Combined sequence-based and genetic mapping analysis of complex traits in outbred rats

Rat Genome Sequencing and Mapping Consortium*

Genetic mapping on fully sequenced individuals is transforming understanding of the relationship between molecular variation and variation in complex traits. Here we report a combined sequence and genetic mapping analysis in outbred rats that maps 355 quantitative trait loci for 122 phenotypes. We identify 35 causal genes involved in 31 phenotypes, implicating new genes in models of anxiety, heart disease and multiple sclerosis. The relationship between sequence and genetic variation is unexpectedly complex: at approximately 40% of quantitative trait loci, a single sequence variant cannot account for the phenotypic effect. Using comparable sequence and mapping data from mice, we show that the extent and spatial pattern of variation in inbred rats differ substantially from those of inbred mice and that the genetic variants in orthologous genes rarely contribute to the same phenotype in both species.

Unraveling the complex relationship between phenotype and genotype poses a formidable challenge for biomedical science. Despite considerable success in identifying genetic loci that contribute to quantitative variation and disease susceptibility in humans[1], in most organisms, the causal genetic variants at loci that contribute to complex phenotypes remain unclear[2]. Finding the responsible molecular changes would allow an understanding of how phenotypic variation arises and would confirm the identity of relevant genes.

In this report, we present results from an outbred rat heterogeneous stock (hereafter, NIH-HS) in a combined sequence-based and genetic mapping analysis of 160 phenotypes. The NIH-HS, established in the 1980s at the US National Institutes of Health (NIH), is descended from eight inbred progenitors[3]—BN/SsN, MR/N, BUF/N, M520/N, WN/N, ACI/N, WKY/N and F344/N—containing segregating variation representative of that found in commonly used laboratory rats.

Heterogeneous stocks have three characteristics suited to genetic mapping: (i) quantitative trait loci (QTLs) can be resolved to megabase resolution; (ii) the complete sequence of genotyped heterogeneous-stock animals can be imputed with high accuracy from the progenitor genomes; and (iii) the population has a well-defined haplotype space that can be exploited to determine whether genetic association is caused by single sequence variants or by haplotypes[4–6]. The distinction between haplotypic and single-marker association is fundamental to understanding the signals from genome-wide association studies (GWAS), where it is unknown how often causality can be attributed to a single variant. In natural populations, it is rarely feasible to test for haplotypic effects because of the difficulty of estimating the large number of unknown rare haplotypes[7].

Here we describe the sequence of the 8 progenitor strains, the development of a rat SNP array, the genotyping and phenotyping of 1,407 outbred NIH-HS rats and the mapping of hundreds of QTLs. We use the haplotypic properties of the NIH-HS to investigate the molecular basis of these QTLs.

RESULTS

Sequence analysis

We generated SOLiD sequence data for the eight NIH-HS inbred founder strains equivalent to an average of 22× base coverage. After mapping sequence to the reference strain (BN/NHsdMcwi)[8], we report our results with respect to the accessible genome, which represents ~88% of the reference genome (**Table 1**). We identified 7.2 million SNPs (containing 19.8 million genotypes differing from the reference in at least 1 strain), 633,000 indels (<10 bp, with the majority consisting of 1-bp (79.3%) or 2-bp (12.3%) changes) and 44,000 structural variants.

We assessed the sensitivity and specificity of variant calls by comparison with 2.1 Mb of DNA from one non-reference strain, LE/Stm, finished to an estimated accuracy of 1 error per 100,000 bp[9]. Although LE/Stm is not an NIH-HS progenitor strain, it is one of the few non-reference rat strains cloned into a library of BACs (and thus suitable for highly accurate clone-based sequencing)[9] and one that similarly diverged from the reference strain (BN/NHsdMcwi). Comparison of SOLiD and capillary sequencing variant calls showed that 2.7% of SNPs, 2.2% of indels and 16.7% of structural variants were false positive calls. These error rates were independently confirmed in the NIH-HS strains by analysis of a randomly selected subset of variants using PCR-based resequencing, which confirmed all selected SNPs (84/84) and indels (80/80) and most structural variants (53/54). In contrast, false negative rates were much higher: 17.2% for SNPs, 41.4% for indels and 65% for structural variants. Most false negative SNPs and indels are next to repeats (77.9% and 80.8%, respectively).

We summarized the variation in each strain (**Table 1**). Excluding BN/SsN (which is a substrain of the reference and consequently has far fewer differences than the other strains), the average number of SNPs per strain was 2.8 million.

*A full list of authors and affiliations appears at the end of the paper.

Received 2 November 2012; accepted 25 April 2013; published online 26 May 2013; doi:10.1038/ng.2644

Table 1 Sequence variation in the eight progenitor strains of NIH-HS rats

Strain	Mapped data (Gb)	Coverage	Inaccessible genome (%)	SNPs	Private SNPs	Indels	Private indels	Structural variants	Private structural variants
ACI/N	65.9	26.3	12.6	2,883,405	228,468	166,425	12,646	19,499	756
BN/SsN	54.4	21.7	9.4	71,038	563,308	0	14,839	27	4,203
BUF/N	62.3	24.9	12.7	2,748,633	125,202	172,934	7,195	22,176	1,002
F344/N	77.9	31.1	11.8	2,831,144	97,951	157,522	5,007	25,257	1,003
M520/N	72.5	28.9	12.3	2,836,898	89,277	170,031	5,008	24,090	915
MR/N	62.4	24.9	12.3	2,664,124	223,514	151,099	12,005	18,306	1,004
WKY/N	63.4	25.3	12.1	3,088,953	496,327	164,634	23,979	28,270	3,357
WN/N	62.3	24.9	12.2	2,698,493	249,563	154,769	13,541	18,563	700

Shown for each strain is the amount of sequence mapped to the reference, the coverage, the percent of the genome deemed inaccessible and the counts of the three classes of variants compared to the reference strain. Private variants are variants that distinguish a specified strain from all others; most of the alleles private to BN/SsN are reference alleles.

Nucleotide diversity in NIH-HS progenitors

We examined sequence diversity among the NIH-HS progenitors (**Fig. 1**), identifying the following characteristics of this diversity. First, diversity between all pairs of strains was similar, such that there were no strains that were extremely sequence divergent (**Supplementary Fig. 1**). Second, in total, 29% of 7.2 million SNPs were private to a particular strain; hence, unique haplotypes are relatively common in the NIH-HS. Third, regions of low diversity were small (median of 400 kb), with no blocks over 35 Mb in length (**Fig. 1a**). Within divergent regions, there was a median of 151 differences per 100 kb (**Fig. 1b**).

In comparison with the eight inbred strains that founded the mouse heterogeneous stock[4,10], the rat founders were less diverse (10.2 million SNPs in the mouse founders), but diversity was more homogeneous: in the mouse genomes, long tracts of identical haplotypes alternate with segments of much greater diversity (**Fig. 1a**,**b**).

Phenotypes and genotypes

NIH-HS rats were phenotyped with a protocol that includes six disease models (anxiety, diabetes, hypertension, aortic elastic lamina ruptures, multiple sclerosis and osteoporosis) and measures of risk factors for common diseases (for example, lipid and cholesterol levels and cardiac hypertrophy)[11] (**Table 2**). In total, 160 phenotypes were measured (**Supplementary Table 1**). We selected 1,407 animals and 198 non-phenotyped parents for genotyping together with the heterogeneous stock founders.

We designed a high-density Affymetrix SNP genotyping array (RATDIV), using sequences from 13 inbred strains, which interrogated 803,485 SNPs. The SOLiD and RATDIV calls agreed at 99.98% of the 560,000 SNPs segregating in the 8 NIH-HS founders. We genotyped the NIH-HS with this array and reconstructed the mosaics of NIH-HS founder haplotypes from 265,551 polymorphic high-quality SNPs. In the NIH-HS, the mean minor allele frequency (MAF) was 22% (**Fig. 1c**), and linkage disequilibrium (LD) fell below 0.2 (median r^2) within 1 Mb of autosomal SNPs (**Fig. 1d**). Four pairs of loci showed high interchromosomal LD, owing to misassembly of the reference sequence used here (Rnor3.4); these loci were excluded from the analysis (**Supplementary Table 2**).

QTLs

The NIH-HS contains individuals of varying relatedness that generate population structure in the genotypes and, hence, false positive genetic associations. We evaluated two strategies for dealing with relatedness: mixed models in which the genotypic similarity matrix between individuals modeled their phenotypic correlation[12] and resampling methods to identify loci that replicate consistently across multiple QTL models fitted on subsamples of the mapping population[13]. In both strategies, QTLs were detected by haplotype association[14].

We compared the methods by simulation to determine which best controlled the false positive rate while retaining power. Mixed models performed better than resampling when phenotypes were simulated to have a normal distribution, but the reverse was true for phenotypes that did not have a normal distribution (that is, binary phenotypes and those with a negative binomial distribution). Because these methods have different advantages, we mapped all traits with both, but we only report those QTLs detected at false discovery rate (FDR) of 10% by the method that performed best for each trait (thresholds are given in **Supplementary Table 1**). A genome scan for one phenotype (platelet aggregation) is shown (**Fig. 2**) in which three loci were identified with FDR of 10%.

We identified 355 QTLs for 122 phenotypes, with a mean of 2.9 QTLs per phenotype (**Supplementary Table 3**). The number of QTLs per phenotype and the QTL effect sizes (**Fig. 1e**) have markedly skewed distributions, with a median effect size of 5% (mean effect size of 6.5%). Large-effect QTLs were rare: only 22 QTLs explained more than 15% of the variance. We identified 28 QTLs that explained less than 2.5% of the phenotypic variance.

