Human Molecular Genetics, 2026, 35(22), ddag115
https://doi.org/10.1093/hmg/ddag115
Published: 31 October 2026
Original Article

OXFORD

Identifying fundamental properties of neuromuscular disorder genes through genomic feature analysis

Alexandria Martin[1,2,3,4], M. Alejandra Llanes-Cuesta[1,3], Jessica N. Hartley[2,4], Patrick Frosk[2,4,5], Britt I. Drögemöller[2,4], Galen E.B. Wright[1,2,3,4,*]

[1]Department of Pharmacology and Therapeutics, Rady Faculty of Health Sciences, University of Manitoba, 753 McDermot Avenue, Winnipeg, Manitoba R3E 0T6, Canada
[2]Department of Biochemistry and Medical Genetics, Rady Faculty of Health Sciences, University of Manitoba, 745 Bannatyne Avenue, Winnipeg, Manitoba R3E 0J9, Canada
[3]PrairieNeuro Research Centre, Kleysen Institute for Advanced Medicine, Health Sciences Centre and Rady Faculty of Health Sciences, University of Manitoba, 710 William Avenue, Winnipeg, Manitoba R3E 0Z3, Canada
[4]Children's Hospital Research Institute of Manitoba, 15 McDermot Avenue, Winnipeg, Manitoba R3E 3P4, Canada
[5]Department of Pediatrics and Child Health, Rady Faculty of Health Sciences, University of Manitoba, 840 Sherbrook Street, Winnipeg, Manitoba R3A 1S1, Canada

*Corresponding author. Department of Pharmacology and Therapeutics, University of Manitoba, Winnipeg, MB R3E 0T6, Canada. E-mail: galen.wright@umanitoba.ca

Abstract

Neuromuscular disorders (NMDs) encompass conditions that affect the peripheral nervous system. They are often caused by genetic alterations that impair skeletal muscle function. Molecular diagnosis remains challenging because some causal genes have not yet been identified. We therefore used feature-selection approaches to characterize genomic properties that distinguish NMD genes from other protein-coding genes. Curated genes implicated in NMDs ($n = 639$; GeneTable of NMDs) were merged with 134 genomic features for autosomal protein-coding genes spanning gene complexity, genetic variation, expression patterns, and other gene traits. Machine-learning-based feature selection using Boruta was combined with statistical analyses to compare NMD genes with other protein-coding genes. We also evaluated NMD genes against high-confidence non-NMD disease genes, examined variation across NMD subgroups, and assessed genes added to the NMD catalogue between 2020 and 2024. Compared with other protein-coding genes, NMD genes showed higher expression in disease-relevant tissues, including skeletal muscle and heart; greater transcript complexity and population genetic constraint, and stronger conservation across several evolutionary measures. Of 33 features that distinguished NMD genes from both high-confidence non-NMD disease genes and other autosomal genes, 28 were expression measures. Feature profiles varied across NMD subgroups, while genes added between 2020 and 2024 showed profiles similar to established NMD genes across all Boruta-confirmed features. This study identified genomic features that distinguish NMD genes from high-confidence disease genes and broader autosomal protein-coding genes. These findings provide insight into the gene architecture, expression, constraint, and evolutionary conservation of NMD genes and may inform candidate-gene prioritization in genetically unresolved NMDs.

Keywords **Bioinformatics, disease gene prioritisation, genomic feature analysis, genomics, neurogenetics, neuromuscular disorders, rare disease**

Introduction

Neuromuscular disorders (NMDs) are a diverse group of conditions that affect the peripheral nervous system, including motor neurons, neuromuscular junctions, and skeletal muscle [1]. Globally, an estimated 15 million people are affected by NMDs, corresponding to a prevalence of approximately one in 1000 individuals, though this rate varies by region [2–4]. The majority of NMDs have a genetic etiology, and advances in sequencing technologies have led to the identification of over 600 genes that are known to cause NMDs [5, 6].

Despite this progress, the diagnostic yield for these conditions remains below 60% [7, 8], and the continued expansion of the NMD disease-gene catalogue indicates that additional causal genes remain to be identified. Many individuals remain undiagnosed due to limitations of available gene panels and exome sequencing, the complexity of variant interpretation, and the presence of causal variants in genes and genomic regions not yet linked to NMDs [7, 9]. As a result, affected individuals often face a prolonged 'diagnostic odyssey,' which can delay access to targeted interventions, clinical support resources, and genetic counselling [10, 11]. Diagnostic

Received: April 22, 2026. **Revised:** September 29, 2026. **Accepted:** October 20, 2026
© The Author(s) 2026. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

strategies that span larger genomic regions, such as whole-genome sequencing, are increasingly being applied for unresolved cases [12]. These approaches employ computational filtering techniques to prioritize candidate variants, but are limited by the large number of detected variants and the difficulty of distinguishing pathogenic variants from benign background variation [13]. Since the human genome contains nearly 20 000 protein-coding genes, systematic approaches to identify previously unknown causal NMD genes are crucial for informing genetic diagnoses and for understanding disease mechanisms and neurobiology [14].

An approach for improving disease gene discovery is to analyze annotated genomic features to identify shared biological properties that distinguish disease-associated genes from the broader genome [15–18]. In recent years, multi-omic annotations have become increasingly available due to technological and methodological advances, enabling comprehensive exploration of gene properties [9, 16]. Thus, using these features can aid in prioritizing novel candidate disease genes and enhance the interpretation of gene-phenotype relationships [18]. This strategy has become more feasible with the widespread availability of public resources, such as the GeneTable of NMDs, which catalogues the growing number of genes robustly linked to these disorders [5].

Similar computational approaches have been applied to more restricted disease groups, highlighting the need to extend these analyses to broader disease contexts. For example, an investigation used multi-omic annotations to characterize genes implicated in hereditary ataxia, a subtype of NMDs [9]. Using computational approaches, the study revealed shared patterns across childhood- and adult-onset ataxia genes, including underappreciated contributors such as short tandem repeat (STR) density. However, these methods have not been applied across the larger set of NMD genes, creating an important knowledge gap.

Machine-learning-based feature selection provides one means of integrating these functional genomic annotations. Boruta is a random forest-based feature-selection algorithm. A random forest combines many decision trees, each built from a different sample of the data, to identify variables that consistently help distinguish groups. Boruta then compares each real variable with randomized *'shadow'* versions to determine whether it is more informative than expected by chance [19]. Boruta has previously been used in various human genetic studies, including identifying discriminative features associated with gain- and loss-of-function pathogenic variants, as well as prioritizing relevant genes from genetic variant data in Alzheimer's disease [20, 21].

We therefore hypothesized that applying a genome-wide feature selection approach across all NMD subtypes could reveal shared biological features and provide insight into their genetic architecture. To investigate this, we compiled an extensive, curated list of genomic annotations from several public datasets and integrated them with a list of known NMD genes. We then applied Boruta feature selection along with complementary analytical approaches to identify the most important features for distinguishing NMD genes from other protein-coding genes (Fig. 1). We further compared NMD genes with other high-confidence disease genes, examined variation across NMD subgroups, and assessed whether the identified feature patterns were also present among genes added to the NMD catalogue more recently. This analysis aimed to gain insight into the biological characteristics of NMD genes and establish a basis for testing the utility of these features in future candidate-gene prioritization studies.

Results

Genomic features distinguish NMD genes from other autosomal protein-coding genes

We evaluated the discriminative potential of genomic features to distinguish NMD genes ($n = 639$) from other protein-coding genes ($n = 17\,880$) using the Boruta machine-learning-based feature selection algorithm and complementary statistical techniques. Of the 134 features across four categories, Boruta classified 107 features (79.9%) as *'confirmed,'* indicating that they were important for distinguishing between the two gene lists (Supplementary Fig. S1). This result indicates that differences between NMD genes and other autosomal protein-coding genes extend across multiple genomic feature classes.

