



# Decoding Neuromuscular Disorders Using Phenotypic Clusters Obtained From Co-Occurrence Networks

Elena Díaz-Santiago<sup>1</sup>, M. Gonzalo Claros<sup>1,2,3,4\*</sup>, Raquel Yahyaoui<sup>3,5</sup>, Yolanda de Diego-Otero<sup>3</sup>, Rocío Calvo<sup>3,5</sup>, Janet Hoenicka<sup>2,6</sup>, Francesc Palau<sup>2,6,7</sup>, Juan A. G. Ranea<sup>1,2,3†</sup> and James R. Perkins<sup>1,2,3†</sup>

<sup>1</sup>Department of Molecular Biology and Biochemistry, Universidad de Málaga, Málaga, Spain, <sup>2</sup>CIBER de Enfermedades Raras (CIBERER), Madrid, Spain, <sup>3</sup>Institute of Biomedical Research in Malaga (IBIMA), IBIMA-RARE, Málaga, Spain, <sup>4</sup>Institute for Mediterranean and Subtropical Horticulture "La Mayora" (IHSM-UMA-CSIC), Málaga, Spain, <sup>5</sup>Laboratory of Metabopathies and Neonatal Screening, Málaga Regional University Hospital, Málaga, Spain, <sup>6</sup>Sant Joan de Déu Hospital and Research Institute, Barcelona, Spain, <sup>7</sup>Hospital Clinic and University of Barcelona School of Medicine and Health Sciences, Barcelona, Spain

## OPEN ACCESS

### Edited by:

Silvia Bottini,  
Université Côte d'Azur, France

### Reviewed by:

Tao Huang,  
Shanghai Institute of Nutrition and  
Health (CAS), China  
Filippo M. Santorelli,  
Fondazione Stella Maris (IRCCS), Italy  
Yu-Hang Zhang,  
Brigham and Women's Hospital,  
United States

### \*Correspondence:

M. Gonzalo Claros  
claros@uma.es

†These authors have contributed  
equally to this work

### Specialty section:

This article was submitted to  
Biological Modeling and Simulation,  
a section of the journal  
Frontiers in Molecular Biosciences

**Received:** 29 November 2020

**Accepted:** 15 February 2021

**Published:** 19 April 2021

### Citation:

Díaz-Santiago E, Claros MG, Yahyaoui R, de Diego-Otero Y, Calvo R, Hoenicka J, Palau F, Ranea JAG and Perkins JR (2021) Decoding Neuromuscular Disorders Using Phenotypic Clusters Obtained From Co-Occurrence Networks. *Front. Mol. Biosci.* 8:635074. doi: 10.3389/fmolb.2021.635074

Neuromuscular disorders (NMDs) represent an important subset of rare diseases associated with elevated morbidity and mortality whose diagnosis can take years. Here we present a novel approach using systems biology to produce functionally-coherent phenotype clusters that provide insight into the cellular functions and phenotypic patterns underlying NMDs, using the Human Phenotype Ontology as a common framework. Gene and phenotype information was obtained for 424 NMDs in OMIM and 126 NMDs in Orphanet, and 335 and 216 phenotypes were identified as typical for NMDs, respectively. 'Elevated serum creatine kinase' was the most specific to NMDs, in agreement with the clinical test of elevated serum creatinine kinase that is conducted on NMD patients. The approach to obtain co-occurring NMD phenotypes was validated based on co-mention in PubMed abstracts. A total of 231 (OMIM) and 150 (Orphanet) clusters of highly connected co-occurrent NMD phenotypes were obtained. In parallel, a tripartite network based on phenotypes, diseases and genes was used to associate NMD phenotypes with functions, an approach also validated by literature co-mention, with KEGG pathways showing proportionally higher overlap than Gene Ontology and Reactome. Phenotype-function pairs were crossed with the co-occurrent NMD phenotype clusters to obtain 40 (OMIM) and 72 (Orphanet) functionally coherent phenotype clusters. As expected, many of these overlapped with known diseases and confirmed existing knowledge. Other clusters revealed interesting new findings, indicating informative phenotypes for differential diagnosis, providing deeper knowledge of NMDs, and pointing towards specific cell dysfunction caused by pleiotropic genes. This work is an example of reproducible research that i) can help better understand NMDs and support their diagnosis by providing a new tool that exploits existing information to obtain novel clusters of functionally-related phenotypes, and ii) takes us another step towards personalised medicine for NMDs.

**Keywords:** neuromuscular disorders, rare disease, phenotype, network analysis, cluster, co-occurrence analysis

**Abbreviations:** ECM, extracellular matrix; GO, gene ontology; HPO, human phenotype ontology; HyI, hypergeometric index; KEGG, kyoto encyclopedia of genes and genomes; MD, muscular dystrophy; NMD, neuromuscular disease; non-NMD, non-neuromuscular disease; OMIM, online Mendelian inheritance in man; PMID, PubMed identifier; SLC, solute carrier.

## 1 INTRODUCTION

Neuromuscular disorders (NMDs) encompass a range of pathologies affecting muscle function (Roy et al., 2015) that can be caused by problems in spinal motor neurones, peripheral nerves, muscles, and neuromuscular junctions. They affect 6–8 million people worldwide (Scotton et al., 2014) and lead to elevated morbidity and mortality (McCormack et al., 2013). Many result from genomic mutations (Laing, 2012), although they can also be caused by autoimmune disorders and infections (Kraker and Zivković, 2011). Around half manifest during childhood and classification is often based on the affected area (Turakhia et al., 2013). Updated details of known mutations associated with NMDs are published yearly (Benarroch et al., 2019) (<http://www.musclegenetable.fr/index.html>). **Table 1** illustrates the main types of NMDs and shows their high heterogeneity in terms of clinical manifestation.

Although diagnosis of NMDs has been aided in recent years by advances in whole-exome/genome sequencing (Brown and Meloche, 2016), it still requires a high level of medical specialisation, due to the high phenotypic and pathophysiology diversity, and large number of causal genes (McDonald, 2012; Roy et al., 2015). Moreover, given that individual NMDs tend to be rare, it can be hard to find sufficient patients to conduct well-powered studies. As such, a range of diagnostic tests (including electrophysiology, tissue biopsies, and measuring levels of certain enzymes such as elevated serum creatinine kinase), and molecular imaging (Thavornatanaburt et al., 2018) must be used, and diagnosis can be slow (Spuler et al., 2011). There is no cure for most NMDs but rather symptomatic treatments to delay progression. Current research is focused on gene therapies and investigating new medications (Scoto et al., 2018). However, efforts are hampered by their etiological heterogeneity and phenotype diversity.

Further work is needed to better understand how NMDs are related in terms of phenotypic overlap and underlying genes and mechanisms, in order to facilitate diagnosis and improve treatment. Although not specific to NMDs, previous studies have compared phenotypic profiles between different diseases to build clusters of related phenotypes (Sirota et al., 2009; Bagley et al., 2016); others have focused on phenotype-similarity based on co-morbidity across multiple diseases (Rzhetsky et al., 2007; Hidalgo et al., 2009). Such studies enable us to identify patterns between groups of diseases and phenotypes by showing how they tend to co-occur. This has multiple potential uses for disease classification and diagnosis. Further studies have investigated the connection between clinical manifestations in disease by integrating gene-disease and protein-protein interaction data (Yang et al., 2011; Hwang et al., 2012; Zhou et al., 2014), as well as connecting proteins with phenotypes through the use of phenotypic clusters based on similarity and by predicting proteins associated with the phenotypes through machine learning (Ren et al., 2020). These studies lead the way towards explaining the co-occurrence of phenotypic patterns across diseases through common underlying mechanisms.

There are several resources that map known diseases to their pathological phenotypes and associated genes. For example, MENDELIAN (<https://www.mendelian.co/es/>) allows the symptom-guided search of rare diseases. OMIM (Online Mendelian Inheritance in Man) gathers data obtained *via* curation of the biomedical literature (Hamosh, 2004; Amberger et al., 2019) to provide a clinical synopsis of all known Mendelian traits and disorders, describing genes, allelic variants and pathological phenotypes. Orphanet (<http://www.orphandata.org> and <https://www.orpha.net/>) is based on expert knowledge, gathering information about rare diseases specifically, with the aim of collecting and unifying the scarce knowledge available for such disorders, including NMDs of genetic origin. There are several tools that exploit these databases to associate disease-related phenotypes with genes, including *Phen2Gene* (Zhao et al., 2020), *AMELIE* (Birgmeier et al., 2020), *Phevor* (Singleton et al., 2014), *Phenolizer* (Yang et al., 2015) and *Phenomizer* (Köhler et al., 2009). Such studies require the use of common phenotype terminology in such a way that makes the information amenable to computational analysis (Hoehndorf et al., 2015). The Human Phenotype Ontology (HPO, <https://hpo.jax.org/>), provides such a standardised vocabulary to describe phenotypic abnormalities associated with more than 7 800 diseases (Köhler et al., 2021).

However, few studies have combined phenotype and gene information for groups of heterogeneous diseases to look for related phenotypes shared across multiple disorders with common underlying mechanisms. Such an approach was recently conducted by our group using patients with largely undiagnosed rare disorders (Díaz-Santiago et al., 2020), taking advantage of the rare-disease database DECIPHER (Firth et al., 2009). This work showed that by using biomedical networks and systems medicine approaches our understanding of rare diseases can be improved based on phenotype co-occurrence patterns. It also showed the power of the re-analysis of existing data from public databases to obtain new knowledge, something that is recommended in the research community whenever possible (Kovalevskaya et al., 2016; Tan et al., 2020). In this work, the automated workflow *PhenoClusters* is used to investigate phenotype co-occurrence across NMDs and produce functionally coherent clusters of phenotypes with similar underlying biological functions. This can help differentiate diagnosis (‘elevated serum creatine kinase’ is the most significant NMD phenotype) and provide a better understanding of NMDs (many clusters gather typical phenotypes and functions of NMDs) based on the specific cell functions (unanticipated phenotypes such as ‘macroglossia’ and ‘arthrogryposis’ point to cell dysfunctions involved in an NMD), including those affected by pleiotropic/multi-functional genes.

## 2 MATERIALS AND METHODS

### 2.1 *PhenoClusters* Workflow

The *PhenoClusters* workflow described in this study is based on OMIM, Orphanet and HPO data and is outlined in **Figure 1**. The following sections describe its modules in more detail.

**TABLE 1** | Classification of NMDs including some examples.

Type	Description	Examples
Muscular dystrophies (MDs)	Diseases causing weakness and degeneration of the skeletal muscles	Myotonic dystrophy; facioscapulohumeral MD; EDMD; Duchenne MD; Becker MD; LGMDs; congenital MDs
Myopathies	Muscle diseases in which the muscle fibres do not function properly, resulting in muscle hypotonia and weakness	Congenital myopathies; distal myopathies; endocrine myopathies; mitochondrial myopathies; metabolic myopathies
Peripheral nerve diseases	Diseases where motor and sensory nerves that connect the brain and spinal cord to the rest of the body are affected, causing impaired sensations, movements and muscular weakness	Charcot-Marie-Tooth disease; Giant axonal neuropathy
Motor neurone diseases	Diseases where motor neurones progressively lose function, causing the muscles they control to become weak and eventually non-functional	Hereditary spastic paraplegias; spinal muscular atrophy; spinal-bulbar muscular atrophy
Ion channel diseases	Diseases associated with defects in ion channels, typically marked by muscular weakness, absent muscle tone, or episodic muscle paralysis	Andersen-Tawil syndrome; hyperkalemic periodic paralysis; hypokalemic periodic paralysis; myotonia congenita; paramyotonia congenita; potassium-aggravated myotonia
Neuromuscular junction diseases	Neuromuscular junction disorders that result from the destruction, malfunction or absence of one or more key proteins involved in the transmission of signals between muscles and nerves	Congenital myasthenic syndromes; Lambert-Eaton myasthenic syndrome; myasthenia gravis

EDMD: *Emery-Dreifuss muscular dystrophy*; LGMDs: *limb girdle muscular dystrophies*.

