.08	148.92	147.88	110.12	108.28	89.16	76.56	65	52.8
Fragment length max:	661.84	762.76	855.64	856.2	736.16	803.84	777	704.92	738.72	616.56	561	434.92
Fragment length SD:	102.8	123.92	137	140.68	104.68	117.88	112.92	97.16	115.64	94.56	74.6	54.52

Supplementary Table 8: ACMG criteria for classifying variants (Richards et al., 2015)

Category	Explanation	Evidence of pathogenicity
PVS1	Null variant (nonsense, frameshift, canonical +/−1 or 2 splice sites, initiation codon, single or multi-exon deletion) in a gene where loss of function (LOF) is a known mechanism of disease	Very strong
PS1	Same amino acid change as a previously established pathogenic variant regardless of nucleotide change	Strong
PS3	Well-established in vitro or in vivo functional studies supportive of a damaging effect on the gene or gene product	Strong
PM1	Located in a mutational hot spot and/or critical and well-established functional domain (*e.g.* active site of an enzyme) without benign variation	Moderate
PM2	Absent from controls (or at extremely low frequency if recessive) in Exome Sequencing Project, 1000 Genomes or ExAC	Moderate
PM4	Protein length changes due to in-frame deletions/insertions in a non-repeat region or stop-loss variants	Moderate
PM5	Novel missense change at an amino acid residue where a different missense change determined to be pathogenic has been seen before	Moderate
PM6	Assumed de novo, but without confirmation of paternity and maternity	Moderate
PP2	Missense variant in a gene that has a low rate of benign missense variation and where missense variants are a common mechanism of disease	Supporting
PP3	Multiple lines of computational evidence support a deleterious effect on the gene or gene product (conservation, evolutionary, splicing impact, etc)	Supporting
PP4	Patient's phenotype or family history is highly specific for a disease with a single genetic etiology	Supporting
BS1	Allele frequency is greater than expected for disorder	Strong evidence of benign impact
BS4	Lack of segregation in affected members of a family	
BP1	Missense variant in a gene for which primarily truncating variants are known to cause disease	Supporting evidence of benign impact
BP4	Multiple lines of computational evidence suggest no impact on gene or gene product (conservation, evolutionary, splicing impact, etc)	

Supplementary Figure 1: Roche 454 Sequencing method

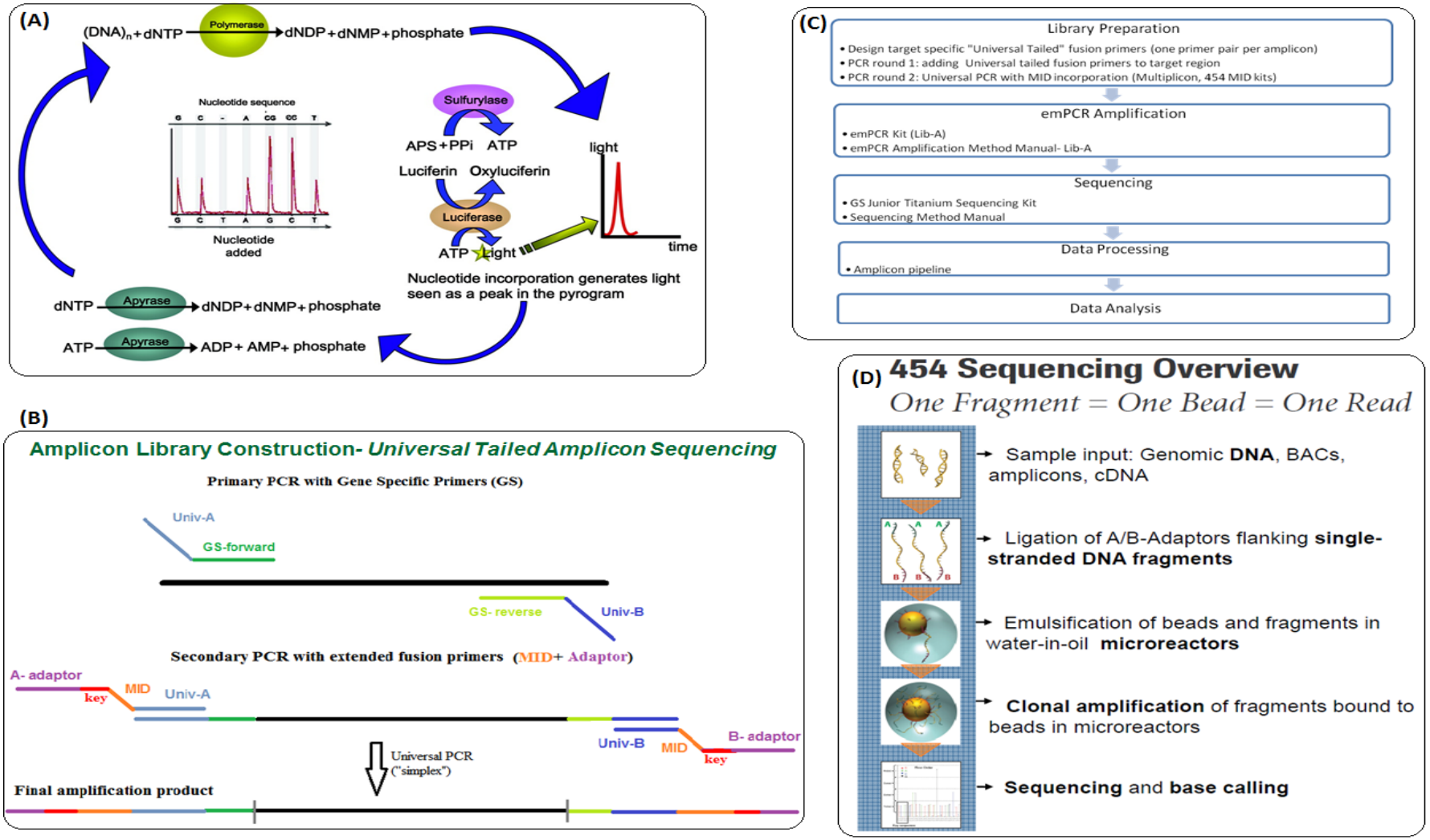
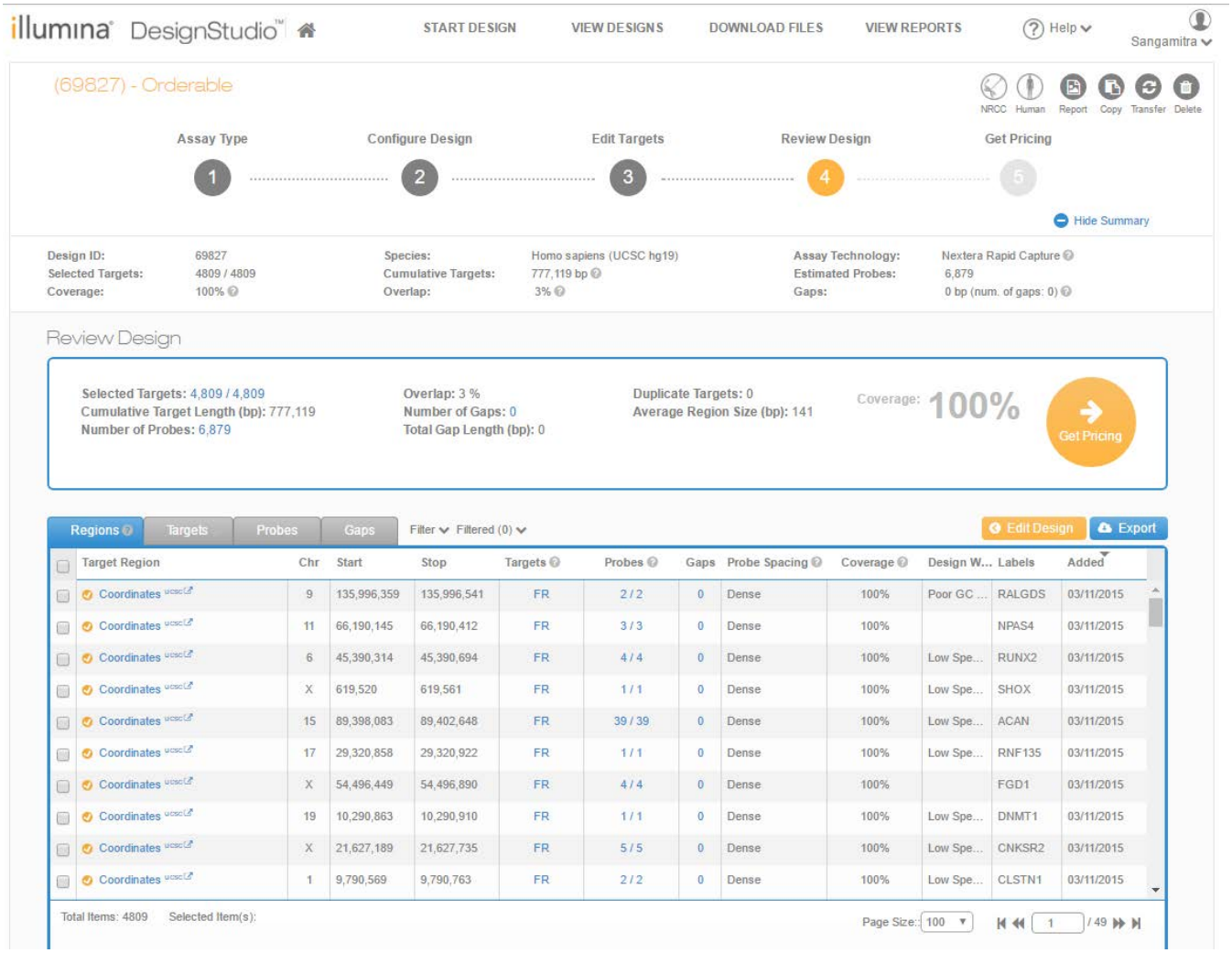


