GWAS replication rates are lower than expected due to over-optimistic assessments of effect sizes

**Introduction**

Despite increasing concerns about the rate and predictors of reproducibility in many areas of scientific research1,2, replication is not routinely carried out in mouse and rat genetic studies of complex traits. It is widely accepted in genome-wide association studies (GWAS) of human complex traits that replication in two or more independent cohorts is best practice in order to robustly establish an association between a phenotype and a genetic variant 3. The mouse genetics community has not adopted this practice, even though rodents appear to share the same highly polygenic genetic architecture observed for complex traits in humans.

The extent to which genetic findings replicate is expected to relate to the allele frequency, effect size and the sample size: more common alleles, larger effect sizes and larger samples all increase the chances of replication. Power calculations use these parameters to provide estimates of the chances of replicating a finding; however, because discovery samples will over-estimate effect sizes (“Winner’s curse”), power to replicate will be systematically overestimated. The extent to which empirical findings from replication studies are consistent with expectations is not often investigated, despite indications that replication rates are lower than might be expected. In a study of twenty-two single gene mutations evidence of between center heterogeneity was observed in 38% of genotype-phenotype comparisons 4. The causes of this heterogeneity were not identified, but may include gene by environmental interactions, which have been proposed to account for differences in the behaviour of inbred strains tested in different laboratories 5. In other instances, the lack of replication has been interpreted as arising for purely genetic reasons: an analysis of the lack of overlap in variants identified from mapping complex traits in two cohorts taken from the same population of Drosophila was attributed to epistasis 6.

In this paper we investigated the replication of loci contributing to complex traits in two experiments that obtained outbred mice from the same commercial provider (Charles River) 9 over the same period of time. One experiment was carried out at the University of Oxford (OX), the other at the University of Chicago (UC). By chance, 20 similar or identical behavioral and physiological phenotypes were assayed in both studies. Thus, it was possible to investigate the degree of replication within and between the two studies. Our results thus provide a broad test of the reproducibility of genome-wide genetic associations in model organisms.

**Results**

We identified 20 phenotypes measured in both studies: muscle weight (5), bone mineral and morphology (2), other morphometric traits (3), fasting blood glucose (1), and behavior (9) (Supplemental). The OX cohort sample size was 1,831 ( X male and Y female), whereas the UC cohort consisted of 1,030 male mice. The two studies used very different genotyping strategies. UC employed genotyping-by-sequencing, in which a portion of the genome was sequenced to high coverage (average X.X coverage over Y.Y% of the genome). In contrast, OX used ultra low-coverage sequencing across the entire genome, yielding 0.15x coverage. We obtained a common set of 1,523,299 SNPs for mapping using a common method to convert sequence data to gentoypes (Supp Figure shows quality metrics).

In this paper we report results from two experiments, one that explored the replication between two studies, and one that combined both studies to increase confidence in loci found in a single study and to identify new loci that, due to low power, were invisible to either component study.

*Replication*

To determine the extent to which QTLs identified in one data set could be found in the second, we first mapped each trait in the two data sets to the autosomes, and estimated the false discovery rate by permutation. At an FDR threshold of 5%, we identified 87 QTLs in the OX data set and 26 in the UC data set (Table 1), consistent with the larger number sample size in OX. To control for differences due to the laboratory or to the genotyping method we also split the OX sample, treating one half (OX1; 871 mice) as discovery and one half (OX2; 873 mice) as replication. We identified 25 loci at a 5% FDR in the discovery set OX1 (Table 1). Supplemental Fig X shows quantile-quantile plots demonstrating that the distribution of P-values shows no sign of inflation; in addition, we confirmed that the distribution of permuted P-values (where no truly significant effects occur) are uniformly distributed.

We used three measures of replication: (i) a *p*-value less than 0.05 in a replication sample (with an effect in the same direction as the discovery sample); (ii) a replication Bayes factor10 where values greater than 1 support replication and values less than 1 support non-replication. (iii) the proportion of results that are truly associated, denoted by π1 (where π1 is 1 - total number of truly null features (*m*0) divided by the total number of p-values (*m*)).

At a P-value of less than 0.05, 36 of the OX discovery sample QTLs replicated in the UC set (41%); using the UC results as a discovery sample, 15 QTLs replicated in OX (57%). In the split Oxford sample 13 replicated in OX2 (52%). These figures are lower than expected from power estimates, which predict replication at over 80% of loci (Table 1).

As an alternative way of weighing evidence in favor of replication we calculated, for each QTL, a replication Bayes factor10. Table 1 shows the number of loci with a Bayes factor greater than 1; the numbers agree closely with the 5% threshold in each experiment.

Our third estimate of replication is an estimate of the proportion of results that are truly associated, relying on the fact that with no true effect on the phenotype the distribution of their P-values should be uniform on the interval (0,0.95). Figure 1 is a histogram of the replication P-values from the OX data set. If there were no true associations, the distribution would be flat, as appears to be the case with P-values > 0.5, but appears not to be true of P-values less than 0.5.

Using a Fisher exact test, we examined the deviation of failed-replication P-values (P > 0.05) from a null distribution. We binned the P-values into 11 groups, from a discovery FDR <0.01 up to <0.5. If there were no signal from true positives, then adding in additional P-values would increasingly dilute the signal seen in the group from FDR < 0.05. Figure 2 shows that this is not the case. The significance of the Fisher exact test continues to increase as P-values from increasingly liberal FDR thresholds are added. This indicates that even at much lower FDR levels than typically reported there remains some true associations.

We estimated the overall proportion of true *p* values using Storey and Tibsharani’s method {Storey, 2003 #42249}. Assuming that P-values are well calibrated, then the proportion of true findings can be estimated from the point at which the distribution of P-values shown in figure 1 becomes uniform. Beyond that point, the results are consistent with the null hypothesis of no effect. We estimated π1 at different degrees of FDR for OX, UC and the divided OX sample, and give the results in table 2. Table 2 shows the proportion of true findings. For the 5% FDR between 70 and 80% of findings can be replicated. Table 2 also shows considerable variation in the π1 estimates, due to the relatively small numbers of P-values. Table 2 therefore also shows π1 for P-values from all studies, which shows a more consistent relationship to FDR than the individual study results. These analyses indicate that the proportion of true findings is likely larger than suggested by our first two measures of replication.