The correlation between heritability and the total variance explained jointly by the detected QTLs is shown (**Fig. 1f**). On average, the QTLs explained 42% of the heritable phenotypic variance. When considering QTLs mapped in other rat crosses in the Rat Genome Database, there was significant overlap with NIH-HS QTLs for the number of arterial elastic lamina ruptures, total cholesterol levels and heart weight (at a nominal P value of 0.05; **Supplementary Table 4**).

We estimated the confidence intervals for QTL locations by simulating a large number of QTLs throughout the genome with various effect sizes, and we calculated the distribution of the widths of the confidence intervals as a function of their significance (**Supplementary Fig. 2**). The median size of the 90% confidence interval was 4.5 Mb, on average containing more than 40 genes.

Incorporation of sequence with mapping data

We investigated the extent to which our near-complete catalog of segregating sequence variants would identify genes and causative mutations. The heterogeneous stock permits a test, called merge analysis[6], of whether a variant is responsible for phenotypic variation, under the assumption that a single imputed variant or variants on a single progenitor haplotype are causal. Because genetic variation segregates in the form of progenitor haplotypes in the heterogeneous stock, QTLs can always be explained by variation in the haplotypes. When a QTL corresponds to a single variant though, genotypic variation at that variant will explain phenotypic variation better than progenitor haplotypes. To measure whether a single variant explained a QTL, we calculated difference (d) as $\log P_{merge} - \log P_{haplotype}$, where $\log P_{haplotype}$ is the maximum negative $\log_{10} P$ value of the haplotype test of no association and $\log P_{merge}$ is the maximum of all merge $\log_{10} P$ values of

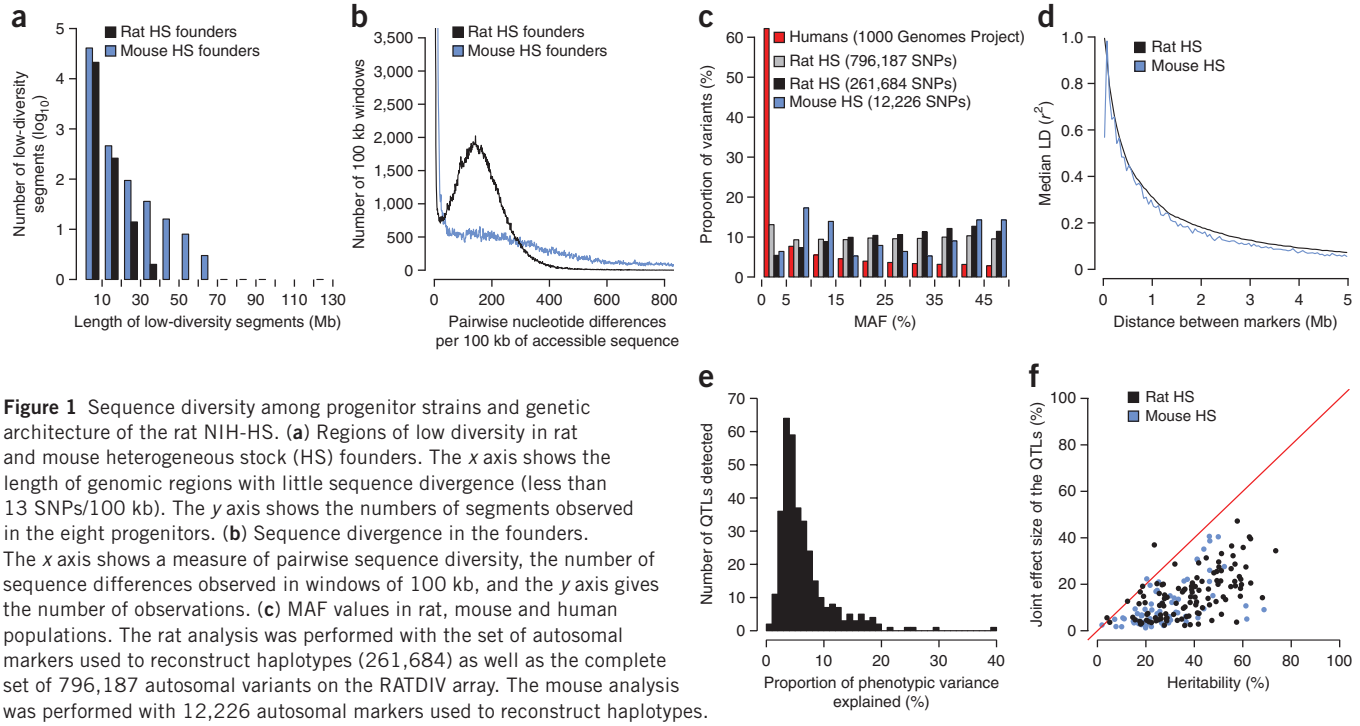


Figure 1 Sequence diversity among progenitor strains and genetic architecture of the rat NIH-HS. (**a**) Regions of low diversity in rat and mouse heterogeneous stock (HS) founders. The *x* axis shows the length of genomic regions with little sequence divergence (less than 13 SNPs/100 kb). The *y* axis shows the numbers of segments observed in the eight progenitors. (**b**) Sequence divergence in the founders. The *x* axis shows a measure of pairwise sequence diversity, the number of sequence differences observed in windows of 100 kb, and the *y* axis gives the number of observations. (**c**) MAF values in rat, mouse and human populations. The rat analysis was performed with the set of autosomal markers used to reconstruct haplotypes (261,684) as well as the complete set of 796,187 autosomal variants on the RATDIV array. The mouse analysis was performed with 12,226 autosomal markers used to reconstruct haplotypes. (**d**) The extent of LD (r^2) in the rat NIH-HS. Distances between pairs of autosomal markers were binned (*x* axis). The *y* axis shows the median of the corresponding distribution of LD values. (**e**) The distribution of effect sizes for the 343 loci mapped by mixed models in the rat NIH-HS. The *x* axis shows the proportion of phenotypic variance attributable to each locus. (**f**) The proportion of heritability that can be explained by the joint effect of the QTLs detected for each phenotype. Each data point represents a single phenotype, with the *x* axis showing the heritability and the *y* axis showing the joint QTL effect for that phenotype.

variants included within the QTL. Any imputed variant with a merge $\log_{10} P$ value that exceeded the maximum haplotype $\log_{10} P$ value was termed a candidate variant. If d was <0, then no candidate variants existed at the QTL. We investigated the characteristics of candidate variants at 343 QTLs mapped using mixed models: at 131 QTLs (38%) we identified at least 1 candidate variant (**Supplementary Table 3**).

There are three ways in which focusing on these candidate variants helps identify genes at a QTL. First, we increase resolution by ruling out a causal role for the great majority of sequence variants (usually over 90%) within most QTLs. We found 28 QTLs at which only a single gene contained candidate variants (**Table 3**). One example was *Ctnnd2* (encoding catenin δ2) at a QTL for an anxiety-related phenotype (**Fig. 3a**). CTNND2 is a protein found in complexes with cadherin cell adhesion molecules at neuronal synapses[15].

Table 2 Summary of phenotypes collected

Phenotype	Disease model	Number of measures	Age (weeks)
Coat color		4	7
Wound healing		1	7, 17
Fear-related behaviors	Anxiety	10	8–10
Glucose tolerance	Type 2 diabetes	6	11
Cardiovascular function	Hypertension	2	12
Body weight	Obesity	1	13
Basal hematology		26	13
Basal immunology		34	13
Induced neuroinflammation	Multiple sclerosis	11	13–17
Bone mass and strength	Osteoporosis	43	17
Arterial elastic lamina ruptures		6	17
Serum biochemistry		15	17
Renal agenesis		1	17

Another example involved a locus influencing heart weight, where, out of 82 coding genes within the QTL, only *Shank2* contained candidate SNPs (**Fig. 3b**). *Shank2* encodes a synaptic protein[16] not previously associated with cardiovascular physiology.

Second, merge analysis identifies some candidate variants in coding regions. Those predicted to affect protein structure are more likely to be causal. Thus, we identified a potential causal nucleotide variant in a QTL for antibody recognition of CD45RC on $CD4^+$ and $CD8^+$ T cells (**Fig. 3c**). The antibody used binds to the CD45RC isoform, which expresses a C domain, encoded by exon 6, in which we found a candidate variant changing an amino acid (p.Arg114His).

At 43 out of 91 nonsynonymous candidate variants, where similar protein structures were available[17], we predicted the structural consequences of mutations (for a further 48 candidate variants, there were no homologies with known protein structures). Nine genes (**Table 3**) contained candidate variants for which structural evidence suggested that protein structure or interactions might be altered.

An example is shown (**Fig. 3d**) for the protein Tbx21, encoded by a gene within a QTL influencing the proportion of $CD4^+$ cells with high expression of CD25. Here the candidate variant changed glycine to arginine (p.Gly175Arg). The substitution with arginine could alter the DNA-binding characteristics of this protein.

The crystal structure of human ABCB10, a mitochondrial transporter induced by GATA1 during erythroid differentiation[18,19], is shown (**Fig. 3e**). The candidate variant p.Thr233Met, predicted to influence mean red blood cell volume, mapped to a position in the protein structure where the side chain of the residue points to the center of the transporter channel (**Fig. 3e**). Threonine has a polar, uncharged side chain, whereas methionine has a hydrophobic side chain, and the difference between their structures probably results in altered transporter function.

