1. Executive Summary:

Inherited Neuromuscular Disorders (NMDs) form a very large and heterogeneous group of genetic diseases that cause progressive degeneration of the muscles and/or motor nerves. The overall prevalence of NMDs is very difficult to evaluate, but one can estimate that around 1 out of 1000 people may have a disabling inherited neuromuscular disease, namely Duchenne/Becker muscular dystrophies (DMD/BMD), limb girdle muscular dystrophies (LGMD), congenital muscular dystrophies (CMD), and hereditary motor-sensory neuropathies or Charcot-Marie-Tooth neuropathies (CMT). The precise diagnosis of NMDs requires a conjunction of extensive clinical examination and targeted complementary tests: biological analyses, electromyography, imaging, and histological analysis of biopsies. Thus, a differential molecular genotyping by a gene by gene approach is required, which is until now highly complex, cost-effective and time consuming (two weeks to one year). As a consequence, many patients remain devoid of genetic confirmation of their disease. To date, this proportion amounts to 30 to 40 % of NMDs. The aim of NMD-Chip project, initiated by French Association against Myopathy (AFM) and several European diagnostic laboratories, is to design, develop and validate new sensitive high throughput DNA arrays to efficiently diagnose patients affected by NMDs. The tools originating from this project are designed to assess all known genes implied in a group of disease at one time, as well as to efficiently analyse chip data through optimised read-out bioinformatics tools, within 72hrs to one week. Beside the development of these new high throughput molecular diagnostics tools, NMD-Chip also fosters the knowledge of NMDs by accelerating new disease causing mutations discovery, using a candidate gene approach. Indeed, the scientific strategy was to design 4 types of chips, 2 for known NMD genes testing, and 2 for candidate genes exploration. In each case, both CGH arrays to detect CNV (insertions or deletions), as well as Sequence Capture arrays for massive re-sequencing and point mutation detection, have been set up.

Today, the two first generations of CGH NMD-chips including all the known genes have been validated, and are proposed to be spread to diagnostics workflow to progressively replace the current techniques. The capture chips have been compared with the "in-solution" captures tools that have emerged as an alternative solution during the two past years. This part of the work has not completely been achieved, and it is still unclear whether one approach is better than the other one. The CGH chips have been validated on different previously characterised DNA with several deletions or insertions in LAMA2, DYSF, CAPN, DMD, PMP22, COL6A genes and others. One problem remains after a round of validation among the partners for some particular genes (for instance, EMD): the design must be improved since there is a lot of a variation in the results obtained so far. However from this part of the project, several uncharacterised patients have been provided with a molecular diagnosis, thus making the project a success. For the research part of the project (candidate gene exploration), the CGH chips have been designed and validated too. The successes obtained with control DNA have convinced one partner to explore a cohort of 33 patients with an uncharacterised LGMD, with at the moment one new CNV candidate being further analysed. As regards the DNA captures, different approaches have been initialised, combining the commercial options (exome) with dedicated probes, from different providers. This part of the work has been firmly validated only with control known variations, which is a success already.

In the meantime, a large Reference Material database has been created to collect the most relevant samples from each partner to be used as positive controls on the chips.

2. Project Context and Objectives:

Inherited Neuromuscular Disorders (NMDs) form a large and very heterogeneous group of genetic diseases that cause progressive degeneration of the muscles and/or motor nerves that control movements. Most NMD types result in chronic long term disability posing a significant burden to the patients, their families and public health care. Premature death may result from cardiac and respiratory muscle involvement.

These pathologies are present in all populations, affecting children as well as adults. The overall prevalence of NMDs is difficult to evaluate, but one can estimate that, given the incidence of every different type, around 1 out of 1500 people may have a disabling inherited neuromuscular disease. It is thus crucial to improve molecular diagnosis of these pathologies, as well as to apply the new technologies developed at the whole genome scale to the NMD field.

2.1 NMD-Chip concept and objectives

The NMD-Chip project was built considering the expectations of the EU call HEALTH-2007-1.2-6 concerning "High throughput molecular diagnostics in individual patients for genetic diseases with heterogeneous clinical presentation". Indeed, the aim of NMD-Chip is to design, develop and validate new sensitive high throughput DNA arrays to efficiently diagnose patients affected by NMDs, namely Duchenne/Becker muscular dystrophies (DMD/BMD), limb girdle muscular dystrophies (LGMD), congenital muscular dystrophies (CMD), and hereditary motor-sensory neuropathies or Charcot-Marie-Tooth neuropathies (CMT). The new sensitive and reliable tools (reliability from 95 to >99%) originating from this project is aimed allow assessing all known genes implied in a group of disease at one time (2,100,000 probes) and analyzing efficiently chip data through optimized read-out bioinformatics' tools, within 72hrs to one week. This approach to diagnosis will thus be cheaper than any "gene by gene" approach.

Besides the development of these new high throughput molecular diagnostics tools, NMD-Chip will also foster the knowledge of NMDs by identifying new disease causing mutations using a gene candidate approach.

2.1.1 NMD-Chip concept

The currently available molecular diagnostic methods only permit a gene by gene exploration and are thus cost intensive and time consuming. Given the unexpected molecular genetic heterogeneity of NMDs (more than 100 genes implied), with the recent identification of a large number of novel genes responsible for very rare subtype of NMDs on the one hand, and the rapid development of gene therapies in the NMDs field on the other, there is an urgent need for the development of new high throughput, cost effective molecular diagnostic tools for NMDs. At the moment of project submission, DNA arrays seemed a highly promising technology for the development of such tools.

Indeed, while current diagnostic methods restrict genetic testing of large cohorts to a single exon or a limited group of exons, DNA-Chips would give the unique opportunity to test all pertinent exons of a single patient (and even from several patients) at one shot, thus providing the patient with a faster and more accurate diagnosis.

Some other groups have recently worked on high-throughput technologies for diagnosis purpose. For instance, one group at the Emory University in Atlanta (USA) developed a specific DMD/BMD-chip, enabling detection of all rearrangements as well as point mutations in the dystrophin gene. By testing such combinatorial method, they found 100% concordance within 30 characterized patients, bringing a proof of concept that such a chip-based approach is relevant in diagnosis of NMDs.

Based on these findings and on the recognized consortium expertise in NMDs, (8 partners are members of TREAT-NMD) and quality assessment (one partner is a member of EuroGenTest), NMD-Chip has developed Sequence Capture (SC) and Comparative Genome Hybridization (CGH) arrays for high throughput diagnosis of LGMD, CMD and CMT neuromuscular diseases.

2.1.2 NMD-Chip objectives

This project was designed to:

1. increase detection rate (95 to 99%) of molecular diagnosis, leading to a dramatic improvement in term of quality of life for patients and families,
2. the identification of novel genes involved in NMDs,
3. dramatically decrease NMD diagnosis costs (by a factor of 10);
4. pave the way to new therapies by giving a global comprehension of the groups of NMDs.

In total, the NMD-Chip project, which is based on human genome knowledge and an advanced read-out technology, aims to give patients an easy access to molecular diagnosis and will thus allow them to benefit from personalized cutting edge therapies which are currently developed.

The scientific objectives of the Consortium were:

1. To collect all known gene mutations for CMT, CMD and LGMD and create a database;
2. To identify novel disease causing mutations for CMT, CMD and LGMD;
3. To develop bioinformatics' tools to discriminate between causative mutations and silent polymorphisms;
4. To characterize new NMD patients

In the meantime, some technical objectives were also assessed:

1. To design oligonucleotides probes to spot on the chips to assess known mutations for CMT, CMD, LGMD;
2. To technically validate CMT, CMD and LGMD chips;
3. To assess the quality of DMD, CMT, CMD and LGMD Chips;
4. To design chips to assess candidate genes for CMT, CMD and LGMD;
5. To develop novel bioinformatics' tools to accurately and quickly analyze relevant data obtained from Chips.

2.1.3 NMD-Chip scientific approach

The challenge lies in increasing detection rate as well as abbreviating the time-to-diagnosis (down to 72h to one week) for patients and families via characterization of all mutation types and reducing analysis costs by using platforms with high diagnostic capacities. These two goals will allow us to characterize the genotype also in rare and atypical phenotypes, in genetically ambiguous sporadic cases and in NMDs whose pathophysiology is multi-allelic or multi-genic. DNA-Chips really correspond to that "one-shot" technology. On the other hand, only patients with a well characterized pathology, both on clinical and genetic sides, are eligible for clinical trials or protocols.

To achieve NMD-Chip's aims, it was planned:

1. To design specific Sequence Capture DNA arrays containing all the genes already known to be involved in LGMD, CMD, congenital myopathies, and CMT. We have decided on an original chip design, clinically driven. Two chips have actually been designed, one containing all the sequences for DMD, LGMD and CMD known sequences (a "muscle" capture chip), the other for all the CMT sequences (the "neuro" capture chip).

2. To design a whole gene CGH array containing all the genes already known to be involved in LGMD, CMD, congenital myopathies and CMT. Three subtypes of chips have been designed, one for DMD/LGMD, one for CMD and one for CMT.

3. To develop bioinformatics' tools to accurately and quickly analyze DNA-Chip data.

4. To assess the quality of these chips. Several hybridization tests have been performed to assess a good reproducibility and a strong efficiency.

