Folkhälsan Research Center
Department of Medical and Clinical Genetics,
Medicum

Faculty of Biological and Environmental Sciences
Doctoral Programme in Integrative Life Science
University of Helsinki

Improving CNV detection from short-read MPS data in neuromuscular disorders

Salla Välipakka

Doctoral thesis, to be presented for public examination with the permission of the Faculty of Biological and Environmental Sciences of the University of Helsinki, in Lecture Hall 1, Haartman Institute, on the 4th of September, 2020 at 13 o'clock.

Thesis Supervisors

Professor Bjarne Udd, MD, PhD
Folkhälsan Research Center and Department of Medical and Clinical Genetics, Medicum, University of Helsinki, Helsinki, Finland
Neuromuscular Research Unit, Department of Neurology, University Hospital and University of Tampere, Tampere, Finland
Department of Neurology, Vaasa Central Hospital, Vaasa, Finland

Docent Peter Hackman, PhD
Folkhälsan Research Center and Department of Medical and Clinical Genetics, Medicum, University of Helsinki, Helsinki, Finland

Thesis Committee

Mari Kaunisto, PhD
Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Helsinki, Finland

Docent Samuel Myllykangas, PhD
Blueprint Genetics, Helsinki, Finland
Institute of Biomedicine, University of Helsinki, Helsinki, Finland

Reviewers

Professor Matti Nykter, PhD
Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland

Docent Csilla Sipeky, PhD
Institute of Biomedicine, University of Turku, Turku, Finland

Opponent

Professor Paweł Stankiewicz, MD, PhD
Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, USA

Custos

Professor Juha Partanen, PhD
Molecular and Integrative Biosciences Research Programme, Faculty of Biological and Environmental Sciences, University of Helsinki, Helsinki, Finland

The Faculty of Biological and Environmental Sciences uses the Urkund system (plagiarism recognition) to examine all doctoral dissertations.

Dissertationes Scholae Doctoralis Ad Sanitatem Inevstigandam Universitatis Helsinkiensis 55/2020

ISBN 978-951-51-6389-9 (print) ISSN 2342-3161 (print)
ISBN 978-951-51-6390-5 (online) ISSN 2342-317X (online)
http://ethesis.helsinki.fi

Cover image: Duplication of PMP22 and deletion in DMD against normal samples as illustrated with CoNIFER, and code from scripts developed in this study.

Unigrafia Oy
Helsinki 2020

“All sorts of things can happen when you’re open to new ideas and playing around with things.”
Stephanie Kwolek

To my family

CONTENTS

LIST OF ORIGINAL PUBLICATIONS

This thesis is based on the following original publications:

I. Välipakka, S., M. Savarese, M. Johari, L. Sagath, M. Arumilli, K. Kiiski, A. Saenz, A.L. de Munain, A.M. Cobo, K. Pelin, B. Udd, and P. Hackman. 2017. Copy number variation analysis increases the diagnostic yield in muscle diseases. *Neurol.Genet.* 3:e204.

II. Välipakka, S., M. Savarese, L. Sagath, M. Arumilli, T. Giugliano, B. Udd, and P. Hackman. 2020. Improving Copy Number Variant Detection from Sequencing Data with a Combination of Programs and a Predictive Model. *J.Mol.Diagn.* 22:40-49.

III. Savarese, M., S. Välipakka, M. Johari, P. Hackman, and B. Udd. 2020. Is Gene-Size an Issue for the Diagnosis of Skeletal Muscle Disorders? *J.Neuromuscul Dis.* 7(3):203-216.

The publications are referred in the text by their Roman numerals. Unpublished data are also presented.

AUTHOR CONTRIBUTIONS

I.	Conceptualization	**SV**, MS, MJ,
	Data acquisition	**SV**, MS, MJ, LS, MA, KK
	Data analysis tool development	**SV**, MS, MA
	Formal analysis	**SV**, MS, MJ, LS, MA, KK, AS, AL, AC
	Writing – original draft	**SV**
	Writing – review & editing	**SV**, MS, MJ, LS, MA, KK, KP, BU, PH
	Project administration	KP, BU, PH
II.	Conceptualization	**SV**, MS
	Data acquisition	**SV**, LS, TG, MA
	Data analysis tool development	**SV**, MS, LS, TG, MA
	Formal analysis	**SV**, LS, TG
	Writing – original draft	**SV**
	Writing – review & editing	**SV**, MS, LS, TG, BU, PH
	Project administration	BU, PH
III.	Conceptualization	MS, **SV**, MJ, BU, PH
	Data curation	MS, **SV**, MJ
	Writing – original draft	MS, **SV**, MJ
	Writing – review & editing	MS, **SV**, MJ, PH, BU
	Project administration	MS, PH, BU

ABBREVIATIONS

ACMG = American College of Medical Genetics
ALS = amyotrophic lateral sclerosis
ASD = autism spectrum disorder
BAC = bacterial artificial chromosome
BAM = binary alignment/map
BIR = break-induced replication
BMD = Becker muscular dystrophy
bp = base pair
BWA = Burrows-Wheeler Aligner
CBS = circular binary segmentation
cDNA = complementary DNA
CGH = comparative genomic hybridization
CHN = congenital hypomyelinating neuropathy
CMT = Charcot-Marie-Tooth disease
CNM = congenital centronuclear myopathy
CNV = copy number variant
cPAL = probe-anchor ligation
cPAS = probe-anchor synthesis
ddPCR = droplet digital PCR
dHMN = distal hereditary motor neuropathy
DMD = Duchenne muscular dystrophy
DNA = deoxyribonucleic acid
dNTP = deoxyribonucleotide triphosphate
dSMA = distal spinal muscular atrophy
emPCR = emulsion PCR
FISH = fluorescent in situ hybridization
FoSTeS = fork-stalling and template-switching
FSHD = facioscapulohumeral dystrophy
GATK = Genome Analysis Toolkit
GOF = gain of function
HI = haploinsufficiency index
HMERF = hereditary myopathy with early respiratory failure
HMM = hidden Markov model
HNPP = hereditary neuropathy with liability to pressure palsies
HSAN = hereditary sensory and autonomic neuropathy
HSP = hereditary spastic ataxia
indel = insertion/deletion
IPN = inherited peripheral neuropathy
IQR = interquartile range
kb = kilo base pair (1,000 bp)
LCR = low-copy repeat
LGMD = limb-girdle muscular dystrophy
LOF = loss of function
LOH = loss of heterozygosity
MAF = minor-allele frequency
Mb = mega base pair (1,000,000 bp)
MLPA = multiplex ligation-dependent probe amplification
MMBIR = microhomology-mediated break-induced replication
MMEJ = microhomology-mediated end-joining mechanism
MND = motor neuron disease
MPS = massively parallel sequencing
mRNA = microRNA
mtDNA = mitochondrial DNA
NAHR = non-allelic homologous recombination
NEM = nemaline myopathy
NHEJ = non-homologous end joining
NMD = neuromuscular disorders
OMIM = Online Mendelian Inheritance in Man
PAC = phage artificial chromosome
PacBio = Pacific Biosciences
PCR = polymerase chain reaction
pLI = probability of being LOF intolerant
qPCR = quantitative PCR
RNA = ribonucleic acid
RPKM = reads per kilobase per million mapped reads

SAM = sequence alignment/map
SD = segmental duplication
sIBM = sporadic inclusion body myositis
SMRT = single-molecule real-time
SNP = single nucleotide polymorphism
SNV = single nucleotide variation
SOLiD = sequencing by oligonucleotide ligation and detection
SPG = spastic paraplegia
STR = short tandem repeat
TAD = topologically associating domain
TMD = tibial muscular dystrophy
TS = triplosensitivity
UMD = Ullrich muscular dystrophy
UMI = unique molecular identifier
UTR = untranslated region
WDM = Welander distal myopathy
WES = whole exome sequencing
WGS = whole genome sequencing
VNTR = variable number tandem repeat

In addition, standard abbreviations of nucleic acids and approved symbols of human genes and proteins are used.

ABSTRACT

Neuromuscular disorders (NMD) are highly heterogenic with around 1000 reported different subtypes. Most are genetic in origin, and some 500 genes are currently identified to cause NMDs. Massively parallel sequencing (MPS) approaches have been widely used to increase the cost-effectiveness and diagnostic yield in the work-up of the genetic molecular diagnosis and to speed up the process. Copy number variants (CNVs), deletions and duplications larger than 50 base pairs, explain approximately 10% of the Mendelian disorders. No best practices pipelines have been developed yet for CNV analysis from MPS data. Therefore, the detection and verification of CNV findings has often involved complementary methods, such as array comparative genomic hybridization (array CGH), multiplex ligation-dependent probe amplification (MLPA) and quantitative PCR approaches. Recently, various CNV detection programs have been developed, but for widely different types of designated research settings, which complicates choosing the correct approach for NMDs. These individual programs have generally exhibited less than ideal sensitivity and specificity for CNV detection.

Our aim was to develop a comprehensive pipeline for the detection and annotation of CNVs with high accuracy from targeted gene panel sequencing and whole exome sequencing (WES) data of patients with NMDs. Four different CNV analysis programs were chosen for this study: CoNIFER, XHMM, ExomeDepth and CODEX. The targeted gene panel MYOcap includes 349 genes for myopathic disorders and MNDcap 302 genes for neurogenic disorders in their current panel versions. 2359 samples were sequenced with MYOcap, 942 samples with MNDcap and 262 samples with WES. This included for the targeted gene panels 24 positive control samples with previously characterized CNVs and 31 negative control samples with certain genes verified to not have CNVs. A detection sensitivity of 100% and specificity of 100% were reached for these control samples. Previously undetected CNVs from MYOcap or MNDcap sequenced samples were verified as true positive detections in 36 cases with MLPA, PCR or array CGH, and eight CNVs were verified as false positive detections. These and the positive control samples were utilized in validation of a predictive logistic regression model. *In silico* CNV generation into MYOcap sequenced samples provided 18,677 specific and 3892 unspecific CNV detections to initially train the model. The model was trained to differentiate true positive detections from false positive detections in order to increase the specificity of the CNV detection pipeline.

The advantage of using four different CNV detection programs compared to using them individually, or with any other combination, was demonstrated by CNV detection sensitivity from the set of *in silico* CNVs. The predictive model with variables from all four programs provided the highest sensitivity (96.6%) and specificity (87.5%) for predicting CNV detections correctly, indicating an accuracy of 95.5% (95% CI 87.3–99.1%). The CNV detection pipeline together with the predictive model was validated for WES samples with control samples with 235 previously characterized CNVs. For CNVs spanning at least three exons, the detection

sensitivity was 97.3% and the sensitivity of the predicative model was 99.3% after adjusting the model threshold for WES data.

The CNV annotation platform cnvScan was expanded to contain the most recent CNV population databases as well as in-house CNV databases for all the sequenced sample sets. CNV detection results were filtered by < 1% frequency with reciprocal overlap of 90% in the common CNV population databases, with both it and < 5% frequency with 50% reciprocal overlap in the in-house CNV database, and by the true positive prediction with the model. These procedures significantly decreased the workload (with 3–13% of the original CNV detections preserved) in evaluating the CNVs further regarding clinical significance. The added value, i.e. the additional diagnostic yield from CNVs for both the targeted gene panel sequenced samples and WES samples was estimated to be 1.9%. Altogether 39 final genetic diagnoses were solved with these CNV findings. In addition, 18 patient cases had a likely pathogenic finding, and five had a heterozygous CNV likely pathogenic for a recessive disease without association to the patient's phenotype. The clarified cases included six different *DMD* deletions or duplications causing dystrophinopathies. In three sequenced familial cases, the detected CNVs in *CACNA1A*, *SGCD* and *TTN* genes co-segregated with the disease. One case had two separate genetic diseases, tibial muscular dystrophy (TMD) and BMD, caused by the founder mutation FINmaj in the gene *TTN* and a deletion in *DMD*. Some of the solved cases had novel findings: the second ever reported large intragenic deletion in *NEB* causing dominant disease, and the first CNV, an intragenic deletion, in *TIA1* in a patient diagnosed with Welander distal myopathy (WDM).

Some of the genes associated with NMDs are challenging to analyze from short-read sequencing data due to homology or repetitive regions. An additional script was thus written to differentiate copy numbers of the highly homologous genes, *SMN1* and *SMN2*. Two *SMN1/SMN2* copy number 0/3 control cases were successfully recognized, and five cases were identified with a possible exon 7 conversion in *SMN1* and a compatible spinal muscular atrophy phenotype. The latter findings were considered likely pathogenic and are awaiting further validation on the genomic level. Comparison of CNV detections within the in-house CNV database revealed divergences in the CNV detections within the triplicate repetitive region of *NEB* with potentially clinically significant changes. One array CGH validated change correlated well with the nemaline rod pathology observed in the patient.

CNV analysis utilizing MPS data from targeted gene panels and WES samples provided increased diagnostic yield as reported also in other studies on NMDs. Our multi-algorithm and -platform approach decreased the workload in variant analysis and provided more insight into the many difficult to analyze genomic regions involved in NMDs. In the future, whole genome sequencing and long-read sequencing will likely provide higher resolution for CNV detections and reveal an even wider spectrum of structural genomic variants, together with other emerging comprehensive methods, such as optical mapping.

TIIVISTELMÄ

Lihastaudit ovat hyvin heterogeenisiä, ja niistä on kuvattu noin tuhat alatyyppiä. Suurin osa on perinnöllisiä tauteja, ja tähän mennessä on tunnistettu noin 500 eri lihastauteja aiheuttavaa geeniä. Massiivista rinnakkaissekvensointia (MPS) on käytetty laajalti perinnöllisten tautien diagnostisen prosessin nopeuttamiseksi, kustannustehokkuuden parantamiseksi ja lopullisen geeniperäisen diagnoosin saavuttamiseksi. Kopiolukumuutokset, yli 50 emäsparin deleetiot tai duplikaatiot, aiheuttavat arviolta 10 % Mendelin mukaisesti periytyvistä taudeista. Kopiolukumuutosten havaitsemiseen sekvensointidatasta ei ole vielä kehitetty yleisesti hyväksyttyjä ja suositeltuja käytänteitä. Kopiolukumuutosten havaitsemiseksi ja varmistamiseksi käytetäänkin usein täydentäviä menetelmiä, kuten vertaileva genominen hybridisaatio sirulla (aCGH), rinnastettu ligaatio-riippuvainen alukemonistus (MLPA) ja kvantitatiivinen PCR. Kopiolukumuutosten havaitsemiseen sekvensointidatasta on kehitetty useita työkaluja vaihtelevissa tutkimusasetelmissa, mikä hankaloittaa oikean lähestymistavan valitsemista lihastaudeille. Yksittäisten ohjelmien on todettu tuottavan usein epätäsmällisiä ja herkkyydeltään vaihtelevia tai riittämättömiä havaintoja.

Tämän tutkimuksen tavoitteena oli kehittää kattava menetelmä kopiolukumuutosten havaitsemiseen ja annotointiin suurella tarkkuudella kohdennetun geenipaneelin ja koko eksomin (WES) sekvensointidatasta lihastautipotilailta. Tutkimukseen valittiin neljä kopiolukumuutosanalyysin työkalua: CoNIFER, XHMM, ExomeDepth ja CODEX. Kohdennetuista geenipaneeleista MYOcap kattaa 349 geeniä lihaspainotteisille taudeille ja MNDcap 302 hermopainotteisille taudeille nykyisissä paneeliversioissa. MYOcap:lla sekvensointiin 2359 näytettä, MNDcap:lla 942 ja WES:llä 262. Kohdennetuilla geenipaneeleilla sekvensointiin 24 positiivista kontrollinäytettä, joissa on aiemmin tunnistettu kopiolukumuutos, ja 31 negatiivista kontrollinäytettä, joissa tietyt geenit oli varmistettu kopiolukumuutoksia sisältämättömiksi. Kontrollinäytteille saavutettiin kehittämällämme menetelmällä 100 % havaitsemisherkkyys ja 100 % tarkkuus. MYOcap:lla tai MNDcap:lla sekvensoiduista näytteistä havaituista kopiolukumuutoksista 36 varmistettiin todellisiksi havainnoiksi MLPA:lla, PCR:lla tai aCGH:llä ja kahdeksan varmistettiin vääriksi positiivisiksi. Nämä ja positiiviset kontrollinäytteet sisällytettiin logistiseen regressioon perustuvan tilastollisen mallin validointiin. Erottelumallin kehitysvaiheessa MYOcap-sekvensoituihin näytteisiin tehtiin *in silico* kopiolukumuutoksia, mikä tuotti 18677 spesifiä ja 3892 ei-spesifiä kopiolukumuutoshavaintoa mallinnukseen. Malli kehitettiin erottelemaan todelliset kopiolukumuutoshavainnot vääristä positiivista havainnoista havaintomenetelmän tarkkuuden lisäämiseksi.

Neljän ohjelman havaintojen käyttämisen paremmuus verrattuna ohjelmien käyttämiseen yksittäin tai muilla yhdistelmillä todennettiin *in silico* kopiolukumuutosten havaitsemisen herkkyyden tuloksilla. Erottelumalli, jossa oli muuttujia kaikilta neljältä ohjelmalta, saavutti korkeimman herkkyyden (96,6 %), täsmällisyyden (87,5 %) ja tarkkuuden 95,5 % (95 % CI 87,3–99,1 %) kopiolukumuutosten erottelulle. Kopiolukumuutoshavaitsemismenetelmä ja

erottelumalli validoitiin WES-kontrollinäytteillä, joissa oli 235 aiemmin tunnistettua kopiolukumuutosta. Havaitsemisherkkyys kopiolukumuutoksille, jotka sisältävät vähintään kolme eksonia oli 97,3 %, ja erottelumallin herkkyys oli 99,3 % kunhan mallin arviointiraja oli uudelleensäädetty WES-datalle.

Kopiolukumuutosten annotaatiotyökalu cnvScan laajennettiin sisältämään uusimmat kopiolukumuutospopulaatiotietokannat ja talonsisäinen kopiolukumuutostietokanta kaikista sekvensointinäytejoukoista. Alkuperäiset kopiolukumuutoshavainnot neljältä ohjelmalta suodatettiin 1 % enimmäisyleisyyden ja vastavuoroisen 90 % muutoksen kattamisen vaatimuksella yleisissä kopiolukumuutospopulaatiotietokannoissa, tällä sekä 5 % enimmäisyleisyyden ja vastavuoroisen 50 % muutoksen kattamisen vaatimuksella talonsisäisessä tietokannassa, ja lisäksi erottelumallilla todellisiin havaintoihin. Nämä toimenpiteet vähensivät merkittävästi työmäärää kliinisen merkityksen arvioinnille kopiolukumuutoksille säästäen 3–13 % alkuperäisistä havainnoista.

Lisääntyneiden diagnoosien määrä kopiolukumuutoshavaintojen myötä sekä kohdennetuilla geenipaneeleilla että WES-sekvensoiduilla näytteillä oli noin 1,9 %. Kopiolukumuutoshavainnoilla saavutettiin 39 lopullista geneettistä diagnoosia potilaille. Lisäksi 18:lla tutkitulla oli todennäköisesti patogeeninen löydös, ja viidellä tutkitulla havaittiin heterotsygoottinen kopiolukumuutos, jonka arvioitiin olevan patogeeninen peittyvästi periytyvän taudin variantti ilman yhteyttä potilaan taudinkuvaan. Selvitettyihin tapauksiin sisältyi kuusi eri *DMD*-geenissä olevaa deleetiota tai duplikaatiota, jotka aiheuttivat dystrofinopatioita. Kolme potilasta, joilla oli oireisia perheenjäseniä, sekvensointiin perhetapauksina, ja havaitut kopiolukumuutokset geeneissä *CACNA1A*, *SGCD* ja *TTN* segregoituivat yhdessä taudin kanssa. Yhdellä tutkitulla havaittiin kaksi perinnöllistä tautia, tibiaalinen lihasdystrofia (TMD) ja BMD, joiden aiheuttajina olivat perustajamutaatio FINmaj *TTN*-geenissä ja deleetio *DMD*-geenissä. Osalla selvitetyistä tapauksista oli ennen havaitsemattomia löydöksiä: *NEB*-geenissä toinen koskaan raportoitu iso geeninsisäinen deleetio, joka aiheuttaa vallitsevasti periytyvän taudin, sekä *TIA1*-geenin geeninsisäinen deleetio, joka on ensimmäinen havaittu kopiolukumuutos *TIA1*:ssä Welanderin distaalimyopatiaa (WDM) sairastavalla potilaalla.

Jotkin geeneistä, jotka on liitetty lihastauteihin, ovat haastavia analysoitavia lyhytlukuisesta sekvensointidatasta homologian ja toistojaksojen takia. Hyvin homologisille geeneille *SMN1* ja *SMN2* kehitettiin erillinen ohjelma erottelemaan geenien kopiolukumäärät. Kaksi kontrollitapausta tunnistettiin onnistuneesti *SMN1* ja *SMN2* kopiolukumäärillä 0 ja 3, ja lisäksi tunnistettiin viisi tapausta, joilla on mahdollisesti eksonin 7 konversio *SMN1*:ssä ja yhteensopiva spinaalinen lihasatrofia. Jälkimmäiset löydökset luokiteltiin todennäköisesti patogeeniseksi, ja ne odottavat genomista lisävarmistusta. Kopiolukumuutoshavaintojen vertailu *NEB*-geenin triplikaattitoistoalueella talonsisäisessä tietokannassa paljasti eroavaisuuksia, joilla on potentiaalisesti kliinisesti merkitystä. Yksi aCGH:llä varmistettu muutos korreloi selkeästi nemaliinisauvakappalepatologian kanssa, joka potilaalla oli havaittu.

Kopiolukumuutoshavainnointi käyttäen sekvensointidataa kohdennetusta geenipaneelista tai WES-näytteistä lisäsi diagnoosien määrää kuten aiemmissa vastaavissa tutkimuksissa lihastaudeille. Käyttämämme usean algoritmin ja alustan lähestymistapa vähensi varianttianalyysin työmäärää ja tarjosi lisää tietoa useista hankalasti analysoitavista genomisista alueista, jotka on liitetty lihastauteihin. Tulevaisuudessa koko genomin sekvensointi ja pitkälukuinen sekvensointi tarjonnevat paremman resoluution kopiolukumuutoksille ja paljastavat enemmän rakenteellisia genomin muutoksia yhdessä muiden kehitteillä olevien kattavien menetelmien kanssa, kuten optinen kartoitus.

1 INTRODUCTION

Massively parallel sequencing (MPS) has enabled surveying of the human genome at an unprecedented scale and throughput. This method with different applications has brought more insight on the genomic organization, function and variance both in health and disease (Goodwin et al., 2016). Most of all, MPS methods are becoming the standard for the diagnosis of genetic diseases and for research on new genetic defects and disorders (Lappalainen et al., 2019).

The most common variant types in the human genome are single nucleotide variants and small insertions and deletions sized 1–1000 base pairs (bp) (1000 Genomes Project Consortium et al., 2015; Mills et al., 2006). Structural variants are much less numerous but encompass genomic sequence more by multitudes compared to the smaller variation (Chaisson, M. J. P. et al., 2019). Structural variants include copy neutral events, such as translocations and inversions, and copy number variable changes with deletions and duplications, which can span from 50 bp to aneuploidies of whole chromosomes (Harel and Lupski, 2018). Structural variants have been recognized as disease-causing genetic entities (Stankiewicz and Lupski, 2010). Diseases associated with structural variation include multifactorial disorders, such as schizophrenia and autism (Stefansson et al., 2008; Krumm et al., 2015) and various Mendelian disorders (Truty et al., 2019; Pfundt et al., 2017). Approximately 10% of Mendelian disorders are estimated to be explained by copy number variants (CNV) (Truty et al., 2019). The first comprehensive databases for population-wide structural variation have recently been released (Collins et al., 2020; Ruderfer et al., 2016).

Neuromuscular disorders are predominantly genetic in origin and one of the most heterogeneous group of diseases (Bonne et al., 2018). With overlapping phenotypes and varying disease presentations, discovery of the molecular genetic defect is often required for achieving a definitive diagnosis (Laing, 2012). MPS approaches have been especially pivotal in advancing the diagnostics of such diseases with unambiguous phenotypic presentations and genetic backgrounds. They have both increased the diagnostic yield in neuromuscular disorders and decreased costs and workload (Bacquet et al., 2018; Ankala et al., 2015). Some well-known pathogenic CNVs have been documented for neuromuscular disorders, such as the reciprocal deletion and duplication of the whole gene *PMP22* (Chance et al., 1993; Lupski et al., 1991). CNVs affecting other genes have also been confirmed to cause neuromuscular disorders (Kiiski, K. et al., 2016; Bacquet et al., 2018; Hiraide et al., 2019). However, no conclusive recommendations for best practices for CNV analysis from MPS data have been achieved.

Currently, multiple CNV detection tools for MPS data are available for widely differing settings, and they commonly suffer from low sensitivity and specificity (Kosugi et al., 2019; Roca et al., 2019). Therefore, CNV analysis has often been conducted with complementary methods to MPS methods or neglected completely (Nishikawa et al., 2017; Dohrn et al., 2017). However, only accurate molecular genetic diagnosis allows for correct management, genetic

counseling and prognosis for the patients, and in some cases is the basis for direct therapeutic interventions (Carter et al., 2018). CNV analysis from MPS data requires extensive procedures for validation and improvement in accuracy in order to be applied in a routine genetic diagnostic setting.

2 REVIEW OF THE LITERATURE

2.1 Structural genomic variation

2.1.1 Extent of genomic variation

The most frequent variants in the human genome are single nucleotide polymorphisms (SNP), a difference in one single DNA nucleotide (or base pair, bp) compared to the reference genome. SNPs occur with an average of 3.5 to 4.5 million per genome (1000 Genomes Project Consortium et al., 2015; Lappalainen et al., 2019). The next most frequent are small insertions and deletions (indels, variants sized 1–1000 bp) with an average of 550,000 to 625,000 per genome in different populations (1000 Genomes Project Consortium et al., 2015; Mills et al., 2006). A typical genome contains approximately 27,000 distinct structural variants, but their estimated impact on genomic sequence is higher by multiple orders of magnitude as compared to SNPs (1000 Genomes Project Consortium et al., 2015; Lappalainen et al., 2019; Chaisson, M. J. P. et al., 2019). Structural variants include copy number variable events (copy number variant, CNV), such as deletions, tandem duplications, dispersed duplications, and higher-grade amplifications (triplications, quadruplications, etc.), as well as novel insertions. Structural variation can also occur as copy number neutral events, which lead to changes in genomic segment orientation or localization without associated gain or loss of DNA. These variants include inversions, translocations, and complex combinations of the previous (Carvalho, C. M. and Lupski, 2016; Redin et al., 2017; Hurles et al., 2008). The total differences from the human reference genome are estimated to originate by 0.1% for the SNPs and by 0.8–1.2% for the CNVs and indels, the latter having greater genomic content regardless of their lower frequencies (Pang et al., 2010; Conrad et al., 2010; Redon et al., 2006).

The first genomic structural variants detected were at minimum 3 mega base pairs (Mb, 1,000,000 bp) in size, such as aneuploidies, some rearrangements and fragile sites (small breaks or constrictions in a chromosome visible under specific cell-culture conditions) (Feuk et al., 2006). The improvements in studying elongated prometaphase chromosomes allowed the observation of more discrete structural variants, such as reciprocal translocations, deletions, duplications, insertions and inversions (Feuk et al., 2006). Studies by Sebat et al. and Iafrate et al. during 2004 provided the first large scale maps for CNV prevalence globally in the human genome, and others followed (Iafrate et al., 2004; Sebat et al., 2004; Tuzun et al., 2005). For the duration of almost the whole first decade of large scale CNV studies, CNVs were defined to be at least 1 kilo base pairs (kb, 1,000 bp) in size (Conrad et al., 2010; Korbel et al., 2007; Hwang et al., 2015). Then the detection developed enough to discern smaller structural variation (Mills et al., 2011; Collins et al., 2020). Consequently, the CNV size scale is currently defined as varying from the size of an average exon, 50 to 200 bp, to Mbs of DNA (Carvalho, C. M. and Lupski, 2016; Conrad et al., 2010; Alkan et al., 2011; Collins et al., 2020).

Estimations on CNV content in the human genome have varied depending on the number of subjects and ethnicities included in the studies and the CNV detection methods used (Redon et

al., 2006; Itsara et al., 2009; Conrad et al., 2010; Iafrate et al., 2004; Feuk et al., 2006; Sebat et al., 2004; Zarrei et al., 2015; Chaisson, M. J. P. et al., 2019; Collins et al., 2020). According to some recent estimations, 4.8 to 9.5% of the genome is affected by CNVs (Zarrei et al., 2015). The amount and types of CNVs vary between populations; some CNVs are entirely unique for an individual, whereas some are common polymorphisms shared across populations (Redon et al., 2006; Itsara et al., 2009; Mills et al., 2011; Conrad et al., 2010; Sebat et al., 2004). According to the most recent comprehensive studies for structural variant prevalence, most structural variants are less than 1 kb in size with a median of 331 bp, and rare with approximately 92% of the detected variants presenting with less than 1% frequency (Chaisson, M. J. et al., 2015; Collins et al., 2020; Chaisson, M. J. P. et al., 2019).

2.1.2 Formation mechanisms of structural variation

The locus-specific mutation frequency is orders of magnitude greater for structural variation compared to point mutations, especially during meiosis (Hastings, Lupski et al., 2009). A recent rough estimation for *de novo* structural variation emergence is 0.29 variants per generation (Collins et al., 2020). Compared to SNPs and indels, the formation of structural variants requires disruption of the DNA sugar-phosphate backbone (Carvalho, C. M. and Lupski, 2016). Genomic regions surrounding the breakpoints of different structural variant types have only the degree of conservation in common, mirroring their differing origins (Abyzov et al., 2015).

Copy number variable regions are significantly associated with segmental duplications (SD) and some other types of repetitive genomic sequences (Redon et al., 2006; Itsara et al., 2009; Hastings, Lupski et al., 2009; Bailey et al., 2002). Over 50% of the human genome consists of repeat sequences, which include mobile elements (*Alu*-processed pseudogenes being the most prevalent), simple sequence repeats, tandem repeat sequences (predominantly in centromeres, telomeres and ribosomal gene clusters), and low-copy repeats (LCR), which include SDs (Carvalho, C. M. and Lupski, 2016; Lander et al., 2001; Stankiewicz and Lupski, 2002b). Definitions for the different classes of repetitive variants have remained arbitrary and overlapping (Lappalainen et al., 2019). Short Tandem Repeats (STR), or microsatellites, are consecutive expansions of repeat units of 1 to 6 bp (Dashnow et al., 2018; Lappalainen et al., 2019). Variable Number Tandem Repeats (VNTR), or minisatellites, have 6 to 100 bp repeating units (Bakhtiari et al., 2018; Lappalainen et al., 2019). Both VNTRs and STRs cover approximately 3% of the human genome (Dashnow et al., 2018; Bakhtiari et al., 2018).

LCRs share over 97% sequence identity and typically occur only twice or a few times as units of 10–400 kb in the highly complex regions of the genome (Stankiewicz and Lupski, 2002b). LCRs can contain single or multiple genes, gene fragments, pseudogenes, endogenous retroviral sequences or other paralogous fragments of direct or inversely oriented sequences (Bailey et al., 2002; Harel and Lupski, 2018; Stankiewicz and Lupski, 2002a). SDs share at least 90% sequence identity and are at least 1 kb in length (Carvalho, C. M. and Lupski, 2016; Bailey et al., 2002). LCRs and SDs themselves are often present in variable copy number in addition to participating in their formation, as discussed further, and can thus be counted as

CNVs (Feuk et al., 2006; Iafrate et al., 2004; Stankiewicz and Lupski, 2010). Transposable elements can also form structural variants; most of the mobile element insertions are *Alu* elements of 300 bp or L1/LINEs of 6 kb (Mills et al., 2011; Abyzov et al., 2015).

The relative orientation, size, degree of homology and distribution of local repeat and other sequences help predict the types of structural variants the region is susceptible to, and their possible formation mechanisms (Carvalho, C. M. and Lupski, 2016; Stankiewicz and Lupski, 2002a). DNA repair and replication mechanisms, which occasionally lead to structural variants, leave mutational signatures behind on the genomic regions surrounding the breakpoints. These have been studied to explain the complex array of polymorphic human structural variants and their origin mechanisms (Abyzov et al., 2015; Austin-Tse et al., 2018). For example, the formation of duplications seems to be more sequence dependent than deletions with more breakpoint-associated sequence motifs (Conrad et al., 2010). The emergence rate for triplications from duplications is over 100X higher than the rate of *de novo* duplications, which means that an existing duplication significantly predisposes the region to further amplifications (Liu, P. et al., 2014). Inverted duplications tend to participate in complex rearrangements, which occur also more often in tandem (Newman et al., 2015). Generally, different types of structural variants have different meiotic versus mitotic risks to form. The timing of replication is also associated with the mechanisms, which could generate the different rearrangements (Carvalho, C. M. and Lupski, 2016; Abyzov et al., 2015). As an example, recurrent CNVs seem to be more likely produced on regions of early replication, whereas non-recurrent CNVs occur more frequently on regions replicated later (Carvalho, C. M. et al., 2015).

2.1.2.1 Recurrent structural variation

Recurrent structural variants can be detected with the same size, breakpoints and genomic content in different, unrelated individuals (Figure 1) (Carvalho, C. M. and Lupski, 2016; Austin-Tse et al., 2018). These variants tend to have long stretches of homology at their breakpoint junctions provided by long highly identical glancing interspersed paralogous repeats, most often LCRs. These elements provide homology for the mechanisms producing the structural variants (Hastings, Lupski et al., 2009; Carvalho, C. M. and Lupski, 2016). Early on it was recognized that LCRs predispose genomic regions to frequent genomic rearrangements (Lupski, 1998; Stankiewicz and Lupski, 2002a).

Structural aberrations can originate from processes related to DNA recombination, replication or repair, and the recombination-based changes were the first to be observed (Carvalho, C. M. and Lupski, 2016; Stankiewicz and Lupski, 2002a). Homologous recombination is normally used in the repair of double-strand DNA breaks or broken replication forks with single double-strand ends. Homologous recombination takes also part in ordered segregation and allelic recombination in meiosis (Hastings, Lupski et al., 2009). Homologous recombination is the basis for accurate DNA repair with a requirement for as much as 300 bp of homology and the sister chromatid as the preferred recombination pair (Hastings, Lupski et al., 2009).

In non-allelic homologous recombination (NAHR), a non-allelic homologous segment is used as template. This leads to a loss of heterozygosity (LOH) without change in the quantity of the involved genomic sequence. If the homologous sequence originates from different chromosomal location, a translocation, inversion, duplication or deletion can be the outcome (Stankiewicz and Lupski, 2002b; Hastings, Lupski et al., 2009). NAHR requires substrates with near-perfect homology, such as LCRs, SDs or repetitive sequences, and also with suitable length and within a certain distance (Stankiewicz and Lupski, 2002a; Abyzov et al., 2015). A typical formation mechanism for recurrent structural variation is NAHR between intrachromosomal (or occasionally interchromosomal) at least 10 kb LCRs within 10 Mb distance (Carvalho, C. M. and Lupski, 2016; Stankiewicz and Lupski, 2002a; Stankiewicz and Lupski, 2010). Therefore, NAHR occurs mostly near telomeres and recombination hotspots, where susceptible LCRs misalign during mitosis or meiosis (Mills et al., 2011). Most recurrent CNVs seem to be produced during meiosis through NAHR (Watson et al., 2014).

Generally, NAHR forms recurrent rearrangements with clustered breakpoints and leaves behind long homologous sequences at breakpoints (Carvalho, C. M. and Lupski, 2016; Stankiewicz and Lupski, 2002a; Abyzov et al., 2015). NAHR tends to produce longer CNVs compared to other mechanisms (Redon et al., 2006; Conrad et al., 2010). Depending on whether the repeats seeding the rearrangement are located on the same chromatid or on different chromatids or chromosomes, and the orientation of the repeats, NAHR can lead to deletions, duplications or inversions of the intervening genome (Lupski, 1998; Hastings, Ira et al., 2009; Stankiewicz and Lupski, 2002a). NAHR can also generate arrays of tandem duplications with varying sizes, and reciprocal deletions and duplications (Redon et al., 2006). The regions with LCRs or SDs were early on linked to recurrent genetic diseases, such as microdeletion and microduplication syndromes with more than 20 different currently recognized (Carvalho, C. M. and Lupski, 2016; Bailey et al., 2002; Watson et al., 2014).

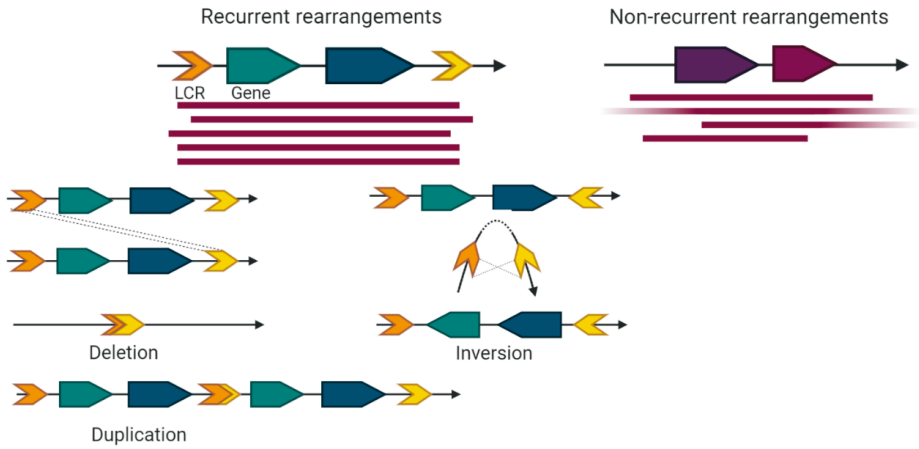

Figure 1: Recurrent and non-recurrent rearrangements. In recurrent rearrangements, the breakpoints of the variants are grouped in the same genomic location, while non-recurrent rearrangements have more varied breakpoint locations. Recurrent rearrangements are often mediated by flanking LCRs, the orientation of which affects whether deletions, duplications or inversions are formed. *(All figures in the Review of the literature are made by Salla Välipakka and created with BioRender.com.)*

2.1.2.2 Non-recurrent and complex structural variation

Non-recurrent rearrangements are unique in size and genomic content. Nevertheless, individuals with overlapping clinical phenotypes may be found to have the same dosage sensitive gene or genes affected by structural variation in the smallest region of overlap, representing a breakpoint grouping region in these copy number variable regions (Carvalho, C. M. and Lupski, 2016; Zarrei et al., 2015). Most non-recurrent rearrangements tend to have either simple blunt ends or microhomologies of < 35 bp at their breakpoints, and they can be considerably complex in structure (Carvalho, C. M. and Lupski, 2016; Lee, J. A. et al., 2007; Hastings, Lupski et al., 2009). These junctions can display a mixture of amplified, inversed, deleted and unchanged sequence (Hastings, Lupski et al., 2009).

The mechanisms forming non-recurrent structural variants include both non-homologous end joining (NHEJ) used in the repair of DNA, and replication related mechanisms, such as break-induced replication (BIR), microhomology-mediated break-induced replication (MMBIR), and fork stalling and template switching (FoSTeS) (Carvalho, C. M. and Lupski, 2016; Harel and Lupski, 2018; Abyzov et al., 2015). NAHR and NHEJ are both recombination-based mechanisms occasionally leading to rearrangements, with NHEJ accounting for the majority (Hastings, Lupski et al., 2009). According to one study, NHEJ was involved in the formation of 56% of structural variants, retrotransposition 30% and NAHR 14% (Korbel et al., 2007), but older studies may be biased because of limited detection of structural variants, which will be discussed later.

The DNA replication-based mechanisms require at least microhomology as a primer for replication. NHEJ can utilize *Alu*-elements with notably less sequence identity (as low as 75%) as compared to SDs (Carvalho, C. M. and Lupski, 2016; Lee, J. A. et al., 2007). If the microhomology originates from another chromosome, then deletions, duplications, inversions or translocations can be formed. Annealing with the homologous chromosome rather than the sister chromatid leads to LOH for the affected region (Hastings, Lupski et al., 2009). NHEJ is also utilized in repairing double-stranded DNA breaks alternatively with microhomology-mediated end-joining mechanisms (MMEJ), which can function on sites without extensive sequence homology (Abyzov et al., 2015; Hastings, Lupski et al., 2009). NHEJ can lead to small indels of 1 to 4 bp, while MMEJ needs homology of 5 to 25 bp and can lead to larger deletions between stretches with microhomology (Hastings, Lupski et al., 2009). NHEJ can either generate blunt CNV breakpoints or leave short homology or small insertions or deletions of random nucleotides at the breakpoints, much like MMEJ (Carvalho, C. M. and Lupski, 2016; Abyzov et al., 2015).

The breakpoints of non-recurrent rearrangements tend to occur in LCR-rich regions with complex genetic architecture (Hastings, Lupski et al., 2009; Stankiewicz et al., 2003). Therefore, LCRs are thought to mediate recurrent rearrangements by NAHR and also stimulate non-recurrent rearrangements by offering homology or microhomology for the DNA repair-based mechanisms (Liu, P. et al., 2011; Stankiewicz et al., 2003). The occurrence of non-

recurrent CNVs close to LCRs could also be explained by the tendency of these regions to form secondary DNA structures, leading to replication fork stalling or collapse, which may provide single-stranded regions as starting points for some of the formation mechanisms (Hastings, Lupski et al., 2009; Carvalho, C. M. and Lupski, 2016). In template switching, the single-stranded DNA template is switched during replication to either another template with microhomology within the same replication fork, or to a template originating from a different replication fork (short and long-distance template switch) (Carvalho, C. M. and Lupski, 2016). Template switching occurs more frequently within a chromosome, but interchromosomal rearrangements can occur as well (Carvalho, C. M. and Lupski, 2016). Generally, these disturbances in replication can also lead to the switching of replicative polymerases to more efficient but error-prone low-processivity polymerases, which in turn can generate structural variation (Carvalho, C. M. and Lupski, 2016).

In BIR, homologous recombination is utilized to repair single-end double-stranded DNA breaks left behind by collapsed or broken replication forks (Hastings, Lupski et al., 2009; Carvalho, C. M. and Lupski, 2016). Sometimes a replication fork collapses or stalls under stress without the proteins or long enough homology required by BIR. Additionally, a collapsed replication fork may leave only one single-stranded DNA end, in which case NHEJ with the requirement for double-stranded breaks cannot function either (Hastings, Lupski et al., 2009). In these cases, microhomology-mediated BIR is used instead. In this process, the 3' end from the collapsed fork anneals with any single-stranded template with microhomology, initiating DNA synthesis and a low-processivity replication fork anew (Hastings, Lupski et al., 2009; Carvalho, C. M. and Lupski, 2016). Depending on whether the new fork is located upstream or downstream, deletion or duplication occurs, and whether the leading or lagging strand is used as a template decides the orientation of the incorporated fragment (Stankiewicz and Lupski, 2010; Hastings, Ira et al., 2009).

MMBIR may be the main mechanism for the formation of non-recurrent structural variation as it can lead to multiple genomic consequences: deletions, inversions and translocations, and in particular duplications, triplications and complex rearrangements (structural variants with more than two breakpoint junctions), as well as segmental uniparental disomy, or LOH (Hastings, Lupski et al., 2009; Carvalho, C. M. and Lupski, 2016; Liu, P. et al., 2011; Hastings, Ira et al., 2009). Since MMBIR has less stringent requirements for homology in recombination compared to BIR, it leads more probably to LOH (Carvalho, C. M. et al., 2015; Hastings, Ira et al., 2009). MMBIR can also explain the frequently observed microhomology and inserted short segments at non-recurrent structural variation breakpoints, since the template switch can occur multiple times (Hastings, Lupski et al., 2009; Carvalho, C. M. and Lupski, 2016; Liu, P. et al., 2011; Hastings, Ira et al., 2009).

The replication-error mechanism FoSTeS produces non-recurrent complex rearrangements in a replication-based manner. The replication fork can stall at DNA lesions or at closely located LCRs and jump back and forth on the genomic length. Skipping of segments leads to deletions

and re-replicating them generates duplications (Lee, J. A. et al., 2007). This jumping may also target a single-stranded DNA with microhomology on a nearby replication fork to reinitiate DNA synthesis (Stankiewicz and Lupski, 2010). Thus, FoSTeS can also generate deletions with micro-insertions originating from elsewhere in the genome (Lee, J. A. et al., 2007). These events may mix and occur multiple times in the same region during a single replication event, thus generating complex structural variants with deletions and/or duplications interrupted by normal sequence or triplicated segments (Lee, J. A. et al., 2007; Stankiewicz and Lupski, 2010; Liu, P. et al., 2011). FoSTeS occurs especially at lagging-strand template, sites of frequent transcription, and at sites prone to secondary DNA structures (Hastings, Lupski et al., 2009). Furthermore, FoSTeS and MMBIR can both participate in forming LCRs themselves, which predisposes genomic regions to structural changes, invoking a cycle to form complex rearrangements (Stankiewicz and Lupski, 2010; Hastings, Lupski et al., 2009).

Chromothripsis, a chromosome-shattering event leading to numerous genomic rearrangements with dozens to hundreds of breakpoints, was first described in cancers (Weischenfeldt et al., 2013; Liu, P. et al., 2011). These rearrangements can include deletions, duplications, triplications, translocations and inversions (Liu, P. et al., 2011). Chromothripsis affects typically one or two chromosomes (Turajlic et al., 2019). A constitutionally acquired similar event, chromoanasynthesis (chromosome re-assortment) has been described as well (Liu, P. et al., 2011; Pellestor and Gatinois, 2018). Both of these mechanisms are estimated to be based on similar events with error-prone DNA repair on the cellular level (Liu, P. et al., 2011). Chromothripsis seems to be driven by random events of chromosomal shattering stitched together by NHEJ, while chromoanasynthesis is thought to be based on a replication-based process involving FoSTeS or MMBIR (Pellestor and Gatinois, 2018). The rearrangements appear to be formed in one single event, since the multiple rearrangements are localized on a certain genomic region usually involving only one chromosome. The individual rearrangements are also present at equal amounts, rather than mosaic, which would hint towards the rearrangements appearing at different times (Liu, P. et al., 2011).

DNA repair-mediated non-recurrent rearrangements are more likely to be formed during mitosis, and therefore contribute more to diseases arising from somatic mutations, such as cancers (Carvalho, C. M. and Lupski, 2016). Generally, genomic rearrangements in genomic disorders are less complex compared to the ones in cancer. This probably reflects different selective forces, since organisms would not survive large-scale and widespread genomic changes occurring early in development (Liu, P. et al., 2011).

The different types of structural variants discussed above are illustrated in Figure 2.

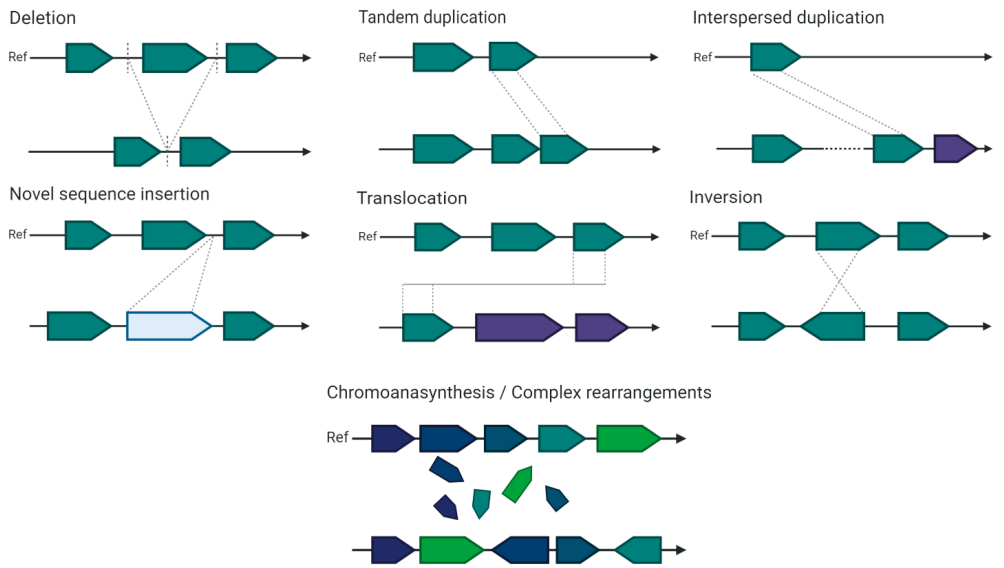

Figure 2: Different types of structural variants. Structural variants include unbalanced copy number variable events, such as deletions and duplications, and balanced events, such as translocations and inversions.

2.1.3 Impact of structural variation

One of the first disorders studied in depth and revealed to be caused predominantly by CNVs was Charcot-Marie-Tooth disease type 1A (CMT1A) caused by duplication of the gene *PMP22* (Lupski et al., 1991; Raeymaekers et al., 1991). CMT1A is an autosomal dominant progressive demyelinating peripheral neuropathy characterized by weakness and atrophy of distal limb muscles (Raeymaekers et al., 1991). Soon following this, hereditary neuropathy with liability to pressure palsies (HNPP) was found to be caused by a deletion encompassing the same gene. HNPP causes episodic focal pressure neuropathies with mild disability and occasional peripheral neuropathy manifestations (Chance et al., 1993). The deletion and tandem duplication that cause these disorders were discovered to be recurrent reciprocal recombination products on a genomic region, where flanking CMT1A-REP repeats serve as substrates for NAHR (Pentao et al., 1992; Reiter et al., 1998; Stankiewicz and Lupski, 2002b). Following these findings, genomic disorders were defined in a seminal paper as entities originating from changes in the genome architecture leading to a loss or gain or disruption of integrity of a gene or genes with dosage sensitivity (Lupski, 1998). This description has been expanded during the following two decades of studying structural variants and their effects on phenotype and involvement in genetic disorders (Stankiewicz and Lupski, 2010; Harel and Lupski, 2018; Collins et al., 2020).

2.1.3.1 Evolutionary point of view

According to genome-wide surveys, most structural variants seem to be neutral from an evolutionary point of view (Hurles et al., 2008). Nevertheless, their distribution is unequal throughout the genome. Namely, structural variants are biased away from genes and other functional elements (Redon et al., 2006). CNVs are more pronouncedly located in

pericentromeric and subtelomeric regions (Bailey et al., 2002; Schuster-Bockler et al., 2010; Collins et al., 2020). For example, changes in the repeat size of VNTRs lead most commonly to small structural variants packed at centromeres (Mills et al., 2011; Collins et al., 2020). Additionally, ultraconserved elements in all mammalian genomes have less structural variants (Zarrei et al., 2015). Deletions seem to be more biased away from genes as well as evolutionarily conserved genomic regions than duplications (Redon et al., 2006; Abyzov et al., 2015). Genes involved in disorders and genes with high genic intolerance score and haploinsufficiency index (HI) have less structural variants compared to others (Zarrei et al., 2015).

For haploinsufficient genes, a decrease in gene dosage is detrimental, whereas dosage sensitive genes are affected adversely by either an increase or a decrease of the dosage (Weischenfeldt et al., 2013). The HI score is based on the integration of genomic, evolutionary and functional information (Huang et al., 2010). Genes with high genic intolerance score are statistically depleted of protein sequence affecting (non-synonymous) variation (Zarrei et al., 2015; Lek et al., 2016). Genic intolerance score has also been calculated separately for CNVs and provides similar values both for deletions and duplications (Ruderfer et al., 2016; Collins et al., 2020). Genic intolerance score for CNVs correlates with the HI calculated for SNVs and indels, and the correlation is stronger for deletions than duplications (Ruderfer et al., 2016; Lek et al., 2016; Collins et al., 2020).

Haploinsufficient genes have higher expression levels during early development, more interaction partners, and statistically higher association with dominant diseases than other genes (Huang et al., 2010). Genes with intolerance for LOF mutations are also generally more expressed and present more widely in different tissues (Karczewski et al., 2020). Additionally, constitutive regulatory elements in non-coding regions are more dosage sensitive than cell-type specific regulators (Abel et al., 2018). Evidently, structural variation is biased away from genes involved in protein phosphorylation, signal transduction, protein degradation, transcriptional machinery and regulation, intracellular transport, development, differentiation and cell cycle (Zarrei et al., 2015). Purifying selection acts more strongly against deletions than duplications, which could also explain the tendency of deletions to be shorter than duplications (Itsara et al., 2009; Redon et al., 2006; Hastings, Lupski et al., 2009; Conrad et al., 2010; Ruderfer et al., 2016). Individuals have less deletions compared to duplications in their genome, especially on whole gene level (Truty et al., 2019; Ruderfer et al., 2016). Benign CNVs tend to be substantially shorter compared to pathogenic, and even when corrected for length, pathogenic CNVs include more genes than benign (Rice and McLysaght, 2017).

On the contrary, genes with functions in immune responses, responses to biotic stimuli such as olfactory receptors, drug and steroid metabolism, starch and sucrose metabolism, pregnancy-specific adhesion molecules, ER, vesicle and Golgi apparatus are enriched with structural variation, especially duplications (Bailey et al., 2002; Feuk et al., 2006; Zarrei et al., 2015; Schuster-Bockler et al., 2010; Mills et al., 2011; Redon et al., 2006). Therefore, structural

variation may have a role in adaptability in response to external pressure; duplication followed by functional specialization could provide the variability required for adaption and for the potential to evolve (Bailey et al., 2002; Feuk et al., 2006). CNV rich regions may drive evolution through frequent duplications, which occasionally get fixed and lead to the evolution of gene families (Schuster-Bockler et al., 2010). An example of this is *AMY1*, salivary amylase, involved in the digestion of starch. Individuals can possess 1–10 copies of the gene, which affects the protein levels and is associated with population-specific differences in starch consumption (Hurles et al., 2008; Perry et al., 2007). On the other hand, complete deletion of some genes seems to have no apparent phenotypic effect. Individuals have on average 11 genes inactivated by homozygous deletions (Collins et al., 2020). These genes could belong to a gene family, have redundant function or cause late-onset phenotypes, and thus not affect fitness (Zarrei et al., 2015).

2.1.3.2 Mechanisms for influencing gene function

The most apparent mechanism for CNVs to alter gene function is by altering the dosage of the genes they encompass (Schuster-Bockler et al., 2010; Stankiewicz and Lupski, 2010). Evolutional constraint of a gene against gain or loss suggests dosage sensitivity, and these genes tend to have more pathogenic CNVs (Rice and McLysaght, 2017; Schuster-Bockler et al., 2010). Genes constrained by dosage sensitivity may need to maintain stoichiometric balance with other genes or act in a coordinated concentration-dependent manner, such as developmental morphogenes or co-factors (Weischenfeldt et al., 2013). Their protein products may be prone to form aggregates toxic to cells or bind non-physiologically in high concentrations (triplosensitive genes). Alternatively, a minimum amount of protein may be required to achieve effect, such as with many transcription factors and developmental genes (haploinsufficient genes) (Wu et al., 2015; Schuster-Bockler et al., 2010; Rice and McLysaght, 2017). Pathogenic CNVs are enriched especially in developmental and neurodevelopmental genes, reflecting probably the dosage sensitivity of the process (Rice and McLysaght, 2017; Redin et al., 2017). Some of these diseases will be described and discussed further.

Haploinsufficiency and triplosensitivity affect more likely genes controlling the expression of other genes and their protein products, which are often involved in complexes (Schuster-Bockler et al., 2010; Papp et al., 2003; Cabrejo et al., 2006). Less duplications have been detected in genes partaking in protein complexes, which supports the fact that a stoichiometric balance needs to be retained in protein complexes (Schuster-Bockler et al., 2010). Excess of one sub-unit could form harmful homodimers as compared to the natural and functional heterodimer, disturb the balance in regulatory subunits of the protein complex, or form toxic aggregates (Papp et al., 2003). Enzymes tend to be more resistant to CNVs but could be haploinsufficient through functioning as a rate limiting factor in a biochemical pathway (Schuster-Bockler et al., 2010; Harel and Lupski, 2018).

Change in gene dosage can occur in various ways and at various stages. In most cases (80%), the copy number is positively correlated with gene expression, but the opposite has also been

shown (Hurles et al., 2008; Schuster-Bockler et al., 2010). Change in mRNA level may be inconsistent with the change in protein level since the latter is affected by additional post-transcriptional mechanisms, translational control, and protein folding and stability (Weischenfeldt et al., 2013). Dosage compensation mechanisms, such as epistatic interactions, regulatory feedback mechanisms or phenotypic buffering with genes with redundant functions could be in effect in the more starkly unintuitive cases (Weischenfeldt et al., 2013).

Homo- and hemizygous deletions, whether partial with one or both of the CNV breakpoints within the gene or encompassing the whole gene, often result in a total loss of gene function (Alkuraya, 2015; Gambin, Akdemir et al., 2017). Most duplications (83%) seem to result in the copy being in tandem with the original gene and in direct orientation (Newman et al., 2015). In these cases, one normal copy of the gene may be preserved, whereas inverted and inserted duplications may disrupt genes at breakpoint junctions, affecting also the original gene (Newman et al., 2015). CNVs may also generate chimeric or completely novel fusion genes if the combined genes are in the same orientation and the reading frame is preserved (Bailey et al., 2002; Conrad et al., 2010; Harel and Lupski, 2018; Korbel et al., 2007). Balanced chromosomal aberrations, such as inversions, could have deleterious effects on gene function through direct disruption of genes at breakpoints, or they may cause long-range regulatory changes by altering the chromosomal structure. This has also been utilized to discover new disease genes (Redin et al., 2017). Compensatory effects are also possible: an asymptomatic father of a DiGeorge syndrome patient was detected to carry both a deletion and a duplication on the 22q11.2 region, leading to normal gene dosage (Carelle-Calmels et al., 2009).

In some cases, pathogenic CNVs have been detected to encompass non-coding regions, which affect gene expression, such as intergenic sequences, non-coding elements within protein-coding genes, as well as non-coding RNAs, such as microRNAs (miRNA) and long non-coding RNAs (Zhang, F. and Lupski, 2015; Szafranski et al., 2013). Structural variation can also alter gene dosage by affecting regulatory elements, such as enhancers or repressors, boundary elements or the intervening sequences to their targets (Hurles et al., 2008; Gheldof et al., 2013; Stankiewicz and Lupski, 2010; Stankiewicz et al., 2005). For example, adopting an enhancer or losing a repressor can lead to a gain of function (GOF) effect through regulatory change (Redin et al., 2017). Structural variation may also affect gene expression locally or genome-wide by repositioning a genomic region within the nucleus or by altering the chromatin architecture. Perturbation of chromatin loops or topologically associating domains (TADs) can affect also genes without CNVs (Zhang, F. and Lupski, 2015; Harel and Lupski, 2018; Gheldof et al., 2013; Stankiewicz et al., 2005).

TADs supposedly represent structural scaffolds, where enhancers and promoters interact separated by boundary regions, which limit the distance and direction of their operation areas (Lupianez et al., 2015). The boundary regions have usually binding sites for specific factors, which block interaction between adjacent TADs (Dixon et al., 2012). Changes in the orientation or location of these boundary elements can form new separate TADs or fuse them (Lupianez et

al., 2015). Misplacement of an enhancer can have tissue- or developmental-stage-specific effects (Lupianez et al., 2015; Weischenfeldt et al., 2013). Comparably, intragenic CNVs can disrupt the reading frame of a specific gene isoform with effects in a certain developmental stage (Newman et al., 2015). Mobile-element insertion by retrotransposition has the potential to disrupt or reorder genes and regulatory elements (Masson et al., 2020), but this has not been recognized as a major genomic disorder causing factor (Kazazian and Moran, 2017; Lappalainen et al., 2019).

CNVs in STRs or VNTRs tend to have deleterious effects by local repeat expansion, which leads to gene product with structural features that disrupt normal cellular processes (Mirkin, 2007). Alternatively, some are located on regions where the expansion disturbs promoters or other elements affecting gene expression (Dashnow et al., 2018; Lappalainen et al., 2019). The expansions occur in *cis,* and the mutations can be dynamic, becoming increasingly unstable after a certain threshold (Mirkin, 2007). They can thus diverge from the principles of classical genetics of mutations stably transmitting through generations. Increase in repeat length can affect disease penetration, severity and/or age of onset, with increasing severity and earlier age of onset termed anticipation (Mirkin, 2007). Most expanding repeats are trinucleotide units such as polyglutamine and polyalanine mutant protein stretches with deleterious aggregation propensity (Mirkin, 2007).

If replicative repair is completed using homologous segments as templates, or mitotic crossing-over occurs between homologs or sister chromatids and extends for multiple kilobases, no copy number change will occur, but the segment will show loss of heterozygosity (LOH) (Carvalho, C. M. et al., 2015; Campbell et al., 2016; Hastings, Lupski et al., 2009). Especially complex rearrangements and triplications are associated with regions of LOH, or segmental uniparental disomy (Campbell et al., 2016; Feuk et al., 2006). Uniparental disomy is a non-Mendelian human disease-causing genetic mechanism based on either an involvement of an imprinted loci or an exposure of a recessive trait (Spence et al., 1988; Carvalho, C. M. et al., 2015). By definition, the expression of an imprinted gene is determined by parental origin, and problems can arise if these gene loci are not inherited from both parents intact (Weischenfeldt et al., 2013). Alternatively, LOH may unmask a recessive disease variant and result in homozygosity for that locus with only one parent being a carrier, thus distorting Mendelian expectations (Carvalho, C. M. et al., 2015). Similarly, deleted regions can unmask otherwise phenotypically silent recessive alleles (Albers et al., 2012; Harel and Lupski, 2018).

The variable mechanisms for structural variation to affect gene function are illustrated in Figure 3.

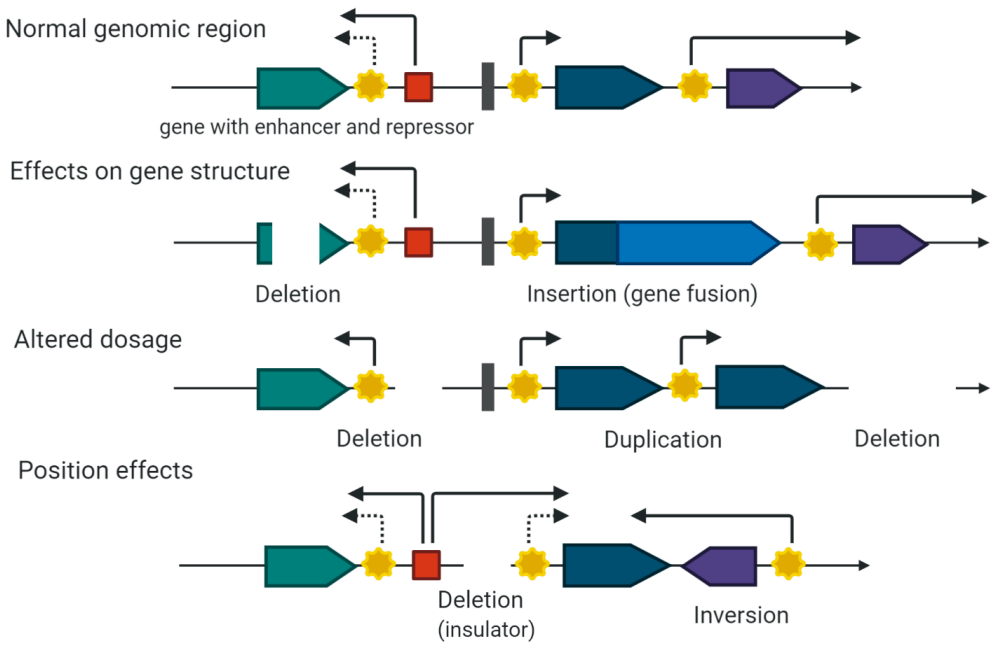

Figure 3: Effects of structural variants on genes and genomic regions. Generally, structural variation can either affect gene structure or gene dosage. The mechanisms include the direct disturbance of genomic elements (genes, regulators of gene expression, regulatory units), or alteration in their copy numbers.