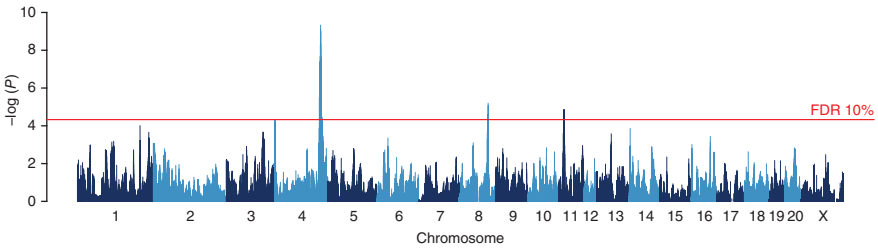


Figure 2 Genome scan for platelet aggregation. The scan shows the results of a haplotype-based mixed model. The *y* axis shows the negative log *P* values for association with variation in platelet aggregation. The association peak on chromosome 4 harbors the von Willebrand factor gene that was identified through sequence analysis as the causative gene.

Third, merge analysis eliminates candidate genes at a QTL that are distant from any candidate variant. This approach confirmed a well-established relationship between a cluster of apolipoprotein genes at a QTL on chromosome 1 and cholesterol biosynthesis (high-density lipoprotein (HDL), low-density lipoprotein (LDL) and total cholesterol). Similarly, merge analysis identified a locus influencing platelet aggregation on chromosome 4 that harbors the von Willebrand factor gene (*Vwf*), encoding a key glycoprotein involved in blood coagulation.

Merge analysis also contributed to an understanding of the pathogenesis of experimental autoimmune encephalomyelitis (EAE), an autoimmune neuroinflammatory disease with clinical and pathological similarities to multiple sclerosis[20]. The major histocompatibility complex (MHC) class II region on chromosome 20 (*Eae1*) is known to influence EAE susceptibility. However, attempts to identify the responsible gene have had limited success. In this study, the two variants most likely to underlie the QTL effect on chromosome 20 (with the highest merge $\log_{10} P$ value) were a variant in an intron of *Btnl2* and a variant 274 bp upstream of *RT1-Db1*, both in the MHC class II region. The human ortholog of *RT1-Db1*, *HLA-DRB1*, is associated with multiple sclerosis, with risk allele *HLA-DRB1*15:01* (ref. 21).

Single variants rarely account for NIH-HS QTL genetic effects

Unexpectedly, 212 QTLs (62%) had no candidate variant (**Fig. 4a**). We considered four explanations for this observation: (i) causative variants were missing from the sequence catalog; (ii) haplotype mapping was biased toward QTLs without candidate variants; (iii) the merge analysis underestimated statistical significance compared to single-marker association analysis; and (iv) there were multiple causal variants at a single QTL.

First, causal variants may have been missed because our sequence data were incomplete. Despite LD extending over a few megabases, not all variants were tagged by a nearby variant with identical strain distribution pattern (SDP) in the founders. For example, only 50% of the structural variants were tagged by a SNP lying within 1 Mb of the variation.

However, because only a limited set of possible SDPs exist in the heterogeneous stock, we can test whether missing genotypes are responsible for the inability to detect candidate variants. We generated SDPs for all possible diallelic and triallelic variants at every locus within the 212 QTLs and tested each by merge analysis to determine how many would have been candidate variants. Only 44 QTLs had candidate diallelic variants, and 165 had diallelic or triallelic variants. Thus, if the effect for each QTL were attributable to a single diallelic variant that we had not sequenced, there would still be 168 QTLs (49%) without a candidate variant. If the effect were attributable to a diallelic or triallelic variant, the fraction would be reduced to 14%. However, triallelic SNPs are very uncommon and are therefore unlikely to explain the large number of QTLs without candidate variants.

Second, haplotype mapping might simply not be powerful enough to detect candidate variants or might be biased toward QTLs without candidate variants. We addressed the first possibility by simulation (**Fig. 4a**). We report the distribution of the *d* values for the differences between maximum $\log_{10} P_{merge}$ and $\log_{10} P_{haplotype}$ values, where, for QTLs where candidate variants exist, $d > 0$. When simulated QTLs arose from single causal variants, merge analysis did indeed identify candidate variants at almost all QTLs placed in random regions of the genome as well as at QTLs simulated in the same locations as the detected QTLs.

We also considered the performance of the method at QTLs where it was highly probable that a single variant was the causal variant, namely at *cis*-acting expression QTLs (eQTLs)[22,23]. We tested 1,398 eQTLs detected in the hippocampus of heterogeneous stock mice[24], finding that the merge analysis identified variants with *P* values that exceeded those of the haplotype-based test at 97% of QTLs (**Fig. 4b**). Notably, when we carried out the same analysis on *trans* eQTLs, the distribution of *d* values was similar to that seen for the rat phenotypic QTLs (**Fig. 4b**). This difference between *cis* and *trans* eQTLs was true across all log *P* values, indicating that the difference is not due to lower power to detect *trans* eQTLs.

Because mapping QTLs using haplotype analysis might bias results toward loci without candidates (a winner's curse is likely to operate), we used merge analysis to map QTLs across the genome. The two methods did not identify the same QTLs (152 were unique to the merge method), but the merge method identified 16% fewer QTLs than the haplotype method. Notably, only 9% of the merge-identified QTLs had no candidate variants (**Supplementary Fig. 3**). Consequently, we conclude that haplotype mapping overestimates the number of QTLs without a candidate variant, whereas merge analysis underestimates the number of these QTLs. Therefore, our best estimate of the proportion of QTLs without candidate variants is obtained from combining both methods. In the set of QTLs identified by either merge or haplotype mapping, we found that 44% of QTLs could not be explained by single causal variants (compared to 62% when only the haplotype-based QTLs were considered). Thus, although a winner's curse does operate in favor of the haplotype analysis, it cannot account for all QTLs without a candidate variant.

The third explanation was that the merge analysis underestimates statistical significance. We compared the performance of the merge analysis with that of single-marker association analysis at genotyped SNPs. Across all phenotypes, r^2 correlation between log *P* values was 0.9; agreement was strongest for the most highly associated SNPs. This result indicates that merge analysis performs as well as SNP analysis.

Finally, we investigated the extent to which multiple variants at QTLs would account for our findings. We investigated the consequences of a variety of complex QTL architectures by simulation (**Fig. 4a**). Simulating multiple causal variants on different haplotypes reduced the frequency at which any single variant exceeded the maximum haplotype log *P* value, although this simulated complexity was still insufficient to mimic the observed frequency of QTLs without causal variants (**Fig. 4a**). Simulating irreducible haplotypic effects arising from the reconstructed haplotype mosaics in the heterogeneous stock (rather than from a selection of sequence variants) also led to fewer QTLs with candidate variants (**Fig. 4a**), although, again, the simulated proportion of QTLs without variants did not match that observed with the real QTL set. Our simulations suggest that the presence of multiple causal variants at a locus accounts in part for the inability to identify candidate causal variants.

Table 3 Summary of genes identified at QTLs and potential functional variants

Measure	Chr.	QTL location (Mb)	Gene	Gene description	Only gene with candidate variants in QTL	Amino acid change with potential effect	Location of the residue, potential effect
Mean response latency	2	80.23–84.83	*Ctnnd2*	Catenin δ2	+	None	–
Femur neck width	1	156.27–160.9	*Fchsd2*	FCH and double SH3 domains protein 2	+	None	–
Distal femur total density	2	152.74–157.22	*Kcnab1*	Voltage-gated potassium channel subunit β1	+	None	–
Femoral neck total density	5	4.03–8.22	*Eya1*	Eyes absent homolog 1	+	None	–
Femur midshaft cortical density	6	38.24–41.52	*Lpin1*	Phosphatidate phosphatase LPIN1	+	None	–
Femur midshaft total area	2	43.96–48.57	*Ndufs4*	NADH dehydrogenase (ubiquinone) iron-sulfur protein 4, mitochondrial	+	None	–
Femur work to failure	8	21.57–26.17	*Dpy19l1*	Protein dpy-19 homolog 1	+	None	–
Lumbar trabecular area	20	21.1–25.75	*F1LW02_RAT*	Uncharacterized protein	+	None	–
Heart weight	1	202.15–206.63	*Shank2*	SH3 and multiple ankyrin repeat domains protein 2	+	None	–
Area under glycemia curve over baseline	2	80.5–85.11	*Ctnnd2*	Catenin δ2	+	None	–
Hemoglobin concentration	12	1.62–5.77	*Insr*	Insulin receptor subunit α, insulin receptor subunit β	+	None	–
Mean platelet mass	1	193.98–197.88	*Dock1*	Dedicator of cytokinesis protein 1	+	None	–
Mean platelet mass	9	52.53–88.11	*ErbB4*	Receptor tyrosine protein kinase erbB–4ERBB4 intracellular domain	+	None	–
Platelet clumps	8	100.57–104.81	*Clstn2*	Calsyntenin-2	+	None	–
Platelet count	11	14.47–18.54	*Hspa8*	Heat shock 70-kDa protein 8	+	None	–
Absolute $CD25^+CD4^+$ cells	19	50.71–54.96	*Galnt2*	Polypeptide N-acetylgalactosaminyltransferase 2	+	None	–
Absolute $CD8^+$ T cells	20	1.00–8.90	*RT1-Db2*	RT1 class II, locus Db2	+	None	–
Proportion of B cells in white blood cells	10	27.1–31.59	*D3ZTU5_RAT*	Uncharacterized protein	+	None	–
Proportion of B cells in white blood cells	20	1.00–2.66	*Olr1687*	Olfactory receptor Olr1687	+	None	–
Proportion of $CD4^+$ cells expressing CD45RC	13	36.86–62.54	*Ptprc*	Receptor-type tyrosine protein phosphatase C	+	None	–
Proportion of $CD4^+$ cells in T cells	20	14.83–19.43	*RGD1559903*	Uncharacterized protein	+	None	–
Proportion of $CD8^+$ cells expressing CD45RC	13	50.49–55.97	*Ptprc*	Receptor-type tyrosine protein phosphatase C	+	None	–
Proportion of $CD8^+$ cells with high expression of CD25	19	52.29–56.8	*Sipa1l2*	Signal-induced proliferation-associated 1–like protein 2	+	None	–
Lowest weight	3	121.45–126.25	*Pak7*	Serine/threonine protein kinase PAK 7	+	None	–
Weight loss compared to day 0	2	169.79–174.4	*Fam198b*	Protein FAM198B	+	None	–
Serum alkaline phosphatase	3	18.49–23.11	*Lrp1b*	Low-density lipoprotein–related protein 1B (deleted in tumors)	+	None	–
Serum chloride concentration	9	32.72–36.5	*Uggt1*	UDP-glucose:glycoprotein glucosyltransferase 1	+	None	–
Serum triglycerides	4	74.8–79.28	*Dfna5*	Deafness, autosomal dominant 5	+	None	–
Weight loss compared to day 0	20	2.48–7.07	*RT1-Da*	RT1 class II histocompatibility antigen Da chain	–	p.Thr182Ala	Surface exposed, disturbed intermolecular interactions
Weight loss compared to day 0	20	2.48–7.07	*RT1-Da*	RT1 class II histocompatibility antigen Da chain	–	p.Thr182Met	Surface exposed, disturbed intermolecular interactions
Weight loss compared to day 0	20	2.48–7.07	*RT1-Bb*	RT1 class II histocompatibility antigen, B-1β chain	–	p.His200Arg	Surface exposed, disturbed intermolecular interactions
Weight loss compared to day 0	20	2.48–7.07	*RT1-Bb*	RT1 class II histocompatibility antigen, B-1β chain	–	p.Thr165Met	Surface exposed, disturbed intermolecular interactions
Weight loss compared to day 0	20	2.48–7.07	*RT1-Bb*	RT1 class II histocompatibility antigen, B-1β chain	–	p.Gln162Arg	Surface exposed, disturbed intermolecular interactions

