Hypothesis

Not peer-reviewed version

When size *really* matters: the eccentricities of dystrophin transcription and the hazards of quantifying mRNA from very long genes

John Hildyard * and Richard Piercy

Posted Date: 29 June 2023

doi: 10.20944/preprints202306.2167.v1

Keywords: DMD, Dystrophin, gene expression, transcription, mRNA, RNAseq

Hypothesis

When size *really* matters: the eccentricities of dystrophin transcription and the hazards of quantifying mRNA from very long genes

John C.W. Hildyard [1]* and Richard J. Piercy [1]

[1] Comparative neuromuscular disease laboratory, Department of Clinical Science and Services, Royal veterinary college, London NW1 0TU

* Corresponding author: jhildyard@rvc.ac.uk, ORCID: 0000-0003-2283-2118

Abstract: At 2.3 megabases in length, the dystrophin gene is enormous: transcription of a single mRNA requires approximately 16 hours. Principally expressed in skeletal muscle, the dystrophin protein product protects the muscle sarcolemma against contraction-induced injury, and dystrophin deficiency results in the fatal muscle-wasting disease, Duchenne muscular dystrophy. This gene is thus of key clinical interest, and therapeutic strategies aimed at eliciting dystrophin restoration require quantitative analysis of its expression. Approaches for quantifying dystrophin at the protein level are well established, however study at the mRNA level warrants closer scrutiny: measured expression values differ in a sequence-dependent fashion, with significant consequences for data interpretation. In this manuscript we discuss these nuances of expression and present evidence to support a transcriptional model whereby the long transcription time is coupled to a short mature mRNA half-life, with dystrophin transcripts being predominantly nascent as consequence. We explore the effects of such a model on cellular transcriptional dynamics, and then discuss key implications for the study of dystrophin gene expression, focussing both on conventional (qPCR) and next-gen (RNAseq) approaches.

Keywords: DMD; Dystrophin; gene expression; transcription; mRNA; RNAseq

Introduction

The dystrophin gene

The dystrophin gene is one of the largest in the mammalian genome: spanning approximately 2.3 megabases of the X chromosome, this single gene accounts for almost one thousandth of total genomic DNA (fig 1A). Much of this sequence is noncoding: the majority of the 79 exons of the canonical full-length dystrophin mRNA are shorter than 150 bases, typically interspersed with introns that are thousands (or even hundreds of thousands) of bases long, and all of which are transcribed in full. Accordingly, transcription of a single full-length mRNA requires approximately 16 hours, and splicing occurs co-transcriptionally [1], with some long introns spliced in a sequential, multi-stage process [2]. The mature dystrophin mRNA is ~14kb in length (fig 1B, C): less than 1% of the size of the dystrophin gene, yet still a large transcript (even the 3.5kb dystrophin 3' UTR is larger than most mRNAs). The translated dystrophin protein is commensurately large, at 427kDa, and is thus also known as dp427 (dystrophin protein 427kDa). Dp427 is functionally complex; it has three principal domains: an actin binding N-terminus, a central 'rod domain' containing 24 spectrin-like repeats, and a dystroglycan-binding C-terminus (fig 1D). In skeletal muscle, where dp427 is chiefly expressed, dystrophin is closely associated with the sarcolemma (fig 1E). Here it acts as a 'bridge', forming a physical link between cytoskeletal actin and, through α- and β-dystroglycan, the extracellular matrix environment: this is proposed to act as a 'shock absorber', buffering the membrane stresses associated with muscle fibre contraction. The protein also carries four proline-rich 'hinges' conferring flexibility upon the rod domain, and the spectrin-like repeats themselves are functionally distinct: repeats 11-17 form a secondary actin binding domain (which can partly substitute for the N-terminal domain [3]); repeats 16-17 further constitute a binding site for neuronal nitric oxide synthase (nNOS), allowing muscle contraction to elicit vasodilatory increases in blood flow through NO signalling [4, 5]; repeats 20-23 interact with microtubules (and consequently help organise the cytoskeletal microtubule network [6, 7]). Finally, the C-terminus is rich with protein:protein interaction domains, recruiting multiple binding partners (both soluble and membrane-bound) in addition to dystroglycan, including the syntrophins, dystrobrevin, sarcoglycans and

doi:10.20944/preprints202306.2167.v1

sarcospan [8]. Collectively, dp427 thus forms a core component of the dystrophin-associated glycoprotein complex (DAGC), a multimeric, multifunctional sarcolemmal assembly essential for maintenance of muscle fibre integrity (fig 1E).

This picture is further complicated by the fact that dystrophin expression is unconventional: the gene has seven distinct promoters, three of which generate full-length (427kDa) protein products, and four of which are internal, generating N-terminally truncated proteins that contain distinct subsets of the full-length functional milieu [9]. All these isoforms differ only in their unique first exons, each contributing between 100 and 400 unique bases of sequence (predominantly 5′ UTR), while all remaining downstream sequence is shared (fig 1C, D). Like dp427, these isoforms are denoted by molecular weight of the protein product, thus dp260, dp140, dp116 and dp71, with the three full-length isoforms further delineated by principal site of expression, giving dp427c (cortical), dp427m (muscle), dp427p (Purkinje). The major isoform in adults is dp427m, expressed in essentially all skeletal, smooth and cardiac muscle, however the brain also expresses dystrophin [10-13], with dp427c (and to a small extent, dp427p and dp427m) found alongside dp140 and dp71, with dp140 especially abundant within the cerebellum [14, 15]. Dp260 is found within the retina [16], and dp116 expression is associated with the Schwann cells of the peripheral nervous system [17]. Dp71 is widely expressed (with the notable exception of skeletal muscle) but is particularly abundant in endo- and epithelial lineages [9, 18]. These isoforms are also expressed more widely during embryonic development: dp140 is found within the developing kidney [19], but also within the developing central nervous system (suggesting a role in axonal migration [20]), and we have further shown dystrophin isoform expression during both limb development and tooth maturation [21]. This widespread involvement in fundamental developmental processes suggests an ancient origin for this gene, and this is indeed the case: dystrophin orthologs are found in animal lineages from worms and flies, to fish, birds and to humans, and the gene likely predates the arrival of the metazoan kingdom. Interestingly, conservation does not extend to all isoforms: lineage tracing suggests a piecemeal acquisition of shorter isoforms via neofunctionalization. Dp71 is found in all vertebrates, but dp260 and dp140 are tetrapod-specific (for an excellent overview, see [22]).

Dystrophin and muscular dystrophy

The bulk of postnatal dystrophin expression is the dp427m isoform, within skeletal muscle. As discussed above, here dystrophin holds both structural and signalling roles, forming an essential sarcolemmal buffer against the stresses of muscle contraction. Loss of dp427m leaves muscle fibres vulnerable to contraction-induced injury, and results in the muscle-wasting disease Duchenne muscular dystrophy (DMD). DMD is the single most common fatal monogenic disorder, affecting approximately 1 in 3500-5000 new-born boys every year [23]. The disease currently has no cure, and is characterised by repeated cycles of muscle fibre degeneration and compensatory regeneration, with the persistent damage and associated inflammation resulting in progressive replacement of muscle tissue with fibrotic scar tissue and fat, loss of ambulation, and ultimately death via cardiac or respiratory failure.

The sheer size of the dystrophin gene renders it susceptible to point mutations through simple probability: this gene represents 0.1% of the entire genome, thus with ~50-100 *de novo* mutations per generation [24, 25], approximately one in ten individuals will carry a new mutation within the dystrophin gene. The gene is also vulnerable to mutational insertions, duplications and deletions (with the latter being common in DMD patients [8]) and moreover has mutational ‘hot-spots’: regions apparently more prone to mutation than others (the major hot-spot being between exons 44 and 53 [26, 27]). Mutations that generate a premature termination codon (PTC), either via point mutation (such as in the classic animal model of DMD, the *mdx* mouse [28, 29]) or through frameshift following deletion or duplication of one or more exons, result in essentially no detectable dystrophin protein, and consequently a DMD phenotype (notably, seven of the ten exons within the 44-53 region are vulnerable to frameshift). Mutations causing internal truncations that otherwise preserve the reading frame instead result in the milder condition, Becker muscular dystrophy (BMD). BMD patients often retain muscle function well into adulthood, and some are effectively asymptomatic [30, 31]. The extent of internal truncation varies greatly, but even very large deletions (such as exons 13-41) can result in only mild disease [32]. In essence, the dystrophin N- and C-termini appear critical for protective function, while much of the internal rod domain is dispensable: this latter observation has driven several potential therapeutic strategies aimed at restoring dystrophin, effectively converting the DMD phenotype to BMD. These therapies take two principal approaches: exogenous expression of a dystrophin transgene, using "mini-dystrophin" constructs with N- and C-termini but minimal rod domain (allowing the transgene to be packaged within a viral vector); or reframing via exon 'skipping', eliciting the exclusion of one or more additional exons from endogenous dystrophin to restore the reading frame (either at the transcript level via antisense oligonucleotides, or at the genomic level via CRISPR/Cas9-mediated gene editing). The application of this latter approach is contingent on the specific patient mutation, but some exons are more viable targets than others: several mutations within the exon 44-53 hot-spot region can be rescued by skipping of exon 51, for example. Several of these therapies are now approved for human medicine (or have entered clinical trials), and the accelerating pace of development places increasing emphasis on dystrophin quantification, at both protein and mRNA levels.

Quantifying dystrophin protein
Restoration of dystrophin protein is a primary metric for assessing therapeutic efficacy, as only protein confers resistance to contraction-induced damage (high efficiency of dystrophin correction at gene- or transcript-level is of little value if this does not translate to protein). Quantification of dystrophin protein is comparatively straightforward: there are multiple well-validated antibodies to different dystrophin epitopes, and as dystrophic muscle typically contains no detectable dystrophin protein, western blotting or capillary electrophoresis (or indeed immunoaffinity mass spectrometry [33]) can be employed to provide both qualitative and quantitative data [34]. Use of standard curves prepared using known ratios of healthy and dystrophin-negative dystrophic muscle tissue further allows the extent of restoration to be evaluated with precision. It is important to note, however, that such measurements should still be corroborated by immunohistochemistry, allowing the extent of dystrophin restoration to be put into spatial context. Establishing that dystrophin protein is (for example) '15% of WT levels' is insufficient, as this could be achieved via 15% restoration in 100% of muscle fibres, or 100% dystrophin in only 15% of muscle fibres, or indeed anywhere between these extremes. Even comparatively low levels of dystrophin have been shown to protect myofibres [35, 36], thus the modest but global protection offered by the former scenario is likely to prove of more therapeutic benefit than the profound but focal protection of the latter. These approaches also extend to study of dystrophin isoforms: the very short unique N-terminal sequences mean that isoform-specific antibodies are challenging to generate, however C-terminal antibodies will detect all isoforms, which can thus be distinguished electrophoretically on basis of size (for example, to confirm which isoforms are lost in the dystrophic brain [15, 34]).

Quantifying dystrophin mRNA and transcript imbalance
Quantification at the mRNA level is commonly employed to determine efficiency of exon skipping, but can also be used to demonstrate restoration of stable dp427 mRNA via other means (such as gene editing). Study of dystrophin at the transcript level is however more challenging than at the protein level. Dystrophin mutations producing a DMD phenotype are typically those generating PTCs, which preclude viable translation of protein product (see above). Presence of a PTC also flags the offending mRNA for prompt degradation via nonsense mediated decay (NMD), and thus levels of such dystrophin mRNA should ostensibly be low. Reported levels of dystrophin mRNA in dystrophic cells and tissues are however often higher than would be consistent with such NMD-mediated clearance, a finding that, combined with the apparent nuclear enrichment of dystrophin mRNA, has led some to propose that mechanisms other than NMD operate on dystrophin transcripts [37]. It is crucial to recognise, however, that presence of a PTC does not prevent transcription: the pioneer round of translation (the checkpoint for presence of PTCs) occurs after nuclear export, not before [38], and nascent mRNAs (currently being transcribed, but not yet completed) are thus not subject to NMD-mediated clearance. For most conventional genes, this distinction is of little consequence: transcription times are sufficiently short such that nascent transcripts represent only a transient, minority population. For mRNAs with lengthy transcription times, however (such as dystrophin), substantial numbers of transcripts might be present in nascent form, detectable via most measures of gene expression (and indeed exclusively within the nucleus), but not necessarily representative of mature mRNA behaviour.

