Supplementary Note

We have divided the supplementary note into four sections: (1) acquisition, care and behavioral and physiological phenotyping of the mice; (2) genotyping-by-sequencing (GBS) assays and genotype calling; (3) assessment of genotype quality, and characterization of key genetic properties of CFW mice; (4) Mapping loci for behavioral and physiological traits, quantification of gene expression levels in multiple tissues using RNA-seq assays and statistical procedures to map QTLs and eQTLs. Relevant data and the R code implementing our statistical analyses are available at http://github.com/pcarbo/cfw.

1.1 Overview of phenotype data pre-processing

Once we collected and recorded all the phenotype data, we took several steps to prepare the data for QTL mapping. First, we tried to identify relevant covariates (e.g., age, weight, box/test apparatus). For those that explained at least some of the phenotypic variance, we included them in the linear regression models of the phenotype (see Supplementary Table 1). We also checked for batch effects—that is, whether the phenotypic measurements were different in one of the shipments of mice. In cases where we detected such effects (e.g. batch 16 for some of the musculoskeletal traits) we corrected for the batch effect by including a binary indicator for the batch as a covariate in the model (Note that all of our mice are males, so there is no need to check for effects of sex on the phenotypes.) Second, we considered whether the data could be made to more closely resemble a normal distribution using an appropriate transformation. For example, all the fear conditioning phenotypes were proportions (numbers between 0 and 1), so we used the logit function (inverse of the sigmoid function) to convert the proportions to numbers distributed on the real line. Third, we inspected the phenotype residuals, which were obtained by removing linear effects of covariates. We checked whether the residuals had empirical quantiles that closely matched expected quantiles under the normal distribution. If necessary, we removed outlying data points. Details specific to each phenotype are given in the next sections. Investigators were blind to mouse genotype during the collection of all phenotypes.

1.2.1 Fasting glucose levels

One day after the mice arrived at University of Chicago (Supplementary Figure 1), we measured glucose levels 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 milligrams per deciliter (mg/dL), 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 they were tested for methamphetamine sensitivity (Supplementary Figure 1).

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

1.2.2 Methamphetamine sensitivity phenotypes

Sensitivity to the locomotor stimulant effects of methamphetamine (MA) is a heritable trait that utilizes neurocircuitry also associated with the rewarding effects of drugs. The primary objective of this test was to identify genes that influenced initial sensitivity to MA. Specifically, we measured the extent to which MA

stimulated locomotor activity. We followed a 3-day testing protocol: on the first two days, the mice were injected with saline and placed in one of 12 open field arenas to assess baseline response to a novel environment; on the third day, mice were injected with a dose of MA, and locomotor activity was again recorded. Our testing protocols closely followed those previously described[1–4].

Mice were approximately 52 days of age at the beginning of testing (range = 46–55 days). Testing was conducted over three consecutive days during the light phase, between 08:00 and 16:00 hours. Mice were transported from the adjacent vivarium and then allowed to habituate to the procedure room for 30 minutes in their home cages. Each chamber consisted of a clear acrylic arena (40 x 40 x 30 cm) placed inside a frame containing evenly spaced infrared photo beams 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 distance traveled (cm). Each activity chamber was encased within a sound-attenuating PVC/lexan environmental chamber (AccuScan). Overhead lighting in each chamber provided dim illumination (~80 lux), and a fan provided both ventilation and masking of background noise.

On the first and second days of testing, mice were removed from their home cages, weighed, and placed in individual holding cages filled with clean bedding. Mice were then given an intra-peritoneal (i.p.) injection of physiological saline, and were immediately placed in individual activity chambers where locomotor activity was recorded for 30 minutes. On the third day of testing, mice received an i.p. injection of 1.5 mg/kg MA, and were immediately placed in the activity chambers to measure locomotor activity over 30 minutes. The 1.5 mg/kg dose was intended to produce locomotor stimulation without inducing stereotyped behaviors. All systemic injections were administered in a volume of 0.01 ml/g body weight.

On all three days, mice were returned to their home cages immediately after the 30-minute test. Activity chambers were cleaned with 10% isopropanol between tests. We returned the mice to the vivarium at the end of each day.

Locomotor activity was measured as the distance traveled (in cm) over the 30-minute interval immediately after receiving an injection. We measured distance travelled using automated Versamax activity chambers (AccuScan). Beam breaks were recorded on a computer and converted into distance travelled.

For assessing sensitivity to MA, the main phenotype of interest was locomotor activity on day 3. MA response was defined as the total distance travelled on day three during the 30 minute test beginning immediately after administration of the drug. Activity and time spent in the center of the arena on the first day of testing was also recorded, as it provided a measure of baseline response to a novel environment; differences in response to novelty have been associated with high levels of subsequent drug self-administration in rats[5].

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

1.2.3 Conditioned fear phenotypes

Approximately 12 days after testing for MA sensitivity (see Supplementary Figure 1), at a mean age of 63 days (range = 58–69 days), we tested mice for conditioned fear (CF). CF is a classic Pavlovian learning paradigm in which an aversive unconditioned stimulus is paired with a previously neutral stimulus and recall of the fearful memory is measured. The conditioned fear 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 test chamber (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. Our conditioned fear testing protocol was similar to

those described previously[6–10]. Pilot studies did not identify an effect of prior methamphetamine testing on this behavior (results not shown); since all mice had identical prior exposure we do not expect it to be a major confounding factor.

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 the first day, 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 ("pre-training freezing") 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.

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). 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.

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 on this occasion they were not paired with shocks. We recorded freezing over two periods: the same 30-to-180-second period as on days 1 and 2, and during presentation of the four 30-second tones (at 180–210, 240–270, 300–330 and 360–390 seconds).

In summary, our phenotypes from the 3-day fear conditioning tests consisted of five measurements of immobility: average proportion of freezing on day 1 during the pre-training interval (30-180 seconds) before exposure to tones and shocks ("pre-training freezing"); average proportion of freezing on the first day during exposure to the conditioned stimulus ("freezing to tone on day 1"); average proportion of time freezing in the 30–180 second interval on the second day in conditions identical to the first day ("freezing to same context"); average proportion of time freezing over the 30-180 interval on the third day in an altered context; and average proportion of time freezing on the third day in the altered setting during the 30-second intervals in which the tones we presented ("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[11]. To avoid extremely small or extremely large values after the transformation, the proportions were projected onto range [0.01, 0.99].

For analysis of the day-2 and day-3 freezing measures, we corrected for freezing on day 1 ("freezing to tone on

day 1") in order to attenuate the effect of baseline variation in freezing that was not relevant to learned fear behavior. Freezing to tone on day 1 explained 25% of variation in freezing to the same context on day 2, 10% of variation in freezing to the altered context on day 3, and 20% of variation in freezing to the tones on day 3. For all the fear conditioning traits, our 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 fear conditioning phenotypes. Further, the fear conditioning 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.

1.2.4 Prepulse inhibition (PPI) phenotypes

We tested mice for PPI approximately 9 days after the final day of CF testing (Supplementary Figure 1). Mice were tested at a mean age of 75 days (age range = 69–79 days). PPI is a reduction in startle response that occurs when a non-startling lead stimulus ("prepulse") precedes a startling stimulus[12], and is considered to be an endophenotype for schizophrenia and possibly other psychiatric disorders as well[13,14]. During PPI, the mice were exposed to loud pulses (120 dB) that caused them to exhibit the startle response. The exposure to the loud pulse was sometimes preceded by a barely perceptible "prepulse" (3–12 dB over background levels), which sometimes inhibited the startle response. Our PPI testing procedures follow protocols detailed in previous papers[12,15–18].

Immediately before testing, mice were transferred from the vivarium to the testing room, one cage at a time. The mice were weighed, and then placed into one of the 5 cylindrical Plexiglas containers (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 the recorded it on a computer. Before the start of each test day, the chambers 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 pseudo-random order. The startle response during each trial was recorded for 65 ms beginning at the start of the 120-dB stimulus, and at the start of all "no stimulus" trials. 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.

The startle inhibition phenotypes were defined as the difference of the average startle amplitude during the 3, 6 or 12-db prepulse trials to the average startle amplitude during the pulse-alone trials, normalized by the pulse-alone amplitude:

$$PPI = (SA_{pulse} - SA_{prepulse}) / SA_{pulse}.$$

Here, we defined SA_{pulse} to be the average startle amplitude measured in the pulse-alone trials in testing blocks 2 and 3, and $SA_{prepulse}$ to be the average startle amplitude averaged across all prepulse trials of a given sound pressure level (3, 6 or 12 dB). Our data showed that higher prepulse levels yield greater amounts of inhibition, as expected. Note that startle amplitudes are expressed in arbitrary units.

In addition to startle inhibition, we measured two startle response phenotypes: baseline startle response, which we defined as the average startle amplitude during the pulse-alone trials in blocks 2 and 3 (average startle to pulses); and habituation to the pulses (habituation to pulses), which we defined as the average startle amplitude during the fourth pulse-alone trials after removing the effect of average startle response during the first pulse-alone trials.

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 value. 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 did not respond to the 120 dB pulses, suggesting the possibility that they were deaf (Supplementary Figure 2). These "deaf" mice added a disproportionate amount of variance to the PPI phenotypes because the denominator (SA_{pulse}) was close to zero, 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. Therefore, we included all PPI testing chamber indicators as covariates in analysis of the PPI phenotypes. Body weight was expected to be correlated with startle amplitude, but this correlation disappeared once we normalized by the average response during the pulse-alone trials.

1.2.5 Bone-mineral density

Bone-mineral density (BMD) in each mouse was measured by Dr. Cheryl Ackert-Bicknell at Jackson Laboratories (now at the University of Rochester). 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).

To obtain a normal distribution, we transformed the BMD measurements, which were ratios, to the (base 10) log-scale. A concern with areal BMD and bone-mineral content (BMC) 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 BMD.

In comparison to the Hybrid Mouse Diversity Panel (Supplementary Figure 3), a substantial fraction of the CFW mice exhibited abnormally high BMD, 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 BMD fell on the long tail of the observed distribution (refer to Supplementary Figure 3), which we defined as any value of real BMD 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.

1.2.6 Musculoskeletal traits

Following sacrifice, one leg was cut off just below pelvis, placed in a tube and transferred into a -80 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 (Z22855, OWIM GmbH & Co).

The examined muscles differed in size, and in properties of the fibers that constituted the individual muscles. Since there are important functional and developmental differences between fiber types[19], it was expected that muscles with different composition of fiber types were affected by distinct genetic mechanisms. For example, soleus muscle is a slow-twitch muscle dominated by type 2A (most abundant), type 1 and 2X fibers (least abundant), and contains only traces of type 2B fibers[20]. On the other hand, the fast-twitch EDL and plantaris muscles are dominated by type 2B, 2X and 2A fibers. TA and gastroc are larger fast-twitch muscles that express the entire range of fiber types. Different fiber types were not evenly distributed in these muscles, but instead were arranged in a surface-to-core gradient, in which type 2B fibers comprised the superficial region, type 2A and type 1 comprised the core, and 2X fibers were found in between. Soleus and EDL were similar in size, while plantaris was approximately 2 times larger (~20 mg).

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. To illustrate this, we show in Supplementary Figure 4 the association statistics (*p*-values) for TA muscle weight with and without conditioning on the linear effect of tibia length. We observed that including tibia length as a covariate led to stronger support for the TA muscle weight QTL on chromosome 5, and another QTL on chromosome 2 was detected only after controlling for the effect of tibia length. Support for a QTL on chromosome 11 decreased substantially after controlling for tibia length (Supplementary Figure 4), but this locus showed some evidence for being directly associated with tibia length. Further, isolating the variance in muscle weight that was not explained by tibia length allowed us to conclude that variants at the same locus on chromosome 12 separately affected EDL muscle weight and tibia length (see Supplementary Figure 15 and Supplementary Table 2). In our data, tibia length explained 12–17% of the variance in the muscle weights.

We 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 better 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, we 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-mineral density (below), we 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.

1.2.7 Body weight: Body weight was recorded at various time points, including (1) upon arrival in Chicago when the tail was snipped for glucose testing, (2) on three consecutive days of the methamphetamine sensitivity tests, (3) during PPI testing, and (4) when the mice were sacrificed. Since body weights measured on different days were highly correlated with each other, we only mapped QTLs for body weight measured when mice were sacrificed. 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 included age as a covariate for the initial body weight measurement.

1.2.8 Testis weights: Testis weights were measured at the time of sacrifice, when the mice were approximately

91 days of age (range was 86–99 days). The testis were removed, weighed on a digital scale, placed in tubes containing RNALater (Ambion) and then stored in a –80 C freezer for possible future studies.

1.2.9 Tail length: Tail length (in cm) was measured at the time of sacrifice. We measured tail length as the base of the tail to the tip of the tail.

1.3 Measurement of gene expression using RNA-Seq

We collected RNA-Seq gene expression data in three brain tissues to augment our QTL mapping efforts for behavioral traits. As shown in previous studies, gene expression is an informative intermediate phenotype for many behavioral phenotypes, including behavioral traits such as conditioned fear and methamphetamine sensitivity[4,7]. In the following sections, we describe the samples, tissues, and preprocessing steps for collecting RNA-Seq experimental data.

1.3.1 Tissues and samples

Samples of the brain were removed from one mouse per cage immediately after sacrifice. The mouse to be sampled was chosen in advance in a randomized fashion, and was always the first mouse to be removed from the cage. The whole brain was incubated in a pool of chilled RNALater (Ambion) for one minute before individually dissecting the prefrontal cortex, hippocampus and striatum. Individual tissues were placed in 1 mL of RNALater buffer and immediately put on ice. Samples were kept in a –80°C freezer prior to RNA extraction.

1.3.2 Extraction and sequencing

Brain samples were removed from RNALater and homogenized in 1 mL QIAzol lysis reagent (Qiagen). RNA was isolated using a standard phenol-chloroform procedure and purified using GenCatchSpin mini spin columns (Epoch Life Sciences). Total RNA was quantified with the Quant-It Ribogreen RNA Assay Kit (Invitrogen) and checked for quality using the Bioanalyzer RNA 6000 Nano Kit (Agilent). Only samples with a RIN score above 8.0 were processed for sequencing.

Library preparation was performed with the TruSeq RNA Sample Kit (Illumina). mRNA was purified from 1 μg of total RNA. cDNA libraries were quantified using the Quant-It Picogreen DNA Assay Kit (Invitrogen) and checked for quality using the Bioanalyzer DNA 1000 Kit (Agilent).

We sequenced the RNA libraries on an Illumina HiSeq 2000 sequencer, using single-end 100-bp reads, to a target depth of 60-fold coverage per sample. The sequenced reads were demultiplexed using Illumina's in-house processing pipeline, "CASAVA" (http://support.illumina.com/sequencing/sequencing_software/casava/documentation.html). The output of this processing pipeline is the set of short reads for each sample.

1.3.3 Quantification of RNA levels

We processed the RNA-Seq read data using the *Tuxedo* software suite[21]. First, we aligned the short reads to NCBI release 38 of the Mouse Genome Assembly (mm10) using *bowtie2* (version 2.2.2)[22]. Since this alignment can fail for reads that span splice junctions, we improved the alignment using *tophat2* (version 2.0.11), which is able to align short reads to known splice junctions. To compile a list of known splice junctions, we used the gene models from Ensembl release 68, which includes all known genes, novel genes and pseudogenes at the time of the release (July 2012). Note that we did not use *tophat* to discover unknown isoforms or splice sites since that was beyond the scope of our study. With the alignments obtained from *tophat* that included reads spanning multiple exons, we used *cufflinks* (version 2.1.0)[23] to calculate relative quantities of each isoform per sample, with the same gene models used for alignment. We also calculated a gene-level measure of expression in *cufflinks*, in terms of reads per kilobase per million mapped (RPKM). Importantly, the RPKM measure is invariant to both the transcribed sequence length of genes, and the sequencing depth of samples. We focused on

this gene-level measure of expression for subsequent analyses, including eQTL mapping and tests for allele-specific expression.

2. Genotyping-By-Sequencing (GBS)

In this section, we detail the various steps that were taken to obtain high confidence genotypes for all individuals using a genotyping-by-sequencing approach. These steps include the DNA extraction, digestion and library preparation followed by the bioinformatics analyses to genotype the samples at loci proximal to the restriction enzyme cut sites.

2.1 DNA isolation, digestion and GBS library preparation

Genomic DNA was isolated from the spleen using a standard salting-out protocol. 1 μg of DNA was digested with the Type-II restriction endonuclease *PstI*, following the protocol described in Grabowski *et al.*[24]. *PstI* recognizes a 6-bp motif, CTGCAG, and generates a 3' overhang of 4 bp, TGCA. GBS libraries were prepared using a set of 48 indexed adapters designed by Grabowski *et al.* [24].

Two types of adapters were used in library preparation for GBS: an indexed, or barcoded adapter, and a common adapter[25]. The GBS libraries were derived from DNA fragments that were ligated to an indexed adapter at one end and to a common adapter at the other end. Fragments ligated to the same type of adapter at both ends were excluded from the final library because they contained a binding site for only one of the two primers, and these primers are required for amplification and sequencing on the Illumina platform. Each indexed adapter contained a unique sequence on the 3' end of its bottom strand (5'- XXXX - AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT-3'), whereas the top strand terminated with the reverse complement of the index (YYYY) followed by the reverse complement of the overhang produced by *PstI* (5'- ACACTCTTTCCCTACACGACGCTCTTCCGATCT-YYYY-TGCA-3'). The common adapter did not contain an index. Its bottom strand terminated with the complement of the *PstI* overhang (5'-CTCGGCATTCCTGCTGAACCGCTCTTCCGATC-TGCA-3') and its top strand contained only the sequence required by the Illumina platform (5'-GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3').

2.2 Sequencing of pooled GBS libraries

GBS is a reduced-representation sequencing method that aims to sequence only the regions that are proximal to a restriction enzyme cut site. Since only a fraction of the genome is sequenced, GBS allows many more samples to be sequenced on a single flowcell compared to whole-genome sequencing methods. There are approximately 1 million *PstI* cut sites across the mouse genome. We estimated this using an *in silico* digestion of the mouse reference genome (mm10). In practice we found that we obtained sequencing data at only one third of these cut sites, perhaps in part due to size selection (see below). We pooled 12 barcoded GBS samples per sequencing lane. At this level of pooling, we expected ~15-fold coverage across the GBS cut sites. The pooled samples were sequenced with the Illumina HiSeq 2000 sequencing platform using 100-bp single-end reads, across 101 flowcell lanes. We obtained and average of 4.8 million reads per sample. The distribution of coverage across the genome averaged across all the samples that were sequenced is shown if Supplementary Figure 24.

2.3 GBS variant discovery and genotyping

The 100-bp single-end short reads were aligned to Mouse Reference Assembly 38 from the NCBI database (mm10) using the short read aligner *bwa* (version 0.5.5-r16/0.7.4)[26]. Our protocol for discovering variants and obtaining genotype probabilities using the GBS short reads was largely derived from the GATK best practices pipeline for whole-genome sequencing. Given that there are some important differences between GBS and whole genome sequencing, we modified some steps of the "best practices" pipeline, as explained below.

2.3.1 Indel realignment

Prior to variant discovery, we used *picard* (version 1.129) to generate GATK-compatible sequence alignment files. We realigned the reads around known indels using GATK's indel realignment tool. Since there is no catalog of known indels for CFW mice, we used the set of indels discovered in lab strains for the Mouse Genomes Project[27].

2.3.2 Variant discovery and filtering

Following indel realignment, we used the GATK Unified Genotyper (version 2.6-5), in discovery mode to identify polymorphisms and obtain genotype probabilities at these variant sites. Initially, we used liberal thresholds in order to identify variants at a high level of sensitivity (at the cost of a high rate of false discovery). In this first step, we obtained 5.42 million variants proximal to the cut sites of the restriction enzyme *PstI*. Since we expected most of these variants to be sequencing errors, we used a two-step filtering process.
In the first step, we discarded variants that did not meet thresholds for missingness (95%) and MAF (< 1%). We used a non-stringent threshold for missingness since we expected to be able to accurately impute genotypes even when the genotypes were called only in a small fraction of the samples. Second, we filtered out variants that had a minor allele frequency less than 1% as we expected very few rare variants, and we would not have good power for an association study with these variants. Note that, as a result of imputation, some SNPs in our final dataset have a MAF less than 1%. Filtering out variants using these thresholds left us with approximately 1.05 million variants, many of which are still expected to be false positives.

2.3.3 Variant quality score recalibration

In the second step of variant filtering, we ran the GATK Variant Recalibrator algorithm to recalibrate quality scores so that they more accurately reflected the probability that the genetic variant identified was a true variant. This was accomplished by applying various annotations to the markers, and using these annotations to classify these markers. In addition to the standard Variant Quality Score Recalibration (VQSR) annotations, such as Quality-by-depth (QD), haplotype score, rank sum of the mean quality at variant sites and mapping quality, we also included annotations for minor allele frequency and inbreeding coefficient. These two annotations in particular led to a substantial increase in recall, yielding a larger set of variants at the same level of specificity (results not shown). We explicitly excluded annotations based on read-depth or read-position because they are not informative; GBS yields high variation in coverage, and produces reads that always map to the same start position (the start position corresponds to the restriction enzyme cut sites).

VQSR calibrates the quality scores against a known, or "ground-truth," set of SNPs. One option would be to use the variants discovered in the lab strains as part of the Mouse Genomes Project[27] as a proxy for the set of ground-truth polymorphic sites in the CFW mice. However, as the history of CFW mice is not well documented, it is unknown whether the diversity of the lab strains could reasonably capture genetic polymorphisms in the CFW population. In light of this concern, we instead generated our own ground-truth training set for VQSR using whole-genome sequencing (WGS) data for a small number of CFW mice (full details on the whole-genome sequencing are provided in the next section). In addition to the WGS variants, we used the variants identified in the Wellcome Trust lab strains and variants in dbSNPv137 as "training" or "known" sets of SNPs. We applied VQSR to the set of about 1 million variants that were retained after the MAF and missingness threshold. We used 4 different VQSR tranches at 50%, 80%, 90% and 99%, *i.e.*, different thresholds for separating true from false variants based on the number of known variants that are rediscovered (ranging from 50–100%). Of the 1 million variants that we began with, we classified 18,264, 64,093, 121,723 and 85,6091 variants as true variants at increasing levels of leniency (50%, 80%, 90% and 99%). Based on different metrics such as the ratio between known and novel variants and transitions to transversions (Ti/Tv), we decided to use the variants classified as true variants at the 90% tranche. Among the 121,723 variants meeting this cutoff, there were 92,734 bi-allelic SNPs, 1,016 multi-allelic SNPs and 27,973 indels. Since we were not confident about the discovery or genotyping of indels and multi-allelic SNPs using GBS, we excluded

them from all further analyses. Reassuringly, for the 92,734 bi-allelic SNPs used for all downstream analyses, the ratio of Ti/Tv was 1.97, which was close to the expected ratio of 2. The Ti/Tv ratio for newly discovered SNPs was 0.71, suggesting that these SNP consisted a mixture of true and false positives. The genome-wide distribution of GBS-discovered SNPs in the CFW mice, and the corresponding density of SNPs in the Wellcome Trust lab strains, are shown in Supplementary Figure 5.

2.4 Whole-genome sequencing, and variant discovery using WGS data

From the mice that were genotyped using the GBS protocol described above, we selected 12 of these mice for low-depth whole-genome sequencing (WGS). The main reason for sequencing these mice was to obtain a catalog of variants in the CFW population that could then be used to calibrate variant discovery from the GBS data in the larger number mice, without biasing variant discovery toward the variants segregating in lab strains.

The 12 mice were shotgun-sequenced using single-end 100-bp reads on an Illumina HiSeq 2000 sequencer. Three samples were multiplexed per flowcell lane, resulting in an expected 3-fold coverage across the genome. The raw reads were aligned to the mm10 reference genome using *bwa*[26]. We used the GATK "best practices" pipeline to identify a set of high-confidence variants, as detailed in the next paragraph. This set of variants was subsequently used to score variants identified using GBS.

We preprocessed the alignments using *picard* so that they were in the form required by GATK. We then used the GATK Indel Realigner to realign reads overlapping known indels. For the reference set of known indels, we retrieved the indels identified in lab strains as part of the Mouse Genomes Project[27]. Using GATK's Unified Genotyper, we identified variant sites, and called genotypes at these sites. Similar to the GBS procedure, we used a liberal threshold to acquire an inclusive set of variants resulting in an initial set of 8.94 million SNPs and indels. Next, using GATK's Variant Recalibrator, we calibrated the quality scores of the WGS variants against the variants identified in the Mouse Genomes Project lab strains. To evaluate the recalibration step, we computed quality metrics using a different variant panel—the set of variants present in the dbSNP database (release 137) and the variants discovered as part of the Wellcome Trust effort to sequence the different lab strains. To minimize the number of false variants in the training set used for variant discovery in the GBS pipeline, we retained only variants that passed VQSR at our most stringent recalibration tranche, where only 50% of the training variants were recovered. Prior to the variant recalibration step, 1.47 million variants were filtered out due to low quality. A further 4.59 million SNPs and indels were filtered by the variant recalibration step, leaving us with a final set of 2.88 million SNPs from the 12 mice that were whole-genome sequenced. These 2.88 million SNPs were used as part of the recalibration step for GBS genotyping; however, we did not expect to observe nearly this many because GBS is a reduced-representation sequencing method.

2.5 Imputing SNP genotypes

GBS yields highly variable coverage across samples at the same cut site, and therefore highly heterogeneous call rates. Like most DNA sequencing technologies, read coverage in GBS varies across the genome. However, unlike shotgun sequencing, we cannot identify PCR duplicates in GBS, which is an important source of variation in read coverage. Our inability to filter out PCR duplicates in GBS means that variability in coverage is higher than would be expected when using comparable technologies such as shotgun sequencing. This resulted in some sites with very high call rates, and other sites in which only a fraction of the samples had sufficient coverage to call genotypes.

We used imputation to estimate GBS SNP genotypes that were not called due to insufficient coverage, and even improved the accuracy of other genotypes that were not called with high confidence. In addition, we use the called genotypes along with LD information to estimate the genotypes at sites that were not called from the sequence reads. We used IMPUTE2[28] to estimate missing genotypes, while improving the accuracy of genotypes that were estimated with low confidence using the GBS sequence reads. We recorded allele dosages (expected allele counts) from imputation instead of deterministic genotype calls. These dosage estimates

captured genotype uncertainty in imputation, and therefore allowed us to account for uncertainty in downstream analyses.

3. Analysis of GBS data

We used the genotype data to characterize useful genetic properties of the CFW mouse colony. We investigated confounding factors including population stratification, excess inbreeding and cryptic familial relationships. We used principal component analysis (PCA) and t*reemix* (version 1.12)[29] to investigate population structure, and *relate*[30] to estimate tracts inherited identical by descent (IBD). As part of this investigation, we also compared the GBS genotype data against genotypes from a custom SNP array to flag samples that were mislabeled or mishandled over the course of the study. We detail our procedures and findings in the following sections.

3.1 Sample mislabeling

Because of the numerous steps including behavioral testing, spleen tissue collection, DNA extraction, GBS library preparation, and sequencing, we were concerned about the possibility that samples had been inadvertently mislabeled or switched. Therefore, we sought to identify potential sample mixups by independently validating genotypes.

3.1.1 Sequenom Genotyping

We compared genotypes from tail tissue that were collected when the mice first arrived in Chicago to the GBS genotypes obtained from spleen tissue that was collected at time the mice were sacrificed (approximately 2 months later). When the genotypes from spleen and tail did not agree, we concluded that the sample had been mislabeled; as described below, such samples were excluded from our GWAS. We genotyped 669 samples using DNA obtained from the tail samples at 50 SNPs using a custom Sequenom genotyping microarray. The 50 SNPs were chosen because they had allele frequencies near 0.5 in a small discovery sample genotyped using the Mouse Diversity Array.

We discarded 3 out of the 50 Sequenom SNPs that had low call rates. For each of the remaining 47 SNPs, we compared the Sequenom genotypes against a nearby marker in the GBS panel with the highest correlation (estimated in our sample). Out of the 47 Sequenom SNPs, 14 did not have any GBS SNP with a correlation coefficient greater than 0.7; we excluded these 14 SNPs from further analysis. For each of the 33 SNPs that remained, we fit a linear model of the Sequenom genotype given the genotype at the nearby GBS SNP, and calculated the sum of absolute deviations between observed and fitted values of the Sequenom genotype across all the 33 Sequenom SNPs for each sample. We used this sum of absolute deviations as a measure of the difference between the tail and spleen genotypes for each sample. To assess whether these deviations might be expected if they were from different samples, we repeated the same analysis with permuted Sequenom labels to generate a null distribution of this statistic (i.e. the null here captures the distribution of errors when the GBS and Sequenom genotypes come from different samples). The observed and null distribution of this statistic are shown in Supplementary Figure 25. By inspecting the null distribution, we defined samples with a deviation score above 15 as being mislabeled.

3.1.2 OpenArray Genotyping

To further investigate sample mislabeling in our entire sample (not just the samples for which Sequenom genotyping was performed), we used a custom Life Technologies OpenArray genotyping array. We genotyped all mice in the study using DNA from spleen. We also included the DNA from the tail for a subset of animals. The OpenArray genotyping array consisted of 16 SNPs that were also genotyped by GBS, with minor allele frequencies over 0.3. In this phase we had DNA from both the tail and spleen that were purportedly from the same subject. By comparing genotypes from the tail and spleen samples, we were able to identify possible errors in sample collection or labeling.

Out of the 16 markers that were genotyped, two were excluded because they had poor clustering. Another four were excluded because of Hardy-Weinberg errors. As a result, only 10 SNPs were used for analysis (Supplementary Figure 26). Similar to the method used for Sequenom genotyping, we fit a linear model to compare the OpenArray and GBS genotypes, then we summarized the discrepancy between the GBS and OpenArray genotypes as the mean of the absolute differences in the genotypes. We calculated the null distribution of this statistic by randomly permuting the sample labels for the OpenArray genotypes (Supplementary Figure 27). For genotypes obtained from the tail, the null distribution of the mean absolute difference was indistinguishable from the observed data, suggesting widespread errors. By contrast, we found that ~90% of the spleen samples had low discrepancies in comparing GBS and OpenArray genotypes. We flagged 110 samples for exclusion because the GBS genotypes had high deviation from the OpenArray genotypes for the same sample (defined as an average mismatch of 0.5 or greater). We concluded that these samples had been mislabeled during GBS library preparation. Analysis of individual batches or flowcells did not identify any discernable patterns, suggesting multiple small errors involving a small number of samples each, rather than a single isolated error that impacted many samples.

3.2 Estimation of IBD, and genotyping error rates

We measured error rates in GBS genotyping by comparing called genotypes within segments that were estimated to be identical by descent (IBD). We focused on the 100 pairs of mice with the highest estimated k_2, and estimated shared IBD segments in these mice using software *relate*[30] (example shown in Supplementary Figure 10). We used these tracts to confirm the kinship coefficient estimates from *plink* (Supplementary Figures 28, 11). For these pairs, we detected few tracts in which both alleles were inherited IBD ("IBD-2 tracts").

At a posterior probability of 0.75, *relate* identified 2,224 IBD segments in the 100 pairs of mice (within all of these IBD segments, only one pair of alleles is shared IBD). Using these IBD tracts, we estimated the error rate of the GBS genotypes within these tracts. In an IBD tract in which both alleles were inherited IBD, SNPs at which the genotypes were opposite homozygotes (AA and BB) or one sample was a homozygote and the other was a heterozygote constituted an error. In a tract in which one pair of alleles was inherited IBD, any instances where the samples were opposite homozygotes was an error (homozygous mismatch). From both types of tracts, we counted 124 homozygous mismatches out of 7,953 total genotypes included in the IBD tracts. This yields an estimate of 1.55% for the error rate in GBS genotypes. Note that this error rate estimate should be considered a lower bound as it depends on the parameters used to detect the IBD tracts, such as error tolerance ($\varepsilon = 0.01$) and the minimum posterior probability (0.75) needed to identify a tract as an IBD tract.

3.3 Genotype validation using MegaMUGA array genotypes

We genotyped a subset of our samples on the MegaMUGA genotyping array[31] to estimate the discordance rates between MegaMUGA and GBS. We selected 48 animals from the different batches and shipments. Of the 48 samples, three failed due to technical issues, and three additional samples were excluded due to suspected sample mislabeling. We used the remaining 42 samples to estimate the discordance rate between GBS and MegaMUGA genotypes. We obtained genotypes at 77,808 SNPs on the MegaMUGA array for the 42 samples. Since the SNPs on the MegaMUGA array are reported on the previous version of the mouse reference genome (mm9), we lifted the positions of these SNPs over from mm9 to mm10 using the *liftover* tool from the UCSC genome browser suite of tools. Only 77,664 SNPs could be successfully lifted over. Out of the 77,664 SNPs that were genotyped on the MegaMUGA array, 37,611 SNPs were polymorphic in CFW mice. Among these 37,611 SNPs, 612 SNPs were also discovered using GBS. As part of the initial quality filtering of the genotypes obtained from MegaMUGA, we discarded SNPs that failed the Hardy-Weinberg equilibrium test at a *p*-value threshold of 0.05 (60 SNPs excluded). Subsequently, we identified variants where the alleles on the MegaMUGA array did not match with the alleles discovered using GBS (21 SNPs excluded). Because calculating discordance required hard-call genotypes, we only used SNPs with imputation certainty scores > 0.98 (242 SNPs excluded). Finally, using the hard-called genotypes at these 234 SNPs, we obtained a

discordance rate of 3.04% (289/9499 genotypes were discordant). Of the 289 discordant genotype pairs, there were 167 cases where MegaMUGA called a heterozygote genotype and GBS called a homozygote genotype, 87 cases where GBS called a heterozygote genotype and MegaMUGA called a homozygote genotype and 35 cases where the methods called opposite homozygotes. Further, the discordance rates were higher when GBS genotype was a heterozygote, compared to when the GBS genotype was a homozygote as shown in Supplementary Table 4.

3.4 PCA, inbreeding and relatedness

Since the breeding history of the CFW mouse colony is unknown, we used the GBS genotype data to investigate factors such as population structure and cryptic relatedness that might confound our tests for phenotype-genotype association. First, we verified that the distribution of the minor allele frequency (MAF) for the variants identified by GBS matched our expectations. Given what we knew about the breeding history of this population--a narrow bottleneck followed by maintenance as an outbred population with a small (~100) number of breeding pairs, we expected the MAF spectrum to be biased towards common alleles, i.e. we expected a reduction in the number of sites with low minor allele frequencies when compared to a population with constant population size. Supplementary Figure 6 shows that the MAF distribution is mostly flat for MAF > 0.1 and has relatively few sites with low MAF (< 0.05). We used PCA to investigate batch effects, population structure and inbreeding in our samples. The samples do not cluster after projection onto the first 2 principal components (Supplementary Figure 7), suggesting that there are no widespread batch effects or obvious population stratification (see also Supplementary Figure 8).

We estimated inbreeding coefficients for all mice using *plink* [32]. Supplementary Figure 8 shows the samples projected onto the first two principal components, in which the samples are colored according to their inbreeding coefficient. As expected, the first principal component correlates with the inbreeding coefficient. The distribution of the inbreeding coefficient, shown in Supplementary Figure 9, does not indicate excessive inbreeding in our sample. This suggests that the vendor maintained a mating scheme that was close to random, as claimed by the vendor.

To identify closely related pairs of mice, we calculated pairwise kinship coefficients for all pairs of mice using *plink*. The distribution of kinship coefficients for all pairs of mice in the sample is shown in Supplementary Figure 10. Most of this distribution lies near the origin, indicating that only a small number of samples are closely related.

3.5 Linkage disequilibrium

To produce the LD decay plot, we used the `--r2` option in *plink* 1.9[32], which computes the pairwise LD statistic r^2 [33] from maximum-likelihood estimates of haplotype frequencies. This r^2 statistic, averaged over SNP-distance intervals, was used to draw the curves in Figure 2B. Specifically, once we obtained LD estimates between pairs of SNPs on the same chromosome, we computed Monte Carlo estimates of average r^2 within SNP-distance intervals of 100 kb, ranging from 0–100 kb to 4.9–5 Mb. The Monte Carlo estimate for each SNP-distance interval was computed from approximately 10,000 pairs of SNPs. Rather than select pairs of SNPs uniformly at random, we selected pairs of SNPs that have similar frequencies (specifically, the selected pairs of SNPs do not have frequencies that differ by more than 5%). This "frequency matching" approach[34] provides a measure of LD decay that is less severely affected by differences in allele frequencies between SNPs[34] (see Hernandez *et al.*[35] for an illustration of this approach for comparing LD decay patterns between primate populations). Therefore, this approach provides a more useful metric for assessing overall resolution of association mapping in different populations. Additionally, since power to map QTLs increases with minor allele frequency, LD decay rates between more common SNPs is a more broadly relevant measure of mapping resolution, so we restricted the LD decay calculations to SNPs with MAF of 20% or greater.

To generate the CFW LD decay curve, we used the genotype data for 1,150 male CFW mice at 92,734 SNPs on

autosomal chromosomes. In these calculations, we only included genotypes called or imputed with high confidence—precisely, genotypes with maximum genotype probability at least 98%. (Overall, 80% of genotypes had maximum genotype probability of 98% or greater.) SNP base-pair positions for the CFW mice were based on NCBI release 38 of the mouse genome assembly.

For comparison, we also computed LD in several other mouse populations for which genotype data were available online, or through our collaborators: heterogeneous stock (HS) mice[36]; 30 inbred laboratory strains[37–39]; Diversity Outbred mice[31]; the Hybrid Mouse Diversity Panel (HMDP)[40,41]; and F_{34} crosses from an advanced intercross line (AIL)[6,42].

To estimate LD decay in the HS population, we downloaded genotype data for 1,940 HS mice from http://mus.well.ox.ac.uk/mouse/HS. We used the base-pair positions for 9,416 SNPs on autosomal chromosomes (out of 11,745 genotyped SNPs) from the sex-averaged genetic map that was available from the same webpage. These base-pair positions were based on NCBI release 34 of the mouse genome assembly.

The panel of inbred lab strains ("30 inbred strains" in Fig. 2B) is from our recent work on mapping QTLs for prepulse inhibition[39]. The 30 strains are a subset of inbred laboratory strains from the Mouse Phenome Database "priority strains" list for which dense, genome-wide genotype data are publicly available. We retrieved genotype data from the Mouse Diversity Array (MDA) web resource at the Jackson Labs Center for Genome Dynamics (http://cgd.jax.org/datasets/popgen/diversityarray/yang2011.shtml). We retained the 297,329 SNPs on autosomal chromosomes that were polymorphic in the 30 strains and had genotypes called in all 30 strains. All genomic positions were based on mouse genome assembly 37 from the NCBI database.

For the HMDP[40,41], we computed LD using genotype data from http://mouse.cs.ucla.edu/mousehapmap/emma.html. These data contained genotypes for 251 strains at 130,908 SNPs on autosomal chromosomes. To simplify the calculations, we set the small number of heterozygous genotype calls to missing.

To estimate genome-wide LD in the DO population, we used genotypes from the MegaMUGA genotyping platform[31], called using the Illumina Bead Studio algorithm. The data used in our LD calculations for the DO were genotypes at 137,402 SNPs on chromosomes 1–19 for 2,074 mice from DO outbreeding generations 16–19. The SNP base-pair positions are based on mouse genome assembly 38. These genotypes were sampled from later generations of the DO and have not yet been published, or made available online; they were kindly provided by Daniel Gatti at the Jackson Laboratory in Bar Harbor, Maine.

Finally, LD estimates for the advanced intercross line were from genotypes at 4,524 SNPs on autosomal chromosomes for 687 F_{34} crosses (353 males and 334 females) of the LG/J and SM/J inbred mouse strains[6]. All genomic positions in this data set were based on mouse genome assembly 37. These genotype data are available for download at http://github.com/pcarbo/lgsmfear.

4.1 Mapping loci for behavioral and physiological traits

4.1.1 Linear mixed model for QTL mapping. We did not find any obvious patterns of population structure by inspecting principal components in the genotypes, nor did we find evidence for highly related pairs of mice by inspecting estimated identity coefficients. However, it is possible that there exists subtle population structure that is more difficult to characterize, often called cryptic relatedness, in which some mice are more related than others. Such varying degrees of genetic sharing among individuals can confound tests for association in QTL mapping, leading to inflation of spurious associations[43–47]. We compared QTL mapping with and without correcting for population structure, and we found that any population structure that might exist in the mouse population had at most a small impact on association tests. Thus, in practice, the QTL mapping results would likely not change much if we ignored confounding due to population structure. Despite this finding, we did account for population structure in order to provide stronger guarantees that the QTLs we detected were not

confounded by cryptic relatedness. We used an approach based on linear mixed models (LMMs) to map QTLs for all phenotypes. LMM-based approaches for QTL mapping have emerged as a robust strategy to account for confounding due to population structure[42,43,47–52].

We used the LMM approach implemented in the software GEMMA (genome-wide efficient mixed-model association)[53]. We had two reasons for using GEMMA over other LMM-based approaches: (1) the numerical computations in GEMMA scale well to large numbers of markers and samples; (2) unlike other implementations of LMMs for genome-wide mapping (e.g. EMMAX[48], GRAMMAR[54]), GEMMA avoids making approximations which lead to a reduction of power to detect QTLs in certain circumstances. GEMMA yields calculations that are equivalent to exact test statistics (e.g. as implemented in EMMA[44]]), but computes these statistics much faster. An alternative resource that we expect would work equally well for QTL mapping is Fast-LMM[49,50].

For a given SNP, GEMMA is identical to a standard linear regression, in which the quantitative trait (Y) is modeled as a linear combination of the genotype (X) and the covariates (Z), except that it includes an additional random or polygenic effect capturing the covariance structure in the phenotype that is attributed to genome-wide genetic sharing:

$$y_i = \mu + z_{i1}\alpha_1 + \cdots + z_{im}\alpha_m + x_{ij}\beta_j + u_i + \varepsilon_i$$

Here, y_i is the ith phenotype sample, z_{ik} is ith sample of covariate k, α_k is the regression coefficient corresponding to covariate k, m is the number of covariates included in the regression (covariates are specified in sections above), x_{ij} is the genotype of sample i at SNP j, β_j is the regression coefficient corresponding to SNP j, u_i is the polygenic effect for the ith sample, ε_i is the residual error, and μ is the intercept in the linear regression. The genotype, x_{ij}, is encoded as the observed alternative allele count (0, 1 or 2), or the expectation of this count when it is estimated using genotype imputation, and therefore β_j is the additive effect of the allele count on the phenotype. The residuals e_i are assumed to be *i.i.d.* normal with zero mean and covariance σ^2, whereas the polygenic effect $u = (u_1, \ldots, u_n)^T$ is a random vector drawn from the multivariate normal distribution with mean zero and $n \times n$ covariance matrix $\sigma^2\lambda K$, where n is the number of samples. K is a matrix determined by genotype data, or by identity coefficients estimated from a pedigree. Parameters σ^2 and λ specify the contribution of the polygenic and residual variance components to the variance of Y, respectively and these quantities were estimated from the data.

All the phenotypes we investigated were continuously valued, with one exception, the abnormal BMD trait (defined above). The abnormal BMD phenotype should be modeled as a binary trait using, for example, a logistic or probit regression. Instead we applied the linear model to this trait. This can be considered a reasonable approximation under certain circumstances[55], but in practice it leads to a reduction in power to detect QTLs with larger effects.

4.1.2 Dominance effects. In this model we included only one effect per SNP (the additive effect). We did not include an additional coefficient per SNP to capture dominance effects. While it is typical in human genetic association studies to model only additive effects, in mice it is more common to allow for dominance effects, since it leads to greater power to detect dominant or recessive loci. In the CFW mice, however, we were faced with a different situation than most experimental mouse crosses because we did not expect to have SNPs that were in complete LD with the causal genetic variant. In this situation, nearby SNPs will have effects that appear to be additive, even if the functional locus is purely dominant or purely recessive[56]. Ignoring dominance effects will surely lead to some reduction in power to detect QTLs that are largely dominant or recessive, but we expect this reduction to be relatively small. The upside of ignoring dominance effects is that we avoid a reduction in overall power to detect QTLs because we do not introduce an extra degree of freedom to the model that would increase thresholds for significance.

4.1.3 Tests for association. We report support for a QTL at each SNP using the p-value calculated from the

likelihood-ratio test. To calculate the p-value, GEMMA uses the result that the likelihood-ratio test statistic under the null hypothesis follows the chi-square distribution with 1 degree of freedom[53]. Although other association tests are also implemented in GEMMA (the Wald test and the score test), there are a few advantages to using the likelihood-ratio test over other tests: (1) the λ parameter is fitted separately under the null and alternative models by efficiently maximizing the likelihood under these two models; (2) the likelihood-ratio test is well-behaved in that it controls for type I error; and (3) it exactly follows the chi-square distribution under the null model.

4.1.4 Marker-based estimates of genetic sharing. To specify the $n \times n$ covariance matrix of the polygenic effect, we must provide the relatedness matrix, K. We use the genotypes of the genetic markers to specify this matrix. (Previous work has shown that marker-based and pedigree-based estimates of genetic sharing yield QTL mapping results that broadly agree, at least in some experimental crosses[6,57]). The expression for K can be derived in different ways, depending on whether one takes an expectation with respect to the (unknown) genetic effect, and conditions on the (observed) IBD status, or whether one takes an expectation with respect to (unknown) alleles and conditions on the genetic effect[58,59]. In practice, these different choices for relatedness matrices yield similar association results in non-inbred populations[50]. We specify the covariance matrix using the realized relationship matrix $K = XX^T/p$, where p is the number of SNPs, and X is the $n \times p$ genotype matrix with entries x_{ij}. This formulation can be derived from a polygenic model of the phenotype in which all SNPs help explain variance in the phenotype, and the contributions of individual SNPs are *i.i.d.* normal[50,53,59]. Note that this formulation implicitly assumes that the non-additive genetic contributions to the phenotype are negligible, which is often a reasonable assumption in outbred populations[58]. A useful feature of the realized relationship matrix in our mouse population is that it has a close interpretation to kinship coefficients due to the historical founder bottleneck; that is, identity-by-state (IBS) should closely recover identity-by-descent (IBD). We note that we cannot rule out the possibility of unknown genetic contamination from outside mice, which would degrade the IBS-IBD connection.

4.1.5 Revised LMM approach to address proximal contamination. A critical consideration for LMM-based association mapping is that including a genetic marker in the genetic similarity matrix K can deflate the test statistic for this marker, leading to a loss of power to detect a QTL. This phenomenon has been called proximal contamination[50]. The basic intuition for this loss of power is that including the candidate SNP in K makes the null log-likelihood higher when the SNP helps to explain variance in the trait. In human GWAS with smaller sample sizes, this loss in power is expected to be small[51]. In our study, however, we expected that proximal contamination would have a larger impact on QTL detection because of the extended patterns of linkage disequilibrium and higher effect sizes.

The general strategy to avoiding this loss of power is to remove SNPs from K that are in close proximity to the SNP being interrogated[50]. We adopted a variation of this strategy by analyzing each chromosome separately; using a realized relationship matrix estimated using all SNPs outside the chromosome[57]. Therefore, in the QTL mapping we used a slightly different polygenic covariance matrix for each chromosome[6].

To illustrate the benefits of avoiding proximal contamination in this way, we assessed support for soleus weight QTLs using three different methods (Supplementary Figure 13). The first method was based on a simple linear regression, and therefore did not attempt to correct for confounding due to hidden relatedness. Comparing the expected and observed association test p-values (Panel A), it appeared that there may be an inflation of small p-values caused by hidden relatedness in the CFW mice. This observation is supported by Panel B, where we observed a dramatic reduction in inflation after using the LMM-based approach to control for hidden relatedness. However, Panel C suggests that most of this correction for population structure is an artificial reduction in p-values caused by proximal contamination. Our method that addresses proximal contamination yields p-values that are broadly similar to the simple linear regression, suggesting that there was no systematic population structure in the CFW mice that could produce a large number of false positive associations in the traits we studied.

Still, the results shown in Supplementary Figure 13 raised two possible concerns. One possible concern is that our LMM approach that analyzes each chromosome separately might not adequately correct for population structure. However, this is unlikely because the marker-based estimates of pairwise relatedness that include all SNPs outside the chromosome should not be much different from the relatedness estimates that include all SNPs. Therefore, the ability to control for population structure should be similar for both LMM analyses. A second related concern is that our QTL mapping approach has not sufficiently reduced genomic inflation (Panel C), and in light of this we may need to use an additional correction of the test statistics, such as genomic control[60,61]. In response to this concern, we point out that a large degree of inflation is expected for traits such as soleus that exhibit strong association signals in loci with long-range LD patterns; a large number of SNPs in such loci will have more extreme test statistics than we would expect under the null, and this provides an explanation for the overall inflation we observed.

4.1.6 Determining significance of *p*-values. To determine whether or not a *p*-value at a given SNP constitutes significant support for a QTL, we calculate a threshold for significance by estimating the distribution of minimum *p*-values under the null hypothesis (Supplementary Figure 21), then we took the threshold to be the $100(1 - \alpha)$th percentile of this distribution, with $\alpha = 0.1$. A common approach to estimating the null distribution is to randomly permute the phenotype observations while keeping the genotypes the same (e.g., Broman and Sen[62]). However, such a procedure is technically not appropriate here because it fails to account for the lack of exchangeability among the samples[58]. For this reason, various permutation tests have been developed that account for the effect of familial relationships or population structure[42,54,57,63,64]. However, based on our discussion above about population structure, cryptic relatedness does not appear to have a major impact on association tests, hence a naive permutation test that assumes independence of the samples should provide an acceptable means to estimate the rate of false positive associations[58]. We tried different numbers of permutation replicates, and found that our estimates of the $\alpha = 0.1$ percentile stabilized after 1000–2000 replicates, similar to what others have previously found for QTL mapping in experimental crosses[65].

Our final calculation for the *p*-value significance threshold is approximately 2×10^{-6} for all phenotypes, or a $-\log_{10}p$-value of about –5.7 (Supplementary Figure 21; At $\alpha = 0.05$, or the 95th percentile of the null distribution, the *p*-value threshold based on the same permutation replicates would be approximately 8.8×10^{-7}, or a $-\log_{10}p$-value of about –6.1.) To provide an independent validation of this significance threshold, which is based on the assumption that the samples are exchangeable, we compare it against a simple Bonferroni correction. The Bonferroni correction often leads to an overly stringent significance threshold because it ignores correlations between the markers (*i.e.* the association tests are not independent). Therefore, we reduce the effective number of tests in the Bonferroni correction by first pruning the set of markers so that no pair of markers has a (Pearson's) correlation coefficient greater than $\rho = 0.95$. This reduces number of tests from about 80,000 to 64,000. Using this as our number of tests, *p*-values of 1.6×10^{-6}, or $-\log_{10}p$-values of –5.8, yield Bonferroni-adjusted *p*-values of 0.1, so the Bonferroni threshold is only slightly more stringent than our significance threshold based on a naive permutation test (*p*-values of 7.8×10^{-7} yield Bonferroni-adjusted *p*-values of 0.05). Note than in many cases the observed associations vastly exceeded any of these thresholds (Supplementary Table 2).

4.1.7 Assessment of multiple QTLs. For several phenotypes, the initial genome-wide scan yielded support for QTLs that are sometimes nearby each other on the same chromosome. However, testing each marker one at a time cannot tell us whether we have support for multiple QTLs. To address this question, we took a stepwise regression approach, in which we included the SNP with the lowest *p*-value as a covariate and then recomputed *p*-values for all other SNPs on the same chromosome, using the LMM in the same way as before. We repeated this procedure for each of the QTLs identified in the initial genome-wide scans.

Occasionally, including a QTL in the linear model of the phenotype increased support for association at other loci. The most striking example of this was the soleus QTL near 9.7 Mb on chromosome 13 (Supplementary Table 2). Initially, we obtained modest support for an association at this locus, with a *p*-value of 1.57×10^{-5} for SNP rs46826545, which was well below our significance threshold. To assess support for multiple QTLs on the

chromosome, we included the genotype of the most strongly associated SNP, rs222759307. In this second analysis, we observed increased support for association at many SNPs on chromosome 13 (Supplementary Figure 14), including rs46826545, with a revised *p*-value of 2.83×10^{-6} that approaches our significance threshold.

4.1.8 Novel SNPs and QTLs

Among the SNPs included in our QTL mapping analyses, 13,450 were novel SNPs discovered by our GBS pipeline. 12 of these SNPs had strong associations with one or more traits, such as tibia length, bone-mass density and anxiety-like behavior. Of course, these associated SNPs were often correlated with other SNPs that were not novel. When multiple SNPs at a locus had very low *p*-values, we always chose the non-novel SNP (with an rsID) so that the association was more easily cross-referenced with mouse genetics databases. As a result, no novel SNPs are listed in Supplementary Table 2. For further discussion of the discoveries at novel SNPs, see section 4.3.6 “Novel SNPs and cis-eQTLs”.

4.1.9 Effect of QTL alleles across time

For some of the behavioral traits we studied, phenotypic data was acquired over discrete time bins. In Supplementary Figure 18, we explored the effects of SNPs implicated in these traits (see Supplementary Table 2) at 5-minute intervals to determine whether the QTLs exerted their effects in a uniform or time-specific manner. Notably we observed some alleles that appeared to influence behavior only early in the test (Supplementary Figure 18G, H) whereas others appeared to exert their influence only later in the test (Supplementary Figure 18B).

4.2 Estimating the proportion of phenotypic variance explained by available genotypes

A quantity that is often used to summarize the genetic contribution to a phenotype is the narrow-sense heritability, h^2, defined as the maximum variance in a given phenotype that can be explained by a linear combination of the allele counts[66]. If the available genetic markers completely capture, or tag, all the genetic variants that contribute to variation in the phenotype, then h^2 estimated using available SNPs should closely recover the true narrow-sense heritability. Our SNPs do not completely tag all casual variants, so our estimates of h^2 underestimated the true narrow-sense heritability; our estimates of h^2 therefore represent the proportion of variance in the phenotype that is explained by available genotypes, sometimes called the SNP heritability[67].

The most commonly used approach to estimating h^2 using marker data is based on the assumption that all genetic markers make a small contribution to variation in the trait, and that these contributions are normally distributed with the same variance[68–70]. This polygenic model (i.e. ridge regression[71]) is equivalent to expressing the covariance of the phenotype measurements as $Cov(Y_1, \ldots, Y_n) = \sigma^2 H$, where $H = (I + \sigma_a^2 K)$, I is the $n \times n$ identity matrix, K is the $n \times n$ realized relatedness matrix, σ_a^2 is the variance of the additive genetic effects, and σ^2 is the variance of the residuals. Under this formulation, σ_a^2 represents the relative contribution of the additive genetic variance, and we can use this parameter to provide an estimate for h^2:

$$h^2 = \sigma_a^2 \, s_a / (\sigma_a^2 s_a + 1)$$

where s_a is the mean sample variances of all the available SNPs, or the mean of the diagonal entries of K assuming that the columns of X are centered so that each of the columns has a mean of zero.

Instead of fitting the model parameters σ^2 and σ_a^2 using the maximum likelihood or REML estimate (e.g.[69]), we analytically integrated out the residual variance parameter σ^2 assuming the standard non-informative prior $p(\sigma^2) = Z_0/\sigma^2$, where Z_0 is a constant ensuring that the probability density integrates to 1, then we evaluated the likelihood$w(h^2) \equiv p(y \vee X, \sigma_a^2) = Z_1/\sqrt{y^T H^{-1} y H}$ for different choices of h^2, in which σ_a^2 in this

expression is a function of h^2, $|A|$ is the determinant of matrix A, and Z_1 is another constant ensuring that the probability density integrates to 1. One advantage of this approach is that it allowed us to easily quantify uncertainty in our estimate of h^2 without making additional assumptions. If there were any covariates included in the linear model of the phenotype, we analytically integrated out the linear effects of these covariates assuming non-informative (flat) priors for the regression coefficients, following the calculations described in Chipman[72].

If we evaluate the likelihood over a regular grid of values for h^2, then the weights $w(h^2)$ are proportional to posterior probabilities of h^2 *assuming a uniform prior for* h^2 [53]. Therefore, we used these weights to calculate, for example, the posterior mean estimate of h^2. To obtain posterior quantities, we evaluated the weights at equally spaced grid points $h^2 = 0.01, 0.02, \ldots, 0.99$. We assessed uncertainty in the estimates of h^2 by calculating the 95% posterior credible interval[73], which we defined as the smallest contiguous interval about the posterior mean containing 95% of the posterior mass. Note that this credible interval is at the same resolution as the grid points. Also note that the credible interval is not necessarily symmetric about the mean; see [Gelman *et al.* [73]] for alternative definitions of credible intervals.

Although the accuracy of the polygenic estimate of h^2 hinges on the validity of the assumption that all additive genetic effects are small, in practice this estimate is quite robust to deviations from this assumption[68]. In cases where we identified QTLs in the genome-wide mapping that explained a large proportion of variance in trait (e.g. testis weight), we easily improved the accuracy of the estimate by first removing the additive effects of these QTLs from the polygenic model. Thus, in traits for which individual variants explained a substantial portion of variance in these traits, our estimate of h^2 represents the proportion of variance explained by the remaining available SNPs, plus the proportion of variance explained by the selected variants with large effects.

Finally, we remark that this treatment only applies to continuously-valued phenotypes. Thus, our estimates of h^2 for the binary trait abnormal BMD should only be considered approximate calculations. More precise estimates can be obtained using modifications to this approach based on, for example, the liability threshold model[74,75].

4.3 Mapping expression QTLs (eQTLs)

4.3.1 Quality control and filtering

In the first step of our eQTL analysis, we excluded genes with low expression and genes with no variability. In each tissue, we computed the mean RPKM values for each gene. Genes which had a mean RPKM less than 1 were discarded. Although the threshold is low and would result in an increase in computation time, we do not expect the inclusion of some low expression genes to affect our ability to detect eQTLs. Further, we removed genes that showed no variability in their expression levels across samples. We interpret this as a technical artifact since we expect some biological variance across samples. After filtering for both low expression and lack of variability, we retained 43,414 genes for eQTL analysis (14,575 genes in the hippocampus, 14,476 genes in the prefrontal cortex and 14,363 genes in the striatum).

4.3.2 Quantile normalization

One of the assumptions of the LMM that we used to detect variants associated with the expression traits is that the residuals are normally distributed. One way to address this concern would be to treat the expression outcomes like the other quantitative physiological and behavioral traits in the study and remove outliers and/or identify transformations to ensure that the traits are normally distributed. This approach is not feasible for expression traits due to the large number of regressions performed in this analysis. We ensured that the expression traits satisfy the normality assumption of the linear mixed model by quantile normalizing the expression values of every gene in each tissue to a standard normal, *i.e.* we transformed the ranks of the expression traits in each gene to the quantiles of a *N(0,1)* distribution.