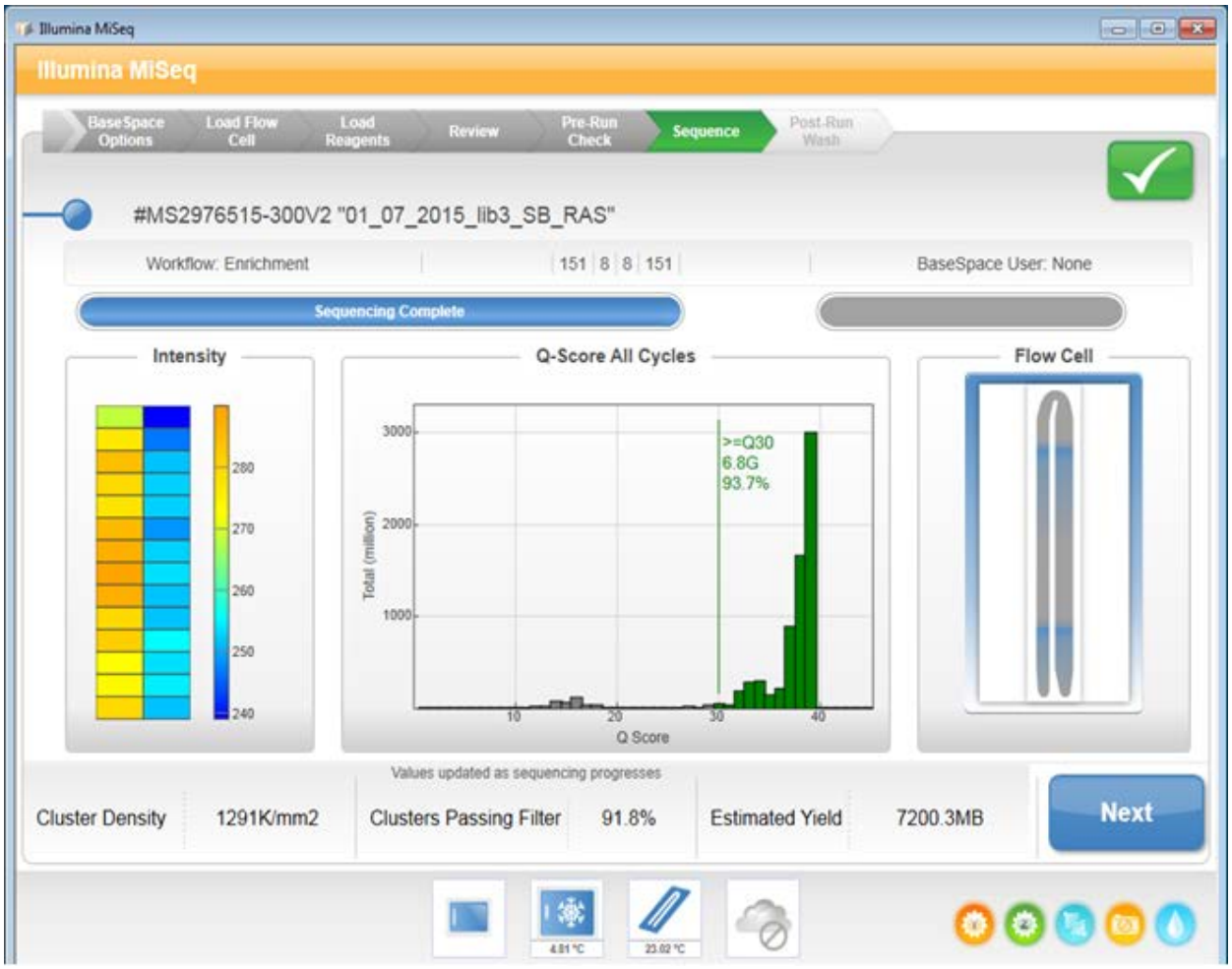
Figure A1: **(A)** Pyrosequencing chemistry: Biochemical reactions and enzymes involved in the generation of light signals by DNA pyrosequencing [adopted from Marsh S, 2007]. dNTP, deoxynucleoside triphosphate; dNDP, deoxynucleoside diphosphate; dNMP, deoxynucleoside monophosphate; PPi, pyrophosphate; APS, adenosine 5'-phosphosulfate. **(B)** Overview of the two-step amplification strategy. In the primary PCR the target sequence is amplified using amplicon specific primers flanked by universal tails. The secondary PCR introduces the patient-specific identifier (Multiplex Identifier: MID) and the 454 adaptors A and B by means of the universal tail. **(C)** Workflow of a 'Universal Tailed' amplicon sequencing experiment **(D)** Workflow of 454 sequencing overview (Figures for 454 sequencing are reproduced with the permission of Roche diagnostics and remain their copyright ©Roche Diagnostics. (http://www.gsjunior.com/))

Supplementary Figure 2: Screenshot of DesignStudio® software after completion of the target probe design of the gene panel, displaying the summary and review of the design with the regions selected.



Target Region - A target region comprises a pair of coordinates that define a continuous sequence in the genome (for example, from X to Y) for which DesignStudio® designs amplicons.
Chr – Chromosome numbers followed by start and stop position of the chromosome.
Targets - FR (Full region) targets.
Probes - Column values consist of 2 numbers divided by a forward slash (/). The first number denotes the number of attempted targets added to your design. The second number shows the total number of targets in the target region.
Gaps - Number of sub regions that DesignStudio could not cover, resulting in gaps of coverage.
Probe Spacing - The spacing of the enrichment probes within the region of interest. (Intermediate—180 bp center to center) (Dense—120 bp center to center)
Coverage - Displayed as a percentage, coverage is the total number of non-overlapping bases covered by the attempted amplicons divided by the total number of bases in design.
Design Warnings - Warnings, if any, regarding the design are mentioned here.
Labels - A text description by user applied when the target or region was added to the design followed by date the target or region was added.