Dystrophin expression also exhibits a phenomenon that has been termed 'transcript imbalance': measured levels of muscle dystrophin mRNA differ according to the region of the transcript targeted. Sequence lying toward the 3' end of the long dp427 transcript is typically detected at substantially lower levels than that toward the 5′ end (figure 2). This phenomenon has been noted by us [39] and others [40], and indeed was first recognised by Tennyson and colleagues in the 1990s [1, 41]. Of particular relevance, while this 5′-3′ imbalance is more pronounced in dystrophic muscle than in healthy (figure 2B, C), this phenomenon is nevertheless unarguably present within healthy tissue, suggesting it represents a normal facet of dystrophin expression, rather than a disease-specific consequence of aberrant transcription.

One proposed explanation for transcript imbalance is premature transcription termination (PTT), where the RNA polymerase complex dissociates and its transcript thus simply fails to be completed. Some mutations in non-coding regions of dystrophin produce a dystrophic phenotype by eliciting PTT [42], and a similar mechanism might also operate under healthy conditions: in essence, production of dystrophin transcripts might simply have a low success rate. Such proposals are not without merit: while transcription of more conventional genes is typically completed within minutes, the length of the dystrophin gene, and therefore its transcription, might well increase the chances of polymerase dissociation. As noted above, transcription of the entire 2.3Mb dystrophin gene requires 16 hours, suggesting an average transcription rate of ~40 bases a second. This value accords well with those reported for other long genes (40-60 bases per second [43, 44]) and implies that this lengthy transcription time cannot be attributed to pausing, but instead represents essentially continuous RNA polymerase activity. Spontaneous dissociation of the polymerase complex (and consequent PTT) might be extremely rare, but only a single event is needed to disrupt transcription: the requirement for 16 hours of uninterrupted processivity could represent an upper biological limit. Notably, the frequency of stochastic dissociation would increase as a function of transcriptional distance, thus progressive decline in sequence toward the 3′ end might be an expected outcome.

Another candidate explanation is less intuitive: that transcription is not subject to significant PTT, but that instead mature dystrophin transcripts might have half-lives substantially below that of the 16-hour transcription time. As dystrophin is co-transcriptionally spliced [1], 5′ sequence emerges long before transcript completion, and should then persist within the nucleus until polyadenylation and export: levels of this sequence thus reflect both nascent and mature dystrophin mRNA. 3′ sequence is conversely not transcribed until relatively late and thus more closely reflects mature transcripts only. If the lifespan of mature transcripts is comparatively brief (i.e. less than 16 hours), while the initiation of transcription is concerted, then most dystrophin mRNA will, at steady state, be nascent. Such a model was proposed by Tennyson *et al* [41] but could not be empirically confirmed under the technical constraints at the time.

The dystrophin transcriptional model

Using a single-transcript multiplex fluorescence *in situ* hybridisation (RNAscope FISH) approach with probes to 5′ (exons 2-10), middle (exons 45-51) and 3′ (exons 64-75) regions of the dystrophin dp427 mRNA (figure 3A), we recently explored this phenomenon at single-molecule resolution [21, 39]. ISH in adult skeletal muscle (figure 3B) robustly detected all three probes colocalising within the sarcolemma as punctate foci, consistent with triplex-labelling of individual mature dp427m transcripts (figure 3B inset i, arrowheads). Within myonuclei, where probes would bind to nascent transcripts, a strikingly different labelling behaviour was observed: here our triplex approach consistently revealed large, intense foci of 5′ probe; slightly smaller, less intense foci of middle probe; and rare, punctate foci of 3′ probe (figure 3B insets i and ii). A similar pattern was found within developing myotubes of mouse embryos (figure 3C). Here sarcoplasmic mature transcripts appeared more abundant, but nuclear labelling again revealed large 5′ foci, moderate middle probe foci and small, punctate 3′ foci (figure 3C inset iii, and inset schematic). This is consistent with the strong 5′ labelling of myonuclei reported by others [37], and we have shown that these labelling patterns are also found in non-muscle cells expressing full length dystrophin, such as neurons within embryonic and adult brain of both mice and dogs [15, 21], implying that this is canonical behaviour for expression at the dystrophin locus. This phenomenon is consistent with the model proposed by Tennyson *et al* [41], and is summarised in figure 3D: dystrophin transcriptional initiation is robust and continuous, with co-transcriptional splicing ensuring that viable 5′ probe binding sequence emerges rapidly. Transcription time is long, thus high numbers of nascent molecules are present within myonuclei at any given time: of these, most will bind the 5′ probe, approximately half will bind the middle probe, while very few will bind the 3′ probe. Mature transcripts (able to bind all three probes) are rapidly exported, but have modest half-lives and are soon degraded: nascent mRNAs consequently represent the bulk of dystrophin transcripts. Dystrophin mRNA appears to be predominantly nuclear precisely because most dystrophin mRNAs are still being transcribed.

In further support of this model, in tissues where dystrophin expression is expected to be composed predominantly of the shorter isoforms (such as dp140 within the developing nervous system), nuclear labelling is altered accordingly: in the embryonic murine spinal cord (figure 3E), strong nuclear 5′ foci are found only rarely (associated with modest, sporadic expression of dp427), however prominent nuclear labelling with the middle probe is retained, indicating nuclei expressing dp140. The transcriptional start-site for dp140 lies between exons 44 and 45, and thus nascent mRNAs carrying middle probe sequence (but not exons 2-10 required for the 5′ probe) are present within the nucleus for the ~8-hour dp140 transcription time (see model, figure 3F). Finally, as we have shown previously [21, 39], cells expressing the short isoform dp71 label with the 3′ probe only, and exhibit no marked nuclear accumulations: consistent with the short ~1hr transcription time required for this isoform, and analogous to the behaviour of more typically-sized genes.

This simple model (robust transcriptional initiation, long transcription time and short mature transcript half-life) adequately accounts for these observed phenomena, with the corollary that this same model is sufficient to explain the observed transcript imbalance, obviating any requirement for premature termination. We cannot exclude the possibility that PTT also occurs, however, but this process typically results in rapid degradation of the incomplete, non-polyadenylated message (up to and including 5′ sequence). Polymerase dissociation should occur stochastically as a function of length, and thus be more likely to occur toward the 3′ end: longer transcripts would then indeed be underrepresented. As a consequence of PTT-associated degradation, however, every sequence element of these transcripts (from 5′ to 3′) will be equally underrepresented, essentially resulting in an *en bloc* reduction in measured levels of dystrophin mRNA, regardless of position. In other words, whether PTT occurs or not, any observed discrepancy in 5′ vs 3′ sequence can still wholly be accounted for by the combined effects of high transcriptional initiation, long transcription time, and short mature transcript half-life.

As we have previously noted [39], this transcriptional model is counterintuitive: the implication is that myonuclei continuously initiate expression from the dystrophin locus (our data is consistent with 20-40 nascent mRNAs per nucleus), with each new transcript requiring 16 hours of continuous transcription, only for the sarcoplasm to degrade these same transcripts some four hours after completion. This arrangement, while ostensibly wasteful, is likely of negligible metabolic cost compared to the energetic requirements of muscle activity, and interestingly also appears to be conserved among mammals: transcript imbalance is found in humans, mice and dogs [39, 40, 45]. The enormous size of the dystrophin gene is also largely conserved in other vertebrate lineages, suggesting that this transcriptional model, counterintuitive or not, is likely widespread.

Our proposed explanation is that this arrangement permits circumvention of otherwise absolute biological limits: conventional regulation of expression (supply coupled to demand via control of transcriptional initiation) would be entirely adequate under steady-state conditions, and would moreover increase efficiency dramatically (figure 4A), but here any increase in demand would unavoidably incur a 16-hour delay before response, and then a concomitant 16-hour lag to return to basal levels (Figure 4B, C). Conversely, continuous overproduction matched with post-transcriptional control via degradation (i.e. a supply constitutively in excess of normal demand) results in considerable ongoing waste under steady-state conditions, but permits changes in transcript levels (both up and down) to occur over more rapid timescales (figure 4D-F). Whether cellular demand for dystrophin does indeed change markedly under healthy conditions is not presently known, however there are two scenarios under which demand is unarguably high: embryonic myogenesis and muscle repair. In both these conditions, dystrophin levels start at zero, but must reach functional levels comparatively rapidly. A system capable of delivering sufficient mRNA to meet these early needs might entail subsequent inefficiency as an inescapable trade-off.

The biochemical and therapeutic ramifications of this transcriptional model merit more comprehensive examination elsewhere: we discuss some aspects below, but the focus of this work is to instead address the profound consequences of the unconventional dystrophin expression programme at the level of basic investigation, from sample preparation to quantitation and data analysis.

Quantitative measurement of unconventional dystrophin expression

cDNA synthesis

The generation of copy DNA (cDNA) from RNA via reverse transcription is an essential element of most studies of gene expression. Reverse transcriptases require short oligonucleotide primers as initiators: typically, investigators use oligo dT, random hexamers/nonamers, or both. Priming via oligo dT alone allows reverse transcription to commence from the polyA tail (restricting cDNA synthesis to mRNA and avoiding otherwise abundant ribosomal sequences), however polyadenylation occurs only upon transcript completion: nascent mRNAs lack these tails, and will consequently be excluded. Furthermore, the low processivity of reverse transcriptase (even highly optimised recombinant enzymes incorporate only ~1500 bases in a single binding event) means that complete reverse transcription of long transcripts requires consecutive cycles of binding and dissociation: 3′ sequence is consequently more readily captured than 5′ sequence in cDNA libraries, and the extent of this 3′ bias increases as a function of mRNA length. For most mRNAs these nuances are of minimal consequence, but as shown in figure 5, for dystrophin this distinction is critical: mRNA isolated from dp427-expressing cells will contain a mixture of transcripts at different stages of maturity (figure 5A-C), only a fraction of which will be polyadenylated. Random priming allows capture of representative (albeit fragmented) sequence (figure 5D); use of oligo dT priming not only excludes all nascent transcripts, but also overrepresents 3′ sequence of all mature transcripts (indeed we have shown that use of oligo dT priming alone can bias prominently against 5′ sequence even in the ~4.5kb short dystrophin isoform dp71 [46], where nascent transcripts are a minority). In essence, reliance on oligo dT priming gives a representation of dystrophin expression that reflects the exact opposite of biological reality (fortunately, the widespread recognition of 3′ bias means that most investigators recognise the challenges presented by the ~14kb mature dystrophin mRNA, and employ random priming as a matter of course).

