Towards development of a statistical framework to evaluate myotonic dystrophy type 1 mRNA biomarkers in the context of a clinical trial

Adam Kurkiewicz[1*], Anneli Cooper[3¤], Emily McIlwaine[3], Sarah A. Cumming[3], Berit Adam[3], Ralf Krahe[4], Jack Puymirat[5], Benedikt Schoser[6], Lubov Timchenko[7], Tetsuo Ashizawa [8], Charles Thornton [9], Simon Rogers[2☯], John McClure[1☯], Darren G Monckton[3☯]

1 Institute of Cardiovascular and Medical Sciences, University of Glasgow, Glasgow, United Kingdom

2 School of Computing Science, University of Glasgow, Glasgow, United Kingdom

3 Institute of Molecular Cell and Systems Biology, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, United Kingdom

4 Department of Genetics, University of Texas, MD Anderson Cancer Center, Houston, TX, USA

5 Laboratory of Human Genetics, CHUL Medical Research Centre, University of Laval, Quebec City, QC, Canada

6 Friedrich Baur Institute, Department of Neurology, Ludwig Maximilians University, Munich, Germany

7 Department of Pediatrics, Division of Neurology, Cincinnati Children's Hosptial, University of Cincinnati, College of Medicine, Cincinnati Ohio, USA

8 VA Medical Center, Houston, Texas, USA

9 University of Rochester, Medical Center School of Medicine and Dentistry, Rochester, New York, USA

☯These authors contributed equally to this work.

¤Institute of Biodiversity, Animal Health and Comparative Medicine, University of Glasgow, Glasgow, United Kingdom

* Adam.Kurkiewicz@glasgow.ac.uk

Abstract

Myotonic dystrophy type 1 (DM1) is a rare genetic disorder, characterised by muscular dystrophy, myotonia, and other symptoms. DM1 is caused by the expansion of a CTG repeat in the 3'-untranslated region of *DMPK*. Longer CTG expansions are associated with greater symptom severity and earlier age at onset. The primary mechanism of pathogenesis is thought to be mediated by a gain of function of the CUG-containing RNA, that leads to *trans*-dysregulation of RNA metabolism of many other genes. Specifically, the alternative splicing (AS) and alternative polyadenylation (APA) of many genes is known to be disrupted. In the context of clinical trials of emerging DM1 treatments, it is important to be able to objectively quantify treatment efficacy at the level of molecular biomarkers. We show how previously described candidate mRNA biomarkers can be used to model an effective reduction in CTG length, using modern high-dimensional statistics (machine learning), and a blood and muscle mRNA microarray dataset. We show how this model could be used to detect treatment effects in the context of a clinical trial.

Introduction

Myotonic dystrophy type 1

Myotonic dystrophy type 1 (DM1) is an autosomal dominant trinucleotide repeat disorder, caused by an expanded CTG repeat in the 3' UTR of the *dystrophia myotonica protein kinase* (*DMPK*) gene [1]. Transcription of *DMPK* in affected individuals produces a toxic, GC-rich mRNA molecule, which results in dysregulation of several RNA binding factors, including proteins MBNL1, MBNL2, MBNL3, CELF1 (CUGBP1), HNRNPH1 and STAU1 (Staufen1) [2, 3]. The pathomechanism of the dysregulation of splicing factor MBNL1 is perhaps best understood, with MBNL1 sequestration to the toxic *DMPK* RNA product resulting in alternative splicing defects of pre-mRNAs of multiple genes, including the chloride channel (*CLCN1*), brain microtubule-associated tau (*MAPT*) and insulin receptor (*INSR*) [4]. Such alternative splicing (AS) defects are generally believed to be a major contributing factor of clinical symptoms of DM1, such as myotonia (*CLCN1*) or abnormal glucose response (*INSR*),

and have been postulated to play a role in cardiac conduction defects (*RYR2*, *SERCA2*, *TNNT2*) [5]. AS defects may underlie other clinical symptoms of DM1, including muscle wasting, cataracts, hypersomnia, gastrointestinal abnormalities, as well as premature baldness and testicular atrophy in males [2, 6]. The severity of symptoms is positively correlated with CTG repeat length [2].

The toxic *DMPK* transcript in DM1-affected individuals has been identified as an active target for theraupetic intervation [7], and it is expected that breakdown of the toxic mRNA will result in at least partial reversal of DM1-induced AS changes and other known and unknown DM1-induced biomolecular pathologies.

Spliceopathy of DM1 is an active area of research, with novel splicing defects being continuously reported. Nakamori *et al.* [8] identified a set of 41 genes, which are mis-spliced in DM1, suggesting these genes as potential biomarkers of DM1. Batra *et al.* [9] identified 80 genes, whose expression indicates disrupted APA, reusing the (human) dataset of Nakamori *et al.* and using other datasets, including data from mouse models.

Predictors in genetics research

A widespread paradigm in biological and clinical research is the case-control study, using frequentist statistics tools focusing on hypothesis testing (inference). Examples of such designs include Genome Wide Association Studies (GWAS) [10], placebo and active control clinical trial designs [11], non-inferiority designs [12] or heredity designs based on twin studies. It is reported that designs of as many as 70% of studies published in leading medical journals use at most the following three statistical tests as part of their design: Student's t-test, Fisher's exact test, and the Chi-square test [13].

A possible alternative to the focus on hypothesis testing is building predictors or classifiers, which produce a numerical estimate of a given trait (height, size of the DM1 trinucleotide expansion) or effect size, predict participant's category (such as affected/unaffected), or estimate the effect size of the treatment, given a set of independent variables (e.g. genotypes, mRNA profiles, *etc.*). If necessary, the efficacy of these predictors/classifiers can be evaluated using traditional frequentist tools, such as p-values.

Predictors have been successfully applied in genetics research. For example, Lello *et al.* [14] report predicting human height from genotype data, obtained using human Single Nucleotide Polymorphism (SNP) microarrays, to within a few centimeters for most participants in their sample. This level of accuracy is achieved due to a very large sample size of nearly half a million individuals. A review by Van Raden *et al.* [15] reports ability to predict dairy output of certain cattle breeds with R^2 of 49%, using non-linear models based on SNP microarrays. The work of Azencott *et al.* [16] allows one to incorporate prior information about biological networks into the predictive model, increasing prediction accuracy. In one of the direct motivations behind our research here, Lee *et al.* [17] report being able to predict most of Huntington disease trinucleotide repeat size using mRNA profiling of lymphoblastoid cell lines. Here, we demonstrate a further application of predictors in genetics research, by constructing a predictor, which can produce a numerical estimate of a participant's DM1 CTG repeat length (measured from blood) from an mRNA profile (obtained from muscle), and demonstrate its usefulness in the context of a hypothetical clinical trial of a DM1 treatment.