4.3.3 Correcting for unknown confounders

Various technical artifacts and biological processes might induce correlation in the expression levels of sets of genes in some individuals. This correlation structure in the data would result in a lack of power to identify eQTLs by increasing the residual variance. Several methods have been proposed for removing these effects without excluding any true signals[76,77]. As shown in Pickrell *et al.*[78] , these correlations can be accounted for using PCA.

In each of the three brain tissues, we calculated the principal components (PCs) of the *K x N* matrix of expression values, where *K* is the number of genes whose expression is measured and *N* is the number of samples that were included in the study. For each sample, we obtained the loadings for the PCs. We used linear regression to remove the effects of the first *p* PCs from the quantile normalized expression values. For each tissue, the number of PCs to correct for, *p*, was chosen as the number of PCs that resulted in the largest number of eQTL discoveries. The PC-corrected expression values are given by

$$y_{ij,new}^{k} = y_{ij}^{k} - \sum \hat{\beta}_{li}^{k} \, x_{lj}^{k}$$

where $y_{ij,new}^{k}$ and y_{ij}^{k} are the PC-corrected and the original quantile-normalized expression values for the *i*-th gene in sample *j* in tissue *k*, $\hat{\beta}_{lj}^{k}$ is the estimated regression coefficient for the *l*-th PC for gene *i* and x_{lj}^{k} is the loading for the *l*-th PC on sample *j*. We repeated the quantile-normalization procedure on the PC-corrected expression values to ensure normality of these residuals. These quantile-normalized PC-corrected expression values were used for eQTL mapping.

4.3.4 eQTL mapping using LMM

Similar to our approach to mapping QTLs for the physiological and behavioral traits, we used an LMM to map eQTLs to account for any cryptic relatedness or residual correlations between samples that had not been accounted for by the PCs. In a given tissue, we fit the expression values for gene *i* using

$$y_{ij} = \mu_i + \beta_l x_{lj} + u_j + \epsilon_{ij}$$

Where y_{ij} is the expression level for the gene *i* in sample *j* in the tissue, μ_i is the mean expression of gene *i* in the tissue, u_i is the polygenic component for sample *i*, x_{lj} is the allele dosage of the alternate allele at SNP *l* for sample *j* and ϵ_{ij} is the residual error. The standard assumption is that the residual errors are *i.i.d* normal with variance σ^2, whereas the polygenic components, ($u_1, u_2, \ldots, u_n$) are distributed as a multivariate normal with variance-covariance matrix given by $\sigma^2 \lambda K$, in which *K* is the relatedness matrix estimated from the genotype data.

We fit the linear mixed model for each gene in each tissue using GEMMA. As we did with the models for behavioral and physiological phenotypes, we used the realized relationship matrix estimated from the GBS genotypes as an estimate of the correlation matrix *K*. In contrast to the physiological and behavioral phenotypes, we used the same *K* matrix for all the expression phenotypes irrespective of the gene's chromosome, *i.e.* we did not correct for proximal contamination when mapping eQTLs.

4.3.5 Choice of *cis*- window and significance thresholds

We used the p-value computed using the likelihood ratio test from GEMMA to report the strength of association between a SNP and an expression trait. We fit the LMM for all SNPs with MAF > 5% against the expression of each gene which survived our initial filtering. For the *cis*-eQTL analysis, we only analyzed SNPs that were in the 1Mb *cis*-region of the gene (within the span of the gene or 1 Mb upstream or downstream of it). The choice of 1Mb as the region for *cis*-association, was based on the window size that gave us the most number of *cis*-eQTLs (Supplementary Figure 19).

We permuted a set of 1000 genes and repeated the analyses to compute the significance threshold for each gene. In the permutation analysis, we treated the samples as independent, i.e. we dropped the polygenic component and fit only a linear model instead of a LMM. We computed the *p*-value for the likelihood ratio test under this linear model, testing each SNP for association with the permuted expression phenotypes. We calculated the significance threshold for each gene as the 5th percentile value in the distribution of the most significant results in each of the 1000 permutations for the SNPs that are included in the 1 Mb *cis*-region of the chosen gene. Over all three tissues, we identified *cis*-eQTLs for 6,045 genes, the number of significant *cis*-eQTL findings in each tissue and the overlap among the tissues is shown in Supplementary Table 3 and Figure 4B (in the main manuscript).

4.3.6 Novel SNPs and *cis*-eQTLs

We considered whether the novel SNPs identified by GBS were useful for finding eQTLs. Of the 6,045 significant cis-eQTL, 945 (15.6%) of them showed the strongest association with a novel SNP. Since 14% of the SNPs identified by GBS were novel, this result provided strong evidence that the novel SNPs played a meaningful role in our study. (Also see section 4.1.8 “Novel SNPs and QTL”).

4.3.7 Comparison of cis-eQTLs in CFW with cis-eQTLs from another study

We compared our findings with the eQTL results from Park *et al.*[79]; who used microarrays to measure gene expression in the HMDP. We replicated 20 of their top 100 genes in the hippocampus and 12 of their top 100 genes in the striatum. The incomplete overlap can be attributed to various factors including different alleles in the two populations, type 1 and type 2 errors in both studies, different techniques for measuring gene expression and environmental differences such as age and diet.

4.4 Identifying t*rans*-eQTLs

For the *trans*-eQTL analysis, we performed an analysis similar to the *cis*-eQTL analysis, wherein we included all the SNPs identified using GBS, instead of restricting ourselves to the 1 Mb *cis*-region around the gene. In each tissue, for each gene, we designated the SNP that had the best correlation with the gene expression as the *trans*-eQTL SNP for that gene. In order to alleviate the problem of *trans*-QTLs tagging the *cis*-eQTL signal for the same gene, we excluded the SNPs that lay within a 2 Mb *cis*-region of the gene (within the extent of the gene or 2 Mb upstream or downstream of it). We chose to use a 2 Mb window by empirically determining the window outside which the number of *trans*-eQTLs plateau (Supplementary Figure 20).

4.4.1 *trans*-eQTL significance thresholds

In each tissue, the significance thresholds for the *trans*-eQTLs were determined using the permutations for 1,000 randomly chosen genes. For each of these 1,000 genes, we randomly permuted the expression values and performed a genome wide scan for *trans*-eQTLs. We calculated the genome-wide *p*-value threshold for a significance level of 0.05 using the distribution of the *p*-values for the best associated SNP for each of these 1,000 genes. Note that we did not discard SNPs in the *cis*-region of the genes while computing the permutation *p*-value threshold, implying that the computed threshold is more conservative than its nominal significance level of 0.05. For a nominal significance level of 0.05, the permutation-based *p*-value threshold is 7×10^{-7}. The number of significant *trans*-eQTL findings in each tissue is given in Supplementary Table 3 and the transband plots showing the distribution of eQTLs across the entire genome for all the genes that were assayed are shown in Supplementary Figure 21.

4.4.2 *trans*-eQTL findings

Over all three tissues, we find 2278 genes that have a *trans*-eQTL in the genome at a permutation derived significance threshold of 0.05 (7×10^{-7}). Note that, while this permutation-derived threshold corrects for the number of SNPs that are tested (92,734 GBS SNPs), it does not account for the multiple testing burden due to the number of genes that are tested across the three tissues. Since we are testing 43,414 genes in total across the three tissues, we expect 2,170 significant findings at the 0.05 level. So a large fraction of our *trans*-eQTL findings are expected to be found under the null hypothesis. Supplementary Figure 22 shows the QQ plot from the *trans*-eQTL analysis in the three tissues as well as a QQ plot for the findings combined across all three tissues. The QQ plots show an excess of low *p*-values in the observed data, indicating the presence of true signals for some of our *trans*-eQTL findings.

4.5 Allele specific expression analysis

In addition to the traditional eQTL analyses detailed above, we performed an allele specific test to find genes with ASE eQTLs.

4.5.1 Genotyping using RNA-Seq reads

Since the RNA-Seq reads overlap with protein-coding sequences at a high level of coverage (~60X), we were able to use these reads to call genotypes at variant sites in the gene transcripts. To call genotypes, we used a procedure similar to the variant-calling pipeline for the GBS data.

We used the read alignments generated in the quantification step to call variants. For a given sample, we combined the aligned reads from all three brain tissues (in the cases where all three tissues had been collected from the same animal). We used the GATK Indel Realigner to correct for alignment errors caused by the presence of indels at the ends of short reads. Following indel realignment, we used the GATK Unified Genotyper to call genotypes at known polymorphic sites. We obtained the list of known polymorphic sites from the Wellcome Trust mouse genomes project. We recognized that this set of variants might not contain all the variation segregating in the CFW population. Since we did not have a genome-wide catalog of variants in this population, and the variants discovered by GBS are sparsely spread across the entire genome, we limited our genotyping to sites that were known to be variable in the 17 lab strains that were sequenced as part of the Wellcome Trust project.

Using the SNPs from the Wellcome Trust strains and limiting our genotype calling to exonic sequences in our gene models (Ensembl release 68, on mm10), we genotyped 94 samples that had RNA-Seq coverage at 325,422 variant sites. In contrast to the GBS data, the deeper coverage from RNA-Seq provided us with higher quality genotypes, so we did not use genotype imputation to estimate missing genotypes or to improve the accuracy of low-confidence genotypes. Any samples that had fewer than 10 reads covering the SNPs being genotyped were discarded. Furthermore, we filtered out genotypes that were assigned a quality score below 30. The resulting genotypes were used for the ASE analysis.

We only tested for ASE for a given tissue if we had 10 or more heterozygous (informative) mice. For the hippocampus, 38,492 SNPs in 6,439 genes passed our filters. For the striatum, 32,997 SNPs in 5,769 genes passed our filters. For the prefrontal cortex, 32,313 SNPs in 5,680 genes passed our filters. A little more than a quarter of the genes contained only one SNP that passed our filters and was thus used to test for ASE. The distribution of the number of SNPs that could be used to test for ASE for each gene is shown in Supplementary Figure 29.

4.5.2 Modelling ASE read counts

For genes which had at least one variant with 10 high confidence heterozygote calls, at each SNP in the gene, we modeled the read counts at both the alleles in each individual as binomially distributed with a proportion of reads from the reference allele *p*. To account for over-dispersion in the read counts due to unaccounted

confounding factors, we used a beta-binomial model with a different p_i for each sample, with a mean proportion of p. The likelihood of the data under the beta binomial model is shown below.

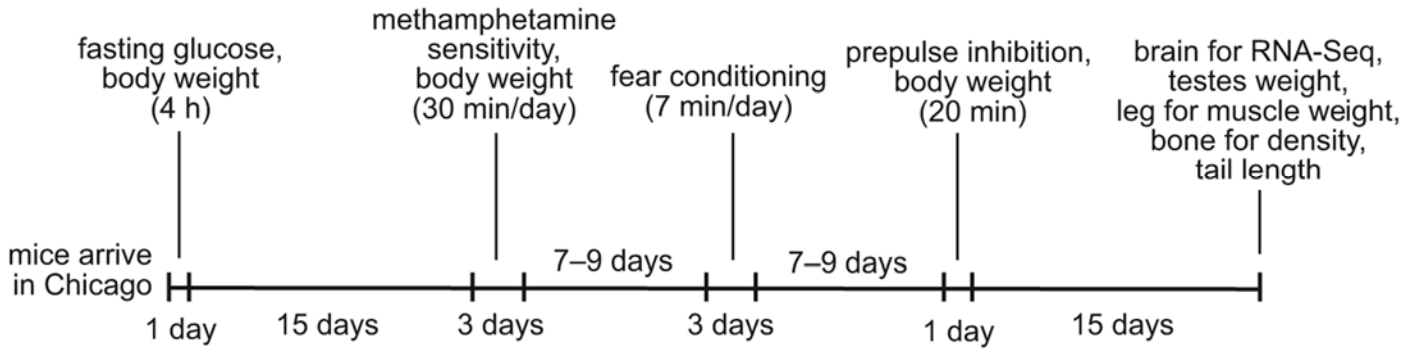
$$P(D \vee \alpha, \beta) = \prod_{i=0}^{n} P(D_i \vee \alpha, \beta)$$

$$P(D_i \vee \alpha, \beta) = \binom{n_i}{r_i} \frac{B(r_i+\alpha, n_i-r_i+\beta)}{B(\alpha,\beta)}.$$

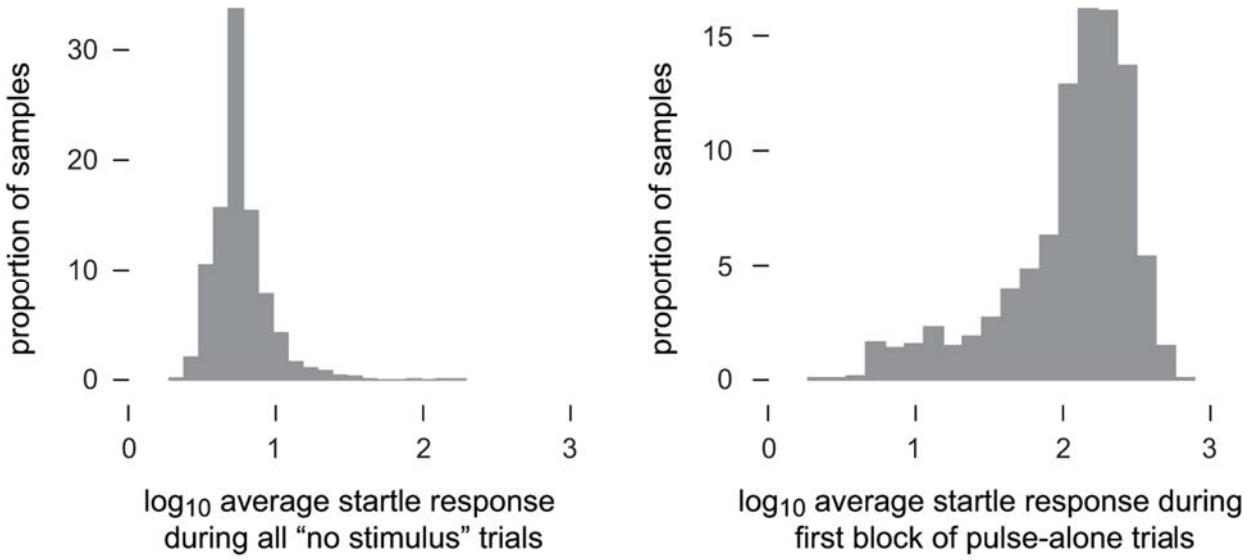
Here, D_i is the data at the SNP for individual i, which consists of n_i, the total number of sequenced bases at the SNP and r_i is the number of reference bases at the SNP, B() is the beta function, α and β are the parameters of the beta distribution that define the proportion p and the extent of over-dispersal in this proportion. Under the beta binomial model, p comes out as α/β and the variance of p is given by $\alpha\beta/(\alpha + \beta)^2(\alpha + \beta + 1)$. This allows us to account for the inter-individual variation in the proportion of reads from the reference base.

4.5.3 ASE significance threshold and findings

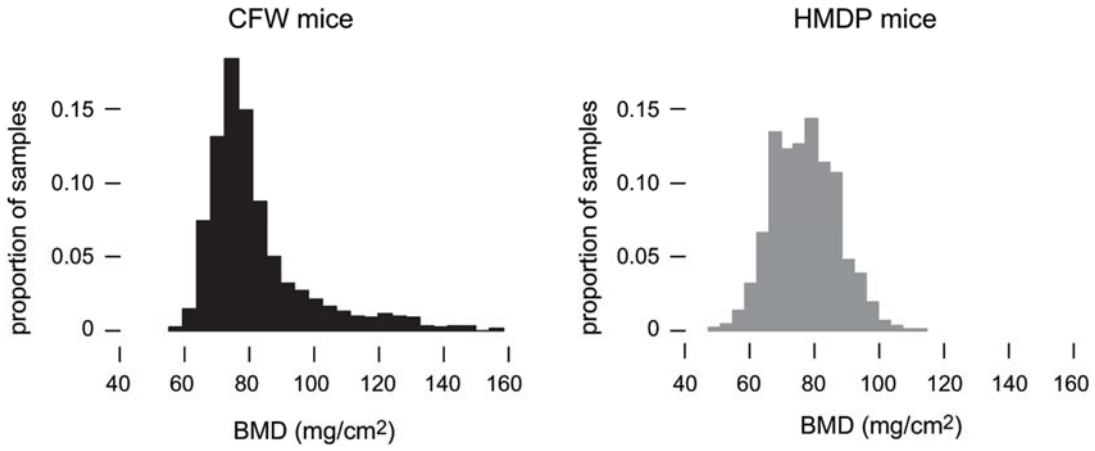
We used a likelihood ratio test to determine if the mean proportion of reference alleles across all the samples was 0.5 as would be expected in the case of no ASE, i.e., under the null model, we fixed the mean proportion p to 0.5 (H_0: $\alpha = \beta$) and found the maximum likelihood fit. Under the full model, we found the maximum likelihood fit for all the parameters (H_0: $\alpha \neq \beta$). The difference of the log-likelihoods from the two different models is distributed as a χ^2 with one degree of freedom. We used the asymptotic distribution to identify genes where the mean proportion differed significantly from 0.5. If at least one variant in the gene has a significant difference (p-value threshold = 1×10^{-6}, $\alpha \approx 0.05$) in the expression of its two alleles, we declare the gene to display ASE. Using this method in the three different tissues, we found 655 genes that had at least one SNP that showed a signal of imbalance in the expression of the two different alleles at the SNP. The number of findings for each tissue is shown in Supplementary Table 3.



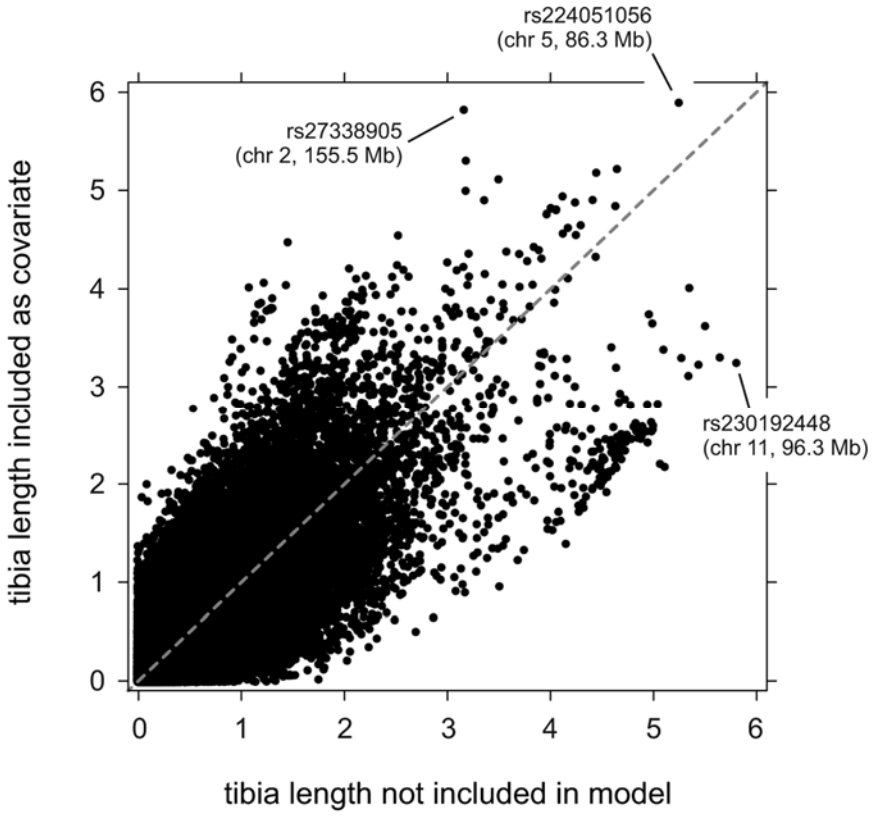
Supplementary Figure 1: Timeline for physiological and behavioral phenotyping. Times in parentheses give the amount of time a mouse was subjected to testing.



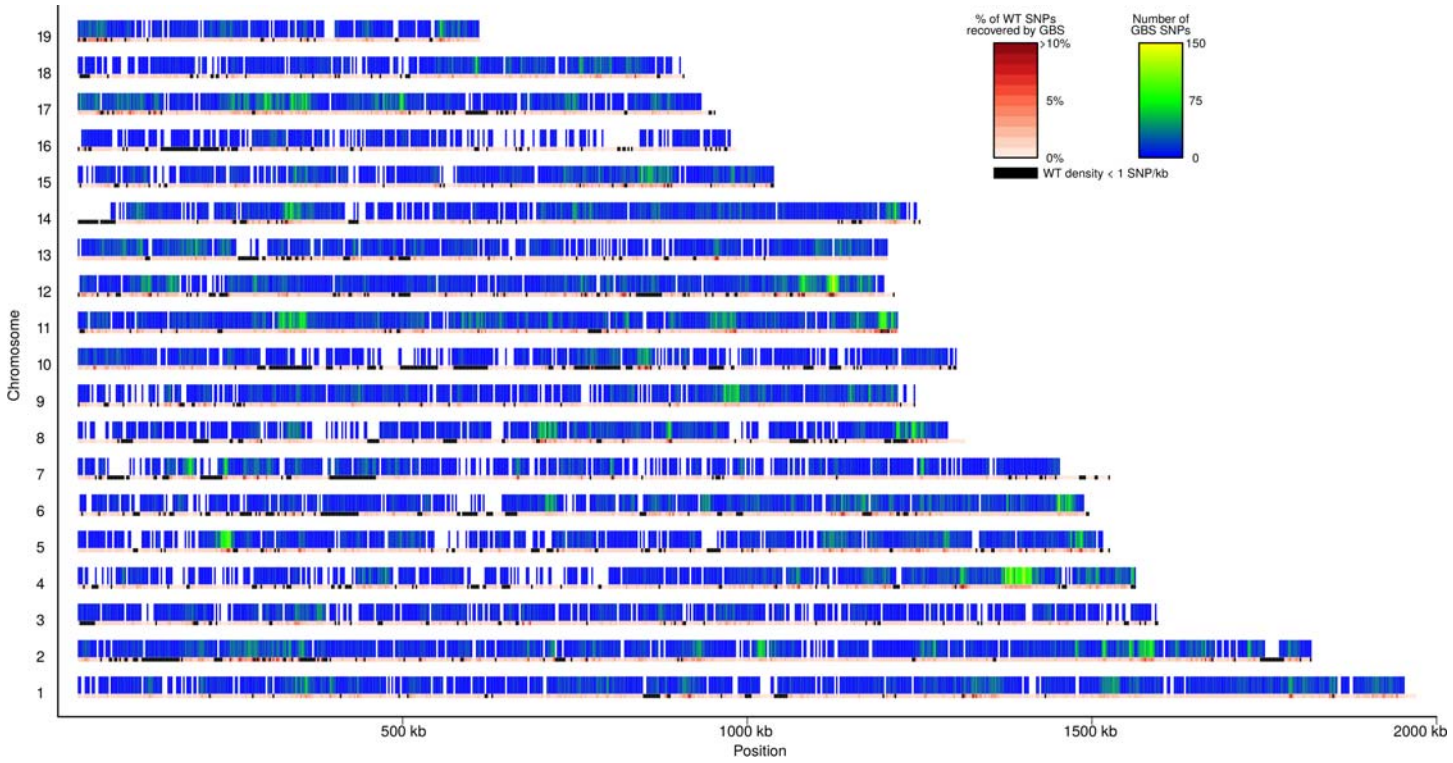
Supplementary Figure 2: Comparison of startle response to pulses against response during "no stimulus" trials. *Left-hand panel:* Distribution of average startle response during 8 "no stimulus" trials. *Right-hand panel:* distribution of average startle response during the first block of pulse-alone trials. Data shown for 1148 samples. Comparison of these histograms suggests that samples falling within the tail of the startle response distribution (in the right-hand panel) are not responding to the startle cues, possibly because they are deaf.



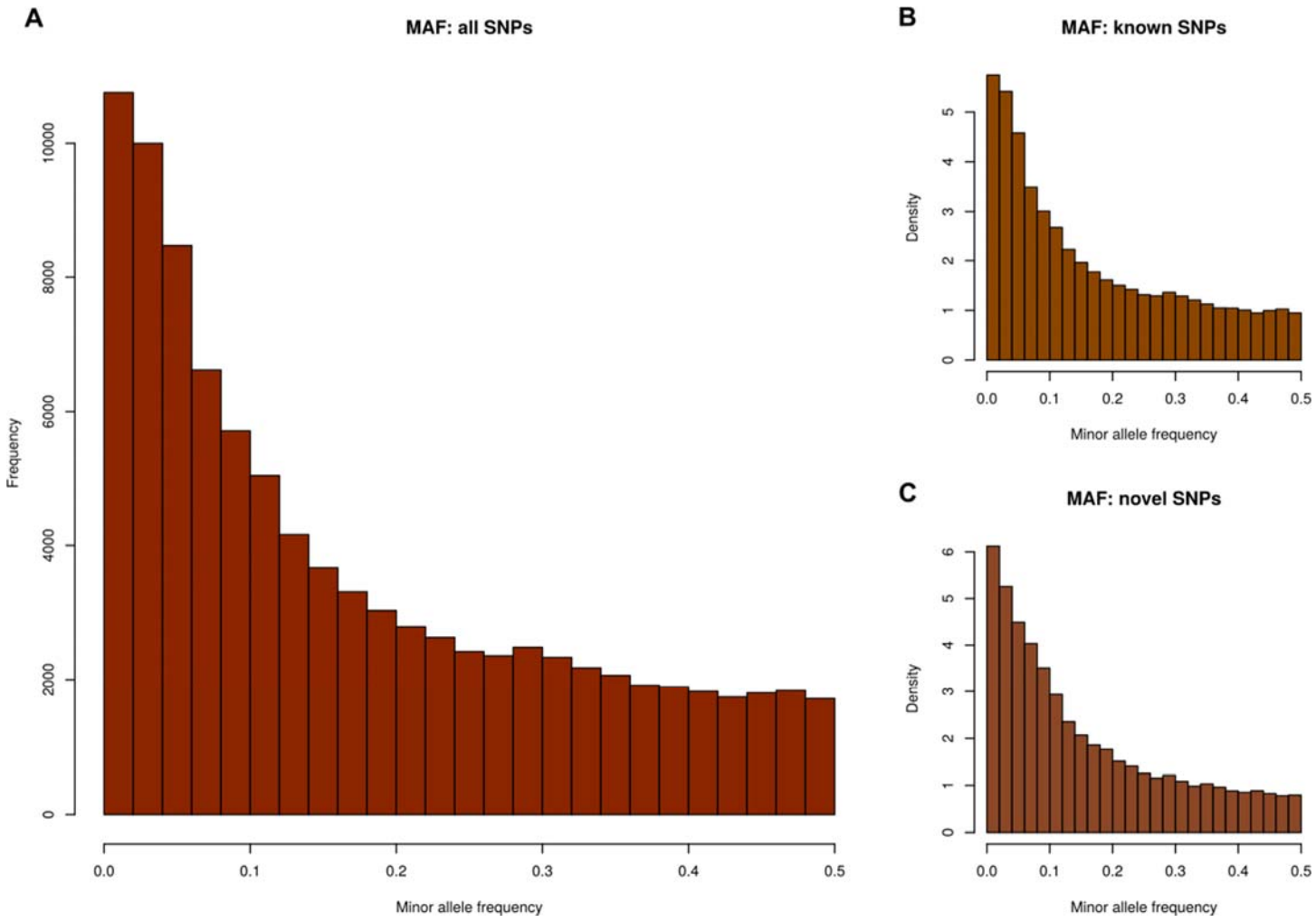
Supplementary Figure 3: Comparison of bone-mineral density (BMD) in CFW mice and mice from Hybrid Mouse Diversity Panel. Left-hand panel shows distribution of areal BMD measured in femurs of 1,057 CFW mice. Right-hand panel shows distribution of areal BMD measured in femurs of 878 mice from 97 HMDP strains[80]. Compared against the HMDP mice, a substantial fraction of the CFW mice exhibit abnormally high BMD.



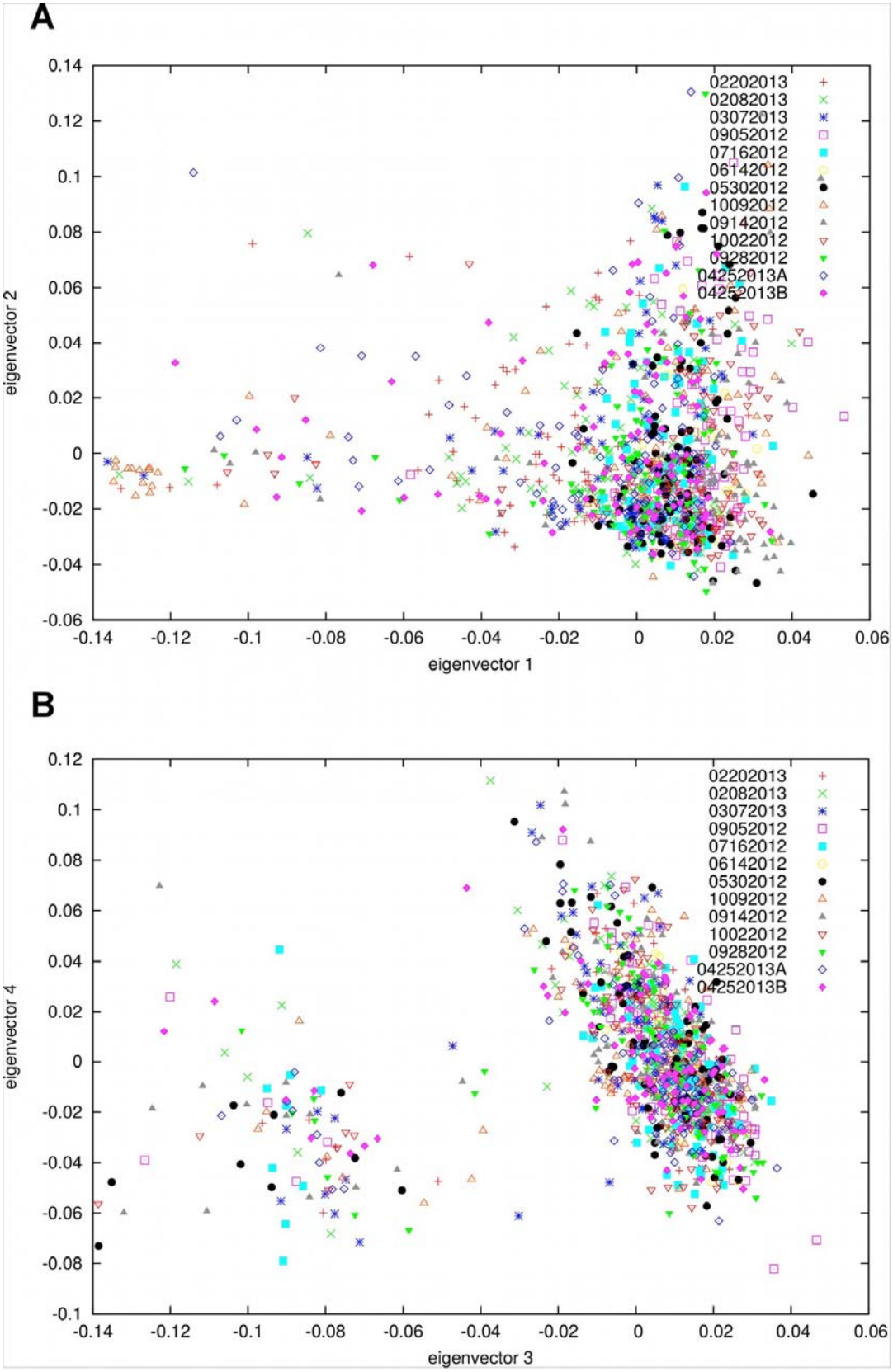
Supplementary Figure 4: Support for TA QTLs with and without conditioning on tibia length. Scatterplot shows $-\log_{10}$ *p*-values for all SNP association tests with TA muscle weight when tibia length is included in the regression model (vertical axis), and when tibia length is not included in the model (horizontal axis).



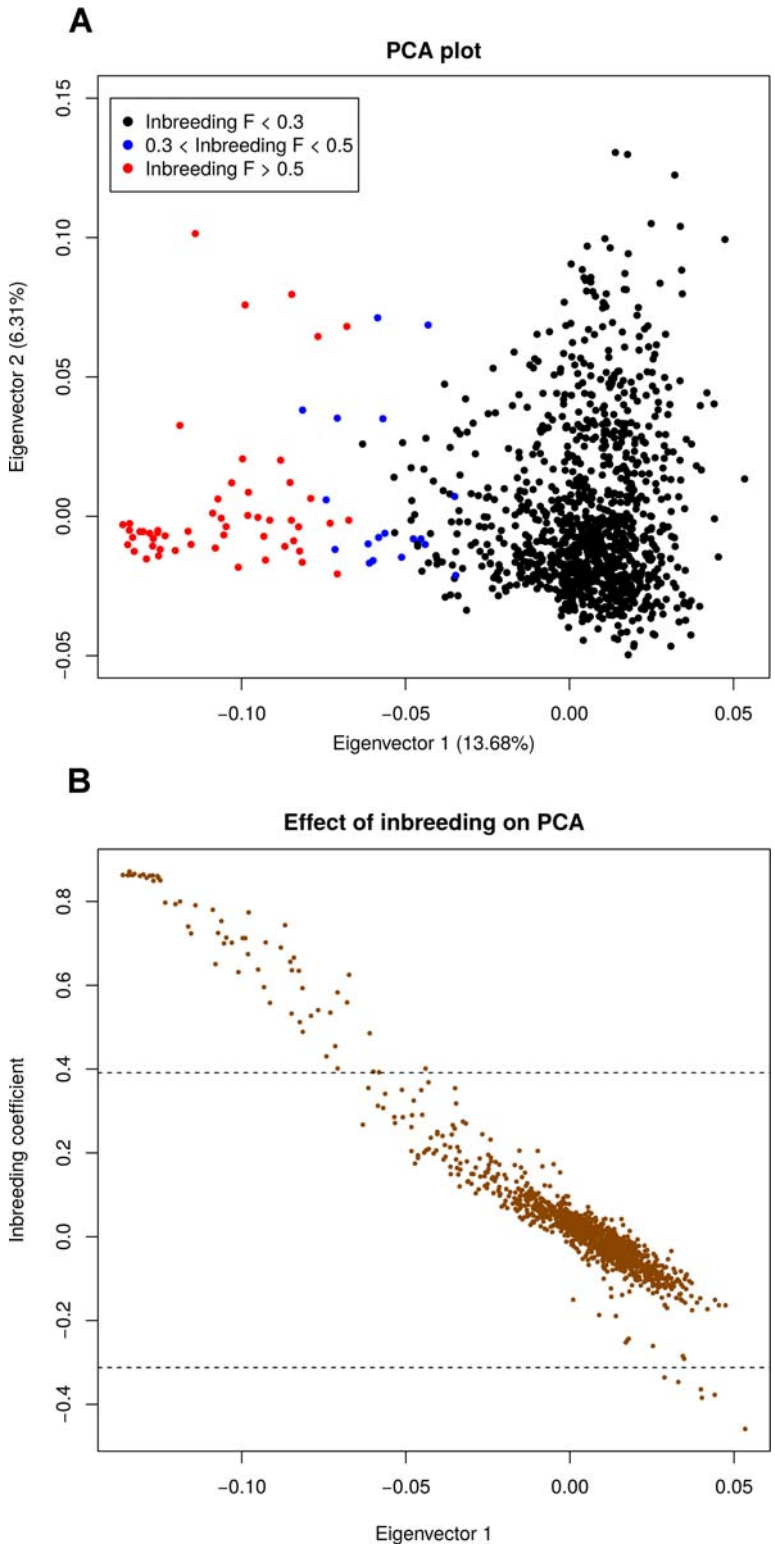
Supplementary Figure 5: Density of SNPs discovered by GBS in the CFW population. The colored bars represent the number of SNPs discovered by GBS in the CFW population. Each bar represents a 250-kb window. The color of the bar shows the number of SNPs that were found in that 250kb window, with blue signifying no SNPs to yellow signifying 150+ SNPs. Areas of the genome that do not have any GBS coverage, either due to lack of restriction cut sites or inability to map due to repeats in the reference genome, are shown in white with no bars at all. The smaller track underneath the colored bars gives the proportion of Wellcome Trust SNPs that were discovered by GBS in the same window. In windows where the density of Wellcome Trust SNPs is less than 1 SNP/kb, the window is marked using a black bar. For example, the beginning of chromosome 14 does not have any GBS SNPs discovered in it, but this region also has a low density of Wellcome Trust SNPs.



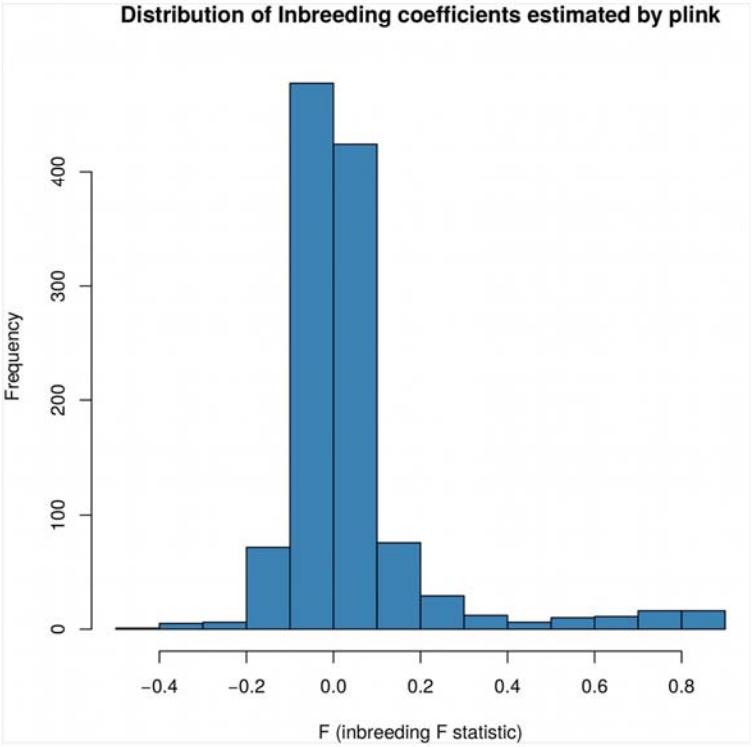
Supplementary Figure 6: Minor allele frequency distribution. The minor allele frequency distribution of the variants discovered using GBS in the CFW mice. Panel (A) shows the distribution for all SNPs discovered by GBS, whereas panels (B) and (C) show the minor allele frequency distribution for known and novel SNPs among the SNPs discovered by GBS.



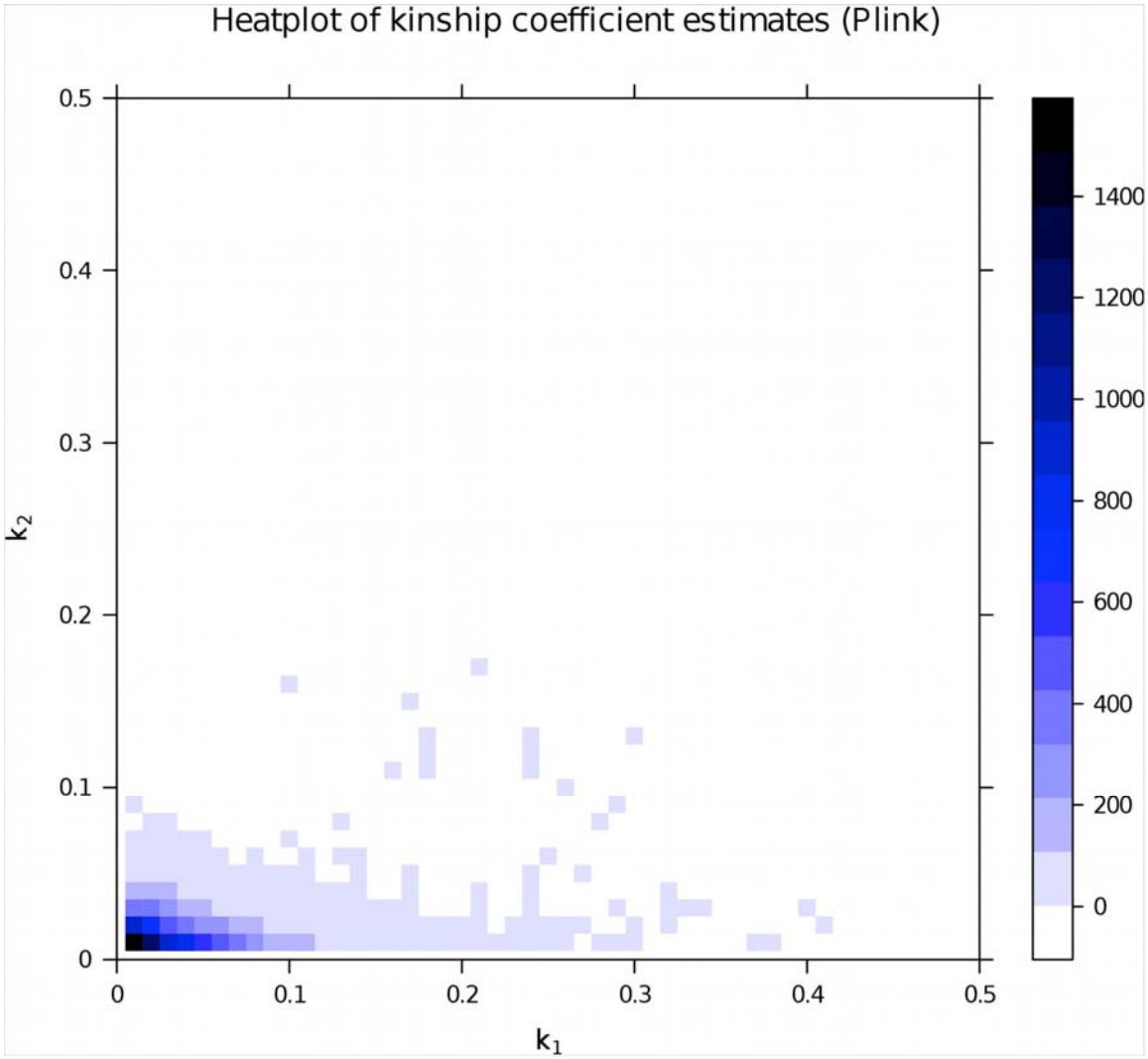
Supplementary Figure 7: Principal components analysis of the genotype dosages. Panels A and B show the first 4 principal components (PCs) of the genotype dosage matrix. The different symbols and colors indicate different shipment batches in which we received the CFW mice from vendors. None of the 4 PCs are correlated with inclusion in a batch.



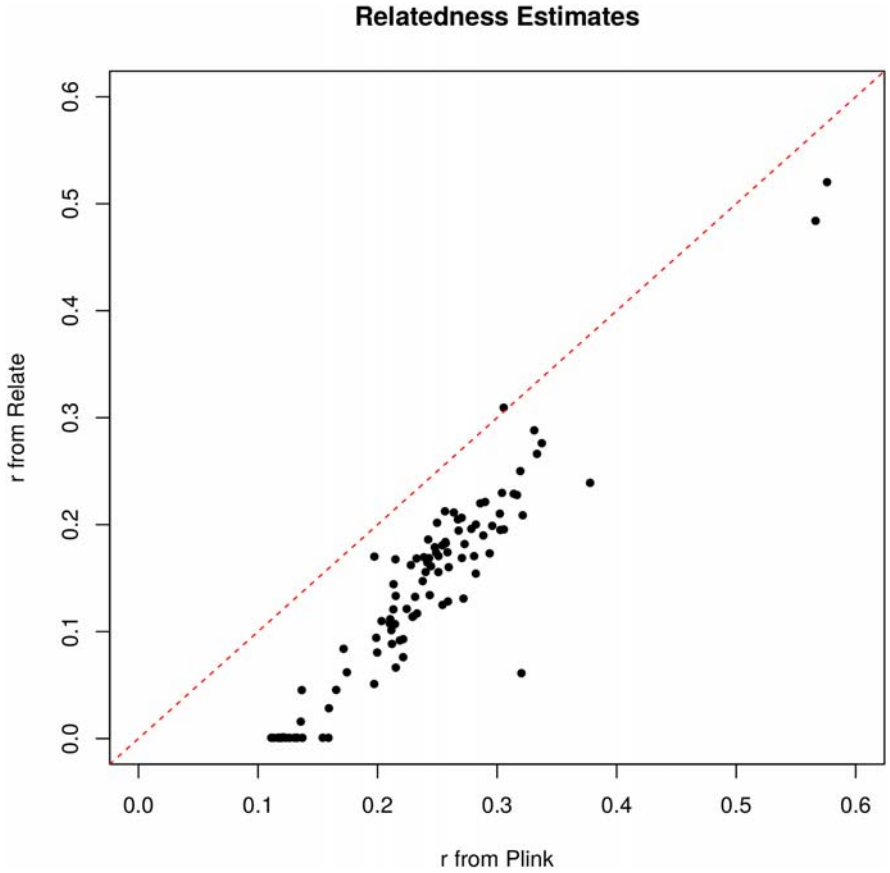
Supplementary Figure 8: Inbreeding coefficient vs PCs. Panel (A) shows the first two PCs of the genotype dosage matrix. The samples are colored by their inbreeding coefficient. Panel (B) shows the relationship of the first PC to the inbreeding coefficient.



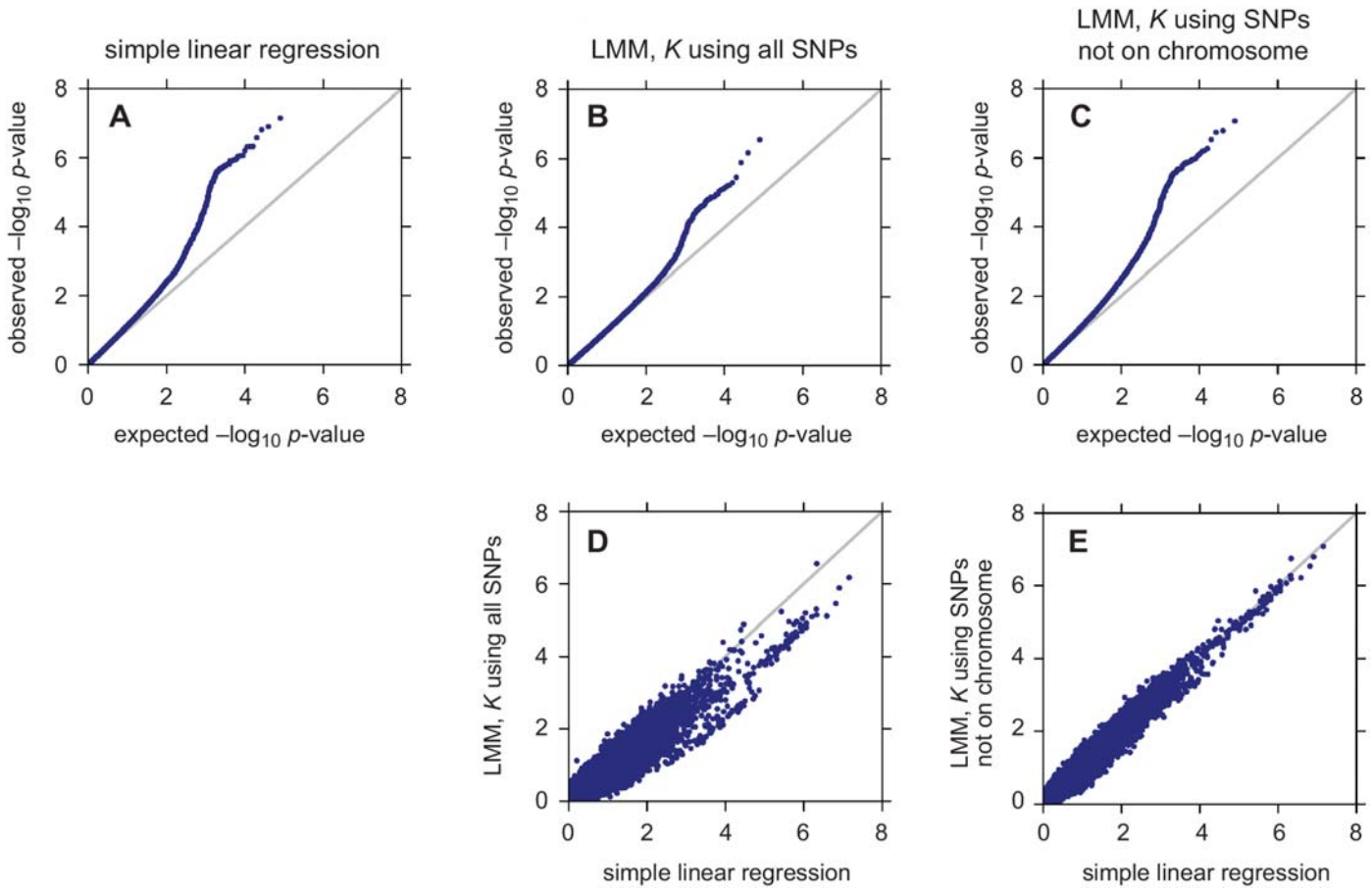
Supplementary Figure 9: Distribution of inbreeding statistic. The plot shows the distribution of the inbreeding statistics computed using *plink*. Most individuals have inbreeding coefficients near 0, indicating that the mating scheme used by Charles River has successfully maintained an outbred population.



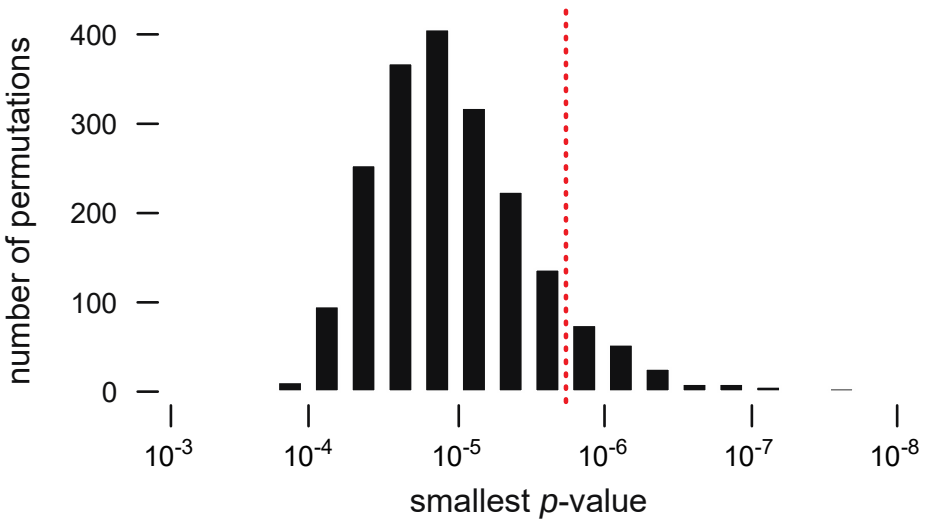
Supplementary Figure 10: Distribution of identity-by-descent coefficients. IBD coefficients k_1 and k_2 for all pairs of samples are shown as a density plot. Both IBD coefficients are rounded to the nearest 0.01. Most of the mass lies near the origin, indicating that most samples are not closely related.



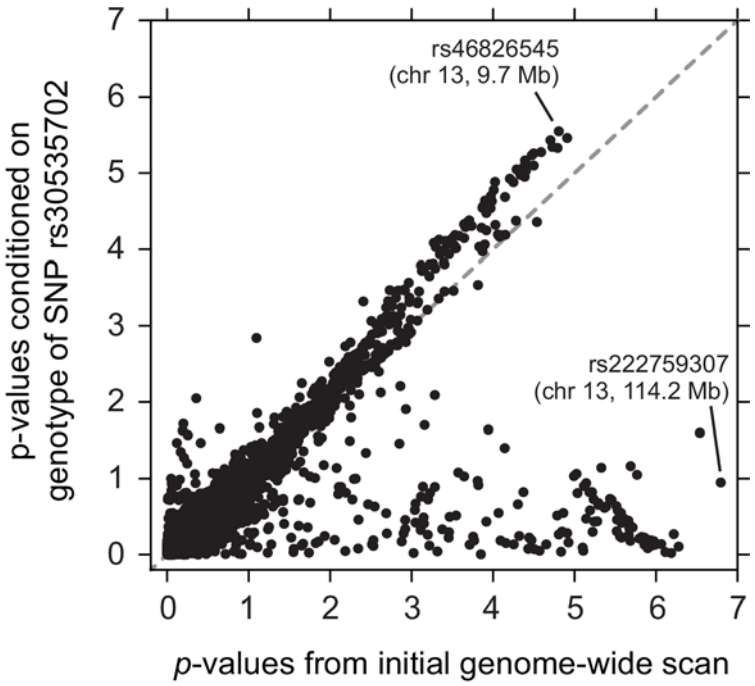
Supplementary Figure 11: Relatedness estimates from plink and relate. The relatedness estimates from *plink* and *relate* are plotted against each other for 100 pairs of mice with the highest estimates of relatedness. This plot shows an overall agreement between the two methods with some overestimation from *plink*.



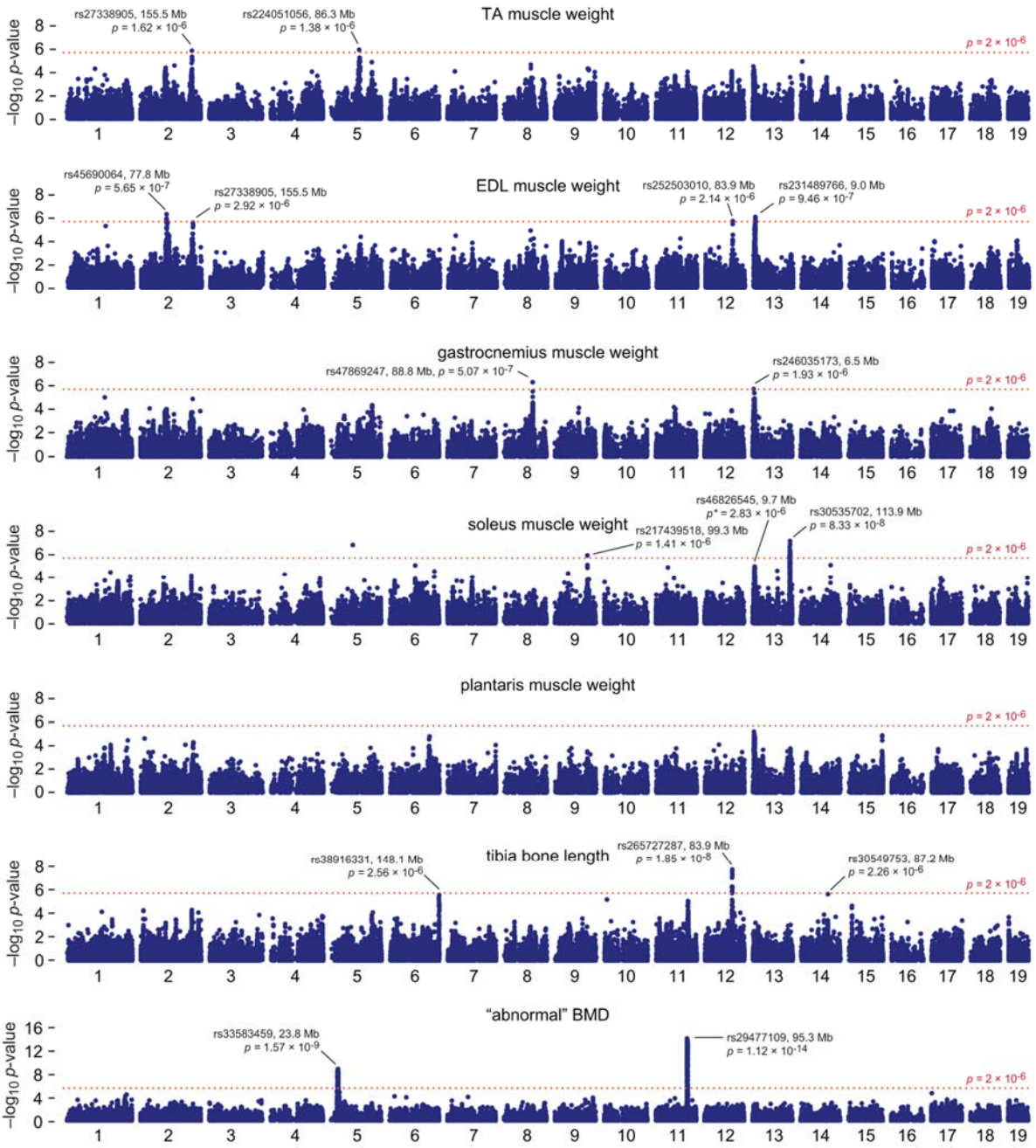
Supplementary Figure 12: Comparison of support for soleus weight QTLs using different methods to account for population structure. Panels A–C are genomic inflation plots[47,61] comparing quantiles of the calculated *p*-values (vertical axis) against the expected quantiles under the null (uniform) distribution of *p*-values. These are *p*-values for soleus weight QTLs. The *p*-values are calculated using three different methods: a simple linear regression that does not attempt to correct for population structure (Panel A); an LMM that includes all SNPs in the relatedness matrix *K*, and therefore does not address proximal contamination (Panel B); and an LMM that addresses proximal contamination by specifying a separate *K* for each chromosome (Panel C). Panels D and E show scatterplots that directly compare the *p*-values obtained using the three different QTL mapping approaches.



