Genome-wide analysis in UK Biobank identifies over 100 QTLs associated with muscle mass variability in middle age individuals

Authors: Ana I Hernandez Cordero,[1] Jennifer S Gregory,[1] Alex Douglas,[2] and Arimantas Lionikas[1,*]

[1] School of Medicine, Medical Sciences and Nutrition, University of Aberdeen, Aberdeen, AB24 3FX, UK

[2] Institute of Biological and Environmental Sciences, University of Aberdeen, Aberdeen, AB24 3FX, UK

[*]Corresponding Author a.lionikas@abdn.ac.uk

Abstract

Muscle bulk in humans is highly variable even after accounting for differences in height, age and sex, hence the impact of aging-related muscle loss known as sarcopenia varies in a similar fashion. Although heritability estimates are 40-80%, only a small number of muscle mass affecting genes have been identified.

In 95,545 genotyped individuals of predominantly British ancestry, aged 37-48 years, body composition was assessed using bioelectrical impedance. We aimed to identify the genetic architecture underlying variability in appendicular lean mass (ALM), a proxy for muscle mass.

A genome wide-association study (GWAS) in the “Discovery” cohort containing 60% of individuals identified 209 single nucleotide polymorphisms (SNP) with significant ($P < 5 \times 10^{-8}$) effects on ALM. We confirmed 62% of the SNPs ($P < 2 \times 10^{-5}$) in the remaining sub-set, the “Replication” cohort. A subsequent GWAS in the Combined cohort revealed 132 significant quantitative trait loci (QTLs) that collectively explained 6.75% of ALM phenotypic variance. Sixteen novel genes with missense and frameshift polymorphisms and expressed in skeletal muscle emerged (*THBS3, CEP120, STC2, WSCD2, CCDC92, AKAP13, CPNE1, PPM1J, ITSN2, C2orf16, FGFR4, MLXIPL, PTCH1, BDNF, NCOR2, WDR90*). Furthermore, non-coding SNPs located in 74 regulatory elements suggest gene expression being an important contributor to variation in muscle mass.

In conclusion, we identified genetic architecture explaining a significant fraction of the phenotypic variation in middle age human ALM. The highlighted genes and regulatory elements will help understand the mechanisms underlying differences in skeletal muscle mass that can affect the risk and impact of sarcopenia.

Introduction

The skeletal muscle plays key roles in locomotion, respiration, thermoregulation, maintenance of glucose homeostasis and protection of bones and viscera. The loss of muscle due to aging, known as sarcopenia, affects mobility and can lead to frailty and deterioration of quality of life (1). The risk of disability is 1.5 to 4.6 times higher in older individuals affected by sarcopenia than in the age matched individuals with normal muscle mass (2). Skeletal muscle tissue, particularly in the limbs of the body, is the most abundant component of lean mass assessed in humans non invasively by DXA or bioelectric impedance analysis. These analyses demonstrated that lean mass varies extensively, differing by more than two fold between healthy adult individuals of same sex, age and height (3). It is conceivable that accretion of lower muscle mass by middle age could significantly influence the risk of sarcopenia and frailty later in life.

Genetic factors contribute substantially to the variability in lean body mass in humans, with heritability estimates of 40 – 80 % (4). A continuous distribution of the trait and data obtained from animal models (5-7) indicate a polygenic nature to this complex trait. However, thus far, genome-wide association studies (GWAS) have implicated fewer than a dozen genes, explaining only a small fraction of this heritability (8, 9). A limited sample size in early studies (10-14), the confounding effect of subjects age (8) and restrictive analytical models might have hindered detection of genes previously.

Muscle mass depends on the girth (15) and length of the muscles in the body (16). The number and cross-sectional area of constituent muscle fibres determine the former, whereas the size of skeleton directly influences the latter. The enlargement of the cross-sectional area and the axial growth of muscle fibres depend on fusion of satellite cells that supplement the pool of existing myonuclei in the fibre. Satellite cells are a lineage of postnatal myogenic progenitors originating from the Pax3-expressing embryonic progenitors (17). However, the same embryonic progenitors give rise to at least some precursors of white and brown adipose tissue (18). Therefore, innate differences in muscle mass might stem from the intrinsic properties of the myogenic lineage, differences in bone length, and can be developmentally linked to the quantity of adipose tissue. For a comprehensive understanding of the underlying mechanisms, it is important to consider all these scenarios.

The aim of the study was to identify the genetic architecture underlying variability in appendicular lean mass (ALM) in middle age adults. We addressed this question in a large cohort of densely genotyped individuals of middle age from the UK Biobank (UKB), which permitted superior QTL detection power and resolution.

Results

Genetic architecture of lean appendicular mass

Discovery cohort analyses

The ALM varied extensively in the Discovery cohort ranging between 11.8 and 41.6 kg in females, and 15.3 and 42.5 kg in males (Table 1). There was a strong and positive phenotypic correlation between ALM, size of the skeleton and the whole body fat (WBF) content (Supplementary Material, Table S1). The proportion of ALM

phenotypic variance explained by the all genotyped single nucleotide polymorphisms (SNP) was 0.40 ± 0.006.

Genome wide association analysis in the Discovery cohort revealed a total of 209 SNPs significantly associated ($P < 5 \times 10^{-8}$) with ALM after all four fixed effect combinations were explored (see Materials and Methods section). Due to the polygenicity of a complex trait, $\lambda > 1$ is expected (Supplementary Material, Figure S1). Indeed, the obtained λ were 1.14 for model A (sex adjusted) and B (sex and whole body fat (WBF) adjusted), 1.13 for model C (sex, WBF and leg length) and 1.12 for model D (sex and leg length). Many SNPs that showed a significant association were in close proximity to each other (Supplementary Material, Figure S2 and Figure S3) and often in linkage disequilibrium (LD) (Supplementary Material, Figure S4), hence these were considered part of the same QTL. We found that the genetic architecture of ALM in the Discovery cohort consisted of 47 QTL (Supplementary Material, Table S2) and their cumulative proportion of phenotypic variance explained (PVE) was 3.72 %. A number of these QTLs, 35 out of 47, were detected in one but not the other fixed effect models (Figure 1; Supplementary Material, Figure S5 – Figure S8), whereas the remaining 12 loci were consistently captured regardless of the models used (Supplementary Material, Table S2).

Table 1. Phenotypic variability and SNP heritability in the Discovery and Replication cohort

Cohort	Trait	N records	MIN	MAX	MEAN	SD	SNP $h^2 \pm SE$
Discovery	Appendicular lean mass (kg)	Females = 31,030	11.80	41.60	20.00	2.60	0.40 ± 0.006
		Males = 26,266	15.30	42.50	29.89	3.77	
Replication	Appendicular lean mass (kg)	Females = 20,313	12.20	40.4	20.04	2.62	-
		Males = 17,535	17.50	42.50	29.91	3.80	

Columns: 1) Cohort, 2) Trait, 3) N records: number of subjects grouped by sex, 4) MIN: minimum value of appendicular lean mass within the records of each cohort, 5) MAX: maximum value of appendicular lean mass within the records on each cohort, 6) MEAN: average appendicular lean mass within each cohort, 7) SD: standard deviation, 8)SNP h^2: SNP heritability based on all genotyped SNP(n = 524,481).

