Title: Genome-wide identification of dominant polyadenylation hexamers for use in variant classification

Authors: Henoke K. Shiferaw[1], Celine S. Hong[1], David N. Cooper[2], Jennifer J. Johnston[1], NISC[3], Leslie G. Biesecker[1]

Affiliations: 1. Center for Precision Health Research, 2. Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff CF14 4XN, United Kingdom, 3. NIH Intramural Sequencing Center, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, 20892 USA, 3.

Correspondence to:

Henoke Shiferaw

National Human Genome Research Institute

50 South Drive Room 5138

Bethesda, MD, USA

20892

Email address: henoke.shiferaw@nih.gov

ABSTRACT

Polyadenylation is essential for the stabilization and export of mRNAs to the cytoplasm. The polyadenylation signal hexamer (herein referred to as hexamer) motif plays a key role in this process. As yet, however, only 14 Mendelian disorders have been associated with hexamer variants. This is likely a serious under-ascertainment as hexamers are poorly defined and not routinely examined in molecular analysis. To facilitate the interrogation of putatively pathogenic hexamer variants, we set out to define functionally important hexamers genome-wide as a resource for research and clinical testing. We identified predominant polyA sites (herein referred to as pPAS) and putatively predominant hexamers across protein coding genes (PAS usage >50% per gene, 15,767 sites). As a measure of the validity of these sites, the population constraint of 4,532 predominant hexamers was measured. The predominant hexamers had fewer observed variants compared to non-predominant hexamers and the trimer controls, whilst the CADD scores for variants in these sites were significantly higher than in controls. Exome and transcriptome data from 76 individuals were interrogated to identify 65 variants in predominant hexamers. 3’ RNA-seq data showed that these variants resulted in alternate [LES: alternative?] polyadenylation events (38%) and in elongated mRNA transcripts (12%). Our list of pPAS and predominant hexamers are available in the UCSC genome browser and as BED files in GitHub. We suggest that this list of predominant hexamers may be used to interrogate exome and genome data. Variants in these predominant hexamers should be considered candidates for pathogenic variation in human disease, and to that end we suggest pathogenicity criteria for classifying hexamer variants.

INTRODUCTION

Most genes use polyadenylation (polyA) to maintain the stability of nascent mRNA and to export mRNA to the cytoplasm. This process is directed by the polyadenylation signal (PAS) which comprises a set of sequences in the 3' untranslated region. These sequences initiate cleavage of the 3' end of the mRNA at the polyA site followed by addition of a polyA tail [1]. The best defined component of these sequences is the polyadenylation signal hexamer (herein referred to as hexamer) motif, typically located 10-30 bases upstream of the polyA site [2]. The canonical hexamer AATAAA (and variants thereof) is critical in facilitating this process [3].

Relatively few hexamer variants have been associated with disease. In the 2021.2 version of the HGMD [4] database, only 31 hexamer variants in 14 genes associated with Mendelian disorders were listed (Table 1). By comparison, in HGMD overall, there were 1,058 disease-associated variants in initiator methionine (AUG) codons and 323,661 disease-associated variants of any type. We postulated that disease-associated hexamer variants are under-ascertained because of the manifold challenges in identifying such variants.

Whilst there are numerous polyA site and hexamer databases, these databases invariably provide lists of all potential sites without considering either their importance or their functionality [5-8]. For example, the PolyASite 2.0 database specifies >560,000 PAS and >860,000 hexamers based on 3' end sequencing of mRNA molecules for 32,494 genes. A major challenge of hexamer analysis is that genes can have numerous candidate PAS that may not be biologically relevant. Additionally, each polyA site can have multiple hexamer motifs that represent candidate hexamers. Although several efforts have been made to globally assess hexamers and the consequences of variation in those sites, those efforts were focused on common variants associated with susceptibility traits, rather than Mendelian, single gene disorders [9, 10]. To identify hexamers that are candidates for association with Mendelian disease, we postulated that such hexamers would 1) be associated with the pPAS within a given gene, 2) exhibit a strong population constraint signal, 3) demonstrate aberrant scores from *in silico* evaluations of variants, and 4) a sample set of such variants would be associated with perturbations in mRNA processing. We set out to define a set of PAS and corresponding hexamers that met these criteria which could be used as a candidate list for Mendelian gene-associated pathogenic variation. We also propose variant classification criteria for clinical genomic testing laboratories.

METHODS

Identification of pPAS and predominant hexamers for protein-coding genes

To identify pPAS, we used the PolyASite 2.0 database (https://polyasite.unibas.ch/atlas/) [5]. Only the PAS with gene annotations as protein-coding genes defined by Ensembl release 96 were examined. Quality control included re-annotation and removal of intergenic PAS (see Supplementary methods).

To measure the relative usage of PAS within a gene, we used the quantification of polyA site abundance from PolyA Site 2.0, which was quantified in Transcripts Per Million (TPM). For each polyA site, TPM values were averaged across 221 samples. Usage for each polyA site was calculated as $\frac{TPM\ of\ a\ site}{Total\ TPMs\ of\ all\ sites\ in\ a\ gene}$. The pPAS were defined as >50% polyA site usage. If no polyA site reached >50% usage, the gene was not further considered.

To identify the candidate hexamers for each pPAS, hexamers occurring within the -60 to +10 region around each pPAS representative position (the position that had the highest number of supporting reads in PolyASite 2.0) were examined. The predominant hexamers were selected based on the strength of the hexamer sequence in polyadenylation [11], relative distance from the representative position, and the distribution of the distances at which the hexamer sequence occurred as the lone hexamer (Figure 1, see Supplementary methods).

Constraint analysis of hexamer variants

To assess the population constraint within the predominant hexamers, we compared gnomAD v3.0 single nucleotide variant (SNV) frequencies in predominant hexamers with the SNV frequencies in control regions. The control regions were defined as Secondary hexamers (hexamers upstream of pPAS that were not the predominant hexamers), Other hexamers (hexamers upstream of non-pPAS), and trimer controls as described in the Methods (Figure 4A). For each of the predominant hexamers, two trimers from the 3'UTR upstream of the predominant hexamers, with the same nucleotide composition as the predominant hexamers, were selected as controls (see Supplementary methods, Figure 4). The distribution of variant frequencies in all positions (inclusive of positions with 0% VAF) within predominant hexamers, non-predominant hexamers, and the trimer controls were compared using the Mann-Whitney U test and the Kolmogorov-Smirnov test.

Box1: Glossary

Polyadenylation sites (PAS)

The position in the mRNA that is cleaved and polyadenylated.

Predominant polyadenylation site (pPAS)

The polyadenylation site with >50% usage in a given gene.

Non-predominant polyadenylation site (Non-pPAS)

A polyadenylation site with ≤ 50% usage in a given gene.

Hexamers

The six nucleotide polyadenylation signal sequence motif within 60 nucleotides upstream of a PAS.

Predominant hexamers

The hexamer that is associated with the pPAS.

Other hexamers

Hexamers that are associated with Non-pPAS of a gene.

Other strong hexamers

A subgroup of other hexamers that are either AATAAA or ATTAAA sequence motif.