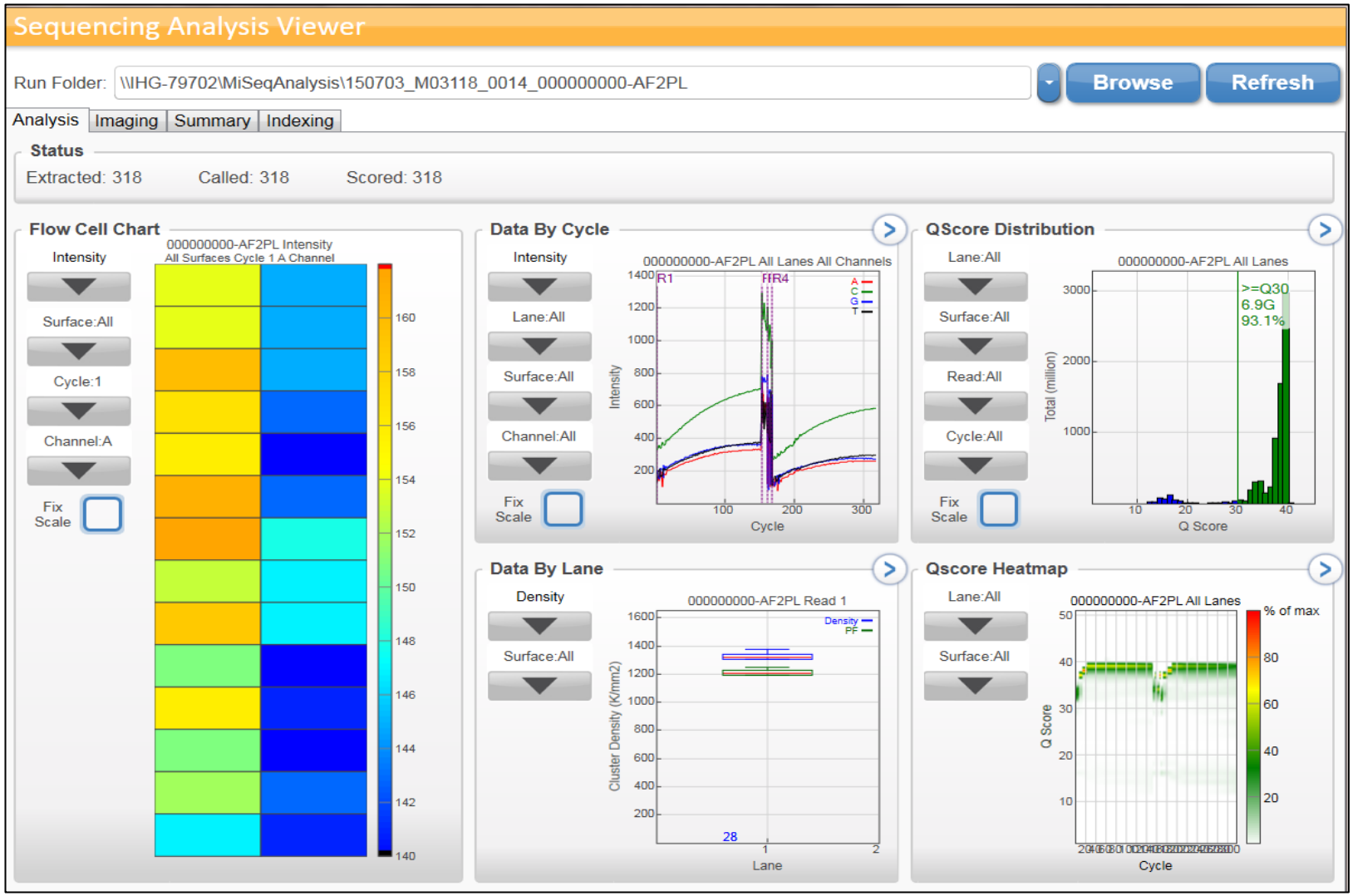
Supplementary Figure 3: Screenshot of the Illumina MCS after completion of the sequencing reaction of a single run displaying the cluster density, cluster passing filter with the Q30 value. (Lower-right corner of interface screen shows the activity indicators and base of interface screen shows Sensor indicators).



After sequencing begins, the following metrics appeared at the indicated cycles: Cycle 1-4: Intensity; Cycle 5-25: Intensity and Cluster Density; Cycle 26 through run completion: Intensity, cluster Density, %PF, Yield and Q-scores.

- **Intensity** – shows the value of cluster intensities of the 90th percentile of each tile. Here it is two columns graphic representing the flow cell image on the top and bottom surface.
- **Q-score All cycles** – shows the average percentage of bases greater than Q30 (quality score measurement). A Q-score is a prediction of the probability of a wrong base call. A Q30 score is the error probability of 1 in 1000 wrong call of a base.
- **Cluster Density (K/mm^2)** – Shows the number of clusters per square millimeter on the flow cell for each run.
- **Clusters Passing Filter (%)** - Shows the percentage of clusters passing filter based on the default chastity filter settings giving a measure of quality.
- **Estimated Yield (MB)** – Shows the projected number of bases called for the run measured in megabases.

Supplementary Figure 4: Screenshot of the Illumina SAV- Analysis tab after completion of the sequencing reaction of a single run displaying the flow cell intensity chart, plots showing the data by lane and cycle, Q score distributions plot with the Q30 value and Q score heatmap.



The Flow Cell Chart shows color-coded quality metrics per tile for the entire flow cell.

The Data by Cycle pane shows plots that allow you to follow the progression of quality metrics during a run.

The Data by Lane pane shows plots that allow you to view quality metrics per lane.

The Q-score Distribution pane shows plots that allow you to view the number of reads by quality score.