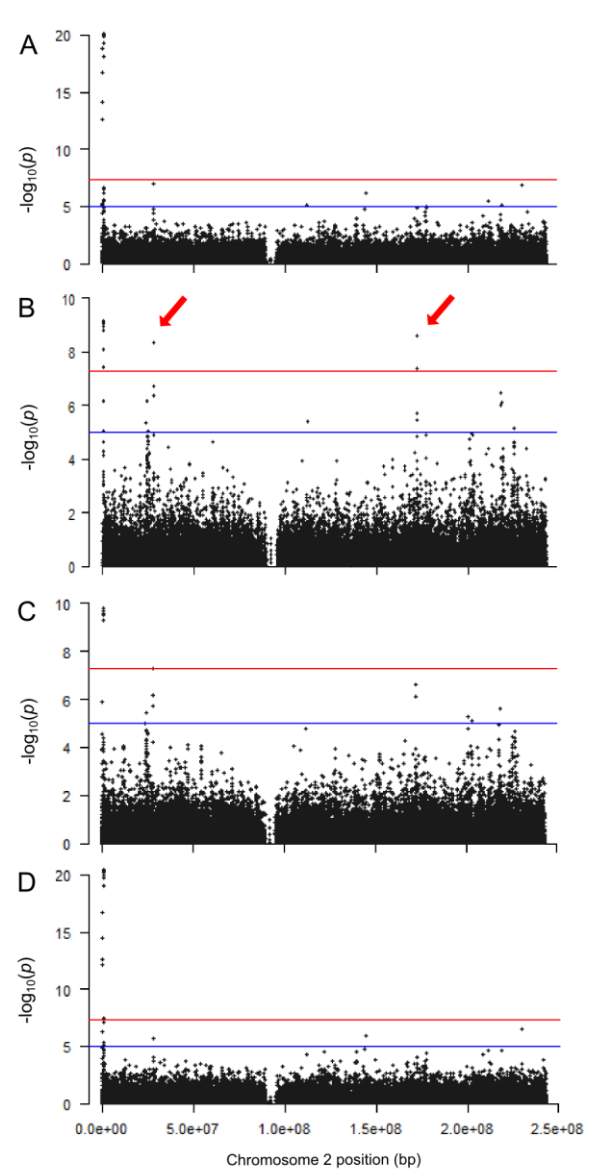

Figure 1. Model-specific loci on Chromosome 2. Letters A, B, C and D represent model used in the association analyses. In addition to 4 principal components, fixed effects in each model included: A) sex; B) sex and whole body fat (WBF); C) sex, WBF and leg length; D) sex and leg length.

Replication cohort analyses

The SNPs with the lowest p-value at each QTL and their effect size, beta, estimated in the Discovery cohort (Table S2) were used to test the effect of the sarcopenia risk

score (i.e. cumulative effect of all decreasing alleles) conferred by this genetic architecture on ALM in the Replication cohort. We employed Mendelian randomisation approach for this purpose and partitioned individuals of this cohort into five quantiles based on sarcopenia risk score (see Methods for details). The quantile effect explained a significant ($X^2 = 307.6$, df = 4, $P = 2.49 \times 10^{-65}$) proportion of the ALM variance in the Replication cohort; quantile with the highest sarcopenia risk score, Q5 (23.80 ± 0.07 kg), had on average a 1.47 kg lower ALM ($P = 9.81 \times 10^{-58}$) compared to that with the lowest sarcopenia risk score, Q1 (25.27 ± 0.07 kg). There was no significant difference between quantiles of ALM when using the negative control markers (Figure 2).

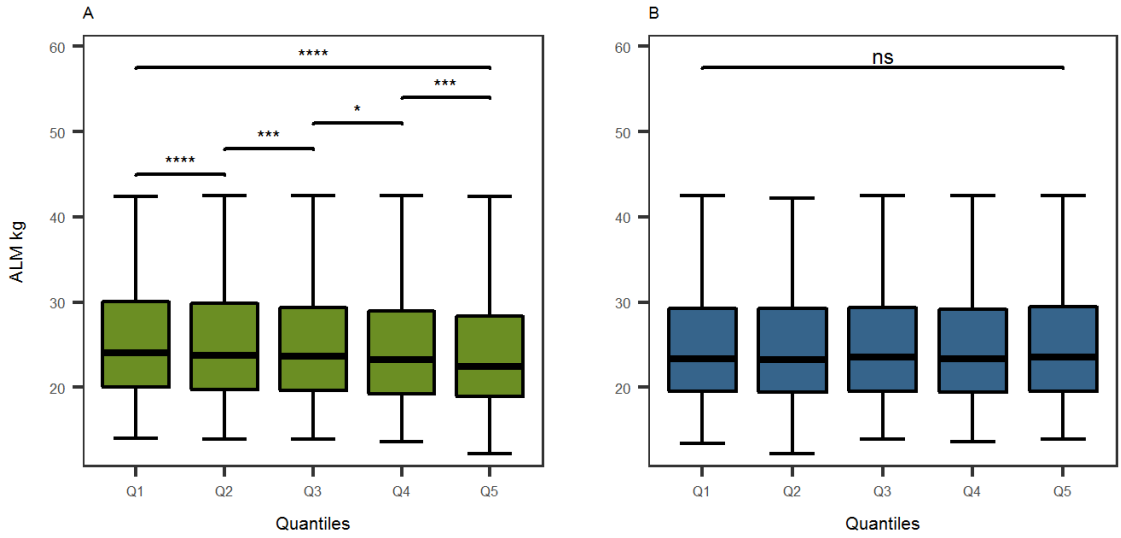

Figure 2. Genetic architecture identified in the Discovery cohort predicts muscle mass in the Replication cohort (n=38,000). Vertical axis represents the appendicular lean mass in kilograms. Horizontal axis represents the five quantiles grouped based on the sarcopenia risk score. Whiskers represent minimum and maximum values, distance between a whisker and the top or bottom of the box contains 25% of the distribution, the box captures 50% of the distribution, bold horizontal line represents

median. Group test for the replicating SNPs ($X^2 = 307.6$, df = 4, $P = 2.49 \times 10^{-65}$) in panel A, and negative control SNPs (*P = 0.67*) in panel B. Pairwise comparison: *= $P < 0.05$, **= $P < 0.01$, ***= $P < 0.001$, ****= $P < 0.0001$, ns: $P > 0.05$.

After replicating the cumulative effect of the genetic architecture on ALM, we examined individual SNPs. The SNPs significantly associated with ALM in the Discovery cohort were selected and tested in the Replication cohort. Of the 209 SNPs identified, 129 SNPs were replicated based on the threshold showed in Table 2 and by the sign of the beta (Supplementary Material, Table S2). These SNPs clustered in 27 replicated QTLs (Table 2). The majority of SNPs were replicated across more than one model, but SNPs in 10 QTLs were only identified in models with a specific fixed effect combination (Figure 3 and Supplementary Material, Table S2). No negative control SNPs showed association with ALM. Integration of the findings obtained by four analytical models across the Discovery and Replication cohorts indicates that we identified a truly polygenic, reproducible and robust genetic architecture for ALM.

Table 2. Summary of replicated appendicular lean mass QTLs.

