Supplementary Information: β-catenin is central to DUX4-driven network rewiring in Facioscapulohumeral muscular dystrophy

Christopher R. S. Banerji[1,2,3,4,*], Paul Knopp[4], Louise Moyle[4], Simone Severini[2], Richard W. Orrell[5], Andrew E. Teschendorff[1,6], Peter S. Zammit[4]

1. Statistical Cancer Genomics, Paul O'Gorman Building, UCL Cancer Institute, University College London, London WC1E 6BT, UK.
2. Department of Computer Science, University College London, London WC1E 6BT, UK.
3. Centre of Mathematics and Physics in the Life Sciences and Experimental Biology, University College London, London WC1E 6BT, UK.
4. King's College London, Randall Division of Cell and Molecular Biophysics, New Hunt's House, Guy's Campus, London SE1 1UL, UK.
5. Department of Clinical Neuroscience, Institute of Neurology, University College London, Rowland Hill Street, London NW3 2PF, UK.
6. CAS-MPG Partner Institute for Computational Biology, Chinese Academy of Sciences, Shanghai Institute for Biological Sciences, 320 Yue Yang Road, Shanghai 200031, China.

Supplementary Methods

Comparing Possible Transformed Pearson Correlations in Step 1 of InSpiRe

The choice of the transformation of the Pearson correlations to construct the stochastic matrix described in the Materials and Methods must ensure non-negativity, and that interpretation of a high edge weight indicative of an increased likelihood of interaction between connected proteins is valid. We considered two possible transformations:

$$w_{ij} = \frac{1}{2}(1 + C_{ij}) \tag{1}$$

and

$$w_{ij} = |C_{ij}| \tag{2}$$

where C_{ij} denotes Pearson correlation in the gene expression profiles of protein i and protein j across samples corresponding to the phenotype considered. Equation (1) assumes that signal transduction flows preferably along paths of proteins with positively correlated gene expression profiles and has been employed previously [1, 2]. It was noted, however, that this was only an approximation [1], and there is mounting evidence that negative correlations play an important role in signal transduction [3], consequentially the transformation described by equation (2), which assigns both strong positive and strong negative correlations a high weight may be considered more realistic for this interpretation.

We utilised both transformations independently and upon performing analysis described in the methods to identify proteins significantly rewiring between FSHD and control skeletal muscle, we found that the transformation (2) provided a substantially better discriminator than (1) as judged by p-value histograms [4]. In fact, all p-value histograms utilising the transformation described in (1) were flat, implying that in utilising this transformation one is unable to reliably determine differences between the phenotypes. Consequentially all results described in this work were obtained utilising the more realistic transformation (2).

Statistical Analysis via the Jackknife

The jackknife procedure was employed previously to analyse differential local flux entropies [1], the methodology is considered superior to the bootstrap estimation (for our purposes) which is known to artificially inflate correlations [1, 5].
Jackknife estimation is performed as follows: given a quantity X (e.g. a differential local flux entropy), we first estimate X from our entire data set, consisting of n samples, denoting this estimate by $\hat{X}$. We next compute n subsequent estimators $(X_i)_{i=1}^{n}$, from the data set, by removing each sample, one at a time, and re-estimating X. We then compute an estimate for the mean X_μ and the variance X_σ of X via:

$$X_\mu := n\hat{X} - \frac{(n-1)}{n}\sum_{i=1}^{n} X_i \tag{3}$$

$$X_\sigma = \frac{Var[n\hat{X} - (n-1)X_i]}{n-1}. \tag{4}$$

We then compute a Z statistic

$$Z = \frac{X_\mu}{\sqrt{X_\sigma}} \sim \mathcal{N}(\mu, 1). \tag{5}$$

which can be used to test the hypotheses on the mean of the quantity X. In our analyses the X will either be a differential local flux entropy a differential correlation or a Kullback-Leibler divergence, hence the null hypothesis will be that the mean of the quantity if 0. Statistical significance is assessed at the 5% level.

Comparing methodologies

To evaluate the performance of InSpiRe relative to other methodologies, we applied InSpiRe, NetWalk [6] and GSEA [7] on differentially expressed genes to each FSHD data set independently, and evaluated the enrichments of identified genes.