2.1.3.3 Disease causativity

CNVs have been detected to cause autosomal dominant, recessive, and X-linked or Y-linked diseases (Stankiewicz and Lupski, 2002a). Depending on the number of genes affected, structural variation can result in a Mendelian disease, a contiguous gene syndrome or a chromosomal disorder. The contiguous gene syndromes include Williams–Beuren syndrome from a common 1.6 Mb deletion on 7q11.23, DiGeorge syndrome from a 3 Mb or 1.5 Mb deletion on 22q11.2, and Smith-Magensis syndrome from a deletion on 17p11.2, with all these regions possessing flanking LCRs (Stankiewicz and Lupski, 2002a; Potocki et al., 2003; Stankiewicz et al., 2003). CNVs have been associated with diseases in variable categories, such as neurodevelopmental disorders including schizophrenia (Stefansson et al., 2008; Stankiewicz and Lupski, 2010), autism spectrum disorders (ASDs) (Krumm et al., 2015; Stankiewicz and Lupski, 2010), Parkinson's disease (Singleton et al., 2003), some complex disorders (Lappalainen et al., 2019) and various Mendelian disorders (Truty et al., 2019; Pfundt et al., 2017). CNVs affect also genes with pharmacogenetic implications, such as the genes encoding cytochrome P450 enzymes (Meijerman et al., 2007; Santos et al., 2018). Regardless of disease category, some 10–11% of Mendelian diseases are estimated to be explained by pathogenic CNVs (Clark et al., 2019; Truty et al., 2019). Somatic CNVs are significant factors in tumorigenesis of various cancers, such as breast cancer, glioma and lung cancer. Deletions or duplications of some cancer related genes in tumor samples can direct treatment decisions and help in diagnosis (Kim, H. Y. et al., 2017; Hehir-Kwa et al., 2018). Recurrently amplified regions with oncogenes include *MYC* and *GLI2* (Hehir-Kwa et al., 2018). Also germline CNVs

have been detected in genes such as *BRCA1, MSH2*, and *TP53* associated with cancer disposition syndromes (Hehir-Kwa et al., 2018).

Structural variants associated with complex phenotypes seem to frequently intersect; *de novo* duplications at the 7q11.23 locus are associated both with ASDs and schizophrenia (Weischenfeldt et al., 2013; Stankiewicz and Lupski, 2010). Multiple deletions, duplications and aneuploidy of chromosome Y have been associated with ASDs characterized by neurodevelopmental abnormalities, social impairment, and a restricted range of behaviors and interests (Stankiewicz and Lupski, 2010; Krumm et al., 2015). Schizophrenia is a severe psychiatric disorder with different combinations of hallucinations, delusions and cognitive deficits. It was first associated with microdeletions at 1q21.1, 15q11.2 and 15q13.3 loci, and later also with more variable CNVs similar to ASDs (Stankiewicz and Lupski, 2010; Stefansson et al., 2008). Parkinson's disease is a neurodegenerative disorder of the brain, which impairs motor functions and speech. Triplications and duplications of the *SNCA* gene, which encodes the main component of the aggregated protein detectable in the disease, Lewy bodies, cause Parkinson's diseases of different severity (Stankiewicz and Lupski, 2010).

CNVs may cause the same disease when surrounding a gene, or different diseases depending on CNV location and state (duplication or deletion) (Zhang, F. and Lupski, 2015). For example, both deletions and duplications of the gene *PLP1* cause the same disease, Pelizaeus-Merzbacher X-linked recessive hypomyelinative leukodystrophy (Weischenfeldt et al., 2013). Both 190 kb duplications approximately 33 kb upstream and 150 kb duplications 136 kb downstream of *PLP1* have been associated with spastic paraplegia and axonal neuropathy (Zhang, F. and Lupski, 2015). A common mechanism for the latter diseases may be the disruption of a boundary element leading to an enhancer being placed next to the gene with a GOF effect (Lupianez et al., 2015). On the contrary, deletions and duplications upstream or downstream of *SOX9* result in various different clinical phenotypes. This suggests that *SOX9* is surrounded by long-range *cis*-regulatory elements (Zhang, F. and Lupski, 2015; Stankiewicz et al., 2005). Reciprocal deletions and duplications at the same genomic location have also been associated with different clinical phenotypes (Carvalho, C. M. and Lupski, 2016; Watson et al., 2014). Sometimes they are mirror traits, such as early onset underweight and obesity from the duplication and deletion of 16p11.2, or microcephaly and macrocephaly associated with deletions and duplications at 1q21 (Weischenfeldt et al., 2013; Watson et al., 2014). On the contrary, CMT1A and HNPP are both peripheral neuropathies (with different presentations) resulting from deletion and duplication of the whole gene *PMP22* (Stankiewicz and Lupski, 2010).

2.2 Methods for genome variant detection

Genome-scanning technologies and comparative DNA sequence analysis are the two general approaches in studying genome variation (Chaisson, M. J. P. et al., 2019; Goodwin et al., 2016).

2.2.1 Karyotyping, FISH and optical mapping

Disorders involving whole or partial chromosomal abnormalities are detectable by karyotyping (Weischenfeldt et al., 2013). In Giemsa banding, Giemsa stained heterochromatin reveals a distinct pattern of bands for each chromosome and their form and size (Feuk et al., 2006). Spectral karyotyping involves staining of each chromosome differentially with DNA labeling probes, which can reveal rearrangements involving different chromosomes (Feuk et al., 2006). In fluorescent *in situ* hybridization (FISH), fluorescently labeled DNA probes are hybridized to interphase cells or metaphase chromosomes to detect relative location and presence of the targeted sequences (Feuk et al., 2006). Karyotyping allows detection of variation by the scale of chromosomal aneuploidies, and translocations and CNVs involving over 5–10 Mb of DNA, and FISH with resolution of 500 kb (Carvalho, C. M. and Lupski, 2016; Weischenfeldt et al., 2013; Alkan et al., 2011).

FISH is still commonly utilized: as an *in situ* method it can provide accurate location for genomic copies. Advanced imaging and computational methods as well as automatization have streamlined and standardized the analysis workflow (Onozato et al., 2019). Nevertheless, only a few dozen genetic loci can be inspected with a single FISH assay (Onozato et al., 2019). Single-cell sequencing will probably overtake FISH as the standard *in situ* method when its throughput, accuracy and cost-effectiveness have been improved enough (Onozato et al., 2019).

In optical mapping, certain repeating DNA sequences in the genome are labeled by fluorescent markers. After imaging, the sequences can be arranged and aligned, and reference genome is usually used to guide the mapping (Goodwin et al., 2016; Alkan et al., 2011). BioNano Genomics optical mapping platform produces fragments up to 1 Mb in length and enables the detection of commonly elusive variants, such as inversions, novel sequence insertions and translocations. The method enables also locating of extra copies of copy number alterations (Goodwin et al., 2016; Alkan et al., 2011). However, copy number neutral events with breakpoints on centromeres or leading to LOH cannot be detected (Neveling et al., 2020). Since the method lacks base-pair level resolution, it is currently used as a low-cost genome-wide screening approach or as a complementary method to help build a scaffold for *de novo* genome assembly (Goodwin et al., 2016). Following platform improvements with increased throughput, decrease in costs and improved resolution, validation studies for the detection of structural variants in clinical settings are emerging (Neveling et al., 2020).

2.2.2 PCR-related methods

Multiple PCR-based methods have been developed to inspect genomic structure and variation. In digital PCR, the quantity of a single template molecule is measured with the use of fluorescently labeled probes and inspection of the amounts of amplification products compared to test samples (Vogelstein and Kinzler, 1999). Digital PCR allows investigation of both genomic variants and changes in gene expression since both DNA and mRNA can be used as source material. Utilization of one DNA molecule as a starting point for each reaction enables

allele-specific resolution (Salk et al., 2018). High-throughput applications of digital PCR involve distribution of molecules into separate wells, or more commonly into droplets (Weaver et al., 2010; Ito et al., 2019). In droplet digital PCR (ddPCR), each sample is partitioned into droplets for compartmentalized and multiplexed amplification (Ito et al., 2019; Amr et al., 2018). This method enables highly accurate quantification of DNA copy number for amplifications. Low-frequency mutations and closely related pseudogenes can also be differentiated (Ito et al., 2019; Amr et al., 2018; Harel and Lupski, 2018).

In real-time PCR (or real-time quantitative PCR, qPCR), the PCR reaction was conducted originally with gene-specific primers (TaqMan FRET, fluorescence resonance energy transfer), where a probe usually labeled with a fluorescent dye is annealed to the PCR product. After signal detection, the probe is degraded by the Taq polymerase, allowing product amplification to be followed and quantified in real-time (Holland et al., 1991). SYBR Green is a newer and cheaper option for probes needed in the TaqMan assay (Ponchel et al., 2003). SYBR Green dye binds undiscriminatingly into double-stranded PCR products, allowing a signal representing the amount of the product to be monitored after each cycle (Ponchel et al., 2003). Real-time qPCR has higher throughput than digital PCR, but provides only the relative amount of target molecules and is generally suitable for detecting only single deletions and duplications (Weaver et al., 2010; Feuk et al., 2006). qPCR is currently used for genotyping, gene expression analysis, CNV assays and pathogen detection (Goodwin et al., 2016).

In multiplex ligation-dependent probe amplification (MLPA), each probe specific for the target sequence consists of two oligonucleotides (Schouten et al., 2002). Exact match in sequence is required for them to hybridize to adjacent sites on a target sequence. After this, the halves are ligated and only these continuous molecules can be amplificated in PCR, thus discriminating also single nucleotide differences. Universal primer sites in the probes enable PCR amplification with one primer pair, and one of the primers is labeled with a fluorescent dye to enable detection (Figure 4) (Schouten et al., 2002). MLPA provides semi-quantitative multiplexing, since each probe can be incorporated with a stuffer sequence. The resulting different sized products can be differentiated in gel electrophoresis. Up to 50 targets can be investigated in one assay (Schouten et al., 2002; Shen and Wu, 2009; Kerkhof et al., 2017). MLPA does not provide real-time information, but the amounts of amplicons generated correspond to the proportions of the original templates, similar to ddPCR (Shen and Wu, 2009).

MRC Holland (Amsterdam, the Netherlands) provides MLPA kits for most Mendelian disease genes, but custom probe design may also be needed (Shen and Wu, 2009). Because SNPs and secondary structure formation need to be avoided and the probes need to behave similarly in the reaction setting, probe design can be complicated (Shen and Wu, 2009). Additionally, as for all presented PCR methods, the requirement for probes prevents detection of novel sequences. The location or orientation of the amplified regions cannot be resolved either. Polymerase errors limit the sensitivity of the method, so repetitive regions or completely homologous sequences cannot be targeted (Shen and Wu, 2009; Amr et al., 2018; Kerkhof et

al., 2017). The assays are generally cost-effective for a single sample and gene, but only moderately scalable for multiple loci and samples (Ito et al., 2019).

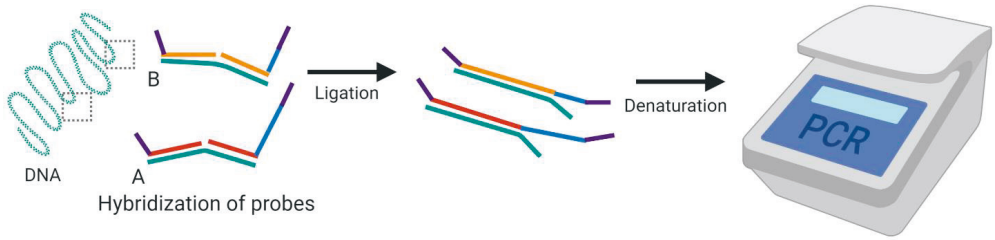

Figure 4: Multiplex ligation-dependent probe amplification (MLPA). Pairs of probes specific to targets A and B are hybridized to adjacent sites in DNA, which allows the probes to be ligated into continuous molecules. The ligation products are flanked by universal PCR initiation sites and contain stuffer sequences of different lengths. After denaturation and PCR, the products can be differentiated by size by gel electrophoresis.

2.2.3 Array-based hybridization methods

Comparative genomic hybridization (CGH) was conventionally performed with metaphase chromosomes, limiting the resolution to 2–5 Mb (Kallioniemi et al., 1992; Shen and Wu, 2009). The basic idea of CGH remains the same regardless of the used target for hybridization and source of DNA material (Figure 5). DNA is extracted from a test sample, such as blood or skin, and a normal reference DNA is required as well. The samples are labeled differently with two fluorescent dyes and after denaturation they are hybridized as single-stranded DNA into a panel of DNA targets. Differences in hybridization between the test DNA and the normal reference, regions of loss or gain, can be detected as changes in the ratio of the fluorochrome intensities. They are captured and quantified with digital imaging systems. Extremely high-copy amplifications, such as in cancer, can be detected as a change in the intensity of one color (Kallioniemi et al., 1992; Pollack et al., 1999). In signal analysis, the background noise needs to be subtracted from the results, and the ratio calculations are normalized across the array. Alternatively, a platform-specific reference is used to account for platform-specific artefacts and biases (Pollack et al., 1999; de Leeuw et al., 2011).

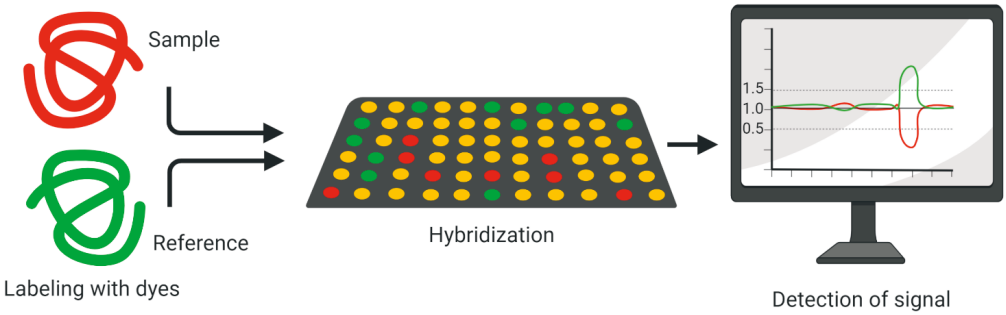

Figure 5: Array comparative genomic hybridization. The test and reference DNA are labeled with different fluorescent labels and hybridized to target DNA on array. The fluorescence intensities are imaged and analyzed for differences, which reveal regions of loss or gain.

As an advantage of CGH, DNA did not have to originate from cells undergoing division (Pollack et al., 1999). Sources for hybridization targets in CGH include genomic clones from

bacterial or phage artificial chromosomes (BAC and PAC), cDNA, PCR products and oligonucleotides (Feuk et al., 2006). Adaptation of CGH into an array enabled higher resolution, easier procedures for sample preparation and analysis, and parallelization into multiple inspected genomic loci (Solinas-Toldo et al., 1997). The first microarrays based on cosmids, PAC, BAC, or cDNA had an average resolution of 1–1.5 Mb (Pollack et al., 1999; Carvalho, B. et al., 2004; Schena et al., 1995). They were designed for pre-known targets with locally high resolution (> 40 kb) (Pinkel et al., 1998). Compared to the first utilized target molecules, synthetic oligonucleotides were cheaper and faster to produce and could be flexibly targeted for any part of the genome (Carvalho, B. et al., 2004).

Array CGH can be used to analyze structural variants, DNA-protein interactions or expression levels by measuring gene-specific cDNA (Goodwin et al., 2016). The first large-scale human genome CNV prevalence studies were performed with oligonucleotide and BAC-based microarrays (Iafrate et al., 2004; Sebat et al., 2004). With these approaches, small variants (< 1 kb) were largely missed, exact breakpoints were indiscernible and many CNVs were overestimated in size (Alkan et al., 2011; Zarrei et al., 2015). Since the CNVs were mostly assessed from healthy subjects and lacked resolution for definite breakpoints, inferring the exact genomic content and clinical significance for the discovered variants was challenging (Alkan et al., 2011).

SNP array is a variation of the array-based CGH approach. Hybridization signal intensities from spotted oligonucleotides on SNP arrays are compared with average values from controls (de Leeuw et al., 2011). SNP arrays provide higher resolution than array CGH. They are commonly used to identify polymorphisms associated with diseases and phenotypes, and in genome-wide association studies (Goodwin et al., 2016; Zhou et al., 2018; Roca et al., 2019). SNP arrays have lower signal-to-noise ratio than array CGH, but the method provides also genotype information, thus revealing regions of loss of heterozygosity, which could signify a deletion or uniparental disomy (Harel and Lupski, 2018; de Leeuw et al., 2011; Alkan et al., 2011). Therefore, the population under study needs to be taken into account more closely in the probe design for SNP arrays than for array CGH (Alkan et al., 2011). Array CGH and SNP arrays are not suitable for detecting the location or orientation of amplificated genomic stretches, absolute differences in higher levels of amplification, exact breakpoints at a base-pair level, copy-neutral rearrangements or low-level mosaicism (South et al., 2013; Harel and Lupski, 2018; Alkan et al., 2011). Completely new inserted sequences not represented in the existing human reference genome cannot be detected either (South et al., 2013; Alkan et al., 2011). Generally, deletions seem to be easier to detected with these approaches than duplications, especially if they are small in size (Zarrei et al., 2015; Alkan et al., 2011).

Array target designs may be distributed genome-wide evenly, targeted to a certain region of interest, or include a combination of these two with varying overall probe distribution and resulting local resolutions (South et al., 2013; Sagath et al., 2018). The high-accuracy microarrays used routinely in clinical diagnostics (> 100 kb resolution) have a limited resolution

to detect smaller CNVs (Redin et al., 2017; Yao et al., 2017). Most accurate off-the-shelf arrays can reach > 1 kb resolution but with an average detection limit of 20 kb (Zhou et al., 2018; Roca et al., 2019; Whitford et al., 2019; Marchuk et al., 2018). Resolution of 500 bp is achievable with highly specific designs - specifically, local increase in the amount of probes - but arrays are still not sensitive enough to detect smaller CNVs (Trost et al., 2018). Nevertheless, microarray-based CNV analysis remains the first-tier approach in clinical investigations of unsolved cases with developmental delay or intellectual disability, autism spectrum disorders, multiple congenital anomalies, and cancer (Zhou et al., 2018).

2.3 Massively parallel sequencing

In Sanger sequencing, dye-labeled deoxyribonucleotide triphosphates (dNTP) and dideoxy-modified dNTPs are included in a mix with an appropriate ratio. During the progress of a standard PCR, some elongating strands get incorporated with dideoxy-modified dNTPs preventing further elongation. The resulting mixture is fractionated by size by gel electrophoresis. The terminal base in each strand is identified by laser excitation and spectral emission analysis, revealing the DNA sequence (Sanger et al., 1977). Nowadays, Sanger sequencing is largely automated (Ambardar et al., 2016). Similarly to the many presented other methods, Sanger sequencing is usually limited to small genomic regions and challenging to scale or transfer to another loci (Salk et al., 2018).

Beginning in 2005, release of the first high-throughput sequencing platforms enabled human genome sequencing with over 50,000-fold drop in the costs compared to the Human Genome Project (Goodwin et al., 2016; Margulies et al., 2005). Other novel approaches for sequencing were developed soon after (Quail et al., 2012; Eid et al., 2009; Rothberg et al., 2011; Bentley et al., 2008). For a while, “next generation sequencing” was the standard term for these sequencing platforms. Some newer platforms, with sequencing of a single DNA template the first common factor, were called “third generation sequencers” (Ambardar et al., 2016; Zhao et al., 2013; Alkan et al., 2011). However, the release papers describing the first platforms used the term massively parallel sequencing (MPS), which is a more descriptive term for these methods (Rothberg et al., 2011; Bentley et al., 2008). The current trend is to again use MPS and other more descriptive terms for the different sequencing methods. Therefore, the term “next generation sequencing” will not be used further.

The basic definition for MPS is sequencing of multiple DNA templates from the same sample in a single run. Since no sequencing approach is flawless in performance yet, each genomic location is sequenced multiple times to provide a strong signal against background noise and to distinguish variants from errors, thus increasing variant detection sensitivity and accuracy (Lappalainen et al., 2019; Ambardar et al., 2016; Goodwin et al., 2016). Sequencing depth, or read depth refers to the number of reads, a layer, covering a certain genomic position (Salk et al., 2018). A consensus sequence is acquired by aligning sequencing reads and determining the most likely base at each position (Goodwin et al., 2016). With whole genome sequencing

(WGS), all genomic sequences are targeted for sequencing, while whole exome sequencing (WES) is usually limited to the coding part of the genome, the exons in genes. In targeted gene panel sequencing, a certain set of genes is targeted for sequencing, and usually only exons (Goodwin et al., 2016; Evila et al., 2016).

For the sequencing step itself, multiple chemistries and platforms have been developed (Ambardar et al., 2016; Goodwin et al., 2016; Lappalainen et al., 2019). The succession of more longer-lived and popular different sequencing platforms and providers include GS FLX from 454 Life Sciences/Roche Diagnostics, Genome Analyzer, HiSeq, MiSeq, NextSeq and Novaseq from Illumina, SOLiD from ABI, Ion Torrent from Life Technologies, SMRT Sequencing with PacBio RS and PacBio RS II from Pacific Biosciences, Nanopore sequencing from Oxford Nanopore, Complete Genomics from Beijing Genomics Institute, Qiagen GeneReader, and GnuBIO from BioRad (Ambardar et al., 2016; Goodwin et al., 2016; Lappalainen et al., 2019). Since the release of the original MPS platforms, huge improvements in sequencing speed and decrease in the cost of sequencing have been achieved (Ambardar et al., 2016). For the study in this thesis, sequencing was performed with target enrichment with hybridization for targeted gene panels or WES, and exclusively Illumina platforms were used.

2.3.1 Short-read sequencing platforms

Short-read MPS together with resolving the human reference genome provided faster and more comprehensive means to study the genome and genomic variants at a relatively lower cost (Bentley et al., 2008). The platforms generally provide reads of dozens to several hundreds of nucleotides in length (Salk et al., 2018; Goodwin et al., 2016). The conventional clinical short-read sequencing approaches include extraction of DNA from samples, DNA quality assessment, normalization of DNA concentration, and sequencing library preparation with DNA fragmentation to provide overlapping fragments with random distribution. This is followed by ligation of sequencing and amplification primers required for the initiation of the sequencing reactions, library amplification typically with PCR, and library quality assessment before sequencing (Clark et al., 2019; Ma et al., 2019; Salk et al., 2018; Goodwin et al., 2016). The template can originate from one DNA strand for single-end sequencing or from both strands to provide paired-end reads (Goodwin et al., 2016).

In targeted approaches, a system of capture or amplification isolates or enriches the targeted regions (Hodges et al., 2007; Goodwin et al., 2016). For DNA enrichment by hybridization, high-molecular weight DNA is first fragmented, the ends are repaired into blunt ends and phosphorylated, and the strands are denatured and captured. The selected fragments are enriched with PCR together with ligation of sequencing adapters at the latest in this stage (Hodges et al., 2007). Target enrichment by hybridization is more time consuming but generally easier to design and more widely used than amplicon-based methods with target isolation based on targeted PCR (Salk et al., 2018).

An index sequence, a DNA nucleotide code, may be attached to all molecules within a DNA

sample to allow for multiplexing different samples in a single sequencing run (Schirmer et al., 2016; Salk et al., 2018). A molecular barcode or unique molecular identifier, UMI, can be added to the individual DNA molecules to recognize copies, which originate from the same founder molecule in PCR amplification. This enables consensus-based error correction (Salk et al., 2018). In practice, the molecular barcodes are artificial sequences incorporated into sequencing adapters or PCR primers and alternatively or combined with random shearing of the DNA molecule ends (Salk et al., 2018). UMI incorporation into one adapter strand is one of the most easily implemented and popular approaches for consensus building in sequencing (Salk et al., 2018).

The DNA template is provided for short-read sequencing as groups of localized monoclonal clusters generated through amplification (Hodges et al., 2007). This amplification is generally based on PCR approaches: either bead-based with emulsion PCR, where emulsion droplets sequester templates during the process, or solid-phase, where templates are bound to a surface (Goodwin et al., 2016). In the Illumina bridge amplification, the DNA template is bound to a solid surface with a free end, which interacts with a nearby primer. This forms a bridge structure and enables PCR to create a complementary strand (Figure 6) (Goodwin et al., 2016; Bentley et al., 2008). After amplification, the template is immobilized and ready for sequencing reactions carried with fluid reagents streamed and flushed away sequentially (Bentley et al., 2008). Patterned flow cells increase sequencing throughput by having better spatial resolution, which enables higher density for DNA templates while maintaining clonal integrity (Goodwin et al., 2016).

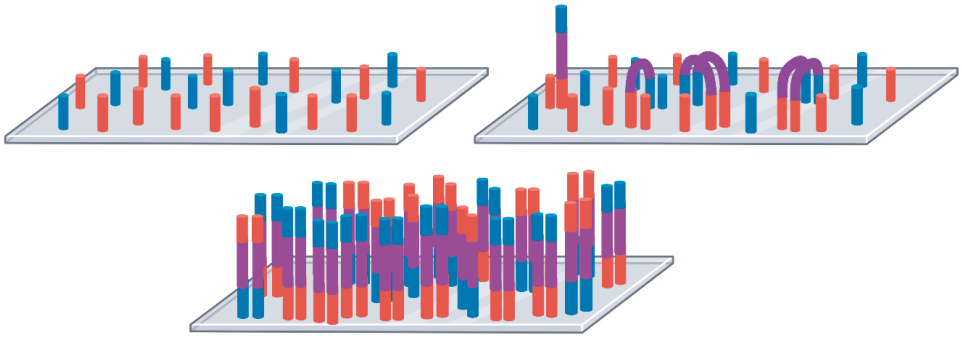

Figure 6: Bridge amplification of DNA templates. Primers allowing amplification in both directions are bound to a flow cell surface. The DNA templates have adapters attached to both ends, which recognize these primers and pair with them forming bridges, which allows template extension by PCR.

Emulsion PCR (emPCR) offers droplet compartments for multiple simultaneous PCR reactions without molecule exchange between the droplets (Shao et al., 2011). In bead-based amplification, one adapter attached to the template is complimentary to an anchor on the surface of a bead, and the other initiates emPCR. Each clonal DNA fragment remains immobilized on the same single bead (Figure 7) (Shao et al., 2011). The beads can be enriched for example with a magnetic process, after which the beads are usually distributed into wells. The well depth restricts the bead amount to one per well (Rothberg et al., 2011). Originally, emPCR was the

first novel method for isolating and amplifying DNA fragments *in vitro* as an alternative to subcloning in bacteria (Margulies et al., 2005). Bacterial cloning hosts were intolerant to some extreme base-compositions, genes or inverted repeats, introducing eventual sequencing bias (Aird et al., 2011).

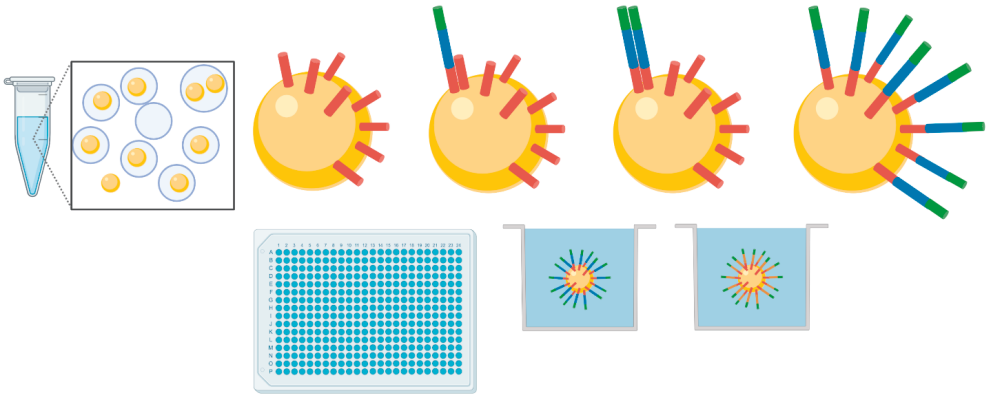

Figure 7: Bead-based amplification. Anchors are attached to the surface of beads suspended in emulsion droplets. The DNA templates have an adapter, which recognizes an adapter on the bead surface, and a primer to initiate PCR. After amplification, templates are removed to leave single-stranded templates, which are distributed into wells for sequencing.

Two general approaches have been applied for short-read sequencing: sequencing by ligation and sequencing by synthesis (Goodwin et al., 2016; Ambardar et al., 2016).

2.3.1.1 Sequencing by ligation

In sequencing by ligation, an anchor fragment in the probes is complementary to an adapter sequence attached to the DNA template, which provides a site for initiating ligation. The other part of the probe encodes one base or a base-pair, which ligates to the DNA template and carries a detectable signal source, such as a fluorophore (Goodwin et al., 2016; Ambardar et al., 2016).

In sequencing by oligonucleotide ligation and detection (SOLiD) method by Applied Biosystems, dinucleotide probes with fluorophores are utilized with each signal corresponding to two bases (Goodwin et al., 2016; Valouev et al., 2008). The library preparation employs emPCR on beads, which are covalently bound to a glass plate (Valouev et al., 2008). Circularization of the DNA fragments provides a paired-end library, which makes the SOLiD platform suitable both for single-end and paired-end sequencing (Valouev et al., 2008). The newest SOLiD platforms utilize solid-phase template walking for template amplification, which resembles Illumina bridge amplification (Goodwin et al., 2016). The sequencing process involves repeated rounds of hybridization and ligation of a primer with a tail of degenerate bases, hybridization of a di-base probe labeled with a fluorophore, detection by fluorescence imaging, then cleavage of the fluorophore and some of the degenerate part (Figure 8). This leaves the probe and a tail in place, allowing identification of two out of every five bases. After rounds of extension, all primers are removed and the cycle starts over with an offset anchor in the beginning to cover a new section (Goodwin et al., 2016; Valouev et al., 2008).

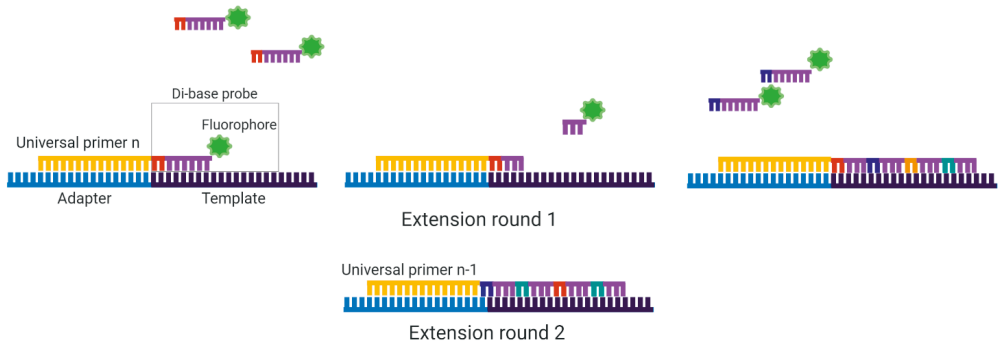

Figure 8: Sequencing by ligation, SOLiD di-base probe approach. In the first round, a universal primer hybridizes to the adapter, providing a free end for a di-base probe for ligation. After probe hybridization, the fluorophore signal is imaged, fluorophore and a part of the probe is cleaved, and the next di-base probes are provided. After the first round, the probe sequence is cleaved off and a universal primer with different offset is provided to allow sequencing of the other sections.

BGI (Beijing Genomics Institute) utilizes solution enrichment with The Complete Genomics technology, where double-stranded DNA template is iteratively ligated, circularized and cleaved to create a circular template with four adaptors (Goodwin et al., 2016; Drmanac et al., 2010; Xu et al., 2019). Then, rolling circle amplification is used to produce multiple copies of the single-stranded template head-to-tail formed into discrete DNA nanoballs, which can be distributed onto a patterned slide (Figure 9) (Goodwin et al., 2016; Drmanac et al., 2010). For sequencing, Complete Genomics uses combinatorial probe-anchor ligation (cPAL), or probe-anchor synthesis (cPAS). cPAL is based on unchained hybridization and ligation technology, where degenerate anchors and a probe are used to read bases adjacent to adapter sites at several locations of a DNA nanoball template simultaneously (Drmanac et al., 2010). After imaging, the probe-anchor complexes are removed and in the subsequent cycles new combinations are hybridized, where the known base is included in the probe part (Goodwin et al., 2016; Drmanac et al., 2010). Compared to many other MPS sequencing approaches, cPAL has lower reagent costs, since the process is not dependent on preceding incorporated nucleotides and tolerates low-quality base incorporations, thus also avoiding error accumulation (Drmanac et al., 2010). cPAS, polymerase-based cycle sequencing, provides longer reads than cPAL (Fehlmann et al., 2016). However, in availability cPAL is limited to a service platform for WGS and cPAS BGISEQ-500 in mainland China (Goodwin et al., 2016).

Figure 9: Rolling circle amplification. DNA template is first circularized, and then four different adapters are attached. With continuous PCR of the circularized template, a concatemer with multiple DNA templates is produced and distributed as nanoballs into a slide with cohesive forces keeping the templates together.

Sequencing by ligation techniques used by SOLiD and Complete Genomics systems have generally a high accuracy of approximately 99.99% (Goodwin et al., 2016). The greatest disadvantage for both of the techniques is the read length, with maximum of 75 bp for SOLiD and 28–100 bp for Complete Genomics, limiting their usage especially in genome assembly and structural variation detection (Goodwin et al., 2016; Ambardar et al., 2016). Additionally, SOLiD has comparably long runtime of several days (Goodwin et al., 2016; Ambardar et al., 2016). Because of these shortcomings and advances in other sequencing technologies, manufacturing of SOLiD has been discontinued (Salk et al., 2018).

2.3.1.2 Sequencing by synthesis

In sequencing by synthesis, DNA is synthetized with a DNA polymerase, and a signal representing incorporation of each nucleotide into the growing strand can be detected (Goodwin et al., 2016; Ambardar et al., 2016). Roche 454 was the first true MPS platform and utilized pyrosequencing (Margulies et al., 2005). The platform was seminal as it was released at a time when the cost estimation for sequencing a human genome was between $10 and $25 million dollars (Margulies et al., 2005). For this platform, the DNA template is prepared with emPCR on beads and then loaded into individual wells on a slide, referred as picoliter-scale sequencing reactors (Margulies et al., 2005). Four bases are provided and removed sequentially, and nucleotide incorporation is detected indirectly by an enzymatic cascade from the release of inorganic pyrophosphate with an eventual bioluminescence signal. At homopolymer regions, the incorporation of multiple nucleotides results in a proportional increase in the signal (Margulies et al., 2005). The instrument provided reads for single-stranded template DNA with an eventual read length of 700 bp (Margulies et al., 2005; Goodwin et al., 2016). However, the problem of de-synchronization in sequencing results in dominating indel errors and inaccurate homopolymer sequencing of more than five identical nucleotides (Margulies et al., 2005; Ambardar et al., 2016). Eventually, the sequencing platform did not keep up with other technologies with comparatively high sequencing cost and was discontinued in 2016 (Ambardar et al., 2016; Goodwin et al., 2016).

The Ion Torrent platform originated from the idea that DNA sequencing was limited by requirements for imaging technology, modified nucleotides and electromagnetic intermediates, such as light (Rothberg et al., 2011). The platform implements a method of DNA sequencing (called Ion Torrent sequencing or semiconductor sequencing) based on direct sensing of hydrogen ion release in template-directed DNA polymerase synthesis. Therefore, no altered bases, enzymes or optical detection are needed (Rothberg et al., 2011). The DNA templates are prepared with emPCR and binding of sequencing primers and DNA polymerase on the surface of beads, which are loaded into proton-sensing wells (Rothberg et al., 2011). In the sequencing reaction, the four nucleotides are introduced sequentially, and the pH shift indicative of base incorporation is converted to a digital signal by off-chip electronics (Rothberg et al., 2011). The change in pH is limitedly proportional to the number of nucleotides incorporated, allowing for a limited accuracy with homopolymer lengths (Rothberg et al., 2011). Indel errors dominate in Ion Torrent sequencing data with the deletion rate increasing with the homopolymer length

(Ambardar et al., 2016; Ross et al., 2013).

Much like pyrosequencing, Ion Torrent sequencing is based on identification of a signal representing incorporation of a nucleotide into the elongating strand. With this approach, each of the four nucleotides must be presented separately to ensure only one incorporation event, and the nucleotides are not blocked to allow elongation (Goodwin et al., 2016). Ion Torrent platform provides relatively long reads of 400 bp similar to pyrosequencing, which is an advantage in applications focusing on repetitive or complex DNA (Goodwin et al., 2016). But unlike pyrosequencing, Ion Torrent platform has kept up better with yield and economic efficiency since the platform uses non-modified bases and a single polymerase enzyme (Goodwin et al., 2016; Ambardar et al., 2016). Additionally, Ion Torrent prevails as a relatively fast sequencing platform well-suited for point-of-care clinical applications, such as gene-panel and transcriptome sequencing and splice site identification (Goodwin et al., 2016).

Illumina sequencing utilizes cyclic reversible termination similar to Sanger sequencing with blocked ribose 3' hydroxyl group in the incorporated nucleotides preventing elongation (Figure 10) (Bentley et al., 2008). The basic Illumina protocol involves DNA template preparation with fragmentation, end-fixing, attachment of adapters, and template denaturation into single-stranded DNA for annealing to complementary oligonucleotides on a flow cell surface. This is followed by solid-phase bridge amplification to produce high-density template colonies (Bentley et al., 2008). The sequencing chemistry involves four reversible terminators with different fluorophores and 3'-modified ends to avoid over-incorporation (Bentley et al., 2008). Therefore, all nucleotides can be provided simultaneously in each cycle to incorporate a single nucleotide into each strand followed by washing of unbound nucleotides and label imaging with fluorescence microscopy (Bentley et al., 2008). The incorporation is carried out with a modified DNA polymerase, which improves incorporation of modified nucleotides (Bentley et al., 2008). After imaging, the fluorescent dye is removed and the 3' hydroxyl group is regenerated to start a new cycle (Bentley et al., 2008).

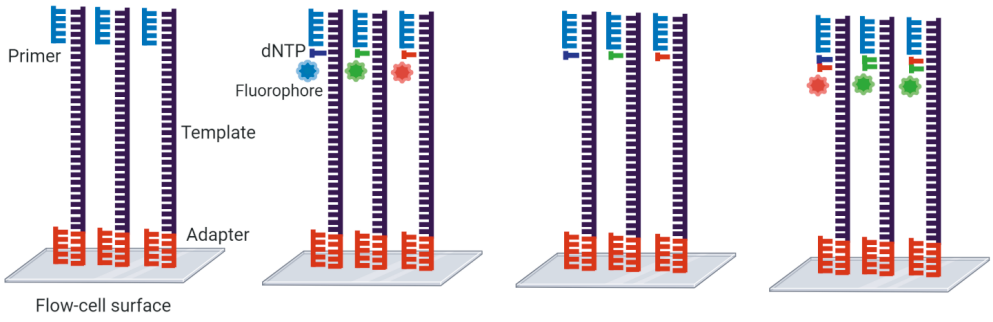

Figure 10: Illumina sequencing with cyclic reversible termination. DNA templates are attached to the surface of a flow cell with an adapter in one end and a primer in the other end enabling nucleotide incorporation. The four nucleotides are labeled with different fluorophores and can be provided simultaneously. After nucleotide incorporation, fluorophore signal is imaged, and then the fluorophore is cleaved to provide a free end for the next nucleotide incorporation.

Most Illumina platforms use nucleotides labeled with four different fluorophores requiring as many different imaging channels, whereas NextSeq, MiniSeq and Novaseq use a two-fluorophore system, where two bases have one fluorophore (C and T), one has either of them (A) and one has no fluorophore. This enables their differentiation with two imaging channels and thus faster and cheaper overall sequencing with fewer cycles (Goodwin et al., 2016; Bentley et al., 2008; Ambardar et al., 2016). The two-channel system tends to have slightly more errors and underperformance for low-diversity regions due to more ambiguous base discrimination. This can be compensated for in computational steps (Goodwin et al., 2016). Originally, Illumina sequencing provided only a read length of 35 bp, but one of its advantages over the other platforms was the ability to provide paired-end reads from the start (Bentley et al., 2008; Ambardar et al., 2016). For paired-end sequencing, the template DNA is first converted into double-stranded DNA, and then the original strand is removed to provide a template for the complementary strand (Bentley et al., 2008). Illumina sequencing is less susceptible to homopolymer errors observed with nucleotide addition approaches and provides an accuracy of > 99.5% but displays more substitution errors (Goodwin et al., 2016).

Some short-read sequencing platforms aim for certain applications or a comprehensive, concise workflow, rather than investing in high throughput and accuracy. Qiagen GeneReader with a several-day runtime is focused on cancer gene panels and intended to be a clinical device, similar to Illumina MiSeq, but with potentially lower cost per sequencing yield unit (Goodwin et al., 2016). This system has QIAcube sample preparation system and the Qiagen Clinical Insight variant analysis platform integrated and thus incorporates all steps from sample preparation to analysis. The platform uses also labeled nucleotides but without sequential incorporation, enabling just high enough accuracy for achieving identification (Goodwin et al., 2016). GnuBio is a droplet-based DNA sequencing platform, which utilizes microfluidics and emulsion technology. A single GnuBio instrument provides a streamlined workflow from library preparation to DNA sequencing, and the analysis is conducted inside the same droplets, reaction vesicles, decreasing reagent costs (Ambardar et al., 2016).

Illumina platforms have long dominated the market for sequencing instruments (Rieber et al., 2013; Goodwin et al., 2016). Illumina's keys to success include refined technology with up to 300 bp read length, cross-platform compatibility and versatility of platforms in capacity, runtime, read structure and read length, from both small low-throughput benchtop units to large ultra-high-throughput instruments (Goodwin et al., 2016). Currently, Illumina HiSeq and NovaSeq outdo the other approaches in cost, ease of use and accuracy. Consequently, most human genetics studies turn to these short-read platforms (Lappalainen et al., 2019). The popularity of Illumina platforms in MPS research has raised concerns about systematic errors in the sequencing data (Schirmer et al., 2016). In some new approaches, multiple sequencing methods with complementary strengths are integrated, involving and providing also longer read lengths (Goodwin et al., 2016).

2.3.2 Long-read sequencing platforms

Two general approaches have been developed for producing long sequencing reads: single-molecule real-time sequencing, and synthetic approaches based on existing short-read technologies followed by construction of long reads *in silico* (Goodwin et al., 2016). Long reads from these platforms are generally several kbs to tens of kbs in read length (Salk et al., 2018; Goodwin et al., 2016). Compared to prevailing short-read MPS technologies, current true long-read sequencing platforms provide longer reads but generally with a notably lower raw read accuracy, higher costs and lower throughput (Goodwin et al., 2016; Salk et al., 2018; Lappalainen et al., 2019). Therefore, either higher read depth or combining with short-read information is needed for increasing the accuracy (Madoui et al., 2015). Single-molecule approaches lack the step of amplification of DNA into a clonal population, which makes use of indices or UMIs unnecessary. Additionally, no pause or chemical cycling is required, which enables real-time sequencing (Salk et al., 2018; Goodwin et al., 2016).

2.3.2.1 Synthetic long-read sequencing

In synthetic approaches, no actual long reads are generated. During library preparation, barcodes are attached through ligation or amplification to each single read, which after sequencing with existing short-read sequencers allows computational reassembly into the original larger fragment (Goodwin et al., 2016). 10X Genomics has developed GemCode and a newly released Chromium platform for pre-sequencing reactions in synthetic long-read sequencing (Marks et al., 2019). In these microfluidic instruments, the DNA starting material is distributed as up to 100 kb fragments to droplets containing beads, adapters and one unique barcode (Marks et al., 2019; Goodwin et al., 2016; Ambardar et al., 2016). The barcoded short-read libraries can be assembled into original long molecules, linked-reads, after sequencing with a standard Illumina short-read platform (Marks et al., 2019). Aligning and stacking linked-reads from the same loci provides continuous coverage, which can span 50 kb (Marks et al., 2019; Goodwin et al., 2016). Data output is limited partially by the number of barcodes used. Additionally, inefficient partitioning can lead to an excess of DNA fragments within a droplet, complicating sequence deconvolution. Together, these lead to ambiguity in positioning reads sharing the same barcode (Goodwin et al., 2016).

Illumina has developed the TruSeq synthetic long-read method (McCoy et al., 2014). With this method, pre-sequencing reactions are partitioned to microtiter plate wells containing individual barcodes. Thus, the pre-sequencing reactions can be performed without special instruments (McCoy et al., 2014). After DNA pooling and sequencing with standard short-read pipelines, the data is demultiplexed *in silico* by barcode sequences to trace the molecules and assemble synthetic long-reads (McCoy et al., 2014). This local assembly of short-read data has provided read lengths of 1.5–8.5 kb (McCoy et al., 2014). The method has a low error rate of 0.03% per base due to inherent consensus build step, which is a lower error rate than for Illumina short reads (McCoy et al., 2014).

2.3.2.2 Single-molecule real-time sequencing and nanopore sequencing

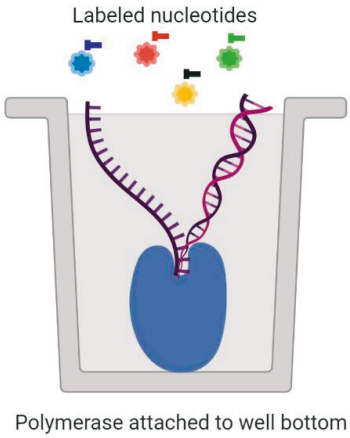

Figure 11: Single-molecule real-time (SMRT) sequencing utilized by PacBio platforms. DNA polymerase is bound to the bottom of a well and performs uninterrupted (real-time) template-directed synthesis. All nucleotides are labeled differently with fluorophores and provided simultaneously.

The Pacific Biosciences (PacBio) platform is based on single-molecule real-time (SMRT) sequencing (Eid et al., 2009). PacBio utilizes rolling circle templates, where the strand-displacing capability of the polymerase and two hairpin adapters connecting double-stranded DNA templates allow continuous circular sequencing (Eid et al., 2009; Goodwin et al., 2016). Rolling circle amplification is more time-consuming and laborious than amplification on a flow cell or bead surface, but linear DNA amplification prevents error accumulation as compared to exponential DNA amplification (Fehlmann et al., 2016). The rolling circle templates can be sequenced in a single run, providing a consensus sequence from one single DNA molecule source (Eid et al., 2009; Lappalainen et al., 2019). A modified DNA polymerase molecule is bound to a single-stranded DNA template and attached to the bottom of a zero-mode waveguide, a transparent picolitre well, along with immobilized enzymes (Figure 11). The polymerase performs uninterrupted template-directed synthesis with each incorporation of the fluorescently labeled nucleotides directly observed as color and duration of emitted light (Eid et al., 2009). A phosphodiester bond catalyzed by the polymerase releases the fluorophore, leaving natural DNA behind (Eid et al., 2009).

With PacBio, DNA can be sequenced in a native form without cloning or amplification. The platform provides typically reads exceeding 50 kb read length, with an average of 10–15 kb (Chaisson, M. J. et al., 2015; Goodwin et al., 2016). PacBio has the potential to directly detect some DNA changes, such as DNA binding proteins, DNA polymerase inhibitors and base methylation (Goodwin et al., 2016; Eid et al., 2009). However, the single pass error rate can be as high as 15% with indel errors more prominent (Goodwin et al., 2016; Quail et al., 2012; Eid et al., 2009). Inaccurate quantification of the incorporation event intervals presents as deletion errors. Disassociation of the nucleotide from the active site prior to phosphodiester bond formation or non-incorporated nucleotides remaining in the active site result as insertion errors, which is the dominant error for PacBio (Eid et al., 2009; Ambardar et al., 2016; Quail et al., 2012). A sufficiently high coverage with approximately 99.999% accuracy reached with ten reads overcomes this since the errors are randomly distributed (Goodwin et al., 2016; Quail et al., 2012; Eid et al., 2009). Read depth as high as 40X has also been used for error correction (Chaisson, M. J. et al., 2015). PacBio RSII has been shown to be ideal for full-length transcript sequencing, *de novo* genome assembly and resolving complex long-range genomic structures, but the platform has high running costs and a limited throughput (Goodwin et al., 2016; Quail et al., 2012; Chaisson, M. J. et al., 2015).

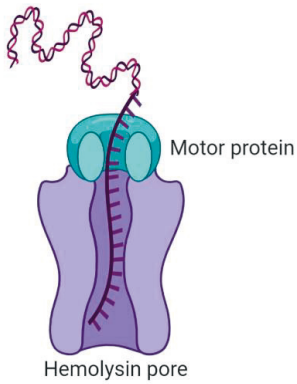

Figure 12: Oxford nanopore sequencing. DNA and current are fed through alpha-hemolysin with a motor protein regulating the process. Shifts in voltage signal different DNA sequence structures.

The first nanopore sequencer platform released by Oxford Nanopore Technologies was MinION (Goodwin et al., 2016). No secondary signals, such as light, color or pH are followed in this sequencing approach. Specifically, DNA composition of a native single-stranded DNA molecule is detected directly (Kasianowicz et al., 1996). Current and DNA are passed through a large biological pore capable of sensing DNA, alpha-hemolysin. The magnitude and duration of the shifts in voltage for the DNA sequence located in the pore represent overlapping consecutive k-mers of DNA (sizes range from 3 to 6 bases) (Figure 12) (Goodwin et al., 2016; Kasianowicz et al., 1996). The instrument can recognize more than 1000 possible k-mers and some modified bases as well, such as 5-methylcytosine (Goodwin et al., 2016; Jain et al., 2018). A leader-hairpin library structure attached during the library preparation links two strands of DNA and enables passing both through the pore, which generates paired-end reads (Salk et al., 2018; Goodwin et al., 2016). An additional leader sequence added in the library preparation interacts with the pore together with a motor protein to direct DNA movement (Cherf et al., 2012).

Nanopore has a high raw sequence error rate of over 30% for one read with indel errors overrepresented. This is probably attributable to the notable amount of distinct signals (Goodwin et al., 2016; Madoui et al., 2015; Ambardar et al., 2016). The platform is also inaccurate for homopolymers longer than the k-mer size with difficulties in distinguishing them (Goodwin et al., 2016). Additionally, some modified bases pose problems, but improvements in the chemistry and base calling algorithms are under development (Goodwin et al., 2016). Nevertheless, the USB-based MinION is a small and a relatively low-cost device and can be run off a personal computer. This gives it superior portability over other sequencing platforms (Goodwin et al., 2016; Madoui et al., 2015). Even cell lysate can be directly sequenced in real-time with no PCR amplification or chemical labeling steps needed in the library construction (Ambardar et al., 2016). Therefore, this platform may have most utility in rapid clinical responses and field locations, such as in rapid pathogen profiling (Goodwin et al., 2016). Another Nanopore platform, PromethION, can compete in throughput with Illumina HiSeq X (Goodwin et al., 2016). Recently, up to 882 kb ultra-long reads have been produced with a Nanopore platform, made possible through improvements to the pore, library preparation techniques, sequencing speed and software (Jain et al., 2018). Future plans include enabling passing of a DNA molecule back and forth through the pore to increase accuracy by more rigorous consensus building (Salk et al., 2018).

2.3.3 MPS data error sources and computational data analysis

Precise MPS data error rates and types vary depending on the sample properties, such as DNA damage, the sequencing platform and chemistry as previously described, and the sequencing

data analysis protocol (Salk et al., 2018). Sequencing errors have remained approximately 0.1–1.0% for all short-read sequencing platforms (Ma et al., 2019; Salk et al., 2018; Ambardar et al., 2016; Lappalainen et al., 2019). For example, Illumina NovaSeq has a similar error profile as compared to older HiSeq platforms (Ma et al., 2019). With coverage bias the reads are non-uniformly distributed across the targeted genome region (Ross et al., 2013).

Numerous steps in DNA preparation for sequencing contain error sources. Tissue processing and storage, DNA isolation or DNA fragmentation can cause DNA damage and nucleotide conversions. The sources include both normal cell processes and environmental exposures, such as chemical extraction, heating or clinical sample stabilization methods, such as formalin fixation (Salk et al., 2018; Ma et al., 2019). These changes can be alleviated to some extent with treatments, which excise or fix the modified bases, but this may decrease amplification efficiency on the regions, and repair enzymes themselves can introduce errors (Salk et al., 2018; Chen et al., 2017).

After DNA extraction, DNA fragmentation can be performed with physical or mechanical methods, such as ultra-sonication or nebulization with compressed nitrogen or air (Tanaka et al., 2020). Ultra-sonication creates even cuts across DNA. This simplifies fragment size control, but oxidative damage associated with the treatment can cause base conversions (Tanaka et al., 2020; Ma et al., 2019; Salk et al., 2018). Several commercial library preparation kits utilize restriction enzyme digestion with endonucleases or transposases for DNA fragmentation (Quail et al., 2012; Tanaka et al., 2020). Usually, high-fidelity polymerases are used in the amplification steps of library preparation, but lower-fidelity polymerases used for repair and A-tailing may present errors (Salk et al., 2018; Tanaka et al., 2020). Additionally, enzymatic DNA fragmentation may produce low-level artefacts with nicks or incomplete cleavage, which leave non-blunt ends more prone to damage and copying errors by lower-accuracy polymerases in the subsequent steps (Salk et al., 2018). Correcting algorithms could be used to ignore some of these artefacts in the sequencing data analysis stage (Tanaka et al., 2020).

PCR is possibly the most error prone step in sequencing protocols, especially in library amplification (Aird et al., 2011). PCR amplification steps may introduce bias with incorporation errors, and most notably, GC coverage bias, discussed in depth further (Aird et al., 2011; Benjamini and Speed, 2012).

Target enrichment can introduce sequencing bias affecting eventual target coverage. Preferential capture of reference alleles in the hybridization and capture, which has been observed, could be mitigated by using alternate allele target probes, but the problem would persist for rare variants (Meynert et al., 2014; Sulonen et al., 2011). Longer fragments may be captured with higher specificity, and some off-target sequences are usually captured as well (Hodges et al., 2007). For regions with less unique genomic alignment, designing probes for accurate capture is also challenging (Meynert et al., 2013). In emPCR, the relative frequency of templates can be distorted by multiple beads per droplet bias or removal of multitemplated

beads (Valouev et al., 2008). False index pairs may be generated during oligonucleotide synthesis, template amplification, template colony formation or sequencing (Kircher et al., 2012). They may arise from spontaneous index swapping, cross-contamination or PCR jumping between samples (Kircher et al., 2012). Double-indexing with UMIs in both adapters or physical linkage of complementary DNA strands help to identify and exclude sequences with mixed indices (Kircher et al., 2012; Salk et al., 2018).

In sequencing by synthesis methods, relative sequencing efficiency can be affected by the usage of engineered polymerases and modified bases (Schirmer et al., 2016). Error rates may increase towards the ends of the reads because of chemical molecule residues. They may perturb DNA polymerase, impair stability of the DNA and hinder substrate recognition and primer extension, limiting the possible read length (Schirmer et al., 2016). Interruptions in enzyme function in stepwise sequencing systems are thought to be deleterious for sequencing accuracy (Eid et al., 2009).

2.3.3.1 GC bias

If steps with PCR are included in the sample preparation protocol for sequencing, as is usually the case, a notable bias originating from differences in GC content has been detected in the sequencing coverage (Hodges et al., 2007; Benjamini and Speed, 2012). GC coverage bias, the correlation between coverage and GC content, manifests as lower or less uniform coverage on both GC-rich and poor regions (below 10% and above 75% GC content) as compared to regions with more balanced base composition (Benjamini and Speed, 2012; Quail et al., 2012; Ross et al., 2013; Meynert et al., 2014). Therefore, regions with simple repeats, such as long ATAT motifs, low complexity repeats, CpG islands and satellites, which tend to have high GC content, are usually less well covered (Ross et al., 2013; Meynert et al., 2014; McCoy et al., 2014). The degree of this bias varies somewhat for different sequencing platforms (Rieber et al., 2013; Quail et al., 2012).

Because the problem is so widespread, it has been closely studied. Regions with higher GC content appear to remain tightly annealed as double-stranded DNA in various settings. This reduces the amplification efficiency of these regions by affecting the probability of fragmentation and by preventing access of primers and probes (Veal et al., 2012; Benjamini and Speed, 2012). Partly for this reason, DNA fragmentation is employed to shorten and separate stretches of tightly annealing GC-rich elements before amplification and target capture (Veal et al., 2012). Denaturing can also be stimulated with enhancers, such as dimethyl sulphoxide (DMSO) or single-strand DNA binding proteins (Veal et al., 2012). Additionally, PCR protocol optimization with adjustments to the polymerase, temperatures or step times enables more complete denaturation of some of the tightly annealed fractions, but usually at the expense of the opposite extreme in GC content (Aird et al., 2011; Veal et al., 2012). GC bias could also be partly alleviated by embedding the DNA fragments into larger constructs with adapters and such to decrease the amount of genomic DNA and GC bias in each sequenced stretch (Rieber et al., 2013; Benjamini and Speed, 2012).

Avoiding PCR in the library preparation could eliminate some of the GC coverage bias but this is challenging with current sequencing methodology (Quail et al., 2012). Depending on the protocol, PCR is usually involved in amplicon production, capture steps and pre-capture amplification (Ma et al., 2019). Especially amplicon sequencing requires always multiple cycles of PCR as a part of the library preparation process (Schirmer et al., 2016). However, GC coverage bias is detectable also in WGS data, albeit less than from methods which require more amplification steps (Mallawaarachchi et al., 2016). PCR is usually required in WGS to enrich fragments carrying adapters on both ends and to avoid needing a large amount of input DNA (Aird et al., 2011; Mallawaarachchi et al., 2016). Single-molecule sequencing technologies need to improve considerably as well before sequencing without template amplification is possible (Salk et al., 2018).

Additional GC bias may be introduced in downstream steps with cluster amplification and sequencing by synthesis involving also primer extension by DNA polymerase (Aird et al., 2011; Mallawaarachchi et al., 2016). Even a PacBio trial without template amplification in library preparation displayed coverage reduction for extremely high GC coverage regions (> 75%) and for lowest GC content regions, probably attributable to dissociation of fragments in adapter ligation, which concerns also other technologies (Ross et al., 2013). Additionally, AT-rich sequences are under-represented in coverage since they disturb sequencing by ligation (Valouev et al., 2008). Therefore, GC coverage bias, which is especially detrimental for variant analysis approaches relying on read depth information, such as detection of structural variants discussed further, seems to be unavoidable in sequencing data for now. This needs to be taken into account by normalization in computational analysis to reveal the original signal (Benjamini and Speed, 2012).

2.3.3.2 Sequencing data pre-analysis

The popularly utilized computational pipeline for MPS short-read sequencing data, Genome Analysis Toolkit (GATK) Best Practices workflow, is based on variant detection by comparison to reference genome with local assembly (DePristo et al., 2011; McKenna et al., 2010). The pipeline is adaptable for multiple sequencing platforms and experimental designs.

Burrows-Wheeler aligner (BWA) and its application BWA-MEM are utilized in most modern MPS pipelines to align reads to the human reference genome as the first step of the pipeline (Lappalainen et al., 2019; Regier et al., 2018; Trost et al., 2018; Li, Heng, 2013; Li, H. and Durbin, 2009). BWA-MEM is based on a similar seed-and-extend approach as the previous alignment tool, where the longest exact matches found are extended (Li, Heng, 2013). BWA-MEM is more advanced than BWA, which was developed for read lengths of less than 50 bp. With longer reads (100 bp or more), allowing longer gaps is important to reveal potential structural variants and to speed up alignment (Li, Heng, 2013). Many of the sequencing platforms support paired-end sequencing, which can be ultimately used to resolve orientation and distance between sequences (Bentley et al., 2008; McCoy et al., 2014). This enables higher coverage and more accurate read alignment and consensus building, which are especially

beneficial in resolving repeats (Bentley et al., 2008; McCoy et al., 2014).

In the data cleanup, molecular PCR duplicates are eliminated, which can comprise as much as 5–20% of the original reads. This minimizes issues from the PCR amplification bias and improves variant calling (Sulonen et al., 2011; Meynert et al., 2014). Sequencing artefacts, such as adapter and primer sequences and low-confidence sequences of insufficient quality can be computationally removed to reduce background noise (Salk et al., 2018). Filtering may be used to remove also some off-target or unmappable reads attributable to sequencing errors, too many non-reference bases, or multiple mapping positions in the genome (Meynert et al., 2013). Finally, local realignment on indels is done to improve indel calling, and base qualities are recalibrated to eliminate sequencer-specific bias. After these steps, the data is ready for further applications in a technology-independent reference file format, Sequence alignment/Map (SAM) (DePristo et al., 2011).

SAM is a format for sequence data compression and storing read alignments against reference sequences (Li, H. et al., 2009). The developers of the data format provide also SAMtools for data analysis, with various tools for post-processing alignments in the SAM format, such as converting between alignment formats, indexing, sorting, merging, variant calling and alignment (Li, H. et al., 2009). Binary Alignment/Map format (BAM) is a binary representation of SAM with the same information but compressed into a BGZF library. An indexed position-sorted BAM file allows applications to process specific genomic regions without need to load the entire file into memory (Li, H. et al., 2009). Sequencing data file size compromises mostly from base qualities (Li, H. et al., 2009). Per base quality score contains information on probability that the called base has been sequenced correctly (DePristo et al., 2011). Phred quality score, which measures probability of a base being identified erroneously on a logarithmic scale, was originally developed for Sanger sequencing and adapted for image-based MPS platform outputs (Salk et al., 2018). For example, Illumina sequencing utilizes quality value scaled by the phred algorithm, and > Q30 (> 99.99%) base calling accuracy represents high quality (Bentley et al., 2008).

2.3.4 Whole genome, whole exome, targeted gene panels - current views

Reasons for preferring targeted sequencing with WES or targeted gene panels over WGS include (originally notable) difference in costs, and coding sequences being the most informative and also most well annotated (Sulonen et al., 2011). Limitation of targets requires smaller amount of input DNA and enables more samples to be sequenced with one run (Sulonen et al., 2011; Goodwin et al., 2016). Consequently, the main source for the cost difference between targeted sequencing and WGS is raw sequencing, but includes also data storage requirements and computing time required for data analysis (Meynert et al., 2014). Size is also an obstacle for WGS data in transfer speeds for data access and sharing (Regier et al., 2018). Especially data from targeted gene panel sequencing require less storage space, are time- and cost-effective in analysis and offer higher accuracy for variant detection with deeper coverage (Povysil et al., 2017). On the other hand, the cost of sequencing has been reportedly brought

down to 1000$US for the whole human genome, increasing its appeal in research and clinics (Goodwin et al., 2016; Volk and Kubisch, 2017). Additionally, the informatics challenge for data storage and computation requirements for WES and WGS data are being solved with new platforms, such as cloud computing resources (Abel et al., 2018; Hehir-Kwa et al., 2018).

With WES, 3' and 5' untranslated regions or intronic regions are usually not covered. Targeting them might be especially important in studying a complex disease with variation expected in other than protein coding regions and when expecting variants affecting splice sites (Sulonen et al., 2011; Gulilat et al., 2019; Mallawaarachchi et al., 2016). Additionally, variants missed by WES or targeted gene panel sequencing may include commonly unexpected variants, such as novel disease genes, non-coding RNA genes, or regions with poor coverage in WES data due to technical issues (Lionel et al., 2018). For genes with closely similar pseudogenes or homologous regions, designing short probes in WES to capture only the intended targets and not their counterparts is challenging, but WGS requires no separate target capture step (Mallawaarachchi et al., 2016). WGS prevails also in resolving repetitive regions and revealing genomic rearrangements by enabling more accurate read alignment (Mallawaarachchi et al., 2016; Meynert et al., 2014).

Although WGS has been predicted to replace targeted approaches in MPS studies for years (Meynert et al., 2014; Lappalainen et al., 2019; Harel and Lupski, 2018), targeted MPS approaches are still popular with their time- and cost-effectiveness (Gulilat et al., 2019). The approach of semi-comprehensive WES sometimes called "Mendeliome" covers some thousands of genes (4000–6000, for example Illumina TruSight One and the Agilent SureSelect Focused Exome) with clinical significance (Pengelly et al., 2020). This approach has been applied for diagnosis of some genetically heterogeneous disease groups (Marelli et al., 2016). WES has also been implemented with a flexible and popular approach of *in silico* gene panels, where a region or regions of interest are separated from the WES data to be analyzed in a more time-effective manner while also avoiding incidental findings (discussed further) (Hehir-Kwa et al., 2018; Pfundt et al., 2017). WES is currently more expensive than most targeted gene panels, but the virtual panel approach brings down the costs to similar lever to panels and allows analyzing unsolved cases further from the same data (Hartley et al., 2018).