QTL id	top SNP	CHR	BP position	Alleles	MAF	Model	PVE	Discovery			Replication			Gene
								Beta	SE	p value	Beta	SE	p value	
ALM6	rs17391694	1	78623626	T/C	0.13	A	0.00058	0.16	0.03	2.40E-09	0.23	0.04	1.10E-10	Intergenic
ALM9	rs905938	1	154991389	C/T	0.26	C	0.00062	0.10	0.01	4.50E-11	0.12	0.03	2.40E-06	*DCST2*, *ZBTB7B*
ALM12	rs543874	1	177889480	G/A	0.20	B	0.00081	0.15	0.02	1.80E-12	0.20	0.03	4.80E-12	*SEC16B*
ALM15	rs2867125	2	622827	T/C	0.17	B	0.00145	-0.22	0.02	3.40E-21	-0.25	0.03	3.30E-16	Intergenic
ALM17	rs1260326	2	27730940	T/C	0.38	C	0.00044	-0.08	0.01	4.50E-09	-0.08	0.02	6.20E-06	*GCKR*
ALM19	rs72885917	2	172416376	C/A	0.23	C	0.00056	-0.09	0.02	2.60E-09	-0.11	0.02	4.10E-08	*CYBRD1*
ALM32	rs16895971	4	17884986	C/T	0.14	C	0.00143	-0.16	0.02	3.40E-20	-0.16	0.03	1.80E-06	*LCORL*
ALM51	rs2814993	6	34618893	A/G	0.14	A	0.00078	0.18	0.03	2.20E-11	0.18	0.03	6.80E-08	*C6orf106*
ALM54	rs2206277	6	50798526	T/C	0.18	B	0.00078	0.16	0.02	9.30E-13	0.13	0.03	7.20E-06	*TFAP2B*
ALM57	rs4549631	6	126966308	T/C	0.50	C	0.00058	-0.07	0.01	8.70E-09	0.09	0.02	2.00E-07	*PRELID1P1*
ALM63	rs2282978	7	92264410	C/T	0.32	A	0.00048	0.12	0.02	3.60E-09	0.17	0.02	1.40E-11	*CDK6*
ALM69	rs7846385	8	78160179	C/T	0.28	C	0.00092	0.10	0.01	1.90E-12	0.09	0.02	2.80E-07	Intergenic
ALM77	rs2274224	10	96039597	C/G	0.44	D	0.00129	0.10	0.01	7.60E-18	0.10	0.02	2.20E-09	*PLCE1*
ALM79	rs11042725	11	10325325	A/C	0.47	C	0.00076	-0.08	0.01	1.70E-10	-0.07	0.02	5.40E-06	*ADM*, *AMPD3*

ALM80	rs11030084	11	27643725	T/C	0.18	B	0.00055	-0.14	0.02	2.90E-09	-0.14	0.03	1.70E-06	*BDNF-AS*
ALM89	rs7132908	12	50263148	A/G	0.37	A	0.00059	0.12	0.02	1.00E-09	0.13	0.02	6.30E-08	*FAIM2*
ALM90	rs1351394	12	66351826	T/C	0.47	C	0.00126	0.11	0.01	3.90E-18	0.17	0.02	7.90E-14	*HMGA2*
ALM91	rs1498707	12	93997128	A/G	0.27	C	0.00056	0.08	0.01	2.50E-08	0.08	0.02	1.10E-05	Intergenic
ALM97	rs1326122	13	50722895	A/C	0.02	C	0.00061	0.26	0.05	2.60E-09	0.32	0.06	4.80E-08	*DLEU1*
ALM98	rs3118914	13	51116901	T/G	0.20	C	0.00128	-0.14	0.02	9.60E-18	-0.16	0.03	1.20E-08	*DLEU1*
ALM100	rs8024628	15	84586463	A/G	0.46	D	0.00057	0.07	0.01	6.90E-10	0.08	0.02	3.40E-07	*ADAMTSL 3*
ALM106	rs11642046	16	29993108	G/C	0.46	A	0.00066	0.12	0.02	1.30E-10	0.14	0.02	2.20E-09	*TAOK2*
ALM107	rs1121980	16	53809247	A/G	0.42	B	0.00205	0.19	0.02	2.90E-27	0.25	0.02	1.10E-26	*FTO*
ALM116	rs2070776	17	62007498	A/G	0.36	C	0.00134	-0.11	0.01	1.80E-16	-0.12	0.02	4.30E-13	*CD79B*
ALM120	rs10871777	18	57851763	G/A	0.24	A	0.00184	0.22	0.02	1.80E-24	0.24	0.03	1.40E-18	Intergenic
ALM126	rs6087577	20	32955423	A/G	0.49	C	0.00105	0.09	0.01	2.50E-12	0.11	0.02	2.80E-06	*ITCH*
ALM127	rs2425052	20	33880729	C/A	0.44	C	0.00126	0.09	0.01	1.20E-13	0.13	0.02	1.70E-14	*FAM83C*

QTL id column refers to appendicular muscle mass QTLs catalogued in Table S2. Minor allele frequency (MAF) and effect size (beta) was estimated for the first allele (A1) presented in the “Alleles” column (A1/A2). PVE: proportion of variance explained by QTLs. In addition to 4 principal components, fixed effects in each set included: A) Sex; B) Sex and whole body fat (WBF); C) Sex, WBF and leg length; D) Sex and leg length. P value thresholds for replication tests and fixed effects were (A) 2.04×10^{-05}; (B) 1.93×10^{-05}; (C) 1.96×10^{-05}; (D) 1.93×10^{-05} (see Methods for threshold derivation details).

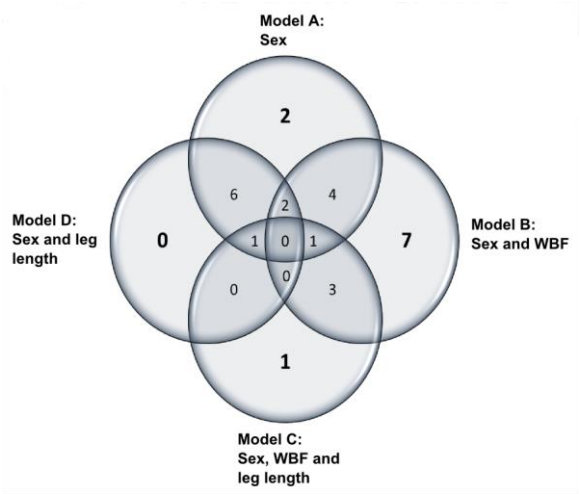

Figure 3. Summary of replicated appendicular lean mass QTLs. The linear mixed model included (A) sex; (B) sex and whole body fat (WBF); (C) sex, WBF and leg length; (D) sex and leg length fixed effects.

Combined cohort association analysis

In order to increase QTL detection power we performed a GWAS in the Combined cohort (n =95,545). This analysis revealed 573 SNPs significantly associated *(P* < 5 × 10^{-8}) with ALM which were clustered in 132 QTLs (Figure 4), nearly tripling the number identified in the Discovery cohort alone. Collectively these 132 QTLs accounted for 6.75 % of the phenotypic variability, equalling to 17% of the SNP heritability. The allelic effect of individual SNPs ranged between 0.04 ± 0.01 kg to 1.13 ± 0.20 kg (Supplementary Material, Table S2). The SNP with the largest effect size, rs148833559, was characterised by MAF of 0.001 and PVE of 0.000376. This association was identified only in the combined analysis under model A, B and C. Similar to the findings in the Discovery cohort alone, the majority of the QTLs in the Combined cohort, 79, were identified by more the one analytical model, however, a

number of QTLs, 5, 32, 12 and 4, appeared specific to only A, B, C or D model, respectively (Supplementary Material, Tables S2). In the next stage we explored genetic content of 132 ALM QTLs.