To gain further insight into the features most informative for distinguishing NMD genes, we focused on the top 10 confirmed features ranked by median Boruta importance (Fig. 2A). Boruta determines feature relevance by comparing observed importance scores with those of randomized shadow features, which were used in the analysis but are not shown in Fig. 2. Skeletal muscle gene expression [quantified using the Genotype-Tissue Expression (GTEx) *t*-statistic] ranked as the most important feature, followed by gene expression in the heart left ventricle. Higher GTEx t-statistics indicate relatively greater expression in the indicated tissue. Other highly ranked features included heart *Tau* (tissue-specificity score), transcript count, evolutionary measures based on omega values for the rhesus macaque (*Macaca mulatta*) and marmoset (*Callithrix jacchus*), expression in the heart atrial appendage, mouse (*Mus musculus*) homolog percent identity, and measures of population genetic constraint based on gnomAD loss-of-function (LoF) and missense observed-to-expected (*o/e*) ratios. Boruta also *'rejected'* several features, including neuronal survival metrics, differentially developmentally regulated (DDG) values, expression in several less disease-relevant tissues (e.g. adrenal gland, thyroid, lung, etc.) and some region-specific STR measures.

We complemented the Boruta analysis by calculating Wilcoxon absolute effect sizes for each feature to quantify differences between NMD and other autosomal protein-coding genes, using Wilcoxon rank-sum tests with multiple-testing correction to assess significance (Bonferroni-adjusted P-value < 0.05; Supplementary Fig. S2). Of the 134 features analyzed, 72 differed significantly between the gene lists. The features with the largest effect sizes included transcript count (absolute $r = 0.097$), several exon-count metrics (absolute $r = 0.070$–0.086), and maximum transcript length (absolute $r = 0.076$), all of which were higher in NMD genes. *M. musculus* homology percent identity also had a relatively large effect size ($r = 0.077$) and was higher in NMD genes. Features with higher Boruta importance generally showed larger absolute effect sizes, indicating agreement between the machine-learning-based and traditional statistical analyses. We therefore compared the rankings of features by Boruta importance with their rankings by Wilcoxon absolute effect size (Fig. 2B). Overall, we observed a modest positive correlation between Boruta importance rank and effect size rank (Spearman $\rho = 0.54$, $P = 2.6 \times 10^{-11}$). This indicates that the two analysis methods capture overlapping but non-identical aspects of the data. Top-ranked features, including GTEx skeletal muscle, GTEx heart left ventricle, and transcript count, ranked relatively highly in both analyses, while features such as GTEx lung, GTEx spleen, GC content and mean 3'UTR length were discordant between rankings.

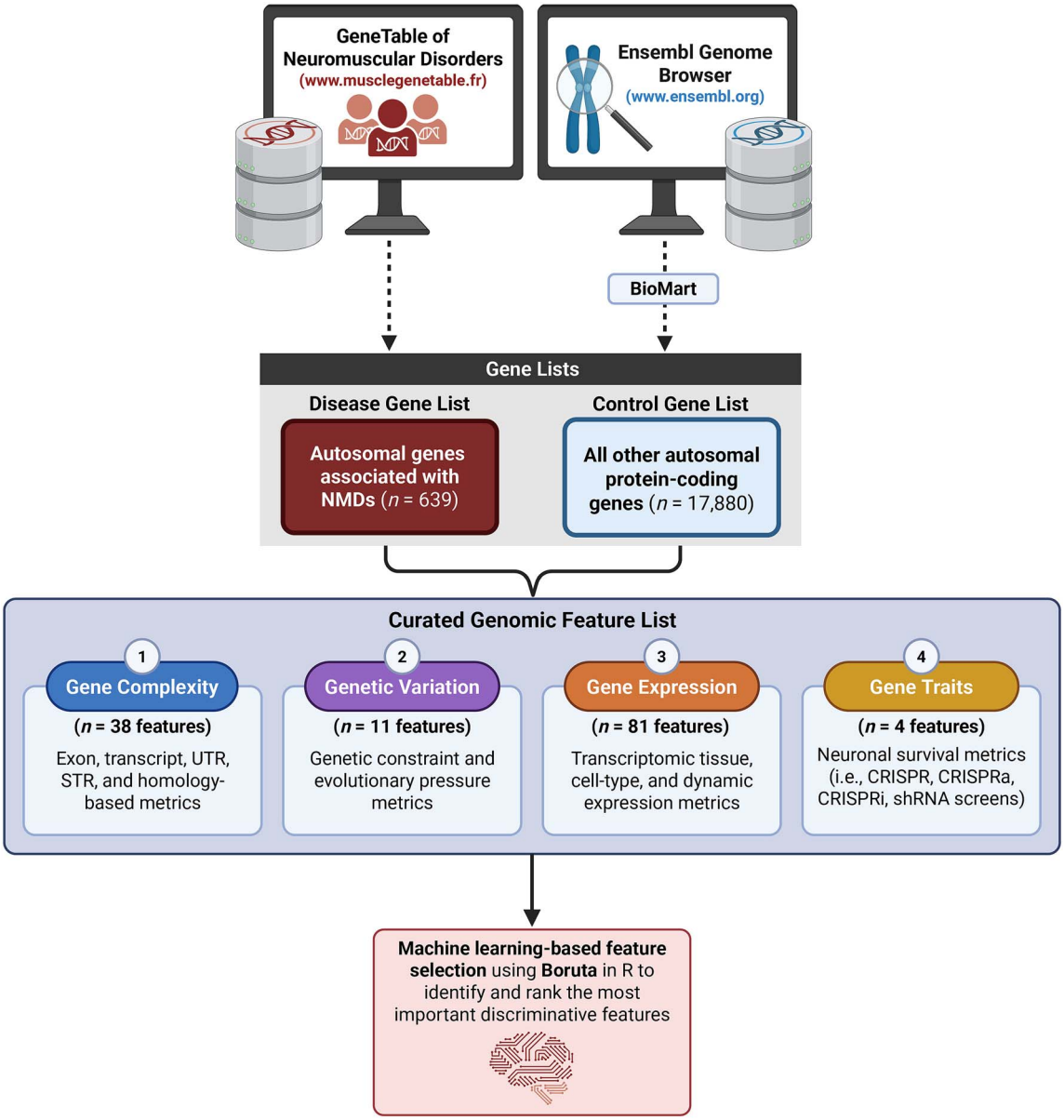


Figure 1 Overview of study design for identifying discriminative features of NMD genes. A list of genes associated with NMDs ($n = 639$) was obtained from the GeneTable of NMDs and merged with a curated genomic feature dataset for other human autosomal protein-coding genes ($n = 17\,880$) from Ensembl and other sources. The two gene sets were evaluated across 134 genomic features within four categories: Gene complexity ($n = 38$), genetic variation ($n = 11$), gene expression ($n = 81$), and gene traits ($n = 4$). Machine-learning-based feature selection using Boruta was applied to identify the most important features distinguishing NMD genes from other genes in the genome. Abbreviations: CRISPR, clustered regularly interspaced short palindromic repeats; CRISPRa, CRISPR activation; CRISPRi, CRISPR interference; NMDs, neuromuscular disorders; shRNA, short hairpin RNA; STR, short tandem repeat; UTR, untranslated region.