### 2.1.1 Module A: NMD Datasets With Genes, Phenotypes and Typical NMD Phenotypes

The OMIM and Orphanet databases (as of 15th-Nov-2019) were used in parallel since they gather information from different origins and with different goals. OMIM was searched with NMD related keywords (*muscular dystrophy, myopathy, myasthenic, myasthenia, neuropathy, amyotrophic lateral sclerosis, spinal muscular atrophy, spinal and bulbar muscular atrophy, myotonia, periodic paralysis, myotonic dystrophy, mitochondrial cytopathy, necrotizing encephalomyelopathy, mitochondrial DNA depletion*) to obtain a list of NMDs as complete as possible. Since Orphanet organises diseases based on ORDO (Orphanet Rare Disease Ontology) (Vasant et al., 2014), all diseases descendant of the category neuromuscular diseases (Orphanet:68381) were used. Hence, diseases in OMIM and Orphanet were assigned to “All NMDs” or “All non-NMDs” (grey rectangles in Module A of **Figure 1** and **Table 2**). These were used to query HPO (v1.2; 15th-Nov-2019) to retain only those diseases (“Non-NMDs” and “NMDs” in **Table 2**, corresponding to red rectangles in Module A of **Figure 1**) for which both phenotype (white rectangles in Module A of **Figure 1**) and gene (teal rectangles in Module A of **Figure 1**) information was available.

Finally, phenotype frequencies between “NMDs” and “non-NMDs” in **Table 2** and **Figure 1** were compared using Fisher’s exact test (one-tail) with a threshold of  $P \leq 0.05$ . A list of “Typical NMD phenotypes” (**Table 2** and **Figure 1**) was thus obtained.

### 2.1.2 Module B: Bipartite Network to Obtain Phenotype Co-Occurrences

All diseases (“NMDs” and “non-NMDs”) as well as their corresponding phenotypes were used to construct a standard bipartite network (Pavlopoulos et al., 2018) of disease-phenotype pairs to find common (co-occurring) phenotypes across diseases. Phenotype co-occurrences (grey rectangle in Module B of **Figure 1**) were extracted using *NetAnalyzer* (Rojano et al., 2017) (see below). To recover only significant associations for

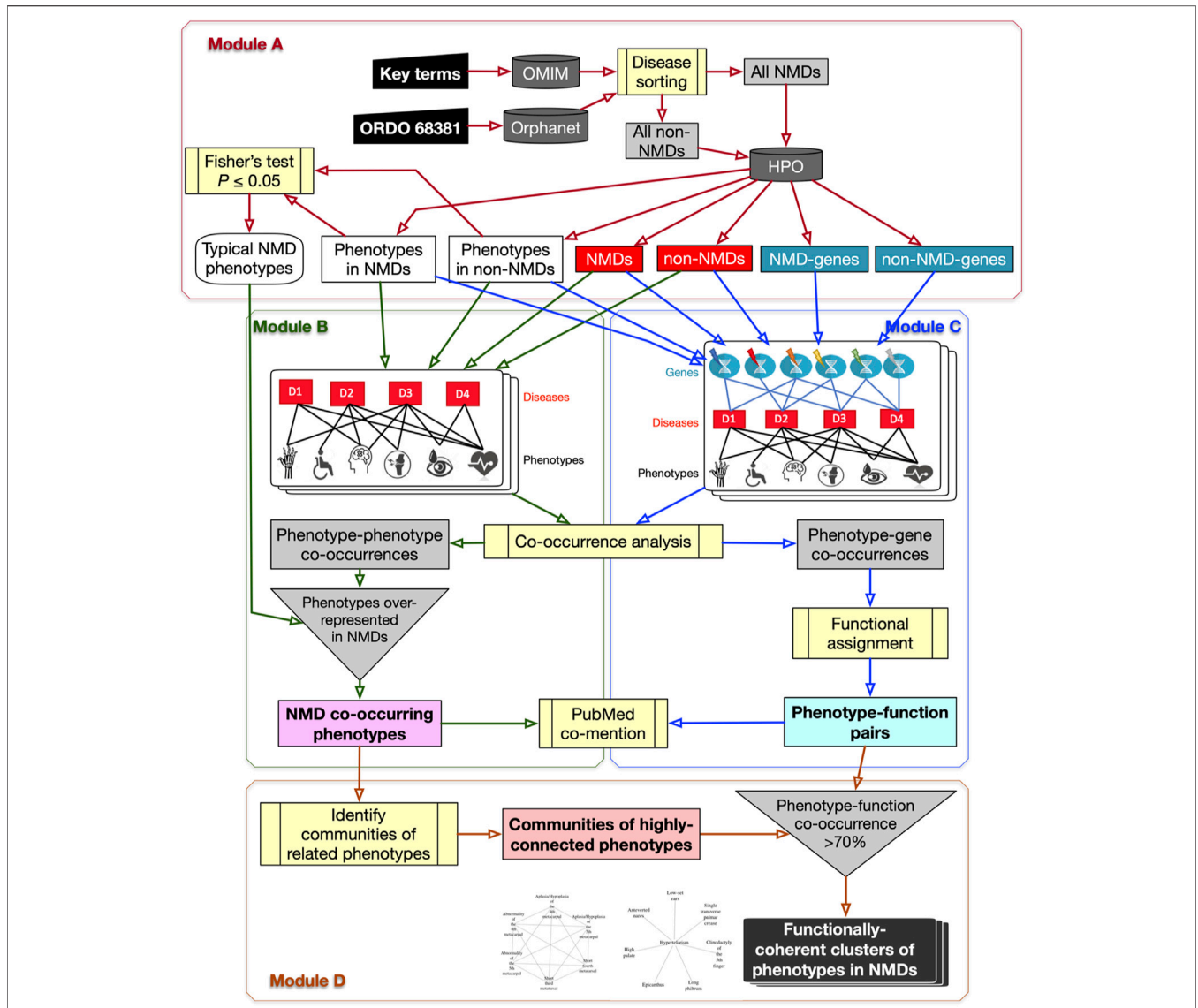
NMDs (pale magenta rectangle of “NMD co-occurring phenotypes” in Module B of **Figure 1**), all phenotype-phenotype co-occurrences were filtered using the “Typical NMD phenotypes” obtained from Module A.

### 2.1.3 Module C: Tripartite Network to Obtain Phenotype-Function Pairs

A tripartite network of all diseases (NMDs and non-NMDs), their phenotypes and the affected genes was constructed to link, in first instance, phenotypes with genes, based on their co-occurrence across diseases. The network was constructed as described in (Rojano et al., 2017) and then investigated using *NetAnalyzer* to find significant phenotype-gene co-occurrences (grey rectangle in Module C of **Figure 1**). Phenotype-gene pairs were converted into phenotype-function pairs based on the biological process sub-ontology from Gene Ontology (GO), KEGG (Kyoto Encyclopaedia of Genes and Genomes) pathways and Reactome pathways. To reveal those enriched functions significantly associated with a given phenotype, all genes associated with that phenotype were obtained and this gene list was used for functional enrichment, based on over-representation analysis (Yu et al., 2012). Association between phenotypes and genes/functions was performed separately for OMIM and Orphanet. This resulted in significant “phenotype-function pairs” (pale cyan rectangle in bold in Module C of **Figure 1**).

### 2.1.4 Module D: Functionally-Coherent Clusters of Phenotypes in NMDs

The typical NMD phenotypes obtained in Module B were used to detect communities of related, highly interconnected phenotypes (grey rectangle at bottom centre of **Figure 1**) using the R package *linkcomm*. Following the rationale proposed in (Díaz-Santiago et al., 2020), phenotype communities for which at least 70% of the constituent phenotypes shared the same functional annotations obtained from the phenotype-function pairs of Module C were



**FIGURE 1 |** Flowchart of the *PhenoClusters* workflow. Module (A): all diseases in OMIM and Orphanet were classified as “NMDs” or “non-NMDs”. HPO is then queried with those disease, retaining only those diseases (red boxes) for which both a gene (teal boxes) and a phenotype (white boxes) are known. Module (B): a bipartite network was constructed and analysed to obtain the typical “NMD co-occurring phenotypes” that tended to occur in NMDs. Module (C): a tripartite network enabled the obtention of new, significant phenotype-gene pairs that were translated to “phenotype-function pairs”. Module (D): communities of highly related phenotypes were filtered based on shared function to obtain the final set of “functionally coherent clusters of phenotypes in NMDs”. PubMed co-mention was used to assess the reliability of the co-occurrence approach. Dark grey cylinders correspond to databases; yellow boxes are relevant analyses; inverted triangles indicate process merging; light grey rectangles are intermediate sets of results; coloured rectangles with text in bold mark relevant results.

retained. These were considered “functionally coherent clusters of phenotypes in NMDs” (final dark rectangle in Module D of **Figure 1**) and constitute the major result of the workflow. OMIM clusters were tagged with “\*”, while Orphanet clusters were tagged with “#”.

## 2.2 PubMed Co-Mention to Validate the Co-Occurrence Approach

Phenotypes co-occurring in NMDs from Module B and phenotype-function pairs from Module C were investigated in

terms of co-mention in the scientific literature, by comparing how many PubMed abstracts mention both terms in a given pair, to how many abstracts mention only one of them. To do so, the NCBI Entrez Programming Utilities API was used as previously described (Díaz-Santiago et al., 2020). In brief, the different textual descriptions for a given i) phenotype as described in HPO, ii) GO term from the biological process vocabulary, iii) KEGG pathways, and iv) Reactome pathways, were retrieved. All these terms were queried separately in PubMed (the complete database) to obtain the lists of PMIDs (PubMed identifiers) of abstracts mentioning each given term. PMID lists were then



**TABLE 2 |** Summary of diseases found in OMIM and Orphanet databases and the retained diseases, phenotypes and genes after the HPO query.

	OMIM			Orphanet		
	Diseases	HPO		Diseases	HPO	
		Phenotypes	Genes		Phenotypes	Genes
Total	26 943	-	-	3 431	-	-
All non-NMDs	26 387	-	-	3 204	-	-
All NMDs	556	-	-	227	-	-
Complete information in HPO	5189	6385	4015	2020	5430	2730
Non-NMDs	4765	6183	4015	1894	5312	2730
NMDs	424	1314	345	126	1007	222
Common	0	1112	345	0	889	222
Typical in NMDs	-	335	0	-	216	0

"-": not considered or not relevant for the analysis.

compared for each pair of terms (phenotype-phenotype, phenotype-GO, phenotype-KEGG or phenotype-Reactome) using Fisher's exact test to detect the significantly ( $P \leq 0.05$ ) co-mentioned pairs. The numbers of significant pairs obtained for each set was then compared to the number of significant pairs obtained from random models, by calculating whether the probability of finding as many significant pairs in the random dataset was at least as high as that detected using real data.