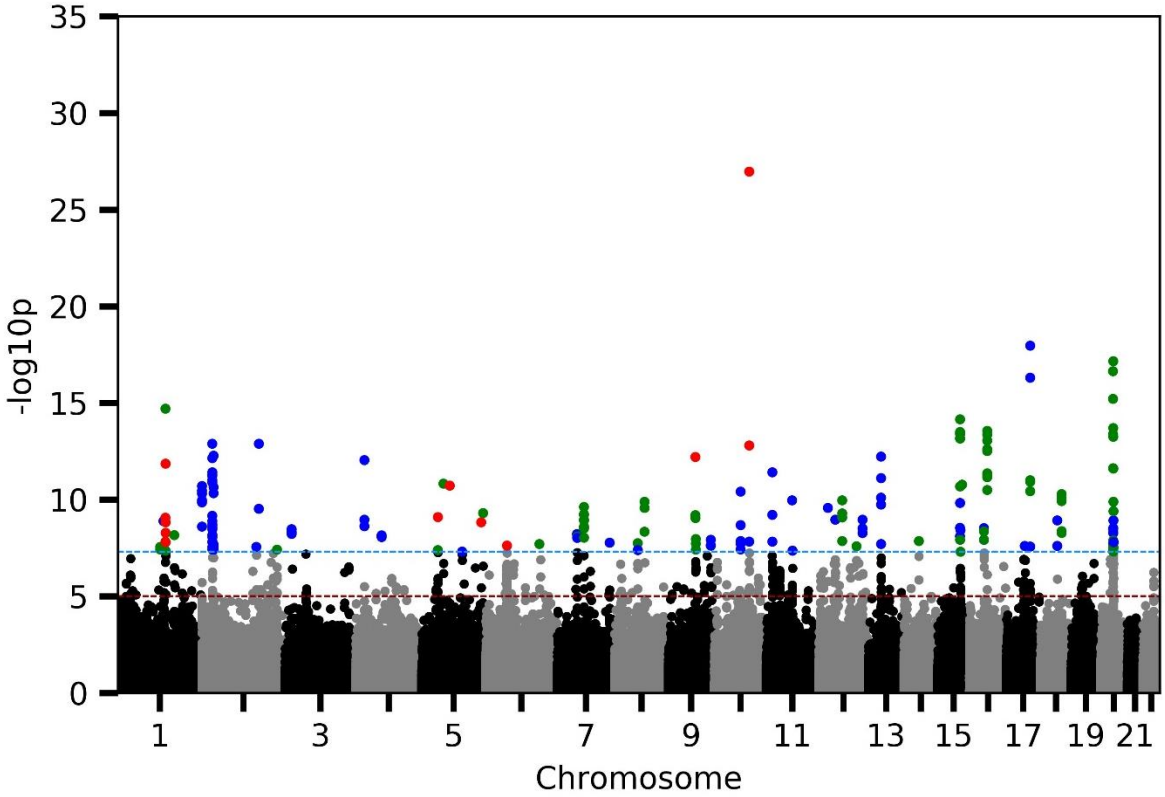

Figure 4. Manhattan plot for genome wide association analysis of appendicular lean mass in the Combined cohort (n =95,545) of 37-48 year old individuals. Horizontal blue and red lines represent significance ($P < 5 \times 10^{-8}$) and suggestive thresholds ($P < 1 \times 10^{-5}$). The significant SNPs are colour coded depending on their beta: blue (effect size < 0.05), green (0.05 < effect size < 0.09) and red (effect size > 0.09). Note that all fixed effects were included in this analysis (Model C), Manhattan plots of model A, B, and D are provided in the Supplementary Materials, Figure S9 - Figure S12.

SNP-tagged regularoty regions and genes

Bioinformatics analysis of 573 significant SNPs revleaed that the majority, 85%, were located in either regulatory regions or coding and immediately adjacent regions (UTRs, introns, upstream and downstream) of the genome.

A set of 78 SNPs mapped to 74 regulatory regions: 17 promoters, 22 promoter flaking regions, 19 open chromatin regions, 6 enhancers, 6 transcriptional repressor CTCF binding sites and 4 transcription factor binding sites (Supplementary Material, Table S3). The majority of these regulatory regions were located in the proximity of one or more genes, however 7 regulatory regions mapped to the intergenic regions.

A total of 430 SNPs identified in the Combined cohort analyses were located within 236 known genes, of which, 138 are expressed (0.4 – 1352.7 reads per kilobase million) in skeletal muscle (19). Twenty three of these genes were identified in the Discovery cohort and confirmed in the Replication cohort (Table 2). Approximately 15% of 430 SNPs might affect the structure and/or function of the protein as they were non-synonymous 63, and frameshift, 1, variants identified in 54 annotated genes (Supplementary Material, Table S3). Expression Atlas analyses indicated that 48 of those genes are expressed in human skeletal muscle tissue at low-to-medium levels (0.5 – 230 transcripts per million).

The PolyPhen2 (20) and SIFT (21) algorithm predictions were used to screen the 54 genes based on the predictive SNPs consequence on the protein structure and

function. Damaging effects for 7 genes (*THBS3, CEP120, STC2, WSCD2, CCDC92, AKAP13, CPNE1*) were predicted by both algorithms, and at least one algorithm predicted damaging consequence for 10 more genes (*PPM1J, ITSN2, C2orf16, VCAN, FGFR4, MLXIPL, PTCH1, BDNF, NCOR2, WDR90*). All 17 genes are expressed in skeletal muscle tissue. Exploration of the effect of a frameshift rs147019139 variant, resulting from one nucleotide insertion between the position 20:35627312 and 20:35627313 in *CPNE1* gene, revealed that premature stop would result in a truncated protein of 426 amino acids; the wildtype protein is 542 amino acids long.

Collectively these analyses revealed a large number of novel candidate genes for appendicular lean mass. They also indicated that many of these genes might be intrinsic to skeletal muscle tissue and that their effects might be mediated by the coding sequence variation or alterations in expression levels.

Discussion

The key feature in the present report is a substantially larger genetic architecture of lean mass than in any previous attempts in human studies. In total we mapped 132 QTLs that collectively explain 17% of SNP heritability for ALM. The study revealed a large number of novel candidate genes implicated in human ALM variability. These findings help understanding of the aetiology of conditions associated with body composition, particularly those stemming from the properties of the most abundant component of lean mass, skeletal muscle tissue.

We argue that two main factors contributed to the increase in the number of detected QTLs compared to previous research. The first one was correction for more than one set of biologically relevant variables. Although SNP heritability estimate, 0.42, recently published by Zillikens and colleagues in a large GWAS on lean mass was very similar to the present report, 0.40, only 3 QTLs were successfully discovered and replicated in that study (8). The analytical approach that we have taken permited a substantial reduction of “missing heritability”. Inclusion of fixed effects in the linear mixed model (LMM) is a common practice used to control for potentially confounding effects of the factors contributing to the phenotypic variability. However, a single set of confounding effects is unlikely to address all complex interactions. Considering the effects on the skeletal muscle mass, sex plays a substantial role primarily due to the influence of testosterone on growth and hypertrophy of muscle fibres (22). Size of the skeleton, bone length in particular, can influence muscle mass through the effect on the axial dimension of muscle morphology. The link between muscle mass and body fat appears more complex; an antagonistic relationship between the two types of tissue with respect to metabolism and energy balance (23) is not supported by the positive correlation between skeletal muscle and whole body fat (Supplementary Material, Table S1). The latter phenomenon is likely to stem from the common developmental roots and might be due to the fact that muscle and white adipose tissue cells, at least in some fat deposits, are derived from the same progenitor cells expressing markers, such as Pax3 and Myf5, that until recently were believed to be muscle tissue specific (18). Hence, phenotypic variability in muscle mass could be due to genetic influences affecting muscle fibres and/or their precursors directly, or due to secondary effects of hormonal status, bone growth or a combination of all above. In order to account for these possible and independent scenarios we

conducted analyses using four different models of fixed effects. As illustrated by the analyses of the most robust, replicated QTLs (Figure 3), none of the models captured all QTLs. A similar pattern reemerged in the analyses of the Combined cohort with substantially higher QTL detection power. Fifty three out of 132 QTLs appered model specific. In each of the four models, there were one or more model-specific QTLs with a candidate gene expressed in skeletal muscle (Supplementary Material, Table S3) suggesting tissue-specific rather than systemic effects. We interpret these observations as showing that all models revealed variants contributing to variability in muscle mass, albeit perhaps in different ways, via different and independent mechanisms. While the interplay between different models remains to be understood, it is clear that a large part of the genetic architecture would have been missed by focussing on only one analytical model.

