The genetic regulatory signature of type 2 diabetes in human skeletal muscle

Laura J. Scott1,*, Michael R. Erdos2,*, Jeroen R. Huyghe1,*, Ryan P. Welch1,*, Andrew T. Beck1, Brooke N. Wolford2, Peter S. Chines2, John P. Didion2, Narisu Narisu2, Heather M. Stringham1, D. Leland Taylor2,3, Anne U. Jackson1, Swaroopanari Vadlamudi4, Lori L. Bonnycastle2, Leena Kinnunen5, Jouko Saramies6, Jouko Sundvall5, Ricardo D’Oliveira Albanus7, Anna Kiseleva7, John Hensley7, Gregory E. Crawford8,9, Hui Jiang1, Xiaoquan Wen1, Richard M. Watanabe10,11, Timo A. Lakka12,13,14, Karen L. Mohlke4, Markku Laakso15,16, Jaakko Tuomilehto17,18,19,20, Heikki A. Koistinen5,21,22, Michael Boehnke1,*, Francis S. Collins2,*, & Stephen C.J. Parker7,23,*

Type 2 diabetes (T2D) results from the combined effects of genetic and environmental factors on multiple tissues over time. Of the >100 variants associated with T2D and related traits in genome-wide association studies (GWAS), >90% occur in non-coding regions, suggesting a strong regulatory component to T2D risk. Here to understand how T2D status, metabolic traits and genetic variation influence gene expression, we analyse skeletal muscle biopsies from 271 well-phenotyped Finnish participants with glucose tolerance ranging from normal to newly diagnosed T2D. We perform high-depth strand-specific mRNA-sequencing and dense genotyping. Computational integration of these data with epigenome data, including ATAC-seq on skeletal muscle, and transcriptome data across diverse tissues reveals that the tissue-specific genetic regulatory architecture of skeletal muscle is highly enriched in muscle stretch/super enhancers, including some that overlap T2D GWAS variants. In one such example, T2D risk alleles residing in a muscle stretch/super enhancer are linked to increased expression and alternative splicing of muscle-specific isoforms of ANK1.
The prevalence of diabetes in 2015 reached 415 million adults worldwide, and is projected to increase to 642 million by 2040 (www.idf.org/diabetesatlas). T2D accounts for ~90% of these individuals. Pancreatic islet beta-cell dysfunction, accentuated by insulin resistance in skeletal muscle and other peripheral tissues, are the hallmarks of T2D. To examine the relationship between T2D and related traits and muscle metabolism, we obtained blood samples and skeletal muscle biopsies, and performed clinical phenotyping, including oral glucose tolerance tests (OGTT), on 271 Finnish individuals. Study subjects were chosen to span the range of glucose tolerance from normal to newly diagnosed (not on glucose-lowering drug therapy) T2D.

Because most (>90%) T2D and related trait GWAS single-nucleotide polymorphisms (SNPs) reside in non-coding regions, we aim to refine the molecular mechanisms underlying these associations by identifying the target gene(s) and direction of effect the risk allele has on target gene expression. To accomplish this, we further perform dense and diverse molecular profiling on this collection of skeletal muscle biopsies at the genome, epigenome and transcriptome level. Integration of these data types allowed us to nominate multiple T2D GWAS SNP effector transcripts. These genetic, epigenomic and transcriptomic results represent the largest study of the regulatory landscape in human muscle and reveal how it relates to T2D.

Results

T2D insights from mRNA-seq maps and clinical phenotypes.

We sequenced the mRNA from study participants (Fig. 1a, Supplementary Table 1) to a mean depth of 91.3M strand-specific reads, the most comprehensive human skeletal muscle transcriptome catalogue to date. We tested for (1) differential gene expression between individuals with T2D and normal glucose tolerance (NGT) and (2) association between gene expression and three T2D-related traits (fasting glucose, fasting insulin and body mass index (BMI)) (Fig. 1b), adjusting for covariates (see Methods). At a false discovery rate (FDR) of 5%, we found three genes differentially expressed with T2D status; most strongly differentially expressed was a positive regulator of senescence, ubiquinolin 1 (UBN1) ($\beta = 0.83$, $q$-value $= 0.0089$). At FDR 5%, we detected a wide range in the number of gene–trait associations, from 38 for fasting glucose to 6,080 for fasting insulin.

To probe the underlying biology of differential gene expression, we performed gene ontology (GO) term analysis for each trait using the strength and direction of trait–gene association ($-\log_{10}(P$ value) signed for the direction of association) as a predictor of GO term membership. We display a pruned list of the most strongly associated GO terms, selected separately for terms enriched for genes with positive and negative trait association (Fig. 1c, Supplementary Fig. 1).

With T2D status, and with increases in fasting glucose, fasting insulin and BMI, we observed lower expression of genes involved in endoplasmic reticulum protein localization and translational elongation. For T2D, the most significant trends were for decreased expression of cellular respiration genes ($q$-value $= 1.4 \times 10^{-35}$), consistent with previous observations in skeletal muscle samples from T2D and NGT individuals following hyperinsulinemic–euglycemic clamp. Mitochondrial regulatory protein PGC-1alpha (PPARGC1A) was identified by Mootha et al. as a potential master regulator of mitochondrial expression. We observed lower, non-significantly different expression levels of $PPARGC1A$ ($\beta = -0.24$, $q$-value $= 0.57$) in individuals with T2D. Decreased mitochondrial function is a component of the mTOR pathway which is dysregulated in metabolic diseases; downregulation of the pathway shifts cells away from protein synthesis and cell growth and towards protein catabolism.

Consistent with this, for T2D, we observed lower expression of genes involved in generation of precursor metabolites, translational elongation and higher expression of genes involved in protein polyubiquitination (Fig. 1c).

Interestingly, higher expression levels of genes for leucocyte activation were strongly associated with higher levels of fasting insulin ($q$-value $= 1.3 \times 10^{-14}$) and less strongly with other traits. This could suggest that muscle samples from individuals with insulin resistance have increased inflammatory cell infiltration. However, this association remained when we estimated and then adjusted for the fraction of white blood cells (WBCs) and lymphocytes in the muscle samples (see Methods) in the trait–gene expression analysis ($q$-value $= 2.5 \times 10^{-14}$, Supplementary Fig. 2).

Genetic and epigenomic regulatory architecture of muscle.

To understand how genetic variation may influence muscle gene expression and thus T2D and related traits, we successfully genotyped 267 of the 271 samples using the Illumina Omni2.5 array and imputed 8,406,237 SNPs with minor allele frequency > 0.01 using the Genetics of Type 2 Diabetes (GoT2D) reference panel (Methods). We tested for association of variants within 1Mb of the transcription start site (TSS) with gene expression (cis-eQTL). At FDR 5%, we detected 2,104,118 cis-eQTLs in 19,697 (92%) of 21,420 tested genes (Supplementary Table 2), of which 168,633 cis-eQTLs remained when pruned on a per-gene basis by linkage disequilibrium (LD) at $r^2 < 0.2$. cis-eQTLs cluster at the TSS (Supplementary Fig. 3), as previously noted. Within 15,449 tested protein-coding genes, 14,479 (93.6%) had $\geq 1$ cis-eQTLs. Using a complementary approach, we tested for allelic-specific expression (ASE) in protein-coding genes that had a heterozygous-transcribed SNP in $\geq 10$ individuals; we detected ASE at FDR 5% in 7,404 (80.2%) of 9,228 tested genes (see Supplementary Fig. 4 for overlap with cis-eQTL). We compared the total set of our cis-eQTL results to those of 361 genotype-tissue expression (GTEx) skeletal muscle samples for 95.5M SNP–gene pairs tested in both studies (Supplementary Fig. 5). Of these SNP–gene pairs, in our muscle sample at FDR 5% we detected 3.5M SNP–gene pairs, of which 83.8% had a concordant direction in GTEx. Comparing our catalogue to the 1.1M GTEx SNP–gene pairs (FDR 5%), 813K SNP–gene pairs were significantly associated in both studies; 99% of these had concordant effect directions.

