An Extended Targeted Copy Number Variation Detection Array Including 187 Genes for the Diagnostics of Neuromuscular Disorders

Lydia Sagatha,∗, Vilma-Lotta Lehtokari, Salla Välipakka, Bjarne Udda, Carina Wallgren-Pettersson, Katarina Pelin, and Kirsi Kiiski

The Folkhälso Institute of Genetics and the Department of Medical and Clinical Genetics, Medicum, University of Helsinki, Finland

Neuromuscular Research Center, Tampere University and University Hospital, Finland

Department of Neurology, Vaasa Central Hospital, Finland

Faculty of Biological and Environmental Sciences, Molecular and Integrative Biosciences Research Programme, University of Helsinki, Finland

Abstract

Background: Our previous array, the Comparative Genomic Hybridisation design (CGH-array) for nemaline myopathy (NM), named the NM-CGH array, revealed pathogenic copy number variation (CNV) in the genes for nebulin (NEB) and tropomyosin 3 (TPM3), as well as recurrent CNVs in the segmental duplication (SD), i.e. triplicate, region of NEB (TRI, exons 82–89, 90–97, 98–105). In the light of this knowledge, we have designed and validated an extended CGH array, which includes a selection of 187 genes known to cause neuromuscular disorders (NMDs).

Objective: Our aim was to develop a reliable method for CNV detection in genes related to neuromuscular disorders for routine mutation detection and analysis, as a much-needed complement to sequencing methods.

Methods: We have developed a novel custom-made 4 × 180 k CGH array for the diagnostics of NMDs. It includes the same tiled ultra-high density coverage of the 12 known or putative NM genes as our 8 × 60 k NM-CGH-array but also comprises a selection of 175 additional genes associated with NMDs, including titin (TTN), at a high to very high coverage. The genes were divided into three coverage groups according to known and potential pathogenicity in neuromuscular disorders.

Results: The array detected known and putative CNVs in all three gene coverage groups, including the repetitive regions of NEB and TTN.

Conclusions: The targeted neuromuscular disorder 4 × 180 k array-CGH (NMD-CGH-array v1.0) design allows CNV detection for a broader spectrum of neuromuscular disorders at a high resolution.

Keywords: Array comparative genomic hybridisation (aCGH), copy number variation (CNV), nebulin, NEB, titin, TTN, NM-CGH, NMD-CGH

INTRODUCTION

Congenital myopathies are hereditary disorders within the heterogeneous group of neuromuscular disorders (NMDs). They may be caused by mutations, including copy number variations (CNVs), in genes...
coding for structural proteins of the sarcomere, genes for membrane or matrix proteins, or genes affecting muscular development by playing a role in e.g. protein turnover or muscle differentiation [1]. The wide array of causative genes and mutations leads to a widely varying group of disorders; they may be passed on according to all modes of inheritance and they display a multitude of clinical phenotypes, with variation in progression, severity and distribution of muscle weakness [2, 3].

It is important to note that the phenotypes of congenital myopathies may overlap with each other and with other neuromuscular disorders, and that different mutations in the same gene may cause various phenotypes, thus challenging diagnostics. Specifically designed follow-up and specific putative future therapies can be administered only when the correct molecular diagnosis and the underlying causative mechanisms of the disorder have been determined [3, 4].

CNVs are structural changes of the genome, in which a certain genomic sequence is present in the studied genome at an aberrant copy number compared with the reference genome. CNVs were initially described as variations of at least 1 kb in size [5], but modern technology has allowed CNV detection at a higher resolution, and the minimum size for a deleted or duplicated sequence has been suggested to be as small as 50 bp [6]. Estimations are that de novo locus-specific CNVs have a higher mutation rate compared with nucleotide substitutions [5].