The second important factor was the restriction of subjects age. The ALM in the extremities of human body primarily consists of the skeletal muscle tissue, the bone and the skin. Skeletal muscle is the most abundant of those tissues. The majority of identified candidate genes are expressed in skeletal muscle tissue and might influence it directly. Furthermore, when we compared the list of ALM candidate genes discovered in this study with the candidate genes affecting muscle weight in laboratory mouse models, where individual muscles were dissected and weighed (5, 6, 24), there was an overlap of 68 genes (Supplementary Material, Table S3). Therefore the genetic artchitecture identified in this study to a large extent is genetic architecture for appendicular muscle mass. Skeletal muscle is a dynamic tissue reaching its peak mass by late 20s, then a trend of decline emerges after 40s and accelerates about two decades later (1). An estimated 30-50% decline in muscle

mass can be expected between 40 and 80 years of age (25). In Zillikens and colleagues study, the age of subjects ranged from 18 to 100 years (8) obstructing detection of genetic architecture even though age was included in their model. In the present study, restriction of age to a narrow range, 37 – 48 years, permitted us to assemble the largest to date, developmentaly stable study cohort, which improved QTL detection power.

Having built-in replication cohorts has been a common GWAS practice aimed at curbing the false positives and highlighting the most robust genetic markers. Mendelian randomisation in the Replication cohort indicated that increasing sarcopenia risk score was associated with a lower ALM providing support for robustness of the Discovery cohort findings. Furthermore, an increasing body of evidence indicates high replicability of GWAS findings (26). Therefore we explored the underlying genetic mechanisms for all QTLs indentified in the Combined cohort. Effects of 4 genes, *VCAN*, *ADAMTSL3*, *CALCR* and *FTO*, captured in these QTLs overlaped with the significant or suggestive findings reported by Zillekens and colleagues (8). Importantly, our study indentified 16 novel genes (*THBS3, CEP120, STC2, WSCD2, CCDC92, AKAP13, CPNE1*, *PPM1J, ITSN2, C2orf16, FGFR4, MLXIPL, PTCH1, BDNF, NCOR2, WDR90*) expressed in skeletal muscle with potentially damaging non-synonymous or frameshift variants. We provide an overview of selected candidates below.

The *FTO* gene has been strongly linked to body mass index and accumulation of adipose tissue (27-29) and considered as a potential target for anti-obesity

medicines (30-32). However, our analyses using four different models suggested that this genomic region not only affects fat tissue as has been established in the past (33), but is also related to muscle mass; the strength of the association between the *FTO* gene and ALM decreased when controlling for body fat in models B and C of our analysis. This observation is consistent with a scenario that allelic variants at the *FTO* locus have the same effect on skeletal muscle and adipose tissue, and has important implications; it suggests that the *FTO* targeting medications might have an adverse side effect on skeletal muscle mass.

The *CPNE1* gene emerged as a novel and interesting candidate gene for muscle mass, harbouring two potentially consequential variants. One of the variants (rs147019139) results in a frameshift due to insertion of one nucleotide that would lead to a truncated protein. Importantly, the mutant allele was associated with larger muscle mass suggesting for this gene a suppressor role in muscle development and/or growth. The gene is expressed in human (19) and mouse (34) skeletal muscle and in proliferating and differentiating myoblasts (35), hence could have intrinsic effects in this tissue. The *CPNE1* gene is involved in calcium mediated process (36), and was found to play a role in neuronal progenitors cell differentiation (37) and affect proliferation of osteosarcoma cells (38). There are no reports linking rs147019139 to any phenotype or condition, however, a variant neighbouring *CPNE1* (rs2425143) was recently associated with blood protein levels (39).

A rare variant located in the first exon of the *STC2* gene, rs148833559, is predicted to have a detrimental effect on the protein. It had the largest effect size (1.13 ± 0.20

kg) but exceeded the threshold of significance only in the Combined cohort analyses. This may be due to the relatively small number of participants with the minor allele (MAF = 0.001). No previous studies have identified an association of rs148833559 with lean mass; however, the *STC2* gene is associated with height, prostate cancer, urate levels and immunoglobulin A vasculitis (40-43). In addition, *STC2* negatively regulates growth, its overexpression (44) results in reduced bone and muscle growth in mice (45).

Both *CPNE1* and *STC2* genes merit the follow up analyses and assessement of their potential as target for treatments of sarcopenia and other conditions accompanied by muscle loss. Diverse experimental strategies and models will be required for studying these genes and the remaining genetic architecture. It will be of particular relevance to determine which of the genes influence properties of individual muscle fibres such as cross-sectional area and type of expressed myosin, quantity of muscle fibres in the muscle, or both.

In conclusion, the present study sheds light on a substantial portion of previously undiscovered genetic architecture of muscle mass. We provide evidence for a large number of novel genes affecting muscle mass variability in middle aged addults. These new findings have the potential to help further understand the aetiology of sarcopenia and offer targets for development of pharmacological interventions.

Material and Methods

Study cohorts

Discovery and Replication sub-sets

The cohort used in the study consisted of 99,065 adult individuals between 37 - 48 years of age (project #26746). This cohort was selected because muscle mass is fully matured yet not affected by aging-related decline. This population was drawn from the UKB project (46), which is a large study of more than 500,000 individuals of the ages between 37 and 48 years at the time of recruitment (Field ID: 21022). All participants recruited were identified from the UK National Health Service (NHS) records and attended a baseline visit assessment between 2006 and 2010. At this visit the participants gave written consent, answered a questionnaire, and gave a face to face interview regarding health and lifestyle. Blood samples and anthropometric measurements were collected from each participant. The assessments were conducted at 22 assessment centres in Scotland, England and Wales. We excluded from the analyses participants (n = 3,520) that were reported to be ill with cancer, pregnant, or having undergone a leg amputation procedure and individuals with discordant sex records between genetic and self reported. This resulted in a total of 95,545 adult individuals (51,394 females and 44,151 males) for the analyses.

We used the sample function implemented in R software (47) to randomly divide the data set into a Discovery (n = 57,545) and a Replication (n = 38,000) cohort with the purpose of replicating the genetic architecture of the trait. This process is fully described in Analyses of the Replication cohort section of the materials and

methods. The overall sex (females ~ 54 %, males ~ 46 %,) and ethnicity (~ 90 % British ancestry) was maintained in both cohorts.

Body composition traits

Records on standing (UKB field ID: 50) and sitting (UKB field ID: 20015) heights, WBF (UKB field ID: 23100), and arms (UKB fields ID: 23121 and 23125) and legs lean mass (UKB fields ID: 23113 and 23117) were used for genome wide association analyses. The standing and sitting heights (distance from rump to crown when sitting) were recorded in cm, and the difference between them was used to calculated leg length of each individual. The body composition measures (see Web Resources for details) were obtained using bioelectrical impedance. It has been shown that bioelectrical impedance and magnetic resonance imaging measurements are comparable methods for estimation of lean mass (48). The ALM was calculated as the sum of lean mass in two arms and legs. In addition, we only used phenotypic records within ± 3 SD from the ALM mean in the analyses. This is a common practice aimed at eliminating samples falling outside of the the normal distribution range. Application of this threshold resulted in the removal of 299 individuals. Normality of residuals was assessed using QQ-plots after fitting a linear model for all traits with sex as an explanatory variable.