WGS is expected to be the end goal for genetic testing in diagnostics. It has been continuously predicted that higher costs and requirements for data analysis and storage with WGS will be outweighed by diagnostic gain compared to WES (Lionel et al., 2018). Evidently, WGS seems to provide higher diagnostic yield than conventional targeted genetic testing especially in clinically heterogeneous disorders (Lionel et al., 2018). On the other hand, it has also been reported that WGS provides limited advancement over WES in clinical diagnostics (Alfares et al., 2018). Nevertheless, WGS is still not the standard first approach test in congenital nor oncogenetic testing, and WGS is estimated to remain substantially more expensive for a while still than WES or array approaches (Hehir-Kwa et al., 2018; Marchuk et al., 2018).

2.4 Variant detection from MPS data and its applications

Before large-scale methods were available for genomic analysis, a hypothesis of the genetic cause for a certain disease or phenotype had to be formulated in order to target the genetic inspection to a certain region (Yang et al., 2013). This often led to multiple unsuccessful diagnostic tests: chromosomal microarray analysis, DNA methylation studies, single-gene sequencing tests, mitochondrial genome sequencing, enzyme analysis and biochemical analytic studies, with corresponding costs (Yang et al., 2013). With current MPS, the hypothesis of a phenotype can, if needed, be formulated after analyzing the sequence data, which then directs the investigation into a certain direction, rather than the other way around (Alkuraya, 2015). MPS methods have been used to study genetic variation in humans and animal models in health and disease, and increasingly to understand the organization, regulation and function of the genome (Goodwin et al., 2016). Since short-read sequencing platforms still dominate in these applications, other sequencing technologies with their current and emerging usages will be discussed later.

MPS can be used in the analysis of modified bases, such as methyl-seq to inspect methylation of bases. Different approaches are based either on capture of the methylated regions, their selective digestion, or on modification of the methylated bases to introduce a SNP into the sequence (Goodwin et al., 2016). In ChIP-seq, immunoprecipitation is followed by sequencing to capture regions of genome covered by inspected entities, such as proteins (Goodwin et al., 2016). Chromosome conformation capture followed by sequencing enables studying of genomic interactions (Dixon et al., 2018). For RNA-seq, mRNA capture is followed by complementary DNA (cDNA) synthesis (Ambardar et al., 2016). Different non-coding RNA classes can be studied as well, enabling the recognition of new miRNAs and miRNA targets (Fehlmann et al., 2016).

Most commonly, MPS methods are utilized for calling variants from the genome. They have enabled resolving spatial and temporal genetic heterogeneity in tumors (Salk et al., 2018; Turajlic et al., 2019; Garg et al., 2020). CNVs and SNVs in genes encoding drug metabolizing enzymes, membrane transporters and drug targets (together called pharmacogenes) necessitating altered pharmacotherapy have also been recognized (Gulilat et al., 2019). Most of all, MPS methods have been utilized to detect pathogenic variants in Mendelian disorders, and also for the study of common diseases based on enrichment of more common variants (Salk et al., 2018; Goodwin et al., 2016). Generally, diagnostic yield for rare Mendelian diseases with WES and WGS varies between 25 and 50% (Cummings et al., 2017; Ellingford et al., 2016; Srivastava et al., 2019).

2.4.1 Variant-calling from MPS data: SNVs and indels

SNVs and indels are the easiest type of variants to detect from short-read MPS data. Algorithms for detecting SNVs and indels compare the sequencing data to the reference genome and detect discrepancies (Lappalainen et al., 2019; DePristo et al., 2011). Indels can be detected within a

single MPS sequencing read as one defining factor for their size of 50–100 bp (Carvalho, C. M. and Lupski, 2016). The sensitivity and specificity of the variant detection can be adjusted: for discovering rare variants in Mendelian diseases, sensitivity can be emphasized with a higher false positive error rate, while in population study settings specificity can be given more weight (DePristo et al., 2011). With more aggressive filtering the risk of excluding true variants increases (Salk et al., 2018).

Commonly utilized evaluation metrics for sequencing data quality include mean average coverage and the percentage of the targeted sequence covered by at least 20X coverage, with approximately 95% threshold regarded sufficient (Charng et al., 2016; Boone et al., 2013; Gulilat et al., 2019; Yang et al., 2013). The required mean average coverage has been estimated with studies on required read depth with WES and WGS to detect variants. The estimations vary and sometimes overlap depending on variant type and scope (Meynert et al., 2014). WGS covers usually a higher proportion of targeted bases sufficiently to enable detection of variants with lower read depth than WES. This is mostly due to avoiding capture bias and more uniform read coverage (Mallawaarachchi et al., 2016; Meynert et al., 2014). One estimation for detecting heterozygous SNPs on coding regions with WES is 39–41X average read depth to reach the same high 95% detection sensitivity as achieved with 14–18X read depth with WGS (Meynert et al., 2014). Typical coverages for standard WES vary between 40 to 100X both in research and clinical diagnostic settings (Yang et al., 2013; Gambin, Akdemir et al., 2017; Gulilat et al., 2019). Current WGS studies usually aim to a read depth of minimum of 20X (Lappalainen et al., 2019; Regier et al., 2018). Recently, high sensitivity and specificity of > 99.7% for SNVs and > 95% for indels have been reached from MPS data (Regier et al., 2018; Goodwin et al., 2016; Lappalainen et al., 2019). The detection of low allelic frequency variants requires deeper sequencing and specialized data analysis algorithms (Chen et al., 2017). Even low false-assignment rates can cause problems in studying rare somatic mutations in cancer or for the study of mitochondrial heteroplasmy (Kircher et al., 2012).

2.4.2 Variant calling from MPS data: structural variation

Soon after the first comprehensive CNV studies with array CGH, the utility of MPS in analyzing structural variation was hypothesized. MPS provided advantages in detecting balanced variants and novel sequence insertions, estimation of absolute copy number, and higher resolution, even without *a priori* knowledge of the variants (Feuk et al., 2006; Mills et al., 2011; Alkan et al., 2011). Additionally, MPS enabled detection of structural variation together with other variant types, thus avoiding the need for other analysis methods and increase in costs (Mallawaarachchi et al., 2016). Paired-end sequencing generally enables more accurate detection of genomic rearrangements, repetitive sequence elements, gene fusions and novel transcripts (Ambardar et al., 2016).

The first part of this review focuses on structural variation detection tools designed for short-read sequencing data. A plethora of detection programs are currently available but with just a few shared algorithmic approaches (Figure 13).

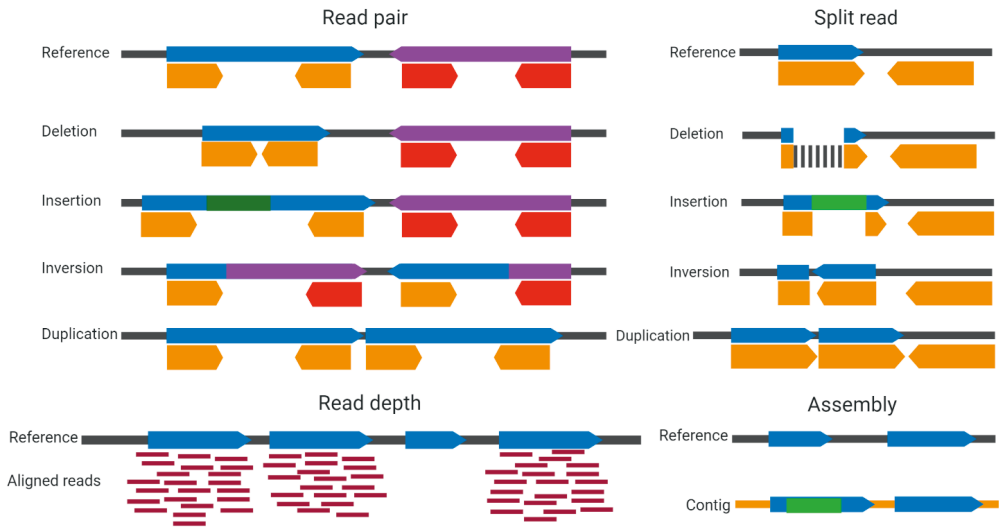

Figure 13. Approaches for structural variant detection from MPS data. Read pair method investigates the relative location and orientation of read pairs (orange and red arrows). Split read method investigates reads, which fail to map or map only partially, split and realign them to reveal variation. Read depth method investigates and compares the coverage of genomic regions, and assembly method builds contigs of sequenced data and compares them to the reference genome.

Read pair location and orientation were employed as the first signals representing structural variants in MPS data (Korbel et al., 2007; Tuzun et al., 2005). Read pairs located too far indicate deletions and too close insertions, tandem duplications and dispersed duplications map read pairs to more distant loci, differences in orientation indicate inversions, and successful mapping of only one read could indicate a novel insertion (Korbel et al., 2007; Tuzun et al., 2005; Alkan et al., 2011). Read pair method cannot provide accurate estimation of copy number, is inapplicable to insertions larger than the insert size and cannot resolve low-complexity regions (Tuzun et al., 2005; Zhao et al., 2013). Longer insert size enables higher coverage and higher overall structural variation detection power, but shorter insert size enables more accurate calling of smaller variants (Zhou et al., 2018). Read pair method performance is affected by whether reads with multiple mapping positions are discarded or retained – most of all, this defines how well the structural variants located in repetitive regions can be analyzed (Zhao et al., 2013). Read pair method is thus suitable only for paired-end sequencing data. At the time of developing the approach, this was not a limitation with the prevalence of paired-end short-read sequencing data, but, as discussed before, some of the emerging sequencing approaches so far produce only single-end sequencing data.

Split read method was first applied to Sanger sequencing to map indels in the human genome (Mills et al., 2006). The method is based on read pairs, where one aligns to the reference genome and the other fails to map or only partially maps during alignment. The unmapped reads are split, and each fragment is realigned independently (Mills et al., 2006; Zhao et al., 2013). Stretch of gaps in the unmapped reads indicates deletion, and in the reference genome an

insertion (Alkan et al., 2011). Split read method enables detection of small deletions and insertions with unprecedented breakpoint resolution and anchors insertions (Alkan et al., 2011). The method is only suitable for unique genomic regions and limited in detectable variant size by library size more than the read pair method (Zhao et al., 2013).

Assembly method allows potentially fine-scale discovery of all structural variant types, including novel sequences evasive for the previously presented methods, and enables resolving of complex genomic regions (Simpson et al., 2009; Li, Y. et al., 2011). For computationally feasible true *de novo* sequence assembly, long and accurate reads surpassing current technologies would be required. Therefore, *de novo* local assembly of short reads to generate contigs and their extension by comparison to reference genome is a common approach to improve computational efficiency and contig accuracy (Simpson et al., 2009; Alkan et al., 2011; Zhao et al., 2013). In its current state, this approach has a possible supportive role in discovering novel sequence insertions and improving breakpoint estimation (Alkan et al., 2011; Li, Y. et al., 2011). The method is insensitive for structural variants in highly repeated regions, such as tandem repeats, because of their tendency to collapse in assembly into one (Alkan et al., 2011; Li, Y. et al., 2011).

Read depth method investigates differences in read depth distribution to discover duplications and deletions, and works thus both for single-end and paired-end data, but only for copy number variants (CNV) (Yoon et al., 2009). The workflow involves calculation of read depth in regions divided into windows, normalization of the reads locally in each window to correct biases (mainly GC bias and batch effect), detection of regions with divergent copy number, and segmentation to merge adjacent regions with the same copy number into one detection (Zhao et al., 2013; Yoon et al., 2009; Teo et al., 2012). The baseline assumption with this method is that the sequencing data is uniform in coverage. Reads mapped to a region are assumed to follow Poisson distribution with a suitable window size and a high enough read depth (originally suggested 100 bp for 30X coverage) (Yoon et al., 2009). Larger window size limits precision of breakpoint-calling and detection of smaller CNVs, since the signal originates from fewer windows, but the assumption of normal distribution wavers with smaller window sizes (Yoon et al., 2009; Zhou et al., 2018). Either predefined or dynamic window sizes are used with an attempt to keep the number of reads mapped to each window within a certain threshold (Zhao et al., 2013). For segmentation, statistical models such as circular binary segmentation (CBS) or Hidden Markov Model (HMM) are commonly utilized (Zhao et al., 2013). CBS was originally developed for array CGH to discover segments with the same copy number among noisy intensities (Zhao et al., 2013).

Tools based on read depth evaluation tend to have different study designs enabling analysis of either single samples, paired case/control samples or a large set of pooled samples (Kadalayil et al., 2015). With comparison to only the reference genome, areas may appear normal on regions which are incorrectly mapped in the reference genome or have the same allele. Therefore, comparison of read depth between multiple samples was originally recommended to

clearly recognize polymorphic events (Yoon et al., 2009). A matched control or analysis within a sample set lowers false positive count and allows discovery of smaller events and higher statistical power (Zhao et al., 2013). The case-control setting alleviates also the effects of GC bias, since it is expected to match between the two samples (Zhao et al., 2013). In the pooled data of sample sets, CNVs are usually detected as deviations from the average read count of the batch (Kadalayil et al., 2015). With this approach, the sample set should be processed with the same sequencing pipeline to minimize technical bias (Jiang et al., 2015). This is a popular approach, since normal samples can be often unavailable, incomplete or unmatched.

Compared to other methods, only the read depth method enables the detection of some segmental duplications (SDs) and inferring of absolute copy number (Alkan et al., 2009). Generally, ambiguously mapping reads in repetitive regions cannot be accurately resolved. On the contrary, estimation of coverage for a region with unequivocally mapped reads is possible with read depth method, unlike the read pair and split read methods (Alkan et al., 2011; Yoon et al., 2009). Discarding ambiguously mapped reads completely may limit analysis to only unique regions and increase the amount of false positive deletion detections (Teo et al., 2012). Some tools assign these reads randomly to one of the equally possible positions, which enables the analysis of repetitive regions. However, this leads to the increase of false positive detections and a diluted signal. Soft clustering approaches allow multiple good mappings, and it may be the only approach for retaining signals from SDs, which are often enriched in CNVs (Zhao et al., 2013; Teo et al., 2012).

The read depth method provides a poor breakpoint resolution (Alkan et al., 2011; Yoon et al., 2009). Coverage loss from GC bias affects somewhat negatively the other three CNV detection methods, since these regions may have insufficient information to discern variants or to enable contig assembly. Read depth analysis is inherently the method most affected by GC coverage bias (Teo et al., 2012). Balanced rearrangements or novel insertions missing from the reference genome are not detectable by the read depth method, and a precise location cannot be provided for insertions (Yoon et al., 2009; Zhao et al., 2013). Detection of duplications is less sensitive because of smaller relative difference in read depth and thus weaker signal (Teo et al., 2012; Marchuk et al., 2018). The read depth method has a persisting size limitation and detects accurately only CNVs bigger than 1000 bp in size (Yoon et al., 2009; Trost et al., 2018).

Combining read depth method and paired-end read information in a program owing to their complementarity in performance is an old idea (Yoon et al., 2009). The read pair method excels at detecting small deletions (< 1 kb) with a high breakpoint resolution, while the read depth method detects larger variants better, but is based on a less powerful signal and lacks potential to call accurate breakpoints (Tan et al., 2014; Mills et al., 2011). Since both read pair method and split read method utilize only position information, exact copy numbers cannot be inferred with them (Zhao et al., 2013). Read depth method and read pair method have remained the most popular combination in combinatory algorithms (Zhao et al., 2013) and in algorithms overall (Zhang, L. et al., 2019).

In practice, CNV detection algorithms differ on multiple aspects in addition to their detection method, such as programming language and operating system. Popularly used languages include R and C/C++, which are convenient for incorporating statistical models and distribution across multiple operating systems (Zhao et al., 2013). The programs also have differing memory requirements and run times, sometimes with a span from minutes to hours for the same sample set (affected mostly by support for parallelized analysis) (Zhang, L. et al., 2019; Tan et al., 2014; Povysil et al., 2017). For example, the read depth method tends to consume more memory in structural variant analysis than the other approaches (Kosugi et al., 2019). The tools based on read depth method usually accept BAM and SAM files, and tools inspecting read pairs and split reads settle for FASTQ-files, since the latter do not utilize read depth information (Zhao et al., 2013). Some programs can also utilize both read depth and SNP zygosity information (both BAM and VCF files required): information for B-allele frequency reveals possible regions of LOH (Gambin, Akdemir et al., 2017). Genotyping information generally increases detection sensitivity and specificity (Kosugi et al., 2019; Teo et al., 2012).

The tools are usually developed for a specific setting, such as for detecting rare or common variation (Alkan et al., 2011), detecting CNVs especially on duplicated and complex genomic regions (Alkan et al., 2009) or detecting small homo- or hemizygous deletions (Gambin, Akdemir et al., 2017). The programs can also be specified for the detection of germline or somatic structural variation, for specific sequencing data type, or for sample batch or control-sample comparison settings. The programs can also have different modes for different study settings available, such as an option for a virtual gene panel included in the CNV calling program (Zhang, L. et al., 2019; Povysil et al., 2017). Most of the tools are designed for hybridization sequencing data with smoother coverage, but some have been designed for amplicon enriched sequencing data as well (Talevich, E. et al., 2016).

All CNV detection tools are mostly standalone distributions rather than available to be run online (Zhao et al., 2013). Only a few tools have both code and graphical user interface versions available for users without programming experience (Povysil et al., 2017). Therefore, most of the tools require at least moderate programming skills to install the required packages and execute commands (Roca et al., 2019). The existing algorithms have also been modified or combined, which has been enabled by the policy of freely available source codes (Samarakoon et al., 2014).

2.4.2.1 Structural variants and the different short-read sequencing data sources

Discontinuous sequence data in WES and targeted gene panel sequencing approaches prevents the capture of most CNV breakpoints as they land mostly in introns, which limits the resolution of these CNVs (Kadalayil et al., 2015). In practice, this leaves read depth method as the only possible approach and limits the detections to the CNV class of structural variation. Therefore, even if breakpoints are located in exons, their accurate detection is challenging because of the inherent limitations of the read depth method (Truty et al., 2019). Overall, discovering and genotyping CNVs from WES data with read depth method is more challenging than from WGS

data, with more uniform coverage as the key difference (Tan et al., 2014). GC content affects exome capture, amplification by PCR and efficiency of sequencing and skews read depth distribution more in WES than WGS data (Kadalayil et al., 2015). Therefore, deep and uniform coverage is generally required for accurate CNV detection from WES data, with average coverages in CNV detection studies varying between 60X and 90X (Kerkhof et al., 2017; Yao et al., 2017; Sadedin et al., 2018; Marchuk et al., 2018), but also 350X has been used with targeted gene panels (Truty et al., 2019).

Structural variants are common in cancer genomes (Turajlic et al., 2019). Challenges for identifying somatic CNVs in cancer include tumor heterogeneity, complexity and contamination from normal tissue. Cancer genome may have changes in ploidy, which compromises comparison to reference and estimation of absolute copy number (Hehir-Kwa et al., 2018; Whitford et al., 2019; Eijkelenboom et al., 2019). Inferring the correct copy number is more difficult for increased multiplications, which are also more common in cancer (Turajlic et al., 2019). Resolving the breakpoints and structure of complex rearrangements, such as in chromothripsis, is generally very challenging with short-read sequencing technologies (Hehir-Kwa et al., 2018). The programs specified for somatic CNV detection utilize often both read depth and B allele frequency information to detect CNVs and regions of LOH more accurately, and take possible tumor impurity into account (Liu, X. et al., 2018). These tools require usually a tumor-control sample pair for analysis to reach better accuracy (Kim, H. Y. et al., 2017; Kadalayil et al., 2015), and higher average read depth is generally required for calling CNVs from tumor samples (100–300X) depending on normal tissue contamination (Liu, X. et al., 2018).

The sensitivity of CNV calling from WES data for clinically relevant germline CNVs with at least three exons has been estimated to be very high, 96% (Gambin, Akdemir et al., 2017), and even a detection sensitivity of 100% and specificity of 99.8% has been reached for gene panels (Kerkhof et al., 2017; Ellingford et al., 2017). On the other hand, less than ideal sensitivity of 86.5% and specificity of 78% has also been demonstrated (Gambin, Akdemir et al., 2017). Current consensus is that CNV analysis from WES data provides notably lower sensitivity compared to SNVs and indels (Hehir-Kwa et al., 2018).

Initially, algorithms based on analyzing read depth were more frequent due to WGS still being rare (Zhao et al., 2013). Currently, it is estimated that only 5–10% of structural variants can be detected solely by read depth analysis from short-read sequencing data, even with WGS data (Lappalainen et al., 2019). Nevertheless, read depth methods are still popular for WGS data by being conceptually the simplest to use and by superiority in detecting large CNVs (the minimum limit of > 1 kb persisting), which are more likely to have clinical significance (Trost et al., 2018).

The increase of WGS studies has led to the development of new tools for analyzing structural variation from WGS data with other methods than read depth (Zhang, L. et al., 2019). WGS

data provides multiple sources of evidence to detect structural variants: relative read depth, read pair location and split reads spanning the breakpoints (Mallawaarachchi et al., 2016). Breakpoint analysis enables detection of smaller events and copy number balanced variants compared to WES data (Neerman et al., 2019). Generally, read pair methods display the most balanced sensitivity and specificity for analyzing different structural variant classes from WGS data (Whitford et al., 2019). The combination of read depth and read pair analysis software for CNV analysis from WGS data is common (Zhou et al., 2018). WGS data provided 96% clinical sensitivity for detecting pathogenic structural variants in a recent study (Neerman et al., 2019).

Only structural variants with distinct breakpoints (approximately 75%) can be detected from short-read sequencing data, thus generally excluding mobile element insertions, short tandem repeats, SDs (such as MHC clusters), balanced translocations in non-uniquely mappable areas, and multi-allelic CNVs (Lappalainen et al., 2019; Abel et al., 2018; Zhou et al., 2018; Lionel et al., 2018; Neerman et al., 2019). Also inversions tend to be flanked by highly identical sequence stretches, making their alignment and identification challenging from short-read sequencing data (Chaisson, M. J. P. et al., 2019). Approximately 7–8% of the balanced chromosomal rearrangement breakpoints undetectable from short-read sequencing data are located near centromere heterochromatic regions or within SDs (Redin et al., 2017). Moreover, the different algorithms have less concordance in detecting duplications from WGS data (which will be discussed more further), so these remain more difficult to detect (Trost et al., 2018).

Transposable elements can exhibit high sequence identity, high copy number or complex genomic arrangements, making their detection by short-read sequencing challenging (McCoy et al., 2014). Highest mapping scores for reads with primarily repeat content are usually obtained from the longest repeat locus with the same repeat unit in the reference genome, which may not be the original source of the expansion stretch (Dashnow et al., 2018). Some tools have been developed for the specific detection of STRs and VNTRs (Dashnow et al., 2018; Bakhtiari et al., 2018), mobile element and viral element insertions from short-read MPS data, but comprehensive studies for these variant types are limited (Lappalainen et al., 2019; Kosugi et al., 2019). Regardless of these advances, repeat expansions detectable in MPS data are still being examined manually to reach accurate detection results (Gulilat et al., 2019). Some of the diseases caused by repeat expansions are genetically heterogeneous and would thus benefit from a comprehensive genomic test (Dashnow et al., 2018).

2.4.2.2 Lack of call concordance, sensitivity and specificity, and solutions

Early on, it was recognized that some structural variant types are uniquely detected by different detection approaches (Alkan et al., 2011). For example, different programs and program combinations have been shown to be more accurate in calling variants of specific state (deletions or duplications) or size (Tan et al., 2014; Zhang, L. et al., 2019; Kosugi et al., 2019). Furthermore, poor concordance has been observed with structural variants detected by different tools in multiple studies. The programs suffer from low sensitivity and specificity, generally with a trade-off between the two (Yao et al., 2017; Samarakoon et al., 2014; Rajagopalan et al.,

2020; Povysil et al., 2017; Neerman et al., 2019; Kadalayil et al., 2015; Whitford et al., 2019; Tan et al., 2014; Chaisson, M. J. P. et al., 2019; Regier et al., 2018).

Combining results from algorithms with the same detection method increases specificity (Kosugi et al., 2019), whereas combining results from two similarly designed and well performing programs does not improve overall sensitivity (Whitford et al., 2019). By contrast, combining algorithms with different CNV calling methods has been demonstrated to increase both sensitivity and specificity (Kosugi et al., 2019). Program comparison studies are becoming more comprehensive both for WGS and WES settings and usually also include evaluations of different program combinations. For example, in some latest studies 10 to 69 programs have been evaluated in a single study (Zhang, L. et al., 2019; Kosugi et al., 2019; Roca et al., 2019; Trost et al., 2018). In most of them, the conclusive recommendation has been to use more than one tool regardless of the study setting (Sadedin et al., 2018). In one systematic study, the best detection accuracy was achieved with a combination of nine tools (Roca et al., 2019). This approach may be restricted in practicality by increasing computing time and costs, and by the requirement for additional complex steps of merging variant calls into a consensus set (Lappalainen et al., 2019; Trost et al., 2018). Tools for combining CNV calls are not commonly available and the process also lacks consensus: criteria of 50–80% overlap with and without requirement for reciprocality have been used (Hwang et al., 2015; Zarrei et al., 2015; Trost et al., 2018; Kosugi et al., 2019; Zhang, L. et al., 2019).

Tool performance estimations differ greatly between studies, which further complicates tool performance evaluation (Sadedin et al., 2018). One of the reasons may be in conducting the comparison studies in different settings without optimization of parameters (Sadedin et al., 2018). Variation in sequencing data metrics could also explain some of the discrepancies in these studies (Kadalayil et al., 2015). Evidently, in multiple evaluation studies only the default parameters of the programs have been utilized (Tan et al., 2014; Yao et al., 2017; Hwang et al., 2015; Zhang, L. et al., 2019; Trost et al., 2018; Roca et al., 2019). As an example, some algorithms are optimized for a certain read depth or read length as default, with parameter adjustments recommended to be done according to the data (Zhang, L. et al., 2019; Povysil et al., 2017).

As an additional explanation for the discrepancies, standard protocols and quality control measures for analyzing MPS data for structural variants have long been missing (Teo et al., 2012). The evaluation settings often involve varying sets of self-produced real and simulated sequencing samples and structural variants (Kosugi et al., 2019; Liu, X. et al., 2018; Roca et al., 2019; Sedlazeck et al., 2018; Neerman et al., 2019). Some studies have also used non-transferrable evaluation metrics, such as positive predictive value, which is dependent on CNV prevalence and cannot be transferred from one disease setting to another (Kadalayil et al., 2015). Especially WES and targeted gene panel sequencing data with inherently varied settings are still lacking high quality reference samples and a gold standard CNV set (Rajagopalan et al., 2020; Roca et al., 2019). An ideal reference sample or a sample set would contain numerous

known variants with a variety of types and sizes preferably processed with the same platform and protocols as test samples. The latter would be especially important for read depth methods with more significant sequencing batch effects (Rajagopalan et al., 2020). But because of challenges in meeting the presented requirements, CNV detections are often validated with complementary methods, such as arrays, which could distort the validation results due to their own platform specific biases (Yao et al., 2017; Kadalayil et al., 2015; de Ligt et al., 2013).

Nevertheless, some programs have been included in multiple validation studies and are more commonly used than others, such as ExomeDepth for WES data (Plagnol et al., 2012). It was developed almost a decade ago but is still considered one of the most sensitive CNV detection algorithms (Rajagopalan et al., 2020; Hehir-Kwa et al., 2018). Copy Number Inference From Exome Reads (CoNIFER) (Krumm et al., 2012) was the first program developed to detect rare CNVs. CoNIFER has still relatively high specificity and it has been used as the gold standard in some recent comparisons (Hehir-Kwa et al., 2018; Yao et al., 2017). EXome hidden Markov Model (XHMM) (Fromer et al., 2012) provides useful quality scores both for CNV detections and the detected breakpoints separately and was used to build the ExAC CNV database (Ruderfer et al., 2016). The program for comparing the performances of CNV detection tools, Ximmer, contains CoNIFER, XHMM, ExomeDepth, CODEX and cn.MOPS, which are thought to be commonly used and reliable tools (Sadedin et al., 2018).

For WGS settings, more reference samples are available from public collections or single studies (Neerman et al., 2019). The so-called Genome in a Bottle samples with structural variants called from WGS data have also become available (Neerman et al., 2019; Trost et al., 2018). One of the most widely used and information rich reference sample to test structural variant calling has lately been the NA12878. This sample has genome data provided from different platforms, such as Illumina HiSeq and PacBio RS, parental samples available, and DGV, NCBI dbVar and GIAB provided structural variant sets (databases discussed further) (Neerman et al., 2019; Trost et al., 2018; Kosugi et al., 2019; Sedlazeck et al., 2018; Zhang, L. et al., 2019). However, some duplications and inversions have been found to be missing from the original call sets (Kosugi et al., 2019; Whitford et al., 2019), and they also include some apparent false positive detections (Zhang, L. et al., 2019). Therefore, even this sample and its variant call sets have been considered to lack in quality (Neerman et al., 2019).

In addition to differences in tool validation protocols, the workflows for detecting the actual structural variation vary with no common consensus or best practices pipeline available (Trost et al., 2018). For example, the most cited algorithm for structural variation calling from WGS data has only < 12% of total citations (Trost et al., 2018). Furthermore, different MPS pipelines seem to lead to differences especially in structural variant call sets (Regier et al., 2018).

2.4.2.3 Emerging MPS approaches and detection of structural variants

Longer read length allows sequencing of complex repetitive regions or structural rearrangements since they can be covered with a single continuous read and assembled locally

(Salk et al., 2018; Goodwin et al., 2016). Long reads from both emerging long-read sequencing technologies and linked-read approaches can correct for amplification biases and mapping errors. This strengthens statistical power and sensitivity of variant detection with read depth method, increases detectable size of CNVs for split read method and improves assembly quality (Zhao et al., 2013; Hehir-Kwa et al., 2018). Evidently, compared to a PCR-free setting, WGS data from PCR-based libraries had less uniform read depth, which increased false positive structural variant detections (Trost et al., 2018). Long-read sequencing provides higher breakpoint resolution even for inversions, complex insertions and long tandem repeats, which are the most difficult variant types to detect with short reads (Hehir-Kwa et al., 2018; Chaisson, M. J. et al., 2015). Genotyping with long-read information for structural variants detected from short-read sequencing data also improves call accuracy (Audano et al., 2019). RNA sequencing (RNA-seq) provides further insight on some genomic regions or variants elusive for DNA sequencing, as will be discussed further. Although separate tools or existing tools with updates are being developed for structural variant detection from RNA-seq data, the tools have not been widely validated yet and are not in routine use (Talevich, Eric and Shain, 2018; Serin Harmanci et al., 2020).

High costs owing to high read depth required to cover for high error rate and low throughput limit the general use of long-read sequencing methods (Kosugi et al., 2019; Hehir-Kwa et al., 2018; Sedlazeck et al., 2018). Additionally, meeting the strict sample requirements for high molecular weight DNA for long-read sequencing technologies is challenging in many settings (Hehir-Kwa et al., 2018). Nevertheless, some algorithms have already been developed for structural variant calling from long-read and linked-read sequencing data (Sedlazeck et al., 2018; Kosugi et al., 2019; Hehir-Kwa et al., 2018). Recent novel structural variant discoveries from long-read data compared to short-read data approaches have mostly been small in size (500 bp on average) and more often insertions (Audano et al., 2019; Chaisson, M. J. et al., 2015). The programs have especially resolved complex CNVs with nested structures (Sedlazeck et al., 2018). Novel structural variants have also been found on regions with extreme GC content (Audano et al., 2019), and the increase in read length improves inferring bi- and multiallelic variants and multi-site variants (Campbell et al., 2016).

Long reads have enabled more accurate discrimination of pseudogenes and detection of repeat expansions, repeat associated structural variants and CNVs in polymorphic regions, such as the HLA locus (Hehir-Kwa et al., 2018; Chaisson, M. J. P. et al., 2019). Successful analysis of repetitive genomic regions for structural variants has increased the estimation of structural variant prevalence per genome over six-fold, from 4442 to 27,662 (Abel et al., 2018; Chaisson, M. J. P. et al., 2019). However, long-read sequencing technologies provide an excess of false positive small deletions (Nanopore), insertions, and duplications (PacBio), the majority of which are located in homopolymers or other simple repeats (Chaisson, M. J. et al., 2015; Sedlazeck et al., 2018; Dashnow et al., 2018). Although long reads can span whole repeat loci with higher probability, inferring the repeat unit count accurately is challenging because of an excess of these indel errors. This is somewhat alleviated in cases of notable differences in the

lengths of normal and pathogenic repeat unit counts (Dashnow et al., 2018; Bakhtiari et al., 2018).

Structural variant detection from short-read MPS data is today common in clinical research settings. Analysis of structural variants tends to add several hours to library preparation and software steps (Clark et al., 2019), but increases the diagnostic yield even in patient cohorts with multiple preceding non-conclusive diagnostic tests. Nevertheless, array approaches (SNP array and array CGH) have persistently been suggested to remain as the first-choice gold standard method for CNV detection in clinical diagnostic settings, with CNV detection from MPS data suggested to be used in initial screening (Yao et al., 2017; Marchuk et al., 2018). CNVs detected from MPS data are recommended to be verified before reporting (Eijkelenboom et al., 2019; Rajagopalan et al., 2020). Evidently, CNV detections from MPS data are still being validated with varying complementary methods, such as MLPA (Austin-Tse et al., 2018; Marchuk et al., 2018; Pfundt et al., 2017; Truty et al., 2019), PCR (Truty et al., 2019; Gambin, Akdemir et al., 2017), qPCR (Marchuk et al., 2018), ddPCR (Austin-Tse et al., 2018), Sanger sequencing (Charng et al., 2016; Truty et al., 2019), array CGH (Charng et al., 2016; Gambin, Akdemir et al., 2017; Austin-Tse et al., 2018; Samarakoon et al., 2014; Pfundt et al., 2017; Truty et al., 2019) or even FISH (Marchuk et al., 2018).

In the latest studies, more clinically relevant CNVs have been discovered with WGS compared to arrays (Trost et al., 2018; Zhou et al., 2018). Generally, array approaches have lower resolution for small CNVs and for determining exact breakpoints compared to detection of structural variants from MPS data (Kosugi et al., 2019; Rajagopalan et al., 2020). CNV analysis from low-coverage WGS data has lower costs compared to the current higher density arrays (Zhou et al., 2018). Sequencing with a targeted gene panel provides approximately 70% cost reduction compared to Sanger sequencing and MLPA approaches and significantly shorter turn-around time with similar diagnostic yield (Kerkhof et al., 2017). However, evaluating WES and WGS approaches in performance over older methods is not straightforward, since they are often the last-straw test, not first-tier, with pathogenic variants depleted from the patient cohorts (Lionel et al., 2018). The aim of the clinical validation of CNV detection from MPS data is to eventually avoid a need for complementary techniques, which increases costs and work-load in the diagnostic workflow (Kerkhof et al., 2017).

2.4.3 Variant annotation

Apart from advancements in variant detection methods, the most prominent bottleneck in diagnostic MPS approaches is the inability to interpret much of the variation (Harel and Lupski, 2018). Without accurate diagnosis, potential treatments or the risk of recurrence cannot be identified, and prognosis cannot be provided (Yang et al., 2013). Therefore, accurate assessment of clinical significance of discovered genetic variants, annotation, is of utmost importance to the patients, their relatives and the healthcare system. The proposed annotation guidelines for diagnostic MPS differ for somatic variants and germline variants (Matthijs et al., 2016; Li, M. M. et al., 2017; Richards et al., 2015; Riggs et al., 2019). The American College

of Medical Genetics and Genomics (ACMG) has defined the now widely-used classification terms of pathogenic, likely pathogenic, variant of uncertain significance (VUS), likely benign and benign to describe variants in Mendelian disease genes (Richards et al., 2015). Numerous criteria need to be evaluated before annotating the variants with any of these statuses.

The common consensus is that the most informative resources for prioritizing variants are variant-disease and gene-phenotype associations, population frequency and information on segregation (Neerman et al., 2019; Richards et al., 2015). Variant population frequency databases include for example NHLBI Exome Sequencing Project (ESP), dbSNP, 1000 Genomes, ExAC and gnomAD. These databases are either disease specific, such as ESP for heart, lung and blood disorders, or contain variants from putatively healthy individuals (Li, M. M. et al., 2017). ExAC provides exonic variant information from 60,000 individuals (Lek et al., 2016). The most comprehensive database of these, gnomAD, contains variants from over 125,700 exomes and 15,700 genomes spanning multiple ancestries (Karczewski et al., 2020).

Deleterious variants are often rare due to purifying selection. Thus, allele frequency in the general population can be used to predict potential phenotypic effects and clinical significance (Lappalainen et al., 2019). However, no agreed-on cutoff exists for classification of variation as polymorphic and thus probably benign. Most commonly, rare variation is defined to have a minor-allele frequency (MAF) of < 1%, common variation has a MAF of > 5%, and low-frequency variants fall in between. Ultra-rare variants refer often to singleton variants identified in only one person in a large study, but they may be described also with an arbitrary frequency of MAF < 0.01% (Li, M. M. et al., 2017; Lappalainen et al., 2019; Abel et al., 2018). Most identified variants are ultra-rare, but most (> 95%) variation in an individual is common. These are ancient polymorphisms which have arisen early in human development and present now in all major ancestry groups (Lappalainen et al., 2019). However, the databases with supposedly only non-deleterious variants are known to contain some pathogenic variants with e.g. incomplete penetrance or late manifesting phenotypes (Richards et al., 2015; Neerman et al., 2019). For example, some variants in neuromuscular disorders cause a mild, possibly late-onset dominant disease, while carriers of variants for recessive disease are often normal phenotypically (Laing, 2012). Utilizing ethnically matched MAFs is recommended (Li, M. M. et al., 2017), but most current databases are European-centric and would need increased ancestral diversity (Karczewski et al., 2020; Sirugo et al., 2019). For example, one study revealed a rare pathogenic variant in combination with different functional common alleles (leading to a surpassed pathogenic threshold) depending on the patient ethnicity (Wu et al., 2015).

Germline variant databases include for example Human Gene Mutation Database (HGMD) and ClinVar. HGMD collects variant annotations in the published literature (Richards et al., 2015). DECIPHER is a molecular cytogenic database, which links genomic microarray data with phenotypes (Richards et al., 2015; Firth et al., 2009). ClinVar provides information on clinical and phenotypic significance of rare germline variants from pathogenic to benign and the

relevant clinical and experimental evidence (Landrum et al., 2014; Richards et al., 2015). The Online Mendelian Inheritance in Man (OMIM) database describes comprehensively Mendelian genes and genetic conditions with examples of disease-associated genetic variants (Richards et al., 2015; Amberger et al., 2015). Somatic databases, such as My Cancer Genome and COSMIC, list the prevalence of variants across different cancer types (Li, M. M. et al., 2017). Many of the presented databases contain also relevant literature for reference (Li, M. M. et al., 2017). Some of the databases are disease or gene-specific (Richards et al., 2015).

Some problems have been recognized with the disease variant databases: as many as 25% of variants listed as pathogenic may be either sequencing errors, common polymorphisms, or remain singular cases with no further evidence of pathogenicity (Volk and Kubisch, 2017). The disease and gene-specific databases may also contain incorrectly associated variants (Wenger et al., 2017; Richards et al., 2015). For example, the same affected individuals or families may be reported in multiple different studies leading to false increase in variant frequency (Richards et al., 2015). Most of all, the variant databases are in need of standardization to increase their clarity and usability (Richards et al., 2015). This will most likely involve fixed encoding for patient phenotypes and measuring the relevance of variants for a phenotype in a standardized manner (Wenger et al., 2017).

In cases of novel or rare variants, which have not been incorporated into databases, *in silico* prediction programs can illuminate variant effect on gene function. Commonly used predictive software include SIFT, PolyPhen2, CADD and MutationTaster (Richards et al., 2015). These tools are based either on predicting the effects of single variants or on the calculation of hypothetical scores for every base change. Especially the latter enables annotation of novel SNVs (Ganel et al., 2017). Variant effect on transcript and protein level is often determined in two categories: whether the variant is deleterious to protein function or structure, or whether it affects splicing (Richards et al., 2015). Since the tools generally demonstrate only moderate specificity and tend to overestimate the deleteriousness of variants, their predictions are not recommended to be used as the only source of evidence for clinical significance (Li, M. M. et al., 2017).

Additional tools include variant tolerance estimation scores for genes. ExAC genetic intolerance constraint is based on comparison of predicted rare variation per genes and the observed variation, with depletion interpreted as intolerance. The probability of being LOF intolerant (pLI) score is significant for all known haploinsufficient genes (Lek et al., 2016). More recently, a loss-of-function score, loss-of-function observed/expected upper bound fraction (LOEUF) with low LOEUF scores indicating genes depleted for pLOF variation, was calculated from the combination of individual datasets from both gnomAD and ExAC (Karczewski et al., 2020). These scores can help to evaluate effects of a knockout on a gene, since these events are too frequent in different populations to be unequivocally deleterious (Alkuraya, 2015). A second knockout event or a milder genetic modifier may be required for a deleterious effect to present, complicating the analysis of these variants (Wu et al., 2015).

Discovery and inspection of novel genetic variants just on genomic DNA level may not be enough to enable conclusive determination of their clinical significance. RNA sequencing (RNA-seq) provides functional information on transcriptional perturbations and helps the inspection of large genes, which statistically harbor multiple putatively pathogenic variants (Cummings et al., 2017; Richards et al., 2015). RNA-seq can both help to validate discovered variants for clinical significance and enable the discovery of perturbations unidentifiable from genomic DNA sequencing data alone (Cummings et al., 2017). Additionally, expression levels can be quantified, and alternatively spliced isoforms can be detected. These may reveal regulatory alterations by variants in the promoter or in deep intronic regions (Volk and Kubisch, 2017). Both exonic and intronic variants can lead to exon skipping or exon extension, and either generate new splice sites or activate cryptic weaker splice sites (Cummings et al., 2017). Source material from the diseased tissue is generally required for RNA-seq, but some tissues are difficult to access (Cummings et al., 2017). In these cases, functional studies could be conducted with protein studies, and by animal models (Laing, 2012).

The most relevant of the presented information sources have been gathered into a widely utilized tool, ANNOVAR. Annotate variation (ANNOVAR) assists in annotation and filtration of SNVs and indels by examining their functional effect according to gene-based and region-based datasets and by comparison to variant databases for frequency filtration (Wang, K. et al., 2010). For example, synonymous and non-frameshift variants are often first excluded to provide a list of more potentially causative variants (Wang, K. et al., 2010). A variant can be rated higher in significance evaluation if it occurs in a gene in which mutations have been previously reported to cause the same disease as in the examined patient (Yang et al., 2013). Rare and deep intronic variants recently detected from WGS data remain challenging to interpret since the first comprehensive WGS variant databases have been just released (Volk and Kubisch, 2017; Austin-Tse et al., 2018). Regulatory or compensatory non-coding variation may explain phenotypic variability associated with the same pathogenic variants in different individuals, and these variants are currently scarcely documented (Zhang, F. and Lupski, 2015). On the other hand, patient phenotype description may be incomplete with some aspects not recorded or emerged yet, which confounds the diagnostic efforts (Wenger et al., 2017). In these cases, genetic studies on the family members could be informative.

Trio analysis, typically with an affected child and healthy parents, can be used to identify *de novo* variants, phase and prioritize variants, and analyze segregation (Wenger et al., 2017). Phasing information helps deduce whether potentially pathogenic variants are in *cis* or in *trans* to interpret their allele-specific impact on gene expression (Marks et al., 2019; Richards et al., 2015). Closely related individuals, such as siblings, are expected to share many genetic variants. Therefore, in dominant disease cases examining distantly related affected family members may be more informative (Gorokhova et al., 2015). For recessive disorders, examining the parents is informative, since both are expected to carry one pathogenic mutation (Gorokhova et al., 2015). A theoretical pitfall in these studies is that the segregating variant thought to be pathogenic could be in linkage disequilibrium with the true pathogenic variant (Richards et al.,

2015). Pathogenic variants with incomplete penetrance or variable expressivity are problematic for reporting since predicting their phenotypic consequences in the patient or their family members is challenging (Marchuk et al., 2018).

Incidental findings can be medically actionable conditions unrelated to the indication for testing, which might necessitate treatment or surveillance for the patient and their relatives. They may also alternatively involve autosomal recessive carrier status genes or pharmacogenetic findings (Yang et al., 2013; Green et al., 2013). ACMG has published recommendations for reporting incidental findings and estimate that 1% of cases will have such a finding (Green et al., 2013). Currently, the ACMG list for reportable incidental findings includes 59 genes (Kalia et al., 2017).

Regardless of comprehensive genetic tests, some patients remain without pathogenic genetic findings. In these cases, the disease could have non-genetic etiology, or the causative variant could be present in mosaic quantities challenging to detect (Marchuk et al., 2018). Some diseases are multifactorial; for example, compound inheritance of minor effect variants may be required to decrease the gene dosage beyond a certain threshold for deleterious effect to manifest (Weischenfeldt et al., 2013; Wu et al., 2015). Environmental factors could also influence disease presentation (Weischenfeldt et al., 2013), and penetrance can vary between populations with alternative risk haplotypes depending on ancestry (Rosenfeld et al., 2013; Wu et al., 2015).

Increasing knowledge and evolving phenotype may alter or add to the diagnosis of some patients (Yang et al., 2013). Approximately 10–11% new diagnoses have been reached with reanalysis of either WGS or WES data (Costain et al., 2018; Wenger et al., 2017). No clear agreement or recommendation has been made on the timing for this (Richards et al., 2015). Some studies have suggested periodic reanalysis of genomic data every 1–3 years, or sooner if the phenotype evolves (Costain et al., 2018; Wenger et al., 2017). Approximately 250 new gene-disease associations are being reported annually in OMIM and 9200 new variant-disease associations in HGMD (Wenger et al., 2017). However, a gap of years may precede the update of relevant databases after publication of the primary literature (Wenger et al., 2017). This necessitates long-term data storage in diagnostic centers. Alternatively, decreasing costs of sequencing and advancements in sequencing technology may make re-sequencing of DNA the more attractive option in the future (Costain et al., 2018).

Regardless of recommended best practices, human genetics still lacks clearly defined instructions that would be utilized widely and unequivocally in comprehensive analysis of MPS data. This would allow for unequivocal and robust comparison of sequencing data and results from different diagnostic groups (Campbell et al., 2016). This shortcoming concerns especially the more recently recognized variant types for which both the initial analysis and annotation methods and databases are still under development, such as CNVs and large-scale genomic rearrangements (Campbell et al., 2016).

2.4.3.1 Special aspects with CNV annotation

Fewer guidelines have been proposed for detection and reporting of CNVs than for SNVs and indels (Eijkelenboom et al., 2019). Therefore, interpretation of CNVs has been more ambiguous. Although the genome contains far less CNVs than SNVs or indels, 70% of individuals have at least one rare CNV in a gene, and a mean of four to ten genes altered by structural variants (Abel et al., 2018; Ruderfer et al., 2016; Collins et al., 2020). This prevents using frequency as the only criterion for differentiating pathogenic CNVs. In one study, all deletion events were interpreted as disruptive (Pfundt et al., 2017), while homozygous deletions unassociated with disease have been detected in genes tolerant to LOF variants (Gambin, Akdemir et al., 2017). Just recently, ACMG has provided recommendations for CNV clinical significance interpretation. With this, CNV classification was upgraded to the same five-tier system as for other variants (Riggs et al., 2019; Abou Tayoun et al., 2018). This part of the literature review will walk through that workflow and the available resources for adhering to them.

Generally, the new recommendations seem to consider still a relatively narrow spectrum of CNVs in Mendelian dominant disease genes (Riggs et al., 2019). The first part of the workflow inspects gene content of the affected genomic region and certain annotations for those genes. CNVs with no gene content are estimated to be likely benign. However, CNVs on gene-deficient regions can be deleterious through effects on regulatory functions (Szafranski et al., 2013). The workflow involves then scoring of CNVs by amount of genes encompassed by the variant, with significance given for gene counts of over 25 for deletions and 35 for duplications (Riggs et al., 2019). In the next step, evaluation of CNV pathogenicity is strongly based on curated HI scores and triplosensitivity (TS) scores provided by The Clinical Genome Resource (ClinGen) Dosage Sensitivity Map catalog for each gene (Riggs et al., 2012). Whole gene deletions are considered most probably pathogenic if the gene has a LOF mechanism, while whole gene duplications are considered generally benign unless the gene has a validated TS score (Riggs et al., 2019). Evidently, duplications spanning a whole gene have been detected to be less likely pathogenic (Truty et al., 2019). However, neither of these scores are yet comprehensively available for all the genes in The ClinGen Dosage Sensitivity Map catalog, which is updated daily (Riggs et al., 2019).

Bioinformatically calculated constraint metric of genic intolerance score (pLI) has been calculated separately for CNVs in genes according to prevalence of rare CNVs in the ExAC CNV database (Ruderfer et al., 2016). Additionally, a similar new score has been calculated for the gnomAD CNV set (pLOF), but the workflow does not take this into account (Collins et al., 2020). Rather, in addition to pLI, the LOEUF scores based on gnomAD SNVs and indels (Karczewski et al., 2020) as well as the older HI index (Huang et al., 2010) are considered to have predictive value for genes. Estimation for pathogenicity is increased only for deletions encompassing the haploinsufficient genes marked by these scores, since the scores are not designed to describe the effect of increased dosage (Riggs et al., 2019). The theoretical oversight here is that only the effect of deletions was originally evaluated in calculating these

scores, whereas CNVs can have LOF effect also by other means, as described earlier (Huang et al., 2010). CNVs can also lead to a GOF disease causing mechanism, although apparently very rarely (Truty et al., 2019). These tolerance scores are not based on WGS data or variants in other than coding regions and thus provide no information for non-coding regions (Abel et al., 2018).

In the next step of the workflow, the probable effect of CNV on gene structure integrity is inspected. The pathogenicity scoring differs for CNVs encompassing the whole gene, partially the gene with either 3' or 5' end involved, and for completely intragenic CNVs. Intragenic CNVs and CNVs overlapping the 5' end get higher scores (Riggs et al., 2019). Partial gene duplications involving the terminal coding exon are often non-deleterious because a functional copy of the gene with intact structure is preserved (Riggs et al., 2019; Truty et al., 2019). On the other hand, duplications completely within a gene can be deleterious. Most intragenic CNVs seem to encompass only internal exons (Truty et al., 2019). Therefore, an additional part of the workflow for intragenic CNVs (Abou Tayoun et al., 2018) is likely needed in many of the cases. According to these guidelines, pathogenicity of intragenic CNVs is determined by whether the reading frame is disrupted (Abou Tayoun et al., 2018). Duplicated copies in tandem (as most are) are considered less likely pathogenic than those interspersed from the above presented reason (Abou Tayoun et al., 2018). CNV effects on the reading frame are difficult to infer without detailed information on breakpoints and gene structure, as is available for dystrophin (*DMD*) (Bladen et al., 2015). The alternative of functional or RNA sequencing studies may provide too much workload for routine diagnostic settings. This tends to leave variants with status of VUS if their effect on the reading frame is unclear (Truty et al., 2019).

Inferring breakpoints is more challenging for duplications than deletions; the additional copies could display different orientation or location compared to the original, and bioinformatics tools may even drop these unambiguous detections (Mills et al., 2011; Newman et al., 2015; Campbell et al., 2016). In a recent study more than half of the duplications were considered VUS (Truty et al., 2019). Duplications can also form fusion genes with various possible consequences, which is not considered in the workflow (Newman et al., 2015). The workflow concerns only CNVs with change in copy number, while especially WGS has enabled the detection of balanced chromosomal aberrations (Redin et al., 2017). Detection of the exact genes and regions their breakpoints disrupt would be important to deduce the functional effect for these variants (Redin et al., 2017). Complex structural variants (approximately 5% of cases) have multiple or intertwined breakpoints, complicating the interpretation of their structure and clinical significance (Lappalainen et al., 2019). Therefore, precise breakpoint level analysis and information on exact location and orientation would be needed to enable the interpretation of the genetic consequences of many types of CNVs, but difficult to achieve with current methods (Conrad et al., 2010; Newman et al., 2015). Even for structural variants detected from WGS data many breakpoints (28% in one dataset) were not successfully mapped to single-base resolution (Abel et al., 2018).

The workflow does not consider or comment on the possibility of higher amplification of CNVs. Triplications of the same disease region may cause more severe phenotype (Liu, P. et al., 2014). The exact degree of amplification may be of specific diagnostic importance also in cancer associated genes. For example, a threshold of 10 copies of *MET* gene has been used for tyrosine kinase inhibitor treatment to be initiated (Eijkelenboom et al., 2019). Compensatory effect is also possible, as discussed previously for an asymptomatic father of a DiGeorge syndrome patient (Carelle-Calmels et al., 2009). This may be one of the mechanisms explaining incomplete penetrance. Recurrent CNVs with identical breakpoints have been associated with incomplete penetrance and variable clinical expressivity even within families (Newman et al., 2015; Rosenfeld et al., 2013; Stefansson et al., 2008). However, in the DiGeorge case the offspring has 100% risk of unbalanced outcome. Therefore, this finding would have significance in genetic counseling.

In the next steps of the workflow, detailed information on a straightforward gene-phenotype relationship would provide stronger indication for pathogenicity. Therefore, the workflow appears less suitable for disorders with high genetic and phenotypic heterogeneity (Riggs et al., 2019). At least 3–4 verified segregations in an affected family are considered significant proof towards pathogenicity (Riggs et al., 2019). Thus, small studies involving commonly only trios would not provide enough segregations. *De novo* CNVs are considered to be more likely pathogenic, if the status is verified (Watson et al., 2014; Riggs et al., 2019).

In the last step, CNVs are inspected for overlap with established benign CNVs, which is a strong indicator against pathogenicity (Riggs et al., 2019). Supposedly benign CNVs in the human genome, mostly from studies in healthy individuals with microarray and MPS methods, have been provided by the 1000 Genomes Project, DECIPHER, ExAC, and Database of Genomic Variants (DGV) (Ruderfer et al., 2016; 1000 Genomes Project Consortium et al., 2015; Firth et al., 2009; MacDonald et al., 2014). The most comprehensive database, gnomAD, has nearly 435,000 structural variant calls from almost 15,000 WGS samples (Collins et al., 2020). However, structural variants and larger indels are generally underrepresented in the current variant databases, and their diversity in ancestry groups is more limited than for other variants (Lappalainen et al., 2019). GnomAD includes structural variants from only 170 samples representing subpopulations, and 46% of the studied individuals are of European ancestry (Collins et al., 2020). Non-coding variation is not generally represented in these databases, while pathogenic non-coding CNVs have been also recognized (Szafranski et al., 2013).

The older CNV databases integrate detections originating from array CGH, Sanger sequencing and targeted short-read MPS studies, while in the newer databases detections originate from short-read and long-read WES and WGS studies (Zarrei et al., 2015; Karczewski et al., 2020). CNV detection with older platforms, such as BAC CGH, tends to overestimate CNV sizes due to lower resolution, and the smallest variants are missed (Zarrei et al., 2015). Therefore, Zarrei and colleagues provided a curated CNV map for DGV, the first comprehensive CNV database (Zarrei et al., 2015). Many of the filtered CNVs were singleton detections and extremely rare,

possibly false positive or late onset pathogenic variants, which are common problems for databases (Zarrei et al., 2015). Deletions tend to be easier to detect with any platform and are probably overrepresented in the databases (Hwang et al., 2015; Zarrei et al., 2015). Lack of exact coordinates for the structural variants in some databases makes their utilization systematically in annotation challenging (Neerman et al., 2019). Exact solving of breakpoint junctions (which is becoming more feasible with some emerging sequencing technologies) could consolidate common CNVs. They tend to appear different in databases due to usage of variable CNV detection platforms and inaccurate breakpoint detection with current methods (Newman et al., 2015). Just recently, the first studies with structural variant detection from long-read sequencing data including various nationalities and even subpopulations such as the Finnish have been conducted, but these sets are not yet widely available (Audano et al., 2019).

A population frequency of MAF > 1% is accepted also for CNVs as a threshold for common variation (Riggs et al., 2019). MAFs of 2% and 0.5% have been used as well, so the policy is more variable for CNVs than SNVs and indels (Neerman et al., 2019; Collins et al., 2020). However, the workflow does not provide exact technical recommendations for comparison of CNVs to databases. Comparison of CNVs is not as unequivocal as for SNVs and indels. Deletions have a more straightforward genomic structure and their comparison to databases is thus easier (Newman et al., 2015). Overlap requirements of 50% to 80%, either reciprocal or non-reciprocal, have been utilized in the different CNV detection studies discussed previously (Neerman et al., 2019; Hwang et al., 2015; Kosugi et al., 2019; Tan et al., 2014; Trost et al., 2018; Zhang, L. et al., 2019). These differences complicate comparison. Also possible different effects of deletions and duplications need to be considered, which necessitates differentiation for the state in the databases and database searches (Liu, P. et al., 2014).

Non-recurrent rearrangements have scattered breakpoint locations but usually a recognizable minimal overlapping region conferring the similar phenotypes (Carvalho, C. M. and Lupski, 2016; Lupianez et al., 2015). Nevertheless, the prevalence of unique rearrangements in disease-affected individuals makes the interpretation of their clinical effect more difficult (Carvalho, C. M. and Lupski, 2016; Rice and McLysaght, 2017). Then again, patients with overlapping phenotypes and non-identical breakpoints in CNVs can help delineate genomic regions causative for phenotypes (Weischenfeldt et al., 2013). Some patients were first recognized and classified based on a finding of a common overlapping genetic lesion rather than clinical features (Szafranski et al., 2013; Watson et al., 2014). This represents also a transition from phenotype-first approach to genotype-first approach.

Additional aspects of CNVs missing from the workflow need to be considered in certain settings. The ACMG recommendations for reporting incidental findings do not take into account disorders commonly caused by repeat expansions or CNVs as the primary cause (Green et al., 2013). Approximately 0.15% of cases in studies with different cohorts seem to have a CNV finding in a gene from the ACMG incidental findings gene list (Pfundt et al., 2017; Collins et al., 2020). As a second aspect, biallelic CNVs may cause disease through homozygous

deletions, which is more probable in consanguineous populations, or contribute to pathogenesis via compound heterozygosity (Harel and Lupski, 2018; Boone et al., 2013; Pfundt et al., 2017). CNVs can also present with SNVs or indels in compound heterozygosity, necessitating combination of different types of variant detection methods and detection results for a conclusive genetic diagnosis (Charng et al., 2016). Each gene associated with a recessive disease may have different frequency of CNVs and SNVs, with an average of 13.5 times more SNVs than CNVs (Boone et al., 2013). Alternatively, in one study a structural variant was missed since two SNVs were detected first, which would have been enough to explain the phenotype (Neerman et al., 2019). This exemplifies how the analysis always has to be comprehensive in order to reveal all variants affecting the phenotype or conferring carrier status. Depending on the disease cohort, it is estimated that 4–6% of cases could be compound heterozygotes for pathogenic mutations in different genes leading to multiple diagnoses (Lionel et al., 2018; Yang et al., 2013).

Heterozygous CNV encompassing multiple genes can confer carrier status for multiple recessive conditions, sometimes combinatory complex syndromes, in contrast to carrier point mutations (Boone et al., 2013). A single heterozygous CNV may be causative if the genes it affects are part of the same pathway, which increases the mutational load on the functional level (Boone et al., 2013). Structural variants encompassing several genes are thought to exert their pathogenic effect largely through this effect on gene dosage rather than by positional effects (Weischenfeldt et al., 2013). Even when not associated with disease-causing alleles, structural variation affects the number of total alleles in a study. The resulting distortion to significance calculations is comparable to a change in sample size (Schuster-Bockler et al., 2010). Therefore, structural variants make interpretation of SNVs in possible compound heterozygosity more difficult by influencing relative SNV frequencies (Turajlic et al., 2019).

Considering current CNV annotation programs, ANNOVAR can annotate previously reported CNVs and highlights overlapping genes but does not provide pathogenicity estimations on the genomic level (Ganel et al., 2017). Only rudimentary prediction tools are available for structural variants, such as SVScore, which aggregates SNP pathogenicity scores per base (Ganel et al., 2017). Some programs for CNV annotation were developed only for specific diagnostic settings. Anaconda is a tool integrating both detection and annotation of somatic CNVs in tumor WES data (Gao et al., 2017). Biofilter is an annotation tool, which groups CNV detections by biological pathways utilizing databases with genes, pathways and protein families (Kim, D. et al., 2016). The tool also differentiates between genes with surplus of rare or common CNVs and allows mainly CNV enrichment analysis. Most of the free programs are not up-to-date with the most recent CNV population databases (Samarakoon et al., 2016), whereas the most competent CNV annotation programs tend to be integrated into commercial diagnostic MPS pipelines (Neerman et al., 2019; Lassuthova et al., 2016).

2.4.4 Future directions for variant detection and annotation

Short-read WGS is estimated to become the standard in diagnostic screening (Collins et al.,

2020). With existing widely used tools (GATK), approximately only 72% of the genome can be scanned effectively from short-read WGS data. More specifically, these are the unique genomic regions (Regier et al., 2018; Goodwin et al., 2016; Lappalainen et al., 2019). Genomic regions prone to misalignment with short-read sequencing data include segmental duplications and high-copy repeats, which provide highly discordant SNV and indel calls (Regier et al., 2018). This problem is probably not solvable with improvements in data analysis and algorithms, but increase in read length and involving intronic sequences could enable more accurate alignment (Lappalainen et al., 2019).

The utility of Nanopore has been demonstrated for genome *de novo* assembly (Madoui et al., 2015) and phasing highly similar *MHC* haplotypes (Jain et al., 2018). Both Nanopore and PacBio long reads have enabled addition of novel sequence to the human reference genome. These sequences have mostly consisted of short and long tandem repeats (Chaisson, M. J. et al., 2015; Jain et al., 2018). Since typical target enrichment approaches are unsuitable for Nanopore sequencing, completely novel applications involving CRISPR-Cas9 enrichment or enabling selective sequencing of targets in real-time are being developed, which increase the specificity of the platform and enable more accurate variant calls (Kovaka et al., 2020). The per-base error rates of these platforms are still high, which limits their usage especially in SNP and indel calling (Lappalainen et al., 2019; Hehir-Kwa et al., 2018). Additionally, no applications have yet been developed in long-read sequencing to measure transcript levels. By contrast, transcriptomic structures and novel splice isoforms can be recognized with higher resolution since entire mRNA transcripts can be covered by a single read (Goodwin et al., 2016).

Combining sequencing data from different platforms in variant calling has already been generally proposed to increase variant calling specificity by covering for platform specific errors and biases (Rieber et al., 2013; Ross et al., 2013). Recently, inaccurate long reads (Nanopore) have been used to phase and thus to increase specificity of SNPs called with more accurate short reads (Illumina) (Jain et al., 2018; Turajlic et al., 2019). Also linked reads have been used to reconstruct haplotypes and to identify complex structural variation and balanced events (Marks et al., 2019). However, these synthetic long reads have brought only modest improvements to reference-based variant detection and have so far been limited to specific regions and variant types (Marks et al., 2019; Lappalainen et al., 2019).

It is estimated that with long-read sequencing 10% of the genome is still inaccessible for structural variant analysis. These involve mostly large and complex regions, which would require even longer reads to span them, together with improvements in computational tools for assembly (Audano et al., 2019). In a previous study, stretches of segmental duplications and inversions longer than the read length of 20 kb remained unresolved (Chaisson, M. J. et al., 2015). Therefore, it is estimated that a combination of both variant calling algorithms and genome analysis platforms is needed to capture all variation (Chaisson, M. J. P. et al., 2019; Collins et al., 2020). In contrast to sequencing based methods, optical mapping can capture

repetitive and unknown sequences but has a low resolution of 1 kb. Algorithms for calling variants from this data are also still in development (Hehir-Kwa et al., 2018). Nevertheless, in one recent study short-read and long-read sequencing, synthetic long-read sequencing and optical mapping were used together to increase structural variant detection sensitivity and to cross-validate findings (Chaisson, M. J. P. et al., 2019). In a study for somatic variation, complex structural variants were resolved and phased by combining high-throughput chromosome conformation capture, optical mapping and WGS (Dixon et al., 2018).