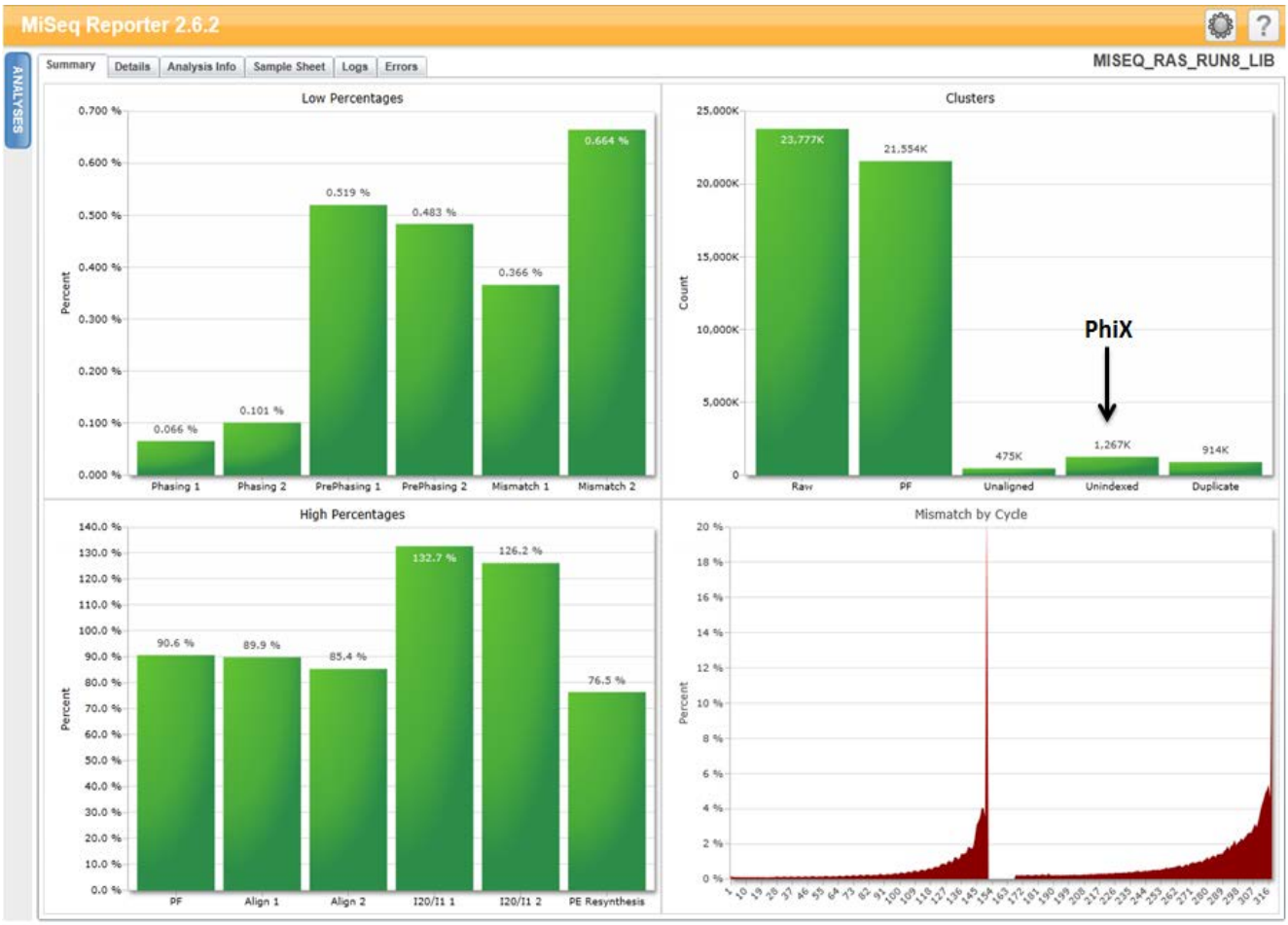
The Q-score heat map shows plots that allow you to view the Q-score by cycle.

The Imaging tab lists detailed data and metrics for the run.

The Summary tab leads to tables with basic data quality metrics summarized per lane and per read.

The Indexing tab lists count information for indexes used in the run.

Supplementary Figure 5: Screenshot of the Illumina MSR- Summary tab of a single run displaying the Percentage charts, plots showing the cluster density count, and mismatch by cycle chart.



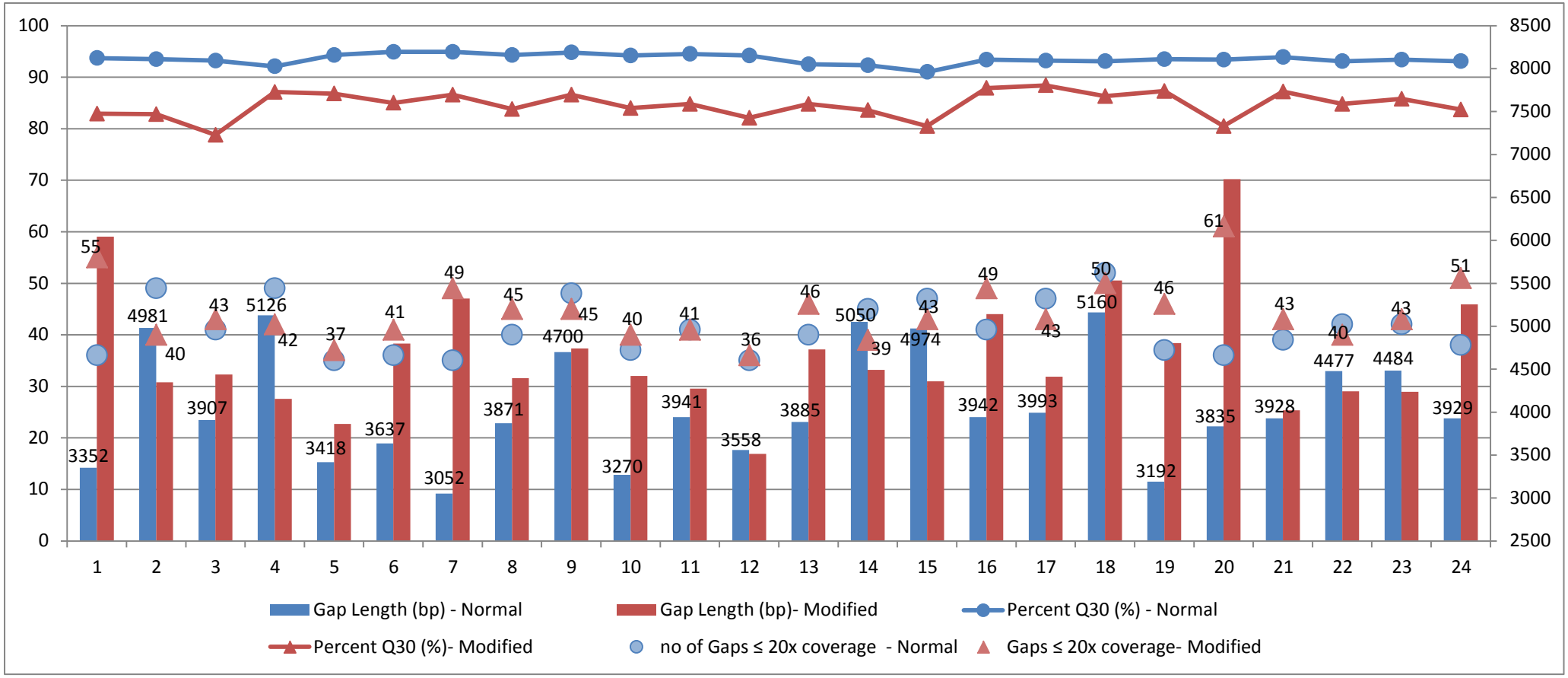
Low percentages graph - shows phasing, pre-phasing and mismatches in percentages. Low percentages graph indicates good run statistics.

High percentages graph - shows clusters passing filters, alignment to the references, and intensities in percentages. High percentages graph indicates good run statistics.

Clusters graph - Shows number of raw clusters, clusters passing filters, clusters that did not align, clusters that are not associated with the index (PhiX) and duplicates.