Genotypes

DNA extraction and genotyping of the data was conducted by the UKB using the Axiom array from Affymetrix (Santa Clara, California, USA). This array contains approximately 800,000 markers and was specifically designed to include markers

with known associations, coding variants and to provide coverage across the whole genome (see Web Resources for details). The genotypic data used for this project was obtained on June 2017, we used the un-imputated version of the genotypes due to a reported inaccuracy on the imputation version of the genotypes at the time of the analyses in accordance with guidance from UKB. In addition to the quality controls applied by the UKB project, we applied further filters using PLINK 1.9 (49) and only used genotypes that had a MAF > 0.001, marker missing rate < 0.05, and individual missing rate < 0.1 (see Web Resources). This resulted in a total of 524,481 SNPs used in this study.

GWAS in the Discovery cohort

The GWAS analyses of the ALM in the Discovery cohort were carried out by using the LMM approach implemented in BOLT-LMM 2.3 software(50); this approach is capable of handling large populations, while taking into account relatedness (cryptic relatedness and/or population stratification). Nevertheless, we assessed the population structure of the Discovery cohort by performing a principal component analysis (PCA) on the genotypes using the software PLINK 1.9 (49); the PCA was performed on the 524,481 SNP that were kept after quality control of the genotypes (described above). Based on the PCA we included the first four principal components (PC) as fixed effects in the model; the use of PC was also recommended by the BOLT-LMM 2.3 software (50).

We aimed to capture all possible loci that are directly or indirectly (e.g. through the effect on bone growth or fat content) associated with lean mass variability. To

identify these genomic regions we explored the effect of sex, leg length and WBF on the ALM by using four different LMMs. In model A, which was adjusted for sex, we expected to capture loci associated with musculoskeletal growth that might refect systemic effects influencing growth since no other adjustment was included. In model B (sex and WBF) we aimed to find loci associated with lean mass while accounting for variation in WBF. The same embryonic progenitor gives rise to satellite cells (postnatal myogenic progenitor) and some of the white and brown adipose tissue (18), therefore it is likely that some of the individual differences in both tissues, muscle and fat, are a consequence of the same genes. However, some differences are due to muscle but not fat affecting genes, and such genes would be captured in model B. Model C controlling for sex, WBF and leg length aimed to capture muscle-specific genetic associations, ajusted for influences caused by developmental and growth relationships with fat tissue and the skeleton (18). Finally in model D (sex, leg length), we should capture genes associated with muscle mass independently of the size of skeleton. The axial length of the muscle and hence its mass is directly influenced by the size of the skeleton (16). Whilst, using different models results in a substantial increase in the complexity of the analyses, failing to consider all these scenarios would result in underestimation of QTLs influencing lean mass.

Variance components and SNP heritability estimates of the ALM were obtained using the BOLT-REML method described by Loh and colleagues (51) and implemented in BOLT-LMM (50). The SNP heritability was calculated based on the 524,481 SNP used for the association analyses.

In order to identify the QTLs, we visually examined the Manhattan plot resulting from each association analysis and defined a QTL as a series of adjacent SNPs (or a single SNP from a series of adjacent SNPs) exceeding the threshold of significance ($p < 5 \times 10^{-8}$; Supplementary Materials, Figure S3). We characterised each QTL by position of its peak SNP and its effect size (beta ± SE). In the instances where threshold-exceeding SNPs spanned large distances (>1 Mb) an r^2 matrix between all ALM significantly associated SNPs was calculated (by chromosome) using PLINK 1.9 (49) and we examined if one or more LD blocks were present. If more than one LD block was present, we treated the region as two (or more) separate QTLs (Supplementary Materials, Figure S4).

Analyses of the Replication cohort

Replication of genetic architecture

We adopted the Mendelian randomization strategy initially proposed by Katan (52) and Gray and Wheatley (53) to test the effect of the genetic architecture identified from the analyses of the Discovery cohort. The SNP with the lowest p value at each defined QTL was identified and selected for this analysis in Replication cohort. We then estimated a “sarcopenia risk score” of the genetic architecture as follows. First, contribution of the effect of each SNP to an individual’s ALM was estimated as a product of the absolute value of SNP effect size β (in kilograms), obtained from BOLT-LMM, and the number of the ALM decreasing alleles of the marker; e.g. 2, in case of homozygosity for decreasing alleles, 1, in heterozygous, and 0, if homozygous for increasing alleles. Second, the sarcopenia risk score was calculated

for each individual as sum of the products for all selected SNPs. Finally, the resulting distribution was ranked in ascending order and partitioned into five quantiles and the difference in median ALM between the quantiles was tested using a Kruskal-Wallis test. Subsequent pairwise comparisons between quantiles were conducted using a Wilcoxon test. A negative control test was conducted by randomly selecting a subset (n = 66) of non-significant SNPs in the Discovery cohort and generating 'sarcopenia risk score' as described above. All values reported for the replication analyses were means ± SE, unless instated otherwise.

LMM in the Replication cohort

For the replication of the SNPs identified in the discovery cohort, all significant SNPs ($p < 5\ e^{-8}$) were selected and a LMM was fitted in the Replication cohort using the same fixed effects structure. A set of non-significant SNPs ($p > 0.99$) selected across all chromosomes were used as negative control as described above. The SNPs were considered replicated if they passed a Bonferroni adjusted p value threshold.

GWAS of the Combined cohort

In order to increase QTL detection power, we re-analysed the ALM in the Discovery and Replication cohorts together (Combined cohort). The association analyses were performed using the same parameters established for the Discovery analyses. Significant associations were reported at $p < 5 \times 10^{-8}$. The QTLs were defined as previously described for the Discovery cohort.

Phenotypic variance explained by QTLs.

The SNP with the lowest p value within a QTL (top SNP) was used to estimate the PVE by locus. We first estimated phenotypic residuals by fitting a linear model with the same combination of fixed effects used in the corresponding association analysis. These residuals were then used in a linear model which included the top SNP genotypes as independent variables. The resulting coefficients of determination (r^2) was reported as PVE. The PVE was estimated and reported for all QTLs identified in the Discovery and Combined cohorts.

SNPs and tagged genes

To further explore significant SNPs ($p < 5 \times 10^{-8}$) we used a BioMart data mining tool (See Web Resources) to identify SNPs located within gene sequence (introns, exons and UTRs). The expression of genes tagged by significant intronic, exonic or UTR SNPs was investigated in skeletal muscle using the Expression Atlas (See Web Resources). Using the BioMart tool implemented in Ensembl (http://www.ensembl.org/index.html) we examined if non-coding variants were located within known regulatory regions. Non-synonymous SNPs were evaluated for their predicted effect on the protein using Polyphen2 (20) and SIFT algorithms (21).

Web Resources

UK Biobank Body composition protocol, DNA extraction, genotyping and quality control protocol: http://www.ukbiobank.ac.uk/ .PLINK 1.9 software: www.cog-genomics.org/plink/1.9/. BOLT-LMM software: https://data.broadinstitute.org/alkesgroup/BOLT-LMM/. BioMart:

https://www.ensembl.org/info/data/biomart/index.html. Expression Atlas: https://www.ebi.ac.uk/gxa/home/

Data availability

The data used for this study can be obtained upon application to the UK biobank (www.ukbiobank.ac.uk/).