## 2.3 Tissue Expression Testing

Genes in phenotype-gene associations were further analysed to see if they were more likely to be expressed in neuronal or muscular tissues, both highly relevant for NMDs. To do so, expression location of genes paired with NMD phenotypes was compared to that of genes paired with non-NMD phenotypes. Expression locations were obtained from the normal tissue expression dataset (*normal\_tissue.tsv.zip* file) at The Human Protein Atlas version 20.0 (Uhlén et al., 2015) (<http://www.proteinatlas.org>). A gene was considered as expressed in neuronal or muscular tissue if it showed medium or high expression in one of the following tissue types: cerebellum, cerebral cortex, hippocampus, caudate neuronal cells, skeletal muscle or heart muscle. The relative proportions of genes from each set that showed expression in neuronal or muscular tissue were compared using Fisher's exact test.

## 2.4 PhenoClusters Architecture and Execution Details

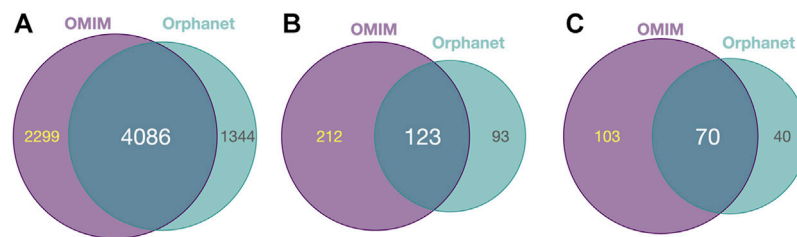
All the above calculations have been implemented as an automated workflow named *PhenoClusters* that is based on *PhenCo* (Díaz-Santiago et al., 2020) and uses additional scripts from (Jabato et al., 2020). The Picasso supercomputer of University of Malaga was used for code implementation and testing. It consists of an OpenSUSE LEAP 12.3 with Slurm queue system and Infiniband network (54/40 Gbps) containing 216 nodes with Intel E5-2670 2.6 GHz cores for a total of 3 456 cores and 22 TB of RAM. The code is available from GitHub at <https://github.com/Elenadisa/PhenoClusters>. The main requirements are Python 3, Ruby 2.4.1, R 4.0.0 or higher, Bioconductor 3.4 (Huber et al., 2015) or higher, scripts from *sys\_bio\_lab* (Jabato et al., 2020), and *Anaconda Individual edition* (<https://docs.anaconda.com/anaconda/install/>) as package and environment manager. The workflow was managed using *AutoFlow* (Seoane et al., 2016).

For co-occurrence analysis, *PhenoClusters* requires *NetAnalyzer* (Rojano et al., 2017). Although many co-occurrence methods are available, previous work suggests that the hypergeometric index (*HyI*) (Fuxman Bass et al., 2013), which can be considered analogous to a contingency table based approach, is the most suitable for co-occurrence studies (Rojano et al., 2017; Bueno et al., 2018). Hence,  $HyI \geq 2$  was set to calculate associations within the layers of a network and obtain phenotype-phenotype or phenotype-gene associations that were considered significant co-occurrences (Díaz-Santiago et al., 2020).

Functional enrichment was performed using the genes associated with each phenotype. The Bioconductor package *clusterProfiler* 3.18.0 (Yu et al., 2012) was used to extract the biological process subontology from GO and the KEGG pathways and conduct over-representation analysis for these resources. The Bioconductor package *ReactomePA* 1.34.0 (Yu and He, 2016) was used to extract Reactome pathway data and perform over-representation analysis for this resource. Both packages were executed with default parameters ( $pvalueCutoff = 0.05$ ,  $pAdjustMethod = "BH"$ ,  $universe$ ,  $qvalueCutoff = 0.2$ ,  $minGSSize = 10$ ,  $maxGSSize = 500$ ) and only the assignments with  $P \leq 0.05$  after Benjamini-Hochberg multiple testing adjustment (Benjamini and Hochberg, 1995) were considered significant.

Highly interconnected phenotype communities were obtained using the R-CRAN package *linkcomm* 1.0.13 (Kalinka and Tomancak, 2011) with default parameters for undirected networks. R-CRAN packages required for graphical representations and data management are *ggplot2* 3.3.2, *RColorBrewer* 1.1.2, *igraph* 1.2.6, *dplyr* 1.0.2 and *VennDiagram* 1.6.20. The final, user-friendly HTML reports were produced using R markdown packages *rmarkdown* 2.6, *knitr* 1.30 and *kableExtra* 1.3.1.

Details about *PhenoClusters* execution can be found at <https://github.com/Elenadisa/PhenoClusters>. As can be seen, diseases from OMIM and Orphanet were considered separately in different scripts. One run usually takes at least 6 h 45 min without the co-mention verification, since literature analysis



**FIGURE 2** | Common and distinct disease-associated HPO-phenotypes in OMIM and Orphanet described in **Table 2**. **(A)**: Distribution of the 6 385 HPO phenotypes for OMIM and the 5 430 phenotypes for Orphanet for all NMDs and non-NMDs. **(B)**: Distribution of the typical 335 OMIM and 216 Orphanet phenotypes significantly ( $P < 0.05$ ) associated with NMDs. **(C)**: Same as B, but using the increased significance ( $P < 0.001$ ) NMD-associated phenotypes (172 for OMIM and 110 for Orphanet).

takes a very long, heterogeneous amount of time depending on the PubMed server overload. The resulting reports are named *omim\_report.html* (**Supplementary Material S1**) and *orphanet\_report.html* (**Supplementary Material S2**). To simplify cluster inspection, they are also saved as independent files in *omim\_clusters\_report.html* and *orphanet\_clusters\_report.html* including the cluster ID, the phenotype graph together with HPO IDs and descriptions, as well as the associated functions and the genes identified within them. Clusters showing functional coherence are given at the top of the file. For the sake of simplicity, “\*” is used to tag OMIM clusters, while Orphanet clusters are tagged by “#”, as indicated in Module D.

## 3 RESULTS

### 3.1 Typical NMD Phenotypes Reflect Neuromotor Impairments

Diseases in OMIM and Orphanet were classified as NMD or non-NMD (**Table 2**). All downstream analysis treated the resources separately (**Supplementary Material S1, S2**). The total numbers of diseases in each resource for which genes and pathological phenotype information was available are shown in **Table 2**; full details can be found in **Supplementary Material S3, S4**. It can be seen that relatively few diseases were considered NMDs with known genes and phenotypes in HPO (424 and 125 in OMIM and Orphanet, respectively). Many phenotypes were shared by both NMDs and non-NMDs (row “Common” in **Table 2**). Fisher’s exact test was then performed to compare the relative occurrence of each phenotype between disease groups to find those relatively more common in NMDs. This resulted in 335 and 216 typical NMD phenotypes for OMIM and Orphanet. Overlap between the two is shown in **Figure 2**, where Panels B and C show that the proportion of common phenotypes in both sets increases with decreasing cutoff  $p$ -values. This reinforces the suitability of the co-occurrence approach to extract typical NMD phenotypes.

The top over-represented phenotypes are shown in **Supplementary Material S5**. The most significant in both resources is ‘elevated serum creatine kinase’, which reflects the important role of muscle decay in NMDs, given that the normal

location of this enzyme is the cytoplasm and mitochondria, not serum (Moghadam-Kia et al., 2016). This result is reassuring with respect to the approach of *PhenoClusters*, as elevated serum creatine kinase is routinely tested when diagnosing NMDs as a consequence of muscle injury (Thavorntanaburt et al., 2018).

The top phenotypes (**Supplementary Material S5**) seem to be more indicative of muscular impairment than neuronal dysfunction, pointing to proximal or distal muscular weakness, both very typical in the clinical history of NMD patients (McDonald, 2012; Mukherjee et al., 2019). Other important phenotypes are related to specific diseases, such as Duchenne/Becker muscular dystrophy (‘calf muscle hypertrophy’, ‘Gowers sign’), where muscle hypertrophy can exaggerate postural instability and joint contracture (Kornegay et al., 2012). In contrast, the absence of ‘hypotonia’-related phenotypes in this top ten list is notable. However, it should be made clear that such phenotypes were also significantly over-represented among NMDs, just not included in the top ten (**Supplementary Material S3, S4**). To conclude, the *PhenoClusters* approach to obtain statistically over-represented phenotypes for NMDs extracts phenotypes characteristic of these disorders.

### 3.2 Typical NMD Phenotypes Tend to Co-Occur Across Diseases

Bipartite networks were created (one for OMIM diseases and another for Orphanet), with nodes representing diseases and phenotypes, and edges representing known relations in OMIM or Orphanet respectively. The network was analysed to identify pairs of typical NMD phenotypes that tend co-occur across diseases (Module B in **Figure 1**). Around a quarter of all phenotype pairs co-occurring in at least one disease showed significant co-occurrence (27.61% in OMIM and 21.20% in Orphanet). Only a small subset ( $\leq 1\%$ ) of all phenotype pairs represent significant co-occurrence between NMD phenotypes. However, this value was far lower for the random sets of non-NMD phenotypes, being almost negligible (**Table 3**), showing that pairs of NMD phenotypes are far more likely to significantly co-occur across multiple diseases than pairs of randomly selected phenotypes. These results point to the presence of comorbidity between NMD phenotypes.

**TABLE 3** | Numbers of pairs of co-occurring phenotypes for the OMIM and Orphanet databases. Significantly co-occurring pairs defined as having a hypergeometric index value  $\geq 2$ .

Category	OMIM	Orphanet
Total co-occurring phenotype pairs	455 379	467 899
Significant co-occurring pairs ( $Hyl \geq 2$ )	125 720	99 193
Significant co-occurring pairs, NMD phenotypes	4 579	2 314
Significant co-occurring pairs, non-NMD phenotypes <sup>a</sup>	655 $\pm$ 46.4	173 $\pm$ 16.43

<sup>a</sup>This result is expressed as the average  $\pm$  SD for 100 sets of 335 randomly chosen OMIM non-NMD phenotypes and 216 random Orphanet non-NMD phenotypes; the non-NMD phenotypes were selected to have a similar prevalence to the NMD phenotypes.

### 3.3 Co-Occurring Phenotypes Tend to be Co-Mentioned in Research Papers

As well as showing that the typical NMD phenotypes tended to co-occur across diseases much more frequently than non-NMD phenotypes, further validation was performed based on co-mention analysis for each phenotype pair within PubMed abstracts. The NMD phenotype pairs were shown to be co-mentioned in the literature much more frequently than equal numbers of pairs of randomly chosen non-NMD phenotypes (Table 3). Fisher's exact test analysis revealed that 1 173 OMIM pairs and 689 Orphanet pairs (Supplementary Tables S1, S2) were significantly co-mentioned. Conversely, for non-NMDs, the numbers of significant pairs were much lower ( $333.08 \pm 34.63$  in OMIM and  $110.0 \pm 13.47$  in Orphanet, Supplementary Tables S1, S2, respectively). These differences were also supported by the distribution of their  $p$ -values (Supplementary Figures S1, S2) that were clearly more dense at  $P \leq 0.05$  for NMDs than in non-NMDs. These results demonstrate the validity of the *PhenoClusters* approach to find pairs of NMD-related phenotypes that tend to co-occur together.