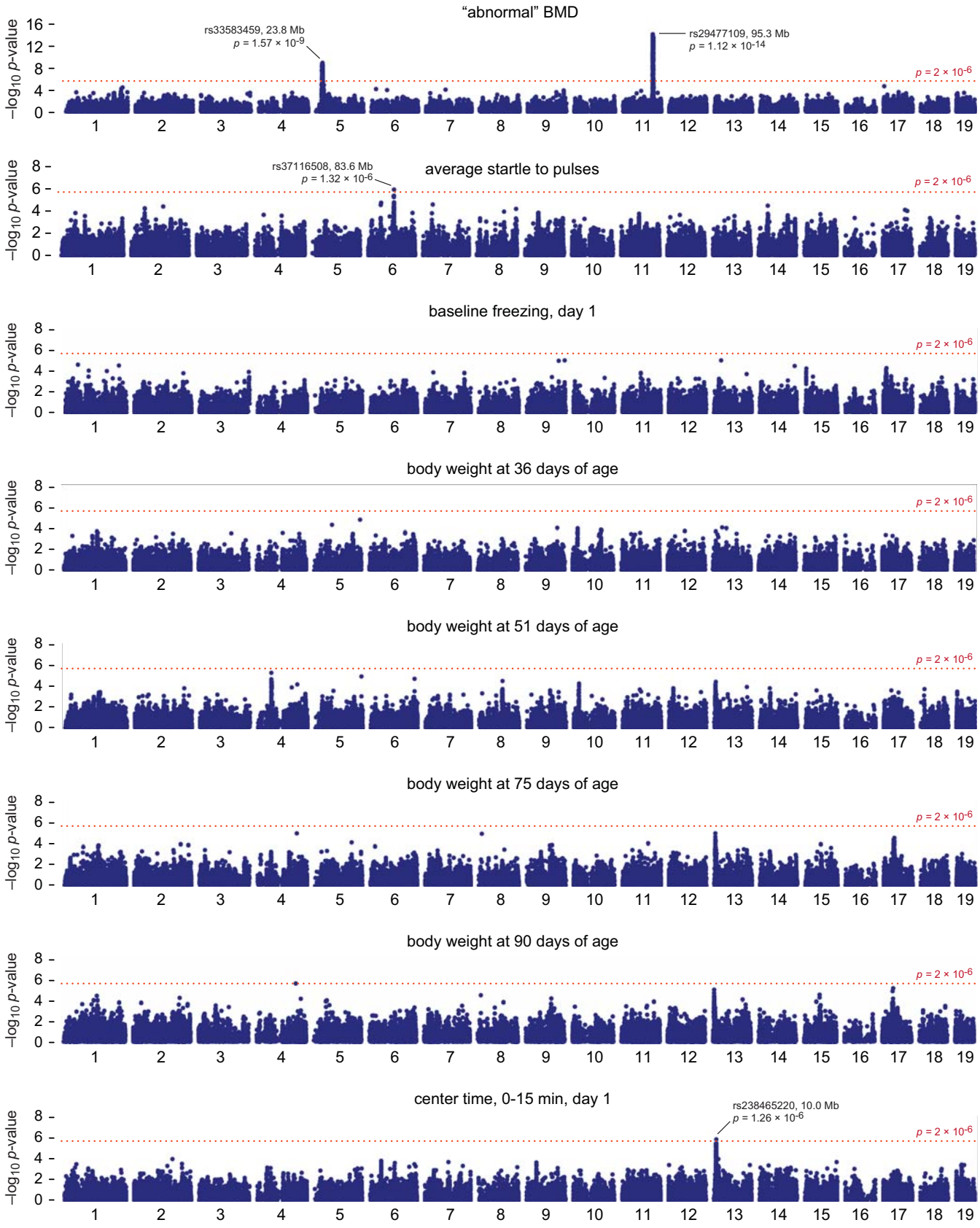
Supplementary Figure 13: Results of permutation analysis for QTL mapping. This plot summarizes the distribution of minimum *p*-values from mapping QTLs in 1,000 permuted data sets. The minimum *p*-value in each permuted data set is the smallest *p*-value from the 92,734 SNPs tested for association. This permutation test is meant to simulate the distribution of *p*-values under the null hypothesis. The 90th percentile of this distribution, which is approximately 2×10^{-6}, is depicted here by the dotted red line.



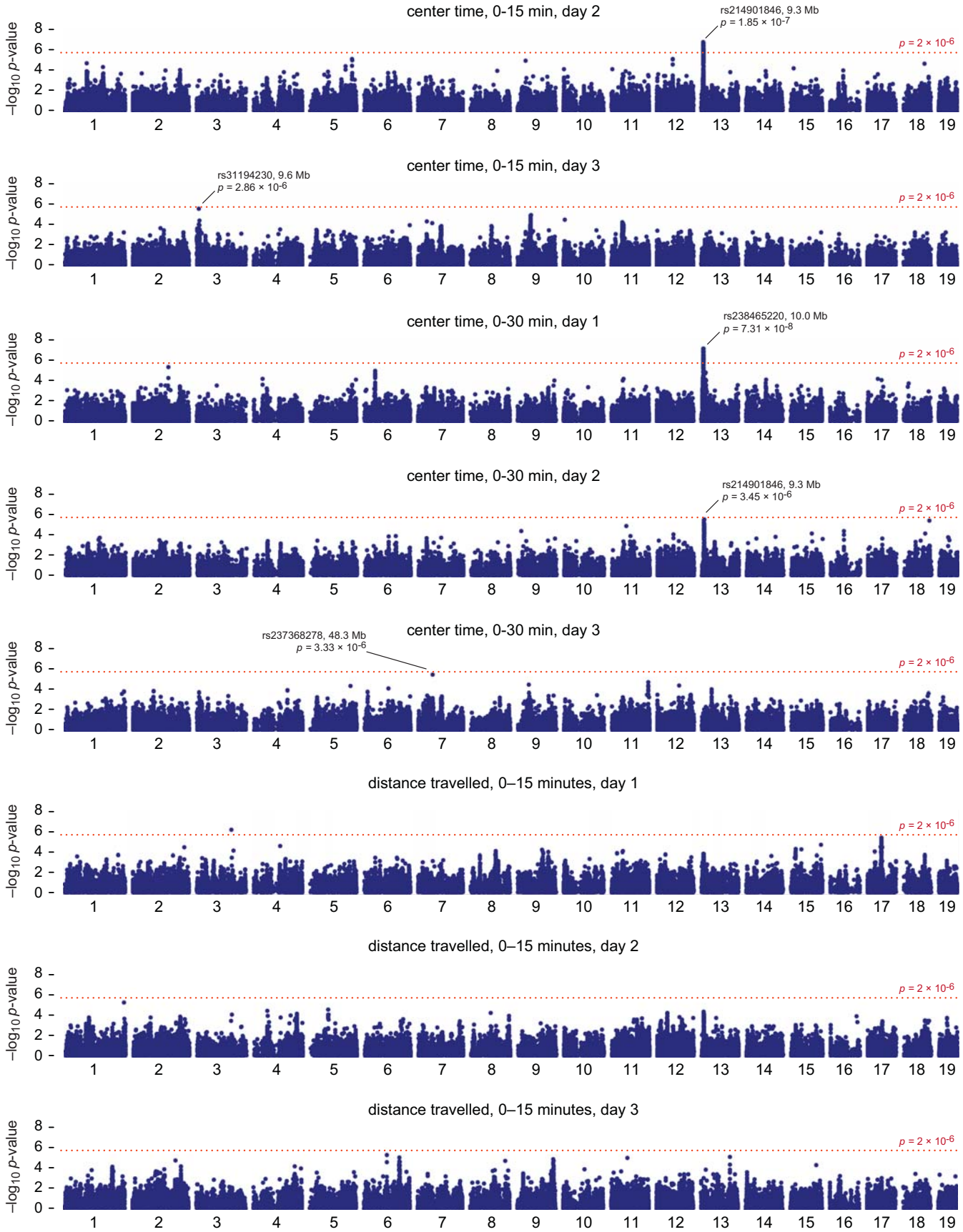
Supplementary Figure 14: Support for soleus QTLs on chromosome 13 before and after conditioning on SNP rs30535702 (chromosome 13, 113.9 Mb). The horizontal axis shows *p*-values obtained in the initial QTL analysis, and the vertical axis shows *p*-values obtained after including the genotype of SNP rs30535702 as a covariate in the linear regression of soleus muscle weight.

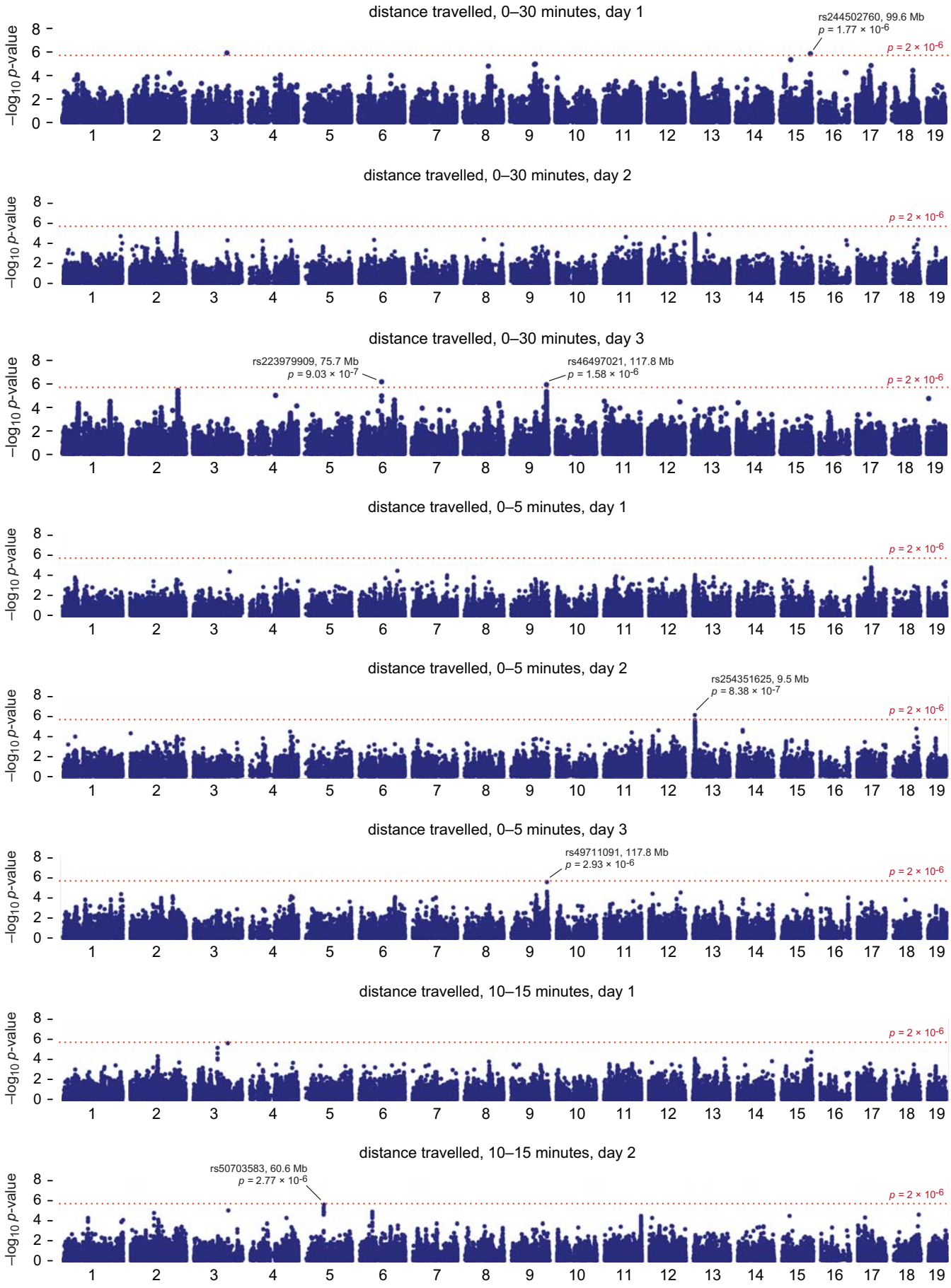


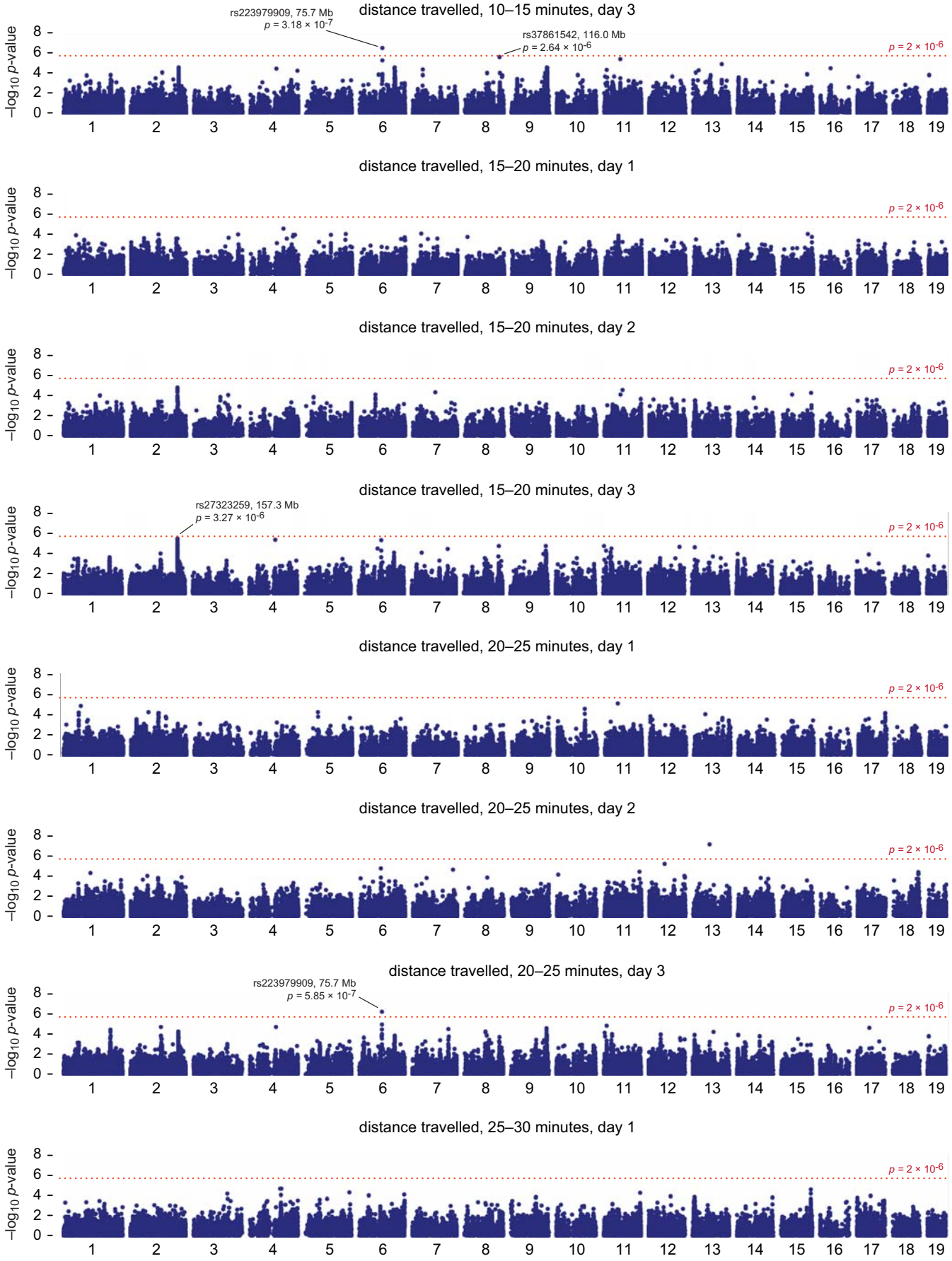
Supplementary Figure 15: Genome-wide scans for musculoskeletal traits. *p*-values quantify support for a QTL at each of the 92,734 candidate SNPs on autosomal chromosomes. *p*-values are calculated from the likelihood-ratio test using a linear mixed model that corrects for hidden relatedness among the mice, in which relatedness is estimated using the SNP genotypes. The threshold for reporting significant *p*-values is shown as a dotted red line. QTLs approaching or exceeding 2 x 10^{-6} are highlighted in by showing information about the SNP in the QTL region with the strongest support for association with the phenotype; refer to Supplementary Table 2 for additional information on these QTLs. Some SNPs exceeding the significance threshold are not highlighted (e.g. a SNP mapped on chromosome 5 for soleus weight) because it has a much smaller *p*-value than all other nearby SNPs, and the SNP at that base-pair position does not have a corresponding entry in the dbSNP database, suggesting the possibility that the genetic variant is a false positive. Some highlighted loci do not meet genome-wide significance in the initial analysis, but show stronger evidence for association after assessing support for multiple QTLs on the same chromosome; *p*-values marked with an asterisk (*) are obtained after conditioning on one or more QTLs on the same chromosome (see Supplementary Table 2, methods section for details).

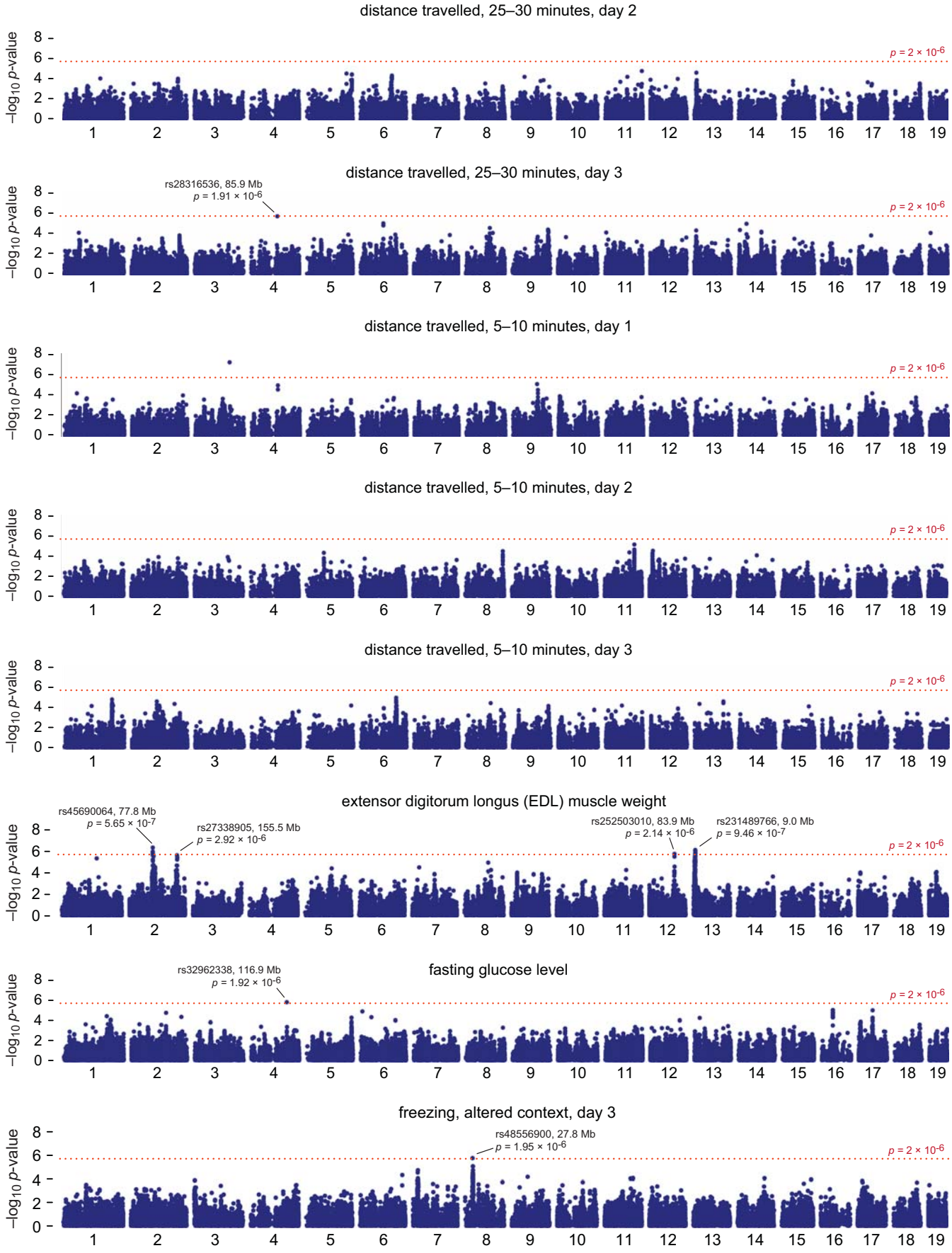


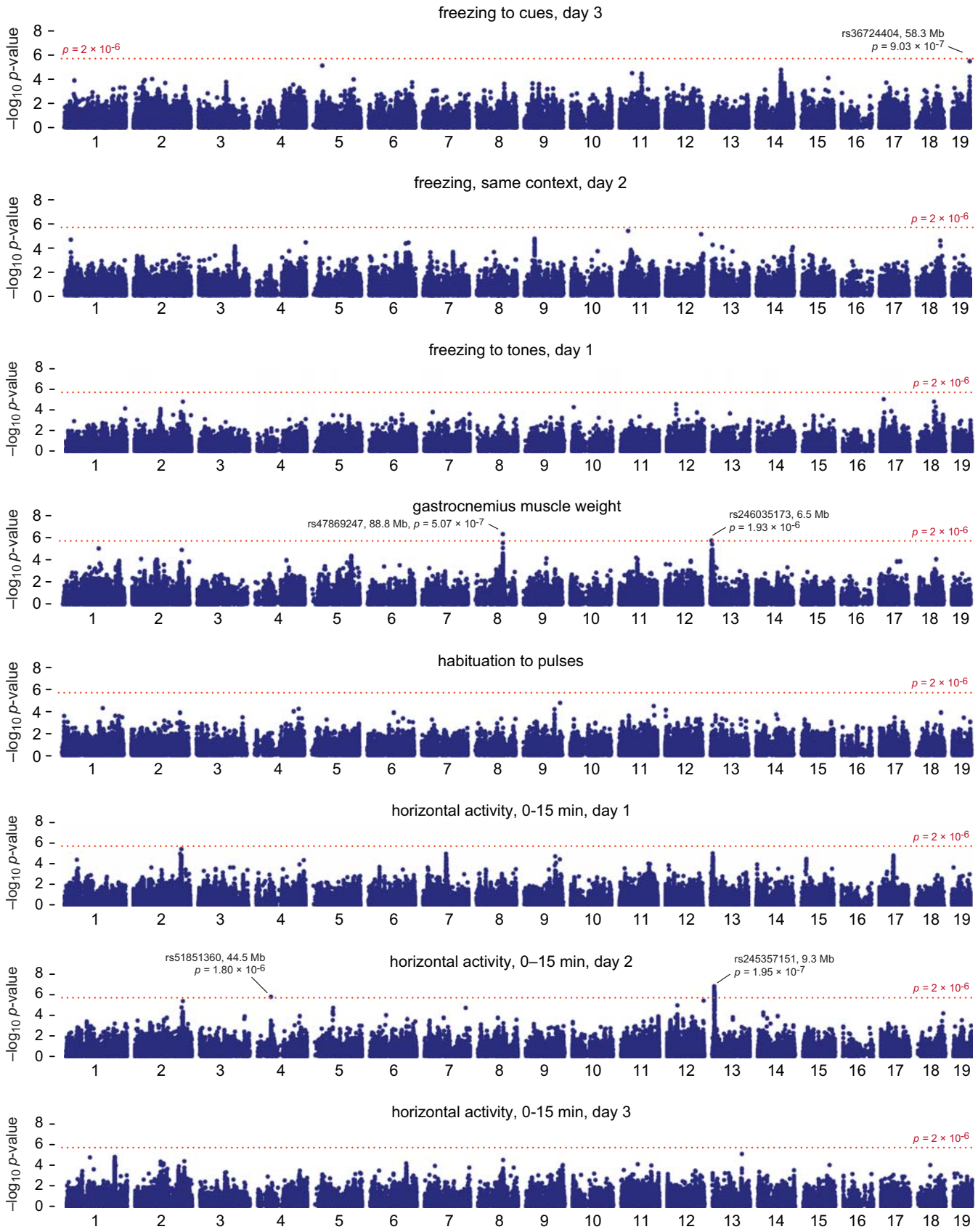
Nature Genetics: doi:10.1038/ng.3609

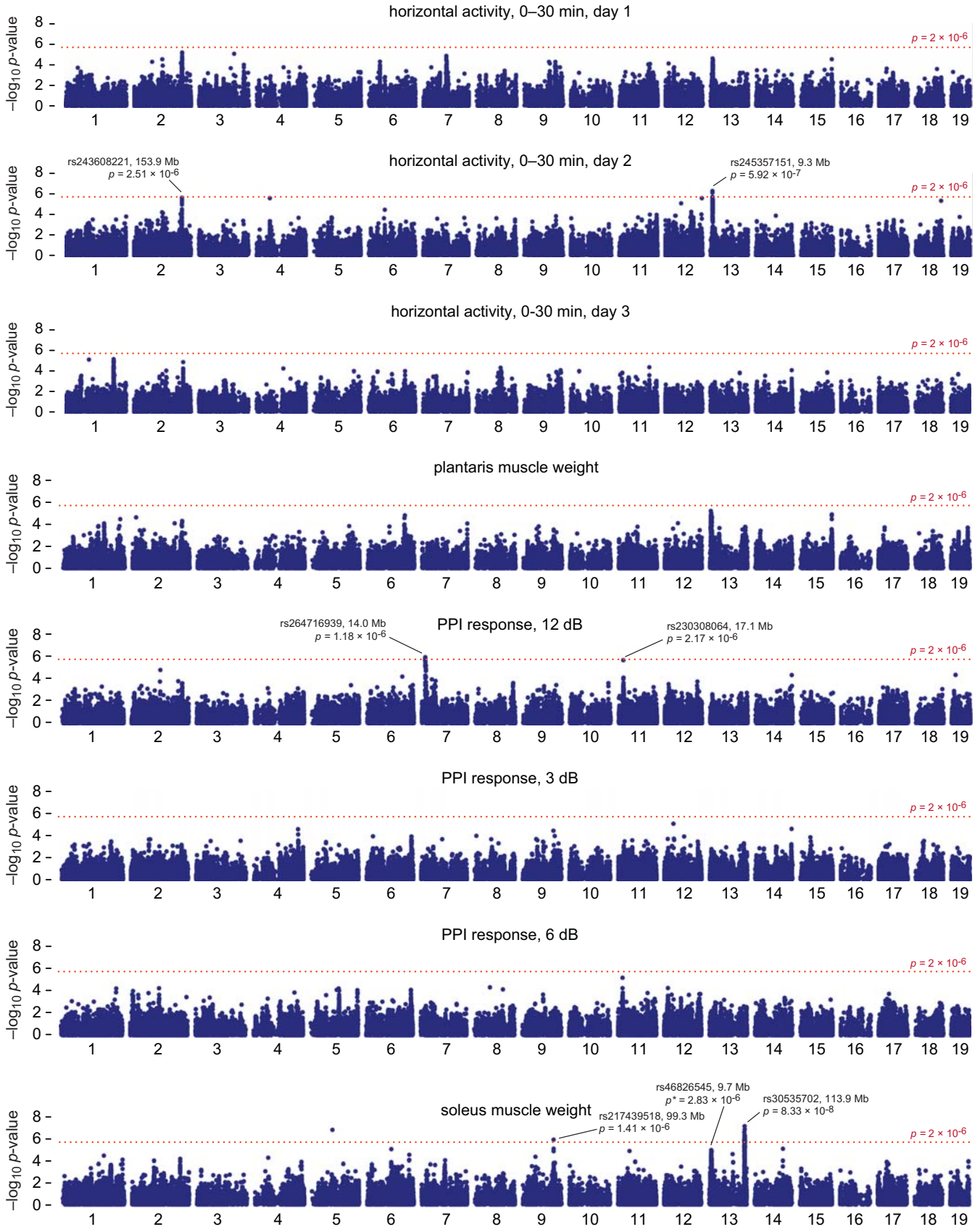

center time, 0-15 min, day 2
rs214901846, 9.3 Mb
p = 1.85 × 10-7
p = 2 × 10-6
center time, 0-15 min, day 3
rs31194230, 9.6 Mb
p = 2.86 × 10-6
p = 2 × 10-6
center time, 0-30 min, day 1
rs238465220, 10.0 Mb
p = 7.31 × 10-8
p = 2 × 10-6
center time, 0-30 min, day 2
rs214901846, 9.3 Mb
p = 3.45 × 10-6
p = 2 × 10-6
center time, 0-30 min, day 3
rs237368278, 48.3 Mb
p = 3.33 × 10-6
p = 2 × 10-6
distance travelled, 0–15 minutes, day 1
p = 2 × 10-6
distance travelled, 0–15 minutes, day 2
p = 2 × 10-6
distance travelled, 0–15 minutes, day 3
p = 2 × 10-6
−log10 p-value
0
2
4
6
8
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19

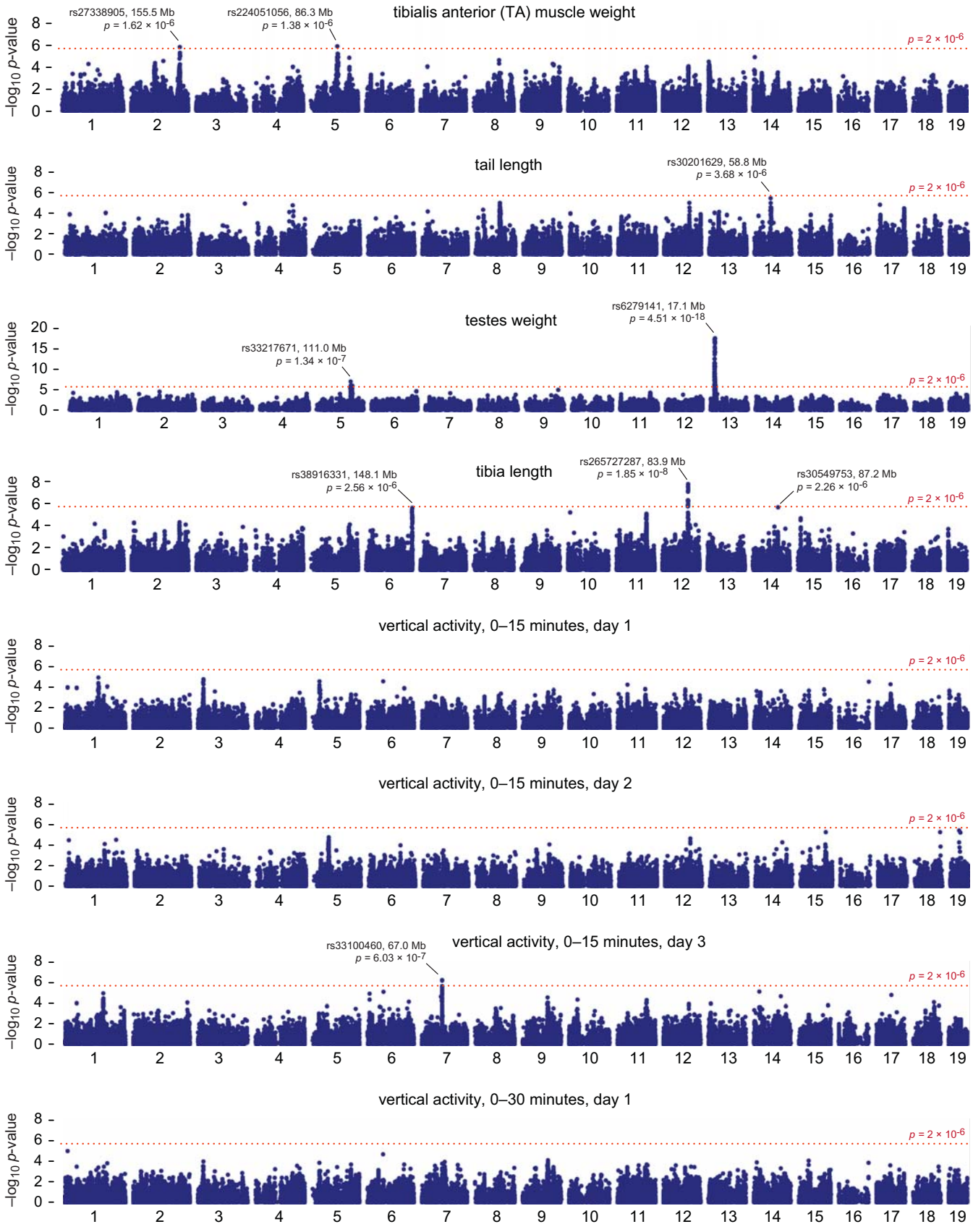

distance travelled, 0–30 minutes, day 1
rs244502760, 99.6 Mb
p = 1.77 × 10-6
p = 2 × 10-6
distance travelled, 0–30 minutes, day 2
p = 2 × 10-6
distance travelled, 0–30 minutes, day 3
rs223979909, 75.7 Mb
p = 9.03 × 10-7
rs46497021, 117.8 Mb
p = 1.58 × 10-6
p = 2 × 10-6
distance travelled, 0–5 minutes, day 1
p = 2 × 10-6
distance travelled, 0–5 minutes, day 2
rs254351625, 9.5 Mb
p = 8.38 × 10-7
p = 2 × 10-6
distance travelled, 0–5 minutes, day 3
rs49711091, 117.8 Mb
p = 2.93 × 10-6
p = 2 × 10-6
distance travelled, 10–15 minutes, day 1
p = 2 × 10-6
distance travelled, 10–15 minutes, day 2
rs50703583, 60.6 Mb
p = 2.77 × 10-6
p = 2 × 10-6
−log10 p-value
0
2
4
6
8
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19


distance travelled, 10–15 minutes, day 3
rs223979909, 75.7 Mb
p = 3.18 × 10-7
rs37861542, 116.0 Mb
p = 2.64 × 10-6
p = 2 × 10-6
distance travelled, 15–20 minutes, day 1
distance travelled, 15–20 minutes, day 2
distance travelled, 15–20 minutes, day 3
rs27323259, 157.3 Mb
p = 3.27 × 10-6
distance travelled, 20–25 minutes, day 1
distance travelled, 20–25 minutes, day 2
distance travelled, 20–25 minutes, day 3
rs223979909, 75.7 Mb
p = 5.85 × 10-7
distance travelled, 25–30 minutes, day 1
−log10 p-value
0
2
4
6
8
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19

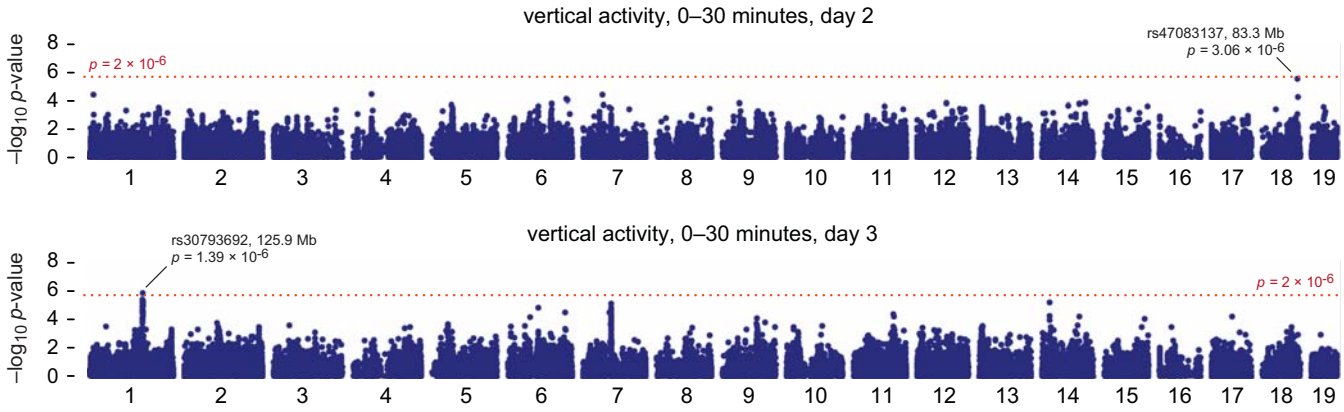

distance travelled, 25–30 minutes, day 2
distance travelled, 25–30 minutes, day 3
rs28316536, 85.9 Mb
p = 1.91 × 10-6
distance travelled, 5–10 minutes, day 1
distance travelled, 5–10 minutes, day 2
distance travelled, 5–10 minutes, day 3
extensor digitorum longus (EDL) muscle weight
rs45690064, 77.8 Mb
p = 5.65 × 10-7
rs27338905, 155.5 Mb
p = 2.92 × 10-6
rs252503010, 83.9 Mb
p = 2.14 × 10-6
rs231489766, 9.0 Mb
p = 9.46 × 10-7
fasting glucose level
rs32962338, 116.9 Mb
p = 1.92 × 10-6
freezing, altered context, day 3
rs48556900, 27.8 Mb
p = 1.95 × 10-6
p = 2 × 10-6
−log10 p-value
0
2
4
6
8
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19


freezing to cues, day 3
rs36724404, 58.3 Mb
p = 9.03 × 10-7
p = 2 × 10-6
freezing, same context, day 2
p = 2 × 10-6
freezing to tones, day 1
p = 2 × 10-6
gastrocnemius muscle weight
rs47869247, 88.8 Mb, p = 5.07 × 10-7
rs246035173, 6.5 Mb
p = 1.93 × 10-6
p = 2 × 10-6
habituation to pulses
p = 2 × 10-6
horizontal activity, 0-15 min, day 1
p = 2 × 10-6
horizontal activity, 0–15 min, day 2
rs51851360, 44.5 Mb
p = 1.80 × 10-6
rs245357151, 9.3 Mb
p = 1.95 × 10-7
p = 2 × 10-6
horizontal activity, 0-15 min, day 3
p = 2 × 10-6
−log10 p-value
0
2
4
6
8
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19


horizontal activity, 0–30 min, day 1
p = 2 × 10-6
horizontal activity, 0–30 min, day 2
rs243608221, 153.9 Mb
p = 2.51 × 10-6
rs245357151, 9.3 Mb
p = 5.92 × 10-7
p = 2 × 10-6
horizontal activity, 0-30 min, day 3
p = 2 × 10-6
plantaris muscle weight
p = 2 × 10-6
PPI response, 12 dB
rs264716939, 14.0 Mb
p = 1.18 × 10-6
rs230308064, 17.1 Mb
p = 2.17 × 10-6
p = 2 × 10-6
PPI response, 3 dB
p = 2 × 10-6
PPI response, 6 dB
p = 2 × 10-6
soleus muscle weight
rs46826545, 9.7 Mb
p* = 2.83 × 10-6
rs30535702, 113.9 Mb
p = 8.33 × 10-8
rs217439518, 99.3 Mb
p = 1.41 × 10-6
p = 2 × 10-6
−log10 p-value
0
2
4
6
8
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19



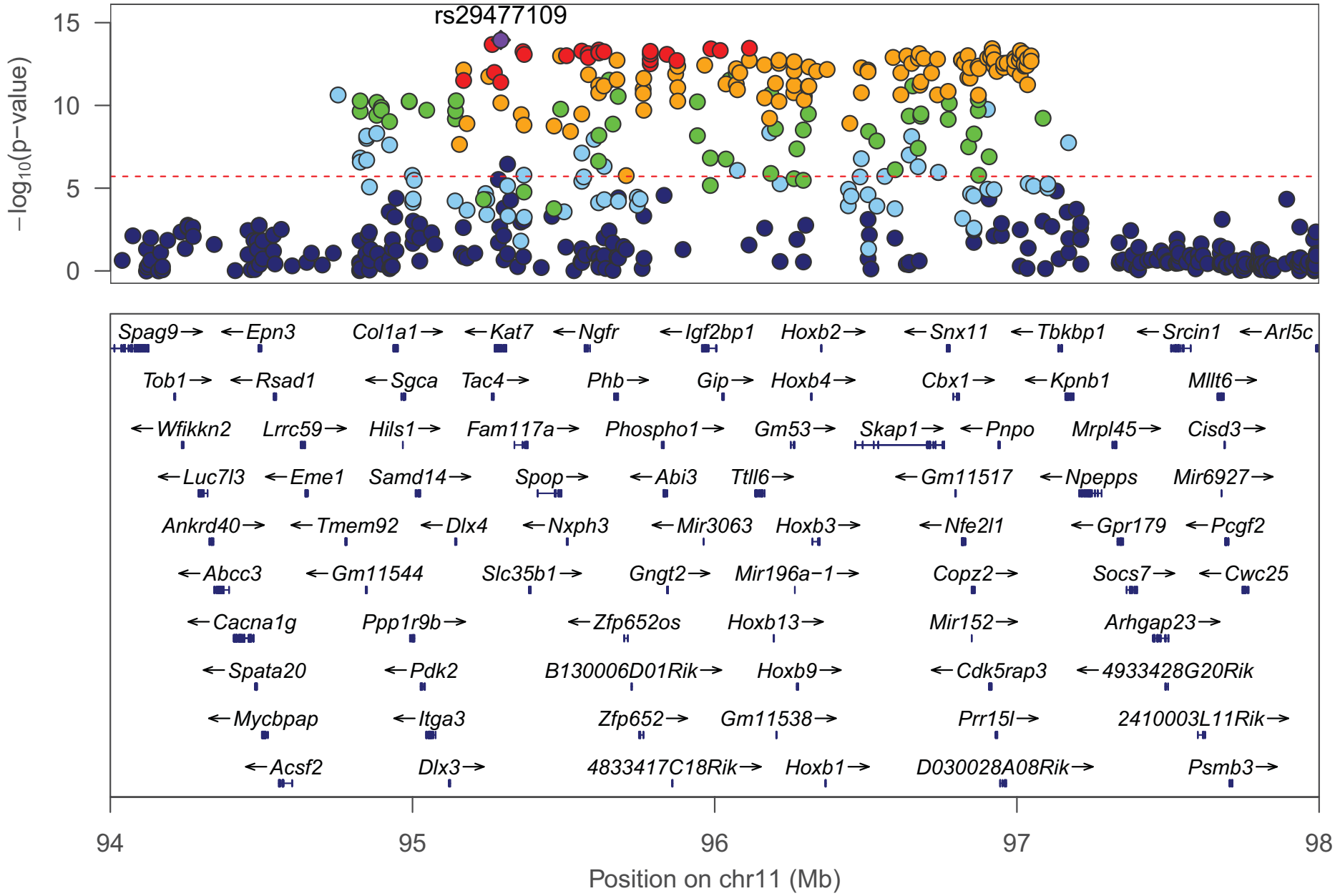
Nature Genetics: doi:10.1038/ng.3609


vertical activity, 0–30 minutes, day 2
rs47083137, 83.3 Mb
p = 3.06 × 10-6
p = 2 × 10-6
–log10 p-value
8
6
4
2
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
vertical activity, 0–30 minutes, day 3
rs30793692, 125.9 Mb
p = 1.39 × 10-6
p = 2 × 10-6
–log10 p-value
8
6
4
2
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19

Supplementary Figure 16: Genome-wide scans for all the phenotypes analyzed in our study. *p*-values quantify support for a QTL at each of the 92,734 candidate SNPs on autosomal chromosomes. *p*-values are calculated from the likelihood-ratio test using a linear mixed model that corrects for hidden relatedness among the mice, in which relatedness is estimated using the SNP genotypes. The threshold for reporting *p*-values is shown as a dotted red line. QTLs approaching 2×10^{-6} are also shown in some cases. Refer to Supplementary Table 2 for additional information on these QTLs.