Previous studies of blood and lymphoblastoid cell lines have shown enrichment of cis-eQTLs that overlap enhancer and promoter regions. To explore this relationship in muscle, we created chromatin state maps using ChromHMM for skeletal muscle and 30 diverse cell or tissue types, including adipose, liver and pancreatic islets (Supplementary Fig. 6; see Methods). To assess the relationship between our skeletal muscle cis-eQTLs and chromatin states, we calculated enrichment statistics for how these different features overlap while controlling for minor allele frequency, distance to TSS and the number of SNPs in LD. We observed enrichment of skeletal muscle cis-eQTLs in genomic regions with active (promoter, enhancer or transcribed) chromatin states and a depletion in those with inactive (low signal) chromatin states across a broad range of cell types (Fig. 2a). This general enrichment pattern may reflect the predominance of cis-eQTLs from widely expressed housekeeping genes and mask a smaller set of muscle-specific genes.

To develop a measure of muscle gene expression specificity, we analysed additional RNA-seq data representing a set of diverse tissues from the Illumina Human Body Map 2.0 project. We used an information theory approach to score genes based on
muscle expression level and specificity relative to the panel of 16 diverse tissues; we refer to this score as the muscle expression specificity index (mESI; Methods) (Fig. 2b). We divided the genes into mESI deciles, where the 1st decile represents genes that are expressed at uniformly low levels across all tissues and the 10th decile represents genes that are highly and specifically expressed in muscle (Fig. 2c). cis-eQTL SNPs for genes in the 10 deciles were collected and partitioned into corresponding bins, which showed a larger range in skeletal muscle chromatin state enrichments (Supplementary Fig. 7) compared with the general non-mESI-based enrichments (Fig. 2a).

We previously defined chromatin ‘stretch enhancers’ as long (>3 kb) regions of tissue-specific active chromatin, and found they are highly enriched for GWAS variants in disease-relevant tissues15,16. Here we observed increasing cis-eQTL enrichment with increasing mESI decile in skeletal muscle stretch enhancers, but neither in muscle ‘typical enhancers’ (<800 bp; median enhancer size) nor in stretch and typical enhancers from other tissues.

Figure 1 | Molecular profiling maps of skeletal muscle combined with dense phenotyping reveals insights about T2D. (a) To understand the full spectrum of genetic variation and regulatory element usage in T2D-relevant tissue and across disease progression, we obtained skeletal muscle biopsies from the vastus lateralis of 271 well-phenotyped Finnish individuals with normal and impaired glucose tolerance, impaired fasting glucose or newly diagnosed, untreated T2D. (b) Correlogram of Spearman rank correlation coefficients for key metabolic traits. (c) Heatmap of GO terms for differentially expressed genes. For each trait, the 20 GO terms most significantly enriched for positive expression-trait association and the 20 GO terms most significantly enriched for negative expression-trait association were selected. GO terms were pruned to eliminate redundant terms. The terms were hierarchically clustered using the GO term enrichment beta. Darker red, stronger positive gene expression–trait association; darker blue, stronger negative association. Circle size represents number of significant GO terms.
tissues (Fig. 2d, Supplementary Fig. 8). Enrichment remained at different length thresholds (longest 10, 5, 1%) for calling stretch enhancers (Supplementary Fig. 9). Super enhancers \textsuperscript{17,18}, defined based on the abundance of acetylated lysine 27 on histone H3 (ref. 19), substantially overlap stretch enhancers (Supplementary Fig. 10) and show similar muscle cis-eQTL enrichment trends (Supplementary Fig. 11). These results suggest that the genetic regulatory architecture of muscle-specific gene expression is preferentially encoded in muscle stretch/super enhancers relative to typical enhancers.

Regulatory information encoded in DNA is activated through the binding of transcription factors (TFs) that can alter nucleosomal architecture and increase local chromatin accessibility. Because the stretch and super enhancer maps are based on modified histone data sets, it is difficult to detect these local TF-binding events. To construct higher-resolution regulatory maps across skeletal muscle and specifically within the larger stretch enhancer regions, we performed open chromatin accessibility mapping in frozen skeletal muscle using ATAC-seq\textsuperscript{20}. We found that ATAC-seq performed on frozen skeletal muscle is reproducible using 10 and 2 mg tissue inputs (Fig. 3a, Supplementary Fig. 12), and therefore we combined these replicates for further analyses. Comparison of our skeletal muscle ATAC-seq maps with similar maps from adipose\textsuperscript{21}, the

Figure 2 | The genetic regulatory architecture of muscle-specific gene expression. (a) Muscle cis-eQTL enrichment across chromatin states in diverse cell or tissue types\textsuperscript{11,15,63-64}. (b) Genes that are both highly expressed in skeletal muscle and highly tissue-specific fall into muscle expression specificity index (mESI) decile 10; genes which are lowly and ubiquitously expressed fall in decile 1. (c) As mESI decile increases, genes have greater expression in skeletal muscle but not in other tissue types. (d,e) cis-eQTL genes stratified by mESI decile, cis-eQTLs which fall into muscle stretch enhancers (≥ 3 kb) are significantly enriched for genes expressed highly and specifically in muscle versus cis-eQTLs in typical enhancers (≤ 800 bp).

such maps can be used to identify potentially causal TFBSs that drive cis-eQTL signals.

**Linking GWAS SNPs to effector transcripts in muscle.** We and others previously demonstrated that stretch/super enhancers in disease-relevant tissues are highly enriched for GWAS-disease-associated SNPs\(^{15,19}\), and a recent study identified autoimmune GWAS SNPs that reside in a T-cell super enhancer and act as cis-eQTL.\(^{24}\) However, no T2D GWAS cis-eQTLs in stretch/super enhancers have been identified in any tissues. To identify genetic regulatory signatures that may contribute to the diabetes phenotype, we assessed the overlap of our muscle cis-eQTL catalogue with 225 GWAS SNPs associated with T2D and 7n T2D-related traits (see Methods). Of the 220 GWAS SNPs assessed in our study, 99 SNPs in 218 GWAS SNP–gene pairs (of a total 4,545 GWAS SNP–gene pairs) had \( \geq 1 \) significantly associated genes. We performed iterative conditional analysis to identify GWAS cis-eQTL SNPs likely to have substantial stronger expression associations in the same gene (see Methods). 53 variants in 78 GWAS SNP–gene pairs (59 unique genes) remained associated (FDR \(< 5\%\) for the conditional analysis); of these 38 of the 53 variants remained after pruning at \( r^2 < 0.2\). The top conditional cis-eQTL signals for GWAS SNPs are shown in Table 1; the full list is shown in Supplementary Data 1.