Secondary hexamers

Hexamers upstream of pPAS that are not predominant hexamers.

Non-predominant hexamers

All hexamers that are not predominant hexamers (secondary hexamers plus other hexamers).

Sequencing and data processing

Two types of RNA sequencing were performed on 76 samples from the ClinSeq® study to study the effects of predominant hexamer variants. RNA-sequencing (TruSeq Stranded mRNA kit, Illumina) was performed to examine gene expression and extension of mRNA, and 3'-end sequencing (Quantseq_REV, Lexogen, Greenland, NH) was performed to identify and measure the PAS usage. Total RNA was isolated from whole blood using the PAXgene Blood RNA system (Qiagen, Gaithersburg, MD), prepared as described in the Supplementary methods, and sequenced at the NIH Intramural Sequencing Center (NISC) on a NovaSeq 6000 with v1.0 reagents (Illumina).

FASTQ files were aligned to hg19 using STAR v.2.7.3a [12]. For RNA sequencing, BAM files were aligned to the transcriptome. Duplicate reads were marked with Picard v.2.22.2 [13], and gene expression was quantified using RSEM v.1.3.2. For 3' end sequencing, the PolyASite 2.0 pipeline was used to identify PAS and hexamers (see Supplementary methods).

Assessing the effect of hexamer SNVs

The effect of predominant hexamer SNVs was compared to selected control SNVs as described in the Supplementary methods. For each SNV, the following were examined: 1) the usage of alternative polyadenylation (APA), 2) mRNA extension, 3) the effect of the SNV on gene expression. The usage of APA and the extension of mRNA due to the SNV was examined on the UCSC Genome Browser. The gene expression analyses of samples with and without the SNVs were compared by TPM values of RNA sequencing data using ANOVA. Usage of APA was compared to baseline APA for the gene. A gene was considered to have baseline APA when a second peak was observed at a non-pPAS, in ≥10% of control samples, with the peak height ≥10% of the height of the primary peak in the UCSC Genome Browser. In cases where peaks overlapped two identified PAS, that peak was counted as a single peak.

Known allele-specific expression variants and eQTLs were determined using GTEx data. The data used for the analyses described in this manuscript were obtained from: the GTEx Portal (https://gtexportal.org/home/datasets) on 04/30/2018 and dbGaP accession number phs000424.v8.p2 on 08/20/2019. eQTLs were retrieved from the GTEx Portal v.8 https://storage.googleapis.com/gtex_analysis_v8/single_tissue_qtl_data/GTEx_Analysis_v8_eQTL.tar) and were intersected with hexamers using bedtools. WASP-corrected ASE expression matrices were downloaded from dbGAP.

Assessing deleteriousness of predominant hexamer variants using *in silico* tools

To assess the potential pathogenicity of variants in our predominant hexamers compared to variants in control regions, we compared the score distributions from FATHMM-MKL [14] and CADD [15, 16] *in silico* prediction tools. The control regions were defined as secondary hexamers, strong (AATAAA and ATTAAA) hexamers upstream of non-pPAS, other hexamers upstream of non-pPAS, and the trimer controls as described in the Methods (Figure 4A). Each group was filtered and lifted over as described in the Methods. Control hexamers that overlapped the regions of pPAS and predominant hexamers were removed. Phred-scaled CADD v1.6 and non-coding FATHM-MKL scores were retrieved using Ensembl VEP v103 [17]. Every possible combination of SNVs at each unique position was included.

To assess the ability of FATHMM-MKL and CADD tools to predict the pathogenicity of a predominant hexamer variant, we defined a set of pathogenic/likely pathogenic predominant hexamer variants and non-pathogenic predominant hexamer variants and compared the scores by calculating the likelihood ratio. Predominant hexamer variants listed as disease mutations in HGMD [4] v.2021.3 were classified for pathogenicity as described below (Table 1). Ten variants that attained a classification of pathogenic/likely pathogenic without the use of bioinformatic evidence (PP3) were used in this analysis. Non-pathogenic predominant hexamer variants were defined as variants in predominant hexamers in gnomAD v 2.1.1 with MAF >1%, excluding variants with MAF >50%. As FATHMM-MKL predicts deleteriousness scores for SNVs only, eight SNVs (excluding two indel variants) were included for non-coding FATHMM-MKL, and all ten SNVs and indels were included for CADD prediction scores.

Determination of clinically significant genes

HGMD release 2020.1 variants (https://my.qiagendigitalinsights.com/bbp/ last accessed April 6, 2020), OMIM [18] (https://omim.org/downloads), and CGD [19] (https://research.nhgri.nih.gov/CGD/download/CGD.txt) were downloaded and converted to HGNC IDs. The list of clinically significant genes included OMIM genes with an inheritance pattern and phenotype, HGMD genes with a ‘DM’ variant annotation, and all CGD genes.

Classification of variant pathogenicity using ACMG/AMP pathogenicity criteria

Hexamer variants designated as “DM, Disease Mutations” in HGMD were classified using our adaptation of the American College of Medical Genetics and Genomics and Association for Molecular Pathology (ACMG/AMP) rules for pathogenicity classifications, Table 1. Specifications to the ACMG/AMP criteria can be found in Table 2. We determined that several criteria did not apply to hexamer variants including PVS1 for loss of function variants, PS1 for same amino acid changes, PM4 for protein length changes, PP2 for missense variants in a gene with a low rate of missense variation, BP1 for a missense variant in a gene where loss of function is the known mechanism of disease, BP3 for in-frame insertions or deletions, and BP7 for synonymous variants. Several other criteria were specified for variants in hexamers. For PS3/BS3, functional studies that showed reduced RNA or protein levels in patient cells can be used as evidence to support pathogenicity. For genes where haploinsufficiency is a known mechanism of disease, PS3_moderate can be awarded when three or more heterozygous cell lines from unrelated individuals show a >25% reduction in mRNA and/or protein levels as compared to wild-type; PS3_Supporting can be awarded when one or two heterozygous cell lines from unrelated individuals show a >25% reduction in mRNA and/or protein levels as compared to wild-type. If protein and/or mRNA levels in cell lines from two unrelated individuals are shown to be comparable to wild-type, BS3_Moderate can be awarded. If protein and/or mRNA levels in a cell line from a single individual is shown to be comparable to wild-type, BS3_Supporting can be awarded. For PM5, typically used for missense variants in the same codon, it was determined that a variant in a hexamer where a previous variant had been classified as pathogenic can be awarded PM5 at a supporting level assuming the new hexamer sequence is predicted to have equivalent or less polyadenylation activity as compared to the previously classified pathogenic variant (See Table 3). Insertion and deletion variants that recreate a variant hexamer should be interpreted the same as single nucleotide variants, insertion and deletion variants that disrupt the hexamer can be awarded PM5_Supporting. PM1, presence in a well-established functional domain, is awarded at supporting for [LES: ??] variants in predominant hexamers. PP3 can be awarded for variants with a CADD score ≥10. A CADD score <5 can be used in support of benign status, BP4. Recent guidance by ClinGen has focused on setting criteria strength levels using likelihood ratios based on representative benign and pathogenic variants. An adequate truth set of variants are not available for hexamer variants so these suggested criteria should in time be revisited as more pathogenic hexamer variants are identified.