### 3.4 Phenotype-Function Associations are Robust

Associations between NMD phenotypes and genes and functions were calculated in order to uncover potential underlying mechanisms leading to the expression of a pathological phenotype. This was done by initially connecting phenotypes with genes using a tripartite network based approach (Module C in Figure 1) by combining all disease-phenotype pairs and all disease-gene pairs in OMIM and Orphanet separately. The networks were then analysed in order to connect genes with phenotypes based on co-occurrence across multiple diseases.

Using a threshold of  $Hyl \geq 2$ , 34 139 and 36 218 phenotype-gene associations were obtained for OMIM and Orphanet, respectively; from these, a total of 1 078 and 611 distinct genes were associated with NMD phenotypes, respectively. There was significant overlap between OMIM and Orphanet in terms of genes associated with NMD phenotypes ( $P < 2.2 \times 10^{-16}$ , Fisher's exact test using all phenotype-associated genes as background). Moreover, a significantly greater proportion of the NMD associated genes (84.41% in OMIM, 86.73% in Orphanet) were shown to be expressed in neuronal or muscle tissue than the non-NMD associated genes (70.32% in OMIM, 72.88% in Orphanet), which is a statistically

significant difference ( $P = 2.2 \times 10^{-16}$  and  $P = 6.3 \times 10^{-11}$ , for OMIM and Orphanet, respectively). The same kind of test was performed for the remaining non-relevant tissues, revealing in this case that there was no significant difference between NMD and non-NMD genes after a Fisher's test ( $P > 0.9$  for both databases).

The genes associated with each phenotype were then used for functional enrichment analysis using *clusterprofiler* and *ReactomePA* to find over-represented ( $P \leq 0.05$ ) GO terms (biological process ontology), and KEGG and Reactome pathways (Table 4). As with gene-based analysis, there was significant overlap between OMIM and Orphanet ( $P < 2.2 \times 10^{-16}$ ,  $6.842 \times 10^{-8}$ , and  $2.381 \times 10^{-14}$ , respectively, using Fisher's exact test with all phenotype-associated functions used as background). Supplementary Material S6 shows the overlap between the different resources using Venn diagrams. Therefore, genes associated with NMD phenotypes are significantly expressed in neuronal and muscle tissue, the most relevant for NMDs, supporting the robustness of the phenotype-function pairs obtained by *PhenoClusters*.

In spite of the robustness of these novel associations between phenotypes and functions, additional validation was performed based on co-mention analysis using PubMed abstracts as a sign of relevance. Term co-occurrence of a given phenotype alongside its corresponding function was looked for within the same abstract (Supplementary Figures S1, S2). Only a subset of the phenotype-function pairs obtained from the network were significantly co-mentioned ("Confirmed" columns in Table 5). However, this subset was several times larger than the number of co-mentioned pairs within the "random" pairs, and this difference was significant for all comparisons ( $P < 2.2 \times 10^{-16}$ ) using Fisher's exact test. Consistent results were found for both OMIM and Orphanet: respectively i) 5% and 4% of phenotype-GO pairs, ii) 10.5% and 9% of KEGG pairs, and iii) 4% and 3% of Reactome pairs were significantly co-mentioned. As such, *PhenoClusters* (Figure 1) appears to produce robust and reliable associations between NMD phenotypes and functions based on phenotype and gene co-occurrence across diseases. Even though only a small fraction could be confirmed in PubMed abstracts, this was several times more than would be expected by chance. The analysis showed much higher co-mention for the KEGG associations as a proportion of the total pairs. This supports our previous work suggesting that information in KEGG is more reliable for functional studies in certain situations (Luque-Baena et al., 2014).

**TABLE 4 |** Numbers of genes and functions associated with phenotypes. Total: all genes in each dataset. "Any": genes/functions associated with any phenotype. "Only NMD": genes/functions associated with a typical NMD phenotype.

Feature	OMIM-based				Orphanet-based				OMIM-Orphanet overlap			
	Genes	GOs	KEGG	Reac	Genes	GOs	KEGG	Reac	Genes	GOs	KEGG	Reac
Total	4015	-	-	-	2730	-	-	-	2452	-	-	-
Any	3870	6398	214	1397	2700	6281	209	1356	2374	6260	208	1348
Only NMD	1078	3747	142	821	611	3618	123	631	360	2666	102	451

"-": not considered or relevant in the analysis.

**TABLE 5 |** Co-mention validation of phenotype-function pairs. "All" corresponds to all pairs including a function from **Table 4**. "Confirmed" refers to the number of these pairs that were significantly co-mentioned in PubMed. "Random" refers to the number of pairs in a randomised list based on "All" pairs that were significantly co-mentioned in PubMed; average of 100 random datasets  $\pm$  SD is shown in this case.

Paired function	Phenotypes					
	OMIM			Orphanet		
	All	Confirmed	Random <sup>a</sup>	All	Confirmed	Random <sup>a</sup>
GO	567 721	26 841	10 814 $\pm$ 80.5	535 389	21 863	9 613 $\pm$ 93.4
KEGG	17 679	1 858	627 $\pm$ 22.1	17 104	1 556	628 $\pm$ 21.2
Reactome	82 826	3 278	1001 $\pm$ 166.3	78 409	2 402	816 $\pm$ 28.7

<sup>a</sup>The randomised phenotype-function pair set was formed by shuffling the links between the pairs in each list, keeping the total number of links per phenotype/function unchanged. This sampling procedure was repeated to produce 100 different replicas of randomised phenotype-function pairs. These sets were used in the corresponding Fisher's exact tests.

**TABLE 6 |** Overview of communities (numbers and average sizes) generated with the co-occurrent phenotype pairs for OMIM and Orphanet. Values shown for communities obtained using pairs of NMD phenotypes and equal numbers of randomly generated pairs of non-NMD phenotypes (average  $\pm$  SD).

Community	OMIM		Orphanet	
	NMD	non-NMD <sup>a</sup>	NMD	non-NMD <sup>a</sup>
Total number	231	94.98 $\pm$ 9.6	150	23.85 $\pm$ 4.293
Phenotypes per community	16.75	5.55 $\pm$ 0.6	13.63	4.15 $\pm$ 0.278
Functionally coherent clusters	40	41.21 $\pm$ 7.72	72	14.51 $\pm$ 3.189

<sup>a</sup>'non-NMD' results correspond to the grouping obtained using the 100 random sets described in **Table 3**.

### 3.5 Clusters Formed Between NMD Phenotypes are Functionally Coherent

The pairs of significantly co-occurring NMD phenotypes (pale magenta rectangle of Module B in **Figure 1**) were processed to extract communities of highly interconnected phenotypes (Module D in **Figure 1**). As summarised in **Table 6**, the number of communities and average number of phenotypes per community is higher using the modules based on NMD phenotypes compared to non-NMDs. This indicates that the co-occurring NMD phenotypes tend to form groups of related phenotypes.

The phenotype communities were combined with the phenotype-function associations obtained previously (light cyan rectangle of Module C in **Figure 1**) to evaluate their functional coherence. In total, 40 OMIM communities and 72 Orphanet communities (**Table 6**) showed coherent function, defined as having shared functional enrichment for at least 70% of their constituent phenotypes; they constituted the

"functionally coherent clusters". The complete list of functionally coherent clusters is provided at the end of the reports given in **Supplementary Material S1, S2**. In terms of cluster properties, some have high interconnection (\*193, \*138, \*219, #115, #150), others show spoke-hub structure, with a central node connected to a number of other nodes which do not then connect to each other (clusters \*58, \*1, \*3, \*4, \*91, \*26, #144, #2, #3, #21, #60, #22), and others have near linear topology (\*146, \*85, #97). Compared to their respective random non-NMD models **Table 6**, OMIM presents a similar number of clusters, in stark contrast with Orphanet, for which 5 times more clusters were found. The coherence threshold must be decreased to 50% to find a substantially different number of coherent clusters in OMIM compared to the random datasets (**Supplementary Figure S1**). This difference may be due to the OMIM clusters being larger than for Orphanet (16.75 vs. 13.63 phenotypes, respectively) or related to the nature of the diseases considered NMDs in each dataset. These results support the approach presented here in which the information from the different databases is considered separately, with Orphanet arguably providing more confident results with respect to how well the typical NMD phenotypes in these clusters represent realistic and useful groupings of co-occurrent phenotypes that reflect co-morbidity with shared underlying cellular functions.

### 3.6 Clinical Application I: Functionally Coherent Clusters Help Direct Diagnosis

The functionally coherent clusters found by our approach (**Supplementary Material S1, S2**) were inspected by clinicians



**TABLE 7** | Known associations between NMDs, clusters and underlying cell function, sorted by topological and pathophysiological levels.

Topological or pathophysiological level	Group of NMD	Cluster ID: main cell function involved <sup>a</sup>	References
<i>Peripheral nervous system</i>			
2 <sup>nd</sup> motor neuron in the anterior horn of the spinal cord	Spinal muscular atrophy	*82: R-HSA-191859 snRNP Assembly *82: R-HSA-194441 Metabolism of non-coding RNA	Bäumer et al. (2009), Rossoll and Bassell (2009)
Peripheral nerve	Peripheral neuropathies	*3: GO:0042552 – myelination *23: GO:0008366 – axon ensheathment *67: GO:0042552 – myelination *200: GO:0008366 – axon ensheathment *47: GO:0008366 – axon ensheathment	Kamil et al. (2019), Zhou and Notterpek (2016)
Neuromuscular junction	Myasthenic syndromes	*97: GO:0007274 – neuromuscular synaptic transmission *166: GO:0007271 – synaptic transmission, cholinergic *166: GO:0007528 – neuromuscular junction development	Rodríguez Cruz et al. (2018), Souza et al. (2016)
Muscle	Congenital or developmental myopathies	*15: GO:0048747 – muscle fiber development *6: GO:0051146 – striated muscle cell differentiation *6: GO:0048747 – muscle fiber development *193: GO:0003012 – muscle system process	Sarnat (1994), Cassandrini et al. (2017) Meola and Cardani (2015)
	Myotonic syndromes/dystrophies		
	Muscular dystrophies (non-congenital)	#13: GO:0048747 – muscle fiber development #13: GO:0007517 – muscle organ development #112: GO:0007517 – muscle organ development #112: GO:0048747 – muscle fiber development #60: hsa05414 – dilated cardiomyopathy #60: hsa05410 – hypertrophic cardiomyopathy #60: R-HSA-390522 – Striated Muscle Contraction	Lovering et al. (2005)
<i>Mitochondrial target</i>			
Mitochondrial genome maintenance	Mitochondrial myopathy or disease	*205: GO:0000002 – mitochondrial genome maintenance	Pfeffer and Chinnery (2013)
Mitochondrial inheritance	Mitochondrial myopathy or disease	*48: GO:0006119 – oxidative phosphorylation *48: GO:0022900 – electron transport chain *48: GO:0045333 – cellular respiration	