We observed a top T2D GWAS SNP cis-eQTL signal for the muscle-specific (mESI decile = 10) gene ANK1 (Fig. 4a), where the T2D risk allele at rs516946 resulted in increased gene expression (Table 1). Although the underlying molecular mechanisms were unknown at the time, this locus was first reported as being associated with T2D\(^{25,26}\), the results we present here help define the impact on skeletal muscle gene expression. This cis-eQTL SNP resides in a skeletal muscle stretch/super enhancer (Fig. 4b, Supplementary Fig. 14). Notably, there are no amino-acid-altering variants in strong LD \((r^2 \geq 0.8)\) with rs516946, but several overlap active chromatin marks in skeletal muscle tissue and human skeletal muscle myoblast (HSMM) cells (Supplementary Fig. 15). One such SNP, rs508419, is flanked by skeletal muscle stretch enhancers, resides in an active promoter specific to skeletal muscle and HSMM, and overlaps an ATAC-seq peak in our muscle data (Fig. 4c). This active promoter ATAC-seq peak SNP is immediately upstream of the TSS for highly expressed short isoforms of ANK1 (Fig. 4b, see inset bar plot) and is predicted by our SNP-aware motif scans to disrupt a TR4-binding site, consistent with what is reported in the HaploReg database\(^{27}\).
We tested each allele of rs508419 in an electrophoretic mobility shift assay (EMSA) using nuclear extract from human skeletal muscle cells (SkMC). We found that the non-risk allele is preferentially bound compared with the risk allele (Fig. 4d). This allelic effect was replicated in both human (HSMM) and mouse (C2C12) myoblast cells (Supplementary Fig. 16). Our ATAC-seq footprinting results indicate that TR4 is bound at this position (CENTIPEDE posterior probability = 1), and we observe an allele-specific supershift in the EMSA using a TR4 antibody (Fig. 4d). Together, these results suggest that TR4 binding at the non-risk allele is linked to transcriptional repression, consistent with previous reports about TR4 activity28,29, and that TR4 binding and repression is disrupted by the T2D risk allele rs508419, which results in increased ANK1 expression.

Further splicing (sQTL) and exon usage (exonQTL) investigation (see Methods) of the muscle mRNA-seq data revealed complex splicing patterns that are not apparent when considering bulk gene expression. sQTL analysis of the four short ANK1 isoforms with a TSS near the active muscle promoter rs608419 SNP (Fig. 5a) showed that abundance for two of the isoforms (ENST00000522543.1 and ENST00000314214.8) is significantly affected by rs508419 genotype, whereas the other two (ENST00000457297.1 and ENST00000522231.1) appear to be invariant (Fig. 5b,c). Interestingly, the two variable transcripts change expression by rs508419 T2D risk allele copies in opposite directions: ENST00000522543.1 decreases expression while ENST00000314214.8 increases it (Fig. 5c). We confirmed this sQTL analysis using an independent exonQTL analysis, quantitative reverse transcription–PCR and droplet digital PCR (ddPCR) (Supplementary Figs 17–20).

### Discussion

Protein products of the small ANK1 isoforms we identified as linked to a T2D GWAS SNP through our skeletal muscle cis-eQTL map interact with obscurin30,31, a critical component of, and required for, proper sarcoplasmic reticulum (SR) assembly32,33. The SR is involved in insulin action on glucose uptake through the regulation of GLUT4 translocation to the plasma membrane (reviewed in ref. 34). Recently, ANK1 was shown to interact with IRS1 in skeletal muscle35. IRS1 is indispensable in insulin action on glucose uptake in human SkMCs36. Thus, ANK1 is linked to glucose uptake in muscle, and alterations to its expression might perturb this process leading to an insulin-resistant state. However, how the different small ANK1 isoforms influence insulin-stimulated glucose uptake is presently unclear. Functional studies to better characterize this process are needed.

In summary, we report here on a genome-wide genetic and mRNA-seq analysis of the largest set of clinically characterized human muscle samples described to date. We observed lower expression in T2D individuals for genes involved with cellular respiration, consistent with a previous smaller study7. We demonstrated that the genetic regulatory architecture of muscle-specific gene expression is highly and specifically enriched in muscle stretch/super enhancers. We identified T2D and related trait GWAS SNPs as cis-eQTLs for several genes, including complex transcriptional and splicing regulation of the muscle-specific isoforms of ANK1 that is associated with SNPs that reside within a muscle stretch enhancer. Together, these studies define links between GWAS SNPs and their target genes in skeletal muscle, providing functional insights with potential precision therapeutic implications for T2D.

### Methods

#### Sample recruitment

We attempted to contact still-living FUSION spouses and offspring who participated in FUSION study visits between 1994 and 1998 (ref. 37), individuals who had participated in the population-based Savitaipale Prospective Diabetes Study38, the FINRISK 2007 survey, the Dose Responses to Exercise Training (DR’s EXTRA) study39 and the Metabolic Syndrome in Men (METSIM)
Figure 4 | T2D GWAS SNPs in a muscle-specific stretch enhancer of ANK1 provide mechanistic insights into T2D pathophysiology in skeletal muscle.

(a) Regional association plot showing the significance of SNPs (points) associated with expression of ANK1 (highlighted with red rectangle), where the best cis-eQTL rs516946 (purple point) is a T2D GWAS SNP. (b) UCSC genome browser view of chromatin states near ANK1 isoforms. The chromatin states between skeletal muscle and other T2D relevant cell types (adipose, liver, islets) are markedly different. ANK1 is associated with hereditary spherocytosis, a disease of the red blood cells, which is consistent with the transcribed chromatin states in K562, a myelogenous leukaemia line of the erythroleukemia type. T2D and related trait GWAS SNPs (dark green) and SNPs in strong LD ($r^2 > 0.8$; blue) are found within muscle-specific stretch enhancers. ANK1 expression is shown for each isoform and normalized so that the sum over all isoforms is 1. (c) The chromatin states across T2D relevant tissues or cells for SNPs in strong LD ($r^2 > 0.8$) with T2D GWAS SNP rs516946. Location of the ATAC-seq peak is noted. Asterisks denote SNPs that reside in stretch ($\geq 3$ kb) enhancers. Colour-coding of chromatin is as shown in b. (d) EMSA using human skeletal muscle cell (SkMC) nuclear extract demonstrates allele-specific binding (see lower horizontal arrow on the left side of the gel) for the non-risk allele (A) of rs508419. A supershift (see upper horizontal arrow on the left side of the gel) using the TR4 antibody shows that TR4 participates in the allele-specific binding.

study. Additional subjects were recruited by newspaper advertisements. We excluded individuals: (1) with drug treatment for diabetes, (2) with diseases that might be expected to confound the analyses (for example, cancer, skeletal muscle diseases, acute or chronic inflammatory diseases), (3) with diseases that increase haemorrhage risk during biopsy (for example, von Willebrand’s disease, haemophilia, severe liver diseases), (4) taking medications that need to be taken daily and increase haemorrhage risk in the biopsies including warfarin (patients on acetylsalicylic acid were instructed to stop for 7 days prior to biopsy), (5) taking medications that could confound the analyses (for example, oral corticosteroids, other anti-inflammatory drugs such as S-ASA, infliximab or methotrexate), and (6) of age < 18 years. The study was approved by the coordinating ethics committee of the Hospital District of Helsinki and Uusimaa. A written informed consent was obtained from all the subjects.

Clinical visit. We performed clinical visits in Helsinki, Savitaipale and Kuopio. 279 individuals participated in both clinical and muscle biopsy visits (see below). The clinical visit included a 2-h, four-point OGTT and other phenotypes measured after a 12-h overnight fast, and health history, medication and lifestyle questionnaires. The clinical visit took place for an average of 14 days before the biopsy visit (90% of clinical visits ≤ 32 days before biopsy; range 89 days before to 15 days after). We defined glucose tolerance categories of NGT, impaired glucose tolerance (IGT), impaired fasting glucose (IFG) and T2D using World Health Organization (WHO) criteria. Analysis of Reimbursement.