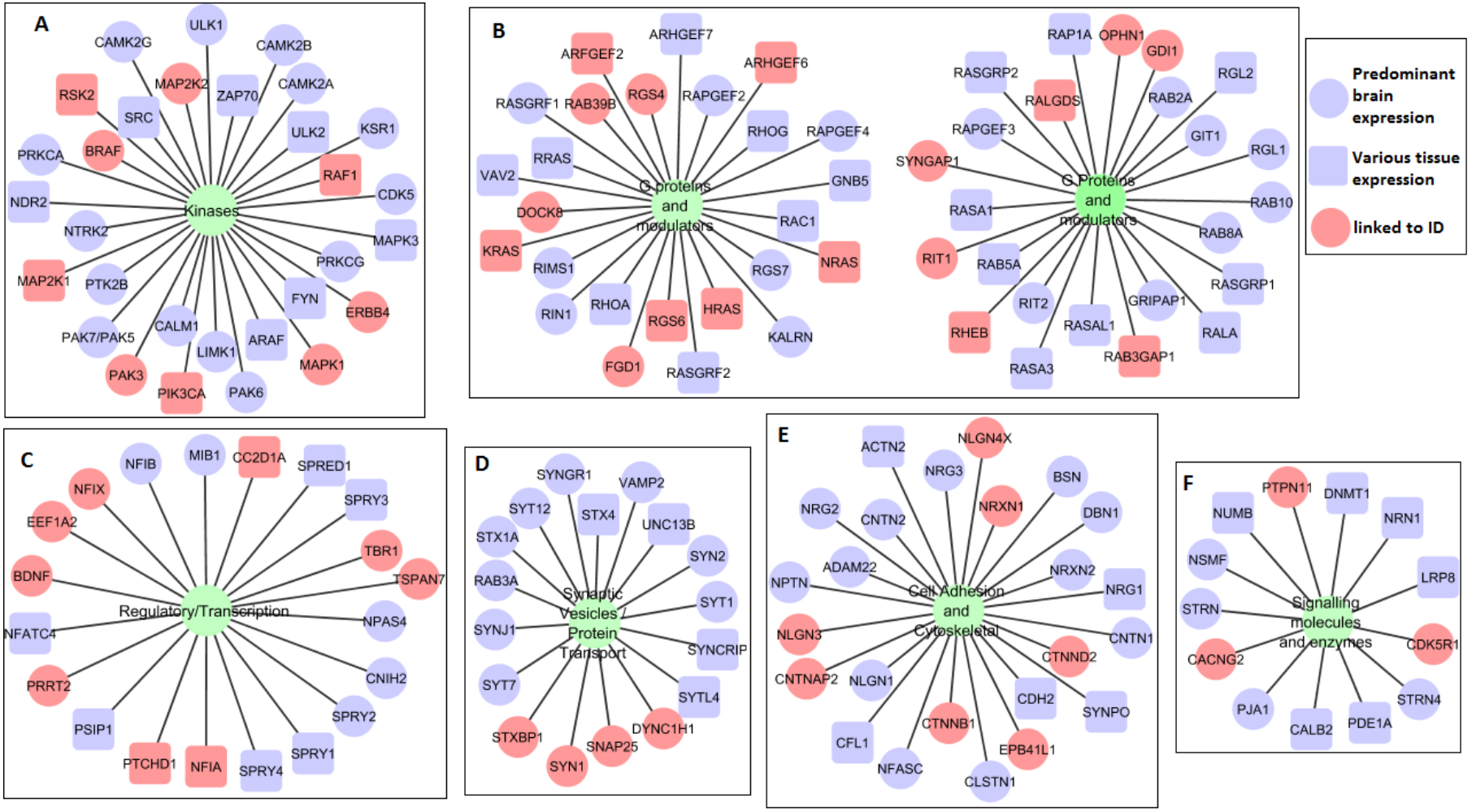
Mismatch by cycle graph - plots the mismatches between a sequence read and a reference genome after alignment. (Cycle: Plots the % mismatches for all clusters in a run versus cycle)

Supplementary Figure 6: Comparison graph between standard and modified protocols



Summary statistics showing comparison between normal (standard) protocol and modified protocol for two individual runs performed in the current study displaying the Q-scores, number of gaps and gap length in bp along for each corresponding sample in individual run. The parameters between normal and modified protocols are indicated in detail by color coding.

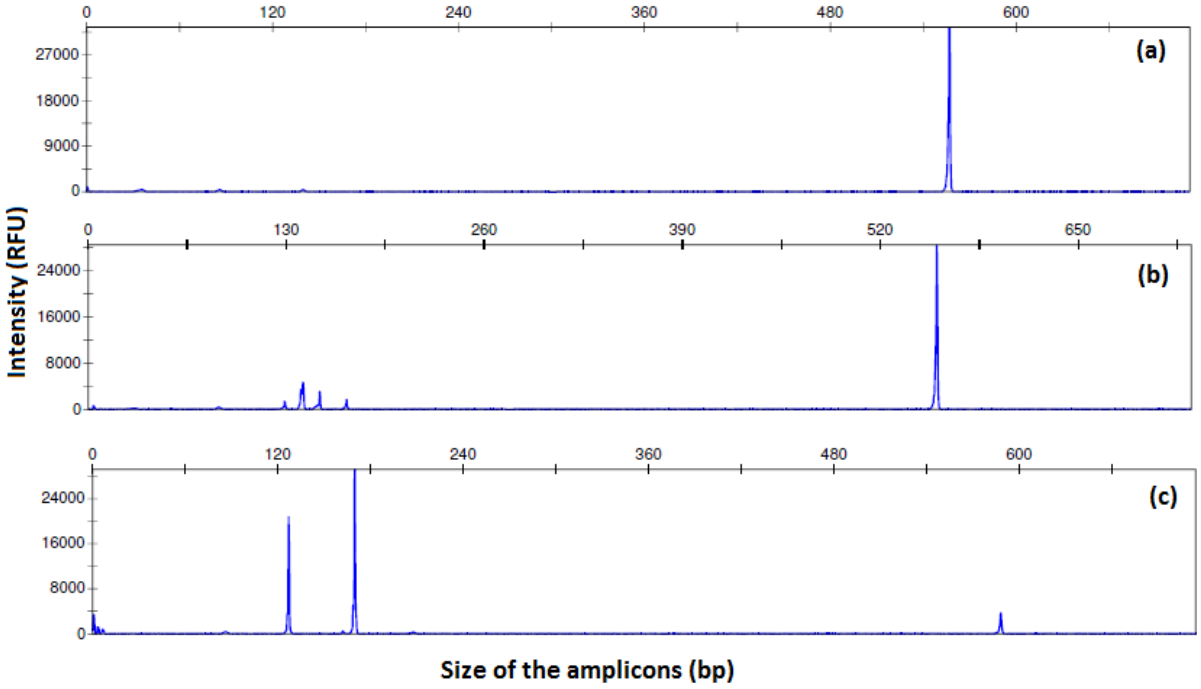
Supplementary Figure 7: Different protein classes with their corresponding genes, related to RAS/MAPK pathway, classified according to their function in the current project (prepared by using Cytoscape software, version 3.3)



The 221 selected genes related to the RAS pathway were broadly divided into protein classes according to their function – (A) Kinases (30 genes); (B) G Proteins and modulators (49 genes); (C) regulatory / transcription factors (20 genes); (D) Synaptic Vesicles / Protein Transport (17 genes); (E) Cell Adhesion and Cytoskeleton (24 genes) and (F) Signalling molecules and Enzymes (13 genes).

Supplementary Figure 8: Fragment Analysis

The fragment analysis served as an important quality measure in the NGS runs in the samples tested for PIK3CA-related overgrowth study. Two independent fragment analyses runs were done on all the amplicons. Each amplicon was analysed individually before pooling, checking for its amplification at the desired length. The presence of short fragments was also detected for each amplicon which helped in further assessing for either additional purification step of the amplicon or its removal from the NGS run. In the second NGS run, amplicons with shorter fragments especially FFPE samples were removed which improved the run quality compared to the first.



Example of fragment analysis method showing the amplicons for three different samples.