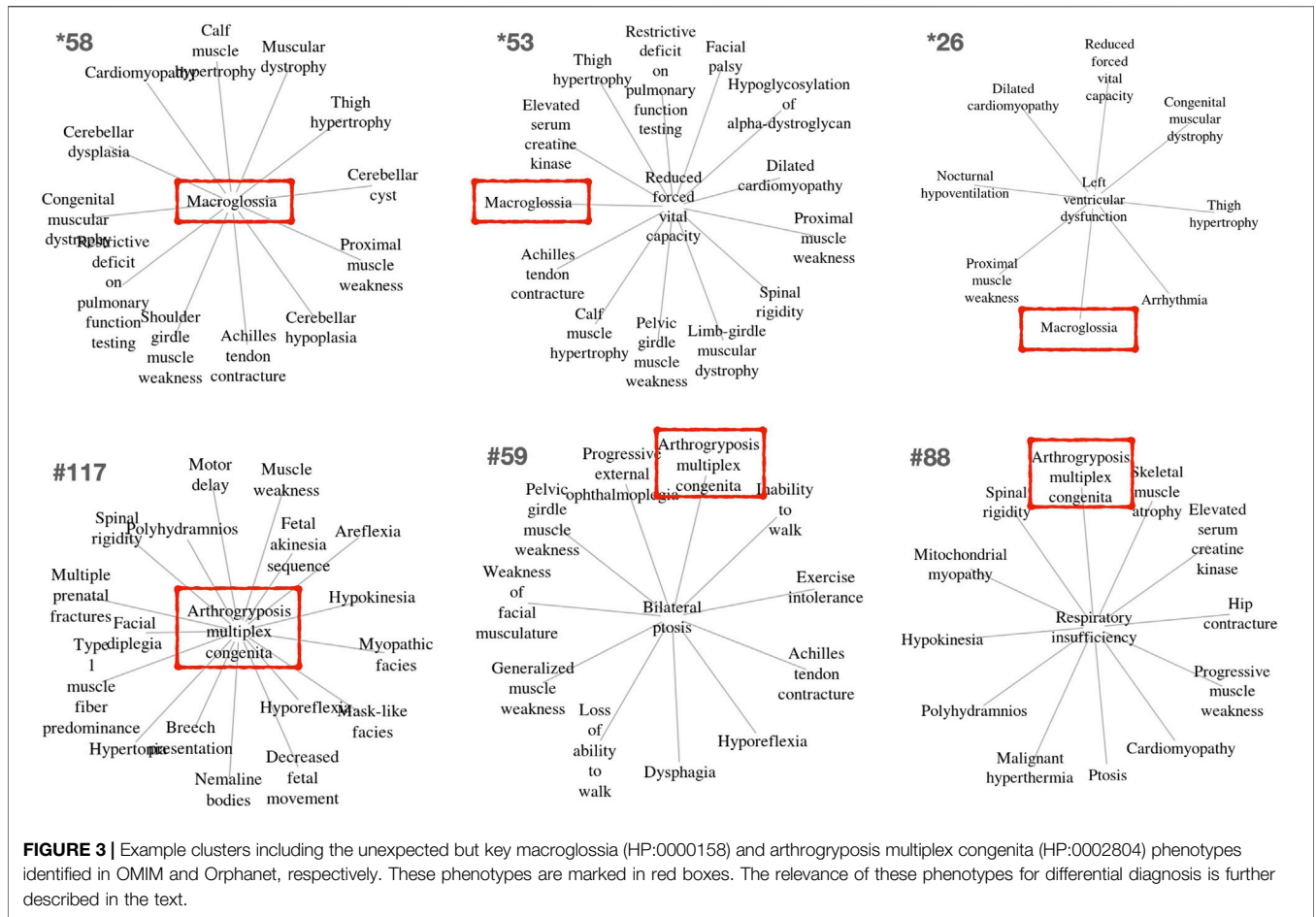
<sup>a</sup>\*\*\* tag for OMIM clusters. “#” tag for Orphanet clusters.

with expertise in paediatric neuromuscular diseases, finding clusters with sets of expected phenotypes for NMDs, and others with unexpected phenotypes. Clusters are grouped in **Table 7** based on the area in which the constituent phenotypes manifest, followed by further subdivision by group of NMD. Since they represent direct links between NMDs, their symptoms and their underlying mechanisms, they could be used for clinical diagnosis, by suggesting novel phenotypes to test for, should a patient present other phenotypes within a cluster. They also suggest potentially affected genes, useful for directing genetic analysis, and potentially affected functions. Although this analysis necessarily has a subjective aspect, given that clinicians differ in training and experience, it is remarkable that *PhenoClusters* produces functionally coherent clusters easily recognised by clinicians. In summary, there are multiple clusters that fit in with our current knowledge about NMDs as shown in **Table 7**,

which reinforces once again the coherence of the results generated by the presented approach.

### 3.7 Clinical Application II: Unexpected Phenotypic Relationships in NMDs

Other clusters provided by *PhenoClusters* contained more unusual groupings of phenotypes. Clinicians were surprised by the presence of ‘macroglossia’ (HP:0000158) and ‘arthrogryposis’ (HP:0002804) (**Figure 3**). Macroglossia is usually associated with non-neuromuscular syndromes and in metabolic disorders such as glycogen storage disease type II (Pompe disease), mucopolysaccharidosis, oligosaccharidosis, mucopolipidoses, sphingolipidoses and galactosidosis. However, macroglossia appeared as a hub node in the cluster \*58 (**Figure 3**), strongly associated with highly representative NMD phenotypes, such as ‘congenital muscular dystrophy’ and ‘achilles tendon contracture’.



This phenotype also appears in clusters \*26, and \*53, with macroglossia occupying a peripheral position (Figure 3). In macroglossia-containing clusters, 70%–80% of phenotypes show significant enrichment for the “protein O-linked mannosylation” biological process (GO:0035269), suggesting that the clinical observation of macroglossia in patients with NMDs may prompt practitioners to ask for O-linked mannosylation tests. This contrasts with cluster \*211 (Supplementary Material S1) that also contains congenital muscular dystrophy and ankle flexion contracture, as well as ‘increased endomysial connective tissue’, but not macroglossia, the functions for this cluster being ‘focal adhesion’ (hsa04510) and ‘ECM proteoglycans’ (R-HSA-3000178) but not “protein O-linked mannosylation”. Hence, cluster \*211 seems to be related to processes involved in physically connecting (‘focal adhesion’) cells to the extracellular matrix (‘ECM proteoglycans’). Therefore, functionally related clusters may help to discriminate disorders presenting with the macroglossia phenotype from other muscular dystrophies caused by the disruption of other cell functions. This finding shows the potential clinical utility of our approach to identify phenotypes whose presence alongside a given set of other phenotypes can indicate distinct underlying processes, with implications for diagnosis and treatment.

‘Arthrogryposis’ represents another phenotype of potential use for NMD diagnosis (Figure 3). Arthrogryposis multiplex

congenita consists of contractures in multiple body areas (Ambegaonkar et al., 2011). Cluster #117 shows a range of clinical phenotypes with arthrogryposis as the central node (Figure 3), while #59 and #88 are cases where this phenotype appears in a peripheral position. Interestingly, cluster #115 (Supplementary Material S2) presents almost all phenotypes observed in cluster #117, except arthrogryposis. Hence, it can be inferred that when arthrogryposis is present, the pathways associated with the majority of phenotypes in the cluster are more specific to muscle fibre-related processes including contraction dysfunction. However, when arthrogryposis is absent, the cluster phenotypes share additional pathways related to cardio-myopathies and the regulation of pseudopodium assembly, indicating a broader aetiology in the genetic systems involved. As such, the presence or absence of arthrogryposis alongside other symptoms may help inform NMD diagnosis and treatment, by indicating distinct underlying processes.

### 3.8 Research and Clinical Application: Clarifying Gene Involvement

The genes underlying the enriched functions were further investigated. It was noted that some of them occur in multiple

clusters but are linked to different functions depending on the affected phenotypes. For example, *DAG1* is a pivotal component of the dystrophin-glycoprotein complex and its dysfunction is related to many muscular dystrophies, amongst other diseases (Durbeej et al., 1998; Sciandra et al., 2003). It can be considered a pleiotropic gene due to its role in glycosylation, a process which is involved in myriad cellular functions. For NMDs, it seems that glycosylation pathways are affected when *DAG1* appears in clusters \*1, #29 and #113, in agreement with previous studies (Barresi and Campbell, 2006). However, the involvement of *DAG1* in clusters \*198, #2, #144, #112, #103, #118, #11, #89 and #63 (**Supplementary Material S1, S2**), appears to be related to cardiomyopathy (Michele et al., 2009), while it is more focused on 'ECM proteoglycan components' in cluster \*211. Hence, depending on the cluster in which *DAG1* appears, it seems to be involved in different pathways and functions, showing how *PhenoClusters* facilitates the identification of genes with multiple putative roles in NMDs depending on context.

Other clusters suggest that members of the solute carrier (SLC) family can perform many roles within NMDs. SLCs mediate the transport of a wide range of essential nutrients and metabolites, performing many different functions in cells and tissues (Zhang et al., 2019). SLC proteins were found to be associated with nutrient supply in cluster #93; metabolic transformation in clusters \*85, \*128, and \*218; energy homeostasis in cluster \*138; oxidative stress in cluster \*205; and neurological regulation in clusters \*97 and \*166. These findings show how the phenotypic components of the functionally coherent clusters defined using *PhenoClusters* can point to the disruption of different cell functions in relation to NMDs involving genes from the SLC protein family.

## 4 DISCUSSION

### 4.1 PhenoClusters Generates Reproducible Results and can be Extended to Other Diseases

As the use of bioinformatics analysis becomes more routine within biomedicine, it is crucial that published studies are accompanied by workflows that allow the analysis to be recorded and reported in a reproducible manner that can be applied to similar datasets, if applicable. *PhenoClusters* is based on phenotype co-occurrence, incorporating randomisation using matched non-NMD phenotypes to generate control datasets and including literature co-mention for validation of the approach. It also uses known NMD-causing genes to provide functional support to the co-occurring phenotypes. It produces a HTML report that can be easily interpreted, following the proposed Reproducible Research System (Mesirov, 2010; Piccolo and Frampton, 2016). Although there are many articles in the bioinformatics and systems biology fields presenting modular software tools and packages, in many cases code and data are not provided, despite the importance of reproducible research in health sciences (Harris et al., 2018). As it is essential for the present study, the necessary code to conduct

the workflow as described in **Figure 1** has been provided in full. Since the analysis starts by downloading diseases from OMIM and Orphanet, it can be adapted to study phenotype/gene/function co-occurrences in other diseases. *PhenoClusters* is thus an example of the re-analysis of previously published data using a new approach that produces new results and confirms already known facts. As such, it has achieved the goals of reproducible research (Mesirov, 2010; Piccolo and Frampton, 2016) in the health sciences (Harris et al., 2018) and exploiting already published results (Kovalevskaya et al., 2016).