The likely explanation why the first two measures provided such poor estimates of reproducibility is due to over-estimates of the effect size at loci in discovery samples (a “winner’s curse”). Table 1 shows that effect sizes are larger in the discovery sample and that this difference is significant (from counting the number where the discovery is larger than the replication and applying a binomial test).

We attempted to assess the extent to which replication failures were due to “winner’s curse“. By dividing the OX cohort in half, our experiment includes three equally sized samples, OX1, OX2 and the UC sample (884 mice) We used one sample for QTL discovery, one for QTL replication and a third to provide an unbiased assessment of effect size, providing altogether six permutations of these sets for discovery, replication and effect size estimates (thus UC->OX1->OX2, UC->OX2->OX2, OX1->UC->OX2 etc). Since all phenotypes were quantile-normalised prior to genetic analysis the effect-size estimates are dimensionless and therefore comparable across phenotypes.

Figure 2a plots the empirical cumulative distribution for effect size estimates against FDR, and shows that estimates from the second (replication) and third studies are approximately similar, confirming that discovery effect sizes are over-estimated but that replication effects are less exaggerated. For true QTLs, larger effects will be estimated more accurately than smaller, as Figure 2b shows for one three-way comparison.

We used the tri-partite design to investigate the extent to which our initial FDR was over-optimistic because of “winner’s curse”. Taking all SNPs whose discovery FDR was less than a given threshold, we identified replicated QTLs (in which SNPs in the replication set had a P-value < 0.05 and a direction of effect consistent with the discovery). We then estimated the extent of "winner's curse" in the discovery sample by dividing the set 3 replicated SNP effect sizes by the corresponding discovery effect sizes. Finally, we removed the “winner’s curse” from the power estimates by dividing the discovery effect size by the over-estimate, and re-calculated an empirical FDR from discovery to replication cohorts.

Table 2 shows empirical FDRs. The discrepancy between theoretical and empirical FDR based on discovery effect size (left panel of the figure) is due to both “winner’s curse” and errors in the theoretical FDR. Discrepancy between theoretical and empirical FDR based on the recalibrated power (right side of the table) is due to errors in theoretical FDR, assuming that we have adequately removed “winner’s curse”.

*Mega-analysis*

We analyzed a combined sample of the OX and UC cohorts. Overall in this mega-analysis we identified 163 loci at an FDR less than 5%. This is almost twice that identified from the OX sample alone and six times the number identified in the UC cohort. In some cases [how many] the increase in sample size led to diminished support for the presence of a discovery QTL from one of the component cohorts. Nevertheless, overall the mega-analysis of the two samples appears to have increased the total number of discoveries, though clearly the FDR should not be taken at face value.

A table of newly discovered loci is given in Supp Table X. Example plots are shown in Figure 5. The mega analysis implicates *Sh3glb1*, *Prkab2*, *Adrb2* and *Fmnl2* genes in skeletal muscle phenotypes. Bax-interacting factor 1 protein encoded by *Sh3glb1* involved in muscle autophagy (Khan et al., 2014) is a strong functional candidate to influence tissue mass. The *Prkab2* gene encodes beta 2 regulatory subunit of AMP-activated protein kinase (AMPK), a sensor of cellular energy status, which is upregulated in resistance training induced muscle hypertrophy (Wojtaszewski et al., 2005). Agonist stimulated beta 2 adrenergic receptor encoded by *Adrb2* induces muscle hypertrophy in rodents (Joassard et al., 2013;Sato et al., 2008). Whereas the formin family protein encoded by *Fmnl2* gene plays a role in development and repair of myofibrils in the cardiac muscle (Rosado et al., 2014).

**Discussion**

We report a surprisingly low rate of replication between two large genetic analyses of complex traits in mice. Using a harmonized set of phenotypes and genotypes from the same population of outbred of mice, obtained from the same commercial breeder, replication rates were on average 47%. We explored a number of explanations for this finding, as discussed below.

The result is not simply an artifact of the way we defined replication. We chose a relatively simple definition of replication: that the P-value of replication locus should be less than 5% and the effect was in the same direction. This strategy is appealing for being easily understood, and is in-line with common definitions of replication. Applying a Bayesian replication factor did not substantially alter the number of loci for which we evidence for replication.

We expected the extent to which QTLs replicated to be explained in part by differences between phenotypes. Variation in the way phenotypes were assessed might mean that equally heritable phenotypes would arise from different loci. In fact we found little evidence that phenotypic differences were to blame. Replication rates in phenotypes assessed by the same group of investigators using the same protocol were not significantly different from replication rates observed in phenotypes measures independently in the two laboratories. When we carried out a within cohort replication the rates were similar to between cohort replication (52% compared to 55% in the comparison between OX male mice and UC). Genetic correlations between phenotypes were high, also making it less likely that measurement differences contribute. While we cannot exclude the contribution of phenotypic differences, we believe that contribution to be small.

One important reason for the low replication rates is the operation of “winner’s curse”, that is the over-estimated effect size obtained from the discovery sample. This results in over-optimistic estimates of power and thus in falsely high expectations of replication success. When we applied a corrected effect size estimate, obtained from a joint analysis, then the expected number of replicated loci fell closer to the number observed, but still did not fully account for the low replication rates.

Failure to replicate does not mean the loci are not true findings. We found some evidence that the non-replicated loci contain true signals. The distribution of replication P-values that were more than 0.05 was not uniform, and including replication P-values from discovery FDRs up to 0.5, the deviation from uniformity continued to be significant (indeed became more significant), indicating the presence of additional true signal. We drew the same conclusion from the relatively poor correlation between Bayes factors and the discovery P-values. These observations indicate that it would be incorrect to simply discard “non-replicated” loci.

But what of the need to obtain robustly identified loci, whose association is not in doubt? Our findings make the following relevant points. We noted that a 100% replication rate for loci with a discovery logP greater than 8.5. Simply setting a high enough discovery threshold will ensure a robust finding. Because of the difficulties in accurately assessing effect sizes, establishing this threshold is likely to need the sort of empirical justification we have provided in this paper. Similar arguments led physicists to insist on a five sigma threshold for declaring a finding significant.

Second, our findings demonstrate the value of large discovery sample size. Although going from the OX sample to replicate in the UC sample incurred the penalty of lower replication power, the total number of loci discovered is more than twice that found proceeding the other way around (36 against 15). Starting with the larger sample size (OX) and replicating in a smaller sample also provides a better discovery estimate of the effect size. Even more strikingly, using the combined sample of both UC and OX mice more than doubled the total number of QTLs identified at an FDR of 5%. It identified 17 loci with logP greater than 8 (compare to ten in the OX alone sample).