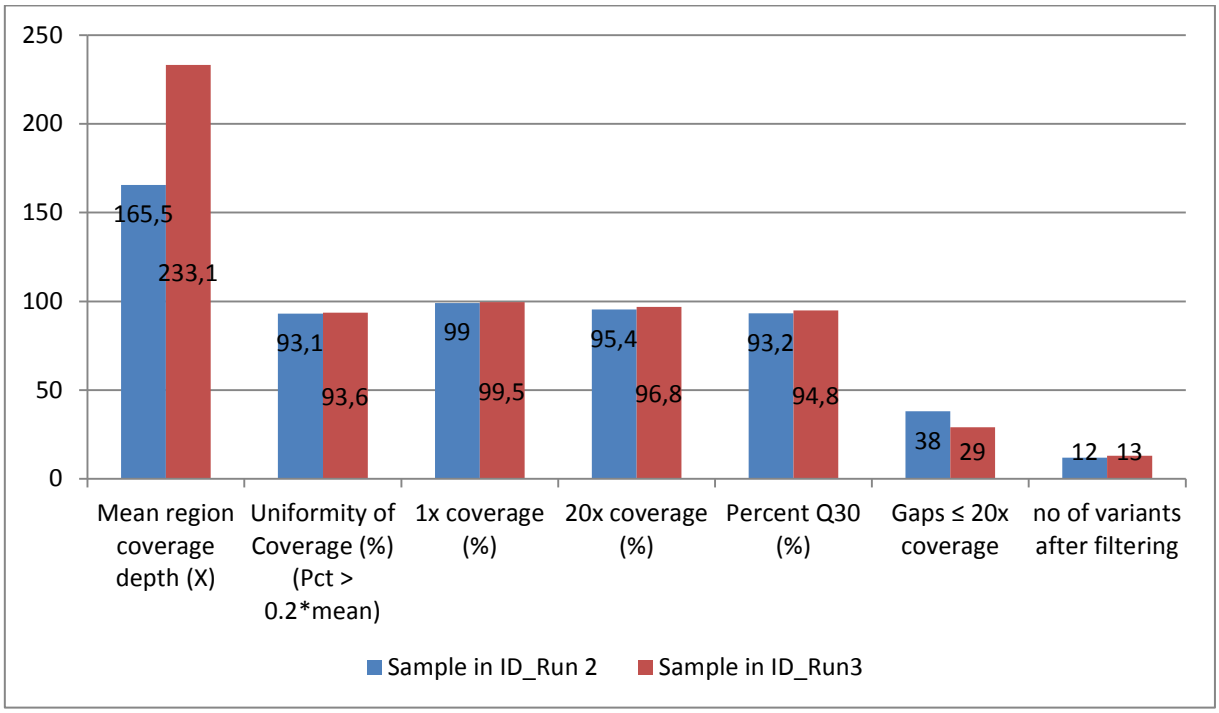
(a) Sample number 27 (Skin) with desired peak size of 556 bp (Exon 21).

(b) Sample number 45 (FFPE- Tumor tissue) with desired peak size of 556 bp (Exon 21). The FFPE sample shows the presence of short fragments around 150bp and also reduced intensity in peak size.

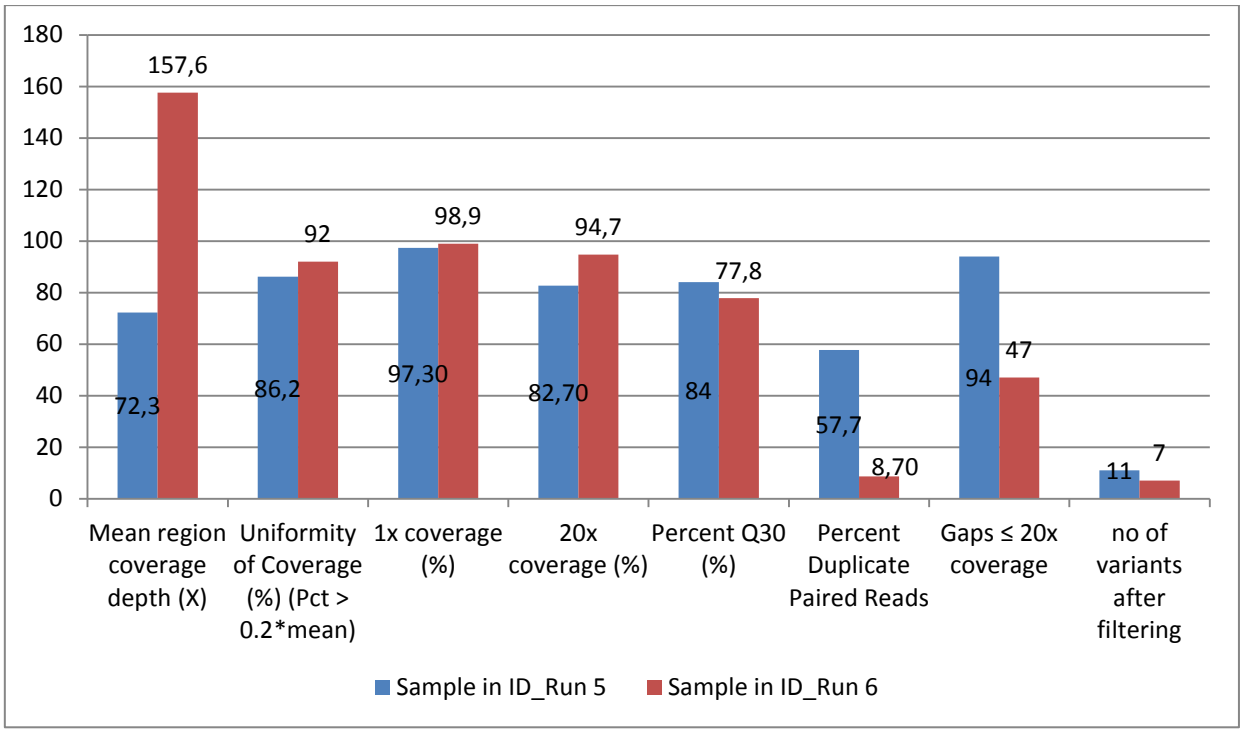
(c) Sample number 20 (Cartilage) with desired peak size of 587 bp (Exon 10). The sample was excluded from NGS run2 due to the presence of high short fragments at around 150 bp than the desired amplicon.

Supplementary Figure 9: Reproducibility of the Nextera® Rapid Capture Enrichment (NRCE) assay.

(a) Example showing Reproducibility of the **standard** NRCE assay.
Reproducibility of the standard NRCE assay has been tested by using the same sample in two different runs under the same assay conditions and produced identical results.



(b) Example showing the same sample repeated in the **modified** NRCE assay.
Reproducibility of the modified NRCE assay has been tested by using the same sample in two different runs under the same assay conditions. In run 5, the sample was poorly covered with high percentage of paired duplicate reads which on repetition in different run with the same modified protocol produced promising results.



Appendix I: DNA Quantification – Promega QuantiFluor®-ST Method

Warm all assay components to room temperature before use. The QuantiFluor® dsDNA Dye is dissolved in 100% DMSO and frozen at or below 4°C. Prior to dilution, thaw dye at room temperature, protected from light.

- Prepare 1X TE buffer by diluting the 20X TE Buffer, 20-fold with nuclease-free water. For example, add 1ml of 20X TE Buffer to 19ml of Nuclease-Free Water and mix.
- Prepare the QuantiFluor® dsDNA Dye working solution by diluting the 200X QuantiFluor® dsDNA Dye with 1X TE buffer.
- Use the following worksheet to determine the volume of Dye working solution for your quantitation assay:

Number of standard samples _____ + Number of unknown samples _____ + 4 (for pipetting error) = _____ × 50µl = Volume of QuantiFluor® dsDNA Dye working solution needed.

Mix well by vortexing and store in the dark.

- In new 1.5ml tubes, dilute the DNA samples in TE buffer to ≤1ng/µl, since the maximum measurement capacity of the QuantiFluor® is 1ng/µl. Mix well by vortexing for ~1min.

For example: if the DNA conc is around 300-400ng/µl, then a 1:1000 dilution is made (1µl sample + 999µl TE).

- In new 1.5ml tube, dilute the λ standard DNA sample in TE buffer to 1:1100 dilution (since the standard value is 900). (1µl λ standard DNA sample + 1100µl TE).
- Mix each unknown or standard sample with an equal volume of QuantiFluor® dsDNA Dye working solution prepared in step 1.
- In new 1.5ml brown tubes, add 50µl of diluted sample and to it add 50µl of QuantiFluor® Dye working solution (prepared in Step 1) to each tube containing unknown or standard sample, and mix briefly by pipetting.
- Incubate assays for 5 minutes at room temperature, protected from light
- Calibration of the device has to be done with every new experiment. One calibrated the samples are measured.
- Measure samples in triplicates individually. Place the cuvette with 100µl sample & press read 3 times, noting each time its value. Then an average of the 3 values is calculated, noted and compared with the NanoDrop measurements.