Prior identification of alternative splicing and alternative polyadenylation events in muscles of DM1-affected individuals

Our primary reference is the work of Nakamori *et al.* [8], who identified 42 genes exhibiting AS defects in DM1. Briefly, the methodology of the study was as follows: muscle tissue (patients: four biceps, two quadriceps, one tibialis anterior, one diaphragm; controls: eight vastus lateralis) were sampled post-mortem from eight DM1 affected individuals and eight unaffected controls. Selection of the postmortem DM1 samples were based on high integrity of RNA present in the sample and presence of splicing misregulation of *INSR* and *AP2A1*, and then compared to quadriceps biopsy samples from eight healthy controls.

mRNA was extracted from the samples, purified and hybridized to GeneChip™ Human Exon 1.0 ST microarrays. Putative AS defects were identified using a mixture of existing methods, such as Affymetrix's PLIER, DABG and Alternative Transcript Analysis Methods for Exon Arrays, and new methods proposed and described by the authors. Identified putative defects were validated using RT-PCR in 50 DM1 subjects,

yielding 42 genes with confirmed AS defects.

Nakamori *et al.* report several technical obstacles with this approach, with the initial version of their pipeline suffering from as many as 80% putative AS defects failing to replicate with Reverse Transcription-PCR. Further analysis suggested that this occurred when "signal intensity for the entire transcript or a particular exon was low", "overall expression of a transcript was strongly up- or down- regulated in DM1 relative to normal controls", or "signal intensity of an exon was inappropriately high relative to other exons in the same transcript" [8].

Batra *et al.* [9] re-used the dataset of Nakamori *et al.* [8] and used similar techniques to filter down the data, in a search for genes with dysregulated APA. Their selection criteria focused on probesets with over 2-fold change in DM1 or DM2 vs. unaffected controls, excluding genes that were represented by ≤ 5 probesets, retrogenes and non-protein coding genes, which resulted in pre-selection of 438 probesets. The authors performed visual inspection of all pre-selected probesets identifying 123 APA events belonging to 80 genes.

Evaluating Potential Biomarkers of DM1

We propose that predictors are a suitable statistical tool, which can find applications in DM1 research and clinical practice. As described before, DM1 case/control status [8, 9] leaves a discernible pattern in the mRNA profiles of muscle samples. In this research, rather than to work with the DM1 status as a binary variable, we look at DM1 as a spectrum disease, severity of which is quantified by the length of the DM1 CTG repeat in any individual patient. We propose that it is possible to capture the effect, which the length of this repeat has on mRNA expression in muscle into a simple statistical model, based on linear regression. Using the model we can predict the size of the DM1 CTG repeat from the mRNA profile significantly better than a random predictor. We propose that the model can serve as a valuable tool in evaluating efficacy of any treatment for DM1 as such treatment enters pre-clinical or clinical trials, by enabling investigators to directly quantify the treatment effect as measured by effective reduction of DM1 CTG repeat length.

We would like to stress here that "effective reduction" in the case of most candidate

treatments will not be an actual reduction in the repeat length (with notable exception of candidate treatments based on gene editing). Rather, an "effective reduction" is a reduction of "effective repeat length", i.e. repeat length as judged by the degree of splicing changes. This can be demonstrated with an example of a patient with a certain pre-treatment repeat length, and both physiological and molecular symptoms characteristic for that repeat length. If such patient were to undergo an effective treatment we would expect these symptoms to be partially reversed, and our predictive framework to predict a shorter repeat length than the patient's actual repeat length. Subject to correctness of our understanding of the molecular pathophysiology of the splicing changes, which we rely on to predict the effective repeat length, we expect that the reversal of the splicing changes would occur immediately after the release of inactivated AS and APA factors, such as MBNL1. This release should in turn happen immediately after a candidate therapeutic were to reach a clinically significant level in the relevant tissue. We expect this to happen on a timescale of days to weeks, unlike the reversal of physiological symptoms, which we would expect to happen on longer timescales.

Materials and methods

The dataset

As part of the *Dystrophia Myotonica* Biomarker Discovery Initiative (DMBDI) a dataset was obtained from 35 participants, including 31 DM1 cases and four unaffected controls. All DM1 cases in this research were heterozygous for the abnormally expanded CTG repeat. The mode of the length of the DM1 CTG expansion (Modal Allele Length, MAL) was determined by small-pool PCR of blood DNA for 35/36 patients [18]. For this work we did not attempt to measure the repeat length from muscle, due to a very high degree of repeat instability in muscle cells [3] and associated difficulties in its experimental measurement. One patient refused blood donation. For each of the 35 blood-donating patients mRNA expression profiling of blood was performed using Affymetrix GeneChip™ Human Exon 1.0 ST microarray. For 28 of 36 patients a successful quadriceps muscle biopsy was obtained. The muscle tissue was mRNA

profiled using the same type of microarray. In total, a complete set of samples (blood and muscle) was obtained for 27 of 36 patients. mRNA profiling was carried out by the GeneLogic service lab (on a fee-for-service basis) using standard Affymetrix hybridisation protocol.

Principal Component Analysis (PCA) was performed on both blood and muscle profiles, and visual inspection was carried out to detect the presence of outliers. Zero outliers in the blood batch and four outliers in the muscle batch were identified and expression profiling was repeated for these patients. For the analysis we used only the repeated profiles.

It should be noted that our dataset differs from the dataset collected by Nakamori *et al.* in the following ways:

1. We did not perform confirmatory RT-PCR analyses.
2. Our dataset includes both blood and muscle tissue samples.
3. The number of mRNA-profiled participants is about twice the size of the discovery dataset of Nakamori *et al.*, however, we did not perform a follow-up validation study.
4. Unlike Nakamori *et al.* we have additional information to participant's case-control status, specifically, the mode of the CTG repeat length from blood (MAL). We also estimate the repeat length at birth (Progenitor Allele Length, PAL) using a previously developed method [19].
5. We sample from the same muscle group (quadriceps), as opposed to from a wide range of muscle groups (biceps, quadriceps, tibialis anterior, diaphragm, vastus lateralis) for all study participants, which eliminates a potentially important confounding effects.
6. All our participants are alive at the time of sample collection, which eliminates another potential confounder, but other potential confounders still remain, see Limitations.

The dataset is deposited to Array Express with accession number E-MTAB-7983.

Affymetrix GeneChip™ Human Exon 1.0 ST microarray

The microarray chip used was the Affymetrix GeneChip™ Human Exon 1.0 ST microarray, which contains over 5 million probes, *i.e.* short cDNA sequences, which target transcribed regions with high specificity. Each probe in the HuEx chip contains precisely 25 nucleotides. Consequently, transcribed regions of interest, which are shorter than 25 nucleotides, for example some of the short exons, are not targeted. This is an important limitation in the context of DM1, as some AS events previously described in DM1, such as AS of exon 5 in cardiac troponin T (*TNNT2*) [20] cannot be detected.

Each continuous section of DNA can be targeted by a collection of up to four probes, referred to as probeset. DNA sections targeted by the chip include known and suspected coding exons in known and suspected genes, as well as non-coding genomic features, including 5' and 3' UTRs, and various other types of transcribed or hypothetically transcribed DNA (miRNA, rRNA, pseudo-genes, *etc.*). All probes targeting such a region belong to a single probeset. Probesets are further grouped into transcription clusters, which correspond to the entire genes.