Finally, this increase in sample size led to the identification of a number of novel loci, and candidate genes.. For example, the mega analysis implicates Sh3glb1, Prkab2, Adrb2 and Fmnl2 genes in skeletal muscle phenotypes.

Our examination of the factors that contribute to replication highlights the need to be cautious over claims from a single study. But it would be wrong to insist that replication alone is necessary to establish the robustness of an initial discovery. There appear to be a number of factors, not just study design and the application of appropriate thresholds for the discovery phase, that contribute to whether a finding replicates or not. Some of these lie in technical and artefactual details, but others reflect biological differences, which contribute to phenotypic differences. While different analytical approaches will alter replication rates, we do not believe they alone will produce acceptable levels of replication. Rather, more nuanced analyses are required that consider the context of the experiment and also the nature of genes and other features at the QTLs. We suspect there may be unexpected biological insights to be found from an unexpected lack of replication.

Methods

*Subjects*

Every two weeks, 48 male CFW mice were transported by bus from the Charles River Labs colony in Portage, Michigan to UC for phenotyping. UC requested that CRL send only one mouse from each litter to avoid siblings (close relatives reduce power to map QTLs, and complicate analysis). The average age of the mice upon arrival was 35 days (ranging from 34 to 46 days), and their average weight was 25.5 g (ranging from 13.4 g to 38.7 g). Mice were housed 4 per cage and given ~15 days to adapt to their new environment prior to any behavioral testing. Standard lab chow and water were available *ad libitum*, except during the fasting glucose tests and behavioral procedures. Mice were maintained on a standard 12:12h light-dark cycle (lights on at 06:30). All phenotyping occurred during the light phase between 08:00 and 16:00 hours, over the period of August 2011 to December 2012. All procedures were approved by the UC Institutional Animal Care and Use Committee (IACUC) in accordance with National Institute of Health guidelines for the care and use of laboratory animals.

*Phenotypes*

Blood Glucose

One day after the mice arrived at UC (Mean Age = 36.6 days, SD = 2.5 days), glucose levels were measured using a glucometer after a 4-hour fasting period. In the morning, between 08:00 and 09:00, mice were ear tagged, then transferred into new cages with clean bedding and their food was removed, but water was available *ad libitum*. After that, mice were transported to a testing room, and they were left there for 4 hours. After 4 hours mice were weighed, and a small piece of the tip of the tail was cut using a razor blade and saved for subsequent DNA extraction. This produced a small amount of blood that could be analyzed with glucose strips (Bayer Contour TS Blood Glucose Test Strips) and a glucometer (Bayer Contour TS Blood Glucose Monitoring System). Units are millimoles/liter, or mmol/L, which is the SI unit for measuring glucose in blood. Immediately after all mice within a single cage were tested, food was returned to the cage, and the cage was returned to the colony room where they remained for approximately two weeks before behavioral testing.

Fasting blood glucose levels were partially explained by body weight (proportion of variance explained = 5.6%), so for QTL mapping body weight was included in all regression models of fasting glucose levels. Batches 1 and 11 showed a considerable departure in fasting glucose levels from other batches, so we included 1 and 11 batch indicators in the linear regression models to control for effects of these batches.

Locomotor Activity

Mice were ~51 days of age at the time of test (SD = 4). Mice were transported from the adjacent vivarium, and then allowed to habituate to the procedure room for 30 minutes in their home cages. After 30 minutes of acclimation, mice were removed from the home cage and placed into individual holding cages for 5 minutes. They were weighed individually, injected i.p. with physiological saline (0.01 ml/g body weight), and immediately placed in the center of the OF chamber (AccuScan Instruments, Columbus, OH). Each chamber consisted of a clear acrylic arena (40 x 40 x 30 cm) placed inside a frame containing evenly spaced infrared photobeams from the front to the back and from the left to the right of the arena. Beam breaks were recorded on a computer and converted into horizontal activity. Each activity chamber was encased within a sound attenuating PVC/lexan environmental chamber. Overhead lighting in each chamber provided dim illumination (~80 lux), and a fan provided both ventilation and masking of background noise. Activity chambers were cleaned with 10% isopropanol between tests. A total of eight phenotypes were collected: total activity recorded between 0 and 5 minutes, total activity recorded between 25 and 30 minutes, total activity recorded between the last 10 minutes, decay in activity over time, total activity for the entire 30 minutes, total activity between 5 and 25 minutes, total activity for the first 15 minutes, and total activity for the last 15 minutes.

We checked recorded measurements from all 12 chambers used in these tests to see whether the phenotypes measured using any given chamber differ noticeably from the others. Only one chamber, number 7, had a noticeable effect on the phenotypes, so we included a binary indicator for this cage in all regression models of the locomotor phenotypes.

Conditioned Fear

Mice were tested for conditioned fear (CF) at ~ 63 days of age (SD = 2.9). The CF paradigm was performed over three days: on the first test day, mice were conditioned to associate a test chamber and a tone with a shock; on the second test day, the mice were re-exposed to the same context, but no tones or shocks were given; on the third day, mice were exposed to the conditioned stimulus (the tones), but in a different environment. Immobility, or “freezing” behavior, was interpreted as a measure of learned fear.

We tested mice in 4 chambers obtained from Med Associates (St. Albans, VT, USA). These chambers had inside dimensions of 29 cm x 19 cm x 25 cm. Each chamber had a stainless steel floor grid, metal sides, clear plastic ends and a ceiling, all housed within a sound-attenuating enclosure. A fluorescent light on the top of the chamber provided dim illumination (~3 lux), and a fan provided a low level of masking background noise. Chambers were cleaned with 10% isopropanol between animals. We recorded freezing behavior by analyzing digital video with Freeze Frame software (Actimetrics, Evanston, IL, USA). The freezing data was exported in 30-second blocks, and then averaged into summary measures (detailed below).