5. To validate these DNA-arrays on pre-diagnosed patient samples and test their robustness on undiagnosed samples.

6. To design distinct candidate genes SC- and CGH-chips for LGMD, CMD and CMT.

7. To test patients with unidentified gene mutations with candidate gene chips. This last step is to provide us with information on the reliability of the tools developed.

The developed chips consist first of a series of chips dedicated to sequence capturing of all known genes implied in a given group of NMDs. This, coupled with a high throughput sequencing technology (pyrosequencing), will bring a quick molecular diagnosis to patients. Design and validation of one CGH-array chip to scan candidate genes for large rearrangements, deletions or insertions is also intended. Then, if no deleterious mutation is found with the first run on chips for known genes causing NMD, a second series of chips dedicated to candidate genes will be hybridized with patient's DNA.

That means that every gene implied in a given NMD group will be checked at a glance, whereas until now, diagnostics laboratories have to sequence one gene after another until the mutation is found. If deleterious mutations are identified in known genes, the delay to diagnosis will be reduced to less than a week.

Project Results:

2.1 State of the art: current diagnosis of NMD

2.1.1 Neuro-Muscular disorders

NMD are typically progressive, disabling, often life-threatening genetic diseases, which affect adults and children. They have a strong impact on the quality of life of patients due to muscle weakness and subsequent related loss of autonomy, disease-related morbidity, and limited life span (for several forms of NMDs).

NMDs are characterized by their wide genetic and clinical heterogeneity. Indeed, more than 100 monogenic disorders have been identified, a vast majority of them being possible to classify in several nosologic groups according to the muscle or nerve impairment. The most complex of these groups, also being the most frequently involved are those of the muscular dystrophies which include Duchenne/Becker muscular dystrophies (DMD/BMD), the Limb-girdle muscular dystrophies (LGMD) and the congenital muscular dystrophies (CMD); second are the hereditary motor and sensory neuropathies - HMSN, also named Charcot-Marie-Tooth (CMT) diseases. Together, they represent approximately 80% of all patients affected with inherited NMDs.

NMDs are mainly inherited diseases. Strong research efforts have been made to characterize the genetic features of NMDs, and many genes and mutations have been identified in the past 20 years. Nevertheless, new mutations in genes are regularly discovered, and many remain to be identified.

Below follows a brief description of the major NMD groups included in the chips, altogether estimated to represent the diagnosis in 80% of patients with NMD:

DMD/BMD

Duchenne and Becker muscular dystrophies (DMD and BMD) are X-linked recessive neuromuscular disorders affecting approximately 1 in 3,500 and 1 in 30,000 live male births, respectively. DMD and BMD are both characterized by progressive symmetrical muscular weakness often with calf hypertrophy. DMD symptoms typically present before age five with wheelchair dependency reached by age 12. Both DMD and BMD are caused by mutations in the dystrophin gene. The human dystrophin gene is the largest human gene, spanning >2,200 kb on the X chromosome and occupying roughly 0.1% of the genome. It is composed of 79 exons that together account for only 0.6% of its sequence (Abbs et al, 2010; Hegde et al., 2008).

LGMD

The Limb-Girdle Muscular Dystrophies (LGMD) is an important subgroup of MD, responsible for up to a third of the number of DMD cases. The prevalence of LGMD is 1/12,000 (Piluso et al. 2005). LGMD are progressive myopathies grouped together on the basis of common clinical features: they all primarily and predominantly affect proximal muscles around the scapular and the pelvic girdles. However, some LGMD are severe, some are benign and others exhibit a large spectrum of severity and, in addition, the symptoms can appear anytime from childhood to adulthood. LGMDs are divided into autosomal dominant (LGMD1), which represent around 15% of all LGMDs, and autosomal

recessive (LGMD2) forms, with a prevalence of 1/15,000 (Piluso et al. 2005), which imply at least 19 different genes at the beginning of the project (Daniele et al. 2007, Gene Table Neuromuscular Disorders, Volume 18, Issue 1, January 2008, Pages 101-129).

CMD

Congenital muscular dystrophies (CMDs) are a heterogeneous group of inherited myopathies, most of them with autosomal recessive transmission. They are characterized clinically by early onset hypotonia, muscle weakness, and delayed motor development, and, morphologically, by the finding of dystrophic changes on the muscle biopsy. Spinal deformities and joint contractures are frequently observed in the course of the disease. At least 15 genes have been proven to be implied. The diseases can be classified into 3 major groups based on the affected genes and the location of their expressed proteins: abnormalities of extracellular matrix proteins (LAMA2, COL6A1, COL6A2, COL6A3), abnormalities of membrane receptors for the extracellular matrix (fukutin, POMGnT1, POMT1, POMT2, FKRP, LARGE, and ITGA7), and abnormal endoplasmic reticulum protein (SEPN1) (Yanagisawa et al. 2007; Mendell et al. 2006, Gene Table Neuromuscular Disorders, Volume 18, Issue 1, January 2008, Pages 101-129). The prevalence for CMD is 4-5/100,000 (Welbury 2001, Elke 2000).

Congenital myopathies

Congenital myopathies are a group of rare neuromuscular disorders normally present at birth as muscular hypotonia, marked by a characteristic but not pathognomonic structural abnormalities in muscle fibers. Similar to congenital muscular dystrophies, their incidence is approximately 6:100 000 live births. Congenital myopathies present with a wide spectrum of clinical severities. Recent advances in molecular genetics research have allowed accurate genetic diagnosis in many of these disorders.

CMT

Charcot-Marie-Tooth disease (CMT), also named hereditary motor and sensory neuropathies, includes a clinically and genetically heterogeneous group of disorders affecting the peripheral nervous system, with an overall prevalence of about approximately 1/2500 (Skre et al. 1974, Shy et al. 2005). Traditionally, the different classes of CMT have been divided into demyelinating forms (CMT1, CMT3, and CMT4) and axonal forms (CMT2), a clinically very useful distinction. CMT1 can be distinguished from CMT2 by measuring motor nerve-conduction velocities (NCVs): patients affected with CMT1 show reduced NCVs (< 38 m/s), whereas patients affected with CMT2 show NCVs > 38 m/s; the normal value is > 48 m/s. Genetically, CMT disease is characterized by a great heterogeneity. Different modes of transmission have also been evidenced for both demyelinating and axonal types. At the beginning of the project 4 years ago, about 50 different loci were shown to be linked to a form of CMT, among them more than 30 genes were identified (Immerman et al., 2006) (Neuromuscular disease centre, http://www.neuro.wustl.edu/neuromuscular/time/hmsn.html).

These genes encode proteins of various functions: myelin structural proteins; transcription factors involved in myelin gene regulation; proteins involved in: protein sorting synthesis or degradation,

transport processes, cytoskeleton and mitochondrial dynamics (Niemann, Berger et al. 2006). Despite this heterogeneity, about half of all CMT cases and 70-90% of demyelinating CMTs (CMT1) are due to a 1.5 Mb duplication of a chromosomal segment at chromosome 17p11.2 (PMP22 duplication, CMT1A, (Lupski, de Oca-Luna et al. 1991; Valentijn, Bolhuis et al. 1992). Except for this large rearrangement and its reciprocal deletion, causing Hereditary Neuropathy with Liability to pressure palsies (HNPP), large genomic rearrangements are not known as frequent defects in CMT. It is estimated that around 40% of patients with CMTs are not diagnosed at the molecular level, either because the mutations they carry in already known genes escape the diagnostic methods currently used, or, more likely for most patients, because their disease causing mutations lie in not yet identified genes.

2.1.2 Therapies for NMD

There is currently no curing treatment available for these diseases. Promising experimental therapies are being developed mainly consisting of exon skipping, AAV-mediated gene transfer, cell therapies or nonsense read-through. Performing clinical trials of such therapies will require cohorts of patients with accurate genetic diagnosis (Hoffman et al., 2011).

2.1.3 Current diagnosis of NMD

LGMD, CMD, CMT and DMD/BMD show clinical and genetic heterogeneity as their hallmark, even within the same group, making their genetic diagnosis highly challenging. This situation leads to frequent mis-diagnosis or delay and lack of diagnosis in many patients/families.

At the present time, physicians involved in NMDs must face three distinct diagnosis situations:

- The preliminary test results (clinical examination, histology and / or biochemistry) are informative or characteristic for a specific disorder, they suggest immediately the first gene to be sequenced is the good one and the diagnostic process is rather quick, straightforward and fast (with regards of the different steps duration, as described below). In that case, from the consultation to the identification of the mutation, between 2 weeks to 2 months may have elapsed.

- No deleterious variation is identified in the first candidate gene tested; a more complicated phase of the diagnosis begins. It consists in reconsidering the clinical signs, all the complementary data, and to determine which other potential genes could be responsible of the phenotype. This can take months, sometimes years. But the molecular cause of the pathology is found in the end.

- No mutation is identified, although many genes have been checked. Reconsidering clinical diagnosis is an important step to determine which other potential gene(s) could be responsible for the phenotype. Searching for the disease causing gene in such situations is sometimes a matter of months/years, however with no certainty of success. Failure to be successful could arise from an absence of identification of a mutation due to technical reasons, to a mutation's position within an unscreened region of the gene (promoter region, flanking intron), but also from the fact that the disease is associated to a yet unknown or unexplored gene. Indeed, some of the largest known genes involved in LGMD, CMT, and CMD are not explored by any

diagnostics laboratory, or, like Dysferlin, by only one laboratory per country. This situation thus remains the most expensive and time-consuming among diagnosis situations and it is the most frustrating for patients, even if it is less frequent than the other 2 ones.