Nature Genetics: doi:10.1038/ng.3609

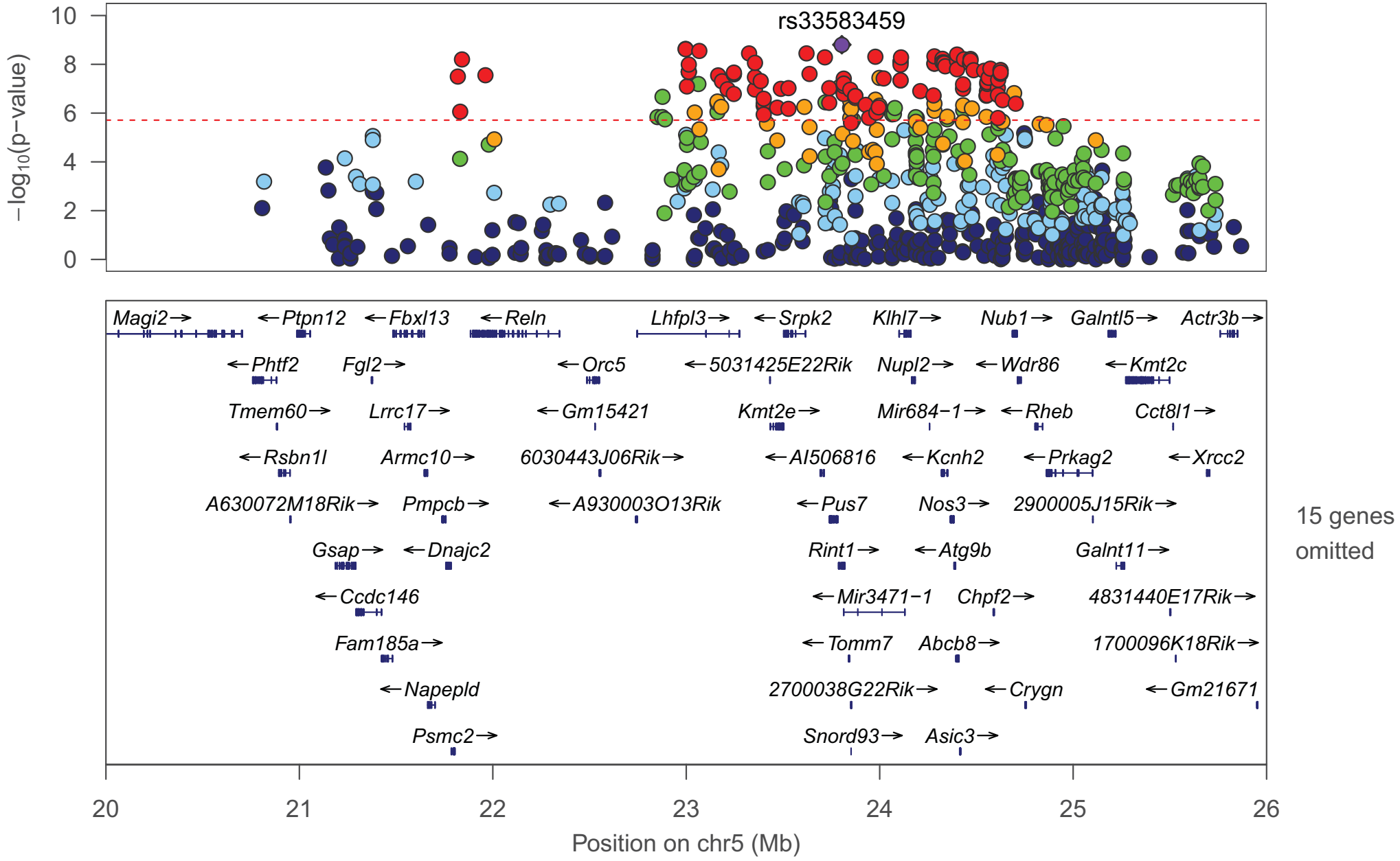
'Abnormal' BMD



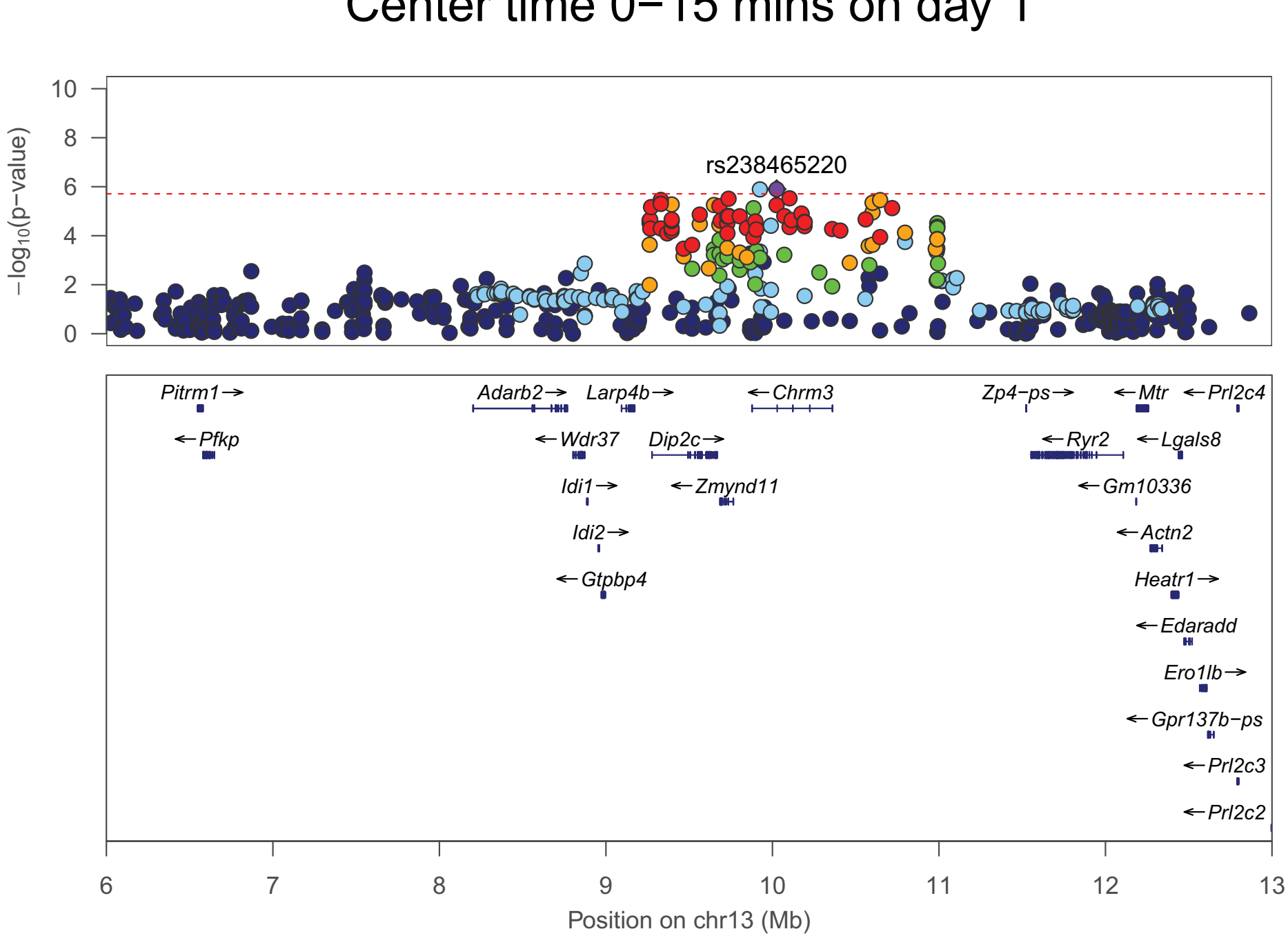
36 genes omitted

Nature Genetics: doi:10.1038/ng.3609

'Abnormal' BMD

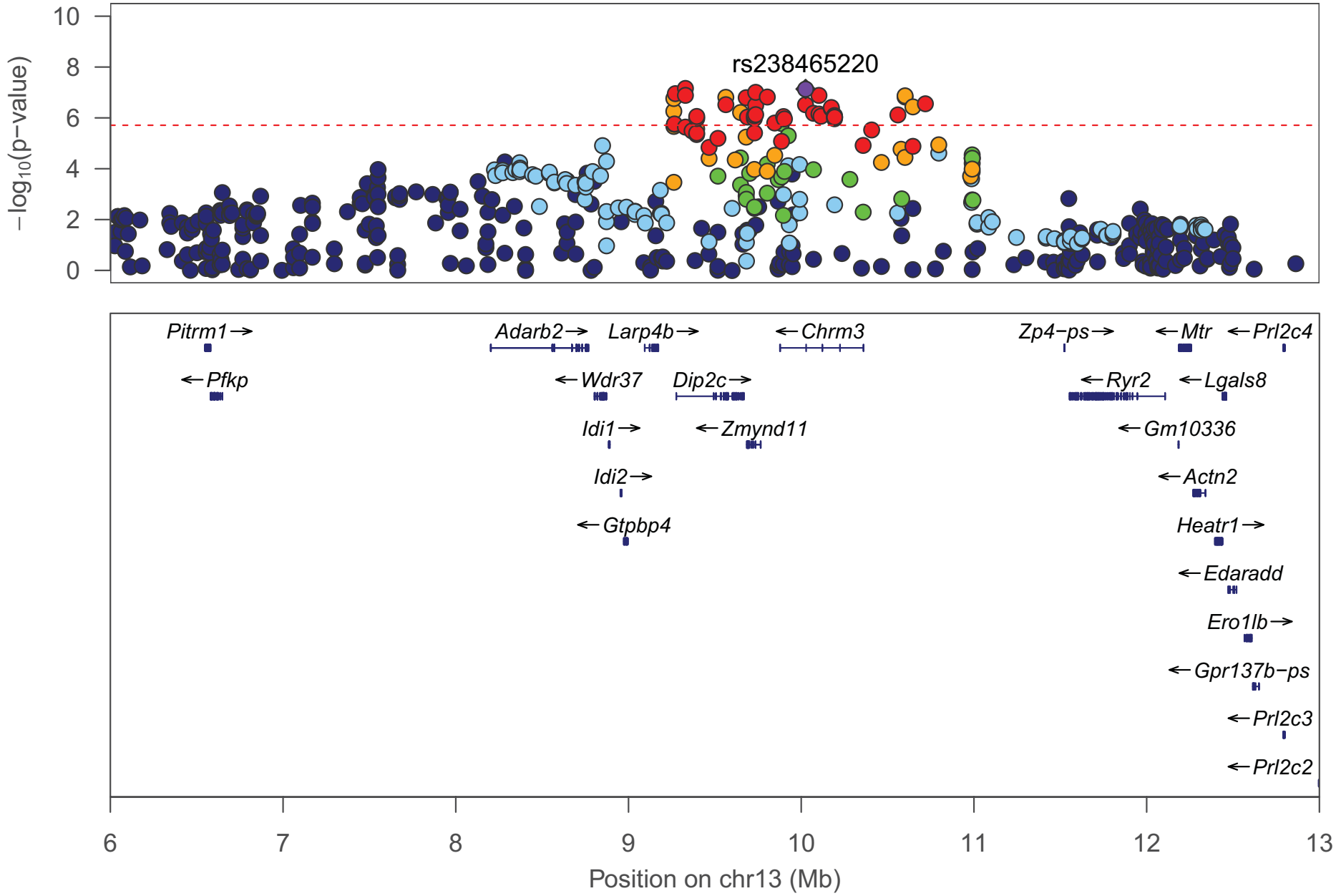


Nature Genetics: doi:10.1038/ng.3609



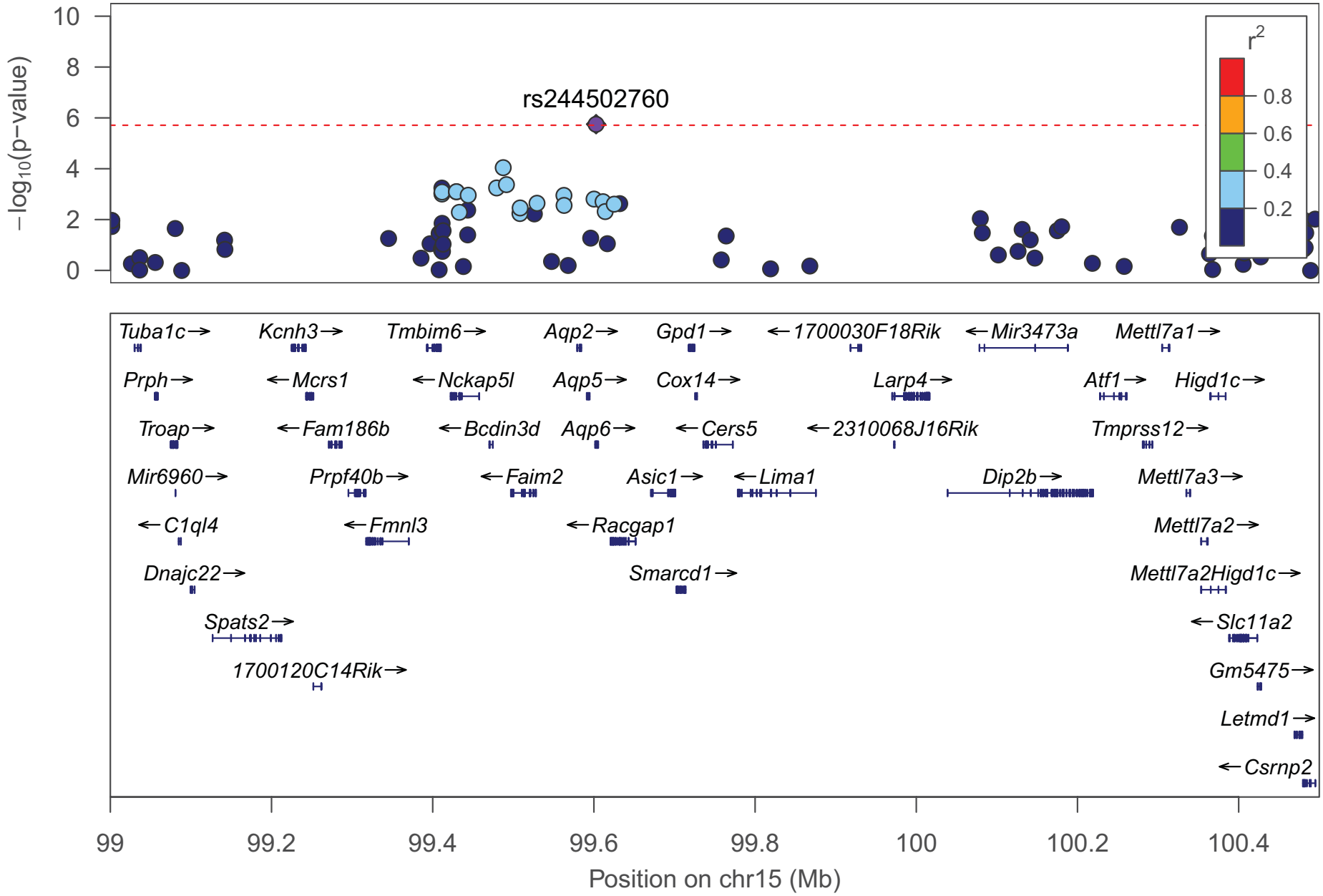
Nature Genetics: doi:10.1038/ng.3609

Center time 0−30 mins on day 1



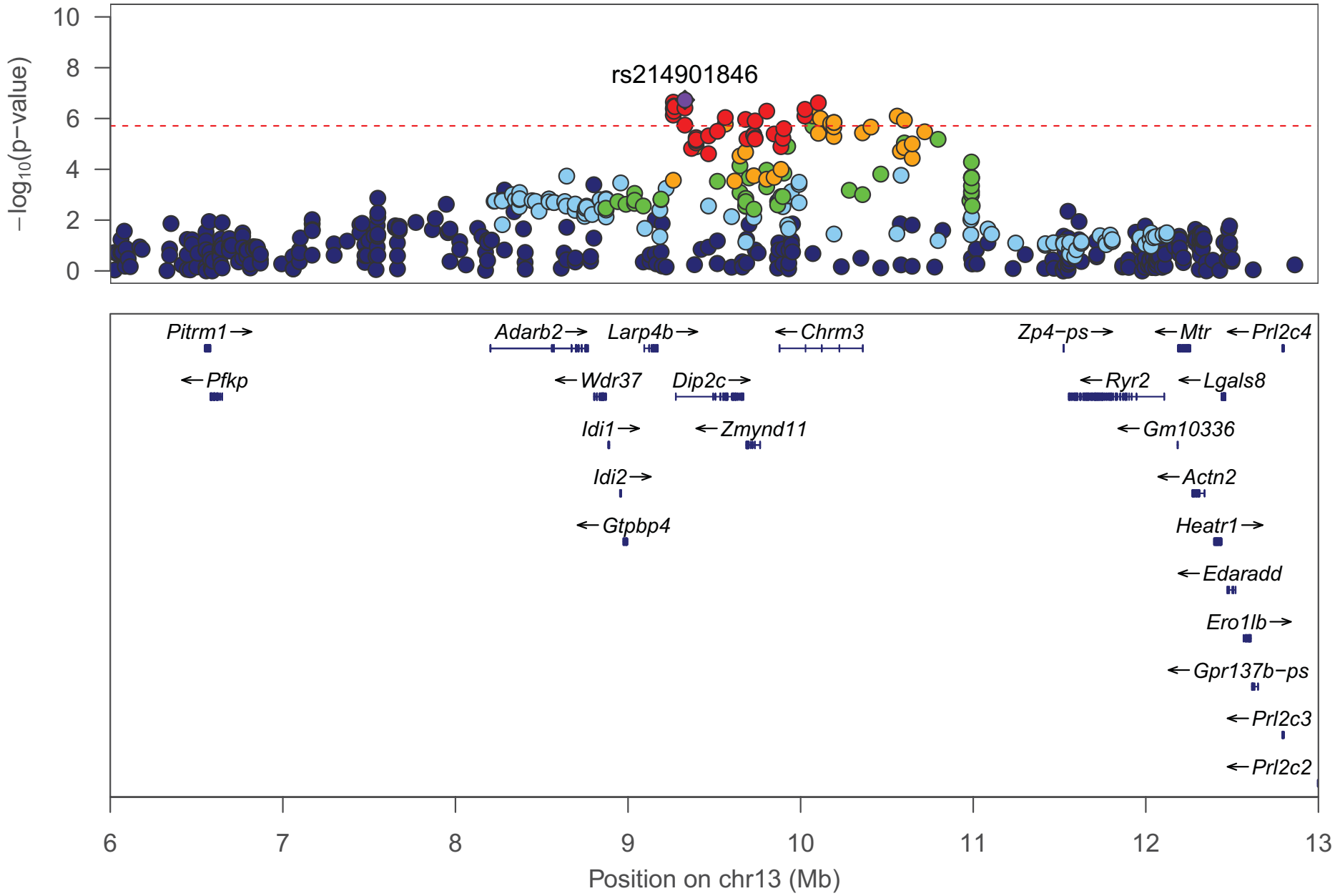
Nature Genetics: doi:10.1038/ng.3609

Total distance traveled on day 1



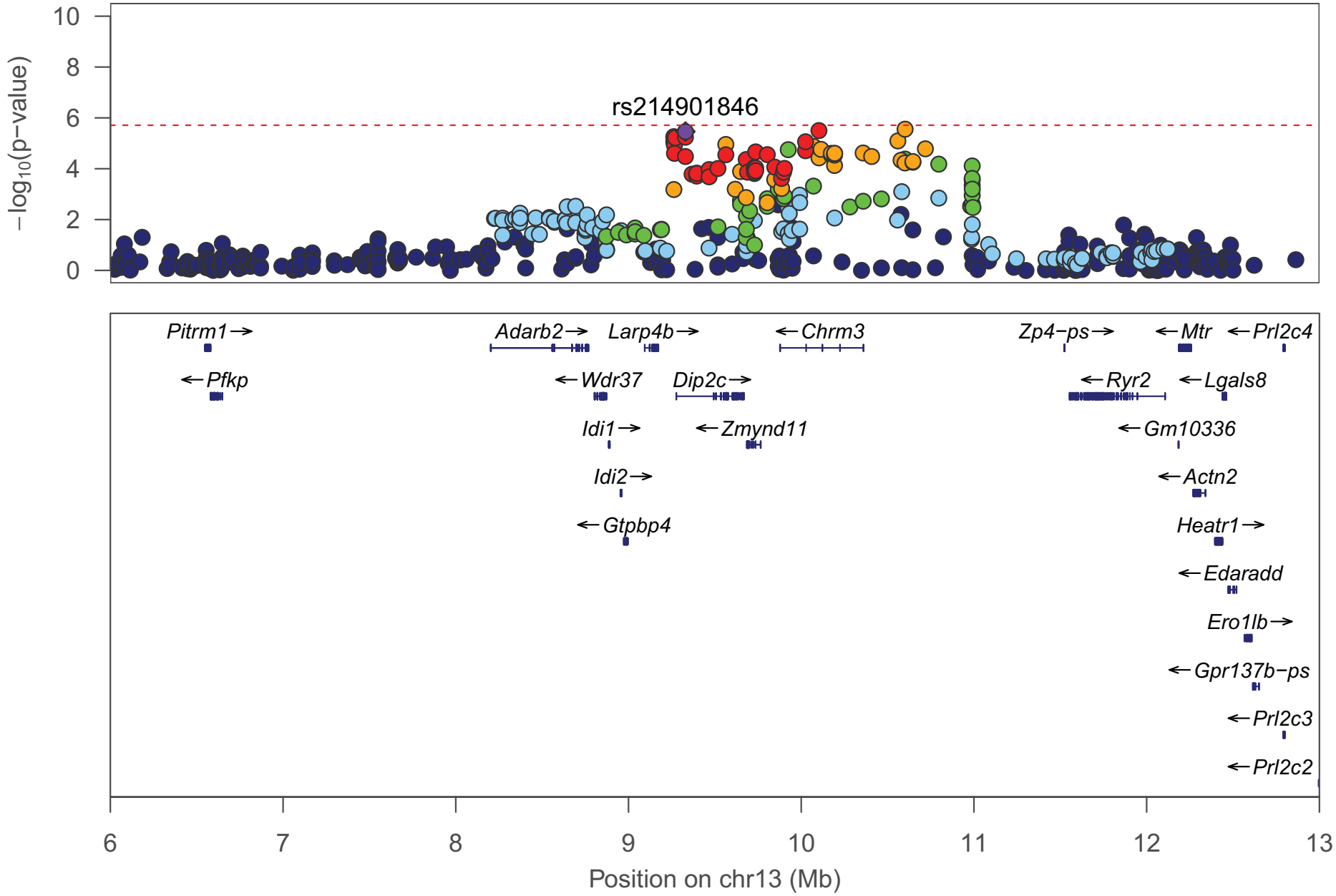
Nature Genetics: doi:10.1038/ng.3609

Center time 0−15 mins on day 2



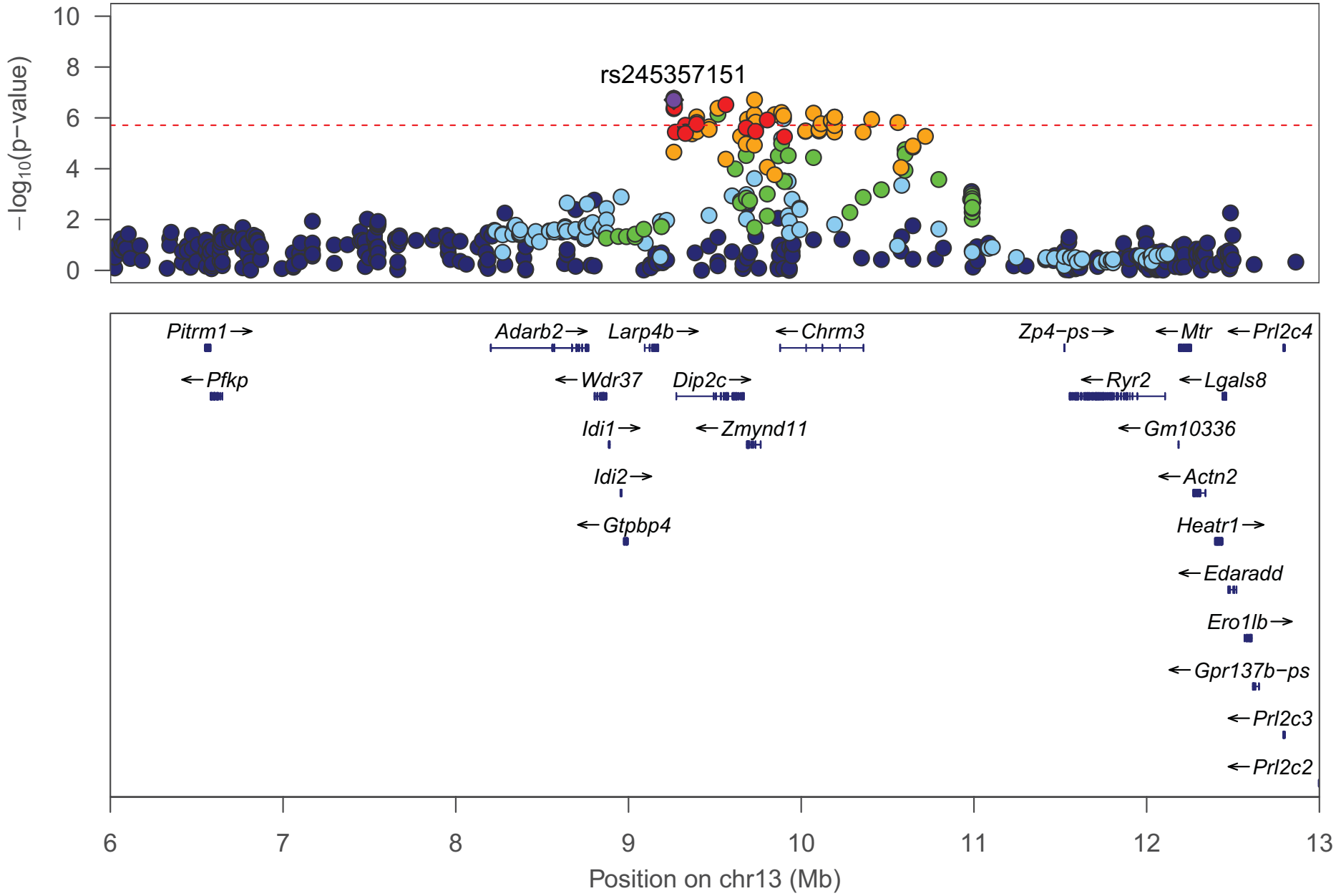
Nature Genetics: doi:10.1038/ng.3609

Center time 0−30 mins on day 2



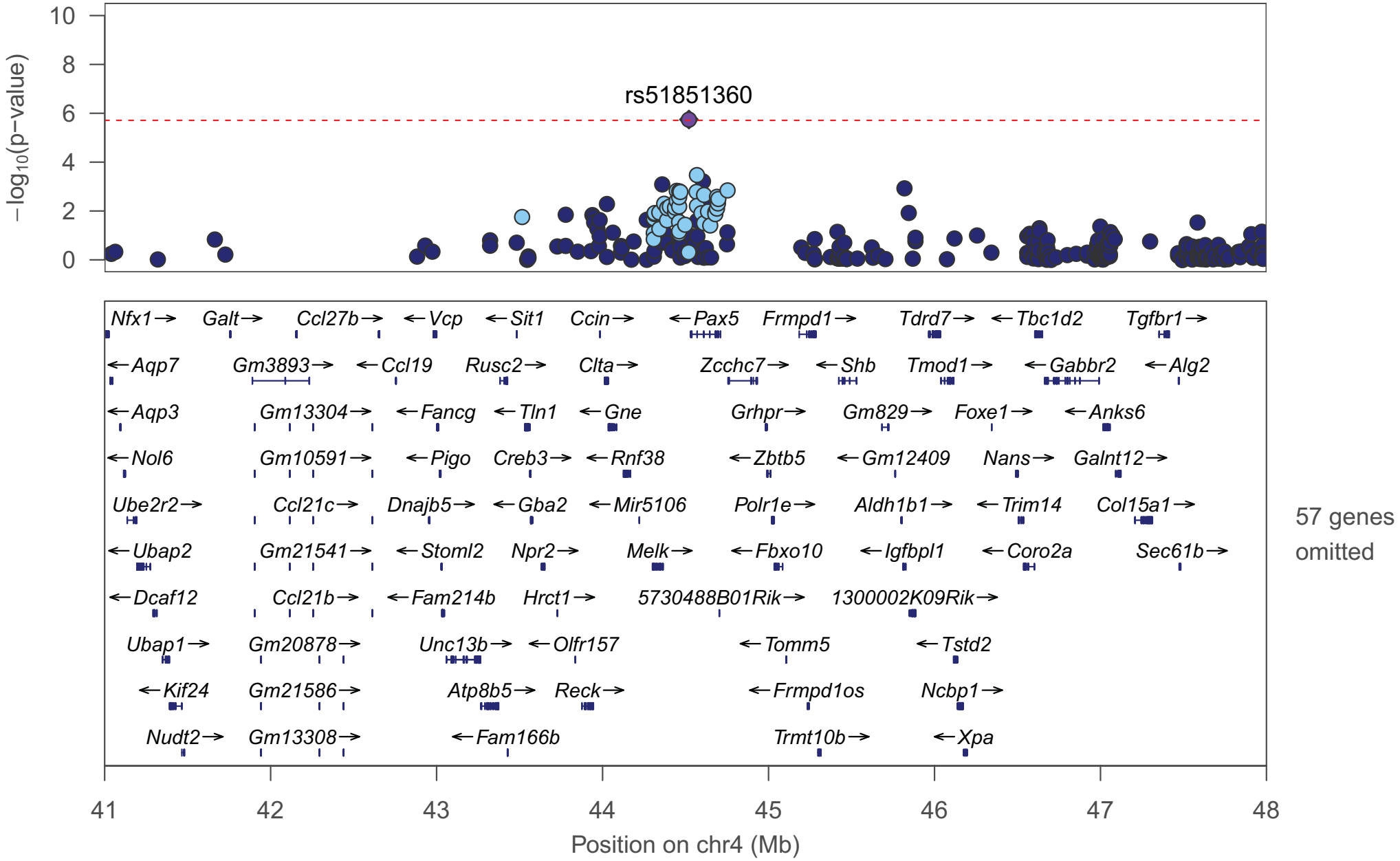
Nature Genetics: doi:10.1038/ng.3609

Horizontal activity 0−15 mins on day 2



Nature Genetics: doi:10.1038/ng.3609

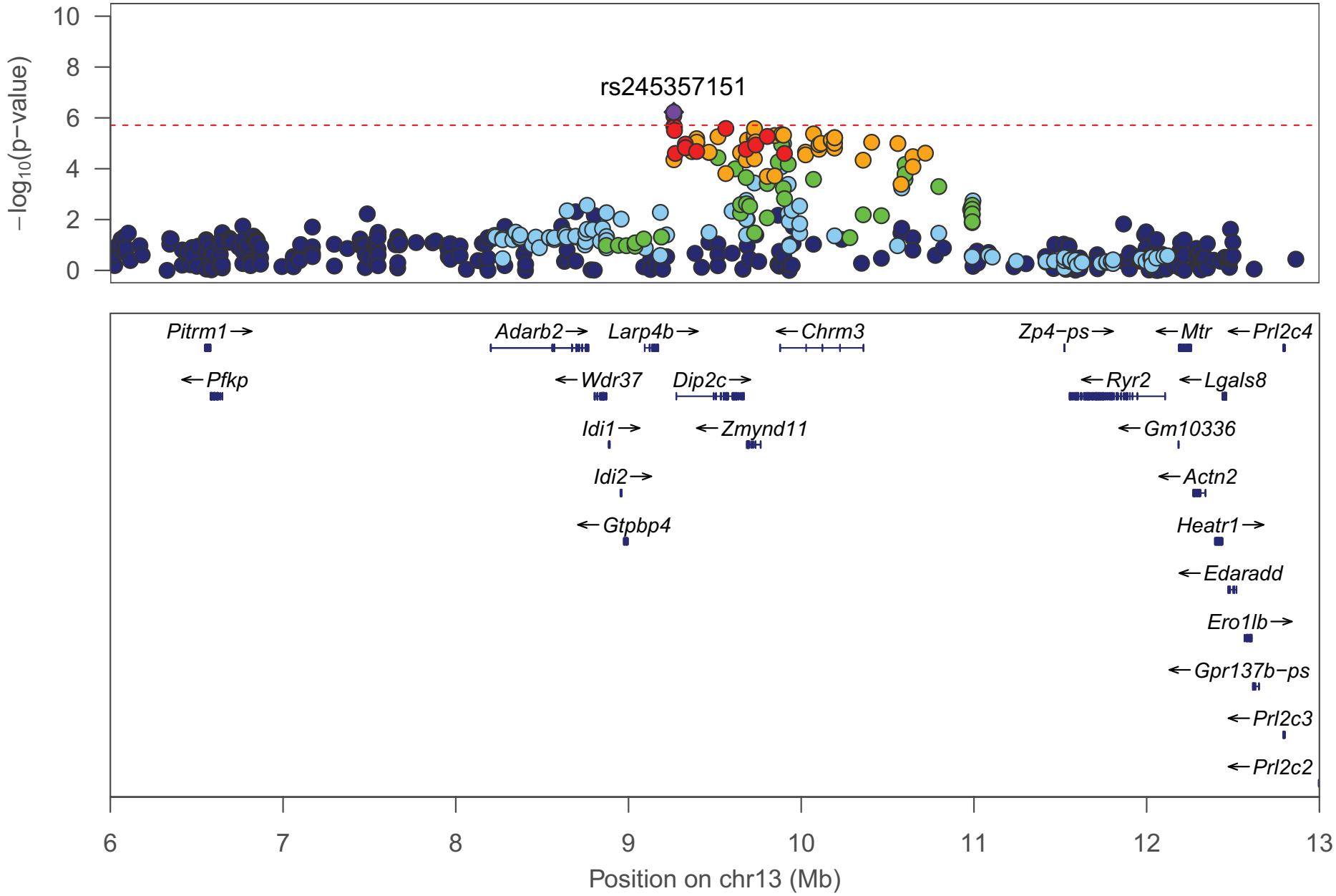
Horizontal activity 0−15 mins on day 2



57 genes omitted

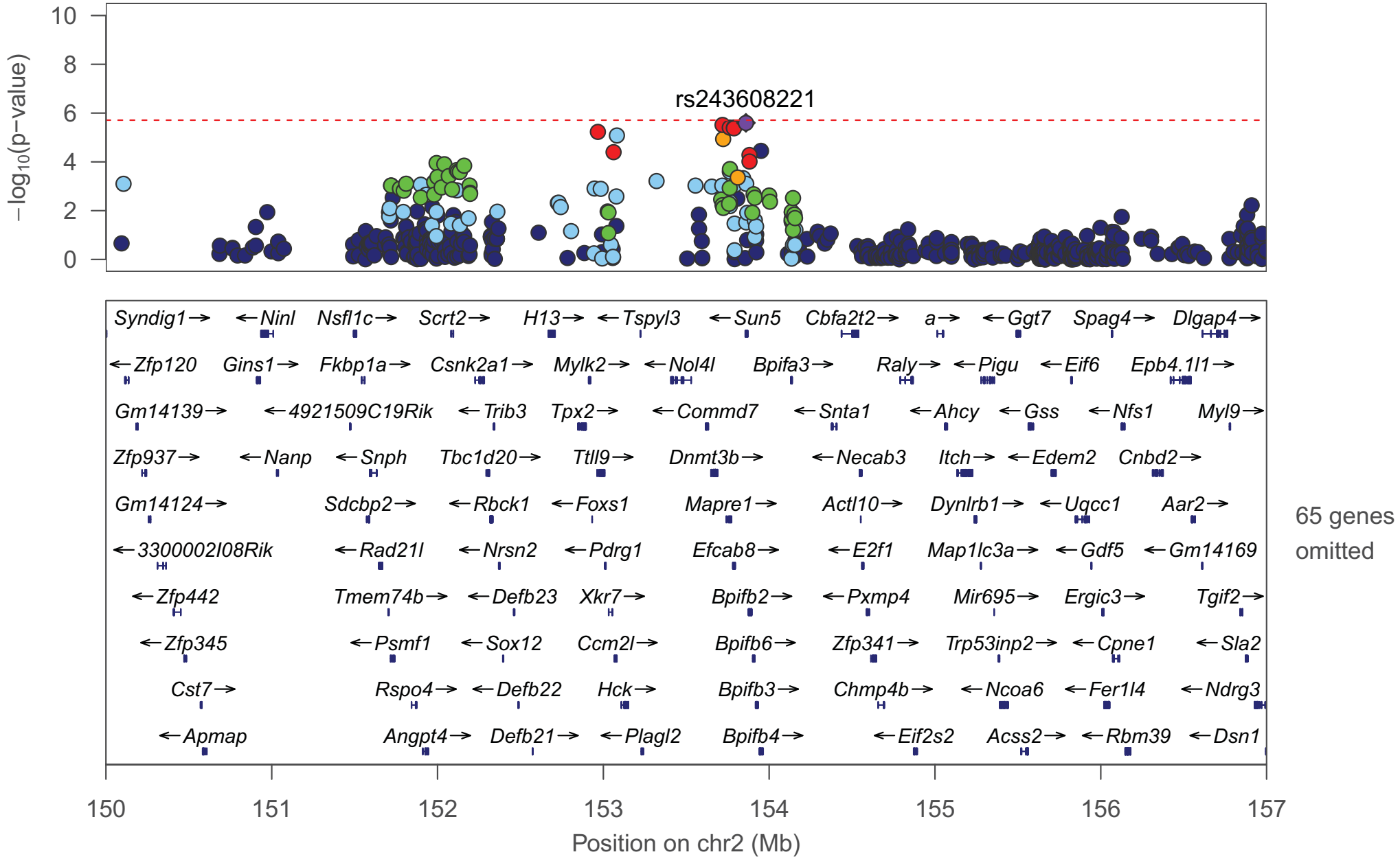
Nature Genetics: doi:10.1038/ng.3609

Horizontal activity 0−30 mins on day 2



Nature Genetics: doi:10.1038/ng.3609

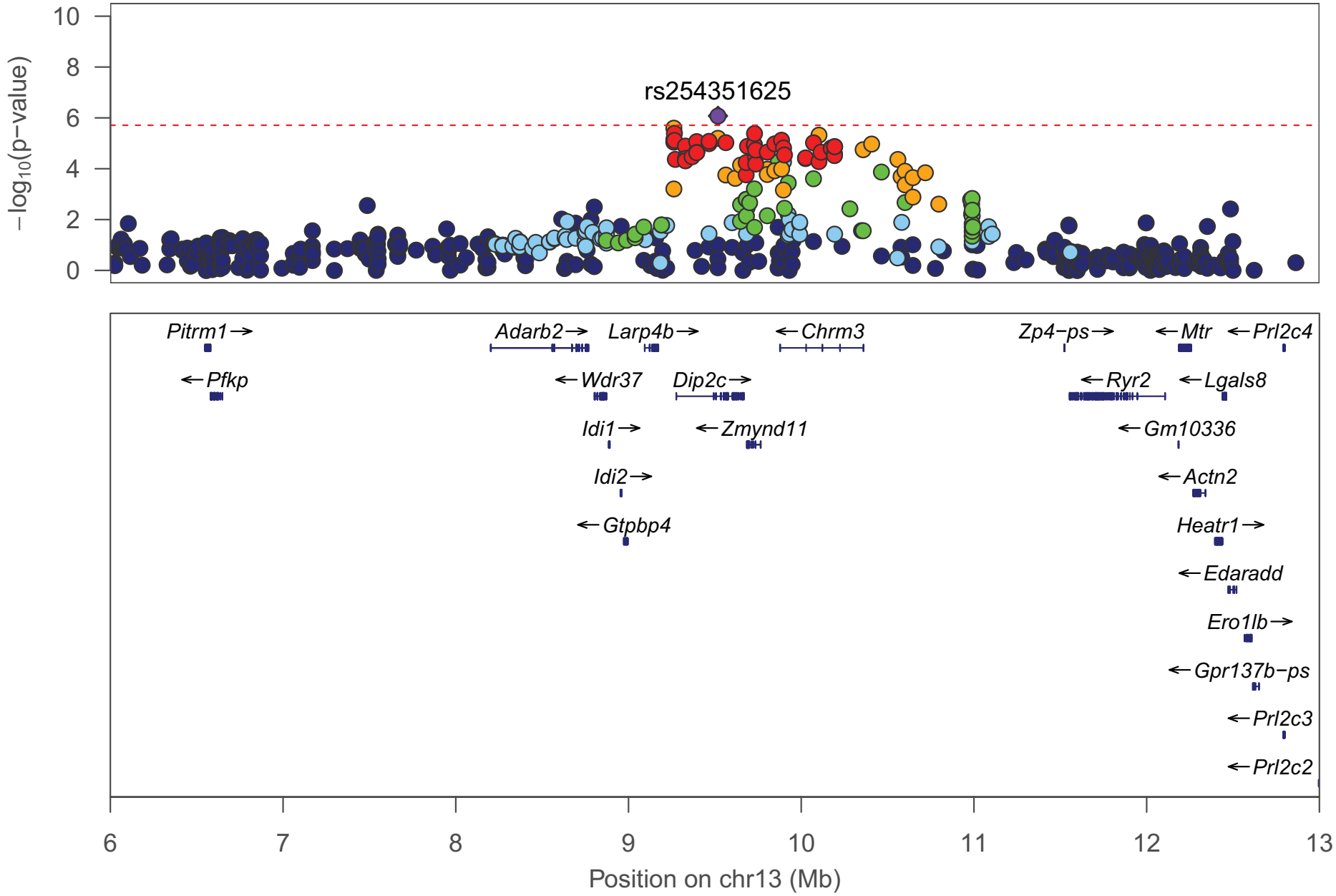
Horizontal activity 0−30 mins on day 2



65 genes omitted

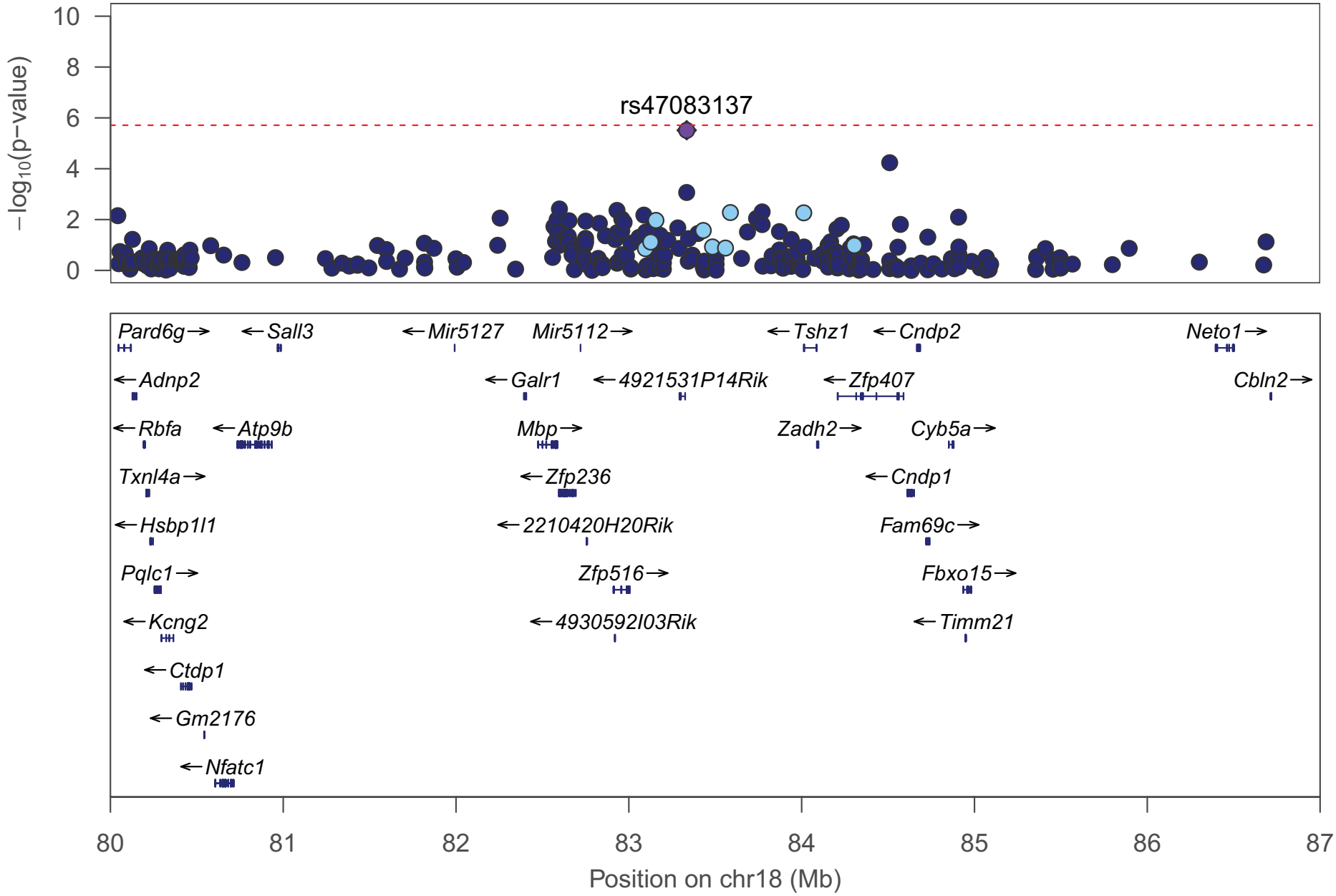
Nature Genetics: doi:10.1038/ng.3609

Distance travelled 0−5 mins on day 2



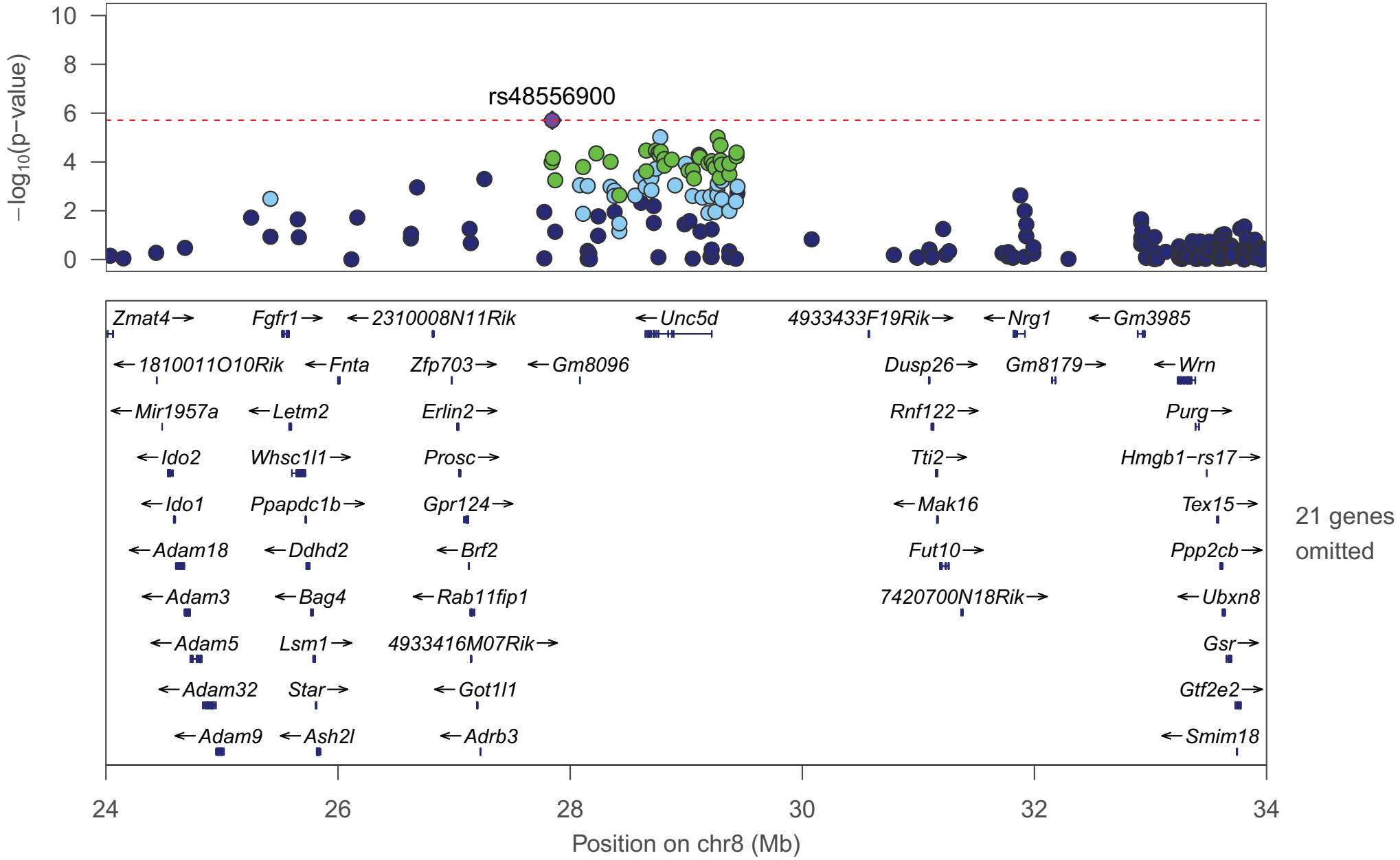
Nature Genetics: doi:10.1038/ng.3609

Vertical activity 0−30 mins on day 2



Nature Genetics: doi:10.1038/ng.3609

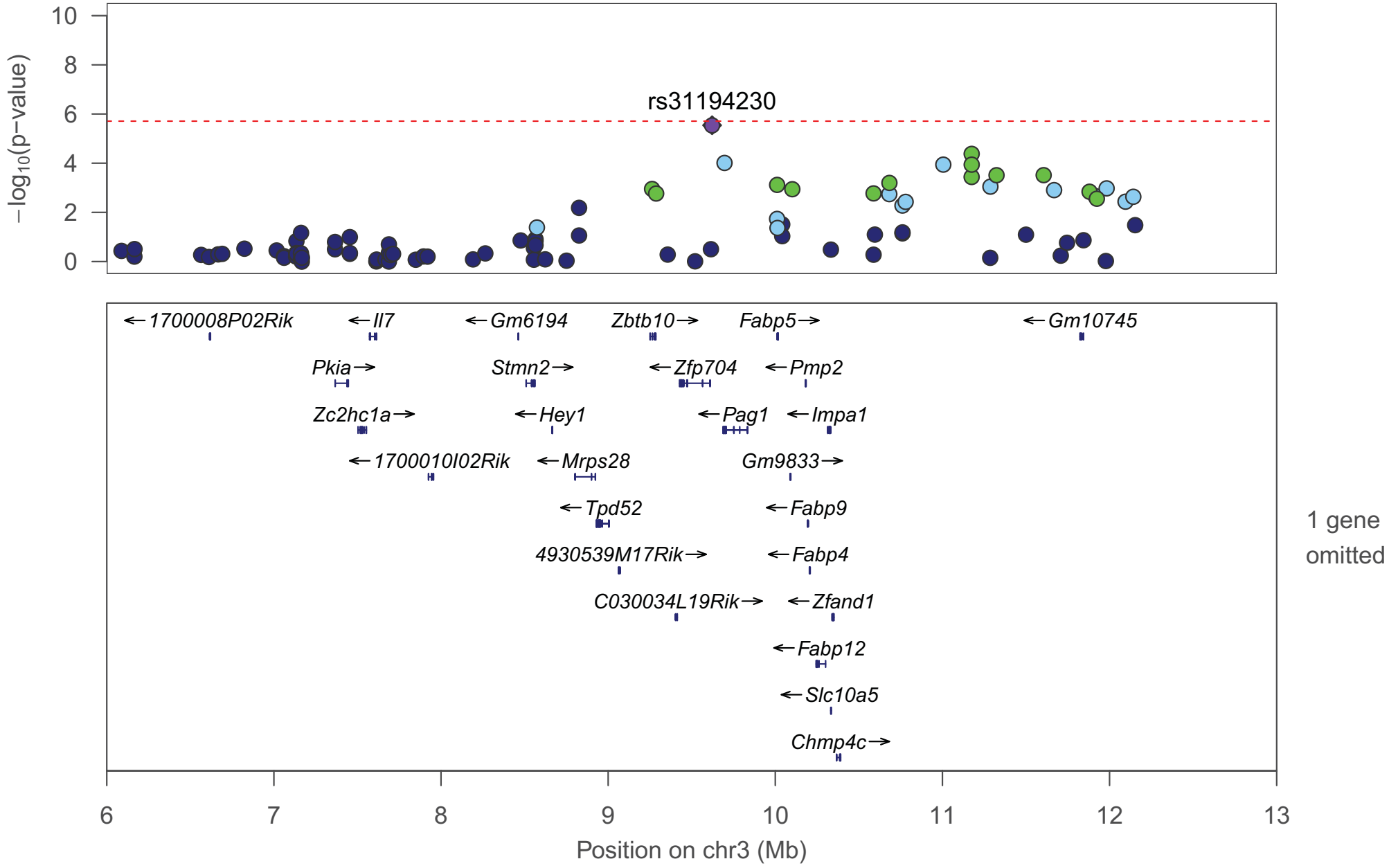
Freezing to altered context on day 3



21 genes omitted

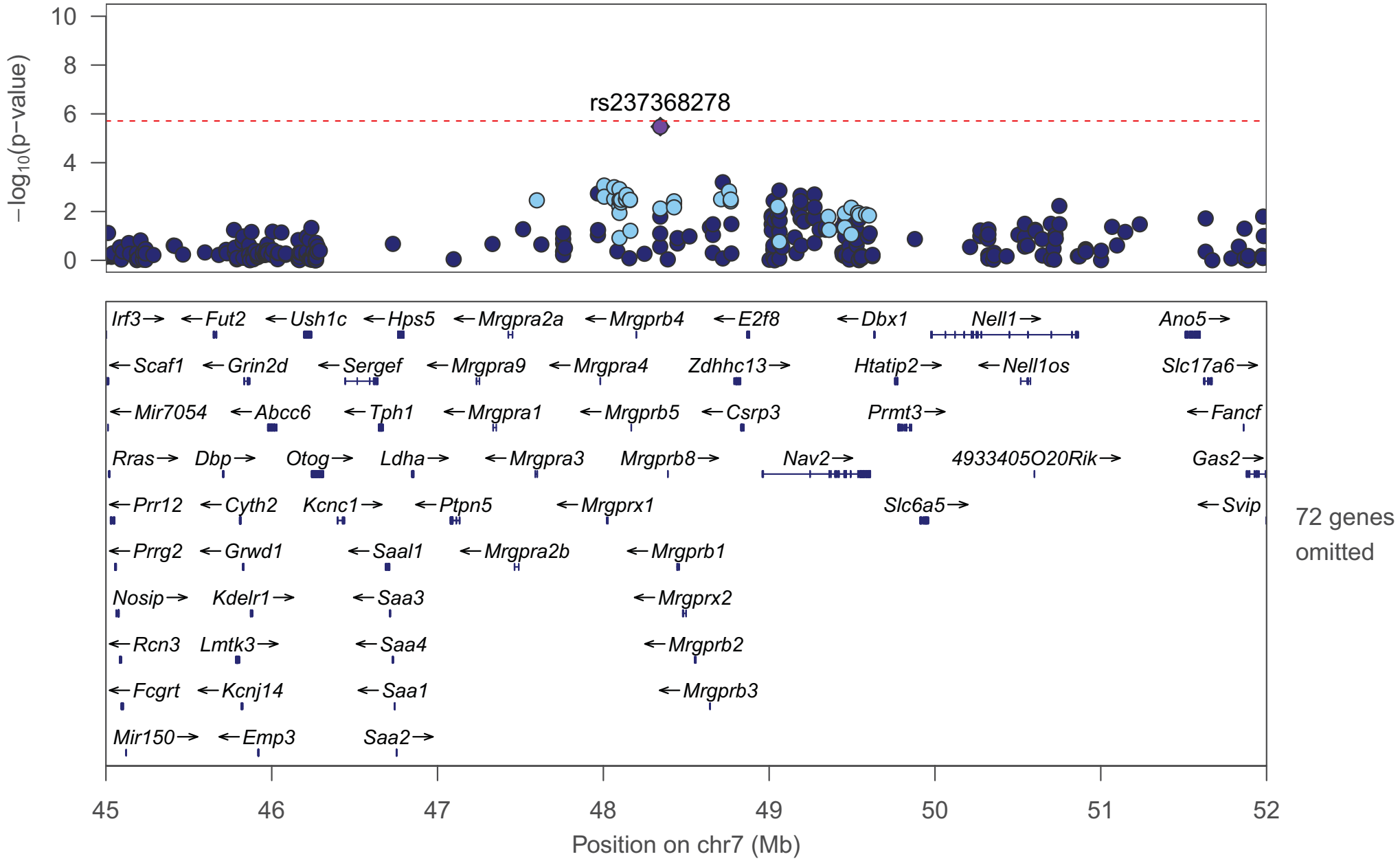
Nature Genetics: doi:10.1038/ng.3609

Center time 0−15 mins on day 3



Nature Genetics: doi:10.1038/ng.3609

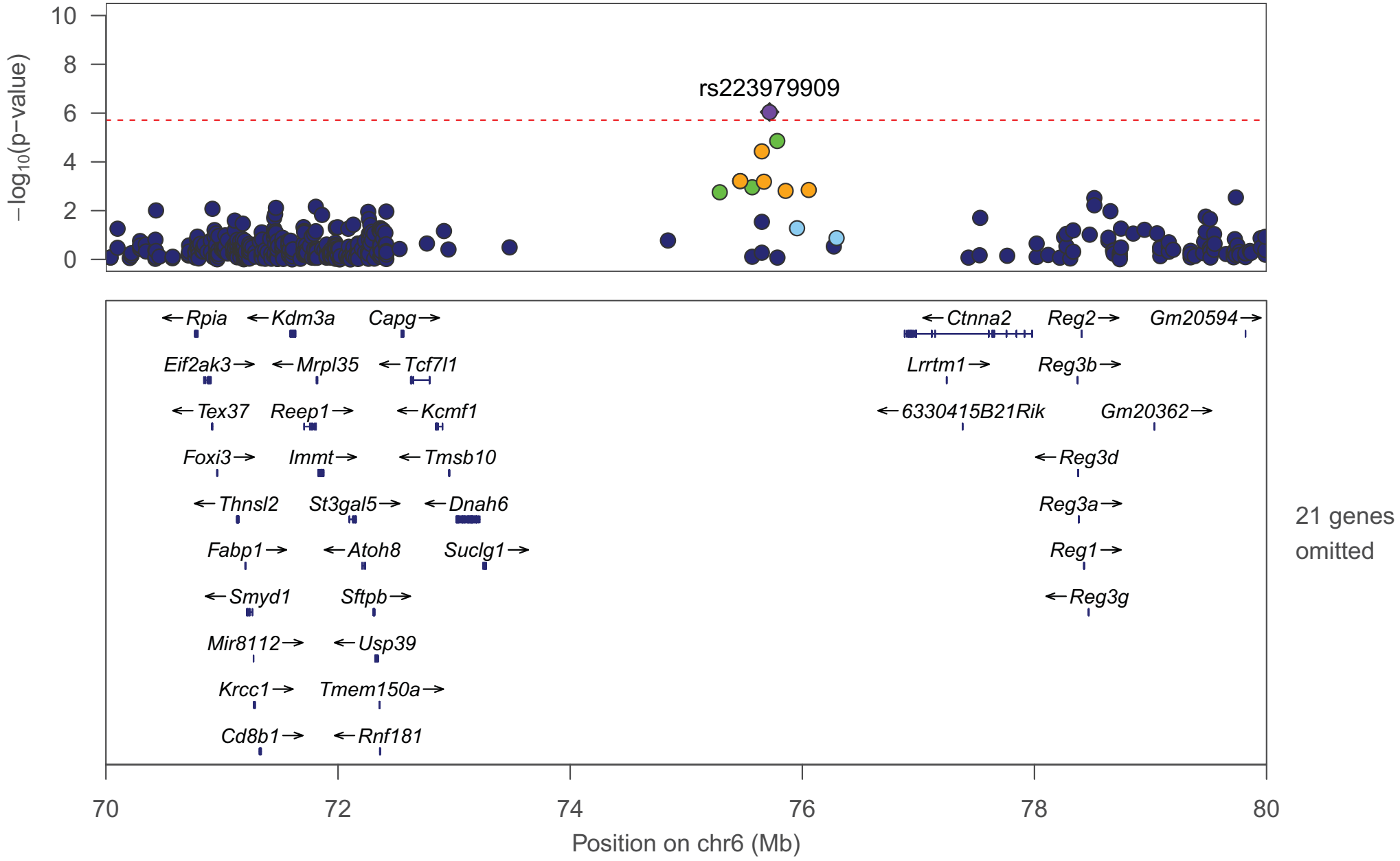
Center time 0−30 mins on day 3



72 genes omitted

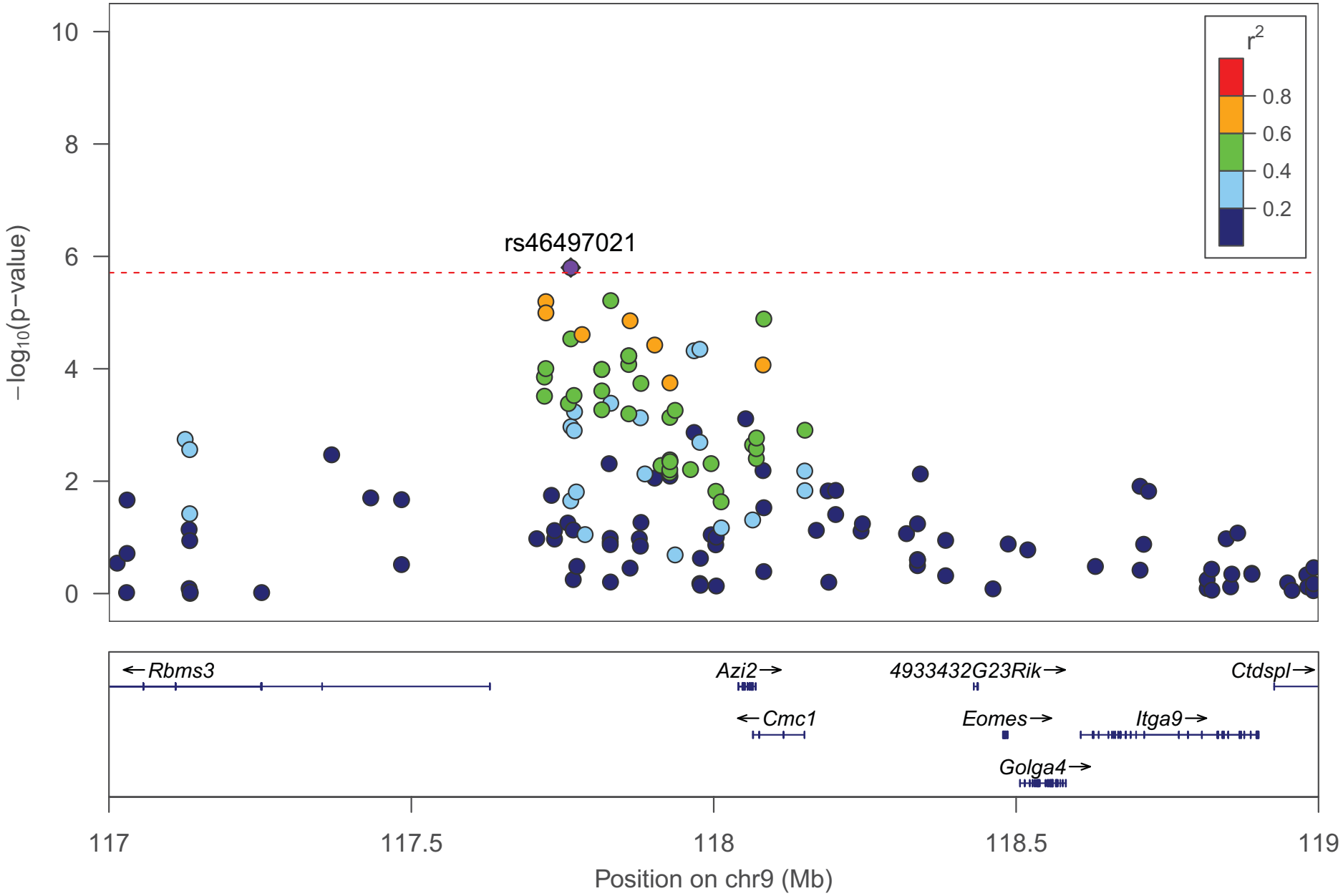
Nature Genetics: doi:10.1038/ng.3609

Total distance traveled on day 3



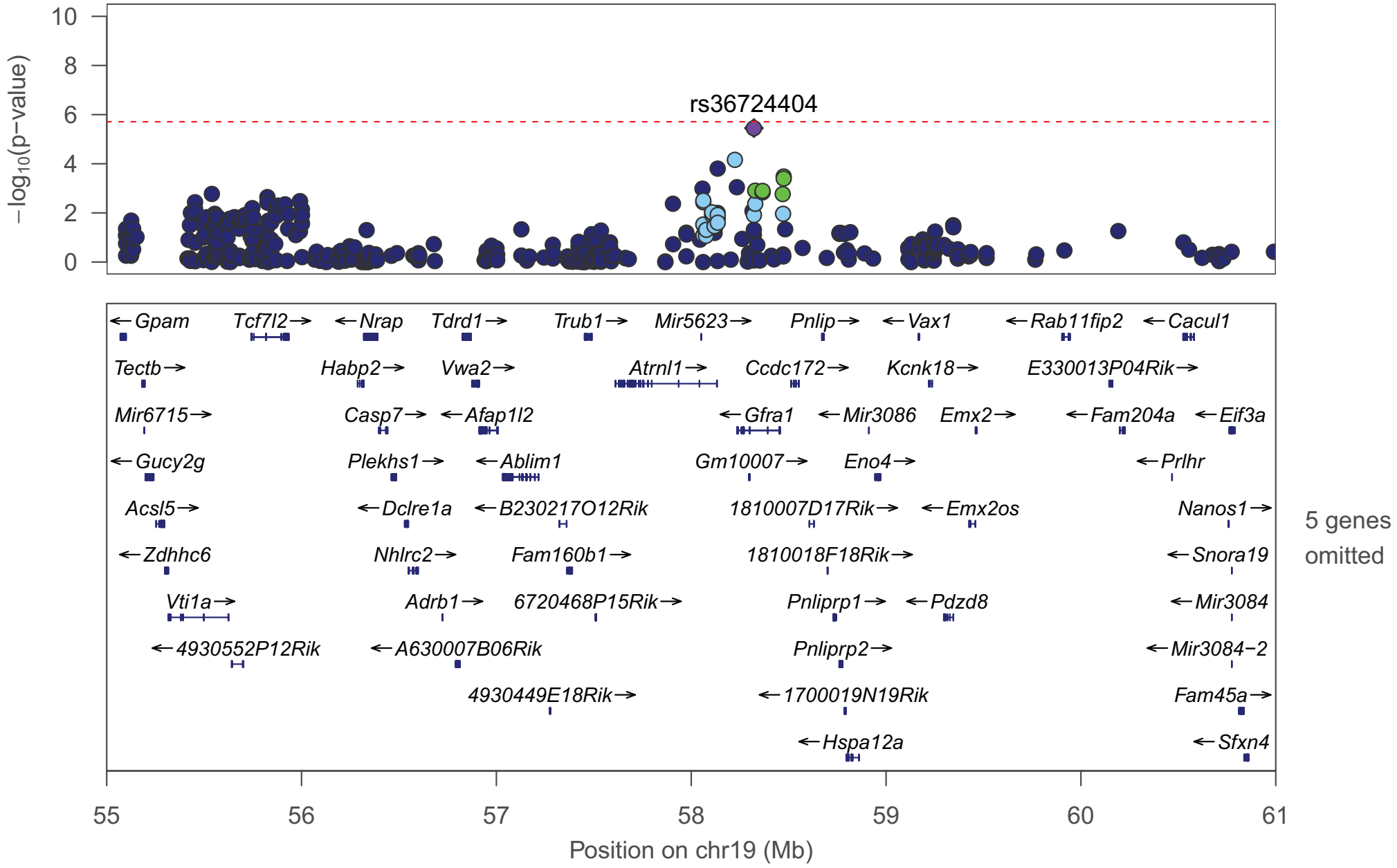
Nature Genetics: doi:10.1038/ng.3609