Statistical analyses and liftover

Statistical analyses (Mann-Whitney U test, Kolmogorov-Smirnov test, ANOVA) and graph generation was performed using R v.3.6.2 [20]. Box plots, violin plots, and Q-Q plots were generated using R library ggplot2 v.3.3.1 [21]. Likelihood ratios were calculated as described [22]. Liftover was performed using Crossmap v.0.5.4 [23].

Assessment of associated phenotype in individuals with predominant hexamer variants

Exome data were available from 1,514 ClinSeq® participants and were assessed for rare variants in predominant hexamers (<0.2% in gnomAD popmax, 194 variants). Variants were further filtered so as to remove variants in genes primarily associated with disorders demonstrating autosomal recessive inheritance (69 variants), variants in genes associated with severe disease unlikely to be present in the cohort (32 variants), variants common in the cohort (four variants) and variants in genes with limited evidence for loss of function as a mechanism of disease (54 variants). Thirty-five variants were determined to be variants of interest. The personal and family history of participants with a variant of interest were assessed for associated disease.

RESULTS

Selection of pPAS and predominant hexamers

The workflow for identifying pPAS and their associated predominant hexamers is outlined in Figure 1. To identify pPAS, we used all PAS in PolyASite 2.0 as the starting point. This database provides a comprehensive list of PAS identified by performing a meta-analysis of 29 3' end sequencing studies. The total number of unique PAS after filtering and re-annotating was 280,148. The average number of PAS per protein-coding gene was 14.2, with 85% of protein coding genes having more than one.

We used relative TPM values to assess the usage of each PAS for all sites identified in a gene model. For genes that had >50% usage of one site, that PAS was defined as the pPAS. Most PAS (>77%) had <1% usage (Figure S1, Additional file 1). Across 18,580 protein coding genes, 15,767 pPAS were identified in 16,004 genes and 197 of these were associated with more than one gene (Figure 1). For 2,576 protein coding genes, no pPAS was identified (Figure S1, Additional file 1).

Next, we aimed to identify a predominant hexamer(s) for each pPAS. For 15,767 pPAS, 45% had more than one candidate hexamer, with an average of 2.5 candidate hexamers for each pPAS (Figure 2A). To identify the predominant hexamers, we considered both the strength of the hexamers and the proximity of the hexamers to the pPAS [11, 24].

In total, 15,212 predominant hexamers were identified for 15,165 pPAS and no hexamers were identified for 602 pPAS. We identified the canonical AATAAA hexamer motif as the most common (61.6%) with the next most common ATTAAA being 15%. The remaining 23.4% of the predominant hexamers comprised one of 16 other hexamer motifs (Figure 2B).

Of 5,376 clinically important genes, we identified pPAS for 4,532 genes (84%) (Figure 3A-B). For the remaining 844 genes, 744 (88%) had an average of >30 PAS per gene with an average highest PAS usage of 37.6%. The remaining 100 genes did not have any identified PAS.

Predominant hexamers are less tolerant to variation than control sequences

We postulated that predominant hexamers would be critical for gene function and hence less tolerant to population variation. To test this, we compared the occurrence of variants in gnomAD in predominant hexamers versus non-predominant hexamers and versus upstream control sequences (see Methods). There were 7,150 predominant hexamers, 141,561 non-predominant hexamers and 5,144 sets of trimers included in this analysis. Q-Q plots showed significant deviation from the expected distributions (Figure 4B-C). The distribution of allele frequencies in the predominant hexamers were significantly lower than in the non-predominant hexamers (*Mann-Whitney U test p-value*$<2.22 \times 10^{-16}$*, Kolmogorov-Smirnov test p-value*$=1.08\text{x}10^{-09}$) and the trimer controls (*Mann-Whitney U test p-value*$=1.93 \times 10^{-09}$*, Kolmogorov-Smirnov test p-value*$=1.36\text{x}10^{-07}$). This indicated that the predominant hexamers are likely to be functionally important.

Predominant hexamer variants show higher deleteriousness scores

We postulated that a non-coding *in silico* metapredictor would yield significantly higher deleteriousness scores for variants in predominant hexamers compared to non-predominant hexamers and trimer controls. To test this, we compared CADD and non-coding FATHMM-MKL scores of all possible combinations of SNVs occurring in predominant hexamers, secondary hexamers, other hexamers, other strong hexamers, and control trimers as defined in the Methods (Figure 4A). Other 'strong' hexamers were a subset of other hexamers that only included AATAAA and ATTAAA hexamers. This was to ensure that the deleteriousness scores for the strong hexamers were not diluted due to the inclusion of other weak hexamers. After quality control, 7,148 predominant hexamers, 3,401 secondary hexamers, 134,104 other hexamers, 50,710 other strong hexamers, and 5,142 trimer control sets were analyzed.

Both CADD and FATHMM-MKL scores were higher for variants in predominant hexamers when compared to each of the control groups (p-values for comparisons had Mann-Whitney U test and Kolmogorov-Smirnov test p-values $<2.22 \times 10^{-16}$) (Figure 4D-E). The median CADD scores for predominant hexamers were 14.2, whereas the median scores for secondary hexamers, other hexamers, other strong hexamers, and trimer controls were 12.7, 6.4, 6.4 and 7.4, respectively. This pattern was also observed for FATHMM-MKL predictions, with scores >0.5 suggesting deleteriousness and scores <0.5 suggesting neutral effect [14]. The median scores for variants in predominant hexamers were 0.93, suggesting deleteriousness of variants. The median scores for secondary hexamers, other hexamers, other strong hexamers, and the trimer controls were 0.89, 0.20, 0.20 and 0.23, respectively. Low CADD and FATHMM-MLK scores for other hexamers, other strong hexamers, and the trimer controls suggest neutral effects of the variants in these regions. Interestingly, both predictions showed that the secondary hexamers in pPAS may be functionally important.

Effects of predominant hexamer variants in RNAseq data

To understand the effects of variants occurring in predominant hexamers, exome sequencing data from 76 individuals were examined for variants in 11,656 predominant hexamers where the wild-type hexamers were either AATAAA or ATTAAA. A total of 121 predominant hexamer variants were identified. Two types of RNA-sequencing (3' end sequencing and standard RNA sequencing) were performed. RNA sequencing data that passed QC were available for 65 variants (Table S1, Additional file 2). PolyA site usage was determined by 3' end sequencing and analysis of extended RNA products (non-polyadenylated) whereas overall gene expression was determined by RNA-sequencing. For controls, we randomly selected 60 non-hexamer 3' UTR variants from the same dataset.