2.2 Towards a better diagnosis of NMD

Providing new genomic diagnostics tools is a mandatory step towards

- improving genetic counselling
- designing targeted therapeutic strategies to improve disease specific management, and
- identifying patient cohorts amenable to therapeutic trials.

In recent years, new technologies have been developed for the rapid and parallel identification of genomic variations.

2.2.1 Sequence Capture (SC) array coupled with pyrosequencing

While at first, a resequencing array strategy was envisaged for point mutation detection in this project, very fast and important developments of the spotting technologies and hybridization strategies have driven us to consider a more relevant approach. This consists in a two-shot approach, combining the specific and powerful capacity of array-based DNA fragments capture, with the high-throughput pyrosequencing techniques. This approach was driven by the very new technological developments of the Roche-Nimblegen company, which was the only one to propose a very high density and completely flexible spotting strategy. From the spring of 2008, Roche-Nimblegen did commercialize a 2.1 million probe chip. In a first step, DNA obtained from the patient blood sample has to be extracted by classical methods. Then, it has to be fragmented, and a linker must be linked to its both ends. In a third step, DNA is hybridized on the customized sequence capture (SC)-array for 60 hours. Then, the slide is quickly washed, the non-matched DNA is eluted. In the last step, conserved fragments are also eluted, then pyrosequenced using a Solexa or GS-Flx system. This method differs from the "historical" Sanger's sequencing one because free nucleotides are not added altogether to the target, but one after the other. When the correct base is added, a luciferase-coupled polymerase will produce a signal that is measured by a CCD camera. Thus, pyrosequencing is a "real-time" reading of the target sequence, avoiding all usual bias such as CG rich regions linked problems.

The major advantage of this approach is the lack of any PCR based genome amplification before hybridization, thus resolving any bias introduced by enzyme activity.

It has to be noted that pyrosequencing technology remains expensive, the apparatus is still hard to acquire for a small/medium size laboratory. But as this technology quickly spread through the laboratories, prices should fall down in the coming years.

2.2.2 CGH Chips

To detect copy number variations (CNV), as well as deep intronic and non-coding region mutations in the targeted genes, CGH-arrays (comparative genome hybridization arrays) represent an innovative technology able to increase the detection rate of the genetic analysis. Evidences of the urge for quick CNV detection can be easily given. The most common form of CMT, CMT1A (estimated to represent 40-50% of all CMTs, Pareyson et al. 2006), is due to the duplication of the PMP22 containing locus on chromosome 17, but haploidy for that gene is also frequent and causes HNPP, another inherited neuropathy with liability to pressure palsies (Suter et Patel 1994). Large deletions have also been identified in several LGMD genes as Dysferlin in LGMD2B (Bartoli & al., 2011), but since they are considered to be much less frequent than point mutations, they are not systematically and specifically searched and, in most cases, are not detectable by standard approaches. By the use of long distance RT-PCR, exon deletions in COL6A1 has been described in patients with dominant severe or moderate form of Ullrich type CMD (Pepe & al., 2006). The prevalence of deletions in Ullrich type CMD, and more generally in NMDs, may thus be underestimated because this type of mutation is not detectable by standard genomic DNA analyses (as PCR) or may be missed by using short-distance RT-PCR. The resolution of CGH-array increased significantly with the development of in situ-synthesized 60 to 80-mer arrays, which made possible the detection on small or large CNVs throughout the whole genome in a single step (Dhami & al. 2005). With this technique, the genomic DNAs of patient and reference samples are isolated and labeled, without further amplification, with red and green fluorescent dyes, respectively. Each of the labeled DNAs is subjected to competitive hybridization with metaphase chromosomes from normal cells. The ratio of red and green fluorescent signals is measured along the longitudinal axis of each chromosome. Hybridization of repetitive sequences is blocked by adding Cot-1 DNA. The chromosomal regions involved in deletion or amplification in test DNA appear red or green, respectively, but the chromosomal regions equally represented in test and reference DNAs appear yellow. This technology, covering the whole genomic loci of NMD known genes, will directly give the deletion (or insertion) milestones, which advantage is not achieved by "traditional" exon sequencing.

3. Description of the main scientific and technical results

3.1 Collecting of sequences

This has been the first scientific step of the whole project. The NMD gene list constantly evolves, so the actual NMD-chip list is not as complete as the actual Gene Table published. Actually, two separate lists have been defined: one for the "muscular" pathologies, namely LGMD, CMD, DMD, and the other one for more "neurological" pathologies, namely CMT.

These lists have been extracted based on the list published in the Gene Table from Neuromuscular Disorders, Volume 18, Issue 1, January 2008, Pages 101-129, and updated with additional personal information, as well as with new data published in the interval before the chip design, up to May 2010.

A bioinformatics team has developed software tools and methods to help design various versions of CGH chips and capture chips. Various design files for capture chips have been designed. We also initiated the collection of reference sequences to be able to generate so-called LRGs which will become the long term reference system for LSDBs.

3.2 Design of the diagnosis arrays (known NMD genes)

3.2.1 Design of the "muscle" CGH array

We initiated the design of the first NMD chips. RefSeqGene, position, size and exon number were retrieved for the 45 genes involved in progressive NMD, congenital myopathies and congenital muscular dystrophies. A first version of design was reviewed to assess gene coverage and distribution of the probes. Upon approval, 12-plex arrays (135 000 probes/sub-array) was manufactured for the attention of NMD-chip partners.

Given the results of the first hybridisations performed on that muscular CGH arrays, it was decided to design a new set of probes for LGMD / CMD. This 2d generation array has been designed using the listed known genes previously established, introducing slight modifications for the tilling, the coverage of backbone regions and the control regions included. In brief, the design for the array comprises:

- a 12-plex array format (135000 probes/sub-array),
- a homogenous tilling design of the genes, plus a region of 2000 bp around the 5' and 3' terminal exons (1 probe each 50 bp),
- a minimum number of backbone probes representing at least 25% of the total probes on the chip (1 probe each 1000 bp on average),
- three sequences known to be highly polymorphic in the general population (Copy Number Variants (CNV)) to be used as positive controls (two in Chr8 and Chr14, and a new additional one in Chr22).

The final set of probes included in CGH-array design was refined with the help of small software. Proposed probes by Roche-Nimblegen pipeline were alternatively kept, or reversed complemented, in

order to target both DNA strands. Moreover, each probe specificity was re-analysed by the Consortium, and possibly improved or removed.

3.2.2 Design of the CMT CGH array

A list of 43 genes known to be involved in CMT or related neuropathies (dCMT, HSAN, HAN) has been established and exon coordinates have been retrieved from the hg18 build using UCSC genome browser.

Exon coordinates were communicated to the Roche-Nimblegen design team for probe design representing all genes of interest. Upstream and downstream regions +/2- 2 kb were also covered with probes with a coverage described below.

After several rounds of design/verifications, the following final design was approved and 2.1 M 12-plex Nimblegen arrays (135 k each array) were manufactured according the following rules:

1. 43 genes distributed on a total of 18 chromosomes
2. Probes covering gene +/- 2 kb
3. Alternated probes on (+) and (-) strands
4. Tiling:
    - 10bp tiling in exonic regions and intron-exon boundaries (150 bp upstream and downstream of the exon)
    - 30 bp tiling in 3' and 5' UTR
    - one probe for 100 bp in introns.
5. backbone probes each 6 kb
6. Total number of probes: 137 207
    - Gene probes (exon, intron, 5' and 3'-UTR): 69 570
    - Backbone probes: 67 637 probes (~one each 30 kb)

For validation purposes, two positive controls were evaluated during the CGH-array analysis.

- Analysis of CNV sequences. Three regions were analysed; Chr8:39369000-39465000, Chr14:105405000-105489000 and Chr22:22671942-22725328.
- Detection of two mutations on CAPN3 and DYSF genes. For CGH-array validation, two samples with; a characterized CAPN3 partial deletion (c.309+4469_1116-1204 del), and a partial deletion in the DYSF gene (c.89-643_4474-2493del), respectively were used. These samples are referenced in the Reference Material database as Marseille-6318 (for CAPN3) and Marseille-6319 (for DYSF). The DNAs were derived from EBV-transformed cell lines and were extracted using a QIAGEN protocol.

In addition to the two patients, we used DNA from a Human cell line GM12878 as reference sample for CGH. GM12878's DNA was purchased from the Coriell Institute.