Data preparation and analysis

We designed and built a pipeline, programmed in Python, which has the following data preparation capabilities: Reading raw Affymetrix CEL v4 files (peer reviewed and merged into Biopython [21, 22]); quantile normalisation and log2 transformation of intensity data; strict annotation of Affymetrix probes using GENECODE v26 lift 37 through selecting probes corresponding to annotated GENECODE transcripts of type "protein_coding", annotated genes of type "protein_coding" and exons of type "CDS" or "UTR". Appendix S1 Appendix gives full source code, user manual and additional explanation of each step of this pipeline.

The pipeline's final output are two directories: "experiment_muscle" and "experiment_blood", each containing 19,826 files, whose filenames correspond to HGNC gene names. The following is a two-line excerpt from one of such files, "experiment_blood/TNNI1". Data for several patients has been removed to enhance clarity:

gene_name	probeset_id	seq5to3plus	chrom	strand

```
    ↪   genecode_left   genecode_right  x       y
    ↪ patient_111747589       patient_117440822       ...
    ↪ patient_896445336
TNNI1   2450836 TGGCCTGTGCCTCGCCGTAGACTGC       chr1    -
    ↪        201390827        201390851       195     793
    ↪ 6.69115092487   6.44442470429   ...     6.86989190909
```

In each file, the first row is a header containing tab-separated names of data or metadata types contained in a given column. Below we give a brief description of some of the data or metadata types:

- seq5to3 – unlike Affymetrix we always report the sequence in 5' to 3' direction, and always with regards to the plus strand, even if the coding sequence is contained on the minus strand.
- genecode_left, genecode_right – these are genomic coordinates as reported by a reference assembly (GRCh37). Following convention, first coordinate is 1-based, and coordinates are left-, right- inclusive.
- x, y give the x and the y coordinates of probes on the chip.
- patient_* – these are quantile normalised and log2-transformed intensities at the given probe for the given study participant.

Each subsequent row contains data and metadata for a single probe.

We develop a predictive model, which closely follows that of Lee *et al.* [17]. We work with pre-selected sets of genes that act as candidate biomarkers. For this purpose, we look at the following collections of genes:

1. A previously identified selection of genes, listed in S2 Appendix and identified by Nakamori et al. [8] as genes whose AS is disrupted in DM1. We codename these genes "DM1-AS".
2. A previously identified selection of genes, listed in S3 Appendix and identified by Batra et al. [9] as genes whose AP is disrupted in DM1. We codename these genes "DM1-APA". The overlap of this list of genes with DM1-AS is a list of two genes: LDB3, MBNL2.

3. A single gene (which also belongs to DM1-APA), Troponin I1, slow skeletal type, TNNI1. We codename this single-gene collection "TNNI1". The gene was chosen post-hoc.

4. All human genes, as identified in data preparation step. We codename this collection "ALL".

Then we execute eight separate statistical analyses (four muscle and four blood analyses), based on these groups of genes. Each analysis was carried out as follows:

1. We randomly split our data into two sets, the training and the testing set, each containing 70% and 30% of participants respectively. We restricted our analysis only to genes, which belonged to a particular group of genes being studied (DM1-APA, DM1-AS, TNNI1, ALL)

2. Following Lee *et al.*, we select (up to) 500 probes whose intensities across all patients in the training set are most correlated with their corresponding MAL. We work at the level of individual probes, which allows us to circumvent issues around GC-correction and probeset aggregation. Probe data are fed directly into the model.

3. Again using the training set, we trained a 2-dimensional Partial Least Squares Regression (PLSR) model on selected probes as features and corresponding MAL as the model output, which we later used to predict MAL in the testing set, again following Lee *et al.*.

4. We repeated steps 1 to 3 10,000 times.

5. We report coefficient of determination (R^2) of the predicted MAL with the measured MAL across all folds obtained in step 4.

6. We simulated a distribution of R^2 of a random predictor, and obtained a p-value for the prediction of R^2 in step 5.

To confirm that any observed signal is not a by-product or an artifact of the mathematical model used, or its implementation, we carried out the same kind of analysis using three other mathematical models:

1. lasso
2. random forest regression
3. linear regression

To relate our findings to a potential clinical setting, we present power analysis relating various potential treatment effect sizes to their detectability in future clinical trials. The power analysis was performed as follows:

1. We simulated the distribution of MAL prediction error for the best predictor.
2. We simulated the distribution of MAL post-treatment, with the assumption that MAL would be reduced by 10%, 20% or 50%, which correspond to, respectively, small, medium and large treatment effect size.
3. We established the rejection region for a null hypothesis of "no treatment effect" at $\alpha = 0.05$.
4. We simulated the power, (1 - β), for studies involving 10 to 200 participants, in increments of 10 participants.

Results & Discussion

Performance of PLSR-based MAL predictors in muscle and blood

Using our predictive model based on PLSR and a selection of candidate biomarkers: DM1-AS; DM1-APA; TNNI1 and ALL, we can report the following capabilities to predict MAL from mRNA profiles.

In muscle we can detect strong signal for some of the selected candidate biomarker sets, with the two strongest predictors, DM1-AS and TNNI1, giving us R^2 and p-values equal to 0.289, 0.320 and 0.00434, 0.00249 respectively. All values of R^2 and p-values are given in Table 1. In particular, DM1-AS looks promising as a set of biomarkers, giving weight to the findings of Nakamori *et al.* [8].

Based purely on numerical analysis of obtained results, one could declare TNNI1 as the best predictor of MAL in muscle, but care must be taken, as we chose this

particular gene post-hoc and its true predictive value might have been influenced by issues related to multiple hypotheses testing. An additional validation study would have to be performed before drawing any conclusions on the performance of TNNI1 as a DM1 biomarker.

A safer belief can be assigned to the predictive value of "DM1-AS" and "DM1-APA" as both sets of genes have been previously implicated in DM1 [9], and we can detect strong signal from both sets of genes, with models based on these genes capturing respectively 30% and 15% of MAL variance in our study group.

One has to note an interesting observation relating to the "curse of dimensionality" and the performance of our PLSR-based model. Although TNNI1 is a subset of DM1-APA, TNNI1 on its own is a much better predictor than DM1-APA, as increasing the number of genes (features) increases the dimensionality of our data and worsens the prediction delivered by the PLSR model.

A single run of 10,000 repetitions of a simulation can be visualised by plotting the predicted value of MAL against the actual, adding a small amount of random noise on the x-axis. For DM1-AS in muscle, such visualisation is given in Fig 1.

In blood we cannot detect significant signal for most candidate biomarkers, with R^2 not significantly respectively greater and lower than would be expected by chance, except in one case, DM1-APA, whose R^2 and p-value are 0.15 and 0.0564 respectively. Full R^2 and (uncorrected) p-values are given in Table 2. Poor performance of blood data acting as a predictor of repeat length is perhaps not unexpected, as all candidate biomarkers evaluated in this study are based on prior analyses of muscle data, and might not be expressed, or only lowly expressed in blood, or that expression profiles might be very different from expression profiles in muscle. Only DM1-APA can potentially carry some predictive value, however, not as strongly as in muscle.

Table 1. 10000 repetitions of a simulation predicting MAL from muscle cross-validated with a testing set separate from the training set.

	DM1-AS	DM1-APA	TNNI1	ALL
R^2	0.289	0.150	0.320	0.084
p-value	0.00434	0.0470	0.00249	0.143

Table 2. 10000 repetitions of a simulation predicting MAL from blood cross-validated with a testing set separate from the training set.