Differential expression analysis

Differential expression analysis was performed on normalised data sets matched to the protein interaction network using the limma package in R [8]. Gene set enrichment analysis (GSEA) [7] was then performed against the gene sets of the Molecular Signatures Database [9], using the t-scores output by the limma analysis to rank the genes. Gene sets identified by GSEA as displaying $p < 0.05$ and $FDR < 0.25$ were considered significantly enriched.

NetWalk analysis

NetWalk is a network based algorithm which considers the stationary distribution of a weighted random walk on a network of compiled interactions. Weights on network nodes are data derived and bias walker visitation in a biologically relevant manner. We implemented NetWalk on normalised data sets using the NetWalker software [10], and employing the compiled Knowledgebase provided as the underlying network for implementation; functional annotation of identified edges was performed using the FunWalk option. The i^{th} element of the weight flux vector, $(\mathbf{w})_{i=1}^{N}$, where N is the number of genes in the microarray, for a given data set, was defined as the ratio of the mean expression the i^{th} probe across samples corresponding to the phenotype examined (disease, aged, atrophic) to the mean expression across control samples. This selection is a recommended option [10]. In order to ensure the findings of NetWalk were statistically robust, we utilised the jackknife re-sampling procedure (see above) to assess significance of edge visitation ratios, and functional annotation ratios.
As with the methodology developed in this paper, NetWalk was run independently on each data set to produce a list of significant functional terms, and each FSHD data set identified around 3000 significant functional terms. The intersection of the significant functional terms in the FSHD data set consisted of 266 terms, and contained several terms associated to oxidative stress, apoptosis and mitochondrial dysfunction. When terms also associated with age, atrophy and other diseases were removed from this intersection however, only 19 terms remained, none of which had a strong justification to association with FSHD in the literature.

Functional annotation for InSpiRe implicated genes

Functional annotation of InSpiRe implicated genes was performed on the significant genes ($p < 0.05$) implicated by local flux entropy or local symmetrised KL divergence, using the DAVID Bioinformatics Resources 6.7 [11]. The p-value cut off for the Fisher's exact test (EASE score) employed by the DAVID software for implicating enriched pathways was 0.05. The background gene set utilised consisted of all the genes in the protein interaction network.

Comparison

To compare the relevance of the results output by the various methodologies, we considered 14 FSHD associated pathways: Wnt signalling, TNF or MAPK related signalling, vasculature development, calcium signalling, oxidative stress response, cell cycle, apoptosis, mitochondrial dysfunction, asymmetrical development, muscle structure, nuclear envelope, muscle differentiation, histone modification and actin cytoskeletal signalling. For each pathway we gave each method a score from 0 to 4 corresponding to the number of FSHD data sets it was capable of detecting the pathway in. Of the three methodologies considered InSpiRe was the most successful, achieving an average score of 3.5, NetWalk achieved 3.29 and GSEA on differentially expressed genes achieved 1.86. A summary of the results is provided in Figure S5.

The FSHD Network

The FSHD network is provided as a Supplementary file: *FSHDNetwork.cys*, this file can be opened and the network examined using the freely available software Cytoscape [12], available to download from *http://www.cytoscape.org/*.

Supplementary Table Captions

Table S1 A full summary of local flux entropy and Kullback-Leibler divergence statistics for each FSHD dataset considered for the core 164 genes significantly rewiring specifically in FSHD muscle and not attributable to atrophy, ageing or other muscle diseases.