Comparison with other high-confidence disease genes identifies NMD-associated features

Because some genomic features associated with NMD genes may also be common to disease genes, we compared NMD genes with high-confidence disease genes from the Gene Curation Coalition (GenCC) [22]. High-confidence GenCC genes were defined as those with *'definitive'* or *'strong'* gene-disease classifications, after excluding genes already included in the NMD set. This resulted in 639 NMD genes, 3468 high-confidence GenCC non-NMD genes, and 14 412 other autosomal protein-coding genes (Fig. 3). Across all features, 37 differed significantly between NMD and GenCC non-NMD genes after Benjamini-Hochberg correction (FDR < 0.05). To identify features that more specifically distinguished NMD genes from other disease genes, we additionally required that features differ significantly between NMD genes and other autosomal protein-coding genes and show a concordant direction of effect across both comparisons. Thirty-three features met these FDR-based criteria (Supplementary Fig. S3). Most were gene-expression measures (28/33), with four gene-complexity features and one gene-trait feature; no genetic-variation features met the NMD-distinguishing criteria.

We next examined the highest-ranked Boruta features in this comparison. GTEx skeletal muscle and heart left ventricle expression

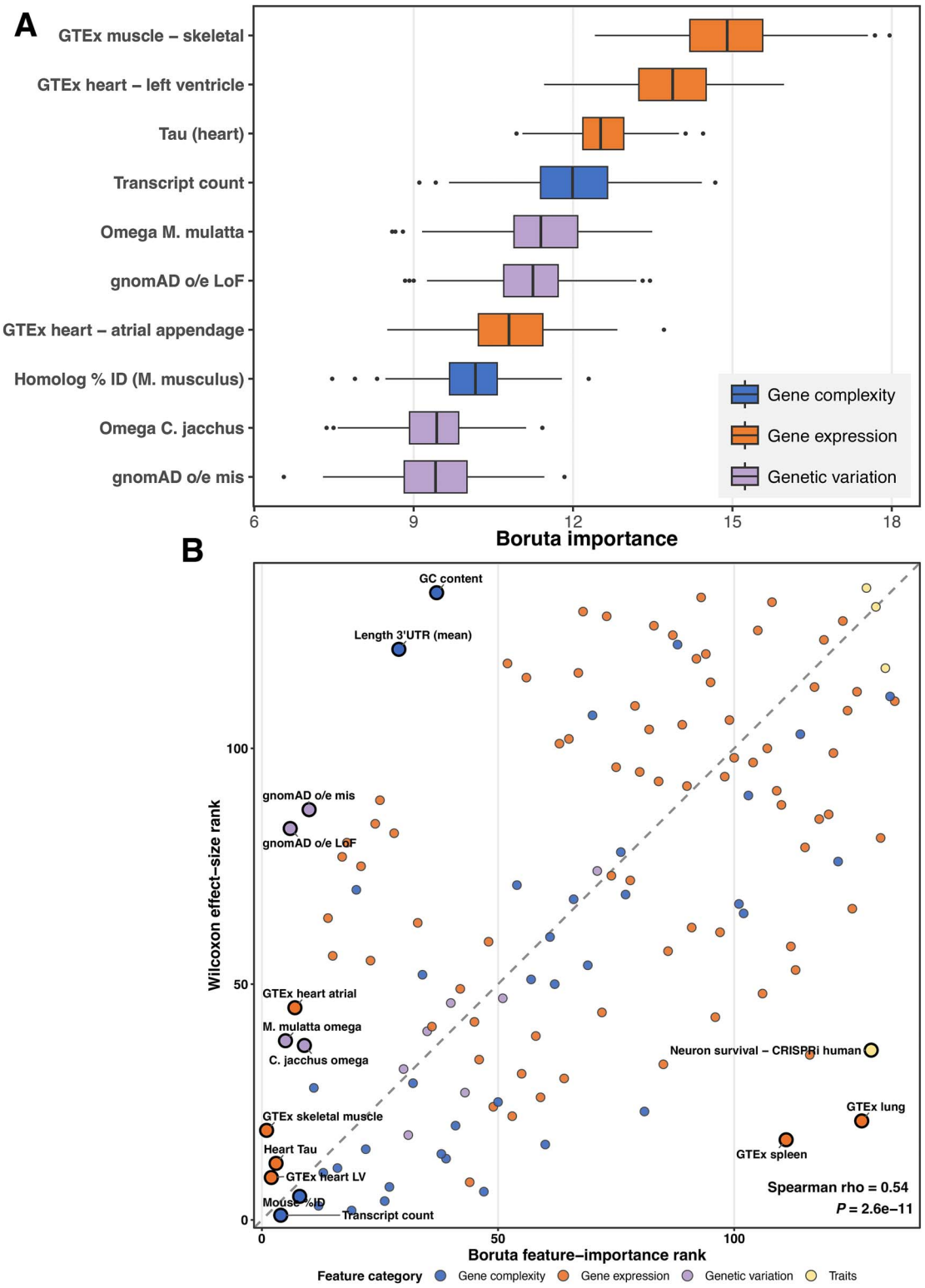


Figure 2 Boruta identifies genomic features that distinguish NMD genes, with concordance between Boruta importance and Wilcoxon effect-size rankings. (A) Boruta importance distributions across iterations for the 10 highest-ranked confirmed features distinguishing NMD genes ($n = 639$) from other autosomal protein-coding genes ($n = 17\,880$). Features were ranked by median Boruta importance and are coloured by category: Gene complexity (blue), gene expression (orange), and genetic variation (lavender). Randomized shadow features were included in the Boruta analysis but are not shown in the figure. Boxplots show the median and interquartile range; whiskers extend to 1.5 times the interquartile range, with more extreme values shown as points. (B) Comparison of Boruta feature-importance rank, based on median importance across iterations, with rank based on absolute Wilcoxon effect size for 134 genomic features. The dashed diagonal indicates equal ranks between methods. Colours indicate feature category. The 10 highest-ranked Boruta features and the five features with the largest differences between the two rankings are labelled. Rankings were positively correlated (spearman $\rho = 0.54$, $P = 2.6 \times 10^{-11}$). Abbreviations: GTEx, genotype-tissue expression; LoF, loss-of-function; mis, missense; o/e, observed/expected; UTR, untranslated region.