Phenotype measurements and laboratory analysis. We measured height and weight in light clothing. Waist circumference was measured midway between the lower rib margin and the iliac crest. We determined OGTT plasma glucose (fluoride citrate plasma) concentrations by hexokinase assay (Abbott Architect analyzer, Abbott Laboratories, Abbott Park, IL, USA) and serum insulin by chemiluminescent microparticle immunoassay (Architect analyzer). Glucose and insulin analyses were done at a certified core laboratory at the National Institute for Health and Welfare, Helsinki, Finland.

Muscle biopsy visit. Biopsies were performed using a standardized protocol and one physician (T.A.L.) trained all doctors performing biopsies. We instructed participants to avoid strenuous exercise for at least 24 h prior to biopsy. Following overnight fast, we obtained ~ 250 mg vastus lateralis skeletal muscle using a...
conchotome, under local anesthesia with 20 mg ml\(^{-1}\) lidocaine hydrochloride without epinephrine. Altogether 9 experienced and well-trained physicians collected 331 muscle biopsies in 2009–2013 in 3 different study sites (Helsinki, Kuopio and Savitaipale). Three physicians, one in each site, performed most of the biopsies (237). All physicians were trained to perform the biopsy in an identical manner. The muscle samples were cleaned of blood, fat and other non-muscle tissue by scapel and forceps, rinsed with NaCl 0.9% solution, and frozen in liquid nitrogen. Samples were frozen within 30 s after sampling. Muscle samples were then stored at −80 °C for a duration of 0–4 years before analysis. Overall, the biopsy procedure was well-tolerated. Apart from a few expected cases of bruising, numbness at the biopsy site and vasovagal reactions, there were no clinically significant adverse sequelae.

### RNA isolation and mRNA sequencing

We visually dissected 30–50 mg of each frozen muscle biopsy sample to avoid adipose tissue. Total RNA was extracted and purified with Trizol (Invitrogen, Carlsbad, CA). RNA integrity numbers ranged from 7.2 to 9.4 (median 8.5). To minimize and quantify batch effects, we randomly queued samples for sequencing using a 24-sample barcode-pooling approach and targeted proportional representation of the OGTT states (NGT, IGT, IFG and T2D) in each sequencing batch. External RNA Controls Consortium (ERCC) RNA controls were spiked prior to barcoding to facilitate library quality control (QC). Poly(A)–selected RNA samples were sequenced by the NIH Intramural Sequencing Center (NISC) using the Illumina TruSeq directional mRNA-seq library protocol. About 80 million 100 bp paired-end reads per sample. In total, 41,512,000 41,522,000 41,514,000 41,516,000 41,518,000 41,520,000 41,522,000 reads mapping to the mitochondrial genome, with up to 52% mitochondrial reads in some sequencing batches. We examined library degradation within affected batches by inspecting gene body coverages for individual highly expressed single-isoform genes and ERCC spike-in transcripts. This did not show any evidence of RNA degradation, but did reveal the presence of systematic batch effects with patterns of read coverage heterogeneity along genes or spike-in transcripts, typically highly similar within batches with noticeable systematic differences between some batches. This result strongly suggests we can correct for these in downstream analyses. Further analysis revealed that batches with high-mitochondrial read fractions had lower estimated RNA fragment lengths.

To address variability in mitochondrial read fraction, we calculated fragments per kilobase transcript per million mapped reads (FPKMs) separately for nuclear genes, mitochondrial genes and ERCC transcripts. PCA on the nuclear genome FPKM matrix and colour labelling by sequencing batch also revealed the presence of systematic batch effects. We used the PEER framework for correcting these batch effects. In our primary trait–gene expression association analysis, we correct for these observed batch effects. In the cis-eQTL analyses, we correct for these observed batch effects, as well as unknown technical confounders, by using the PEER framework. PCA after PEER correction effectively removed batch effects. In our primary trait–gene expression association analysis, we correct for these observed batch effects. In the cis-eQTL analyses, we correct for these observed batch effects, as well as unknown technical confounders, by using the PEER framework. PCA after PEER correction effectively removed batch effects (Supplementary Fig. 21). We further investigated possible technical confounders in mRNA-seq by checking for sample-specific GC-content biases, which have been reported in early RNA-seq studies. We found no evidence for such effects in our data.

We used verifyBamID\(^{47}\) to check for sample swaps and contamination. We asked if reads in the RNA-seq BAM files matched the SNP chip genotype data in transcribed regions for each individual, and determined whether BAM files were contaminated and comprised of reads derived from more than one individual. We identified two pairs of sample swaps which we were able to correct; one of these samples had an estimated contamination of 8% and was excluded. As an additional check for sample swaps and to detect outliers, we verified reported sex by examining expression of the XIST gene and mean Y chromosome gene expression. PCA on the FPKM matrix identified four samples as outliers; we removed them from further analyses. One additional sample was an outlier with respect to age (20 years) and was excluded. After all QC exclusions, 271 samples remained and were used for trait-associated expression analyses; of these, 267 had genotype data available and were used for cis-eQTL analyses.

**Figure 5** | Truncated ANK1 isoforms are expressed in muscle and have differential splicing associated with rs508419 genotype. (a) The four short isoforms of ANK1. (b) The mean expression of each isoform normalized by the total ANK1 expression is stratified by genotype; error bars represent ±1 s.d. N = 8/82/174 for risk allele dosages 0/1/2. (c) The direction of effect (β) for splicing associated with the rs508419 risk allele (y-axis labels are the same as in b).
Expression quantification. To study a wide spectrum of regulatory variation, we performed analyses at three levels: gene, exon and transcript. Definitions for all transcriptome features were based on GENCODE v19 (ref. 43). We counted fragments mapping to genes using htsqecount v0.5.4 (ref. 48) (http://www-huber.embl.de/users/anders/HTSeq/doc/count.html) and calculated FPKM values for each gene. For differential expression analysis and cis-eQTL mapping, we filtered for genes with five or more counts in ≥25% of samples. We counted reads in exonic parts of genes using dsexqcov count v1.0.2 (ref. 49) and calculated exon-level FPKMs for all transcripts in the GENCODE v19 comprehensive annotation. To avoid double-counting of fragments in the quantification of exon abundance, we clipped overlapping read pair mates using the ClipOverlap function of BamUtil (http://genome.sph.umich.edu/wiki/BamUtil:_clipOverlap). We estimated transcript abundance using rseqC (http://www-personal.umich.edu/~jiangshu/rseqC/) which is based on a Poisson regression model50. This model uses information from the insert length distribution inferred using the aligned read fragments which has been shown to help improve estimation. To reduce the number of transcripts per gene to avoid identifiability issues and to restrict analysis to high-confidence transcripts, we estimated transcript expression values for the subset of GENCODE transcripts with the tag ‘basic’ in the GTF file.

Trait–gene expression association. We use ‘trait’ to refer to T2D status or a related quantitative trait. First, we describe trait–gene expression analyses with adjustment for known covariates. Second, we describe an analysis that also includes adjustment for unknown factors learned from the gene expression data. Our primary analysis was without adjustment for unknown factors since we were concerned that the unknown factors might include biological signal, as well as expression differences due to technical issues. We compare the results of the two analyses.

T2D–gene expression association. For individual $i$ and gene $j$, let $T_i$ denote T2D status ($0 =$ NGT, $1 =$ T2D) and $Z_i = [Z_{i1}, Z_{i2}, ..., Z_{it}]^T$ the vector of $t$ covariates. Let $Y_{ij}$ denote the rank-based inverse normalized FPKM, where inverse normalization is performed for each gene, randomly breaking ties. We tested for association between $Y_{ij}$ and T2D status using the linear regression model:

$$Y_{ij} = \alpha_j + \beta_j T_i + \gamma_j Z_i + \epsilon_{ij}$$  

(1)

where $\alpha_j$ is the intercept, $\beta_j$ is the regression coefficient for T2D status on gene $j$, $\gamma_j$ a vector of coefficients for the covariates and $\epsilon_{ij}$ is a normally distributed error term with mean 0 and variance $\sigma^2$. We included as covariates age, sex and experimental batch.