The UCSC browser was used to view aligned 3' end sequencing data and the presence of APA was manually assessed in the 65 genes with predominant hexamer variants and in 60 genes with non-predominant hexamer 3' UTR variants. Overall, >70% of genes used a single PAS for polyadenylation and APA was observed in <30% of the genes (Figure 5A); this did not differ between the two sets of genes. This suggested that the majority of genes use a single PAS. The changes in PAS usage were determined for the 65 predominant hexamer variants versus 60 control 3' UTR variants (Figure S8, Additional File 1). Twenty-five genes (25/65, 38%) with predominant hexamer variants showed changes in APA (Figure 5B-C, Table S1, Additional file 2). The changes in polyadenylation were observed irrespective of whether a gene typically used a single

PAS or APA was common for the gene. By contrast, no changes in polyadenylation were observed in the 60 controls (Figure 5B-C). We conclude that changes in polyadenylation are associated with predominant hexamer variants.

The effect of hexamer variation on RNA transcript extension and gene expression was also determined for these genes and variants. Extended RNA transcripts were observed for eight predominant hexamer variants (8/65, 12%) but in none of the genes with control 3' UTR variants (0/60) (Figure 5D). Interestingly, these extended transcripts were not supported by additional PAS in 3' end sequencing data. This suggested that the predominant hexamer variants resulted in loss of polyadenylation, with the longer non-polyadenylated transcript being subject to nonsense-mediated decay or degradation by miRNA.

No change of expression level was noted for any of the predominant hexamer variants in the dataset (Figure 5E). However, the predominant hexamer variants identified in *MS4A6A, TP53, TRAPPC3, SHISA5, ATP5F1E, ERAP1*, *DHRS7, IK, SPTLC1* and *TMEM176A* were previously identified eQTL variants and/or variants showing allele-specific expression [25]. This suggests that our cohort size of n=76 may have been underpowered to detect small expression changes from eQTL variants.

Of the 65 variants identified in the 76 samples with RNA sequence data, only three variants were in genes associated with a disorder inherited in an autosomal dominant pattern (*PMP22, SPTLC1* and *TP53*) and all three variants were too common in gnomAD (>1% popmax) to be classified as pathogenic.

Prediction of pathogenic predominant hexamer variants using in silico tools

Next, we sought to assess if FATHMM-MKL and CADD can be used to discriminate pathogenic variants versus common variants in predominant hexamers. We collected 32 variants from the HGMD database that were considered to be hexamer variants; one variant in *IGF1* was found not to be in the vicinity of the 3' end of the mRNA and was therefore removed from all analyses. Each variant was manually examined for evidence of pathogenicity (see methods). Ten variants had sufficient evidence to be classified as likely pathogenic or pathogenic without considering PP3 (bioinformatic prediction of pathogenicity). For control variants, gnomAD variants in predominant hexamers regions with MAF >1% were included.

Although the number of pathogenic variants is small, we observed higher CADD and FATHMM-MKL scores for pathogenic variants compared to the control variants (Figure 4F-G). Using a CADD score of ≥10, the likelihood ratio was 2.5 with 100% sensitivity and 60% specificity. For FATHMM-MKL, the likelihood ratio at a score of ≥0.7 was 2.86 with 100% sensitivity and 65% specificity. This suggested that *in silico* prediction can aid in the identification of pathogenic variants at these score thresholds.

Classification of reported hexamer variants according to the adapted ACMG/AMP criteria

Thirty-one hexamer variants identified in HGMD were annotated with evidence that supported pathogenicity according to the classification criteria specified for polyadenylation variants (*IGF1* variant not considered owing to insufficient evidence of a hexamer). Nine of the 31 variants had sufficient evidence to support a likely pathogenic classification whilst four had sufficient evidence to support a pathogenic classification; eight of these variants were reported in multiple unrelated cases. Of the 18 variants that had insufficient evidence, the majority were single case reports with limited case data. Nine variants would move from VUS to likely pathogenic with the addition of moderate functional data.

Phenotypes in individuals with predominant hexamer variants

Thirty-four individuals in this dataset had predominant hexamer variants considered to be variants of interest based on their frequency in gnomAD, inheritance pattern of associated disease and disease mechanism. Analysis of personal and family history taken at the time of enrollment did not identify associated phenotypes for 31 of the participants. For three participants, potentially related conditions were evident in the personal or family history. An individual with a variant in the *ALK* gene, associated with susceptibility to neuroblastoma, had a benign brain tumor. An individual with a variant in the *NBN* gene, associated with uterine smooth muscle tumors and ovarian cancer (among other phenotypes), had a family history of uterine fibroids and uterine cancer. Finally, an individual with a variant in the *ABCA1* gene, associated with abnormal cholesterol, had high cholesterol.

DISCUSSION

The underrepresentation of hexamer variants associated with Mendelian disease is likely due to several interrelated factors. First, functionally relevant hexamers are not well-defined for clinical or research laboratories to interrogate. Second, many exomes do not include 3’ UTR regions. Finally, it may be the case that few genes are biologically sensitive to the changes caused by hexamer variants. While this may be a truly uncommon class of pathogenic genetic variation, we strongly suspect that there are more than 14 genes that are susceptible to this type of variant. Under-ascertainment could be mitigated by genome sequencing or with the addition of 3’ UTR capture probes in exome sequencing. However, the lack of well-defined functionally relevant hexamers hinders the discovery of pathogenic hexamer variants and holds back our understanding of the role of these variants in disease. Here we have defined a set of predominant hexamers as candidates for Mendelian disease-associated variation, reasoning that predominant hexamers are more likely to be functionally relevant. We have shown predominant hexamers to be constrained, have higher deleteriousness scores by *in silico* prediction tools, and demonstrated that variants in these predominant hexamers in the ClinSeq® cohort contribute to APA and mRNA extension. The key resource we have provided is a set of predominant hexamers that can be interrogated in both clinical and research sequencing. We have also suggested adaptations of the ACMG criteria to support the pathogenicity classification of hexamer variants.

Several databases provide information on PAS and hexamers without consideration of their importance or functionality [5-8]. While identification of all potential PAS and hexamers is useful for understanding the mechanism of RNA processing and regulation, for the purpose of interrogation of exome and genome sequence data for association with Mendelian diseases, focusing on highly relevant hexamers is required to facilitate the assessment of variant pathogenicity. To define a list of hexamers for clinical interrogation we used the PolyASite 2.0 database for identification of all PAS and corresponding hexamers and then used a 50% threshold to define a set of pPAS and hexamers. The large tissue diversity in the PolyASite 2.0 database (221 samples across multiple tissue types) allowed the identification of pPAS and related hexamers across tissue types for the majority of protein-coding genes. We suggest that this set of highly used PAS, and the corresponding hexamers, are likely to be functionally relevant and therefore useful for clinical interrogation. Indeed, of the 31 variants identified in HGMD associated with Mendelian diseases, 30 were in predominant hexamers. This list of predominant hexamers should be considered tissue-agnostic with the understanding that hexamers that are tissue-specific may be underrepresented. However, work by Shulman et al. showed that 69% of polyadenylation QTLs affected more than one tissue with consistent effects across tissues, which may mitigate this limitation [9]. It is likely, however, that some biologically relevant PAS and hexamers were missed in our analyses because they did not meet our threshold of >50% use (2,576 or 12.7% of protein coding genes) or genes were not expressed in the tissues in the PolyASite 2.0 database (1,664 or 8.2% of protein coding genes). The former could occur if the gene has high APA. As our understanding of gene regulation increases, these PAS will be further investigated.