(continued)

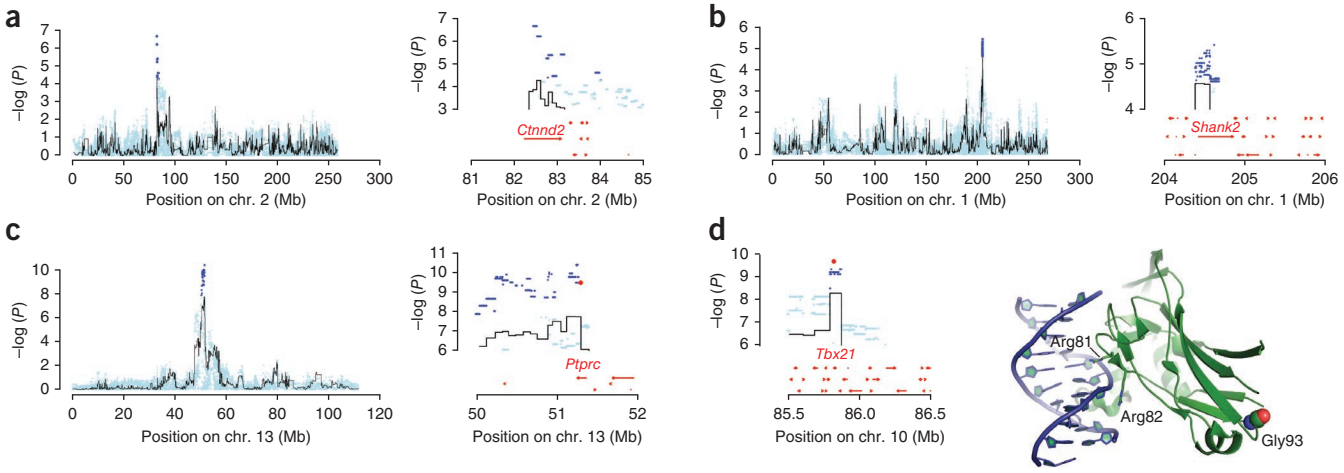
Table 3 Continued

Measure	Chr.	QTL location (Mb)	Gene	Gene description	Only gene with candidate variants in QTL	Amino acid change with potential effect	Location of the residue, potential effect
Expression on RT1B on B cells	17	26.63–27.55	*Tbc1d7*	TBC1-domain family member 7	–	p.Ser116Leu	Surface exposed, disturbed intermolecular interactions
Proportion of B cells in white blood cells	1	182.36–186.67	*Itgal*	Integrin αL	–	p.Asn890Ser	Abolished glycosylation
Proportion of CD4+ cells with high expression of CD25	10	84.27–87.32	*Tbx21*	T-box transcription factor TBX21	–	p.Gly175Arg	Surface exposed, additional interactions with DNA
Ratio of T cells to B cells	1	183.58–187.41	*Rabep2*	Rab GTPase–binding effector protein 2	–	p.Ile336Thr	Partially buried, disturbed oligomerization
Ratio of T cells to B cells	1	183.58–187.41	*Itgal*	Integrin αL	–	p.Leu806Ser	Surface exposed, disturbed intermolecular interactions
Mean corpuscular red blood cell volume	19	53.11–55.80	*Abcb10*	ATP-binding cassette, sub-family B (MDR/TAP), member 10	–	p.Thr233Met	Transport channel exposed, altered transport
Platelet count	12	1.00–7.47	*Rfc3*	Replication factor C (Activator 1)	–	p.Pro173Ala	Surface exposed, alteration of the α helix
Proportion of monocytes in white blood cells	1	250.37–254.00	*Pdcd11*	Protein RRP5 homolog	–	p.Glu160Gly	Surface exposed

Shown are the phenotype measured, the chromosome (chr.), the start and stop coordinates of the QTL, gene symbol and description, whether the gene is the only one at a QTL with candidate variants, whether a variant alters an amino acid and, if so, the residue changed and the potential consequences.

Concordance between species

It is often assumed that the genetic loci underlying a phenotype in one species are homologous to those underlying the same phenotype in another and that natural variation within these loci will map to the same genes[25–27]. However, there have been no genome-wide tests of this hypothesis for natural variation. Our data allowed us to examine whether



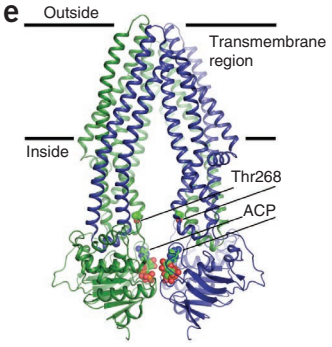


Figure 3 Merge analysis to identify causative genes and sequence variants. (**a**–**c**) Analysis was performed for phenotypes of anxiety (**a**), heart weight (**b**) and the proportion of CD4+ T cells with high expression of CD25 (**c**). Left, whole-chromosome scans for each phenotype; the black lines represent the haplotype-based analysis, and the blue data points represent the results of merge analysis testing for association with all sequence variants identified in the progenitor strains. Right, enlargement of the highest peak showing the location of candidate variants and genes. Candidate variants are those whose significance in merge analysis exceeds that of the haplotype analysis (dark-blue data points above the highest value of the black line). Genes are shown by red arrows. (**d**) Shown are candidate variants on chromosome 10 for the proportion of CD4+ cells with high expression of CD25. The variant with the highest significance lies in the Tbx21 protein. The crystal structure of human TBX5-DNA complex (Protein Data Bank (PDB) 2X6V) maps the location of the rat Tbx21 p.Gly175Arg alteration to the DNA-binding domain. The structure of TBX5 (green) complexed with DNA (blue) is shown in ribbon representation. Gly93 is shown as spheres (green, carbon; red, oxygen; blue, nitrogen). Gly93 and the corresponding Gly175 residue in rat are conserved. The side chains of two arginine residues that mediate interactions with DNA are shown as sticks. (**e**) Shown is a candidate variant encoded in the *Abcb10* gene on chromosome 19 for a locus influencing mean red blood cell volume. The structure of homodimeric human ABCB10 (PDB 4AYT) is shown in ribbon representation, with the monomers colored blue and green. Two ATP analogs (ACP) and the side chains of Thr268 are shown as spheres (green, carbon; red, oxygen; blue, nitrogen; orange, phosphorus). Thr268 in the human protein corresponds to the conserved Thr233 residue in the rat protein. The rat Abcb10 Thr286 alteration lies in the central cavity of the translocation pathway. Amino acid sequence identity of rat and human ABCB10 proteins is 84% (587 aligned residues). Black lines delineate the transmembrane region.

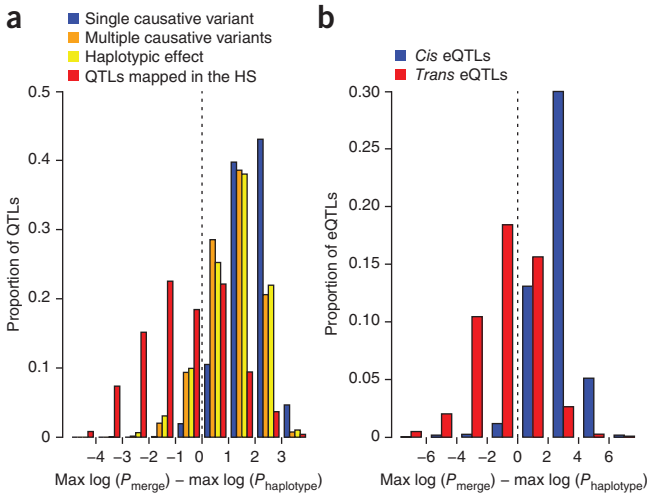


Figure 4 Simulation of causal variants. (**a**) Plotted are the differences between the maximum negative log *P* values for association of imputed variants and the maximum haplotype-based log *P* values for the rat simulated and real QTLs. In cases where there is a single causal variant at a QTL, the log *P* values of some imputed variants will exceed the significance values from the haplotype analysis, such that the mean of the distribution of the differences between these log *P* values will be greater than zero (blue histogram). The distribution observed for the phenotypic QTLs (red histogram) has a mean less than zero. The results of simulating haplotypic effects are shown in yellow, and the consequence of simulating multiple causative variants are shown in orange. (**b**) Plotted is a set of 1,386 *cis*-acting and 7,464 *trans*-acting eQTLs mapped in a mouse heterogeneous stock. The distribution of the differences in log *P* values for the *cis* eQTLs resembles that seen when simulating single causative variants. The distribution for the *trans* eQTLs is most similar to that for the phenotypic QTLs.

genes and QTLs identified in the NIH-HS overlapped those found to underlie the same phenotypes in a mouse heterogeneous stock[10].