Comparing healthy and dystrophic transcripts

This transcriptional model also influences interpretation of expression under dystrophic conditions. As shown by the phenomenon of transcript imbalance, however, careful choice of target site allows nuanced assessment of transcriptional dynamics. Mutations eliciting a DMD phenotype predominantly introduce premature termination codons (PTCs) into the dystrophin transcript: these will be promptly degraded via nonsense mediated decay (NMD), but as noted above, this step occurs only upon nuclear export (i.e. after transcript completion -figure 5F). Consequently, in dystrophic muscle, mature transcripts are greatly reduced as expected, but nascent mRNAs remain (indeed *mdx* muscle myonuclei retain strong 5′ probe foci under ISH [39]). If transcriptional initiation remains otherwise unchanged, RT-qPCR from healthy and dystrophic muscle (figure 5G, H i-iii) will thus suggest different extents of NMD depending on the precise region of the transcript measured. 5′ sequence (exons 1-2) is present in essentially all transcripts (both nascent and mature) and will show only modest fold changes, while 3′ sequence (exons 62-63) is present chiefly in mature transcripts subject to NMD, so here strong reductions will be reported (figure 5I). Note this also complicates interpretation of isoform expression (for example within dystrophic brains): here first exon sequence is the only means to distinguish isoforms (and thus might not reflect the true effects of NMD), while 3′ sequence is common to all isoforms (and thus will not reflect isoform specific behaviour). Conversely, changes in transcriptional initiation (in the absence of NMD) will result in *en bloc* alterations in transcript numbers (figure 5J-L), leading to more consistent fold changes regardless of region measured (figure 5M). These two scenarios are not merely of academic interest: *in vivo,* both might be present simultaneously, and use of target sites that distinguish predominantly nascent mRNAs (exons 1-2) from predominantly mature (exons 62-63) permits contributions of NMD and transcriptional initiation to be assessed effectively independently. We and others [37, 39] have shown that in the *mdx* mouse, reductions in transcriptional initiation do indeed occur alongside NMD (see figure 2): levels of mature transcripts are markedly lower than in healthy muscle (consistent with degradation) but nascent transcripts levels are also reduced, and by a greater factor than can be attributed to NMD (accordingly under ISH, nuclear foci are prominent, but reduced in intensity). Notably, this phenomenon does not appear to be a generalised feature of dystrophic mammals: in the DE50-MD dog model of

DMD loss of mature transcripts to NMD is profound, while transcriptional initiation remains essentially unchanged [45].

Measuring transcriptional changes over time

Temporal studies of gene expression (both *in vitro* and *in vivo*) are also necessarily subject to constraints due to this lengthy transcription period. Use of 5′ sequence to study dystrophin transcription during myogenic differentiation would suggest expression begins earlier (and reaches higher final levels) than would be reported if 3′ sequence were used instead. This also extends to pharmacological modulation of transcriptional behaviour under steady-state conditions: changes in transcriptional initiation will be detectable relatively swiftly if analysis is focussed on 5′ sequence, but will remain undetectable at the 3′ end for more than half a day. Reversion to canonical expression patterns (such as following pharmacological washout) will similarly exhibit substantial lag, necessitating extreme caution in data interpretation. A 6-hour pharmacological blockade of transcriptional initiation followed by a 6-hour washout, for example (figure 6A-E), would generate wildly different findings depending on the specific regions of the transcript analysed. 5′ sequence might respond largely as expected, but 3′ sequence would suggest no response to either treatment or washout over this timeframe. Such a treatment/wash protocol would create a ‘bubble’ of altered transcriptional behaviour moving along the dystrophin locus, and the most profound effects on mature transcripts might not be detected until a full day after treatment began (figure 6E). Inhibition of post-transcriptional degradation, conversely, would increase measured levels of all sequence regions equally (figure 6F), essentially applying an *en bloc* linear transformation (as has been reported following cycloheximide treatment [37]). Expressed as fold change, this increase would first manifest most prominently at the 3′ end where basal levels are low, and these increases would be gradual, commensurate with the slow but continuous production of dp427 mRNAs (5′-3′ differences should also lessen as mature transcripts accumulate, but the prolonged inhibition times necessary for this effect are likely to be incompatible with cell viability).

Quantifying exon skipping

Of particular therapeutic interest, this model influences assessment of exon skipping (targeted exclusion of one or more exons to restore the dystrophin reading frame): if most dystrophin mRNA is nascent, and moreover co-transcriptionally spliced, many skipped transcripts will be immature at the time of measurement (figure 7A). Primers spanning the target site (such as exons 22 and 24 for the *mdx* mouse) permit absolute quantification of skipped transcripts regardless of maturity, however expressing this as a fraction of overall dystrophin levels is nontrivial. Quantifying sequence 5′ to the region of interest (for example, exons 1:2, for ‘total’ dystrophin) underestimates skipping efficiency, as this necessarily includes nascent transcripts that do not yet even contain the target site. Use of a site more 3′ correspondingly overestimates efficiency, as fewer transcripts in total contain such 3′ sequence: indeed, for a skipping site at exon 23, normalising to 3′ sequence could give efficiency estimates in excess of 100%. Sequence closest to the target site gives greatest accuracy (or alternatively sequence unique to unskipped transcripts: measuring across the exon 23:24 splice, for example), however regardless of site, it should be remembered that all measurements are additionally subject to survivorship bias: unskipped mature transcripts are degraded (and thus underrepresented), while successfully skipped transcripts are not (figure 7A, i-iii).

This picture is further complicated by the dystrophin locus itself: while dp427 mRNA takes ~16 hours to complete, individual exons are not spaced equidistantly along the gene (figure 7B): some thus exhibit more closely matched transcriptional behaviour than others. The first ten exons, for example, are sparsely distributed across some ~800 kilobases of genomic DNA, while the following 30 exons occupy less than half that. Consequently, specific sequence regions do not emerge in a smooth gradient, but in a more ‘burst-like’ fashion (see figure 7C), with corresponding consequences for comparison of skipped/unskipped transcripts (or healthy/dystrophic). Exon 10 is transcribed almost 6 hours after initiation, but the entire following sequence to exon 41 then emerges in only ~2 hours: all exons within this 10-41 region thus represent similar fractions of total dystrophin mRNA (figure 7D), and will exhibit comparable fold differences between healthy/dystrophic (figure 7E).

The most rigorous approach consequently would be to quantify multiple regions independently, in both treated and untreated samples: comparison of early 5′ sequence would assess influence of treatment on transcriptional initiation, comparison of skipped sequence with adjacent representative sequence would assess skipping efficiency, and comparison of late 3′ sequence would establish the resultant fold enrichment of mature, viable mRNAs.

Dystrophin in the transcriptomic era

Next-generation high-throughput transcriptomic approaches (such as RNAseq) are increasingly popular and accessible even to those with modest budgets. As such, the eccentricities of dystrophin expression must also be extended to these techniques, here applying caveats both to sample preparation and data analysis/interpretation. For sample preparation, standard RNAseq pipelines first employ oligo dT column purification (eliminating ubiquitous non-coding RNAs such as ribosomes that might otherwise dominate sequencing). As noted in figure 5, even this simple step excludes all nascent dystrophin transcripts, a bias then potentially compounded by oligo dT-directed reverse transcription, which favours 3′ sequence. Following fragmentation, ligation and sequencing to FastQ format, analysis first requires mapping of sequence data to a reference genome to establish the genomic location of each read: these (large) BAM files of aligned reads are then subsequently compared to an annotated feature file to determine which transcripts each genomic location corresponds to, ultimately producing a simple value of ‘reads per transcript’ (or more accurately, ‘reads per feature’). Post-hoc corrections can be applied to normalise for transcript length (longer mRNAs generate more reads), but conventionally, the specific location of each read within a transcript is not considered relevant to downstream analysis. For a typical RNAseq pipeline, therefore: mRNA isolation eliminates nascent transcripts, reverse transcription biases against 5′ sequence of whatever remains, and downstream sequencing analysis then hides these concerns from the investigator. For dystrophin in particular, this approach also precludes analysis of isoform-specific expression behaviour: even when isoform variants are present within the feature file, all reads that map to the *Dmd* gene feature are flagged simply as ‘Dmd’.

Addressing sample preparation concerns is challenging but not impossible: use of ribodepletion instead of oligo dT purification allows retention of nascent transcripts, and similarly (as in figure 5), use of random priming (rather than polyA-directed oligo dT priming) allows capture of sequence without 3′ bias (this approach can also be employed to investigate intronic sequence [42]). These modifications cannot be applied post-hoc, however, and are thus of no benefit to sequencing data already obtained using conventional polyA purification/priming.

Data analysis is conversely more tractable to post-hoc reassessment. Genome feature files (.GFF or .GTF format) store genomic coordinates of each exon, and the gene ID of the corresponding spliced transcript, but there is no *a priori* reason feature mapping cannot be conducted down to the level of individual exons. This can be done by opening aligned BAM files in a genome browser (such as IGV) and manually counting reads for each exon [47], though this approach is somewhat painstaking. Alternatively, commonly used programs such as Htseq-count can conduct exon-level analysis innately, however this approach necessarily applies to all exons across the entire feature file (generating a vastly excessive dataset), and moreover the nomenclature used to designate individual exons is unwieldy, especially for transcripts with multiple listed variants (such as *Dmd*). For dystrophin-focussed investigations, a more practical approach would be advantageous. As gene IDs are simple text fields, and genome feature files can be readily edited (via excel or text editor), we therefore modified the GRCmm39 mouse genome feature file to add each dystrophin exon as a unique, distinct gene ID (*Dmd_exon3*, etc) alongside the conventional ‘Dmd’ assignment. This allows mapped reads to be evaluated both overall (total ‘Dmd’), and at single-exon resolution (figure 8A). We further assigned distinct IDs to unique isoform first exons, extending resolution to individual full length isoforms (dp427c, m, p) and to dp260, dp140, dp116 and dp71 (we note that a similar approach was elegantly employed by Doorenweerd *et al.* to identify isoform-specific expression patterns in human brain RNAseq datasets [20]).

We first assessed public repository FastQ data taken from a comprehensive transcriptional profiling study of muscle types in healthy mice (generated by Terry *et al* [48]), examining three commonly-studied pelvic limb muscles, the tibialis anterior (TA), the extensor digitorum longus (EDL) and the soleus (SOL). These datasets were then subjected to a standard analysis pipeline using a standard mouse feature file or our modified version (see methods). This dataset should reflect mature transcripts only: as discussed above, use of polyA purification excludes nascent transcripts prior to reverse transcription (figure 5E). Conventional analysis confirmed the authors’ original findings: as expected, myosin heavy chains represented a high percentage (~5%) of total reads yet also revealed differences associated with the characteristic roles of each muscle type. Reads per million (RPM) for the very fast IIB myosin heavy chain (*MYH4*) were high (~50,000) in the faster TA and EDL muscles but lower in the slower SOL, with the TA also exhibiting higher levels of the fast IIX isoform (*MYH1*). Conversely, the fast IIA (associated with fast oxidative fibres) and slow Ib MHCs (*MYH2, 7*) were elevated in SOL but not TA or EDL (figure 8B). RPM values for dystrophin (*Dmd*) showed no prominent muscle-specific behaviour, being similar across all samples. Counts for this transcript were also markedly lower (~200 RPM) as expected for a low abundance transcript.

Reanalysis of these data using our exon-specific dystrophin feature file added substantial context to these findings. The dystrophin 3′ UTR (exon 79) is 2.7kb in length, almost 20% of the mature transcript: one would thus expect reads to exon 79 to be over-represented. Oligo dT primed reverse transcription also biases in favour of 3′ sequence, potentially compounding this over-representation. As shown (figure 8B, *Dmd* exon 79) this was indeed the case: reads to the (predominantly untranslated) exon 79 represented 30-60% of total *Dmd* reads and were consequently highly comparable to *Dmd* values obtained via conventional analysis. Reads to all other exons were markedly lower, but 3′ bias was still evident: RPM values at the 3′ end numbered in the hundreds, gradually diminishing to mere tens as exons become more 5′ (figure 8C). We then adjusted per-exon RPM for exon length, effectively obtaining an "RPM per base" value for each exon: normalising both for the enormous size of the 3′ UTR, and for single reads assigned to multiple exons (see figure 8A). Expressed in this manner (figure 8D) the 3′ bias was rendered substantially more obvious, and moreover illustrated the highly consistent read counts between muscles. Finally, we plotted these corrected counts against transcript length, using the midpoint positions of each exon within the long dp427 mRNA as X-axis coordinates: as the efficiency of reverse transcription declines essentially as a first-order function of length, plotting in this manner allowed the 3′ bias of oligo dT-primed reverse transcription to be empirically calculated (figure 8E). For this dataset, we observed a mean gradient of ~-0.0005 log2(RPM).base^{-2}, i.e. a 2-fold drop in per base sequence capture for every 2000 bases of distance from the 3′ terminus: again, this gradient was remarkably consistent across muscles, suggesting that this could be used as a dataset-wide correction factor for 3′ bias. Finally, ~95% of isoform-specific first exon counts (figure 8F) mapped to dp427m, indicating (as expected) that essentially all sequence data represents the muscle isoform of full-length dystrophin. Low levels (<4 counts) of dp116 and dp71 were detected in some samples, consistent with minor contributions from peripheral nervous tissue and vasculature, respectively, but reads to all other isoforms were essentially absent.