Some emerging sequencing-based applications require more comprehensive validation before a wide-spread adaptation for clinical usage (Salk et al., 2018). Methods involving identical barcoding of all molecules from single cells with flow sorting or droplet compartmentalization are becoming more common both for DNA and RNA applications. The combination of high-throughput compartmentalization and ultra-long-read single-molecule sequencing will enable WGS for large populations of single cells (Salk et al., 2018). Single-cell sequencing would be particularly informative in cancer studies since contamination from stromal cells could be avoided, and the measurement of genotype and phenotype from the same cell would be possible (Turajlic et al., 2019). Sequencing of cell-free DNA of tumor origin in liquid biopsy samples is a promising non-invasive longitudinal approach for determining genetic makeup of a tumor and monitoring for response to treatment (Salk et al., 2018; Turajlic et al., 2019). Likewise, fetal DNA in the mother's blood can be collected in a non-invasive prenatal testing for MPS analysis (Salk et al., 2018).

As a drawback, most current genome analysis methods and the ones in development are based on comparison to the reference genome. The current human reference genome is a mosaic, and no canonical human reference genome with the most common haplotypes is yet available (Audano et al., 2019). Regardless of sequencing platform, alignment of reads tends to be inaccurate in regions of high genomic diversity, such as *MHC* locus, *KIR* genes and *DYP2D6* (Lappalainen et al., 2019). The detection of a CNV in all long-read sequenced subjects indicates that the current reference genome carries a minor allele or an error at this location (Audano et al., 2019). Increased human reference sequence accuracy would improve mapping and thus enable more accurate variant discovery (Audano et al., 2019). Sequencing pipelines are moving towards using the GRCh38 version of the reference genome (Regier et al., 2018). This enhanced version has closed gaps, localized some orphan sequences and includes multiple alternative loci (Hehir-Kwa et al., 2018). Nevertheless, the older GRCh37 version is still widely in use in various research centers and databases (Regier et al., 2018).

Rapid diagnostic workflows are needed for some severe diseases, where diagnosis can affect treatment and alter outcome. These include some metabolic diseases, aggressive cancer or infections (Clark et al., 2019; Goodwin et al., 2016). Ineffectiveness of optimization of the steps from sample receipt to output of genomic variants has revealed variant interpretation as the bottleneck for speed gain in the workflow (Clark et al., 2019). Nevertheless, diagnostic workflow with WGS, automated phenotyping and variant interpretation has been accomplished

in less than a day (Clark et al., 2019). However, the current attempts for automatic variant interpretation have run into trouble with variants with conflicting interpretations in variant databases (Clark et al., 2019). Clinical and phenotypic information needs to be standardized across all of the variant, disease and gene databases to increase their accuracy and usability (Weischenfeldt et al., 2013). Universal structured phenotypic features, such as International Classification of Diseases (ICD) codes or diagnosis-related group (DRG) codes are still too sparse and unspecific for clinical phenotype description (Clark et al., 2019). Human Phenotype Ontology (HPO) terms with a hierarchical reference vocabulary have demonstrated better variability (Clark et al., 2019). Automated extraction of HPO terms from unstructured text has been promisingly even more accurate than manual interpretation (Clark et al., 2019).

In the near future, supervised autonomous systems may become an effective first-tier diagnostic approach and leave more time to concentrate on unsolved and difficult cases (Clark et al., 2019). Automated systems could also be used to reanalyze unsolved cases periodically, as previously discussed, and to standardize analysis pipelines. Machine learning approaches have been a promising development for *in silico* prediction programs for variants and their genomic effects, such as some previously elusive intronic variants and their effects on splicing (Lappalainen et al., 2019; Jaganathan et al., 2019). As opposed to the comprehensive automated variant analysis systems, the significance of gene-specific annotation guidelines (as compared to the ACMG guidelines) has been recently realized in increasing the amount of significant variant annotation results (Rivera-Muñoz et al., 2018). For some genes, the ACMG guidelines can be too generic: for example, the variant effect prediction tools can be misleading for very large and complex genes (Savarese et al., 2020). Both individual research groups and multi-institutional workgroups, such as the Clinical Genome Resource variant curation panels, are developing new gene and disease-level variant interpretation recommendations (Rivera-Muñoz et al., 2018; Savarese et al., 2020).

As stated earlier, differences and batch effects between MPS studies with different workflows are a problem when comparing variant calls with public databases and between studies (Lappalainen et al., 2019). A common pipeline would enable aggregate joint calling of variants from increased sample size and increase statistical power (Regier et al., 2018). The current best practices pipelines may be at a turning point. Regardless of the original supposed flexibility of the programs, BAM format has been unable to handle the newly available ultra-long reads. SAM or CRAM formats together with a separate long-read specialized aligner have been utilized to decrease computing times and RAM requirements (Jain et al., 2018; Sedlazeck et al., 2018). Therefore, current pipelines should be updated for new data types (long reads), file formats and tools, if they become widely used and approved (Regier et al., 2018). However, the trend seems to be that mainly commercial providers have enough resources to develop validated comprehensive pipelines with variant detection, filtering and annotation included (Neerman et al., 2019; Lassuthova et al., 2016). Technically, these pipelines have the potential to become gold standard approaches, but in practice they are not as easily available as BWA and GATK have been. Especially in the annotation of structural variants, internal restricted databases from

these providers are used to evaluate overlap with normal and pathogenic variants, the latter of which are scarcely available in public databases (Neerman et al., 2019; Lassuthova et al., 2016). These are both progressive and worrying developments, since data-sharing is important to facilitate future discovery of novel disease-causing variants and genes.

2.5 Genetic diagnosis of neuromuscular disorders

Neuromuscular disorders (NMDs) are one of the most heterogeneous group of disorders both clinically and genetically. Most are genetic in origin, and the Gene Table of Neuromuscular Disorders has over 1000 listed diseases and 587 different genes identified so far (Bonne et al., 2018) (http://www.musclegenetable.fr/index.html). NMDs are currently categorized into 16 main groups with different forms of muscular dystrophies (two), myopathies (five), myotonic syndromes, ion channel muscle diseases, malignant hyperthermia, congenital myasthenic syndromes, motor neuron diseases, hereditary ataxias, paraplegias, motor and sensory neuropathies and other neuromuscular disorders (Table 1).

Table 1. Neuromuscular disorders as categorized in The Gene Table of Neuromuscular Disorders.

1. Muscular dystrophies	**5. Other myopathies**	**9. Metabolic myopathies**	**13. Hereditary ataxias**
2. Congenital muscular dystrophies	**6. Myotonic syndromes**	**10. Hereditary cardiomyopathies**	**14. Hereditary motor and sensory neuropathies**
3. Congenital myopathies	**7. Ion channel muscle diseases**	**11. Congenital myasthenic syndromes**	**15. Hereditary paraplegias**
4. Distal myopathies	**8. Malignant hyperthermia**	**12. Motor neuron diseases**	**16. Other neuromuscular disorders**

NMDs affect primarily the peripheral nervous system, muscle tissue or neuromuscular junctions by damaging development, function or survival of their cellular components (Efthymiou et al., 2016; Ankala et al., 2015). Diseases affecting predominantly muscles include myopathies, dystrophies, ion channel diseases and malignant hyperthermia (Efthymiou et al., 2016; Vasli and Laporte, 2013). Diseases affecting primarily nerves include Charcot-Marie Tooth disease (CMT), motor neuron disease and hereditary spastic paraplegia, which frequently are present in combinations. Myasthenic syndromes affect neuromuscular junctions, which can be both pre- and postsynaptic (Efthymiou et al., 2016; Vasli and Laporte, 2013). However, myogenic disorders can also affect the innervating nerves as they progress, and some genes have functions both in muscle and nerve, leading to a mixed phenotype (Laing, 2012). All of the separate diseases are rare and often severe; they can lead to muscle weakness and wasting, cramps, numbness and respiratory and cardiac involvement (Vasli and Laporte, 2013). In most cases, NMDs are caused by genetic defects with autosomal recessive or dominant or X-linked inheritance, and some are caused by mitochondrial DNA defects (Laing, 2012).

Muscular dystrophies vary in prevalence, age of onset, severity, spectrum of affected muscles, and other features (Laing, 2012). Duchenne muscular dystrophy (DMD) is the most common inherited muscle disease in childhood, leads to severe comprehensive muscle weakness and

wasting, and has a prevalence of 8/100,000 (Carter et al., 2018). For adults, the most common forms are myotonic dystrophies with muscle wasting and myotonia, and facioscapulohumeral dystrophies (FSHD), with prevalence of 11/100,000 and 3/100,000 (Carter et al., 2018). Limb-girdle-muscular dystrophies (LGMDs) can present either in early childhood, adolescence or later in adulthood (Carter et al., 2018). Currently, four forms of autosomal dominant LGMDs (LGMDD1-D4) and 25 types of autosomal recessive (LGMDR1-R25) have been recognized. LGMD subtypes are variable in their age of onset, progression and severity. LGMDs lead to progressive and predominantly proximal muscle weakness (Straub et al., 2018).

DMD is an X-linked degenerative muscle disease affecting approximately 1 in 5000 males (Shieh, 2018). It is caused by mutations in the *DMD* gene, which encodes a sarcolemmal protein dystrophin. Large deletions (affecting more than one exon) can be found in 68% of the patients, with exons 45 and 55 being the most often involved. In the rest of the cases, 11% have large duplications and 3–11% have other mutation types (Bladen et al., 2015). Most (93%) of the CNVs affecting exons follow the reading frame hypothesis: preservation of the reading frame results in a milder Becker muscular dystrophy (BMD) phenotype (Bladen et al., 2015).The UMD/TREAT-NMD DMD database displays variants recorded in *DMD* and reading frame effect predictions for all possible duplications and deletions (http://umd.be/TREAT_DMD/, (Bladen et al., 2015)). Emerging treatment approaches for DMD include gene therapy with minidystrophin gene delivery or synthetic antisense oligonucleotides for exon skipping, which would restore the reading frame. Both aim to retain partial protein function (Shieh, 2018).

The prevalence of congenital muscular dystrophy varies globally, with Ullrich congenital muscular dystrophy (UMD) the most common form among them (Carter et al., 2018). Congenital myopathies are characterized by hypotonia (muscle weakness) at birth or in the first year of life (Nishikawa et al., 2017). They are traditionally subdivided into different categories based on histopathology findings, such as nemaline myopathy, central core disease and centronuclear myotubular myopathy (Nishikawa et al., 2017; Sewry et al., 2019). Metabolic myopathies can display clinically heterogeneous symptoms, such as muscle weakness, exercise intolerance or rhabdomyolysis (Nishikawa et al., 2017). They can be caused by defects in enzymes in glycogen or lipid energy metabolism, or by mitochondrial abnormalities affecting the energy delivery (Nishikawa et al., 2017).

Inherited peripheral neuropathies (IPNs) are the most common inherited neurological disorders (Cutrupi et al., 2018). Over 1000 mutations in more than 90 genes have been associated to IPNs (Cutrupi et al., 2018; Lassuthova et al., 2016). Onset is typically in the later childhood, but IPNs include also congenital and late onset adult forms (Antoniadi et al., 2015). The symptoms manifest classically as progressive symmetric wasting and weakness of distal limb muscles, mild to moderate loss of sensory function, areflexia in the upper and lower limbs, and foot deformities (Dohrn et al., 2017). IPNs are classified by clinical phenotype, inheritance mode, age of onset, electrophysiological studies and causal mutation (Antoniadi et al., 2015). IPNs can be broadly divided into three subtypes; they affect either motor nerves with pure motor

involvement (distal hereditary motor neuropathy, dHMN), sensory nerves with neuropathic pain and/or autonomous symptoms (hereditary sensory and autonomic neuropathy, HSAN), or both (hereditary motor and sensory neuropathy, or Charcot-Marie Tooth disease, CMT) (Bacquet et al., 2018; Hartley et al., 2018). IPNs include also hereditary neuropathy with a liability to pressure palsy (HNPP), and congenital hypomyelinating neuropathy (CHN) (Nam et al., 2016).

Hereditary motor and sensory neuropathy, or CMT is one of the main IPN subtypes with a prevalence of 1/2500 (Bacquet et al., 2018). It has high genetic and phenotypic heterogeneity (Antoniadi et al., 2015). CMT is divided into additional subtypes of either demyelinating (type I) or axonal (type II), and the inheritance can be autosomal dominant, autosomal recessive or X-linked (Antoniadi et al., 2015). In CMT type I, loss of myelination of the peripheral nerves leads to loss in nerve conduction velocity. In CMT type II, axonal loss in peripheral nerve fibers leads to decreased motor and sensory nerve action potentials (Bacquet et al., 2018). Patients can also present with both symptoms and thus have an intermediate form of CMT. Clinically, most patients develop distal motor and sensory weakness (Bacquet et al., 2018). CMT1A is the most common IPN, and it is caused by a 1.5 Mb duplication of *PMP22*, as discussed earlier. *PMP22* duplication accounts for 70–80% of CMT1 cases and 50% of all CMT cases (Pehlivan et al., 2016; Cutrupi et al., 2018).

NMDs without genetic causes, or acquired muscle disorders, include idiopathic immune-mediated myopathies. They can be categorized into necrotizing autoimmune myopathy, inclusion body myositis, dermatomyositis, polymyositis and nonspecific myositis (Milone, 2017). These diseases are characterized by muscle weakness, elevated creatine kinase levels and myopathic findings (Milone, 2017). Sporadic inclusion body myositis (sIBM) is the most common acquired muscle disease presenting in later age (over 50 years) (Needham and Mastaglia, 2016). It is separated from the other inflammatory myopathies by rimmed vacuolar myopathy, selective and often asymmetric muscle weakness and wasting, and by slowly progressive clinical course (Needham and Mastaglia, 2016). The pathogenic basis of sIBM is thought to involve both an inflammatory and a degenerative process (Milone, 2017). Perceived variability in prevalence has been thought to originate from differences in population frequencies of the *HLA-DRB1*03:01* allele, which is strongly associated with the disease (Needham and Mastaglia, 2016; Johari et al., 2017). Pathogenesis of sIBM is complex and likely multi-factorial, with increasing evidence for polygenic susceptibility of HLA and non-HLA gene involvement (Needham and Mastaglia, 2016; Johari et al., 2017).

2.5.1 Challenges in diagnosing NMDs and pre-MPS approaches

NMDs are an exemplary group of rare diseases with often unspecific symptoms, variable phenotypes, incomplete family history and regionally centered expertise (Dohrn et al., 2017). The diagnosis is especially challenging in early disease stage when a classical differentiating phenotype has not yet developed fully. On the other hand, some cases become challenging to distinguish in advanced stages, such as electrophysiological differentiation of demyelinating

and axonal neuropathy (Dohrn et al., 2017). Since many of the myopathies are late-onset, parent samples are often unavailable (Gorokhova et al., 2015). The mode of inheritance is challenging to decide with small families or without family history (Efthymiou et al., 2016). Altogether, with most disease subtypes patients will undergo multiple invasive and expensive tests (Ankala et al., 2015; Marelli et al., 2016).

Since NMDs are highly phenotypically heterogenic, molecular diagnosis is the gold standard for their diagnosis (Laing, 2012; Dohrn et al., 2017). Accurate molecular genetic diagnosis allows for prognosis, genetic counseling and development of therapeutic trials (Carter et al., 2018). Difficulties for molecular diagnosis of NMDs include genetic heterogeneity; multiple genes can be associated with a disease, and multiple diseases with one gene (Antoniadi et al., 2015). Mutations in a single gene can lead to different phenotypes with different modes of inheritance (Antoniadi et al., 2015). Furthermore, the age of onset, severity, progression and prognosis can vary for patients even with the same genetic defect (Carter et al., 2018). Thus, a mutation segregating in one family can be associated with phenotypically varying disorders, such as with a *DYSF* mutation causing either LGMD or distal myopathy (Illarioshkin et al., 2000). Neuromuscular disorders have also high prevalence of *de novo* events, making their diagnosis and estimation of prevalence even more difficult (Chae et al., 2015).

Often, genes to be tested can be narrowed down with protein staining by immunohistochemistry and immunoblotting (Ankala et al., 2015). Genes can also be chosen according to genotype-phenotype correlation (Bacquet et al., 2018). Originally, combination of multiple techniques including Sanger sequencing, MLPA and other PCR-based approaches, and array CGH were utilized to inspect genes for different types of mutations (Efthymiou et al., 2016). However, heterogenic and unspecific clinical and histopathological features in the initial stages of many of the disorders complicated this candidate gene search (Vasli and Laporte, 2013). DMD and BMD are the few diseases with a distinct disease gene and high diagnostic yield from a single molecular genetic test (Chae et al., 2015). Additionally, variants in four genes, *PMP22*, *GJB1*, *MFN2* and *MPZ,* explain approximately 90% of common CMT subtypes (DiVincenzo et al., 2014).

Genes and proteins associated with NMDs involve some of the largest humans have; many are building blocks essential for muscles (Laing, 2012). These giant muscle genes include *TTN*, *NEB*, *RYR1* and *DMD* (Vasli and Laporte, 2013). *NEB* has 183 exons and multiple differently spliced isoforms (Donner et al., 2004). Mutations in *NEB* are the most common cause for recessive nemaline myopathy (Pelin et al., 1999). Most of these mutations have been detected in compound heterozygosity (Lehtokari et al., 2014). Altogether 12 disease genes are currently known to cause nemaline myopathy (Sewry et al., 2019). *DMD* is genomically the largest human gene with 79 exons, some of the largest introns and a total span of 2.3 Mb (Koenig et al., 1987). However, *TTN* has the largest coding sequence, produces the largest human protein and has 363 exons in its longest isoform (Savarese et al., 2016; Vasli and Laporte, 2013). *TTN* has been associated with various cardiomyopathies and skeletal muscle diseases. These include

late-onset autosomal dominant tibial muscular dystrophy (TMD), limb-girdle-muscular dystrophy type R10 (LGMDR10), hereditary myopathy with early respiratory failure (HMERF) and congenital centronuclear myopathy (CNM) (Savarese et al., 2016). Dominant late onset disease and young or early adult onset recessive disease are caused by mutations in different specific exons of the gene (Savarese et al., 2016). For example, TMD is a mild adult-onset distal myopathy with autosomal inheritance (Udd et al., 1993). A dominant founder mutation in the Finnish population, FINmaj, is an 11 bp insertion-deletion variant and responsible for relatively high prevalence of 2/10,000 of TMD in Finland (Udd et al., 1993; Hackman et al., 2002).

Originally, inspection of these large genes required multiple Sanger sequencing runs, which was costly (Laing, 2012). Alternatively, the genes were not fully screened (Vasli and Laporte, 2013). For example, the approach of multiplex PCR was targeted to only certain hotspot exons of *DMD* (Vasli and Laporte, 2013). Especially *TTN* with its 363 exons was rarely screened completely with Sanger sequencing (Gorokhova et al., 2015), although phenotypic variability of muscular dystrophies caused by *TTN* mutations has been explained in some cases with findings of additional, modifying *TTN* mutations (Evila et al., 2014).

2.5.2 Advancements in diagnosis of NMDs with MPS approaches

WES was originally recommended for patients with nonspecific or unusual disease presentations, or with genetically heterogeneous diseases (Yang et al., 2013). Thus, MPS analysis is beneficial for many of the NMD categories (Nigro and Piluso, 2012). CMT was one of the first disease groups for which comprehensive analysis by MPS approaches was used (Laing, 2012). MPS is advantageous also for dystrophinopathies since *DMD* is a large gene with a wide spectrum of point mutations and CNVs recognized (Bladen et al., 2015). Compared to gene-by-gene screening, MPS approaches have provided increases of 2- or 3-fold in diagnostic yield for NMDs (Bacquet et al., 2018; Ankala et al., 2015). Overall, MPS approaches have enabled effective, cost-and time-saving diagnosis in many NMD subgroups (Bacquet et al., 2018; Ankala et al., 2015; Cordoba et al., 2018; Nam et al., 2016; Dohrn et al., 2017; Antoniadi et al., 2015). Early diagnosis before onset or worsening of clinical phenotype may enable treatment of later complications and trials for intervention rather than treating of patients with already significant disease progression (Alkuraya, 2015).

With MPS approaches, novel disease-causing mutations have been found in patients who had waited for diagnosis for more than a decade (Vasli and Laporte, 2013). MPS approaches have enabled discovery of multiple new NMD disease genes. First such discoveries were made with MPS as a follow-up approach after a region of interest had been defined with genome-wide linkage studies (Nigro and Piluso, 2012; Gorokhova et al., 2015). Patients with matching clinical findings have been sequenced together to screen multiple genes simultaneously and to find the common genetic cause (Nigro and Piluso, 2012; Gorokhova et al., 2015). Genes from mitochondrial components can be inspected as well and MPS approaches have brought more insight into genetics of mitochondrial disorders (Chae et al., 2015; Efthymiou et al., 2016). MPS studies have also corrected wrong initial diagnoses in diseases with closely similar phenotypes

and unexpected genetic findings (Dohrn et al., 2017). Incomplete or misleading clinical information or misclassified mode of inheritance can prevent initial correctly aimed pre-MPS gene screening, and hence discovery of pathogenic variants by MPS also in diseases with distinct prioritized single genes (Lassuthova et al., 2016).

MPS studies have widened phenotypic spectrum for known disease genes (Hartley et al., 2018). In a study cohort of neuropathies, the causative gene was not primarily associated with inherited peripheral neuropathy in 10% of the diagnosed cases (Antoniadi et al., 2015). New allelic diseases have been discovered especially for large genes, such as *TTN, MYH7* an *RYR1* (Gorokhova et al., 2015). In fact, routine sequencing of the entire *TTN* gene in research and diagnostics was impossible preceding emergence of MPS techniques (Savarese et al., 2016). Due to the size of the gene and previous neglect in whole gene analysis, *TTN* variants identified in MPS studies have rarely been previously characterized (Savarese et al., 2016). This has led to a troublesome situation; mutations in *TTN* have been ignored in diagnostics and research because many of them are scarcely annotated (Gorokhova et al., 2015). Nevertheless, some of the most recent phenotype additions for genetic defects in *TTN* were discovered through MPS studies (Gorokhova et al., 2015).

By enabling comprehensive investigation of pathogenic variants, MPS approaches have revealed cases of simultaneous presentation with two rare muscle disorders. One such case was a combination of BMD caused by a *DMD* deletion, and rippling muscle disease caused by a *CAV3* mutation (Hiraide et al., 2019). In cohorts involving neuropathies and myopathies, 2–5% of the patients have been commonly found with a pathogenic variant in more than one gene, leading to complex phenotypes (Posey et al., 2017; Antoniadi et al., 2015). Mixtures of different phenotypes may be erroneously interpreted as a new clinical entity or a phenotypic expansion for one disease (Volk and Kubisch, 2017; Karaca et al., 2018). Pathogenic variants can also be overlooked if the first finding explains the phenotype partially (Gorokhova et al., 2015). However, finding multiple genetic variants has implications for recurrence risk and genetic counseling, in addition to allowing more accurate prognosis and treatment (Antoniadi et al., 2015). Pathogenic variants in more than one disease gene can also explain phenotypic variation within families (Karaca et al., 2018).

Targeted gene panels have been a common approach for clinical diagnostics of NMDs. They include typically dozens to hundreds of genes and either for specific NMD subtypes or for NMDs generally (Evila et al., 2016; Kitamura et al., 2016; Chae et al., 2015; Bacquet et al., 2018; Antoniadi et al., 2015; Savarese et al., 2016; Nam et al., 2016; Lassuthova et al., 2016; Nishikawa et al., 2017; Zenagui et al., 2018). This targeted approach allows for strong optimization of coverage on relevant targets (Dohrn et al., 2017). Incidental findings can also be avoided, which in NMDs may include certain heart conditions and malignant hyperthermia, even on gene panels (Gorokhova et al., 2015). One drawback is the rapid discovery of new disease genes, which necessitates continuous update of the panels (Volk and Kubisch, 2017; Zenagui et al., 2018). For example, in a WES study with a NMD cohort many of the causative

defects were discovered in genes recognized during the last 10 years, which may not be represented in gene panels (Waldrop et al., 2019). However, many of the gene panels are already designed with inclusion of candidate genes, alleviating this drawback (Lassuthova et al., 2016; Evila et al., 2016). During one study for IPNs, the panel was redesigned according to emerging knowledge during the course of the study, and the added genes explained 10% of cases (Lassuthova et al., 2016).

Utility of RNA sequencing has been demonstrated also for NMDs. In a cohort with rare muscle disorders, RNA-seq provided a diagnostic yield of 35% (Cummings et al., 2017). Splice-site disrupting mutations were validated and deep intronic variants were revealed to create novel splice sites or activate alternate cryptic splice sites (Cummings et al., 2017). Additionally, a splice site affecting mutation was identified in an exon, which was not adequately covered in WES data (Cummings et al., 2017). However, a majority of the commonly disrupted NMD genes are not expressed in high quantities in blood or fibroblasts, which necessitates often invasive sampling of the affected tissues (Cummings et al., 2017; Volk and Kubisch, 2017). For IPNs, obtaining the disease relevant tissue is impossible (Cutrupi et al., 2018), while muscle biopsy is one of the easier tissue sources (Volk and Kubisch, 2017). However, muscle biopsies may be contaminated by skin or fat, which could originate also from late-stage degenerative muscle pathology (Cummings et al., 2017).

Diagnostic yield for NMD patients with targeted gene panel or WES has remained approximately 40% (between 26% and 48%) (Vasli and Laporte, 2013; Waldrop et al., 2019; Kitamura et al., 2016; Chae et al., 2015; Bacquet et al., 2018; Cordoba et al., 2018; Antoniadi et al., 2015; Punetha et al., 2016; Marelli et al., 2016; Savarese et al., 2016; Lassuthova et al., 2016; Nam et al., 2016). The diagnostic rate has generally been higher in homogeneous groups and cohorts carefully pre-selected for muscle symptoms (Chae et al., 2015). Originally, MPS studies were rarely used as the first-tier diagnostic test, which also affects the achievable diagnostic yield (Gorokhova et al., 2015; Volk and Kubisch, 2017). In a rare comparison study with utilization of four different targeted gene panels (for muscular dystrophy, congenital myopathy, congenital myasthenic syndrome, metabolic myopathy and myopathy with protein aggregations/rimmed vacuoles), the highest diagnostic yield of 46% was achieved for muscular dystrophies (Nishikawa et al., 2017). WGS approaches have not yet been utilized in routine clinical diagnostics of NMDs (Efthymiou et al., 2016).

MPS studies tend to provide multiple variants of uncertain significance. In the case of NMDs, deciphering clinical significance for these requires often interdisciplinary collaboration with clinicians and neuropathologists in addition to the molecular geneticists (Bacquet et al., 2018). Precise phenotyping with clinical, electrophysiological and sometimes histopathological patterns is usually required to verify diagnosis and separate disorders with phenotypic heterogeneity (Dohrn et al., 2017). Functional studies can include biochemical analysis of mutated protein from muscle biopsy, protein assays, splicing assay and animal models (Gorokhova et al., 2015). Albeit commonly too laborious for routine clinical diagnostics, these

studies will also be essential to increase knowledge on the pathogenesis of the diseases (Vasli and Laporte, 2013).

2.5.3 Detection of CNVs in NMDs with MPS approaches

Preceding MPS approaches, array CGH has been used to detect CNVs in NMD genes (Vasli and Laporte, 2013). Currently, more comprehensive arrays are being designed for NMD genes, revealing causative CNVs in genes such as *SGCG* and *LAMA2* (Piluso et al., 2011; Sagath et al., 2018; Giugliano et al., 2018). Originally, it was thought that MPS had not enough potential for detecting CNVs (Laing, 2012). In many of the older MPS studies for NMDs, no CNV analysis was performed, neither from MPS data nor with complementary methods (Kitamura et al., 2016; Chae et al., 2015; Evila et al., 2016; Nishikawa et al., 2017). In some studies, CNVs have been detected with complementary methods, such as MLPA (Dohrn et al., 2017; DiVincenzo et al., 2014) or array CGH (Pehlivan et al., 2016; Ankala et al., 2015). Recently, CNV analysis from MPS data has been included more routinely in many of the studies, using generally much evaluated programs such as ExomeDepth (Bacquet et al., 2018), CoNIFER (Pfundt et al., 2017) or XHMM (Hiraide et al., 2019).

In a study with multiple disorder groups, CNVs were found to be pronounced as disease causing entities especially in neurological diseases. This included CMT, neuropathies, muscular dystrophy, neuromuscular disorders and spastic paraplegia, and more specifically genes *SMN1*, *PMP22* and *DMD* (Truty et al., 2019). Deletions and duplications of *PMP22* are one of the first CNVs which were actively inspected from MPS data (Antoniadi et al., 2015; Nam et al., 2016). Generally, CMTs are explained by *PMP22* duplications in 43–57% of the cases and by *PMP22* deletions in 22–27% of the cases (DiVincenzo et al., 2014; Lassuthova et al., 2016). Atypical CNVs involving only parts of *PMP22* have also been reported (Cutrupi et al., 2018) and rarely (< 3% of cases) CNVs outside *PMP22* have been found as causative for CMT (Pehlivan et al., 2016). A separate tool has been developed to screen for the many CNVs of *DMD* from MPS data, with detection on exon-level a requisite for correct diagnosis and a qualification for some therapeutic trials (Kozareva et al., 2018).

Pathogenic deletions and duplications in *SPAST* explain 8–41% of autosomal dominant spastic paraplegia 4 (SPG4) cases (Boone et al., 2014). SPG4 is characterized by spasticity and progressive weakness in lower limbs (Boone et al., 2014). *SPAST* has *Alu*-rich genomic architecture, which predisposes the region to different genomic rearrangements (Boone et al., 2014). Both deletions and duplications in *SPAST* have been reported to affect various exons and most often the final or the first (Boone et al., 2014). Some deletions which span beyond *SPAST* to nearby genes are predicted to form novel chimeric genes (Boone et al., 2014). This could for example explain a family with co-segregation of spastic paraplegia and dementia (Boone et al., 2014).

Variant types challenging to detect from MPS data and involved in NMDs involve repeat expansions and copy number changes in highly homologous genes and repeated regions (Volk

and Kubisch, 2017; Zenagui et al., 2018). Both *TTN* and *NEB* harbor a so-called triplicate repeat region with repetitive blocks of exons: *TTN* between exons 172 and 198 with nine exon units, and *NEB* between exons 82 and 105 with eight exon units (Bang et al., 2001; Kiiski, K. et al., 2016). Copy number changes affecting more than one repeat block (i.e. triplicate repeat copy number of 4/6 or 8/6 or more) in the same allele on *NEB* have been detected to be deleterious (Kiiski, K. et al., 2016), but similar findings have not yet been made for *TTN*.

Myotonic dystrophy 1 (DM1) and some spinocerebellar ataxias are caused by trinucleotide repeat expansions and myotonic dystrophy 2 by a tetranucleotide repeat expansion (Mirkin, 2007; Volk and Kubisch, 2017). In myotonic dystrophies, the size of the expansion correlates with the severity of the disease and through further repeat expansion in the generation leads to anticipation in DM1 (Laing, 2012; Tsilfidis et al., 1992). Normally, the CTG repeat in *DMPK* associated with DM1 consists of 5 to 27 copies, but affected individuals can have from 50 copies in a mild disease form to several kb of repeat in more severe cases (Brook et al., 1992). Repeat retraction in the D4Z4 microsatellite causes FSHD type 1, and a hexanucleotide repeat expansion in *C90rf72* causes amyotrophic lateral sclerosis (ALS) (Volk and Kubisch, 2017). Single gene testing by PCR-based fragment length evaluation or other PCR approaches are the standard methods for these cases (Volk and Kubisch, 2017). MPS-based approaches can potentially replace these techniques, but generally technological, bioinformatic and cost-effectiveness issues need to be solved first (Volk and Kubisch, 2017). Especially long-read sequencing or synthetic long reads may resolve these challenging to detect variants, as previously discussed. Nevertheless, some repeat expansion defects can be detected already from short-read MPS data, as demonstrated in a cohort of ataxia patients (Marelli et al., 2016; Dashnow et al., 2018).

Different forms of autosomal recessive proximal spinal muscular atrophy are caused by a homozygous deletion of *SMN1* (Lefebvre et al., 1995). *SMN2* is a closely related pseudogene for *SMN1* but with a critical C>T change in exon 7, which impairs the function of an exonic splicing enhancer and reduces the amount of functional full-length *SMN2* transcript (Lorson et al., 1999). Homozygous deletion of *SMN1* is the dominant disease cause, but also smaller variants, such as SNVs or indels, and partial gene conversions producing *SMN1/SMN2* hybrid genes have been detected (Lorson et al., 1999; Feng et al., 2017). Increased copy number of *SMN2* provides partial functionality and affects the severity of the disease resulting from loss of *SMN1*; two copies of *SMN2* lead to presentation as SMA type I and four copies to SMA III (Lefebvre et al., 1995). MLPA has been used to differentiate these two genes, but MPS approaches are emerging (Feng et al., 2017). A fast and comprehensive analysis method would be needed to achieve early diagnosis to initiate a newly approved treatment in time for maximized response. This first approved treatment is based on increasing the amount of functional *SMN2* transcript by preventing exon skipping with antisense oligonucleotide administration (Mercuri et al., 2018).

3 AIMS OF THE STUDY

The aim of this study was to develop an accurate and standardized copy number variant (CNV) detection and annotation method for targeted gene panel and whole exome sequenced data. The developed method was utilized to increase the diagnostic yield in a cohort of patients with neuromuscular disorders. The prevalence of CNVs was evaluated in the genes associated with neuromuscular disorders.

The specific objectives in each of the subprojects were:

I. Validation of a CNV analysis program combination based on complementary detection sensitivity and specificity to attain highly sensitive CNV detections. Evaluation of sensitivity and specificity of the CNV detections and the accuracy of the detected regions by verification with complementary methods of either array comparative genomic hybridization, PCR or multiplex ligation-dependent probe amplification. Development of additional scripts to differentiate difficult to analyze regions: homologous genes *SMN1/SMN2* and *NEB* triplicate repeat region.

II. Improving the specificity of the CNV detection approach by developing a logistic regression model to differentiate detections predicted to be true positive. A set of *in silico* CNVs was generated into sequenced samples to train and test the model. Positive control samples and validated CNV detections from the preceding subproject were utilized in model validation.

III. Development of a comprehensive annotation pipeline for CNV detections by utilizing and expanding an existing program cnvScan. The most recent CNV population databases were included and an in-house CNV database was constructed to enable filtration of results by frequency for clinical significance evaluation.

IV. Validation of the developed pipeline for CNV detection sensitivity, predictive model performance and CNV annotation for WES samples.

4 MATERIALS AND METHODS

4.1 Sequencing data preparation

4.1.1 Subjects

DNA was extracted from peripheral blood samples from patients with neuromuscular disorders or their unaffected relatives. 2359 samples were received and sequenced with the targeted gene panel MYOcap (described below). The sample set included also 16 unaffected relatives and 55 DNA samples sent as CNV control samples by foreign collaborators (clinicians and researchers). The control samples included 24 samples with a known previously validated CNV (positive control samples) and 31 samples with certain genes verified to not contain CNVs (negative control samples). For MNDcap, 942 samples were received, one of which was an unaffected relative, and two had a known CNV (positive control samples). For WES, 262 samples were received for sequencing, 22 of which were from unaffected relatives and 85 of which were from sIBM patients. Written informed consent was collected from patients or their representatives. Samples were obtained according to the Declaration of Helsinki, and the study was approved by the Coordinating Ethics Committee of the Hospital District of Helsinki and Uusimaa (HUS 195/13/03/00/2011 §32). Samples received in collaboration were received as extracted DNA samples.

4.1.2 Targeted gene panels and sequencing

In this study, six versions of the customized gene panel called MYOcap and three versions of MNDcap were utilized (Evila et al., 2016). The gene panels included exonic regions, UTRs and +/- 15 bp intronic regions for genes and candidate genes related to myopathies (MYOcap) and neuropathies (MNDcap). MYOcap covered also certain specific intronic regions known to contain some well-known causative variants. In addition, the MYOcap v.6 panel covered the whole *TTN* gene with intronic probes included. The latest versions of the panels contain sequences of 341 (MYOcap) and 301 (MNDcap) genes (Supplemental Table 1 and Table 2 for gene lists for all the panel versions).

DNA target capture, enrichment and sequencing for the targeted gene panels were performed either at FIMM (Institute for Molecular Medicine Finland, Helsinki, Finland), FuGU (Biomedicum Functional Genomics Unit, Helsinki, Finland), or Oxford Genomics Centre (Oxford Genomics Centre, Oxford, UK) (Table 2). Hybridization-based target capture was performed with different versions of a custom-designed SeqCap EZ Choice Library (Roche Sequencing, Pleasanton, USA). Sequencing was performed with the Illumina HiSeq 1500–4000, MiSeq or Novaseq platform (Illumina Inc., San Diego, CA) to 150 bp, 100 bp or 75 bp paired-end read lengths. The achieved average target read depths varied from 110X to > 1000X. MYOcap sequencing was performed with batches of 17 to 160 samples with an average of 58 samples per run. 25 samples were sequenced twice, totalling 2384 MYOcap sequenced samples. With MNDcap, a total of 948 samples were sequenced in batches of 18 to 96 samples with an average of 43 samples per run, and six samples were sequenced twice. Additionally, 25 samples

were sequenced both in MYOcap and MNDcap.

Table 2: Target set version and sequencing information for the targeted gene panels MYOcap and MNDcap.

Panel version	Genes in panel	Batches	Samples	Provider(s) and platform(s): number of batches	Average achieved read depth* and read length
MYOcap v.1	218	2	96	FIMM HiSeq1500: 2	110X | 100 bp
MYOcap v.2	238	7	337	FIMM HiSeq1500: 1	130X | 100 bp
				FuGu MiSeq: 2	150X | 75 bp
				Oxford HiSeq2000: 4	560X | 100 bp
MYOcap v.3	315	12	716	FIMM HiSeq2500: 1	140X | 100 bp
				Oxford HiSeq4000: 11	650X | 75 bp
MYOcap v.4	300	8	384	Oxford HiSeq4000: 8	700X | 75 bp
MYOcap v.5	332	10	464	Oxford HiSeq4000: 10	580X | 75 bp
MYOcap v.6	349	5	387	Oxford HiSeq4000: 3	470X | 75 bp
				NovaSeq: 2	1010X | 150 bp
MNDcap v.1	278	1	24	Oxford HiSeq1500: 1	230X | 100 bp
MNDcap v.2	278	7	352	FIMM HiSeq2500: 5	220X | 100 bp
				Oxford HiSeq4000: 2	670X | 75 bp
MNDcap v.3	302	14	572	Oxford HiSeq4000: 11	660X | 75 bp
				Novaseq: 3	1060X | 150 bp

*Average achieved read depth has been rounded to nearest 10.

4.1.3 Whole exome sequencing

DNA samples were processed and whole exome sequenced either at FIMM, FuGu, ATLAS (ATLAS Biolabs, Berlin, Germany), BGI (Beijing Genomics Institute Copenhagen, Denmark), GATC (now part of Eurofins Genomics, Ebersberg, Germany) or Blueprint Genetics (Blueprint Genetics, Espoo, Finland). The target capture kits and sequencing protocols (with paired-end read lengths and average read depths) are listed in Table 3. Providers for the targets used for WES included Agilent Technologies (Agilent Technologies, CA, USA), Axeq Technologies (Axeq Technologies Inc., MD, USA), Roche and Illumina. Like for the targeted gene panels, only different Illumina platforms (HiSeq 1500–4000, NextSeq, Novaseq) were utilized for sequencing. A total of 262 samples were sequenced with 29 samples on average per run, including two samples sequenced twice. Additionally, 114 samples sequenced with MYOcap and 17 samples sequenced with MNDcap were also whole exome sequenced.

Table 3: Sequencing information for WES batches.

Target kit	Samples	Provider and sequencing platform	Average achieved read depth*	Read length
Agilent SureSelect v2	30	BGI HiSeq 1500	80X	90 bp
Agilent SureSelectXT v6	28	GATC HiSeq 1500	30X	125 bp
Agilent SureSelectXT v6	19	BGI HiSeq 4000	50X	100 bp
NimbleGen MedExome	62	FIMM HiSeq 2500	140X	100 bp
NimbleGen MedExome	24	FIMM Novaseq	150X	100 bp
Axeq TruSeq exome	21	FuGU NextSeq	80X	75 bp
NimbleGen v2.0	27	FIMM HiSeq 1500	50X	90 bp
Nimblegen SeqCap EZ Exome v2	28	Atlas HiSeq 1500	100X	100 bp
BpG+xgen-exome-research-panel	23	Blueprint Genetics Novaseq	280X	150 bp

*Average achieved read depth has been rounded to nearest 10.

4.1.4 Sequencing data pre-analysis

Data pre-processing into BAM files starting from FASTQ files was performed according to the Genome Analysis Toolkit (GATK) (Broad Institute, Cambridge, MA) best practices basic protocol (DePristo et al., 2011; Van der Auwera et al., 2013; McKenna et al., 2010). Namely, reads were aligned to the Human Reference Genome version GRCh37/hg19 with Burrows-Wheeler aligner (version 0.7.10, 0.7.12, 0.7.15, or 0.7.17)(Li, H. and Durbin, 2009), duplicated reads were removed with Picard tools (version 1.119 or 2.18.10, Broad Institute, Cambridge, MA), misaligned reads were realigned around indels with SAMtools (version 1.2, Genome Research Ltd., Cambridge, UK (Li, H. et al., 2009)), and finally base quality recalibration was done with GATK (version 3.3, 3.7 or 4.1.0.0). The tool versions (Table 4) for this part of the workflow and others discussed further were varied because the time scale for sample acquisition, sequencing and analysis spanned from 2012 for the oldest WES batch until the latest batches sequenced and analysed in 2019.

Table 4: Program version and source information. Different bioinformatics tools used in this study, versions used and their internet reservoirs at the time of writing this thesis, 03/2020.

Program	Versions	Web source (as of 03/2020)
Burrows-Wheeler aligner / BWA	0.7.10, 0.7.12, 0.7.15, 0.7.17	http://bio-bwa.sourceforge.net/
Picard tools	1.119, 2.13.2, 2.18.10	http://broadinstitute.github.io/picard
SAMtools	1.2, 1.4	https://github.com/samtools/samtools
Genome Analysis Toolkit / GATK	3.3, 3.7, 4.1.0.0	http://www.broadinstitute.org/gatk/

CoNIFER	0.2.2	http://conifer.sourceforge.net/download.html
XHMM	1.1	https://atgu.mgh.harvard.edu/xhmm/download.shtml
ExomeDepth	1.1.10, 1.1.12	https://cran.r-project.org/web/packages/ExomeDepth/index.html
CODEX	1.10.0, 1.12.0	https://www.bioconductor.org/packages/release/bioc/html/CODEX.html
CNVkit	0.8.5	https://cnvkit.readthedocs.io/en/stable/
SavvyCNV	1.0*	https://github.com/rdemolgen/SavvySuite
BEDtools	2.26.0	https://bedtools.readthedocs.io/en/latest/
cnvScan	1.0	https://github.com/PubuduSaneth/cnvScan
CoverView	1.4.4	https://github.com/RahmanTeamDevelopment/CoverView
ExomeCQA	1.0*	http://exomecqa.sourceforge.net

*Program version was not specified by the developers.

Control samples for WES validation were received from Blueprint Genetics as ready-aligned BAM files. A commercial "sentieon-genomics-201711.01" pipeline based on GATK best practices was used for all the sequencing data preparation steps, with no specific details available.

4.2 CNV calling programs and pipeline

Multiple studies with program development and comparisons were evaluated to decide on the best approach for CNV calling from targeted gene panel and WES data (Table 5). The pre-set properties of the study setting, which restricted program selection, were both the aim to detect rare germline variants and the lack of normal control samples for comparison (de Ligt et al., 2013; Tan et al., 2014). For example, GATK recently provided a CNV calling tool, but it requires a panel of normal samples for normalization and thus was unsuitable for this setting (Hehir-Kwa et al., 2018). According to comparison studies, different programs seemed to provide complementary variant calls in specificity, sensitivity and variant types, and the use of multiple programs was recommended to achieve most comprehensive results. Therefore, we selected four programs, which were designed to detect rare CNVs from a sample batch without controls. The programs CoNIFER, XHMM, ExomeDepth and CODEX were selected based on the following criteria and resources comparing the four programs to each other or against other programs:

Table 5: Conclusions from comparison studies involving CoNIFER, XHMM, ExomeDepth and/or CODEX.

Program	CNV calling sensitivity/specificity	CNV size*	CNV state (del/dup)
ExomeDepth	High sensitivity, low specificity (Hwang et al., 2015), (Samarakoon et al., 2014), (Tan et al., 2014), (Roca et al., 2019)	Small (Hwang et al., 2015), (Kadalayil et al., 2015), (Tan et al., 2014)	Bias to deletions (Hwang et al., 2015), (Tan et al., 2014)

XHMM	Higher specificity than sensitivity (Yao et al., 2017)	Mostly big (Yao et al., 2017), (Samarakoon et al., 2014)	Balanced (Tan et al., 2014)
CoNIFER	High specificity, low sensitivity (Samarakoon et al., 2014), (de Ligt et al., 2013), (Sadedin et al., 2018)	Only big (Tan et al., 2014), (de Ligt et al., 2013), (Yao et al., 2017), (Samarakoon et al., 2014)	Bias to duplications (Tan et al., 2014)
CODEX	Balanced (Jiang et al., 2015), (Sadedin et al., 2018)	Balanced (Jiang et al., 2015), (Gambin, Akdemir et al., 2017), (Kim, H. Y. et al., 2017)	Bias to deletions (Roca et al., 2019)

*small CNV affects fewer than three exons here, del = deletion, dup = duplication

Some other programs published later during this study were evaluated as well for gene panel sequencing data. CNVkit and SavvyCNV seemed to provide additional complementarity and algorithmic approaches compared to the four programs discussed first. For example, both utilize information from off-target reads (Talevich, E. et al., 2016; Laver et al., 2019). However, the initial four programs remained our choice in the further steps.

4.2.1 Program descriptions and utilized parameters

4.2.1.1 CoNIFER

Copy Number Inference From Exome Reads (CoNIFER) (Krumm et al., 2012) is written entirely in Python. The workflow involves first transformation of reads into per kilobase per million mapped reads (RPKM) for input BAM files (all the other programs start from BAM files as well). This step normalizes the read count for targets against the total read coverage in a sample. This normalization corrects for low sample coverage and enables more linear comparison of samples in a batch. CoNIFER uses singular value decomposition (SVD) to correct for systematic biases and filter common variation to transform RPKMs into standardized Z-scores. Singular values are plotted into a “screeplot” to identify experimental noise and to aid in choosing of the SVD value (an inflection point in the plot) for component removal. The final corrected SVD-ZRPKM values represent normalized copy numbers for each exon in a sample. The program cannot be utilized to detect aneuploidies, but rare variants on chromosome X can be detected. Therefore, males and females can be analysed together in one batch. CoNIFER is designed to identify CNV calls from at least three consecutive exons and requires at least eight samples in a batch to run. In several comparisons, CoNIFER seems to achieve a high specificity, but on the cost of a reduced sensitivity (Table 5). Additionally, the detected CNVs seem to be overly large (Kadalayil et al., 2015; Krumm et al., 2012).

Here, CoNIFER version 0.2.2 was used with default settings for all samples in a batch analysed together, since the program normalizes differences in chromosome X for males and females. The SVD value was chosen for each batch according to the screeplot, as instructed by the developers. A lower SVD value was selected in unclear cases to reach higher sensitivity.

4.2.1.2 XHMM

EXome hidden Markov model (XHMM) (Fromer et al., 2012) is written in C++ and R. First, BAM coverage is calculated with GATK Depth-of-Coverage, and then principal-component analysis (PCA) is used for noise reduction and normalization. In a pre-normalization step, some outliers are filtered out, such as targets with extreme GC content (< 0.1 or > 0.9), and targets with diverging coverage. This step homogenizes the sample set prior to PCA analysis. The developers recommend adjusting the thresholds for filtration in this step according to the study setting. PCA removes high-variance components originating mainly from batch effects and population-related effects rather than being related to read depth change. The default for PCA is derived from a calculation of 0.7/sample amount. The normalized data is then used to train and run a hidden Markov model (HMM) to discover CNVs. The model takes into account the parameters for exome-wide CNV rate and exon target length distribution and distances. The developers recommend at least 50 samples to be analysed in a batch. XHMM reports Phred-scaled detection specific scores with QScore (a global quality threshold) for the whole CNV detections and for each break point. XHMM annotates general CNV state compared to reference without differentiation of homozygous deletions or duplications.

In this study, XHMM version 1.1 was used, and more lenient settings were selected for targeted gene panel sequenced samples compared to default settings. Namely, sequencing data from targeted gene panels surpassed the default thresholds with higher average read depth and greater standard variation between samples. These thresholds for the pre-filtering step were adjusted as follows: maximum target size (maxTargetSize) to 10,000, maximum mean target read depth (maxMeanTargetRD) to 1500, maximum mean sample RD (maxMeanSampleRD) to 1000, maximum standard deviation sample RD (maxSdSampleRD) to 350, and maximum SD target RD (maxSdTargetRD) to 80. The exome-wide CNV rate was also increased from 1e-8 to 1e-3. For WES samples, the default exome-wide CNV rate 1e-8 was used, and parameters for the pre-filtering step were closer to the default values: maxTargetSize 10,000, minMeanTargetRD 10, maxMeanTargetRD 1100, minMeanSampleRD 10, maxMeanSampleRD 1000, maxSdSampleRD 350 and maxSdTargetRD 30. Females and males were analysed separately in each batch as recommended by the program developers. This division by gender was done also for the rest of the programs. For MYOcap, this produced female batches of eight to 33 samples with an average of 17 samples per analysis, and male batches of five to 76 samples with an average of 28 samples per analysis. For MNDcap, female analysis batches were sized from eight to 33 samples with an average of 17 samples, and male batches from nine to 63 samples with an average of 26 samples. For WES, the numbers were for female batches three to 33 samples with an average of 13 samples, and for male batches five to 29 samples with an average of 15 samples.

The normalization methods utilized by CoNIFER and XHMM are based on Gaussian noise structure estimation. These methods, SVD and PCA, seem to effectively eliminate the strongest components from the read depth signal in order to remove noise and reveal rare variation.

However, the methods require large enough sample batches to avoid false negative detections from exclusion of real signals as batch effect (Tan et al., 2014).

4.2.1.3 ExomeDepth

ExomeDepth (Plagnol et al., 2012) is written in R. The developers argue that Poisson and binomial distributions fit WES data inadequately, so instead a beta-binomial model is used for coverage estimation. ExomeDepth builds an internal optimized aggregate control sample set to maximize detection power. Variance can be limited by increasing the number of samples in the reference set, but if the samples included are less correlated, then bias increases. The optimum reference set size was tested to be 10, which is also indicative for the required sample batch size in the analysis. After building the reference set, the program provides a likelihood value for the state of each exon and combines these values with an HMM across multiple exons to provide read count ratio at exon level. ExomeDepth calculates the detection quality score by Bayes Factor, a likelihood ratio for CNV versus normal copy number state. The program has the potential to detect very small deletions of one or two exons. The output contains a ratio for the expected and detected read depth, which allows the detection of copy number changes beyond heterozygous variants. ExomeDepth seems to require high read depth and high correlation between the samples (Kadalayil et al., 2015). It also tends to over-segmentate detections into multiple parts more than other programs (Hwang et al., 2015).

Here, ExomeDepth version 1.1.10 or 1.1.12 was used with default parameters both for targeted gene panel sequenced and WES samples. This is because ExomeDepth seems to achieve high sensitivity and surpass other tools in various settings already with default parameters (Sadedin et al., 2018). The developers recommend that samples with < 0.97 correlation to the reference sample set should be removed from analysis, but we did not follow this, since according to control sample CNV detections some samples with significant results would have been excluded.

4.2.1.4 CODEX

COpy number Detection by EXome sequencing (CODEX) (Jiang et al., 2015) is written in R and uses Poisson log-linear decomposition and Poisson likelihood-based segmentation to model the read depth data. This approach is claimed to resolve especially high variance in read depth between exons better than other models. Preceding read count modelling, the program executes a quality-filtering step for mappability, exon size and minimum coverage. The normalization step with log-linear decomposition removes bias originating from extreme GC content, exon length differences and technical artefacts from capture and amplification. The tool utilizes a Poisson likelihood-based circular binary segmentation algorithm to detect differences in copy numbers. CODEX annotates CNV detections with a simple quality score of lratio: likelihood ratio of CNV versus copy neutral event.

CODEX version 1.10.0 or 1.12.0 was used in this study. For the targeted gene panel sequenced samples, the tool version with integer mode and segmentation by target, and maximum K value

of nine was used. For WES samples, segmentation was done by chromosome, and maximum K value was set to eight. In both settings, the selected coverage threshold of 20 to 4000 differed from the default. For two batches (one male sample set in MYOcap and one female sample set in WES) with the number of samples less than the K value, the K value was decreased (to minimum of five).

The general workflow for all the algorithms is described in Figure 14.

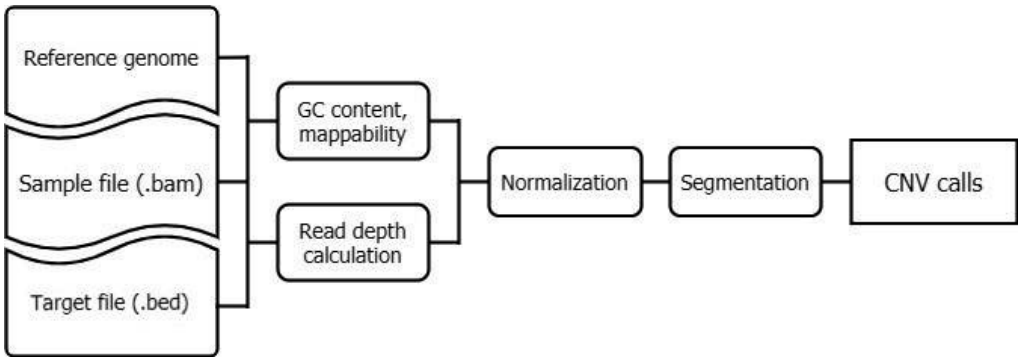

Figure 14: General workflow for CNV detection with the read depth method. Sample file(s) and a file with sequencing targets are needed, and some programs utilize also the reference genome. After read depth calculation and normalization steps, different segmentation algorithms can be used to make the final CNV calls. (*Figure created by Salla Välipakka with diagrams.net.)*

4.2.2 Algorithm for *SMN1/SMN2* differentiation

Due to the clinically significant *SMN1* gene having a highly similar pseudogene, copy number for *SMN1* was estimated with a different and more specific approach for all the MNDcap sequenced samples. Based on the mathematical model in the paper by Feng et al. (Feng et al., 2017), a new algorithm was constructed. For each sample in a batch of samples, the read depth was calculated exactly at the two exonic nucleotides, which differ between *SMN1* and *SMN2*: exon 7 c.840 chr5:70247773 with C for *SMN1* and corresponding chr5:69372353 with T for *SMN2* (the functionally significant nucleotide change as described previously), and exon 8 c.233 chr5:70248501 with G for *SMN1* and chr5:69373081 with A for *SMN2*. All the genomic coordinates here and later adhere to the Human Reference Genome version GRCh37/hg19. The following pipeline was used separately for both c.840 and c.233 calculations.

Samples with *SMN1* to *SMN2* ratio between 0.8 and 1.2 according to the ratio of the c.840/c.233 nucleotide counts from the previous step were grouped together to identify the median sample, which is expected to represent a case with exactly two copies of both *SMN1* and *SMN2*. Then the average read depth for each exon in *SMN1* and *SMN2* was calculated for all the samples. These exon coverages were then normalized with the average coverage from the median samples. For deciding the final *SMN1* and *SMN2* copy numbers for each sample, the following formula was used:

$n1 = rd1/(rd1+rd2)*\Sigma c/\chi c*4$

where rd1 and rd2 are the read depths of the c.840/c.233 nucleotide at *SMN1* and *SMN2* (or *SMN2* and *SMN1* when calculating copy number for *SMN2*), Σc is the combined exon 7/exon 8 coverage of *SMN1* and *SMN2* for the sample, and χc is the median of all the calculated Σc in the analysed samples. Here, copy number results from the c.840 calculation were given priority if in disagreement with c.233. In case of a result with indication of *SMN1* copy number 0, the sample bam file was inspected visually with the Integrative Genetics Viewer (IGV, Broad Institute, Cambridge, MA) (Thorvaldsdottir et al., 2013). In ambiguous cases, a pseudoreference approach was attempted with concealing of *SMN2* from the reference genome to force alignment only into the *SMN1* locus.

4.2.3 Algorithm for *NEB* triplicate region differentiation

A similar approach as above for *SMN1/SMN2* genes was attempted for copy number calculation in the triplicate repeat region of *NEB*. *NEB* has eight repeating exons in three units (TRI): TRI1 ex 82–89, TRI2 ex 90–97, and TRI3 ex 98–105 (Kiiski, K. et al., 2016). The copy number for each triplicate unit was estimated based on single nucleotide positions with differing nucleotide in the corresponding exon in other repeat units (Table 6). None of the nucleotide positions differ between all the triplicate units, but some nucleotides are unique for a repeat unit, thus enabling their differentiation.

Table 6: Genomic locations and changes for nucleotides with differences between *NEB* triplicate units.

TRI unit	Genomic location	Change	Genomic location	Change	Genomic location	Change
TRI1	chr2:152463200	A>G	chr2:152460241	T>C		
TRI2	chr2:152448640	G>A	chr2:152448563	C>T	chr2:152447862	C>T
TRI3	chr2:152435919	A>G				

TRI = Triplicate

4.2.4 CNV analysis from mitochondrial DNA

CNV analysis was attempted from mitochondrial DNA (mtDNA) with the four described CNV analysis programs, and CNVkit version 0.8.5 (Talevich, E. et al., 2016). CNVkit has a mode originally designed for tumor samples for calling CNVs from samples with contamination from a genetically differing source. In a training batch of mtDNA sequenced samples, two samples had a known single large deletion, sized 2 kb and 7.5 kb, with a heteroplasmy rate of 40%. Three samples had multiple smaller deletions. The first four tools were utilized with settings as presented, and CODEX was additionally utilized with the “fraction” mode designed for detection of somatic CNVs in cancer from heterogeneous samples. CNVkit was utilized with the mode for tumor samples, and different rates for expected tumor/normal contamination were given as input parameter. The tumor sample modes with CNVkit and CODEX were expected to compensate for sample heteroplasmy bias with mtDNA.

4.3 CNV control samples

4.3.1 Control samples for the targeted gene panels

Our CNV analysis method was validated (for other than *SMN1/SMN2* genes) with a heterogeneous sample set with known and previously characterized CNVs in 24 samples (Table 7). The samples were received from other research groups in collaboration. The control samples were sequenced together with patient samples in the MYOcap batches and analysed with the same sequencing data pre-analysis and CNV analysis pipelines.

Table 7: Known and previously characterized CNVs in positive control samples. The CNV is presented as previously published, and the publication is listed if available.

CNV type	Gene	Targets	CNV as previously published	Published previously in
Het del	*CAPN3*	ex 2–8	DEL het chr15 5' breakpoint boundary 40439563–40463933 3' breakpoint boundary 40472217–40473795	(Piluso et al., 2011)
	CSRP3	ex 4–7	het del: exons 4–7 and 3'UTR	(Giugliano et al., 2018)
	DMD	ex 44	-	-
	DMD	ex 47–52	-	-
	LAMA2	ex 13–14	del: exons 13–14	(Giugliano et al., 2018)
	LAMA2	ex 13–37	DEL Het chr6 5' breakpoint boundary 129555720–129612876 3' breakpoint boundary 129756115–129764002	(Piluso et al., 2011)
	MTM1	large deletion with *MTM1*	min deletion region: chrX: 149591931–149841591; max deletion region: chrX: 149526823–149844072	(Savarese et al., 2016)
	MYPN	ex 4–7	het del: *MYPN* exons 3–5	(Giugliano et al., 2018)
	NEB	ex 14–81	chr2:g.(152454645_152456955)_(152554712_152561404)del (GRCh37)	(Kiiski, K. J. et al., 2019)
	NEB	ex 43–45	del ex 43–45	(Lehtokari et al., 2014)
	SGCD	ex 1	het del: first coding exon	(Giugliano et al., 2018)
	SGCG	ex 7	del: exon 7	(Giugliano et al., 2018)
	SGCG	large deletion with *SGCG*	DEL Het chr13 5' breakpoint boundary 21593561–22464962 3' breakpoint boundary 23788143–23962786	(Piluso et al., 2011)
Hom del	*SGCB*	ex 6	hom del: last 12 codons in exon 6 and 3' UTR (U)	(Giugliano et al., 2018)
Hemiz del	*DMD*	ex 5–7	-	-
	DMD	ex 45–47	-	-
	DMD	ex 45–49	hem del: exons 45–49 (U)	(Giugliano et al., 2018)

Het dup	*DMD*	ex 45–49	-	-
	DMD	ex 2-7	-	-
	DMD	ex 1	-	-
	LAMA2	ex 5–12	dup: exons 5–12	(Giugliano et al., 2018)
	LAMA2	ex 21–55	dup: exons 21–55	(Giugliano et al., 2018)
	NEB	ex 72–81	-	-
Hom dup	*NEB*	ex 72–81	-	-

del = deletion, het = heterozygous, hom = homozygous, hem = hemizygous, dup = duplication

4.3.2 Initial CNV detections and CNV verifications

Batches with the CNV control samples were analysed to evaluate the initial CNV detection sensitivity and performance of the programs with this setting. Additional interesting or potentially clinically significant CNV findings (according to patient phenotype and familial segregation) were verified. The CNV verifications and segregation studies were performed when possible with suitable methods including MLPA (10 samples), array CGH (14 samples), and PCR (14 samples, followed by successful Sanger sequencing for solving exact breakpoints for four of these samples). MLPA verifications and array CGH were performed elsewhere by collaborators. CNV detections from these verifications, both true positive and false positive detections, were included in the development of a logistic regression model both as target regions for *in silico* CNVs, and as validation detections for model validation.

For PCR, primers were designed using Primer3 v4.0.0 (primer3.ut.ee) (Koressaar and Remm, 2007; Untergasser et al., 2012) (Supplemental Table 3 for primer designs). The primers were either designed into first exons expected to be deleted and last exons expected to be retained to reveal hemi- and homozygous deletions on exon-level, or to produce shorter amplification products with heterozygous deletion compared to normal samples. PCR was performed using DreamTaq DNA Polymerase (Thermo Fisher Scientific, Waltham, MA) and a touchdown PCR program. The program included an initial denaturing step at +95 °C for 5 min, followed by four cycles of three repetitions of denaturation at +95 °C for 30 s, annealing at +67 °C (decreased by 3 °C after each three repetitions) for 30 s, and elongation at +72 °C for 1 min, so 12 cycles in total. The remaining 25 cycles had an annealing temperature of +55 °C with denaturation and elongation as above, and lastly 10 min of extra elongation at +72 °C. The products were run and purified from agarose gel, and Sanger sequencing was attempted to capture CNV breakpoints accurately.

Array verifications were performed with custom-made array CGHs, either on a 8×60k NM-CGH array (Kiiski, K. et al., 2013) with eight genes (one sample), or on a 4×180k array (Sagath et al., 2018) with 87 genes (13 samples), which were also included in MYOcap. The Cytosure Software v.4.6.85 (Hg19) (Oxford Gene Technology Ltd) was used for the graphic analysis and visual inspection of the data, and a minimum of five probes was used as a threshold to make a

call (Sagath et al., 2018).

The kit used for MLPA was SALSA MLPA P034 and P035 DMD kit (MRC-Holland, Amsterdam, NL). The analysis was performed according to the manufacturer's protocol (Giugliano et al., 2018).