Quantitative trait–gene expression association. To define a transformed quantitative trait $X_i$ we (1) inverse normalized the raw quantitative trait, (2) adjusted for age, sex and experimental batch by linear regression, and (3) inverse normalized the resulting residuals. To define the gene expression value, $Y_{ij}$ for each gene we (1) inverse normalized FPKM, (2) performed linear regression of age, sex and experimental batch on the inverse normalized FPKM, and (3) inverse normalized the resulting residuals. We then tested for association between transformed gene expression $Y_{ij}$ and each transformed quantitative trait $X_i$ using linear regression model:

$$Y_{ij} = \alpha_j + \beta_j X_i + \epsilon_{ij}$$  

(2)

where $\beta_j$ is the regression coefficient for $X_i$ on gene $j$.

We used FDR51 to account for multiple testing and considered as significant associations with FDR ≤ 5%.

Association analysis adjusting for unobserved confounders. To examine the effect of potential unobserved technical confounders on gene expression–trait association, we modified our definition of transformed gene expression $Y_{ij}$ by (1) inverse normalizing FPKM, (2) performing factor analysis via PEER52 on the inverse normalized FPKMs specifying from 1 to 15 factors, together with covariates age, sex and experimental batch, and (3) inverse normalizing the residuals. We used the transformed $Y_{ij}$ to perform differential expression analysis for each trait as above for T2D (equation (1)) and quantitative traits (equation (2)). For each trait, we selected the number of factors that maximized the number of differentially expressed genes. We ran GO term analysis and compared the results to those obtained without adjustment (see below)

Association analysis adjusting for tissue heterogeneity. The presence of non-muscle cells/tissue within the muscle biopsies may influence the trait and quantitative trait-expression analysis. To investigate this, we estimated tissue heterogeneity, via tissue component transcripts from Illumina Body Map 2.0, randomly subsampling reads to an equal amount and calculating FPKMs. For each reference tissue, we calculated the expression specificity index (see below) of each gene and selected the top 500 tissue-specific genes per tissue to use as the reference set. Using the combined tissue-specific gene set, we estimated the tissue heterogeneity of each skeletal muscle biopsy sample. In our samples, we estimated < 0.1% adipose contamination, 9–27% WBC, 0–18% lymphocyte and 64–86% skeletal muscle across samples.

To adjust for the effects of non-muscle tissue, we included the estimated percentages of WBCs and lymph as additional covariates in our trait and quantitative trait-expression analysis. We ran GO term analysis and compared the results with those obtained without adjustment (Supplementary Fig. 2).

GO term enrichment analysis. For each trait, we performed GO term enrichment analysis using RNA-Enrich53. For each trait, we define $P$ as the signed log ($\log(P$ value), signed as ‘+’ for trait–gene expression association in which higher values of gene expression are associated with T2D, and ‘−’ for lower gene expression associated with T2D. We used the logistic regression model

$$\logit(\pi_{ij}) = \alpha + \beta T_i + \gamma_j Z_i$$  

(3)

where $\pi_{ij}$ is the probability of GO term membership for gene $j$, $\alpha$ is the intercept, $\beta$ is the regression coefficient for GO term membership with $P$, the signed −log($P$ value), and $\gamma_j$ is the regression coefficient for the GO term membership with $L$, the log($L$ gene length). We include $L$ in the model to account for the potential confounding effect of gene length on the enrichment test; longer genes tend to have higher power for expression–trait association, and many GO terms contain set of genes that are substantially longer or shorter than average.

To present the GO term results, for each of the 4 traits tested we retained the 20 most statistically significant GO terms in which genes showed positive association with GO term and the 20 in which genes showed negative association with the trait. Within each trait and direction of association, we ranked the GO terms from 1 (most significant) to 20. We combined the 40 × 4 = 160 GO terms and assigned the lowest rank when GO terms appeared more than once on the list. We pruned redundant GO terms from the combined lists of GO terms, preferentially retaining GO terms with lower rank54. We hierarchically clustered the pruned GO terms based on the regression coefficients (βs) from equation (2) from the four traits using complete linkage and the Euclidean distance measure in four-dimensional space.

To assess the robustness of our findings, we repeated RNA-Enrich analysis using the PEER factor adjusted trait–gene expression association results (see above). Results were very similar.

Sample and genotype QC. We extracted DNA from blood. DNA samples were genotyped at the Genetic Resources Core Facility (GRCF) of the Johns Hopkins Institute of Genetic Medicine on the HumanOmni2.5-4v1_H BeadChip array (Illumina, San Diego, CA, USA). We mapped the Illumina array probe sequences to the hg19 genome assembly using the Burrows-Wheeler Aligner (BWA)55. We excluded SNPs with probe alignment problems, known variants in the 3′ end of probes, call rates <95%, minor allele count (MAC) <1 or Hardy–Weinberg equilibrium $P$ value <10^{-6}, leaving 1,642,012 SNPs for subsequent analysis. All alleles were oriented relative to the reference.

Of the 271 individuals that passed RNA-seq QC, we genotyped 267 samples, of which all were successful, with minimum call rate >98.7%. Based on 10 duplicate samples (from a larger set of genotyped samples), overall genotype concordance was 99.993%. We identified two unexpected pairs of first-degree relatives using KING (http://people.virginia.edu/~we9c/KING/). Each was an NGT-IGT pair; from each pair we excluded the NGT participant. We performed principal components analysis using SMARTPCA56 on 156,416 SNPs with MAF >5% and in near linkage equilibrium (r2 <0.2), after excluding SNPs from regions of high LD57. No population outliers were identified.

Imputation. We performed SNP imputation using a two-step strategy58. As reference panel we used the haplotypes from 2,737 European individuals sequenced in the GoT2D project. To improve phasing quality given the small target sample set, we pre-phased our 267 individuals together with the GoT2D reference panel samples using ShapeIT version 2 (https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html). We then imputed genotypes with Minimac2 (ref. 59). For chromosome X, we performed pre-phasing and imputation separately for the pseudo-autosomal and non-pseudo-autosomal regions. For cis-eQTL analysis, we included 8,406,237 variants with imputation quality $r^2 > 0.3$ and MAC >5.

cis-eQTL analysis. We performed cis expression quantitative trait (eQTL) analysis of SNPs within 1 Mb of the most upstream TSS of each gene using Matrix eQTL60 separately at the levels of gene, exon and transcript isoforms. To generate the gene expression value $Y_{ij}$, we (1) inverse normalizing FPKM, (2) performed factor analysis via PEER on the inverse normalized FPKM (separately for genes, exons and transcript isoforms; specifically, we performed an analysis of the first 1000 principal components based on age, sex, and normalizing to genes with five or more counts in ≥25% of samples. To test for association between a candidate eQTL $c$ and gene expression, we estimated the following regression model

$$Y_{ij} = \alpha + \beta_c G_{ic} + \epsilon_{ij}$$  

(4)

where $G_{ic}$ is the imputed allele count for SNP $c$ for individuals $i$, and $\beta_c$ is the...
regression coefficient of the imputed allele count for SNP \(p\) on transformed gene expression \(Y\).

We used FDR to account for multiple testing and considered as significant associations with FDR \(\leq 5\%\).

We present results based on 60 PEER factors since we expect that removing technical and biological variation will increase power to detect cis-eQTLs and we observed very little increase in the number of cis-eQTLs (FDR \(\leq 5\%\)) for >60 factors.