Control positive CNVs: In both control patients, all CNVs were successfully identified with the CGH-array. The control CNV region on chromosome 22 was shown to contain only one copy in Marseille-6318 control sample but two copies in the Marseille-6319 sample. In the control CNV region on chromosome 8, both control samples showed a single copy fragment embracing the entire CNV region. Regarding CNV region on chromosome.14, Marseille-6319 was shown to have two copies, but Marseille-6318 control sample displayed large variations across the region, showing alternative gain or loss. Validation of all three CNV by qPCR is currently in progress

Control positive mutations: In both control patients, both mutations were successfully identified with the CGH-array; The DYSF deletion was detected with an fragment average log2(ratio) of -0.8. Breakpoint analysis using CGHweb showed that the DYSF control deletion occurred for 5' breakpoint at position 71561362, and for 3' breakpoint at position 71721298. Accuracy in 5' breakpoint definition was of +484bp, and in 3' breakpoint was of - 503bp. The CAPN3 deletion was detected with an fragment average log2(ratio) of -0.3. Breakpoint analysis showed that the CAPN3 control deletion occurred for 5' breakpoint at position 40444450, and for 3' breakpoint at position 40475048. Accuracy in 5' breakpoint definition was of -22bp, and in 3' breakpoint was of -37bp.

3.2.3 Design of the Sequence Capture chips

This technology allows selecting regions of interest, among the whole genomic DNA, replacing amplification techniques, and then followed by Next Gen Sequencing. Initially, the Consortium planned to use "on-chip" DNA capture for NextGen sequencing. However, during the project, it appears that "in solution" DNA capture was an alternative. We thus extensively tested target sequence enrichment both using array-based and in-solution sequence capture. The data obtained show that this methodology can be used to successfully isolate sequences of interest from a full human genome. All known variants were successfully retrieved as well as some new variants, either overlooked using Sanger sequencing or from the regions not covered by the latter technology. Comparison shows clear advantages for in-solution capture, both regarding ease of use and a significantly higher enrichment of targeted sequences.

SC-chip design has been initiated as soon as CGH-array chips have been validated.

SC-chips for the 50 known LGMD/DMD/CMD genes are designed as follows. The selected regions are based on the corresponding CGH chips sequences. All the exons have been selected, with the 5' UTR, but not the 3'UTR, with 200bp flanking intronic sequences. This represents 1227 different genomic regions in total, covering 940 kb of genomic sequences, which is suitable for a 385K format SC array.

SC results for LGMD/DMD/CMD sequence selection have been compared to in solution capture technique SureSelect from Agilent.

In parallel, SC-chips for the 45 known CMT genes as been ordered, with some slight modifications:

1. 385 K Roche-Nimblegen array
2. 45 genes distributed on a total of 18 chromosomes

3. 594 regions spanning 627kb
4. Exonic regions and intron-exon boundaries (200 bp upstream and downstream of the exon)
5. 1 kb upstream and downstream of each gene (5' and 3' UTR)
6. Probes specificity re-analyzed as described upper
7. Alternated probes on (+) and (-) strands

According to the total length, it was decided to stay on 385K array format (not use the HD2 format). We sent the coordinates of all regions of interest to the design team of the manufacturer and worked together to finalize the list of probes to be set on the 385K arrays. This step has been done by our bioinformatics team, in charge of the probes review (discard remaining non-specific probes, reversing probes to capture both strands).

3.3 Design of research arrays (candidate gene chips)

3.3.1 Selection of CMT candidate genes

The selection of candidate genes for CMT was based on data from the literature, as well as on data extracted from databases (AmiGo, Aceview, MGI). The selected genes include paralogues, genes in pathways where already known CMT genes have been involved (myelin structure, polyphosphoinositide signalling for example), and interacting partners of proteins mutated in CMT. Mouse models presenting relevant phenotypes were also taken into consideration, when selecting candidate genes.

Several software were used for prediction: Genomatix BiblioSphere PathwayEdition, MatInspector; TargetScan, Miranda. Prediction was verified by literature data mining.

A final list of 301 candidate genes was build up. Two additional genes (AARS and FAM134B) were included in the CGH list since these genes have been recently implicated in newly described CMT after the building-up of the WP2 chips. Candidate genes are listed in the deliverable D3.1 contractual document.

3.3.2 Selection of CMD candidate genes

Candidate genes list for CMD was build up, mainly selected by using the web based tool, Endeavour (http://homes.esat.kuleuven.be/~bioiuser/endeavour). ENDEAVOUR is a software application for the computational prioritization of candidates genes, based on a set of training genes (trimset). It is made up of three stages: training, scoring and fusion. In the first stage, information about the training genes (genes already known to play a role in the process under study) are retrieved from numerous data sources in order to build models. As output, it gives a global prioritization score and specific prioritizations scores for mined database. The following genes were used as trimset: DAG1, FKRP, FKTN, POMT1, POMT2, POMgnt1, LARGE, LAMA2, ITGA7, PLEC1, SEPN1, COL6A1-3, SYNE1, TTN, BAG3, FHL1, DPM3 and CNTN1.

An additional set of genes was included on the basis of evidences from animal models or functional studies that these genes could be involved in the pathogenesis of CMD.

A final list of 345 gene candidates was selected, who has been reported in the deliverable D3.2_update contractual document.

3.3.3 Selection of LGMD candidate genes

The selection of LGMD candidate genes have been performed according to several strategies.

Search and prioritization for LGMD candidates was performed mainly at Genethon. The selection of LGMD candidate genes have been performed by according to several strategies The underlying assumption for the choice of candidate genes is that genes involved in the same disease share a link with each other. This could mean that their protein sequences present similarities that their proteins are physically interacting within cells, that they share functional regulation motifs, or participate in similar pathway for instance. The starting point for identification of related LGMD candidates was the list of the known LGMD proteins, namely for LGMD1: Myotilin, Lamin A/C and Caveolin 3 and for LGMD2: Calpain 3, Dysferlin, Gamma, Alpha, Beta and Delta Sarcoglycans, Telethonin, TRIM32, FKRP, Titin, POMT1, Anoctamin5, Fukutin, POMT2 and POMGNT1.

A final list of 476 first priority gene candidates was established. Using these abovementioned strategies, we have been able to prioritize LGMD candidate genes in four levels of prioritization from A1 to A4; 245, A1, 76 A2, 57 A3, 72 A4. In addition we have included all 26 MIRgenes. Candidate genes are listed in the deliverable D3.6 contractual document.

3.3.4 CGH Chip design

Following experience from the known gene chips, and given the number of sequences to target, a definitive 3-plex 2.1M array format (2.1 million probes, 720000 probes/sub-array) chip was chosen instead of the 12-plex.

From the candidate gene list defined above, the targeted sequences were retrieved from the USCS Genome Browser (http://genome.uscs.edu) using the HUGO GeneSymbol Identifier. All gene isoforms were considered and overlapped sequences were merged in a unique sequence prior to probe design using a Pearl-script.

We established a basic tiling density to favour homogenous coverage of all exonic and intronic regions, with a homogenous tilling density of 1/50bp for CMD and LGMD chips, and 1/60 for CMT chip.

Furthermore, to avoid artefactual results observed at the gene extremities, we decided to use a probe density of 1/50 in a extent of about 2000bp around the 5' and 3' terminal exons for each gene.

The total number and the average density of backbone probes were reduced compared to known gene arrays, defining a minimum total number of 25% of the total probes on chip. The estimated average density tilling for backbone coverage is between 1/6000 and 1/12000bp. Gene desert regions have not been tiled for backbone probes.

As a technical control to validate these CGH chip, we included sequences with known polymorphic Copy Number Variants (CNV). Two regions were initially considered, because of their inclusion in the Known Gene chips: Chr8:39369000-39465000 and Chr14:105405000-105489000. From previous experience, Chr14:105405000-105489000 showed abnormal results imputed to the proximity of the telomere of Chr14. We proposed another non telomeric CNV region; chr22:22,671,942-22,725,328 (including GSTT genes) to complete technical controls. For CNV regions, we used a homogenous tilling of 1/100 probes covering the region. In addition, given the importance of technical controls, we included two known disease causing genes; CAPN3 (genomic sequences from exon 3 and 5) and SGCG (complete genomic sequence) for CMD and LGMD chips, and PMP22 (complete genomic sequence) for CMT chip. These sequences will be used to perform a validation with a LGMD2A patient carrying the characterized CAPN3 mutation IVS4+404Δ5328, as well LGMD2C patients, carrying small/large deletions in the SGCG gene and PMP22 deletion carriers. A summary of the initial tilling density is indicated below.

Type of sequence	Density of tilling (probes by bp)
Exons+ Intron	1/50 for LGMD/CMD and 1/60 for CMT
5' and 3' gene borders (2000 bp)	1/50 for LGMD/CMD and 1/60 for CMT
CNV and control gene regions	1/100
Backbone sequences	>1/6000

From the list of candidate genes and the proposed design, Roche-NimbleGen has generated a draft probe set. Their specificity criteria is that the selected probe must contain at least 20 successive nucleotides that match only once on the genome. Uncovered regions will be redesigned with lower stringency criteria (default criteria, 38 successive nucleotides with unique match on the genome). In order to obtain a better accuracy for the baseline signal, the remaining probes will then be designed alternatively, from both sense and antisense complementary DNA strand and filtered after BLAST each probe to Human genome (removing all probes with >5 hits with 60% of sequence identity). This defines the final set of probes, and this second selection process has been done by NMD-chip partner PhenoSystems.

3.3.5 SC-chip design

This task has suffered from an important delay in its execution dates. During one of the Consortium meeting, a major change was decided in the DNA capture method to be used. In solution capture method was preferred against retention on SC-chips for the DNA enrichment in target sequences. Indeed, in solution sequence capture methods developed for next-generation-sequencing (NGS) has revealed itself as more powerful, robust and cheaper than capture based-array methods.