Total distance traveled on day 3



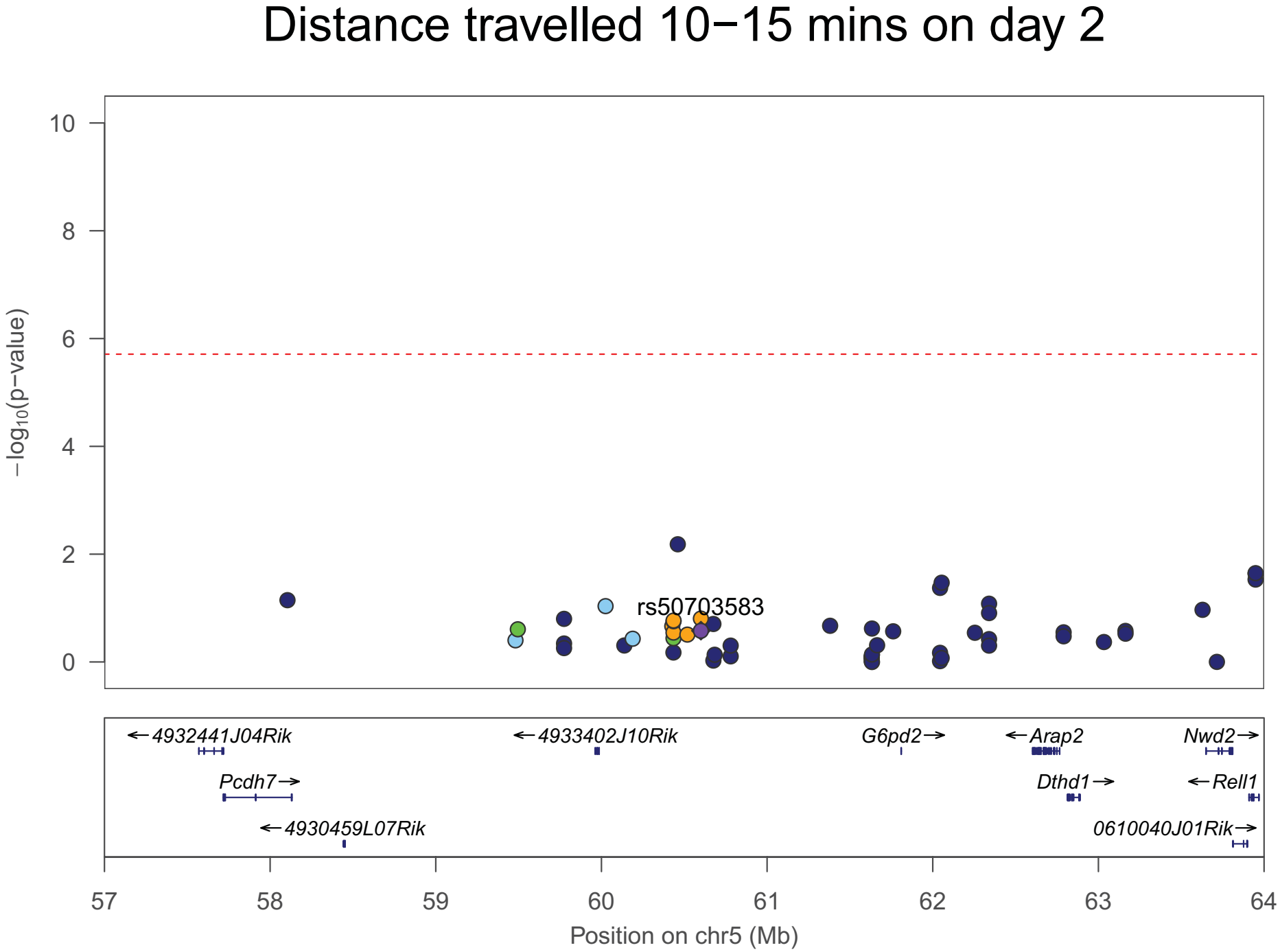
Nature Genetics: doi:10.1038/ng.3609

Freezing to tone on day 3



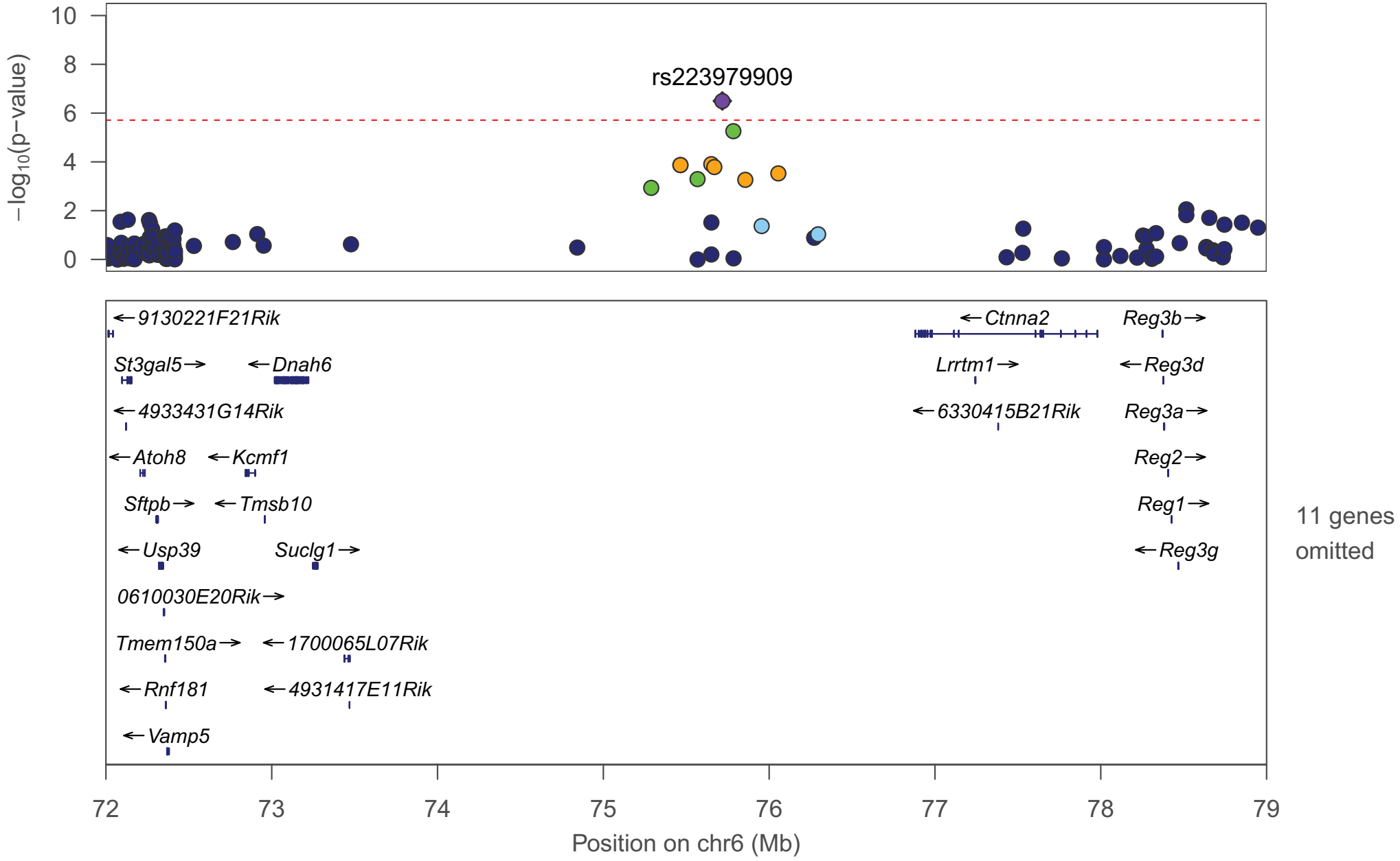
5 genes omitted

Nature Genetics: doi:10.1038/ng.3609



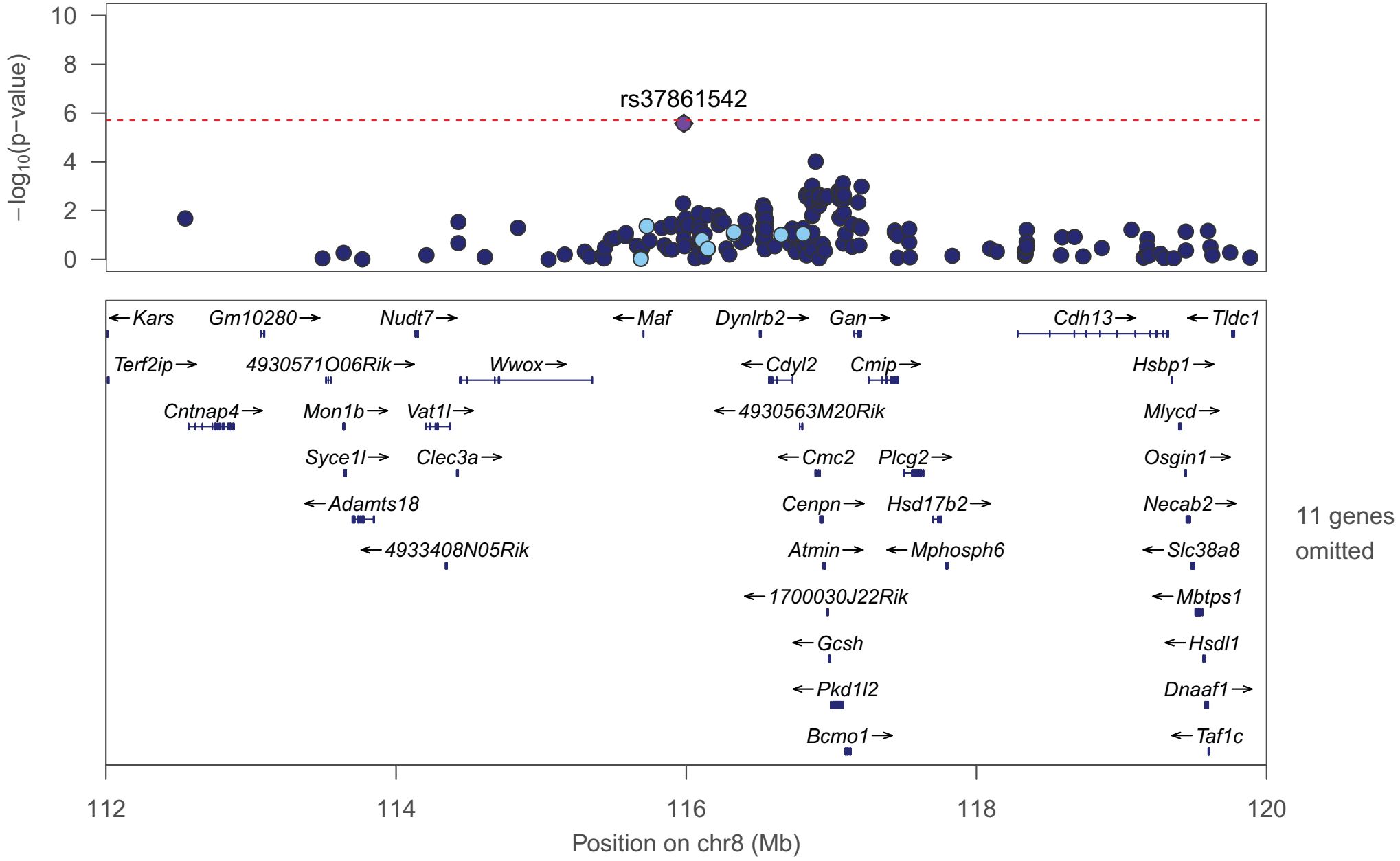
Nature Genetics: doi:10.1038/ng.3609

Distance travelled 10−15 mins on day 3



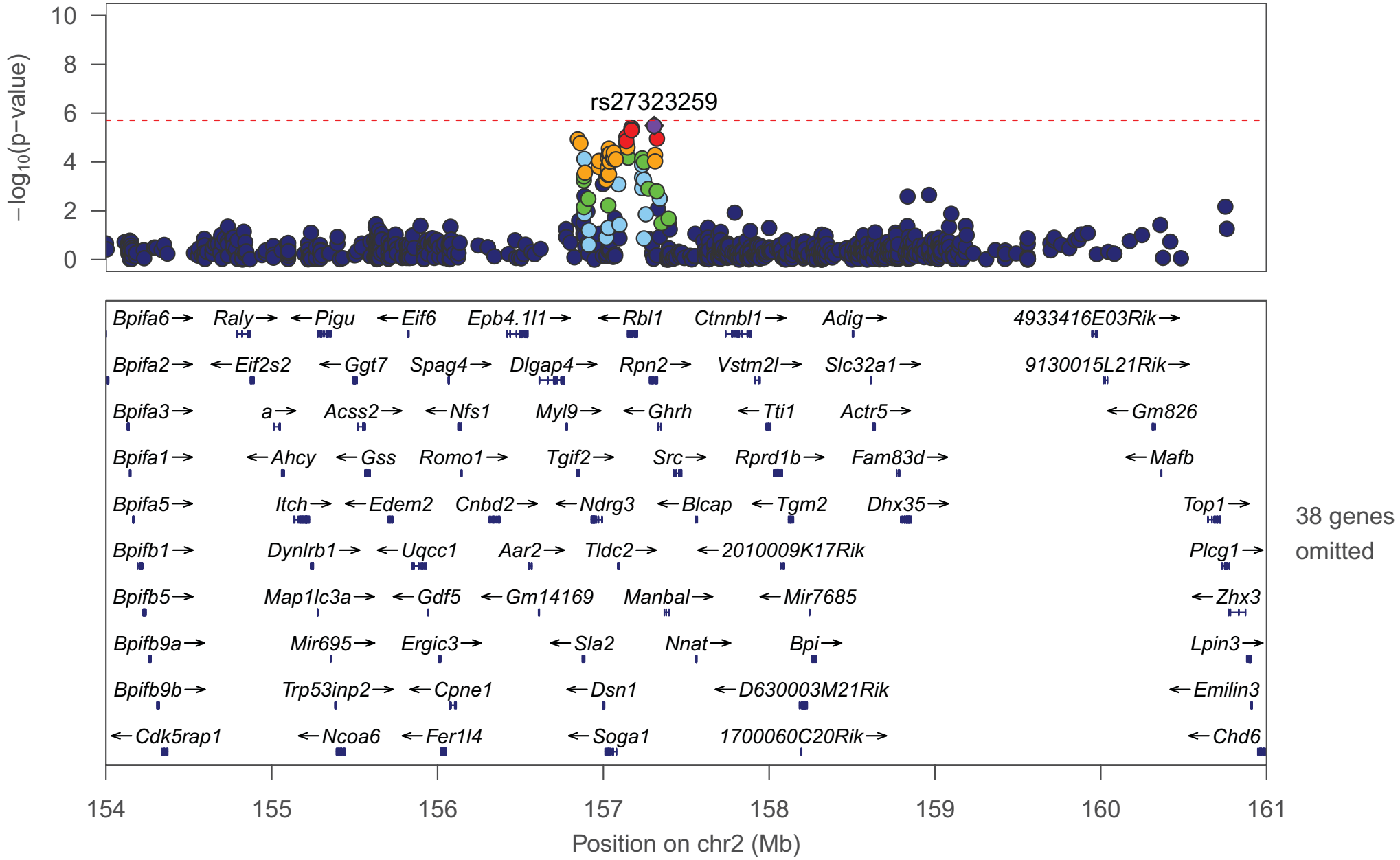
Nature Genetics: doi:10.1038/ng.3609

Distance travelled 10−15 mins on day 3



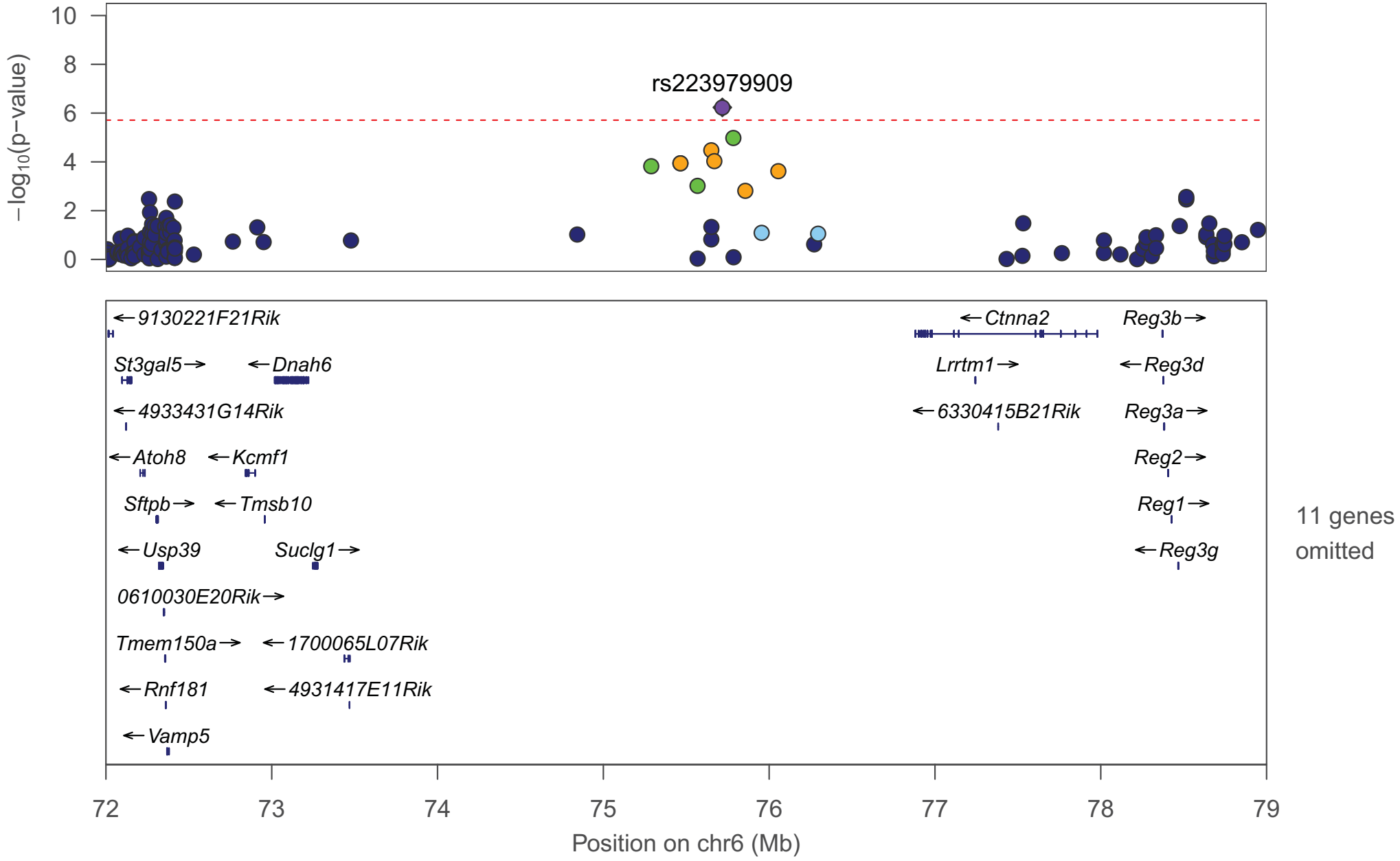
Nature Genetics: doi:10.1038/ng.3609

Distance travelled 15−20 mins on day 3



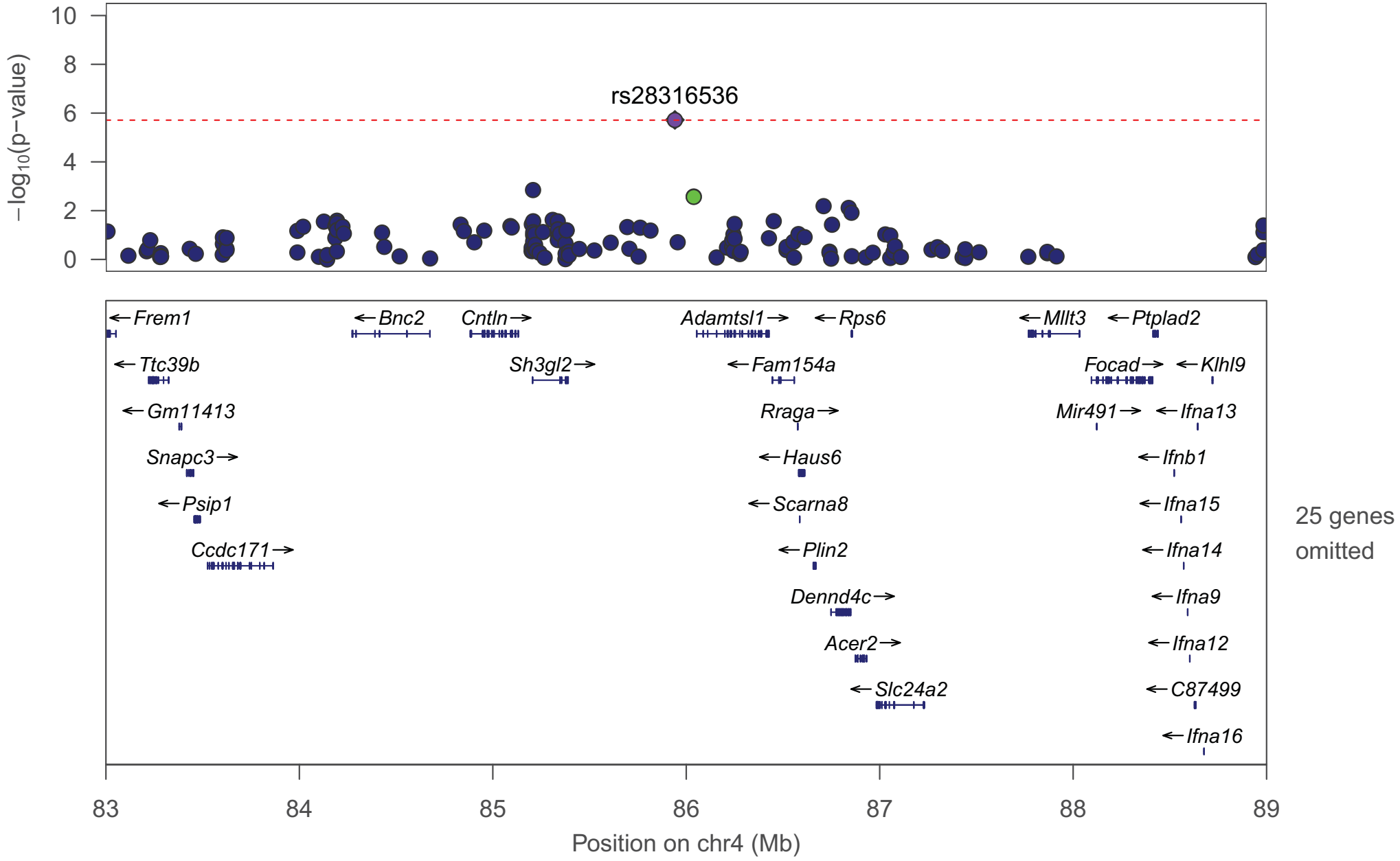
Nature Genetics: doi:10.1038/ng.3609

Distance travelled 20−25 mins on day 3

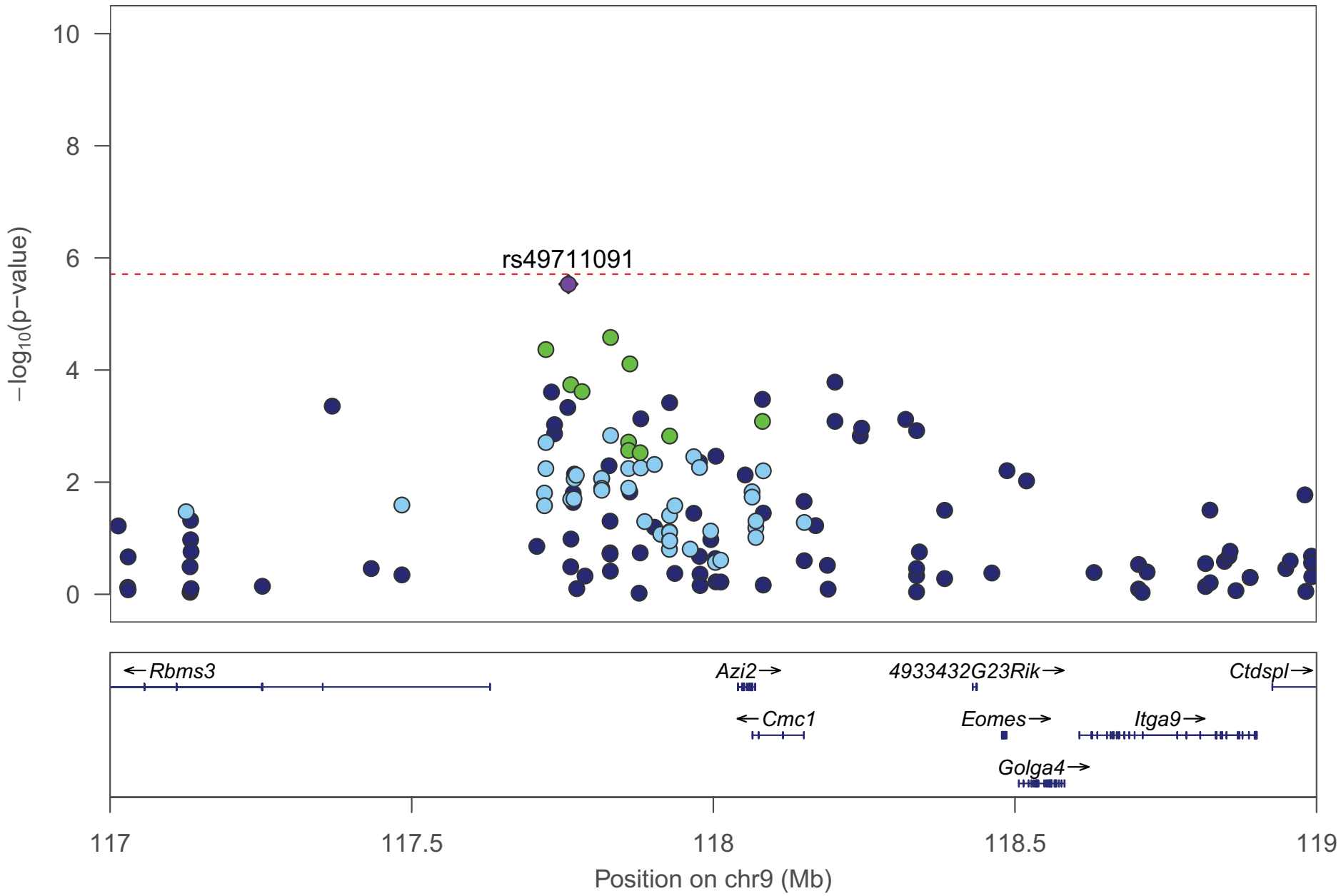


Nature Genetics: doi:10.1038/ng.3609

Distance travelled 25−30 mins on day 3

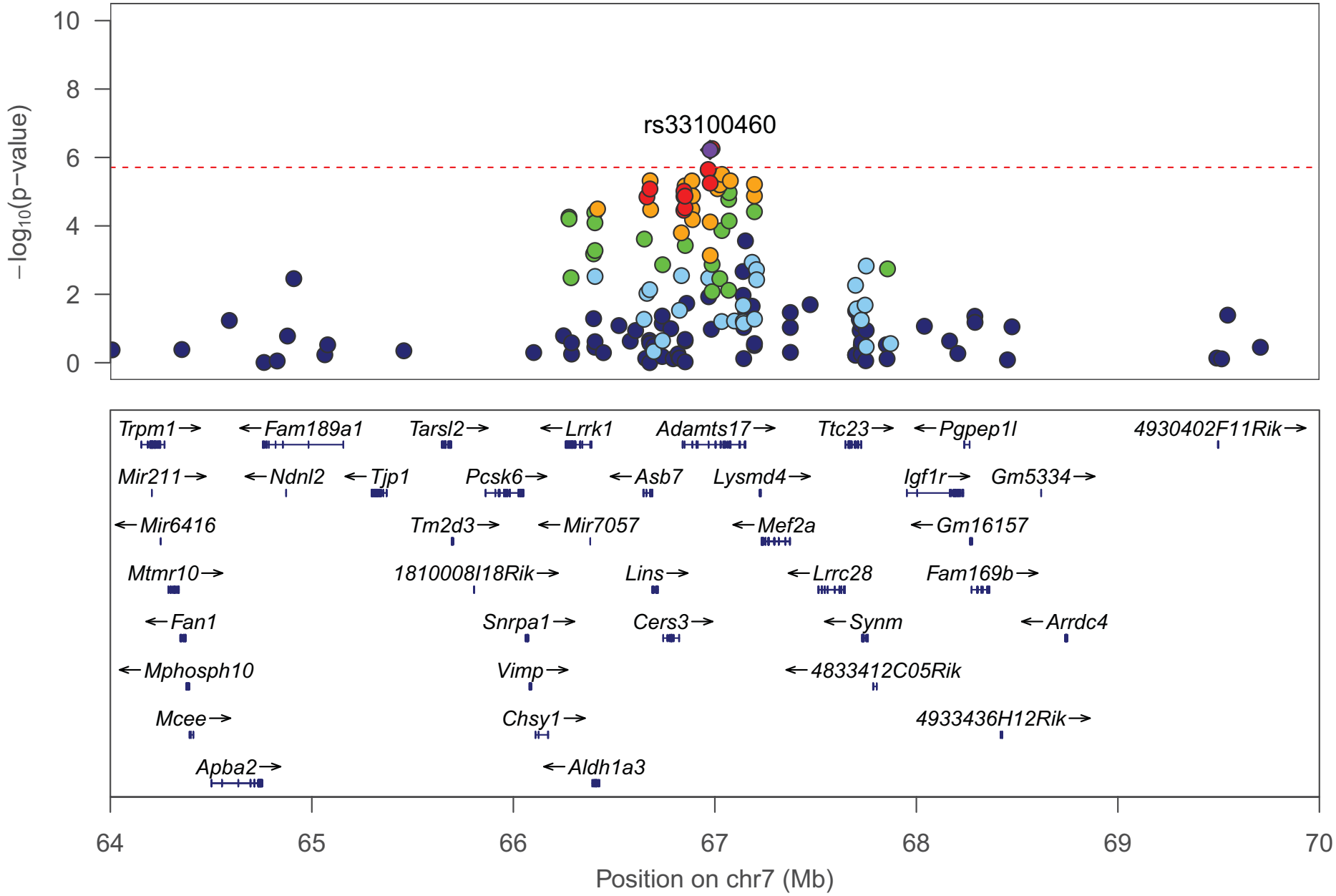


Nature Genetics: doi:10.1038/ng.3609



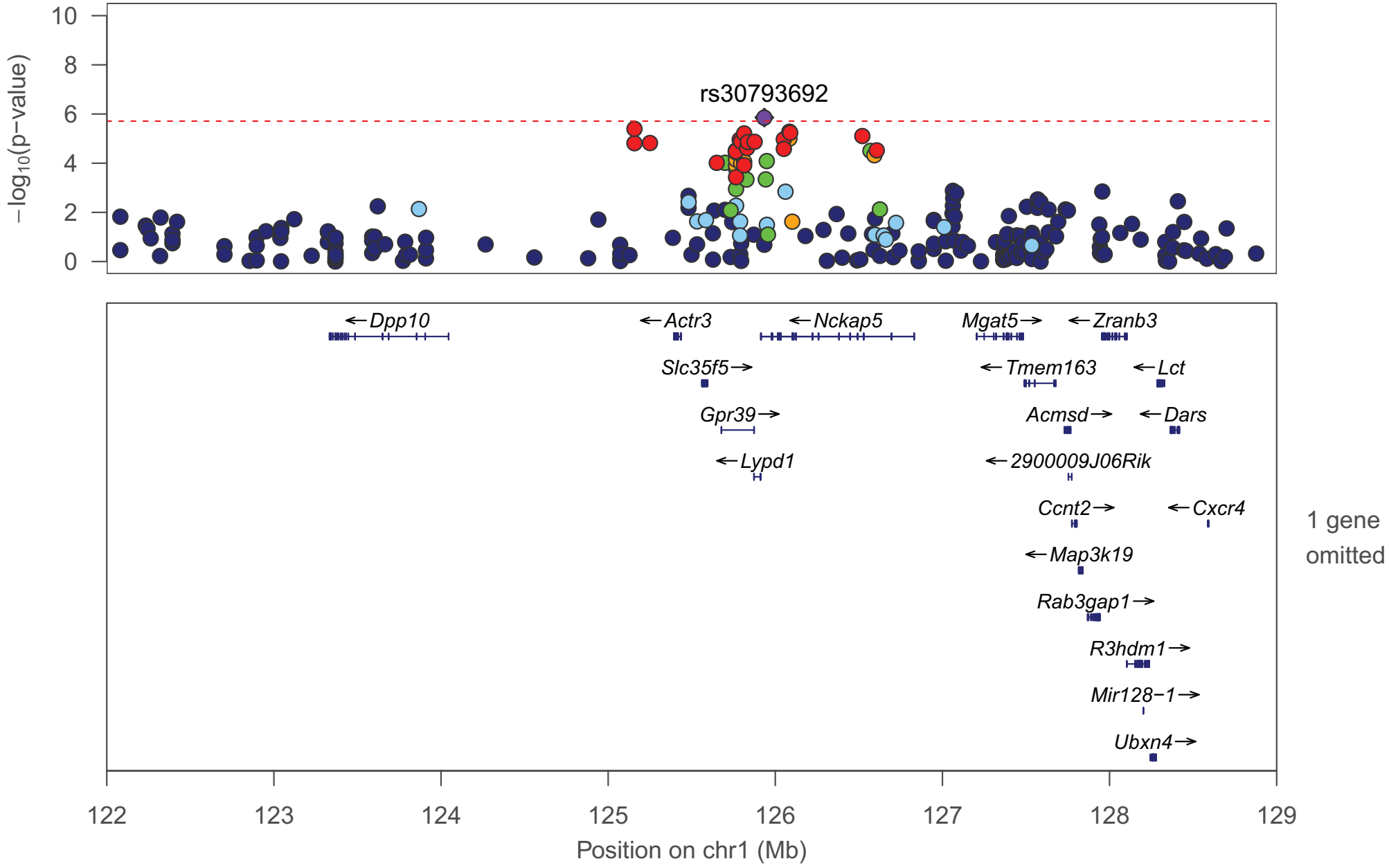
Nature Genetics: doi:10.1038/ng.3609

Vertical activity 0−15 mins on day 3



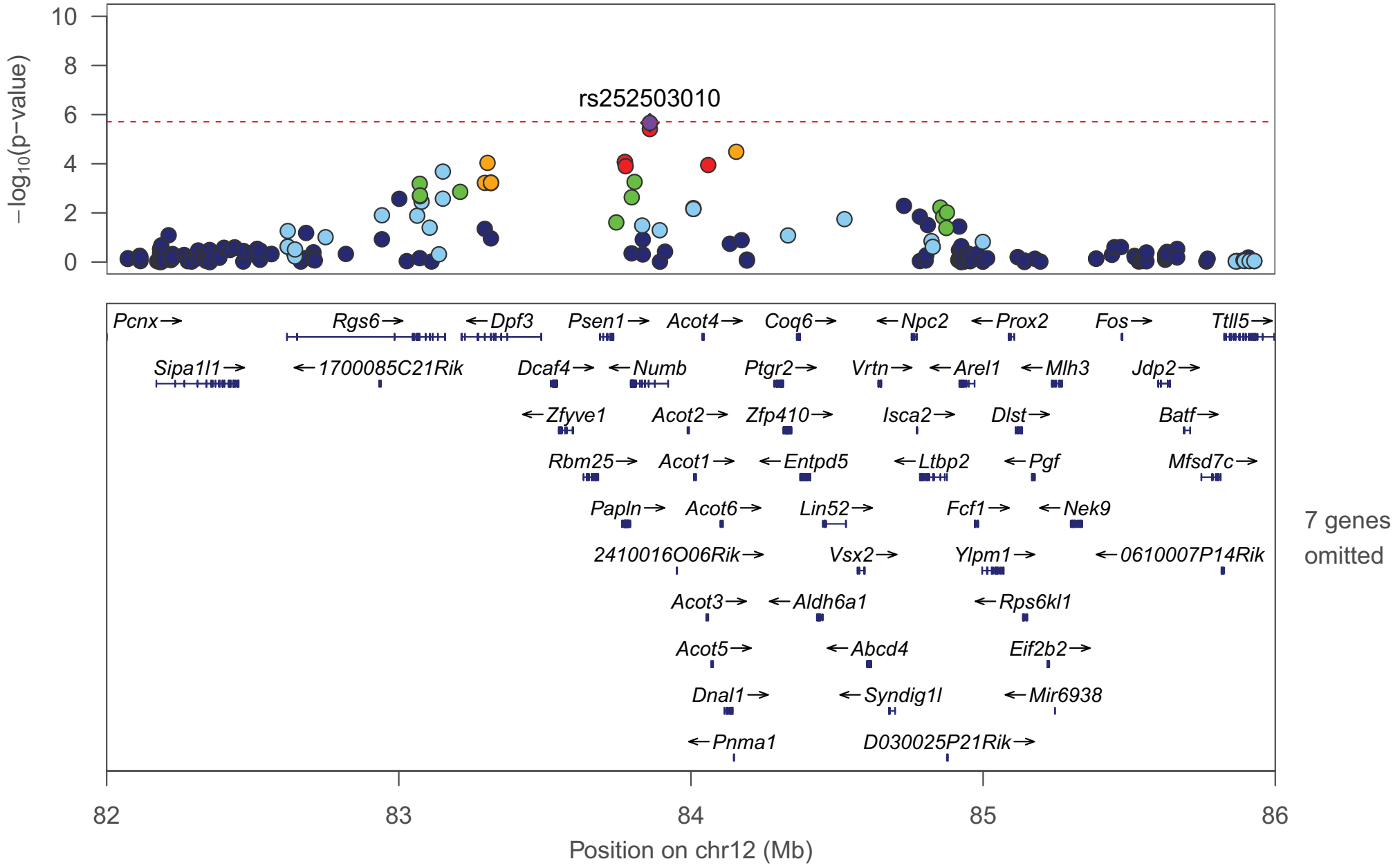
Nature Genetics: doi:10.1038/ng.3609

Vertical activity 0−30 mins on day 3



Nature Genetics: doi:10.1038/ng.3609

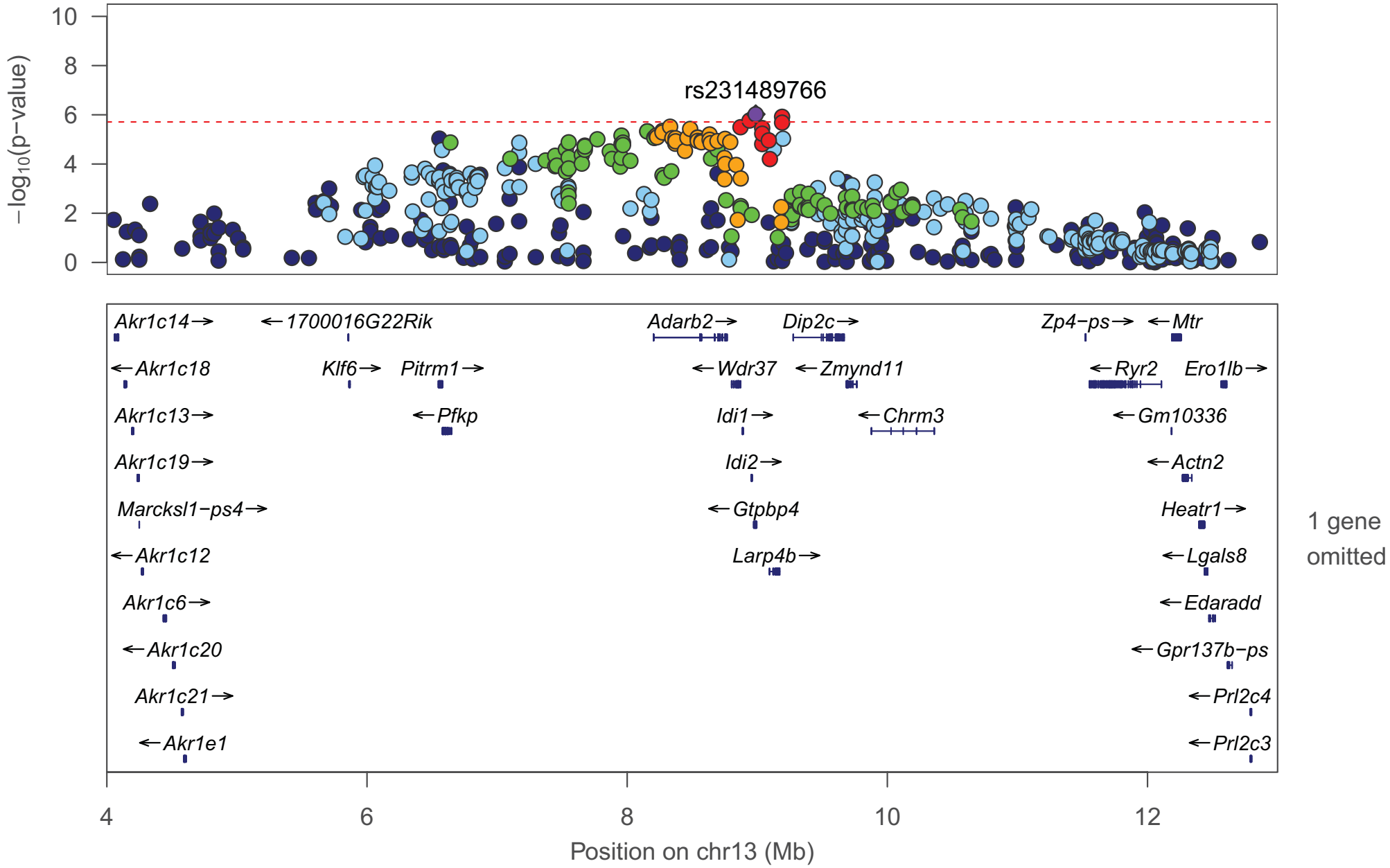
EDL muscle weight



7 genes omitted

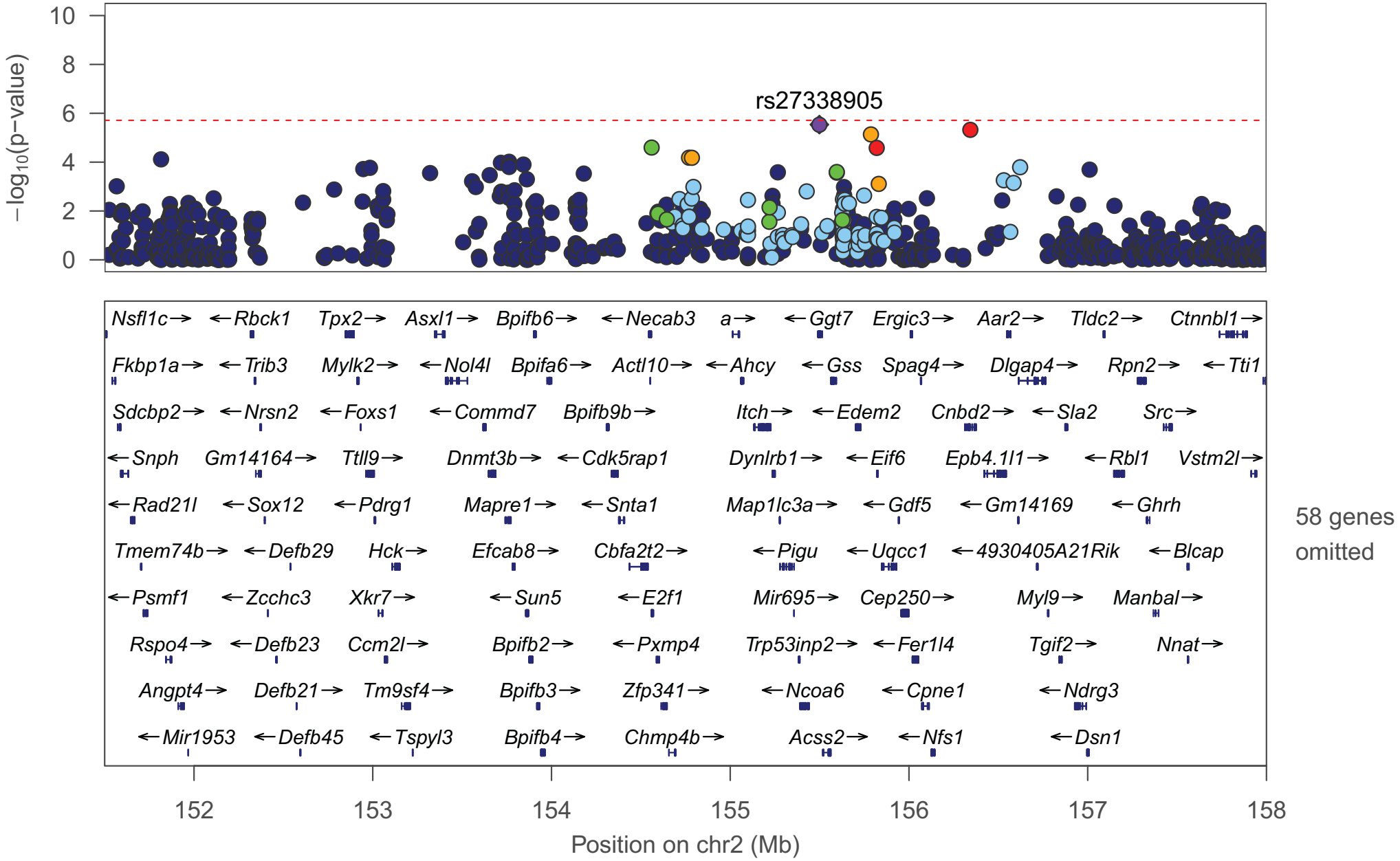
Nature Genetics: doi:10.1038/ng.3609

EDL muscle weight



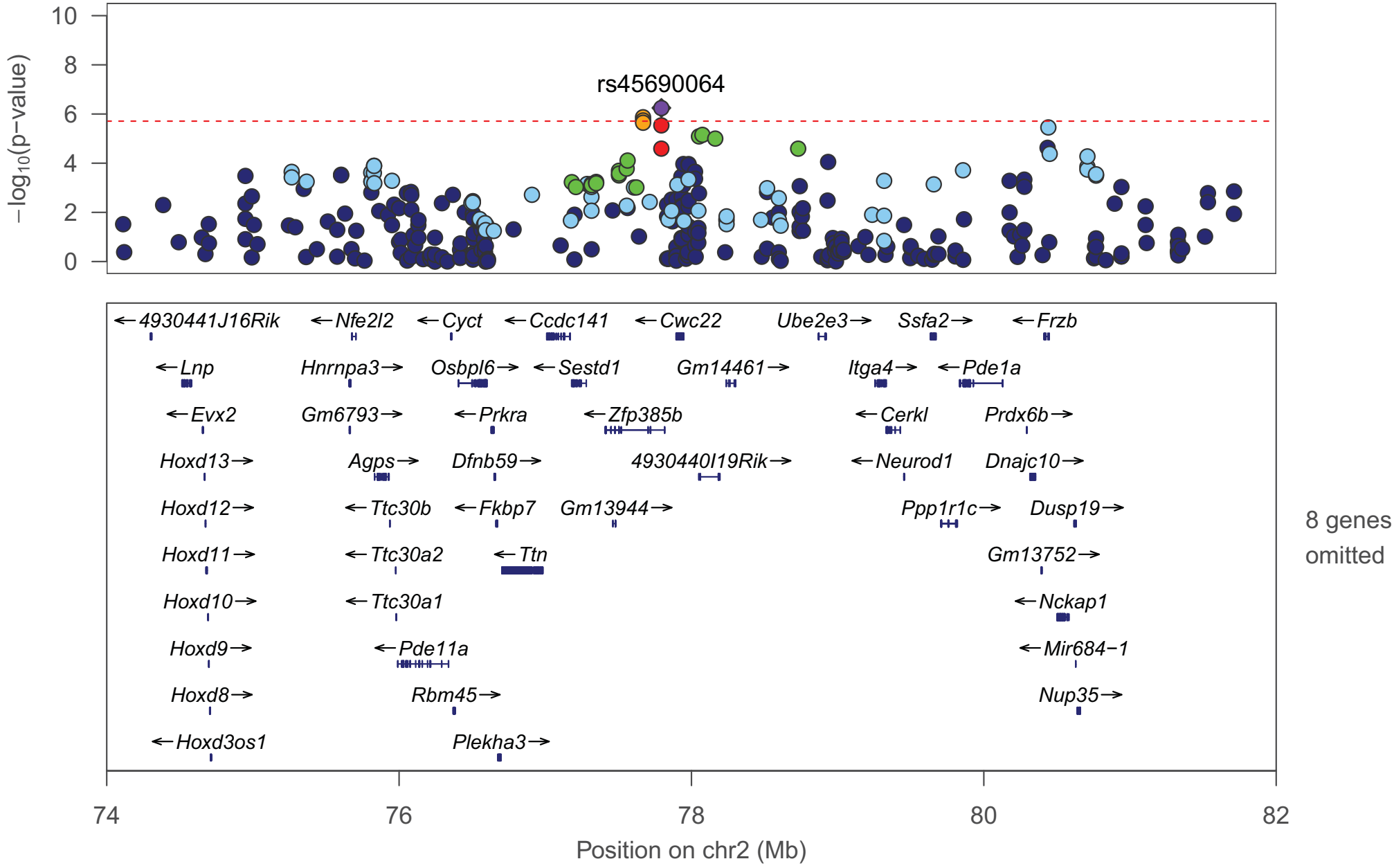
Nature Genetics: doi:10.1038/ng.3609

EDL muscle weight



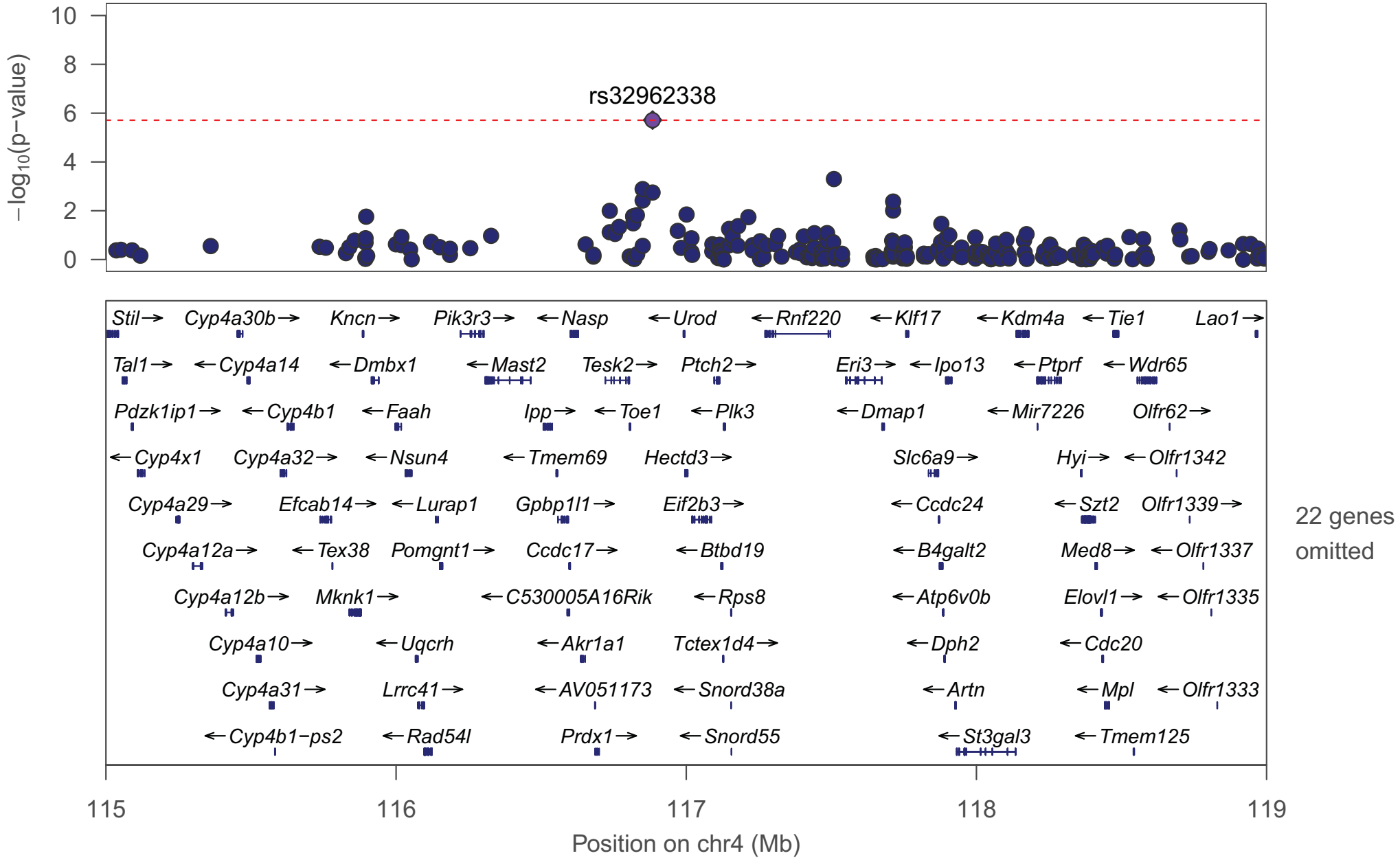
Nature Genetics: doi:10.1038/ng.3609

EDL muscle weight



Nature Genetics: doi:10.1038/ng.3609

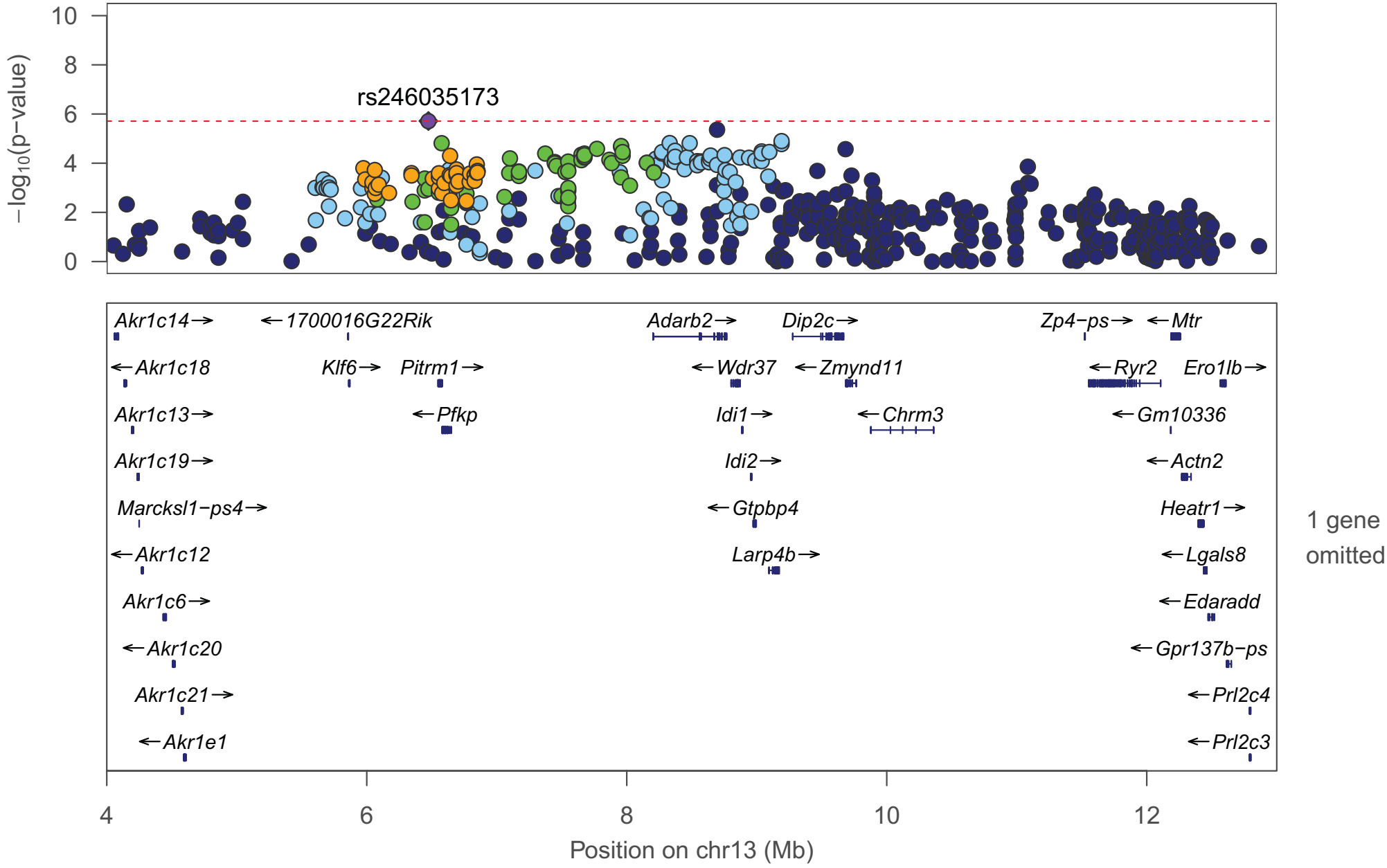
Fasting glucose



22 genes omitted

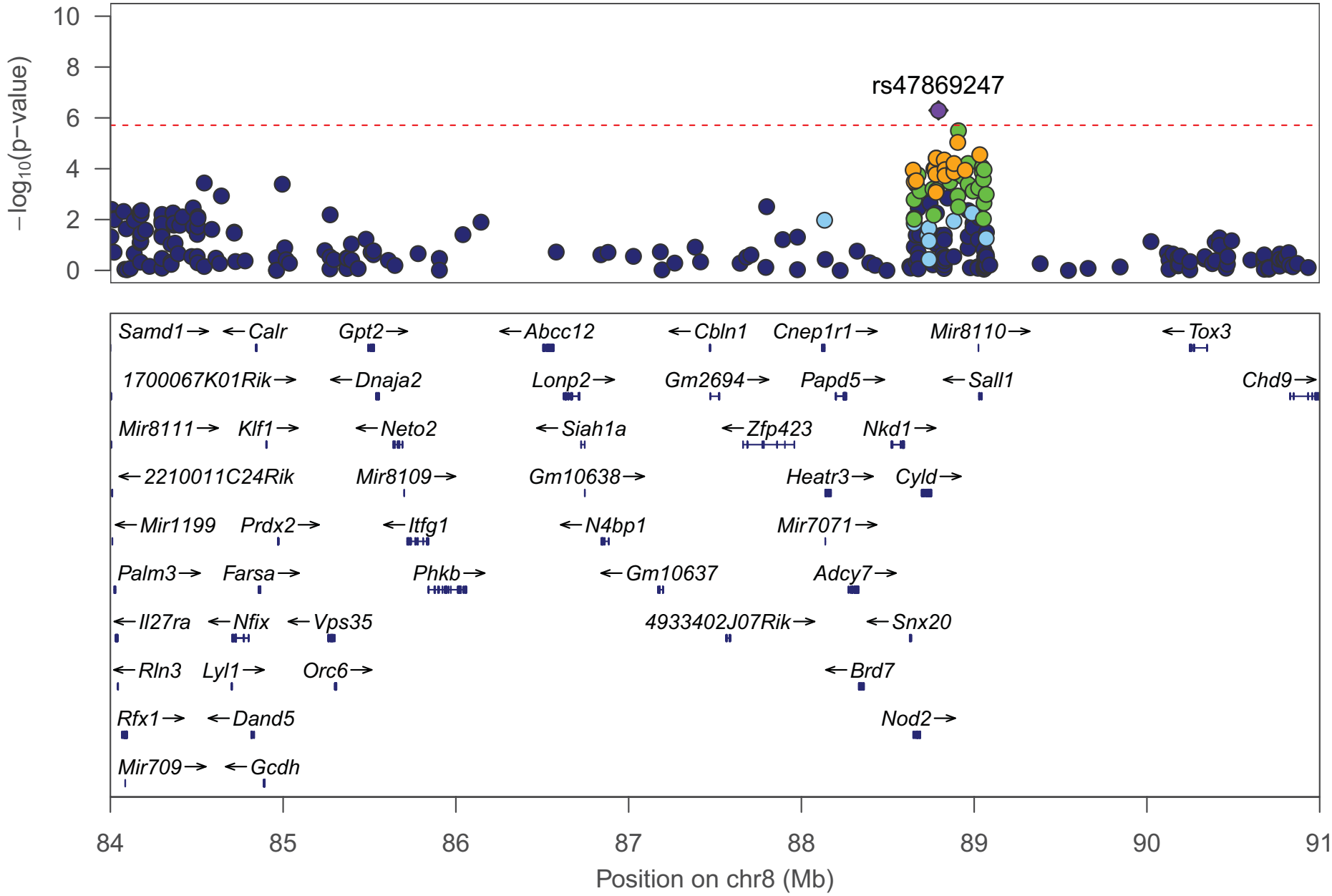
Nature Genetics: doi:10.1038/ng.3609

Gastrocnemius muscle weight



Nature Genetics: doi:10.1038/ng.3609

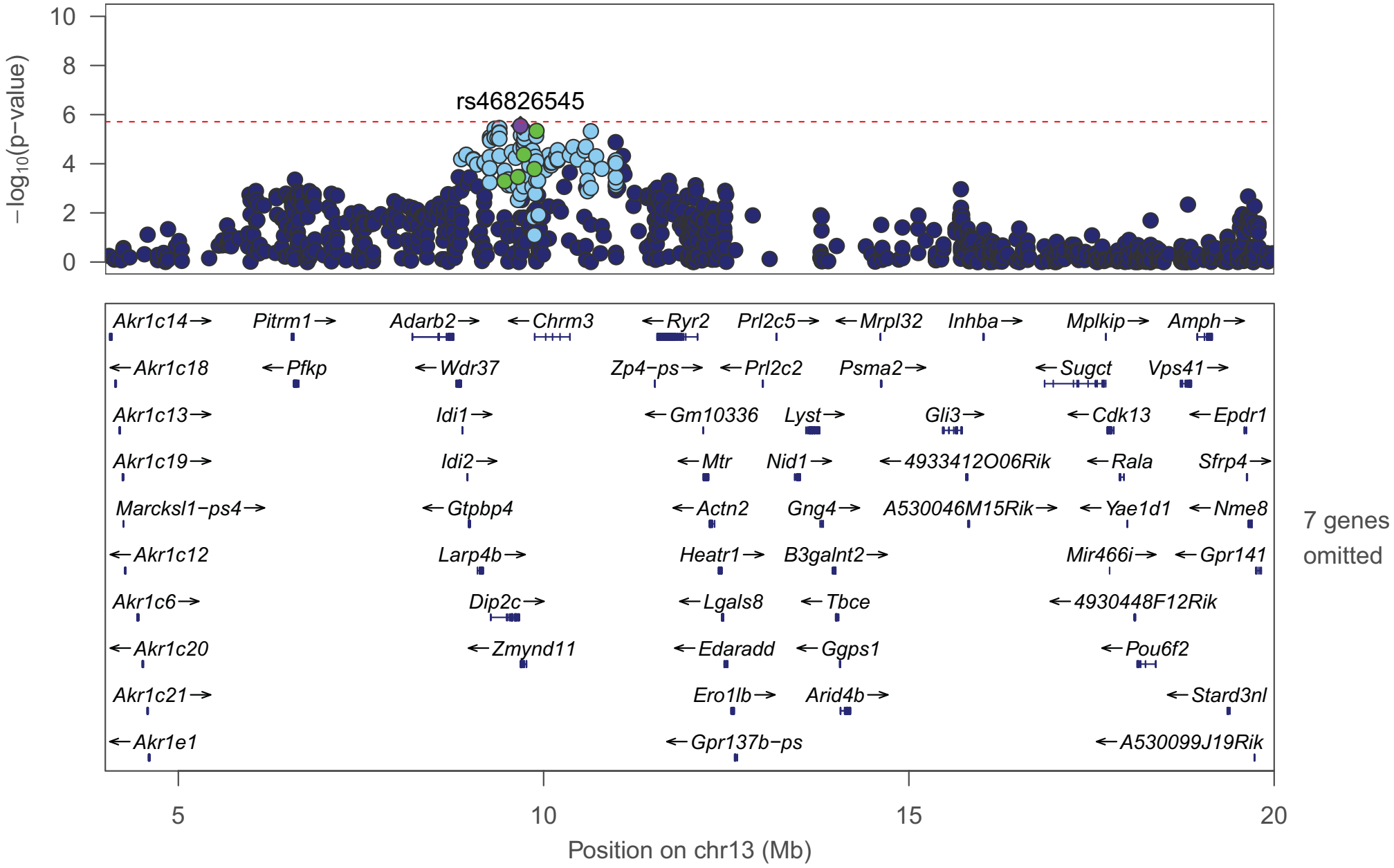
Gastrocnemius muscle weight



47 genes omitted

Nature Genetics: doi:10.1038/ng.3609

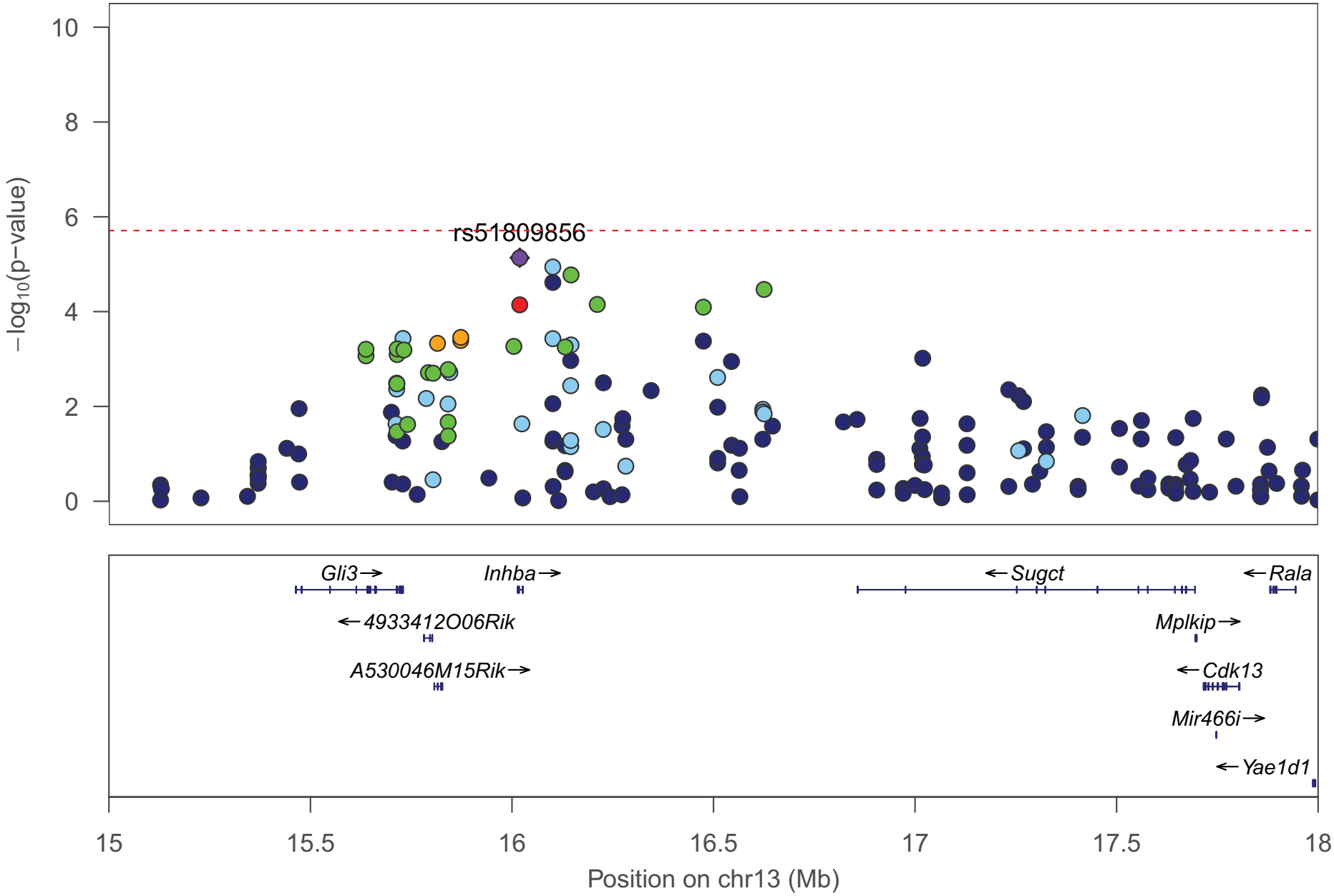
Soleus muscle weight conditioned on rs30535702 genotype