While thousands of regions are prone to benign CNVs [7], variants that interfere with normal gene expression or function can be potentially pathogenic [8]. For example, Duchenne muscular dystrophy (DMD; MIM ID: DMD #310200) is commonly caused by CNVs in the DMD gene [9, 10]. Furthermore, pathogenic CNVs have been shown to contribute to roughly 6.4% of nemaline myopathy cases (NM; MIM IDs: NEM1 #609284, NEM2 #256030, NEM3 #161800, NEM4 #609285, NEM5 #605355, NEM6 #609273, NEM7 #610687, NEM8 #615348, NEM9 #615731, NEM10 #616165, NEM11 #617336) [11].

Array Comparative Genomic Hybridisation (aCGH) constitutes the standard method for research and diagnostics into disorders with pathogenetic mechanisms involving CNVs. Copy number variation analysis methods based on Next-Generation Sequencing (NGS) are evolving [12], but the precision and reliability of aCGH remains to be matched by computer algorithms.

Especially regions of segmental duplication (SD) pose a challenge to NGS based CNV analysis methods [13], and are in fact often excluded from commercial CGH arrays and sequencing panels. Such regions are found both in the nebulin gene (NEB) [14] and the titin gene (TTN) [15], the products of which are enormous sarcomeric proteins. In NEB, this region holds a threefold repetition of an 8-exon block (exons 82–89, 90–97, 98–105), referred to as the triple-locus (TRI). The NEB TRI region is known to harbour CNVs – gains and losses of one 8-exon block are considered benign, while gains of more than one 8-exon block are considered to be pathogenic [16]. In TTN, the SD constitutes a threefold repetition of 9 exons (exons 166–174, 175–183, and 184–192) within the PEVK segment of the gene, followed by fourth copies of the first two exons of the repeated block [15]. CNVs in this type of region are all but rare, and may potentially be pathogenic [16].

Here we describe our custom targeted neuromuscular disorder CGH array (NMD-CGH array v1.0), targeting 187 different genes related to NMDs. As NM can be caused by CNVs [11, 17, 18], especially common in NEB, we hypothesised that CNVs may be disease-causing also in other genes related to NMDs – consequently, that there probably are undiscovered pathogenic CNVs to be identified in these genes.

The array targets a wide variety of genes related to NM and other NMDs. The design includes our previously published, targeted 8 x 60k nemaline myopathy CGH-array (NM-CGH array) [11] in an updated version. In addition to the 11 published and one unpublished NM-causing gene; ACTA1, NEB, TPM3, TPM2, TNNT1, CFL2, KBTBD13, KLHL40, KLHL41, LMOD3, MYPN and YBX3, the array includes 175 carefully chosen genes related to other NMDs – amongst these, the TTN, dystrophin (DMD), the ryanodine receptor (RYRI), and obscurin (OBSCN) genes.

**METHODS**

**Samples**

Forty-two DNA samples from 19 different clinicians in 11 countries were acquired for the validation of the array. The 42 DNA samples included 3 positive controls with known CNVs, 9 blind cross-validation samples, and 30 patient samples. In addition, we acquired 4 healthy control samples, of which 3 from the Finnish Red Cross Blood Service and one commercial reference sample (Male reference
DNA, Promega Corporation, Madison, WI, USA). (Table 1).

The study was approved by the local Ethics Committees and samples were obtained according to the Declaration of Helsinki of 1975.

**Microarray design, protocol and data analysis**

The CGH array was designed with the support of Oxford Gene Technology IP Limited Ltd (Oxford, UK) (Human reference sequence, GRCh37/Hg19). The array covers 187 genes associated with muscular disorders, divided into 3 different categories of highly targeted levels.

The array includes the known NM genes (ACTA1, CFL2, KBTBD13, KLHL40, KLHL41, LMOD3, NEB, TNNT1, TNNT3, TPM2, TPM3, MYPN) and 176 known or candidate neuromuscular disorder genes, including those of the 1st version of the targeted NGS MyoCap panel [19]. The current number of targeted probes chosen for the design is 150,300/180,000, leaving space for future additions and updates.