Next we applied this approach to dystrophic muscle, assessing RNAseq data generated by Chemello *et al* [49] using the ΔEx51 mouse model of DMD (a gene-edited mouse model that lacks dystrophin exon 51). Conventional analysis of 4-week old TA muscle samples showed *Dmd* was expressed in healthy samples at levels highly comparable to the mouse muscle data above, but was markedly decreased in ΔEx51 samples, as would be expected (loss of exon 51 causes frameshift, leading to transcript degradation by NMD). Exon-specific analysis again showed that exon 79 reads were over-represented, accounting for ~30% of total (figure 9G), but also demonstrated that ΔEx51-associated decreases in reads were comparable across the entire length of the transcript (reassuringly, zero reads mapped to the absent exon 51 in ΔEx51 samples -figure 9H). This consistent decrease suggests these ΔEx51sequences do indeed correspond to mature mRNAs, perhaps captured prior to nuclear export or *en route* to degradation (note that first exon sequences were near exclusively dp427m, figure 9I). Interestingly, 3′-5′ bias in this instance was consistent with substantially higher efficiency reverse transcription (2-fold drop every ~7000 bases) potentially explaining the more modest enrichment of 3′ UTR sequence in this dataset.

Finally, to explore the broader utility of exon-based interpretations, we examined data generated by Schmitt *et al* [50], using murine brain samples collected from embryonic day 15.5 (E15.5) to 29 days post birth (P29). Given developmental expression of dystrophin (especially within the brain), this data should report multiple isoforms, and moreover, sample preparation for this dataset used ribodepletion and random priming, so should be more representative of both nascent and mature transcription. Conventional analysis readily identified expression of *Dmd*, with expression increasing ~2-fold from E15.5 to P29 (figure 10A). Exon-specific analysis again demonstrated the strong representation of the exon 79 3′ UTR, but here this exon contributed a smaller fraction than in the studies assessed above (~20%, figure 10B). Examination of unique first exon reads moreover added considerable nuance to this data, demonstrating that expression of *Dmd* was indeed distributed across multiple isoforms (figure 10C). Both the cortical full-length isoform dp427c and the short dp71 isoform were robustly detected, and at comparable levels, with expression patterns that broadly mirrored overall *Dmd* expression (though age-associated increases in dp427c were more dramatic than in dp71). Dp140 was the other major contributor, but here expression declined with age: a finding at odds with the behaviour of 'Dmd' under conventional analysis, but one that reflects the involvement of this isoform in earlier rather than later neural development. Both muscle and Purkinje full-length isoforms (dp427m, p) were initially essentially absent, but both were present at very low levels by P22-29, while expression of retinal dp260 and peripheral nerve dp116 remained consistent with stochastic noise. Expression of individual exons across the entire dystrophin locus was also markedly different from the patterns revealed above (figure 10D-I): read counts here were biased in the opposite fashion, exhibiting a clear 5′ enrichment rather than 3′, alongside periodic 'spikes' in expression corresponding to distinct initiation events from the dp140 and dp71 promoters (upstream of exons 45 and 63, respectively). The magnitude of these 'spikes' differed according to relative isoform abundance within a sample (note the exon 45/dp140-associated spike is barely detectable by P29 -figure 10I), and which moreover similarly decline from 5′-3′. These data are wholly consistent with the predominantly nascent expression model proposed here: 5′ sequence should be more abundant than 3′ regardless of isoform, and the contributions of multiple isoforms expressed within a sample (dp427, dp140, dp71, figure 10J-L) should thus overlap to create a 'saw-tooth' distribution of exon abundance (figure 10M).

In summary, this brief overview, using established repository-located RNAseq datasets, supports our transcriptional model and moreover demonstrates the limitations of polyA-purification and oligo dT priming in the generation of cDNA samples and sequencing libraries. This work focusses only on dystrophin expression, but these findings likely extend to other long transcripts or genes (such as titin and obscurin). We also illustrate the utility of applying a more nuanced exon-level analysis to dystrophin within RNAseq data: this approach offers insights into transcriptional dynamics and isoform expression and can potentially allow broader evaluation of dataset-wide bias.

Discussion

The dystrophin transcriptional model

In this manuscript we present a transcriptional model that accounts for previously reported eccentricities of dystrophin expression observed biochemically [1, 37, 40, 41], and for the behaviour shown here and previously [15, 21, 39] via multiplex FISH. We further show how dystrophin-focussed analysis of next-generation transcriptomic datasets also reveals evidence in support of this model. As noted above, similar models have been proposed historically, but have typically been viewed with caution, or considered to be transcriptional aberrations. As we discuss here, our data suggests that this model is indeed correct, and moreover extends to expression of some N-terminally truncated dystrophin isoforms. We argue that this model represents normal, functional expression for the dystrophin gene under healthy conditions. As dystrophin expression is restricted to specific cell types, control at the level of transcriptional initiation clearly occurs, but this is essentially on/off Boolean control: with an unavoidable 16-hour transcription time, more responsive expression simply cannot be achieved. Continuous overproduction coupled with rapid post-transcriptional degradation thus represents the only way to control expression over meaningful biochemical timescales. Whether such short-term fine control is essential remains an open question. As discussed above, during embryonic development and muscle regeneration (where initial dystrophin levels are zero) overproduction would allow high early translational demands to be met in a timely fashion, with post-transcriptional breakdown rates subsequently increasing to bring levels to steady state. There might however be other, more subtle scenarios under which marked increases in dystrophin supply are required over the shorter term, such as membrane repair. Muscle fibre membranes are subject to considerable stresses even under normal muscle activity, which can cause microtears: the resultant calcium influx initiates a repair cascade that rapidly seals the initial injury and then mediates the remodelling necessary for complete repair [51, 52]. Evidence suggests dystrophin protein is essentially immobile once localised to the sarcolemma [36], implying that full repair of the sarcolemmal environment requires *de novo* dystrophin synthesis. At eukaryotic translation rates of ~5 amino acids per second [53], production of a single dystrophin protein requires ~12 minutes assuming mRNA is readily available: a modest delay that remains compatible with remodelling-associated repair. A 16-hour delay might not be so well-tolerated, and our proposed model would thus represent an effective (if wasteful) solution to this physiological dilemma.

A further question is how transcript degradation is controlled. The extensive 3′ UTR is likely to play a role: studies have shown that the dystrophin 3′ UTR can influence the stability of luciferase constructs [54] (and indeed long 3′ UTRs are known to promote degradation innately [55]). The 3′ UTR of dystrophin is also highly conserved across species, and mutations affecting this UTR can produce both BMD and DMD phenotypes [56], suggesting that its contributions to stability are complex. It is also key to note that while the 3′ UTR (and thus susceptibility to degradation) is shared across dystrophin isoforms, transcription times vary extensively: cells could thus readily achieve higher levels of short dp71 than long dp427, even if both transcripts are subject to the same constant rate of degradation. Degradation could also be modulated as a consequence of disuse: translation factors such as PABP and eIF-4E compete with mediators of mRNA decay [55, 57], thus translational activity inherently increases mRNA stability. Given the slow turnover of dystrophin at the protein level (half-life of weeks to months [58, 59]), demand at the mRNA level might be so minimal under healthy conditions that most dp427 transcripts are not translated at all, and thus promptly degraded as a consequence. This would explain the reported persistence (days to weeks) of therapeutically skipped transcripts in dystrophic muscle [59]: rare, corrected mRNAs would be in high translation demand in dystrophin-negative context, and therefore continually protected from degradation. This could also explain the domain-restricted behaviour reported previously, where dystrophin protein remains apparently confined to myonuclear territories [36]. Under this model, in healthy muscle, once sufficient dystrophin protein is established within the immediate sarcolemmal territory of a myonucleus, subsequent mRNAs from that nucleus will predominantly be degraded as surplus long before diffusion or trafficking can carry them beyond the domain boundary. This could similarly underpin the marked differences in the pattern of dystrophin restoration reported by Morin *et al*, depending on therapeutic approach [36]: here CRISPR/Cas9-mediated genomic correction of *mdx* muscle restored discrete, separate but prominently dystrophin-positive domains, whereas antisense oligonucleotide-mediated transcript correction elicited widespread but more modest restoration. Under this model, genomic correction (via CRISPR) would result in re-establishment only of a local domain arrangement, with these immediately proximal ‘healthy’ levels of dystrophin protein ensuring rapid turnover of excess corrected mRNAs (with concomitant failure to restore dystrophin protein more distally); conversely, distributed low-level correction at the transcript level (via ASO) would restore only modest levels of dystrophin protein, but more widely, and these ‘beneficial but sub-normal’ levels of protein would not wholly satisfy demand, potentially serving to protect corrected transcripts by virtue of rendering them more translationally active. If dystrophin mRNA stability is indeed strongly influenced by translational activity, ‘not enough’ might prove markedly more effective than expected (this rationale could moreover be applied to dystrophin positive ‘revertant fibres’ [60, 61]: viable transcripts produced by rare aberrant splicing events would be rendered highly stable and thus capable of generating substantial dystrophin protein over time). Factors that interact with the 3′ UTR and mediate stability would be broadly beneficial to multiple therapeutic approaches, and thus represent potential therapeutic targets, though we note that within this transcriptional framework, gene therapies utilising mini- or micro-dystrophin constructs (which do not carry the long 3′ UTR) would not be subject to these constraints, and thus transgene mRNAs might be effective even distant from transduced myonuclei.

Dystrophin expression and RNAseq

We show here how this transcriptional model informs interpretation of RNAseq datasets, and illustrate the advantages gained by use of ribodepletion and random priming for reverse transcription (as opposed to polyA purification and oligo dT-directed priming). The analysis approach used here permits more detailed assessment of dystrophin transcription, but we acknowledge that this analysis remains somewhat simplistic. The HTSeq package is primarily intended for differential expression studies, and use of this software for mapping to the level of individual exons is thus not without caveats. A sequence that maps to two exons is counted as a ‘read’ to both exons even if one is represented by only a single base: our per-base read counts are thus not truly reflective of exonic read depth, and very short exons might be overcounted purely because such short sequences can be partly present in more reads (for example, the 39-base exon 71 exhibited consistently higher corrected reads than adjacent exons). Similarly, sequences at transcript termini (5′ unique first exons and exon 79) might be undercounted to some extent, as reads to these

doi:10.20944/preprints202306.2167.v1

sequences can only overlap from one end. The substantial length of exon 79 moreover means most reads to this feature map *only* to this feature, potentially compounding this under-representation (we note that per-base corrected RPM values for exon 79 were consistently lower than adjacent exons -see figures 8 and 9). Our approach also does not consider alternate splicing: while full-length dystrophin is not held to be alternately spliced in mature skeletal muscle [62], exon 78 is omitted from dp427m at high frequency during embryogenic expression [63, 64], and there are multiple splice variants reported for dp71 [65]. Splice events can be identified by eye (using genome viewers), and splice-aware mapping approaches are also available, but these analyses can be challenging, particularly for highly variably spliced transcripts, and consequently fall beyond the scope of this manuscript. A further potential source of variability is random priming itself: although necessary to capture representative expression along the entire dystrophin locus, this method assumes that specific hexamer sequences occur with approximately equal frequencies in eukaryotic genomes, and that these same hexamers are equally efficient as primers for reverse transcription. Neither assumption is correct [66], and while this is of only minimal consequence for whole mRNAs, when assessed at the level of individual exons, some regions might well be more readily captured than others. A more nuanced approach would factor in these biases, count mapped reads down to the true individual base level, and ideally incorporate identification and enumeration of alternative splicing events: future investigations might address these current limitations (indeed it would be of considerable interest to repeat this approach comparing polyA/oligo dT and ribodepleted/random primed datasets generated from the same underlying RNA samples).