**Comparison to GTex skeletal musculagenic cis-eQTL results.** We downloaded the entire set of skeletal muscle SNP–gene association tests (V6) for 361 GTEx skeletal muscle samples from the GTEx portal (www.gtexportal.org). We called GTEx SNP–gene association tests with FDR \(\leq 5\%\) as significant. About 95,528,846 SNP–gene pairs (19,038 genes and 6,694,033 SNPs) were tested in common in the 2 studies and had concordant alleles. We oriented the association results in each study to the same effect allele.

**Genic cis-eQTL for GWAS variants for T2D and related traits.** We compiled a list of 225 GWAS variants for T2D, fasting glucose, fasting insulin and 2-h glucose, each with and without adjustment for BMI, and for fasting proinsulin, from the NHGRI GWAS catalogue and carried out manual curation of the literature to create a comprehensive list that was up-to-date as of May 2014. We crossed this list of 225 GWAS variants with a list of gene-based cis-eQTL results, resulting in a total of 220 SNPs and 4,545 tested GWAS SNP–gene pairs. To identify GWAS variant cis-eQTLs, we performed iterative conditional analysis on each of the 4,545 SNP–gene pairs. For each such pair, we predicted \(Y\) (as defined for cis-eQTL analysis) starting with the GWAS SNP genotype in the model, and then performed step-wise forward selection of SNPs within 1 Mb of the most upstream TSS, with a stopping threshold of a \(P\) value of 0.0019 (corresponding to the \(P\) value threshold for gene-based cis-eQTLs with FDR \(\leq 5\%\)).

**ASE measurements.** Our goal was to quantify ASE at each protein-coding SNP. Based on the calls from our DNA genotyping and imputation, for each individual we identified all sites at which a SNP in a gencode V19 annotated protein-coding exonic region was called as a heterozygote. At these sites, we quantified the strand-specific read coverage using samtools mpileup (version 0.1.18) to process the aligned RNA-seq read BAM files. We required a minimum mapping quality of 255, minimum base quality of 20 and reads mapped in a proper pair. We excluded reads that failed vendor quality checks or that were not the primary alignment. ASE was defined as the reference allele read count divided by reference + alternate read count, termed fracRef.

**ASE filtering to adjust for read mapping bias.** We implemented multiple filtering steps to identify and exclude SNPs susceptible to mapping errors that could bias ASE quantification. To identify SNPs with mapping biases, we simulated reads as previously described, except we used 101 bp reads, and after mapping the simulated reads excluded SNPs with a total simulated coverage of \(<193\) and \(>202\) SNPs based on additional filters. Although mono-allelic expression occurs at some sites, we required representation of both alleles in each SNP per individual by requiring 0.01 < fracRef < 0.99. In addition, we excluded SNPs in regions blacklisted by the ENCODE Project Consortium because of poor mappability or the presence of collapsed repeat regions. We excluded any SNP within 100 bp of an indel greater than 4 bp or overlapping an indel of any length.

**ASE statistical analysis.** fracRef varies systematically by the reference and alternate allele of the SNP. As previously described, starting with the filtered SNP list, for each sample, for each SNP reference and alternate allele pair (for example, AG and GA are separate allele pairs), we estimated the expected fracRef. The sample-specific and allele-pair-specific expected fracRef was calculated as the sum of the reference allele counts divided by the sum of the total allele counts across all SNP–gene pairs and alternative pair for an individual. To prevent SNPs with high coverage from biasing the estimated fracRef, SNPs with read count coverage in the top 25th percentile were down-sampled to 30 \(x\) coverage and the down-sampled reference allele and total count were used. For each individual and for each SNP, we performed a two-sided binomial test, using the observed sample-specific or allele-pair-specific fracRef as the true fracRef under the null hypothesis of no ASE. Of the set of SNPs for each individual, we call ASE significant if the Storey's FDR \(q\)-value \(\leq 0.05\). The forward and reverse stranded ASE quantifications were combined after filters and before statistical analysis for accurate sample-specific adjusted expectation calculation. To increase power to detect ASE, at each SNP that meets the above thresholds and occurred in \(\geq 10\) samples, we combined the binomial test \(P\) values using Fisher's combined probability test.

**Analysis of muscle-specific expression.** We used an information theory approach to score genes based on muscle expression level and specificity relative to the panel of 16 diverse Illumina Human Body Map 2.0 tissues. We first calculated expression \((e)\) in FPKM values for all Gencode v19 genes across each of the 16 tissues in the Body Map 2.0 data. We calculated the relative expression of each gene \((g)\) in skeletal muscle compared with all 16 tissues \((t)\) as:

\[
p_{s,\text{muscle}} = \frac{x_{s,\text{muscle}}}{\sum_{t} x_{s,t}}
\]

We next calculated the entropy for expression of each gene across all 16 tissues as:

\[
H_{g} = -\sum_{t} p_{s,t} \log(p_{s,t})
\]

Following previous studies, we defined muscle tissue expression specificity \((Q)\) for each gene as:

\[
Q_{g,\text{muscle}} = H_{g} - \log(p_{s,\text{muscle}})
\]

To aid in interpretability, we divided \(Q\) for each gene by the maximum observed \(Q\) and subtracted this value from 1 and refer to this new score as the mESI:

\[
m\text{ESI}_g = 1 - \frac{Q_g}{Q_{\text{max}}}
\]

mESI scores near zero represent low and/or ubiquitously expressed genes, and scores near 1 represent genes that are highly and specifically expressed in skeletal muscle. We note that although these calculations were performed using expression measured by FPKM, we obtained similar results when using transcripts per million.

**Chromatin state analyses.** We performed read mapping and integrative chromatin state analyses as in our previous report. Chromatin states were learned jointly by applying the ChromHMM (version 1.10) hidden Markov model (HMM) algorithm at 200 bp resolution to six data tracks (Input, K27ac, K27me3, K36me3, K4me1, K4me3) from each of the cell or tissue types. We collected cell or tissue ChIP-seq reads from a diverse set of publically available data representing 31 cells/tissues. We ran ChromHMM with a range of possible states, and settled on an 11-state model as it accurately captured information from higher state models and provided sufficient resolution to identify biologically meaningful patterns in a reproducible way. This process was similar to our previous analysis. To determine how our learned states relate to previously published states from nine published studies, we performed enrichment analyses comparing our states with the published states in each cell type (Supplementary Fig. 6). We also performed gene body feature overlaps, and TSS proximity analyses. This process led to a clear state assignment (Supplementary Fig. 6), which we used for all subsequent analyses.

We measured enrichment of cis-eQTLs at various levels of significance (FDR thresholds of 5, 1 and 0.1%) to overlap chromatin states by retaining the single best (lowest value) cis-eQTL per gene, and using the GREGOR tool to calculate enrichment relative to MAF, TSS-distance and number of LD neighbours-matched null SNP sets. The enrichment trends were consistent across the different FDR thresholds having slightly more pronounced trends. We report here the results for the FDR = 0.1% set. We used the following GREGOR parameters, which are what was reported in the original publication: \(r^2\) threshold = 0.8, LD window size = 1 Mb and minimum neighbour number = 500.