The genes have been divided into three gene coverage groups of varying coverage according to the known and potential NMD-causing pathogenicity harboured by the genes. Coverage group 1 includes the NM genes at the same ultra-highly tiled coverage as the previously validated and published 8 × 60 k NM-CGH array [11]. Coverage group 2 consists of 29 other characterised muscle disorder genes (ANO5, BAG3, BIN1, CRYAB, DES, DNM2, DYSF, FLNC, GNE, HSPB8, KLHL9, LDB3, MTM1, MYBPC1, MYBPC2, MYH1, MYH13, MYH2, MYH3, MYH4, MYH7, MYH8, MYOT, OBSCN, OBSL1, RYR1, SEPN1, TIA1, TTN) at a very high coverage. The remaining 147 genes (Supplementary Table 1) of coverage group 3 were also targeted at a high level of coverage, with approximately 4 probes covering each exon. The coverage group 2 and 3 genes are either known neuromuscular disorder genes or strong candidates for such. Of these, we selected genes coding for structural proteins of the sarcomere to be targeted at group 2 level as their putative impact on muscular structure is clearer than that of regulatory genes such as transcription factors, which were targeted at group 3 level. Exceptions to this rule are genes that are known to harbour pathogenic or putative pathogenic CNVs in group 2.

In all the groups, a minimum of one probe per exon and per intron was targeted in each gene. Flanking regions were included 25 kb upstream and downstream of the targeted genes. The coverage level of the three gene groups is presented in Table 2.

In addition, the backbone coverage was increased from approximately 1 probe per Mb to 1 probe per 300 kb, in order to allow for the detection of large genomic rearrangements, as well as better alignment throughout the genome. The 12,300 backbone probes are spaced evenly across the non-targeted regions.

The USCS’ Genome Browser (https://genome.ucsc.edu/) online application was used to visualise the probe distribution in the designing process. All genes were targeted at least at the exon level. In addition to the Oxford Gene Technology in-house probes, new custom probes were designed for the regions of interest, especially in the regions of repetitive sequences.

Labelling, hybridisation, scanning and analysis of the array was performed according to the manufacturer’s protocol (Oxford Gene Technology Ltd, ISCAY array, Cytosure Genomic DNA labelling kit protocol, 4 × 180 k format, code 020020). We used a commercial male DNA reference for all the studied samples (Promega Corporation). The array was scanned using the Agilent DNA Microarray Scanner G2505C with 2 μm resolution, Agilent Scan Control version A.8.5.1 (Agilent Technologies Inc, Santa Clara, CA, USA) and normalised and further analysed in Feature Extraction software v.12.0.0.7 (Agilent Technologies). The Cytosure Software v.4.6.85 (Hg19) (Oxford Gene Technology Ltd) was used for graphic analysis of the data. A minimum of 5 probes was used as a threshold for making a positive call for an aberration. The array profiles and aberration calls were also manually checked.

**Validation**

The initial validation of the NMD-CGH array was made using 4 healthy controls and 3 positive controls (identified with the NM-CGH array). In addition, we performed a blind cross-validation process.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>The representative numbers of samples and families included in the validation process</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Samples</strong></td>
</tr>
<tr>
<td>Healthy controls</td>
<td>4</td>
</tr>
<tr>
<td>Positive controls</td>
<td>3</td>
</tr>
<tr>
<td>Cross-validation samples</td>
<td>9</td>
</tr>
<tr>
<td>Patient samples, of which</td>
<td></td>
</tr>
<tr>
<td>First mutation previously identified</td>
<td>10</td>
</tr>
<tr>
<td>Candidate mutation previously identified</td>
<td>7</td>
</tr>
<tr>
<td>Both mutations unidentified</td>
<td>13</td>
</tr>
<tr>
<td>Altogether</td>
<td>46</td>
</tr>
</tbody>
</table>
with 9 samples. In these, CNVs had been detected with 4 different NGS-based CNV detection algorithms (CoNIFER v0.2.2, XHMM v1.1, ExomeDepth v1.1.10 and CODEX v1.4.0). The CNV detection by algorithms was performed as described in Välipakka et al., 2017 [20].