Supplementary Figures

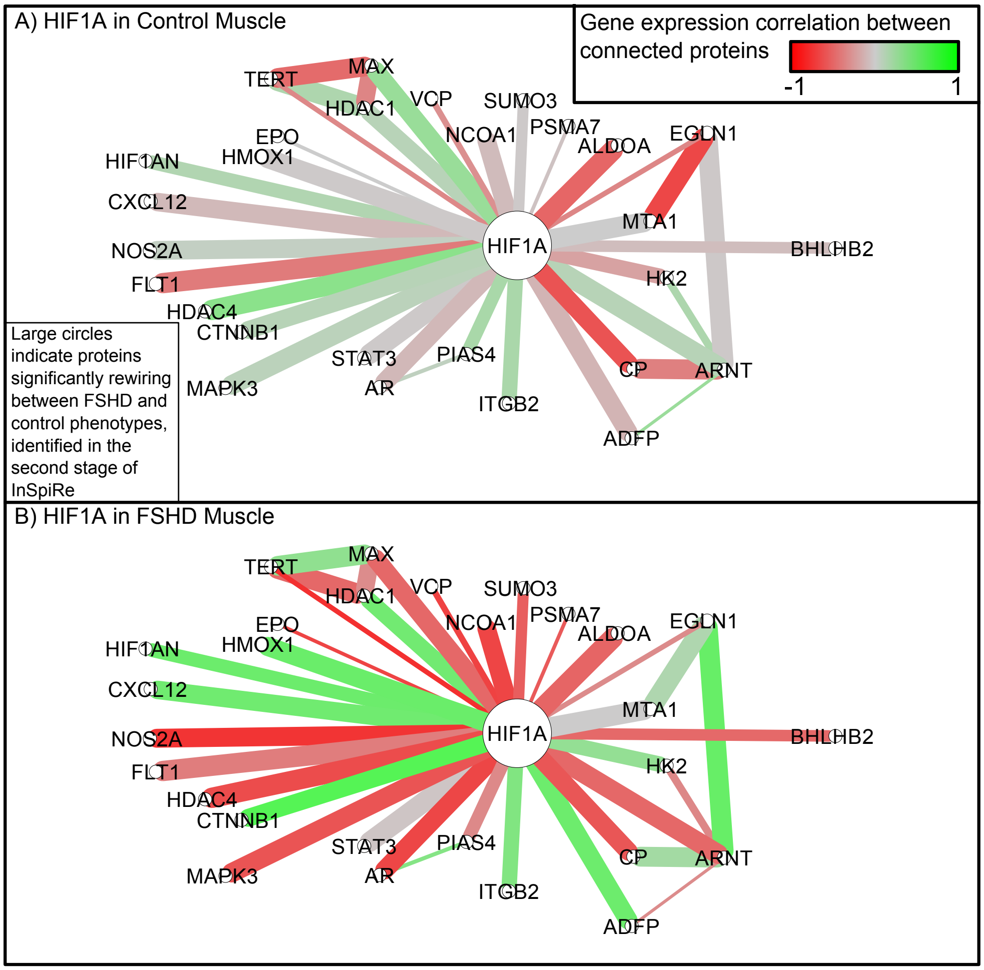
A) HIF1A in Control Muscle
Gene expression correlation between connected proteins
-1
1
TERT
MAX
HDAC1
VCP
SUMO3
PSMA7
NCOA1
ALDOA
EGLN1
EPO
HMOX1
HIF1AN
CXCL12
MTA1
NOS2A
HIF1A
BHLHB2
FLT1
HK2
HDAC4
CTNNB1
STAT3
PIAS4
CP
ARNT
MAPK3
AR
ITGB2
ADFP
Large circles indicate proteins significantly rewiring between FSHD and control phenotypes, identified in the second stage of InSpiRe
B) HIF1A in FSHD Muscle

Figure 1: The neighbourhood of *HIF1A* in the FSHD network. Interactions are coloured proportional to the Pearson correlation in gene expression between connected genes across control samples (A) and FSHD samples (B). Red edges are negatively correlated, grey edges uncorrelated and green edges positively correlated. The thickness of edges is proportional to $1-p$ where $p \in (0, 0.05]$ is the p-value of the statistical analysis performed to determine whether the correlation in gene expression between connected edges is different between FSHD and controls. Large nodes belong to the core set of 164 high confidence FSHD specific rewiring genes. Note the strong increase in correlation between *HIF1A* and *CTNNB1*.