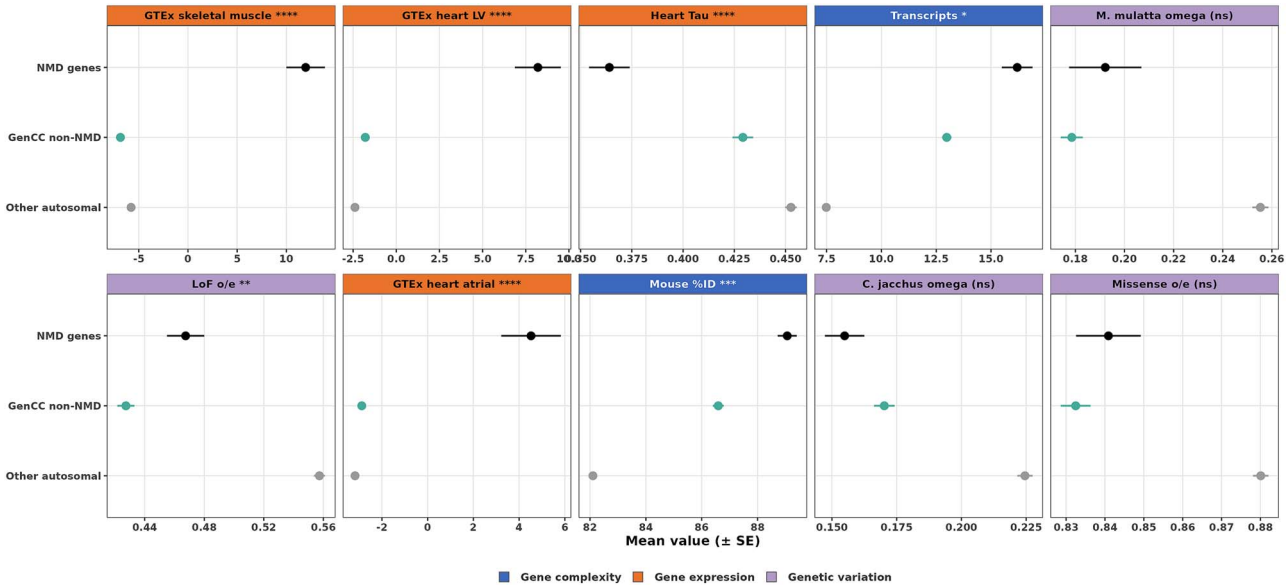


Figure 3 Comparison of top-ranked genomic features between NMD genes and other high-confidence disease genes. Mean values (± standard error, SE) are shown for the 10 highest-ranked Boruta features across NMD genes ($n = 639$), high-confidence GenCC disease genes not associated with NMDs ($n = 3468$), and other autosomal protein-coding genes ($n = 14\,412$). High-confidence GenCC genes were defined as those with *'definitive'* or *'strong'* gene-disease classifications after excluding NMD genes. Group sizes indicate the number of genes assigned to each group; plotted means and SEs use available values for each feature. Significance refers to Wilcoxon rank-sum comparisons between NMD and GenCC non-NMD genes, with Benjamini-Hochberg correction across features: *FDR < 0.05, **FDR < 0.01, ***FDR < 0.001, ****FDR < 0.0001; ns, not significant. Abbreviations: FDR; false discovery rate, GenCC; gene curation coalition, GTEx, genotype-tissue expression; LV, left ventricle; LoF, loss-of-function; mis, missense; o/e, observed/expected; SE, standard error.

were higher in NMD genes compared to other high-confidence disease genes (effect size $r = 0.130$ and 0.133, respectively), as was heart atrial appendage expression ($r = 0.089$), whereas heart *Tau* was lower ($r = -0.096$) (Fig. 3). Mouse homolog percent identity ($r = 0.065$) and transcript count ($r = 0.042$) were also significantly higher in NMD genes. In contrast, the primate *omega* measures and gnomAD missense *o/e* ratio did not differ significantly between NMD and GenCC non-NMD genes. The gnomAD LoF *o/e* ratio was higher in NMD genes than in GenCC non-NMD genes. These results indicate that tissue-expression measures comprised most of the features distinguishing NMD genes from other high-confidence disease genes, whereas several constraint and evolutionary features identified in the initial analysis are shared more broadly across disease-associated genes.

Genomic features vary across NMD disease subgroups

Subgroup analyses showed that the top-ranked genomic features varied across NMD categories (Fig. 4). The clearest differences involved tissue expression: skeletal muscle and cardiac expression were highest in myopathy and cardiomyopathy groups, whereas several neurological NMD groups (e.g. ataxias, paraplegias, neuropathies) showed lower relative values in these tissues. In contrast, transcript count was elevated across a broader range of categories, including both muscle and neurological disorders, suggesting that increased transcript complexity is not restricted to a single clinical subgroup. Constraint and primate evolutionary metrics showed smaller and less consistent differences across NMD subgroups and, consistent with the GenCC comparison, appeared to reflect broader disease-gene properties rather than features specific to individual NMD classes (Fig. 4).

Selected expression, gene-complexity, and homology features are shown in greater detail in Supplementary Fig. S4. Whole-blood expression was lower across several NMD groups, while transcript count, mean transcript length, and total STR count showed distinct subgroup distributions. Cross-species homology also varied by disease group. Mouse homolog percent identity was generally higher in NMD genes than in non-NMD genes, whereas zebrafish and *Drosophila* homology showed more heterogeneous patterns. Metabolic myopathy genes showed relatively high conservation in both zebrafish and *Drosophila*, while distal myopathy genes showed particularly low *Drosophila* homology. The proportion of genes with extreme NMD-like feature values also varied across disease groups, with subgroup-specific differences in the proportion of genes falling within the direction-specific 5% and 1% tails of the non-NMD distribution (Supplementary Fig. S5).

Genomic feature profiles are similar between established and more recently added NMD genes

To assess whether these genomic feature patterns were also present among more recently identified NMD genes, we compared genes established by July 2020 ($n = 550$) with genes added to the NMD gene list between 2020 and 2024 ($n = 89$). Median feature values, expressed as percentiles relative to other autosomal protein-coding genes, were positively correlated across all 107 Boruta-confirmed features (Spearman $\rho = 0.77$, $P = 5.2 \times 10^{-22}$) and across the top 10 features ($\rho = 0.88$, $P = 8.1 \times 10^{-4}$; Fig. 5A). The top-ranked features showed similar profiles between established and more recently added NMD genes (Fig. 5B; Supplementary Fig. S6). These results indicate that the genomic feature patterns identified in established NMD genes are also observed in more recently added genes.