7 genes omitted

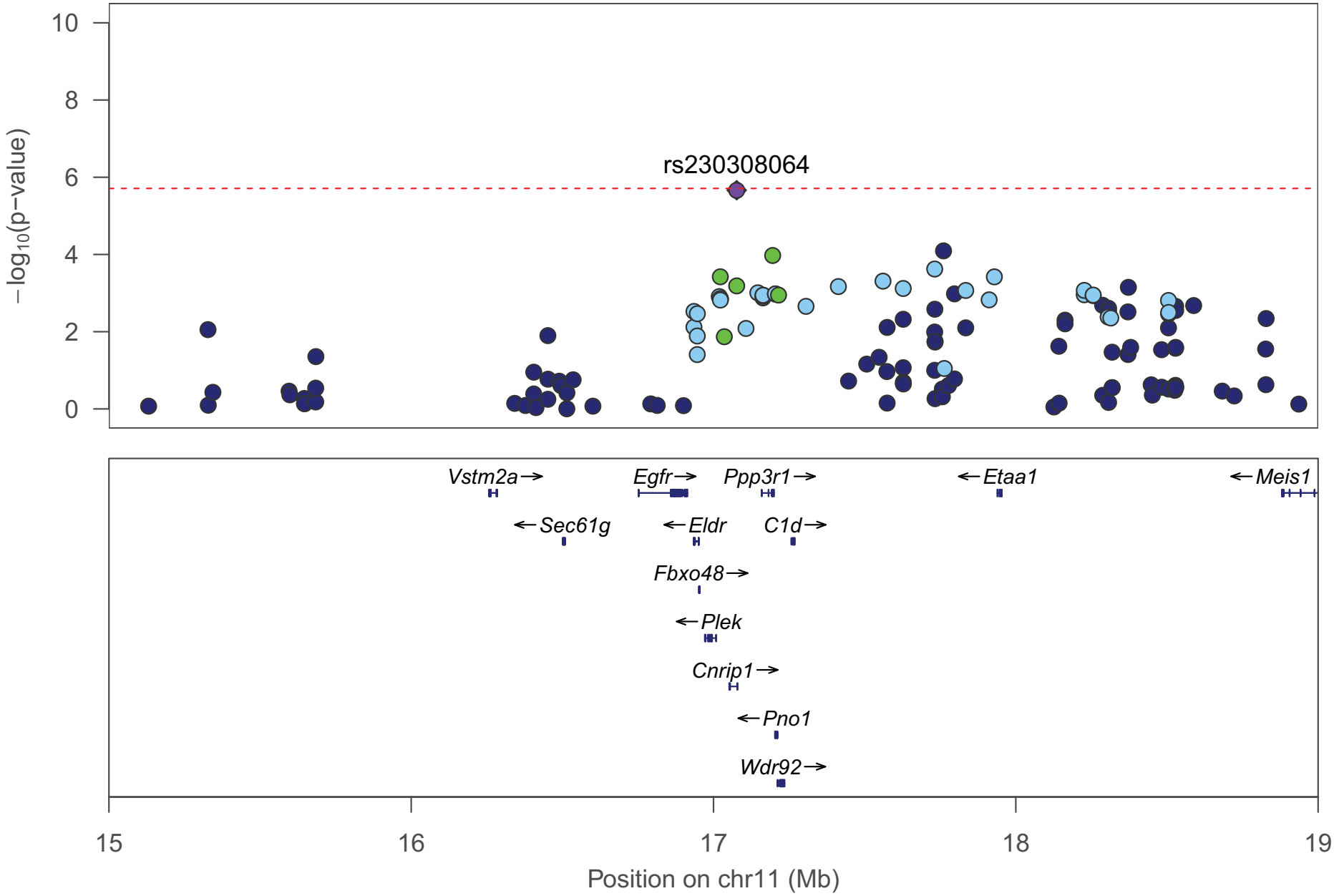
Nature Genetics: doi:10.1038/ng.3609

Testes weight conditioned on rs6279141 genotype



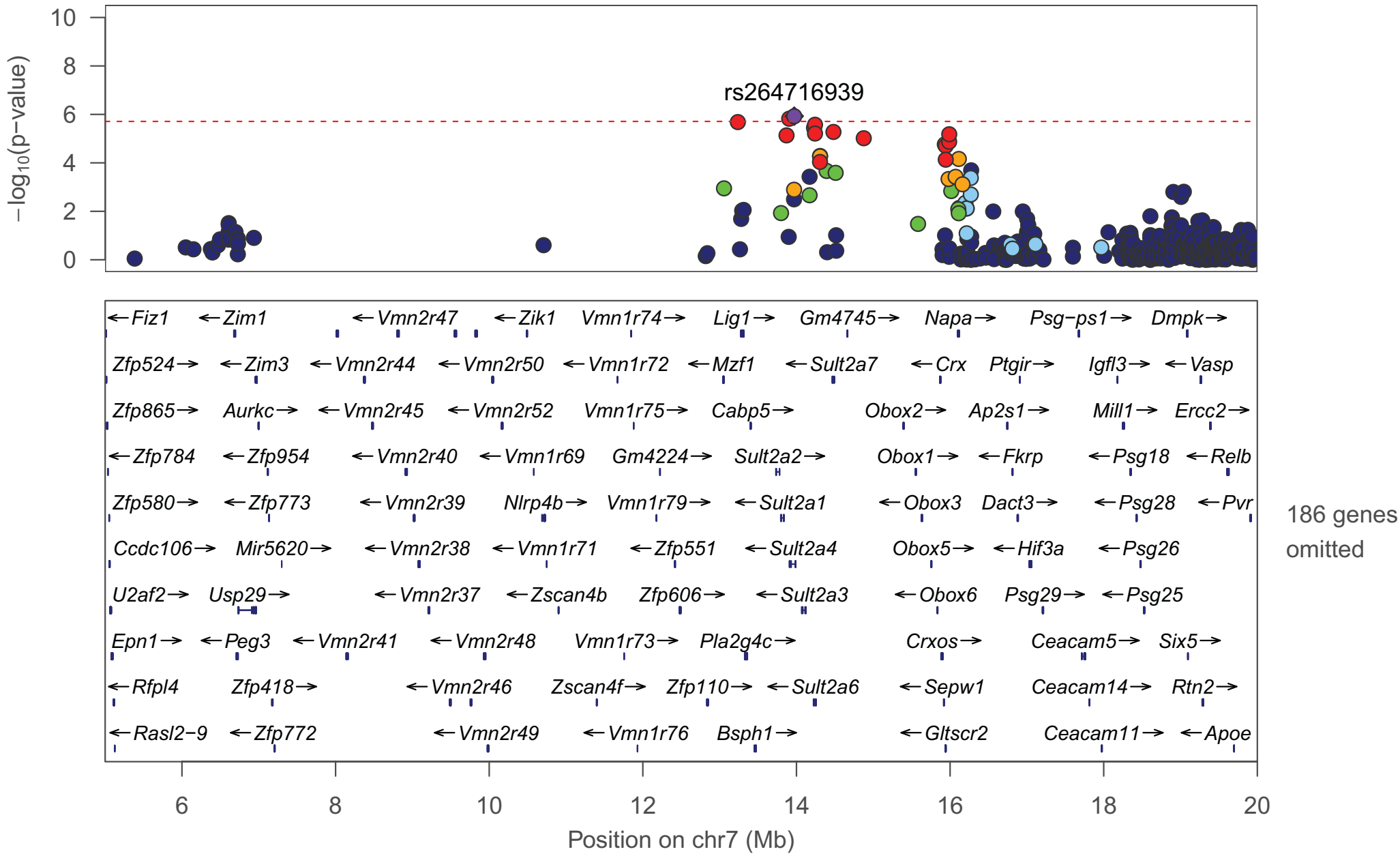
Nature Genetics: doi:10.1038/ng.3609

Pre−pulse inhibition with 12 decibel pulse



Nature Genetics: doi:10.1038/ng.3609

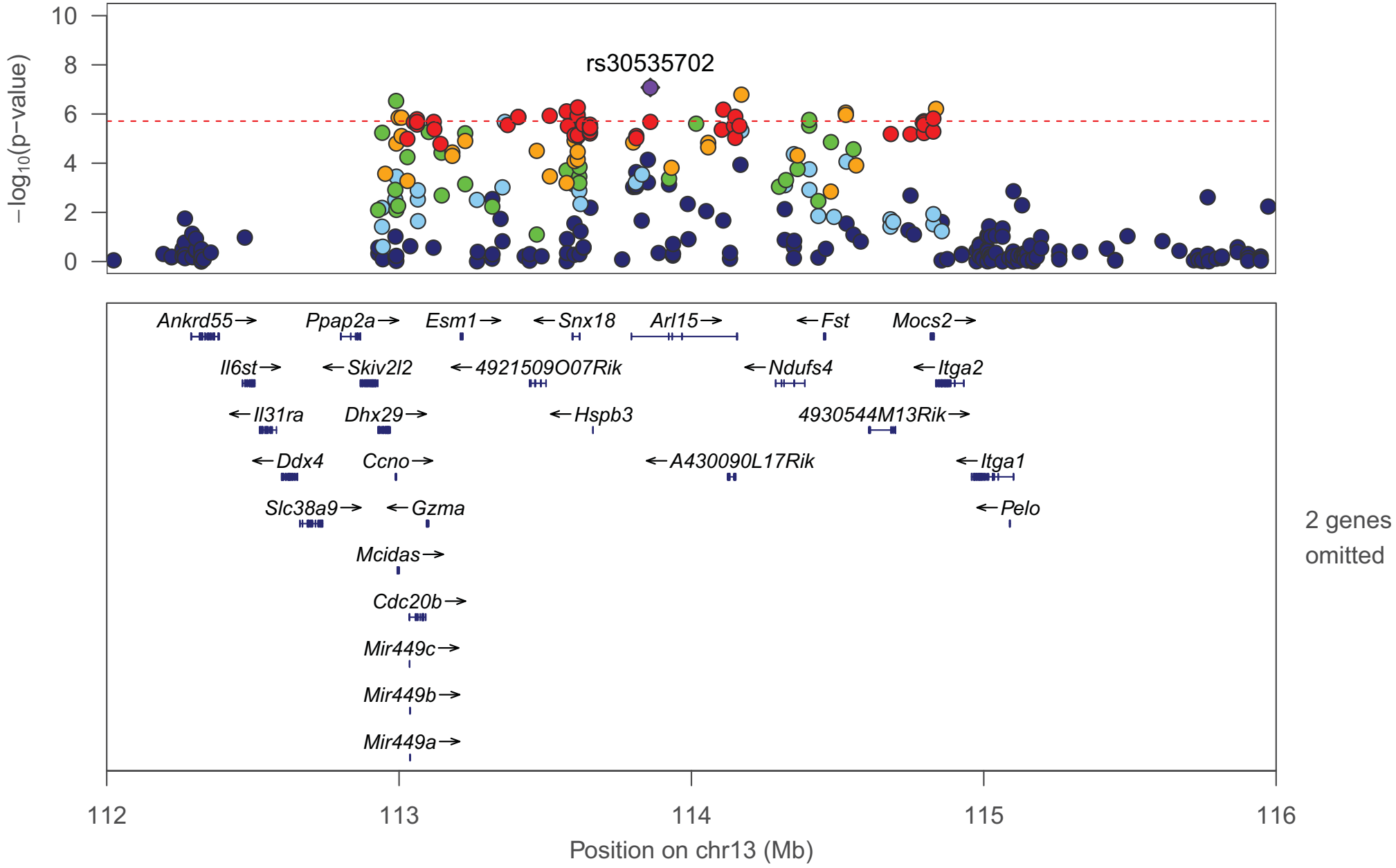
Pre−pulse inhibition with 12 decibel pulse



186 genes omitted

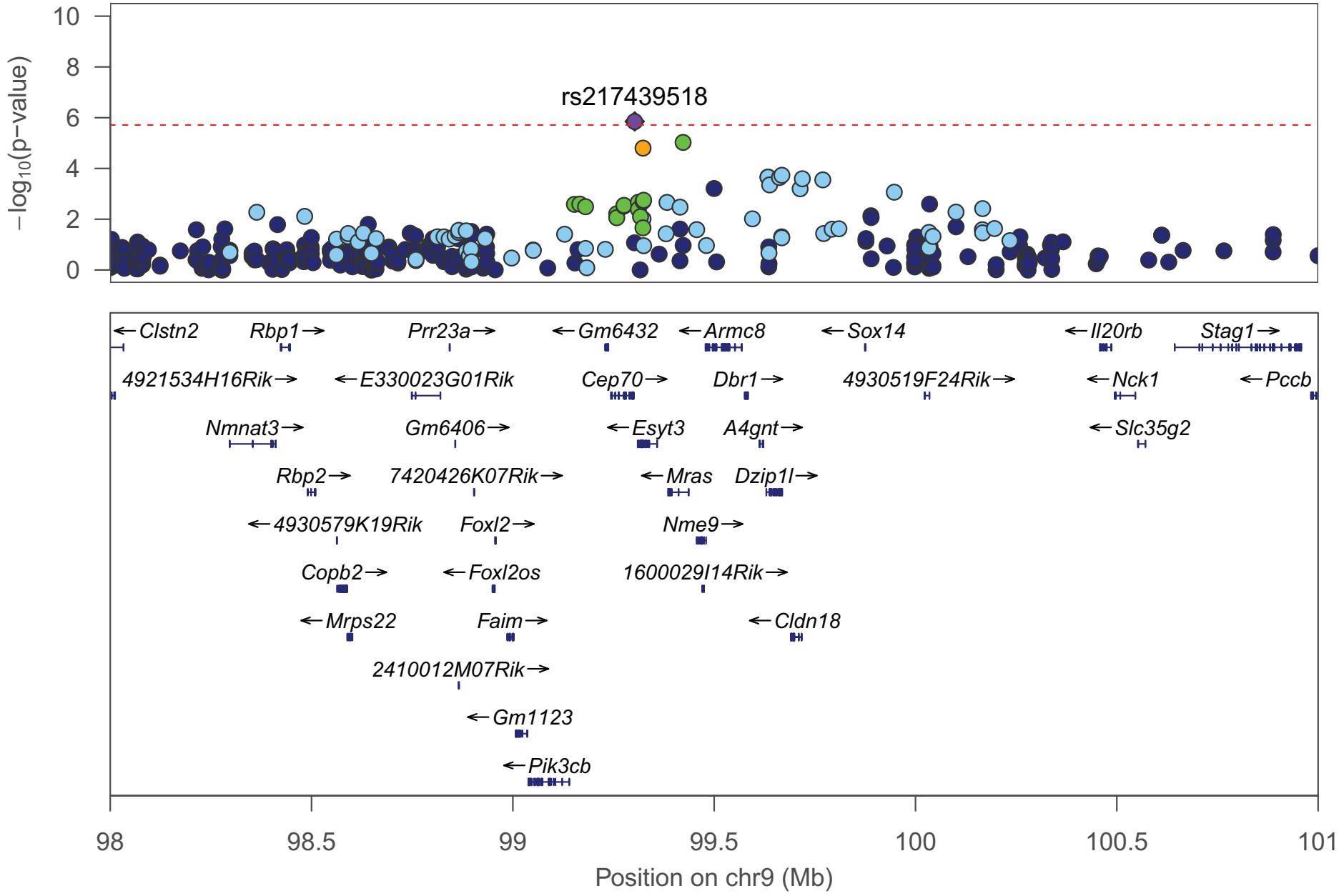
Nature Genetics: doi:10.1038/ng.3609

Soleus muscle weight



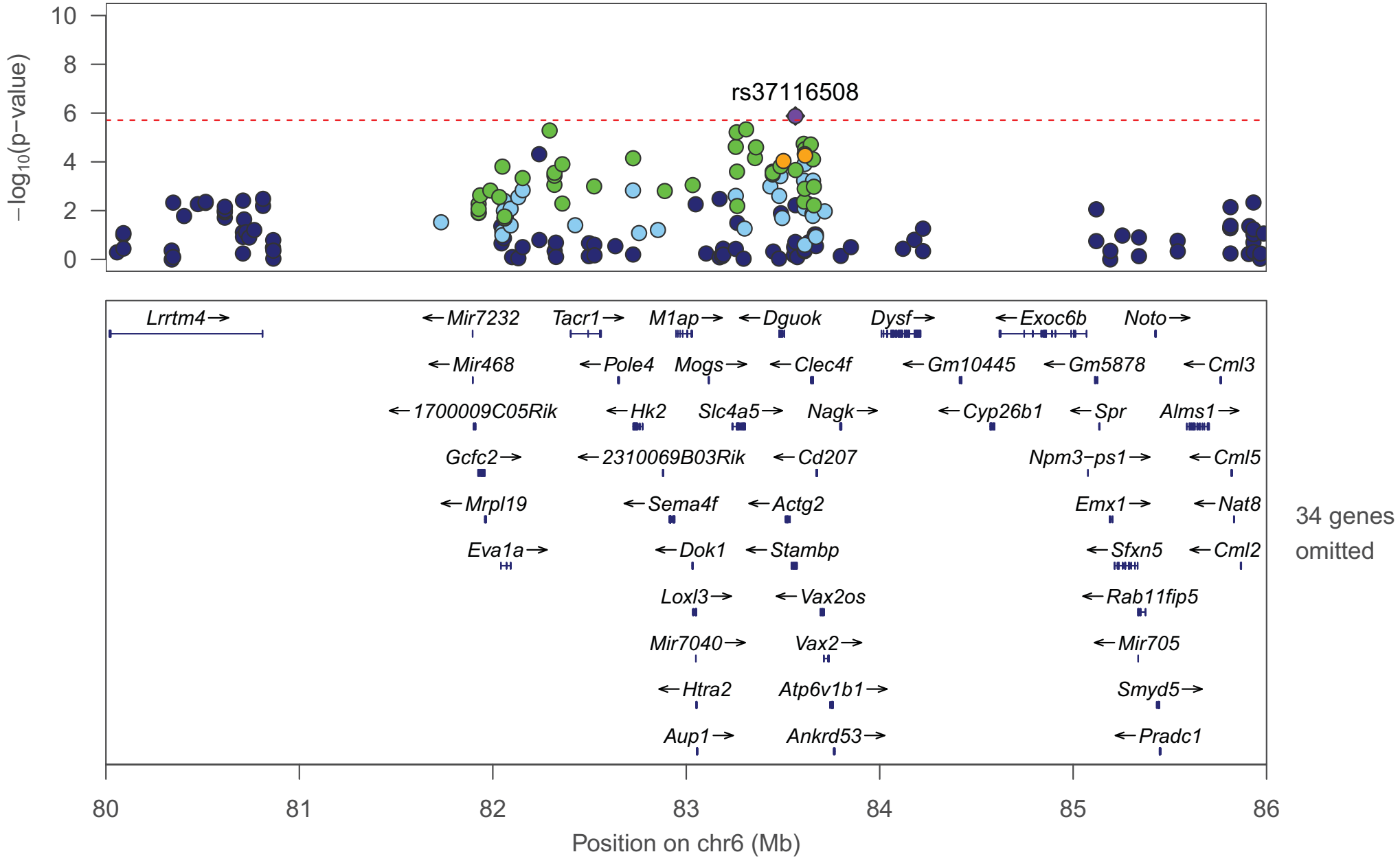
Nature Genetics: doi:10.1038/ng.3609

Soleus muscle weight



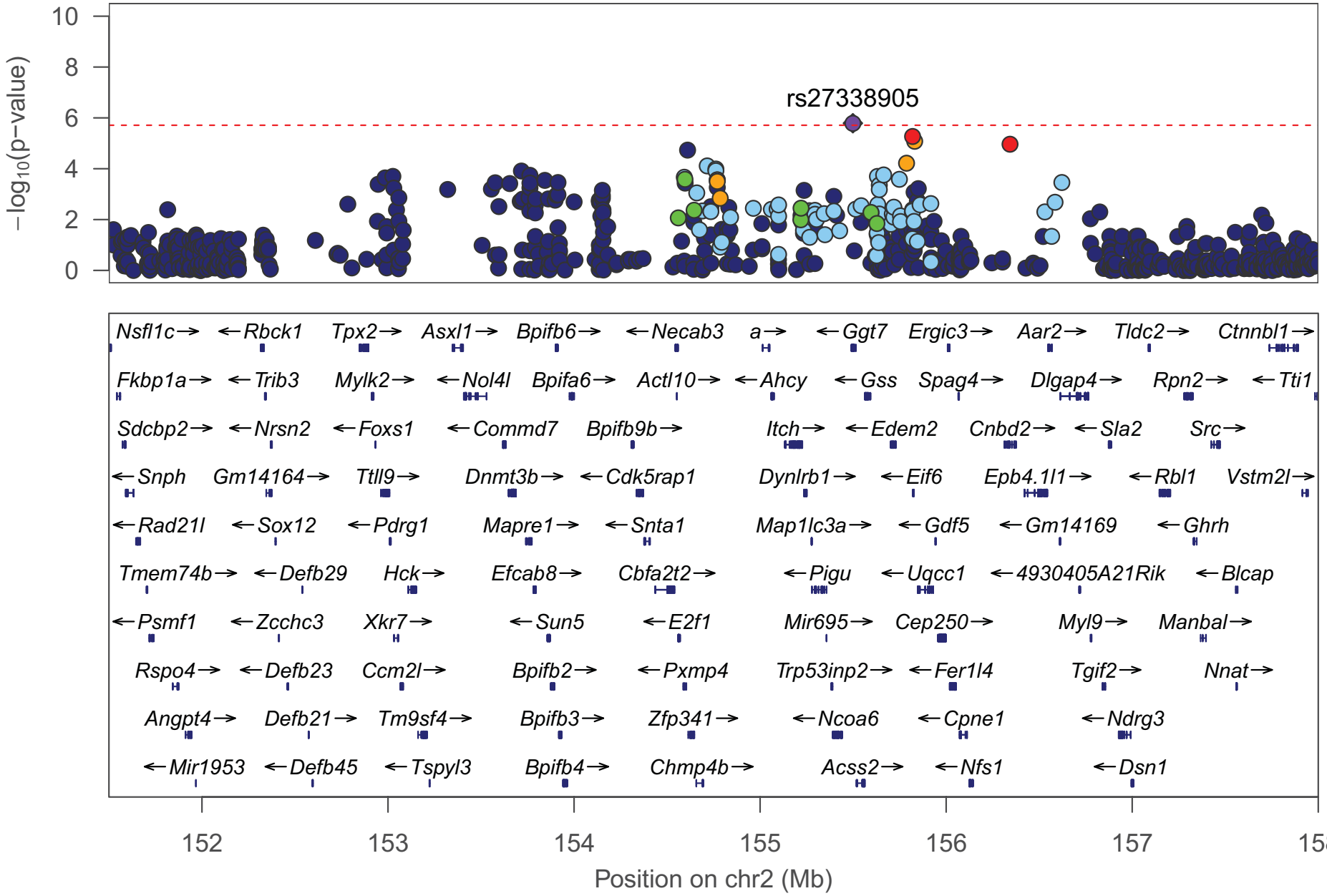
Nature Genetics: doi:10.1038/ng.3609

Startle response



Nature Genetics: doi:10.1038/ng.3609

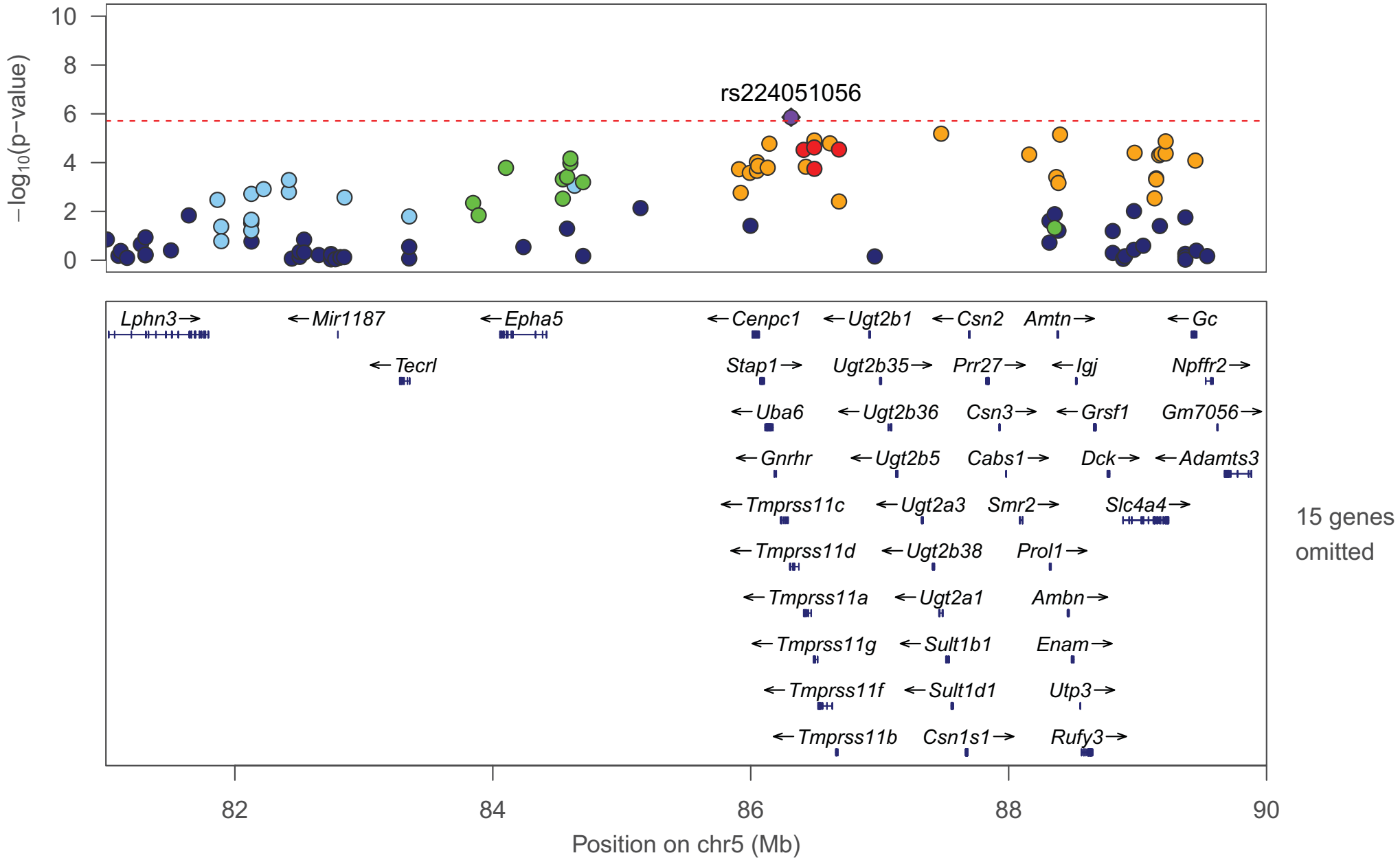
TA muscle weight



58 genes omitted

Nature Genetics: doi:10.1038/ng.3609

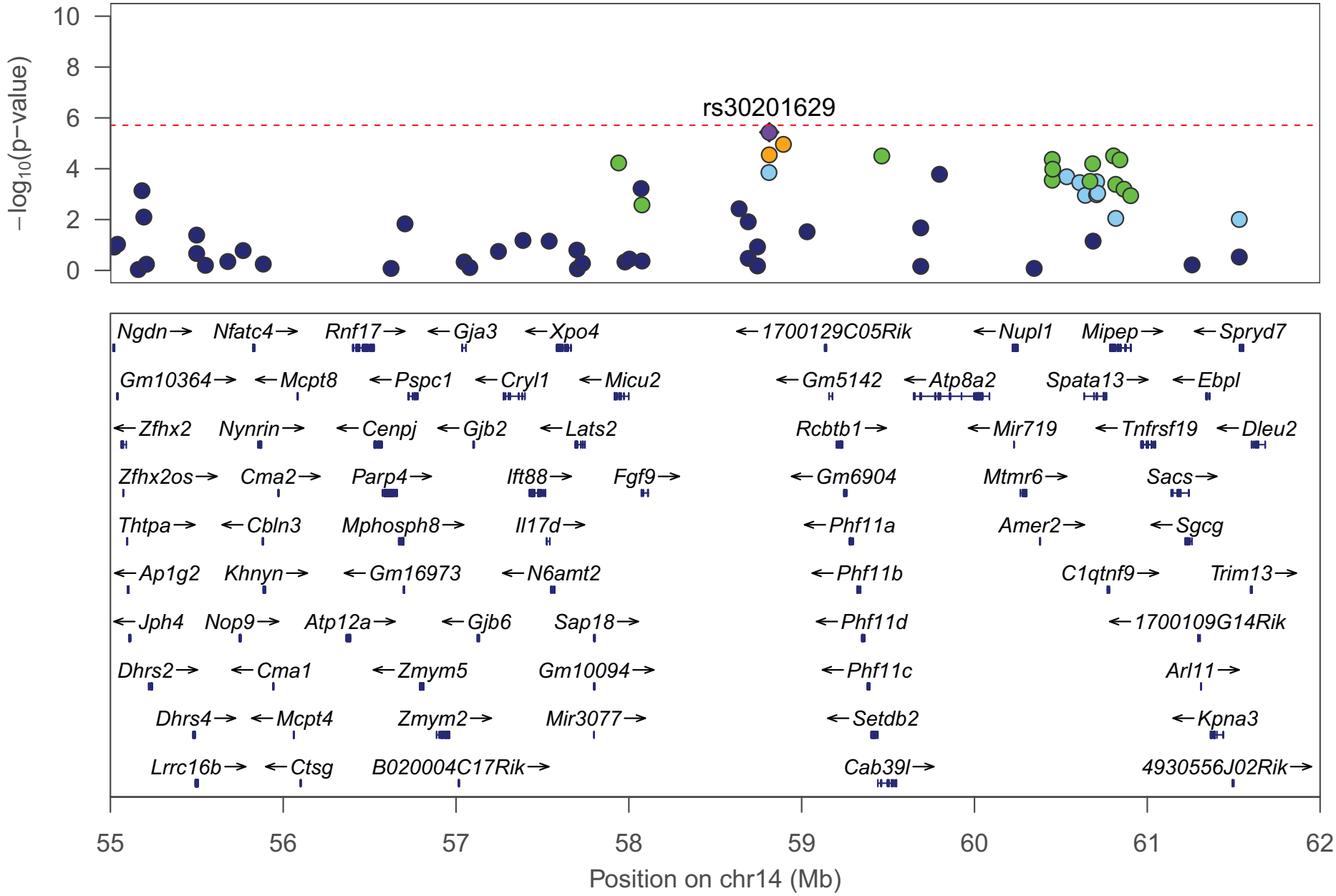
TA muscle weight



15 genes omitted

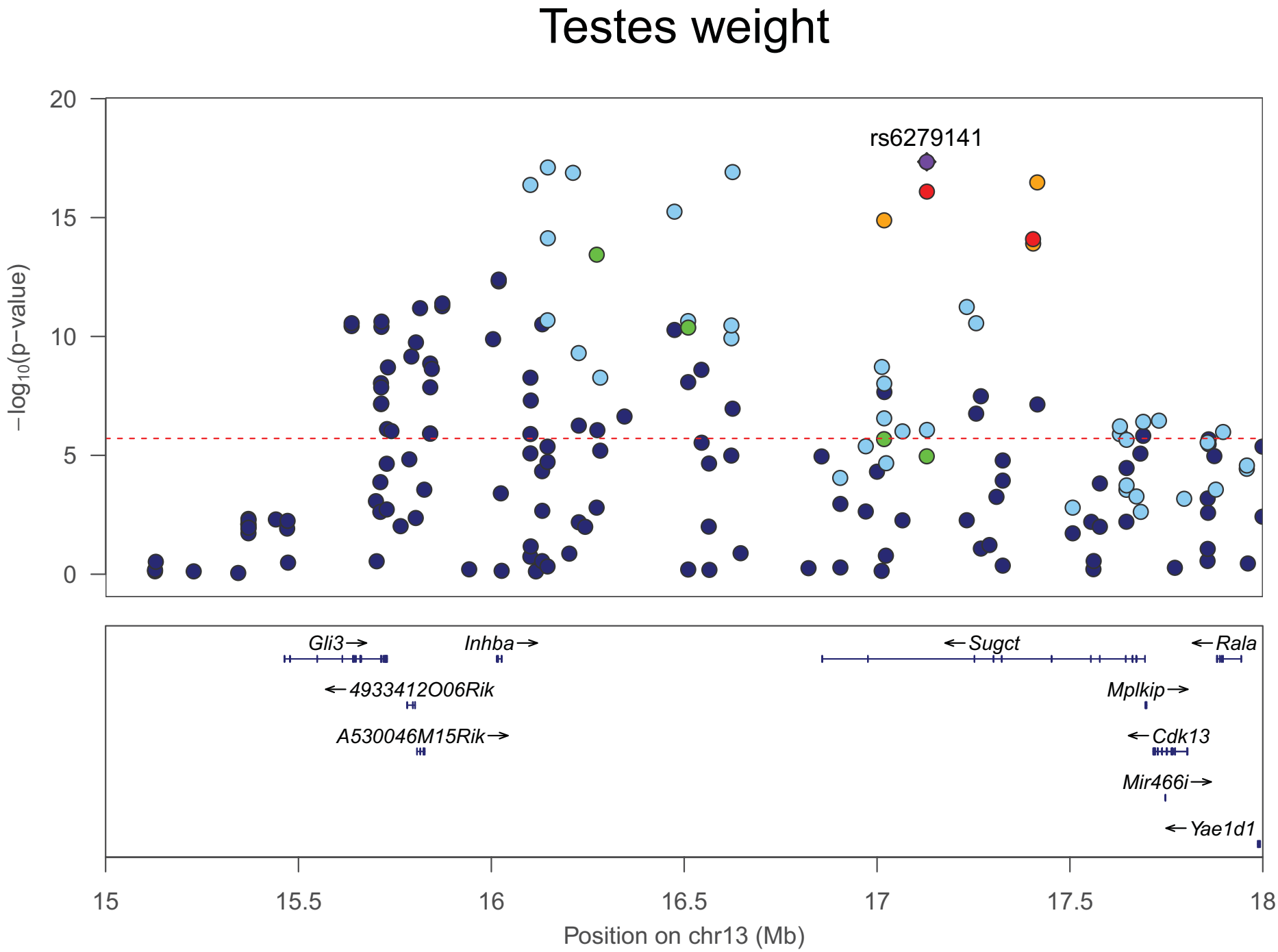
Nature Genetics: doi:10.1038/ng.3609

Tail length



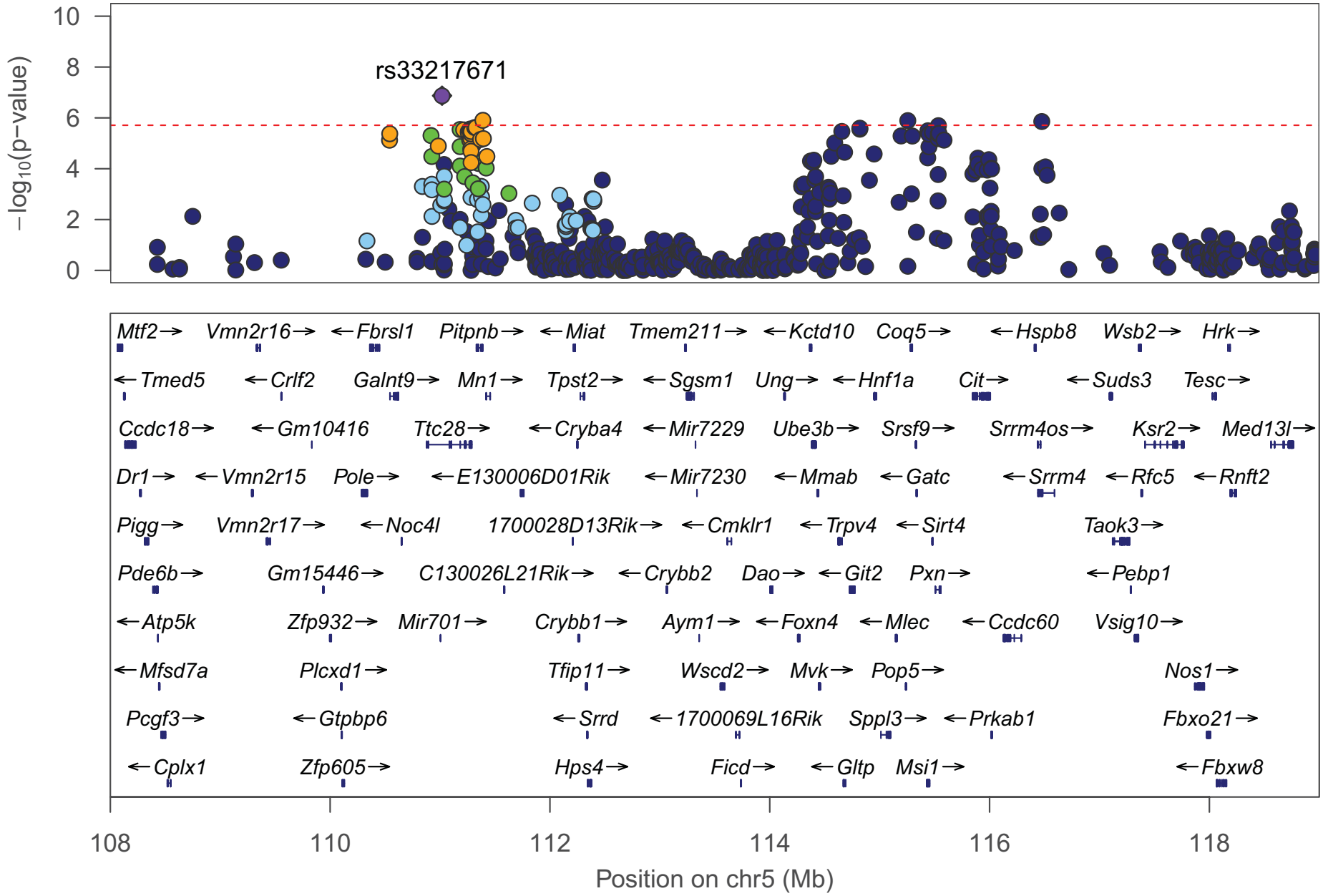
48 genes omitted

Nature Genetics: doi:10.1038/ng.3609



Nature Genetics: doi:10.1038/ng.3609

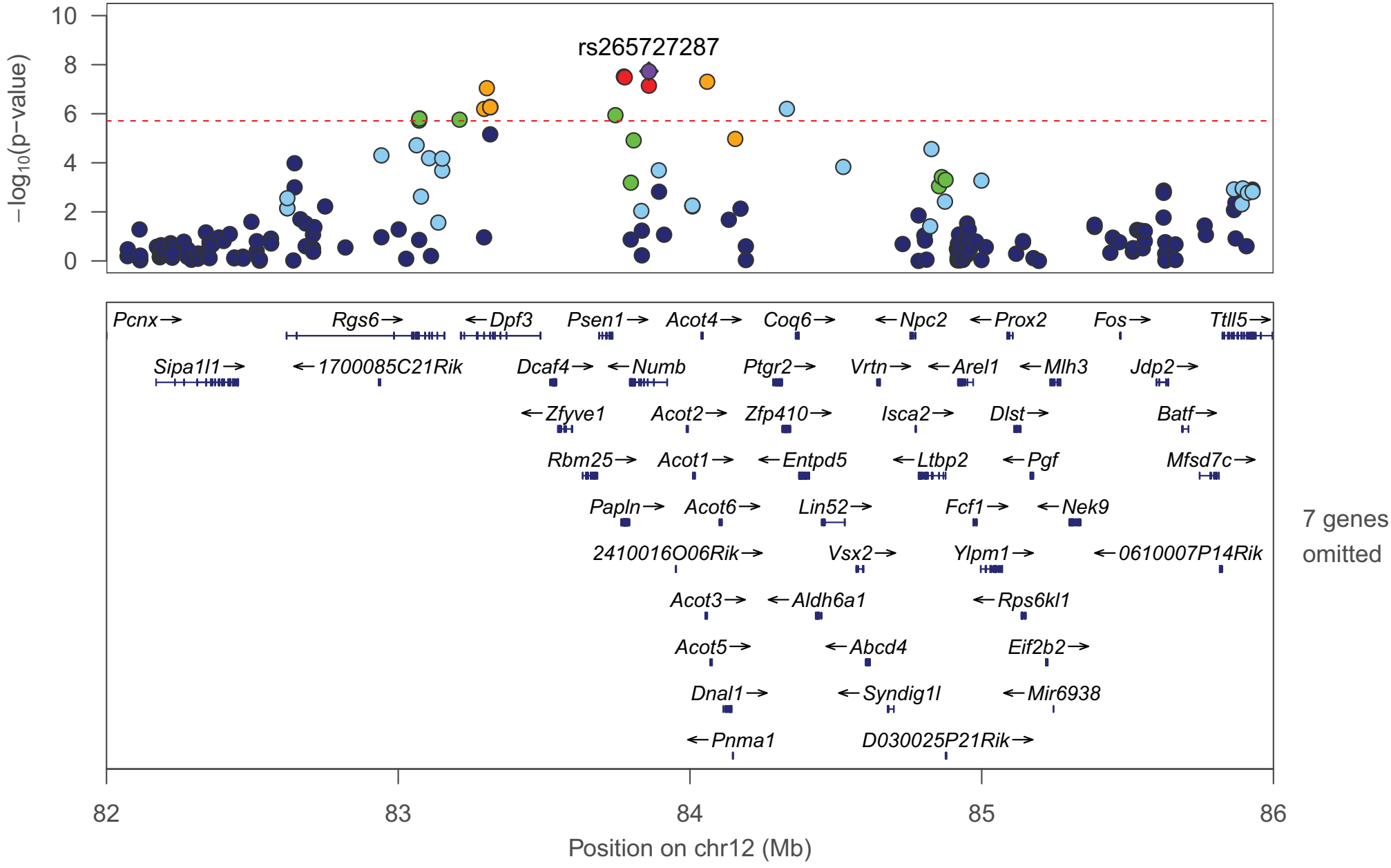
Testes weight



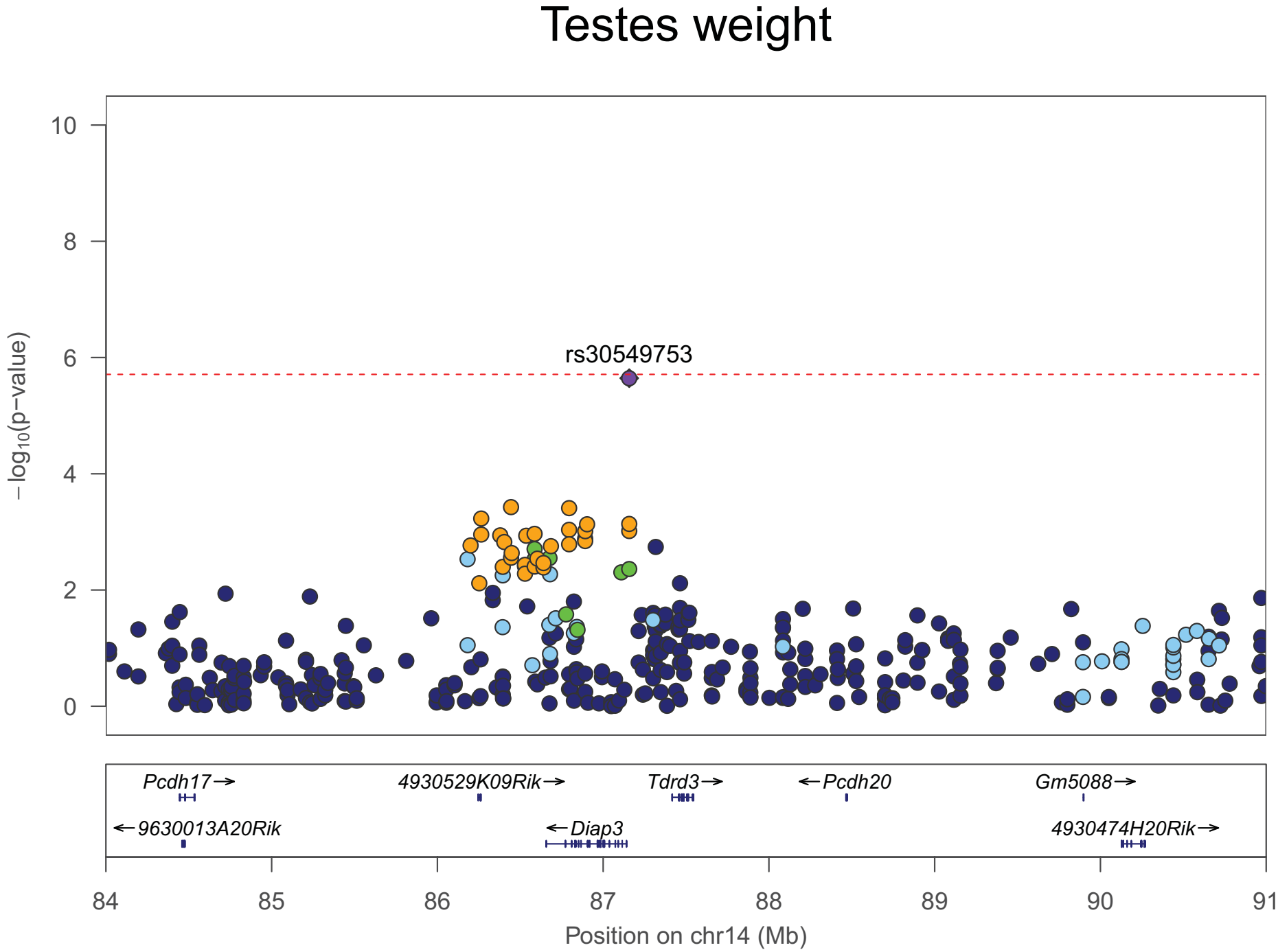
85 genes omitted

Nature Genetics: doi:10.1038/ng.3609

Tibia length

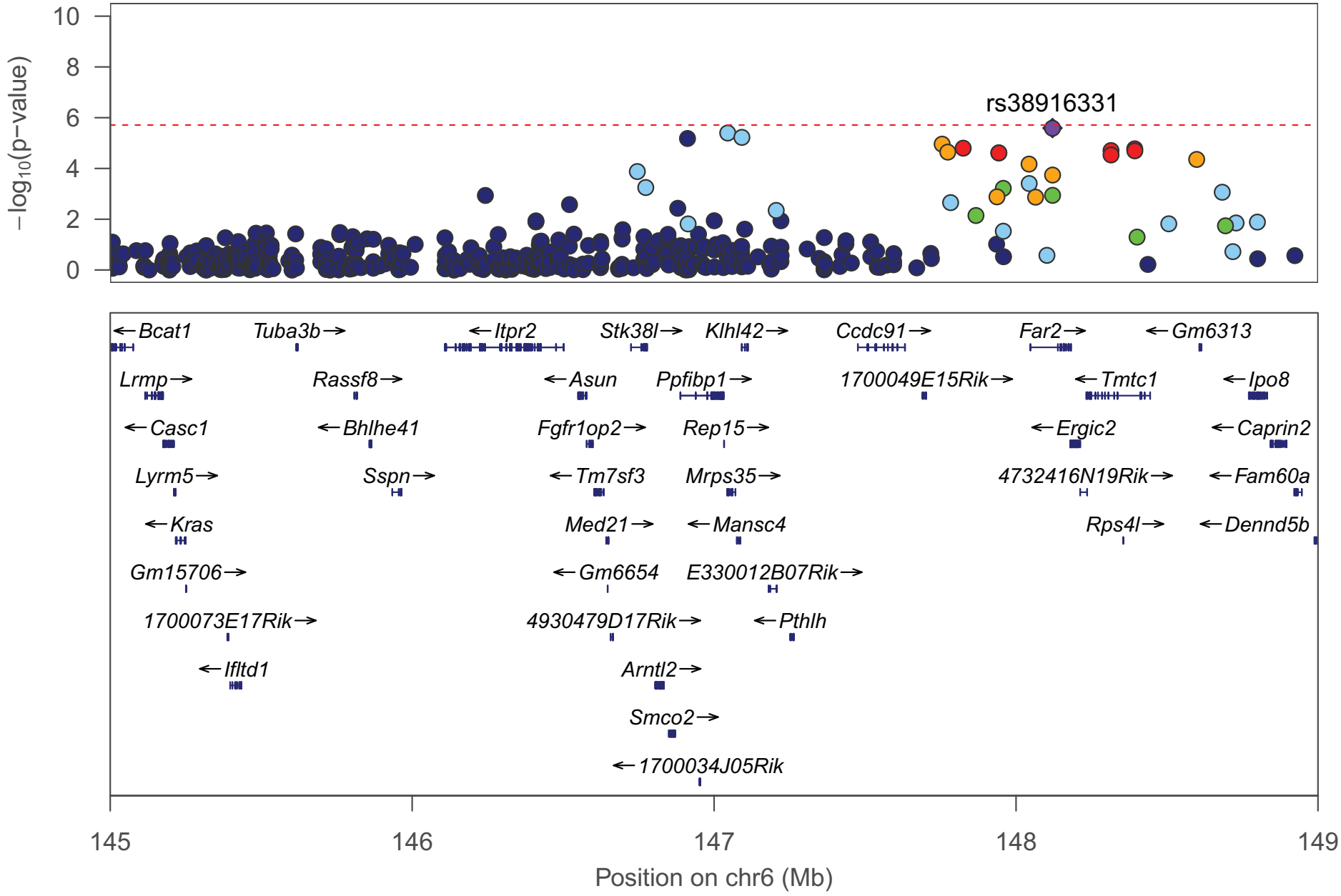


Nature Genetics: doi:10.1038/ng.3609



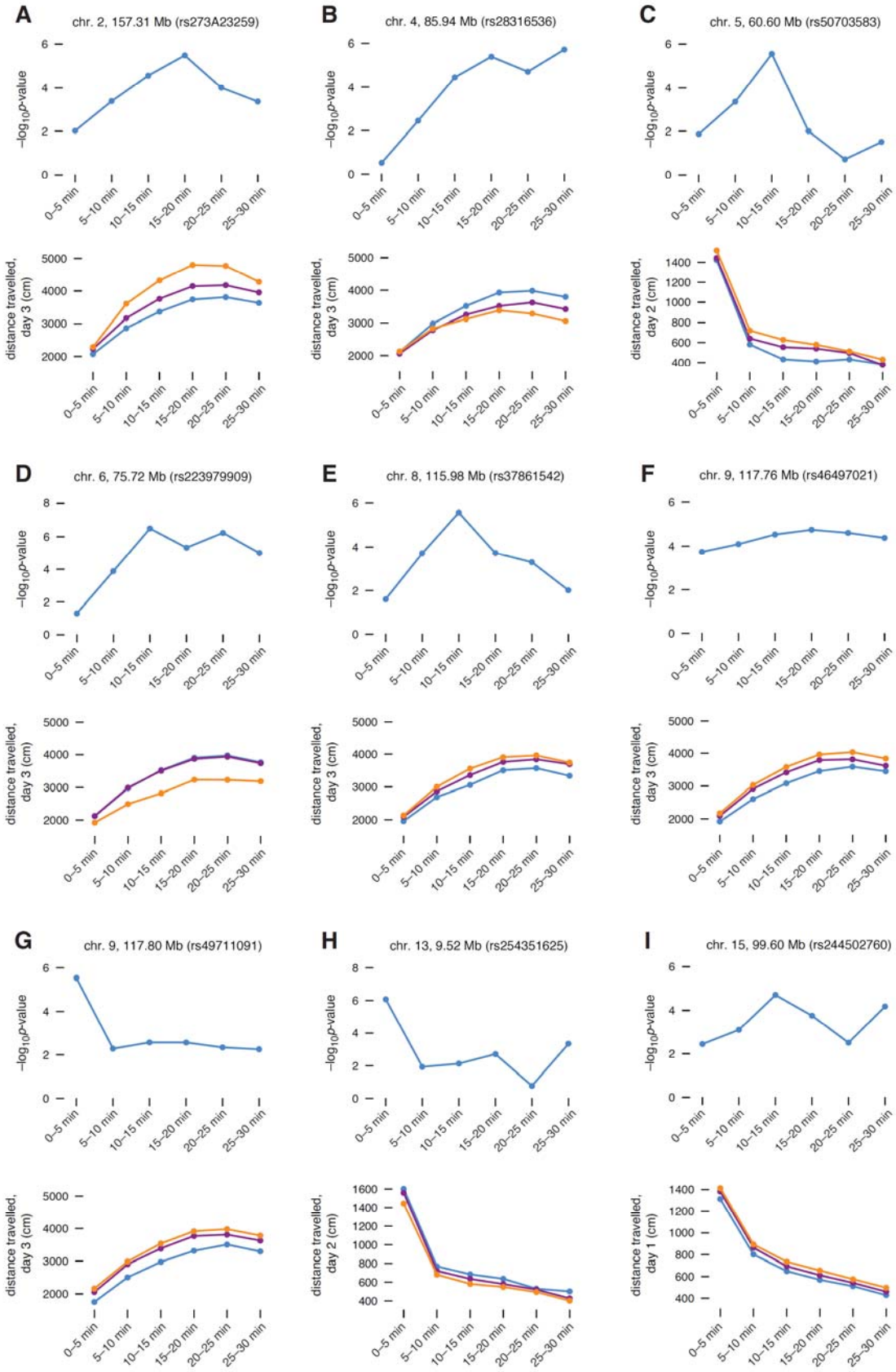
Nature Genetics: doi:10.1038/ng.3609

Tibia length



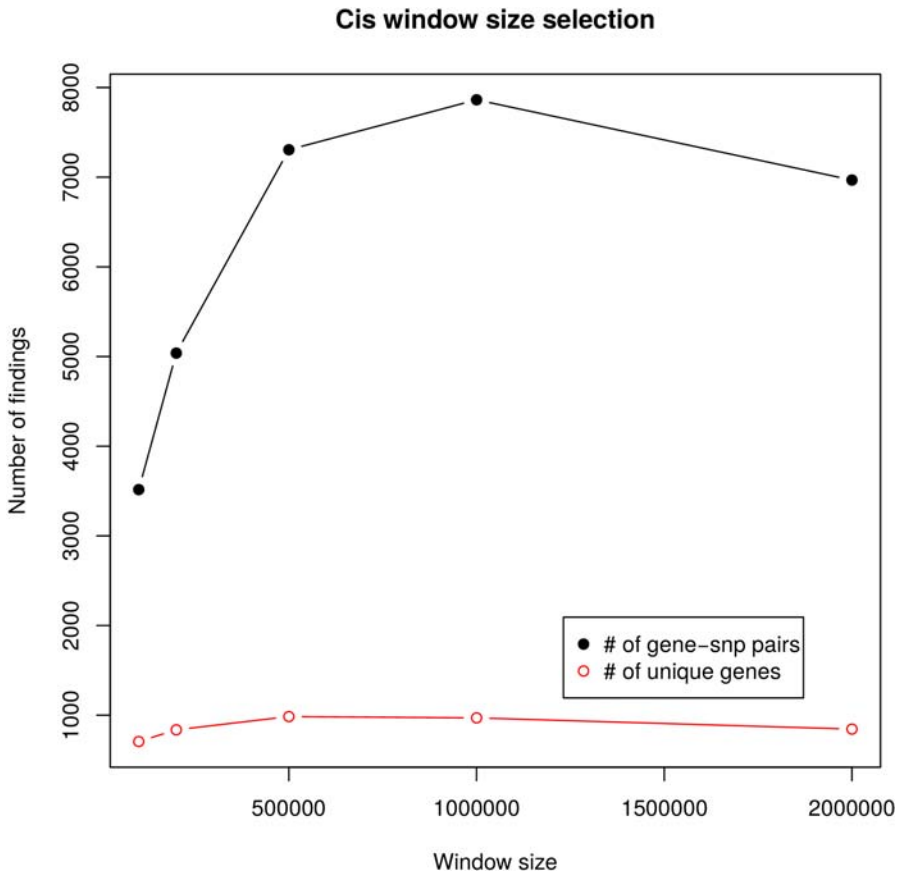
Nature Genetics: doi:10.1038/ng.3609

Supplementary Figure 17: Locuszoom plots for genome-wide significant QTL findings. The *locuszoom* plots showing the zoom in of the 50 QTL findings reported in Supplementary Table 2. The points for each SNP are colored by the level of the linkage disequilibrium (r^2) with the index SNP, the SNP with the highest association to the quantitative trait.