All statistical analyses were conducted in the R statistical environment (47) unless otherwise stated.

Conflict of Interest Statement

The authors declare no competing interests.

Acknowledgments

The authors would like to acknowledge the support of the Maxwell compute cluster funded by the University of Aberdeen, and Elphinstone and IMS studentship for funding AIHC.

References

1 Lauretani, F., Russo, C.R., Bandinelli, S., Bartali, B., Cavazzini, C., Di Iorio, A., Corsi, A.M., Rantanen, T., Guralnik, J.M. and Ferrucci, L. (2003) Age-associated changes in skeletal muscles and their effect on mobility: an operational diagnosis of sarcopenia. *J. Appl. Physiol.*, **95**, 1851-1860.

2 Janssen, I., Shepard, D.S., Katzmarzyk, P.T. and Roubenoff, R. (2004) The healthcare costs of sarcopenia in the United States. *J. Am. Geriatr. Soc.*, **52**, 80-85.

3 Kim, J., Wang, Z., Heymsfield, S.B., Baumgartner, R.N. and Gallagher, D. (2002) Total-body skeletal muscle mass: estimation by a new dual-energy X-ray absorptiometry method. *Am. J. Clin. Nutr.*, **76**, 378-383.

4 Livshits, G., Gao, F., Malkin, I., Needhamsen, M., Xia, Y., Yuan, W., Bell, C.G., Ward, K., Liu, Y. and Wang, J. (2016) Contribution of Heritability and Epigenetic Factors to Skeletal Muscle Mass Variation in United Kingdom Twins. *J. Clin. Endocrinol. Metab.*, **101**, 2450-2459.

5 Nicod, J., Davies, R.W., Cai, N., Hassett, C., Goodstadt, L., Cosgrove, C., Yee, B.K., Lionikaite, V., McIntyre, R.E., Remme, C.A. *et al.* (2016) Genome-wide association of multiple complex traits in outbred mice by ultra-low-coverage sequencing. *Nat. Genet.*, **48**, 912-918.

6 Hernandez Cordero, A.I., Carbonetto, P., Riboni Verri, G., Gregory, J.S., Vandenbergh, D.J., J, P.G., Blizard, D.A. and Lionikas, A. (2018) Replication and discovery of musculoskeletal QTLs in LG/J and SM/J advanced intercross lines. *Physiol. Rep.*, **6**.

7 Carbonetto, P., Cheng, R., Gyekis, J.P., Parker, C.C., Blizard, D.A., Palmer, A.A. and Lionikas, A. (2014) Discovery and refinement of muscle weight QTLs in B6 x D2 advanced intercross mice. *Physiol. Genomics*, **46**, 571-582.

8 Zillikens, M.C., Demissie, S., Hsu, Y.-H., Yerges-Armstrong, L.M., Chou, W.-C., Stolk, L., Livshits, G., Broer, L., Johnson, T., Koller, D.L. *et al.* (2017) Large meta-analysis of genome-wide association studies identifies five loci for lean body mass. *Nat. Commun*, **8**, 80.

9 Urano, T. and Inoue, S. (2015) Recent genetic discoveries in osteoporosis, sarcopenia and obesity. *Endocr. J.*, **62**, 475-484.

10 Hai, R., Pei, Y.-F., Shen, H., Zhang, L., Liu, X.-G., Lin, Y., Ran, S., Pan, F., Tan, L.-J. and Lei, S.-F. (2012) Genome-wide association study of copy number variation identified gremlin1 as a candidate gene for lean body mass. *J. Hum. Genet.*, **57**, 33.

11 Liu, X.G., Tan, L.J., Lei, S.F., Liu, Y.J., Shen, H., Wang, L., Yan, H., Guo, Y.F., Xiong, D.H., Chen, X.D. *et al.* (2009) Genome-wide association and replication studies identified TRHR as an important gene for lean body mass. *Am. J. Hum. Genet.*, **84**, 418-423.

12 Guo, Y.-F., Zhang, L.-S., Liu, Y.-J., Hu, H.-G., Li, J., Tian, Q., Yu, P., Zhang, F., Yang, T.-L. and Guo, Y. (2013) Suggestion of GLYAT gene underlying variation of bone size and body lean mass as revealed by a bivariate genome-wide association study. *Hum. Genet.*, **132**, 189-199.

13 Ran, S., Liu, Y.-J., Zhang, L., Pei, Y., Yang, T.-L., Hai, R., Han, Y.-Y., Lin, Y., Tian, Q. and Deng, H.-W. (2014) Genome-wide association study identified copy number variants important for appendicular lean mass. *PloS ONE.*, **9**, e89776.

14 Urano, T., Shiraki, M., Sasaki, N., Ouchi, Y. and Inoue, S. (2014) Large-scale analysis reveals a functional single-nucleotide polymorphism in the 5′-flanking region of PRDM16 gene associated with lean body mass. *Aging cell*, **13**, 739-743.

15 Legrand, D., Adriaensen, W., Vaes, B., Mathei, C., Wallemacq, P. and Degryse, J. (2013) The relationship between grip strength and muscle mass (MM), inflammatory biomarkers and physical performance in community-dwelling very old persons. *Arch Gerontol Geriatr.*, **57**, 345-351.

16 Hogrel, J.-Y., Decostre, V., Alberti, C., Canal, A., Ollivier, G., Josserand, E., Taouil, I. and Simon, D. (2012) Stature is an essential predictor of muscle strength in children. *BMC Musculoskel Disord*, **13**, 176.

17 Bentzinger, C.F., Wang, Y.X. and Rudnicki, M.A. (2012) Building muscle: molecular regulation of myogenesis. *Cold Spring Harb. Perspect. Biol.*, **4**.