Subsequently, we ran 30 patient samples from 20 families, which had not previously been analysed for CNVs by any method.

RESULTS

The validation runs on the 7 control samples confirmed the coverage, specificity and resolution of the array and the quality of the probes chosen. In the 4 healthy controls (samples 9–12), we identified a gain in the *TTN* SD region (coverage group 2) in 1 sample. In the 3 positive controls (samples 1-2 and 8), we identified 1 one-copy loss in the *NEB* TRI region (coverage group 1) and 3 losses in the *TTN* SD region, as well as the known pathogenic changes in *NEB* (coverage group 1) and *DMD* (coverage group 3).

In the blind cross-validation samples (16–23, 46), we were able to detect the expected CNVs. The changes include a partial duplication of the adjacent coverage group 2 genes *MYH6* and *MYH7* (Fig. 1A) and duplications in coverage group 3 genes *MYOM* and *PYGM* (Fig. 1B-C). Moreover, deletions in the genes *TTN* (Fig. 1D), *DMD*, *LDHB*, and the upstream region of *COL6A3* were detected (all but *TTN* belong to coverage group 3). A supposedly benign gain was detected in the *TTN* SD region of one sample (Fig. 1F). The NGS-based analysis programmes had flagged one change that was not detected on the array and was concluded to be a false positive. The cross-validation process confirmed the ability of the array to detect CNVs in all gene coverage groups.

The diagnostic runs performed on patient samples (samples 3–7, 13–15, 24–45) revealed two aberrations of unknown significance in coverage group

---

Table 2

The coverage of exon, intron and flanking regions in the three subgroups

<table>
<thead>
<tr>
<th>Coverage by gene coverage group</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exon</td>
<td>10 bp tiling, +/-</td>
<td>Adjacent probes, +/-</td>
<td>~4 probes/exon</td>
</tr>
<tr>
<td>Intron</td>
<td>20 bp tiling, +/-</td>
<td>Adjacent probes, +/-</td>
<td>~1–3 probes/intron</td>
</tr>
<tr>
<td>Flanking</td>
<td>±25 kb</td>
<td>±25 kb</td>
<td>±25 kb</td>
</tr>
</tbody>
</table>

*+/- signifies probe targeting on both the forward and the reverse strand. ± signifies the region covered both upstream and downstream of the targeted genes.*

---

Fig. 1. A demonstration of the array results. *NEB* is targeted at Group 1 level, *TTN* and *MYH7* are targeted at Group 2 level and *MYOM1* and *PYGM* are targeted at Group 3 level.
The prevalence of gains and losses in the NEB triplicate region (TRI) and TTN segmental duplication (SD) region in the different sample categories.

<table>
<thead>
<tr>
<th></th>
<th>Positive controls</th>
<th>Healthy controls</th>
<th>Patients and blind validation samples</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEB TRI gain</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>NEB TRI loss</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TTN SD gain</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>TTN SD loss</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>6</td>
</tr>
</tbody>
</table>

2 genes, one upstream deletion of ANO5 and one upstream duplication of OBSLI partially covering the gene DNPEP. Benign variation of the NEB TRI region was detected in 2 samples altogether, an example of which is presented in Fig. 1E. Copy number variation in the TTN SD region was seen in 8 diagnostic samples, of which 3 were losses and 5 were gains.

In addition to the known and putative pathogenic aberrations and known benign variations detected, the novel NMD-CGH array revealed previously unrecorded variation in the TTN SD region (Table 4). Variation in the TTN SD region was seen in 13 samples altogether. These findings warrant further studies, and will be the subject of another publication.

All changes identified, including the controls and cross-validation samples, are presented in Table 3. To date, we have run around 410 samples from approximately 300 families with NM or related disorders using the NM-CGH or NMD-CGH-array. The yield of pathogenic CNVs in these samples has been altogether around 13%, of which 4.5% are pathogenic CNVs within the NEB TRI region.