	DM1-AS	DM1-APA	TNNI1	ALL
R^2	0.0465	0.106	0.0479	0.0465
p-value	0.213	0.0564	0.207	0.213

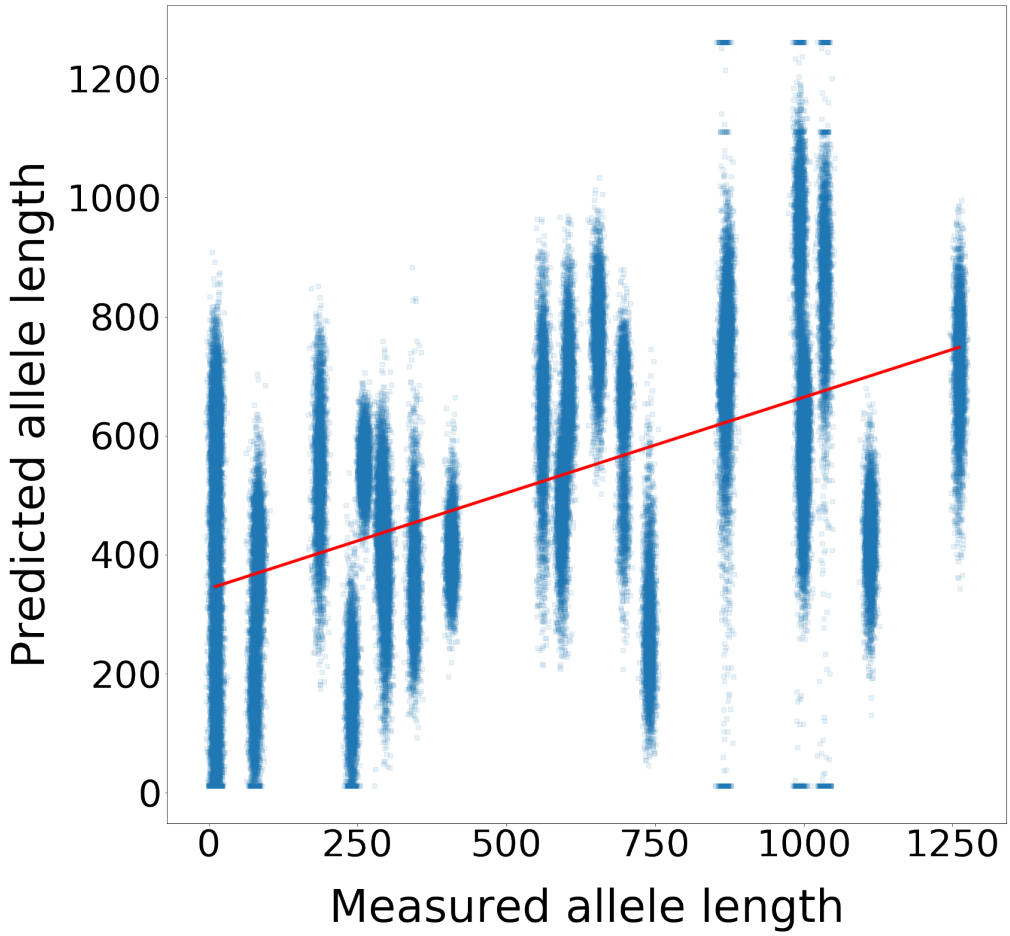

Fig 1. DM1-AS muscle MAL prediction. 10,000 repetitions of cross-validated MAL prediction from genes labeled DM1-AS from muscle for 18 training samples.

Other models

As described in Data preparation and analysis, we have used a model based on feature selection from sets of candidate mRNA biomarkers to predict the MAL of DM1 CTG repeat using PLSR. As reported in Performance of PLSR-based MAL predictors in muscle and blood, the set of genes DM1-AS is the strongest predictor if we limit our consideration to predictors chosen a-priori (*i.e.* excluding TNNI1).

A potential source of criticism could be that the effect observed is a technical effect due to the choice or the implementation of the mathematical model used (PLSR). We

thus re-ran the analysis as described before in Data preparation and analysis using three mathematically distinct models: lasso, random forest regression and linear regression.

We report the performance of these models in Table 3:

Table 3. 10000 repetitions of a simulation predicting MAL from muscle, using DM1-AS as a predicting set and a selection of mathematical models

	linear regression	PLSR	lasso	random forest
R^2	0.291	0.289	0.286	0.149
p-value	0.00418	0.00434	0.00454	0.0478

It should be noted that all models pick up statistically significant signal with both PLSR, lasso and linear regression performing almost equally well, and random forest performing about two times poorer (but still significantly better than a random predictor), which allows us to conclude that the effect observed is unlikely to be a modelling artifact.

Applying the model in a clinical setting

Let us now consider the potential application of this predictive model in the context of evaluating efficacy of a DM1 treatment. DM1 patients would undergo muscle biopsy before starting the treatment, and another biopsy after the treatment had been started and necessary biological changes to reverse DM1 symptoms had occurred. Both biopsies would be mRNA profiled, and the resulting profiles would be used to perform MAL predictions. We expect that pre-treatment prediction would correspond to the actual MAL of any given participant. We expect that post-treatment prediction of MAL would correspond to an "effective MAL", which we would expect to be lower than the "actual MAL" in affected participants, as long as the treatment is effective and DM1-induced disruption of AS or APA, as measured by DM1-AS or DM1-APA biomarkers is measurably reversed in obtained mRNA profiles. Pre-treatment and post-treatment predictions could be combined into a statistic that, given enough patients, would allow us to quantify the efficacy of the treatment at the molecular level. We discuss this idea further in the chapter Power Analysis.

In some respects, such a study could allow for better performance of the model, conceptually, it should be easier to capture DM1 specific expression changes in a setting where noise due to varied genetic backgrounds of participants can be reduced by looking

at pairs of measurements of a single participant. There are a number of details in such study design, which need to be discussed by the community and decided upon, among others:

1. The mathematical basis of the model used. We propose a PLSR-based model, and demonstrate that models based on lasso and linear regression perform similarly, but other models can also be considered, in particular the work of Azencott *et al.* [16] in the context of L_1-penalised regression (lasso) looks promising as it allows to incorporate prior biological knowledge, in the form of protein-protein interaction networks or other types of graph ontology, into the model, through the introduction of additional penalties based on discrete Laplacian, or apply alternative modelling strategies based on network flow, thereby increasing its predictive power.

2. The training/testing pre-treatment/post-treatment data split and which genes should be included in the model input.

3. The size of the study. Predictions in our study are based on 24 participants for blood and 18 for muscle. How much would the models' predictive power improve with a larger dataset?

4. Establishing the clinically relevant effect size. Pandey *et al.* [7] report various efficacies of a candidate DM1 treatment ISIS 486718 to lower toxic *DMPK* concentrations in wild-type and transgenic animal models and a range of tissues, starting with the efficacy of about 50% in cardiac muscle, through about 70% in skeletal muscle, up to about 90% in liver and skeletal muscle. However, measuring DMPK levels may not necessarily directly correspond to the efficacy of treatment to reverse symptoms, as the relationship between the quantity of the toxic transcript, splicing disruption and eventual clinical symptoms may be complex and non-linear. Conservatively, we need to expect the rate of symptom reduction to be lower than the reported 50% to 90%. A difficult open question is what minimum treatment efficacy we are willing to accept as clinically significant?