Appendix II: DNA Quantification – Qubit® 3.0 Fluorometer Method

Warm all assay components to room temperature before use. Prior to dilution, thaw dye at room temperature, protected from light.

- Set up two Assay Tubes for the standards and one tube for each user sample.
- Prepare the Working Solution by diluting the reagent 1:200 in buffer. Prepare 200 µl of Working Solution for each standard and sample.

- Prepare the Assay Tubes* according to the table below.

	Standard Assay Tubes	User Sample Assay Tubes
Volume of Working Solution (from step 2) to add	190 µl	199 µl
Volume of Standard (from kit) to add	10 µl	—
Volume of User Sample to add	—	1 µl
Total Volume in each Assay Tube	200 µl	200 µl

* Use only thin-wall, clear 0.5 mL PCR tubes.

- Vortex all tubes for 2–3 seconds.
- Incubate the tubes for 2 minutes at room temperature
- Insert the tubes in the Qubit® 2.0 Fluorometer and take readings.
- Measure samples in triplicates individually. Place the tubes with 200µl sample & press read 3 times, noting each time its value. Then an average of the 3 values is calculated, noted and compared with the NanoDrop measurements.

Note: The same protocol is followed for both the Broad range DNA kit and High sensitivity DNA kit.

Appendix III: NRCE Index Adapter Sequences

Nextera® Rapid Capture Enrichment (NRCE) Index Adapter Sequences

A dual indexing strategy uses two 8 base indexes, Index 1 (i7) next to the P7 sequence and Index 2 (i5) next to the P5 sequence. Dual indexing is enabled by adding a unique Index 1 (i7) and Index 2 (i5) to each sample.

Index 1 (i7)	**Sequence**		**Index 2 (i5)**	**Sequence**
N701	TAAGGCGA		E502	CTCTCTAT
N702	CGTACTAG		E503	TATCCTCT
N703	AGGCAGAA		E504	AGAGTAGA
N704	TCCTGAGC		E505	GTAAGGAG
N705	GGACTCCT		E506	ACTGCATA
N706	TAGGCATG		E507	AAGGAGTA
N707	CTCTCTAC		E508	CTAAGCCT
N708	CAGAGAGG		E517	GCGTAAGA
N709	GCTACGCT			
N710	CGAGGCTG			
N711	AAGAGGCA			
N712	GTAGAGGA			

- N refers to Nextera®
- E refers to enrichment
- 7 refers to Index 1 (i7)
- 5 refers to Index 2 (i5)
- 01–12 refers to the Index number

Appendix IV: Additional Gene panels

The total number of genes is further divided into two panels. Second gene panel contains the list of genes shown to have at least one monogenic disorder in humans. Third gene panel contains list of genes which have no monogenic disorder described yet in humans.

		Second Panel genes - 106						
ACAN	CDK5R1	FGFR4	HESX1	INSR	MYC	POU1F1	RIMS1	SYNJ1
ACTN2	CNKSR1	GATA1	HHIP	IRS2	NFIA	PRKCA	RUNX2	TBR1
ADAMTS17	CNKSR2	GDF5	HMGA1	JAK2	NIN	PRKCG	SPRY2	TBX4
ADIPOQ	CNTN1	GHR	HMGA2	KALRN	NOG	PROP1	SHOX	TP53
AKT2	CNTN2	GHRHR	HOMER2	LHX4	NRG1	PTCH1	SPRY4	TRIP11
AR	CYP19A1	GHSR	IGF1R	LRP8	NSMF	PTPN1	SRC	VDR
BMP2	DNMT1	GLI2	IGF2	LRRC7	NTRK2	RALGDS	STAT1	WNT4
BMP4	DYM	GRB10	IGFALS	MAPK1	OTX2	RASA1	STAT2	WNT7A
CALM1	EFEMP1	GRB2	IGFBP7	MC3R	PCLO	RASGRP2	STAT3	YWHAE
CAV1	ESR1	GRIA2	IHH	MC4R	PCNT	RBP3	STAT4	ZAP70
CDK4	FBLN5	GRID2	IL6	MIB1	PIK3R1	RGS4	STAT5B	
CDK5	FGF3	GRM5	INPPL1	MPDZ	POMC	RGS6	SYN2	

		Third panel - 147 genes					
ADAM22	CNIH2	GRIK3	LIN7A	PAPPA	RAPGEF3	SHC1	SYT7
ADAMTSL3	CYR61	GRIK4	MAGI2	PDE1A	RAPGEF4	SHC2	SYTL4
AKAP5	DBN1	GRIK5	MAPK3	PIK3CB	RASA3	SIPA1L1	TGFA
ANAPC13	DLG1	GRIPAP1	NETO1	PIK3CG	RASAL1	SLC2A4	TRPC5
ARAF	DLG4	GRM2	NETO2	PIK3R3	RASGRF1	SOCS2	ULK1
ARHGEF7	DLGAP1	GRM3	NFASC	PJA1	RASGRF2	SPRY1	ULK2
ARRB1	DOT1L	GRM4	NFATC4	PPP1CC	RASGRP1	SPRY3	UNC13B
BAIAP2	FGF2	GRM7	NFIB	PRKG2	RGL1	STAT5A	VAMP2
BMP6	FOS	GRM8	NLGN1	PSIP1	RGL2	STAT6	VAV2
BSN	FUBP3	HOMER1	NPAS4	PTK2B	RGS7	STK38L	VIPR2
CABP1	FYN	IGFBP1	NPTN	RAB10	RHOA	STRN	YWHAB
CALB2	GAB1	IGFBP3	NRG2	RAB2A	RHOG	STRN4	YWHAH
CAMK2A	GAB2	IGFBP4	NRG3	RAB3A	RIN1	STX1A	YWHAZ
CAMK2B	GIT1	IRS4	NRN1	RAB5A	RIT2	STX4	ZBTB38
CAMK2G	GNB5	KCNA4	NRXN2	RAB8A	RRAS	SYNCRIP	
CDH2	GPRIN1	KSR1	NUMB	RAC1	SCMH1	SYNGR1	
CDKN1A	GRIA1	LIMK1	PACSIN1	RALA	SDCBP	SYNPO	
CFL1	GRIA4	LIN28A	PAK6	RAP1A	SEMA4C	SYT1	
CLSTN1	GRIK1	LIN28B	PAK7	RAPGEF2	SHARPIN	SYT12	

Glossary

Candidate gene - A gene that has been selected on the basis of a perceived match between the known or presumed function of the gene and the biologic characteristics of the disease in question

De novo mutation – Any DNA sequence change that occurs during replication, such as a heritable gene alteration occurring in a family for the first time as a result of a DNA sequence change in a germ cell or fertilized egg

Transoposomes - Transposomes integrate into genomic DNA as they have free DNA ends and insert randomly into DNA in a 'cut and paste' reaction.

Tagmentation – The first step in library prep is the tagmentation reaction, which involves the transposon cleaving (fragmentation) and tagging of the double-stranded DNA with a universal overhang.

Massively parallel (or next generation) sequencing - DNA sequencing that harnesses advances in miniaturization technology to simultaneously sequence multiple areas of the genome rapidly and at low cost

Read - a sequence "read" refers to the data string of A, T, C, and G bases corresponding to the sample DNA or RNA.

Dot read - no incorporation of a nucleotide in a sequencing cycle

Mixed read – presence of more than one fragment or two beads in a well or two beads in an emulsion bubble

Q score – A quality score (Q-score) is a prediction of the probability of an error in base calling. Higher Q scores indicate a smaller probability of error. Lower Q scores can result in a significant portion of the reads being unusable. They may also lead to increased false-positive variant calls, resulting in inaccurate conclusions.