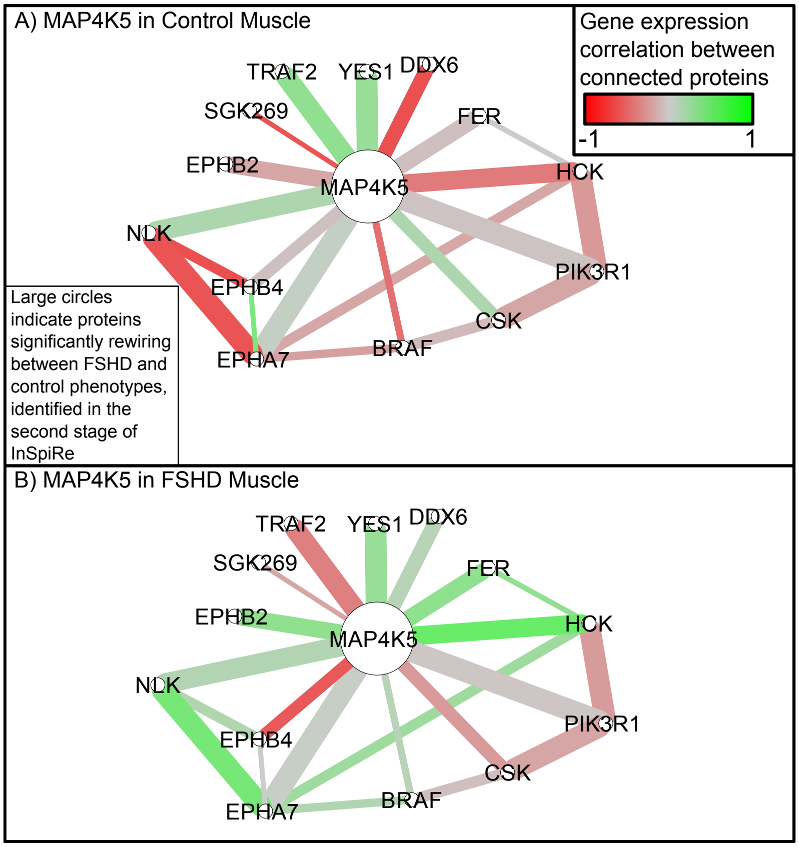
A) MAP4K5 in Control Muscle
Gene expression correlation between connected proteins
-1
1
TRAF2
YES1
DDX6
SGK269
FER
EPHB2
MAP4K5
HCK
NLK
PIK3R1
EPHB4
CSK
BRAF
EPHA7
Large circles indicate proteins significantly rewiring between FSHD and control phenotypes, identified in the second stage of InSpiRe
B) MAP4K5 in FSHD Muscle
TRAF2
YES1
DDX6
SGK269
FER
EPHB2
MAP4K5
HCK
NLK
PIK3R1
EPHB4
CSK
BRAF
EPHA7

Figure 2: The neighbourhood of *MAP4K5* in the FSHD network. Interactions are coloured proportional to the Pearson correlation in gene expression between connected genes across control samples (A) and FSHD samples (B). Note the strong increase in correlation between *MAP4K5* and *TRAF2*.