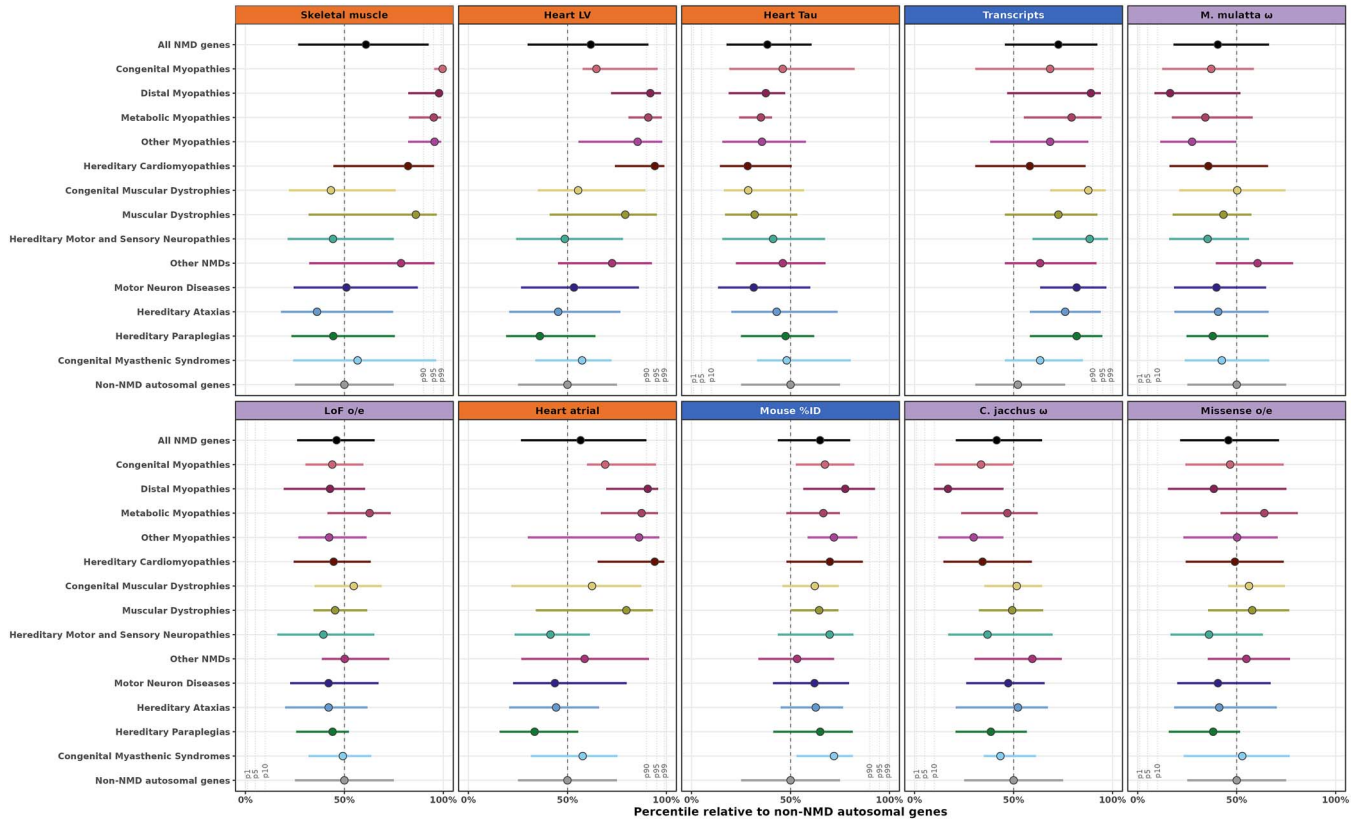


Figure 4 Top-ranked genomic features vary across NMD disease groups. Median feature percentiles and interquartile ranges are shown for the 10 highest-ranked Boruta features across 13 NMD disease groups containing at least 10 genes, all NMD genes, and non-NMD autosomal genes. Feature values were converted to percentiles relative to non-NMD autosomal genes. The dashed vertical line indicates the 50th percentile. Dotted lines indicate the 90th, 95th, and 99th percentiles for features with higher values in NMD genes and the 1st, 5th, and 10th percentiles for features with lower values in NMD genes. Direction was defined from the overall NMD versus non-NMD comparison. Genes associated with more than one NMD disease group were retained in each relevant group. Colours identify disease groups; all NMD genes are shown in black and non-NMD autosomal genes in grey. Abbreviations: GTEx, genotype-tissue expression; LV, left ventricle; LoF, loss-of-function; mis, missense; o/e, observed/expected.

Discussion

Our study identified genomic features that help distinguish NMD genes from other autosomal protein-coding genes and clarified which of these features are shared more broadly with disease-associated genes. Across 134 genomic features, NMD genes differed in disease-relevant tissue expression, gene and transcript complexity, population genetic constraint and cross-species conservation. Comparison with high-confidence GenCC disease genes separated properties shared more broadly across disease-associated genes from features that more specifically characterize NMD genes, which were dominated by tissue-expression measures. These patterns varied across NMD subgroups and were also evident among genes added to the NMD catalogue after 2020.

Tissue-expression measures were among the strongest discriminators of NMD genes, especially in tissues most directly affected by these disorders. Skeletal muscle and left ventricular heart expression ranked highest in the Boruta analysis and differed significantly between NMD and other autosomal protein-coding genes, whereas whole-blood expression was lower. The prominence of skeletal-muscle and cardiac expression is expected given the phenotypes represented in the NMD gene set, but these differences were also evident when NMD genes were compared with high-confidence GenCC non-NMD disease genes. Of the 33 features meeting the NMD-distinguishing criteria, 28 were expression measures, including higher expression in skeletal muscle and cardiac tissues. The subgroup analyses further showed that these patterns were not uniform across NMDs, with the strongest skeletal-muscle and cardiac profiles in myopathy and cardiomyopathy groups [23, 24]. These findings indicate that disease-relevant expression distinguishes NMD genes not only from the broader genome, but also from other established disease genes, while retaining clear differences across NMD subtypes.

Disease-relevant expression profiles can provide information for genetic variant interpretation that is not apparent from genome sequencing alone [25]. When direct sampling of affected tissues is impractical, disease-relevant cellular models may provide complementary evidence for evaluating candidate genes and variants. In such cases, more accessible sources such as peripheral blood mononuclear cells or skin fibroblasts can be reprogrammed into patient-specific induced pluripotent stem cells (iPSCs), which can then be differentiated into relevant cell types or organoids to capture disease-specific gene expression profiles [26, 27]. This approach has already been employed in NMD research; for instance, studies have produced organoids from patient-derived iPSCs that serve as platforms for drug testing and recapitulating disease pathology in conditions such as muscular dystrophies and amyotrophic lateral sclerosis [28–31].