The conditioned fear tests consisted of three 7-minute trials over three consecutive days. On all days, the mice were allowed to habituate to a sound-proof room for 30 minutes in their home cages prior to testing. The mice were then transferred to chambers in the testing room from within their individual holding cages. Thirty seconds after being placed in the test chambers, we recorded a baseline measurement of freezing (“freezing before tone, training Day (%)”) ending at the 180-second mark. After this pre-training period, mice were exposed four times to the conditioned stimulus (CS), an 85 dB, 3 kHz tone which lasted 30 seconds, and co-terminated with the unconditioned stimulus (US), a 2-second, 0.5-mA foot shock delivered through the stainless steel floor grid. After each CS-US pairing, there was a 30-second period in which no stimulus was delivered to the subject. We also measured average percent time spent freezing to the first tone prior to shock (“freezing to tone during training (%)”). Finally, in order to attenuate the effect of baseline variation in freezing that was not relevant to unconditioned freezing, we corrected for baseline freezing on the average percent time freezing to the first tone prior to shock (“corrected unconditioned freezing”).

Test day 2 began exactly 24 hours after the start of testing on day 1. The testing environment was identical to day 1, except that neither tones nor shocks were presented to the mice. We measured freezing in response to the test chamber during the same period of time as pre-training freezing (30–180 seconds, “freezing to context, test day (%)”). We chose this time period for two reasons: one, to allow for direct comparisons to the pre-training freezing scores on day 1; two, to avoid measuring freezing behavior during the latter part of the trial in which the mice may have anticipated shocks based on the previous days’ test. In order to attenuate the effect of baseline variation in freezing that was not relevant to learned fear behavior, we also obtained a measurement of baseline context freezing that was corrected for baseline freezing during training on day 1 (“corrected freezing to same context”).

Test day 3 began exactly 24 hours after the start of test day 2. On day 3, we altered the context in several ways: (1) a different experimenter conducted the testing, and she wore a different style of gloves; (2) the transfer cages had no bedding; (3) the metal shock grid, chamber door and one wall were covered with a different material, a hard white plastic; (4) we changed the lighting by placing yellow film over the chamber lights; (5) we cleaned the chamber and plastic surfaces with 0.1% acetic acid solution; and (6) the vent fan was partially obstructed to alter the background noise. On day 3, the tones were presented at the same times as on day 1, but in this occasion they were not paired with shocks. We recorded freezing over the same 30-to-180-second period as on days 1 and 2 by measuring percent time freezing at baseline during the cue test before the first tone (“fc.cue.baseline”). In addition, we measured average percent freezing during presentation of the four 30-second tones (at 180–210, 240–270, 300–330 and 360–390 seconds; “freezing to tone cue (%)”), as well as the average percent freezing during presentation of the four 30-second tones corrected for baseline freezing during training on day 1 (“corrected freezing to cue”).

In summary, UC phenotypes from the 3-day fear conditioning tests consisted of eight measurements of immobility: average proportion of freezing on day 1 during the pre-training interval (30-180 seconds) before exposure to tones and shocks ("freezing before tone, training day (%)"); average proportion of freezing on the first day during exposure to the first conditioned stimulus (“freezing to tone during training (%)”), average proportion of freezing on the first day during exposure to the first conditioned stimulus corrected for baseline freezing (“corrected unconditioned freezing”); average proportion of time freezing in the 30–180 second interval on the second day in conditions identical to the first day (“freezing to context, test day (%)”); average proportion of time freezing in the 30–180 second interval on the second day in conditions identical to the first day corrected for baseline freezing during training on day 1 (“corrected freezing to same context”); average proportion of time freezing over the 30-180 interval on the third day in an altered setting (“fc.cue.baseline”); average proportion of time freezing on the third day in the altered setting during the 30-second intervals in which the tones are presented (“freezing to Tone Cue (%)”); and the proportion of freezing during presentation of the four 30-second tones corrected for baseline freezing during training on day 1 (“corrected freezing to cue”).

All the phenotypes were proportions (numbers between 0 and 1), so to allow for a normal model for these phenotypes we transformed the proportions to the log-odds scale using the (base 10) logit function [Aitchison-1982]. To avoid extremely small or extremely large values after the transformation, the proportions were projected onto range [0.01, 0.99].

For all the conditioned fear traits, UC data showed that the chamber used for testing had an effect on the phenotype, so we included binary indicators for chamber as covariates for all conditioned fear phenotypes. Further, the conditioned fear phenotype measurements in batch 17 had a noticeably different distribution than the other batches, so we included an indicator for batch 17 as a covariate in regression models of all our fear conditioning phenotypes.

Prepulse Inhibition

We tested mice for PPI approximately nine days after the final day of CF testing. Mice were tested at a mean age of 74.5 days (SD = 2.2). During PPI, the mice were exposed to loud pulses (120 dB) that caused them to exhibit the startle response. Occasionally, the exposure to the loud pulse was preceded by a barely perceptible “prepulse” (3–12 dB over background levels), which inhibited the startle response to varying extents. Our PPI testing procedures follow protocols detailed in previous papers [Palmer-2000, Palmer-2003, Palmer-2004, Samocha-2010, Shanahan-2009].

Immediately before testing, mice were transferred from the vivarium to the testing room, one cage (4 mice per cage) at a time. The mice were weighed, and then placed into one of the 5 possible cylindrical Plexiglas containers, each 5 cm in diameter. These containers rested on platforms within a lighted and ventilated chamber (San Diego Instruments, San Diego, CA, USA). We captured mouse movement using a piezoelectric accelerometer, then converted the signal to digital data and recorded on a computer. Before the start of each test day, the apparatuses were calibrated according to the manufacturer's instructions.

Once in the test chamber, mice were presented with 5 minutes of 70-dB white noise. This noise persisted throughout the remainder of the test. The test consisted of the presentation of 62 trials that were a mixture of the following five types: (1) a “pulse-alone” trial, consisting of a 40-millisecond, 120-dB burst; a “no stimulus” trial, in which no stimulus was presented; and three prepulse trials each containing a 20-ms prepulse at 3, 6 or 12 dB above the 70-dB background noise level, followed 100 ms later (onset-to-onset) by a 40-ms, 120-dB pulse. These trials were split into 4 consecutive blocks. The first and fourth blocks consisted of 6 pulse-alone trials. Blocks 2 and 3 consisted of a mixture of 25 trials—6 pulse-alone trials, 4 “no stimulus” trials, and 5 x 3 = 15 prepulse trials—arranged in a pseudorandom order. The startle response during each trial was recorded beginning at 65 ms after the start of the 120-dB stimulus, and at the start of all “no stimulus” trials. Throughout, trials were separated by intervals of 9 to 20 seconds, with an average of 15 seconds.