An informal setting to explore the dataset

As this work is the first presentation of the DMBDI dataset, we recognise that further work might build on the dataset in ways which differ from our approach and cannot be predicted at the current stage of our understanding of DM1-related AS/APA changes. To facilitate this, we would like to propose a tool which on one hand allows for informal, interactive and exploratory analysis of the dataset and on the other allows the flexibility of building a custom analysis – just like the one presented here.

The tool is available online [23], and is implemented as a jupyter notebook with custom visualisation of filtered and normalised DMBDI data. The flexibility of the tool comes at a cost. In order to support arbitrary bioinformatics analyses we have to support arbitrary code execution, which in turn requires protecting the tool with a password. We will share the password with any bona fide researcher upon request. A walk-through video showcasing the capabilities of the tool is available on youtube [24].

The major capability of the tool is the ability to produce "railway plots". Railway plots introduce the idea of a Manhattan plot from genomics community into transcriptomics. Each point represents statistical significance of the change of expression signal at a single probe across DM1 spectrum, as supported by experimental data. Points which belong to the same probeset, are identically coloured. See Figure 2 for an example railway plot.

In a railway plot the y-axis represents negative logs of p-values of a two-tailed tests against a null hypothesis of no expression change at a single probe across the DM1 spectrum. Figure 3 visualises one of such linear regressions for a probe belonging to probeset 245089. The logs of p-values are signed in accordance with the direction of the slope of the regression, with negative values indicating splice-out and positive values indicating a splice-in type of event.

For each of splice-out and splice-in directions in the plot, we show thresholds of statistical significance. The first pair of thresholds (faded red and blue respectively) correspond to statistical significance threshold of 0.05. The second pair of thresholds correspond to $0.05/n$, where n is the total number of probes in the plot (saturated red and blue respectively). This is analogous to the way Manhattan plots are often presented for GWAS.

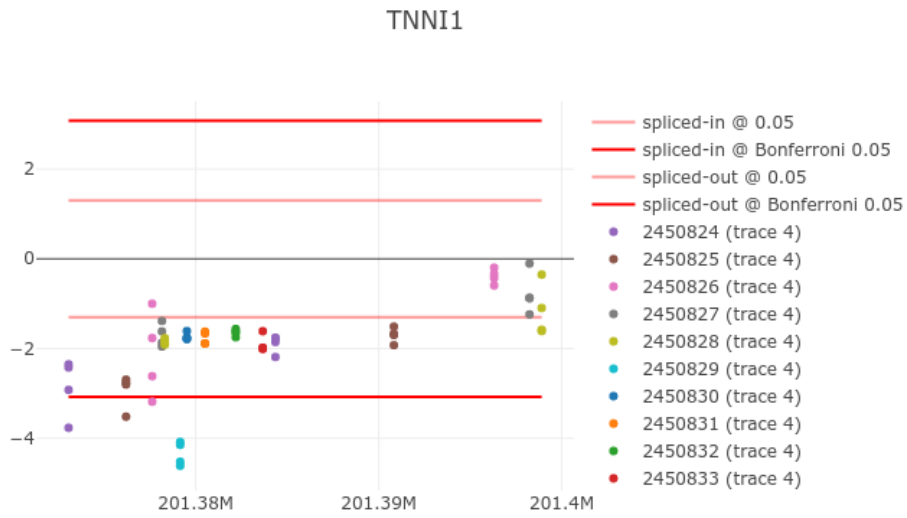

Fig 2. TNNI1 Railway plot shows an APA event at probeset 245089.

Turning our attention from statistics to biology, we can ask about a likely biological interpretation of an observed splice-out event detected at probeset 2450829. Using Ensembl [25], see figure 4), we can identify transcript TNNI1-203, which is the only GENECODE transcript featuring a probe selection region targeted by the probeset 2450829, in form of an alternative 3' UTR. This allows us to suggest that TNNI1-203 is downregulated in participants with longer DM1 repeats. GENECODE annotation further informs us that TNNI1-203 is an APPRIS P2 transcript (i.e. a candidate principal variant, a designation which comes from high level of support for functionality of the isoform), and is not a CDS incomplete transcript, which allows us to strengthen our belief in the fact that this is a biologically functional/ protein-coding transcript, which can play a role in the DM1-related AS/APA changes.

Finally, and returning back to statistics, we can ask whether high significance of the splice-out event is a result of multiplicity effect, given that the gene was chosen post-hoc from a pool of candidate biomarker genes as determined by Batra et al. and Nakamori et al. [8, 9]. A standard approach here would be to combine the data from the discovery dataset with the data from the replication dataset, compute a more powerful test, and apply multiplicity correction. This is not possible in this case as the discovery dataset, underlying both studies is a case-control dataset, whereas our dataset captures DM1

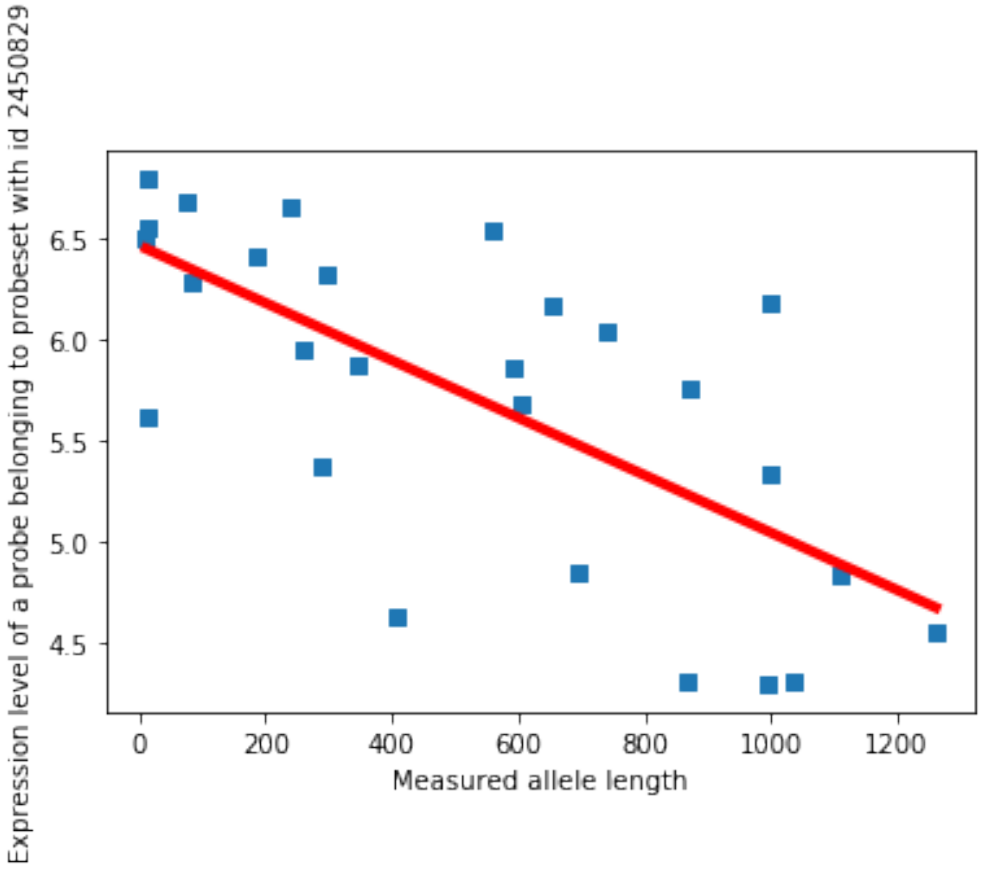

Fig 3. Linear regression of expression intensity at a single probe belonging to probeset 245089 against DM1 repeat length.