The predominant hexamers we identified were significantly constrained with higher deleteriousness scores compared to control sequences supporting the identification of functionally important PAS using TPM across datasets. Unexpectedly, we also observed higher constraint and deleteriousness scores for secondary hexamers compared to other control sequences. This suggested that sequences beyond the predominant hexamers near pPAS are conserved and functionally important. It is possible that these secondary hexamers overlap motifs that play a role in polyadenylation [26-28].

To understand the impact of variants in predominant hexamers, we analyzed exome and RNA sequence data for the effect of such variants on RNA expression and processing. The variants we identified in predominant hexamers did not alter gene expression. However, 45% of variants resulted in either APA or extended RNA suggesting that these predominant hexamers are important in RNA cleavage and polyadenylation. Although it is surprising that variation in APA and/or RNA cleavage were not accompanied by changes in RNA expression, RNA expression may not be necessary for clinically relevant perturbation of gene function. Indeed, a hexamer variant in *STUB1* has been shown to affect protein levels without affecting gene expression and polyadenylation [29]. It is also possible that changes in gene expression were present but were not large enough to be recognized in this study based on the small sample size, as our study was underpowered to detect eQTLs. The remaining 55% of predominant hexamer variants in the sample set did not have a noticeable effect on RNA processing. This may have been due to the creation of an alternative functional hexamer sequence as seen with the *SHISA5* gene (Figure S2, Additional File 1). Of the 40 variants that resulted in no apparent changes, 31 created an alternate hexamer sequence.

The functional relevance of predominant hexamers is supported by our findings that all but one of the 31 ‘DM’ hexamer variants reported in HGMD reside in a predominant hexamer (the variant in *BMP1* is not in the predominant hexamer). A previously proposed set of hexamers for mining clinical variants captured all 31 ‘DM’ hexamer variants reported in HGMD [10], but included 229,014 hexamers. While such a brute force search of hexamer variants may increase yield, manually interpreting a very large number of variants is costly and may not be feasible in a clinical setting. Although such an inclusive list can be useful for advancing our understanding of the polyadenylation mechanism, a more focused list of functionally important and hence clinically relevant hexamers is required. Our approach identified 15,767 pPAS from >560,000 PAS in the PolyASite 2.0 database, removing ~97% of PAS that were less likely to be of clinical importance. Our list of predominant hexamers is efficient for researchers focused on diseases with unexplained etiology to interrogate for clinically relevant variants in their disease cohorts.

To identify predominant hexamer variants with clinical effect, we analyzed exome data from 1,514 individuals. Thirty-five predominant hexamer variants were identified in genes where loss of function variants might contribute to disease. Personal and family history collected at the time of enrollment did not support a contribution to the participant's phenotype for 31 individuals. However, three individuals had phenotypes that could be related to the gene in question. As polyA variants might be expected to be hypomorphic rather than being associated with a complete loss of function, it is possible that resultant phenotypes may be less penetrant or less severe. Looking at the hexamer variants in HGMD, over half of the variants listed are in globin genes (17), where the level of gene products are precisely regulated. Of the remaining 14 variants, eight are on the X-chromosome and may be more sensitive to partial loss of function. We suggest that this list of predominant hexamers may be especially useful in interrogating patient cohorts where phenotypes suggest specific genetic etiologies and the causative variants have yet to be identified, including individuals with disorders inherited in an autosomal recessive manner, where the affected individual has only a single identified pathogenic variant.

To support the pathogenicity classification of hexamer variants, we have suggested specifications [LES: specific modifications…?] to the ACMG/AMP pathogenicity criteria. The ACMG/AMP pathogenicity criteria were intended to be generally applicable to all genes. However, some criteria are not going to be applicable to non-coding variants (PVS1, PS1, PM4, PM5, PP2). As pathogenicity classification is dependent on the number of criteria that can be applied to a given variant, excluding criteria from consideration reduces our ability to classify a variant as pathogenic or likely pathogenic. It is useful in the case of non-coding variants to consider whether certain criteria can be amended rather than dropped. We have suggested a specification for PM5 (same amino acid, specified as same hexamers) that recognizes prior evidence of a different pathogenic variant in the hexamers. Identification of a pathogenic variant in a hexamer supports the importance of that hexamer in gene function. PM1 and PP3 were also specified. Although limited pathogenic variants were available to set a cutoff for a CADD score, comparison of CADD scores of common hexamer variants to pathogenic hexamer variants suggested a cutoff of ≥10 which was implemented. However, as PP3 for noncoding variants considers conservation, which is related to presence in a functional domain (PM1), it was determined that PP3 and PM1 should not both be applied together at full strength until further evidence was available to correctly weight these two criteria for hexamers. We have suggested using PM1 at supporting strength for variants in predominant hexamers. PP3 can be applied for variants with a CADD score ≥10.

Using the ACMG criteria to classify these variants, 13 were classified as pathogenic or likely pathogenic. All four variants classified as pathogenic were awarded PS4 at full strength for multiple unrelated affected individuals with the variant. Twenty-one variants were supported by a single case. Case data including segregation (six variants), phenotype specific for gene (22 variants) and presence with a second pathogenic variant in recessive disease (ten variants) were important in classifying variants as likely pathogenic or pathogenic. Ten variants had functional studies of patient cells that contributed toward their classification. Whilst additional functional data could move eleven variants from VUS to likely pathogenic, the opportunity for *ex vivo* functional studies is limited with many variants being supported by a single case. Thirty variants were present in the predominant hexamers and received PM1_Supporting for presence in a functional domain. Finally, 30 variants received PP3 for a CADD score ≥10. We suggest that classification of hexamer variants as pathogenic or likely pathogenic is limited by patient data and could be aided by *in vitro* functional studies not typically pursued for hexamer variants.

Using the hexamer variants that were classified as pathogenic or likely pathogenic based on our adapted ACMG criteria, we tested CADD and FATHMM-MKL to determine whether *in silico* predictions can predict pathogenic hexamer variants. We suggest that deleteriousness scores with the specified thresholds (CADD score=10, FATHMM-MKL score=0.70) support pathogenicity. Although a CADD score of 10 appears low in comparison to the threshold typically used for coding variants, unlike FATHMM-MKL non-coding scores that only considers non-coding variant properties, CADD considers both coding and non-coding variant properties (e.g., amino acid change), and scores are measured with respect to deleteriousness of all variants [15, 16]. Thus, we expect that the CADD scores for pathogenic hexamer variants will be lower than CADD scores for pathogenic coding variants and will therefore require different thresholds compared to coding variants to be considered as evidence for pathogenicity. When classifying variants using the adapted ACMG criteria, we suggest using a CADD score $\geq$10 in support of pathogenicity. As more pathogenic hexamer variants are identified, the use of CADD versus other bioinformatic predictors, and the exact threshold to be used, can be reassessed.

In summary, we have identified a set of 15,212 hexamers that are candidates for variation that may be associated with Mendelian genetic disorders. These sites are supported by high polyadenylation usage and are conserved and population constrained. We encourage researchers and clinicians who have access to genome sequencing data to evaluate these sites for disease association using the resources we have provided.