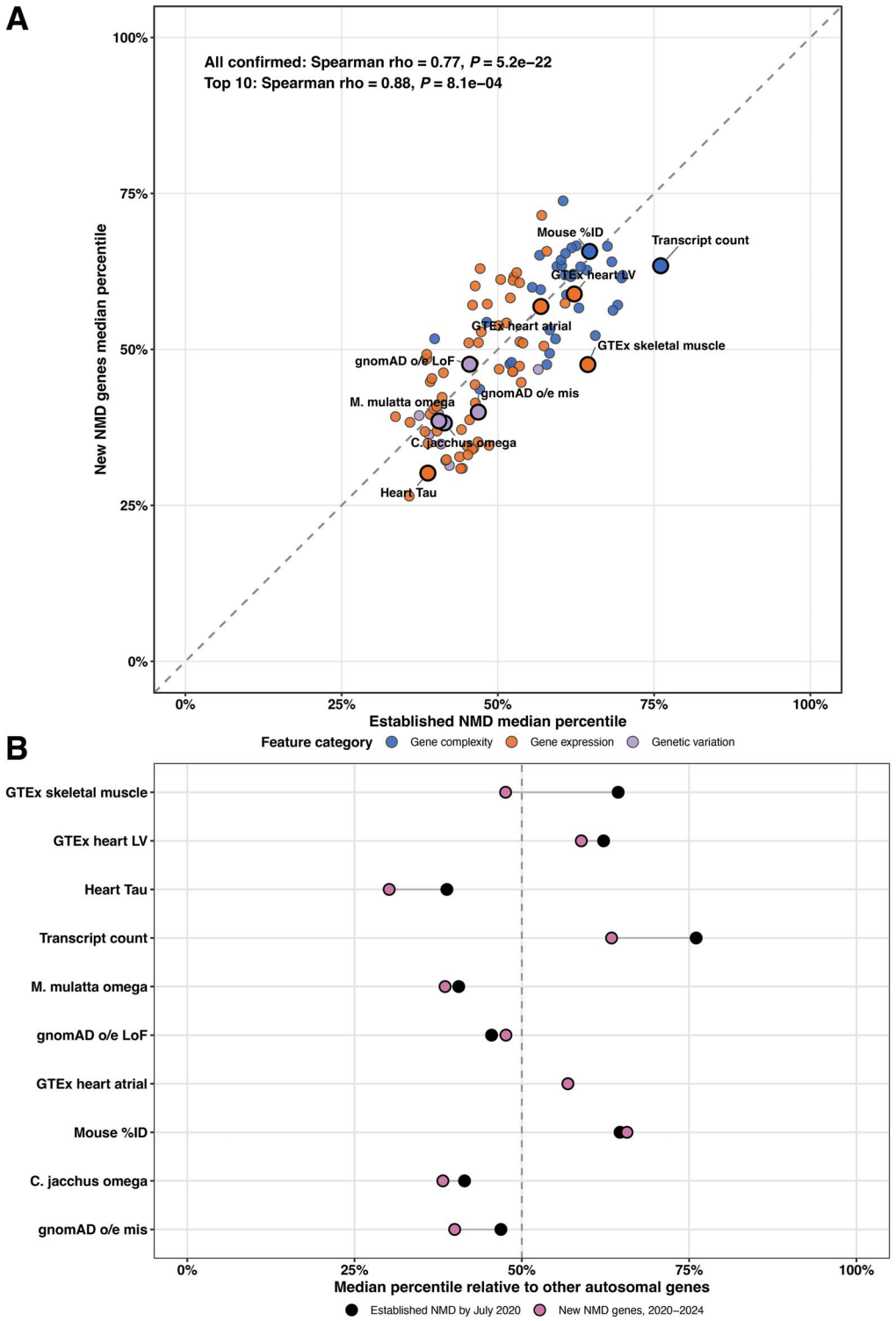


Figure 5 Genomic feature profiles are correlated between established and more recently added NMD genes. (A) Comparison of genomic feature profiles for NMD genes established by July 2020 ($n = 550$) and genes added to the NMD gene list between 2020 and 2024 ($n = 89$). For each feature, the median value within each gene set was expressed as a percentile relative to the distribution among other autosomal protein-coding genes. Points represent the 107 Boruta-confirmed features and are coloured by feature category; the 10 highest-ranked Boruta features are outlined and labelled. The dashed diagonal indicates equal percentiles between the two gene sets. Feature profiles were correlated across all confirmed features (spearman $\rho = 0.77$, $P = 5.2 \times 10^{-22}$) and across the top 10 features (spearman $\rho = 0.88$, $P = 8.1 \times 10^{-4}$). (B) Median feature values expressed as percentiles for the 10 highest-ranked Boruta features in established NMD genes (black) and genes added between 2020 and 2024 (rose). Values for the same feature are connected by grey lines; the dashed vertical line indicates the 50th percentile relative to other autosomal protein-coding genes. Abbreviations: GTEx, genotype-tissue expression; LV, left ventricle; LoF, loss-of-function; mis, missense; o/e, observed/expected.

Gene-complexity measures were also prominent, with NMD genes tending to have more transcripts, longer transcripts and greater STR content than other autosomal genes. These properties are consistent with the involvement of large, extensively spliced genes such as *TTN* and *NEB*, and with the established contribution of alternative splicing to NMDs [32–37]. NMD genes also had more annotated transcripts than high-confidence non-NMD disease genes, whereas transcript length and most exon-count measures showed less evidence of a direct difference between these groups. This suggests that isoform diversity, rather than gene size alone, may be a particularly relevant aspect of NMD gene architecture. Greater transcript diversity may increase the potential for tissue- or isoform-specific variant effects and supports isoform-aware annotation and transcript-resolved approaches such as long-read sequencing.

NMD subgroup analyses illustrated that these genomic feature patterns were not uniform. Transcript counts were elevated across a relatively broad range of muscle and neurological groups, whereas STR counts were highest in hereditary ataxias, hereditary motor and sensory neuropathies, hereditary paraplegias, and hereditary cardiomyopathies, with smaller shifts in most myopathy groups. Increased STR burden should therefore not be interpreted as a general property of every NMD category. Long-read DNA and RNA sequencing may be useful for resolving the transcript structures, isoforms, and repetitive sequences represented by these features [38].

Cross-species homology was also among the genomic features that distinguished NMD genes. Mouse homolog percent identity was among the top-ranked features and was higher in NMD genes than in other autosomal protein-coding genes. It also remained higher when NMD genes were compared with high-confidence GenCC disease genes. The retention of higher mouse homolog percent identity in the GenCC comparison indicates that this conservation signal is not explained entirely by properties shared across high-confidence disease genes. Homology patterns were less consistent in other commonly used NMD model organisms, including zebrafish and *Drosophila*, and varied across NMD subgroups. For example, metabolic myopathy genes showed relatively high conservation in both species, whereas distal myopathy genes showed particularly low *Drosophila* homology. These results suggest that sequence conservation should be considered when selecting model systems for individual candidate genes. In cases where conservation is limited, human cell-based systems, such as iPSCs, may provide a complementary platform for studying disease mechanisms [27].

The comparison of established and more recently added NMD genes also showed that these genomic patterns persisted as the NMD gene catalogue expanded. Feature profiles were strongly correlated between genes established by 2020 and those added between 2020 and 2024, both across all Boruta-confirmed features and among the top-ranked features. Although this retrospective analysis does not test prospective gene prediction, it indicates that the identified genomic characteristics are not restricted to the historically established NMD gene set.

While the analyses presented here offer insight into the biological characteristics of NMD genes, certain limitations should be considered. The comparison set of other autosomal protein-coding genes may include genes that are subsequently linked to NMDs or to incompletely characterized neuromuscular phenotypes. The GenCC analysis helps address this issue by comparing NMD genes with other high-confidence disease genes, although the underlying gene lists remain dependent on current disease annotations. Genes on sex chromosomes were excluded because several feature datasets were incomplete for these chromosomes, limiting the analysis of important X-linked NMD genes such as *DMD*, *MTM1*, and *EMD*. In addition, Boruta importance reflects the ability of a feature to distinguish gene sets rather than an independent or causal contribution to disease, and some features showed differences between Boruta and conventional effect-size rankings.

Future studies could extend the feature set as additional disease-relevant genomic and functional datasets become available. Genome-wide perturbation screens and long-read sequencing could add measures of gene function, transcript structure, isoform diversity, and repetitive sequence content that are not represented in the current analysis [38]. Prospective studies will be needed to determine whether incorporating these features improves prioritization of previously unrecognized NMD genes.

In conclusion, NMD genes show a consistent genomic pattern characterized by disease-relevant tissue expression, gene complexity, and evolutionary conservation. Comparison with other high-confidence disease genes shows that several of these properties are shared more broadly across disease-associated genes, whereas tissue-expression features provide the clearest distinction of NMD genes. Subgroup analyses demonstrate substantial heterogeneity across NMD classes, but the overall feature profile is reproduced among genes added to the NMD catalogue more recently. These findings are limited to autosomal genes because sex-chromosome genes were excluded due to incomplete feature coverage. This consistency provides a basis for testing the value of these features in prospective candidate-gene prioritization studies. If validated in this setting, improved gene prioritization could contribute to higher molecular diagnosis rates in NMDs, with potential benefits for disease monitoring, treatment decisions, genetic counselling and access to support resources [11]. The genomic features identified here may also provide additional context for the interpretation of genetic data in unresolved NMD cases.

Materials and methods

Database construction and NMD gene annotation

We built a comprehensive dataset of genomic features for human protein-coding genes.

To ensure consistency across features, analyses were restricted to autosomal genes as several features were incomplete for sex chromosomes. Genes associated with NMDs were identified using the curated GeneTable of NMDs (https://musclegenetable.fr; accessed November 2024), resulting in 639 NMD genes and 17 880 other autosomal protein-coding genes. Genic information was obtained from Ensembl v.114 and the *biomaRt* package in R (v.2.58.2). Additional data-processing details are provided in Supplementary Methods.

Construction of an extensive genomic feature list

To identify features that distinguish NMD genes, we assembled a set of 134 gene-level features for each gene from publicly accessible multi-omics resources, grouped into four major categories (Fig. 1): (1) gene complexity ($n = 38$), (2) genetic variation ($n = 11$), (3) gene expression ($n = 81$), and (4) other gene traits ($n = 4$).

We obtained genomic features from publicly available resources, including measures of gene and transcript structure, repetitive

sequence content, cross-species homology, population genetic constraint, evolutionary selection, tissue and developmental gene expression, and functional gene essentiality. Detailed definitions, processing procedures, and data sources for individual features are provided in Supplementary Methods and Supplementary Table 1.