Dystrophin transcription: caveats and alternative hypotheses

While the model proposed here is a biophysical inevitability given the size of the dystrophin locus, and is moreover sufficient to explain the transcriptional eccentricities of dystrophin, this model need not be exclusive: other factors might well contribute. As noted above, we and others [37, 39] have shown dystrophin transcriptional initiation is indeed reduced in dystrophic mouse muscle (which has been ascribed to chromatin remodelling [37]), but this notably does not occur in dystrophic dog muscle [45]. This model also describes only bulk behaviour: our data is consistent with ~20-40 nascent transcripts per myonucleus on average, but this could be achieved by modest, continuous 'trickle' delivery, where nascent transcripts are spaced out along the dystrophin locus, or by high, infrequent 'burst' initiation where nascent transcripts form a close-packed transcriptional 'bubble' moving along the locus. We cannot at present distinguish the two empirically, though we favour the former: under a 'burst' system multiplex ISH should reveal substantial numbers of nuclei positive for 5' but not middle probe foci, something we do not observe. At the level of transcriptional elongation, aberrant premature transcriptional termination (PTT) might also play a role, though this would necessarily occur on a background already dominated by nascent transcripts (as discussed above). Potential effects of PTT can be approximated by calculating 'fraction of transcripts completed' for a given dissociation constant: with a transcriptional demand of 2.3 million consecutive base incorporation events, a per-base spontaneous dissociation rate of 10^{-5} renders dystrophin transcription effectively non-viable (and indeed would significantly impact even substantially shorter genes); at 10^{-6}, only ~10% of dp427 transcripts would reach completion; at 10^{-7}, completion rates conversely approach 80%. Put simply, there is only a narrow range of per-base dissociation rates over which PTT might be considered relevant. The spontaneous dissociation rate of the eukaryotic RNA polymerase II complex is not known, though studies in yeast imply high processivity [67]. The error rate (incorporation of incorrect bases) is conversely better studied, and is indeed comparatively high (~10^{-6}-10^{-5} per base [68-70]). The fact these rates are measurable implies dissociation is likely to occur at a markedly lower frequency than such single-base errors. Furthermore, a significant transcriptional failure rate due to PTT would likely be actively deleterious: a biochemical scenario whereby transcription cannot reliably be assumed to proceed to completion is one wherein the success of supply/demand logistics is placed at the mercy of stochastics. Under such a model, even constitutive oversupply cannot necessarily be guaranteed to meet demand, and is indeed equally likely to result in needless excess. Were dystrophin expression to be limited by transcriptional completion rate, one might well expect strong selective pressure for a shorter, less challenging locus size: the fact that this is not the case across multiple genomes strongly implies the enormous size of dystrophin is well-tolerated (one might also expect aberrant overexpression in Becker patients with large internal deletions of the *Dmd* gene, but again this is not reported). Conversely, an overproduction and post-transcriptional control model is not deleterious: wasteful and counter-intuitive, certainly, but as we discuss here, waste might be a necessary trade-off for a system readily capable of meeting both steady-state and variable demand over meaningful biological timescales. One interesting caveat to transcription from such a large locus is that expression is mutually exclusive with replication: 16 hours of uninterrupted transcription cannot be accommodated within a typical mammalian S-phase of 8-10 hours. As we have previously noted [21], expression of full-length dp427 (and indeed dp260 and dp140) is predominantly associated with either muscle or neuronal tissues: both comprised chiefly of post-mitotic cell types. One potential exception is the expression of dp427 in activated satellite cells [71], which occurs prior to asymmetric division. Satellite cells are initially quiescent, however, and satellite cell division itself lags ~15-20 hours behind initial damage-associated activation [72, 73]: this is consistent with a mitotic delay necessary to transcribe and then translate sufficient dp427. Dp427 is conspicuously not expressed within proliferating myoblasts (though these cells do express the short isoform dp71 [74, 75]), and indeed full-length dystrophin expression resumes only once cells have fused to form post-mitotic myotubes.

Finally, we note that dystrophin is not the only long gene: the genes for neurexin-family member *CNTNAP2* and the protein-tyrosine phosphatase *PTPRD* are of comparable size (2-2.3Mb), and approximately 50 human genes are 'very large' (> 1Mb) [76]. Many of the biophysical constraints described here could equally be applied to these other large genes, and indeed the bulk of these 50 genes are associated with post-mitotic muscle or neuronal cell types: it would be of considerable interest to investigate whether these genes also adopt similar transcriptional strategies.

Conclusions

This primary purpose of this work is to consolidate the eccentricities of dystrophin transcription into a cohesive model and explore the consequences that might result. The data presented remains consistent with this model, though further investigations are merited. This work should hopefully serve as a primer to both molecular biologists and bioinformaticians alike, illustrating the complexities of measuring transcription from such a challenging locus and providing a conceptual framework for the interpretation and analysis of gene expression data derived from both traditional and next-generation approaches.

Methods

Sample collection and preparation

All tissues used here were taken from our tissue archive: no animals were killed specifically for this work. Mouse TA muscle samples were collected post-mortem from WT mice (C57Bl/6), mounted on corks with cryoMbed (Bright) in relaxed, longitudinal orientation and snap frozen under liquid nitrogen-cooled isopentane before storage at -80°C. Mouse embryos were collected as described previously [46], fixed in 10% NBF for 24 hours and then processed to wax in sagittal orientation for histological sectioning. Embryo used for the work shown here was collected at embryonic day 16.5 (E16.5).

Muscle samples were cryosectioned at -25°C to 8µm thickness using an OTF5000 cryostat (Bright) and mounted on glass slides (SuperFrost, VWR). Slides were air-dried at -20°C for 1 hour before storage at -80°C.

Wax-embedded embryos were cooled on ice and sectioned at 4µm thickness using a microtome (Leica Biocut), then floated in a waterbath at 48°C and mounted on Superfrost slides as above. Slides were dried at 37°C overnight and stored at room temperature in sealed containers (with silica gel desiccants as recommended) until use.

Multiplex FISH: sample preparation

Single transcript multiplex fluorescence *in-situ* hybridization was conducted as described previously [15, 21, 39], using the RNAScope ISH platform (ACDbio).

For cryosectioned skeletal muscle, slides were removed from -80°C storage and placed immediately into cold (4°C) 10% neutral-buffered formalin, then incubated at 4°C for 1 hour. Slides were dehydrated in graded alcohols then air-dried and baked at 37°C for 1 hour. Sections were ringed using hydrophobic barrier pen (Immedge, Vector Labs) and then treated with RNAscope hydrogen peroxide (15 min) and Protease IV (30 min) as per standard RNAscope protocol.

Paraffin-embedded sections were treated according to the RNAscope protocols for FFPE, with target retrieval using the manufacturer's 'alternative method': slides were immersed slowly in target retrieval buffer (held at a gentle boil) for 15 mins, before cooling directly into room temperature distilled water, followed by ethanol dehydration.

RNAscope multiplex assay

Multiplex assays were performed as suggested by the RNAscope multiplex fluorescent reagent kit v2 (ACDbio) protocols, using our mouse dystrophin probe set: Mm-Dmd (452801), Mm-Dmd-O1-C2 (529881-C2) and Mm-Dmd-O2-C3 (561551-C3) (C1, C2 and C3 probes to 5', 3' and middle sequence of the dp427 transcript, respectively; see figure 3A). Nuclei were stained with Hoechst (1/2000 dilution in wash buffer, 5 min) and slides were mounted in Prolong Gold Antifade mounting medium (Thermofisher) and allowed to dry overnight (room temperature, protected from light).

Fluorophores were assigned as follows: 5' probe (C1), TSA-Cy3; middle probe (C3), TSA-Opal520; 3' probe (C2), TSA-Cy5 (all TSA reagents: Akoya biosciences).

Imaging

Individual images were captured using a DM4000B upright microscope with samples illuminated using an EBQ100 light source and A4, L5, N3 and Y5* filter cubes (Leica Microsystems) and an AxioCam MRm monochrome camera controlled through Axiovision software version 4.8.2 (Carl Zeiss Ltd). Objectives used were 20x HC PL FLUOTAR PH2 (NA=0.5).

RNAseq analysis

All RNAseq analysis used here was conducted within the Galaxy online platform [77], or post-hoc using Microsoft excel. Public repository RNAseq datasets in FastQ format were downloaded and mapped to the GRCm39 mouse genome assembly (RefSeq Acc no. GCF_000001635.27) using HISAT2 [78]. Aligned BAM format files were then mapped to a feature file using htseq-count [79] to determine reads per feature. Use of a custom genome feature file (based on GCF_000001635.27_GRCm39_genomic.gtf) where all dystrophin exons are also assigned unique geneIDs, either according to exon number (exons 2-79) or to isoform (unique first exons) allows this analysis to evaluate both total Dmd counts and counts per individual exon (to allow reads spanning several exons to be assigned correctly, htseq mapping used 'mode: union' and 'nonunique: all'). Custom feature file is available on request. Count files were exported to excel and raw counts per feature extracted: counts were converted to reads per million (RPM) to correct for differences in sequencing depth between datasets, and (where indicated) then corrected for exon length.

Figure legends

Figure 1: the dystrophin gene

The dystrophin gene is located near the centre of the X chromosome (A) and represents ~2% of total X chromosomal sequence. The gene is comprised of 79 canonical exons (B), several of which are interspersed with large introns (>100kb). The gene has seven distinct promoters (C), each of which contributes a unique first exon. Three generate full length dystrophin (dp427c, m, p), while the remaining four are internal, giving rise to N-terminally truncated proteins designated by molecular weight: dp260, dp140, dp116 and dp71. At the protein level (D), full length dystrophin carries an actin-binding N-terminus, a central rod domain of 24 spectrin-like repeats and a C-terminal dystroglycan binding domain. Repeats 11-17 of the rod domain form a secondary actin-binding domain, and 16-17 bind nNOS. Repeats 20-23 confer microtubule-binding activity. The C-terminal domain also mediates interactions with syntrophins, dystrobrevin, sarcospan and sarcoglycans. Each truncated dystrophin isoform carries a subset of the full dystrophin functional milieu. In skeletal muscle, dystrophin is associated with the sarcolemma (E), where it associates with the eponymous, dystrophin associated glycoprotein complex (DAGC), a physical link between actin cytoskeleton and extracellular matrix (chromosomal ideogram adapted from National Center for Biotechnology Information, U.S. National Library of Medicine, other figure elements adapted from Hildyard et al, 2020).

Figure 2: dystrophin transcript imbalance
Transcript imbalance can be detected using primers to 5′ (exons 1-2), central (exons 44-45) and 3′ (exons 62-63) regions of the 14kb dp427 mRNA (A). Used with cDNA prepared from healthy (B) or mdx (C) murine skeletal muscle, these primers reveal markedly greater levels of 5′ sequence than 3′. This phenomenon is not dependent on genotype, but measured levels within dystrophic murine muscle show an en bloc reduction, with 3′ sequence being almost absent (figure adapted from Hildyard et al, 2020).