Data availability

The refined list of pPAS and corresponding predominant hexamers are included in the supplemental material with this publication, publicly available for download in BED file format on our github page (https://github.com/BieseckerLab/PolyAProject), and for view through UCSC public track hubs (insert link here). ClinSeq® sequencing data are available on dbGaP (phs000971.v3.p1).

Acknowledgements

The authors thank the members of the ClinGen Sequence Variant Interpretation Working group for productive discussions on the variant pathogenicity criteria.

Supplementary Material

SupplementaryData01.docx contains supplementary methods as well as additional data on hexamers retrieved from PolyA Site 2.0 and hexamer variants found in ClinSeq®.

SupplementaryTable02.xlsx contains additional data on clinical genes, hexamer variants found in ClinSeq®, as well as a list of pPAS and hexamers in hg19 and hg38.

References

1. Colgan, D.F. and J.L. Manley, *Mechanism and regulation of mRNA polyadenylation.* Genes Dev, 1997. **11**(21): p. 2755-66.
2. Proudfoot, N., *Poly(A) signals.* Cell, 1991. **64**(4): p. 671-4.
3. Keller, W., et al., *Cleavage and polyadenylation factor CPF specifically interacts with the pre-mRNA 3' processing signal AAUAAA.* EMBO J, 1991. **10**(13): p. 4241-9.
4. Stenson, P.D., et al., *The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine.* Hum Genet, 2014. **133**(1): p. 1-9.
5. Herrmann, C.J., et al., *PolyASite 2.0: a consolidated atlas of polyadenylation sites from 3' end sequencing.* Nucleic Acids Res, 2020. **48**(D1): p. D174-d179.
6. You, L., et al., *APASdb: a database describing alternative poly(A) sites and selection of heterogeneous cleavage sites downstream of poly(A) signals.* Nucleic Acids Res, 2015. **43**(Database issue): p. D59-67.
7. Wang, R., et al., *PolyA_DB 3 catalogs cleavage and polyadenylation sites identified by deep sequencing in multiple genomes.* Nucleic Acids Res, 2018. **46**(D1): p. D315-D319.
8. Muller, S., et al., *APADB: a database for alternative polyadenylation and microRNA regulation events.* Database (Oxford), 2014. **2014**.
9. Shulman, E.D. and R. Elkon, *Systematic identification of functional SNPs interrupting 3'UTR polyadenylation signals.* PLoS Genet, 2020. **16**(8): p. e1008977.
10. Chen, M., et al., *Systematic evaluation of the effect of polyadenylation signal variants on the expression of disease-associated genes.* Genome Res, 2021. **31**(5): p. 890-899.
11. Sheets, M.D., S.C. Ogg, and M.P. Wickens, *Point mutations in AAUAAA and the poly (A) addition site: effects on the accuracy and efficiency of cleavage and polyadenylation in vitro.* Nucleic Acids Res, 1990. **18**(19): p. 5799-805.
12. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.
13. *Picard*. Available from: https://broadinstitute.github.io/picard/.
14. Shihab, H.A., et al., *An integrative approach to predicting the functional effects of non-coding and coding sequence variation.* Bioinformatics, 2015. **31**(10): p. 1536-43.
15. Kircher, M., et al., *A general framework for estimating the relative pathogenicity of human genetic variants.* Nat Genet, 2014. **46**(3): p. 310-5.
16. Rentzsch, P., et al., *CADD-Splice-improving genome-wide variant effect prediction using deep learning-derived splice scores.* Genome Med, 2021. **13**(1): p. 31.
17. McLaren, W., et al., *The Ensembl Variant Effect Predictor.* Genome Biol, 2016. **17**(1): p. 122.
18. Medicine, M.-N.I.o.G., *Online Mendelian Inheritance in Man, OMIM®*. 2020.
19. Solomon, B.D., et al., *Clinical genomic database.* Proc Natl Acad Sci U S A, 2013. **110**(24): p. 9851-5.

20. Team, R.C., *R: A Language and Environment for Statistical Computing*. 2017, R Foundation for Statistical Computing: Vienna, Austria.
21. Wickham, H., *ggplot2: Elegant Graphics for Data Analysis*. 2016: Springer-Verlag New York.
22. Pejaver, V., et al., *Evidence-based calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for clinical use of PP3/BP4 criteria.* bioRxiv, 2022: p. 2022.03.17.484479.
23. Zhao, H., et al., *CrossMap: a versatile tool for coordinate conversion between genome assemblies.* Bioinformatics, 2014. **30**(7): p. 1006-7.
24. Gruber, A.J., et al., *A comprehensive analysis of 3' end sequencing data sets reveals novel polyadenylation signals and the repressive role of heterogeneous ribonucleoprotein C on cleavage and polyadenylation.* Genome Res, 2016. **26**(8): p. 1145-59.
25. Consortium, G.T., *The Genotype-Tissue Expression (GTEx) project.* Nat Genet, 2013. **45**(6): p. 580-5.
26. Darmon, S.K. and C.S. Lutz, *Novel upstream and downstream sequence elements contribute to polyadenylation efficiency.* RNA Biol, 2012. **9**(10): p. 1255-65.
27. Nunes, N.M., et al., *A functional human Poly(A) site requires only a potent DSE and an A-rich upstream sequence.* EMBO J, 2010. **29**(9): p. 1523-36.
28. Hall-Pogar, T., et al., *Alternative polyadenylation of cyclooxygenase-2.* Nucleic Acids Res, 2005. **33**(8): p. 2565-79.
29. Turkgenc, B., et al., *STUB1 polyadenylation signal variant AACAAA does not affect polyadenylation but decreases STUB1 translation causing SCAR16.* Hum Mutat, 2018. **39**(10): p. 1344-1348.

Table 1. Pathogenicity Classifications of 31 Known Polyadenylation Signal Hexamer Disease-Associated Variants in 14 Genes