All data sources were merged into a unified genomic feature list. Missing values were imputed using the *MICE* R package (v.3.17.0), using predictive mean matching to generate five imputed datasets ($m = 5$). The imputed feature matrix was used for Boruta feature selection, whereas statistical comparisons, subgroup analyses, GenCC comparisons, and temporal analyses used the original feature values, with missing values excluded on a feature-specific basis.

Statistical analysis

For each genomic feature, we compared NMD genes with other autosomal protein-coding genes using Wilcoxon rank-sum tests with Bonferroni correction for multiple testing. Wilcoxon effect sizes (r) and corresponding 95% confidence intervals were calculated to quantify differences between groups. Effect size direction was assigned according to the direction of the mean differences between groups. Agreement between feature rankings based on Boruta importance and absolute Wilcoxon effect size was assessed using Spearman rank correlation and Kendall rank correlation. All statistical analyses were performed in R (v.4.3.3).

Boruta feature selection

To identify genomic features that most strongly distinguish NMD genes from other autosomal protein-coding genes, we applied the Boruta algorithm [19] to the imputed feature matrix. Feature selection was performed using the R *Boruta* package (v.8.0.0) with a maximum of 250 iterations and a fixed random seed for reproducibility. Tentative feature classifications were then resolved using the *TentativeRoughFix()* function. Feature importance was summarized across Boruta iterations, and confirmed features were ranked by median importance for selection of the 10 highest-ranked features.

Comparison with high-confidence disease genes

To determine whether genomic features associated with NMD genes were also characteristic of disease genes more generally, we compared NMD genes with high-confidence disease-associated genes from the GenCC [22]. GenCC entries classified as *'definitive'* or *'strong'* were retained, and genes already present in the NMD set were excluded from the GenCC comparator group. Genes were assigned to three mutually exclusive groups: NMD genes ($n = 639$), high-confidence GenCC non-NMD genes ($n = 3468$), and remaining autosomal protein-coding genes ($n = 14\,412$).

For each genomic feature, pairwise Wilcoxon rank-sum tests were performed for NMD versus GenCC non-NMD genes, NMD versus other autosomal genes, and GenCC non-NMD versus other autosomal genes. Benjamini-Hochberg correction was applied separately within each comparison, with FDR < 0.05 considered significant. Wilcoxon effect sizes were assigned a direction based on the mean difference between groups. Features were classified as NMD-distinguishing when they differed significantly between NMD and GenCC non-NMD genes and between NMD and other autosomal genes, with concordant directions of effect in both comparisons.

NMD disease subgroup analyses

To assess variation across NMD disease groups, genes were stratified according to the disease categories defined in the GeneTable of NMDs. Groups containing at least 10 genes (Supplementary Table 2) were retained ($n = 13$), and genes associated with more than one disease group were retained in each relevant group. For the top-ranked Boruta features and selected additional features, individual gene values were converted to percentiles relative to the corresponding distribution among non-NMD autosomal genes. Disease-group distributions were summarized using the median percentile and interquartile range.

To examine the contribution of genes with extreme feature values, thresholds were defined from the non-NMD distribution. Feature direction was determined from the overall NMD-versus-non-NMD comparison. For features with higher values in NMD genes, the NMD-like 5% and 1% tails corresponded to values above the 95th and 99th percentiles of non-NMD genes; for features with lower values in NMD genes, the corresponding tails comprised values below the 5th and 1st percentiles. The number and proportion of genes within these direction-specific tails were calculated for each disease group.

Comparison of established and more recently added NMD genes

To assess whether the identified genomic feature patterns were also present among more recently identified NMD genes, we compared the GeneTable of NMDs from July 2020 with the November 2024 version used for the primary analysis. After restricting to autosomal genes represented in the feature dataset and harmonizing updated gene symbols, genes present in the 2020 list were classified as established NMD genes ($n = 550$), while genes present in the 2024 list but absent from the 2020 list were classified as more recently added NMD genes ($n = 89$).

For each of the 107 Boruta-confirmed features, the median feature value within each NMD gene set was converted to a percentile relative to the distribution among other autosomal protein-coding genes. Similarity between established and more recently added NMD genes was assessed using Spearman correlation across all confirmed features and separately across the 10 highest-ranked Boruta features. Mean values and standard errors for the top-ranked features were also calculated for visualization.

Acknowledgements

We gratefully acknowledge the data providers whose contributions, through open-access repositories or direct data sharing, made this genomic feature analysis possible. Limited sections of the manuscript and analysis code were edited for language clarity and code quality with assistance from ChatGPT (OpenAI, San Francisco, CA), without affecting the scientific content.

Author contributions

Alexandria Martin (Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing—original draft, Writing—review & editing), M. Alejandra Llanes-Cuesta (Data curation, Formal analysis, Project administration, Writing—review & editing), Jessica N. Hartley (Conceptualization, Investigation, Writing—review & editing),

Patrick Frosk (Conceptualization, Supervision, Writing—review & editing), Britt Drogemoller (Conceptualization, Funding acquisition, Methodology, Supervision, Writing—review & editing), and Galen Wright (Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Resources, Supervision, Visualization, Writing—original draft, Writing—review & editing).

Supplementary material

Supplementary material is available at *Human Molecular Genetics* online.

Conflicts of interest

The authors report no competing interests.

Funding

This work was supported by the Natural Sciences and Engineering Research Council of Canada Undergraduate Student Research Awards and a Children's Hospital Research Institute of Manitoba Undergraduate Summer Studentship to A.M.; a Natural Sciences and Engineering Research Council of Canada Discovery Grant [RGPIN-2022-04500] and a Tier 2 Canada Research Chair in Pharmacogenomics and Precision Medicine [CRC-2019-00040; CRC-2023-00351] to B.I.D.; and a Natural Sciences and Engineering Research Council of Canada Discovery Grant [RGPIN-2022-04509], a Tier 2 Canada Research Chair in Neurogenomics [CRC-2019-00145; CRC-2025-00075], and a Research Manitoba New Investigator Grant [2022–5381] to G.E.B.W.

Data availability

The data analyzed in this study were obtained from publicly available resources described in the Materials and Methods and Supplementary Table 1. Analysis code is available at: https://github.com/Wright-Lab-Neurogenomics-Research/nmd-genomic-features-hmg.