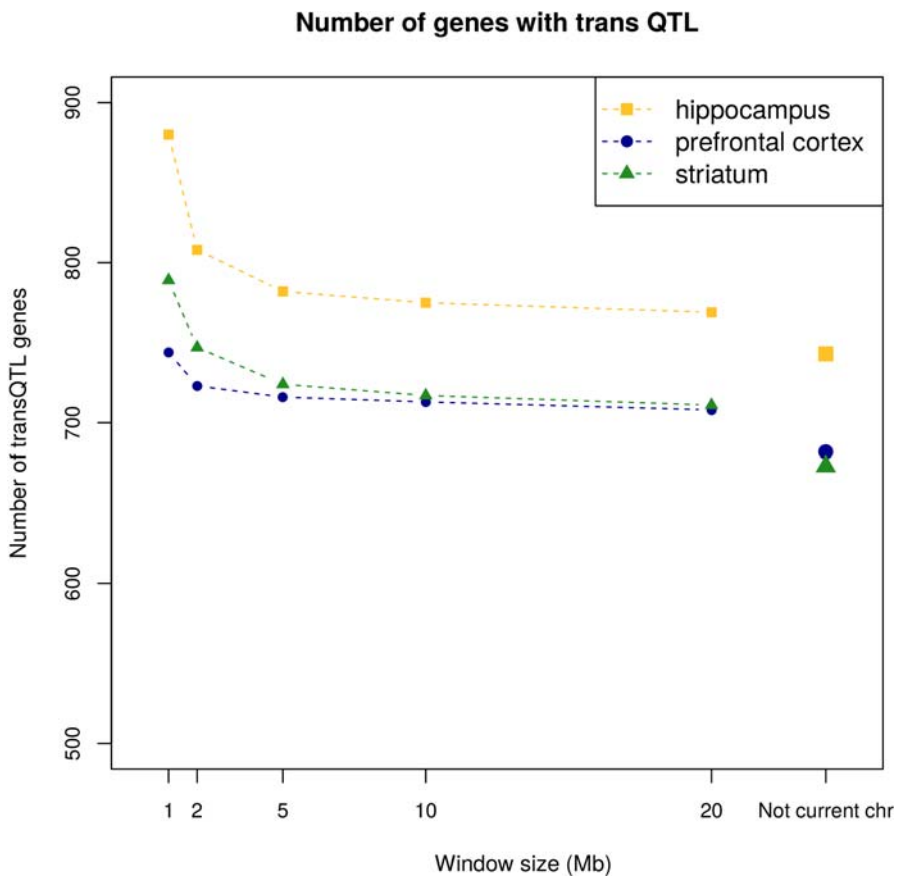


Supplementary Figure 18: *p*-values and mean phenotypes over time for locomotor activity (distance travelled on days 1 and 2) and methamphetamine sensitivity (distance travelled on day 3). The top plot in each panel shows the likelihood-ratio test *p*-value computed in GEMMA for each 5-minute interval during behavioral testing. The bottom plot in each panel shows, for each 5-minute interval, the average phenotype

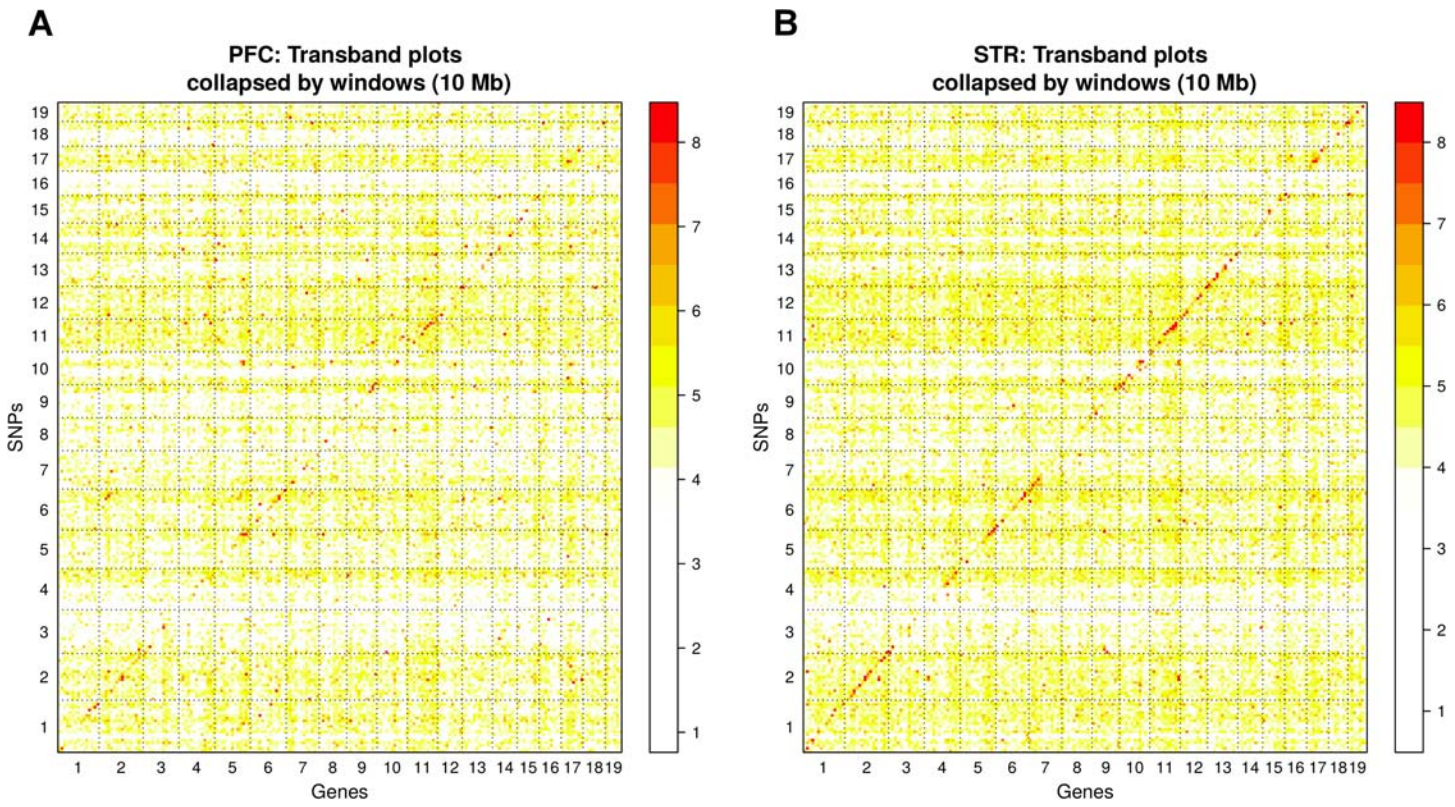
value for individuals that are homozygous for the reference allele ("AA" in Supplementary Table 2, light blue line), heterozygous ("Aa", purple line), and homozygous for the alternative allele ("aa", orange line). The *p*-values for these traits when aggregated over 15 and 30 minutes are shown in Supplementary Table 2; those *p*-values provided the impetus to investigate these particular loci.



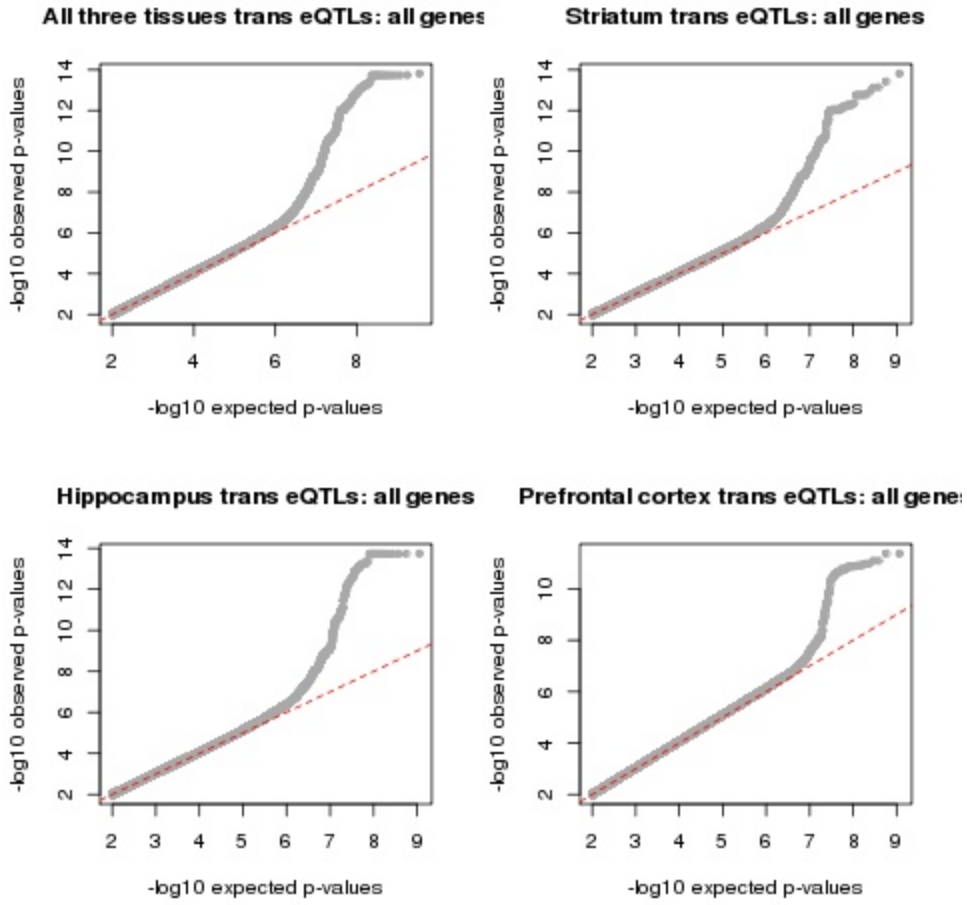
Supplementary Figure 19: Hippocampus *cis*-eQTL window selection. Number of *cis*-eQTL associations found for different choices of window size around a gene for defining *cis*-eQTL analysis region.



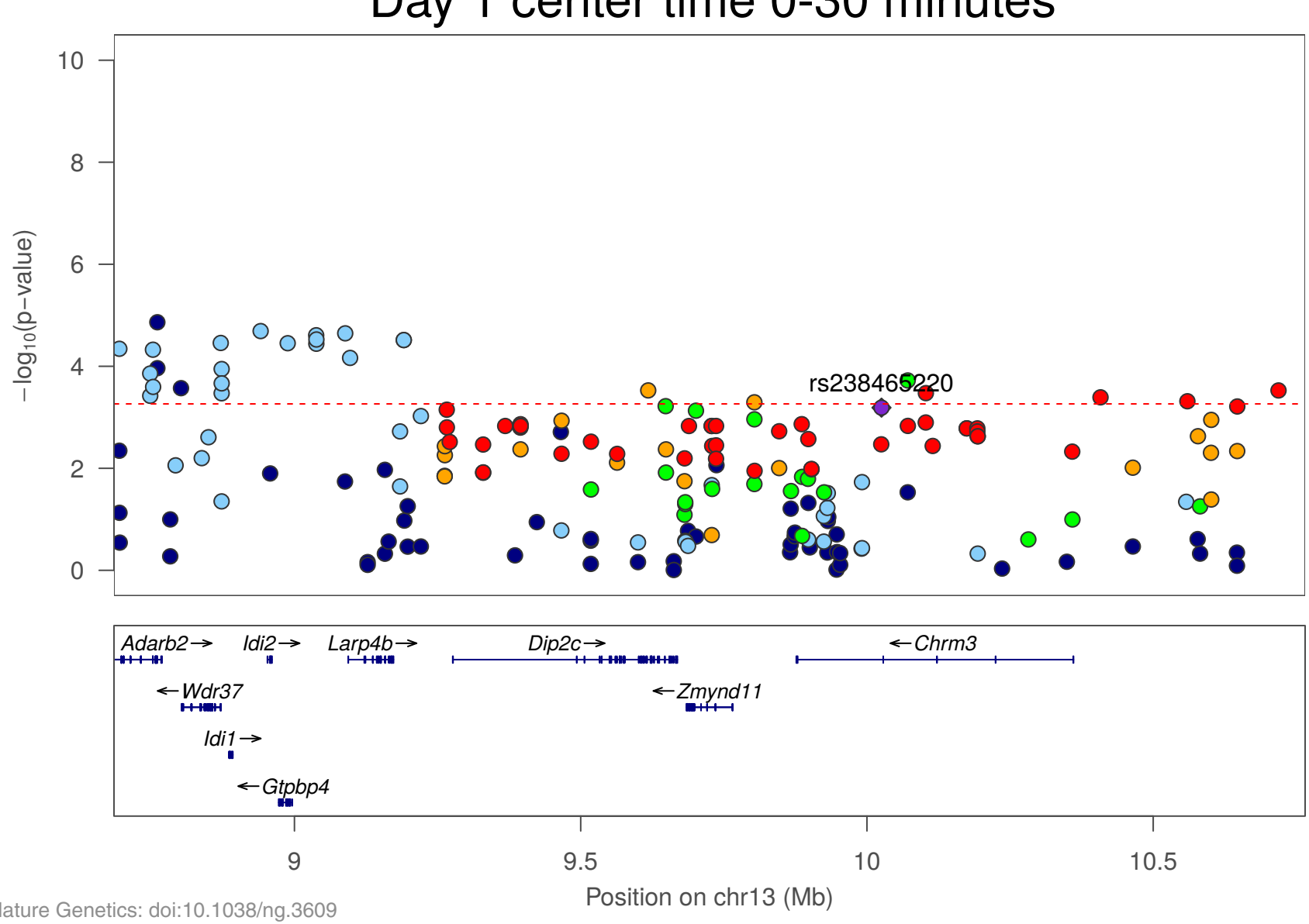
Supplementary Figure 20: Number of *trans*-eQTL genes. The number of genes with *trans*-eQTLs is plotted against the window around a gene outside which a SNP must lie, in order to be considered a *trans* QTL for the expression of the gene. The number of genes with *trans*-eQTLs were computed using a permutation based threshold at a significance level of 0.05.



Supplementary Figure 21: Transband plots for Prefrontal cortex and Striatum. The transband plots for the PFC (A) and STR (B) regions of the brain, where each pixel shows the best expression-genotype association p-value in a 10 Mb x 10 Mb region across the genome.

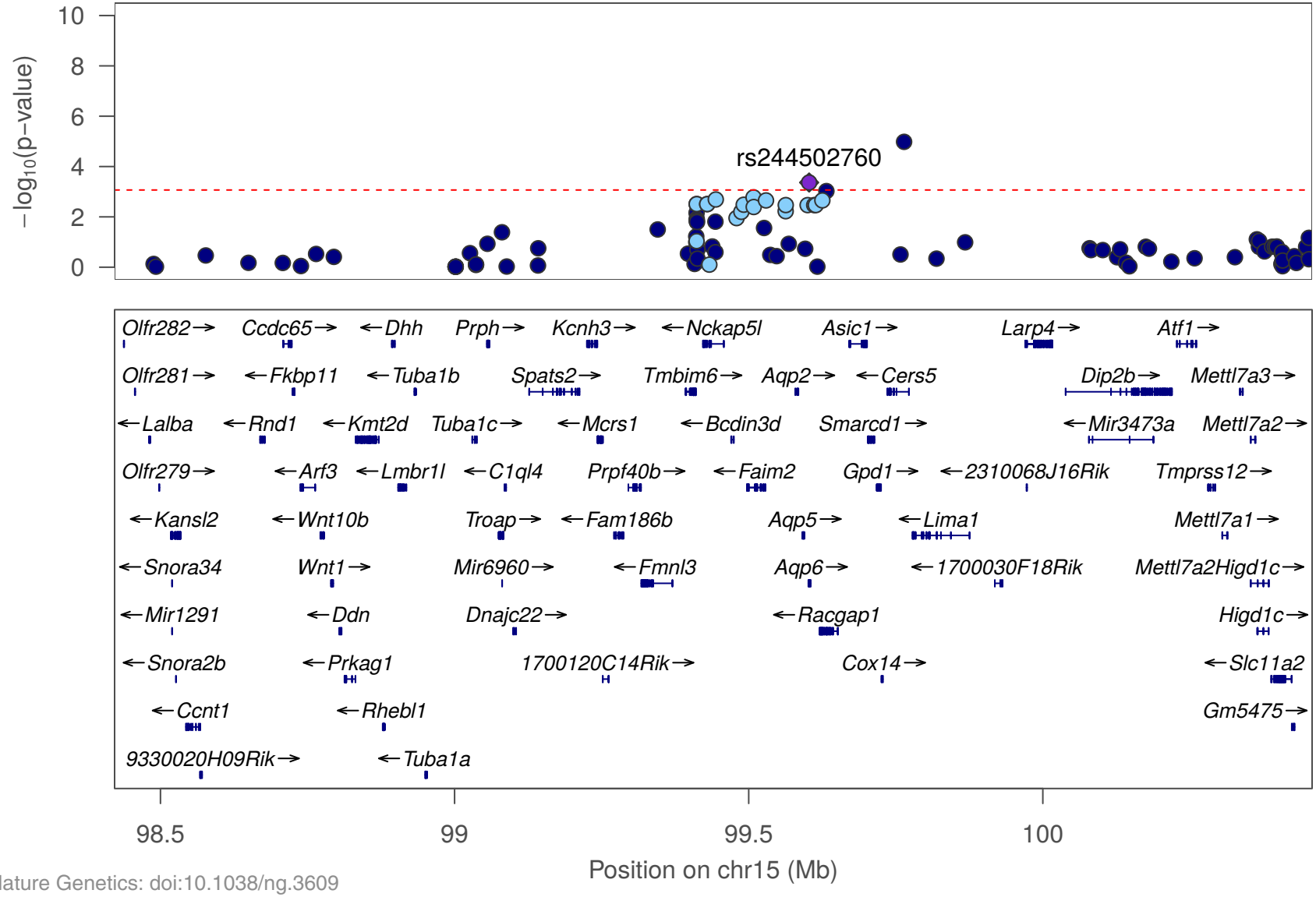


Supplementary Figure 22: Q-Q plots for *trans*-eQTL findings. Each panel shows a quantile-quantile plot for *trans*-eQTLs, comparing quantiles of the observed *p*-values (vertical axis) for the three brain tissues against expected quantiles under the null (uniform) distribution of *p*-values. The top-left panel summarizes the quantiles for all three brain tissues. The remaining three panels summarize the quantiles for each brain region individually. In all four plots, only SNPs with observed *p*-values less than 0.01 are included.



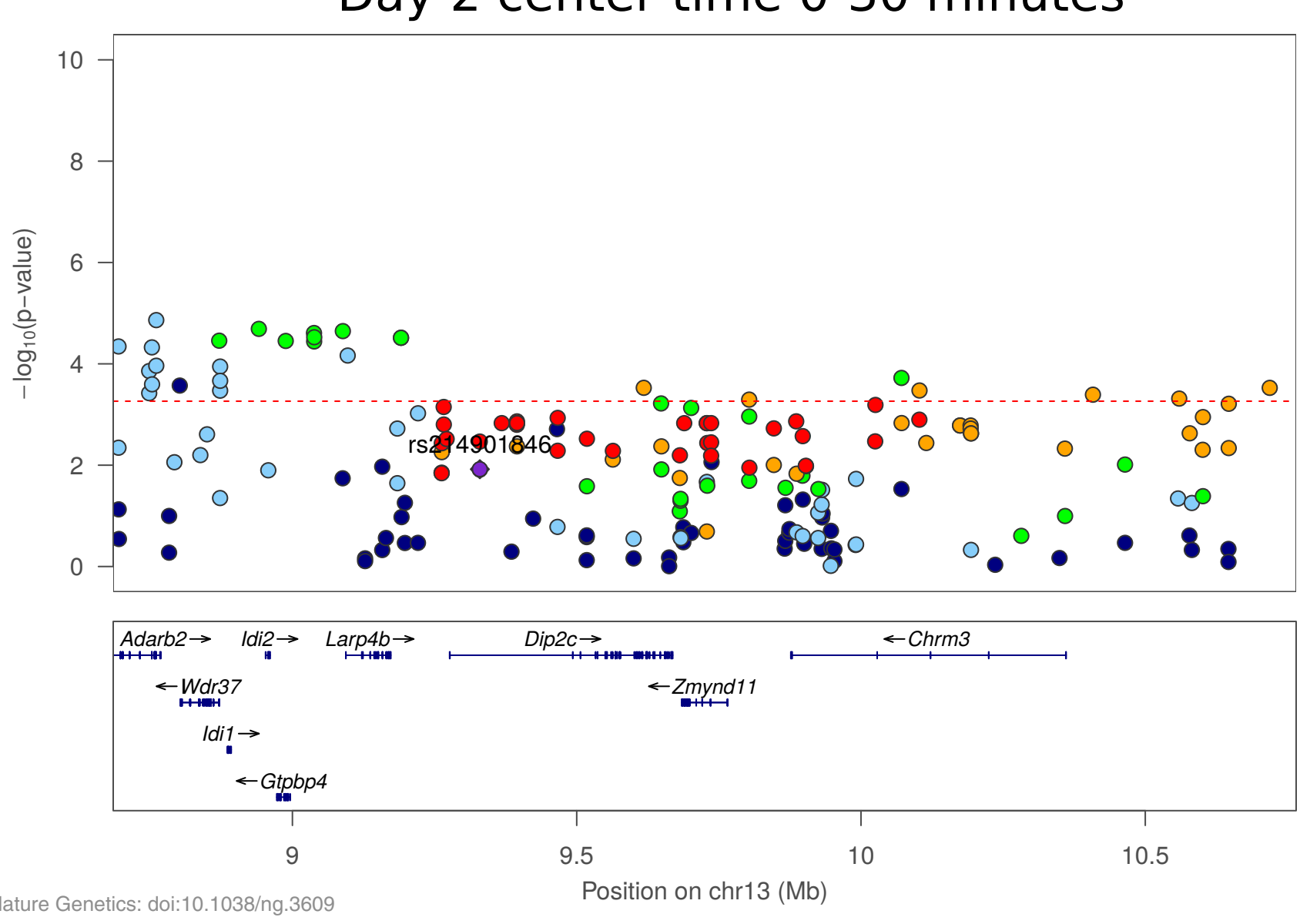
Nature Genetics: doi:10.1038/ng.3609

Nckap5l expression in striatum: Day 1 total distance 0-30 minutes

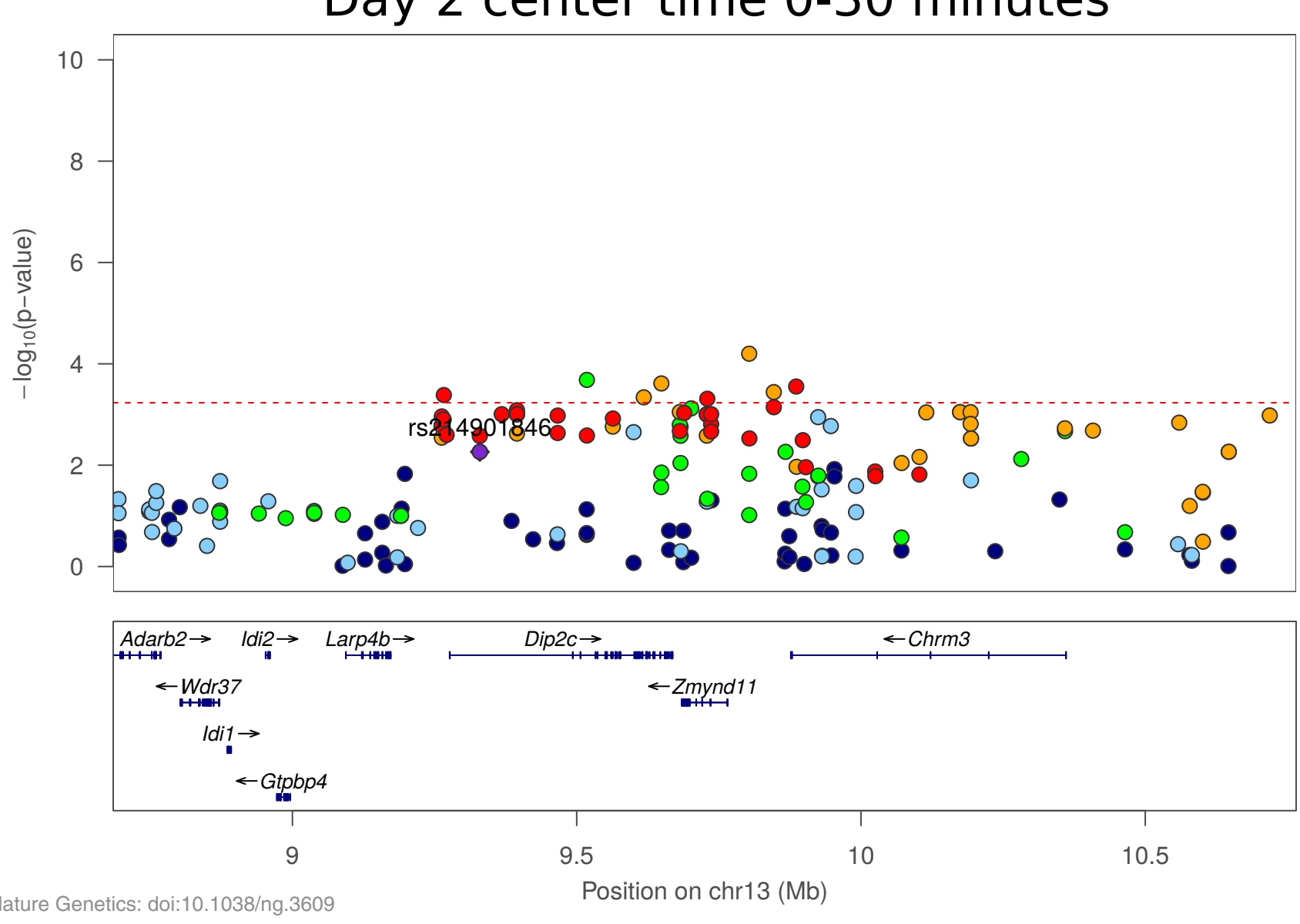


5 genes omitted

Nature Genetics: doi:10.1038/ng.3609

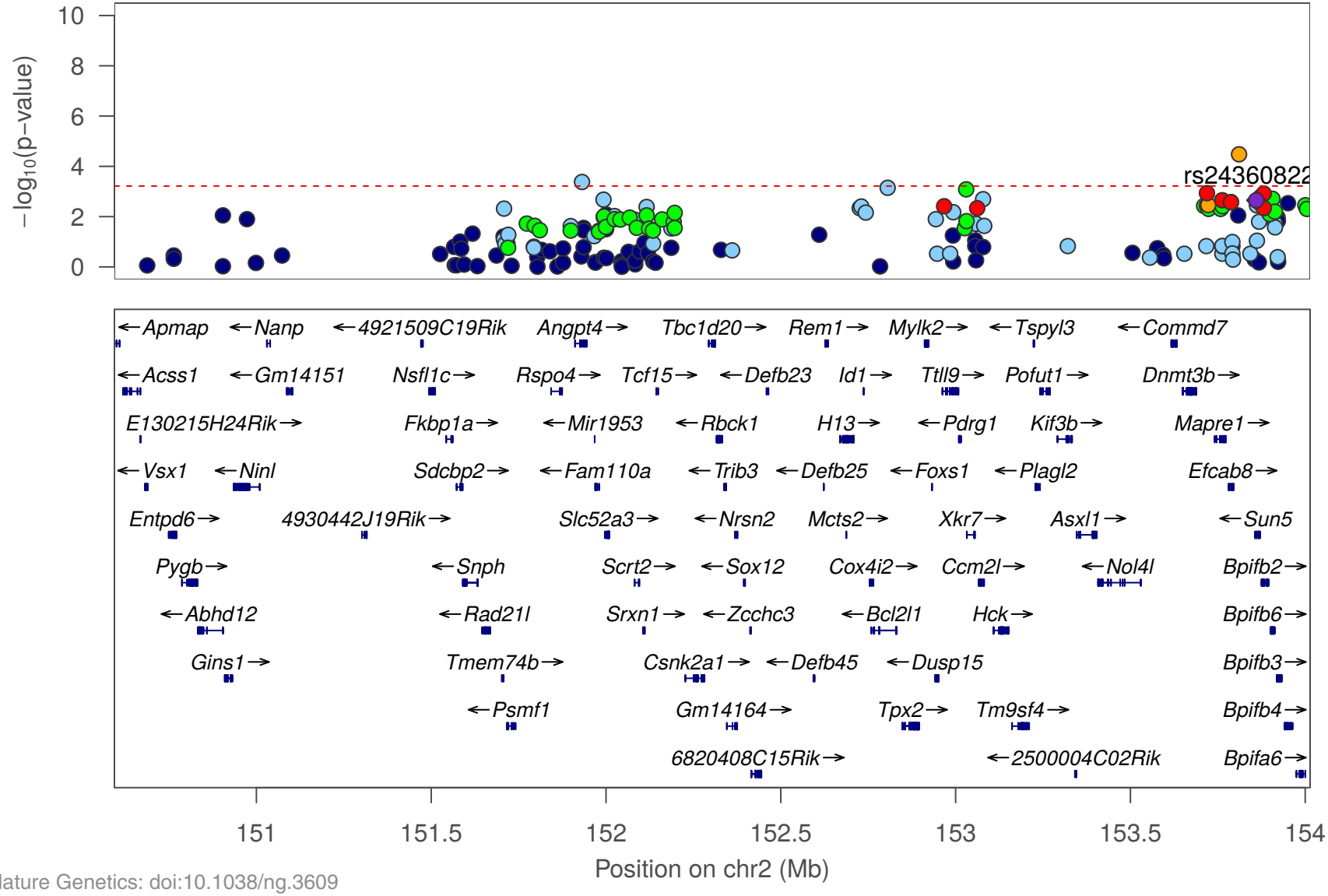


Nature Genetics: doi:10.1038/ng.3609



Nature Genetics: doi:10.1038/ng.3609

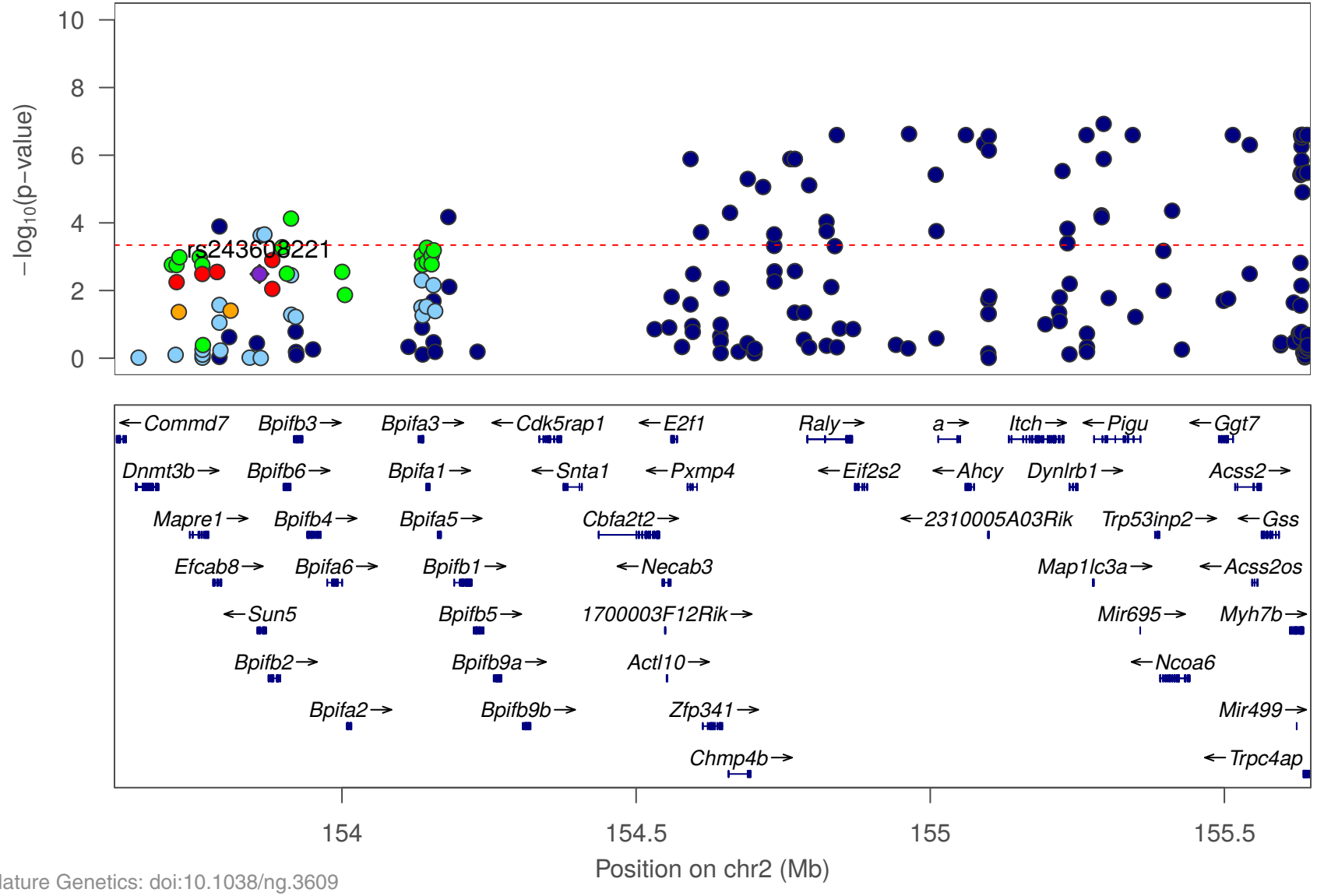
Tbc1d20 expression in prefrontal cortex: Day 2 horizontal activity 0-30 minutes



10 genes omitted

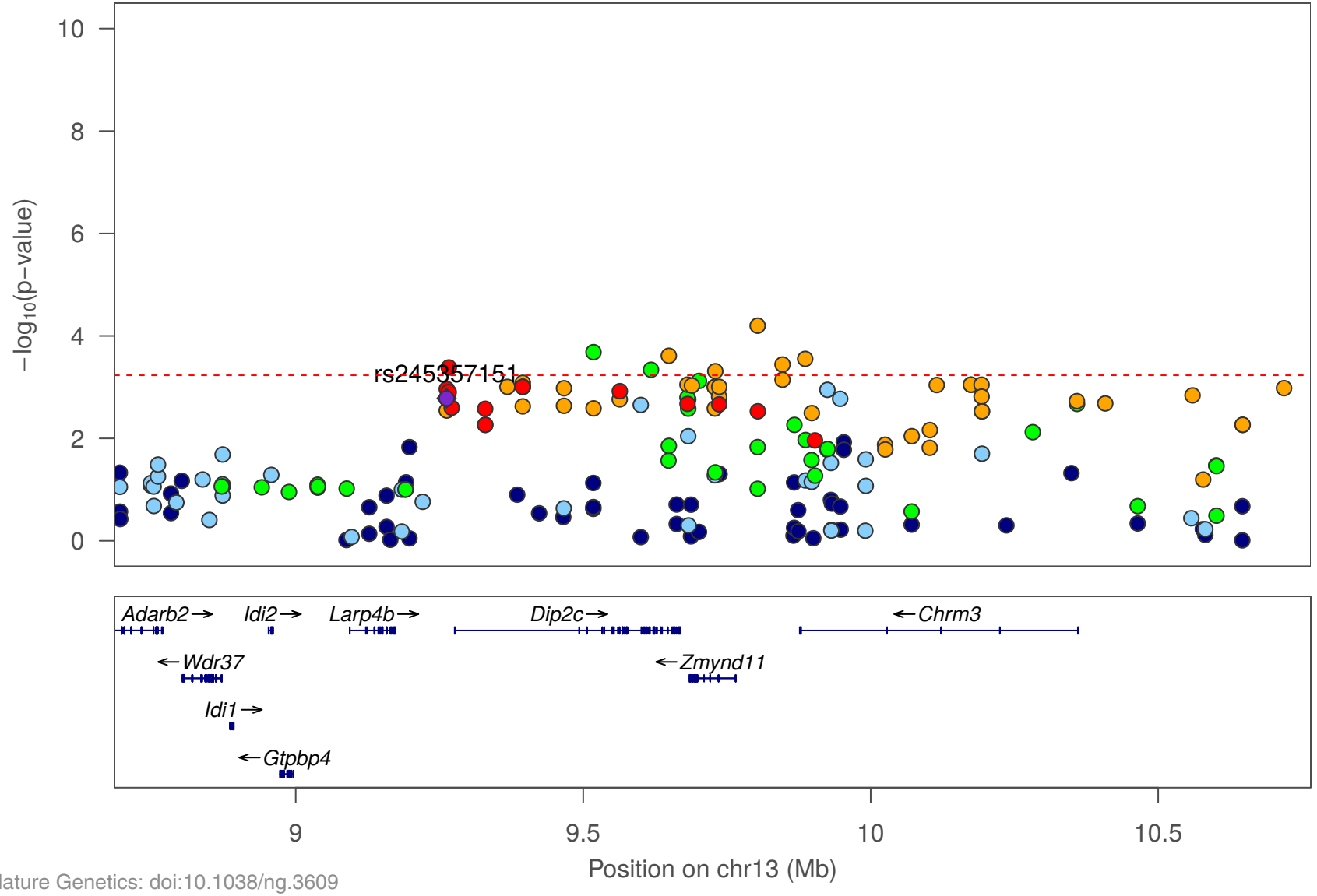
Nature Genetics: doi:10.1038/ng.3609

Zfp341 expression in hippocampus: Day 2 horizontal activity 0-30 minutes



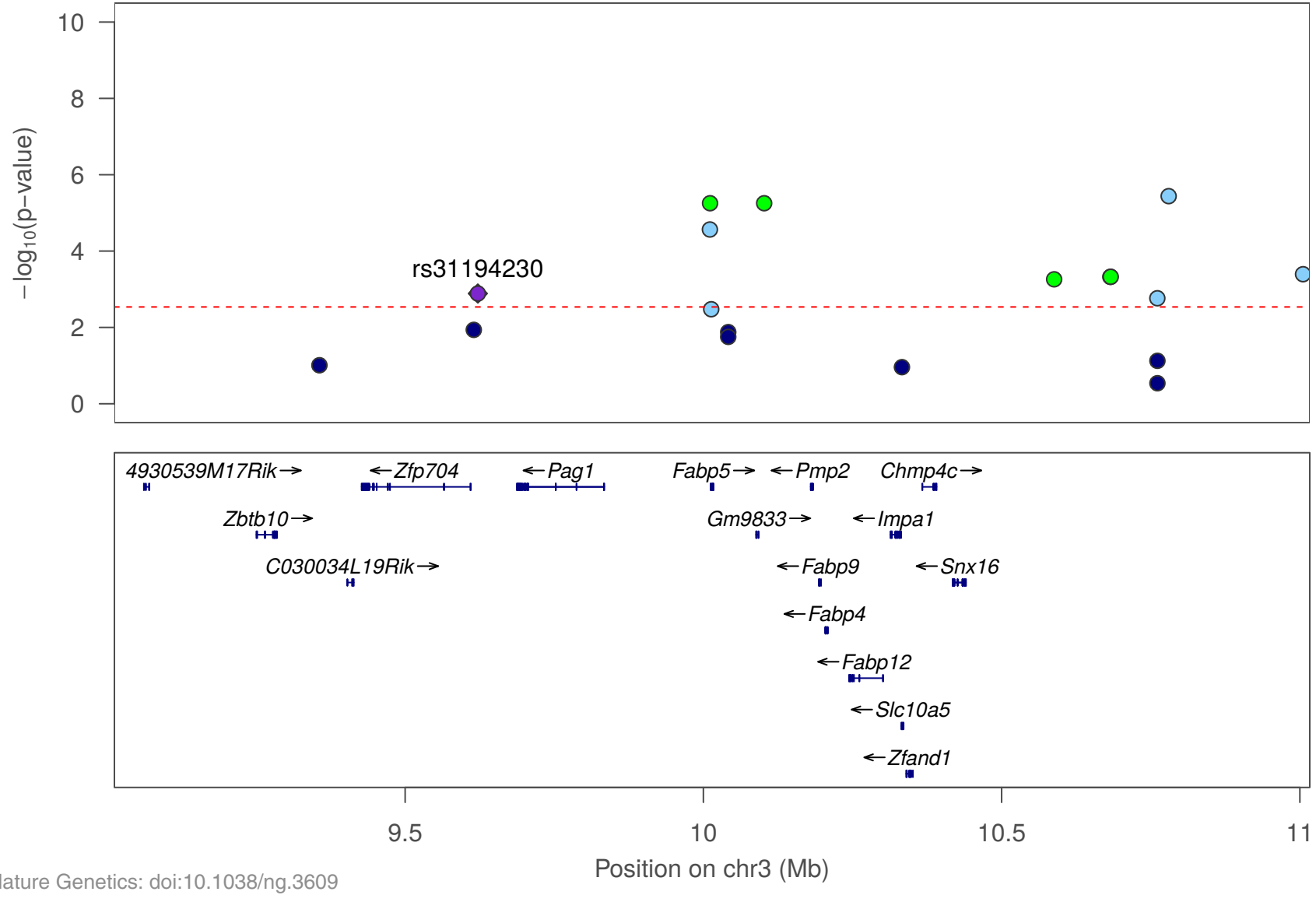
Nature Genetics: doi:10.1038/ng.3609

Zmynd11 expression in prefontal cortex: Day 2 horizontal activity 0-30 minutes

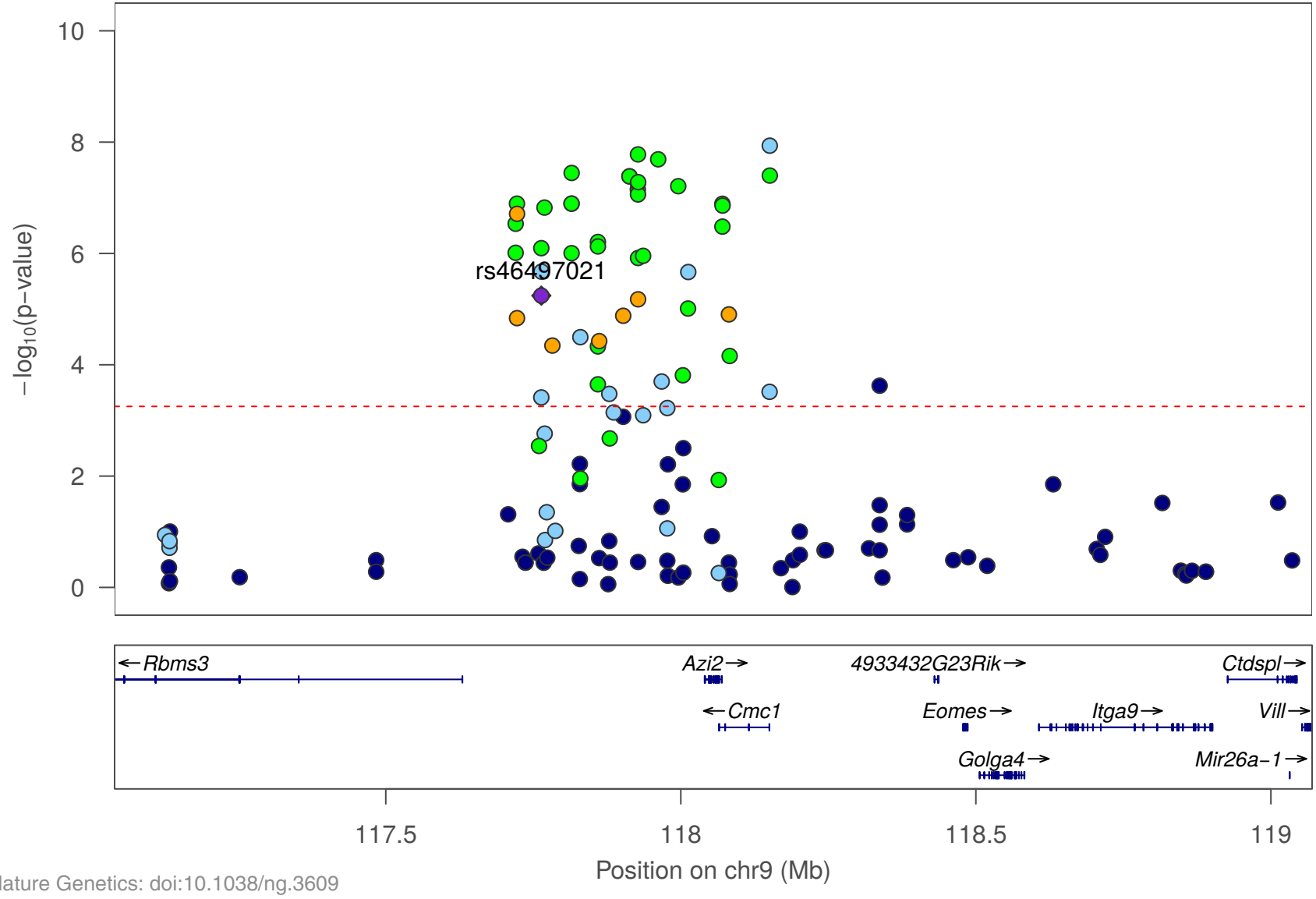


Nature Genetics: doi:10.1038/ng.3609

Fabp5 expression in striatum: Day 3 center time 0-15 minutes

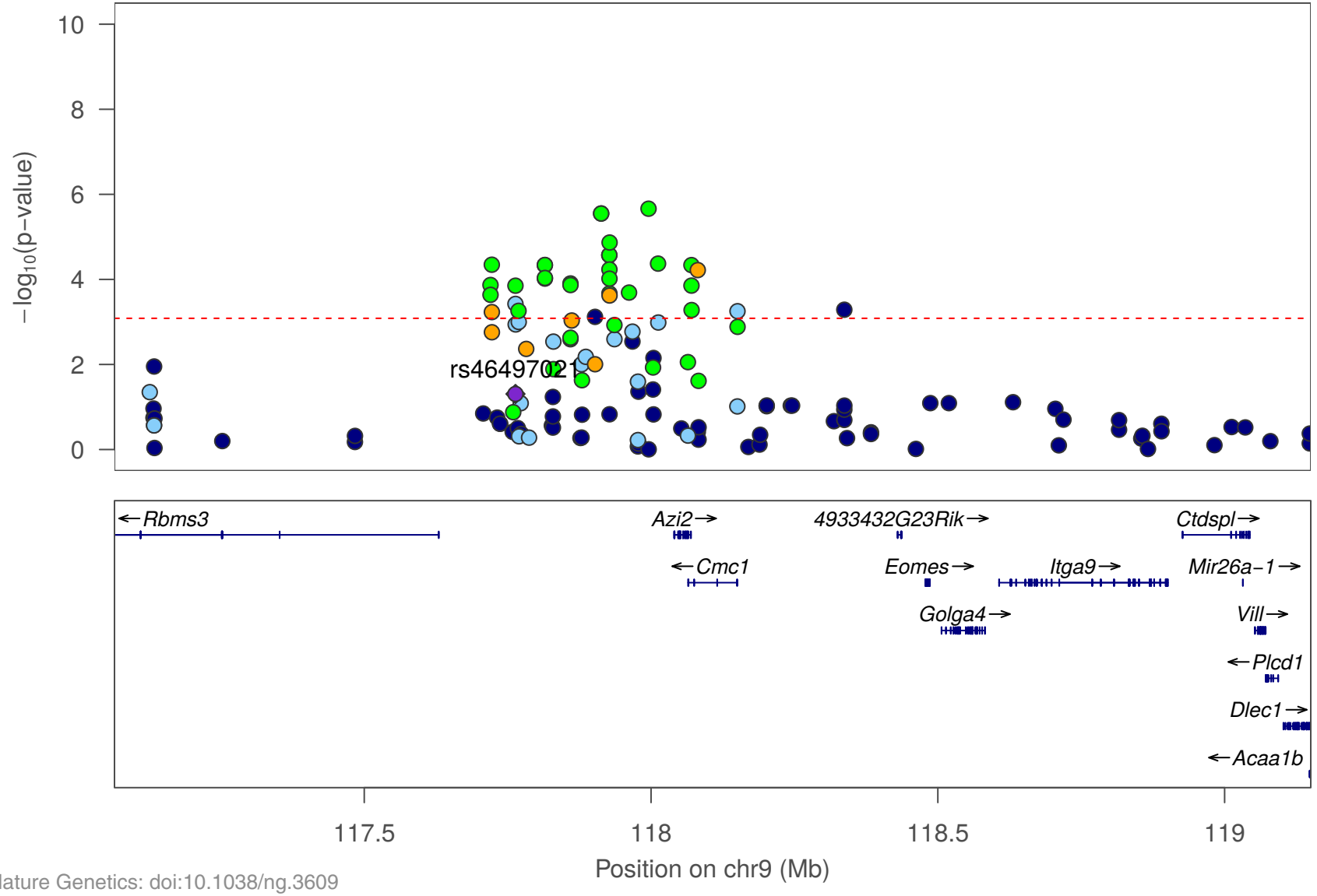


Nature Genetics: doi:10.1038/ng.3609

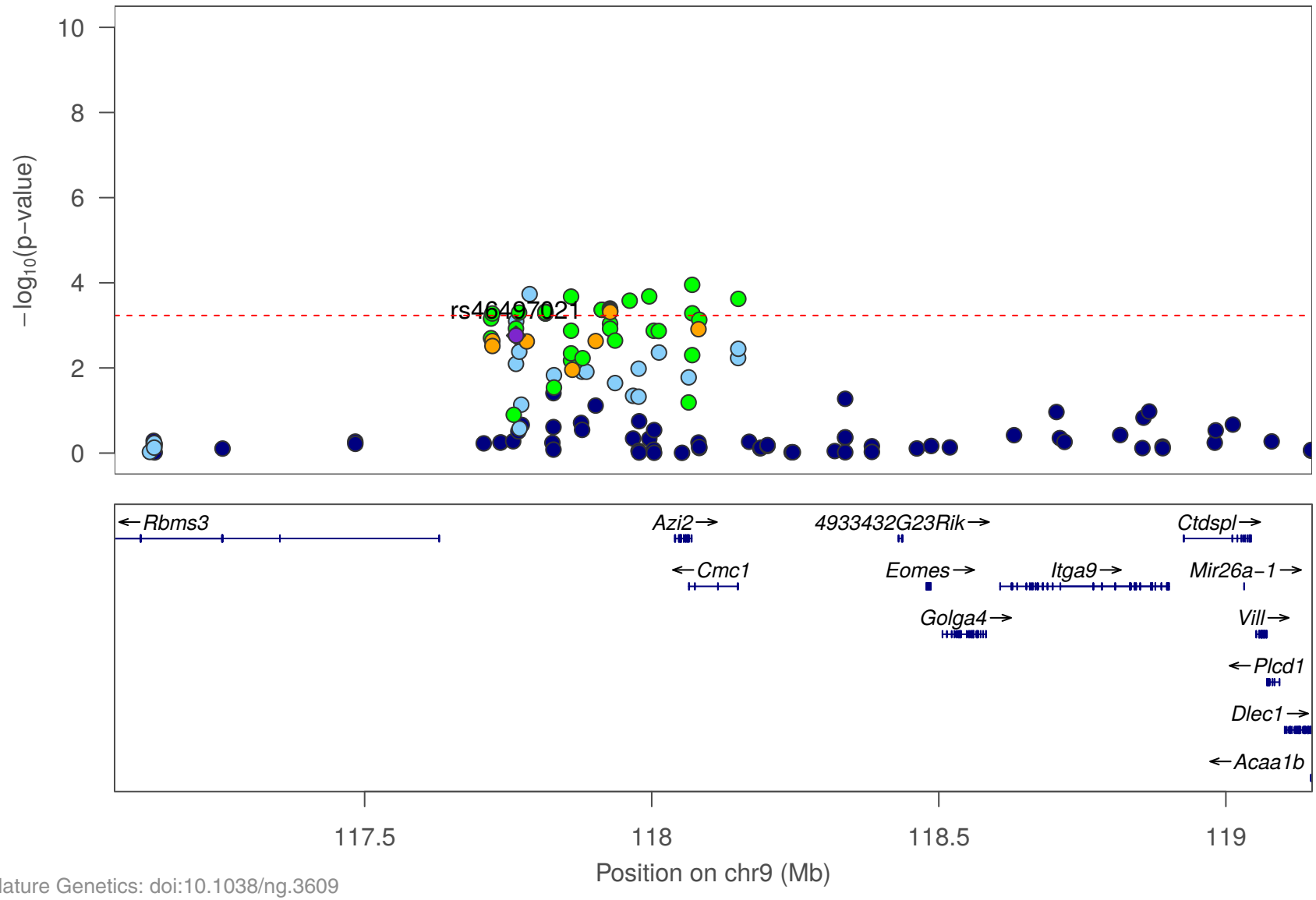


Nature Genetics: doi:10.1038/ng.3609

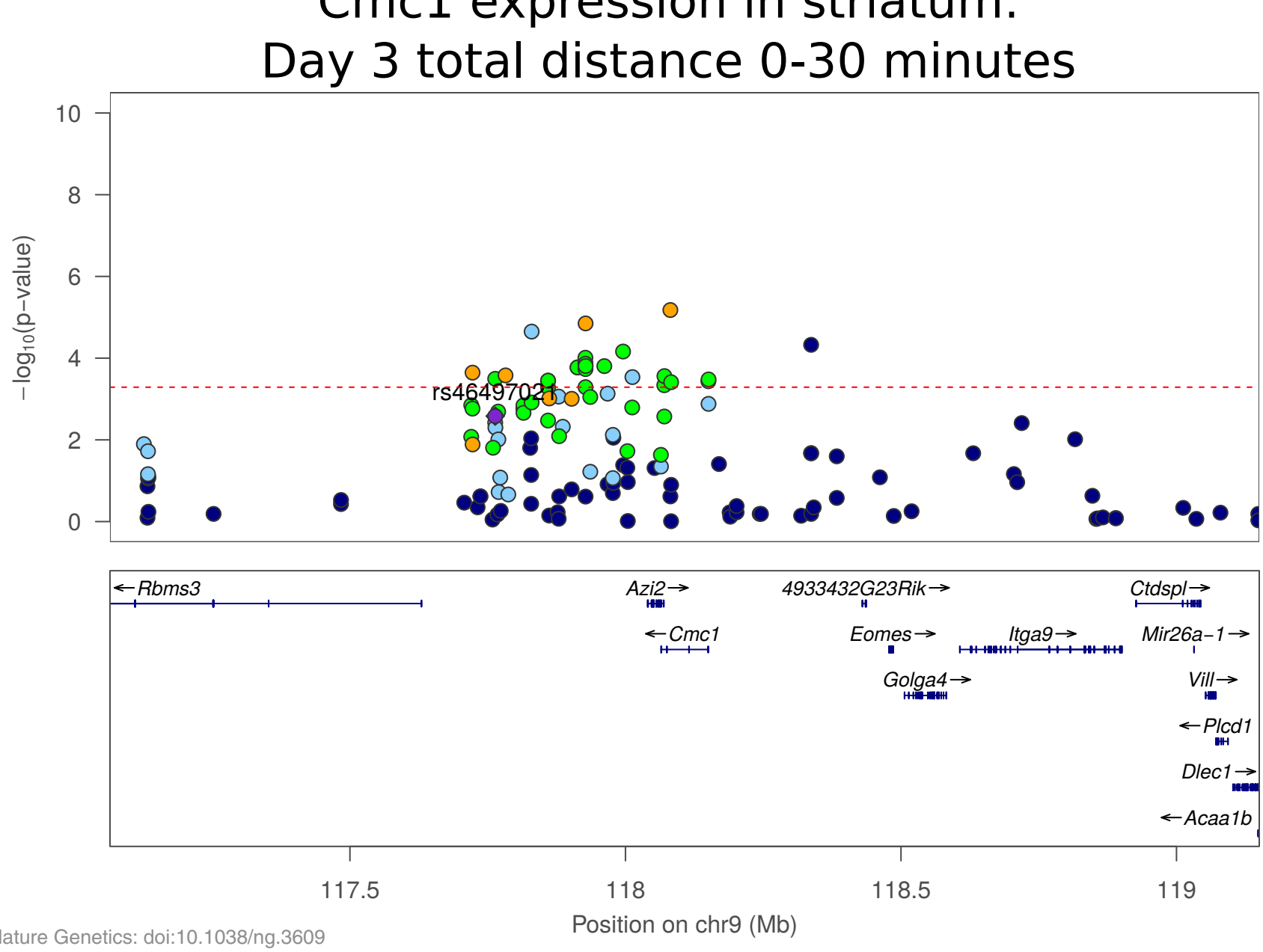
Cmc1 expression in hippocampus: Day 3 total distance 0-30 minutes