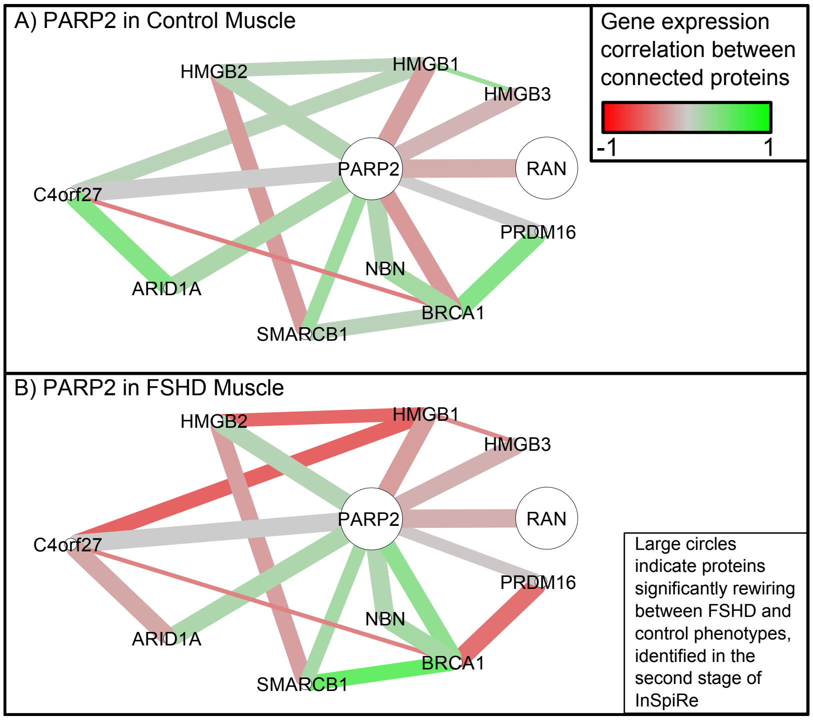
A) PARP2 in Control Muscle
Gene expression correlation between connected proteins
-1
1
HMGB2
HMGB1
HMGB3
PARP2
RAN
C4orf27
PRDM16
NBN
ARID1A
BRCA1
SMARCB1
B) PARP2 in FSHD Muscle
HMGB2
HMGB1
HMGB3
PARP2
RAN
C4orf27
PRDM16
NBN
ARID1A
BRCA1
SMARCB1
Large circles indicate proteins significantly rewiring between FSHD and control phenotypes, identified in the second stage of InSpiRe

Figure 3: The neighbourhood of *PARP2* in the FSHD network. Interactions are coloured proportional to the Pearson correlation in gene expression between connected genes across control samples (A) and FSHD samples (B). Note the altered interaction between *PARP2* and *BRCA1* in FSHD.

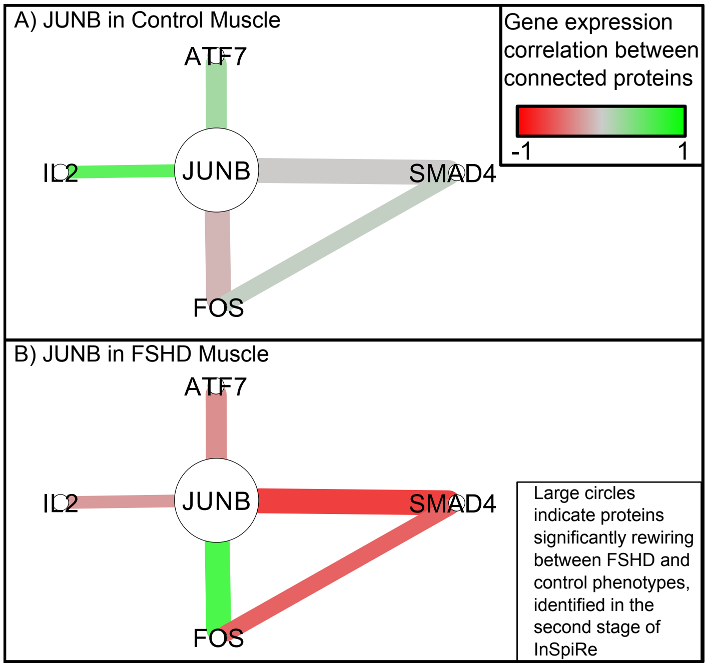
A) JUNB in Control Muscle
Gene expression correlation between connected proteins
-1
1
ATF7
IL2
JUNB
SMAD4
FOS
B) JUNB in FSHD Muscle
ATF7
IL2
JUNB
SMAD4
FOS
Large circles indicate proteins significantly rewiring between FSHD and control phenotypes, identified in the second stage of InSpiRe

Figure 4: The neighbourhood of *JUNB* in the FSHD network. Interactions are coloured proportional to the Pearson correlation in gene expression between connected genes across control samples (A) and FSHD samples (B). The interaction between *JUNB* and *FOS* is significantly altered in FSHD muscle.