References

1. Dowling JJ, Gonorazky D, Cohn RD *et al.* Treating pediatric neuromuscular disorders: the future is now. *Am J Med Genet A* 2018;**176**:804–841. https://doi.org/10.1002/ajmg.a.38418.
2. Deenen JCW, Horlings CGC, Verschuuren JJGM *et al.* The epidemiology of neuromuscular disorders: a comprehensive overview of the literature. *J Neuromuscul Dis* 2015;**2**:73–85. https://doi.org/10.3233/JND-140045.
3. Wilson LA, Macken WL, Perry LD *et al.* Neuromuscular disease genetics in under-represented populations: increasing data diversity. *Brain* 2023;**146**:5098–5109. https://doi.org/10.1093/brain/awad254.
4. Zatz M, Passos-Bueno MR, Vainzof M. Neuromuscular disorders: genes, genetic counseling and therapeutic trials. *Genet Mol Biol* 2016;**39**:339–348. https://doi.org/10.1590/1678-4685-GMB-2016-0019.
5. Benarroch L, Bonne G, Rivier F *et al.* The 2025 version of the gene table of neuromuscular disorders (nuclear genome). *Neuromuscul Disord* 2025;**46**:105261. https://doi.org/10.1016/j.nmd.2024.105261.
6. Laing NG. Genetics of neuromuscular disorders. *Crit Rev Clin Lab Sci* 2012;**49**:33–48. https://doi.org/10.3109/10408363.2012.658906.
7. Koczwara KE, Lake NJ, DeSimone AM *et al.* Neuromuscular disorders: finding the missing genetic diagnoses. *Trends Genet* 2022;**38**:956–971. https://doi.org/10.1016/j.tig.2022.07.001.
8. Yubero D, Natera-de Benito D, Pijuan J *et al.* The increasing impact of translational research in the molecular diagnostics of neuromuscular diseases. *IJMS* 2021;**22**:4274. https://doi.org/10.3390/ijms22084274.
9. Chen Z, Tucci A, Cipriani V *et al.* Functional genomics provide key insights to improve the diagnostic yield of hereditary ataxia. *Brain* 2023;**146**:2869–2884. https://doi.org/10.1093/brain/awad009.
10. Bauskis A, Strange C, Molster C *et al.* The diagnostic odyssey: insights from parents of children living with an undiagnosed condition. *Orphanet J Rare Dis* 2022;**17**:233. https://doi.org/10.1186/s13023-022-02358-x.
11. Dratch L, Azage M, Baldwin A *et al.* Genetic testing in adults with neurologic disorders: indications, approach, and clinical impacts. *J Neurol* 2024;**271**:733–747. https://doi.org/10.1007/s00415-023-12058-6.
12. Mitsuhashi S, Matsumoto N. Long-read sequencing for rare human genetic diseases. *J Hum Genet* 2020;**65**:11–19. https://doi.org/10.1038/s10038-019-0671-8.
13. Najafi A, Caspar SM, Meienberg J *et al.* Variant filtering, digenic variants, and other challenges in clinical sequencing: a lesson from fibrillinopathies. *Clin Genet* 2020;**97**:235–245. https://doi.org/10.1111/cge.13640.
14. Nurk S, Koren S, Rhie A *et al.* The complete sequence of a human genome. *Science* 2022;**376**:44–53. https://doi.org/10.1126/science.abj6987.
15. Capanu M, Ionita-Laza I. Integrative analysis of functional genomic annotations and sequencing data to identify rare causal variants via hierarchical modeling. *Front Genet* 2015;**6**:17. https://doi.org/10.3389/fgene.2015.00176.
16. Ionita-Laza I, McCallum K, Xu B *et al.* A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet* 2016;**48**:214–220. https://doi.org/10.1038/ng.3477.
17. Li X, Li Z, Zhou H *et al.* Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nat Genet* 2020;**52**:969–983. https://doi.org/10.1038/s41588-020-0676-4.
18. Reynolds RH, Hardy J, Ryten M *et al.* Informing disease modelling with brain-relevant functional genomic annotations. *Brain* 2019;**142**:3694–3712. https://doi.org/10.1093/brain/awz295.
19. Kursa MB, Rudnicki WR. Feature selection with the **Boruta** package. *J Stat Soft* 2010;**36**:1–13. https://doi.org/10.18637/jss.v036.i11.
20. Ahmed H, Soliman H, Elmogy M. Early detection of Alzheimer's disease using single nucleotide polymorphisms analysis based on gradient boosting tree. *Comput Biol Med* 2022;**146**:105622. https://doi.org/10.1016/j.compbiomed.2022.105622.
21. Sevim Bayrak C, Stein D, Jain A *et al.* Identification of discriminative gene-level and protein-level features associated with pathogenic gain-of-function and loss-of-function variants. *Am J Hum Genet* 2021;**108**:2301–2318. https://doi.org/10.1016/j.ajhg.2021.10.007.
22. DiStefano MT, Goehringer S, Babb L *et al.* The gene curation coalition: a global effort to harmonize gene-disease evidence resources. *Genet Med* 2022;**24**:1732–1742. https://doi.org/10.1016/j.gim.2022.04.017.
23. Goodwin JF. Cardiomyopathies and specific heart muscle diseases. Definitions, terminology, classifications and new and old approaches. *Postgrad Med J* 1992;**68 S1**:3–6.

24. Nagy H, Veerapaneni KD. Myopathy. In: *StatPearls [Internet]*. Treasure Island (FL): StatPearls Publishing; 2026 Jan. Available from: https://www.ncbi.nlm.nih.gov/books/NBK562290/
25. Cummings BB, Marshall JL, Tukiainen T *et al.* Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci Transl Med* 2017;**9**:eaal5209. https://doi.org/10.1126/scitranslmed.aal5209.
26. El Hokayem J, Cukier HN, Dykxhoorn DM. Blood derived induced pluripotent stem cells (iPSCs): benefits, challenges and the road ahead. *J Alzheimers Dis Parkinsonism* 2016;**6**:275. https://doi.org/10.4172/2161-0460.1000275.
27. Speciale AA, Ellerington R, Goedert T *et al.* Modelling neuromuscular diseases in the age of precision medicine. *JPM* 2020;**10**:178. https://doi.org/10.3390/jpm10040178.
28. Auletta B, Chiolerio P, Cecconi G *et al.* Tissue-engineered neuromuscular organoids. *Commun Biol* 2025;**8**:1074. https://doi.org/10.1038/s42003-025-08484-z.
29. Cheesbrough A, Harley P, Riccio F *et al.* A scalable human iPSC-based neuromuscular disease model on suspended biobased elastomer nanofiber scaffolds. *Biofabrication* 2023;**15**:045020. https://doi.org/10.1088/1758-5090/acf39e.
30. Gao C, Shi Q, Pan X *et al.* Neuromuscular organoids model spinal neuromuscular pathologies in C9orf72 amyotrophic lateral sclerosis. *Cell Rep* 2024;**43**:113892. https://doi.org/10.1016/j.celrep.2024.113892.
31. Maffioletti SM, Sarcar S, Henderson ABH *et al.* Three-dimensional human iPSC-derived artificial skeletal muscles model muscular dystrophies and enable multilineage tissue engineering. *Cell Rep* 2018;**23**:899–908. https://doi.org/10.1016/j.celrep.2018.03.091.
32. Lopes I, Altab G, Raina P *et al.* Gene size matters: an analysis of gene length in the human genome. *Front Genet* 2021;**12**:559998. https://doi.org/10.3389/fgene.2021.559998.
33. Meyer LC, Wright NT. Structure of giant muscle proteins. *Front Physiol* 2013;**4**:1–12. https://doi.org/10.3389/fphys.2013.00368.
34. Savarese M, Välipakka S, Johari M *et al.* Is gene-size an issue for the diagnosis of skeletal muscle disorders? *JND* 2020;**7**:203–216. https://doi.org/10.3233/JND-190459.
35. Pinheiro H, Pimentel MR, Sequeira C *et al.* mRNA distribution in skeletal muscle is associated with mRNA size. *J Cell Sci* 2021;**134**:jcs256388. https://doi.org/10.1242/jcs.256388.
36. Sahakyan AB, Balasubramanian S. Long genes and genes with multiple splice variants are enriched in pathways linked to cancer and other multigenic diseases. *BMC Genomics* 2016;**17**:225. https://doi.org/10.1186/s12864-016-2582-9.
37. Lim WF, Rinaldi C. RNA transcript diversity in neuromuscular research. *J Neuromuscul Dis* 2023;**10**:473–482. https://doi.org/10.3233/JND-221601.
38. Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. *Nat Rev Genet* 2020;**21**:597–614. https://doi.org/10.1038/s41576-020-0236-x.

© The Author(s) 2026. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
Human Molecular Genetics, 2026, Vol. 35, 22, 1–11
https://doi.org/10.1093/hmg/ddag115
Original Article