4.4 Improving CNV detection accuracy: predictive model

Because low specificity was perceived in the CNV detection results, a logistic regression model was developed to filter CNV detection results according to true positive prediction. As discussed in the introduction, *in silico* generated CNVs have been utilized in previous approaches to validate methods and tools. Only deletions have been usually simulated since they are more simple to generate, and the simulation of duplications and inversions with synthesized reads is not thought to comparably well correspond to real genome data (Ellingford et al., 2017; Kosugi et al., 2019; Sadedin et al., 2018). Fewer tools are available for generating simulated reads representing WES data as opposed to WGS data, and also to generate CNVs into the data rather than SNVs or indels (Roca et al., 2019). Furthermore, it has been questioned whether simulated reads resemble real sequencing data enough, especially with specific sequencing designs, such as targeted gene panels (Ellingford et al., 2017; Kadalayil et al., 2015). Therefore, we took a “keep-it-as-real-as-possible” approach for generating simulated CNVs since the amount of real CNV detections at our disposal was not enough for training a statistic model.

Fifteen male and female samples without significant number of CNV detections were selected from the MYOcap v.4 gene panel sequenced samples. These samples were re-analysed for CNVs in separate batches (females and males separately) to gain a list of original CNV detections. Since small CNVs are the most challenging to detect, most of the generated *in silico* CNVs were one-exon or two-to-four exons in size (Marchuk et al., 2018). Both deletions and duplications were generated, but without synthetized reads.

4.4.1 Targets for *in silico* CNVs

GRCh37 (hg19) known canonical transcripts for UCSC genes (University of California, Santa Cruz, http://genome.ucsc.edu and http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database, last accessed June 28, 2017) were intersected with the MYOcap v.4 target set using BEDtools intersect (version 2.26.0). The output file included 5219 exons in 297 genes.

A set of random target regions, 3900 for one-exon CNVs and 5400 for two-to-four-exon CNVs, was generated with exact exon borders from the intersect target file. The same target sets were used both for deletions and duplications. Single-exon targets were at least two exons apart, and two-to-four-exon targets were at least five exons apart with otherwise random distribution except for chromosome X, which was included only for females to generate only heterozygous CNVs. The longer target regions could span genes, which segmented some of them into one-

exon target regions. The CNVs were distributed five per sample involving only one target set and either deletions or duplications, generating four types of sample sets (with single-exon or two-to-four-exon deletions or duplications). The generation of multiple *in silico* CNVs into a sample as opposed to one has been recognized as a sensible and computationally more efficient approach, since even in targeted gene panel sequenced samples detecting more than one rare CNV is not unusual (Sadedin et al., 2018; Kerkhof et al., 2017). The fifth sample set included heterozygous *in silico* CNVs (one per sample) generated according to the coordinates of the real CNVs presented in the control sample CNV table, or some of the verified CNVs. This set included 114 deletions and 20 duplications of different lengths in different genes.

4.4.2 *In silico* CNV generation workflow

For the generation of deletions, reads from the region intended to be deleted were separated from the original BAM file with BEDtools intersect into a new separate BAM file. Another BAM file was generated with the reads from this region removed. The read count in the first BAM file was decreased by 50% using the Picard tool DownsampleSam (version 2.13.2 or 2.18.10). The downsampled reads were combined with the second BAM file (with the downsampled region "emptied" from reads) using the Picard tool MergeSamFiles. For duplications, two BAM files were paired according to similar average read depth. From the donator BAM file, reads were extracted with BEDtools intersect from the region intended to be duplicated. The read count was then decreased with the Picard tool DownsampleSam in a way that an amount corresponding to a 50% addition of reads to the receiver file was preserved. The downsampled reads were then merged with the Picard tool MergeSamFiles with the receiver BAM file adding reads to the target region. The duplication of reads within a single BAM file would have resulted in exclusion of these reads as PCR duplicates in further analysis. The workflow was common for both deletions and duplications from this point, with read groups for the merged files replaced with the Picard tool AddOrReplaceReadGroups to remove notions from two separate original file names. Finally, all the generated BAM files were sorted and indexed with SAMtools (version 1.4) (illustrated outline for *in silico* CNV generation: Manuscript II Figure 1).

4.4.3 Analysis of *in silico* CNV detection sensitivity

CNV detection results from the four programs were converted into BED format and intersected with BEDtools intersect (as in the following steps) with the *in silico* CNV target regions to separate unspecific detections. The detections matching the original CNVs in the samples were removed to reveal the true false positive (FP) detections. CNV detection sensitivity for the *in silico* CNV targets was inspected on exon level with different overlap thresholds, from a minimum of 1 bp overlap up to 99% nonreciprocal overlap. With 99% nonreciprocal overlap, an exon in the CNV had to be covered by the detected CNV with at least 99% overlap on a base pair level.

Before sensitivity evaluation, both the specific and unspecific CNV detections by all four

programs were intersected with all program combinations. Since BEDtools intersect lists overlaps for only one pair at a time, an additional script was written to combine these into a single CNV detection depicting each CNV detected by one to four programs. The overlap requirement was set to 1 bp, rather than commonly used 50%. This was to avoid excluding calls owing to detection inaccuracies: the programs have different tendencies to over-segment CNV detections, and they provide sometimes differing size estimations for CNVs. Some CNV detections consisted of multiple parts due to different matches in over-segmented detections, and these separate matches were deciphered as "CNV detection units". The intersected four-way CNV detections were used in the sensitivity evaluation.

The effect of CNV mosaicism on detection sensitivity was also tested with additional 150 one-exon and 150 two-to-four-exon *in silico* deletions and duplications. These were distributed five per sample into the 30 test samples largely with the previously described procedure. However, the percentage was adjusted with the Picard tool DownsampleSam to cut reads for deletions with percentages of 50%, 40%, 30%, 20%, and 10% to represent a pure heterozygous deletion for comparison and different degrees of mosaicism. Similarly, reads for duplications were increased with percentages of 10% to 50%, with the end of the spectrum representing a pure heterozygous duplication. The detection sensitivity was calculated on exon level with a minimum overlap requirement of 1 bp.

4.4.4 Logistic regression model: training and validation

The specific and unspecific *in silico* CNV detections with 1 bp overlap requirement were combined and converted into comma-separated values (.csv). The minimal overlap category was selected because it was estimated to represent CNV detections and detection evaluation from true samples most closely. The CNV detection units were distributed randomly into five training and test sets for cross-validation, with 80% of the variants in each training set and 20% in each testing set. The effects of various CNV detection features were evaluated for each CNV detection unit in different combinations as variables in a logistic regression model for differentiating true positive (TP) detections from FP detections. CoNIFER detections were converted into a binary format (1 = detected, 0 = not detected, feature 1) and for the other three programs, both a CNV detection-specific score (features 2 to 4) and a median score in the in-house CNV detection database from MYOcap samples (at the time with CNVs from 1956 samples) were included (features 5 to 7). The number of targets (exons) and detection length in base pairs were included both in the prioritized detection order of CODEX > ExomeDepth > XHMM > CoNIFER (in accordance to previous estimations for the program accuracies in breakpoint detection) and as a mean from all program detections (features 8 to 9 for targets and 10 to 11 for length). State (deletion or duplication) was included as feature 12.

Firstly, 15 model versions with different combinations of CNV specific scores (features 1–4) were tested. Then, additional features 8 to 12 were added as variables to the best models from the first stage and tested in five model versions. Lastly, seven model versions were tested with CNV detection specific scores switched to in-house median CNV scores (features 5 to 7).

Multicollinearity between variables was avoided with calculation of the variance inflation factor, a measure of inflation in variance of the estimated regression coefficient attributable to correlation among variables in a model. Variance inflation factor was required to be less than two for every variable in any model version for the model to be valid (Lin, 2008).

The model performances were evaluated with the area under the receiver operating characteristic curve (AUC), calculated with a method for cross-validated estimates (LeDell et al., 2015). The best models with AUCs > 0.90 were validated with real CNV detections, consisting of the control sample CNVs and newly validated CNVs. This was 66 positive control samples and 8 samples with false positive CNV detection in total. Each sample had only one CNV, but some were divided into multiple CNV detection due to over-segmentation and thus more CNV detection combinations between the different programs. Therefore, prediction results were evaluated in two different categories: 66 CNVs matching the number of samples and 74 CNV detections with eight additional CNV detection units from CNVs on five samples. True positive rate [TP/(TP + FN)] (or sensitivity), true negative rate (or specificity) [TN/(TN + FP)] and overall accuracy [(TP + TN)/(TP + TN + FN +FP)] were calculated for the validation tests, as has been done in previous studies (Zhang, L. et al., 2019). The threshold for TP/FP status prediction was set separately for each different model version to maximize accuracy in each setting. 95% CIs were calculated for overall accuracy measurements with the exact Clopper-Pearson method. All statistical calculations were performed with R version 3.4.3 (The R Foundation).

4.4.5 Control samples for WES validation

The CNV detection method and predictive model were validated with WES samples with known CNVs. Samples received from Blueprint Genetics for this purpose had been targeted and enriched with either “xgen exome research panel probes with custom additions” targeting the whole exome (designated as “WES”, with three batches and 83 samples altogether) or “mendelome probes” targeted for 6399 clinically significant genes (designated as “mendelome”, with three batches and 130 samples altogether). The samples were sequenced to 150 bp paired-end read depth with Illumina Novaseq with acquired average read depth of 220X for WES batches and 190X for mendelome batches. The sample set included 27 samples from Coriell Institute CNVPANEL01 set with microdeletions and/or microduplications, more than one in some of the samples. These samples were included in one of the WES batches with 24 heterozygous microdeletions, 11 heterozygous microduplications and four homozygous microduplications. The average size for the 24 deletions was 11.3 Mb ± 6.5 Mb (95% CI), and for the 15 duplications 17.2 Mb ± 13.9 Mb. 24 of the Coriell samples were also distributed into the three mendelome batches, including 21 heterozygous deletions, 10 heterozygous duplications and three homozygous duplications. For the 21 deletions, the average size was 9.6 Mb ± 5.7 Mb and for the 13 duplications 16.2 Mb ± 15.7 Mb.

CNVs for the other validation samples (106 in mendelome batches and 56 in WES batches) had been verified previously at Blueprint Genetics with CNVkit or an in-house developed CNV

detection algorithm (unpublished). The sample CNVs have not been published accurately but are illustrated with relevant properties in Table 8. The accuracy for the CNV detections was evaluated on exon level for the listed CNVs, since for most of them the expected CNV was provided with exon-level accuracy.

Table 8: Sizes and states of control CNVs in Blueprint Genetics samples, without Coriell Institute samples.

Mendelome	State	one exon	2–4 exons	5–10 exons	> 10 exons / whole gene	multiple genes
Deletion	Homozygous	5	5			
	Heterozygous	20	31	11	8	7
Duplication	Homozygous			1		
	Heterozygous	1	1	4	6	6
WES						
Deletion	Homozygous	5	3	1	1	
	Heterozygous	21	19	2		
Duplication	Homozygous			1		
	Heterozygous	1	1			1

4.6 CNV annotation

For CNV annotation, the program cnvScan was used as a base-line tool (Samarakoon et al., 2016). The program provides annotations in four general categories: CNV information, such as size, state and genes included, functional information based on the location of the CNV (such as on regions of segmental duplications) and genes included in the CNV (such as haploinsufficiency score), population frequency according to CNV databases, and disease-gene association information. However, since the program was published already in 2016 and has not received updates, it lacks updated databases and some databases completely. Generally, including multiple common CNV population databases is recommended since if a variant is found in multiple subjects in different studies, it is probably more true and not a study-specific repetitive artefact (Zarrei et al., 2015). Moreover, the CNVs were originally compared to variants in databases only with 100% overlap requirement. Therefore, we updated cnvScan (Table 9) by adding more databases (Table 10 for references and web sources) and three different overlap degrees with the bedtools intersect: 1 bp, 50% reciprocal and 90% reciprocal requirements. The prediction of true positive/false positive from the logistic regression model was added as well.

The CNV population frequency database annotations were modified to contain information in the format of population frequency, if information for studied populations and size of studied cohorts was included. The databases for ExAC and DECIPHER CNVs were not originally in this format, as compared to the formats offered by default by gnomAD and 1000g CNV databases. The DGV CNV databases could not be updated with this information. All databases provided deletions and duplications separately, and database matches were evaluated separately

for these for each CNV, since a duplication and a deletion in the same location could have different effects and frequencies (Riggs et al., 2019). In-house CNV detection frequency evaluation was included to allow tracking of detected variants and to provide consistent variant annotations, as has been recommended for variant reporting (Li, M. M. et al., 2017). In-house variant databases are also estimated to reveal especially false positive calls originating from technical bias (Li, M. M. et al., 2017). The in-house CNV databases were built separately for each of the sequencing data types (MYOcap, MNDcap, WES). The in-house CNV databases contained only CNVs evaluated to be true with the model.

Table 9: Updated and newly included databases in cnvScan, and the criteria used for database comparison.

Data	Original database	Updated database	Match criteria
UCSC exons	-	UCSC_exons_modif_canonical_withchr.bed	1 bp
in-house CNV database	-	detected variants per program, BED	three-tier
refGene annotation	-	refGene.txt.gz 25-Nov-2018	1 bp
GENCODE annotation	havana_or_ensembl_gencode.v19.annotation.gtf	-	gene
UCSC segmental duplications	-	genomicSuperDups	two-tier
PhastCon elements	phastConsElements100way	-	gene
Haploinsufficiency index	Dataset_S2	HI_Predictions_Version3	gene
Residual variation intolerance score	SCORES_n12_4NR_v16May15	RVIS_Unpublished_ExACv2_March2017	gene
gnomAD gene LOF intolerance	-	gnomad.v2.1.1.lof_metrics.by_gene	gene
ExAC gene scores for CNV intolerance	-	release0.3.1-cnv-exac-final-cnv.gene.scores071316filtered.bed	gene
UCSC CpG islands	-	cpgIslandExtUnmasked	two-tier
UCSC conserved TF binding sites	-	tfbsConsSites.txt	two-tier
UCSC Conserved mammalian miRNA target sites	-	targetScanS	two-tier
Sanger high resolution CNVs	conrad.et.al.2010_Validated_CNVEs_v5_4Release	removed, included in DGV	
DGV CNVs	GRCh37_hg19_variants_2014-10-16	GRCh37_hg19_variants_2016-05-15	three-tier
curated high quality DGV Inclusive map	s9	Inclusive.Gain+Loss.hg19.2015-02-03	three-tier
curated high quality	s10	Stringent.Gain+Loss.hg19.2	three-tier

DGV Stringent map		015-02-03	
ExAC CNVs by population	-	exac-final.autosome-1pct-sq60-qc-prot-coding.cnvDUPDEL.bed	three-tier
gnomAD CNVs by population	-	gnomad_v2_sv.sites.bed	three-tier
DECIPHER normal population CNVs	-	population_cnv.txt	three-tier
1000 Genomes CNVs by population	union.2010_06.deletions.sites	ALL.wgs.mergedSV.v8.20130502.svs.genotypes	three-tier
OMIM morbid map	morbidmap_formatted_only HGNC.txt	morbidmap.txt 07112017	gene
DECIPHER developmental disorder CNVs	cnvScan_DDG2P_freeze_with_gencode19_genomic_coordinates_20141118	cnvScan_DDG2P_freeze_with_gencode19_genomic_coordinates_20171107.txt	gene
ClinVar HGVS variants	clinvar_20150106	clinvar_20170905	gene
DisGeNET gene-disease annotations	-	all_gene_disease_associations.tsv	gene

With three-tier comparison 1bp, 50% reciprocal and 90% reciprocal overlap requirements were used, and with two-tier only 50% and 90% reciprocal overlap requirements.

Table 10: References and web sources for the data utilized in cnvScan, as of 03/2020.

Data	Reference	Web source (as of 03/2020)
UCSC exons	(Church et al., 2011)	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/
refGene annotation	(Church et al., 2011)	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/
GENCODE	(Harrow et al., 2012)	https://www.gencodegenes.org/human/releases.html
UCSC segmental duplications	(Church et al., 2011)	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/
PhastCon elements	(Siepel et al., 2005)	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/phastCons100way/
Haploinsufficiency index	(Firth et al., 2009)	https://decipher.sanger.ac.uk/about#downloads/data
Residual variation intolerance score	(Petrovski et al., 2013)	http://genic-intolerance.org
gnomAD gene LOF intolerance	(Karczewski et al., 2020)	https://gnomad.broadinstitute.org/downloads
ExAC gene scores for CNV intolerance	(Lek et al., 2016)	https://console.cloud.google.com/storage/browser/gnomad-public/legacy/exacv1_downloads/release0.3.1/cnv/
UCSC CpG islands	(Church et al., 2011)	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/

UCSC conserved TF binding sites	(Church et al., 2011)	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/
UCSC Conserved mammalian miRNA target sites	(Church et al., 2011)	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/
DGV CNVs	(MacDonald et al., 2014)	http://dgv.tcag.ca/dgv/app/downloads?ref=GRCh37/hg19
curated high quality DGV Inclusive map	(Zarrei et al., 2015)	http://dgv.tcag.ca/dgv/app/downloads?ref=GRCh37/hg19
curated high quality DGV Stringent map	(Zarrei et al., 2015)	http://dgv.tcag.ca/dgv/app/downloads?ref=GRCh37/hg19
ExAC CNVs by population	(Lek et al., 2016)	https://console.cloud.google.com/storage/browser/gnomad-public/legacy/exacv1_downloads/release0.3.1/cnv/
gnomAD CNVs by population	(Collins et al., 2020)	https://gnomad.broadinstitute.org/downloads
DECIPHER normal population CNVs	(Firth et al., 2009)	https://decipher.sanger.ac.uk/about#downloads/data
1000 Genomes CNVs	(1000 Genomes Project Consortium et al., 2015)	http://www.internationalgenome.org/phase-3-structural-variant-dataset/
OMIM morbid map	(Amberger et al., 2015)	https://www.omim.org/downloads/
DECIPHER developmental disorder CNVs	(Firth et al., 2009)	https://decipher.sanger.ac.uk/about#downloads/data
ClinVar HGVS variants	(Landrum et al., 2014)	https://www.ncbi.nlm.nih.gov/variation/docs/ClinVar_vcf_files/
DisGeNET gene-disease annotations	***https://doi.org/10.1093/nar/gkw943	http://www.disgenet.org/downloads

*** Disclaimer by DisGeNET: "DisGeNET is a derivative database that integrates gene-disease associations from several public expert curated data sources and text-mining derived associations. We would like to acknowledge all the data sources from where the data are derived: https://doi.org/10.1093/nar/gkw943"

The CNV detections were filtered into rare detections for further inspection according to frequency both in the in-house CNV database and in the population CNV databases. For the population databases, a recommended cutoff of 1% (Richards et al., 2015) was utilized in the 90% reciprocal overlap category, taking into account match in CNV state. In the DECIPHER CNV database, frequencies were only considered in large enough datasets with at least 100 subjects studied. In the in-house CNV database, maximum allowed frequency in the 90% overlap category was 1%, and 5% in the 50% reciprocal overlap category for previously detected CNVs of same state. CNV was labeled unique, if it was not detected in any other sample in the in-house database in any frequency, and not detected with 90% overlap in population databases.

For the filtered CNV detections in WES samples, OMIM, DECIPHER, and DisGeNET annotations for gene-disease connections were evaluated to prioritize genes according to putative disease-causing potential for neuromuscular disorders. Additionally, the WES CNV detections were annotated with gene-disease connections from the 2019 version of the Gene Table of Neuromuscular Disorders (http://www.musclegenetable.fr) to reveal possible CNVs in known disease genes. Genes expressed in skeletal muscle and genes with paralogs associated with suitable phenotypes were also prioritized. Incidental findings were not evaluated. If no interesting CNVs were found among the true predicted CNVs in cases with affected family members sequenced, then also the false predicted detections were evaluated for common CNVs. Sporadic IBM (sIBM) samples sequenced with WES were evaluated separately: the frequency of CNVs detected in sIBM patients was compared with findings in other WES samples. Differences in target set designs in different WES batches were taken into account.

For initially evaluating clinical significance of the filtered and rare CNVs, the most recent ACMG recommendations for CNV significance evaluation were utilized (Riggs et al., 2019; Abou Tayoun et al., 2018). Information from The ClinGen Dosage Sensitivity map catalog (Riggs et al., 2012) for evidence supporting or refuting dosage sensitivity of genes was accepted only with sufficient evidence with a score of 3 as recommended. Intragenic deletions and duplications were evaluated within established haploinsufficient genes according to separate PVS1 rules (Abou Tayoun et al., 2018). ExAC pLI score and DECIPERH HI index (included in cnvScan) were used with the suggested thresholds of simultaneous pLI > 0.9 and gnomAD upper bound of confidence interval < 0.35, and DECIPHER HI index of < 10% to obtain a positive HI predictor score (Riggs et al., 2019). The results from this evaluation workflow were used as additional information but not the only decision supporting source, since the ClinGen Dosage Sensitivity map catalog is still under construction. Additionally, the workflow is unsuitable for genes with mainly recessive disease mechanisms and for disorders with wide phenotypic and genetic heterogeneity, and the recommendations have not been widely validated in practice. Therefore, the underlying pathogenicity of the detected CNVs was substantiated by type of mutation, mode of inheritance and matching phenotype. The sequencing data was evaluated for SNVs and indels as well to reveal cases of possible compound heterozygosity.

For clinically significant CNVs detected from MPS data, HGVS nomenclature with gene name, transcript and exons involved are presented in the results, as recommended (Riggs et al., 2019), but for many the breakpoints could not be solved accurately within the scope of this study. The longest known transcript was selected for variant presentations as suggested (Richards et al., 2015). Variants were reported according to (possible) pathogenicity whether they cause disease or not, such as in carriers (Richards et al., 2015). Location information for the CNVs (regions of segmental duplications, TF binding site etc.) were used only as additional information and not as ranking criteria, since they were not included in the newest recommendations with notable weight for decision-making. For CNVs in the gene *DMD*, The UMD/TREAT-NMD DMD database (http://umd.be/TREAT_DMD/ (Bladen et al., 2015)) was utilized to verify patient phenotype match (DMD versus BMD) with the expected read-frame effect of the CNV.

4.7 Effects of read depth and uniformity on CNV detection

Since the WES batches differed more on technical implementations than MYOcap and MNDcap batches, effects of read depth and read uniformity on CNV detection statistics were evaluated. The first WES batch (BGI HiSeq 1500) was excluded due to failed ExomeDepth analysis and thus non-comparable CNV prediction results. The inspected metrics included number of detected CNVs, false positive and true positive predicted CNVs, rare and unique CNVs on average per sample in each batch, and standard deviations and variances for these. The calculations were normalized with the number of targeted bases in each target set. Read depths for each base in each sample were calculated with CoverView (version 1.4.4) (Munz et al., 2018). Coverage uniformity metric (UE) was calculated with ExomeCQA (version not specified) (Wang, Q. et al., 2017) using CoverView calculations as input. The UE calculation is based on the number, height and width of coverage peaks in target regions. UE of 5 was considered as a threshold for low uniformity, and target coverage of less than 0.2 as described by ExomeCQA was considered low coverage. Correlation between the above mentioned CNV statistics and percentage of targets with low coverage, low uniformity, targets with either, targets with both, and the measure of percentage of bases covered by at least 20X in a batch was calculated using Pearson's correlation coefficient. All statistical calculations were performed with R version 3.4.3 (The R Foundation).

5 RESULTS

5.1 Program performances

The four programs detected CNVs with similar size and state (deletion or duplication) distribution from the samples sequenced with the targeted gene panels or WES (Figure 15 sizes and 16 states).

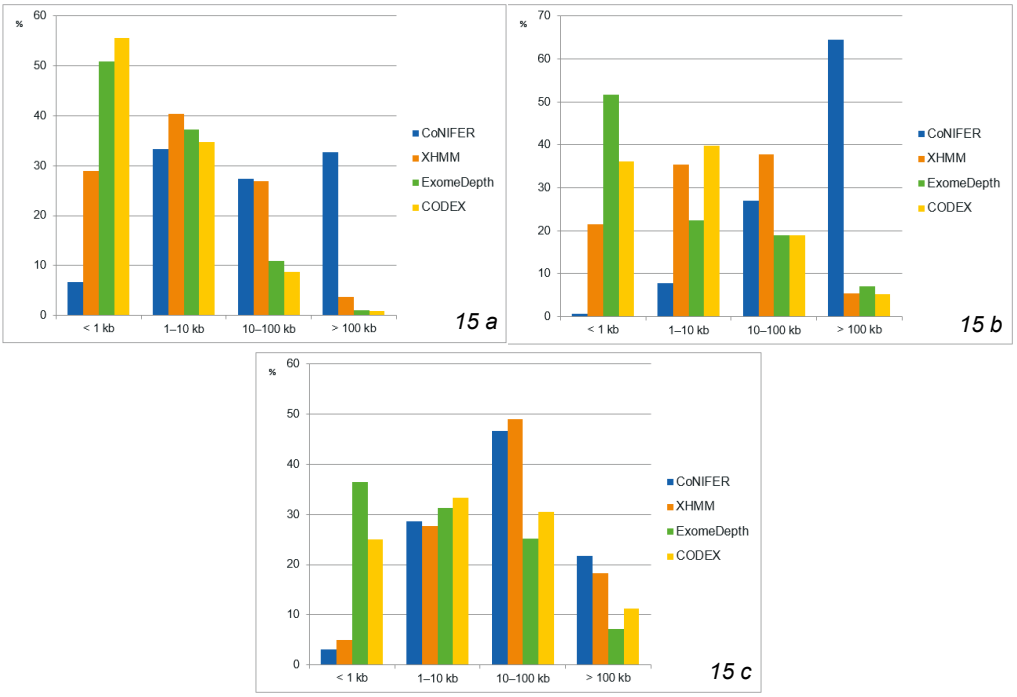

Figure 15: Size distributions of the CNVs detected from MYOcap (15a), MNDcap (15b) and WES (15c) samples by the four programs by percentage of detections in each size category. CoNIFER detects larger CNVs (> 100 kb), and ExomeDepth mostly smaller (< 1kb), with XHMM and CODEX having a more balanced size distribution. *(All the figures in the Results are created by Salla Välipakka with Microsoft Office Excel unless otherwise stated.)*

On average, CoNIFER detected from the targeted gene panel sequenced samples mostly larger CNVs of over 100 kb in size (30–65% of detections), while ExomeDepth detected mostly smaller CNVs of < 1 kb (50%). From WES samples, all the programs had a more balanced CNV detection size distribution, with a cumulative 55–75% of detections for all programs from variants sized 1–10 kb and 10–100 kb. With closer inspection of one MNDcap and a MYOcap batch, approximately 2% of CNV detections were over-segmented into multiple CNV detection units in both. Over-segmentation into smaller parts was displayed the most by ExomeDepth (in 66% of the over-segmented detections), then CODEX (in 44%), and to some extent by XHMM (in 7%).

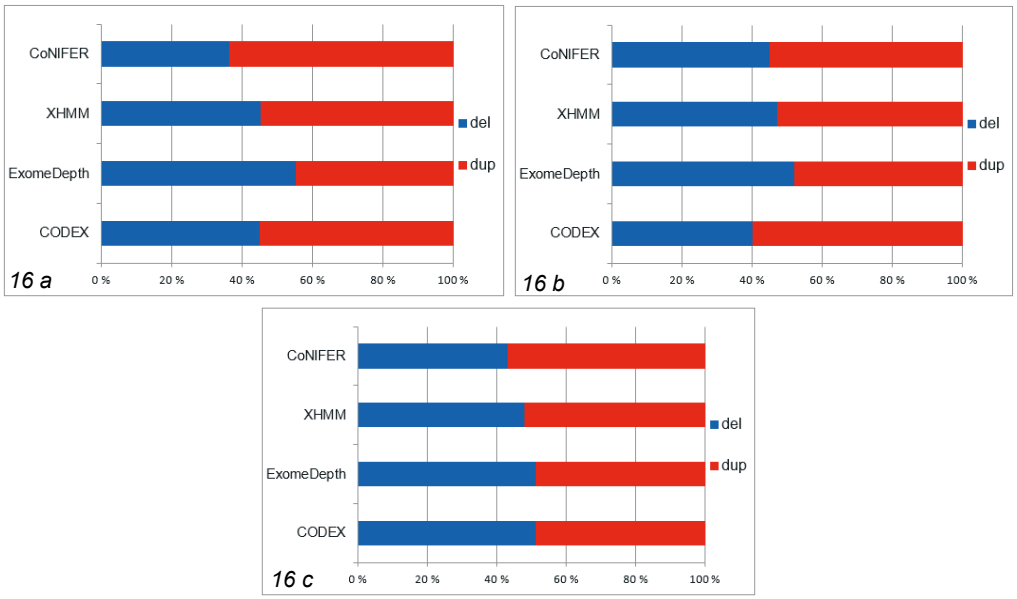

Figure 16: State distributions of CNVs detected from MYOcap (16a), MNDcap (16b) and WES (16c) samples by the four programs. ExomeDepth detected more deletions, while CoNIFER, XHMM and CODEX detected more duplications, and ExomeDepth and XHMM had the most balanced distributions. Del = deletion, dup = duplication.

The state distributions were more uniform across the different sequencing data sets. ExomeDepth detected more deletions (approximately 53% of all detections) and the other three programs detected more duplications: CoNIFER 58%, XHMM 53% and CODEX 54% of all detections. XHMM and ExomeDepth had the most balanced distribution of deletions and duplications detected.

5.1.1 Other CNV detection programs

The other tested programs, CNVkit and SavvyCNV, were used first with default settings and then with the recommended settings for targeted gene panel sequencing data. Either way, they failed to detect the few positive control sample CNVs which were tested and clearly detected by the other four programs. Therefore, these programs were not utilized further, and the initial four programs remained our choice due to their reliability, as discussed in the introduction. Most of the programs received updates from their developers during this study, and allowed for adjusting of parameters, thus providing advantageous flexibility.

5.1.2 Detection of CNVs in positive control samples

All the positive control sample CNVs were detected, and a total sensitivity of 100% was reached (Table 11). Most of the CNVs from the control samples were detected by more than one program (N1 = 3, N2 = 3 N3 = 14, N4 = 4). 17 out of 24 control sample CNVs (71%) were detected exactly as detected before on exon level (which was the accuracy of the known CNV for most of the samples), and an additional four were detected with one or two exon divergence.

In three cases, the divergence was larger. However, the initial information for two of these control sample CNVs, for two wide deletions involving *SGCG* and *MTM1*, was inaccurate and/or the CNV covered more than the gene included in MYOcap.

For regions detected by more than one program, the single detection region was decided for these and other samples by utilizing the breakpoints with highest evidence. Regions detected by CoNIFER were not considered due to the program's known tendency to exaggerate CNV sizes. In the simplest cases, all of the other three programs agreed on the region (usually with a slight 1 bp difference in breakpoints). This was the case for seven of the positive control sample CNV detections. Alternatively, the region similarly detected by two programs was used, either when only two programs had detected the region or if only two programs agreed on the region. The pair of CODEX and ExomeDepth agreeing on the region was the basis for the detected region for six of the 24 positive control samples. If all detected regions were completely different, then the detection by CODEX was prioritized (in three of the samples). This was based on previous observations that CODEX has less tendency to over-segmentate or overestimate the detected CNV compared to the other programs here. If some breakpoints appeared in more than one detection while no whole regions matched, then the breakpoints with the most evidence were utilized. In practice, in a case of detections from XHMM, ExomeDepth and CODEX with breakpoints of chrX:32841412–32841504, chrX:32827610–32841504 and chrX:32827609–32862977, the most confident region for the detection was estimated to be chrX:32827610–32841504. This approach was also used to combine CNV detections into one region if over-segmented. Some common approaches such as using average detected region or the minimal common region were thus not used. Coordinates were selected only if they were concretely among detections.

Table 11: Detected positive control CNVs, and differences to the original detection.

CNV type	Gene	Orig. detection	NP	Detected region	Detected exons	Diff.	Transcript
Het del	*CAPN3*	ex 2–8	4	chr15:42676666–42686546	ex 2–8		NM_000070
	CSRP3	ex 4–7	3	chr11:19203577–19209851	ex 4–7		NM_003476
	DMD	ex 44	3	chrX:32173487–32235180	ex 44		NM_004006
	DMD	ex 47–52	4	chrX:31747748–31950344	ex 46–52	+ 1 exon	NM_004006
	LAMA2	ex 13–14	1	chr6:129571259–129588364	ex 13–16	+ 2 exons	NM_000426
	LAMA2	ex 13–37	3	chr6:129571258–129714400	ex 13–37		NM_000426
	MTM1	ex 1–15/ WG	3	chrX:149764949–149828957	ex 3–13	- 4 exons	NM_000252
	MYPN	ex 4–7	3	chr10:69902698–69909882	ex 4–7		NM_001256267

	NEB	ex 14–81	4	chr2:152466322 –152554162	ex 14–81		NM_001271208
	NEB	ex 43–45	2	chr2:152520062 –152521377	ex 43–45		NM_001271208
	SGCD	ex 1	2	chr5:155756518 –155771687	ex 1–2	+ 1 exon	NM_000337
	SGCG	ex 7	1	chr13:23894777 –23894899	ex 7		NM_000231
	SGCG/ SACS	wide deletion	3	chr13:23853498 –24007867	*SACS* WG & *SGCG* ex 5–8	- 4 exons	*SACS*: NM_014363 *SGCG*: NM_000231
Hom del	*SGCB*	ex 6 (last)	1	chr4:52886863–52890326	ex 6		NM_000232
Hemiz del	*DMD*	ex 5–7	3	chrX:32827610–32841504	ex 5–7		NM_004006
	DMD	ex 45–47	3	chrX:31947713–31986631	ex 45–47		NM_004006
	DMD	ex 45–49	3	chrX:31854835–31986631	ex 45–49		NM_004006
Het dup	*DMD*	ex 45–49	3	chrX:31854837–31986631	ex 45–49		NM_004006
	DMD	ex 2–7	3	chrX:32827610–33038340	ex 2–7		NM_004006
	DMD	ex 1	3	chrX:33038237–33357726	ex 1		NM_004006
	LAMA2	ex 5–12	3	chr6:129468100 –129513998	ex 6–12	- 1 exon	NM_000426
	LAMA2	ex 21–55	4	chr6:129618830 –129802584	ex 21–55		NM_000426
	NEB	ex 72–81	2	chr2:152466323 –152477540	ex 72–81		NM_001271208
Hom dup	*NEB*	ex 72–81	3	chr2:152448548 –152477540	ex 72–95	+ 14 exons	NM_001271208

Orig. = Original, NP = Number of programs, Diff. = Difference in detected exons, Het = heterozygous, Hom = homozygous, Hemiz = hemizygous, del = deletion, dup = duplication, WG = whole gene

None of the four main programs called known CNVs from mtDNA samples. CNVkit provided calls only if the excepted tumor/normal contamination parameter was set to correspond the known mtDNA CNV heteroplasmy.

5.1.3 Negative control samples

No CNVs were detected in the 31 negative control samples in regions previously checked to not contain CNVs. A specificity of 100% was thus achieved.

5.2 Verified novel CNVs

36 CNVs detected by our CNV detection pipeline from targeted gene panel sequencing data were verified true positive (Table 12, the heterozygous *TTN* ex 34–41 deletion and the homozygous *SGCD* ex 1–4 deletion are illustrated against normal samples with CoNIFER and IGV in Manuscript I Figure 1). Additionally, five detected heterozygous whole gene deletions or duplications of the gene *PMP22* were well-known structural rearrangements and matched the patient phenotypes: CMT1 with duplication, HNPP with deletion. These patients had no other explaining genetic findings from previous studies. Therefore, these CNVs were considered true positive detections without need for validation. This was also the case for a compound heterozygous *SACS* deletion in a patient with spastic paraplegia. The clinically significant CNV detections will be discussed in depth further. Eight CNV detections were verified to be false positive. Therefore, the training set included 34 true positive and eight false positive CNV detections.

33 out of the 36 true positive CNVs (92%) were accurately detected from MPS data. Two CNVs were detected with one exon inaccuracy in detection compared to the validated region. One CNV (in the *TTN* TRI region) was verified with array CGH to be much larger. Most of the CNVs were detected from MYOcap sequenced samples. Only the *PMP22* and *SACS* CNVs were detected from MNDcap sequenced samples. Two of the detected CNVs were segmented into multiple detections: a large *MYOM1/MYL12A/MYL12B* duplication into three parts and a large *NEB* deletion into four parts. Some of the validated CNVs (*TIA1* and *CMYA5* deletions) were originally detected after the logistic regression model had been developed, so they were not included in the development of the model. Therefore, these validations provided 34 true positive CNV detections and eight false positive CNV detections for *in silico* CNV target design and model validation.

Table 12: CNVs verified to be true by array CGH, PCR (and Sanger sequencing) or MLPA, or by a diagnostic match. Most were included as true positive (TP) or false positive (FP) detections in the model target design and validation, but two detections were verified to be true after the model had been already developed (Other).

	Gene	CNV type	NP	Detected region	Exons	Tr.	Method	Verified region	Diff.
TP	*CACNA1A*	del het	3	chr19:13325047–13325422	ex 39–40	NM_001127221	PCR	ex 39–40	
	CACNA1A	del het	2	chr19:13325047–13325422	ex 39–40	NM_001127221	PCR	ex 39–40	
	CAPN3	del hom	2	chr15:42676666–42686546	ex 2–8	NM_000070	PCR	ex 2–8	
	COL6A1	del het	2	chr21:47404184–47404383	ex 3	NM_001848	PCR & Sanger	chr21:47403136–47406296	
	COL6A3	dup het	3	chr2:238322596–238323018	ex 1	NM_004369	4x180k aCGH	Chr2:g.(238318105_238322413)_(238346871_238440079)gain	

	COL6A3	del het	4	chr2:238250708–238255208	ex 32–37	NM_004369	PCR & Sanger	chr2:238250123–238255853	
	DMD	del hemiz	4	chrX:31645789–31986631	ex 45–55	NM_004006	PCR	ex 45–55	
	DMD	del het	3	chrX:32456358–32459431	ex 28–29	NM_004006	MLPA	ex 28–29	
	DMD	del hemiz	4	chrX:31645791–32053724	ex 45–55	NM_004006	MLPA	ex 45–55	
	DMD	del hemiz	3	chrX:31893305–32053724	ex 45–48	NM_004006	PCR	ex 45–48	
	DMD	del het	3	chrX:32305646–32408298	ex 31–43	NM_004006	MLPA	ex 31–43	
	DMD	del hemiz	3	chrX:32305646–32328393	ex 42–43	NM_004006	PCR	ex 42–43	
	DMD	del hemiz	4	chrX:31645790–31986631	ex 45–55	NM_004006	PCR	ex 45–55	
	DMD	del hemiz	2	chrX:31947713–32053724	ex 45–47	NM_004006	MLPA	ex 45–47	
	DMD	del het	4	chrX:31144758–31285069	ex 63–78	NM_004006	4x180k aCGH	ChrX:g.(30843289_31086990)_(31313100_31327408)loss	+ 1 ex
	DMD	del hemiz	2	chrX:31947713–32053724	ex 45–47	NM_004006	MLPA	ex 45–47	
	LDHB	del het	4	chr12:21807288–21810905	ex 1–2	NM_002300	4x180k aCGH	Chr12:g.(21803624_21806090)_(21811051_21812528)loss	
	LMOD3	dup het	4	chr3:69156024–69172183	ex 1–3 (WG)	NM_198271	8x60k aCGH	Chr3:g.(68897827_69131170)_(69697959–70536679)gain	
	MYH7	dup het	4	chr14:23881946–23889049	ex 28–40	NM_000257	4x180k aCGH	Chr14:g.(23859986_23859988)_(23889063_23889443)gain	
	MYL5	dup het	4	chr4:671712–673807	ex 1–4	NM_002477	4x180k aCGH	Chr4:g.(652966_654105)_(673829_674004)gain	
	MYOM1	del het	3	chr18:3075414–3075791	ex 35–36	NM_003803	PCR & Sanger	chr18:3074986–3076258	
	MYOM1	del het	4	chr18:3075414–3075791	ex 35–36	NM_003803	PCR & Sanger	chr18:3074986–3076258	
	MYOM1 /MYL12A/MYL12B	dup het	4 *	chr18:31642763278282	ex 1–10/WG/WG	NM_003803/NM_006471/NM_001144945	4x180k aCGH	Chr18:g.(3155151_3155985)_(3653512_3815684)gain	
	NEB	del het	4 *	chr2:152432208–152567053	ex 11–107	NM_001271208	4x180k aCGH	Chr2:g.(152427326_152427830)_(152567183_152567194)loss	

	PYGM	dup het	4	chr11:64513861–64528187	WG	NM_005609	4x180k aCGH	Chr11:g.(64155908_64440343)_(64553037_64758928)gain	
	SGCD	del hom	3	chr5:155753756–156016334	ex 1–4	NM_000337	PCR	ex 1–4	
	TTN	dup het	4	chr2:179522405–179523526	ex 183–189	NM_001267550	4x180k aCGH	Chr2:g.(179516799_179516833)_(179528289_179528492)gain	- 27 ex
	TTN	del het	4	chr2:179631107–179635402	ex 35–41	NM_001267550	4x180k aCGH	Chr2:g.(179630403_179630386)_(179636145–179636177)loss	- 1 ex
	PMP22	dup het	3	chr17:15133918–15168655	WG	NM_000304	Diag. match		
	PMP22	dup het	3	chr17:15133918–15168655	WG	NM_000304	Diag. match		
	PMP22	dup het	3	chr17:15133918–15168655	WG	NM_000304	Diag. match		
	PMP22	del het	3	chr17:15133918–15168655	WG	NM_000304	Diag. match		
	PMP22	del het	4	chr17:15133918–15168655	WG	NM_000304	Diag. match		
	SACS	del het	3	chr13:23902933–24007869	WG	NM_014363	Diag. match		
FP	*ANO5*	del het	2	chr11:22297639–22301311	ex 21–22	NM_213599	MLPA	false positive	
	ANO5	del het	2	chr11:22232809–22249132	ex 3–7	NM_213599	MLPA	false positive	
	ANO5	del het	2	chr11:22225350–22257822	ex 3–8	NM_213599	4x180k aCGH	false positive	
	DMD	del hemiz	2	chrX:32459297–32472949	ex 26–28	NM_004006	PCR	false positive	
	DMD	dup het	1	chrX:31645791–31986631	ex 45–55	NM_004006	MLPA	false positive	
	DMD	dup het	2	chrX:31947713–31986631	ex 45–47	NM_004006	MLPA	false positive	
	DMD	del hemiz	2	chrX:32459296–32466755	ex 27–28	NM_004006	PCR	false positive	
	DMD	del hemiz	2	chrX:32398626–32404582	ex 32–33	NM_004006	MLPA	false positive	
Other	*TIA1*	del het	4	chr2:70451556–70456450	ex 4–7	NM_022173	4x180k aCGH	chr2:g.(70450056–70456538)loss	
	CMYA5	del het	4	chr5:79086794–79089433	ex 11–12	NM_153610	4x180k aCGH	chr5:g.(79086749–79089354)loss	

NP = Number of programs, Tr. = Transcript, Diff. = difference, TP = true positive, FP = false positive, del = deletion, dup = duplication, het = heterozygous, hom = homozygous, hemiz = hemizygous, WG = whole gene, aCGH = array CGH, Diag. match = diagnostic match, * = detected as over-segmented into multiple parts

5.2.1 Accuracy of detection compared to array CGH

The region detected for the CNVs by our pipeline from MPS data was on average 109.1–298.8 kb smaller (95% CI ± 154.5 kb and ± 414.2 kb) than the region detected by array CGH by minimum and maximum array CGH coordinates. When limiting the inspection to array CGH detections with only the inspected gene (since the array CGH designs included probes for some surrounding genes not included in the targeted gene panels), the detected regions were on average 21.2–81.3 kb smaller (95% CI ± 35.7 kb and ± 148.9 kb). The most similar region was detected to have a 1.5 kb difference (*TTN* ex 35–41 deletion with -1 exon difference) and the largest had a 222.5–480.5 kb difference (*DMD* ex 63–78 deletion with +1 exon difference).

5.3 Sensitivity test with *in silico* CNVs

CNV detection sensitivity was calculated separately for small (one-exon and two-to-four-exon) *in silico* deletions, duplications, and the set of different sized CNVs with different overlap criteria (the complete table: Manuscript II Table 2). The combination of CNV *in silico* detections from all the programs provided the highest sensitivity in all *in silico* CNV categories. The sensitivity results surpassed other program combinations and detections from any one program alone (Table 13). Of the programs alone, ExomeDepth had the highest sensitivity for the small *in silico* CNVs (one-exon and two-to-four exon CNVs), but CODEX surpassed it for the larger CNVs based on real CNV detections. In overall sensitivity for all *in silico* CNV sizes, ExomeDepth was the most sensitive, then CODEX, and then XHMM and CoNIFER. Almost all programs detected deletions with higher sensitivity than duplications in each category, but CODEX was an exception with better sensitivity for one-exon duplications than deletions.

Table 13: *In silico* CNV detection sensitivities for the programs with 1 bp overlap requirement.

Programs	One-exon	Two-to-four-exon	Large
CoNIFER	1.1–4.8%	77.5–87.5%	51.2%
XHMM	10.5–61.5%	13.2–60.9%	47.2%
ExomeDepth	70.0–92.3%	83.9–96.9%	54.6%
CODEX	**30.0–38.7%**	83.1–84.7%	56.5%
All	78.3–97.8%	98.5–99.7%	97.2%

For most of the intervals the higher number displayed is for deletions, except for CODEX in one-exon detections with higher sensitivity for duplications (bold text highlight). For large CNVs the sensitivity is combined into one average.

All different program combinations surpassed the detection sensitivity by single programs in the category of large CNVs: the CNV detection sensitivity was increased to at least 71.8%. However, if detections by ExomeDepth were not included in the combination, the program alone surpassed the combinations in some of the *in silico* CNV categories. Therefore, the program combinations which surpassed the individual programs in sensitivity in all categories were all the pairs with ExomeDepth included, and all the trios with ExomeDepth. The highest sensitivity in all of the *in silico* CNV categories was reached with combination of CNV detection results from all the four programs. They provided together 78.3% and 97.8% sensitivity for one-exon duplications and deletions, respectively, 98.5% and 99.7% sensitivity

for two-to-four-exon duplications and deletions, and 97.2% sensitivity for large CNVs with the 1 bp overlap requirement. The sensitivities remained similar until 98% reciprocal overlap requirement: some of the values decreased slightly with 77.6% sensitivity reached for single-exon duplications and 97.1% sensitivity for large CNVs. This *in silico* CNV dataset provided 9514 small deletions, 9032 small duplications and 131 large CNV detections (18,677 combined) for training the model, and altogether 3892 different types of unspecific CNV detections.

5.3.1 Mosaicism sensitivity test with *in silico* CNVs

A smaller batch of *in silico* CNVs was used to test the sensitivity of detecting CNVs in mosaic ratios with the programs individually and with detections combined from all four programs (complete table: Manuscript II Supplementary table 4). In the combined results from all four programs with 1 bp overlap requirement (Table 14), mosaicism of 40% decreased the detection sensitivity mostly for one-exon CNVs compared to heterozygous CNVs. The detection sensitivity at 40% mosaicism was from 85% to close to 100% for deletions and 59–97% for duplications. Detection sensitivity for two-to-four-exon duplications decreased more notably to 83.7% with 30% mosaicism, and for two-to-four-exon deletions the greater drop in sensitivity to 77.2% was perceived with 20% mosaicism. A relatively high sensitivity of 97.5% was retained for two-to-four exon deletions at 30% mosaicism.

The detection sensitivities of the individual programs were affected by mosaicism to different degree with different CNV types. With one-exon CNVs, the most affected by mosaicism were CoNIFER and CODEX since their CNV detection sensitivity was almost zero with 30% mosaicism. CoNIFER and XHMM were the only programs with some CNV detections with 10% mosaicism in the two-to-four exon CNV category. Especially XHMM provided a relatively high sensitivity of 31.3% for two-to-four exon duplications and sensitivity of 8.7% for deletions. Detection sensitivity of ExomeDepth dropped to zero in all categories at 20% mosaicism.

Table 14: Effect of mosaicism on detection sensitivity of *in silico* CNVs at 1 bp overlap requirement with detections combined from four programs.

CNV type	50% (het)	Degree of mosaicism			
		40%	30%	20%	10%
one-exon del	99.3%	85.3%	58.0%	17.3%	1.3%
2–4 exon del	100.0%	99.8%	97.5%	77.2%	17.9%
one-exon dup	78.7%	58.7%	23.3%	17.3%	2.0%
2–4 exon dup	98.9%	97.1%	83.7%	55.9%	31.5%

del = deletion, dup = duplication, het = heterozygous

5.4 Logistic regression model training and validation

Eight of the tested models with different program combinations with features 1–4 (the program specific CNV detection scores) reached at least AUC of 0.90 (Manuscript II Figure 2). Six of these were different combinations with ExomeDepth. Two additional well-performing models were the combination of CODEX and XHMM, and a combination of all the programs except ExomeDepth. The lowest acquired AUC of 0.68 was for XHMM alone. The best combinations with AUC of 0.96 were the combination of all four programs, and the combination with all the programs expect CoNIFER.

Inclusion of other features as variables in the models was restricted by multicollinearity only for CNV detection specific score and in-house median CNV detection score for the same program. Therefore, features 2 and 5, 3 and 6, and 4 and 7 were mutually exclusive pairs. The utilization of median CNV detection scores did not improve the best models beyond AUC of 0.96 compared to CNV detection specific scores. With inclusion of other features, only target count by prioritization increased AUC marginally to 0.97. All possible combinations of the additional features were not tested, since adding them together increased the AUC also to 0.97, and excluding target by prioritization decreased it to 0.96, so this was estimated to be the only variable with some additional value. Standard deviation for the different iterations in each model test varied between 0.001 and 0.006.

5.4.1 Validation of the models with targeted gene panel sequenced real samples

The highest overall accuracy both for real CNV detections and CNV detection units from samples sequenced with MYOcap or MNDcap was achieved with the predictive model with only the CNV detection specific scores as variables (four variables out of the 12 possible) (comprehensive table: Manuscript II Table 3). This model provided a sensitivity of 96.6%, specificity of 87.5% and accuracy of 95.5% (95% CI 87.3–99.1%) for CNV detections, and accuracy of 95.9% (95% CI 88.6–99.2%) for CNV detection units. Some of the other predictive models achieved either the same sensitivity or specificity as the best model, but not the same overall accuracy. The model with median CNV detection scores as variables provided a slightly higher sensitivity of 98.3% for CNV detections, but the specificity was notably lower (62.5%), decreasing the overall accuracy. Generally, inclusion of other features in addition to the CNV detection specific scores decreased the overall accuracy.

The best accuracy for the best model was achieved with a threshold of 0.95 for the CNV detections to be predicted true. This threshold was applied for all the MYOcap and MNDcap sequenced samples, and for the WES samples initially. The true CNVs persistently predicted to be false positive were a homozygous deletion of the last exon of *SGCB* and a two-exon heterozygous deletion in *LAMA2*.

5.4.2 Validation of the model for WES samples

Validation results were evaluated separately for Coriell-samples with microdeletions and microduplications (39 CNVs in the WES batches and 34 in the mendelome batches), and other samples (56 CNVs in the WES batches and 106 in the mendelome batches) with generally smaller CNVs (listed in Methods in Table 8).

5.4.2.1 Coriell-samples

On average, microdeletions and duplications in Coriell-samples were highly over-segmented and detected in five CNV detection units in both WES and mendelome batches. In the WES sample set, this corresponded to one CNV detection unit per every 2.8 Mb, and for the mendelome sample set one detection unit for each 5.3 Mb encompassed by the CNV. The true detected regions were inferred and combined from these units as described before (the most confident breakpoints were selected). In the WES sample set, the average size difference (compared to original information) for detected regions was smaller with an average of 0.52 Mb (95% CI ± 0.4 Mb), and for the mendelome sample set larger with an average of 1.6 Mb (95% CI ± 3.2 Mb).

From the mendelome sample set, three CNVs from the Coriell samples were not detected (91.2% sensitivity), and from the detected all (100%) were correctly predicted to be true. All the Coriell sample CNVs which were not detected had a remark of “ambiguous" in the Coriell dataset, so it is uncertain whether these CNVs really exist. In the WES sample set, all the Coriell-sample CNVs were detected (100% sensitivity) and all but one were predicted true (97.4% sensitivity for predictions).

5.4.2.2 Other CNV control samples

In the mendelome sample set with Coriell-samples excluded (106 CNVs left), all CNVs were detected (100% sensitivity). Initially, 25 of the detections were predicted to be false (76.4% sensitivity for predictions). A clear cutoff was seen in the prediction values, and a new threshold was set to 0.70 for these and other WES samples further. The amount of CNVs predicted to be true increased from 14 on average per sample to 26 on average per sample with this increase of threshold. Twelve CNV detections were still erroneously predicted to be false, providing a sensitivity of 88.7% (94 samples) for true positive predictions. From these, all except two CNVs (97.9%) were detected accurately on exon level; they had a one-exon divergence compared to original information (Table 15).

When excluding one-exon deletions and duplications, from the 80 CNVs left all but six would have been predicted true (92.5% sensitivity for predictions). For the CNVs with at least three exons (N = 56), 100% sensitivity was achieved both for the detection sensitivity and predictions.

Table 15: CNV detection and prediction results for the mendelome samples with Coriell-samples excluded.

Mendelome CNV type		1 exon	2–4 exons	5–10 exons	> 10 exons / whole gene	Multiple genes
Deletion	Hom	5 ->(p) 3	5			
	Het	20 ->(p) 10	31	11 ->(a) 10	8	7
Duplication	Hom			1		
	Het	1	1	4	6 ->(a) 5	6

Hom = homozygous, Het = heterozygous. ->(p) = decrease in the amount of correctly detected and predicted CNVs due to erroneous false positive prediction, ->(a) = decrease due to inaccuracy in detected region on exon level. CNVs in the categories with no arrows were detected and predicted to be true accurately.

In the WES sample set, six of the control sample CNVs (total 56 without Coriell-samples) were not detected at all (89.3% sensitivity for detections). 33 of the detected CNVs were predicted to be true (66 % sensitivity) after the increase of the threshold. This also increased the amount of CNVs predicted to be true from 107 on average per sample to 182 on average. With exclusion of one-exon deletions and duplications from the set (originally 48.2% of the CNVs, now 29 left), 26 of the CNVs were detected (89.7% sensitivity) and all but two of these were predicted to be true (92.3% sensitivity for predictions). From CNVs with at least three exons (19 samples), all but one was detected (94.7% sensitivity), and 100% of the CNV detections were predicted correctly to be true.

From the whole WES sample set, CNVs were detected accurately on exon level in 29 of the true predicted samples (87.9%), and with maximum of one exon divergence in 32 of the samples (97%) (Table 16).

Table 16: CNV detection and prediction results for WES samples with Coriell-samples excluded.

WES		1 exon	2–4 exons	5–10 exons	> 10 exons / whole gene	multiple genes
Deletion	Hom	5 ->(p) 2 ->(a) 1	3 ->(a) 2	1	1	
	Het	21 ->(d) 18 ->(p) 6	19 ->(d) 17 ->(p) 15	2 ->(d) 1 ->(a) 0		
Duplication	Hom			1		
	Het	1	1			1

Hom = homozygous, Het = heterozygous. Decrease in sample number after the arrow ->(d) corresponds to CNVs not detected at all, after arrow ->(p) if not predicted true, and after arrow ->(a) if CNVs were not detected in the samples correctly at exon level accuracy. CNVs in categories with no arrows were detected and predicted to be true accurately.

The Blueprint Genetics sample sets contained altogether 235 CNVs to evaluate. The total detection sensitivity was 96.2% with 226 detected CNVs (if exon level accuracy is not given penalty). Of these, 196 were predicted correctly to be true, providing a sensitivity of 86.7%. For CNVs with at least three exons (N = 148), the detection sensitivity was 97.3% (144 CNVs), and all but one of the predictions were correct, which provided a prediction sensitivity of 99.3%.

5.5 CNVs with implementation of the predictive model and frequency filtration

CNV detection here and later refers to intersected CNV detection results. They have been obtained by intersecting CNV detections from the four programs in all combinations with a minimum of 1 bp overlap, as described in the methods. Multiple detections could be included for some single CNVs due to over-segmentation. On average, ten CNVs were detected per MYOcap sequenced sample, eight of which were false predicted and two true predicted. Filtering for rare variants with cnvScan according to frequency in the CNV databases was applied after this first filtration. For MYOcap sequenced samples rare, true predicted CNVs were discovered on average 0.29 per sample. The filtering steps (model prediction and frequency) preserved 2.8% of the initial MYOcap CNV detections. On average, five CNVs were detected per sample from MNDcap sequenced samples: one was predicted to be true and four false. With filtering for rare variants, 0.26 rare true predicted CNVs were discovered on average per MNDcap sequenced sample. 4.8% of the initial MNDcap CNV detections were preserved after the filtering steps.

The effect of lowering the prediction threshold for WES samples was evaluated as a separate study. After filtering for rare variants by frequency in the in-house database (built from the mendelome sample set CNV detections) and in CNV population databases, on average four potentially clinically significant and true predicted CNVs were left with the old threshold and seven with the new per mendelome sample. For WES samples, the CNV amount after frequency filtration for rare and true predicted CNVs was 11 with the old threshold and 27 with the new on average per sample. The latter values were estimated to be similar for the WES samples in other batches (not received from Blueprint Genetics for pipeline validation study) but were not evaluated separately. For these other WES samples, 312 CNVs were detected in a sample on average, 119 of which were predicted to be true positive detections. In the prediction statistics, the CNV detection results from the first WES batch were excluded due to failed ExomeDepth analysis and non-comparable CNV prediction results. After filtration for rare detections, 41 CNVs were left on average per sample, corresponding to 13.4% of the initial CNV detections.

Average prediction scores by the three programs are presented in Table 17 for true predictions and false predictions for CNVs in MYOcap, MNDcap and WES sequencing data sets. For XHMM, the total average score for true positive predicted CNV detections was approximately 78.9 and for false positive predicted CNV detections 53.4. The corresponding numbers were 87.4 and 13.3 for ExomeDepth, and 229.9 and 63.6 for CODEX.

Table 17: Average scores by XHMM, ExomeDepth and CODEX for true and false predicted CNV detections.

Program	True predicted CNVs			False predicted CNVs		
	MYOcap	MNDcap	Exome	MYOcap	MNDcap	Exome
XHMM	74.3	78.9	83.6	56.1	58.5	45.4
ExomeDepth	98.6	121.4	42.3	17.4	13.0	9.4
CODEX	196.9	378.0	114.9	22.2	16.0	25.4

Most of the polymorphic CNVs (not retained in the rare CNV detections after frequency filtrations) detected and predicted to be true in the MYOcap sequenced samples were on *TTN*, *NEB* and *TNXB*, each explaining 12–40% of the common detections and 65% together. Most of the polymorphic CNV detections in the MNDcap sequenced samples were in the genes *SMN1/SMN2, LOC283683/NIPA1, MAPT* and *MTMR2*, explaining individually 10–28% of the polymorphic detections and 79% together. For WES samples, the following genes or gene groups (gene body followed by star) explained each 1–5.8% of the true predicted polymorphisms, and 32.8% together:
MUC, IGHV*, NBPF*, GOLGA6*, CCZ1, LRRC37A, AMY1A+AMY1B, POTEB, TBC1D3, OR2A1, SPATA31A, AHNAK2, POTEG, PRAMEF, NPIPL1, TRIM49, CLEC18, NUTM2, FAM90A7, SIRPB1, NXF2.*

The program CNV detections overlapped separately for true and false predicted in different sample sequencing sets are described in a set of Venn-diagrams (Figures 17, 18 and 19) Most of the false positive predicted CNVs were detections by single programs, with ExomeDepth providing approximately 8x more false predicted detections for the targeted gene panel sequenced samples and 18x more for WES samples than any other program. CNV detections by ExomeDepth explained 55.5%, 73.7% and 81.9% of the false positive predicted CNVs in the sample sequencing sets. For MYOcap and MNDcap sample sets (with higher predictive model threshold), no detections only by CoNIFER, XHMM or by both programs were predicted to be true, and the same was true in WES samples for detections only by CoNIFER. For targeted gene panels, detections by single programs were predicted to be true more often than for WES samples, explaining 29.6% and 41.9% of the true predicted results in MNDcap and MYOcap, and 23.3% of the true predicted results in WES. ExomeDepth provided the most true predicted CNVs among the single programs, and these were mostly explained by small polymorphic CNV detections, which were discussed above. CNVs detected by multiple different program combinations were predicted to be true in each of the sequencing sets. Detections by the combination of ExomeDepth and CODEX or the three programs without CoNIFER explained 54.6% of the true predicted CNVs in WES batches and 38.0% in MNDcap batches. In MYOcap batches, 38.4% of the true predicted CNVs were detected with the pairing of ExomeDepth and either CODEX or XHMM. Overall, few CNV detections by more than two programs were predicted false in each of the sequencing sets.

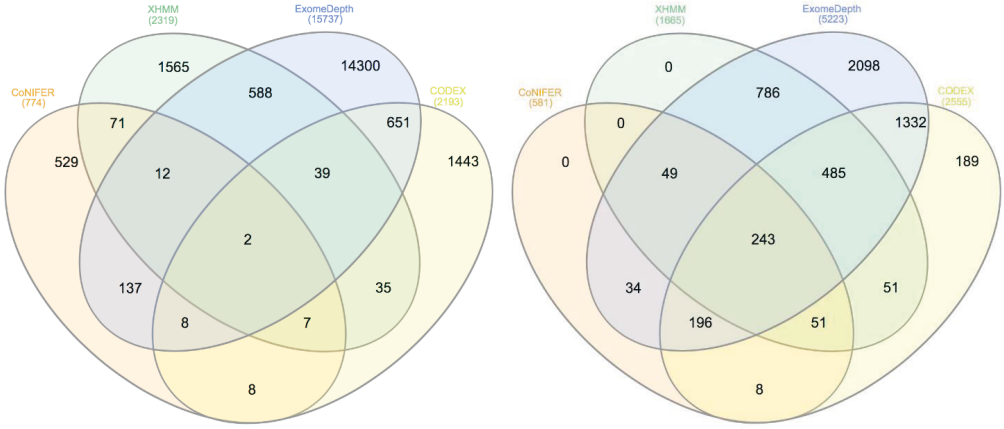

Figure 17: Number of false predicted (left) and true predicted (right) MYOcap CNV detections overlapped between the four programs. *(All the Venn-diagrams in Figures 17-19 are made by Salla Välipakka and generated with InteractiVenn (Heberle et al., 2015).)*

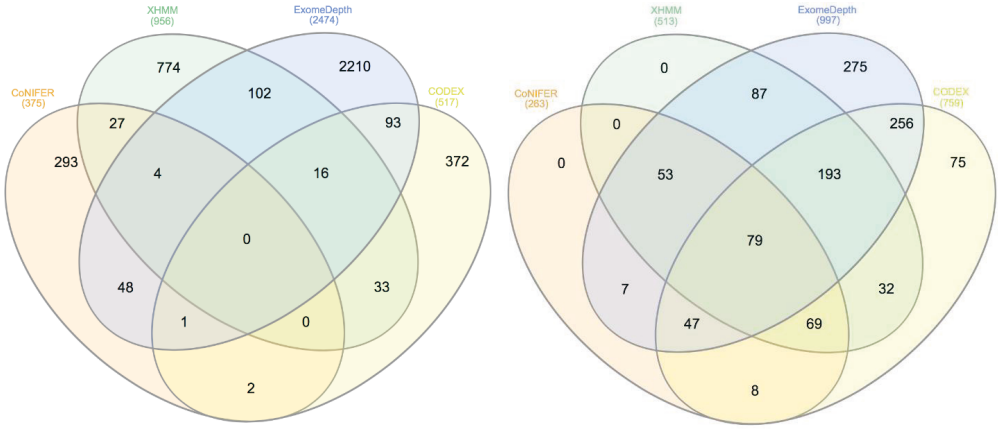

Figure 18: Number of false predicted (left) and true predicted (right) MNDcap CNV detections overlapped between the four programs.