After testing, mice were returned to their home cage, the cylinders were cleaned with soapy water, and the mice were returned to the vivarium. At that point, the next cage of animals was brought into the testing room, and the process was repeated.

Four phenotypes were collected: habituation, startle, pp6.ppi, and pp12.ppi. Habituation was defined as the average startle amplitude during the fourth pulse-alone trials subtracted from the average startle amplitude during the first pulse-alone trials. The second measure of habituation was defined as the average startle amplitude during the fourth pulse-alone trials subtracted from the average startle amplitude during the first pulse-alone trials, then dividing by the average startle amplitude during the first pulse-alone trials and multiplying by 100. Startle was defined as the average startle amplitude to the 120 db pulses during blocks 2 & 3. pp6.ppi was defined as the average pre-pulse inhibition during blocks 1 and 2 to the 6 dB prepulse. pp12.ppi was defined as the average pre-pulse inhibition during blocks 1 and 2 to the 12 dB prepulse.

The way we have defined it, PPI is always a number between 0 and 1, except in the rare case where the mouse startles more during the prepulse trials, in which case we obtained a negative PPI. Therefore, it is appropriate to transform the PPI measurements to the (base 10) log-odds scale using the logit function. To avoid extremely small or extremely large values after the transformation, small (or negative) PPI values less than 0.01 were set to 0.01, and any values greater than 0.99 are fixed at 0.99.

We found that a subset of the mice were not responding in a measurable way to the pulses, suggesting the possibility that they were deaf . These “deaf” mice add a disproportionate amount of variance to the PPI phenotypes, so we removed these samples to improve the quality of the PPI data.

All chambers used in the tests appeared to have some effect on the PPI phenotypes, with the third chamber having a particularly large effect on some phenotypes. We included all PPI testing chamber indicators as covariates in analysis of the PPI phenotypes. Body weight was normally expected to be correlated with startle amplitude, but this correlation disappeared once we normalized by the average response during the pulse-alone trials.

Musculoskeletal Traits

Mice were sacrificed at ~ 90 days of age (M = 91.2, SD = 2.6). Following sacrifice, one leg was cut off just below pelvis, tubed and transferred into a -70 C freezer, then shipped on dry ice to Dr. Arimantas Lionikas at the University of Aberdeen. On the day of dissection, the leg was defrosted and two dorsiflexors, tibialis anterior (TA) and extensor digitorum longus (EDL), and three plantar flexors, gastrocnemius (“gastroc”), plantaris and soleus, were removed under a dissection microscope and weighed to a precision of 0.1 mg on a balance (Pioneer, Ohaus). Then, the soft tissues were stripped off from the tibia, and the length of the tibia was measured to a precision of 0.01 mm with a digital caliper (“Tibia”; Z22855, OWIM GmbH & Co).

Since elongation of bones is associated with longer (and larger) muscles, accounting for tibia length when analyzing muscle mass variation helped to isolate tissue-specific QTLs—that is, genetic factors that directly regulated development of the muscle tissues—from the QTLs that affect growth of multiple tissues.

UC conditioned on body weight when assessing support for tibia length QTLs; body weight explained 16% of variance in tibia length. Since body weight is understood to be a highly complex trait regulated by a complex combination of genetic and environmental factors, we used body weight to effectively isolate the “less complex” variation in tibia length, thereby putting us in a greater position to uncover genetic factors that contributed to this trait. In principle, this rationale also applies to the muscle weights; body weight explained an additional 10–17% of variance in the muscle weights over the linear effect of tibia length. However, UC did not condition on body weight when assessing support for muscle weight QTLs. The reason is that a much higher proportion of body weight is due to muscle mass than bone mass, so conditioning on body weight would likely diminish our power to detect muscle weight QTLs, particularly genetic factors that were not muscle-specific.

For all the musculoskeletal traits, including bone mineralization, UC included a binary indicator for batch 16 as a covariate because the mice from this batch showed a substantial deviation in these traits from the rest of the mice.

Bone Mineralization

Bone mineralization (BM) in UC mice was measured by Dr. Cheryl Ackert-Bicknell at Jackson Laboratories. From each mouse, the hind axial skeleton (left limb, pelvic girdle and lumbar spine) was collected and fixed overnight in 10% neutral buffered formalin (NBF). The NBF was then removed, and hind axial skeletons were placed in 95% ethanol for a minimum of 2 weeks. After that, the femurs were isolated from the surrounding musculature. Areal bone mineral density for the entire isolated femur was assessed by Dual X-ray absorptiometry (DXA) using a GE-Lunar PIXImus II Densitometer (GE-Lunar, Madison, WI).

To admit a normal distribution, we transformed the BM measurements, which were ratios, to the (base 10) log-scale. A concern with areal BM measurements was that they could be impacted by length and geometry of the bone. However, we found that neither tibia length, gastroc muscle weight nor body weight were correlated with BM.

A substantial fraction of the CFW mice (~15%) exhibited abnormally high BM, or excessive bone mineralization. To map loci for excessive mineralization, we created a binary trait (0 or 1 values) that signals “abnormal” or osteopetrotic bones. It was defined as 1 when BM fell on the “long tail” of the observed distribution, which we defined as any value of real BM greater than 90.

As in the musculoskeletal traits, we included a binary indicator for batch 16 as a covariate because the mice in this batch showed substantial deviation in these traits from the rest of the mice.

Body Weight and Tail Length

Body weight was measured at ~91.2 days of age (SD = 2.6). Body weight measurements at sacrifice showed a considerable departure in batch 17, so we included a binary indicator for this batch as a covariate for this phenotype. We also obtained a measure of body mass index calculated with tibia length (“pseudo BMI”).

Tail length (in cm) was also measured at this time. UC measured tail length as the base of the tail to the tip of the tail.

*Population structure across both studies*

Principal components (PCs) generated from principal component analysis (PCA) on a weighted genetic relatedness matrix (GRM) generated using imputed dosages at a the dense set of 1,523,299 SNPs using LDAK (version 5.9, Methods) of the joint set clearly shows that first principal component (PC1) separates the two data sets (Figure 1a). Consistent with differences in imputation accuracy due to the two genotyping strategies, PC1 loaded onto not just a specific part of the genome, but across the genome at all sites the discrepancy between imputation accuracy in the two studies was spread across (Figure 1b). The other PCs shows no evidence of structure in the joint set of genotypes, consistent with the lack of structure detected by PCA on individual studies and the prior knowledge of mice from both studies being collected from the same colony at the same time (Figure 1c).