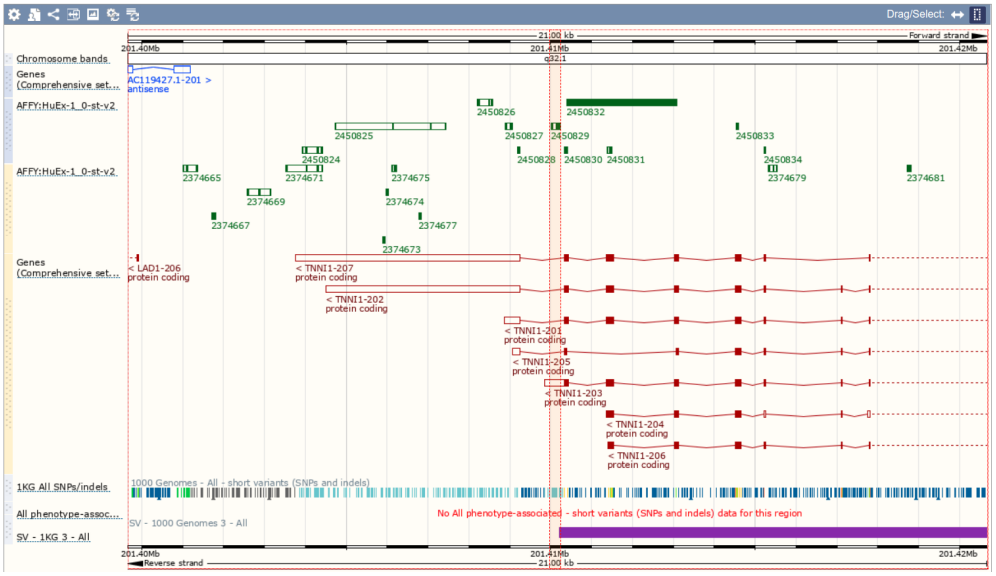

Fig 4. Visualisation of genomic coordinates of TNNI1 transcripts using Ensembl.

status as a continuous variable via DM1 CTG repeat length measurement. An alternative is to combine both p-values using, e.g. Fisher's method [26], which shows that the combined p-value against the null hypothesis of no change of signal intensity at the probeset 2450829 is 4.19×10^{-10}. Computing the Bonferroni correction with multiplicity factor of 1.4 million, equal to the total number of HuEx probesets [27], shows that the combined p-value based on the discovery dataset and our dataset is 5.88×10^{-4}. This is a strong confirmation of the significant correlation of DM1 status with TNNI1-203 downregulation. Details of the computation are available in S1 Appendix.

Power Analysis

Our final contribution, which is of critical significance in the context of any future clinical trial is a power analysis of the current model. We report power for a selection of possible treatment effect sizes (10%, 20% and 50% reduction in effective MAL on a per-patient basis) and a selection of study participants. Our power is defined as $(1 - \beta)$, where β is a supremum of the probability of committing a type II error, with the supremum of the probability of committing type I error (α) kept at a constant 0.05. Table 4 reports power, $(1 - \beta)$, to detect treatment effect of a two-tailed test with p-value cut-off of 0.05 (0.025 per tail), with the statistic simulated from MAL predictions of our model, for varying treatment effect and study sizes. We try to keep our cohort sizes realistic for a rare disease, *i.e.* we allow for patient numbers to range from 10 to 200.

Ideally we would like to be able to achieve power of more than 95%, even with small treatment effect sizes and a small number of patients, but our model, trained on 18 participants, doesn't allow for such level of control over type II error for all but medium or large treatment effects (more than 20% and 50% respectively) and large (more than 140 participants) or medium-sized (more than 30 participants) clinical trials respectively.

However, our results combined with expected improvements of the model performance due to larger training samples, and better gene selections, such levels might be reached for medium treatment effect size (20% reduction in effective MAL) and large clinical sizes.

Table 4. Power analysis. Entries in the table report power to detect treatment effect based on the size of a cohort (from 10 − 200 participants) and the treatment effect of the study to reverse splicing changes (10, 20 and 50%). Entries denoting power greater than 0.95% are presented in boldface.

study size (participants)	treatment effect 10%	treatment effect 20%	treatment effect 50%
10	0.100	0.209	0.635
20	0.131	0.322	0.855
30	0.162	0.423	0.947
40	0.193	0.517	**0.981**
50	0.226	0.596	**0.994**
60	0.259	0.671	**0.998**
70	0.283	0.723	**0.999**
80	0.311	0.772	**0.9998**
90	0.339	0.816	**0.99998**
100	0.370	0.851	**0.99998**
110	0.396	0.881	**0.99999**
120	0.418	0.901	**1.0**
130	0.448	0.924	**1.0**
140	0.467	0.937	**1.0**
150	0.500	**0.952**	**1.0**
160	0.523	**0.962**	**1.0**
170	0.545	**0.969**	**1.0**
180	0.568	**0.977**	**1.0**
190	0.589	**0.981**	**1.0**
200	0.613	**0.986**	**1.0**

Limitations

A source of potential criticism is that muscles of DM1 patients have physiological differences (atrophy, increased fat content), especially when disease is severe. Quite possibly observed changes in AS/APA are partly attributable to these physiological differences in DM1 as opposed to purely biomolecular differences. The structure of this counter-argument could be as follows:

Muscles of DM1 patients have higher fat content than affected controls. Muscle samples collected from DM1 patients have higher ratio of intermuscular adipocytes to myocytes. Adipocytes have different AS/APA profiles than myocytes. Observed AS/APA changes in the DM1 spectrum are mostly derived from differences in adipocyte/myocyte profile. As a result, mRNA study of muscle tissue is no more effective (and possibly less effective) than a blinded study based on pathophysiological inspection of the tissue.

This argument can, of course, be extended to other physiological changes than increased fat content, and other molecular events than AS/APA. Bachinski et al. [28],

for example, propose that splicing changes could be a secondary result of muscle regeneration.

There are multiple ways to address these concerns:

1. Alternative predictive model design based on tissue culture models, where *in vivo* limitations are reduced, with homogeneity of the cellular composition of the model being a big advantage.

2. Investigating methods to correct for potential "physiological" covariates (*e.g.* fat content), using purely statistical techniques to estimate covariate influence from gathered data and existing prior information (*e.g.* mRNA profiles of adipocytes), or biological methods, such as a recent effort to collect higher quality muscle samples through MRI-guided biopsy [29].