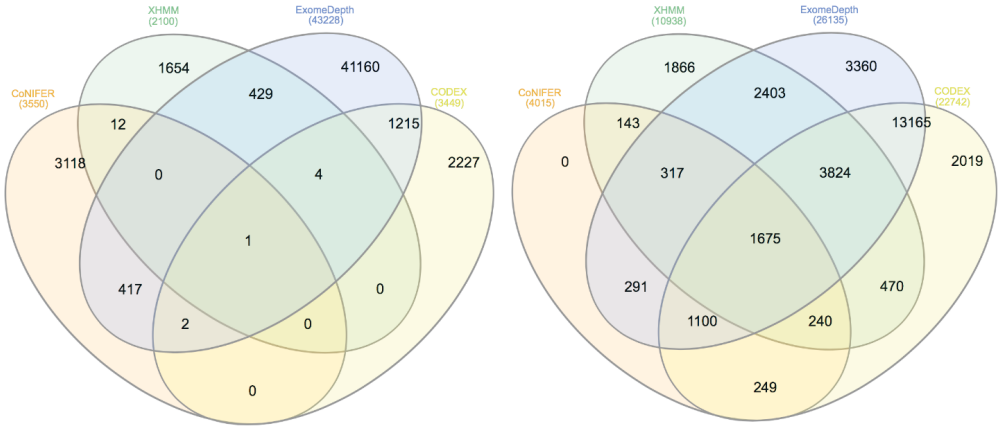

Figure 19: Number of false predicted (left) and true predicted (right) WES CNV detections overlapped between the four programs.

In MYOcap and MNDcap sequenced samples, the median CNV detections predicted to be true were slightly larger than the false predicted. CNV size was measured in base pairs with prioritization method, which was described for the logistic regression model variables. The same size difference was perceived for rare versus polymorphic CNVs in MNDcap batches, both for true predicted and false predicted detections. In MYOcap batches the rare false predicted CNV detections were similarly larger than polymorphic, but true predicted rare detections were slightly smaller than the polymorphic CNV detections (Figure 20). In the WES batches, the median size difference was more notable between true and false predicted CNVs but with same direction as for the targeted gene panels. The size differences for rare and polymorphic CNVs were smaller than for the targeted gene panels but displayed similar trends to MYOcap CNV detections (figure not displayed, but similar to MYOcap).

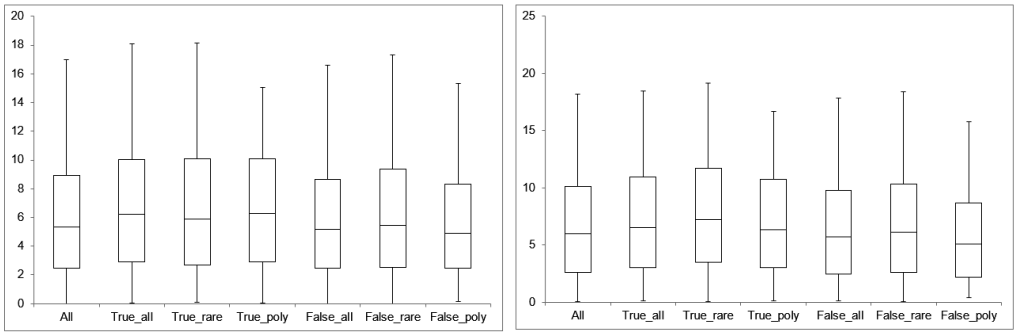

Figure 20: Variability in sizes for CNVs in MYOcap (left) and MNDcap (right) in base pairs in a logarithmic scale for CNVs predicted to be true, false, rare and polymorphic (poly).

5.6 Evaluation of CNVs for clinical significance and solved cases

39 patients sequenced in MYOcap, MNDcap and/or WES were solved with a clinically significant CNV finding (Table 18). This included 33 single cases and six patients in three families. In eight cases including one family with three patients (F1a-c, S4, S15, S16, S30, F3a), the CNV was found in compound heterozygosity with another variant in the same gene, manifesting as a recessive disease. In one of these cases (F3), segregation in an affected family member (F3b, not MYOcap sequenced and not included in solved cases) was verified with PCR and Sanger sequencing (description in Manuscript I, pedigree Figure 2). In one case (S14), the patient was revealed to have a combination of two genetic diseases, which presented as a peculiar phenotype (described more in detail in Manuscript I). In 18 cases, a likely pathogenic CNV was detected, one with a possible compound heterozygosity mechanism (SL5). This brings possibly solved patient cases to 57. In five cases a heterozygous CNV was detected, which would likely be causative if in homozygous state or in compound heterozygosity with another variant. In three of the familial cases (F1-F3), the CNV was detected to segregate with the phenotype in the family. For one of the families with a likely pathogenic finding (FL1), an additional patient case with a similar phenotype but unrelated to the other patients was found to have the same variant (*PGAP1* deletion, FL1a and FL1b and SL7). In the rest of the cases (single cases), the CNV had been previously reported to be causative for the disease (*DMD* and

PMP22) or matched the patient phenotype with no other explaining genetic findings in the patient.

Table 18: CNVs with clinical significance and solved patient cases.

	Families/ single cases	Gene	CNV type	CNV zyg	Exons	Tr.	Seq.	Other variants	Phenotype
P	F1: F1a, F1b, F1c*	*CACNA 1A*	del	het	ex 39–40	NM_001 127221	MYO	*CACNA1A* c.3604_3606 del^	episodic ataxia
	S1	*CAPN3*	del	hom	ex 2–8	NM_000 070	MYO		LGMD
	S2	*COL6A1*	del	het	ex 3	NM_001 848	MYO		Laminopathy
	S3*	*COL6A1*	del	het	ex 9–13	NM_001 848	MYO		Episodic ataxia
	S4	*COL6A3*	del	het	ex 32–37	NM_004 369	MYO	*COL6A3* p.E1386K^	UMD
	S5	*DMD*	del	het	ex 31–43	NM_004 006	MYO		Manifesting DMD carrier
	S6	*DMD*	del	hemiz	ex 42–43	NM_004 006	MYO		DMD
	S7, S8	*DMD*	del	hemiz	ex 45–47	NM_004 006	MYO		BMD
	S9*	*DMD*	dup	het	ex 45–47	NM_004 006	MYO		BMD
	S10	*DMD*	del	hemiz	ex 45–48	NM_004 006	MYO		BMD
	S11*	*DMD*	del	hemiz	ex 45–49	NM_004 006	MYO		BMD
	S12, S13	*DMD*	del	hemiz	ex 45–55	NM_004 006	MYO		BMD
	S14	*DMD*	del	hemiz	ex 45–55	NM_004 006	MYO	*TTN* finmaj het	TMD + BMD
	S15*	*GNE*	del	het	ex 2	NM_005 476	MYO	*GNE* indel^	Distal myopathy
	S16*	*LPIN1*	del	het	ex 18–19	NM_145 693	MYO	*LPIN1* c.2159T>C:p .L720P^	Metabolic myopathy
	S17	*NEB*	del	het	ex 11–107	NM_001 271208	MYO /WE S		NEM
	S18-S24*	*PMP22*	dup	het	ex 1–5/W	NM_000 304	MND		CMT1
	S25-S29*	*PMP22*	del	het	ex 1–5/W	NM_000 304	MND		HNPP
	S30	*SACS*	del	het	ex 1–10/W	NM_014 363	MND	*SACS* c.6827T>C p.L2276P^	HSP
	F2: F2a, F2b	*SGCD*	del	hom	ex 1–4	NM_000 337	MYO /WE S		LGMD

	S31*	*SGCG*	del	het	ex 4–7	NM_000231	MYO		Pathogenic, final dg unknown
	S32*	*SPAST*	del	het	ex 1	NM_014946	MND		SPG
	S33*	*SPAST*	del	het	ex 1–17/W	NM_014946	MND		SPG
	F3: F3a, (F3b)	*TTN*	del	het	ex 34–41	NM_001267550	MYO	*TTN* indel^	TMD
LP	SL1*	*AARS*	del	het	ex 5–9	NM_014946	MND		Neurogenic atrophy
	SL2*	*AR*	dup	het	ex 1	NM_000044	MND		SBMA-like
	SL3*, SL4*	*DHTKD1*	del	het	ex 2	NM_018706	MND		MND
	SL5*	*IGHMBP2*	del	het	ex 3–4	NM_002180	MND	I*GHMBP2* p.C496X^	5q-neg SMA
	SL6*	*OPTN*	dup	het	ex 3–4	NM_001008211	MND		ALS
	FL1: F1a*, F1b*, SL7*	*PGAP1*	del	het	ex 16–17	NM_024989	MND		dSMA
	SL8-11*	*SMN1*	del / conv.	hom	ex 8	NM_000344	MND		dSMA, lower motor neuron disease
	SL12*	*SPAST*	dup	het	ex 1–2	NM_014946	MND		SPG
	SL13*	*SPAST*	del	het	ex 17 (last)	NM_014946	MND		SPG
	SL14*, SL15*	*SPG7*	dup	het	ex 4	NM_003119	MND		SPG
	SL16	*TIA1*	del	het	ex 4–7	NM_022173	MYO		WDM
	SL17*	*FBXO32*	del	het	ex 1–7	NM_058229	MYO		recessive LP
	SL18*	*LDHB*	del	het	ex 1–2	NM_002300	MYO		recessive LP
	SL19*	*POMT1*	dup	het	ex 13–17	NM_001136113	MYO		recessive LP
	SL20*	*CLCN1*	del	het	ex 17–22	NM_000083	MYO		recessive LP
	SL21*	*GLE1*	del	het	ex 4–9	NM_001003722	MYO		recessive LP

P=pathogenic, LP=likely pathogenic, Tr. = transcript, Seq. = sequencing method, zyg = zygosity, del = deletion, dup = duplication, MYO = MYOcap, MND = MNDcap, conv. = conversion, het = heterozygous, hom = homozygous, hemiz = hemizygous, dg = diagnosis, LGMD = limb-girdle muscular dystrophy, UMD = Ullrich muscular dystrophy, DMD = Duchenne muscular dystrophy, BMD = Becker muscular dystrophy, TMD = tibial muscular dystrophy, NEM = nemaline myopathy, CMT1 = Charcot-Marie-Tooth disease type 1, HNPP = hereditary neuropathy with liability to pressure palsies, HSP = hereditary spastic ataxia, SPG = spastic paraplegia, dSMA = distal spinal muscular atrophy, MND = motor neuron disease, ALS = amyotrophic lateral sclerosis, WDM = Welander distal myopathy, * = CNV not verified with additional method. Case in parenthesis not MPS analysed and not involved in diagnostic yield calculations, ^ variants heterozygous and in *trans* with the identified CNV.

The diagnostic yield from the targeted gene panel sequenced samples was estimated according to 1037 samples sequenced in MYOcap: 11 of the samples were positive or negative controls, 8 were unaffected relatives, 294 were solved with findings of another variant types (SNVs and indels, 28.6% of cases) and 20 (1.9% of cases) were solved with CNV findings, including the highly likely pathogenic and verified *TIA1* finding. This provided an additional diagnostic yield of 2.7% with CNV analysis (20 of previously unsolved 730 cases). The rest of the samples from the MYOcap batches and the MNDcap batches were not included in these calculations due to ongoing diagnostic efforts. Among all the unsolved cases in MYOcap (910) and MNDcap (895), five CNVs from five patients were designated VUS in MNDcap (0.6% of unsolved cases), and altogether 42 CNVs in MYOcap from 37 patients (4.1% of unsolved cases).

Some rare but not unique CNV detections from the MNDcap sequenced samples included a partial duplication of the gene *ATM* (N = 8), partial deletion of *ATXN1* (N = 4), different partial or whole gene duplications of *PRPH* (N = 7), and partial deletions of *SCN1A* (N = 5). In the MYOcap sequenced samples, similar detections were a partial deletion of the gene *MYOM1* (N = 11), partial deletion of *CLN3* (N = 11), partial deletion of *ZBTB8B* (N= 9), and partial duplication of *MGME1* (N = 8). All of these CNVs were detected in different samples with almost the same breakpoints and with no phenotypic similarities between the cases. The *MYOM1* deletion was seen with 90% overlap in the ExAC CNV database with 0.1% frequency in the Finnish population. The others were not detected in the CNV population databases either at 90% or 50% overlap. The partial *CLN3* deletion of approximately 300 bps spanning exon 8 and partially 9 (NM_001042432) has not been recognized as a disease causing mutation (Mirza et al., 2019). For now, these findings were designated as likely not pathogenic for the patients, and variants of unknown significance.

The diagnostic yield for WES samples with sIBM samples excluded was 1.9% taking into account only the verified pathogenic variants. The amount of potentially interesting CNVs according to frequency and gene function was calculated separately for singleton cases (N = 62) and familial cases (N = 26, different combinations of cases with sequenced affected or unaffected relatives with total of 64 samples). Altogether, 211 CNVs were considered as potentially interesting for singletons, providing eventually 28 CNVs designated as VUS in 19 different samples, while the rest were evaluated as not interesting for further evaluation for the patient according to lack of putative phenotypic match to gene function. For the familial cases, 80 CNVs were evaluated as interesting, and 13 of these were eventually designated as VUS involving 16 patients in different families, and the rest were designated as probably not putatively matching for the patients. Altogether, 34% of the unsolved patients in WES were detected to have at least one CNV designated as VUS. 1369 unique CNVs were detected from the samples altogether.

The two clinically significant CNVs detected from WES samples, a large heterozygous *NEB* deletion (chr2:152432208–152567053, ex 11–107) and a homozygous deletion in *SGCG* (chr5:155753756–156016334, ex 1–4), were detected from samples included also in the

MYOcap batches. Additionally, one affected relative was detected to have the same *SGCG* deletion in WES. Therefore, these served as validation samples for the CNV detection from WES samples, but also corresponded to 2.1% increase in yield with three of original 140 unsolved WES patients solved. No CNVs with higher representation in sIBM patients compared to other WES patients were perceived, and no overrepresentation of separate CNV detections in a gene or genes was either detected.

5.6.1 Experimental CNV evaluation with the novel ACMG recommendations

The rare and true predicted CNVs detected in the MYOcap and MNDcap sequenced samples were experimentally evaluated according to the new criteria recommended by ACMG for CNVs. For MYOcap, rare CNVs were detected in 73 different genes. For 15 of these genes, a haploinsufficiency (HI) prediction was provided, for 10 genes a triplosensitivity (TS) prediction, and for eight genes both of the predictions from the ClinGen catalog. Additionally, five of the genes were predicted to be haploinsufficient according to the accepted HI predictors, gnomAD pLI score and DECIPHER HI index. One was *DMD*, which had already HI information provided from the ClinGen catalog. In MNDcap, 77 different genes had rare CNV detections. Nine of these had a HI prediction and 16 a TS prediction from the ClinGen catalog, and 10 had both. Additional seven genes were predicted to be haploinsufficient by the two HI predictors. Two of these genes had also HI and TS scores from the ClinGen catalog: *AR* and *SPAST*. For all the single cases, the inheritance was unknown or was assumed to be *de novo* occurrence. In the evaluation workflow, the following evidence were used:

1) **Gene content.** For all detections 1A "Contains protein-coding or other known functionally important elements" was selected. The significance score was not affected by this selection.

2) **Overlap with established/predicted TS/HI regions.** For most of the detected CNVs, the section 2) had to be skipped. They had no information available from ClinGen or by the two HI predictors for the affected gene. CNVs affected 11 MYOcap genes and 13 MNDcap genes with the annotation of autosomal recessive gene, so this section was skipped also for them.

The genes with HI prediction available were mostly detected with a partial intragenic deletion or duplication (*CACNA1A* and *DMD* in MYOcap, and *SCN1A* in MNDcap) leading to evidence 2E. In MNDcap, duplications in *CACNA1A, AR* and *SPAST* covered partially the beginning of the gene (evidence 2K), *PMP22* deletions and duplications the whole gene (evidence 2A), and three different *SPAST* deletions were discovered covering the whole gene (2A). One of them involved the beginning of the gene (2C1), and one partially the end (2D4). For the partial intragenic deletions, the reading frame was expected to be preserved (since no exact information was available), and the altered region was expected to be critical for protein function, giving evidence PVS1_Strong. For the partial intragenic duplications, the copies were presumed to be in tandem (according to likelihood), and no information on effect on reading frame or NMD were available, which provided a N/A score. Some of the genes with deletions had HI predictors surpassing the set threshold, giving them evidence 2H.

3) **Gene number**. All deletions contained less than 25 genes and all duplications less than 35, leading to evidence 3A for all cases.

4) **Detailed evaluation of genomic content.** Only for the CNVs affecting *DMD* or *PMP22* the evidence 4A "The reported phenotype is highly specific and relatively unique to the gene or genomic region" was used, and with assumption of *de novo*. For all the others, the evidence 4C "the reported phenotype is consistent with the gene/genomic region, but not highly specific and/or with high genetic heterogeneity, expected *de novo*" or 4E "Reported proband has a highly specific phenotype consistent with the gene/genomic region, but the inheritance of the variant is unknown" or 4D "the reported phenotype is NOT consistent with the gene/genomic region or not consistent in general" were used, but the selection was mostly arbitrary due to the heterogenic nature of neuromuscular disorders in general. For familial cases, the lowest segregation category 4F was used. Additionally, evidence 4M "Statistically significant increase amongst observations in cases (with a consistent, non-specific phenotype or unknown phenotype) compared to controls" were used for all the evaluated CNVs according to common database frequency filtrations.

5) **Evaluation of Inheritance/patient history.** In this category, for all of the single cases the evidence 5G "Inheritance information is unavailable or uninformative. The patient phenotype is non-specific but is consistent with what has been described in similar cases" was selected since all of the patients are evaluated to have some neuromuscular disorder.

With this evaluation workflow, CNVs evaluated to be pathogenic (all achieving the maximum score of 1.0) would have included the *CACNA1A* deletion in family 1 with route 1A 2E 3A 4C/4E 4M, *DMD* deletions in patients S5-S14 (S9 with *DMD* duplication excluded) with route 1A 2E 2H 3A 4A 4M 5G, *PMP22* deletions and duplications in patients S18-S29 with route 1A 2A 3A 4A 4M 5G (providing the highest score, if calculated beyond 1.0), and all of the *SPAST* deletions with the different routes 1A 2A/2C1/2D4 3A 4C/4E 4M 5G. In total, 27 patient cases were evaluated to carry a pathogenic CNV. The *SPAST* duplication could be given either the score of 0.95 or 1.0, with threshold of pathogenicity at 0.98, by choosing between a value of 0.10 or 0.15 for the evidence 5G from the suggested interval. Two of the *SPAST* cases were evaluated to be eventually pathogenic and two likely pathogenic in our patient cohort.

For all the other CNVs, scores between 0.3 and 0.65 were reached, resulting as evaluation as variant of uncertain significance. The most common score of 0.50 was reached for the individual cases with skipping of step 2) with route 1A, 3A, 4C/4E, 4M, 5G. With families, a score of 0.55 was achieved with route 1A, 3A, 4C/4E, 4F, 4M. Alternatively, if the phenotype was estimated to fit more evidence 4D, the 4C/4E score was replaced with 4D and the scores were 0.3 or 0.35 in the cases described above. If the predictors gave a significant HI evaluation for deletions, the score was 0.65 with the route 1A, 2H, 3A, 4C/4E, 4M, 5G. For example, section 2 was skipped for CNVs in the genes *CAPN3* and *SGCG* predicted to cause recessive diseases, but the patients here had homozygous deletions, and thus the identified CNVs were

designated as pathogenic. In six additional cases, the pathogenicity of the CNV was mediated by compound heterozygosity with another variant on the same gene, which was not considered in the interpretation workflow.

5.6.2 *NEB TRI* region analysis results

In the MYOcap sequenced samples, 14 were detected to have divergent *NEB* TRI region CNV detections compared to others in the in-house CNV database. These were attempted to verify with array CGH (4x180k). One case was solved with a verified 8/6 TRI copy number, which is clinically significant and matched the patient's distal-proximal myopathy with heavy nemaline rod pathology (Kiiski, K. et al., 2016). This patient is included in the diagnostic yield calculations. Two cases had a verified 4/6 TRI copy number, which would be causative if in *cis*, but the phenotypes did not match nebulin pathology (Kiiski, K. et al., 2016). Four of the cases had an ambiguous array CGH result between TRI copy numbers of 7/6 and 8/6, with no conclusive result. Four of the cases had a clinically non-significant TRI copy number of 5/6, and three had a copy number of 6/6, which is normal (Kiiski, K. et al., 2016). The algorithm for differentiating the copy number of the *NEB* TRI repeat blocks from sequencing data did not provide rational results comparable to the verified *NEB* TRI region CNVs. Furthermore, no sample received the expected normal result of copy number estimation of two for each repeat block, so this analysis was not considered a successful approach.

5.6.3 *SMN1/2* analysis results

Most of the samples sequenced with MNDcap had a *SMN1/SMN2* count of 2/2 or 2/1 (Table 19). Two positive control samples were correctly verified to have a *SMN1/SMN2* count of 0/3. Additional three samples were verified to have the same *SMN1/SMN2* 0/3 count, and two samples a 0/4 count. Seven detections of status 0/0 originated from samples with low average read depth (samples with failed sequencing). Detections of higher number of *SMN1* copies than four were rare. Copy number counts such as 4/2 tended to originate from samples with higher read depth than on average in their respective batches: in a batch with 63 samples six out of the 10 samples with the highest coverage in the batch had a *SMN1* copy number count of ≥3. These counts were not verified further and converted (for example to 2/1), since they appeared to not be clinically significant divergences in either case.

Table 19: *SMN1/SMN2* copy number counts for MNDcap samples.

Number of samples	*SMN1/SMN2*	Number of samples	*SMN1/SMN2*
1	2/4, 4/3, >4/0, >4/2, >4/3, >4/4	17	3/0
2	0/4, 1/4, 4/1	22	2/3
3	3/4, 4/4	29	3/2
4	1/3	41	1/1
5	0/3, 1/0, 4/0	52	2/0
6	1/2	75	3/1
9	4/2	262	2/1
16	3/3	374	2/2

All of the *SMN1* copy number 0 detections were investigated in IGV. None of the new 0/4 or 0/3 detections had the whole *SMN1* gene deleted. However, these cases seemed to have a drop in coverage in the clinically significant exon 7 of *SMN1* (Figure 21).

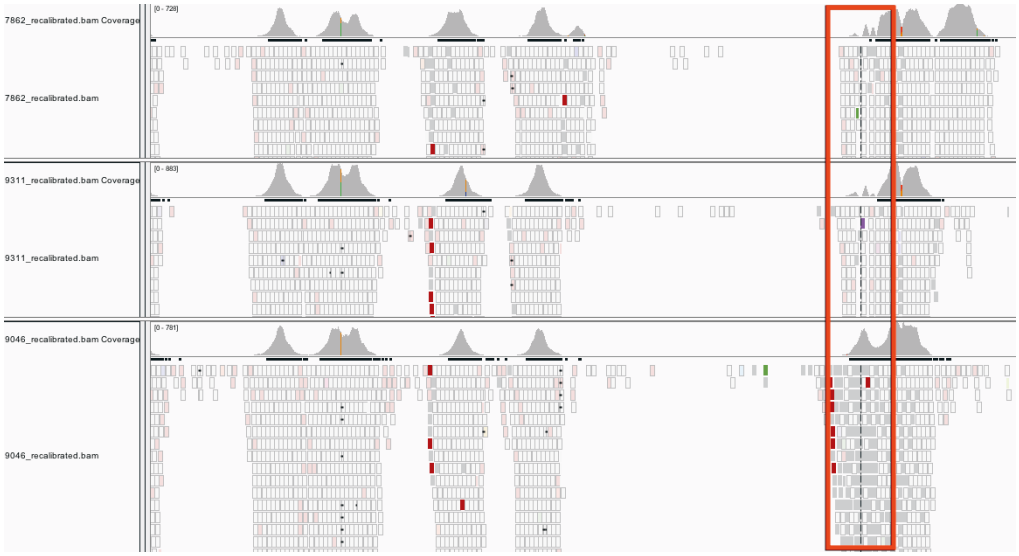

Figure 21: Local drop in coverage of exon 7 of *SMN1* as visualized with IGV, the bottom sample as control.

Additionally, the corresponding region in *SMN2* appeared to have been doubled in coverage compared to the rest of the gene. As an investigation approach, *SMN2* from the reference genome was covered preventing alignment and the reads then re-aligned. According to the known nucleotide mismatches in the intron 6 (chr5:70247724G>A) and exon 7 (chr5:70247773C>T) (Figure 22), all of the reads had the sequence of *SMN2*. A partial conversion was hypothesized to have occurred for the exon 7, which is a known event for the *SMN1/SMN2* homologous genes, and has a potential clinical significance, as described before.

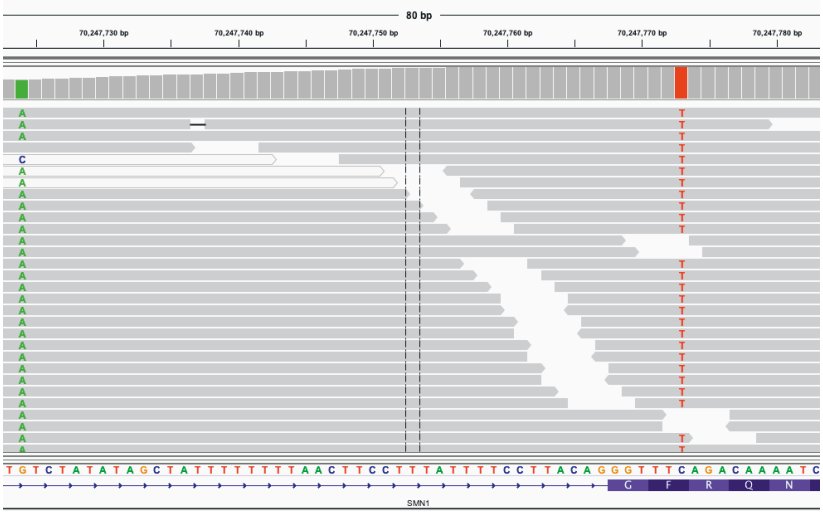

Figure 22: All reads re-aligned to *SMN1* in a potential *SMN1* exon 7 conversion case as visualized with IGV.

5.7 Sequencing data quality and CNV detection accuracy

5.7.1 Inspection of sample and batch quality with false negative CNVs

For the CNVs not detected at all or detected but predicted erroneously false positive, batch, sample and region quality were inspected compared to samples/batches/regions with successful CNV detections and predictions. The evaluated metrics were total read depth and coverage uniformity calculated with CoverView and ExomeCQA, respectively. No apparent differences were detected in sample quality, batch quality or region quality. Interquartile range (IQR) was also evaluated as a metric to calculate read depth uniformity and enable comparison between samples and batches. However, all the samples in an evaluated batch had an IQR higher than 15, which has been defined as a threshold for high IQR for samples by Trost and colleagues (Trost et al., 2018). Therefore, this metric was not evaluated to be descriptive enough, or reliably assignable with an appropriate threshold without further validation steps.

5.7.2 Correlation of read depth and coverage uniformity to CNV detections

Only a few of the tested CNV detection statistics demonstrated statistically significant correlation with some of the tested batch read depth and coverage metrics (Table 20). The CNV statistics with significance were amount of false predicted CNVs and true predicted CNVs on average per sample in a batch, and the standard deviations in these within the batches. The most significant (with a threshold of p-value < 0.05) correlation of -0.90 was detected between the amount of false positive CNV predictions and the ratio of targets with either low coverage or low uniformity with a p-value of 1.27e-05. The measure of % of bases covered by a minimum of 20X in the batch had the second strongest correlation with 0.86 and a p-value of 6.83e-05. All the batch metrics were correlated to some degree with the amount of false CNV predictions. Ratio of targets with either low coverage or low uniformity and the % 20X measurement had somewhat significant correlations with some of the other CNV statistics with p-values between 0.01 and 0.04. The hypothesis that the batch read depth and the coverage uniformity affect the amount of putatively clinically significant and/or unique CNV detections was not confirmed with this evaluation.

Table 20: Significant correlations (above) and p-values (below).

Correlations	LOWeither	LOWc	LOWU	LOWcU	covmin20
nmeanFalses	-0.90	-0.61	-0.63	-0.81	0.86
nmeanTrues	-0.56	-0.44	-0.35	-0.50	0.70
nsdFalses	-0.56	-0.41	-0.39	-0.48	0.56
nsdTrues	-0.57	-0.42	-0.39	-0.51	0.54
p-values	LOWeither	LOWc	LOWU	LOWcU	covmin20
nmeanFalses	**1.27E-05**	**0.02**	**0.02**	**0.0004**	**6.83E-05**
nmeanTrues	**0.04**	0.12	0.22	0.07	**0.01**
nsdFalses	**0.04**	0.15	0.16	0.08	**0.04**
nsdTrues	**0.03**	0.14	0.17	0.06	0.05

LOWeither = ratio of targets with low coverage or low uniformity, c = coverage, U = uniformity, covmin20 = % of bases in the batch covered by minimum of 20X. Significant p-values of < 0.05 with bold text.

6 DISCUSSION

Massively parallel sequencing (MPS) approaches have been widely used in the diagnostic efforts for different disease groups to increase the yield of genetic diagnoses (Ellingford et al., 2016; Garg et al., 2020; Srivastava et al., 2019). However, CNV detection from MPS data has been less common, since no best practices pipelines are available, and different settings require the use of different CNV detection tools. The individual tools have been reported to have low sensitivity and specificity, which prevent their routine use in a diagnostic setting. In neuromuscular disorders, multiple recurrent causative CNVs have been recognized, such as the reciprocal deletion and duplication of the gene *PMP22,* and several CNVs in the gene *DMD* (Truty et al., 2019; Bladen et al., 2015; Giugliano et al., 2018). Generally, neurologic disorders are estimated to have CNVs more represented as genetic causes compared to some other disease groups (Truty et al., 2019).

Here, we set to develop a CNV detection pipeline both for the targeted gene panel sequencing data and WES data in our disease cohort of neuromuscular disorders. We successfully evaluated the performances of the programs and validated a program combination with high CNV detection sensitivity. A predictive model was developed to increase CNV detection specificity and decrease the workload in the bottleneck step of variant effect interpretation. CNV annotation with an existing tool was improved and updated to include the most up-to-date information on CNV clinical significance interpretation available. The diagnostic yield was successfully increased in our patient cohort, along with increase in insight into CNV analysis from the genes associated with neuromuscular disorders.

6.1 Technical aspects

The performance of the utilized programs was generally as expected and documented in earlier comparable studies (de Ligt et al., 2013; Gambin, Yuan et al., 2017; Hwang et al., 2015; Kim, H. Y. et al., 2017; Kadalayil et al., 2015; Roca et al., 2019; Sadedin et al., 2018; Samarakoon et al., 2014; Tan et al., 2014; Yao et al., 2017). CNVs sized 1–100 kb seemed to be the most common, as has been reported in other short-read MPS population CNV studies (Zarrei et al., 2015; Conrad et al., 2010). CoNIFER detected mostly larger sized CNVs, while ExomeDepth detected more CNVs of smaller size. By closer inspection this appeared to be partly explained by ExomeDepth over-segmentating single CNV detections into multiple smaller parts. One explanation for this behavior is that uneven spacing of exons in WES data could lead to over-segmentation (de Ligt et al., 2013). However, the number of over-segmentation events was rather low, involving approximately 2% of CNV detections. CODEX and XHMM were also involved in some over-segmentation events. Therefore, the CNV size distribution in the detections could not be explained only by the differences in the over-segmentation frequency. In the state distribution, ExomeDepth had a bias for deletions, CoNIFER, XHMM and CODEX for duplications, and XHMM and ExomeDepth had the most balanced distribution. Since the genome contains more duplications than deletions, CoNIFER and CODEX may in fact have

the least biased and most reliable detections concerning CNV state distribution. On the other hand, deletions are of greater clinical interest, meaning that a bias in detections, if they are true, should not be considered a problem in a diagnostic setting.

Rare (< 1% frequency in databases) CNVs were in most cases detected to be larger in size than polymorphic, which has also been previously reported (Collins et al., 2020). True predicted CNV detections were generally larger in size than the false predicted detections. This was somewhat surprising, since CNV size as a variable did not significantly affect the performance of the predictive logistic regression model. Eventually, CNV detection specific scores by the programs were the only variables in the predictive model that provided the most accurate results. These scores were expected to be affected by the different CNV properties (state, size) included as features in other model versions. Based on previous comparison studies, an intuitive hypothesis was that higher scores would be achieved for large deletions than the small duplications in the order of ease of detection. However, in contrast to these expectations, no multicollinearity was observed between most of the tested variables, which enabled the testing of more varied model versions.

Most of the CNVs predicted to be false positive were detected uniquely by single programs, and unique detections by ExomeDepth are overrepresented in this group. This is an expected result, since ExomeDepth is the most sensitive of the programs, but also produces more false positive CNV detections. Partially, this observation validates the predictive model for differentiating false positive detections from true positive detections correctly. In one study, ExomeDepth had an average Bayes factor of 45.1 for true positive detections from WES samples (Ellingford et al., 2017), which closely corresponds to our result 42.3 for WES samples with the predictive model. This is an additional level of validation for the model selected for further use. On the contrary, CoNIFER is supposed to provide highly specific CNV detections, but none of the CNVs detected uniquely by CoNIFER were predicted to be true. CoNIFER detections were given an arbitrary binary scale in the predictive model, while for the other programs the provided quality scores for CNV detections were utilized, which may affect these results. Additionally, in this study lenient SVD-values were used for CoNIFER to increase sensitivity, which may have decreased the inherent specificity of the program.

As stated, ExomeDepth lacked in specificity compared to the other programs in this study, which has been repeated also in other studies. In previous studies, the accuracy of ExomeDepth has been increased with pre-modification of data by removing exons with low mappability prior to CNV calling (Rajagopalan et al., 2020). This decreased the false positive rate for ExomeDepth while retaining a high sensitivity (Rajagopalan et al., 2020). However, the removed regions roughly corresponded to retaining only unique regions (Rajagopalan et al., 2020). In another study, ExomeDepth sensitivity was increased by restricting the analysis to non-polymorphic genomic regions (Marchuk et al., 2018). These solutions have not been as automatized or statistically verified as our approach with the logistic regression model. Our

model should not *a priori* exclude the CNVs based on sequencing quality or region repetitiveness, but this must be verified.

According to our CNV detection results in the negative control samples, a specificity of 100% was reached. However, this was not a rational approach for measuring the specificity of the method considering the amount of false positive CNV detections in verification attempts. The first CNVs to be verified were selected among those that had been detected by more than one program. As observed with the variants verified to be false positive, this was not an accurate enough approach for filtering for true positive variants. Nevertheless, this approach has been used in the more recent papers: in one study CNVs were evaluated true if detected by at least three tools with no score evaluation or statistic models used (Roca et al., 2019). In the light of the results of our study, this approach is problematic. In another study, a rudimentary algorithm was used, which scored CNV detections similarly by the number of programs which had detected them. In that study, none of the combinations surpassed the individually most accurate algorithm, and also loss in sensitivity was detected (Trost et al., 2018). The results in our study are clearly different with increased CNV detection sensitivity by most of the program combinations. These results will be further discussed.

The preceding observations led to the development of the predictive logistic regression model. Not enough samples had been verified with true positive or false positive CNV detections to train any statistical model reliably for evaluation of the results. Therefore, *in silico* CNVs were utilized. This is not a new approach, but the exact method we used here has not been utilized in other settings to our knowledge. Adequate statistical power was achieved when the *in silico* CNV detections provided 18,677 true positive and 3,892 false positive (unspecific) detections for the model training. The highest achieved AUC of 0.97 for these *in silico* CNVs also predicted well the model accuracy for true CNV detections, reaching an accuracy of 96%. Therefore, the concerns that *in silico* CNVs may inadequately represent real data were avoided in this study. This was likely owing to the use of reads from real sequenced samples, and the generation of both deletions and duplications (Ellingford et al., 2017; Kadalayil et al., 2015). This approach, although laborious design-wise and computationally, provided the training set with similar properties to the real sample data.

Sensitivity evaluations with the *in silico* CNVs provided additional information on the performance of the programs. ExomeDepth was the most sensitive for the smaller CNVs, as expected. Interestingly, CODEX surpassed it for larger CNVs, although ExomeDepth has been previously reported to have the highest sensitivity also for these CNV types. The sensitivity of the individual programs to detect the large CNVs was surprisingly low even at the most lenient 1 bp overlap requirement: 56.5% for the best performing program CODEX. These programs have also been used individually, such as XHMM for building the whole ExAC CNV database (Ruderfer et al., 2016). Although the specificity of the detections was increased through comparison to array CGH detections, this raises the question whether the sensitivity was adequate for a population level CNV representation. In the novel CNV database of gnomAD,

four programs were used for CNV detection from sequencing data (Collins et al., 2020). This should provide a more comprehensive CNV database according to our study and current common consensus. In our study, combining detections from any of the programs increased the detection sensitivity at least to 71.8%. The minimum sensitivity of 97% achieved for most of the *in silico* CNV categories with detections combined from all four programs was very high, dropping more notably only for one-exon duplications. One-exon CNVs are the most difficult to detect with current read depth methods from MPS data (Marchuk et al., 2018). Thus, the achieved 97.8% sensitivity for one-exon deletions was higher than expected, mainly due to the high sensitivity provided by ExomeDepth. CODEX increased the sensitivity for one-exon duplications to a moderate clinical sensitivity of 78.3%.

The effect of mosaicism tested with *in silico* CNVs was surprisingly variable for the different programs depending on the CNV size and state. Generally, XHMM seemed to be the least affected by mosaicism, and ExomeDepth and CODEX were the most affected. Complementarity in this setting in program performances has not been demonstrated previously to our knowledge. However, the test set included only certain cutoffs for the degree of mosaicism, a smaller *in silico* CNV sample batch and only one or two-to-four exon CNVs, since mosaicism is not a common genetic mechanism in our patient cohort. Nevertheless, the test should be expanded with more variable CNV types and a continuous distribution of the degree of mosaicism to allow proper evaluation of a threshold for CNV detection. Based on our study, the threshold will probably vary for different CNV types. With 40%, mosaicism one-exon deletions could be detected with over 85% sensitivity, whereas one-exon duplications were detected with less than 60% sensitivity.

The average read depth and uniformity of coverage in sequencing batches were shown to not have significant correlation with most of the CNV metrics evaluated in WES batches. Most of all, the amount of rare or unique CNV detections per sample was not affected, as was hypothesized to originate from read depth fluctuations with low coverage uniformity. However, these could be “hidden” in the false positive predicted CNVs, the amount of which per sample achieved the most consistently high and significant correlations. This was also an expected result, since low average read depth and low coverage uniformity should increase the amount of false positive CNV detections due to technical bias. Surprisingly, low uniformity had a comparable effect to low read depth, although uniformity has been estimated more important for CNV analysis from sequencing read depth (Kerkhof et al., 2017). The number of targets with decrease in either showed the highest and most significant correlation with the amount of false positive predicted CNVs. Almost as high correlations were achieved with the traditionally utilized measure of % of bases covered by minimum of 20X, so this would be a simple and informative measure also in the future. However, for deciding proper thresholds for sequencing data for accurate CNV detection, the same samples should be studied with different coverages, and more samples would be needed with CNVs liable for false negative result.

CNVs were detected with high accuracy from each sequencing data set (MYOcap, MNDcap,

WES) both in the positive control samples and in the verified novel cases. The varying technical implementations for sequencing in the different batches did not seem to affect the CNV detection accuracy. As previously demonstrated in a study by Kosugi and colleagues, the read length and insert size did not greatly affect CNV calling accuracy (Kosugi et al., 2019). With the targeted gene panels, 100% sensitivity was reached for CNV detections, 71% of which with exon-level accuracy for the variants with available data on exon-level breakpoints. The decrease in sensitivity with exon-level threshold could be problematic in some settings, such as for CNVs in the gene *DMD* with diagnostic implications. In our cases, the inaccuracies would not have affected the putative diagnoses, since all true hemizygous *DMD* deletions were detected accurately. However, need for additional verification and/or improvement in detection accuracy should be considered.

For the WES validation samples, a sensitivity of 100% was reached for the mendelome validation samples and 95% for the WES sample CNVs with at least three exons. These percentages are high enough for a clinical setting. However, with smaller CNVs included, the sensitivities are notably lower and differ also between the two WES validation sample types. In one previous study, with ExomeDepth alone, a sensitivity of 87% has been reached for CNVs smaller than four exons, and an overall sensitivity 96% from WES data (Rajagopalan et al., 2020). In the mendelome samples, these values were surpassed with the four-program combination, but the results were similar for the WES samples. Some of the inaccurate CNV detections or samples with no CNVs detected could be explained by the affected genes having a highly homologous pseudogene, such as *STRC*. An undetected seven-exon deletion was supposedly located in a gene with a similar homology issue, *IKBKG*.

The samples and batches with false negative control CNV detections could be evaluated for properties affecting the CNV detection sensitivity. Most of the undetected or incorrectly false predicted positive control CNVs from WES data and targeted gene panel data were single exon heterozygous deletions. This is the most challenging CNV type to detect and also most often involved in false positive detections, making deciphering these CNV detections challenging (Marelli et al., 2016; Zenagui et al., 2018). Many of these (6/35 in the WES samples and 1/2 in the MYOcap samples) affected only the first exon of the gene. Accordingly, false negative detections for CNVs from WES data have been observed to be enriched in both the last and first exons, which are often GC-rich and thus challenging to sequence and analyze accurately (Rajagopalan et al., 2020; Zenagui et al., 2018). These small CNVs were persistently predicted to be false positive detections both in WES data and targeted gene panel sequencing data. They appeared to have low program scores and were thus probably liable to be discarded as false detections, although CNV size was supposedly not a significant factor in the predictive model. In a previous study, one-exon deletions were not identified with high intra-sample variation and insufficient coverage for exons and nucleotides (Ellingford et al., 2017). This matches the theory that detecting differences based on one data point with one-exon CNVs allows for little experimental background noise (de Ligt et al., 2013). This could also be the explaining factor for the missed CNVs being mostly small. However, single sample quality metrics have been

also thought to be less informative in estimating performance in CNV detection with the read depth method compared to correlation across the samples in the batch (Plagnol et al., 2012).

The first tests we performed for batch/sample/region read depth and coverage uniformity and CNV detection sensitivity were not informative. For our sample set, the only notable common factor for batches with higher CNV detection and prediction sensitivity was the number of genes in target sets. The mendelome and targeted gene panels had less genes targeted (gene panels < 350 genes, mendelome < 6500 genes) than "real" WES batches. Intuitively, this could provide inherently more uniform distribution of reads. However, the measurement for evaluating coverage uniformity utilized here did not capture this relation, or the inspected variables were not affected by it. In a previous study, high variation in read depth across the sample pool decreased the detection accuracies of ExomeDepth, CoNIFER and XHMM (Samarakoon et al., 2014), but multiple ways to measure coverage uniformity exist. In the same study, small size of the sample batch impaired the program performances as well. However, in our study the samples with false negative CNV detections were included in batches of average size, which should be within the recommendations for the programs. *LAMA2* and *SGCB* are some of the genes, which have been detected to have potentially too low coverage (less than 20X) for variant detection in targeted MPS studies (Ankala et al., 2015), which could be filled in with Sanger sequencing. However, the batches where the control CNVs in these genes with false negative detection results were included did not display low coverage for these genes, and the samples with the CNVs also had high-quality sequencing results with > 99% 20X coverage.

The eventual performance of the predictive model with 96% overall accuracy for targeted gene panels is good enough for a diagnostic setting. For WES samples, the prediction sensitivity varied between 66% and 100% depending on CNV type and sequencing scale, which necessitates checking the unfiltered results in cases of no findings. In one WES familial case, affected siblings, this additional evaluation of unfiltered CNV detections provided a CNV finding designated as VUS, which awaits further clarification for clinical significance. The workload in variant interpretation was relatively unaffected by these occasional evaluations. However, some (mostly small, < 3 exons) control sample CNVs from WES samples were completely missed, highlighting that care needs to be taken in evaluating WES samples as negative for CNV findings.

Although the average sizes of microdeletion and duplication samples included in the mendelome and WES sample sets were similar, the differences in detected CNV sizes compared to the original information were clearly larger in the mendelome set. CNV detections in the mendelome set were generally more accurate on exon level for other validation samples. One affecting factor may be that the Coriell-samples were distributed into all three batches with other samples, which increased variability between the samples. On the other hand, the CNV detections in Coriell-samples were more over-segmented in the WES set compared to the mendelome set, which may be explained by less uniform data. In conclusion, these CNV detection methods cannot provide accurate breakpoints for microdeletions and duplications.

The original breakpoint information for the samples could also be inaccurate. Anyway, these tools have the potential to screen for these larger structural variants.

The window-based read depth estimation has been considered to be unsuitable for detecting aneuploidies (Trost et al., 2018). Nevertheless, potential aneuploidies for chromosome X with XXX for females were detected in our cohort, and chromosome X triplications have been detected also in other studies (Kerkhof et al., 2017). The chromosome X triplications were detected in eight MYOcap sequenced patients, corresponding to a frequency of 0.6%, which is higher than the expected 1/1000 for females (Mayo Clinic, www.mayoclinic.org). Some of these could be mosaic cases or false positive detections. These would need to be verified with a complementary method before further clinical evaluation for manifestations. In addition, some cryptic large partial chromosome X deletions and duplications, which could refer to Turner syndrome (X0) or Klinefelter syndrome (XXY) were observed.

The few additional tested CNV detection tools (CNVkit, SavvyCNV) failed to detect positive control CNVs, which were clearly detected by the four programs used here. We did not succeed in employing some of the other programs (DECoN, DeviCNV) due to lack of documentation and probable platform incompatibilities, which is also a common problem (Bakhtiari et al., 2018). Additionally, CNVkit has been reported to have an unpredictable performance, such as decrease of sensitivity with increase of sequencing read depth (Roca et al., 2019).

None of the four main programs detected CNVs from mtDNA samples, or the CNV analysis was unsuccessful. CNVkit with a mode for analyzing tumor samples was successful for two of the control samples containing a single known deletion only after providing the expected CNV heteroplasmy degree as a parameter. Providing other heteroplasmy degrees led to variable CNV detection results; consequently, the method would be inaccurate for blind samples. An additional reason for the inaccurate and unsuccessful analyses could be that the target set had to be designed with an arbitrary distribution of intervals to cover the mtDNA genomic region. This does not correspond to the original capture with overlapping probes, which was not accepted by the programs as a target file. On the other hand, the same approach of arbitrary distribution was utilized for the *TTN* gene target region covered completely (i.e. UTRs, introns and exons) in the latest MYOcap version, and the CNV analysis was successful according to accurate detection of one known heterozygous deletion. The most relevant differences could be in the genomic properties of mtDNA compared to genomic DNA, and the resulting inability of these CNV analysis programs to measure and normalize read depth for this part of the genome. In a previous study, a notably higher coverage of 1000X was required for the CNV detection with mtDNA heteroplasmy (Kerkhof et al., 2017).

Two of the CNV calls detected from MPS data would not have surpassed the threshold for detection on array CGH. They were verified manually with inspection of the regions expected to be involved in CNVs. These were the deletion on *MYL5* sized 2.1 kb and deletion on *CMYA5* sized 2.6 kb. This can be explained by the difference in potential between the two methods to

detect small CNVs (Trost et al., 2018). Another observation from the verifications with array CGH was that two of the CNV detections with the largest exon count divergences compared to the verified region were both duplications in genes with repeat regions involved, *TTN* and *NEB*. The explaining factors could be both that duplications are more difficult and less accurate to detect from sequencing data, and the properties of the specific repeat regions involved, as will be discussed further. One *DMD* detection showed the highest difference in base pairs, which could be explained by the large introns of *DMD*.

Approximately half of the rare *NEB* TRI region CNV detections selected for verification produced a too ambiguous result on array CGH to enable verification of the actual TRI copy number. Overall, the *NEB* TRI copy number analysis independently from MPS data did not work. For the differentiation algorithm prerequisite there were possibly not enough differences between the triplicate blocks for distinguishing them. The alignment of reads could also be more biased than for *SMN1* and *SMN2*, for which the algorithm seemed to work. In fact, no single one location has a different nucleotide in all the repeats between the repeat blocks. Therefore, attempts to differentiate could provide statistically less reliable results, even if the reads were originally aligned correctly. Similarly, the algorithmic approach was unsuitable for *TTN* even for a trial, because the exons of the replication blocks lacked enough differentiating nucleotides.

6.2 Clinical interpretation

The predictive model and filtering for rare CNVs lowered the workload for the clinical significance interpretation step notably, most clearly for the WES samples. Comparison to the in-house CNV database was also essential for evaluating the putative clinical significance. Some CNV detections could be excluded by having a too high frequency in the in-house database, even if they were not included in the common CNV databases. CNVs shared exclusively by the affected patients in the same family were straightforward to detect with the in-house CNV database. Extensive clinical phenotype comparison of similarities between patients revealed one group with an identical rare CNV detection, which was designated as likely pathogenic. On the other hand, completely different clinical presentations excluded some CNVs as unlikely clinically significant for those patients. However, such CNVs could be pathogenic for a recessive mode of inheritance of other diseases, as listed in the results for some cases. These CNV detections will remain in the in-house CNV database for further significance evaluation and frequency calculations, which has been unattainable from the common CNV population databases.

Analyzing CNV database matches with different overlap requirements was more informative compared to opting for only one overlap degree, both in the in-house CNV database and in population CNV databases. This usually allowed straightforward exclusion of CNV detections on genes with multiple (polymorphic) detections, since high frequency in overlap was visible in all categories. Completely unique CNVs were revealed with the 1 bp overlap requirement.

Utilizing the 90% overlap requirement rather than 100% revealed the detections, which were the same on genomic level, but with slight differences in the detected breakpoints. This originates from the inherent lack of breakpoint resolution of the read depth methods for CNV analysis from MPS data. Additionally, the target design was modified for different batches and this could affect breakpoint accuracy as well. The 50% overlap category enabled the detection of a novel large *NEB* CNV as compared to the previously detected large *NEB* deletion in one of the positive control samples: the 90% reciprocal overlap requirement would have been too strict since the two large *NEB* deletions had a large size difference despite being on the same region.

Most of the patients in our cohort were Finnish. According to recommendations, comparisons to population frequency databases should be preferentially ethnically matched (Richards et al., 2015). However, the CNV database from gnomAD does not contain CNV findings in the Finnish population separately, and the category of "Other" includes less than 200 samples (Collins et al., 2020). A *YARS2* deletion seen in a few of the MYOcap and WES samples was considered rare according to our in-house database, and it was not found at all in gnomAD or many of the other databases but had a frequency of 0.03% with 90% reciprocal overlap in the ExAC-FIN CNV set. Similarly, a *MYOM1* deletion seen in 12 MYOcap sequenced samples had a frequency of 0.1% with 90% reciprocal overlap in ExAC-FIN. Therefore, the most comprehensive CNV database so far, gnomAD, was not as useful as ExAC for filtering CNV detection results for frequency in a cohort with mainly Finnish patients. That is also why an internal database was used, which is generally useful for populations with a separate genetic makeup. This neglect for special genetic characteristics of some populations or subpopulations in databases is a commonly recognized problem (Sirugo et al., 2019). In the new recommendations for CNV interpretation, CNVs without high enough frequency to be considered polymorphisms (> 1%) but observed frequently in the general population can be categorized either as VUS or likely benign, which is a confusing contradiction (Riggs et al., 2019). For now, these detections were considered VUS and are awaiting further inspection for frequency in the in-house database as more samples are sequenced.

Multiple CNVs were detected in the same region or in the same gene in different samples without apparent effects on phenotype. Many of these could also be found in general databases. Although all the utilized programs have been developed to detect rare variation, these were considered to be polymorphisms. These CNV detections were probably not filtered out by the programs because they were not observed on every sample in the batch. Many of them were seen in the same gene but with different breakpoints. In the WES data, many of these polymorphic genes are already well known from previous large-scale CNV sequencing studies such as the amylase alpha 1a and alpha 2a locus (*AMY1A*, *AMY2A*) (Iafrate et al., 2004). This also applies to some rapidly evolving gene families such as *LRRC37*, *GOLGA* and *NBPF* (Alkan et al., 2009), and gene families with variable amino-acid stretches such as *ZNF, NBPF*, and mucins (Audano et al., 2019). The detection results could be directly filtered to exclude these recurrent events in the future, especially in genes with no probable relations to

neuromuscular disorders. However, some of the detections in these genes were rare according to database frequencies (< 1%), which made their effect more difficult to decipher. This could partially be a problem in the accuracy of read alignment and therefore these could be false positive detections, as with the *SMN1* and *SMN2* detections or the CNVs on *TTN* TRI and *NEB* TRI regions. False positive CNV detections have also been recently discovered to be enriched on segmental duplications or other repetitive regions (Rajagopalan et al., 2020).

TTN TRI region CNV detections were overrepresented in MYOcap compared to the *NEB* TRI region CNV detections, although the regions are genomically highly similar with locally repeating groups of blocks. Grouping of the few rarer *TTN* TRI region CNV detections did not reveal any clinical phenotype similarities between the patients, and thus it is unlikely that these CNVs have clinical significance for these patients in the cohort. The *TTN* TRI region may be more polymorphic with benign variation than the *NEB* TRI region. It is also possible that the reference genome is not correctly constructed in this region, which leads to alignment issues, providing an error source already at the data pre-preparation stage (Zenagui et al., 2018). Therefore, many of these *TTN* TRI CNV detections could be false positive. The CNV detection with the highest inaccuracy on exon-level compared to the verified region in array CGH was a CNV involving the *TTN* TRI region. This highly unspecific detection from the MPS data strengthens the hypothesis that the original problem could be caused by inaccurate initial read alignment and reference genome. The same could partially explain another observed large difference with a *NEB* duplication involving partially the *NEB* TRI region. We will move on to use the GRCh38 version soon as is the general trend and appropriate. This version could have a more accurate *TTN* TRI region and therefore decrease the amount of false positive detections.

6.3 CNV detection in a diagnostic setting

A final diagnosis was successfully achieved for several patients in this study. In many of the cases with affected family members included in the study, the clinically significant CNV was detected in a relatively straightforward manner by comparison within the in-house CNV database. This included the siblings with a homozygous partial *SGCD* deletion, and the family with a partial *CACNA1A* deletion. The CNV detection was conclusively validated by verifying correct segregation in the family members. Many of the other cases had recurrent, well-known and comprehensively documented causative CNVs. This category included the duplications and deletions of the whole *PMP22* gene detected in multiple patients in MNDcap sequenced samples and causing CMT1A and HNPP, respectively. Several *DMD* deletions and one duplication were detected in MYOcap samples, in addition to the ones detected in the positive control samples. All of the hemizygous *DMD* deletions were detected with exon-level accuracy, which is essential for achieving the correct diagnosis in dystrophinopathies, as was notified in the new ACMG CNV interpretation recommendations (Riggs et al., 2019). According to those same recommendations, CNVs with well-documented clinical features and significance replicated across multiple independent studies are classified pathogenic. CNVs producing a full extra copy of a gene are often benign unless pre-associated with a triplosensitive disorder as

with *PMP22* (Riggs et al., 2019). Therefore, many of these cases would have been classified pathogenic also according to the new CNV interpretation workflow. In some previous studies, only whole gene deletions and duplications have been screened from MPS data because of their significance in neuromuscular disorders (Antoniadi et al., 2015). However, according to our results, this can exclude clinically significant CNVs of smaller scale in many patients.

One interesting "double-trouble" case with two genetic diseases caused by pathogenic variants in two different genes was identified. Pathogenic variants in more than one gene leading to a complex phenotype have been detected also in other cohorts with neuromuscular disorders (Antoniadi et al., 2015; Hiraide et al., 2019). Since these events have been rarely described, the different frequencies of such events between disease cohorts have not been comprehensively evaluated. However, when considering the number of genes involved and how heterogeneous neuromuscular disorders are, it could be hypothesised that these "double-trouble" patient cases exist more among cohorts of neuromuscular disorders (Posey et al., 2017). Because neuromuscular phenotypes are highly heterogeneous, these cases may be wrongly designated as new diseases or as extended phenotypes of described diseases when the second genetic defect is missed (Antoniadi et al., 2015). In some other disease cohorts, CNVs have had a significant role in correcting incomplete diagnoses "made too early" (Posey et al., 2017). The increasing documentation of such cases from MPS studies should stress the importance of evaluating all genetic findings in a patient, instead of just relying on the first pathogenic variant.

CNVs are a known pathogenic mechanism in the gene *SPAST*, and statistically half of the CNVs detected in *SPAST* span beyond the gene (Boone et al., 2014). Since our targeted gene panel does not encompass the genes flanking *SPAST*, we cannot be certain whether the four CNVs detected in *SPAST* in this study encompass more than the gene, but this is possible since none of the CNVs were clearly only intragenic. The genes flanking *SPAST* beyond the 5' breakpoint are inverted in orientation compared to *SPAST*, but several directly oriented genes are located after the 3' breakpoint. Thus, deletions extending beyond the 3' end can potentially lead to formation of fusion genes (Boone et al., 2014). Other approaches such as RNA-sequencing would be needed to reveal fusion transcripts.

For most of the rare (and presumed pathogenic) CNVs detected, the ACMG CNV interpretation workflow would have given a score corresponding to VUS. The main determining factor in the annotation of CNVs as pathogenic seemed to be the availability of a validated triplosensitivity/haploinsufficiency (TS/HI) prediction score from the ClinGen Dosage Sensitivity map catalog, and the inclusion of familial samples. However, the availability of the TS/HI prediction had more weight, and also in the workflow it is stated that CNVs in genes with no knowledge on dosage sensitivity are mostly likely VUS (Riggs et al., 2019). The list of genes with TS/HI information is updated daily (Riggs et al., 2019). According to the new interpretation workflow, CNVs with unclear effect on the reading frame in a gene are also most likely VUS (Riggs et al., 2019). Comprehensive maps for the effects of all possible CNVs on the reading frame are missing for most genes. Such a map is available for *DMD* with the

UMD/TREAT-NMD DMD database (Bladen et al., 2015). Evaluation of a CNV effect to this extent on the functional level would require additional studies, such as mRNA sequencing or protein studies, which are not feasible in a routine diagnostic setting (Giugliano et al., 2018), and require expert centres for full evaluation. For the novel large intragenic *NEB* deletion detected in this study, this was possible, and protein studies by Western Blotting revealed a truncated product. These results await further inspection and verification of exact clinical significance in collaboration with other groups studying the patient case.

The new CNV interpretation recommendations have been mostly designed for CNVs in genes causing dominant disorders with distinct phenotypes, which is less usual for the heterogeneous NMDs. In conclusion, both the HI/TS database and these recommendations represent our current knowledge on CNVs and their effects on the genomic level. The recommendations will be probably updated to consider more widely different types of disorder groups in the future. For example, if a patient has a phenotype consistent with the described phenotype for the specific observed CNV, this is designated to have supporting pathogenicity evidence, but no comprehensive databases for disease causing CNVs exist yet (Riggs et al., 2019).

Some of the CNVs designated as VUS in this study have been detected in a few patients with similar phenotypes, but their clinical significance could not be confirmed. According to general instructions, variants in multiple unrelated patients with the same phenotype and absence in controls is only a moderate level of evidence for pathogenicity (Richards et al., 2015). Segregation studies are planned for some of these cases to possibly verify the significance of these CNVs. In the few single cases with a unique CNV detection and a more definite gene-phenotype connection, the causal relationship of the CNV had more confidence. In the remaining cases without possibility for segregation studies or too few affected patients, the final results have to await second-tier tests. These include mRNA studies, which could reveal the effect of the CNV on the transcript level (Cummings et al., 2017).

The separate algorithm developed (and based on algorithm in (Feng et al., 2017)) for detecting *SMN1/SMN2* copy numbers as a part of this study and CNV detection pipeline provided independent detection results in our MPS setting. The algorithm had a high accuracy and specificity, up to 100% in high-quality samples. However, only two positive control samples were tested indicating that this validation is yet inconclusive. In five of the newly detected samples with *SMN1* copy number 0 and *SMN2* copy number 3 or 4, a *SMN1* exon 7 conversion to *SMN2* seems to have occurred. This detection result and its clinical significance has yet to be validated. The five patients have a similar phenotype, but an additional validation method would be needed. Different possibilities will be discussed together with methods suggested for *NEB* TRI and *TTN* TRI region copy number validation.

The increase in the diagnostic yield was 1.9%. for both targeted gene panel and WES cohorts in this study. This is well in line with previous studies with 1.6–2% increase in diagnostic yield reported with CNV analysis from WES data in cohorts with various disorders (Pfundt et al.,

2017; Marchuk et al., 2018). In a partially matching targeted gene panel sequenced patient cohort of mostly CMT types I and II and distal hereditary motor neuropathy, 1% of the diagnoses involved CNVs (Bacquet et al., 2018). As in our setting, MPS has been the last method used on a lengthy diagnostic route in other studies as well, which may explain the similarity in the diagnostic yield (Pfundt et al., 2017; Marchuk et al., 2018). This may be the reason for the difference compared to the general estimation of 10% of disorders being explained by CNVs (Truty et al., 2019). However, the comparison of diagnostic yield between studies is not straightforward. This result can be calculated in our study as 1.9% or 2.7% depending on whether the whole cohort or just the originally unsolved cases are taken into account, which is also the case with the presented studies.

In 0.6–4.1% of our targeted gene panel sequenced patient cases CNVs designated as VUS were found, all in known disease genes or potential candidate genes. This is somewhat lower but similar compared to previous studies with > 3% of such findings in the studied cohort (Marchuk et al., 2018). A notably higher percentage (34%) of patients in WES were designated with a rare CNV VUS finding. These were designated based on different criteria with prioritization of genes with putatively relevant gene function and expression in skeletal muscle for the myopathic patients, but these genes have not yet been associated with NMDs. Verifying the clinical significance of these findings would require comprehensive additional work such as functional studies and modeling in animals (Nigro and Piluso, 2012).

In the sIBM patient group, no putative monogenetic causes were found. One unlikely reason for this could be that if all the samples share the same variant, the finding could be filtered from the results, which is a previously discussed problem (Yao et al., 2017). ExomeDepth chooses controls arbitrarily based on a close match with average read depth, and in a previous study it was noticed that if only one sample was used as a control and that included a rare CNV, then that CNV was not detected in the test samples (Povysil et al., 2017). Overall, fewer tools are available for detecting common CNVs from WES data (Zhao et al., 2013), but a variant like this would probably already have been detected in previous studies for sIBM cohorts (Needham and Mastaglia, 2016). It is hypothesized that sIBM could be a multi-factorial disease (Needham and Mastaglia, 2016). The approach of evaluating CNV detections from sIBM patients against CNVs from other patients theoretically resembles a GWAS study, but not nearly enough samples were included for enough statistical power. Some of the seemingly enriched variation was just attributable to different WES target sets.

6.4 Concluding remarks and future prospects

WES includes more potentially significant candidate genes. Accordingly, many more CNVs designated as VUS were detected from WES samples compared to MNDcap or MYOcap sequenced samples. As discussed previously, evaluating the effects for CNVs according to genomic information is less straightforward compared to SNVs and indels. Neuromuscular disorders are individually rare, difficult to diagnose accurately and genetically very

heterogeneous. Finding patients without a familial disease having the same phenotype and with the same genetic defect to increase certainty for variant pathogenicity or to designate a new disease gene is thus challenging (Charng et al., 2016; Gorokhova et al., 2015). In our cohort, many of the samples in WES represented single cases, or siblings. Inclusion of parents or additional family members has increased the diagnostic yield in WES and WGS MPS studies; the traditional trio WES can reduce the number of candidate variants around 10-fold compared to single cases (Volk and Kubisch, 2017). In a previous study with sporadic cases, the diagnostic rate was 18%, and in families with recurrence 26% (Hartley et al., 2018). Currently, we are awaiting information on additional samples from relatives in many patient cases. The diagnostic route is thus moving forward for these patients and will hopefully end with a final genetic diagnosis. Functional studies, RNA sequencing or studies in animal models could be attempted in some other cases to evaluate the effects of the VUS CNVs.

For a more efficient CNV detection filtration the gene level annotation could be improved, which would be especially important for novel CNV detections in possible candidate genes. The more recent workflows for structural variant annotation have intersected the affected genes with known pathogenic small sequence changes (Neerman et al., 2019). Our pipeline did not automatically integrate information on SNVs and indels detected from the patient sequencing data. Combining these with the CNV findings, as in some other studies and tools, would streamline the pipeline and prevent missing possible compound heterozygosity cases more effectively (Neerman et al., 2019; Geoffroy et al., 2018).