### 4.2 Network Analyses Should Rely on More Than One Database

*PhenoClusters* did not mix data from OMIM and Orphanet based on two main facts. The first one is that OMIM and Orphanet are populated in different ways. For example, the variability of the phenotypic series of OMIM was high, with PS253600 (muscular dystrophy, limb-girdle, autosomal recessive) presenting 28 entries, while PS118220 (Charcot-Marie-Tooth disease) has 70 entries. The second one is that OMIM was searched with somewhat arbitrary keywords, whilst Orphanet NMDs were obtained based on its ontology structure. As a result, there are more OMIM-specific phenotypes and less Orphanet-specific phenotypes (**Figure 2**), and fewer functionally coherent clusters in OMIM than in Orphanet (**Table 6**). Besides these discrepancies, both databases rendered a similar ratio of shared phenotype pairs (**Table 3**) and the same top specific phenotype for NMDs, 'elevated serum creatine kinase' (**Supplementary Material S5**), in agreement with the importance of the creatine kinase test often used in NMD diagnosis (Thavorntanaburt et al., 2018). Other common phenotypes include muscle weakness, contractures, altered gait, functional difficulties, and respiratory issues, in agreement with recognised phenotypes for these diseases (Norwood et al., 2009). However, the interesting phenotypes 'macroglossia' and 'arthrogryposis' (**Figure 3**) emerged from database-specific group of phenotypes. In any case, despite the mentioned discrepancies, it is striking that, when the phenotypes were associated with genes and functions, there was high overlap (**Table 4, Supplementary Material S6**). Therefore, we believe it is preferable to use the information from OMIM and Orphanet separately and to compare to contrast the results.

Co-mention validation in **Table 5** revealed that even though GO contains many terms to describe biological functions and Reactome contains more pathway descriptions than KEGG, 2-fold more phenotype-KEGG pathway pairs (percentage-wise) were validated by literature co-mention. Moreover, when comparing functional annotation between OMIM and Orphanet, a proportionally larger overlap was found between the two resources for KEGG pathways (**Supplementary Material S6**). These findings supports our previous finding that KEGG produces better bioinformatic models in genetic algorithms for clinical diagnosis and prognosis (Luque-Baena et al., 2014), as well as for the outcome of diseases (Urda et al., 2018). Hence,

KEGG pathways should always be considered in functional analyses.

### 4.3 Co-Occurrences and Associations Appear Consistent

Due to the etiological heterogeneity of NMDs, each disease is defined by its own set of clinical phenotypes. Phenotypes serve to understand life and disease, but it is not always easy to translate them to molecular mechanisms and vice versa (Yu et al., 2016). Taking advantage of the huge amount of data available in databases, bipartite networks (phenotype-disease, disease-gene) have been used in biomedicine to model factors that influence human diseases, explore their molecular complexity, reveal novel molecular relationships and disease susceptibility genes, uncover the biological significance of disease-associated mutations (Pavlopoulos et al., 2018) and, more recently, discover phenotype/disease clusters that can predict protein-phenotype associations and reveal the underlying mechanisms that link them (Ren et al., 2020). Bipartite gene-disease network analysis was revolutionised with the diseaseome and related studies (Goh et al., 2007) which showed that genes associated with similar disease phenotypes have a higher propensity to interact physically with each other, forming distinct disease-specific functional modules, and that disorders tend to form clusters on the basis of similar pathophysiology (Oti and Brunner, 2007). Other studies revealed that phenotypes and gene co-regulation accurately predict unknown disease-gene relations (Deelen et al., 2019) since genes causing a specific disease or disease symptom often have similar molecular functions or are involved in the same biological process or pathway.

With all these facts in mind, the systems biology approach implemented in *PhenoClusters* lies in bipartite and tripartite networks (Modules B and C in **Figure 1**) to extract co-occurring phenotypes in NMDs (**Supplementary Material S3, S4**) that were consistent with i) clinical histories (Norwood et al., 2009) (**Table 7**), ii) creatine kinase tests (Moghadam-Kia et al., 2016; Thavorntanaburt et al., 2018) (**Supplementary Material S5**), iii) proximal or distal muscular weakness (McDonald, 2012; Mukherjee et al., 2019), and iv) postural instability and joint contracture (Kornegay et al., 2012), as usually found in the clinical history of NMD patients. The genes significantly associated with NMDs by *PhenoClusters* were largely expressed in neuronal and/or muscle-related tissues. For those that were not, it is tempting to speculate that this is due to incomplete information in The Human Protein Atlas, however it may also indicate genes involved in development or other regulatory processes that lead to the manifestation of the NMD phenotypes. There is currently much interest in the relationship between tissue/cell-type specificity and disease progression (Hekselman and Yeager-Lotem, 2020). Functional enrichment of the NMD-associated genes was used to provide phenotype-function associations that were then combined with the co-occurrent phenotype communities to produce functionally coherent clusters. Gene-to-function translation reduced the degrees of freedom of the analysis (several genes were required

to assign a significantly enriched function to a phenotype), increasing statistical power and allowing the detection of shared functions between phenotypes, even if the underlying genes differ. The comparison of NMD-specific associations with randomised data in **Tables 3, 4, 6** demonstrates that the findings were well-founded, as discussed below.

### 4.4 Co-Mention in Abstracts is a Valuable Approach Validation

There is an absence of gold-standard datasets for phenotypic relationships, particularly for rare diseases. As such, Ren et al. (2020) applied systems biology methods to obtain phenotype/disease clusters to seed machine learning models to predict protein-phenotype associations without any objective validation, assuming that they were relevant since authors can find biological support for some of the clusters. Here, abstract co-mention was proposed as a source of external validation for phenotype-phenotype and phenotype-function pairing approach. There were far fewer co-mentions for randomly paired terms than for specific pairs (**Table 5**), even though only a small fraction of phenotype-phenotype and phenotype-function pairs (< 10 % in the second) showed significant co-mention in the literature. This is likely due to a combination of the following reasons: 1) term relationship existed, although not yet described in literature; 2) the co-mention appeared in the main body, but not in abstracts, of articles; 3) phenotypes were co-mentioned with genes rather than functions; and 4) abstract co-mention existed, but using slightly different terms or natural language synonyms that escape the search. In any case, these issues also apply to the randomised/non-NMD associations. In conclusion, although the abstract co-mention is not perfect, the fact that the amount of significant literature co-mention was much higher for NMD-specific pairs than controls (**Table 5**) shows it to provide suitable validation of the described approach and that real relationships were found.

### 4.5 Functionally Related Clusters Provide Insight Into NMDs

Many *PhenoClusters* clusters are in clear agreement with current medical knowledge about the relationships between phenotype comorbidity, NMDs, their pathophysiology and genetic causes (**Table 7**). This serves to further support the bioinformatic approach presented here. Regarding differential diagnosis and treatment, the clusters for which the presence/absence of a specific phenotype (**Figure 3**) could change the nature of the associated functions are particularly interesting: these have potential applications for differential diagnosis and treatment selection. As such, ‘macroglossia’ can be used to identify *O*-mannosylation alterations in a given NMD dysfunction (Spence et al., 2002; Martin, 2007; Moore and Winder, 2010; Dobson et al., 2013; Goody et al., 2015), while ‘arthrogryposis’ (Ambegaonkar et al., 2011) may facilitate NMD diagnosis in other situations, as its presence may indicate that the disease is more specific to muscle fibres and contraction dysfunction.



The advantages of using functions rather than genes only is particularly evident in the findings related to pleiotropic genes and gene families involved in multiple functions. *DAG1* is involved in many diseases because it participates in glycosylation, whose malfunction affects many cellular processes and can produce several NMDs (Durbeej et al., 1998; Sciandra et al., 2003). However, its specific involvement in the proteoglycan components of the ECM seems to be related to cardiomyopathy. Additionally, despite the clear relevance to health and disease of the 190 different SLCs found mutated in human diseases, it is considered an understudied family (César-Razquin et al., 2015). Functionally related clusters from *PhenoClusters* reveal which particular role can be putatively assigned to the present members of the SLC family. Taken together, these results encourage future research lines directed to systematically exploit co-occurrence and clusters of co-occurrent phenotypes and functions in other diseases, as well as further study of a number of presented clusters to obtain more details on the functional implications.

#### 4.6 Clinical Implications of Clusters

It has been shown that functionally coherent clusters have potential utility in NMDs in terms of better understanding clinical presentation in these diseases and obtaining clues as to the underlying molecular genetic mechanisms. The former has important implications for diagnosis and patient classification; the latter will help researchers better understand these diseases and search for potential new therapeutic targets.

Their potential for diagnosis is clear: given a patient that presents a number of phenotypes corresponding to one of the clusters, one can make inferences about other clinical phenotypes that the patient may also suffer and that should be tested for. This will help obtain a better clinical profile for the patient, aiding diagnosis. Moreover, genes and functional terms associated with the same cluster can be of use for guiding genetic diagnosis and indicate further pathological examination. For example, multiple clusters in the peripheral neuropathy group in **Table 7** contain phenotypes associated with myelination and axon ensheathment related genes (Zhou and Notterpek, 2016; Kamil et al., 2019). As such, given a patient showing phenotypes belonging to one of these clusters, but for whom full diagnosis has not yet been achieved, the clinician could refer the patient for the appropriate assays, such as MRI, to look for demyelination and guide genetic studies to focus on related genes.

A further clinical application is related to differential diagnosis based on the phenotypic profile presented by a patient. The presence/absence of phenotypes ‘macroglossia’ and ‘arthrogryposis’ (**Figure 3**) has been thoroughly explained as examples of phenotypes indicating differing underlying mechanisms that lead to disease manifestation. However, many other examples undoubtedly exist among the clusters presented in **Supplementary Material S1, S2**.

Nevertheless, there are several limitations in our bioinformatic approach, largely related to the external databases from which the information is acquired. OMIM and Orphanet are both incomplete resources: there are undoubtedly multiple diseases suffered by people around the world whose description and genetic basis have not made their way into these databases. This means that we are likely to miss co-occurrent phenotypes

as the diseases in which they co-occur have not yet been added. Similarly, the resources used for functional enrichment analysis, used to ascribe function to the NMD phenotypes in this study, are also incomplete. We do not yet know the function of all genes in the genome, and this has effects on the enrichment procedures. As these resources improve, the potential of our workflow to find clusters of co-occurrent phenotypes and fully characterise their underlying basis will no doubt improve with them. However, regarding the information of clusters, this approach can only extract functions from genes known to be involved in NMDs, and cannot hypothesise about other genes even if they are working in the same pathway. And as already mentioned above, there is a limitation concerning literature co-mention validation (**Table 5**), as a higher proportion of confirmed pairs it would be desirable.

In conclusion, our approach and the resultant functionally coherent clusters of NMD phenotypes (**Supplementary Material S1, S2**) can 1) relate phenotype co-occurrence across NMDs to the underlying genes and mechanisms involved in the NMD (or other diseases); 2) provide clinicians with hints about clinical tests to produce a more reliable diagnosis based on the presence or absence of some specific phenotypes that have not yet been reported by the patient or the clinician, in the comorbidity context of the patient; 3) give researchers clues to perform new experiments to discover the underlying biological mechanisms of a disease; and 4) help in selecting optimal treatment. Therefore, *PhenoClusters* can be considered a new tool for more accurate diagnosis and an advance towards personalised medicine for NMDs.

#### DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

#### AUTHOR CONTRIBUTIONS

ED-S conducted the formal analysis and investigation, produces the resources and software, as well as validation and visualisations. MC structured the work, improved the reports and wrote the final manuscript. RY, YD, RC, JH, and FP interpreted the functionally coherent clusters and reviewed the manuscript. JR obtained the founding sources, administrates the project and conceptualised the analysis. JP performs the data curation, investigated the results, selected the methodology, produced the original draft and edited the English wording. JR and JP contributed equally to the research.