*Replication power*

Let R be the study-wide proportion of QTLs that are real (R = 1-FDR). Then, for each of the QTLs from the discovery sample, we have independent but non-identically distributed Bernoulli trials, where the probability of a replication at p=0.05 is R \* P\_i + (1-R)\*T2, where P\_i is the power for QTL i, and T2=0.05 is the type 2 error rate. Let q\_i= R\*P\_i \* (1-R)\*T2

We calculate a likelihood as

L = prod\_{i=1}^N (q\_i)^ x\_i (1-q\_i)^{1-x\_i}

where x\_i are 0 or 1 depending on whether a replication success was observed

We can take the derivative of the log likelihood with respect to R to solve for the MLE estimate of R, and we can calculate a confidence interval using twice the difference between the MLE and the log likelihood for R, which is distributed as a chi squared distribution with 1 degree of freedom.

*Genetic correlation*

We constructed local linkage disequilibrium (LD) weighted genetic relatedness matrices (GRMs) using imputed dosages at the common set of 1,523,299 SNPs for Oxford and Chicago studies separately, and a combined GRM for all mice in both studies, using LDAK (version 5.9) {doi: [10.1016/j.ajhg.2012.10.010](http://dx.doi.org/10.1016/j.ajhg.2012.10.010)}. We then inferred “narrow-sense” heritability from genome-wide SNPs (h2) at 21 phenotypes measured in both studies, and genetic correlations (rG) between the two studies at each phenotype. Estimates for h2 at each phenotype were obtained using the individual study GRMs using restricted maximum likelihood (REML) implemented in LDAK, while rG for each phenotype was estimated with the combined GRM using bivariate REML implemented in GCTA (version 1.24.7) {PubMed ID: 22843982}. All h2 and rG calculations were made with PCs 1 to 20 from PCA on the respective GRMs included as fixed effect covariates.

REFERENCES

1. Button, K.S. *et al.* Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci* **14**, 365-76 (2013).

2. Open Science, C. PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).

3. Studies, N.-N.W.G.o.R.i.A. *et al.* Replicating genotype-phenotype associations. *Nature* **447**, 655-60 (2007).

4. de Angelis, M.H. *et al.* Analysis of mammalian gene function through broad-based phenotypic screens across a consortium of mouse clinics. *Nat Genet* **47**, 969-78 (2015).

5. Crabbe, J.C., Wahlsten, D. & Dudek, B.C. Genetics of mouse behavior: interactions with laboratory environment. *Science* **284**, 1670-2 (1999).

6. Huang, W. *et al.* Epistasis dominates the genetic architecture of Drosophila quantitative traits. *Proc Natl Acad Sci U S A* **109**, 15553-9 (2012).

7. Churchill, G.A. & Doerge, R.W. Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963-71 (1994).

8. Hochberg, Y. & Benjamini, Y. More powerful procedures for multiple significance testing. *Stat Med* **9**, 811-8 (1990).

9. Yalcin, B. *et al.* Commercially available outbred mice for genome-wide association studies. *PLoS Genet* **6**(2010).

10. Verhagen, J. & Wagenmakers, E.J. Bayesian tests to quantify the result of a replication attempt. *J Exp Psychol Gen* **143**, 1457-75 (2014).

**Tables**

Table 1. Summary of QTLs mapped in the discovery and replication experiments

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | OX to OX | OX to UC | OX Males to UC | UC to OX |
| Discovery sample size | 873 | 1763 | 884 | 989 |
| Replication sample size | 871 | 981 | 978 | 1759 |
| Discovery No. of QTLs at 5% FDR | 25 | 87 | 20 | 26 |
| Replication No. of QTLs at 5% FDR | 13(X%) | 36(X%) | 11(X%) | 15(X%) |
| Replication No (Bayes factor > 1) | 13(X%) | 46(X%) | 10(X%) | 15(X%) |
| Discovery effect size | 0.157 | 0.109 | 0.161 | 0.149 |
| Replication effect size | 0.124 | 0.108 | 0.093 | 0.123 |
| P-value comparison of effect size difference | 1.550E-06 | 2.479E-03 | 4.025E-04 | 1.049E-05 |
| Mean power | 97.44% | 82.86% | 96.64% | 99.32% |
| Mean power (mega) | 68.41% | 68.11% | 71.49% | 80.71% |

Cohorts are: OX , Oxford sample (1,831 mice, both sexes), CU, Chicago sample (1,030 male mice) . OX to OX refers to the within Oxford analyses where the cohort was divided into two halves. Sample sizes are shown as means, so are less than the total sample from OX and CU. FDR is the false discovery rate. Replication was scored in two ways: (i) by counting loci that were identified in the replication cohort at P < 0.05 (and the effect in the same direction), (ii) by having a Bayes factor greater than 1. Effect sizes are expressed as correlation coefficients. The P-value reported for the effect size difference is from a binomial test. Power was calculated from the discovery effect size and mean power (mega) refers to power calculated using an effect size estimate from the combined analysis of OX and CU.

Table 2

Estimate of proportion of associations that are true

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| FDR | OX -> UC | UC -> OX | OX1 -> OX 2 | All |
| 0.05 | 0.87 | 0.69 | 0.70 | 0.81 |
| 0.10 | 0.86 | 0.73 | 0.49 | 0.77 |
| 0.15 | 0.62 | 0.32 | 0.37 | 0.53 |
| 0.20 | 0.45 | 0.38 | 0.41 | 0.43 |
| 0.25 | 0.49 | 0.49 | 0.24 | 0.45 |
| 0.30 | 0.48 | 0.51 | 0.03 | 0.42 |
| 0.35 | 0.51 | 0.43 | 0.04 | 0.42 |
| 0.40 | 0.48 | 0.47 | 0.12 | 0.43 |
| 0.45 | 0.39 | 0.46 | 0.21 | 0.38 |
| 0.50 | 0.39 | 0.43 | 0.20 | 0.37 |

The table reports the values of 1 - π0 where π0 is the ratio of the number of truly null findings to the total number of findings. π0 is estimated using the method proposed by Storey and Tibsharani