Figure 3: dystrophin transcription as revealed by multiplex FISH
RNAscope 20-ZZ probes can be designed to the 5′ (exons 2-10), central (exon 45-51) and 3′ (exons 64-75) regions of the dp427 transcript to allow single transcript multiplex FISH (A). Use of these probes in mature skeletal muscle tissue (B) or myotubes of developing (E16.5) embryos (C) reveals a consistent pattern: sarcoplasmic dp427 transcripts generate punctate foci of all three probes (arrowheads) while myonuclei show intense 5′ probe labelling, slightly less intense middle probe foci, and minimal, exclusively punctate labelling with 3′ probe (schematic below C iii). A transcriptional model whereby most dystrophin transcripts are nascent, and mature mRNAs are relatively short-lived, is consistent with this pattern (D), and with transcript imbalance shown in figure 2. Expression of the shorter isoform dp140 within the embryonic (E16.5) spinal cord (E) produces prominent nuclear foci of middle probe, but not 5′ probe, again consistent with a model whereby transcripts are predominantly nascent (F).

Figure 4: Overproduction with post-transcriptional control circumvents transcriptional delay
A conventional model, where transcriptional initiation matches mRNA demand, is sufficient under steady-state, basal conditions even with a 16-hour transcription time (A), but increases in demand cannot be met over shorter timescales (B), and similarly a return to basal demand is also delayed (C). A model where transcription is always active, and always in excess of demand, with levels controlled post-transcriptionally via degradation (D) is constitutively wasteful, but increases in demand (E) can be readily met over rapid timescales simply by reducing degradation. Similarly, a return to basal levels can be rapidly effected by increasing degradation (F).

Figure 5: qPCR quantification under the unconventional dystrophin transcriptional model
cDNA synthesis: Under this dystrophin transcriptional model, most transcripts are nascent rather than mature (A). Following mRNA isolation (B) only the full-length fraction of total dystrophin mRNAs carry polyA tails (C, darker lines), while incomplete transcripts do not (lighter lines). cDNA synthesis via random priming (D) captures all dystrophin sequence (albeit fragmented) while oligo dT-directed priming (E) precludes any capture of nascent sequence, and moreover biases toward polyA-adjacent 3′ sequence. Assessment of NMD: This transcriptional model influences assessment of nonsense-mediated decay, as only mature transcripts are subject to degradation (F). Assuming transcriptional initiation remains unchanged, both healthy (G) and dystrophic (H) RNA isolates will contain large numbers of nascent transcripts (i) which will be retained following random-primed cDNA synthesis (ii), and consequently, qPCR directed to 5′ sequence will report little or no difference in measured expression, while changes in 3′ sequence will be more profound (iii): equivalent WT levels are shown as faint bars (H, iii). These differences become clearer when expressed as fold changes (I). Changes in transcriptional initiation (J) will instead produce en bloc reductions in all measured sequences (K, L), resulting in more consistent fold changes regardless of sequence position (M).

Figure 6: Lengthy transcription times influence responses to pharmacological intervention
Under basal conditions (A), most dystrophin transcripts are nascent and thus measured levels of 5′ sequence are substantially greater than 3′. After 6 hours of transcriptional initiation blockade (B), only measured levels of 5′ sequence report reductions from basal levels: transcripts initiated prior to blockade persist and are not affected, thus levels of central or 3′ sequence are unchanged. After 6 hours of pharmacological washout (C), levels of 5′ and central sequence report changes, while 3′ sequence does not: initiation has resumed but a ‘gap’ in the transcriptional procession persists, and transcripts initiated prior to the beginning of the experiment have still not reached completion. 18 hours after the start of the experiment (D), levels of 5′ sequence remain reduced while changes in central sequence become more marked, and 3′ sequence levels drop profoundly. A full 24 hours after the start of the experiment (E) levels of 5′ and central sequence begin to return to basal values, while 3′ sequence remains markedly lower. Blockade of mRNA degradation (F) will increase fraction of mature transcripts, leading to en bloc increases in all sequence regions. Fold changes will be more prominent in 3′ sequence, reflecting lower initial levels.

Figure 7: Quantifying exon skipping and accounting for exon distribution
(A) Under this transcriptional model, use of antisense oligonucleotides to 'skip' exons at the transcriptional level results in nascent transcripts bearing skipped (green regions) or unskipped (red regions) sequence (i). Only skipped transcripts escape NMD and thus represent the bulk of measured 3′ sequence. This influences measured skipping efficiency (iii): only comparison of skipped sequence with sequence close to the skipping site correctly reflects true efficiency (here 50%: dotted line). Comparison to 5′ sequence underestimates efficiency, while comparison to 3′ sequence will markedly overestimate efficiency. Exons are not distributed equidistantly along the dystrophin locus (B), and thus some sequence regions emerge markedly more rapidly than others. X axis indicates bases of genomic sequence, bars represent exon positions. Bar heights represent exon lengths: first exons are indicated. Non-muscle first exons are shown below the X axis for clarity. (C) timeline for sequence transcription: four hours are required to reach exon 6, while exons 10-41 emerge over ~1 hour. All dp71 sequence (unique first exon and exons 63-79) is transcribed similarly rapidly. This renders some exonic regions more susceptible to transcript imbalance than others (D, E): assuming comparable transcriptional initiation, both healthy (light bars) and dystrophic (dark bars) mRNAs are predominantly nascent, while only mature dystrophic mRNAs are subject to NMD. Dystrophic reductions in 3′ sequence are profound and report dramatic fold changes, while reductions in 5′ sequence might be sufficiently modest to escape detection. The close genomic arrangement of exons 10-41 however results in comparable (modest) transcript imbalance over this entire region (relative exon abundances based on the model used throughout this manuscript where nascent to mature mRNAs are at a ~10:3 ratio).

Figure 8: Exon-level analysis of dystrophin expression in healthy and dystrophic muscle
(A) Individual sequencing reads (blue/purple) are mapped to genomic features, such as exons of the Dmd locus. Conventionally all reads to a given gene are summarised regardless of location (left box), however use of custom feature files allows reads to be mapped on a per-exon basis, giving both overall read counts and counts per exon (right box). Note that reads overlapping multiple exons (blue) count for both. (B) Analysis of RNAseq datasets prepared from different healthy murine muscles: Tibialis anterior (TA, light blue), Soleus (SOL, dark blue) and Extensor digitorum longus (EDL, red), all N=6. Myosin heavy chain expression is consistent with muscle fibre type distribution, with faster MYH genes enriched in faster muscles, while dystrophin expression (Dmd) is comparable regardless of muscle. Counts of Dmd 3′ UTR alone (exon 79) are similar to counts of total Dmd. (C) Exon-level reads (reads per million, RPM) along the Dmd transcript shows that most reads are to the 3′ UTR. (D) Adjusted for exon length (RPM.base^{-1}), 3′ bias in read depth is readily apparent and consistent between muscles, and when plotted against individual exon midpoints along the transcript (E), the processivity of reverse transcription can be estimated (-0.0005 log2(RPM).base^{-2}, R^2=0.89), corresponding to a 2-fold drop in reads for every 2000 bases from the 3′ end. First exon reads (F) are consistent with near-exclusive expression of dp427m. Conventional analysis of healthy (WT, dark blue, N=3) and dystrophic (ΔEx51, red, N=3) mouse muscle RNAseq data shows dystrophy-associated loss in Dmd reads (G), which exon-level analysis again confirms are chiefly represented by exon 79 sequence. A plot of RPM.base^{-1} against transcript position (H) shows loss of Dmd sequence in ΔEx51 muscle is essentially uniform across the entire length of the mRNA (with no reads to exon 51 in dystrophic muscle -see Ex51, shaded region). 3′ bias here is consistent with a 2-fold drop in reads per 7000 bases (-0.00014 log2(RPM).base^{-2}, R^2=0.57), and first exon reads (I) are again predominantly to dp427m.

Figure 9: Exon-level analysis of dystrophin expression in embryonic and neonatal brain
(A) Conventional analysis of Dmd expression in murine brains collected from embryonic day 15.5 (E15.5) to post-natal day 29 (P29) shows a progressive increase in expression (N=2 per time point). Exon-level analysis shows that ~20% of this can be attributed to exon 79 sequence alone (B), while first exon sequences reveal greater transcriptional complexity (C), with expression primarily represented by cortical full-length dystrophin (dp427c), dp140 and dp71. While both dp427c and dp71 show progressive increases in expression with age, expression of dp140 declines. Dashed line and grey box represent read threshold corresponding to stochastic noise (1-2 reads per dataset). Read counts along the transcript (D-I) show no overt 3′ bias, instead demonstrating 5′ enrichment. Read counts increase markedly at exon 45 and exon 63 (shaded regions). This data demonstrates the advantages of ribodepletion and random priming in generation of RNAseq data for analysis of dystrophin expression, and is consistent with a transcriptional model whereby substantial numbers of transcripts are present in nascent form, regardless of isoform: mapping of exonic reads from a mixed sample with expression of dp427 (J), dp140 (K) and dp71 (J) will generate a saw-tooth like pattern of expression (M). Data derived from Schmitt et al [50].

References

1. Tennyson CN, Klamut HJ, Worton RG. The human dystrophin gene requires 16 hours to be transcribed and is cotranscriptionally spliced. Nature genetics. 1995;9(2):184-90. Epub 1995/02/01. doi: 10.1038/ng0295-184. PubMed PMID: 7719347.
2. Gazzoli I, Pulyakhina I, Verwey NE, Ariyurek Y, Laros JF, t Hoen PA, et al. Non-sequential and multi-step splicing of the dystrophin transcript. RNA Biol. 2016;13(3):290-305. Epub 2015/12/17. doi: 10.1080/15476286.2015.1125074. PubMed PMID: 26670121; PubMed Central PMCID: PMCPMC4829307.
3. Warner LE, DelloRusso C, Crawford RW, Rybakova IN, Patel JR, Ervasti JM, et al. Expression of Dp260 in muscle tethers the actin cytoskeleton to the dystrophin-glycoprotein complex and partially prevents dystrophy. Human molecular genetics. 2002;11(9):1095-105. Epub 2002/04/30. doi: 10.1093/hmg/11.9.1095. PubMed PMID: 11978768.
4. Molza AE, Mangat K, Le Rumeur E, Hubert JF, Menhart N, Delalande O. Structural Basis of Neuronal Nitric-oxide Synthase Interaction with Dystrophin Repeats 16 and 17. The Journal of biological chemistry. 2015;290(49):29531-41. Epub 2015/09/18. doi: 10.1074/jbc.M115.680660. PubMed PMID: 26378238; PubMed Central PMCID: PMCPMC4705953.
5. Lai Y, Thomas GD, Yue Y, Yang HT, Li D, Long C, et al. Dystrophins carrying spectrin-like repeats 16 and 17 anchor nNOS to the sarcolemma and enhance exercise performance in a mouse model of muscular dystrophy. The Journal of clinical investigation. 2009;119(3):624-35. Epub 2009/02/21. doi: 10.1172/jci36612. PubMed PMID: 19229108; PubMed Central PMCID: PMCPmc2648692.
6. Belanto JJ, Mader TL, Eckhoff MD, Strandjord DM, Banks GB, Gardner MK, et al. Microtubule binding distinguishes dystrophin from utrophin. Proceedings of the National Academy of Sciences. 2014;111(15):5723. doi: 10.1073/pnas.1323842111.
7. Prins KW, Humston JL, Mehta A, Tate V, Ralston E, Ervasti JM. Dystrophin is a microtubule-associated protein. The Journal of cell biology. 2009;186(3):363-9. PubMed PMID: 19651889.
8. Gao QQ, McNally EM. The Dystrophin Complex: Structure, Function, and Implications for Therapy. Comprehensive Physiology. 2015;5(3):1223-39. Epub 2015/07/04. doi: 10.1002/cphy.c140048. PubMed PMID: 26140716; PubMed Central PMCID: PMCPMC4767260.
9. Muntoni F, Torelli S, Ferlini A. Dystrophin and mutations: one gene, several proteins, multiple phenotypes. Lancet Neurol. 2003;2(12):731-40. doi: Doi 10.1016/S1474-4422(03)00585-4. PubMed PMID: WOS:000186665800016.
10. Boyce FM, Beggs AH, Feener C, Kunkel LM. Dystrophin is transcribed in brain from a distant upstream promoter. Proceedings of the National Academy of Sciences of the United States of America. 1991;88(4):1276-80. Epub 1991/02/15. doi: 10.1073/pnas.88.4.1276. PubMed PMID: 1996328; PubMed Central PMCID: PMCPMC51000.
11. Gorecki DC, Monaco AP, Derry JM, Walker AP, Barnard EA, Barnard PJ. Expression of four alternative dystrophin transcripts in brain regions regulated by different promoters. Human molecular genetics. 1992;1(7):505-10. Epub 1992/10/01. doi: 10.1093/hmg/1.7.505. PubMed PMID: 1307251.
12. Klamut HJ, Gangopadhyay SB, Worton RG, Ray PN. Molecular and functional analysis of the muscle-specific promoter region of the Duchenne muscular dystrophy gene. Molecular and cellular biology. 1990;10(1):193-205. Epub 1990/01/01. doi: 10.1128/mcb.10.1.193. PubMed PMID: 2403634; PubMed Central PMCID: PMCPMC360727.
13. Nudel U, Zuk D, Einat P, Zeelon E, Levy Z, Neuman S, et al. Duchenne muscular dystrophy gene product is not identical in muscle and brain. Nature. 1989;337(6202):76-8. Epub 1989/01/05. doi: 10.1038/337076a0. PubMed PMID: 2909892.
14. Lidov HG, Selig S, Kunkel LM. Dp140: a novel 140 kDa CNS transcript from the dystrophin locus. Human molecular genetics. 1995;4(3):329-35. Epub 1995/03/01. doi: 10.1093/hmg/4.3.329. PubMed PMID: 7795584.
15. Crawford AH, Hildyard JCW, Rushing SAM, Wells DJ, Diez-Leon M, Piercy RJ. Validation of DE50-MD dogs as a model for the brain phenotype of Duchenne muscular dystrophy. Disease models & mechanisms. 2022;15(3). Epub 2022/01/13. doi: 10.1242/dmm.049291. PubMed PMID: 35019137; PubMed Central PMCID: PMCPMC8906169.
16. D'Souza VN, Nguyen TM, Morris GE, Karges W, Pillers DA, Ray PN. A novel dystrophin isoform is required for normal retinal electrophysiology. Human molecular genetics. 1995;4(5):837-42. Epub 1995/05/01. doi: 10.1093/hmg/4.5.837. PubMed PMID: 7633443.