In this study, the breakpoints were resolved only on exon level for most of the pathogenic CNVs, which was enough to clarify their clinical significance and for reporting (Riggs et al., 2019). But if recommendations change, their reporting may require solving the exact breakpoints. Detecting the exact breakpoints would also provide additional information for evaluating the clinical significance, but for now this requires complementary methods such as PCR and Sanger sequencing (Giugliano et al., 2018). Manual checking in IGV could be utilized more routinely to detect read coverage cutoffs representing breakpoints. However, this would probably not provide relevant results in most of the cases, since the breakpoints are most often in introns and thus not covered in WES or targeted gene panel sequencing data (Zenagui et al., 2018). One *TTN* deletion was an exception (variant of unknown significance for now) as the intronic regions were included in the target set for that MYOcap batch. Adding intronic sequences into a capture kit in order to increase accuracy on breakpoint resolution is currently hardly ever used. According to our result from this one sample, the improvement in breakpoint resolution is promising. At the same time, detections from *TTN* TRI region increased exponentially, and it remains to be resolved whether these are true and more accurate calls, or just false positives for an unknown reason.

For the recommended submission of validated CNVs (Richards et al., 2015) to most databases such as ClinVar, the exact breakpoints need to be resolved. However, the current lack of disease specific databases for CNVs should also be overcome. With the increasing amount of MPS

studies and CNV detections in NMDs, such a database would clearly increase our knowledge on CNV prevalence and clinical significance in NMDs, even without breakpoint information. For NMDs, the collaboration of TREAT-NMD (Bushby et al., 2009) has been established to provide patient registries and biobanks, shared tools and expertise, and standardized mutation data among other resources. One major outcome has been the TREAT-NMD DMD Global database (Bladen et al., 2015) described and utilized in this study, and hopefully it is just the first of many.

Some patients had no explaining genetic findings despite excessive search for CNVs, SNVs and indels. An apparent "second hit" seems to be missing for some of them. These could also be coincidental carriers for one heterozygous recessive disease variant, and the phenotype may be due to a defect in another still unknown gene. Searching for another causative variant in another gene is unintuitive, but some diseases with digenic inheritance have been recognized also in neuromuscular disorders (Lee, Y. et al., 2018). However, recognizing novel digenic disease mechanisms and verification of pathogenicity for the variants would likely require larger patient cohorts and functional studies. The tools for evaluating sequencing data for this disease model are still under development (Renaux et al., 2019). The disease mechanism could also be non-genetic, and some diseases may have a multifactorial etiology, as has been proposed for sIBM.

WGS has not been performed for most of the unsolved cases in our cohort. The genetic cause could thus well be something intronic, in the non-coding sequences, or in regions less covered by traditional short-read WES. Repeat expansions cannot either be detected reliably from WES or targeted gene panel data with current methods. WGS could also provide better breakpoint resolution. Along with WGS data, the CNV annotation step would need to be upgraded to take into account the new information sources and possible new disease mechanisms, such as TAD rearrangements. For example, TAD and promoter location information are already being included in the newest annotation tools designed for structural variants detected from WGS data (Geoffroy et al., 2018).

Most research groups will probably move on to WGS at some point. So far, we have sequenced only a handful of samples (< 10) with WGS which is too few to attempt validation of CNV analysis. However, using WGS data is not always straightforward; in one study a high 30X coverage, evaluation of split reads, read pairs and read depth all together were required for the identification of structural variants to base pair resolution (Neerman et al., 2019). Thus, a combination of read depth and read pair tools, either separately or with a script which inherently utilizes both approaches, could probably be utilized. Using more than one tool has also been recommended for WGS data in recent studies (Trost et al., 2018; Collins et al., 2020). According to the latest tool comparison papers, LUMPY and Manta seem to be a promising combination (Regier et al., 2018). Delly and especially LUMPY have been popular and trusted for CNV calling from WGS, and the latter was suggested to be included in a future best practices

pipeline for WGS data (Regier et al., 2018; Zhang, L. et al., 2019; Abel et al., 2018; Dixon et al., 2018).

Additional tools would be needed to detect some variation types, such as repeat expansions. STR expansions in different genes in ataxia patients have been already detected from WGS data (Dashnow et al., 2018; Marelli et al., 2016). As described above, adjustments in the sequencing protocol and analysis tools would be probably needed to enable detection of mtDNA CNVs associated with some NMDs. New tools and approaches are constantly developed, and at some point, a recommended good practices pipeline will be generated. The first comprehensive databases for variants detected from WGS have been just released (Collins et al., 2020). However, they are still short in reporting CNVs, as was shown with the ExAC versus gnomAD comparison.

With lower sequencing costs, the next approach is likely to start utilizing WES and extract MYOcap and MNDcap targets as virtual gene panels from WES data to avoid increased workload and incidental findings. CNVs could be theoretically analyzed also in the virtual gene panel setting with this program combination as they performed well separately both for targeted gene panel sequenced samples and WES samples. In our setting of Mendelian neuromuscular diseases, we do not expect mosaicism to be a main genetic disease-causing mechanism, but some unsolved cases may nevertheless have mosaic variants and detecting them may become challenging with lower read depth. Even in the (simulated) MYOcap samples, the limit for detecting CNVs in mosaic quantities was rather low, especially for duplications, an observation to keep in mind in unsolved cases.

WES samples were evaluated with the same predictive model as was trained for the targeted gene panel sequencing data. The threshold for prediction had to be decreased to account for lower CNV detection scores from WES with lower average read depth. The model could potentially perform with higher accuracy if it was trained separately with *in silico* CNV generation into WES samples. As an example, a tool to detect VNTR expansions had to be trained with simulated reads for the two used sequencing data types separately: short-read and PacBio (Bakhtiari et al., 2018). This could be challenging with the approach we used, since WES samples have notably more CNV detections. Our WES sequencing data were produced with more variable target sets, sequencing platforms and technical specifications than the different MYOcap and MNDcap sequencing batches. Therefore, the sequencing platform and provider would have to be conformed first, because training a model for one setting may not work identically in a different setting. In one study with tiling probes by Agilent, overlapping probes by NimbleGen, and gapped probes by Illumina, inconsistent coverage was revealed between batches, which prevented combining them for CNV calling (Wang, Q. et al., 2017). Although high sensitivity and specificity were achieved with the current predictive model using logistic regression, also other statistical approaches such as random forest might be valuable (Pounraja et al., 2019).

We developed a promising method for differentiating copy numbers of *SMN1* and *SMN2* from MPS data, but the CNVs on *NEB* TRI and especially *TTN* TRI region remained elusive with the same approach. Additionally, some of the *NEB* TRI CNV detections remained unresolved after array CGH. An approach with pseudoreference genome could be possible to force alignment into correct regions. One approach by Dashnow and colleagues for detecting STR expansions involves utilization of specifically designed decoy reference genome stretches (Dashnow et al., 2018). These decoy chromosomes cover all possible versions of 1–6 bp repeat blocks and thus enable mapping of even novel repeat expansion alleles. This provides a more accurate initial setting for detecting variable repeat unit lengths (Dashnow et al., 2018). In another study, a pseudoreference was built for the *TTN* triplicate region, forcing reads to align to one repeat block, and hexaploidy variant calling revealed a significant SNV in the TRI region (Cummings et al., 2017). A homologous region was resolved with a four-allele normalization method taking into account the number of pseudogene copies (Kerkhof et al., 2017). However, MLPA or PCR were needed to determine whether the variant was located in the gene or in a pseudogene. These approaches of forcing all the reads from the repetitive regions into one location would probably produce data impossible to evaluate with the current CNV analysis tools. Most likely, an additional algorithm and comparison to normal samples would be needed, as with the *SMN1/SMN2* algorithm. However, finding enough subjects in each sequencing batch with reliably normal copy number count in the *TTN* TRI and *NEB* TRI region could be challenging. WGS and long-read sequencing may eventually clarify some of these regions with frequent probably false positive CNV detections by allowing more accurate initial read alignment for CNV analysis and enabling multiple algorithmic approaches.

Among the complementary verification methods for unspecific CNV detections from MPS data, ddPCR has been tested. This method could be used to resolve pseudogenes with extremely high similarity (*STRC* has a > 99% similar pseudogene) (Amr et al., 2018). The regions were differentiated by inclusion of unique nucleotides into probe design with TaqMan chemistry (Amr et al., 2018). This is in theory a similar approach to our successful *SMN1*/*SMN2* differentiation algorithm but would be more challenging in the *TTN* TRI and *NEB* TRI regions with less clearly differentiating nucleotides. Probably, both exons and introns should be targeted to increase accuracy (Shen and Wu, 2009). One possibility for inspecting high homology regions is to design longer capture probes for more accurate hybridization-based target enrichment (Gulilat et al., 2019). Long-range sequencing would also work in this verification, but it is still not scalable, simple or cost-effective (Amr et al., 2018). A method with ddPCR to differentiate both *NEB* TRI and *TTN* TRI is under development in a collaborative group, and the samples with putatively interesting findings will be submitted for this analysis to gain further insight.

Although the *NEB* TRI CNV detections from MPS data were inaccurate, they were successfully verified to be diverging and significant in some cases. Therefore, screening from MPS data could be used to select additional confirmatory testing only for possibly positive detections. In some cases, CNV deviations of more than one in block number were clinically non-significant.

The likely explaining mechanism is that the copy number changes occurred in repeat blocks on different alleles, resulting in a one-copy change on each, which apparently is not deleterious (Kiiski, K. et al., 2016), but this was not verified with an additional method. Array CGH cannot differentiate the amount of the copies for the alleles, so ddPCR would also bring more insight into this (Kiiski, K. et al., 2016). Only MYOcap sequencing provided the material for *NEB* TRI comparison. It provides higher average read depth and stronger signal compared to WES samples. The in-house CNV database was large enough with MYOcap samples to compare CNV detections and to detect divergent *NEB* TRI changes. These can now be compared with future CNV detections since *NEB* TRI CNVs are important in NMD cohorts, as was shown for the one patient with a clinically significant *NEB* TRI 8/6 finding.

7 CONCLUSIONS

In this study, the diagnostic bioinformatics workflow for targeted gene panel and WES sequencing data for CNV detection has been optimized.

Objective I

As proven in this study, the setup for accurate CNV detection requires the combined use of more than one program for CNV analysis. This was validated with CNV positive and negative control samples and by verification of newly detected CNVs and *in silico* generated CNVs, as was the aim of the objective I.

Objective II

Besides the combined CNV detections, the optimization of the detection pipeline requires thorough evaluation of the results for true positive detections using a statistical model in a standardized manner to avoid bias and to increase accuracy. As was the aim with the objective II, a predictive logistic regression model was successfully developed and validated to differentiate detections with true positive prediction for this purpose. The detection sensitivity was high for CNVs of more than three exons, and even one exon CNVs could be detected with high enough accuracy. The predictive model also decreased the workload in CNV annotation.

Objective III

CNV annotation with cnvScan was improved with the inclusion of the recently published population CNV databases and other resources, and with different scopes for comparing CNVs to databases. With a frequency filtration in our in-house CNV database and population databases, the number of variants left for the clinical significance evaluation was further decreased. This simplified the diagnostic process, as was the aim of the object III.

Objective IV

The CNV detection and annotation pipeline was also validated for WES samples with slight modifications to reach adequate detection accuracy, as was the aim of the objective IV.

The eventual diagnostic yield for both sequencing data types corresponded well to results from previous studies. Additionally, regions of moderate homology were successfully differentiated from the MPS data with an additional script used for *SMN1* and *SMN2.* However, in order to achieve the ultimate aim: to accept CNV detections without additional verification, the current method needs more validation with true samples with CNVs. A separate method was also needed for the validation of CNV detections from the *NEB* TRI region. Nevertheless, true variations in the copy number of the repeats in the *NEB* TRI region could be validated based on initial screening from the MPS data. Therefore, the validated *NEB* TRI detections in these samples could be used as background data when screening short-read sequencing data for novel detections in this region in the future. However, to increase the total diagnostic yield of CNV analysis, accuracy of CNV detections and to expand the analysis to include structural variants, additional sequencing data sources and methods will be required, such as WGS, long-read sequencing, RNA sequencing, and ddPCR.

8 ACKNOWLEDGEMENTS

This study was conducted for the most part in the group for Neuromuscular Disorders (Prof Udd/Hackman) at Folkhälsan Research Center, the Department of Medical and Clinical Genetics, University of Helsinki, over the years 2017-2020. Professor Anna-Elina Lehesjoki, the current head of the Folkhälsan Research Center is acknowledged for providing excellent resources and facilities and pleasant working environments for research. Part of this research was conducted at Blueprint Genetics, and the Chief Strategy Officer Samuel Myllykangas is thanked for offering and organizing this collaboration and the resources.

I would like to thank the Faculty of Biological and Environmental Sciences, University of Helsinki, for the opportunity to carry out my postgraduate studies, and Professor Juha Partanen for acting as the Custos at my Thesis Defense on such a short notice after the change of schedule. I would like to extend my thanks to the Doctoral School in Health Sciences and the Doctoral Programme in Integrative Life Science for the excellent graduate student education, and especially Liisa Uotila for guidance in studies and Mari Siltala for aid in the practicalities of the last steps of my thesis project. Thanks to Katarina Pelin for guidance in wrapping up my studies and pointing to the correct path to take at the onset for the Clinical Laboratory Geneticist degree. I wish to thank the following foundations and organizations for funding this Doctoral thesis project: The Folkhälsan Research foundation, The Sigrid Jusélius foundation, The Alfred Kordelin foundation, The Jane and Aatos Erkko foundation and the Finska Läkaresällskapet. In addition, I am grateful for the financial support received for participation in international congresses from the University of Helsinki Funds and the World Muscle Society.

I am grateful to the pre-examiners, Professor Matti Nykter and Docent Csilla Sipeky for accepting the task and taking the time to evaluate my thesis. They are thanked for their expertise in reviewing the thesis manuscript and providing me with valuable comments that further strengthened this thesis. I want to thank the external members of my thesis committee, Samuel Myllykangas and Mari Kaunisto, for examining my thesis work over the years and providing constructive feedback and advice. I want to again thank Samuel for the internship at Blueprint Genetics, which was an important step in my PhD project.

My deepest gratitudes are extended to my supervisors Peter Hackman and Bjarne Udd. I admire Bjarne for his profound knowledge in neuromuscular diseases and research, and interest in new methods. I wish to thank Bjarne also as the group leader for offering me this project first as a Master's thesis project and then for allowing me to delve further. Thank you for offering rewarding and educative research environment with independence and responsibility for one's own work. I want to thank Peter as the administrative group leader, for the collaborative connections, and for welcoming me to the group first as my Master's thesis supervisor. Thank you, Peter, for always finding time to discuss results, manuscripts and other matters related to PhD work. I want to thank my supervisors for having confidence in my abilities and providing focus to my project by pulling together all the loose threads. I have been fortunate to have an additional supervisor: I am grateful to Marco Savarese for the everyday advice with this work,

for teaching me how to think critically as a scientist and for sharing my ideas. I appreciate Marco for his vast knowledge on genetics and molecular mechanisms and the field of muscle research, which has greatly helped and inspired me throughout these years and projects. I want to thank Bjarne, Peter and Marco for careful revision of this thesis for language and contents and for their suggestions to improve the flow.

I want to express my gratitude to the honored thesis opponent Professor Pawel Stankiewicz for making my dissertation possible and bearing with us through these turbulent times and change of schedule.

I thank all my past and present co-workers in the Neuromuscular disorders research group for welcoming me and for the pleasant working environment, in addition to my supervisors especially Per Harald Jonson, Jaakko Sarparanta, Anna Naukkarinen, Mridul Johari, Meharji Arumilli, Sampo Koivunen, Merja Soininen and Helena Luque. I want to thank Mridul for energetic and encouraging personality and important impact particularly in the early stages of my project. Thank you Sampo for working out some of the problems of bioinformatics and programming with me and for other immersive discussions (with topics similarly beyond comprehension for many). Thanks to Merja for your friendship and shared breaks with pleasant conversations. I want to thank you for your precise and trustworthy work in keeping the samples and everything else organized. From the Tampere branch of our group, I am grateful for Sini Penttilä, Sara Lehtinen, Tiina Suominen, Johanna Palmio and Manu Jokela for the fun company during WMS and other conference trips and meetings, especially the summer cottage meetings. Specifically, I want to thank you for tackling the specifics of our patient cohort and for your aid in sample logistics.

All the co-authors of the original publications outside our groups are sincerely thanked for their invaluable work and contributions to the projects. I thank all collaborators without whom this work would not have been possible: group leaders Carina Wallgren-Petterson and Katarina Pelin, and Vincenzo Nigro for extending openly their resources for collaboration. I thank Vilma-Lotta Lehtokari, Lydia Sagath, Kirsi Kiiski and Marilotta Turunen for their aid in CNV validations by performing complementary method of array CGH, and Teresa Giugliano by MLPA. Thanks also to Niina Sandholm for ideas and consultancy in statistics. I want to thank Mikko Muona for guidance and expertise at Blueprint Genetics facilities and resources. Finally, I would like to thank all the clinicians, families and patients who have participated in the studies.

Thanks to Marjatta Valkama, Jaana Welin-Haapamäki, Åsa Rehn, Nina Forss and Sebastian Oey for efficiently running the administrative and financial department at the Folkhälsan Research Center. Thanks to Ann-Liz Träskelin for lab management and everlasting interest in my work and other projects with enthusiastic and cheerful attitude.

I want to thank all the Folkhälsan Research Center colleagues I got to share the room with for the most of my stay: Merja, Marilotta, Mira Aronen, Paula Hakala, Sampo, Reetta-Stiina Järvinen, Helena, and also Johanna Lehtonen from the new room with Marco, Mridul, and Jaakko. With the bioinformatics emphasis in my thesis project, I have spent most of my time in

those rooms and got to share many (non)scientific discussions, laughs and frustrations with you in and outside the office. I want to thank all of the FHRC personnel for providing such an open and friendly working environment and all the support and help when I needed it at anything. Coming to work has always been a pleasure, and now a bit sad with the emptiness of the corridors. I want to sincerely thank Peter and the administration of FHRC for ensuring that my thesis work and the work at FHRC has endured and continued safely through this global pandemic of COVID-19 corona virus (the reason for the changes of schedules). I only wish I could have spent more time with the awesome people FHRC houses during these last stages of my work. I am grateful to have been able to start my career in human genetics at FHRC.

I want to thank also my new work community at the Department of HUSLAB Genetics. Thank you for having me as a Clinical Laboratory Geneticist trainee and a part of the team and making me feel so welcome and at home. For this I want to thank especially Kaisa Kettunen, Reetta Vainionpää and Soili Kytölä. And thank you Soili and Anna-Kaisa Anttonen for giving me the opportunity to finish my thesis. I want to mutually thank Peter and Bjarne for flexibility and for allowing me to already take this next step forward while finishing the thesis.

I want to thank my family and friends for their generous support and for keeping me grounded to the life outside of work. Thanks to all my friends for providing completely something else besides the hard work, for all the get-togethers and yearly traditions. Thank you EOL choir for the arts to counterweight the science. Thank you to the many fellow students I have met and for sharing so many enriching and fun discussions and events. I want to thank especially my friends from Viikki for fueling my interest in studying genetics. I am deeply grateful to my parents Juha and Ulla for supporting me in whatever I do, to my grandparents for being supportive and interested in what I do, and for the rest of my family for everything they have done and been for me.

I express my warmest thanks to Juuso for his love and companionship, for sharing the (student) life with me during these past eight years, for understanding the ups and downs of doing research, for peer support in PhD studies, also when the work continued past midnight, (and still) for proofreading this thesis and other texts. Thank you for always being there for me and understanding, but at the same time encouraging me to be the best possible version of myself.

Helsinki, July 2020

9 REFERENCES

1000 Genomes Project Consortium, A. Auton, L.D. Brooks, R.M. Durbin, E.P. Garrison, H.M. Kang, J.O. Korbel, J.L. Marchini, S. McCarthy, G.A. McVean, and G.R. Abecasis. 2015. A global reference for human genetic variation. *Nature.* 526:68-74. doi: 10.1038/nature15393 [doi].

Abel, H.J., D.E. Larson, C. Chiang, I. Das, K.L. Kanchi, R.M. Layer, B.M. Neale, W.J. Salerno, C. Reeves, S. Buyske, et al. 2018. Mapping and characterization of structural variation in 17,795 deeply sequenced human genomes. *bioRxiv.* doi: 10.1101/508515.

Abou Tayoun, A.N., T. Pesaran, M.T. DiStefano, A. Oza, H.L. Rehm, L.G. Biesecker, S.M. Harrison, and ClinGen Sequence Variant Interpretation Working Group (ClinGen SVI). 2018. Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion. *Hum.Mutat.* 39:1517-1524. doi: 10.1002/humu.23626 [doi].

Abyzov, A., S. Li, D.R. Kim, M. Mohiyuddin, A.M. Stütz, N.F. Parrish, X.J. Mu, W. Clark, K. Chen, M. Hurles, J.O. Korbel, H.Y. Lam, et al. 2015. Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. *Nat.Commun.* 6:7256. doi: 10.1038/ncomms8256 [doi].

Aird, D., M.G. Ross, W.S. Chen, M. Danielsson, T. Fennell, C. Russ, D.B. Jaffe, C. Nusbaum, and A. Gnirke. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* 12:R18-2011-12-2-r18. Epub 2011 Feb 21. doi: 10.1186/gb-2011-12-2-r18 [doi].

Albers, C.A., D.S. Paul, H. Schulze, K. Freson, J.C. Stephens, P.A. Smethurst, J.D. Jolley, A. Cvejic, M. Kostadima, P. Bertone, et al. 2012. Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in exon-junction complex subunit RBM8A causes TAR syndrome. *Nat.Genet.* 44:435-9, S1-2. doi: 10.1038/ng.1083 [doi].

Alfares, A., T. Aloraini, L.A. Subaie, A. Alissa, A.A. Qudsi, A. Alahmad, F.A. Mutairi, A. Alswaid, A. Alothaim, W. Eyaid, et al. 2018. Whole-genome sequencing offers additional but limited clinical utility compared with reanalysis of whole-exome sequencing. *Genet.Med.* 20:1328-1333. doi: 10.1038/gim.2018.41 [doi].

Alkan, C., B.P. Coe, and E.E. Eichler. 2011. Genome structural variation discovery and genotyping. *Nat.Rev.Genet.* 12:363-376. doi: 10.1038/nrg2958 [doi].

Alkan, C., J.M. Kidd, T. Marques-Bonet, G. Aksay, F. Antonacci, F. Hormozdiari, J.O. Kitzman, C. Baker, M. Malig, O. Mutlu, et al. 2009. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat.Genet.* 41:1061-1067. doi: 10.1038/ng.437 [doi].

Alkuraya, F.S. 2015. Natural human knockouts and the era of genotype to phenotype. *Genome Med.* 7:48-015-0173-z. eCollection 2015. doi: 48.

Ambardar, S., R. Gupta, D. Trakroo, R. Lal, and J. Vakhlu. 2016. High Throughput Sequencing: An Overview of Sequencing Chemistry. *Indian J.Microbiol.* 56:394-404. doi: 10.1007/s12088-016-0606-4 [doi].

Amberger, J.S., C.A. Bocchini, F. Schiettecatte, A.F. Scott, and A. Hamosh. 2015. OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* 43:D789-98. doi: 10.1093/nar/gku1205 [doi].

Amr, S.S., E. Murphy, E. Duffy, R. Niazi, J. Balciuniene, M. Luo, H.L. Rehm, and A.N. Abou Tayoun. 2018. Allele-Specific Droplet Digital PCR Combined with a Next-Generation Sequencing-Based Algorithm for Diagnostic Copy Number Analysis in Genes with High Homology: Proof of Concept Using Stereocilin. *Clin.Chem.* 64:705-714. doi: 10.1373/clinchem.2017.280685 [doi].

Ankala, A., C. da Silva, F. Gualandi, A. Ferlini, L.J. Bean, C. Collins, A.K. Tanner, and M.R. Hegde. 2015. A comprehensive genomic approach for neuromuscular diseases gives a high diagnostic yield. *Ann.Neurol.* 77:206-214. doi: 10.1002/ana.24303 [doi].

Antoniadi, T., C. Buxton, G. Dennis, N. Forrester, D. Smith, P. Lunt, and S. Burton-Jones. 2015. Application of targeted multi-gene panel testing for the diagnosis of inherited peripheral neuropathy provides a high diagnostic yield with unexpected phenotype-genotype variability. *BMC Med.Genet.* 16:84-015-0224-8. doi: 10.1186/s12881-015-0224-8 [doi].

Audano, P.A., A. Sulovari, T.A. Graves-Lindsay, S. Cantsilieris, M. Sorensen, A.E. Welch, M.L. Dougherty, B.J. Nelson, A. Shah, S.K. Dutcher, et al. 2019. Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell.* 176:663-675.e19. doi: S0092-8674(18)31633-7 [pii].

Austin-Tse, C.A., D.L. Mandelker, A.M. Oza, H. Mason-Suares, H.L. Rehm, and S.S. Amr. 2018. Analysis of intragenic USH2A copy number variation unveils broad spectrum of unique and recurrent variants. *Eur.J.Med.Genet.* 61:621-626. doi: S1769-7212(17)30689-4 [pii].

Bacquet, J., T. Stojkovic, A. Boyer, N. Martini, F. Audic, B. Chabrol, E. Salort-Campana, E. Delmont, J.P. Desvignes, A. Verschueren, et al. 2018. Molecular diagnosis of inherited peripheral neuropathies by targeted next-generation sequencing: molecular spectrum delineation. *BMJ Open.* 8:e021632-2018-021632. doi: 10.1136/bmjopen-2018-021632 [doi].

Bailey, J.A., Z. Gu, R.A. Clark, K. Reinert, R.V. Samonte, S. Schwartz, M.D. Adams, E.W. Myers, P.W. Li, and E.E. Eichler. 2002. Recent segmental duplications in the human genome. *Science.* 297:1003-1007. doi: 10.1126/science.1072047 [doi].

Bakhtiari, M., S. Shleizer-Burko, M. Gymrek, V. Bansal, and V. Bafna. 2018. Targeted genotyping of variable number tandem repeats with adVNTR. *Genome Res.* 28:1709-1719. doi: 10.1101/gr.235119.118 [doi].

Bang, M.L., T. Centner, F. Fornoff, A.J. Geach, M. Gotthardt, M. McNabb, C.C. Witt, D. Labeit, C.C. Gregorio, H. Granzier, and S. Labeit. 2001. The complete gene sequence of titin, expression of an unusual approximately 700-kDa titin isoform, and its interaction with obscurin identify a novel Z-line to I-band linking system. *Circ.Res.* 89:1065-1072. doi: 10.1161/hh2301.100981 [doi].

Benjamini, Y., and T.P. Speed. 2012. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* 40:e72. doi: 10.1093/nar/gks001 [doi].

Bentley, D.R., S. Balasubramanian, H.P. Swerdlow, G.P. Smith, J. Milton, C.G. Brown, K.P. Hall, D.J. Evers, C.L. Barnes, H.R. Bignell, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature.* 456:53-59. doi: 10.1038/nature07517 [doi].

Bladen, C.L., D. Salgado, S. Monges, M.E. Foncuberta, K. Kekou, K. Kosma, H. Dawkins, L. Lamont, A.J. Roy, T. Chamova, et al. 2015. The TREAT-NMD DMD Global Database: analysis of more than 7,000 Duchenne muscular dystrophy mutations. *Hum.Mutat.* 36:395-402. doi: 10.1002/humu.22758 [doi].

Bonne, G., F. Rivier, and D. Hamroun. 2018. The 2019 version of the gene table of neuromuscular disorders (nuclear genome). *Neuromuscul.Disord.* 28:1031-1063. doi: S0960-8966(18)31206-9 [pii].

Boone, P.M., I.M. Campbell, B.C. Baggett, Z.T. Soens, M.M. Rao, P.M. Hixson, A. Patel, W. Bi, S.W. Cheung, S.R. Lalani, et al. 2013. Deletions of recessive disease genes: CNV contribution to carrier states and disease-causing alleles. *Genome Res.* 23:1383-1394. doi: 10.1101/gr.156075.113 [doi].

Boone, P.M., B. Yuan, I.M. Campbell, J.C. Scull, M.A. Withers, B.C. Baggett, C.R. Beck, C.J. Shaw, P. Stankiewicz, P. Moretti, W.E., et al. 2014. The Alu-rich genomic architecture of SPAST predisposes to diverse and functionally distinct disease-associated CNV alleles. *Am.J.Hum.Genet.* 95:143-161. doi: 10.1016/j.ajhg.2014.06.014 [doi].

Brook, J.D., M.E. McCurrach, H.G. Harley, A.J. Buckler, D. Church, H. Aburatani, K. Hunter, V.P. Stanton, J.P. Thirion, and T. Hudson. 1992. Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3' end of a transcript encoding a protein kinase family member. *Cell.* 68:799-808. doi: 0092-8674(92)90154-5 [pii].

Bushby, K., S. Lynn, T. Straub, and TREAT-NMD Network. 2009. Collaborating to bring new therapies to the patient--the TREAT-NMD model. *Acta Myol.* 28:12-15.

Cabrejo, L., L. Guyant-Marechal, A. Laquerriere, M. Vercelletto, F. De la Fourniere, C. Thomas-Anterion, C. Verny, F. Letournel, F. Pasquier, A. Vital, et al. 2006. Phenotype associated with APP duplication in five families. *Brain.* 129:2966-2976. doi: awl237 [pii].

Campbell, I.M., T. Gambin, S. Jhangiani, M.L. Grove, N. Veeraraghavan, D.M. Muzny, C.A. Shaw, R.A. Gibbs, E. Boerwinkle, F. Yu, and J.R. Lupski. 2016. Multiallelic Positions in the Human Genome: Challenges for Genetic Analyses. *Hum.Mutat.* 37:231-234. doi: 10.1002/humu.22944 [doi].

Carelle-Calmels, N., P. Saugier-Veber, F. Girard-Lemaire, G. Rudolf, B. Doray, E. Guerin, P. Kuhn, M. Arrive, C. Gilch, E. Schmitt, et al. 2009. Genetic compensation in a human genomic disorder. *N.Engl.J.Med.* 360:1211-1216. doi: 10.1056/NEJMoa0806544 [doi].

Carter, J.C., D.W. Sheehan, A. Prochoroff, and D.J. Birnkrant. 2018. Muscular Dystrophies. *Clin.Chest Med.* 39:377-389. doi: S0272-5231(18)30004-2 [pii].

Carvalho, B., E. Ouwerkerk, G.A. Meijer, and B. Ylstra. 2004. High resolution microarray comparative genomic hybridisation analysis using spotted oligonucleotides. *J.Clin.Pathol.* 57:644-646. doi: 10.1136/jcp.2003.013029 [doi].

Carvalho, C.M., and J.R. Lupski. 2016. Mechanisms underlying structural variant formation in genomic disorders. *Nat.Rev.Genet.* 17:224-238. doi: 10.1038/nrg.2015.25 [doi].

Carvalho, C.M., R. Pfundt, D.A. King, S.J. Lindsay, L.W. Zuccherato, M.V. Macville, P. Liu, D. Johnson, P. Stankiewicz, C.W. Brown, et al. 2015. Absence of heterozygosity due to template switching during replicative rearrangements. *Am.J.Hum.Genet.* 96:555-564. doi: 10.1016/j.ajhg.2015.01.021 [doi].

Chae, J.H., V. Vasta, A. Cho, B.C. Lim, Q. Zhang, S.H. Eun, and S.H. Hahn. 2015. Utility of next generation sequencing in genetic diagnosis of early onset neuromuscular disorders. *J.Med.Genet.* 52:208-216. doi: 10.1136/jmedgenet-2014-102819 [doi].

Chaisson, M.J., J. Huddleston, M.Y. Dennis, P.H. Sudmant, M. Malig, F. Hormozdiari, F. Antonacci, U. Surti, R. Sandstrom, M. Boitano, et al. 2015. Resolving the complexity of the human genome using single-molecule sequencing. *Nature.* 517:608-611. doi: 10.1038/nature13907 [doi].

Chaisson, M.J.P., A.D. Sanders, X. Zhao, A. Malhotra, D. Porubsky, T. Rausch, E.J. Gardner, O.L. Rodriguez, L. Guo, R.L. Collins, et al. 2019. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat.Commun.* 10:1784-018-08148-z. doi: 10.1038/s41467-018-08148-z [doi].

Chance, P.F., M.K. Alderson, K.A. Leppig, M.W. Lensch, N. Matsunami, B. Smith, P.D. Swanson, S.J. Odelberg, C.M. Disteche, and T.D. Bird. 1993. DNA deletion associated with hereditary neuropathy with liability to pressure palsies. *Cell.* 72:143-151. doi: 0092-8674(93)90058-X [pii].

Charng, W.L., E. Karaca, Z. Coban Akdemir, T. Gambin, M.M. Atik, S. Gu, J.E. Posey, S.N. Jhangiani, D.M. Muzny, H. Doddapaneni, et al. 2016. Exome sequencing in mostly consanguineous Arab families with neurologic disease provides a high potential molecular diagnosis rate. *BMC Med.Genomics.* 9:42-016-0208-3. doi: 42.

Chen, L., P. Liu, T.C. Evans Jr, and L.M. Ettwiller. 2017. DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science.* 355:752-756. doi: 10.1126/science.aai8690 [doi].

Cherf, G.M., K.R. Lieberman, H. Rashid, C.E. Lam, K. Karplus, and M. Akeson. 2012. Automated forward and reverse ratcheting of DNA in a nanopore at 5-A precision. *Nat.Biotechnol.* 30:344-348. doi: 10.1038/nbt.2147 [doi].

Church, D.M., V.A. Schneider, T. Graves, K. Auger, F. Cunningham, N. Bouk, H.C. Chen, R. Agarwala, W.M. McLaren, G.R. Ritchie, et al. 2011. Modernizing reference genome assemblies. *PLoS Biol.* 9:e1001091. doi: 10.1371/journal.pbio.1001091 [doi].

Clark, M.M., A. Hildreth, S. Batalov, Y. Ding, S. Chowdhury, K. Watkins, K. Ellsworth, B. Camp, C.I. Kint, C. Yacoubian, et al. 2019. Diagnosis of genetic diseases in seriously ill

children by rapid whole-genome sequencing and automated phenotyping and interpretation. *Sci.Transl.Med.* 11:10.1126/scitranslmed.aat6177. doi: eaat6177 [pii].

Collins, R.L., H. Brand, K.J. Karczewski, X. Zhao, J. Alföldi, L.C. Francioli, A.V. Khera, C. Lowther, L.D. Gauthier, H. Wang, et al. 2020. A structural variation reference for medical and population genetics. *Nature.* 581:444-451. doi: 10.1038/s41586-020-2287-8.

Conrad, D.F., D. Pinto, R. Redon, L. Feuk, O. Gokcumen, Y. Zhang, J. Aerts, T.D. Andrews, C. Barnes, P. Campbell, et al. 2010. Origins and functional impact of copy number variation in the human genome. *Nature.* 464:704-712. doi: 10.1038/nature08516 [doi].

Cordoba, M., S.A. Rodriguez-Quiroga, P.A. Vega, V. Salinas, J. Perez-Maturo, H. Amartino, C. Vasquez-Dusefante, N. Medina, D. Gonzalez-Moron, and M.A. Kauffman. 2018. Whole exome sequencing in neurogenetic odysseys: An effective, cost- and time-saving diagnostic approach. *PLoS One.* 13:e0191228. doi: 10.1371/journal.pone.0191228 [doi].

Costain, G., R. Jobling, S. Walker, M.S. Reuter, M. Snell, S. Bowdin, R.D. Cohn, L. Dupuis, S. Hewson, S. Mercimek-Andrews, et al. 2018. Periodic reanalysis of whole-genome sequencing data enhances the diagnostic advantage over standard clinical genetic testing. *Eur.J.Hum.Genet.* 26:740-744. doi: 10.1038/s41431-018-0114-6 [doi].

Cummings, B.B., J.L. Marshall, T. Tukiainen, M. Lek, S. Donkervoort, A.R. Foley, V. Bolduc, L.B. Waddell, S.A. Sandaradura, G.L. O'Grady, et al. 2017. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci.Transl.Med.* 9:10.1126/scitranslmed.aal5209. doi: eaal5209 [pii].

Cutrupi, A.N., M.H. Brewer, G.A. Nicholson, and M.L. Kennerson. 2018. Structural variations causing inherited peripheral neuropathies: A paradigm for understanding genomic organization, chromatin interactions, and gene dysregulation. *Mol.Genet.Genomic Med.* 6:422-433. doi: 10.1002/mgg3.390 [doi].

Dashnow, H., M. Lek, B. Phipson, A. Halman, S. Sadedin, A. Lonsdale, M. Davis, P. Lamont, J.S. Clayton, N.G. Laing, D.G. MacArthur, and A. Oshlack. 2018. STRetch: detecting and discovering pathogenic short tandem repeat expansions. *Genome Biol.* 19:121-018-1505-2. doi: 10.1186/s13059-018-1505-2 [doi].

de Leeuw, N., J.Y. Hehir-Kwa, A. Simons, A. Geurts van Kessel, D.F. Smeets, B.H. Faas, and R. Pfundt. 2011. SNP array analysis in constitutional and cancer genome diagnostics--copy number variants, genotyping and quality control. *Cytogenet.Genome Res.* 135:212-221. doi: 10.1159/000331273 [doi].

de Ligt, J., P.M. Boone, R. Pfundt, L.E. Vissers, T. Richmond, J. Geoghegan, K. O'Moore, N. de Leeuw, C. Shaw, H.G. Brunner, et al. 2013. Detection of clinically relevant copy number variants with whole-exome sequencing. *Hum.Mutat.* 34:1439-1448. doi: 10.1002/humu.22387 [doi].

DePristo, M.A., E. Banks, R. Poplin, K.V. Garimella, J.R. Maguire, C. Hartl, A.A. Philippakis, G. del Angel, M.A. Rivas, M. Hanna, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat.Genet.* 43:491-498. doi: 10.1038/ng.806 [doi].

DiVincenzo, C., C.D. Elzinga, A.C. Medeiros, I. Karbassi, J.R. Jones, M.C. Evans, C.D. Braastad, C.M. Bishop, M. Jaremko, Z. Wang, et al. 2014. The allelic spectrum of Charcot-Marie-Tooth disease in over 17,000 individuals with neuropathy. *Mol.Genet.Genomic Med.* 2:522-529. doi: 10.1002/mgg3.106 [doi].

Dixon, J.R., S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J.S. Liu, and B. Ren. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature.* 485:376-380. doi: 10.1038/nature11082 [doi].

Dixon, J.R., J. Xu, V. Dileep, Y. Zhan, F. Song, V.T. Le, G.G. Yardımcı, A. Chakraborty, D.V. Bann, Y. Wang, et al. 2018. Integrative detection and analysis of structural variation in cancer genomes. *Nat.Genet.* 50:1388-1398. doi: 10.1038/s41588-018-0195-8 [doi].

Dohrn, M.F., N. Glockle, L. Mulahasanovic, C. Heller, J. Mohr, C. Bauer, E. Riesch, A. Becker, F. Battke, K. Hortnagel, et al. 2017. Frequent genes in rare diseases: panel-based next generation sequencing to disclose causal mutations in hereditary neuropathies. *J.Neurochem.* 143:507-522. doi: 10.1111/jnc.14217 [doi].

Donner, K., M. Sandbacka, V.L. Lehtokari, C. Wallgren-Pettersson, and K. Pelin. 2004. Complete genomic structure of the human nebulin gene and identification of alternatively spliced transcripts. *Eur.J.Hum.Genet.* 12:744-751. doi: 10.1038/sj.ejhg.5201242 [doi].

Drmanac, R., A.B. Sparks, M.J. Callow, A.L. Halpern, N.L. Burns, B.G. Kermani, P. Carnevali, I. Nazarenko, G.B. Nilsen, G. Yeung, et al. 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science.* 327:78-81. doi: 10.1126/science.1181498 [doi].

Efthymiou, S., A. Manole, and H. Houlden. 2016. Next-generation sequencing in neuromuscular diseases. *Curr.Opin.Neurol.* 29:527-536. doi: 10.1097/WCO.0000000000000374 [doi].

Eid, J., A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, et al. 2009. Real-time DNA sequencing from single polymerase molecules. *Science.* 323:133-138. doi: 10.1126/science.1162986 [doi].

Eijkelenboom, A., B.B.J. Tops, A. van den Berg, A.J.C. van den Brule, W.N.M. Dinjens, H.J. Dubbink, A. Ter Elst, W.R.R. Geurts-Giele, P.J.T.A. Groenen, F.H. Groenendijk, et al. 2019. Recommendations for the clinical interpretation and reporting of copy number gains using gene panel NGS analysis in routine diagnostics. *Virchows Arch.* 474:673-680. doi: 10.1007/s00428-019-02555-3 [doi].

Ellingford, J.M., S. Barton, S. Bhaskar, S.G. Williams, P.I. Sergouniotis, J. O'Sullivan, J.A. Lamb, R. Perveen, G. Hall, W.G. Newman, et al. 2016. Whole Genome Sequencing Increases Molecular Diagnostic Yield Compared with Current Diagnostic Testing for Inherited Retinal Disease. *Ophthalmology.* 123:1143-1150. doi: S0161-6420(16)00030-0 [pii].

Ellingford, J.M., C. Campbell, S. Barton, S. Bhaskar, S. Gupta, R.L. Taylor, P.I. Sergouniotis, B. Horn, J.A. Lamb, M. Michaelides, et al. 2017. Validation of copy number variation analysis for next-generation sequencing diagnostics. *Eur.J.Hum.Genet.* 25:719-724. doi: 10.1038/ejhg.2017.42 [doi].

Evila, A., M. Arumilli, B. Udd, and P. Hackman. 2016. Targeted next-generation sequencing assay for detection of mutations in primary myopathies. *Neuromuscul.Disord.* 26:7-15. doi: 10.1016/j.nmd.2015.10.003 [doi].

Evila, A., A. Vihola, J. Sarparanta, O. Raheem, J. Palmio, S. Sandell, B. Eymard, I. Illa, R. Rojas-Garcia, K. Hankiewicz, et al. 2014. Atypical phenotypes in titinopathies explained by second titin mutations. *Ann.Neurol.* 75:230-240. doi: 10.1002/ana.24102 [doi].

Fehlmann, T., S. Reinheimer, C. Geng, X. Su, S. Drmanac, A. Alexeev, C. Zhang, C. Backes, N. Ludwig, M. Hart, et al. 2016. cPAS-based sequencing on the BGISEQ-500 to explore small non-coding RNAs. *Clin.Epigenetics.* 8:123-016-0287-1. eCollection 2016. doi: 123.

Feng, Y., X. Ge, L. Meng, J. Scull, J. Li, X. Tian, T. Zhang, W. Jin, H. Cheng, X. Wang, et al. 2017. The next generation of population-based spinal muscular atrophy carrier screening: comprehensive pan-ethnic SMN1 copy-number and sequence variant analysis by massively parallel sequencing. *Genet.Med.* 19:936-944. doi: 10.1038/gim.2016.215 [doi].

Feuk, L., A.R. Carson, and S.W. Scherer. 2006. Structural variation in the human genome. *Nat.Rev.Genet.* 7:85-97. doi: nrg1767 [pii].

Firth, H.V., S.M. Richards, A.P. Bevan, S. Clayton, M. Corpas, D. Rajan, S. Van Vooren, Y. Moreau, R.M. Pettett, and N.P. Carter. 2009. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am.J.Hum.Genet.* 84:524-533. doi: 10.1016/j.ajhg.2009.03.010 [doi].

Fromer, M., J.L. Moran, K. Chambert, E. Banks, S.E. Bergen, D.M. Ruderfer, R.E. Handsaker, S.A. McCarroll, M.C. O'Donovan, M.J. Owen, et al. 2012. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am.J.Hum.Genet.* 91:597-607. doi: 10.1016/j.ajhg.2012.08.005 [doi].

Gambin, T., Z.C. Akdemir, B. Yuan, S. Gu, T. Chiang, C.M.B. Carvalho, C. Shaw, S. Jhangiani, P.M. Boone, M.K. Eldomery, et al. 2017. Homozygous and hemizygous CNV detection from exome sequencing data in a Mendelian disease cohort. *Nucleic Acids Res.* 45:1633-1648. doi: 10.1093/nar/gkw1237 [doi].

Gambin, T., B. Yuan, W. Bi, P. Liu, J.A. Rosenfeld, Z. Coban-Akdemir, A.N. Pursley, S.C.S. Nagamani, R. Marom, S. Golla, et al. 2017. Identification of novel candidate disease genes from de novo exonic copy number variants. *Genome Med.* 9:83-017-0472-7. doi: 10.1186/s13073-017-0472-7 [doi].

Ganel, L., H.J. Abel, FinMetSeq Consortium, and I.M. Hall. 2017. SVScore: an impact prediction tool for structural variation. *Bioinformatics.* 33:1083-1085. doi: 10.1093/bioinformatics/btw789 [doi].

Gao, J., C. Wan, H. Zhang, A. Li, Q. Zang, R. Ban, A. Ali, Z. Yu, Q. Shi, X. Jiang, and Y. Zhang. 2017. Anaconda: AN automated pipeline for somatic COpy Number variation Detection and Annotation from tumor exome sequencing data. *BMC Bioinformatics.* 18:436-017-1833-3. doi: 10.1186/s12859-017-1833-3 [doi].

Garg, S., S. Grenier, M. Misyura, M.A. Sukhai, M. Thomas, S. Kamel-Reid, and T. Stockley. 2020. Assessing the Diagnostic Yield of Targeted Next-Generation Sequencing for Melanoma and Gastrointestinal Tumors. *J.Mol.Diagn.* 22:467-475. doi: S1525-1578(20)30014-3 [pii].

Geoffroy, V., Y. Herenger, A. Kress, C. Stoetzel, A. Piton, H. Dollfus, and J. Muller. 2018. AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics.* 34:3572-3574. doi: 10.1093/bioinformatics/bty304 [doi].

Gheldof, N., R.M. Witwicki, E. Migliavacca, M. Leleu, G. Didelot, L. Harewood, J. Rougemont, and A. Reymond. 2013. Structural variation-associated expression changes are paralleled by chromatin architecture modifications. *PLoS One.* 8:e79973. doi: 10.1371/journal.pone.0079973 [doi].

Giugliano, T., M. Savarese, A. Garofalo, E. Picillo, C. Fiorillo, A. D'Amico, L. Maggi, L. Ruggiero, L. Vercelli, F. Magri, et al. 2018. Copy Number Variants Account for a Tiny Fraction of Undiagnosed Myopathic Patients. *Genes (Basel).* 9:10.3390/genes9110524. doi: E524 [pii].

Goodwin, S., J.D. McPherson, and W.R. McCombie. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nat.Rev.Genet.* 17:333-351. doi: 10.1038/nrg.2016.49 [doi].

Gorokhova, S., V. Biancalana, N. Levy, J. Laporte, M. Bartoli, and M. Krahn. 2015. Clinical massively parallel sequencing for the diagnosis of myopathies. *Rev.Neurol.(Paris).* 171:558-571. doi: 10.1016/j.neurol.2015.02.019 [doi].

Green, R.C., J.S. Berg, W.W. Grody, S.S. Kalia, B.R. Korf, C.L. Martin, A.L. McGuire, R.L. Nussbaum, J.M. O'Daniel, K.E. Ormond, et al. 2013. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet.Med.* 15:565-574. doi: 10.1038/gim.2013.73 [doi].

Gulilat, M., T. Lamb, W.A. Teft, J. Wang, J.S. Dron, J.F. Robinson, R.G. Tirona, R.A. Hegele, R.B. Kim, and U.I. Schwarz. 2019. Targeted next generation sequencing as a tool for precision medicine. *BMC Med.Genomics.* 12:81-019-0527-2. doi: 10.1186/s12920-019-0527-2 [doi].

Hackman, P., A. Vihola, H. Haravuori, S. Marchand, J. Sarparanta, J. De Seze, S. Labeit, C. Witt, L. Peltonen, I. Richard, and B. Udd. 2002. Tibial muscular dystrophy is a titinopathy caused by mutations in TTN, the gene encoding the giant skeletal-muscle protein titin. *Am.J.Hum.Genet.* 71:492-500. doi: S0002-9297(07)60330-9 [pii].

Harel, T., and J.R. Lupski. 2018. Genomic disorders 20 years on-mechanisms for clinical manifestations. *Clin.Genet.* 93:439-449. doi: 10.1111/cge.13146 [doi].

Harrow, J., A. Frankish, J.M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B.L. Aken, D. Barrell, A. Zadissa, S. Searle, et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* 22:1760-1774. doi: 10.1101/gr.135350.111 [doi].

Hartley, T., J.D. Wagner, J. Warman-Chardon, M. Tetreault, L. Brady, S. Baker, M. Tarnopolsky, P.R. Bourque, J.S. Parboosingh, C. Smith, et al. 2018. Whole-exome sequencing is a valuable diagnostic tool for inherited peripheral neuropathies: Outcomes from a cohort of 50 families. *Clin.Genet.* 93:301-309. doi: 10.1111/cge.13101 [doi].

Hastings, P.J., G. Ira, and J.R. Lupski. 2009. A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genet.* 5:e1000327. doi: 10.1371/journal.pgen.1000327 [doi].

Hastings, P.J., J.R. Lupski, S.M. Rosenberg, and G. Ira. 2009. Mechanisms of change in gene copy number. *Nat.Rev.Genet.* 10:551-564. doi: 10.1038/nrg2593 [doi].

Heberle, H., G.V. Meirelles, F.R. da Silva, G.P. Telles, and R. Minghim. 2015. InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. *BMC Bioinformatics.* 16:169-015-0611-3. doi: 10.1186/s12859-015-0611-3 [doi].

Hehir-Kwa, J.Y., B.B.J. Tops, and P. Kemmeren. 2018. The clinical implementation of copy number detection in the age of next-generation sequencing. *Expert Rev.Mol.Diagn.* 18:907-915. doi: 10.1080/14737159.2018.1523723 [doi].

Hiraide, T., T. Ogata, S. Watanabe, M. Nakashima, T. Fukuda, and H. Saitsu. 2019. Coexistence of a CAV3 mutation and a DMD deletion in a family with complex muscular diseases. *Brain Dev.* 41:474-479. doi: S0387-7604(18)30594-1 [pii].

Hodges, E., Z. Xuan, V. Balija, M. Kramer, M.N. Molla, S.W. Smith, C.M. Middle, M.J. Rodesch, T.J. Albert, G.J. Hannon, and W.R. McCombie. 2007. Genome-wide in situ exon capture for selective resequencing. *Nat.Genet.* 39:1522-1527. doi: ng.2007.42 [pii].

Holland, P.M., R.D. Abramson, R. Watson, and D.H. Gelfand. 1991. Detection of specific polymerase chain reaction product by utilizing the 5'----3' exonuclease activity of Thermus aquaticus DNA polymerase. *Proc.Natl.Acad.Sci.U.S.A.* 88:7276-7280. doi: 10.1073/pnas.88.16.7276 [doi].

Huang, N., I. Lee, E.M. Marcotte, and M.E. Hurles. 2010. Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet.* 6:e1001154. doi: 10.1371/journal.pgen.1001154 [doi].

Hurles, M.E., E.T. Dermitzakis, and C. Tyler-Smith. 2008. The functional impact of structural variation in humans. *Trends Genet.* 24:238-245. doi: 10.1016/j.tig.2008.03.001 [doi].

Hwang, M.Y., S. Moon, L. Heo, Y.J. Kim, J.H. Oh, Y.J. Kim, Y.K. Kim, J. Lee, B.G. Han, and B.J. Kim. 2015. Combinatorial approach to estimate copy number genotype using whole-exome sequencing data. *Genomics.* 105:145-149. doi: 10.1016/j.ygeno.2014.12.003 [doi].

Iafrate, A.J., L. Feuk, M.N. Rivera, M.L. Listewnik, P.K. Donahoe, Y. Qi, S.W. Scherer, and C. Lee. 2004. Detection of large-scale variation in the human genome. *Nat.Genet.* 36:949-951. doi: 10.1038/ng1416 [doi].

Illarioshkin, S.N., I.A. Ivanova-Smolenskaya, C.R. Greenberg, E. Nylen, V.S. Sukhorukov, V.V. Poleshchuk, E.D. Markova, and K. Wrogemann. 2000. Identical dysferlin mutation in limb-girdle muscular dystrophy type 2B and distal myopathy. *Neurology.* 55:1931-1933.

Ito, T., Y. Kawashima, T. Fujikawa, K. Honda, A. Makabe, K. Kitamura, and T. Tsutsumi. 2019. Rapid screening of copy number variations in STRC by droplet digital PCR in patients with mild-to-moderate hearing loss. *Hum.Genome Var.* 6:41-019-0075-5. eCollection 2019. doi: 10.1038/s41439-019-0075-5 [doi].

Itsara, A., G.M. Cooper, C. Baker, S. Girirajan, J. Li, D. Absher, R.M. Krauss, R.M. Myers, P.M. Ridker, D.I. Chasman, et al. 2009. Population analysis of large copy number variants and hotspots of human genetic disease. *Am.J.Hum.Genet.* 84:148-161. doi: 10.1016/j.ajhg.2008.12.014 [doi].

Jaganathan, K., S. Kyriazopoulou Panagiotopoulou, J.F. McRae, S.F. Darbandi, D. Knowles, Y.I. Li, J.A. Kosmicki, J. Arbelaez, W. Cui, G.B. Schwartz, et al. 2019. Predicting Splicing from Primary Sequence with Deep Learning. *Cell.* 176:535-548.e24. doi: S0092-8674(18)31629-5 [pii].

Jain, M., S. Koren, K.H. Miga, J. Quick, A.C. Rand, T.A. Sasani, J.R. Tyson, A.D. Beggs, A.T. Dilthey, I.T. Fiddes, et al. 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat.Biotechnol.* 36:338-345. doi: 10.1038/nbt.4060 [doi].

Jiang, Y., D.A. Oldridge, S.J. Diskin, and N.R. Zhang. 2015. CODEX: a normalization and copy number variation detection method for whole exome sequencing. *Nucleic Acids Res.* 43:e39. doi: 10.1093/nar/gku1363 [doi].

Johari, M., M. Arumilli, J. Palmio, M. Savarese, G. Tasca, M. Mirabella, N. Sandholm, H. Lohi, P. Hackman, and B. Udd. 2017. Association study reveals novel risk loci for sporadic inclusion body myositis. *Eur.J.Neurol.* 24:572-577. doi: 10.1111/ene.13244 [doi].

Kadalayil, L., S. Rafiq, M.J. Rose-Zerilli, R.J. Pengelly, H. Parker, D. Oscier, J.C. Strefford, W.J. Tapper, J. Gibson, S. Ennis, and A. Collins. 2015. Exome sequence read depth methods for identifying copy number changes. *Brief Bioinform.* 16:380-392. doi: 10.1093/bib/bbu027 [doi].

Kalia, S.S., K. Adelman, S.J. Bale, W.K. Chung, C. Eng, J.P. Evans, G.E. Herman, S.B. Hufnagel, T.E. Klein, B.R. Korf, et al. 2017. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. *Genet.Med.* 19:249-255. doi: 10.1038/gim.2016.190 [doi].

Kallioniemi, A., O.P. Kallioniemi, D. Sudar, D. Rutovitz, J.W. Gray, F. Waldman, and D. Pinkel. 1992. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science.* 258:818-821. doi: 10.1126/science.1359641 [doi].

Karaca, E., J.E. Posey, Z. Coban Akdemir, D. Pehlivan, T. Harel, S.N. Jhangiani, Y. Bayram, X. Song, V. Bahrambeigi, O.O. Yuregir, et al. 2018. Phenotypic expansion illuminates multilocus pathogenic variation. *Genet.Med.* 20:1528-1537. doi: 10.1038/gim.2018.33 [doi].

Karczewski, K.J., L.C. Francioli, G. Tiao, B.B. Cummings, J. Alföldi, Q. Wang, R.L. Collins, K.M. Laricchia, A. Ganna, D.P. Birnbaum, et al. 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature.* 581:434-443. doi: 10.1038/s41586-020-2308-7.

Kasianowicz, J.J., E. Brandin, D. Branton, and D.W. Deamer. 1996. Characterization of individual polynucleotide molecules using a membrane channel. *Proc.Natl.Acad.Sci.U.S.A.* 93:13770-13773. doi: 10.1073/pnas.93.24.13770 [doi].

Kazazian, H.H.,Jr, and J.V. Moran. 2017. Mobile DNA in Health and Disease. *N.Engl.J.Med.* 377:361-370. doi: 10.1056/NEJMra1510092 [doi].

Kerkhof, J., L.C. Schenkel, J. Reilly, S. McRobbie, E. Aref-Eshghi, A. Stuart, C.A. Rupar, P. Adams, R.A. Hegele, H. Lin, et al. 2017. Clinical Validation of Copy Number Variant Detection from Targeted Next-Generation Sequencing Panels. *J.Mol.Diagn.* 19:905-920. doi: S1525-1578(17)30207-6 [pii].

Kiiski, K., L. Laari, V.L. Lehtokari, M. Lunkka-Hytonen, C. Angelini, R. Petty, P. Hackman, C. Wallgren-Pettersson, and K. Pelin. 2013. Targeted array comparative genomic hybridization--a new diagnostic tool for the detection of large copy number variations in nemaline myopathy-causing genes. *Neuromuscul.Disord.* 23:56-65. doi: 10.1016/j.nmd.2012.07.007 [doi].

Kiiski, K., V.L. Lehtokari, A. Loytynoja, L. Ahlsten, J. Laitila, C. Wallgren-Pettersson, and K. Pelin. 2016. A recurrent copy number variation of the NEB triplicate region: only revealed by the targeted nemaline myopathy CGH array. *Eur.J.Hum.Genet.* 24:574-580. doi: 10.1038/ejhg.2015.166 [doi].

Kiiski, K.J., V.L. Lehtokari, A.K. Vihola, J.M. Laitila, S. Huovinen, L.J. Sagath, A.E. Evila, A.E. Paetau, C.A. Sewry, P.B. Hackman, K.B. Pelin, C. Wallgren-Pettersson, and B. Udd. 2019. Dominantly inherited distal nemaline/cap myopathy caused by a large deletion in the nebulin gene. *Neuromuscul.Disord.* 29:97-107. doi: S0960-8966(18)30563-7 [pii].

Kim, D., A. Lucas, J. Glessner, S.S. Verma, Y. Bradford, R. Li, A.T. Frase, H. Hakonarson, P. Peissig, M. Brilliant, and M.D. Ritchie. 2016. Biofilter as a Functional Annotation Pipeline for Common and Rare Copy Number Burden. *Pac.Symp.Biocomput.* 21:357-368. doi: 9789814749411_0033 [pii].

Kim, H.Y., J.W. Choi, J.Y. Lee, and G. Kong. 2017. Gene-based comparative analysis of tools for estimating copy number alterations using whole-exome sequencing data. *Oncotarget.* 8:27277-27285. doi: 10.18632/oncotarget.15932 [doi].

Kircher, M., S. Sawyer, and M. Meyer. 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.* 40:e3. doi: 10.1093/nar/gkr771 [doi].

Kitamura, Y., E. Kondo, M. Urano, R. Aoki, and K. Saito. 2016. Target resequencing of neuromuscular disease-related genes using next-generation sequencing for patients with undiagnosed early-onset neuromuscular disorders. *J.Hum.Genet.* 61:931-942. doi: 10.1038/jhg.2016.79 [doi].

Koenig, M., E.P. Hoffman, C.J. Bertelson, A.P. Monaco, C. Feener, and L.M. Kunkel. 1987. Complete cloning of the Duchenne muscular dystrophy (DMD) cDNA and preliminary genomic organization of the DMD gene in normal and affected individuals. *Cell.* 50:509-517. doi: 0092-8674(87)90504-6 [pii].

Korbel, J.O., A.E. Urban, J.P. Affourtit, B. Godwin, F. Grubert, J.F. Simons, P.M. Kim, D. Palejev, N.J. Carriero, L. Du, et al. 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science.* 318:420-426. doi: 1149504 [pii].

Koressaar, T., and M. Remm. 2007. Enhancements and modifications of primer design program Primer3. *Bioinformatics.* 23:1289-1291. doi: btm091 [pii].

Kosugi, S., Y. Momozawa, X. Liu, C. Terao, M. Kubo, and Y. Kamatani. 2019. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol.* 20:117-019-1720-5. doi: 10.1186/s13059-019-1720-5 [doi].

Kovaka, S., Y. Fan, B. Ni, W. Timp, and M.C. Schatz. 2020. Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED. *bioRxiv.* doi: 10.1101/2020.02.03.931923.

Kozareva, V., C. Stroff, M. Silver, J.F. Freidin, and N.F. Delaney. 2018. Clinical analysis of germline copy number variation in DMD using a non-conjugate hierarchical Bayesian model. *BMC Med.Genomics.* 11:91-018-0404-4. doi: 10.1186/s12920-018-0404-4 [doi].

Krumm, N., P.H. Sudmant, A. Ko, B.J. O'Roak, M. Malig, B.P. Coe, NHLBI Exome Sequencing Project, A.R. Quinlan, D.A. Nickerson, and E.E. Eichler. 2012. Copy number variation detection and genotyping from exome sequence data. *Genome Res.* 22:1525-1532. doi: 10.1101/gr.138115.112 [doi].

Krumm, N., T.N. Turner, C. Baker, L. Vives, K. Mohajeri, K. Witherspoon, A. Raja, B.P. Coe, H.A. Stessman, Z.X. He, et al. 2015. Excess of rare, inherited truncating mutations in autism. *Nat.Genet.* 47:582-588. doi: 10.1038/ng.3303 [doi].

Laing, N.G. 2012. Genetics of neuromuscular disorders. *Crit.Rev.Clin.Lab.Sci.* 49:33-48. doi: 10.3109/10408363.2012.658906 [doi].

Lander, E.S., L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, et al. 2001. Initial sequencing and analysis of the human genome. *Nature.* 409:860-921. doi: 10.1038/35057062 [doi].

Landrum, M.J., J.M. Lee, G.R. Riley, W. Jang, W.S. Rubinstein, D.M. Church, and D.R. Maglott. 2014. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* 42:D980-5. doi: 10.1093/nar/gkt1113 [doi].

Lappalainen, T., A.J. Scott, M. Brandt, and I.M. Hall. 2019. Genomic Analysis in the Age of Human Genome Sequencing. *Cell.* 177:70-84. doi: S0092-8674(19)30215-6 [pii].

Lassuthova, P., D. Safka Brozkova, M. Krutova, J. Neupauerova, J. Haberlova, R. Mazanec, P. Drimal, and P. Seeman. 2016. Improving diagnosis of inherited peripheral neuropathies

through gene panel analysis. *Orphanet J.Rare Dis.* 11:118-016-0500-5. doi: 10.1186/s13023-016-0500-5 [doi].

Laver, T.W., E.D. Franco, M.B. Johnson, K. Patel, S. Ellard, M.N. Weedon, S.E. Flanagan, and M.N. Wakeling. 2019. SavvyCNV: genome-wide CNV calling from off-target reads. *bioRxiv.* doi: 10.1101/617605.

LeDell, E., M. Petersen, and M. van der Laan. 2015. Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates. *Electron.J.Stat.* 9:1583-1607. doi: 10.1214/15-EJS1035 [doi].

Lee, J.A., C.M. Carvalho, and J.R. Lupski. 2007. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell.* 131:1235-1247. doi: S0092-8674(07)01541-3 [pii].

Lee, Y., P.H. Jonson, J. Sarparanta, J. Palmio, M. Sarkar, A. Vihola, A. Evilä, T. Suominen, S. Penttilä, M. Savarese, et al. 2018. TIA1 variant drives myodegeneration in multisystem proteinopathy with SQSTM1 mutations. *J.Clin.Invest.* 128:1164-1177. doi: 97103 [pii].

Lefebvre, S., L. Burglen, S. Reboullet, O. Clermont, P. Burlet, L. Viollet, B. Benichou, C. Cruaud, P. Millasseau, and M. Zeviani. 1995. Identification and characterization of a spinal muscular atrophy-determining gene. *Cell.* 80:155-165. doi: 0092-8674(95)90460-3 [pii].

Lehtokari, V.L., K. Kiiski, S.A. Sandaradura, J. Laporte, P. Repo, J.A. Frey, K. Donner, M. Marttila, C. Saunders, P.G. Barth, et al. 2014. Mutation update: the spectra of nebulin variants and associated myopathies. *Hum.Mutat.* 35:1418-1426. doi: 10.1002/humu.22693 [doi].