Table 3 Recalibrated FDR based on removing Winner’s curse

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Using discovery power** | | | **Using re-calibrated power** | | |
| **FDR** | Empirical | Lower 95CI | Upper 95CI | Empirical | Lower 95CI | Upper  95CI |
| 0.01 | 0.428 | 0.722 | 0.168 | 0.331 | 0.073 | 0.669 |
| 0.05 | 0.435 | 0.655 | 0.231 | 0.317 | 0.027 | 0.502 |
| 0.10 | 0.544 | 0.718 | 0.366 | 0.451 | 0.249 | 0.658 |
| 0.15 | 0.602 | 0.762 | 0.431 | 0.499 | 0.298 | 0.696 |
| 0.20 | 0.676 | 0.814 | 0.519 | 0.572 | 0.377 | 0.750 |
| 0.25 | 0.704 | 0.822 | 0.570 | 0.571 | 0.390 | 0.736 |
| 0.30 | 0.702 | 0.811 | 0.578 | 0.563 | 0.394 | 0.719 |
| 0.35 | 0.768 | 0.855 | 0.668 | 0.626 | 0.480 | 0.758 |
| 0.40 | 0.795 | 0.869 | 0.710 | 0.655 | 0.526 | 0.772 |
| 0.45 | 0.853 | 0.911 | 0.784 | 0.731 | 0.619 | 0.829 |
| 0.50 | 0.851 | 0.904 | 0.790 | 0.737 | 0.642 | 0.822 |

The table shows the empirical false discovery rate (FDR) estimates based first on the initial discovery QTL effect size estimates (left panel) and then after re-calculating power taking Winner’s curse into account (right panel). The first column gives the theoretical FDR