In total, 38 measures were common to both studies and were mapped using the same mixed-model method. Only one measure, the ratio of $CD4^+$ to $CD8^+$ T cells, showed overlap (using an FDR of 10% and looking in the 90% QTL confidence interval), but this overlap was not significant (empirical *P* value of 0.1). We repeated the analysis using QTLs called at a lower significance threshold (20th percentile of the extreme value distribution for each measure) and expanding the width of each QTL to 8 Mb. Overlaps for eight phenotypes, only two of which were significant at an empirical *P* value of 0.05 (serum urea concentration and the ratio of $CD4^+$ to $CD8^+$ T cells), were found (**Table 4**). Overall, genetic variants in orthologous genes rarely contributed to the same phenotype in the two populations.

To test whether QTL overlap existed within similar pathways, for each of the 38 measures we asked whether the same KEGG pathways were enriched for QTL-associated genes in both mouse and rat heterogeneous stocks[28]. For only one measure, the proportion of B cells in the total white blood cell population, were the same pathways enriched in both heterogeneous stocks (corrected *P* value < 0.05). Even at a more relaxed significance threshold of 0.05 (not corrected for multiple testing), only three measures showed the same KEGG pathways enriched in both heterogeneous stocks.

DISCUSSION

Using 1,407 outbred rats, we have mapped 122 phenotypes and identified 355 QTLs at high resolution. We have shown how combining sequence with high-resolution mapping data can lead to the immediate identification of candidate genes and, in some cases, to the identification of candidate causal variants at many QTLs. We highlight two examples here.

The locus on chromosome 10 regulating the frequency of $CD25^+CD4^+$ T cells and the frequencies of $CD4^+$ and $CD8^+$ T cells has previously been shown to control $CD4^+$ and $CD8^+$ T cell frequencies in a cross between ACI and F344 rats[29], both represented in the NIH-HS progenitors. The amino acid substitution at position 175 (p.Gly175Arg) of the Tbx21 protein is a very strong causal candidate in this QTL because the affected protein domain is important for DNA interactions. *Tbx21* has been implicated in the genetic control of regulatory T cells[30], a subset of T cells with high surface expression of CD25, and might indirectly regulate the frequencies of $CD4^+$ and $CD8^+$ T cells through the transcriptional repressor *Sin3a*[31,32].

We implicated *Abcb10* in red blood cell differentiation. Evidence from mouse knockouts indicates that this gene is essential for erythropoiesis[18,19,33]. The p.Thr233Met alteration in Abcb10 positions a larger, bulkier residue in a protein region that is tightly packed in the open-outward conformation of ABC transporters, potentially interfering with the conformational changes that are essential for transport of the substrate.

Two noteworthy features of the genetic architecture of complex traits in the rat emerge from this study: (i) the contrast with findings from human GWAS and (ii) the fact that about half of QTLs cannot be attributed to a single causal variant.

Rat and mouse heterogeneous stock experiments differ from human GWAS in two ways. In rodent GWAS, far fewer subjects are required to detect a significant effect, and fewer loci of larger effect explain more of the variance. In rats, the median proportion of heritability explained by joint QTLs is 39.1% (mean of 42.3%), and, in mice, the median proportion is 32.2% (mean of 42.0%). In humans, the mean proportion of heritability explained is often less than 10%.

One explanation for these differences is the markedly different allele frequencies in these species. Human populations are characterized by a preponderance of rare alleles (with MAF of less than 1%), whereas heterogeneous stock populations have a relatively uniform distribution of MAFs (**Fig. 1c**). However, it is

Table 4 Syntenic QTLs mapped in the rat and mouse heterogeneous stocks for the same measure

Phenotype	Rat chr.	Rat QTL (Mb)	Mouse chr.	Mouse QTL (Mb)	*P* value of overlap
$CD4^+/CD8^+$ cell ratio	2	80.51–88.51	8	71.7–79.7	0.009
$CD4^+/CD8^+$ cell ratio	20	1.00–21.13	17	29.77–37.77	–
$CD4^+/CD8^+$ cell ratio	9	0.16–8.16	17	50.77–58.77	–
Serum urea concentration	3	42.22–50.22	2	62.25–70.25	0.017
Serum calcium concentration	12	32.82–40.82	5	122.62–130.62	0.082
White blood cells	10	57.69–71.77	11	64.92–72.92	0.115
White blood cells	20	47.41–55.24	10	40.74–48.74	–
T cell/B cell ratio	13	76.73–84.73	1	169.63–177.63	0.149
T cell/B cell ratio	20	37.59–45.59	10	36.25–48.68	–
Serum chloride concentration	9	30.61–38.61	13	2.91–15.19	0.22
Monocytes	20	0.17–8.17	17	21.00–29.00	0.301
Serum total cholesterol	4	17.09–25.09	5	12.52–20.52	0.598

Shown are the 8 measures (out of 38) that have syntenic QTLs, the QTL coordinates (chromosome, start and stop) and the *P* value of the overlap (one *P* value per measure).

important to realize that mice and rats differ in the degree of segregating variation (in the rat NIH-HS, there are 7.2 million SNPs compared to 10.2 million in the mouse heterogeneous stock). In rats, there are 2.8 million SNPs per heterogeneous stock strain, whereas the corresponding number in the mouse heterogeneous stock is 4.4 million. In other words, total sequence variation in itself is not a critical determinant of the explanatory power of the QTLs. Furthermore, the heritabilities of homologous phenotypes in the rat NIH-HS and in heterogeneous stock mice are highly correlated ($r^2 = 0.6$; $P = 0.0002$) (**Supplementary Fig. 4**), implying that the greater sequence variation in mice does not result in increased heritability.

The inability to detect a single candidate variant at half of rat QTLs was unexpected. We showed that, although reliance on haplotype-based mapping can underestimate the number of QTLs without candidate variants, after taking this bias into account (by detecting QTLs with both merge and haplotype analysis), there is still a large fraction (44%) of QTLs without candidate variants. The contrast between the 44% figure and the 97% that emerged from an analysis of variants at *cis* eQTLs is striking. It is also notable that the findings from *trans* eQTLs are very similar to those obtained in the analysis of rat phenotypes (**Fig. 4**), suggesting that *cis* eQTLs are atypical. Our simulations indicate but have not proven that, in part, multiple causal variants at single QTLs are to blame. At present, we can only conclude that single causal variants are not always responsible for the genetic signal at a QTL. Whether the lack of single causal variants at many loci is a general feature of loci influencing complex traits remains to be determined.

One simple interpretation of human GWAS is that each locus represents the presence of a single, relatively common functional variant. Our results indicate that more complex models are required. Such alternative hypotheses exist, in which, for example, multiple alleles of varying frequency at the same or closely linked loci contribute to the association signal. Identifying the correct model of genetic action is critical for finding causative variants, as incorrect assumptions about the number and mode of action of genetic variants reduce power and can lead to false positive results[34]. The extent and nature of sequence diversity may be partly responsible for the complex way that sequence variation acts at a QTL.

It was sometimes hoped that loci found in the rat could be typed and identified in humans, thus providing a cost-efficient way to find medically relevant genes. We observe some examples where the same loci act similarly in different species, the most notable example being for variation in the ratio of $CD4^+$ to $CD8^+$ T cells: the locus lies within the MHC in rats, humans[35] and mice[36], and its molecular nature in mice has been identified as a deletion in the promoter of the MHC class II gene *RT1-Da*[36]. However, formal tests for overlap between rats and mice at the gene or pathway level yielded little that was statistically significant. Because the amount of sequence variation segregating within the two heterogeneous stock populations is relatively limited, the inability to detect shared loci may result from sampling. Also, the relatively small number of genes found for each phenotype reduces our power to detect pathways. We suspect that it is currently not possible to accurately assess overlap between the two species.

This study strengthens the rat's role as a model organism in physiology and disease. Our mapping and sequencing data provide an important resource for addressing many biomedical questions.

URLs. Mapping data (see **Supplementary Note** for directions on how to explore the sequence data at each QTL), http://mus.well.ox.ac.uk/gscandb/rat; variant calls and inaccessible regions, http://www.hubrecht.eu/research/cuppen/suppl_data.html; DWAC-Seq v. 0.56, https://github.com/Vityay/DWAC-Seq; 1-2-3-SV v. 1.0, https://github.com/Vityay/1-2-3-SV; Ensembl release 66, http://www.ensembl.org/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Sequence data for the eight heterogeneous stock founders are available from the European Bioinformatics Institute (EBI) Short Read Archive (SRA) under accession ERP001923. The LE/Stm BAC sequences are available in the NCBI Trace Archive (accessions FO181540, FO181541, FO117626, FO181542, FO117624, FO181543, FO117625, FO117627, FO117628, FO117629, FO117630, FO117631 and FO117632).