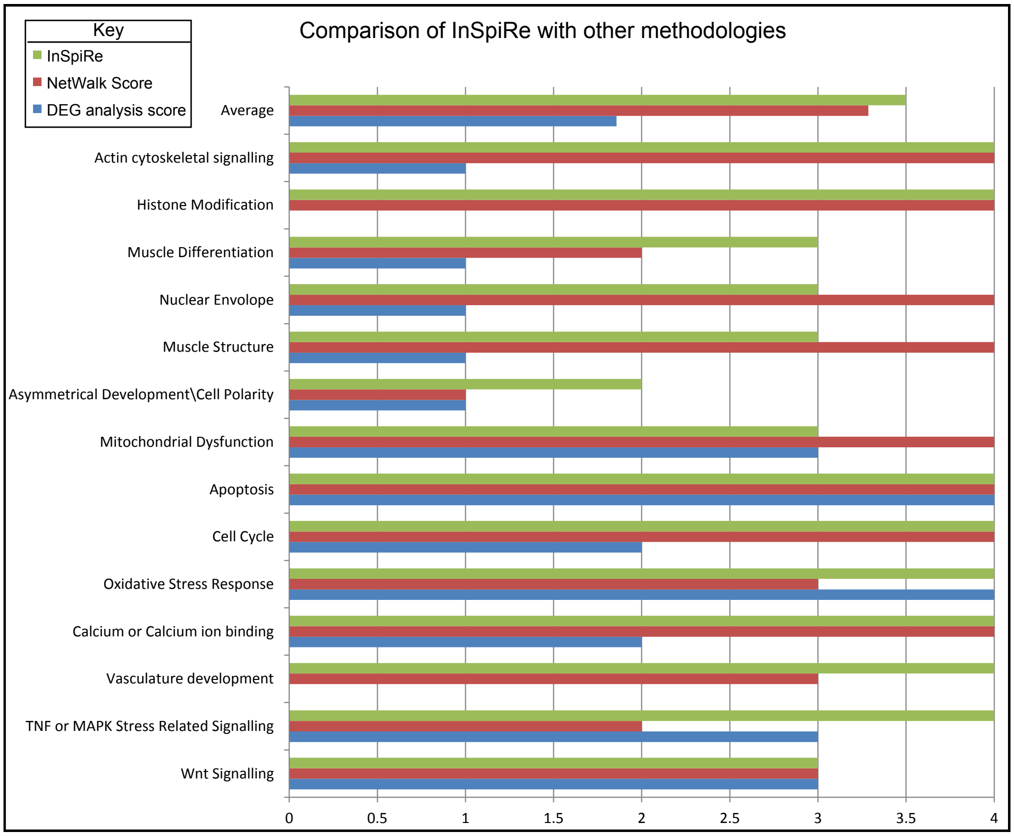

Figure 5: Comparing InSpiRe to NetWalk and GSEA on differentially expressed genes. Each method is scored with the number of data sets considered in which it identifies a given pathway. The average score across all pathways is highest for InSpiRe.

References

[1] West, J., Bianconi, G., Severini, S. & Teschendorff, A. E. Differential network entropy reveals cancer system hallmarks. *Sci Rep* **2**, 802 (2012).

[2] Teschendorff, A. E. & Severini, S. Increased entropy of signal transduction in the cancer metastasis phenotype. *BMC Syst Biol* **4**, 104 (2010).

[3] Zeng, T. & Li, J. Maximization of negative correlations in time-course gene expression data for enhancing understanding of molecular pathways. *Nucleic Acids Res* **38**, e1 (2010).

[4] Pounds, S. B. Estimation and control of multiple testing error rates for microarray studies. *Brief Bioinform* **7**, 25–36 (2006).

[5] van Wieringen, W. N. & van der Vaart, A. W. Statistical analysis of the cancer cell's molecular entropy using high-throughput data. *Bioinformatics* **27**, 556–63 (2011).

[6] Komurov, K., White, M. A. & Ram, P. T. Use of data-biased random walks on graphs for the retrieval of context-specific networks from genomic data. *PLoS Comput Biol* **6** (2010).

[7] Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545–50 (2005).

[8] Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3**, Article3 (2004).

[9] Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–40 (2011).

[10] Komurov, K., Dursun, S., Erdin, S. & Ram, P. T. NetWalker: a contextual network analysis tool for functional genomics. *BMC Genomics* **13**, 282 (2012).

[11] Dennis, J., G. *et al.* DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* **4**, P3 (2003).

[12] Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498–504 (2003).