17. Byers TJ, Lidov HG, Kunkel LM. An alternative dystrophin transcript specific to peripheral nerve. Nature genetics. 1993;4(1):77-81. Epub 1993/05/01. doi: 10.1038/ng0593-77. PubMed PMID: 8513330.

18. Bar S, Barnea E, Levy Z, Neuman S, Yaffe D, Nudel U. A novel product of the Duchenne muscular dystrophy gene which greatly differs from the known isoforms in its structure and tissue distribution. Biochem J. 1990;272(2):557-60. Epub 1990/12/01. doi: 10.1042/bj2720557. PubMed PMID: 2176467; PubMed Central PMCID: PMCPMC1149740.

19. Durbeej M, Jung D, Hjalt T, Campbell KP, Ekblom P. Transient expression of Dp140, a product of the Duchenne muscular dystrophy locus, during kidney tubulogenesis. Developmental biology. 1997;181(2):156-67. Epub 1997/01/15. doi: 10.1006/dbio.1996.8430. PubMed PMID: 9013927.

20. Doorenweerd N, Mahfouz A, van Putten M, Kaliyaperumal R, T' Hoen PAC, Hendriksen JGM, et al. Timing and localization of human dystrophin isoform expression provide insights into the cognitive phenotype of Duchenne muscular dystrophy. Scientific reports. 2017;7(1):12575-. doi: 10.1038/s41598-017-12981-5. PubMed PMID: 28974727.

21. Hildyard JCW, Crawford AH, Rawson F, Riddell DO, Harron RCM, Piercy RJ. Single-transcript multiplex in situ hybridisation reveals unique patterns of dystrophin isoform expression in the developing mammalian embryo. Wellcome Open Research. 2020;5(76). doi: 10.12688/wellcomeopenres.15762.2.

22. Jin H, Tan S, Hermanowski J, Böhm S, Pacheco S, McCauley JM, et al. The dystrotelin, dystrophin and dystrobrevin superfamily: new paralogues and old isoforms. BMC genomics. 2007;8:19-. doi: 10.1186/1471-2164-8-19. PubMed PMID: 17233888.

23. Mendell JR, Shilling C, Leslie ND, Flanigan KM, al-Dahhak R, Gastier-Foster J, et al. Evidence-based path to newborn screening for Duchenne muscular dystrophy. Annals of neurology. 2012;71(3):304-13. Epub 2012/03/28. doi: 10.1002/ana.23528. PubMed PMID: 22451200.

24. Segurel L, Wyman MJ, Przeworski M. Determinants of mutation rate variation in the human germline. Annu Rev Genomics Hum Genet. 2014;15:47-70. Epub 2014/07/09. doi: 10.1146/annurev-genom-031714-125740. PubMed PMID: 25000986.

25. Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, et al. Rate of de novo mutations and the importance of father's age to disease risk. Nature. 2012;488(7412):471-5. Epub 2012/08/24. doi: 10.1038/nature11396. PubMed PMID: 22914163; PubMed Central PMCID: PMCPMC3548427.

26. White SJ, den Dunnen JT. Copy number variation in the genome; the human DMD gene as an example. Cytogenet Genome Res. 2006;115(3-4):240-6. Epub 2006/11/25. doi: 10.1159/000095920. PubMed PMID: 17124406.

27. Ankala A, Kohn JN, Hegde A, Meka A, Ephrem CL, Askree SH, et al. Aberrant firing of replication origins potentially explains intragenic nonrecurrent rearrangements within genes, including the human DMD gene. Genome Res. 2012;22(1):25-34. Epub 2011/11/18. doi: 10.1101/gr.123463.111. PubMed PMID: 22090376; PubMed Central PMCID: PMCPMC3246204.

28. Hoffman EP, Brown RH, Jr., Kunkel LM. Dystrophin: the protein product of the Duchenne muscular dystrophy locus. Cell. 1987;51(6):919-28. Epub 1987/12/24. PubMed PMID: 3319190.

29. Sicinski P, Geng Y, Ryder-Cook AS, Barnard EA, Darlison MG, Barnard PJ. The molecular basis of muscular dystrophy in the mdx mouse: a point mutation. Science (New York, NY). 1989;244(4912):1578-80. Epub 1989/06/30. PubMed PMID: 2662404; PubMed Central PMCID: PMC2662404.

30. Bushby KM, Gardner-Medwin D. The clinical, genetic and dystrophin characteristics of Becker muscular dystrophy. I. Natural history. J Neurol. 1993;240(2):98-104. Epub 1993/02/01. doi: 10.1007/BF00858725. PubMed PMID: 8437027.

31. Comi GP, Prelle A, Bresolin N, Moggio M, Bardoni A, Gallanti A, et al. Clinical variability in Becker muscular dystrophy. Genetic, biochemical and immunohistochemical correlates. Brain. 1994;117 (Pt 1):1-14. Epub 1994/02/01. doi: 10.1093/brain/117.1.1-a. PubMed PMID: 8149204.

32. Morandi L, Mora M, Bernasconi P, Mantegazza R, Gebbia M, Balestrini MR, et al. Very small dystrophin molecule in a family with a mild form of Becker dystrophy. Neuromuscular disorders : NMD. 1993;3(1):65-70. Epub 1993/01/01. doi: 10.1016/0960-8966(93)90043-j. PubMed PMID: 8329891.

33. Farrokhi V, Walsh J, Palandra J, Brodfuehrer J, Caiazzo T, Owens J, et al. Dystrophin and mini-dystrophin quantification by mass spectrometry in skeletal muscle for gene therapy development in Duchenne muscular dystrophy. Gene therapy. 2022;29(10-11):608-15. Epub 2021/11/06. doi: 10.1038/s41434-021-00300-7. PubMed PMID: 34737451; PubMed Central PMCID: PMCPMC9068826.

34. Aartsma-Rus A, Morgan J, Lonkar P, Neubert H, Owens J, Binks M, et al. Report of a TREAT-NMD/World Duchenne Organisation Meeting on Dystrophin Quantification Methodology. J Neuromuscul Dis. 2019;6(1):147-59. Epub 2019/01/08. doi: 10.3233/JND-180357. PubMed PMID: 30614809; PubMed Central PMCID: PMCPMC6398559.

35. Godfrey C, Muses S, McClorey G, Wells KE, Coursindel T, Terry RL, et al. How much dystrophin is enough: the physiological consequences of different levels of dystrophin in the mdx mouse. (1460-2083 (Electronic)).

36. Morin A, Stantzou A, Petrova ON, Hildyard J, Tensorer T, Matouk M, et al. Dystrophin myonuclear domain restoration governs treatment efficacy in dystrophic muscle. Proceedings of the National Academy of Sciences of the United States of America. 2023;120(2):e2206324120. Epub 2023/01/04. doi: 10.1073/pnas.2206324120. PubMed PMID: 36595689; PubMed Central PMCID: PMCPMC9926233.

37. Garcia-Rodriguez R, Hiller M, Jimenez-Gracia L, van der Pal Z, Balog J, Adamzek K, et al. Premature termination codons in the DMD gene cause reduced local mRNA synthesis. Proceedings of the National Academy of Sciences of the United States of America. 2020;117(28):16456-64. Epub 2020/07/04. doi: 10.1073/pnas.1910456117. PubMed PMID: 32616572; PubMed Central PMCID: PMCPMC7368324.

38. Maquat LE, Tarn WY, Isken O. The pioneer round of translation: features and functions. Cell. 2010;142(3):368-74. Epub 2010/08/10. doi: 10.1016/j.cell.2010.07.022. PubMed PMID: 20691898; PubMed Central PMCID: PMCPMC2950652.

39. Hildyard JCW, Rawson F, Wells DJ, Piercy RJ. Multiplex in situ hybridization within a single transcript: RNAscope reveals dystrophin mRNA dynamics. PLoS One. 2020;15(9):e0239467. Epub 2020/09/25. doi: 10.1371/journal.pone.0239467. PubMed PMID: 32970731; PubMed Central PMCID: PMCPMC7514052.

40. Spitali P, van den Bergen JC, Verhaart IE, Wokke B, Janson AA, van den Eijnde R, et al. DMD transcript imbalance determines dystrophin levels. FASEB J. 2013;27(12):4909-16. Epub 2013/08/27. doi: 10.1096/fj.13-232025. PubMed PMID: 23975932.

41. Tennyson CN, Shi Q, Worton RG. Stability of the human dystrophin transcript in muscle. Nucleic acids research. 1996;24(15):3059-64. Epub 1996/08/01. doi: 10.1093/nar/24.15.3059. PubMed PMID: 8760894; PubMed Central PMCID: PMCPMC146056.

42. Waldrop MA, Moore SA, Mathews KD, Darbro BW, Medne L, Finkel R, et al. Intron mutations and early transcription termination in Duchenne and Becker muscular dystrophy. Human mutation. 2022;43(4):511-28. Epub 2022/02/16. doi: 10.1002/humu.24343. PubMed PMID: 35165973; PubMed Central PMCID: PMCPMC9901284.

43. Muniz L, Nicolas E, Trouche D. RNA polymerase II speed: a key player in controlling and adapting transcriptome composition. EMBO J. 2021;40(15):e105740. Epub 2021/07/14. doi: 10.15252/embj.2020105740. PubMed PMID: 34254686; PubMed Central PMCID: PMCPMC8327950.