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We are grateful to T. Serikawa (Kyoto University) for the LE/Stm BAC clones. The Human Genome Sequencing Center sequence production teams at the Baylor College of Medicine produced the Sanger sequencing data for the eight sequenced strains used to define the RATDIV SNP genotyping array (see ref. 8 for a list of Baylor College of Medicine HGSC sequencing contributors). We thank E. Redei for providing the NIH-HS rat colony. The funders we would like to acknowledge are as follows: the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement HEALTH-F4-2010-241504 (EURATRANS); The Wellcome Trust (090532/Z/09/Z, 083573/Z/07/Z, 089269/Z/09/Z); The Swedish Research Council (grant K2008-66X-20776-01-4); the Harald and Greta Jeanssons Foundation; The Swedish Association for Persons with Neurological Disabilities; the Åke Wibergs Foundation; the Åke Löwnertz Foundation; Karolinska Institutet funds; the European Union's Sixth Framework Programme EURATools (grant LSHG-CT-2005-019015); the Bibbi and Nils Jensens Foundation; the Söderbergs Foundation; and the Knut and Alice Wallenbergs Foundation. We also thank the Ministerio de Ciencia e Innovación (reference PSI2009-10532 and the Formación de Personal Investigador fellowship to C.M.-C.); the Fundació La Maratò TV3 (reference 092630); the Direcció General de la Recerca (reference 2009SGR-0051); and the British Heart Foundation (BHFRG/07/005/23633). T.J.A. and S.S.A. acknowledge funding from the Imperial BHF Centre of Research Excellence. M. Johannesson acknowledges support from Prof. Nanna Svartz Foundation, The Swedish Rheumatism Association and The King Gustaf V 80th Anniversary Foundation. D. Gauguier acknowledges support from the Institute of Cardiometabolism and Nutrition (ICAN; ANR-10-IAHU-05). T.M. and E.Y.J. acknowledge support from Cancer Research UK (A10976) and the UK Medical Research Council (G9900061). T.F., D.L.K. and I.A. acknowledge support from the U.S. National Institutes of Health (R01 AR047822).

AUTHOR CONTRIBUTIONS

The writing group included A. Baud, R. Hermsen, V.G., D. Gauguier, P.S., T.O., R. Holmdahl, D. Graham, M.W.M., T.F., A.F.-T., N. Hubner, E.C., R.M. and J.F. The phenotyping group included S.C., D. Gauguier, P.S., M.D., J.O., A.D.B., A.G., N.A., A.O.G.-C., M. Jagodic, T.O., M. Johannesson, J.T., U.N., R. Holmdahl, D. Graham, E.B., N. Huynh, W.H.M., M.W.M., A.F.D., D.L.K., T.F., I.A., S.F., N. Hubner, M.O.-P., E.M.-M., R.L.-A., T.C., G.B., E.V.-C., C.M.-C., S.D.-M., A.T. and A.F.-T. The high-density genotyping array design and analysis group included O.H., D.Z., K.S., G.P., A. Bauerfeind, M.-T.B., M.H., Y.-A.L., C.R., H.S., D.A.W., K.C.W., D.M.M., R.A.G., M.L. and N. Hubner. The sequencing group included R. Hermsen, O.H., N.L., G.P., P.T., F.P.R., E.d.B., H.H., S.S.A., T.J.A., P.F., D.J.A., T.K., K.S., N. Hubner, V.G. and E.C. The protein structure group included T.M. and E.Y.J. QTL data analysis was performed by A. Baud, J.F., D.E. and R.M. The project was coordinated by A. Baud, R.L.-A., A.F.D., N. Hubner, M. Johannesson, R. Holmdahl, T.O., D. Gauguier, A.F.-T., R.M., E.C. and J.F.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Donnelly, P. Progress and challenges in genome-wide association studies in humans. *Nature* **456**, 728–731 (2008).
2. Flint, J. & Mackay, T.F. Genetic architecture of quantitative traits in mice, flies, and humans. *Genome Res.* **19**, 723–733 (2009).
3. Hansen, C. & Spuhler, K. Development of the National Institutes of Health genetically heterogeneous rat stock. *Alcohol. Clin. Exp. Res.* **8**, 477–479 (1984).

4. Keane, T.M. *et al.* Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477**, 289–294 (2011).
5. Talbot, C.J. *et al.* High-resolution mapping of quantitative trait loci in outbred mice. *Nat. Genet.* **21**, 305–308 (1999).
6. Yalcin, B., Flint, J. & Mott, R. Using progenitor strain information to identify quantitative trait nucleotides in outbred mice. *Genetics* **171**, 673–681 (2005).
7. Mayosi, B.M., Keavney, B., Watkins, H. & Farrall, M. Measured haplotype analysis of the aldosterone synthase gene and heart size. *Eur. J. Hum. Genet.* **11**, 395–401 (2003).
8. Gibbs, R.A. *et al.* Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493–521 (2004).
9. Serikawa, T. *et al.* National BioResource Project–Rat and related activities. *Exp. Anim.* **58**, 333–341 (2009).
10. Valdar, W. *et al.* Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nat. Genet.* **38**, 879–887 (2006).
11. Johannesson, M. *et al.* A resource for the simultaneous high-resolution mapping of multiple quantitative trait loci in rats: the NIH heterogeneous stock. *Genome Res.* **19**, 150–158 (2009).
12. Kang, H.M. *et al.* Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709–1723 (2008).
13. Valdar, W., Holmes, C.C., Mott, R. & Flint, J. Mapping in structured populations by resample model averaging. *Genetics* **182**, 1263–1277 (2009).
14. Mott, R., Talbot, C.J., Turri, M.G., Collins, A.C. & Flint, J. A method for fine mapping quantitative trait loci in outbred animal stocks. *Proc. Natl. Acad. Sci. USA* **97**, 12649–12654 (2000).
15. Israely, I. *et al.* Deletion of the neuron-specific protein δ-catenin leads to severe cognitive and synaptic dysfunction. *Curr. Biol.* **14**, 1657–1663 (2004).
16. Berkel, S. *et al.* Mutations in the *SHANK2* synaptic scaffolding gene in autism spectrum disorder and mental retardation. *Nat. Genet.* **42**, 489–491 (2010).
17. Berman, H.M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
18. Shirihai, O.S., Gregory, T., Yu, C., Orkin, S.H. & Weiss, M.J. ABC-me: a novel mitochondrial transporter induced by GATA-1 during erythroid differentiation. *EMBO J.* **19**, 2492–2502 (2000).
19. Hyde, B.B. *et al.* The mitochondrial transporter ABC-me (ABCB10), a downstream target of GATA-1, is essential for erythropoiesis *in vivo*. *Cell Death Differ.* **19**, 1117–1126 (2012).
20. Wallström, E.O.T. Rat models of experimental autoimmune encephalomyelitis. in *Sourcebook of Models for Biomedical Research* (ed. Conn, P.J.) 547–556 (Humana Press, Ottawa, 2007).
21. Sawcer, S. *et al.* Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476**, 214–219 (2011).
22. Degner, J.F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–394 (2012).
23. Veyrieras, J.B. *et al.* High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet.* **4**, e1000214 (2008).
24. Huang, G.J. *et al.* High resolution mapping of expression QTLs in heterogeneous stock mice in multiple tissues. *Genome Res.* **19**, 1133–1140 (2009).
25. Jagodic, M. *et al.* A role for VAV1 in experimental autoimmune encephalomyelitis and multiple sclerosis. *Sci. Transl. Med.* **1**, 10ra21 (2009).
26. Swanberg, M. *et al.* *MHC2TA* is associated with differential MHC molecule expression and susceptibility to rheumatoid arthritis, multiple sclerosis and myocardial infarction. *Nat. Genet.* **37**, 486–494 (2005).
27. Trynka, G. *et al.* Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat. Genet.* **43**, 1193–1201 (2011).
28. Ogata, H. *et al.* KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **27**, 29–34 (1999).
29. Brenner, M., Laragione, T., Yarlett, N.C. & Gulko, P.S. Genetic regulation of T regulatory, CD4, and CD8 cell numbers by the arthritis severity loci *Cia5a*, *Cia5d*, and the MHC/*Cia1* in the rat. *Mol. Med.* **13**, 277–287 (2007).
30. Koch, M.A. *et al.* The transcription factor T-bet controls regulatory T cell homeostasis and function during type 1 inflammation. *Nat. Immunol.* **10**, 595–602 (2009).
31. Chang, S., Collins, P.L. & Aune, T.M. T-bet dependent removal of Sin3A-histone deacetylase complexes at the *Ifng* locus drives Th1 differentiation. *J. Immunol.* **181**, 8372–8381 (2008).
32. Cowley, S.M. *et al.* The mSin3A chromatin-modifying complex is essential for embryogenesis and T-cell development. *Mol. Cell. Biol.* **25**, 6990–7004 (2005).
33. Liesa, M. *et al.* Mitochondrial transporter ATP binding cassette mitochondrial erythroid is a novel gene required for cardiac recovery after ischemia/reperfusion. *Circulation* **124**, 806–813 (2011).
34. Atwell, S. *et al.* Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* **465**, 627–631 (2010).
35. Ferreira, M.A. *et al.* Quantitative trait loci for CD4:CD8 lymphocyte ratio are associated with risk of type 1 diabetes and HIV-1 immune control. *Am. J. Hum. Genet.* **86**, 88–92 (2010).
36. Yalcin, B. *et al.* Commercially available outbred mice for genome-wide association studies. *PLoS Genet.* **6**, pii: e1001085 (2010).