#### FUNDING

The study was funded by grants from the Andalusian Government (Junta de Andalucía) with European Regional Development Fund [UMA18-FEDERJA-102]; the Ramón Areces foundation, which funds project for the investigation of rare disease (National call for research on life and material sciences, XIX edition); the Spanish Ministry of Science and Innovation with European Regional

Development Fund PID2019-108096RB-C21; and the Ramón y Cajal I3 projects through the Research Plan of the University of Malaga (Ayudas del I Plan Propio).

University of Malaga for their provision of computational resources and technical support (<http://www.scbi.uma.es/>).

## ACKNOWLEDGMENTS

The CIBERER is an initiative from the Carlos III Health Institute (Instituto de Salud Carlos III; ISCIII). The authors thank the Supercomputing and Bioinnovation Center (SCBI) at the

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmolb.2021.635074/full#supplementary-material>.

## REFERENCES

- Ambegaonkar, G., Manzur, A. Y., Robb, S. A., Kinali, M., and Muntoni, F. (2011). The multiple phenotypes of arthrogryposis multiplex congenita with reference to the neurogenic variant. *Eur. J. Paediatr. Neurol.* 15, 316–319. doi:10.1016/j.ejpn.2011.01.007
- Amberger, J. S., Bocchini, C. A., Scott, A. F., and Hamosh, A. (2019). Omim.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res.* 47, D1038. doi:10.1093/nar/gky1151
- Bagley, S. C., Sirota, M., Chen, R., Butte, A. J., and Altman, R. B. (2016). Constraints on biological mechanism from disease comorbidity using electronic medical records and database of genetic variants. *PLoS Comput. Biol.* 12, e1004885. doi:10.1371/journal.pcbi.1004885
- Barresi, R., and Campbell, K. P. (2006). Dystroglycan: from biosynthesis to pathogenesis of human disease. *J. Cell Sci.* 119, 199–207. doi:10.1242/jcs.02814
- Bass, J. I. F., Diallo, A., Nelson, J., Soto, J. M., Myers, C. L., and Walhout, A. J. M. (2013). Using networks to measure similarity between genes: association index selection. *Nat. Methods* 10, 1169–1176. doi:10.1038/nmeth.2728
- Bäumer, D., Lee, S., Nicholson, G., Davies, J. L., Parkinson, N. J., Murray, L. M., et al. (2009). Alternative splicing events are a late feature of pathology in a mouse model of spinal muscular atrophy. *PLoS Genet.* 5, e1000773. doi:10.1371/journal.pgen.1000773
- Benarroch, L., Bonne, G., Rivier, F., and Hamroun, D. (2019). The 2020 version of the gene table of neuromuscular disorders (nuclear genome). *Neuromuscul. Disord.* 29, 980. doi:10.1016/j.nmd.2019.10.010
- Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* 57, 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x
- Birgmeier, J., Haeussler, M., Deisseroth, C. A., Steinberg, E. H., Jagadeesh, K. A., Ratner, A. J., et al. (2020). AMELIE speeds Mendelian diagnosis by matching patient phenotype and genotype to primary literature. *Sci. Transl. Med.* 12, eaau9113. doi:10.1126/scitranslmed.aau9113
- Brown, T. L., and Meloche, T. M. (2016). Exome sequencing a review of new strategies for rare genomic disease research. *Genomics* 108, 109–114. doi:10.1016/j.ygeno.2016.06.003
- Bueno, A., Rodríguez-López, R., Reyes-Palomares, A., Rojano, E., Corpas, M., Nevado, J., et al. (2018). Phenotype-loci associations in networks of patients with rare disorders: application to assist in the diagnosis of novel clinical cases. *Eur. J. Hum. Genet.* 26, 1451–1461. doi:10.1038/s41431-018-0139-x
- Cassandrini, D., Trovato, R., Rubegni, A., Lenzi, S., Fiorillo, C., Baldacci, J., et al. (2017). Congenital myopathies: clinical phenotypes and new diagnostic tools. *Ital. J. Pediatr.* 43, 101. doi:10.1186/s13052-017-0419-z
- César-Razquin, A., Snijder, B., Frappier-Brinton, T., Isserlin, R., Gyimesi, G., Bai, X., et al. (2015). A call for systematic research on solute carriers. *Cell* 162, 478–487. doi:10.1016/j.cell.2015.07.022
- Deelen, P., van Dam, S., Herkert, J. C., Karjalainen, J. M., Brugge, H., Abbott, K. M., et al. (2019). Improving the diagnostic yield of exome-sequencing by predicting gene-phenotype associations using large-scale gene expression analysis. *Nat. Commun.* 10, 2837. doi:10.1038/s41467-019-10649-4
- Díaz-Santiago, E., Jabato, F. M., Rojano, E., Seoane, P., Pazos, F., Perkins, J. R., et al. (2020). Phenotype-genotype comorbidity analysis of patients with rare disorders provides insight into their pathological and molecular bases. *PLoS Genet.* 16, e1009054. doi:10.1371/journal.pgen.1009054
- Dobson, C. M., Hempel, S. J., Stalnaker, S. H., Stuart, R., and Wells, L. (2013). O-mannosylation and human disease. *Cell. Mol. Life Sci.* 70, 2849–2857. doi:10.1007/s00018-012-1193-0
- Durbeej, M., Henry, M. D., and Campbell, K. P. (1998). Dystroglycan in development and disease. *Curr. Opin. Cell Biol.* 10, 594–601. doi:10.1016/s0955-0674(98)80034-3
- Firth, H. V., Richards, S. M., Bevan, A. P., Clayton, S., Corpas, M., Rajan, D., et al. (2009). Decipher: database of chromosomal imbalance and phenotype in humans using ensembl resources. *Am. J. Hum. Genet.* 84, 524–533. doi:10.1016/j.ajhg.2009.03.010
- Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabási, A.-L. (2007). The human disease network. *Proc. Natl. Acad. Sci.* 104, 8685–8690. doi:10.1073/pnas.0701361104
- Goody, M. F., Sher, R. B., and Henry, C. A. (2015). Hanging on for the ride: adhesion to the extracellular matrix mediates cellular responses in skeletal muscle morphogenesis and disease. *Dev. Biol.* 401, 75–91. doi:10.1016/j.ydbio.2015.01.002
- Hamosh, A. (2004). Online Mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* 33, D514. doi:10.1093/nar/gki033
- Harris, J. K., Johnson, K. J., Carothers, B. J., Combs, T. B., Luke, D. A., and Wang, X. (2018). Use of reproducible research practices in public health: a survey of public health analysts. *PLoS One* 13, e0202447. doi:10.1371/journal.pone.0202447
- Hekselman, I., and Yeager-Lotem, E. (2020). Mechanisms of tissue and cell-type specificity in heritable traits and diseases. *Nat. Rev. Genet.* 21, 137–150. doi:10.1038/s41576-019-0200-9
- Hidalgo, C. A., Blumm, N., Barabási, A.-L., and Christakis, N. A. (2009). A dynamic network approach for the study of human phenotypes. *PLoS Comput. Biol.* 5, e1000353. doi:10.1371/journal.pcbi.1000353
- Hoehndorf, R., Schofield, P. N., and Gkoutos, G. V. (2015). The role of ontologies in biological and biomedical research: a functional perspective. *Brief. Bioinform.* 16, 1069–1080. doi:10.1093/bib/bbv011
- Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., et al. (2015). Orchestrating high-throughput genomic analysis with bioconductor. *Nat. Methods* 12, 115–121. doi:10.1038/nmeth.3252
- Hwang, T., Atluri, G., Xie, M., Dey, S., Hong, C., Kumar, V., et al. (2012). Co-clustering phenome-genome for phenotype classification and disease gene discovery. *Nucleic Acids Res.* 40, e146. doi:10.1093/nar/gks615
- Jabato, F. M., Seoane, P., Perkins, J. R., Rojano, E., García Moreno, A., Chagoyen, M., et al. (2020). Systematic identification of genetic systems associated with phenotypes in patients with rare genomic copy number variations. *Hum. Genet.* 140, 457. doi:10.1007/s00439-020-02214-7
- Kalinka, A. T., and Tomancak, P. (2011). Linkcomm: an R package for the generation, visualization, and analysis of link communities in networks of arbitrary size and type. *Bioinformatics* 27, 2011–2012. doi:10.1093/bioinformatics/btr311
- Kamil, K., Yazid, M. D., Idrus, R. B. H., Das, S., and Kumar, J. (2019). Peripheral demyelinating diseases: from biology to translational medicine. *Front. Neurol.* 10, 87. doi:10.3389/fneur.2019.00087
- Köhler, S., Gargano, M., Matentzoglou, N., Carmody, L. C., Lewis-Smith, D., Vasilevsky, N. A., et al. (2021). The human phenotype ontology in 2021. *Nucleic Acids Res.* 49, D1207–D1217. doi:10.1093/nar/gkaa1043