After discussion between involved partners, it was pointed out the relevance of performing a single analysis for LGMD and CMD pathologies. These pathologies have an overlap of clinical and pathological characteristics resulting in a high similarity between respective candidate genes list.

Two different approaches were adopted for candidate gene analysis of both LGMD-CMD and CMT group of pathologies. Custom designs for in solution capture for the LGMD-CMD design regroup both candidate and known genes (820 genes) and CMT analysis in a whole-exome approach customized specifically for coverage of all candidate and known CMT genes (351 genes).

The capture of genomic regions focused on CDS and UTRs for selected genes. Genomic coordinates from selected genes were generated using in house bioinformatics tool developed by PCHIP. This tool based on UCSC genome browser, uses the UCSC ID number that is deduced from multiple databases (RefSeq, Uniprot, Genebank, CCDS and comparative genomics) allowing the identification of all exonic sequences thus generating the genomic coordinates for targeted regions. All splice variants are been considered. An update of genomic coordinates from hg18 to hg19 genome version was performed to accommodate with the actual pipeline for manufacture production based on hg19.

Regarding the methodology used, three different custom designs for in solution capture have been designed for candidate genes analysis using either Agilent or NimbleGen technologies, the most advanced and widely used platforms.

For analysis of LGMD-CMD pathologies, two candidate-exon-focused custom designs have been developed from Agilent and Nimblegen; using the Agilent Sure Select and NimbleGen SeqCap EZ Choice in solution capture designs:

- Agilent Sure Select in solution capture design. The design for in solution capture library probe, Agilent Sure Select, was performed with the online design tool eArray from Agilent. The eArray interface generates the probe library needed to capture the regions of interest. The optimization of the design was performed directly by Agilent's core lab that has expertise of library design. The optimized design proposed a library probe design that covered more than 98% of our target sequences and use two times more probes than standard design (115,000 probes) to improve the capture efficiency.

- NimbleGen SeqCap EZ Choice in solution capture design. The design for in solution capture library probe, NimbleGen SeqCap EZ Choice, was performed by NimbleGen's core lab from the coordinates of interest. The total coverage of designed regions was of 96.5%, but rose up to 98.7%, when considering the total length that is also indirectly captured by probes (100bp). NimbleGen solution capture library probe proposed a higher number of probes (385,000) compared to Agilent standard design.

For analysis of CMT, one whole exome customized design has been developed from Nimblegen; using SeqCap EZ XL Choice in solution capture designs. Full exome capture proposed by NimbleGen corresponds to a probe library targeting the 50Mb of exonic sequences in the whole genome. This design has been customized to include both the known and candidate CMT genes using the NimbleGen SeqCap EZ Choice. The customized exome design was performed by NimbleGen's core lab from the coordinates of interest.

3.4. Hybridisation workflow and analysis procedures

3.4.1. Hybridisation

The standard Roche-NimbleGen procedure was used as described in the user guide version 7.0 available on the NimbleGen website (http://www.nimblegen.com/).

Quantification and purity of DNA were assessed using Nanodrop ND-1000 spectrophometer. Absorbance ratios at 260/280 and 260/230, revealed a high purity (respectively >1.8 and > 2). As a quality control before hybridization, DNA integrity was checked in an agarose gel. The quality of hybridization was assessed using two parameters produced by NimbleScan software: Signal-to-Noise ratio (SNR) and Median Absolute Deviation of log2Ratio (Mad.1 dr). Mad.1dr provides a measure of difference between consecutive probes and therefore a surrogate measure of experimental noise in the experiment and was used preferentially as quality standard for hybridization success comparison. SNR and Mad.1 dr values for Marseille 6318 and Cochin 19 samples are 0.86 and 0.103 for Mad.1 dr, and 6.1 and 6.2 for SNR respectively. Data with Mad.1dr <0.15 were considered as good quality. The Reference DNA (GM12878 DNA) was labelled with Cy5 and all test samples with Cy3 dye using the NimbleGen dual colour labelling kit. No dye swap experiment was performed. A total amount of 31 µg was hybridized. The labelled DNA were hybridized on slide for 72 hours and then washed and scanned at 2µm resolution using a NimbleGen MS200 scanner. Following array scanning, the arrays were gridded and sample tracking controls checked using NimbleScan v2.6 software.

3.4.2. Analysis method

A number of algorithms are available for an automatic interpretation of CGH data, thus providing a fast and accurate interpretation of the results. These algorithms can be called through the web interface CGHweb (http://compbio.med.harvard.edu/CGHweb/), a tool that allows the application of a number of algorithms to a single array profile (Lai, Choudhary et al. 2008). CGHweb generates a heat map panel for each method as well as a consensus profile, and produces a tabulated result output. In this analysis, the fused lasso method (Tibshirani and Wang 2008) was the algorithm of choice since it appears to lead to a reduced background noise compared to the other ones. The others algorithms were used to ascertain or compare features of interest a posteriori.

3.5. Design of a high throughput tool to analyse the chip data

3.5.1. Pipeline developed by LUMC

LUMC built the analysis pipeline to score and evaluate variants detected form the high-throughput sequencing data. This pipeline should greatly reduce time and cost involved, especially towards deciding whether these variants might impact the health of the patient.

The first step includes scoring the variants using as main thresholds (1) a minimal coverage, (2) a minimal allele frequency of the 2 variant and (3) presence in reads from both directions. The thresholds to be used here depend on the sequencing technology used, the experiment performed and the location of the variant (e.g. on the X-chromosome in males). To analyze the variants called they are mapped to the current genome build and reported using HGVS nomenclature when they are in a transcribed region (gene). Performance of a module that analysis the 'normalized' data to reveal (large) CNVs (deletions / duplications) based on overall coverage is currently tested.

To evaluate the performance of the software used to map the sequence reads against the reference sequence as well as its performance to call the different types of variants (single nucleotide substitution, small deletion/insertion, etc.) a data simulation tool was made. Based on a sequence of interest, this tool can be used to generate data of the desired sequence read length, paired-end or not and containing model variants to be tested. Using these data, the pipelines constructed can be tested to reveal any hidden weaknesses or errors. The tool seems also very helpful for data analysis courses,

where participants need to learn the strength and weaknesses of the different analysis packages. The model data generated are also used for quality control of the analysis pipeline. Model data are added to the sequence data before sequence analysis starts. At the end, observed/expected outcome of the model data are used to monitor performance of the complete pipeline.

Step two is filtering the variants for those that are not of high interest (no known functional consequences) including:

- those known in the other databases / data sets, including dbSNP database, but not in the HGMD database, 1000 GENOMES, etc.
- those observed before in other experiments in the facility (based on sequence technology, application and sample analyzed) this steps identifies rare local/national variants (not in the 1000 GENOMES) as well as removes recurrent errors (incl. sequencing or mapping) intrinsic for the entire process.

Step three is selecting the variants with possible functional consequences (pathogenic):

- those present in the HGMD database
- those present in existing gene variant databases (LSDBs, collaboration with EU FP7 project Gen2Phen). Note that as part of this step, to further future analysis, the variants called in the sample analyzed are submitted to the respective variant database(s)

Step four is evaluating the remaining variants for those that might have possible functional consequences (pathogenic). Criteria used here include:

- position of the variant with the categories of decreasing interest being exonic protein coding, splice site, exonic 5'/3' UTR, intronic, gene flanking
- predicted consequence of the variant with the categories of decreasing interest being truncating, splicing, missense, silent
- computational predictions using existing tools like PhyloP, UMD Predictor, SIFT, PolyPHen, Panther, AlignGVGD, etc.

3.5.2. Prediction tool developed by INSERM Montpellier

In order to allow a rapid prediction of the pathogenicity of any SNP localized within an exon, we decided to pre compute all possible human mutations corresponding to a single nucleotide substitution. To do so, we had the choice to start from the Human Genome Assembly HG18, which is used by most technologies (microarrays, NGS) or the newly released HG19, which is still frequently updated. We chose to develop the UMD-HTS system from the HG18 genome assembly because it is now a fixed reference sequence that is used by most partners and companies. We designed the system as an easily upgraded system that can integrate data from other HG assemblies in the future.

(Number of SNPs per single chromosome-specific database. Only SNPs resulting in a missense or a synonymous mutation were created while SNPs leading to nonsense mutations were filtered out: See attached).

Data were collected from the Ensembl database hosted at the European Bioinformatics Institute (EBI). They contain data from 32,734 genes, 62,031 transcripts, 524,515 exons and 462,484 introns.

Coding sequences were automatically reconstructed from the crude dataset in order to generate all possible nucleotide substitutions from exonic sequences. Substitutions resulting in stop codons were excluded, as their pathogenicity is usually obvious leading to a total of 179,873,844 substitutions.

In order to perform predictions using the UMD-Predictor algorithm, data were collected at various levels including conservation, physicochemical properties and impact on splicing signals. This led to a total of 3,957,224,568 annotations.

3.5.3. NGS Viewer and analysis framework

The main goal of this part of the project was to create a framework that allows a non-technical person to perform all the steps needed for the analysis of DNA sequences generated by Illumina or Roche NGS systems on a group of genes involved in NMD, from the raw data coming out of the sequencer to the final report that can be saved and exported to databases.