Uniformity of coverage (Pct > 0.2*mean) - the percentage of targeted base positions in which the read depth is greater than 0.2 times the mean region target coverage depth

Coverage - the number of times a sequenced DNA fragment (i.e., a read) maps to a genomic target. The deeper the coverage of a target region (i.e., the more times the region is sequenced), the greater the reliability and sensitivity of the sequencing assay.

Sequencing Depth - Sequencing depth corresponds to the number of nucleotides (reads) that cover a portion of a target sequence. It is related to the concept of coverage. A high sequencing depth corresponds to a high coverage rate.

Clusters - A clonal grouping of template DNA bound to the surface of a flow cell. Each cluster is seeded by a single template DNA strand and is clonally amplified through bridge amplification until the cluster has~1000copies. Each cluster on the flowcell produces a single sequencing read. For example, 10,000clusters on the flow cell would produce 10,000single reads and 20,000paired-end reads.

Flow Cell - The flow cell is the part of the sequencing device from Illumina where the sequencing process takes place. The flow cell is a slide with different lanes that are flooded with the sample DNA library that binds with the adaptor sequence to specific oligo nucleotides coupled on the flow cell. After target amplification, enabled by bride amplification, clusters of the target DNA are formed on the flow cell. During the sequencing process the sequencing chemistry is flooded sequentially over the flow cell.

Contigs - A contig is a single file that merges different fragment files in order to optimize files in as few fragments as possible. In genetics this means single (overlapping) reads are assembled to a larger sequence fragment called contig. This is the first level of higher structured sequences obtained during the sequencing raw data assembling. A variety of different software tools for assembly and contig generation exist.

Cluster Passing Filter - The software performs base calling of raw data to remove any reads that do not meet the overall quality as measured by the Illumina chastity filter. The chastity of a base call is calculated as the ratio of the brightest intensity divided by the sum of the brightest and second brightest intensities. Clusters pass filter (PF) when no more than 1 base call in the first 25 cycles has a chastity of < 0.6.

Re-Sequencing - Re-sequencing is the sequencing of genome regions or whole genomes and the subsequent alignment to a reference genome of a known member of the same species. Re-sequencing is used in order to obtain more detailed information about the genome of interest and to detect genetic variations such as mutations, SNPs, insertions, deletions, inversions or duplications.

Synaptopathy - Refers to brain disorders that have arisen from synaptic dysfunction, including neurodevelopmental (autism spectrum disorders (ASD), intellectual disability (ID), Fragile X syndrome (FXS), Down Syndrome, attention deficit hyperactivity disorder (ADHD), and epilepsy) and neuropsychiatric disorders (bipolar disorder (BPD), schizophrenia (SCZ), and major depressive disorder (MDD)) and neurodegenerative diseases (Alzheimer's disease (AD), Huntington's disease (HD), and Parkinson's disease)

PTVs - Genetic variants predicted to shorten the coding sequence of genes – termed protein-truncating variants (PTVs) – are typically expected to have large effects on gene function. PTVs are defined as single nucleotide variants (SNVs) predicted to introduce a premature stop codon or to disrupt a splice site, small insertions or deletions (indels) predicted to disrupt a transcript’s reading frame, and larger deletions that remove the full protein coding sequence (CDS).

Curriculum vitae

Name : Sangamitra Boppudi
Date of birth : 16 July 1986
Place of Birth : Visakhapatnam, Andhra Pradesh, India
Nationality : Indian

Educational profile:

03/2012 – 12/2017	**Doctoral study (PhD) in Neurobiology** Institute of Human Genetics, University Hospital Magdeburg, Otto-von Guericke-University, Magdeburg and Graduate scholar of Leibniz Institute for Neurobiology, Magdeburg (LGS Synaptogenetics)
10/2009 – 01/2012	**Master of Science (M.Sc.) Integrative Neuroscience** Otto-von Guericke-University, Magdeburg (Germany) Master Thesis: "Mutation screening of new candidate genes for mental retardation by next generation sequencing".
06/2007 – 04/2009	**Master of Science (M.Sc.) Human Genetics** Andhra University, Visakhapatnam (India) Master Thesis: "Genotyping of short tandem repeats of genomic DNA isolated from swabs and stains of sexual assault cases".
06/2004 – 03/2007	**Bachelor of Science (B.Sc.) (Chemistry, Biotechnology, Microbiology)** Dr.L.Bullayya College, affiliated to Andhra University, Visakhapatnam (India)
06/2001 – 03/2003	**XII std- Board of Intermediate Education (Biology, Chemistry, Physics)** Sri Chaitanya Junior College, Visakhapatnam (India)

Poster Presentations

European Human Genetics Conference 2014 Milan, Italy, 2014. Title: Detection of PIK3CA somatic mutations in CLOVES syndrome.

European Human Genetics Conference 2012 Nürnberg, Germany, 2012. Title: Mutation screening of new candidate genes for mental retardation by next generation sequencing.

International Symposium on Genetic and Epigenetic Basis of Complex Diseases Hyderabad, India, 2009, organized by CCMB (centre for cellular and molecular biology), Hyderabad (India). Title: Prevalence of single nucleotide polymorphism A98V of hepatocyte nuclear factor 1 alpha in type 2 diabetes

List of Publications

Hauer NN, Sticht H, **Boppudi S**, Büttner C, Kraus C, Trautmann U, Zenker M, Zweier C, Wiesener A, Jamra RA, Wieczorek D, Kelkel J, Jung AM, Uebe S, Ekici AB, Rohrer T, Reis A, Dörr HG, Thiel CT. Genetic screening confirms heterozygous mutations in ACAN as a major cause of idiopathic short stature (2017). Sci Rep. 2017 Sep 22; 7(1):12225.

Louati R, Bouayed N, **Boppudi S**, Zenker M, Rebai T. Short fragment approach for genotyping KRAS and BRAF genes in Tunisian patients with colorectal cancer (2017). Int J Clin Exp Med 2017; 10(3):5160-5167.

Boppudi S, Bögershausen N, Hove HB, Percin EF, Aslan D, Dvorsky R, Kayhan G, Li Y, Cursiefen C, Tantcheva-Poor I, Toft PB, Bartsch O, Lissewski C, Wieland I, Jakubiczka S, Wollnik B, Ahmadian MR, Heindl LM, Zenker M. Specific mosaic KRAS mutations affecting codon 146 cause oculoectodermal syndrome and encephalocraniocutaneous lipomatosis (2016). Clin Genet. 2016 Oct; 90(4):334-42.

Sangamitra B, Lakshmi V and Bhargavi A. Prevalence of single nucleotide polymorphism A98V of hepatocyte nuclear factor 1 alpha in type 2 diabetes (2010). Journal of Cytology and Genetics 2010. Vol. 11; 43-48.

A Bhargavi, **B Sangamitra** and V Lakshmi. Prevalence of mitochondrial mutation A3242G in type-2diabetics in Visakhapatnam (2009). Indian Journal of Physical Anthropology and Human Genetics, 2009; Vol 28, 213-218.

Erklärung

Hiermit erkläre ich, dass ich die von mir eingereichte Dissertation zum dem Thema " **About the impact of altered RAS-MAPK and PI3K-AKT signalling in human developmental disorders**" selbständig verfasst, nicht schon als Dissertation verwendet habe und die benutzten Hilfsmittel und Quellen vollständig angegeben wurden.

Weiterhin erkläre ich, dass ich weder diese noch eine andere Arbeit zur Erlangung des akademischen Grades doctor rerum naturalium (Dr. rer. nat.) an anderen Einrichtungen eingereicht habe.

Ort, Datum (Sanga Mitra Boppudi)