Gene	**WT Hexamers**	**Hexamers cDNA Position**	**GRCh37 Position & Variation**	**Descriptor**	**Variant hexamers**	**CADD score**	**HGMD ID**	**Classifi-cation**	**Pathogenicity Criteria**
BMP1	AGTAAA	c.*239_*244	Chr8:22,058,957T>C	NM_001199.4:c.*241T>C	AGCAAA	17.97	CR150372	Likely Path	PS3_Mod, PS4_Mod, PM3_Sup, PP3, PP4
F9	AATAAA	c.*1365_*1370	ChrX:138645598A>G	NM_000133.4: c.*1368A>G	AATGAA	2.70	CR005437	VUS	PS4_Sup, PM1_Sup, PP4, BP4
FOXP3	AATAAA	c.*873_*878	ChrX:49106919T>C	NM_014009.4:c.*876A>G	AATGAA	13.02	CR014834	Likely Path	PS4_Mod, PM1_Sup, PP1_Strong, PP3, PP4
			ChrX:49106917T>C	NM_014009.4:c.*878A>G	AATAAG	12.77	CR097218	VUS	PS3_Sup, PS4_Sup, PM1_Sup, PP3, PP4
HBA2	AATAAA	c.*89_*94	Chr16:223690-223691del	NM_000517.4:c.*91_*92del	AAAAGT	12.61	CD2026691	VUS	PS4_Sup, PM1_Sup, PP3, PP4
			Chr16:223691A>G	NM_000517.4:c.*92A>G	AATGAA	11.73	CR920785	Likely Path	PS4_Mod, PM1_Sup, PM3_Sup, PM5_Sup, PP1_Mod, PP3, PF
			Chr16:223692-223693del	NM_000517.6:.*93_*94del	AATAGT	11.84	CD941949	Likely Path	PS4_Mod, PM1_Sup, PM3, PM5_Sup, PP3, PP4
			Chr16:223693A>G	NM_000517.6:c.*94A>G	AATAAG	12.77	CR830007	Pathogenic	PS3_Sup, PS4, PM1_Sup, PM3, PP3, PP4
			Chr16:223693A>C	NM_000517.6:c.*94A>C	AATAAC	12.24	CR106042	Likely Path	PM1_Sup, PM3, PM5_Sup, PP3, PP4
HBB	AATAAA	c.*108_*113	Chr11:5246720T>G	NM_000518.5:c.*108A>C	CATAAA	15.14	CR016252	VUS	PS4_Sup, PM1_Sup, PP3
			Chr11:5246720T>C	NM_000518.5: c.*108A>G	GATAAA	15.27	CR127145	VUS	PS4_Mod, PM1_Sup, PP3, PP4
			Chr11:5246718A>T	NM_000518.5:c.*110T>A	AAAAAA	14.07	CR045224	VUS	PS4_Sup, PM1_Sup, PM5_Sup, PP3, PP4
			Chr11:5246718A>G	NM_000518.5:c.*110T>C	AACAAA	14.17	CR850010	Pathogenic	PS3_Sup, PS4, PM1_Sup, PM3, PM5_Sup, PP1, PP3, PP4
			Chr11:5246718A>C	NM_000518.5:c.*110T>G	AAGAAA	14.03	CR014260	VUS	PS4_Sup, PM1_Sup, PM5_Sup, PP3, PP4
			Chr11:5246717-5256718del	NM_000518.5:c.*110_*111del	AAAAAA	13.15	CD951735	VUS	PS4_Sup, PM1_Sup, PM5_Sup, PP3, PP4
			Chr11:5246714-5246718del	NM_000518.5:c.*110_*114del	AAAACA	14.11	CD920867	Likely Path	PS3_Sup, PS4_Mod, PM1_Sup, PM3_Sup, PM5_Sup, PP3, PF
			Chr11:5246717T>C	NM_000518.5:c.*111A>G	AATGAA	14.04	CR900265	Pathogenic	PS4, PM1_Sup, PM3, PM5_Sup, PP3, PP4
			Chr11:5246716T>C	NM_000518.5:c.*112A>G	AATAGA	14.73	CR900266	Likely Path	PM1_Sup, PM3, PM5_Sup, PP3, PP4
			Chr11:5246716T>A	NM_000518.5:c.*112A>T	AATATA	14.64	CR057232	VUS	PS4_Sup, PM1_Sup, PP3, PP4
			Chr11:5246715T>C	NM_000518.5:c.*113A>G	AATAAG	13.23	CR880076	Pathogenic	PS3_Sup, PS4, PM1_Sup, PM3, PP3, PP4
HBD	AATAAA	c.*106_*111	Chr11:5254085T>A	NM_000519.4: c.*109A>T	AATTAA	13.00	CR109506	VUS	PS4_Sup, PM1_Sup, PP3, PP4
IL2RG	AATAAA	c.*303-*308	ChrX:70327278T>C	NM_000206.3: c.*308A>G	AATAAG	15.05	CR0910465	VUS	PS3_Sup, PS4_Sup, PM1_Sup, PP3, PP4
INS	AATAAA		Chr11:2181023T>C	NM_000207.2:c.*59A>G	AATAAG	12.87	CR101141	VUS	PS3_Sup, PS4_Sup, PM1_Sup, PP3
ITGA2B	AATAAA	c.*163_*168	Chr17:42449567A>G	NM_000419.5c.*165T>C	AACAAA	12.82	CR153724	VUS	PS4_Sup, PM1_Sup, PP3
NAA10	AATAAA	c.*38_*43	ChrX:153195400T>C	NM_003491.4:c.*39A>G	AGTAAA	12.95	CR1913378	Likely Path	PS3_Sup, PS4_Sup, PM1_Sup, PP1_Mod, PP3
			ChrX:153195401T>C	NM_003491.4:c.*40A>G	GATAAA	14.35	CR1913377	VUS	PS4_Sup, PM1_Sup, PP3
			ChrX:153195397T>C	NM_003491.4:c.*43A>G	AATAGA	14.18	CR1913376	Likely Path	PS3_Sup, PS4_Sup, PM1_Sup, PP1_Strong, PP3
RNASEH2C	AATAAA	c.*78_*83	Chr11:65487176T>C	NM_032193: c.*78A>G	GATAAA	16.68	CR2027419	VUS	PS4_Sup, PM1_Sup, PP3
STUB1	AATAAA	c.*238_*243	Chr16:732729T>C	NM_005861.3c.*240T>C	AACAAA	14.22	CR1815154	VUS	PS3_Sup, PS4_Sup, PM1_Sup, PP1, PP3
UROD	AATAAA	c.*57_*62	Chr1:45481230-45481231del	NM_000374.4:c.*62_*63del	AATAAG	12.67	CD122215	VUS	PS4_Sup, PM1_Sup, PP3
GLA	ATTAAA	c.1272_1277	ChrX:100652809-810del	NM_000169.2:c.1277_1278del	ATTAGA	23.7	CD031841	VUS	PS3_Sup, PS4_Sup, PM1_Sup, PP3, PP4

Table 2. Modified ACMG Criteria for Polyadenylation Signal Hexamer Variant Pathogenicity Classification (see supplemental information for full explanations).

Criteria	Criteria Description	Specification
Pathogenic Criteria		
VERY STRONG CRITERIA		
PVS1	Loss of function allele.	Not Applicable
PS2/PM6_ Very Strong	Each proven *de novo* case, 2 points, each assumed *de novo* case, 1 point, ≥8 points	Strength[a]
STRONG		
PS1	Same amino acid change as a previously established pathogenic variant irrespective of nucleotide change	Not Applicable
PS2/PM6_ Strong	Each proven *de novo* case, 2 points, each assumed *de novo* case, 1 point, a total of 4-7 points	Strength[a]
PS3	Well-established functional studies supportive of a damaging effect on protein function • See PS3_Moderate/PS3_Supporting	Strength[a], Variant type Specific
PS4	Prevalence of the variant in affected individuals significantly increased compared with the prevalence in controls • ≥7 unrelated cases with associated condition. If specifications have been provided by an expert panel case, counting should consider their recommendation for strength. • Popmax MAF in gnomAD <0.0006 • For variants with ≥7 unrelated cases and popmax ≥0.0006, an odds ratio can be calculated to determine strength level, an odds ratio ≥18.3 allows for PS4 to be used at strong [LES: ??]	Strength[a]
PP1_Strong	• Co-segregation with disease in ≥7 reported meioses	Strength[a]
MODERATE		
PM1	Located in a mutational hot spot and/or critical and well-established functional domain • Downgraded to avoid overcounting with PP3.	Strength, Variant type specific
PM2	Absence in gnomAD • Incorporated into PS4, do not consider separately unless ClinGen expert panel specifications for PS4 are used and PM2 is incorporated into those specifications	Not Applicable
PM3	For recessive disorders in *trans* with pathogenic variant • Identified with pathogenic variant in *trans*, phase known • Identified in homozygous state in two unrelated affected individuals	None