Those steps include filtering the raw data, aligning it, scanning for variants, performing further analysis on the data, visualizing it and creating a report. The user interface should be comprehensible to a beginner, but also allow an advanced user the flexibility he needs.

For the visualization tool, following discussions with the end-users, the following important features were identified during the design process and were implemented in the released deliverable:

- Creation of a Variant report
- Ability to submit a new variant to LOVD databases.
- Better performance for all features of currently supported file formats (BAM, varscan, wiggle, bed)
- Caching of external resources to work offline
- Speed improvements for slow internet connections
- Paired-end read support
- Various usability improvements
- Bug fixes and stabilization
- Ability to add data (all identified variants) to a local database and increase knowledge with every run.

In order to adequately design the framework, we modelled the diagnostics activity in which process the framework software will be used in the following activity diagram (see attached).

3.6. Validation and results from CGH arrays (both known and candidate genes)

Experiments with the 2nd generation of these chips have showed some improvements in terms of data quality over the first chip generation. However, two major issues are still of concern for use of the chips. First, the choice of the right hybridization conditions is crucial for successful data acquisition. It emerged from the user reports that every lab uses different conditions with respect to initial sample preparation (i.e. sonication) and hybridization times.

The second issue is the software used to analyse the sample data collected in the experiments. Results are highly dependent on the specific program and also on the defined thresholds for calling a variant. The used software can be roughly divided into two types: the software provided by the vendor of the used equipment, and independent software that can be used for data interpretation. An example for the latter is CGHweb, a web-based software that can be used to analyse array data with different algorithms. During the Steering Committee meeting it became clear that most partners use an individual software combination, making a comparison of results very difficult.

It was therefore decided on the Steering Committee meeting M24 that data from different labs should be sent to C. BEROUD who will do an analysis with different software tools. This approach will show how sensible the analysis is towards different lab protocols for hybridization and data acquisition. It will also reveal if there are difficulties with data compatibility between different platforms. Another major issue regarding software is the definition of thresholds for calling variations. Reports revealed that this is a crucial point in data analysis as samples can easily be assigned false positive or false negative if the thresholds are not appropriate.

Here are the conclusions of the validation procedure for candidate gene for LGMD, CMD and CMT CGH-arrays, regarding four points; baseline signal, detection of control positive mutations, detection of control CNV regions, and analysis of reference DNA.

- Baseline: candidate genes CGH-arrays were designed with a homogeneous tilling coverage of all genomic regions plus 2kb each extremity in order to obtain a stable baseline along the targeted genes. Stability and robustness of the baseline was indicated by a visual analysis of the baseline using the SignalMap software and from the self-hybridization tests and duplicate analysis performed with the CMD CG-CGH-array. Furthermore, all analyses indicate the absence of bias of the baseline in the extremities of targeted regions, in contrast to what was observed in previous CGH designs (for earlier known genes CGH), confirming the suitability of the tilling design.

- Control Mutations: the experiments performed with the CMD and LGMD CGH-arrays with the same controls identified both mutations in CAPN3 and SGCG. Definition of the CAPN3 5' breakpoint was similar in both duplicates for the CMD-CGH (+288 and +388bp) but was less precise from that observed with the LGMD-CGH-array (+85 bp). For SGCG analysis, breakpoints are not known for this mutation. However, comparison of breakpoints from CMD and LGMD CGH-array indicates robustness in breakpoint definition. The differences between the two replicates of the CMD CGH-array were 106 and 526 bp for the 5' and 3' breakpoints, respectively. For the LGMD CGH-array, the positions of the breakpoints were defined at about 5 kb for 5'SGCG compared to the CMD CGH-array and a few hundred bp for the 3'SGCG breakpoint. The CMT CGH-array identified the mutations in PMP22 but the sensitivity of the experiment has to be improved. Two major problems were detected: 1) problem for the definition of the

breakpoints with line levels wrongly assigned by Signal Map; 2) log2 ratios are too low in the detection of duplication events (often below 0.3) and might lead to the missing of the event, when using an "automated" analysis, like CGHWeb analysis, with a fixed threshold above 0.3. 0.3 is the minimal value that can be used, otherwise, in candidate gene arrays containing several hundreds of genes, we would have to deal with hundreds of events, most of them being artefacts.

- CNV controls: all three CNV regions analyzed were detected with WP3 CGH-arrays. Analysis of CNV region on chromosome 8, showed that both control mutant samples used in CMD and LGMD-CGH arrays experiment carry a single copy in Chr8, demonstrating the sensitivity of the CG-CGH-array to ascertain with reproducibly one copy versus two copies. This CNV is long and outstretched the entire targeted region, preventing the identification of breakpoint. The other two CNV regions, in chromosome 14 and 22, showed several inconsistencies when comparing the same samples with CMD and LGMD analyses. On chromosome 22, a CNV including the GSTT gene was detected in both arrays with the sample Marseille-6318, but was observed only with sample Cochin-19 in CMD CGH-array. On chromosome 14, no variant was identified in CMD CGH-array analysis, but several regions of loss and gains across the targeted region were found in LGMD CGH-array analysis. Even if these results should be confirmed with a second technique, these CNV regions seem not the most suitable and robust control for array performance.

- Reference DNA: The reference DNA was described to have a heterozygous deletion in ten genes included in the three WP3-arrays design. None of the heterozygous deletions described in the reference DNA has been detected by any of the CG-CGH-array analyses. Even if other variants (gains or losses) were observed in these genes (see CMD analyses), the positions of the identified variants were not consistent with reported genomic coordinates (available from public databases, http://www.genome.gov/ENCODE/) for these heterozygous deletions. These negative results could arise from the quality of the annotation of these variants in the reference DNA, probably being worst for the small variants than for larger fragments. It should be noted that all the ten variants included in the CG-CGH-array design were shorter than 5kb. For the next generation CG-CGH-arrays, the use of other type of internal controls is envisaged to control array performance. One option to ensure a better quality control for both labelling and hybridization steps is the inclusion in the new design of the Labelling and Hybridization controls (LHC) from Roche-NimbleGene.

As an example, highlight is given on NIEH partner's work. 87 DNA samples of NMD patients without molecular diagnosis were examined by LGMD chip and 17 samples by CMT chip. Mutations were found so far in the LARGE, DMD, Col6A3, CAPN3 and SGCD genes. A heterozygous exonic deletion in the LARGE gene was found in a 4 year old child diagnosed with congenital myopathy and has been confirmed by MLPA and qPCR, while the second mutation currently remains undetected by DNA sequencing. One of the parents carries the deletion of the LARGE gene. In addition, a large exonic deletion in the DMD gene has been detected in a patient formerly diagnosed with EDMD and this was later proved by MLPA analysis. Some interesting deletions not described yet were found in several LGMD patients in the SGCD gene which have been confirmed by MLPA. A clinically DMD patient without any dystrophin mutation had a small Col6A3 deletion but no phenotypic relevance could be found. A large heterozygous exonic deletion of the CAPN3 gene was detected in an LGMD

patient who also did not have molecular diagnosis before. Further investigations are needed for other smaller and intronic deletions/duplications; most of them are known non-pathogenic CNVs.

3.7. Preliminary data with Sequence Capture and High throughput sequencing

Sequence capture is used for the enrichment of specific DNA fragments prior to Next Generation Sequencing. One aim of the project was to develop a sequence capture chip specific for genes that are related to neuromuscular diseases. A proof of concept for this chip was presented at the Steering Committee meeting M24 in Wuerzburg, Germany. It was shown that the chip works in principle, but there are still some problems with the capture efficiency of some regions. It was also pointed out that this problem could be avoided with a solution-based capture strategy because of easier handling and better results. Continued testing of the sequence capture chips in parallel with different solution-based approaches was therefore recommended and are still ongoing at the end of the NMD-CHIP project.

3.7.1. Capture experiments from LUMC

1 - Solid phase hybridization using CGH-design 1 and sequence capture

(SC) design 1 arrays (NMD-chip LGMD genes). CGH-design 2 arrays were also used but gave no successful capture results -genomic DNA of 12 patients were pooled (6 samples with known pathogenic variants and 6 with unknown variants), sample prepared for paired end sequencing (Illumina) and hybridized to the different array designs (LGMD genes) -quality control (electrophoresis) showed the presence of genomic DNA after hybridization, washing and elution -for the CGH-array initial sequencing depth (Illumina GA II) was not sufficient to draw definite conclusions, still a range of variants could be detected and confirmed (see previous report). A new sequencing run was just completed on the HiSEQ2000 (Illumina), generating ~90 gb of sequence. Data analysis, (mapping, variant calling) is ongoing. Focus will be on; 1) confirming known variants, 2) detection of unknown variants, 3) determining false positive/false negative levels based on the thresholds set during data analysis The 6 LGMD patient samples with unknown variants have also analyzed using whole-genome in solution capture. Sequence analysis is currently ongoing.

During data analysis we aim to compare the 2 methods applied. For the whole-exome data analysis we will first focus on the genes on the NMD-chip LGMD array. When no conclusive pathogenic variants are found, data analysis will be extended to the whole genome.