Amelie Baud[1], Roel Hermsen[2], Victor Guryev[2,3], Pernilla Stridh[4], Delyth Graham[5], Martin W McBride[5], Tatiana Foroud[6], Sophie Calderari[7], Margarita Diez[4], Johan Ockinger[4], Amennai D Beyeen[4], Alan Gillett[4], Nada Abdelmagid[4], Andre Ortlieb Guerreiro-Cacais[4], Maja Jagodic[4], Jonatan Tuncel[8], Ulrika Norin[8], Elisabeth Beattie[5], Ngan Huynh[5], William H Miller[5], Daniel L Koller[6], Imranul Alam[9], Samreen Falak[10], Mary Osborne-Pellegrin[11], Esther Martinez-Membrives[12], Toni Canete[12], Gloria Blazquez[12], Elia Vicens-Costa[12], Carme Mont-Cardona[12], Sira Diaz-Moran[12], Adolf Tobena[12], Oliver Hummel[10], Diana Zelenika[13], Kathrin Saar[10], Giannino Patone[10], Anja Bauerfeind[10], Marie-Therese Bihoreau[13], Matthias Heinig[10,14], Young-Ae Lee[10,15], Carola Rintisch[10], Herbert Schulz[10], David A Wheeler[16], Kim C Worley[16], Donna M Muzny[16], Richard A Gibbs[16], Mark Lathrop[13], Nico Lansu[2], Pim Toonen[2], Frans Paul Ruzius[2], Ewart de Bruijn[2], Heidi Hauser[17], David J Adams[17], Thomas Keane[17], Santosh S Atanur[18], Tim J Aitman[18], Paul Flicek[19], Tomas Malinauskas[20], E Yvonne Jones[20], Diana Ekman[8], Regina Lopez-Aumatell[1,12], Anna F Dominiczak[5], Martina Johannesson[8], Rikard Holmdahl[8], Tomas Olsson[4], Dominique Gauguier[7], Norbert Hubner[10,21], Alberto Fernandez-Teruel[12], Edwin Cuppen[2], Richard Mott[1] & Jonathan Flint[1]

[1]Wellcome Trust Centre for Human Genetics, Oxford, UK. [2]Hubrecht Institute, Koninklijke Nederlandse Akademie van Wetenschappen and University Medical Center Utrecht, Utrecht, The Netherlands. [3]European Research Institute for the Biology of Ageing, Rijksuniversiteit Groningen, Universitair Medisch Centrum Groningen, Groningen, The Netherlands. [4]Neuroimmunology Unit, Department of Clinical Neuroscience, Centre for Molecular Medicine, Karolinska Institutet, Stockholm, Sweden. [5]British Heart Foundation (BHF) Glasgow Cardiovascular Research Centre, Institute of Cardiovascular & Medical Sciences, Glasgow University, Glasgow, UK. [6]Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, Indiana, USA. [7]Institut National de la Santé et de la Recherche Médicale (INSERM) Unité Mixte de Recherche Scientifique (UMRS) 872, Cordeliers Research Centre, Paris, France. [8]Department of Medical Biochemistry and Biophysics, Division of Medical Inflammation Research, Karolinska Institutet, Stockholm, Sweden. [9]Department of Orthopedic Surgery, Indiana University School of Medicine, Indianapolis, Indiana, USA. [10]Max-Delbruck Center for Molecular Medicine, Berlin, Germany. [11]INSERM U698, Hôpital Bichat, Paris, France. [12]Medical Psychology Unit, Department of Psychiatry & Forensic Medicine, Institute of Neurosciences, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain. [13]Commissariat à l'Energie Atomique, Institut de Génomique, Centre National de Génotypage, Evry, France. [14]Department of Computational Biology, Max Planck Institute for Molecular Genetics, Berlin, Germany. [15]Pediatric Allergology, Experimental and Clinical Research Center, Charité Universitätsmedizin Berlin, Berlin, Germany. [16]Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas, USA. [17]The Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK. [18]Physiological Genomics and Medicine Group, Medical Research Council Clinical Sciences Centre, Faculty of Medicine, Imperial College London, Hammersmith Hospital, London, UK. [19]European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, UK. [20]Division of Structural Biology, Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [21]DZHK (German Centre for Cardiovascular Research), Partner site Berlin, Berlin, Germany. Correspondence should be addressed to A.F.-T. (albert.fernandez.teruel@uab.es), E.C. (e.cuppen@hubrecht.eu), R.M. (rmott@well.ox.ac.uk) or J.F. (jf@well.ox.ac.uk).

ONLINE METHODS

Sequencing of heterogeneous stock founder genomes. *Genome sequencing.* DNA libraries for SOLiD sequencing were generated from genomic DNA from samples of the original rats that were used to create the heterogeneous stock population. Libraries were generated using standard protocols (Life Technologies) and had a median insert size of between 109 and 196 bp. All libraries were sequenced with fragment (50-bp) and paired-end (50+35-bp) runs using SOLiD 4 and SOLiD 5500 sequencers to a depth of at least 22× base coverage for each of the eight heterogeneous stock progenitors and for the LE/Stm strain, which was used to estimate error rates in comparison with hand-finished BAC sequence.

Sequence alignment. Sequence reads were mapped against contigs of the Rnor3.4 rat reference genome assembly (reference strain BN) using Burrows-Wheeler Aligner (BWA) v0.5.9 (ref. 37) with parameters –c –l 25, –k 2 and –n 10. Alignments from different libraries of the same heterogeneous stock progenitor were combined into a single BAM file.

Variant calling. Variant calling was performed independently for each strain. SNPs and short indels (<10 bp) were called using a modified SAMtools[38] pipeline: only unambiguously mapped reads were used. Sites with coverage below 4× or over 2,000× were not used for SNP calling. Nucleotides with base quality of <30 were ignored. Duplicate reads starting at the same position and mapping to the same strand as another read were discarded as probable PCR artifacts. Each of the called alleles had to be supported by at least one read where the variant mapped within the seed part of the read (first 25 bases). Non-reference alleles called with fewer than three reads were set to missing. Variable sites with more than two alleles within one founder were set to missing. The remaining variants were considered to be homozygous non-reference alleles (frequency of the non-reference call of >2/3) or heterozygous alleles (frequency between 1/3 and 2/3); however, we set to missing the small number of heterozygote calls, as these were probably artifacts caused, for example, by unknown duplications. We later attempted to call all the missing genotypes by imputation.

Copy number variants were called using a depth-of-coverage approach implemented in DWAC-Seq v. 0.56 using default parameters. Structural variants were called using discordant-pair mapping implemented in 1-2-3-SV v. 1.0, requiring unambiguous mapping of both paired tags and at least four tag pairs per structural variant. Structural variant calls identified by these tools were merged. Prediction of the functional effect of each variant was performed by the Variant Effect Predictor (VEP 2.1) tool[39].

We defined inaccessible regions of the heterogeneous stock rat genomes in a similar way as for mouse genomes[4]. A base was considered to be accessible if it did not overlap simple, tandem repeats or low-complexity sequence (defined by Dust, source: Ensembl release 66) and was not covered by more than 150 reads and if average mapping quality was at least 40. Nucleotide positions within 15 bp of indels were also considered to be inaccessible for SNP calling.

False positive and false negative rates. Thirteen BACs from the LE/Stm strain were sequenced using capillary methods, assembled and manually edited, producing a total of 2.1 Mb of finished sequence. BAC sequences were aligned using BLAT[40]. For each BAC, a single contiguous alignment was obtained, which was used to extract single-base changes (SNPs), short indels (1–10 bp) and structural variants (100 bp and greater). False positive and false negative rates were estimated using the 1.9 Mb of genome sequence that was syntenic between BACs and the genome assembly, excluding low-quality BAC sequence (as defined by the BAC finishing team) and inaccessible regions. False positive and false negative rates within this 1.9 Mb were estimated on the basis of the discordance between our allele calls and those in the BACs.

Low false positive rates were independently confirmed by analysis of a randomly selected subset of 96 SNPs and 96 indels using PCR-based resequencing. Oligonucleotide primers were selected to amplify 300-bp fragments around the candidate polymorphism. When amplification was successful (SNPs, 84; indels, 80), amplicons were sequenced on an Applied Biosystems ABI 3730XL sequencer using BigDye Terminator technology, and sequences were manually analyzed with PolyPhred software.

For copy number variants and structural variants, 184 variants were selected, and PCR primers were designed in such a way that the presence or absence of a PCR product (depending on the variation type) could confirm the presence of the variation. After PCR, samples were run on agarose gels and analyzed manually. Of the 184 variants, 93 gave a PCR product. Of these 93, a group of 39 variants that were predicted structural variants in the NIH-HS founders were also confirmed by PCR in BN/NHsdMcwi, indicating that these are probably assembly errors in the current reference genome (Rn3.4). Of the remaining 54 variants, 53 gave a banding pattern according to our expectation, and, in one case, the predicted variation type was not correct.

Sequence divergence. Genotypes and genome accessibility data for heterogeneous stock rats (this study) and heterogeneous stock mice[4] were used to characterize patterns of nucleotide diversity in these two panels. We partitioned each genome into non-overlapping windows such that each window contained 100 kb of accessible sequence (defined relative to the rat BN strain or mouse C57BL6 strain). The number of sequence differences per window was calculated for all windows and for all possible pairs of strains.

Low-diversity regions. We found that the spatial distribution of pairwise differences in the rat progenitors was bimodal, with modes at 0 and 150 SNPs per 100-kb window (**Fig. 1b**). On the basis of this distribution, we defined a region of low nucleotide diversity between two strains as consecutive windows with nucleotide diversity below 13 SNPs per 100-kb window.

Phenotyping. *Animals.* The rat NIH-HS originates from a colony established in the 1980s at the NIH[3]. Since its creation, the stock has been bred using a rotational outbreeding regimen to minimize the extent of inbreeding, drift and fixation.

Phenotyping. A full description of the phenotyping protocol is given in the **Supplementary Note**.

All procedures were carried out in accordance with Spanish legislation on the Protection of Animals Used for Experimental and Other Scientific Purposes and the European Community's Council Directive (86/609/EEC) on this subject. The experimental protocol was approved by the Autonomous University of Barcelona ethics committee (permit CEEAH 697).