	• An individual cannot be counted for both PM3 and PS4	
PM4	Protein length change	Not Applicable
PM5	• Previous missense, see PM5_Supporting	Variant type specific
PS2/PM6_ Moderate	Each proven *de novo* case, 2 points, each assumed *de novo* case, 1 point, a total of 2-3 points	Strength[a]
PS3_Moderate	Well-established functional studies supportive of a damaging effect on protein function • For genes where loss of function is a known mechanism of disease, decreased RNA or protein levels (<75% of WT) in three or more cell lines from unrelated individuals who harbor that variant	Strength[a], Variant type Specific
PS4_Moderate	The prevalence of the variant in affected individuals is significantly increased compared with the prevalence in controls • 2-6 unrelated cases with associated condition. If specifications have been provided by an expert panel case, counting should consider their recommendation for strength. • Popmax MAF in gnomAD <0.0006 • For variants with 2-6 unrelated cases and popmax ≥0.0006, an odds ratio can be calculated to determine strength level, an odds ratio ≥4.8 allows for PS4 to be used at moderate [LES: ??]	Strength[a], Disease-Specific
PP1_ Moderate	• Co-segregation with disease in 5-6 reported meioses	Strength[a]
SUPPORTING		
PP1	Co-segregation with disease in 3-4 reported meioses	Strength[a]
PP2	Missense variant in gene with low rate of benign missense variants	Not Applicable
PP3	Computational evidence suggests impact on gene or gene product • CADD score of >=10	Variant type Specific
PP4	Patient's phenotype or family history is highly specific for a disease with a single genetic etiology	None
PS2/PM6_ Supporting	Each proven *de novo* case, 2 points, each assumed *de novo* case, 1 point, a total of 1 point	Strength[a]
PS3_ Supporting	Well-established functional studies supportive of a damaging effect on protein function • For genes where loss of function is a known mechanism of disease, decreased RNA or protein levels (<75% of WT) in cells from an affected individual harboring the variant	Strength[a], Variant type Specific
PS4_ Supporting	The prevalence of the variant in affected individuals is significantly increased compared with the prevalence in controls • One case with associated condition. If specifications have been provided by an expert panel case, counting should consider their recommendation for strength. • Popmax MAF in gnomAD <0.0006	Strength[a], Disease-Specific

PM1_Supporting	Located in a mutational hot spot and/or critical and well-established functional domain • For variants in predominant hexamers, award PM1 at supporting strength	Strength, Variant type specific
PM3_Supporting	For recessive disorders in *trans* with pathogenic variant • Identified homozygous state in an affected individual	
PM5_Supporting	Single nucleotide variant in a hexamer where a different single nucleotide variant was previously determined to be likely pathogenic • New hexamers must be in lower functional group as predicted by Sheets et al. 1990[b] • Previously established likely pathogenic variant must reach a classification of pathogenicity without PM5	Strength, Variant type Specific
Benign Criteria		
STAND ALONE		
BA1	Allele frequency is >0.05 in any general continental [LES: global?] population dataset of at least 2,000 observed alleles and found in a gene without a gene- or variant-specific BA1 modification. • If specifications have been provided by an expert panel, BA1 should be determined as set by the expert panel for that gene.	Disease-Specific
STRONG		
BS1	Popmax allele frequency greater than expected for the disorder • If specifications have been provided by an expert panel, BS1 should be determined as set by the expert panel for that gene.	Disease-Specific
BS2	Observed in a healthy adult individual for a recessive (homozygous), dominant (heterozygous), or X-linked (hemizygous) disorder with full penetrance expected at an early age.	None
BS3	Well-established functional studies show no damaging effect on protein function • See BS3_Moderate	Strength
BS4	Lack of segregation in family members	None
MODERATE		
BS3_Moderate	Well-established functional studies show no damaging effect on protein function • No reduction in RNA or protein level in three or more cell lines from unrelated individuals who harbor the variant.	Strength[a], Variant type Specific
SUPPORTING		
BP1	Missense variant in a gene for which loss of function is a known mechanism of disease	Not Applicable

BP2	Observed in *cis* with a pathogenic variant in any inheritance pattern	None
BP4	Computational evidence suggests no impact on gene or gene product - CADD score of <5.0	Variant type Specific
BP7	A synonymous (silent) variant for which splicing prediction algorithms predict no impact upon the splice consensus sequence nor the creation of a new splice site AND the nucleotide is not highly conserved [LES: evolutionarily or at a population level?]	Not Applicable
BS3_Supporting	Well-established functional studies show no damaging effect on protein function - No reduction in RNA or protein level in cells from an individual harboring the variant.	Strength[a], Variant type Specific

Key:

[a] For criteria that can be assigned different levels of strength based on evidence, only the highest applicable strength level should be used. For example, if PS4 is met, then PS4_Moderate and PS4_Supporting are not used.

[b] See Sheets et al. 1990. Table attached.

[c] Sequence Variant Interpretation Committee, ClinGen.

[d] Cardiomyopathy Expert Panel.

Table 3. Polyadenylation activity values [LES: ± SD?] for 17 variant hexamers as compared to AAUAAA. Hexamers have been grouped according to polyadenylation activity. PS1/PM5 can be awarded at a moderate level when a variant creates a hexamer that falls into the same (or lower) group as the previously classified variant. Sheets et al. [LES: 1990? Explain why reference is cited]

Sequence	**Polyadenylation (% AAUAAA)**	**Classification Group**
AAUAAA	100	Group 6
AUUAAA	77 ± 4.7	Group 5
AGUAAA	29 ± 8.1	Group 4
CAUAAA	18 ± 6.4	Group 3
UAUAAA	17 ± 3.0	Group 3
ACUAAA	11 ± 6.0	Group 2
GAUAAA	11 ± 1.0	Group 2
AAUACA	11 ± 2.3	Group 2
AAUAUA	10 ± 2.3	Group 2
AAGAAA	6.0 ± 1.0	Group 1
AACAAA	4.0 ± 2.0	Group 1
AAAAAA	4.6 ± 3.7	Group 1
AAUGAA	4.3 ± 0.6	Group 1
AAUCAA	4.0 ± 1.7	Group 1
AAUAAC	3.7 ± 1.5	Group 1
AAUAGA	3.3 ± 1.5	Group 1
AAUUAA	2.3 ± 0.6	Group 1
AAUAAG	1.7 ± 0.6	Group 1