NMD-chip partners have visited Leiden (December 2010) and received hands-on experience with array-based capture. Although the training was successful, overall end-results were not great; most samples did not make it to the end. However, this was expected and shows the technical difficulty of the array-based sequence capture requiring a trained, skilled and experienced researcher to be successful

2 - In solution capture

For whole exome in solution capture we have used kits from both Agilent and Nimblegen. The samples analyzed include :

- 10 samples from 5 different families with FSHD2 were captured using Agilent. Sequence data has been obtained for 4 patients, sequencing for the other 6 are currently running. Initial data analysis shows no obvious pathogenic changes although variants in some candidate genes remain. Comparison with the other 6 cases should confirm or exclude their candidate gene status.

- 4 samples from different cases in a CMT family, including a hydrocephaly phenotype, were captured using Agilent. The data show a promising variant in a gene which was a likely candidate based on other CMT-disease genes. We are currently confirming our findings using samples from patient other CMT/hydrocephaly cases. Primer sets were already developed to quickly analyze the gene using standard High-Resolution Melt-curve Analysis (HRMA) and Sanger sequencing

- 30 patient samples from 6 different NMD-chip partners were sample prepped. 8 samples failed (to be repeated). Whole exome capture using Nimblegen was performed for 22 samples. Quality control for 10 samples demanded a re-evaluation, 10 samples passed QC. Sequencing for 8 samples is ongoing, initial data analysis should be completed in ~2-3 weeks.

3- General:

Although experiments were successful in the end, with most data coming in at the moment of writing, we are not very happy with the outcome. We believe however this nicely shows the problems arising with applying new technologies and it shows what needs to be done to apply it successfully in a diagnostic setting. The sequencing equipment, the HiSEQ2000, turned out not to be reliable yet. Some samples were run and failed 5 times during sequencing. For this 3 different machines were used in three different labs across the Netherlands. Reasons for failure were mostly different. The system in Leiden had problems for over 9 months and repairs were not successful, generating continuously new problems. The end result was a serious delay for the project. These experiences indicate that to apply this technology in a diagnostic setting, where time-to-result is often critical, cannot be done without having the direct availability of a second system (to reduce the consequences of system failure).

3.7.2. Capture experiments with candidate genes

DNA from molecularly characterized control patients were selected to test the sensitivity of capture and sequencing methods.

Genethon and Cochin

For LGMD-CMD validation, two patients from the NMD Reference Material Database were selected; for comparison of Agilent Sure Select and NimbleGen SeqCap EZ methods (and to compare the techniques between different partners). In addition to these control DNAs, other uncharacterized patients (16 for Genethon and 8 for Cochin) carrying one known mutation for recessive form of the disease have been sequenced in LGMD-CMD analysis. 2 pools of 10 indexed captured samples from Genethon were sent to IntegraGen for HiSeq 2000 Sequencing and the third pool from Cochin was sent the Pitié Salpêtrière platform. A first result is expected in few days for the first Genethon pool and the others will come in few weeks. For Genethon, 5 additional samples have been sent to LUMC (Leiden) for whole exome sequencing.

InstMyol

Twenty DNA samples were selected for sequence capture array using the in solution agilent technology. Out of these 20, 6 have a known mutation that will be used to validate the sequencing data (2 in COL6A2, 1 in FKRP, 1 in DOK7, 1in DES and 1in TNXB). The Next Generation Sequencing (Illumuna HiSeq) is ongoing on the Pitié Salpêtrière platform. In addition, 5 DNA samples have been selected to be sequenced after whole exome capture (SureSelect, Leiden).

INSERM/Marseille

6 DNA from CMT patients were used for Nimblegen SeqCapEZ in solution capture. 2 pools of 3 indexed captured samples were successfully generated and sent to Integragen for HiSeq 2000 sequencing. A DNA from a CMT patient with one uncharacterized mutation, but multiple characterized SNPs was included in the validation process.

UCL

10 DNA from CMD patients were used for Nimblegen capture and 10 others for Agilent capture. 2 pools of 10 indexed captured samples were successfully generated and sent back to UCL for HiSeq 2000 sequencing. Two DNA were selected with mutations in the nebulin and lamA2 genes.

The technical validation included LGMD, CMD and CMT designs using three different sequence-capture enrichment methods with in particular a comparison of the Agilent and Nimblegen alternative designs. Due to an important delay in sequencing execution for the candidate validation experiments, the validation was carried out with the results obtained for the LGMD-CMD design in ten patients. The mutations present in the control patients were detected unambiguously in all cases thanks to the quality and sequencing coverage of produced libraries. Therefore, these data permit the technical validation of the sequence capture as it was performed.

The overlap of LGMD and CMD pathologies, and the technical possibilities offered, by the library design and the sequencing array, in terms of total length of captured sequence and sequencing array capacity pointed to a single analysis design for both pathologies, for all known and candidate genes. However, the low overlap among LGMD-CMD and CMT pathologies, preclude a single joint design.

Three custom design libraries for in solution capture were designed for WP3 NGS. LGMD-CMD library design regroup both candidate and known genes (820 genes), and CMT library design corresponds to a whole-exome approach customized specifically for coverage of all candidate and known CMT genes (351 genes). From the initial list of genes, genomic coordinates were generated using in house methods. For the 820 LGMD-CMD genes, a list of 13961 fragments was identified representing a total length of 3.9 Mb of sequence. For CMT, a list of 6541 fragments and a total length of 1.75 Mb were identified. The hg19 genomic coordinates was used for library probe manufacture by Agilent and Nimblegen.

The custom design libraries for LGMD-CMD genes were designed using the Agilent Sure Select in solution capture designs and the NimbleGen SeqCap EZ Choice in solution capture designs. Customized exome for CMT analysis was designed using the NimbleGen Customized Exome (EZ XL Choice) in solution capture design.

Potential Impact:

4. DISSEMINATION ACTIVITIES AND EXPLOITATION PLAN

4.1. Dissemination to the scientific community

The NMD-Chip project was officially launched (Kick-off Meeting) at the end of November 2008 and plans for the project website were immediately launched. The domain names http://www.nmd-chip.eu and http://www.nmd-chip.com were purchased, with the .eu domain being used as the primary project website.

An agreement was reached to host the http://www.nmd-chip.eu site on the existing EU FP6 TREAT-NMD Project server at Newcastle University (acting as Project Coordinator of the TREAT-NMD NoE). The shared hosting of the website is the optimal solution, both cost-effective and administratively efficient, due to the similarity between the two projects and to the fact that UNEW is the partner responsible for both websites in both projects.

The website is run using a content management system. The major advantage of such a system is that it can be maintained and edited by non-specialists.

The design of the NMD-Chip website was defined and the site came online in February 2009 (M3). A web committee was established on the occasion of the M6 Steering Committee Meeting to ensure the site is updated with appropriate information. Regular updates are performed, with the team of the Project Coordinator responsible for checking all information to ensure it does not bring up any IP issues.

A special feature is its partner-focused pages, which enable each researcher working on the project to have a short profile with photograph, contact details and summary - this helps improve the profile of individual researchers, which is particularly valuable for young researchers, and forms the basis of the "focus on women" deliverable, showing the strong representation of women within the project.

An Intranet site was also developed (M4), which is restricted to consortium members enabling open discussion of potentially IP-restricted issues. Thanks to this policy of restricting access, we have been able to open a specific and personal access to the "Deliverable" section of the site for the Project Scientific Officer (SO). This was set up before the creation of the SESAM platform, and allowed to avoid many emails with large attached files to circulate. The SO can download the deliverables directly from the site. He is informed by email at each new deliverable uploading. The other parts of the intranet site are closed to the SO.

All the internal documents, such as Steering Committee minutes, meetings minutes, gene lists, internal reporting and WP Leader presentations are shared via this intranet.

To ensure easy email access to all project participants, two email mailing lists have been created: "nmd-chip_all" and "nmd-chip_partners". The first of these enables every person involved in the project, including PEC members and advisory board members, to be contacted by writing to a single email address. The second email list is restricted to members of partner institutions and is thus more appropriate for correspondence where any sensitive IP issues might be discussed. Given the relatively large number of people connected with the project these mailing lists form a practical solution to avoid having to cc a large list and are thus widely used for management issues and consortium-wide communication.

For the long-term success of the project, interaction with leading individuals and groups outside the immediate project consortium is essential. It was therefore decided to host a scientific workshop with external invited speakers. To optimise costs, it was decided to host this as an additional day at the end of the M18 Steering Committee meeting in Ferrara - this enabled all Steering Committee members to be present with minimal additional costs. The input from the external experts was felt to provide consortium members with valuable additional insight that will assist them in future developments. It has also resulted in an opportunity for transatlantic cooperation with US groups that is currently under evaluation and may be valuable for the project provided there is no risk to project IP.

On behalf of NIEH, Dr Veronika Karcagi made an oral presentation about the project at the post gradual course of neurologists. The title of presentation was: "New diagnostic approach by utilisation of DNA-chip in CMT diagnostics".

Other lectures at national conferences, which also focused on the project were: "NMD-Chip Consortium - Involvement of high-throughput techniques in the diagnosis of neuromuscular diseases", "Charcot-Marie-Tooth neuropathies - Clinical and genetic classification and diagnostic possibilities, including microarray technologies", "Genetic background, new diagnostic possibilities and therapeutic measurements in DMD/BMD", "Collaboration between diagnostic laboratory, clinical departments and patient organizations". The laboratory annually participates at the quality assessment schemes in DMD organized by EMQN.