44. Singh J, Padgett RA. Rates of in situ transcription and splicing in large human genes. Nat Struct Mol Biol. 2009;16(11):1128-33. Epub 2009/10/13. doi: 10.1038/nsmb.1666. PubMed PMID: 19820712; PubMed Central PMCID: PMCPMC2783620.

45. Hildyard JCW, Riddell DO, Harron RCM, Rawson F, Foster EMA, Massey C, et al. The skeletal muscle phenotype of the DE50-MD dog model of Duchenne muscular dystrophy. Wellcome Open Res. 2022;7:238. Epub 2023/03/04. doi: 10.12688/wellcomeopenres.18251.1. PubMed PMID: 36865375; PubMed Central PMCID: PMCPMC9971692.

46. Hildyard JCW, Wells DJ, Piercy RJ. Identification of qPCR reference genes suitable for normalising gene expression in the developing mouse embryo. Wellcome Open Res. 2021;6:197. Epub 2022/05/06. doi: 10.12688/wellcomeopenres.16972.1. PubMed PMID: 35509373; PubMed Central PMCID: PMCPMC9024131.

47. Donandt T, Todorow V, Hintze S, Graupner A, Schoser B, Walter MC, et al. Nuclear Small Dystrophin Isoforms during Muscle Differentiation. Life. 2023;13(6). doi: 10.3390/life13061367.

48. Terry EE, Zhang X, Hoffmann C, Hughes LD, Lewis SA, Li J, et al. Transcriptional profiling reveals extraordinary diversity among skeletal muscle tissues. Elife. 2018;7. Epub 2018/05/29. doi: 10.7554/eLife.34613. PubMed PMID: 29809149; PubMed Central PMCID: PMCPMC6008051.

49. Chemello F, Wang Z, Li H, McAnally JR, Liu N, Bassel-Duby R, et al. Degenerative and regenerative pathways underlying Duchenne muscular dystrophy revealed by single-nucleus RNA sequencing. Proceedings of the National Academy of Sciences of the United States of America. 2020;117(47):29691-701. Epub 2020/11/06. doi: 10.1073/pnas.2018391117. PubMed PMID: 33148801; PubMed Central PMCID: PMCPMC7703557.

50. Schmitt BM, Rudolph KL, Karagianni P, Fonseca NA, White RJ, Talianidis I, et al. High-resolution mapping of transcriptional dynamics across tissue development reveals a stable mRNA-tRNA interface. Genome Res. 2014;24(11):1797-807. Epub 2014/08/15. doi: 10.1101/gr.176784.114. PubMed PMID: 25122613; PubMed Central PMCID: PMCPMC4216921.

51. Barthelemy F, Defour A, Levy N, Krahn M, Bartoli M. Muscle Cells Fix Breaches by Orchestrating a Membrane Repair Ballet. J Neuromuscul Dis. 2018;5(1):21-8. Epub 2018/02/27. doi: 10.3233/JND-170251. PubMed PMID: 29480214; PubMed Central PMCID: PMCPMC5836414.
52. Carmeille R, Bouvet F, Tan S, Croissant C, Gounou C, Mamchaoui K, et al. Membrane repair of human skeletal muscle cells requires Annexin-A5. Biochimica et biophysica acta. 2016;1863(9):2267-79. Epub 2016/06/12. doi: 10.1016/j.bbamcr.2016.06.003. PubMed PMID: 27286750.
53. Hershey JWB, Sonenberg N, Mathews MB. Principles of Translational Control. Cold Spring Harb Perspect Biol. 2019;11(9). Epub 2018/07/01. doi: 10.1101/cshperspect.a032607. PubMed PMID: 29959195; PubMed Central PMCID: PMCPMC6719596.
54. Larsen CA, Howard MT. Conserved regions of the DMD 3' UTR regulate translation and mRNA abundance in cultured myotubes. Neuromuscular disorders : NMD. 2014;24(8):693-706. Epub 2014/06/15. doi: 10.1016/j.nmd.2014.05.006. PubMed PMID: 24928536; PubMed Central PMCID: PMCPMC4114305.
55. Hogg JR, Goff SP. Upf1 senses 3'UTR length to potentiate mRNA decay. Cell. 2010;143(3):379-89. Epub 2010/10/30. doi: 10.1016/j.cell.2010.10.005. PubMed PMID: 21029861; PubMed Central PMCID: PMCPMC2981159.
56. Greener MJ, Sewry CA, Muntoni F, Roberts RG. The 3'-untranslated region of the dystrophin gene - conservation and consequences of loss. Eur J Hum Genet. 2002;10(7):413-20. Epub 2002/07/11. doi: 10.1038/sj.ejhg.5200822. PubMed PMID: 12107815.
57. Roy B, Jacobson A. The intimate relationships of mRNA decay and translation. Trends Genet. 2013;29(12):691-9. Epub 2013/10/05. doi: 10.1016/j.tig.2013.09.002. PubMed PMID: 24091060; PubMed Central PMCID: PMCPMC3854950.
58. Verhaart IEC, van Vliet-van den Dool L, Sipkens JA, de Kimpe SJ, Kolfschoten IGM, van Deutekom JCT, et al. The Dynamics of Compound, Transcript, and Protein Effects After Treatment With 2OMePS Antisense Oligonucleotides in mdx Mice. Molecular therapy Nucleic acids. 2014;3(2):e148-e. doi: 10.1038/mtna.2014.1. PubMed PMID: 24549299.
59. Novak JS, Spathis R, Dang UJ, Fiorillo AA, Hindupur R, Tully CB, et al. Interrogation of Dystrophin and Dystroglycan Complex Protein Turnover After Exon Skipping Therapy. J Neuromuscul Dis. 2021;8(s2):S383-S402. Epub 2021/09/28. doi: 10.3233/JND-210696. PubMed PMID: 34569969; PubMed Central PMCID: PMCPMC8673539.
60. Wilton SD, Dye DE, Blechynden LM, Laing NG. Revertant fibres: a possible genetic therapy for Duchenne muscular dystrophy? Neuromuscular Disorders. 1997;7(5):329-35. doi: 10.1016/s0960-8966(97)00058-8.
61. Wilton SD, Dye DE, Laing NG. Dystrophin gene transcripts skipping the mdx mutation. Muscle & nerve. 1997;20(6):728-34. doi: 10.1002/(SICI)1097-4598(199706)20:6<728::AID-MUS10>3.0.CO;2-Q.
62. Bouge AL, Murauer E, Beyne E, Miro J, Varilh J, Taulan M, et al. Targeted RNA-Seq profiling of splicing pattern in the DMD gene: exons are mostly constitutively spliced in human skeletal muscle. Sci Rep. 2017;7:39094. Epub 2017/01/04. doi: 10.1038/srep39094. PubMed PMID: 28045018; PubMed Central PMCID: PMCPMC5206723.
63. Lederfein D, Levy Z, Augier N, Mornet D, Morris G, Fuchs O, et al. A 71-kilodalton protein is a major product of the Duchenne muscular dystrophy gene in brain and other nonmuscle tissues. Proceedings of the National Academy of Sciences of the United States of America. 1992;89(12):5346-50. Epub 1992/06/15. doi: 10.1073/pnas.89.12.5346. PubMed PMID: 1319059; PubMed Central PMCID: PMCPMC49288.
64. Rau F, Laine J, Ramanoudjame L, Ferry A, Arandel L, Delalande O, et al. Abnormal splicing switch of DMD's penultimate exon compromises muscle fibre maintenance in myotonic dystrophy. Nat Commun. 2015;6:7205. Epub 2015/05/29. doi: 10.1038/ncomms8205. PubMed PMID: 26018658; PubMed Central PMCID: PMCPMC4458869.
65. Naidoo M, Anthony K. Dystrophin Dp71 and the Neuropathophysiology of Duchenne Muscular Dystrophy. Mol Neurobiol. 2019. doi: 10.1007/s12035-019-01845-w.
66. Hansen KD, Brenner SE, Dudoit S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic acids research. 2010;38(12):e131. Epub 2010/04/17. doi: 10.1093/nar/gkq224. PubMed PMID: 20395217; PubMed Central PMCID: PMCPMC2896536.
67. Mason PB, Struhl K. Distinction and relationship between elongation rate and processivity of RNA polymerase II in vivo. Mol Cell. 2005;17(6):831-40. Epub 2005/03/23. doi: 10.1016/j.molcel.2005.02.017. PubMed PMID: 15780939.
68. Carey LB. RNA polymerase errors cause splicing defects and can be regulated by differential expression of RNA polymerase subunits. Elife. 2015;4. Epub 2015/12/15. doi: 10.7554/eLife.09945. PubMed PMID: 26652005; PubMed Central PMCID: PMCPMC4868539.

69. Gout JF, Thomas WK, Smith Z, Okamoto K, Lynch M. Large-scale detection of in vivo transcription errors. Proceedings of the National Academy of Sciences of the United States of America. 2013;110(46):18584-9. Epub 2013/10/30. doi: 10.1073/pnas.1309843110. PubMed PMID: 24167253; PubMed Central PMCID: PMCPMC3832031.
70. Lynch M. Evolution of the mutation rate. Trends Genet. 2010;26(8):345-52. Epub 2010/07/03. doi: 10.1016/j.tig.2010.05.003. PubMed PMID: 20594608; PubMed Central PMCID: PMCPMC2910838.
71. Dumont NA, Wang YX, von Maltzahn J, Pasut A, Bentzinger CF, Brun CE, et al. Dystrophin expression in muscle stem cells regulates their polarity and asymmetric division. Nature Medicine. 2015;21(12):1455-63. doi: 10.1038/nm.3990.
72. Schultz E, Jaryszak DL, Valliere CR. Response of satellite cells to focal skeletal muscle injury. Muscle & nerve. 1985;8(3):217-22. Epub 1985/03/01. doi: 10.1002/mus.880080307. PubMed PMID: 4058466.
73. Kaczmarek A, Kaczmarek M, Cialowicz M, Clemente FM, Wolanski P, Badicu G, et al. The Role of Satellite Cells in Skeletal Muscle Regeneration-The Effect of Exercise and Age. Biology (Basel). 2021;10(10). Epub 2021/10/24. doi: 10.3390/biology10101056. PubMed PMID: 34681155; PubMed Central PMCID: PMCPMC8533525.
74. de Leon MB, Montanez C, Gomez P, Morales-Lazaro SL, Tapia-Ramirez V, Valadez-Graham V, et al. Dystrophin Dp71 expression is down-regulated during myogenesis: role of Sp1 and Sp3 on the Dp71 promoter activity. The Journal of biological chemistry. 2005;280(7):5290-9. Epub 2004/11/20. doi: 10.1074/jbc.M411571200. PubMed PMID: 15550398.
75. Farea M, Rani AQM, Maeta K, Nishio H, Matsuo M. Dystrophin Dp71ab is monoclonally expressed in human satellite cells and enhances proliferation of myoblast cells. Sci Rep. 2020;10(1):17123. Epub 2020/10/15. doi: 10.1038/s41598-020-74157-y. PubMed PMID: 33051488; PubMed Central PMCID: PMCPMC7553993.
76. Scherer S. A Short Guide to the Human Genome: Cold Spring Harbor Laboratory Press; 2009.
77. Galaxy C. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. Nucleic acids research. 2022;50(W1):W345-W51. Epub 2022/04/22. doi: 10.1093/nar/gkac247. PubMed PMID: 35446428; PubMed Central PMCID: PMCPMC9252830.
78. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37(8):907-15. Epub 2019/08/04. doi: 10.1038/s41587-019-0201-4. PubMed PMID: 31375807; PubMed Central PMCID: PMCPMC7605509.
79. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166-9. Epub 2014/09/28. doi: 10.1093/bioinformatics/btu638. PubMed PMID: 25260700; PubMed Central PMCID: PMCPMC4287950.