DISCUSSION

The design of the novel NMD-CGH array allows us to reliably identify CNVs in genes associated with muscle disorders, on a resolution varying from circa 70 bp (unpublished data) to the exon level, depending on the gene coverage group. The design can and will be updated based on new findings: we can move genes to a higher-tiled coverage group, or add altogether new genes to the array. Any recurrent, unclear or putatively pathogenic variation identified on our array or by a different method can be basis enough to move a gene to the higher coverage group. It is thus a dynamic method for the research and diagnostics of CNVs in genes causing neuromuscular disorders.

The human genome harbours more than 1000 regions prone to benign variation in terms of copy number, and CNVs have already been proven to constitute part of the mutation spectra behind muscle disorders, e.g. NM [1, 17, 18]. In patients lacking a molecular diagnosis for their disorders or in whom only one recessive mutation has been identified, the NMD-CGH array allows for precise and reliable detection of deletions and duplications in a wide range of NMD genes and NMD candidate genes. The genes were carefully picked out and divided into the coverage groups based on the current knowledge of proven and putative NMD genes, and the resolution at which these genes and their CNVs can be detected on the array.

Despite advancements in NGS data analysis of CNVs, aCGH still remains the gold standard method for CNV detection and analysis, especially in diagnostics, due to the method being well-established and reliable, and its quick sample turnover time. The repetitive nature of the SD regions in certain genes, e.g. NEB and TTN, present challenges for analysis by conventional NGS methods. Algorithms can predict CNVs, but whole-exome, targeted exome, or whole-genome data are not as clear, precise and sensitive as a well-designed CGH array. The number of false positives and false negatives is higher, and the findings still require validation by other methods, such as Sanger sequencing.

The repetitive areas may harbour CNVs in unconventional numbers [16], with the baseline deviating from the normal diploid copy number of two, and these can be difficult to determine using NGS-based methods and read counts only. In addition, the CNV breakpoints may be pinpointed with higher precision using a densely tiled aCGH design compared with sequence-based methods. However, for determining exact breakpoints at the base-pair level, Sanger sequencing is still the only reliable method.

Our array allows for the detection of CNVs, not only in NMD genes in general, but also in the repetitive NEB TRI region [16], and the SD region of TTN. These regions are typically excluded from commercial CGH-array designs because of the difficulty their repetitive nature imposes on the analysis. In terms of the NEB TRI region, our array has been validated [16], and we are currently validating the analysis of the SD region of TTN.

Based on our results so far, we expect the yield of pathogenic CNVs to be similar in range (∼13%) as more samples will be run.
Table 3

Copy number variations identified on the NMD-CGH-array during the validation process and the subsequent diagnostic runs