18 Sanchez-Gurmaches, J. and Guertin, D.A. (2014) Adipocytes arise from multiple lineages that are heterogeneously and dynamically distributed. *Nat. Commun.*, **5**, 4099.
19 Carithers, L.J., Ardlie, K., Barcus, M., Branton, P.A., Britton, A., Buia, S.A., Compton, C.C., DeLuca, D.S., Peter-Demchok, J., Gelfand, E.T. *et al.* (2015) A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreserv Biobank*, **13**, 311-319.
20 Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S. and Sunyaev, S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248-249.
21 Ng, P.C. and Henikoff, S. (2003) SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, **31**, 3812-3814.
22 Egner, I.M., Bruusgaard, J.C., Eftestol, E. and Gundersen, K. (2013) A cellular memory mechanism aids overload hypertrophy in muscle long after an episodic exposure to anabolic steroids. *J. Physiol.*, **591**, 6221-6230.
23 Wolfe, R.R. (2006) The underappreciated role of muscle in health and disease. *Am. J. Clin. Nutr.*, **84**, 475-482.
24 Parker, C.C., Gopalakrishnan, S., Carbonetto, P., Gonzales, N.M., Leung, E., Park, Y.J., Aryee, E., Davis, J., Blizard, D.A., Ackert-Bicknell, C.L. *et al.* (2016) Genome-wide association study of behavioral, physiological and gene expression traits in outbred CFW mice. *Nat. Genet.*, **48**, 919-926.
25 Faulkner, J.A., Larkin, L.M., Claflin, D.R. and Brooks, S.V. (2007) Age-related changes in the structure and function of skeletal muscles. *Clin. Exp. Pharmacol. Physiol.*, **34**, 1091-1096.
26 Marigorta, U.M., Rodríguez, J.A., Gibson, G. and Navarro, A. (2018) Replicability and Prediction: Lessons and Challenges from GWAS. *Trends Genet.*, in press.
27 Akiyama, M., Okada, Y., Kanai, M., Takahashi, A., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., Matsuda, K. *et al.* (2017) Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. *Nat. Genet.*, **49**, 1458-1467.
28 Locke, A.E., Kahali, B., Berndt, S.I., Justice, A.E., Pers, T.H., Day, F.R., Powell, C., Vedantam, S., Buchkovich, M.L. and Yang, J. (2015) Genetic studies of body mass index yield new insights for obesity biology. *Nature*, **518**, 197.
29 Willer, C.J. (2009) Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nature Genet.*, **41**, 25-34.
30 Aik, W., Demetriades, M., Hamdan, M.K., Bagg, E.A., Yeoh, K.K., Lejeune, C., Zhang, Z., McDonough, M.A. and Schofield, C.J. (2013) Structural basis for inhibition of the fat mass and obesity associated protein (FTO). *J. Med. Chem.*, **56**, 3680-3688.
31 Huang, Y., Yan, J., Li, Q., Li, J., Gong, S., Zhou, H., Gan, J., Jiang, H., Jia, G.F., Luo, C. *et al.* (2015) Meclofenamic acid selectively inhibits FTO demethylation of m6A over ALKBH5. *Nucleic Acids Res.*, **43**, 373-384.
32 He, W., Zhou, B., Liu, W., Zhang, M., Shen, Z., Han, Z., Jiang, Q., Yang, Q., Song, C., Wang, R. *et al.* (2015) Identification of A Novel Small-Molecule Binding Site of the Fat Mass and Obesity Associated Protein (FTO). *J. Med. Chem.*, **58**, 7341-7348.
33 Frayling, T.M., Timpson, N.J., Weedon, M.N., Zeggini, E., Freathy, R.M., Lindgren, C.M., Perry, J.R., Elliott, K.S., Lango, H., Rayner, N.W. *et al.* (2007) A

common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science*, **316**, 889-894.

34 Lionikas, A., Meharg, C., Derry, J., Ratkevicius, A., Carroll, A., Vandenbergh, D. and Blizard, D.A. (2012) Resolving candidate genes of mouse skeletal muscle QTL via RNA-Seq and expression network analyses. *BMC Genomics*, **13**.

35 Tomczak, K.K., Marinescu, V.D., Ramoni, M.F., Sanoudou, D., Montanaro, F., Han, M., Kunkel, L.M., Kohane, I.S. and Beggs, A.H. (2004) Expression profiling and identification of novel genes involved in myogenic differentiation. *FASEB J.*, **18**, 403-405.

36 Tomsig, J.L., Sohma, H. and Creutz, C.E. (2004) Calcium-dependent regulation of tumour necrosis factor-alpha receptor signalling by copine. *Biochem. J.*, **378**, 1089-1094.

37 Park, N., Yoo, J.C., Lee, Y.S., Choi, H.Y., Hong, S.G., Hwang, E.M. and Park, J.Y. (2014) Copine1 C2 domains have a critical calcium-independent role in the neuronal differentiation of hippocampal progenitor HiB5 cells. *Biochem. Biophys. Res. Commun.*, **454**, 228-233.

38 Jiang, Z., Jiang, J., Zhao, B., Yang, H., Wang, Y., Guo, S., Deng, Y., Lu, D., Ma, T., Wang, H. *et al.* (2018) CPNE1 silencing inhibits the proliferation, invasion and migration of human osteosarcoma cells. *Oncol. Rep.*, **39**, 643-650.

39 Suhre, K., Arnold, M., Bhagwat, A.M., Cotton, R.J., Engelke, R., Raffler, J., Sarwath, H., Thareja, G., Wahl, A., DeLisle, R.K. *et al.* (2017) Connecting genetic risk to disease end points through the human blood plasma proteome. *Nat. Commun.*, **8**, 14357.

40 He, M., Xu, M., Zhang, B., Liang, J., Chen, P., Lee, J.Y., Johnson, T.A., Li, H., Yang, X., Dai, J. *et al.* (2015) Meta-analysis of genome-wide association studies of adult height in East Asians identifies 17 novel loci. *Hum. Mol. Genet.*, **24**, 1791-1800.

41 Huffman, J.E., Albrecht, E., Teumer, A., Mangino, M., Kapur, K., Johnson, T., Kutalik, Z., Pirastu, N., Pistis, G., Lopez, L.M. *et al.* (2015) Modulation of genetic associations with serum urate levels by body-mass-index in humans. *PLoS ONE.*, **10**, e0119752.

42 Lopez-Mejias, R., Carmona, F.D., Castaneda, S., Genre, F., Remuzgo-Martinez, S., Sevilla-Perez, B., Ortego-Centeno, N., Llorca, J., Ubilla, B., Mijares, V. *et al.* (2017) A genome-wide association study suggests the HLA Class II region as the major susceptibility locus for IgA vasculitis. *Sci. Rep.*, **7**, 5088.

43 Eeles, R.A., Olama, A.A., Benlloch, S., Saunders, E.J., Leongamornlert, D.A., Tymrakiewicz, M., Ghoussaini, M., Luccarini, C., Dennis, J., Jugurnauth-Little, S. *et al.* (2013) Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. *Nat. Genet*, **45**, 385-391, 391e381-382.

44 Chang, A.C., Hook, J., Lemckert, F.A., McDonald, M.M., Nguyen, M.A., Hardeman, E.C., Little, D.G., Gunning, P.W. and Reddel, R.R. (2008) The murine stanniocalcin 2 gene is a negative regulator of postnatal growth. *Endocrinology*, **149**, 2403-2410.

45 Gagliardi, A.D., Kuo, E.Y., Raulic, S., Wagner, G.F. and DiMattia, G.E. (2005) Human stanniocalcin-2 exhibits potent growth-suppressive properties in transgenic mice independently of growth hormone and IGFs. *Am. J. Physiol. Endocrinol. Metab.*, **288**, E92-105.

46 Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M. *et al.* (2015) UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.*, **12**, e1001779.

47 R Core Team. (2017), R Foundation for Statistical Computing. Vienna, Austria., in press.

48 Janssen, I., Heymsfield, S.B., Baumgartner, R.N. and Ross, R. (2000) Estimation of skeletal muscle mass by bioelectrical impedance analysis. *J. Appl. Physiol.*, **89**, 465-471.

49 Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M. and Lee, J.J. (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience.*, **4**, 7.

50 Loh, P.-R., Tucker, G., Bulik-Sullivan, B.K., Vilhjalmsson, B.J., Finucane, H.K., Salem, R.M., Chasman, D.I., Ridker, P.M., Neale, B.M., Berger, B. *et al.* (2015) Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.*, **47**, 284-290.

51 Loh, P.-R., Bhatia, G., Gusev, A., Finucane, H.K., Bulik-Sullivan, B.K., Pollack, S.J., de Candia, T.R., Lee, S.H., Wray, N.R. and Kendler, K.S. (2015) Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance components analysis. *Nat. Genet.*, **47**, 1385.

52 Katan, M.B. (2004) Apolipoprotein E isoforms, serum cholesterol, and cancer. *Int. J. Epidemiol.*, **33**, 9-9.

53 Gray, R. and Wheatley, K. (1991) How to avoid bias when comparing bone marrow transplantation with chemotherapy. *Bone Marrow Transplant.*, **7 Suppl 3**, 9-12.