A brochure published by the TREAT-NMD project also includes a profile of the NMD-chip project, thus helping to raise its profile amongst industry and other interested parties contacting TREAT-NMD. Advantage has been taken of the launch of the new TREAT-NMD website to increase the profile of the NMD-chip project by giving it its own profile page on the TREAT-NMD site at

http://www.treat-nmd.eu/research/related-projects/nmd-chip/. The addition of individual NMD-chip researcher profiles to the TREAT-NMD site to complement the information already available on the chip site is on-going. An example of this is the page for the project coordinator Nicolas Lévy, http://www.treat-nmd.eu/contacts/nicolas.levy.

Finally, the full 18 and 36 month progress updates were added to the website, as well as a lay overview.

4.2. IPR management and Technology transfer

With regard to IPR, most of the work has been done over the latest 9 month period. Several meetings have been held with INSERM Transfer. Letters of invention around chips (5 on chips and 2 on capture libraries) have been written and will be annexed to the patents.

Each partner has nominated a representative that will deal with all the IP aspects related to the NMD-Chip project within his/her organisation and will also participate in the periodic IPC meetings.

The first IPC meeting has been held in Stockholm, and there has been one per year, 3 in total.

At Stockholm's IPC, guidelines for the dissemination and protection of results and models of foreground exploitation have been discussed. Further to this meeting, and in accordance to decisions taken, the agreement with Roche was finalised and a power of attorney was requested at each partner level. The collaboration agreement with Roche is now under discussion between both parts.

On the occasion of the 3rd Steering Committee Meeting held in Ferrara at M18, an IP round-table was organised. The IP follow-up table was introduced to project partners with last updates and fruitful discussions were held. Fundamental rules of the NMD-Chip Consortium Agreement were reminded to partners:

- The beneficiaries shall report on the expected use to be made of foreground, filling extensively the IP Database. Taking into account that it is too early to define precisely any ownership; it was agreed to fill in the IP database with all knowledge of interest produced and to be produced by NMD-Chip, with a commercial impact or not (Standard Operating Protocol, Best Practices, Guidelines to be disseminated, etc.).

- Foreground shall be the property of the beneficiary carrying out the work generating that foreground. Where several beneficiaries have jointly carried out work generating foreground and where their respective share of the work cannot be ascertained, they shall have joint ownership of such foreground. They shall establish a specific agreement defining the allocation and terms of exercising that joint ownership.

- In case of joint ownership of Foreground as long as no joint ownership agreement has been concluded yet, each of the joint owners shall be entitled to use the joint foreground as it sees fit for its proper internal research activities on a royalty-free basis, and to grant nonexclusive licenses to third parties, without any right to sub-license subject to the following conditions: at least 45 Days prior notice must be given to the other joint owner(s); and fair and reasonable compensation must be provided to the other joint owner(s).

Due to strong links between NMD-Chip project and TREAT-NMD network of excellence, as well as good communication reports on the project, several external collaborations have been asked to the coordinator. These requests have been submitted to the Consortium at M17 and M18.

All partners were consulted by email and all agreed on the following guidelines for external collaborations, considering the 3 different cases that can occur:

1. External partners willing to send us DNAs for validation of NMD chips. This would need a MTA to be signed. The MTA is usually written by the sender, what means that each time, the document will be different. In such a case, the concerned partner just has to send the external MTA to the Coordination team and INSERM Transfert for validation, which should not be a problem.

2. External partners willing the Consortium to send them chips for "in situ" testing in their laboratory. This is a come out of the Consortium material. In such a case, the concerned partner has to inform the Coordination team and INSERM Transfert, who will provide him with a draft of a "NMD-Chip" MTA to send to the requesting lab. Such a document has been prepared by the IP leader INSERM Transfert and is available for each Consortium member on the intranet site.

3. External partners willing to exchange both information and material with the Consortium in order to collaborate. This is the most complicated case. We thus would have first to refer to the UE and establish a global collaboration project.

In order to take in account the project evolutions and competition, and in agreement with INSERM Transfert and Cabinet PLASSERAUD (Paris), the redaction of 2 patents were engaged covering the whole patients' diagnosis process with the CGH chips: one patent dedicated to the in vitro diagnostic of DMD, LGMD and CMD Diseases, and the second one dedicated to the in vitro diagnostic of CMT Disease.

Patents will cover the whole process from DNA extraction of a patient to diagnosis including the use of a first Known-gene microarray (probes designed within the project), completed if needed by the successive use of a potential-gene microarray (probes designed within the project), gene capture libraries (probes designed within the project) and high-throughput sequencing. A Patent Cooperation Treaty (PCT) will then plan after 12 months.

Concerning technology transfer, all definitive protocols concerning CGH-microarray hybridizations, and sequence captures were finalized and collected in order to assure technology transfer to the teams who will be in charge of processing NMD-Chip diagnosis.

INSERM Transfert, the private technology transfer subsidiary of INSERM, will be in charge of the management of the patents and of the exploitation plan report built in order to ensure protection and exploitation of results. INSERM will also take over industrial prospection for partners with an interest in exploiting methods. Preference will be given to partners of the NMD-Chip consortium in accordance with the philosophy and rules of FP7 consortium: PCHIP should so be considered as the best candidate as the company has yet a commercial activity in the field of diagnostic services based on chip technology.

The licensing model will then be based on development plan taking in account the products (DNA-Chips, assay, method, kit), the domain (diagnostic of NMD and CMT diseases), territories and, in case of, sublicensing. INSERM Transfert will finally be in charge of Revenue sharing (upfront, milestones, royalties and sub-licenses revenues) between partners based on their contribution. Inventors will be rewarded by their institution based on country law internal rules.

List of Websites:

Project website

http://www.nmd-chip.eu/

Project participants

INSERM U910 Marseille

Nicolas LEVY - Project Coordinator

Patrice BOURGEOIS, scientific assistant of the coordinator

Claudine MAGNE, administrative assistant of the coordinator

Valérie DELAGUE

Marc BARTOLI

Martin KRAHN

Inserm UMR_S910 - Génétique médicale & Génomique Fonctionnelle - Faculté de Médecine de Marseille - 13385 Marseille cedex - FRANCE

INSERM U827 Montpellier

Christophe BEROUD

Gwenaëlle COLLOD-BEROUD

Karine DELETANG

641 avenue du doyen Gaston GIRAUD - 34093 Montpellier - FRANCE

INSERM U567 Paris

Jamel CHELLY

Mireille COSSEE

France LETURCQ

Hôpital Cochin - Paris - FRANCE

INSTITUTE OF MYOLOGY

Thomas VOIT

Gisèle BONNE

Isabelle NELSON

Pascale RICHARD

Valérie ALLAMAND

Hôpital de la Pitié Salpêtrière - Paris - FRANCE

GENETHON

Isabelle RICHARD

Rafael DE CID

1 bis rue de l'Internationale - BP60 - F-91002 Evry Cedex - FRANCE

KAROLINSKA INSTITUTE

Thomas SEJERSEN

Fengqing XIANG

17176 STOCKHOLM - SWEDEN

LEIDEN UNIVERSITY OF MEDICAL RESEARCH

Johan DEN DUNNEN

Bert BAKKER

Ieke GINJAAR

Rowida EL MOMANI

Department of Human Genetics - DMD Genetic Therapy Group - Post-Zone S04-034 - THE NETHERLANDS

NATIONAL INSTITUTE OF ENVIRONMENTAL HEALTH

Veronika KARCAGI

Beata DUDAS

National Center for Public Health - Dept. of Molecular Genetics and Diagnostics - 1097 Budapest,Gyáli út 2-6 - HUNGARY

NEWCASTLE UNIVERSITY

Volker STRAUB

Rachel THOMPSON

Cathy TURNER

Pauline MAC CORMACK

Simon WOODS

Kate BUSHBY

TREAT-NMD Office - Institute of Genetic Medicine - Newcastle University - International Centre for Life - Newcastle upon Tyne - NE1 3BZ - UNITED KINGDOM

PARTNERCHIP

Pascal SOULARUE

Sylvain BAULANDE

Linhda COUVELARD

PartnerChip - Bat G2 - 2, rue Gaston Crémieux - 91000 EVRY - FRANCE

PHENOSYSTEMS SA

David ATLAN

Beat WOLF

SWITZERLAND

TECHNISCHE UNIVERSITAET DRESDEN

Angela HUEBNER

Katrin KOEHLER

Dresden - GERMANY

UNIVERSITY COLLEGE LONDON

Francesco MUNTONI

Sebahattin CIRAK

Michael YAU

Victoria CASTELMAN

Dubowitz Neuromuscular Unit - 1st Floor Institute of Child Health - 30 Guildford Street - London WC1N 1EH - UNITED KINGDOM

UNIFE

Alessandra FERLINI

Marcella NERI

Dipartimento di Medicina - Sperimentale e Diagnostica - Sezione di Genetica Medica - Via Fossato di Mortara, 74 - 44100 - Ferrara - ITALY

UNIVERSITY OF WUERZBURG

Clemens MUELLER-REIBLE

Kai HEINECKE

Würzburg - GERMANY