3. Discovering DM1 biomarkers in blood, as opposed to muscle. Blood, being much more homogeneous tissue than muscle is expected to be less prone to the existence of confounding variables. Additionally, necessity of muscle sampling was highlighted to be a "main drawback" [8]. However, achieving this would require at least two separate, successful studies, one to identify biomarkers and one to evaluate them. Even if blood biomarkers were identified, their clinical utility might be limited, as a reduction of an effective MAL in blood would not be as direct evidence of treatment effectiveness as such reduction in muscle.

4. Introduction of positive controls in experimental designs. Biological samples in the positive control group would be composed of tissue collected from individuals with other muscular dystrophies (e.g. Becker muscular dystrophy, Duchenne muscular dystrophy, facioscapulohumeral muscular dystrophy or tibial muscular dystrophy). These diseases feature dystrophy and increased muscle regeneration program as part of their phenotype, but without disruption of RNA-binding splicing factors. Absence of DM1-specific splicing changes in these positive controls would allow to rule out alternative explanations of mis-splicing mechanisms (e.g. muscle regenration) and strengthen our belief in currently accepted models of molecular pathomechanism of DM1.

5. Confining the analysis to transcripts which are exclusively or predominantly

expressed in skeletal muscle, not fat.

Conclusion

In this study we design and build a model based on PLSR, which can explain as much as 28.9% of the variance in DM1 CTG trinucleotide expansion from mRNA splicing data. Such explainability is only obtained when the model is trained on expression data from genes previously identified by Nakamori *et al.* [8] as having disrupted AS on data obtained from muscle samples. We show how such model could be used in a clinical setting in the context of emerging DM1 treatments, and report power analysis to detect treatment effect depending on size of the treatment effect, type 1 error (α) and potential size of the clinical trial.

Supporting information

S1 Appendix. Code & how to run it. All of the computer programs written for this study can be found on github.

The following instructions should allow one to independently verify results of our simulations (also available in the repository).

We advise that all of this code be run on a machine with 64 GB of RAM or more, given that some parts of the pipeline can use up in excess of 32 GB of RAM. We were able to successfully execute the entire analysis using an AWS "m5.4xlarge" EC2 instance. We found the default amount of storage, 8 GB, to be insufficient to store both the primary data and intermediate computations. We increased the amount of storage to 100 GB. We remove all networking restrictions on the instance, to allow for remote access of jupyter notebooks, which contain our pipeline.

We had to apply the following shell commands to set up the machine:

1. sudo apt-get update
2. sudo apt-get upgrade
3. sudo apt-get install python3-pip
4. git clone https://github.com/picrin/clinical_applications.git

5. pip3 install jupyter

All data and metadata used in this study are available in publicly accessible s3 bucket, with the following paths: "dm1-biomarkers/CEL", "dm1-biomarkers/annotations" respectively. These datasets need to be copied to the root of the "clinical_applications" repository as directories "CEL" and "annotations" respectively.

All required third-party dependencies can be installed from the provided "requirements.txt"

Finally, "jupyter notebook −−ip 0.0.0.0 −−port 8888" can be issued to start the notebook server, which we can access remotely, using a DNS entry allocated for our EC2 instance and provided that communications on our chosen port is configured to be accessible through the AWS firewall.

We now run notebooks in the following order:

1. "01_parse_chip_data.ipynb". This part of the pipeline is responsible for determining probeset ids, sequences of probes and probe coordinates on the chip. It produces an intermediate file with data adhering to the following schema: "probeset, x, y, sequence".

2. "02_parse_csv_annotations.ipynb". Here, we determine "genomic" metadata, i.e. chromosomomal coordinates and strandedness.

3. "03_unpack_CEL_files.ipynb". Here we use our own contribution to Biopython to parse the binary CEL v4 file format, which is what all our microarray data uses.

4. "04_quantile_normalise.ipynb". Here we perform quantile normalisation of our microarray data.

5. "05_reannotate_probeset_level.ipynb". Here we verify Affymetrix's annotation. We determine that over 1% probes are incorrectly annotated. We discard these probes. We limit our attention to probes, which belong to chromosomes chr1-chr22, X, Y and the mitochondrial DNA (M).

6. "06_intervaltrees.ipynb". Here we carry out an exclusive filtering, choosing probes, which are identified by "gencode.v26lift37.annotation.gtf" as having

"transcript_type" equal to "protein_coding", or "gene_type" equal to "protein_coding", as well as the exon type equal to "CDS" (Coding sequence) or "UTR" (5' or 3' untranslated regions).

7. "07_merge_annotation_experiment.ipynb". Here we produce a single file per human gene, as identified in GENCODE v26, with data from all participants for all probes for that gene.

8. "08_predictions.ipynb". Here we run our PLSR model to predict MAL from the microarray data.

9. "10_power_analysis.ipynb". Here we run power analysis presented in Results & Discussion.

10. "13_combine_p_values.ipynb". Here we combine p-values for a probeset of interest using a previously published p-value and one obtained in this study.

S2 Appendix. DM1-AS

ABLIM2, ALPK3, ANK2, ARFGAP2, ATP2A1, ATP2A2, BIN1, CACNA1S, CAMK2B, CAPN3, CAPZB, CLCN1, COPZ2, DMD, DTNA, FHOD1, GFPT1, IMPDH2, INSR, KIF13A, LDB3, MBNL1, MBNL2, MLF1, NFIX, NRAP, OPA1, PDLIM3, PHKA1, RYR1, SOS1, TBC1D15, TTN, TXNL4A, UBE2D3, USP25, VEGFA, VPS39

S3 Appendix. DM1-APA

ABCA1, AGL, ALG3, AMHR2, AP1G1, ARHGEF7, ASPH, ATP5E, BRSK2, BRWD1, CACNA1G, CACNB1, CDC42, CEBPA, CELF1, CHRNA1, CIRBP, CLDND1, COPS4, DAPK2, DES, DNAJB6, DST, DVL3, EZR, FASTK, GPS1, HDAC11, IDH3A, ILF3, KCNK7, KDELR1, KIF1B, KRBA1, LAMP2, LDB3, LMNA, MBNL2, MDN1, MEF2B, MEF2C, MEF2D, MGP, MORC3, MTCH1, MYH6, NDUFB10, NR2F1, NUP43, OSBPL1A, PCBD2, PCM1, PCMT1, PDLIM2, PDLIM5, PEBP4, PFKFB2, PIK3C2B, PLIN2, RAB24, RIN1, RTN2, SAMD4A, SETD3, SLC25A36, SMIM3, SNX1, SPATS2L, SPEG, SPTB, TBL2, TGFBI, TJP2, TMEM38B, TNNI1, TPM1, TPM2, TPM3, TTYH3, U2SURP

Acknowledgments

We would like to thank the Marigold Foundation for organising and funding DMBDI.

We would like to thank Linda Bachinsky and Keith Baggerly for their helpful comments on multiple revisions of this manuscript, as well as performing initial analysis and quality control on blood and muscle microarray samples. We would also like to thank Anna Casasent (nee Unruh) for performing quality control on blood and muscle microarray samples.