**Figures**

Figure1: Distribution of replication P-values compared to uniform P-values

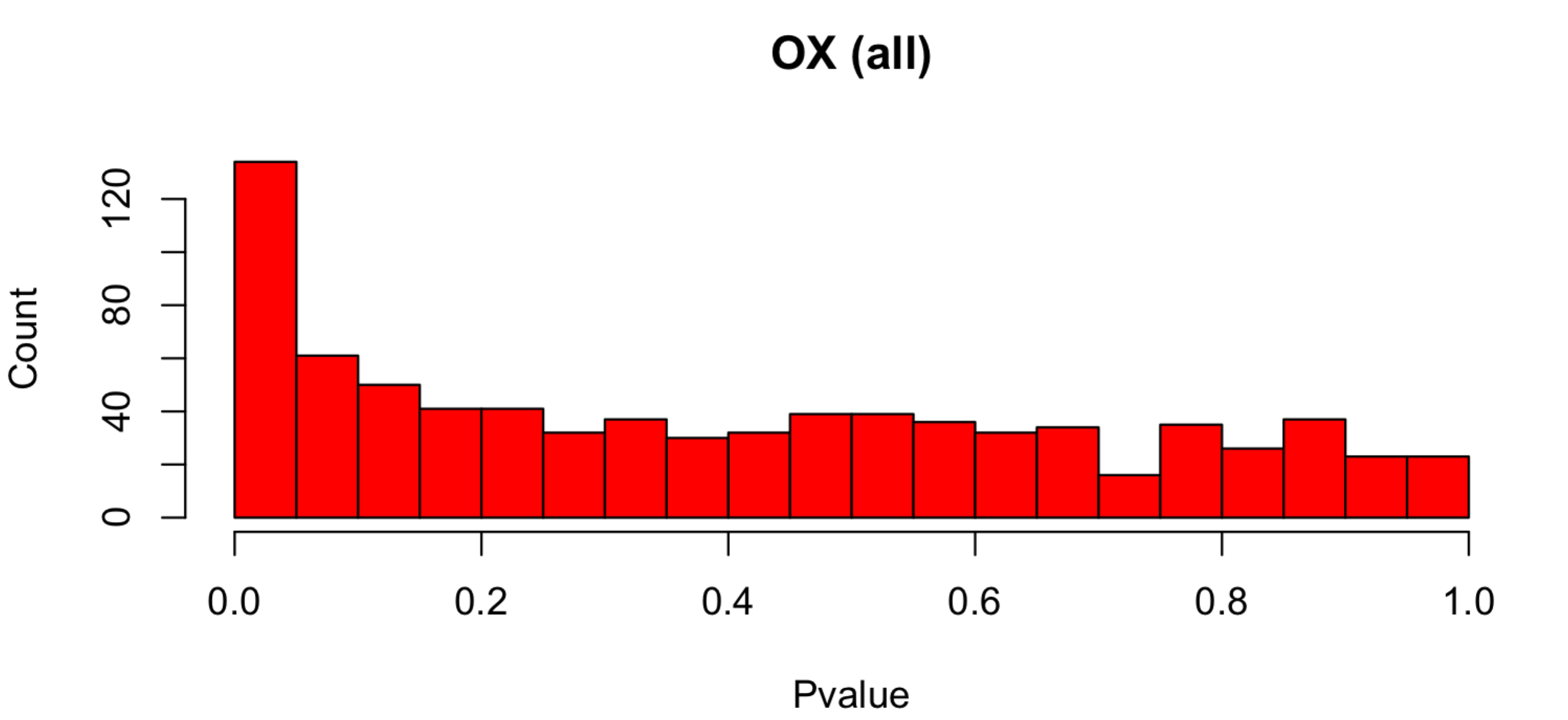
****

Figure 2 Effect size as a function of LogP and empirical cumulative distribution

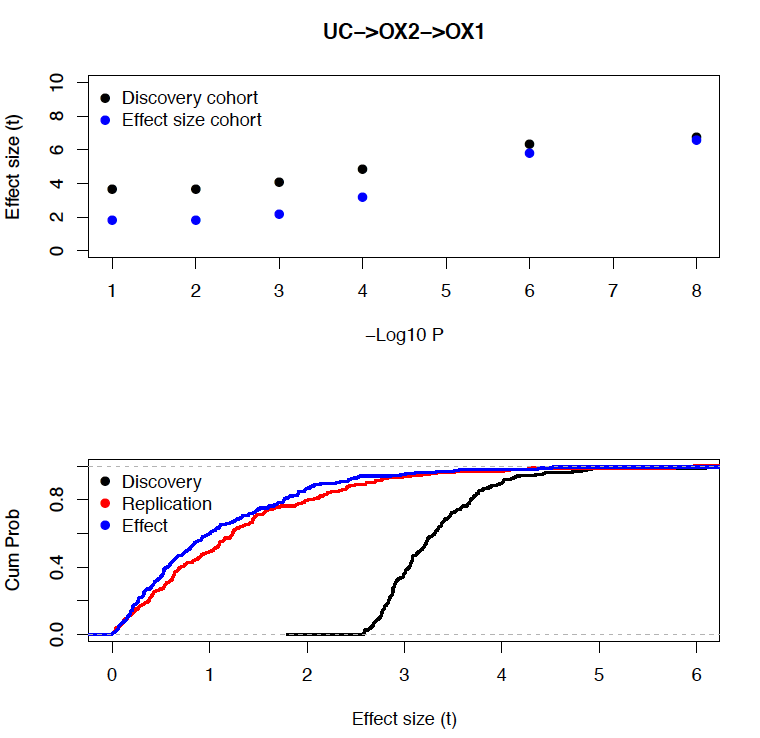
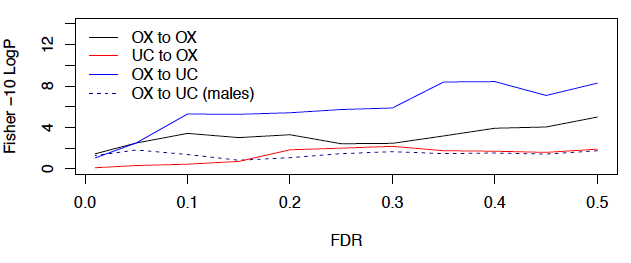


Figure 4 Testing the distribution of non-significant P-values



The horizontal axis is the false discovery rate at which discovery P-values were selected. The vertical axis is the result of a test (expressed as the negative logarithm, base 10 P-value) for deviation from a uniform distribution of the discovery P-values. Discovery P-values were chosen in 11 groups, from all those with an FDR less than 0.01 up to all with an FDR less than 0.5. Note that each group includes P-values from those preceding it, so that the final result (FDR < 0.5) includes all replication P-values.

**Supplementary Table X**. **Novel QTLs detected in the joint mapping.** Four QTLs harbour genes (underlined) implicated in protein turnover and/or associated with muscle hypertrophy. Genes in bold are expressed more abundantly in skeletal muscle compared to other tissues.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| C | Pos | pvalue | beta | A.F. | FDR | Pheno\* | CI,L\*\* | CI,R\*\* | Interval, bp | Genes | Expression (BioGPS)\*\*\* |
| 1 | 119,428,714 | 0.000010 | -0.4 | 0.45 | 0.01 | EDL | 119.2 | 120.4 | 1,122,960 | 28 | *Tmem177, Sctr, Tmem37, Dbi, 3110009E18Rik, Steap3* |
| 2 | 53,221,117 | 0.000043 | 0.23 | 0.32 | 0.04 | Gastroc | 53.0 | 55.4 | 2,352,486 | 13 | *Fmnl2, Prpf40a* |
| 2 | 89,685,916 | 0.000068 | -0.2 | 0.23 | 0.03 | EDL | 88.3 | 90.7 | 2,316,593 | 121 |  |
| 3 | 98,935,351 | 0.000002 | -0.3 | 0.13 | 0.001 | TA | 96.2 | 104.6 | 8,277,225 | 214 | *Hfe2, Polr3gl, Lix1l, Pex11b, Polr3c, Rnf115, Acp6, Chd1l,* ***Prkab2****, Tbx15, Casq2, Vangl1, Csde1, Ampd1, Olfml3, Ap4b1* |
| 3 | 144,138,946 | 0.000035 | -0.3 | 0.09 | 0.02 | Soleus | 143.5 | 145.2 | 1,638,467 | 37 | *Clca3b, Odf2l,* ***Sh3glb1****, Clca1* |
| 18 | 54,628,936 | 0.000058 | -0.3 | 0.11 | 0.03 | EDL | 53.8 | 56.1 | 2,254,451 | 28 |  |
| 18 | 61,300,554 | 0.000021 | -0.2 | 0.27 | 0.02 | EDL | 61.0 | 63,0 | 1,879,648 | 54 | *Pdgfrb, Slc26a2, Ppargc1b, Csnk1a1, Bvht, Afap1l1, Ablim3, Adrb2, Apcdd1* |

\* Phenotypes; weight of EDL, Gastroc, TA and soleus muscles. \*\* confidence interval left and right of the peak SNP was determined as the first SNP which showed no association with the phenotype (p>0.05 FDR) in the series of SNPs spanning for ~1 Mb. \*\*\* genes in skeletal muscle exceeding the median expression level among >90 tissue and cell types in BioGPS mouse database (Wu et al., 2009).

Relevant refs:

Khan M.M., Strack S., Wild F., Hanashima A., Gasch A., Brohm K., Reischl M., Carnio S., Labeit D., Sandri M., Labeit S., and Rudolf R. Role of Autophagy, SQSTM1, SH3GLB1, and TRIM63 in the Turnover of Nicotinic Acetylcholine Receptors. Autophagy 2014: 10: 123-136.

Wojtaszewski J.F., Birk J.B., Frosig C., Holten M., Pilegaard H., and Dela F. 5'AMP Activated Protein Kinase Expression in Human Skeletal Muscle: Effects of Strength Training and Type 2 Diabetes. J Physiol 2005: 564: 563-573.

Joassard O.R., Amirouche A., Gallot Y.S., Desgeorges M.M., Castells J., Durieux A.C., Berthon P., and Freyssenet D.G. Regulation of Akt-mTOR, Ubiquitin-Proteasome and Autophagy-Lysosome Pathways in Response to Formoterol Administration in Rat Skeletal Muscle. Int J Biochem Cell Biol 2013: 45: 2444-2455.

Sato S., Nomura S., Kawano F., Tanihata J., Tachiyashiki K., and Imaizumi K. Effects of the Beta2-Agonist Clenbuterol on Beta1- and Beta2-Adrenoceptor mRNA Expressions of Rat Skeletal and Left Ventricle Muscles. J Pharmacol Sci 2008: 107: 393-400.

Rosado M., Barber C.F., Berciu C., Feldman S., Birren S.J., Nicastro D., and Goode B.L. Critical Roles for Multiple Formins during Cardiac Myofibril Development and Repair. Mol Biol Cell 2014: 25: 811-827.