Nature Genetics: doi:10.1038/ng.3609

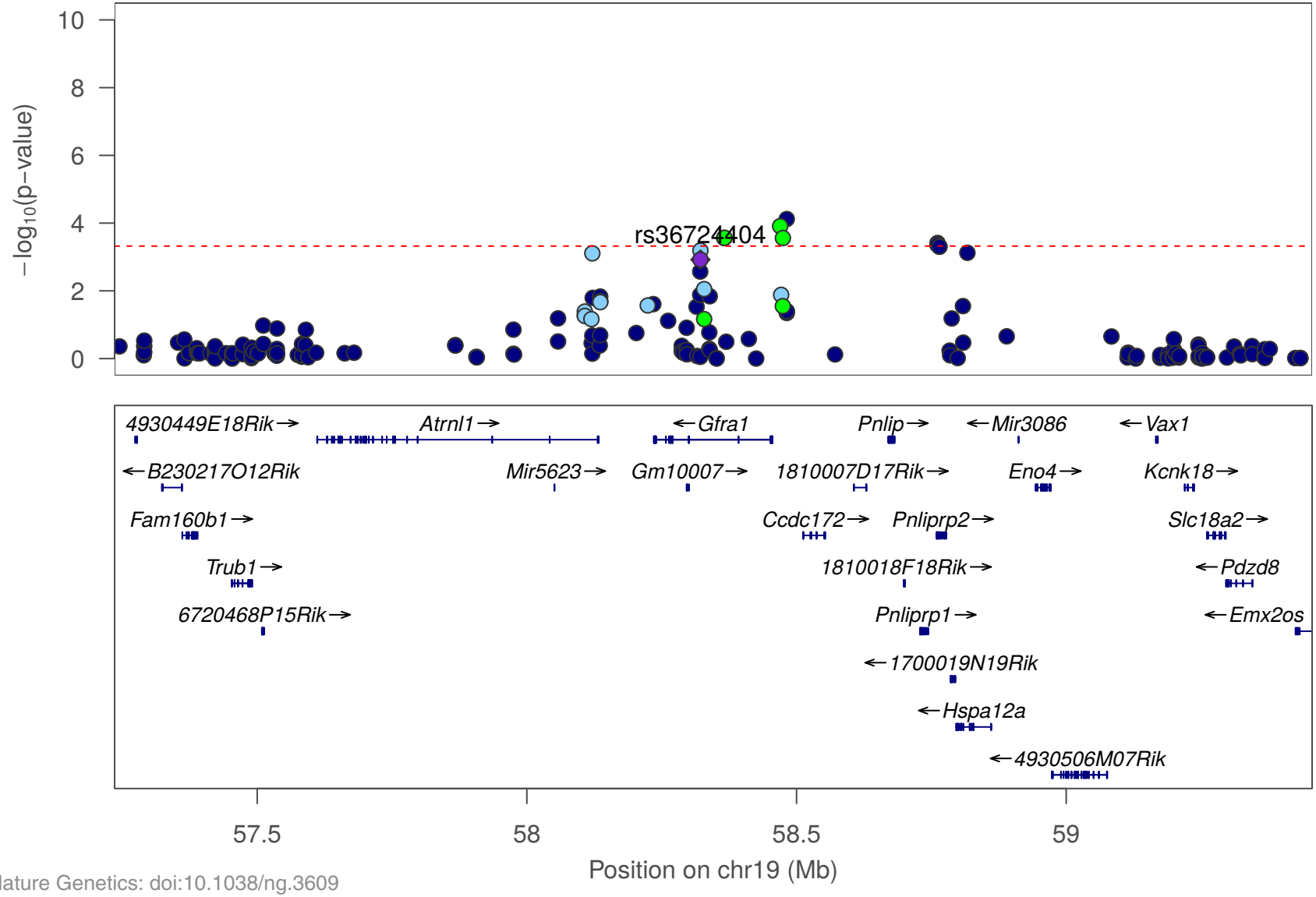


Nature Genetics: doi:10.1038/ng.3609



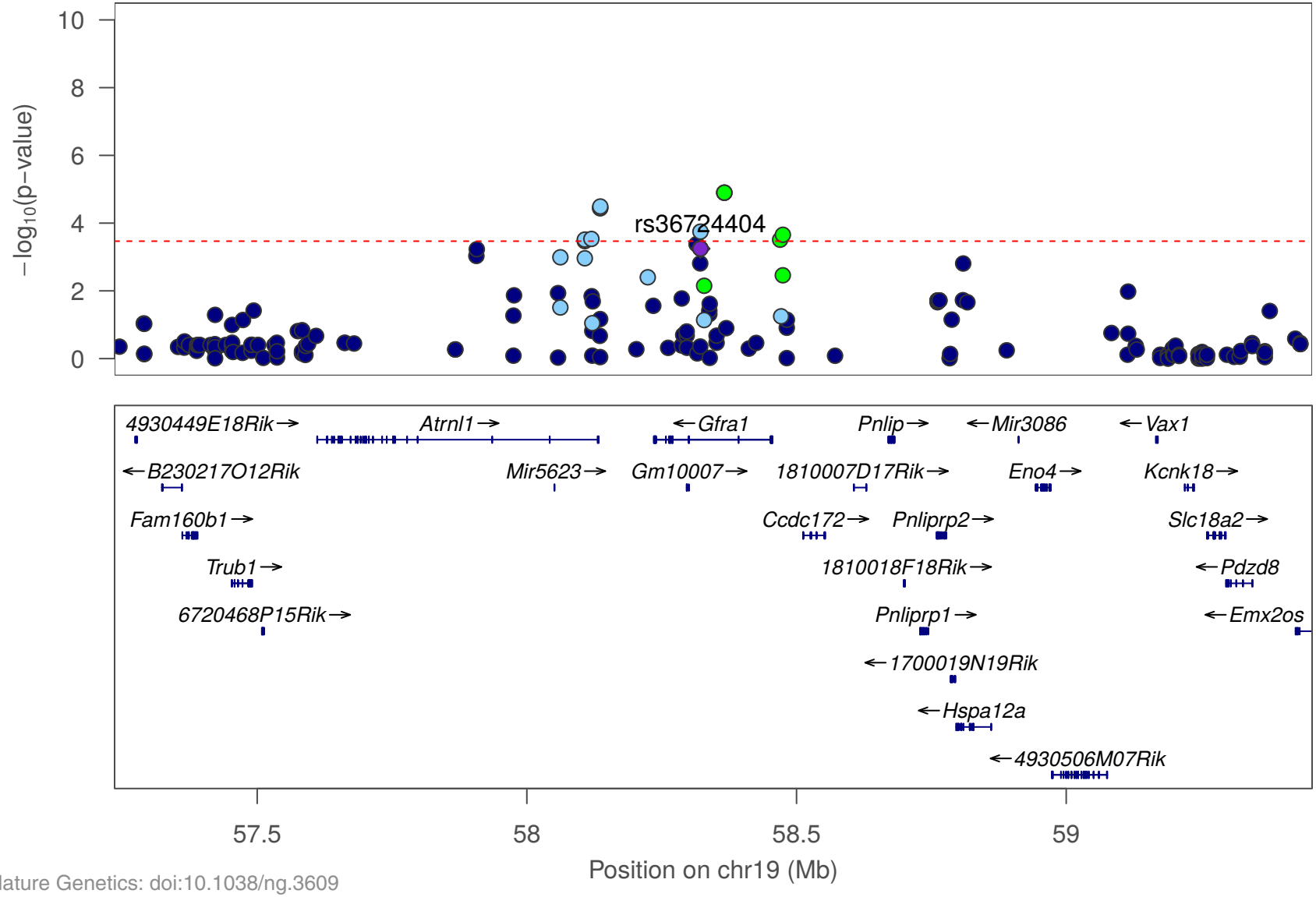
Nature Genetics: doi:10.1038/ng.3609

Gfra1 expression in hippocampus: Day 3 freezing to tone



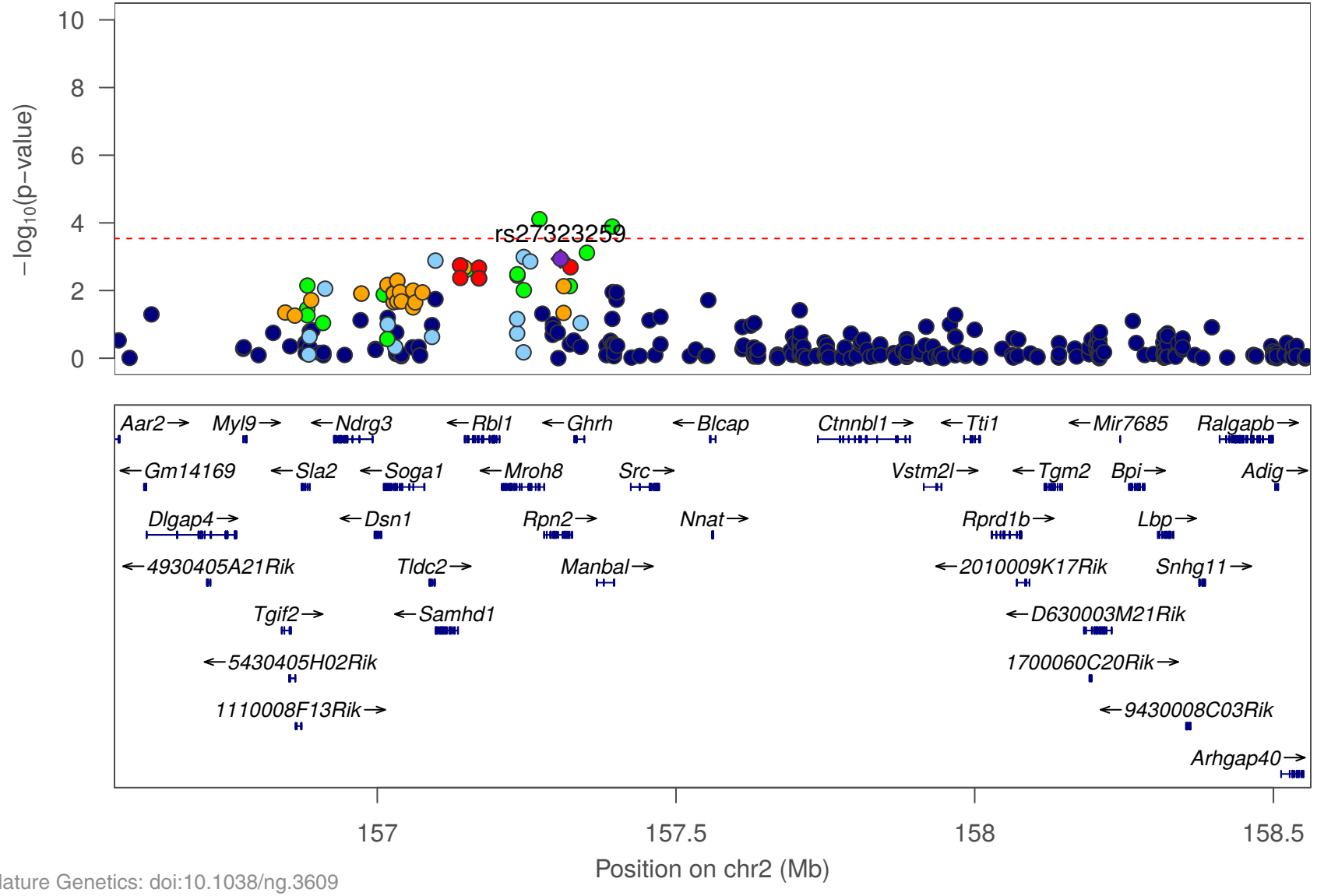
Nature Genetics: doi:10.1038/ng.3609

Gfra1 expression in striatum: Day 3 freezing to tone



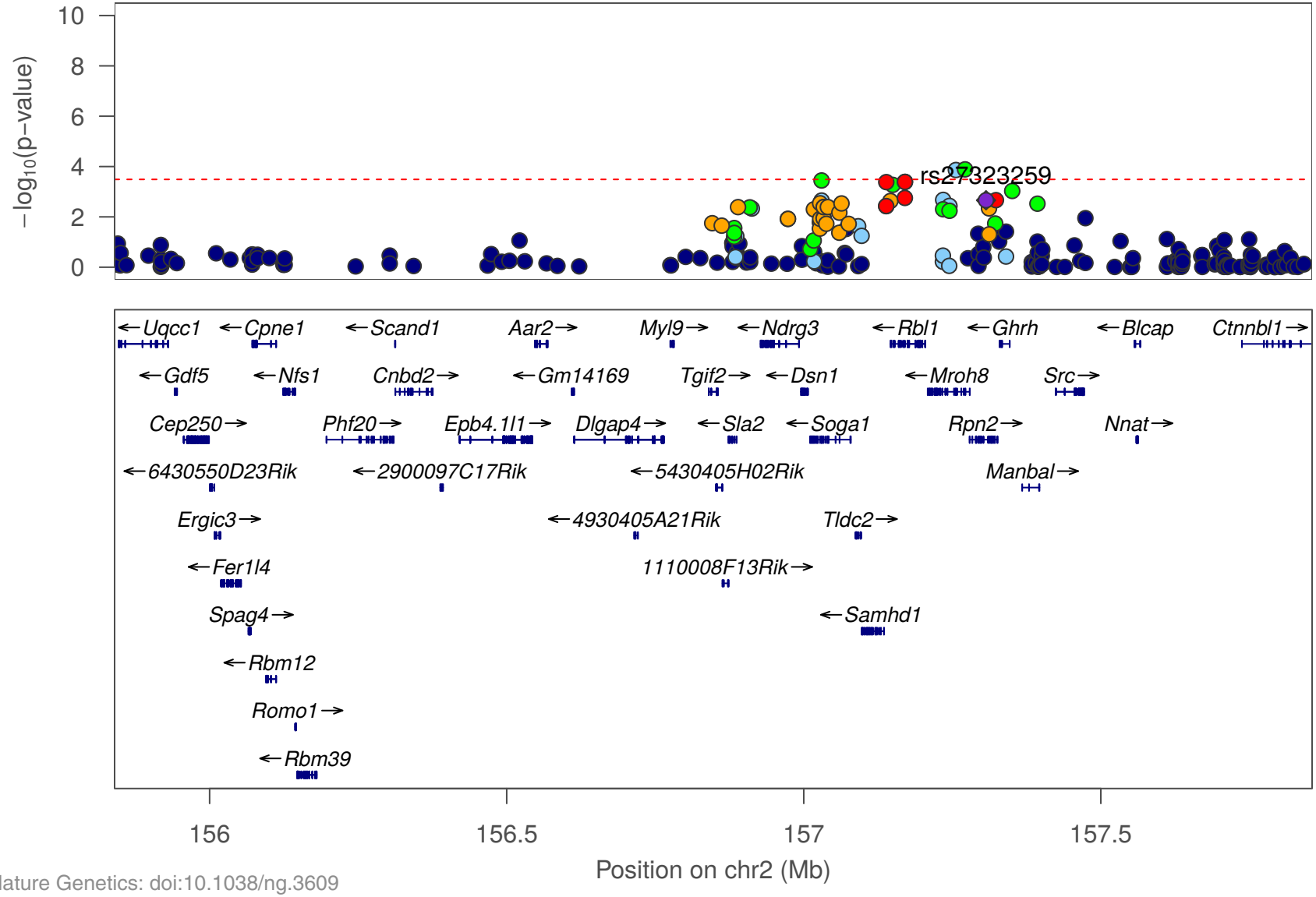
Nature Genetics: doi:10.1038/ng.3609

Nnat expression in hippocampus: Day 3 total distance 0-20 minutes



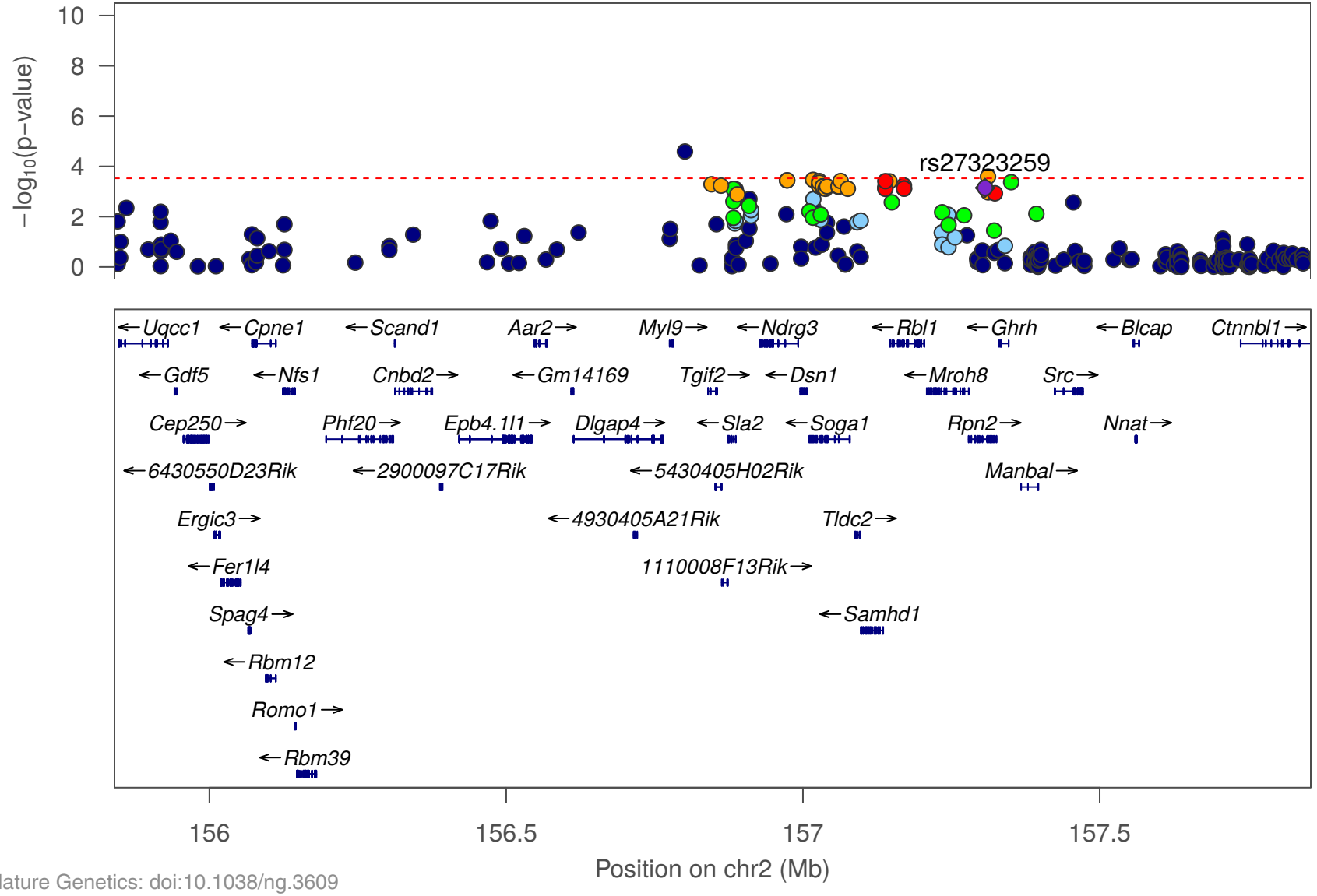
Nature Genetics: doi:10.1038/ng.3609

Tgif2 expression in hippocampus: Day 3 total distance 0-20 minutes



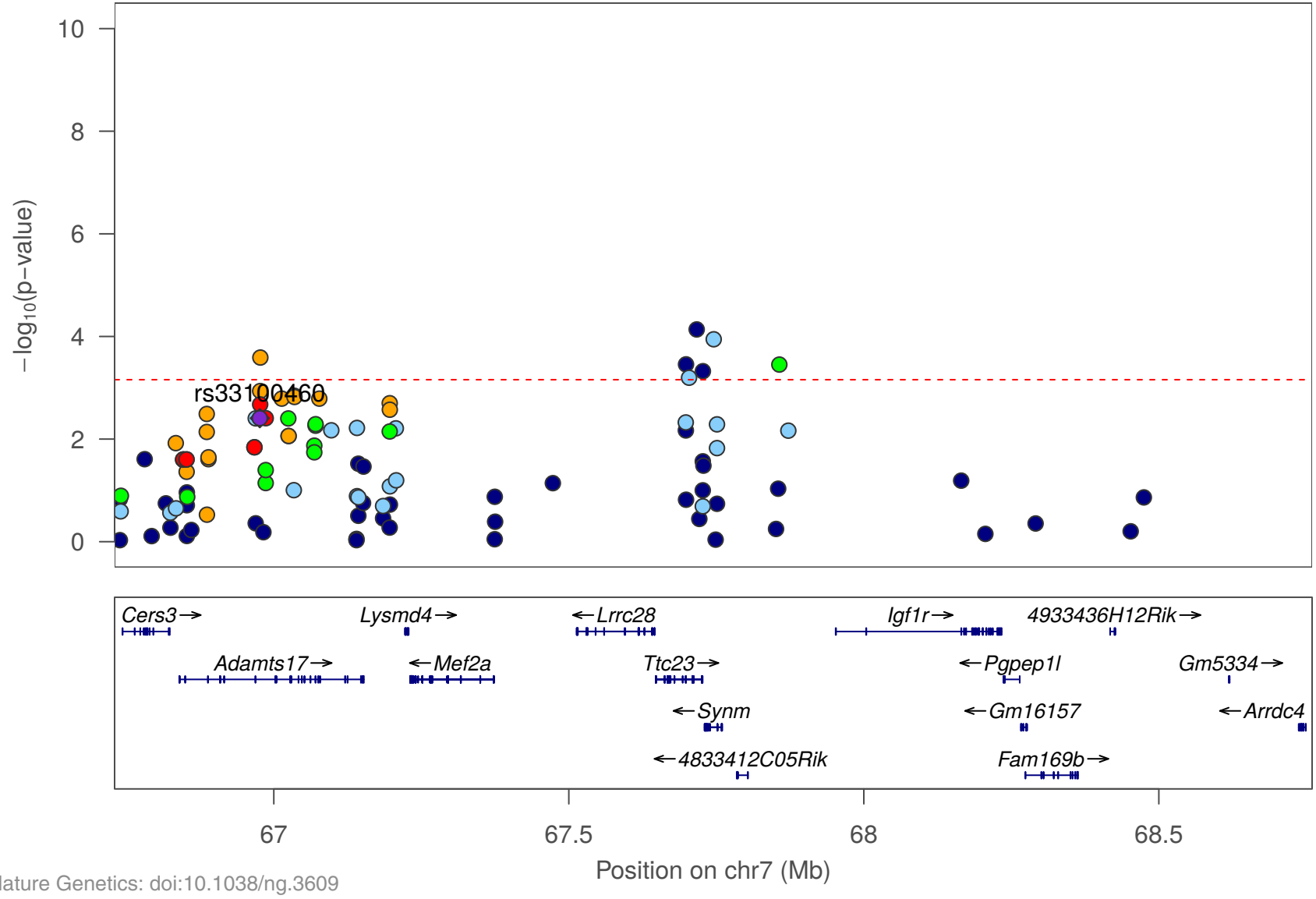
Nature Genetics: doi:10.1038/ng.3609

Tgif2 expression in striatum: Day 3 total distance 0-20 minutes



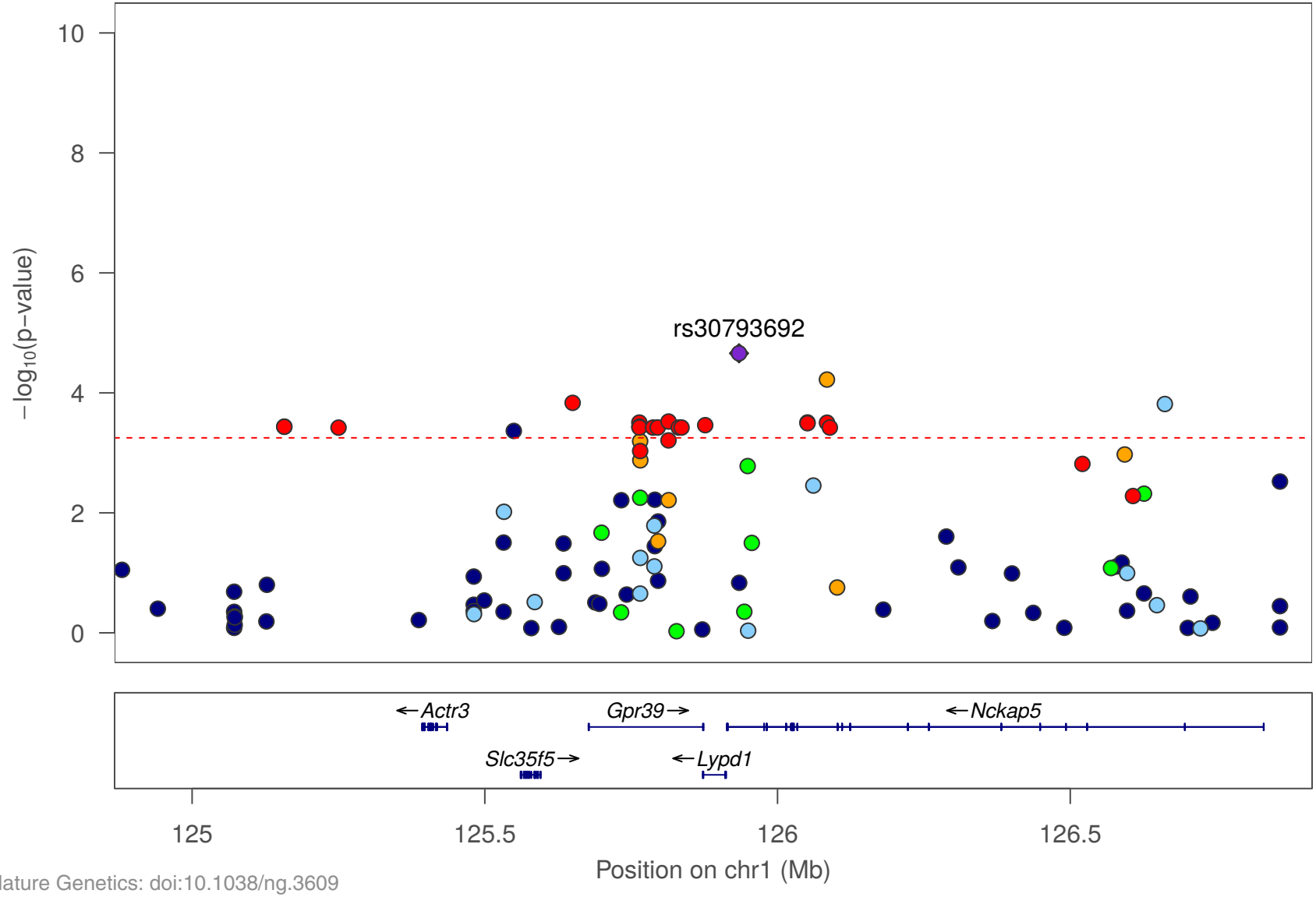
Nature Genetics: doi:10.1038/ng.3609

Synm expression in hippocampus: Day 3 vertical activity 0-15 minutes



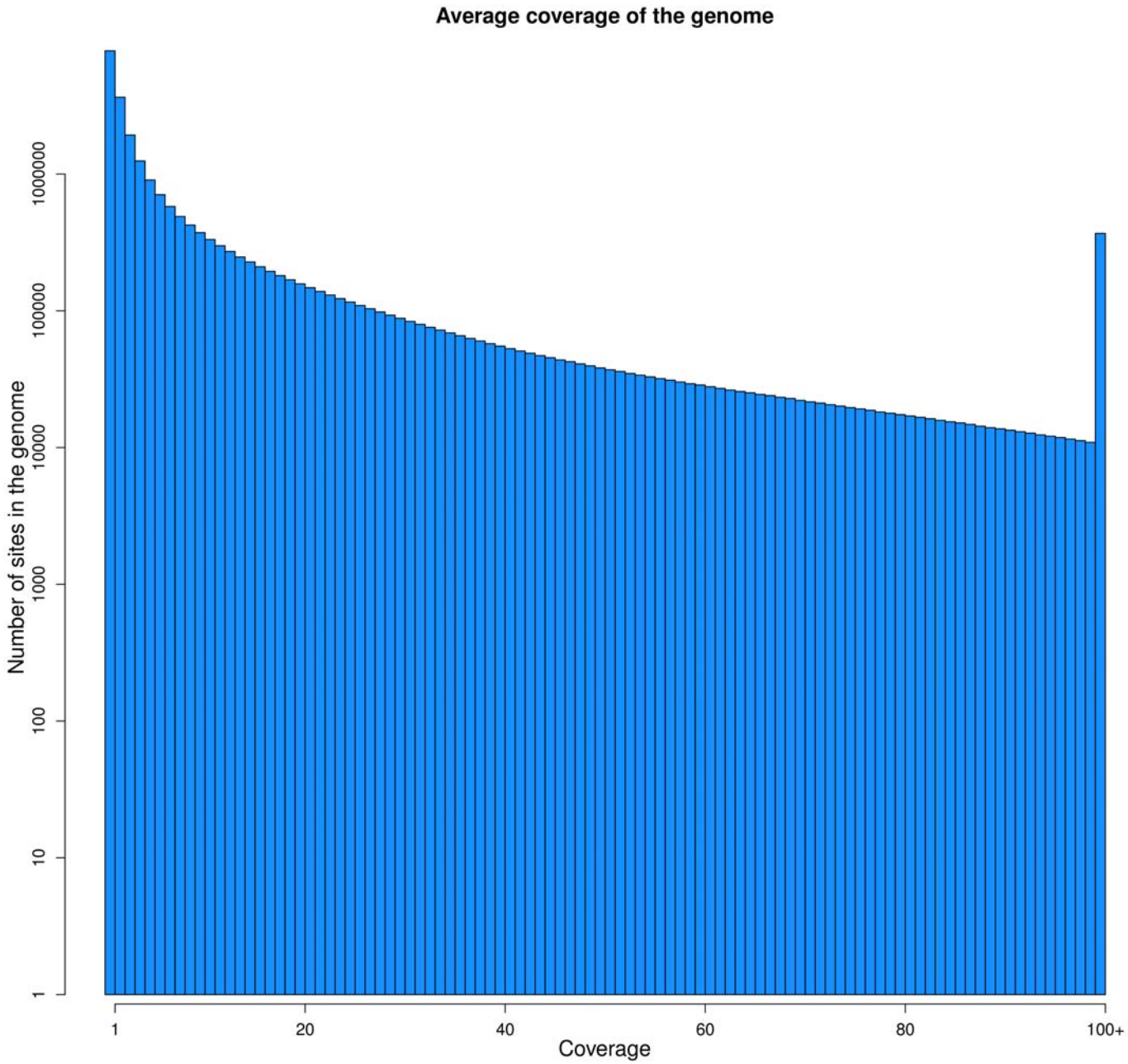
Nature Genetics: doi:10.1038/ng.3609

Lypd1 expression in hippocampus: Day 3 vertical activity 0-30 minutes

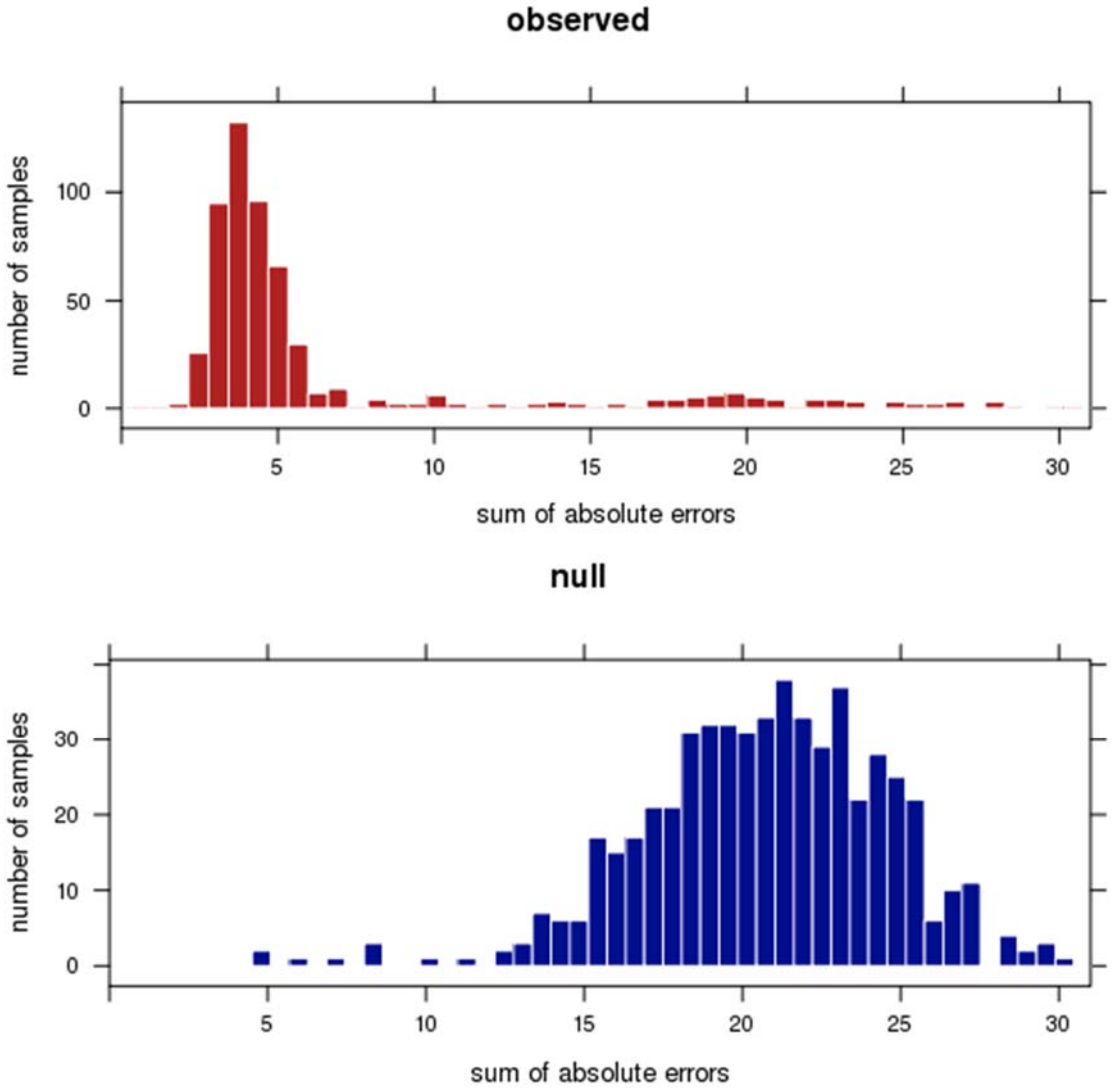


Nature Genetics: doi:10.1038/ng.3609

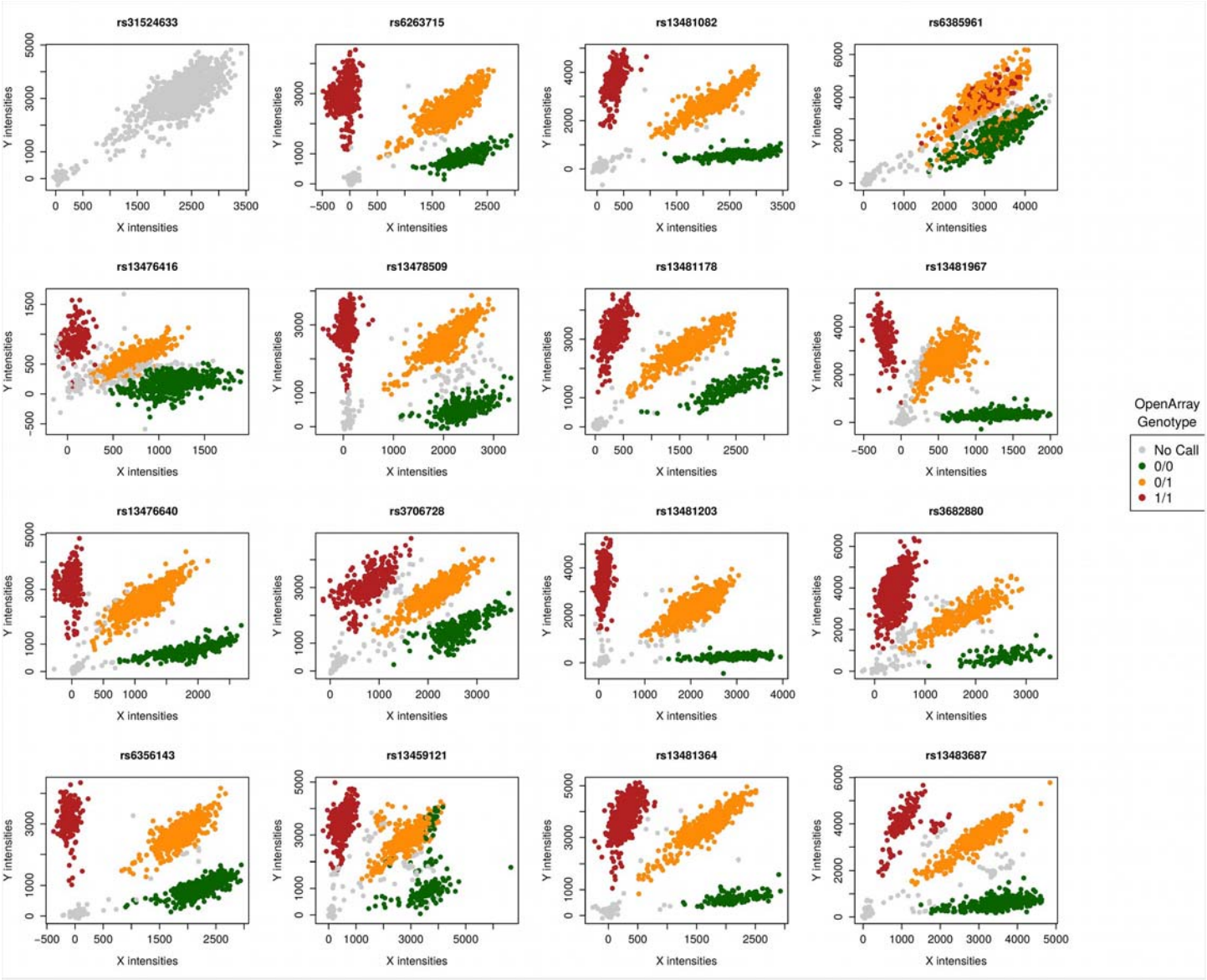
Supplementary Figure 23: Locuszoom plots for the association of the gene expression with index SNPs for behavioral QTLs. The points for each SNP are colored by the level of the linkage disequilibrium (r^2) with the index SNP, the SNP with the highest association to the quantitative trait.



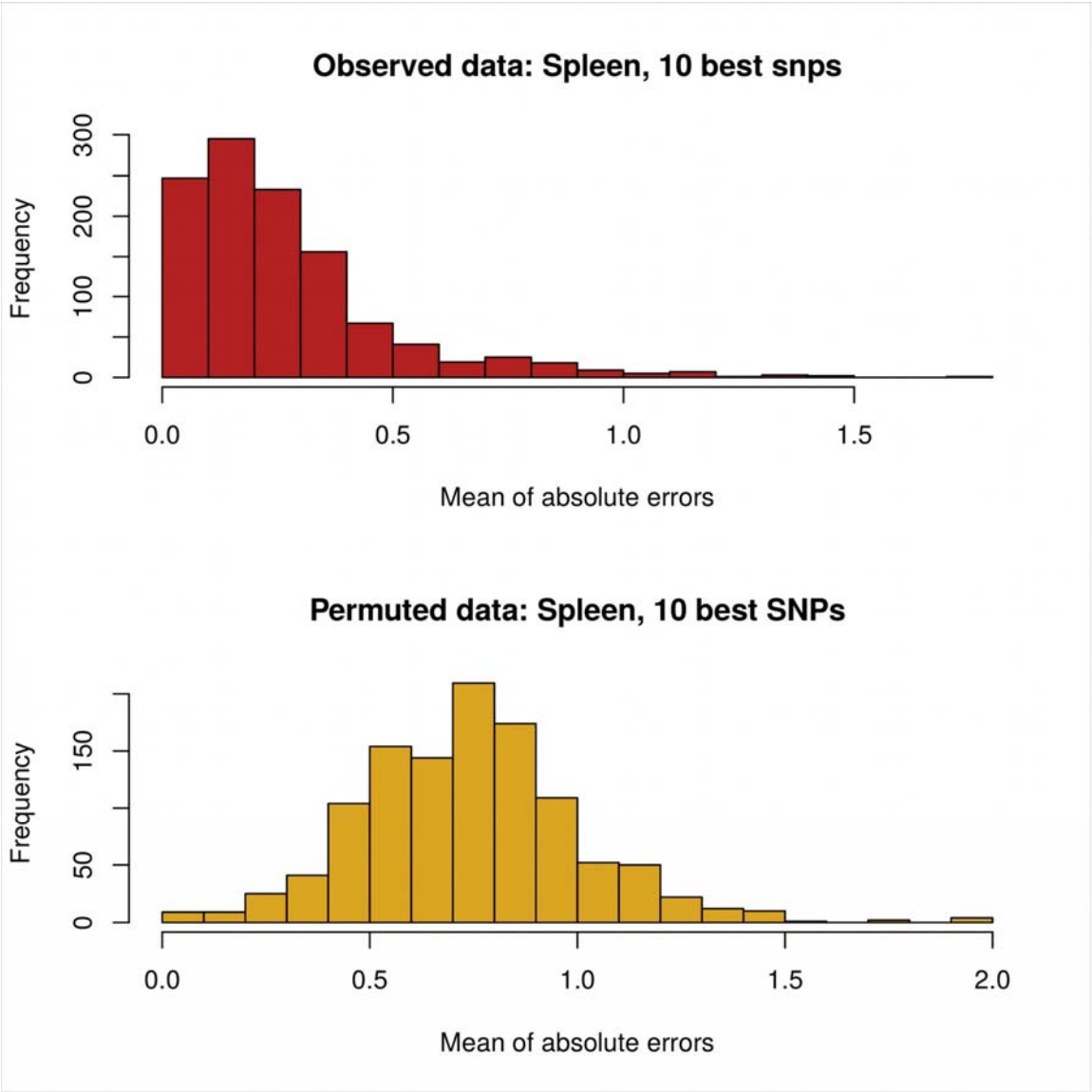
Supplementary Figure 24: Coverage distribution. The distribution of coverage across the genome averaged across all the samples that were sequenced.



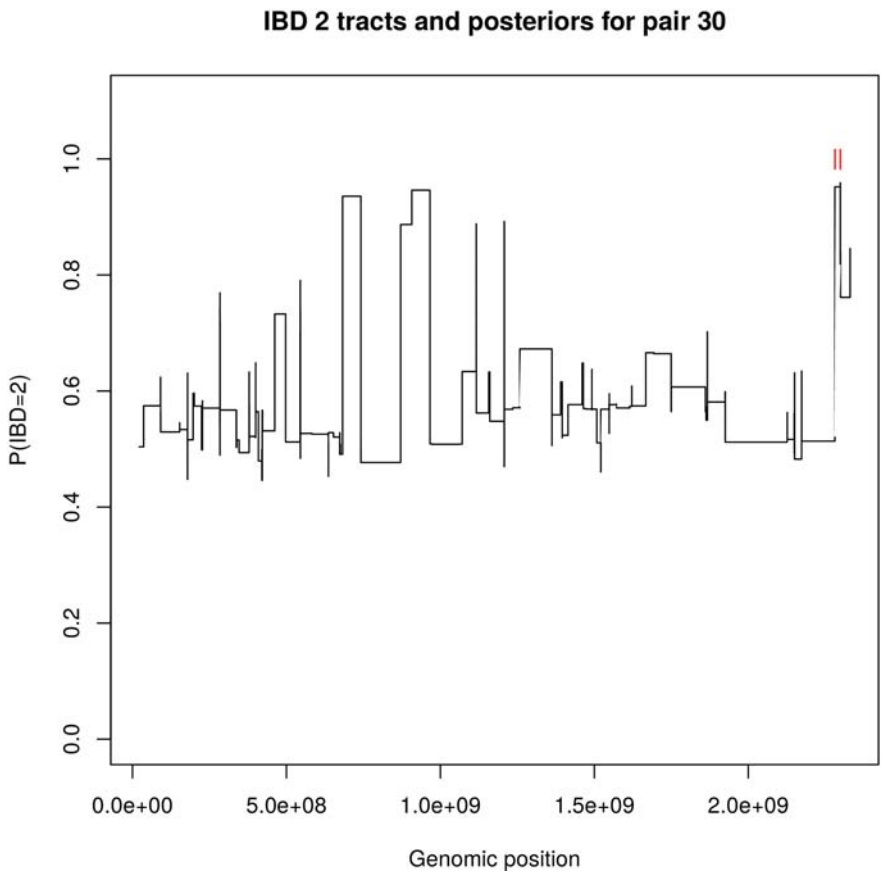
Supplementary Figure 25: Observed and permutation distribution of differences in GBS and Sequenom genotypes for 33 selected SNPs. Distribution of errors, computed as difference in Sequenom genotypes (SNP chip) and GBS dosages, plotted for the real data (red) and permuted data (blue). Most samples have a very low error rate (presumably corresponding to the GBS genotyping error rate), and a few samples exhibit elevated error rates.



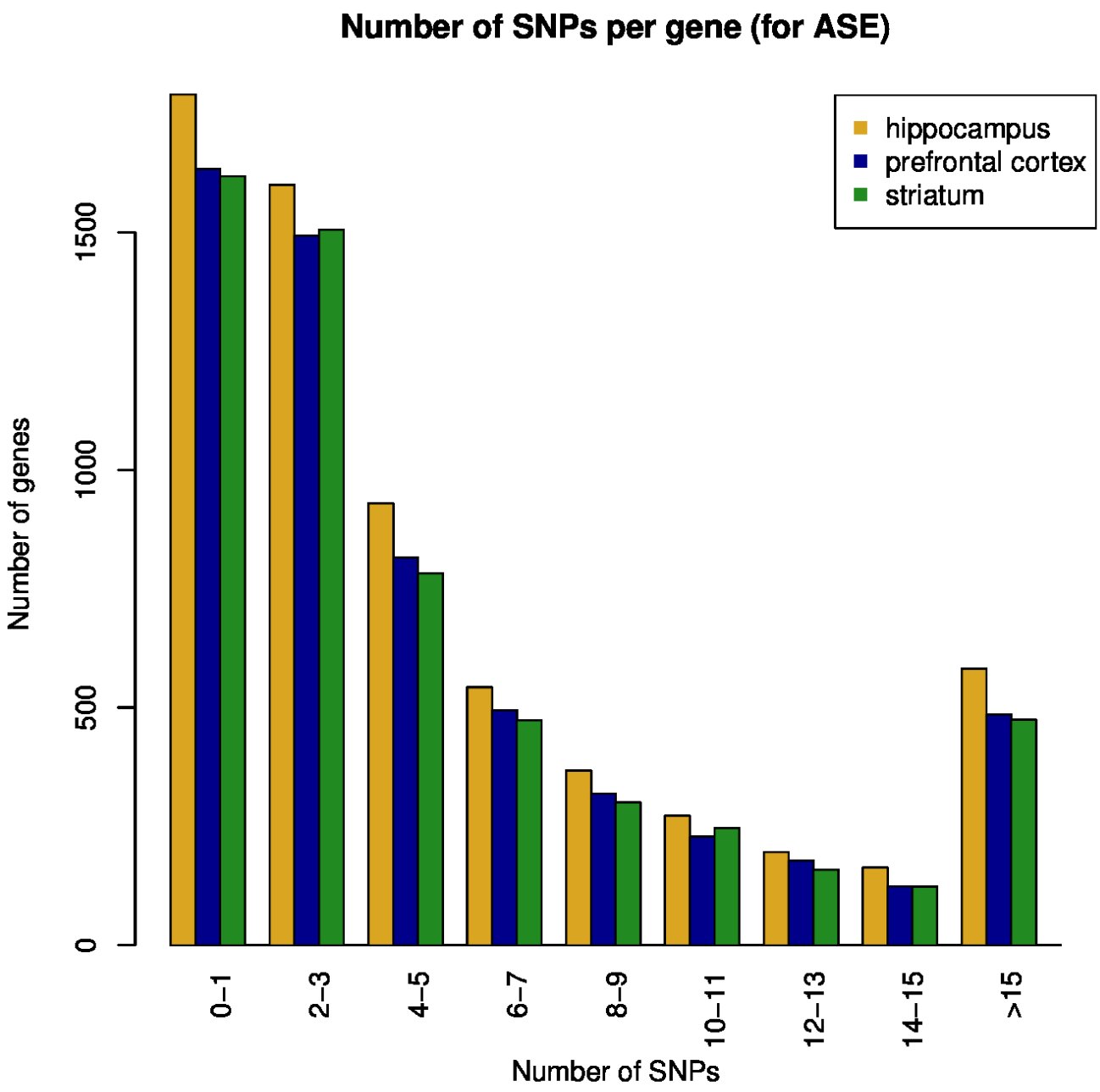
Supplementary Figure 26: The two-color intensity plots from Open Array genotypes for 16 SNPs. The samples have been colored based on the OpenArray genotype calls from the custom clustering algorithm. The first and the fourth SNP (rs31524633 and rs6385961) were excluded due to poor quality clustering and genotyping. Four other SNPs (rs13476416, rs3706728, rs13459121 and rs13483687) were excluded as they were not in Hardy-Weinberg equilibrium.



Supplementary Figure 27: Observed and permutation distribution of differences between OpenArray and GBS genotypes. The top panel shows the distribution of the mean absolute difference between the imputed GBS allele dosage and the OpenArray genotype. The bottom panel shows the same statistic obtained after randomly permuting the sample labels for the OpenArray genotypes.



Supplementary Figure 28: Posterior probability of sharing 2 alleles IBD. The estimated probability of sharing two alleles IBD for one pair of mice in our sample. This is one of the few pairs of mice that has at least one IBD 2 tract in the genome. The regions marked in red show the parts of the genome where the posterior probability is > 0.9.



Supplementary Figure 29: Distribution of number of ASE SNPs per gene. This plot shows the distribution of the number of SNPs per gene that passed the thresholds for the ASE test, i.e., SNPs with at least 10 high confidence heterozygote calls.

Supplementary References

1. Bryant, C. D. *et al.* A major QTL on chromosome 11 influences psychostimulant and opioid sensitivity in mice. *Genes Brain Behav.* **8,** 795–805 (2009).
2. Bryant, C. D. *et al.* A role for casein kinase 1 epsilon in the locomotor stimulant response to methamphetamine. *Psychopharmacology (Berl.)* **203,** 703–711 (2009).
3. Palmer, A. A. *et al.* Gene expression differences in mice divergently selected for methamphetamine sensitivity. *Mamm. Genome Off. J. Int. Mamm. Genome Soc.* **16,** 291–305 (2005).
4. Parker, C. C., Cheng, R., Sokoloff, G. & Palmer, A. A. Genome-wide association for methamphetamine sensitivity in an advanced intercross mouse line. *Genes Brain Behav.* **11,** 52–61 (2012).
5. Piazza, P. V. *et al.* Dopaminergic activity is reduced in the prefrontal cortex and increased in the nucleus accumbens of rats predisposed to develop amphetamine self-administration. *Brain Res.* **567,** 169–174 (1991).
6. Parker, C. C. *et al.* High-resolution genetic mapping of complex traits from a combined analysis of F2 and advanced intercross mice. *Genetics* **198,** 103–116 (2014).
7. Parker, C. C., Sokoloff, G., Cheng, R. & Palmer, A. A. Genome-wide association for fear conditioning in an advanced intercross mouse line. *Behav. Genet.* **42,** 437–448 (2012).
8. Ponder, C. A. *et al.* Selection for contextual fear conditioning affects anxiety-like behaviors and gene expression. *Genes Brain Behav.* **6,** 736–749 (2007).
9. Ponder, C. A., Munoz, M., Gilliam, T. C. & Palmer, A. A. Genetic architecture of fear conditioning in chromosome substitution strains: relationship to measures of innate (unlearned) anxiety-like behavior. *Mamm. Genome Off. J. Int. Mamm. Genome Soc.* **18,** 221–228 (2007).
10. Ponder, C. A. *et al.* Rapid selection response for contextual fear conditioning in a cross between C57BL/6J and A/J: behavioral, QTL and gene expression analysis. *Behav. Genet.* **38,** 277–291 (2008).
11. Aitchison, J. The Statistical Analysis of Compositional Data. *J. R. Stat. Soc. Ser. B Methodol.* **44,** 139–177 (1982).
12. Palmer, A. A. *et al.* Prepulse startle deficit in the Brown Norway rat: a potential genetic model. *Behav.*

Neurosci. **114,** 374–388 (2000).

13. Koch, M. Clinical relevance of animal models of schizophrenia. *Suppl. Clin. Neurophysiol.* **62,** 113–120 (2013).
14. Swerdlow, N. R. *et al.* Deficient prepulse inhibition in schizophrenia detected by the multi-site COGS. *Schizophr. Res.* **152,** 503–512 (2014).
15. Palmer, A. A. *et al.* Identification of quantitative trait loci for prepulse inhibition in rats. *Psychopharmacology (Berl.)* **165,** 270–279 (2003).
16. Palmer, A. A., Printz, D. J., Butler, P. D., Dulawa, S. C. & Printz, M. P. Prenatal protein deprivation in rats induces changes in prepulse inhibition and NMDA receptor binding. *Brain Res.* **996,** 193–201 (2004).
17. Samocha, K. E., Lim, J. E., Cheng, R., Sokoloff, G. & Palmer, A. A. Fine mapping of QTL for prepulse inhibition in LG/J and SM/J mice using F(2) and advanced intercross lines. *Genes Brain Behav.* **9,** 759–767 (2010).
18. Shanahan, N. A. *et al.* Chronic reductions in serotonin transporter function prevent 5-HT1B-induced behavioral effects in mice. *Biol. Psychiatry* **65,** 401–408 (2009).
19. Messina, G. & Cossu, G. The origin of embryonic and fetal myoblasts: a role of Pax3 and Pax7. *Genes Dev.* **23,** 902–905 (2009).
20. Bloemberg, D. & Quadrilatero, J. Rapid determination of myosin heavy chain expression in rat, mouse, and human skeletal muscle using multicolor immunofluorescence analysis. *PloS One* **7,** e35273 (2012).
21. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7,** 562–578 (2012).
22. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
23. Roberts, A., Pimentel, H., Trapnell, C. & Pachter, L. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinforma. Oxf. Engl.* **27,** 2325–2329 (2011).
24. Grabowski, P. P., Morris, G. P., Casler, M. D. & Borevitz, J. O. Population genomic variation reveals roles of history, adaptation and ploidy in switchgrass. *Mol. Ecol.* **23,** 4059–4073 (2014).

25. Elshire, R. J. *et al.* A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS One* **6,** e19379 (2011).

26. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma. Oxf. Engl.* **25,** 1754–1760 (2009).

27. Keane, T. M. *et al.* Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477,** 289–294 (2011).

28. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5,** e1000529 (2009).

29. Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8,** e1002967 (2012).

30. Albrechtsen, A. *et al.* Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genet. Epidemiol.* **33,** 266–274 (2009).

31. Gatti, D. M. *et al.* Quantitative trait locus mapping methods for diversity outbred mice. *G3 Bethesda Md* **4,** 1623–1633 (2014).

32. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).

33. Pritchard, J. K. & Przeworski, M. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* **69,** 1–14 (2001).

34. Eberle, M. A., Rieder, M. J., Kruglyak, L. & Nickerson, D. A. Allele frequency matching between SNPs reveals an excess of linkage disequilibrium in genic regions of the human genome. *PLoS Genet.* **2,** e142 (2006).

35. Hernandez, R. D. *et al.* Demographic histories and patterns of linkage disequilibrium in Chinese and Indian rhesus macaques. *Science* **316,** 240–243 (2007).

36. Valdar, W. *et al.* Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nat. Genet.* **38,** 879–887 (2006).

37. Yang, H. *et al.* A customized and versatile high-density genotyping array for the mouse. *Nat. Methods* **6,**

663–666 (2009).

38. Yang, H. *et al.* Subspecific origin and haplotype diversity in the laboratory mouse. *Nat. Genet.* **43,** 648–655 (2011).
39. Sittig, L. J., Carbonetto, P., Engel, K. A., Krauss, K. S. & Palmer, A. A. Integration of genome-wide association and extant brain expression QTL identifies candidate genes influencing prepulse inhibition in inbred F1 mice. *Genes Brain Behav.* **15,** 260–270 (2016).
40. Bennett, B. J. *et al.* A high-resolution association mapping panel for the dissection of complex traits in mice. *Genome Res.* **20,** 281–290 (2010).
41. Ghazalpour, A. *et al.* Hybrid mouse diversity panel: a panel of inbred mouse strains suitable for analysis of complex genetic traits. *Mamm. Genome Off. J. Int. Mamm. Genome Soc.* **23,** 680–692 (2012).
42. Cheng, R. *et al.* Genome-wide association studies and the problem of relatedness among advanced intercross lines and other highly recombinant populations. *Genetics* **185,** 1033–1044 (2010).
43. Astle, W. & Balding, D. J. Population Structure and Cryptic Relatedness in Genetic Association Studies. *Stat. Sci.* **24,** 451–471 (2009).
44. Kang, H. M. *et al.* Efficient control of population structure in model organism association mapping. *Genetics* **178,** 1709–1723 (2008).
45. Marchini, J., Cardon, L. R., Phillips, M. S. & Donnelly, P. The effects of human population structure on large genetic association studies. *Nat. Genet.* **36,** 512–517 (2004).
46. Newman, D. L., Abney, M., McPeek, M. S., Ober, C. & Cox, N. J. The importance of genealogy in determining genetic associations with complex traits. *Am. J. Hum. Genet.* **69,** 1146–1148 (2001).
47. Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* **11,** 459–463 (2010).
48. Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42,** 348–354 (2010).
49. Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nat. Methods* **8,** 833–835 (2011).

50. Listgarten, J. *et al.* Improved linear mixed models for genome-wide association studies. *Nat. Methods* **9,** 525–526 (2012).
51. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.* **46,** 100–106 (2014).
52. Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38,** 203–208 (2006).
53. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44,** 821–824 (2012).
54. Aulchenko, Y. S., de Koning, D.-J. & Haley, C. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics* **177,** 577–585 (2007).
55. Pirinen, M., Donnelly, P. & Spencer, C. C. A. Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. *Ann. Appl. Stat.* **7,** 369–390 (2013).
56. Kuo, C.-L. & Feingold, E. What's the best statistic for a simple test of genetic association in a case-control study? *Genet. Epidemiol.* **34,** 246–253 (2010).
57. Cheng, R. & Palmer, A. A. A simulation study of permutation, bootstrap, and gene dropping for assessing statistical significance in the case of unequal relatedness. *Genetics* **193,** 1015–1018 (2013).
58. Abney, M., McPeek, M. S. & Ober, C. Estimation of variance components of quantitative traits in inbred populations. *Am. J. Hum. Genet.* **66,** 629–650 (2000).
59. Hayes, B. J., Visscher, P. M. & Goddard, M. E. Increased accuracy of artificial selection by using the realized relationship matrix. *Genet. Res.* **91,** 47–60 (2009).
60. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55,** 997–1004 (1999).
61. Clayton, D. G. *et al.* Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat. Genet.* **37,** 1243–1246 (2005).
62. Broman, K. W. & Sen, S. *A guide to QTL mapping with R/qtl*. (Springer New York, 2009).
63. Peirce, J. L. *et al.* Genome Reshuffling for Advanced Intercross Permutation (GRAIP): simulation and

permutation for advanced intercross population analysis. *PloS One* **3,** e1977 (2008).

64. Zou, F. *et al.* Quantitative trait locus analysis using recombinant inbred intercrosses: theoretical and empirical considerations. *Genetics* **170,** 1299–1311 (2005).
65. Churchill, G. A. & Doerge, R. W. Empirical threshold values for quantitative trait mapping. *Genetics* **138,** 963–971 (1994).
66. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era--concepts and misconceptions. *Nat. Rev. Genet.* **9,** 255–266 (2008).
67. Wray, N. R. *et al.* Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* **14,** 507–515 (2013).
68. Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet.* **91,** 1011–1021 (2012).
69. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42,** 565–569 (2010).
70. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet.* **9,** e1003264 (2013).
71. Hastie, T., Tibshirani, R. & Friedman, J. *Elements of Statistical Learning: data mining, inference, and prediction. 2nd Edition.* (Springer New York, 2009).
72. Chipman, H. *et al.* The Practical Implementation of Bayesian Model Selection. *Lect. Notes-Monogr. Ser.* **38,** 65–134 (2001).
73. Gelman, A. *et al. Bayesian data analysis*. (Chapman & Hall/CRC, 2013).
74. Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88,** 294–305 (2011).
75. Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 1193–1198 (2012).
76. Fusi, N., Stegle, O. & Lawrence, N. D. Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies. *PLoS Comput. Biol.* **8,** e1002330 (2012).

77. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7,** 500–507 (2012).

78. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464,** 768–772 (2010).

79. Park, C. C. *et al.* Gene networks associated with conditional fear in mice identified using a systems genetics approach. *BMC Syst. Biol.* **5,** 43 (2011).

80. Farber, C. R. *et al.* Mouse genome-wide association and systems genetics identify Asxl2 as a regulator of bone mineral density and osteoclastogenesis. *PLoS Genet.* **7,** e1002038 (2011).