Adam Kurkiewicz would like to thank the college of Medical, Veterinary and Life Sciences as well as the Engineering and Physical Science Research Council for financing his doctoral training programme stipend, student fees and his research fund.

References

1. Brook JD, McCurrach ME, Harley HG, Buckler AJ, Church D, Aburatani H, et al. Molecular basis of myotonic dystrophy: Expansion of a trinucleotide (CTG) repeat at the 3′ end of a transcript encoding a protein kinase family member. Cell. 1992;68(4):799–808. doi:10.1016/0092-8674(92)90154-5.

2. Meola G, Cardani R. Myotonic dystrophies: An update on clinical aspects, genetic, pathology, and molecular pathomechanisms. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease. 2015;1852(4):594–606. doi:10.1016/j.bbadis.2014.05.019.

3. Thornton CA. Myotonic Dystrophy. Neurologic Clinics. 2014;32(3):705–719. doi:10.1016/j.ncl.2014.04.011.

4. Holt I, Jacquemin V, Fardaei M, Sewry CA, Butler-Browne GS, Furling D, et al. Muscleblind-Like Proteins. The American Journal of Pathology. 2009;174(1):216–227. doi:10.2353/ajpath.2009.080520.

5. Cudia P, Bernasconi P, Chiodelli R, Mangiola F, Bellocci F, Russo AD, et al. Risk of arrhythmia in type I myotonic dystrophy: the role of clinical and genetic variables. Journal of Neurology, Neurosurgery & Psychiatry. 2009;80(7):790–793. doi:10.1136/jnnp.2008.162594.

6. Udd B, Krahe R. The myotonic dystrophies: molecular, clinical, and therapeutic challenges. The Lancet Neurology. 2012;11(10):891–905. doi:10.1016/s1474-4422(12)70204-1.

7. Pandey SK, Wheeler TM, Justice SL, Kim A, Younis HS, Gattis D, et al. Identification and Characterization of Modified Antisense Oligonucleotides Targeting DMPK in Mice and Nonhuman Primates for the Treatment of Myotonic Dystrophy Type 1. Journal of Pharmacology and Experimental Therapeutics. 2015;355(2):310–321. doi:10.1124/jpet.115.226969.

8. Nakamori M, Sobczak K, Puwanant A, Welle S, Eichinger K, Pandya S, et al. Splicing biomarkers of disease severity in myotonic dystrophy. Annals of Neurology. 2013;74(6):862–872. doi:10.1002/ana.23992.

9. Batra R, Charizanis K, Manchanda M, Mohan A, Li M, Finn DJ, et al. Loss of MBNL Leads to Disruption of Developmentally Regulated Alternative Polyadenylation in RNA-Mediated Disease. Molecular Cell. 2014;56(2):311–322. doi:10.1016/j.molcel.2014.08.027.

10. Bush WS, Moore JH. Chapter 11: Genome-Wide Association Studies. PLoS Computational Biology. 2012;8(12):e1002822. doi:10.1371/journal.pcbi.1002822.

11. Brody T. Biostatistics—Part I. In: Clinical Trials. Elsevier; 2016. p. 203–226. Available from: https://doi.org/10.1016/b978-0-12-804217-5.00009-6.

12. D'Agostino RB, Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues - the encounters of academic consultants in statistics. Statistics in Medicine. 2002;22(2):169–186. doi:10.1002/sim.1425.

13. du Prel JB, Röhrig B, Hommel G, Blettner M. Choosing Statistical Tests. Deutsches Aerzteblatt Online. 2010;doi:10.3238/arztebl.2010.0343.

14. Lello L, Avery SG, Tellier L, Vazquez A, de los Campos G, Hsu SDH. Accurate Genomic Prediction Of Human Height. bioRxiv. 2017;doi:10.1101/190124.

15. VanRaden PM, Tassell CPV, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, et al. Invited Review: Reliability of genomic predictions for North American

Holstein bulls. Journal of Dairy Science. 2009;92(1):16–24. doi:10.3168/jds.2008-1514.

16. Azencott CA, Grimm D, Sugiyama M, Kawahara Y, Borgwardt KM. Efficient network-guided multi-locus association mapping with graph cuts. Bioinformatics. 2013;29(13):i171–i179. doi:10.1093/bioinformatics/btt238.

17. Lee JM, Galkina EI, Levantovsky RM, Fossale E, Anderson MA, Gillis T, et al. Dominant effects of the Huntington's disease HTT CAG repeat length are captured in gene-expression data sets by a continuous analysis mathematical modeling strategy. Human Molecular Genetics. 2013;22(16):3227–3238. doi:10.1093/hmg/ddt176.

18. Monckton DG, Wong LJC, Ashizawa T, Caskey CT. Somatic mosaicism, germline expansions, germline reversions and intergenerational reductions in myotonic dystrophy males: small pool PCR analyses. Human Molecular Genetics. 1995;4(1):1–8. doi:10.1093/hmg/4.1.1.

19. Morales F, Couto JM, Higham CF, Hogg G, Cuenca P, Braida C, et al. Somatic instability of the expanded CTG triplet repeat in myotonic dystrophy type 1 is a heritable quantitative trait and modifier of disease severity. Human Molecular Genetics. 2012;21(16):3558–3567. doi:10.1093/hmg/dds185.

20. Rexiati M, Sun M, Guo W. Muscle-Specific Mis-Splicing and Heart Disease Exemplified by RBM20. Genes. 2018;9(1):18. doi:10.3390/genes9010018.

21. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25(11):1422–1423. doi:10.1093/bioinformatics/btp163.

22. Kurkiewicz A. Parser for Affy CEL file version 4; 2016. Available from: https://github.com/biopython/biopython/commit/6c14a26dda32ad6d3147036a01f3d0c4d306c647.

23. Kurkiewicz A. DMBDI explorer; 2019. Available from: http://dmbdi.adamkurkiewicz.com.

24. Kurkiewicz A. DMBDI explorer walk-through video; 2019. Available from: https://youtu.be/gvs5XtdYwE0.

25. Kersey PJ, Allen JE, Allot A, Barba M, Boddu S, Bolt BJ, et al. Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species. Nucleic Acids Research. 2017;46(D1):D802–D808. doi:10.1093/nar/gkx1011.

26. Fisher RA. Statistical Methods for Research Workers. Oliver and Boyd; 1934.

27. Affymetrix. GeneChip® Exon Array Design; 2005. Available from: http://tools.thermofisher.com/content/sfs/brochures/exon_array_design_technote.pdf.

28. Bachinski LL, Baggerly KA, Neubauer VL, Nixon TJ, Raheem O, Sirito M, et al. Most expression and splicing changes in myotonic dystrophy type 1 and type 2 skeletal muscle are shared with other muscular dystrophies. Neuromuscular Disorders. 2014;24(3):227–240. doi:10.1016/j.nmd.2013.11.001.

29. Lassche S, Janssen BH, IJzermans T, Fütterer JJ, Voermans NC, Heerschap A, et al. MRI-Guided Biopsy as a Tool for Diagnosis and Research of Muscle Disorders. Journal of Neuromuscular Diseases. 2018;5(3):315–319. doi:10.3233/JND-180318.