**ATAC-seq sample processing.** ATAC-seq libraries were prepared using a modified protocol based on previous studies. Frozen human skeletal muscle tissue (Zen-bio, Durham, NC, USA) was cut into a 100-mpg piece and disrupted using liquid nitrogen and a CellCrusher (Cellcrusher, Cork Ireland). Nuclei were isolated by placing disrupted tissue into ice cold nuclei isolation buffer (NIB) (20 mM Tris-HCl, 50 mM EDTA, 5 mM Spermidine, 0.15 mM Spermine, 0.1% mercaptoethanol, 40% Glycerol, pH 7.5). The solution was then filtered through a Miracloth (Calbiochem, San Diego, CA USA) and centrifuged at 1,100g for 10 min at 4 °C. The resultant nuclei pellet was washed with NIB containing 8.5% Triton X and RSB buffer (10 mM Tris-HCl, 10 mM NaCl, 3 mM MgCl2, pH 7.4). The nuclei pellet was resuspended in 50 μl RSB buffer and 1,100 (10 mg total tissue) and 1,500 (2 mg total tissue) dilutions were made. The transposition reactions were performed using a homemade Tn5 preparation (courtesy of the Jacob Titman laboratory). The DNA was then purified using the Qiagen MinElute PCR Purification kit (cat. No. 28004). Next, the libraries were PCR amplified to 9 cycles for the 1:10 dilution and 11 cycles for the 1:50 dilution. A final clean-up was performed with the Qiagen MinElute PCR Purification kit and samples were sent to bioanalyzer for QC.

**ATAC-seq analyses.** We sequenced each ATAC-seq library to a depth of \(\geq 48M\) reads using a 50bp paired-end read layout on an Illumina HiSeq. Adapter sequence was trimmed from the raw ATAC-seq reads using a tool we developed based on the one used in the original ATAC-seq paper. Trimmed reads were then aligned to the hg19 reference with BWA aln (http://bio-bwa.sourceforge.net/) version 0.7.12. Duplicate alignments were marked with Picard Tools (https://broadinstitute.github.io/picard/) version 1.131 and discarded for further analysis. We then filtered for properly paired alignments and for reads uniquely mapped with high quality to autosomal references, using samtools (https://samtools.github.io/) version 1.2 and
flags `-f 3 -F 4 -F 8 -F 256 -F 1024 -F 2048 -q 30'. Finally, we called peaks using MACS2 (https://github.com/tshulin/MACS2) version 2.1.0, with flags `-g hs -n model --shift -100 --extsize 200 -B --broad --keep-dup all' and retained peaks that met a 5% flags `-f 3 -F 4 -F 8 -F 256 -F 1024 -F 2048 -q 30'. Finally, we called peaks using MACS2 (https://github.com/tshulin/MACS2) version 2.1.0, with flags `-g hs -n model --shift -100 --extsize 200 -B --broad --keep-dup all' and retained peaks that met a 5% FDR. For comparative purposes, we performed the same read trimming, alignment, filtering and peak calling steps on publicly available ATAC-seq data in GM12878 (ref. 20) and adipose tissue21.

To detect potential TFBSs using both reference and alternate alleles, we extracted the alternate allele from biallelic SNPs and short indels from 1000 Genomes phase 3 (release v5) along with 29 bp of flanking sequence from the hg19 human reference on each side. We scanned the reference and these extracted alternate allele sequences with position weight matrices (PWMs) from ENCODE65, JASPAR88 and Jenal et al.99 using FIMO89. FIMO was run using nucleotide frequencies from the hg19 reference (40.9% GC), and the default P value cutoff (10^-4).

ATAC-seq footprints were called using CENTIPEDE22. Briefly, for each PWM scan result, we generated a strand-specific (relative to the motif orientation) 5-bp resolution matrix encoding the number of TruS integration events in a region ±100 bp from each motif occurrence. We split the ATAC-seq signal into three different categories based on the diverse fragment length distribution:36 149 bp, 150–324 bp and 325–400 bp. A motif occurrence was considered bound if the CENTIPEDE posterior probability was >0.99 and its coordinates were fully contained within an ATAC-seq peak. We only considered footprints for motifs that were significantly enriched in our data, using an enrichment approach as previously described99.

Electrophoretic mobility shift assay. EMSAs were performed as previously described22. Briefly, biotin-end-labelled complementary oligonucleotides were designed (Integrated DNA Technologies, Coralville, IA USA) around the variant rs508419 (5′-AGGATAGCT/GAGAGGCC-3′). Nuclear protein extract from human SKMCO myoblasts (HSM) (Lonza, Walkersville, MD USA) and mouse muscle myoblasts (C2C12) (ATCC, Manassas, VA USA) was prepared using the NE-PER Extraction Kit (Thermo Scientific, Waltham, MA). The LightShift Chemiluminescent EMSA Kit (Thermo Scientific) was used following the manufacturer’s protocol with binding reactions comprising 1× binding buffer, 1 microgram poly(dI-dC), 6–10 μg nuclear extract and 400 fmol of labelled probe. Reactions were incubated at room temperature for 30 min. Competition reactions included 10 μg nuclear protein extract and 400 fmol of labelled probe. EMSAs were performed and yielded comparable results.

Data availability. RNA-seq, ATAC-seq and genotype data for all samples used in this paper have been deposited in dbGaP with the accession code phs001068.v1.p1 and are available via the repository’s data access request procedures.