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Gene</th>
<th>Variation</th>
<th>Size</th>
<th>Coverage group</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 2, 8, 13</td>
<td><strong>TTN</strong> SD loss</td>
<td>Chr2:hg1:9:g.(179,516,753–179,517,001) (179,528,302–179,528,842)del</td>
<td>11,301–12,089 bp</td>
<td>2</td>
</tr>
<tr>
<td>5, 9, 19, 25</td>
<td><strong>TTN</strong> SD gain</td>
<td>Chr2:hg1:9:g.(179,516,799–179,517,346) (179,528,170–179,529,256)dup</td>
<td>10,824–12,457 bp</td>
<td>2</td>
</tr>
<tr>
<td>29</td>
<td><strong>NEB TR1</strong> loss</td>
<td>Chr2:hg1:9:g.(152,433,608–152,434,494) (152,464,242–152,464,694)del</td>
<td>29,748–31,086 bp</td>
<td>1</td>
</tr>
<tr>
<td>25</td>
<td><strong>NEB TR1</strong> gain</td>
<td>Chr2:hg1:9:g.(152,433,608–152,434,494) (152,466,075–152,466,085)dup</td>
<td>31,581–32,477 bp</td>
<td>1</td>
</tr>
<tr>
<td>3, 38</td>
<td><strong>OBSL1</strong> upstream duplication</td>
<td>Chr2:hg1:9:g.(220,438,234–220,439,293) (220,460,509–220,569,618)dup</td>
<td>21,216–131,384 bp</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td><strong>NEB</strong> partial deletion</td>
<td>Chr2:hg1:9:g.(152,434,836–152,435,622) (152,554,712–152,561,404)del</td>
<td>109,090–126,568 bp</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td><strong>LDBH</strong> partial deletion</td>
<td>Chr12:hg1:9:g.(21,803,624–21,806,090) (21,811,051–21,812,528)del</td>
<td>4,961–8,904 bp</td>
<td>3</td>
</tr>
<tr>
<td>17</td>
<td><strong>DMD</strong> partial deletion</td>
<td>ChrX:hg1:9:g.(32,843,289–31,086,990) (31,313,100–31,327,408)del</td>
<td>226,110–484,119 bp</td>
<td>3</td>
</tr>
<tr>
<td>18</td>
<td><strong>MYOM1</strong> partial duplication</td>
<td>Chr18:hg1:9:g.(3,155,151–3,155,958) (3,653,512–3,815,684)dup</td>
<td>497,527–660,533 bp</td>
<td>3</td>
</tr>
<tr>
<td>20</td>
<td><strong>PYGM</strong> duplication</td>
<td>Chr1:hg1:9:g.(64,155,908–64,440,343) (64,553,037–64,758,928)dup</td>
<td>112,694–603,020 bp</td>
<td>3</td>
</tr>
<tr>
<td>36</td>
<td><strong>ANO5</strong> upstream deletion</td>
<td>Chr11:hg1:9:g.(21,579,750–21,931,390) (22,207,427–22,208,122)del</td>
<td>276,237–628,372 bp</td>
<td>2</td>
</tr>
</tbody>
</table>
CONCLUSIONS AND FUTURE PROSPECTS

The NMD-CGH-array constitutes a dependable CNV analysis method for the diagnostics of NMDs. The use of the NMD-CGH array will continue in our laboratory as a first tier diagnostic tool for patients and families whose disease-causing variants remain unidentified.

The recurrent CNV in the SD region of TTN will be the topic of a further project. Preliminary data suggest that the region is highly prone to variation in a similar fashion as the NEB TRI region. It is possible that this region, too, will be found in some patients to contain modifying or pathogenic CNVs.

ACKNOWLEDGMENTS

We thank Dr. Nina Horelli-Kuitunen for sharing samples for the validation process, Marco Savarese, Mridul Johari, and Merja Soininen for their contributions to the validation process, and Marilotta Turunen and Sampo Koivunen for outstanding technical assistance. This study was supported by grants from Muscular Dystrophy UK, the Sigrid Jusélius Foundation, the Folkhälso Research Foundation, the Finska Läkaresällskapet, the Medicinska understödsforeningen Liv och Hälsa, l’Association Française contre les Myopathies, the Jane and Aatos Erkko Foundation, the Academy of Finland (no. 138491, B.U.), and the Orion Research Foundation sr.

CONFLICT OF INTEREST

The authors declare no conflict of interest to report.

AUTHOR CONTRIBUTIONS

LS ran the array hybridisations, carried out analysis and interpretation of the data and prepared the manuscript. LS and KK designed the extended 4 x 180 k NMD-CGH-array. SV performed the NGS-based CNV-analysis for the blind cross-validation. BU provided samples for the blind cross-validation. KK set up the array method in the laboratory and prepared the manuscript. V-LL, LS, KK and SV selected samples for the array. CW-P was responsible for the project as PI. KP was responsible for the project as Co-PI. All authors have participated in the planning of the study, interpretation of the results, and in the writing of the manuscript.

SUPPLEMENTARY MATERIAL

The supplementary material is available in the electronic version of this article: http://dx.doi.org/10.3233/IND-170298.

REFERENCES