- Köhler, S., Schulz, M. H., Krawitz, P., Bauer, S., Dölken, S., Ott, C. E., et al. (2009). Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am. J. Hum. Genet.* 85, 457. doi:10.1016/j.ajhg.2009.09.003
- Kornegay, J. N., Childers, M. K., Bogan, D. J., Bogan, J. R., Nghiem, P., Wang, J., et al. (2012). The paradox of muscle hypertrophy in muscular dystrophy. *Phys. Med. Rehabil. Clin. N. Am.* 23, 149–172. doi:10.1016/j.pmr.2011.11.014
- Kovalevskaya, N. V., Whicher, C., Richardson, T. D., Smith, C., Grajciarova, J., Cardama, X., et al. (2016). Dnadigest and repovisite: connecting the world of genomic data. *PLoS Biol.* 14, e1002418. doi:10.1371/journal.pbio.1002418
- Kraker, J., and Zivković, S. A. (2011). Autoimmune neuromuscular disorders. *Curr. Neuropharmacol.* 9, 400–408. doi:10.2174/157015911796558000
- Laing, N. G. (2012). Genetics of neuromuscular disorders. *Crit. Rev. Clin. Lab. Sci.* 49, 33–48. doi:10.3109/10408363.2012.658906
- Lovering, R. M., Porter, N. C., and Bloch, R. J. (2005). The muscular dystrophies: from genes to therapies. *Phys. Ther.* 85, 1372–1388. doi:10.1093/ptj/85.12.1372
- Luque-Baena, R. M., Urda, D., Gonzalo Claros, M., Franco, L., and Jerez, J. M. (2014). Robust gene signatures from microarray data using genetic algorithms enriched with biological pathway keywords. *J. Biomed. Inform.* 49, 32–44. doi:10.1016/j.jbi.2014.01.006
- Martin, P. (2007). Congenital muscular dystrophies involving the o-mannose pathway. *Curr. Mol. Med.* 7, 417–425. doi:10.2174/156652407780831601
- Mccormack, P., Woods, S., Aartsma-Rus, A., Hagger, L., Herczegfalvi, A., Heslop, E., et al. (2013). Guidance in social and ethical issues related to clinical, diagnostic care and novel therapies for hereditary neuromuscular rare diseases: “translating” the translational. *PLoS Curr.* 5, 1–17. doi:10.1371/currents.md.f90b49429fa814bd26c5b22b13d773ec
- McDonald, C. M. (2012). Clinical approach to the diagnostic evaluation of hereditary and acquired neuromuscular diseases. *Phys. Med. Rehabil. Clin. N. Am.* 23, 495–563. doi:10.1016/j.pmr.2012.06.011
- Meola, G., and Cardani, R. (2015). Myotonic dystrophies: an update on clinical aspects, genetic, pathology, and molecular pathomechanisms. *Biochim. Biophys. Acta Mol. Basis Dis.* 1852, 594–606. doi:10.1016/j.bbdis.2014.05.019
- Mesirov, J. P. (2010). Computer science. accessible reproducible research. *Science* 327, 415–416. doi:10.1126/science.1179653
- Michele, D. E., Kabava, Z., Davis, S. L., Weiss, R. M., and Campbell, K. P. (2009). Dystroglycan matrix receptor function in cardiac myocytes is important for limiting activity-induced myocardial damage. *Circ. Res.* 105, 984–993. doi:10.1161/CIRCRESAHA.109.199489
- Moghadam-Kia, S., Oddis, C. V., and Aggarwal, R. (2016). Approach to asymptomatic creatine kinase elevation. *Cleve. Clin. J. Med.* 83, 37–42. doi:10.3949/ccjm.83a.14120
- Moore, C. J., and Winder, S. J. (2010). Dystroglycan versatility in cell adhesion: a tale of multiple motifs. *Cell Commun. Signal.* 8, 3. doi:10.1186/1478-811X-8-3
- Mukherjee, A., Dhawan, S. R., and Saini, L. (2019). Muscle hypertrophy in a child with progressive weakness—thinking beyond muscular dystrophy. *J. Clin. Neuromuscul. Dis.* 20, 140–141. doi:10.1097/CND.0000000000000227
- Norwood, F. L. M., Harling, C., Chinnery, P. F., Eagle, M., Bushby, K., and Straub, V. (2009). Prevalence of genetic muscle disease in northern England: in-depth analysis of a muscle clinic population. *Brain* 132, 3175–3186. doi:10.1093/brain/awp236
- Oti, M., and Brunner, H. (2007). The modular nature of genetic diseases. *Clin. Genet.* 71, 1–11. doi:10.1111/j.1399-0004.2006.00708.x
- Pavlopoulos, G. A., Kontou, P. I., Pavlopoulou, A., Bouyioukos, C., Markou, E., and Bagos, P. G. (2018). Bipartite graphs in systems biology and medicine: a survey of methods and applications. *Gigascience* 7, 1–31. doi:10.1093/gigascience/gy014
- Pfeffer, G., and Chinnery, P. F. (2013). Diagnosis and treatment of mitochondrial myopathies. *Ann. Med.* 45, 4–16. doi:10.3109/07853890.2011.605389
- Piccolo, S. R., and Frampton, M. B. (2016). Tools and techniques for computational reproducibility. *Gigascience* 5, 30. doi:10.1186/s13742-016-0135-4
- Ren, X., Wang, S., and Huang, T. (2020). Decipher the connections between proteins and phenotypes. *Biochim. Biophys. Acta - Proteins Proteom.* 1868, 140503. doi:10.1016/j.bbapap.2020.140503
- Rodríguez Cruz, P., Palace, J., and Beeson, D. (2018). The neuromuscular junction and wide heterogeneity of congenital myasthenic syndromes. *Int. J. Mol. Sci.* 19, 1677. doi:10.3390/ijms19061677
- Rojano, E., Seoane, P., Bueno-Amoros, A., Perkins, J. R., and Garcia-Ranea, J. A. (2017). “Revealing the relationship between human genome regions and pathological phenotypes through network analysis,” in *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*, Cham, Switzerland, 197–207.
- Rossoll, W., and Bassell, G. J. (2009). Spinal muscular atrophy and a model for survival of motor neuron protein function in axonal ribonucleoprotein complexes. *Results Probl. Cell Differ.* 48, 289–326. doi:10.1007/400\_2009\_4
- Roy, A. J., Van den Bergh, P., Van Damme, P., Doggen, K., Van Casteren, V., and Van Casteren, V. (2015). Early stages of building a rare disease registry, methods and 2010 data from the Belgian neuromuscular disease registry (BNMDR). *Acta Neurol. Belg.* 115, 97–104. doi:10.1007/s13760-014-0320-0
- Rzhetsky, A., Wajngurt, D., Park, N., and Zheng, T. (2007). Probing genetic overlap among complex human phenotypes. *Proc. Natl. Acad. Sci.* 104, 11694–11699. doi:10.1073/pnas.0704820104
- Sarnat, H. B. (1994). New insights into the pathogenesis of congenital myopathies. *J. Child Neurol.* 9, 193–201. doi:10.1177/088307389400900218
- Sciandra, F., Bozzi, M., Bianchi, M., Pavoni, E., Giardina, B., and Brancaccio, A. (2003). Dystroglycan and muscular dystrophies related to the dystrophin-glycoprotein complex. *Ann. Ist. Super. Sanita* 39, 173–181.
- Scoto, M., Finkel, R., Mercuri, E., and Muntoni, F. (2018). Genetic therapies for inherited neuromuscular disorders. *Lancet Child Adolesc. Health* 2, 600–609. doi:10.1016/S2352-4642(18)30140-8
- Scotton, C., Passarelli, C., Neri, M., and Ferlini, A. (2014). Biomarkers in rare neuromuscular diseases. *Exp. Cell Res.* 325, 44–49. doi:10.1016/j.yexcr.2013.12.020
- Seoane, P., Ocaña, S., Carmona, R., Bautista, R., Madrid, E., Torres, A. M., et al. (2016). AutoFlow, a versatile workflow engine illustrated by assembling an optimised *de novo* transcriptome for a non-model species, such as faba bean (*Vicia faba*). *Curr. Bioinform.* 11, 1–11. doi:10.2174/1574893611666160212235117
- Singleton, M. V., Guthery, S. L., Voelkerding, K. V., Chen, K., Kennedy, B., Margraf, R. L., et al. (2014). Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. *Am. J. Hum. Genet.* 94, 599–610. doi:10.1016/j.ajhg.2014.03.010
- Sirota, M., Schaub, M. A., Batzoglou, S., Robinson, W. H., and Butte, A. J. (2009). Autoimmune disease classification by inverse association with SNP alleles. *PLoS Genet.* 5, e1000792. doi:10.1371/journal.pgen.1000792
- Souza, P. V. S. d., Batistella, G. N. d. R., Lino, V. C., Pinto, W. B. V. d. R., Annes, M., and Oliveira, A. S. B. (2016). Clinical and genetic basis of congenital myasthenic syndromes. *Arq. Neuro-Psiquiatr.* 74, 750–760. doi:10.1590/0004-282X20160106
- Spence, H. J., Chen, Y.-J., and Winder, S. J. (2002). Muscular dystrophies, the cytoskeleton and cell adhesion. *Bioessays* 24, 542–552. doi:10.1002/bies.10098
- Spuler, S., Stroux, A., Kuschel, F., Kuhlmei, A., and Kendel, F. (2011). Delay in diagnosis of muscle disorders depends on the subspecialty of the initially consulted physician. *BMC Health Serv. Res.* 11, 91. doi:10.1186/1472-6963-11-91
- Tan, N. B., Stapleton, R., Stark, Z., Delatycki, M. B., Yeung, A., Hunter, M. F., et al. (2020). Evaluating systematic reanalysis of clinical genomic data in rare disease from single center experience and literature review. *Mol. Genet. Genomic Med.* 8, e1508. doi:10.1002/mgg3.1508
- Thavornthanaburt, S., Tanboon, J., Likasitwattanakul, S., Sangruchi, T., Nishino, I., Ngerancham, M., et al. (2018). Impact of muscle biopsy on diagnosis and management of children with neuromuscular diseases: a 10-year retrospective critical review. *J. Pediatr. Surg.* 53, 489–492. doi:10.1016/j.jpedsurg.2017.06.006
- Turakhia, P., Barrick, B., and Berman, J. (2013). Patients with neuromuscular disorder. *Med. Clin. N. Am.* 97, 1015–1032. doi:10.1016/j.mcna.2013.05.005
- Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., et al. (2015). Tissue-based map of the human proteome. *Science* 347, 1260419. doi:10.1126/science.1260419
- Urda, D., Aragón, F., Bautista, R., Franco, L., Veredas, F. J., Claros, M. G., et al. (2018). Blasso: integration of biological knowledge into a regularized linear model. *BMC Syst. Biol.* 12, 94. doi:10.1186/s12918-018-0612-8
- Vasant, D., Chanas, L., Malone, J., Hanaeur, M., Olry, A., Jupp, S., et al. (2014). Ordo: an ontology connecting rare disease, epidemiology and genetic data. *Phenotype data at ISMB2014*, 1–4. Available at [https://www.researchgate.net/profile/Drashtti-Vasant/publication/281824026\\_ORDO\\_An\\_Ontology\\_Connecting\\_Rare\\_Disease\\_Epidemiology\\_and\\_Genetic\\_Data/links/55f99bc408](https://www.researchgate.net/profile/Drashtti-Vasant/publication/281824026_ORDO_An_Ontology_Connecting_Rare_Disease_Epidemiology_and_Genetic_Data/links/55f99bc408)

- aeafc8ac266edf/ORDO-An-Ontology-Connecting-Rare-Disease-Epidemiology-and-Genetic-Data.pdf
- Yang, H., Robinson, P. N., and Wang, K. (2015). Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nat. Methods* 12, 841–843. doi:10.1038/nmeth.3484
- Yang, P., Li, X., Wu, M., Kwok, C.-K., and Ng, S.-K. (2011). Inferring gene-phenotype associations via global protein complex network propagation. *PLoS One* 6, e21502. doi:10.1371/journal.pone.0021502
- Yu, G., and He, Q.-Y. (2016). ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol. Biosyst.* 12, 477–479. doi:10.1039/c5mb00663e
- Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). ClusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS: A J. Integr. Biol.* 16, 284–287. doi:10.1089/omi.2011.0118
- Yu, M. K., Kramer, M., Dutkowski, J., Srivas, R., Licon, K., Kreisberg, J. F., et al. (2016). Translation of genotype to phenotype by a hierarchy of cell subsystems. *Cell Syst.* 2, 77–88. doi:10.1016/j.cels.2016.02.003
- Zhang, Y., Zhang, Y., Sun, K., Meng, Z., and Chen, L. (2019). The slc transporter in nutrient and metabolic sensing, regulation, and drug development. *J. Mol. Cell Biol.* 11, 1–13. doi:10.1093/jmcb/mjy052
- Zhao, M., Havrilla, J. M., Fang, L., Chen, Y., Peng, J., Liu, C., et al. (2020). Phen2Gene: rapid phenotype-driven gene prioritization for rare diseases. *NAR Genomics and Bioinformatics* 2, lqaa032. doi:10.1093/nargab/lqaa032
- Zhou, X., Menche, J., Barabási, A.-L., and Sharma, A. (2014). Human symptoms-disease network. *Nat. Commun.* 5, 4212. doi:10.1038/ncomms5212
- Zhou, Y., and Notterpek, L. (2016). Promoting peripheral myelin repair. *Exp. Neurol.* 283, 573–580. doi:10.1016/j.expneurol.2016.04.007

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Díaz-Santiago, Claros, Yahyaoui, de Diego-Otero, Calvo, Hoenicke, Palau, Ranea and Perkins. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.