References
1. Scully, T. Diabetes in numbers. Nature 485, 52–53 (2012).
2. Kahn, S. E., Cooper, M. E. & De Prato, S. Pathophysiology and treatment of type 2 diabetes: perspectives on the past, present, and future. Lancet 383, 1068–1083 (2014).
3. Kelldon, S. et al. Expression of phosphofructokinase in skeletal muscle is influenced by genetic variation and associated with insulin sensitivity. Diabetes 63, 1154–1165 (2014).
4. Lindholm, M. E. et al. The human skeletal muscle transcriptome: sex differences, alternative splicing, and tissue homogeneity assessed with RNA sequencing. FASEB J. 28, 4571–4581 (2014).
5. GTEx Consortium. Human genomics. The genotype-tissue expression (GTEx) pilot analysis: multisite gene regulation in humans. Science 348, 648–660 (2015).
6. Lee, C., Patil, S. & Sartor, M. A. RNA-Enrich: a cut-off free functional enrichment testing method for RNA-seq with improved detection power. Bioinformatics (Oxford, England) 32, 1100–1102 (2016).
7. Mootha, V. K. et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat. Genet. 34, 227–232 (2003).
8. Albert, V. & Hall, M. N. mTOR signaling in cellular and organismal energetics. J. Cell Biol. 160, 235–233 (2003).
9. Ackermann, M. A. et al. Integrity of the network sarcoplasmic reticulum in skeletal muscle requires small ankyrin 1. J. Cell Sci. 124, 3619–3630 (2011).
10. Contreras-Ferrat, A., Lavandero, S., Jaimovich, E. & Klip, A. Calcium signaling differences, alternative splicing, and tissue homogeneity assessed with RNA sequencing. J. Cell Sci. 124, D877–D881 (2011).
11. Tanabe, O. et al. The TR2 and TR4 orphan nuclear receptors repress Gata1 transcription. Genes Dev. 21, 2832–2844 (2007).
12. Cui, S. et al. Nuclear receptors TR2 and TR4 recruit multiple epigenetic transcriptional corepressors that associate specifically with the embryonic β-tropin globin promoters in differentiated adult erythroid cells. Mol. Cell. Biol. 31, 3298–3311 (2011).
13. Borzok, M. A., Catino, D. H., Nicholson, J. D., Kontrogianni-Kontopoulous, A. & Bloch, R. J. Mapping the binding site on small ankyrin 1 for obscurin. J. Biol. Chem. 282, 32384–32396 (2007).
14. Willis, C. D., Oashi, T., Busby, B., Mackerell, A. D. & Bloch, R. J. Hydrophobic residues in small ankyrin 1 participate in binding to obscurin. Mol. Membr. Biol. 29, 36–51 (2012).
15. Bagnato, P., Barone, V., Giacomello, E., Rossi, D. & Sorrentino, V. Binding of an ankyrin-1 isoform to obscurin suggests a molecular link between the sarcoplasmic reticulum and myotubules in striated muscles. J. Cell Biol. 160, 245–253 (2003).
16. Ackermann, M. A. et al. Integrity of the network sarcoplasmic reticulum in skeletal muscle requires small ankyrin 1. J. Cell Sci. 124, 3619–3630 (2011).
17. Contreras-Ferrat, A., Lavandero, S., Jaimovich, E. & Klip, A. Calcium signaling differences, alternative splicing, and tissue homogeneity assessed with RNA sequencing. J. Cell Sci. 124, D877–D881 (2011).
18. Caruso, M. et al. Increased interaction with insulin receptor substrate 1, a novel abnormality in insulin resistance and type 2 diabetes. Diabetes 63, 1933–1947 (2014).
19. Bouzakri, K. et al. RNA-based gene silencing reveals specialized roles of IRS-1/Akt2 and IRS-2/Akt1 in glucose and lipid metabolism in human skeletal muscle. Cell Metab. 4, 89–96 (2006).
20. Valle, T. et al. Mapping genes for NIDDM. Design of the Finland-United States Investigation of NIDDM Genetics (FUSION) Study. Diabetes Care 21, 949–958 (1998).
21. Vaātinan, S. et al. Quality of life along the diabetes continuum: a cross-sectional view of health-related quality of life and general health status in middle-aged and older Finns. Qual. Life Res. 23, 1935–1948 (2014).
22. Kouki, R. et al. Diet, fitness and metabolic syndrome—the DR’s EXTRA study. Nutr. Metab. Cardiovasc. Dis. 22, 553–560 (2012).
23. Stančáková, A. et al. Association of 18 confirmed susceptibility loci for type 2 diabetes with indices of insulin release, proinsulin conversion, and insulin sensitivity in 5,527 nondiabetic Finnish men. Diabetes 58, 2129–2136 (2009).
24. landscaped Health Organization (WHO) & International Diabetes Federation (IDF). Definition and Diagnosis of Diabetes Mellitus and Intermediate Hyperglycaemia: Report of a WHO/IDF Consultation. (WHO, Geneva, Switzerland, 2006).
42. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21 (2013).
43. Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 22, 1760–1774 (2012).
44. Hartley, S. W. & Mullikin, J. C. QoRTs: a comprehensive testset for quality control and data processing of RNA-Seq experiments. BMC Bioinformatics 16, 224 (2015).
45. Jiang, L. et al. Synthetic spike-in standards for RNA-seq experiments. Genome Res. 21, 1543–1551 (2011).
46. Stegle, O., Parts, L., Durbin, R. & Winn, J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. PLoS Comput. Biol. 6, e1000770 (2010).
47. Jun, G. et al. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Am. J. Hum. Genet. 91, 839–848 (2012).
48. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics (Oxford, England) 31, 166–169 (2015).
49. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. Genome Res. 22, 2008–2017 (2012).
50. Salzman, J., Jiang, H. & Wong, W. H. Statistical modeling of RNA-seq data. Stat. Sci. 26, 62–83 (2011).
51. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. Proc. Natl Acad. Sci. USA 100, 9440–9445 (2003).
52. Stegle, O., Parts, L., Pipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PERE) to obtain increased power and interpretability of gene expression analyses. Nat. Protoc. 7, 500–507 (2012).
53. Gong, T. & Szustakowski, J. D. DecorRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. Bioinformatics 29, 1083–1085 (2013).
54. Supek, F., Bosnjak, M., Skunca, N. & Smuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS ONE 6, e21800 (2011).
55. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England) 25, 1754–1760 (2009).
56. Patterson, N., Price, A. L. & Reich, D. Population structure and eigeanalaysis. PLoS Genet. 2, e190 (2006).
57. Price, A. L. et al. Long-range LD can confound genome scans in admixed populations. Am. J. Hum. Genet. 83, 132–135 author reply 135–9 (2008).
58. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. 44, 955–959 (2012).
59. Fuchsberger, C., Abecasis, G. R. & Hinds, D. A minimac2: faster genotype imputation. Bioinformatics 31, 782–784 (2015).
60. Shabalinn, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics 28, 1353–1358 (2012).
61. Welter, D. et al. The NHGRI GWAS Catalogue, a curated resource of SNP-trait associations. Nucleic Acids Res. 42, D1001–D1006 (2014).
62. Consortium, T. E. P. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).
63. Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. Nature 518, 317–330 (2015).
64. Mikkelsen, T. S. et al. Comparative epigenomic analysis of murine and human adipogenesis. Cell 143, 156–169 (2010).
65. Kheradpour, P. & Kellis, M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. Nucleic Acids Res. 42, 2976–2987 (2014).
66. Mathelier, A. et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 44, D110–D115 (2016).
67. Johlke, A. et al. DNA-binding specificities of human transcription factors. Cell 152, 327–339 (2013).
68. Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. The MEME suite. Nucleic Acids Res. 43, W39–W49 (2015).
69. Pacis, A. et al. Bacterial infection remodels the DNA methylation landscape of human dendritic cells. Genome Res. 25, 1801–1811 (2015).
70. Fogarty, M. P., Cannon, M. E., Vadlamudi, S., Gaulton, K. J. & Mohlke, K. L. Identification of a regulatory variant that binds FOXA1 and FOXA2 at the qCDC123/CAMK1D type 2 diabetes GWAS locus. PLoS Genet. 10, e1004633 (2014).

Acknowledgements
This research was supported in part by US National Institutes of Health grants 1-ZIA-HG000204 (to F.S.C.), R01DK099240 (to S.C.J.P.), R01DK072193 and U01DK0889256 (to K.L.M.), the American Diabetes Association Pathway to Stop Diabetes Grant 1-14-IN1-07 (to S.C.J.P.), and Academy of Finland Grants 271961 and 272741 (to M.L.) and 258753 (to H.A.K.). We thank Yun Li for helpful discussions about RNA-seq analyses, James Douglas Engel for the TR4 antibody, Jacob Kitzman for Trns transposase preparation, Roger Pique-Regi for assistance with CENTIPEDE, Darryl Leja for artwork in Fig. 1a, all the subjects for participation and the study personnel for excellent technical assistance.

Author contributions
L.J.S., M.R.E., T.A.L., M.L., J.T., H.A.K., M.B., F.S.C. and S.C.J.P. designed the study; L.J.S., J.R.H., P.S.C., J.P.D., N.N., L.L.B., K.L.M., M.L., J.T., H.A.K., M.B., F.S.C. and S.C.J.P. wrote the manuscript; L.J.S., J.R.H., P.S.C., J.P.D., N.N., L.L.B., K.L.M., M.L., J.T., H.A.K., M.B., F.S.C. and S.C.J.P. helped guide experiments and interpretation of results; M.R.E., S.V., L.K., J.Sa., J.Su., A.K., H.A.K. and T.A.L. performed experimental assays; and L.J.S., J.R.H., R.P.W., A.T.B., B.N.W., P.S.C., J.P.D., N.N., H.M.S., D.L.T., A.U.J., H.J., R.d'O.A., J.H., X.W. and S.C.J.P. performed computational and statistical analyses of the data.

Additional information
Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications
Competing financial interests: The authors declare no competing financial interest.
Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/
How to cite this article: Scott, L. J. et al. The genetic regulatory signature of type 2 diabetes in human skeletal muscle. Nat. Commun. 7:11764 doi: 10.1038/ncomms11764 (2016).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/