Lek, M., K.J. Karczewski, E.V. Minikel, K.E. Samocha, E. Banks, T. Fennell, A.H. O'Donnell-Luria, J.S. Ware, A.J. Hill, B.B. Cummings, et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature.* 536:285-291. doi: 10.1038/nature19057 [doi].

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 25:1754-1760. doi: 10.1093/bioinformatics/btp324 [doi].

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 25:2078-2079. doi: 10.1093/bioinformatics/btp352 [doi].

Li, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv.* 1303.

Li, M.M., M. Datto, E.J. Duncavage, S. Kulkarni, N.I. Lindeman, S. Roy, A.M. Tsimberidou, C.L. Vnencak-Jones, D.J. Wolff, A. Younes, and M.N. Nikiforova. 2017. Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer: A Joint Consensus Recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. *J.Mol.Diagn.* 19:4-23. doi: S1525-1578(16)30223-9 [pii].

Li, Y., H. Zheng, R. Luo, H. Wu, H. Zhu, R. Li, H. Cao, B. Wu, S. Huang, H. Shao, et al. 2011. Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome de novo assembly. *Nat.Biotechnol.* 29:723-730. doi: 10.1038/nbt.1904 [doi].

Lin, F. 2008. Solving Multicollinearity in the Process of Fitting Regression Model Using the Nested Estimate Procedure. *Quality {\&} Quantity.* 42:417-426. doi: 10.1007/s11135-006-9055-1".

Lionel, A.C., G. Costain, N. Monfared, S. Walker, M.S. Reuter, S.M. Hosseini, B. Thiruvahindrapuram, D. Merico, R. Jobling, T. Nalpathamkalam, et al. 2018. Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. *Genet.Med.* 20:435-443. doi: 10.1038/gim.2017.119 [doi].

Liu, P., A. Erez, S.C. Nagamani, S.U. Dhar, K.E. Kolodziejska, A.V. Dharmadhikari, M.L. Cooper, J. Wiszniewska, F. Zhang, M.A. Withers, et al. 2011. Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements. *Cell.* 146:889-903. doi: 10.1016/j.cell.2011.07.042 [doi].

Liu, P., V. Gelowani, F. Zhang, V.E. Drory, S. Ben-Shachar, E. Roney, A.C. Medeiros, R.J. Moore, C. DiVincenzo, W.B. Burnette, et al. Mechanism, prevalence, and more severe neuropathy phenotype of the Charcot-Marie-Tooth type 1A triplication. *Am.J.Hum.Genet.* 94:462-469. doi: 10.1016/j.ajhg.2014.01.017 [doi].

Liu, X., A. Li, J. Xi, H. Feng, and M. Wang. 2018. Detection of copy number variants and loss of heterozygosity from impure tumor samples using whole exome sequencing data. *Oncol.Lett.* 16:4713-4720. doi: 10.3892/ol.2018.9150 [doi].

Lorson, C.L., E. Hahnen, E.J. Androphy, and B. Wirth. 1999. A single nucleotide in the SMN gene regulates splicing and is responsible for spinal muscular atrophy. *Proc.Natl.Acad.Sci.U.S.A.* 96:6307-6311. doi: 0286 [pii].

Lupianez, D.G., K. Kraft, V. Heinrich, P. Krawitz, F. Brancati, E. Klopocki, D. Horn, H. Kayserili, J.M. Opitz, R. Laxova, et al. 2015. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell.* 161:1012-1025. doi: S0092-8674(15)00377-3 [pii].

Lupski, J.R. 1998. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet.* 14:417-422. doi: S0168-9525(98)01555-8 [pii].

Lupski, J.R., R.M. de Oca-Luna, S. Slaugenhaupt, L. Pentao, V. Guzzetta, B.J. Trask, O. Saucedo-Cardenas, D.F. Barker, J.M. Killian, C.A. Garcia, et al. 1991. DNA duplication associated with Charcot-Marie-Tooth disease type 1A. *Cell.* 66:219-232. doi: 0092-8674(91)90613-4 [pii].

Ma, X., Y. Shao, L. Tian, D.A. Flasch, H.L. Mulder, M.N. Edmonson, Y. Liu, X. Chen, S. Newman, J. Nakitandwe, et al. 2019. Analysis of error profiles in deep next-generation sequencing data. *Genome Biol.* 20:50-019-1659-6. doi: 10.1186/s13059-019-1659-6 [doi].

MacDonald, J.R., R. Ziman, R.K. Yuen, L. Feuk, and S.W. Scherer. 2014. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.* 42:D986-92. doi: 10.1093/nar/gkt958 [doi].

Madoui, M.A., S. Engelen, C. Cruaud, C. Belser, L. Bertrand, A. Alberti, A. Lemainque, P. Wincker, and J.M. Aury. 2015. Genome assembly using Nanopore-guided long and error-free DNA reads. *BMC Genomics.* 16:327-015-1519-z. doi: 10.1186/s12864-015-1519-z [doi].

Mallawaarachchi, A.C., Y. Hort, M.J. Cowley, M.J. McCabe, A. Minoche, M.E. Dinger, J. Shine, and T.J. Furlong. 2016. Whole-genome sequencing overcomes pseudogene homology to diagnose autosomal dominant polycystic kidney disease. *Eur.J.Hum.Genet.* 24:1584-1590. doi: 10.1038/ejhg.2016.48 [doi].

Marchuk, D.S., K. Crooks, N. Strande, K. Kaiser-Rogers, L.V. Milko, A. Brandt, A. Arreola, C.R. Tilley, C. Bizon, N.L. Vora, et al. 2018. Increasing the diagnostic yield of exome sequencing by copy number variant analysis. *PLoS One.* 13:e0209185. doi: 10.1371/journal.pone.0209185 [doi].

Marelli, C., C. Guissart, C. Hubsch, M. Renaud, J.P. Villemin, L. Larrieu, P. Charles, X. Ayrignac, S. Sacconi, P. Collignon, et al. 2016. Mini-Exome Coupled to Read-Depth Based Copy Number Variation Analysis in Patients with Inherited Ataxias. *Hum.Mutat.* 37:1340-1353. doi: 10.1002/humu.23063 [doi].

Margulies, M., M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, Y.J. Chen, Z. Chen, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature.* 437:376-380. doi: nature03959 [pii].

Marks, P., S. Garcia, A.M. Barrio, K. Belhocine, J. Bernate, R. Bharadwaj, K. Bjornson, C. Catalanotti, J. Delaney, A. Fehr, et al. 2019. Resolving the full spectrum of human genome variation using Linked-Reads. *Genome Res.* 29:635-645. doi: 10.1101/gr.234443.118 [doi].

Masson, E., S. Maestri, D.N. Cooper, C. F\'erec, and J. Chen. 2020. RNA secondary structure mediated by Alu insertion as a novel disease-causing mechanism. *bioRxiv.* doi: 10.1101/2020.01.30.926790.

Matthijs, G., E. Souche, M. Alders, A. Corveleyn, S. Eck, I. Feenstra, V. Race, E. Sistermans, M. Sturm, M. Weiss, et al. 2016. Guidelines for diagnostic next-generation sequencing. *Eur.J.Hum.Genet.* 24:2-5. doi: 10.1038/ejhg.2015.226 [doi].

McCoy, R.C., R.W. Taylor, T.A. Blauwkamp, J.L. Kelley, M. Kertesz, D. Pushkarev, D.A. Petrov, and A.S. Fiston-Lavier. 2014. Illumina TruSeq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements. *PLoS One.* 9:e106689. doi: 10.1371/journal.pone.0106689 [doi].

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M.A. DePristo. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20:1297-1303. doi: 10.1101/gr.107524.110 [doi].

Meijerman, I., L.M. Sanderson, P.H. Smits, J.H. Beijnen, and J.H. Schellens. 2007. Pharmacogenetic screening of the gene deletion and duplications of CYP2D6. *Drug Metab.Rev.* 39:45-60. doi: 770420517 [pii].

Mercuri, E., B.T. Darras, C.A. Chiriboga, J.W. Day, C. Campbell, A.M. Connolly, S.T. Iannaccone, J. Kirschner, N.L. Kuntz, K. Saito, et al. 2018. Nusinersen versus Sham Control in Later-Onset Spinal Muscular Atrophy. *N.Engl.J.Med.* 378:625-635. doi: 10.1056/NEJMoa1710504 [doi].

Meynert, A.M., M. Ansari, D.R. FitzPatrick, and M.S. Taylor. 2014. Variant detection sensitivity and biases in whole genome and exome sequencing. *BMC Bioinformatics.* 15:247-2105-15-247. doi: 10.1186/1471-2105-15-247 [doi].

Meynert, A.M., L.S. Bicknell, M.E. Hurles, A.P. Jackson, and M.S. Taylor. 2013. Quantifying single nucleotide variant detection sensitivity in exome sequencing. *BMC Bioinformatics.* 14:195-2105-14-195. doi: 10.1186/1471-2105-14-195 [doi].

Mills, R.E., C.T. Luttig, C.E. Larkins, A. Beauchamp, C. Tsui, W.S. Pittard, and S.E. Devine. 2006. An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.* 16:1182-1190. doi: 10.1101/gr.4565806 [doi].

Mills, R.E., K. Walter, C. Stewart, R.E. Handsaker, K. Chen, C. Alkan, A. Abyzov, S.C. Yoon, K. Ye, R.K. Cheetham, et al. 2011. Mapping copy number variation by population-scale genome sequencing. *Nature.* 470:59-65. doi: 10.1038/nature09708 [doi].

Milone, M. 2017. Diagnosis and Management of Immune-Mediated Myopathies. *Mayo Clin.Proc.* 92:826-837. doi: S0025-6196(17)30076-9 [pii].

Mirkin, S.M. 2007. Expandable DNA repeats and human disease. *Nature.* 447:932-940. doi: nature05977 [pii].

Mirza, M., A. Vainshtein, A. DiRonza, U. Chandrachud, L.J. Haslett, M. Palmieri, S. Storch, J. Groh, N. Dobzinski, G. Napolitano, et al. 2019. The CLN3 gene and protein: What we know. *Mol.Genet.Genomic Med.* 7:e859. doi: 10.1002/mgg3.859 [doi].

Munz, M., S. Mahamdallie, S. Yost, A. Rimmer, E. Poyastro-Pearson, A. Strydom, S. Seal, E. Ruark, and N. Rahman. 2018. CoverView: a sequence quality evaluation tool for next generation sequencing data. *Wellcome Open Res.* 3:36. doi: 10.12688/wellcomeopenres.14306.1 [doi].

Nam, S.H., Y.B. Hong, Y.S. Hyun, E. Nam da, G. Kwak, S.H. Hwang, B.O. Choi, and K.W. Chung. 2016. Identification of Genetic Causes of Inherited Peripheral Neuropathies by Targeted Gene Panel Sequencing. *Mol.Cells.* 39:382-388. doi: 10.14348/molcells.2016.2288 [doi].

Needham, M., and F.L. Mastaglia. 2016. Sporadic inclusion body myositis: A review of recent clinical advances and current approaches to diagnosis and treatment. *Clin.Neurophysiol.* 127:1764-1773. doi: 10.1016/j.clinph.2015.12.011 [doi].

Neerman, N., G. Faust, N. Meeks, S. Modai, L. Kalfon, T. Falik-Zaccai, and A. Kaplun. 2019. A clinically validated whole genome pipeline for structural variant detection and analysis. *BMC Genomics.* 20:545-019-5866-z. doi: 10.1186/s12864-019-5866-z [doi].

Neveling, K., T. Mantere, S. Vermeulen, M. Oorsprong, R. van Beek, E. Kater-Baats, M. Pauper, G. van der Zande, D. Smeets, D.O. Weghuis, et al. 2020. Next generation cytogenetics: comprehensive assessment of 48 leukemia genomes by genome imaging. *bioRxiv.* doi: 10.1101/2020.02.06.935742.

Newman, S., K.E. Hermetz, B. Weckselblatt, and M.K. Rudd. 2015. Next-generation sequencing of duplication CNVs reveals that most are tandem and some create fusion genes at breakpoints. *Am.J.Hum.Genet.* 96:208-220. doi: 10.1016/j.ajhg.2014.12.017 [doi].

Nigro, V., and G. Piluso. 2012. Next generation sequencing (NGS) strategies for the genetic testing of myopathies. *Acta Myol.* 31:196-200.

Nishikawa, A., S. Mitsuhashi, N. Miyata, and I. Nishino. 2017. Targeted massively parallel sequencing and histological assessment of skeletal muscles for the molecular diagnosis of inherited muscle disorders. *J.Med.Genet.* 54:104-110. doi: 10.1136/jmedgenet-2016-104073 [doi].

Onozato, M.L., C. Yapp, D. Richardson, T. Sundaresan, V. Chahal, J. Lee, J.P. Sullivan, M.W. Madden, H.S. Shim, M. Liebers, et al. 2019. Highly Multiplexed Fluorescence in Situ Hybridization for in Situ Genomics. *J.Mol.Diagn.* 21:390-407. doi: S1525-1578(18)30022-9 [pii].

Pang, A.W., J.R. MacDonald, D. Pinto, J. Wei, M.A. Rafiq, D.F. Conrad, H. Park, M.E. Hurles, C. Lee, J.C. Venter, et al. 2010. Towards a comprehensive structural variation map of an individual human genome. *Genome Biol.* 11:R52-2010-11-5-r52. Epub 2010 May 19. doi: 10.1186/gb-2010-11-5-r52 [doi].

Papp, B., C. Pal, and L.D. Hurst. 2003. Dosage sensitivity and the evolution of gene families in yeast. *Nature.* 424:194-197. doi: 10.1038/nature01771 [doi].

Pehlivan, D., C.R. Beck, Y. Okamoto, T. Harel, Z.H. Akdemir, S.N. Jhangiani, M.A. Withers, M.T. Goksungur, C.M. Carvalho, D. Czesnik, et al. 2016. The role of combined SNV and CNV burden in patients with distal symmetric polyneuropathy. *Genet.Med.* 18:443-451. doi: 10.1038/gim.2015.124 [doi].

Pelin, K., P. Hilpela, K. Donner, C. Sewry, P.A. Akkari, S.D. Wilton, D. Wattanasirichaigoon, M.L. Bang, T. Centner, F. Hanefeld, et al. 1999. Mutations in the nebulin gene associated with autosomal recessive nemaline myopathy. *Proc.Natl.Acad.Sci.U.S.A.* 96:2305-2310. doi: 10.1073/pnas.96.5.2305 [doi].

Pellestor, F., and V. Gatinois. 2018. Chromoanasynthesis: another way for the formation of complex chromosomal abnormalities in human reproduction. *Hum.Reprod.* 33:1381-1387. doi: 10.1093/humrep/dey231 [doi].

Pengelly, R.J., D. Ward, D. Hunt, C. Mattocks, and S. Ennis. 2020. Comparison of Mendeliome exome capture kits for use in clinical diagnostics. *Sci.Rep.* 10:3235-020-60215-y. doi: 10.1038/s41598-020-60215-y [doi].

Pentao, L., C.A. Wise, A.C. Chinault, P.I. Patel, and J.R. Lupski. 1992. Charcot-Marie-Tooth type 1A duplication appears to arise from recombination at repeat sequences flanking the 1.5 Mb monomer unit. *Nat.Genet.* 2:292-300. doi: 10.1038/ng1292-292 [doi].

Perry, G.H., N.J. Dominy, K.G. Claw, A.S. Lee, H. Fiegler, R. Redon, J. Werner, F.A. Villanea, J.L. Mountain, R. Misra, et al. 2007. Diet and the evolution of human amylase gene copy number variation. *Nat.Genet.* 39:1256-1260. doi: ng2123 [pii].

Petrovski, S., Q. Wang, E.L. Heinzen, A.S. Allen, and D.B. Goldstein. 2013. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.* 9:e1003709. doi: 10.1371/journal.pgen.1003709 [doi].

Pfundt, R., M. Del Rosario, L.E.L.M. Vissers, M.P. Kwint, I.M. Janssen, N. de Leeuw, H.G. Yntema, M.R. Nelen, D. Lugtenberg, E.J. Kamsteeg, et al. 2017. Detection of clinically relevant copy-number variants by exome sequencing in a large cohort of genetic disorders. *Genet.Med.* 19:667-675. doi: 10.1038/gim.2016.163 [doi].

Piluso, G., M. Dionisi, F. Del Vecchio Blanco, A. Torella, S. Aurino, M. Savarese, T. Giugliano, E. Bertini, A. Terracciano, M. Vainzof, et al. 2011. Motor chip: a comparative genomic hybridization microarray for copy-number mutations in 245 neuromuscular disorders. *Clin.Chem.* 57:1584-1596. doi: 10.1373/clinchem.2011.168898 [doi].

Pinkel, D., R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W.L. Kuo, C. Chen, Y. Zhai, et al. 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat.Genet.* 20:207-211. doi: 10.1038/2524 [doi].

Plagnol, V., J. Curtis, M. Epstein, K.Y. Mok, E. Stebbings, S. Grigoriadou, N.W. Wood, S. Hambleton, S.O. Burns, A.J. Thrasher, et al. 2012. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics.* 28:2747-2754. doi: 10.1093/bioinformatics/bts526 [doi].

Pollack, J.R., C.M. Perou, A.A. Alizadeh, M.B. Eisen, A. Pergamenschikov, C.F. Williams, S.S. Jeffrey, D. Botstein, and P.O. Brown. 1999. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat.Genet.* 23:41-46. doi: 10.1038/12640 [doi].

Ponchel, F., C. Toomes, K. Bransfield, F.T. Leong, S.H. Douglas, S.L. Field, S.M. Bell, V. Combaret, A. Puisieux, A.J. Mighell, et al. 2003. Real-time PCR based on SYBR-Green I fluorescence: an alternative to the TaqMan assay for a relative quantification of gene rearrangements, gene amplifications and micro gene deletions. *BMC Biotechnol.* 3:18-6750-3-18. doi: 10.1186/1472-6750-3-18 [doi].

Posey, J.E., T. Harel, P. Liu, J.A. Rosenfeld, R.A. James, Z.H. Coban Akdemir, M. Walkiewicz, W. Bi, R. Xiao, Y. Ding, et al. 2017. Resolution of Disease Phenotypes Resulting from Multilocus Genomic Variation. *N.Engl.J.Med.* 376:21-31. doi: 10.1056/NEJMoa1516767 [doi].

Potocki, L., C.J. Shaw, P. Stankiewicz, and J.R. Lupski. 2003. Variability in clinical phenotype despite common chromosomal deletion in Smith-Magenis syndrome [del(17)(p11.2p11.2). *Genet.Med.* 5:430-434. doi: 10.1097/01.gim.0000095625.14160.ab [doi].

Pounraja, V.K., G. Jayakar, M. Jensen, N. Kelkar, and S. Girirajan. 2019. A machine-learning approach for accurate detection of copy-number variants from exome sequencing. *bioRxiv.* doi: 10.1101/460931.

Povysil, G., A. Tzika, J. Vogt, V. Haunschmid, L. Messiaen, J. Zschocke, G. Klambauer, S. Hochreiter, and K. Wimmer. 2017. panelcn.MOPS: Copy-number detection in targeted NGS panel data for clinical diagnostics. *Hum.Mutat.* 38:889-897. doi: 10.1002/humu.23237 [doi].

Punetha, J., A. Kesari, P. Uapinyoying, M. Giri, N.F. Clarke, L.B. Waddell, K.N. North, R. Ghaoui, G.L. O'Grady, E.C. Oates, et al. 2016. Targeted Re-Sequencing Emulsion PCR Panel for Myopathies: Results in 94 Cases. *J.Neuromuscul Dis.* 3:209-225. doi: JND160151 [pii].

Quail, M.A., M. Smith, P. Coupland, T.D. Otto, S.R. Harris, T.R. Connor, A. Bertoni, H.P. Swerdlow, and Y. Gu. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics.* 13:341-2164-13-341. doi: 10.1186/1471-2164-13-341 [doi].

Raeymaekers, P., V. Timmerman, E. Nelis, P. De Jonghe, J.E. Hoogendijk, F. Baas, D.F. Barker, J.J. Martin, M. De Visser, and P.A. Bolhuis. 1991. Duplication in chromosome 17p11.2 in Charcot-Marie-Tooth neuropathy type 1a (CMT 1a). The HMSN Collaborative Research Group. *Neuromuscul.Disord.* 1:93-97. doi: 10.1016/0960-8966(91)90055-w [doi].

Rajagopalan, R., J.R. Murrell, M. Luo, and L.K. Conlin. 2020. A highly sensitive and specific workflow for detecting rare copy-number variants from exome sequencing data. *Genome Med.* 12:14-020-0712-0. doi: 10.1186/s13073-020-0712-0 [doi].

Redin, C., H. Brand, R.L. Collins, T. Kammin, E. Mitchell, J.C. Hodge, C. Hanscom, V. Pillalamarri, C.M. Seabra, M.A. Abbott, et al. 2017. The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. *Nat.Genet.* 49:36-45. doi: 10.1038/ng.3720 [doi].

Redon, R., S. Ishikawa, K.R. Fitch, L. Feuk, G.H. Perry, T.D. Andrews, H. Fiegler, M.H. Shapero, A.R. Carson, W. Chen, et al. 2006. Global variation in copy number in the human genome. *Nature.* 444:444-454. doi: nature05329 [pii].

Regier, A.A., Y. Farjoun, D.E. Larson, O. Krasheninina, H.M. Kang, D.P. Howrigan, B.J. Chen, M. Kher, E. Banks, D.C. Ames, et al. 2018. Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. *Nat.Commun.* 9:4038-018-06159-4. doi: 10.1038/s41467-018-06159-4 [doi].

Reiter, L.T., P.J. Hastings, E. Nelis, P. De Jonghe, C. Van Broeckhoven, and J.R. Lupski. 1998. Human meiotic recombination products revealed by sequencing a hotspot for homologous strand exchange in multiple HNPP deletion patients. *Am.J.Hum.Genet.* 62:1023-1033. doi: 10.1086/301827 [doi].

Renaux, A., S. Papadimitriou, N. Versbraegen, C. Nachtegael, S. Boutry, A. Nowé, G. Smits, and T. Lenaerts. 2019. ORVAL: a novel platform for the prediction and exploration of disease-causing oligogenic variant combinations. *Nucleic Acids Res.* 47:W93-W98. doi: 10.1093/nar/gkz437 [doi].

Rice, A.M., and A. McLysaght. 2017. Dosage sensitivity is a major determinant of human copy number variant pathogenicity. *Nat.Commun.* 8:14366. doi: 10.1038/ncomms14366 [doi].

Richards, S., N. Aziz, S. Bale, D. Bick, S. Das, J. Gastier-Foster, W.W. Grody, M. Hegde, E. Lyon, E. Spector, 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. doi: 10.1038/gim.2015.30 [doi].

Rieber, N., M. Zapatka, B. Lasitschka, D. Jones, P. Northcott, B. Hutter, N. Jager, M. Kool, M. Taylor, P. Lichter, et al. 2013. Coverage bias and sensitivity of variant calling for four whole-genome sequencing technologies. *PLoS One.* 8:e66621. doi: 10.1371/journal.pone.0066621 [doi].

Riggs, E.R., E.F. Andersen, A.M. Cherry, S. Kantarci, H. Kearney, A. Patel, G. Raca, D.I. Ritter, S.T. South, E.C. Thorland, et al. 2019. Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). *Genet.Med.* doi: 10.1038/s41436-019-0686-8 [doi].

Riggs, E.R., D.M. Church, K. Hanson, V.L. Horner, E.B. Kaminsky, R.M. Kuhn, K.E. Wain, E.S. Williams, S. Aradhya, H.M. Kearney, et al. 2012. Towards an evidence-based process for the clinical interpretation of copy number variation. *Clin.Genet.* 81:403-412. doi: 10.1111/j.1399-0004.2011.01818.x [doi].

Rivera-Muñoz, E.A., L.V. Milko, S.M. Harrison, D.R. Azzariti, C.L. Kurtz, K. Lee, J.L. Mester, M.A. Weaver, E. Currey, W. Craigen, et al. 2018. ClinGen Variant Curation Expert Panel experiences and standardized processes for disease and gene-level specification of the ACMG/AMP guidelines for sequence variant interpretation. *Hum.Mutat.* 39:1614-1622. doi: 10.1002/humu.23645 [doi].

Roca, I., L. Gonzalez-Castro, H. Fernandez, M.L. Couce, and A. Fernandez-Marmiesse. 2019. Free-access copy-number variant detection tools for targeted next-generation sequencing data. *Mutat.Res.* 779:114-125. doi: S1383-5742(18)30037-1 [pii].

Rosenfeld, J.A., B.P. Coe, E.E. Eichler, H. Cuckle, and L.G. Shaffer. 2013. Estimates of penetrance for recurrent pathogenic copy-number variations. *Genet.Med.* 15:478-481. doi: 10.1038/gim.2012.164 [doi].

Ross, M.G., C. Russ, M. Costello, A. Hollinger, N.J. Lennon, R. Hegarty, C. Nusbaum, and D.B. Jaffe. 2013. Characterizing and measuring bias in sequence data. *Genome Biol.* 14:R51-2013-14-5-r51. doi: 10.1186/gb-2013-14-5-r51 [doi].

Rothberg, J.M., W. Hinz, T.M. Rearick, J. Schultz, W. Mileski, M. Davey, J.H. Leamon, K. Johnson, M.J. Milgrew, M. Edwards, et al. 2011. An integrated semiconductor device

enabling non-optical genome sequencing. *Nature.* 475:348-352. doi: 10.1038/nature10242 [doi].

Ruderfer, D.M., T. Hamamsy, M. Lek, K.J. Karczewski, D. Kavanagh, K.E. Samocha, Exome Aggregation Consortium, M.J. Daly, D.G. MacArthur, M. Fromer, and S.M. Purcell. 2016. Patterns of genic intolerance of rare copy number variation in 59,898 human exomes. *Nat.Genet.* 48:1107-1111. doi: 10.1038/ng.3638 [doi].

Sadedin, S.P., J.A. Ellis, S.L. Masters, and A. Oshlack. 2018. Ximmer: a system for improving accuracy and consistency of CNV calling from exome data. *Gigascience.* 7:10.1093/gigascience/giy112. doi: 10.1093/gigascience/giy112 [doi].

Sagath, L., V.L. Lehtokari, S. Valipakka, B. Udd, C. Wallgren-Pettersson, K. Pelin, and K. Kiiski. 2018. An Extended Targeted Copy Number Variation Detection Array Including 187 Genes for the Diagnostics of Neuromuscular Disorders. *J.Neuromuscul Dis.* 5:307-314. doi: 10.3233/JND-170298 [doi].

Salk, J.J., M.W. Schmitt, and L.A. Loeb. 2018. Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. *Nat.Rev.Genet.* 19:269-285. doi: 10.1038/nrg.2017.117 [doi].

Samarakoon, P.S., H.S. Sorte, B.E. Kristiansen, T. Skodje, Y. Sheng, G.E. Tjonnfjord, B. Stadheim, A. Stray-Pedersen, O.K. Rodningen, and R. Lyle. 2014. Identification of copy number variants from exome sequence data. *BMC Genomics.* 15:661-2164-15-661. doi: 10.1186/1471-2164-15-661 [doi].

Samarakoon, P.S., H.S. Sorte, A. Stray-Pedersen, O.K. Rodningen, T. Rognes, and R. Lyle. 2016. cnvScan: a CNV screening and annotation tool to improve the clinical utility of computational CNV prediction from exome sequencing data. *BMC Genomics.* 17:51-016-2374-2. doi: 10.1186/s12864-016-2374-2 [doi].

Sanger, F., S. Nicklen, and A.R. Coulson. 1977. DNA sequencing with chain-terminating inhibitors. *Proc.Natl.Acad.Sci.U.S.A.* 74:5463-5467. doi: 10.1073/pnas.74.12.5463 [doi].

Santos, M., M. Niemi, M. Hiratsuka, M. Kumondai, M. Ingelman-Sundberg, V.M. Lauschke, and C. Rodriguez-Antona. 2018. Novel copy-number variations in pharmacogenes contribute to interindividual differences in drug pharmacokinetics. *Genet.Med.* 20:622-629. doi: 10.1038/gim.2017.156 [doi].

Savarese, M., M. Johari, K. Johnson, M. Arumilli, A. Torella, A. Töpf, A. Rubegni, M. Kuhn, T. Giugliano, D. Gläser, et al. 2020. Improved Criteria for the Classification of Titin Variants in Inherited Skeletal Myopathies. *J.Neuromuscul Dis.* 7:153-166. doi: 10.3233/JND-190423 [doi].

Savarese, M., J. Sarparanta, A. Vihola, B. Udd, and P. Hackman. 2016. Increasing Role of Titin Mutations in Neuromuscular Disorders. *J.Neuromuscul Dis.* 3:293-308. doi: JND160158 [pii].

Schena, M., D. Shalon, R.W. Davis, and P.O. Brown. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science.* 270:467-470. doi: 10.1126/science.270.5235.467 [doi].

Schirmer, M., R. D'Amore, U.Z. Ijaz, N. Hall, and C. Quince. 2016. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics.* 17:125-016-0976-y. doi: 10.1186/s12859-016-0976-y [doi].

Schouten, J.P., C.J. McElgunn, R. Waaijer, D. Zwijnenburg, F. Diepvens, and G. Pals. 2002. Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res.* 30:e57. doi: 10.1093/nar/gnf056 [doi].

Schuster-Bockler, B., D. Conrad, and A. Bateman. 2010. Dosage sensitivity shapes the evolution of copy-number varied regions. *PLoS One.* 5:e9474. doi: 10.1371/journal.pone.0009474 [doi].

Sebat, J., B. Lakshmi, J. Troge, J. Alexander, J. Young, P. Lundin, S. Maner, H. Massa, M. Walker, M. Chi, et al. 2004. Large-scale copy number polymorphism in the human genome. *Science.* 305:525-528. doi: 10.1126/science.1098918 [doi].

Sedlazeck, F.J., P. Rescheneder, M. Smolka, H. Fang, M. Nattestad, A. von Haeseler, and M.C. Schatz. 2018. Accurate detection of complex structural variations using single-molecule sequencing. *Nat.Methods.* 15:461-468. doi: 10.1038/s41592-018-0001-7 [doi].

Serin Harmanci, A., A.O. Harmanci, and X. Zhou. 2020. CaSpER identifies and visualizes CNV events by integrative analysis of single-cell or bulk RNA-sequencing data. *Nat.Commun.* 11:89-019-13779-x. doi: 10.1038/s41467-019-13779-x [doi].

Sewry, C.A., J.M. Laitila, and C. Wallgren-Pettersson. 2019. Nemaline myopathies: a current view. *J.Muscle Res.Cell.Motil.* 40:111-126. doi: 10.1007/s10974-019-09519-9 [doi].

Shao, K., W. Ding, F. Wang, H. Li, D. Ma, and H. Wang. 2011. Emulsion PCR: a high efficient way of PCR amplification of random DNA libraries in aptamer selection. *PLoS One.* 6:e24910. doi: 10.1371/journal.pone.0024910 [doi].

Shen, Y., and B.L. Wu. 2009. Designing a simple multiplex ligation-dependent probe amplification (MLPA) assay for rapid detection of copy number variants in the genome. *J.Genet.Genomics.* 36:257-265. doi: 10.1016/S1673-8527(08)60113-7 [doi].

Shieh, P.B. 2018. Emerging Strategies in the Treatment of Duchenne Muscular Dystrophy. *Neurotherapeutics.* 15:840-848. doi: 10.1007/s13311-018-00687-z [doi].

Siepel, A., G. Bejerano, J.S. Pedersen, A.S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L.W. Hillier, S. Richards, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15:1034-1050. doi: gr.3715005 [pii].

Simpson, J.T., K. Wong, S.D. Jackman, J.E. Schein, S.J. Jones, and I. Birol. 2009. ABySS: a parallel assembler for short read sequence data. *Genome Res.* 19:1117-1123. doi: 10.1101/gr.089532.108 [doi].

Singleton, A.B., M. Farrer, J. Johnson, A. Singleton, S. Hague, J. Kachergus, M. Hulihan, T. Peuralinna, A. Dutra, R. Nussbaum, et al. 2003. alpha-Synuclein locus triplication causes Parkinson's disease. *Science.* 302:841. doi: 10.1126/science.1090278 [doi].

Sirugo, G., S.M. Williams, and S.A. Tishkoff. 2019. The Missing Diversity in Human Genetic Studies. *Cell.* 177:1080. doi: S0092-8674(19)30451-9 [pii].

Solinas-Toldo, S., S. Lampel, S. Stilgenbauer, J. Nickolenko, A. Benner, H. Dohner, T. Cremer, and P. Lichter. 1997. Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer.* 20:399-407. doi: 10.1002/(SICI)1098-2264(199712)20:43.0.CO;2-I [pii].

South, S.T., C. Lee, A.N. Lamb, A.W. Higgins, H.M. Kearney, and Working Group for the American College of Medical Genetics and Genomics Laboratory Quality Assurance Committee. 2013. ACMG Standards and Guidelines for constitutional cytogenomic microarray analysis, including postnatal and prenatal applications: revision 2013. *Genet.Med.* 15:901-909. doi: 10.1038/gim.2013.129 [doi].

Spence, J.E., R.G. Perciaccante, G.M. Greig, H.F. Willard, D.H. Ledbetter, J.F. Hejtmancik, M.S. Pollack, W.E. O'Brien, and A.L. Beaudet. 1988. Uniparental disomy as a mechanism for human genetic disease. *Am.J.Hum.Genet.* 42:217-226.

Srivastava, S., J.A. Love-Nichols, K.A. Dies, D.H. Ledbetter, C.L. Martin, W.K. Chung, H.V. Firth, T. Frazier, R.L. Hansen, L. Prock, et al. 2019. Meta-analysis and multidisciplinary consensus statement: exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders. *Genet.Med.* 21:2413-2421. doi: 10.1038/s41436-019-0554-6 [doi].

Stankiewicz, P., and J.R. Lupski. 2010. Structural variation in the human genome and its role in disease. *Annu.Rev.Med.* 61:437-455. doi: 10.1146/annurev-med-100708-204735 [doi].

Stankiewicz, P., and J.R. Lupski. 2002a. Genome architecture, rearrangements and genomic disorders. *Trends Genet.* 18:74-82. doi: S0168-9525(02)02592-1 [pii].

Stankiewicz, P., and J.R. Lupski. 2002b. Molecular-evolutionary mechanisms for genomic disorders. *Curr.Opin.Genet.Dev.* 12:312-319. doi: S0959437X02003040 [pii].

Stankiewicz, P., C.J. Shaw, J.D. Dapper, K. Wakui, L.G. Shaffer, M. Withers, L. Elizondo, S.S. Park, and J.R. Lupski. 2003. Genome architecture catalyzes nonrecurrent chromosomal rearrangements. *Am.J.Hum.Genet.* 72:1101-1116. doi: S0002-9297(07)60639-9 [pii].

Stankiewicz, P., H. Thiele, M. Schlicker, A. Cseke-Friedrich, S. Bartel-Friedrich, S.A. Yatsenko, J.R. Lupski, and I. Hansmann. 2005. Duplication of Xq26.2-q27.1, including SOX3, in a mother and daughter with short stature and dyslalia. *Am.J.Med.Genet.A.* 138:11-17. doi: 10.1002/ajmg.a.30910 [doi].

Stefansson, H., D. Rujescu, S. Cichon, O.P. Pietilainen, A. Ingason, S. Steinberg, R. Fossdal, E. Sigurdsson, T. Sigmundsson, J.E. Buizer-Voskamp, et al. 2008. Large recurrent microdeletions associated with schizophrenia. *Nature.* 455:232-236. doi: 10.1038/nature07229 [doi].

Straub, V., A. Murphy, B. Udd, and LGMD workshop study group. 2018. 229th ENMC international workshop: Limb girdle muscular dystrophies - Nomenclature and reformed classification Naarden, the Netherlands, 17-19 March 2017. *Neuromuscul.Disord.* 28:702-710. doi: S0960-8966(18)30214-1 [pii].

Sulonen, A.M., P. Ellonen, H. Almusa, M. Lepisto, S. Eldfors, S. Hannula, T. Miettinen, H. Tyynismaa, P. Salo, C. Heckman, et al. 2011. Comparison of solution-based exome capture methods for next generation sequencing. *Genome Biol.* 12:R94-2011-12-9-r94. doi: 10.1186/gb-2011-12-9-r94 [doi].

Szafranski, P., A.V. Dharmadhikari, E. Brosens, P. Gurha, K.E. Kolodziejska, O. Zhishuo, P. Dittwald, T. Majewski, K.N. Mohan, B. Chen, et al. 2013. Small noncoding differentially methylated copy-number variants, including lncRNA genes, cause a lethal lung developmental disorder. *Genome Res.* 23:23-33. doi: 10.1101/gr.141887.112 [doi].

Talevich, E., A.H. Shain, T. Botton, and B.C. Bastian. 2016. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput.Biol.* 12:e1004873. doi: 10.1371/journal.pcbi.1004873 [doi].

Talevich, E., and A.H. Shain. 2018. CNVkit-RNA: Copy number inference from RNA-Sequencing data. *bioRxiv.* doi: 10.1101/408534.

Tan, R., Y. Wang, S.E. Kleinstein, Y. Liu, X. Zhu, H. Guo, Q. Jiang, A.S. Allen, and M. Zhu. 2014. An evaluation of copy number variation detection tools from whole-exome sequencing data. *Hum.Mutat.* 35:899-907. doi: 10.1002/humu.22537 [doi].

Tanaka, N., A. Takahara, T. Hagio, R. Nishiko, J. Kanayama, O. Gotoh, and S. Mori. 2020. Sequencing artifacts derived from a library preparation method using enzymatic fragmentation. *PLoS One.* 15:e0227427. doi: e0227427.

Teo, S.M., Y. Pawitan, C.S. Ku, K.S. Chia, and A. Salim. 2012. Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics.* 28:2711-2718. doi: 10.1093/bioinformatics/bts535 [doi].

Thorvaldsdottir, H., J.T. Robinson, and J.P. Mesirov. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform.* 14:178-192. doi: 10.1093/bib/bbs017 [doi].

Trost, B., S. Walker, Z. Wang, B. Thiruvahindrapuram, J.R. MacDonald, W.W.L. Sung, S.L. Pereira, J. Whitney, A.J.S. Chan, G. Pellecchia, et al. 2018. A Comprehensive Workflow for Read Depth-Based Identification of Copy-Number Variation from Whole-Genome Sequence Data. *Am.J.Hum.Genet.* 102:142-155. doi: S0002-9297(17)30496-2 [pii].

Truty, R., J. Paul, M. Kennemer, S.E. Lincoln, E. Olivares, R.L. Nussbaum, and S. Aradhya. 2019. Prevalence and properties of intragenic copy-number variation in Mendelian disease genes. *Genet.Med.* 21:114-123. doi: 10.1038/s41436-018-0033-5 [doi].

Tsilfidis, C., A.E. MacKenzie, G. Mettler, J. Barcelo, and R.G. Korneluk. 1992. Correlation between CTG trinucleotide repeat length and frequency of severe congenital myotonic dystrophy. *Nat.Genet.* 1:192-195. doi: 10.1038/ng0692-192 [doi].

Turajlic, S., A. Sottoriva, T. Graham, and C. Swanton. 2019. Resolving genetic heterogeneity in cancer. *Nat.Rev.Genet.* 20:404-416. doi: 10.1038/s41576-019-0114-6 [doi].

Tuzun, E., A.J. Sharp, J.A. Bailey, R. Kaul, V.A. Morrison, L.M. Pertz, E. Haugen, H. Hayden, D. Albertson, D. Pinkel, et al. 2005. Fine-scale structural variation of the human genome. *Nat.Genet.* 37:727-732. doi: ng1562 [pii].

Udd, B., J. Partanen, P. Halonen, B. Falck, L. Hakamies, H. Heikkila, S. Ingo, H. Kalimo, H. Kaariainen, and V. Laulumaa. 1993. Tibial muscular dystrophy. Late adult-onset distal myopathy in 66 Finnish patients. *Arch.Neurol.* 50:604-608. doi: 10.1001/archneur.1993.00540060044015 [doi].

Untergasser, A., I. Cutcutache, T. Koressaar, J. Ye, B.C. Faircloth, M. Remm, and S.G. Rozen. 2012. Primer3--new capabilities and interfaces. *Nucleic Acids Res.* 40:e115. doi: gks596 [pii].

Valouev, A., J. Ichikawa, T. Tonthat, J. Stuart, S. Ranade, H. Peckham, K. Zeng, J.A. Malek, G. Costa, K. McKernan, et al. 2008. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome Res.* 18:1051-1063. doi: 10.1101/gr.076463.108 [doi].

Van der Auwera, G.A., M.O. Carneiro, C. Hartl, R. Poplin, G. Del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, et al. 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr.Protoc.Bioinformatics.* 43:11.10.1-33. doi: 10.1002/0471250953.bi1110s43 [doi].

Vasli, N., and J. Laporte. 2013. Impacts of massively parallel sequencing for genetic diagnosis of neuromuscular disorders. *Acta Neuropathol.* 125:173-185. doi: 10.1007/s00401-012-1072-7 [doi].

Veal, C.D., P.J. Freeman, K. Jacobs, O. Lancaster, S. Jamain, M. Leboyer, D. Albanes, R.R. Vaghela, I. Gut, S.J. Chanock, and A.J. Brookes. 2012. A mechanistic basis for amplification differences between samples and between genome regions. *BMC Genomics.* 13:455-2164-13-455. doi: 10.1186/1471-2164-13-455 [doi].

Vogelstein, B., and K.W. Kinzler. 1999. Digital PCR. *Proc.Natl.Acad.Sci.U.S.A.* 96:9236-9241. doi: 10.1073/pnas.96.16.9236 [doi].

Volk, A.E., and C. Kubisch. 2017. The rapid evolution of molecular genetic diagnostics in neuromuscular diseases. *Curr.Opin.Neurol.* 30:523-528. doi: 10.1097/WCO.0000000000000478 [doi].

Waldrop, M.A., M. Pastore, R. Schrader, E. Sites, D. Bartholomew, C.Y. Tsao, and K.M. Flanigan. 2019. Diagnostic Utility of Whole Exome Sequencing in the Neuromuscular Clinic. *Neuropediatrics.* 50:96-102. doi: 10.1055/s-0039-1677734 [doi].

Wang, K., M. Li, and H. Hakonarson. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38:e164. doi: 10.1093/nar/gkq603 [doi].

Wang, Q., C.S. Shashikant, M. Jensen, N.S. Altman, and S. Girirajan. 2017. Novel metrics to measure coverage in whole exome sequencing datasets reveal local and global non-uniformity. *Sci.Rep.* 7:885-017-01005-x. doi: 10.1038/s41598-017-01005-x [doi].

Watson, C.T., T. Marques-Bonet, A.J. Sharp, and H.C. Mefford. 2014. The genetics of microdeletion and microduplication syndromes: an update. *Annu.Rev.Genomics Hum.Genet.* 15:215-244. doi: 10.1146/annurev-genom-091212-153408 [doi].

Weaver, S., S. Dube, A. Mir, J. Qin, G. Sun, R. Ramakrishnan, R.C. Jones, and K.J. Livak. 2010. Taking qPCR to a higher level: Analysis of CNV reveals the power of high throughput qPCR to enhance quantitative resolution. *Methods.* 50:271-276. doi: 10.1016/j.ymeth.2010.01.003 [doi].

Weischenfeldt, J., O. Symmons, F. Spitz, and J.O. Korbel. 2013. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat.Rev.Genet.* 14:125-138. doi: 10.1038/nrg3373 [doi].

Wenger, A.M., H. Guturu, J.A. Bernstein, and G. Bejerano. 2017. Systematic reanalysis of clinical exome data yields additional diagnoses: implications for providers. *Genet.Med.* 19:209-214. doi: 10.1038/gim.2016.88 [doi].

Whitford, W., K. Lehnert, R.G. Snell, and J.C. Jacobsen. 2019. Evaluation of the performance of copy number variant prediction tools for the detection of deletions from whole genome sequencing data. *J.Biomed.Inform.* 94:103174. doi: S1532-0464(19)30092-9 [pii].

Wu, N., X. Ming, J. Xiao, Z. Wu, X. Chen, M. Shinawi, Y. Shen, G. Yu, J. Liu, H. Xie, et al. 2015. TBX6 null variants and a common hypomorphic allele in congenital scoliosis. *N.Engl.J.Med.* 372:341-350. doi: 10.1056/NEJMoa1406829 [doi].

Xu, Y., Z. Lin, C. Tang, Y. Tang, Y. Cai, H. Zhong, X. Wang, W. Zhang, C. Xu, J. Wang, et al. 2019. A new massively parallel nanoball sequencing platform for whole exome research. *BMC Bioinformatics.* 20:153-019-2751-3. doi: 10.1186/s12859-019-2751-3 [doi].

Yang, Y., D.M. Muzny, J.G. Reid, M.N. Bainbridge, A. Willis, P.A. Ward, A. Braxton, J. Beuten, F. Xia, Z. Niu, et al. 2013. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N.Engl.J.Med.* 369:1502-1511. doi: 10.1056/NEJMoa1306555 [doi].

Yao, R., C. Zhang, T. Yu, N. Li, X. Hu, X. Wang, J. Wang, and Y. Shen. 2017. Evaluation of three read-depth based CNV detection tools using whole-exome sequencing data. *Mol.Cytogenet.* 10:30-017-0333-5. eCollection 2017. doi: 10.1186/s13039-017-0333-5 [doi].

Yoon, S., Z. Xuan, V. Makarov, K. Ye, and J. Sebat. 2009. Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.* 19:1586-1592. doi: 10.1101/gr.092981.109 [doi].

Zarrei, M., J.R. MacDonald, D. Merico, and S.W. Scherer. 2015. A copy number variation map of the human genome. *Nat.Rev.Genet.* 16:172-183. doi: 10.1038/nrg3871 [doi].

Zenagui, R., D. Lacourt, H. Pegeot, K. Yauy, R. Juntas Morales, C. Theze, F. Rivier, C. Cances, G. Sole, D. Renard, et al. 2018. A Reliable Targeted Next-Generation Sequencing

Strategy for Diagnosis of Myopathies and Muscular Dystrophies, Especially for the Giant Titin and Nebulin Genes. *J.Mol.Diagn.* 20:533-549. doi: S1525-1578(17)30318-5 [pii].

Zhang, F., and J.R. Lupski. 2015. Non-coding genetic variants in human disease. *Hum.Mol.Genet.* 24:R102-10. doi: 10.1093/hmg/ddv259 [doi].

Zhang, L., W. Bai, N. Yuan, and Z. Du. 2019. Comprehensively benchmarking applications for detecting copy number variation. *PLoS Comput.Biol.* 15:e1007069. doi: 10.1371/journal.pcbi.1007069 [doi].

Zhao, M., Q. Wang, Q. Wang, P. Jia, and Z. Zhao. 2013. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics.* 14 Suppl 11:S1-2105-14-S11-S1. Epub 2013 Sep 13. doi: 10.1186/1471-2105-14-S11-S1 [doi].

Zhou, B., S.S. Ho, X. Zhang, R. Pattni, R.R. Haraksingh, and A.E. Urban. 2018. Whole-genome sequencing analysis of CNV using low-coverage and paired-end strategies is efficient and outperforms array-based CNV analysis. *J.Med.Genet.* 55:735-743. doi: 10.1136/jmedgenet-2018-105272 [doi].

SUPPLEMENTAL TABLES

Supplemental table 1: Genes in MYOcap targeted gene panel versions. For the versions after MYOcap v.1, only the newly added genes are displayed.

MYOcap v.1			v.2	v.3	v.4	v.5	v.6
ABHD5	*KCNJ2*	*PABPN1*	*AIFM1*	*ACTN2*	*ADSSL1*	*ACAD9*	*ATP2A2*
ACADS	*KCNQ1*	*PALB*	*ALG13*	*C10orf2*	*ALDOA*	*ACADM*	*INTS11*
ACADVL	*KLHL9*	*PDLIM3*	*ALG14*	*CASQ1*	*AMPD1*	*B4GAT1*	*MICU1*
ACTA1	*KY*	*PDLIM5*	*ALG2*	*CASQ2*	*TBCE*	*CAMSAP1*	*MMP8*
ACTN3	*LAMA2*	*PDLIM7*	*B3GALNT2*	*CELF2*	*B3GNT2*	*CAMSAP2*	*MSTO1*
ACVR1	*LAMP1*	*PFKM*	*B4GALNT1*	*HSPB2*	*BVES*	*CAVIN1*	*MYO18B*
AGL	*LAMP2*	*PGAM2*	*BICD2*	*GYG1*	*CHCHD10*	*CUL3*	*PIP4P1*
ANO5	*LARGE*	*PGK1*	*C20ORF72*	*HNRNPA1*	*CLN3*	*DNAJB2*	*PIP4P2*
ATP2A1	*LDB3*	*PGM1*	*CELF1*	*HNRNPA2B1*	*DNMT3B*	*DNAJC19*	*POPDC3*
B3GNT1	*LDHA*	*PLEC1*	*COL12A1*	*HNRPDL*	*ESRRG*	*DNAJC5*	*SARS2*
BAG3	*LDHB*	*PLEKHG4*	*DCST2*	*HSPB1*	*FAT1*	*DOK7*	*SLC25A32*
BIN1	*LIFR*	*PLN*	*DOLK*	*HSPB3*	*GAPDH*	*FAM111B*	*SLC52A1*
CACNA1A	*LMNA*	*PNPLA2*	*DPAGT1*	*HSPB6*	*GGPS1*	*FLAD1*	*SLC52A2*
CACNA1S	*MATR3*	*POMGNT1*	*DPM1*	*HSPG2*	*HACD1*	*FYCO1*	*SLC52A3*
CAPN3	*MBNL1*	*POMT1*	*ECEL1*	*LMOD3*	*HINT3*	*GOLGA2*	*TANGO2*
CAV3	*MBNL2*	*POMT2*	*EXOSC3*	*LPIN1*	*HNRNPDL*	*HADHA*	*ZNF33A*
CFL2	*MBNL3*	*PRKAG2*	*FBN2*	*SIL1*	*KCNJ18*	*HADHB*	
CHKB	*MEGF10*	*PTRF*	*GATM*	*ORAI1*	*KCNJ5*	*INPP5K*	
CKM	*MSTN*	*PYGM*	*GMPPB*	*POMGNT2*	*KLC1*	*JAG1*	
CLCN1	*MTM1*	*RYR1*	*GNB4*	*PUS1*	*KLHL40*	*LTBP4*	
CMYA5	*MTMR14*	*SCN4A*	*GOSR2*	*SGCE*	*KLHL41*	*OPTN*	
CNBP	*MURF1*	*SEPN1*	*GSN*	*SLC25A4*	*LARGE1*	*PIEZO1*	
CNTN1	*MURF2*	*SGCA*	*HEXB*	*SOD1*	*MB*	*PIEZO2*	
COL6A1	*MYBPC2*	*SGCB*	*HK1*	*SPEG*	*MGME1*	*POGLUT1*	
COL6A2	*MYBPC3*	*SGCD*	*HSPB8*	*SPTB*	*NALCN*	*PYROXD1*	
COL6A3	*MYH1*	*SGCG*	*ITPR1*	*TK2*	*PHKA1*	*SELENON*	
CPT2	*MYH2*	*SLC22A5*	*KBTBD10*	*XIRP1*	*PNPLA8*	*SLC25A42*	
CRYAB	*MYH3*	*SLC25A20*	*KBTBD5*	*XIRP2*	*POMK*	*SPTAN1*	
CSRP3	*MYH4*	*SOX10*	*LAMB1*	*YARS2*	*PREPL*	*SPTBN5*	
CUGBP1	*MYH7*	*SQSTM1*	*LIMS2*		*RRM2B*	*TAZ*	
DAG1	*MYH8*	*SRF*	*MARS*		*SACS*	*TMEM126B*	
DES	*MYL1*	*SYNE1*	*MYBPC1*		*SBDS*	*TP53INP2*	
DHPR	*MYL10*	*SYNE2*	*NEFL*		*SLC4A1*	*TRIP4*	
DMD	*MYL12A*	*SYNE3*	*PDK3*		*SLC7A10*	*TWNK*	
DMPK	*MYL12B*	*SYNPO2*	*PLEC*		*SMPX*	*ZAK*	
DNAJB6	*MYL2*	*TCAP*	*PLEKHG5*		*SPP1*	*ZBTB8B*	
DNM2	*MYL3*	*TIA1*	*POLG*		*SRPK3*	*ASCC2*	
DPM2	*MYL4*	*TMEM43*	*PTPLA*		*SUCLA2*		
DPM3	*MYL5*	*TMEM5*	*QDPR*		*TARDBP*		
DUX4	*MYL6*	*TMOD3*	*RBCK1*		*TPI1*		
DYSF	*MYL6B*	*TNNC1*	*SBF1*				
EMD	*MYL7*	*TNNC2*	*SGK196*				
ENO3	*MYL9*	*TNNI1*	*SLC5A2*				
ETFA	*MYLIP*	*TNNI2*	*SMCHD1*				
ETFB	*MYLK*	*TNNI3*	*STAC3*				
ETFDH	*MYLK2*	*TNNT1*	*STIM1*				
FBXO32	*MYLK3*	*TNNT3*	*STIM2*				
FHL1	*MYLK4*	*TPM1*	*TFG*				
FHL2	*MYLPF*	*TPM2*	*TIAL1*				
FKRP	*MYOM1*	*TPM3*	*TMEM55A*				
FKTN	*MYOM2*	*TRIM32*	*TMEM55B*				
FLNC	*MYOM3*	*TTN*	*TNPO3*				
GAA	*MYOT*	*VCP*	*TNXB*				
GBE1	*MYOZ1*	*VMA21*	*TOR1AIP1*				
GNE	*MYOZ2*		*TPP1*				
GTDC2	*MYOZ3*		*TRAPPC1*				
GYS1	*MYPN*		*TRAPPC11*				
ISCU	*NBR1*		*TRIM54*				
ISPD	*NEB*		*TRIM55*				

ITGA7	*NEBL*		*TRIM63*				
KBTBD13	*NTRK1*		*TTR*				
KCNE1	*OBSCN*		*UBA1*				
KCNE3	*OBSL1*		*VRK1*				

Supplemental table 2: Genes in MNDcap targeted gene panel versions. For the versions after MNDcap v.1, only the newly added genes are displayed (only v.3 got additions).

MNDcap v.1							v.3
AAAS	*ATXN8OS*	*EWSR1*	*INF2*	*NOTCH3*	*SCN4A*	*TNNI2*	*ADCY6*
AARS	*AVIL*	*EXOSC3*	*ITPR1*	*NT5C2*	*SCN9A*	*TNNT3*	*ADGRG6*
AARSD1	*BEAN1*	*FA2H*	*KARS*	*NTRK1*	*SEPT9*	*TOR1A*	*CACNA1B*
ABCA1	*BICD2*	*FAM134B*	*KCNA1*	*OPA1*	*SETX*	*TPM2*	*CAVIN1*
ABCD1	*BSCL2*	*FARSA*	*KCNC3*	*PDK3*	*SFN*	*TRIM2*	*CAVIN4*
ABHD12	*C10orf2*	*FBLN5*	*KIAA0196*	*PEX1*	*SGCE*	*TRPV4*	*CNTNAP1*
ADCK3	*C12orf65*	*FGD4*	*KIF1A*	*PEX7*	*SH3TC2*	*TSFM*	*COL13A1*
AFG3L2	*C9orf72*	*FGF14*	*KIF1B*	*PFN1*	*SIGMAR1*	*TTBK2*	*COQ8A*
AGRN	*CACNA1A*	*FIG4*	*KIF1C*	*PGAP1*	*SIL1*	*TTPA*	*ELP1*
AHNAK	*CACNA1S*	*FLRT1*	*KIF21A*	*PHB*	*SLC12A6*	*TTR*	*GBE1*
AIFM1	*CACNB4*	*FUS*	*KIF5A*	*PHOX2A*	*SLC1A3*	*TUBA8*	*GLDN*
ALDH3A2	*CCT5*	*FXN*	*L1CAM*	*PHYH*	*SLC25A20*	*TUBB3*	*GMPPB*
ALG14	*CD59*	*GALC*	*LAMB2*	*PIP5K1C*	*SLC25A4*	*TYMP*	*KCNJ18*
ALG2	*CHAT*	*GAN*	*LITAF*	*PLEKHG5*	*SLC25A5*	*UBA1*	*MAGEL2*
ALS2	*CHCHD10*	*GARS*	*LMNA*	*PLP1*	*SLC33A1*	*UBQLN1*	*MME*
AMPD2	*CHMP2B*	*GBA2*	*LRSAM1*	*PMM2*	*SLC52A1*	*UBQLN2*	*MRE11*
ANG	*CHRNA1*	*GDAP1*	*MAPT*	*PMP22*	*SLC52A2*	*USP8*	*MYH14*
ANO10	*CHRNB1*	*GFPT1*	*MARS*	*PNPLA6*	*SLC52A3*	*VAMP1*	*MYO9A*
AP4B1	*CHRND*	*GJB1*	*MATR3*	*POLG*	*SLC5A7*	*VAPB*	*PLEC*
AP4E1	*CHRNE*	*GJB3*	*MED25*	*POLG2*	*SMN1*	*VARS*	*PREPL*
AP4M1	*CHRNG*	*GJC2*	*MFF*	*PPP2R2B*	*SOD1*	*VCP*	*RETREG1*
AP5Z1	*COLQ*	*GLA*	*MFN2*	*PRKCG*	*SOX10*	*VPS37A*	*SAFB*
APOA1	*CTDP1*	*GLE1*	*MICAL1*	*PRPH*	*SPAST*	*VRK1*	*SCN11A*
APTX	*CYP7B1*	*GNB4*	*MPV17*	*PRPS1*	*SPG11*	*WDR48*	*SLC18A3*
AR	*DAO*	*GRN*	*MPZ*	*PRRT2*	*SPG20*	*WNK1*	*SMN2*
ARHGEF10	*DCTN1*	*HARS*	*MRE11A*	*PRX*	*SPG21*	*YARS*	*SNAP25*
ARL6IP1	*DDHD2*	*HEXA*	*MTMR2*	*PTRF*	*SPG7*	*ZFR*	*SPART*
ARSA	*DHTKD1*	*HEXB*	*MTPAP*	*QARS*	*SPTBN2*	*ZFYVE26*	*SPART-AS1*
ARSI	*DNAJB2*	*HINT1*	*MTTP*	*RAB3GAP2*	*SPTLC1*	*ZFYVE27*	*TIA1*
ASAH1	*DNM1L*	*HK1*	*MURC*	*RAB7A*	*SPTLC2*		*TWNK*
ATL1	*DNM2*	*HNRNPA1*	*MUSK*	*RAPSN*	*SQSTM1*		*UNC13A*
ATM	*DNMT1*	*HOXD10*	*MYH3*	*REEP1*	*SUCLA2*		*WARS*
ATP1A2	*DOK7*	*HSPB1*	*MYH8*	*RRM2B*	*SYNE1*		*WASHC5*
ATP2B3	*DPAGT1*	*HSPB3*	*NARS*	*RTN2*	*TAF15*		
ATP7A	*DYNC1H1*	*HSPB8*	*NDRG1*	*SACS*	*TARDBP*		
ATXN1	*EGR2*	*HSPD1*	*NEFH*	*SARS*	*TBP*		
ATXN10	*ENTPD1*	*IARS*	*NEFL*	*SBF1*	*TDP1*		
ATXN2	*ERBB3*	*IFRD1*	*NGF*	*SBF2*	*TECPR2*		
ATXN3	*ERLIN1*	*IGHMBP2*	*NIPA1*	*SCN10A*	*TFG*		
ATXN7	*ERLIN2*	*IKBKAP*	*NOP56*	*SCN1A*	*TK2*		

Supplemental table 3: PCR primers. Tm calculated with the Primer3 calculator.

Primer	Primer sequence	Tm (°C)
CACNA1A int40_F	5' CCTTCCAATTCCACGCAGAACTG 3'	62.21
CACNA1A int38_R	5' GTGAGCTATGTTTGTGCCACGG 3'	62.32
CAPN3 ex1_F	5' GACCTTCTGATGGGCTTTCA 3'	57.50
CAPN3 ex1_R	5' CTCTCCTCCCTGCTTCACAC 3'	59.75
CAPN3 ex2_F	5' ACTCCGTCTCAAAAAAATACCT 3'	56.20
CAPN3 ex2_R	5' ATTGTCCCTTTACCTCCTGG 3'	58.0
CAPN3 ex8_F	5' CCCCAGCACACTTGTGATTA 3'	57.80
CAPN3 ex8_R	5' ATCCTTCCTTTCCAGCCAAT 3'	56.77

CAPN3 ex9_F	5' CCTGCTTCCTTAATTCCTCCATTTT 3'	63.80
CAPN3 ex9_R	5' CTCTTCCCCACCCTTACCCTTCT 3'	64.90
COL6A1 int8_F	5' TCCTGCTCCTCCCATGTGTTG 3'	62.07
COL6A1 int13_R	5' AGTGGGTAAACTGAGGCCAATCA 3'	61.85
COL6A3 ex38_F	5' ATGGGTCGATGTTGCAGATGTCT 3'	62.26
COL6A3 ex31_R	5' GGTGAACCTGGGCTAAATGGAAC 3'	61.68
DMD ex41_F	5' AGTTGAGTCTTCGAAACTGAGCA 3'	60.28
DMD ex41_R	5' GGCCCTGTATTGGTTTTGCTCAA 3'	61.88
DMD ex42_F	5' CCATGTGAAAGTCAAAATGCCATCA 3'	60.80
DMD ex42_R	5' ATCACTCATGTCTCACAAGCCCT 3'	61.65
DMD ex43_F	5' ACCCTTGTCGGTCCTTGTACATT 3'	61.83
DMD ex43_R	5' CAACAAAGCTCAGGTCGGATTGA 3'	61.36
DMD ex44_F	5' TTTCCATCACCCTTCAGAACCTG 3'	60.50
DMD ex44_R	5' TGAGAAATGGCGGCGTTTTCATT 3'	62.17
DMD ex45_F	5' TGCCTTTCACCCTGCTTATAATCT 3'	60.08
DMD ex45_R	5' TTGGGAAGCCTGAATCTGCG 3'	60.68
DMD ex46_F	5' CAATGTTATCTGCTTCCTCCAACCA 3'	61.33
DMD ex46_R	5' TTTGTGTCCCAGTTTGCATTAACAA 3'	60.34
DMD ex48_F	5' CCCTACCTTAACGTCAAATGGTCC 3'	61.16
DMD ex48_R	5' CCAGAGCTTTACCTGAGAAACAAGG 3'	61.54
DMD ex49_F	5' GCAAATGTACAACAGGGGAAGCA 3'	61.87
DMD ex49_R	5' GCAGTTCAAGCTAAACAACCGGA 3'	61.85
DMD ex55_F	5' CGGAAATGCCTGACTTACTTGCC 3'	62.27
DMD ex55_R	5' CGAGAGGCTGCTTTGGAAGAAAC 3'	62.20
DMD ex56_F	5' ATGTGAGATACCAGTTACTTGTGCT 3'	60.05
DMD ex56_R	5' TCCGATGATGCAGTCCTGTTACA 3'	61.69
MYOM1 post-ex36R	5' CCACTCGGACAAAGAAGCTGAAT 3'	61.11
MYOM1 pre-ex35F	5' CTTGTCCCACTTGCCTTTCATCC 3'	61.93
SGCD ex1_F	5' GCTGTGTGGAGAATGGCTGAAAA3'	61.86
SGCD ex1_R	5' GACTGCTTTGAAACCGTACTCCG 3'	61.94
SGCD ex5_F	5' CCCCTTGGAGAGTTGTAATG 3'*	55.42
SGCD ex6_F	5' GATGAGACTAATGGTGTTTT 3'*	50.73
SGCD ex6_R	5' AAAATGTACACAGTAGCATC 3'*	51.18
TTN del_F	5' TAGGGAATGCTGGCGATATGGTT 3'	61.84
TTN del_R	5' TCTCCAAGCCACTCACAGATCAG 3'	61.94

* = primer from the LoVD database (https://www.lovd.nl/)