Quality control, covariate analysis and normalization of phenotypes. Phenotype data were uploaded to a database (Integrated Genotyping System)[41] in batches over the 3 years of data collection. All relevant covariates were evaluated for their effect on each measure. The final set of covariates and transformations applied to each phenotype, as well as the number of data points for each measure, is given in **Supplementary Table 1**.

Genotyping. The RATDIV array was developed as a general SNP genotyping array, applicable both to the rat heterogeneous stock project and other populations of laboratory rats. Full descriptions of the development of the rat array and of the selection of the 265,551 SNPs used in this study are given in the **Supplementary Note**.

LD analysis. LD between SNPs in the rat and mouse heterogeneous stocks was calculated using PLINK[42] from the genotypes called for the 261,684 autosomal rat SNPs and 12,226 autosomal mouse SNPs[10]. In the rat heterogeneous stock, eight regions with very high interchromosomal LD were identified and excluded from subsequent analyses (**Supplementary Table 2**). Using the UCSC liftover tool[43], we found that these regions mapped in the new rat reference genome assembly (RGSC 5.0) to the regions with which they were in high LD in the current assembly (Rnor3.4).

QTL mapping. *Reconstruction of heterogeneous stock rat genomes as mosaics of founder haplotypes.* All genetic analysis was performed using R[44]. We used the R HAPPY package[14] to calculate descent probabilities from the 8 heterogeneous stock founders for each animal at each of 265,551 intermarker intervals and then averaged these probabilities over 90-kb windows, such that we eventually worked with 24,196 probability matrices. The density of the 265,000 SNPs was much greater than the density of recombinants in the heterogeneous stock, meaning that averaging did not cause any reduction in mapping resolution (most QTLs mapped to intervals over 1 Mb in length and contained more than 10 90-kb intervals).

Accounting for confounding in the heterogeneous stock. Heterogeneous stock rats with different levels of relatedness were used in this study, including, for example, siblings, half-siblings, cousins, uncles and great-uncles. This unequal genome-wide genetic similarity meant that correlations existed in the heterogeneous stock between distant markers. These long-range correlations (as opposed to short-range correlations due to physical linkage) can be responsible for false positive associations if not accounted for. We used

doi:10.1038/ng.2644

two methods to control for unequal relatedness: Resample Model Averaging (as implemented in BAGPHENOTYPE[13]) for phenotypes with a non-normal distribution and Mixed Models for phenotypes with a normal distribution. Information on the performance of the methods is given in the **Supplementary Note**. Because most of the phenotypes had a normal distribution and the merge analysis was run in the mixed-model framework, we present the mixed models briefly here. These models were implemented in R so that haplotype mapping could be carried out using the descent probabilities from HAPPY[14]. The model used to test for association between the ancestral haplotypes segregating at a locus L and phenotypic variation was

$$y_i = \sum_c \beta_c x_{ic} + \sum_s P_{Li}(s) T_{Ls} + u_i + \varepsilon_i \quad (1)$$

where y_i is the phenotypic value of rat i and β_c is the regression coefficient of covariate c and x_{ic} (the value of the covariate c in rat i). Notably, the covariates include a dummy intercept term. T_{Ls} is the deviation in phenotypic value that results from carrying one copy of a haplotype from strain s at locus L, and $P_{Li}(s)$ is the expected number of haplotypes of type s carried by rat i at locus L output by HAPPY[14]. u_i and ε_i are random effects, with $\mathrm{cov}(u_i, u_j) = \sigma_g^2 K_{i,j}$ and $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = \sigma_e^2 I_{i,j}$ where σ_g^2 and σ_e^2 are estimated in the null model (no locus effect, $T_{Ls} = 0$) using the R package EMMA[12]. K is the genetic covariance matrix and is estimated from the genome-wide genotype data using identity by state (IBS, the proportion of shared alleles between any two animals). The IBS matrix was calculated using the R package EMMA[12]. I is the identity matrix. The total covariance matrix $V = \sigma_g^2 K + \sigma_e^2 I$ can be factorized as $V = A^2$. Writing equation (1) in matrix form, gives

$$y = X\beta + P_L T_L + u + \varepsilon \quad (2)$$

Premultiplying equation (2) by A^{-1} gives a transformed equation

$$(A^{-1}y) = (A^{-1}X)\beta + (A^{-1}P_L)T_L + A^{-1}(u + \varepsilon)$$

in which the variance-covariance structure of the random term $A^{-1}(u + \varepsilon)$ is now proportional to a diagonal matrix and so can be fitted as a standard linear model.

Thresholds and confidence intervals. Calculations of the significance thresholds (when the phenotype was analyzed with mixed models), inclusion probability thresholds (when the phenotype was analyzed by resample model averaging) and confidence intervals are described in the **Supplementary Note**.

Incorporation of sequence into QTL mapping. *Implementation of the merge analysis in the mixed-model framework.* Merge analysis is a form of imputation appropriate to heterogeneous stock–type populations whose genomes are mosaics of known haplotypes. Merge analysis asks two questions at each imputed variant: is the variant associated with the phenotype? (a standard test of association), and is its association as significant as the association in the haplotype-based test in the locality of the variant? We implemented merge analysis[6] in a mixed-model framework by comparing model (2)

$$y = X\beta + P_L T_L + u + \varepsilon$$

and

$$y = X\beta + M_V U_V + u + \varepsilon \quad (3)$$

where V is a sequence variant in interval L and M_V is the merge matrix for the variant, formed by summing those columns of P_L that carry the same allele at V (each column of P_L represents one founder strain). This can be computed efficiently by defining a matrix B_V that encodes the columns to be merged such that $M_V = P_V B_V$. This test is applied at every variable site in the catalog of single-nucleotide variants that segregate between the eight heterogeneous stock founders. From a statistical point of view, there is no difference between two variants with the same strain distribution pattern at a locus; they will give the same merge analysis result.

Because models (2) and (3) are nested, the best possible fit (in terms of variance explained) is obtained with haplotype model (2). If the QTL arises from variation at a single variant V, the fit of merge model (3) for variant V will be as good as the fit of model (2), and its significance will be greater, owing to the fewer number of degrees of freedom (for a diallelic variant, there is 1 degree of freedom for the merge model compared to the 7 degrees of freedom for the haplotype model). The merge model is fitted by multiplying by A^{-1}.

Simulating all possible strain distribution patterns at a QTL. For each QTL lacking variants with a merge log P value exceeding the haplotype log P value, we looked for unobserved causal variants that might not have been sequenced. We simulated candidate variants with every possible SDP (127 possible SDPs for diallelic variants and 1,094 possible SDPs when allowing for 3 alleles). Simulated variants were repeated within each QTL interval.

Simulating different QTL architectures. To investigate the hypothesis that the inability to detect candidate variants by merge analysis reflected complex architecture of the QTLs, we simulated QTLs arising from a single causal variant, QTLs arising from multiple causal variants within the same locus and/or multiple causal variants at linked loci, and QTLs arising from haplotypic effects not reducible to individual variants. In all cases, the phenotypes were simulated from three components: a genetic random effect explaining 20% of phenotypic variation, uncorrelated errors explaining 75% of phenotypic variation and a single QTL explaining 5% of phenotypic variation. When multiple causal variants were simulated, each explained the same proportion of phenotypic variation (5% divided by the number of causal variants). The effect sizes calculated *a posteriori* could be quite different from their target values owing to correlations between the different components of the simulated phenotypes. For the simulations reported in **Figure 4a**, either a single causal variant was simulated or nine causal variants were simulated in three linked loci (with each locus within 2 Mb of the central locus and distant by at least 200 kb from each other locus). Alternatively, the P_L probabilities were used to simulate irreducible QTLs. We analyzed each simulation by merge analysis, and, when log ($P_{\mathrm{haplotype}}$) was between 4 and 6 (to have a similar distribution of log P values to that of the rat QTLs), we calculated d as max log (P_{merge}) – max log ($P_{\mathrm{haplotype}}$). We compared the distributions of d from the different simulation sets to determine the probable genetic architecture of the QTLs.

eQTL mapping and merge analysis in the mouse heterogeneous stock. Hippocampus expression levels in 460 heterogeneous stock mice measured using 12,000 probes on the Illumina Mouse WG-6 v1 BeadArray[24] were mapped to the mouse ancestral haplotypes in the mixed-model framework. QTLs were called in the same way as for the rat QTLs but using a confidence interval of 8 Mb and a significance threshold of 4. *Cis* eQTLs were defined as being within 2 Mb of the beginning of the probe, and *trans* eQTLs were defined as being on a different chromosome than that of the probe or being more than 10 Mb away from it on the same chromosome. Merge analysis was carried out at each eQTL, and the difference between the maximum merge log P value and the maximum haplotype log P value was calculated.

Homology modeling. To assess the potential effects of mutations on protein structure, homology models of target proteins were constructed and analyzed. Amino acid sequences of target proteins were retrieved from the Ensembl or UniProt databases[45] and were analyzed using the HHPred[46] web server to identify structures with similar amino acid sequences in PDB[17] for homology modeling with MODELLER[47]. The potential locations of the mutation-affected side chains (buried or surface exposed) and effects on the structure-function relationship (for example, disturbed hydrophobic core) were evaluated manually in PyMOL (The PyMOL Molecular Graphics System, Version 1.5.0.4).

Genetic architecture. *Heritability.* Heritability was defined as the ratio of the genetic variance component to the sum of the variance components estimated in the null mixed model (covariates but no QTL).

QTL effect sizes and joint effect sizes. Effect sizes were defined as the ratio between the fitted sum of squares and the total sum of squares in a model with covariates and without genetic random component. Joint effect sizes were defined as the ratio between the fitted sum of squares and the total sum of squares in a model without genetic random component, including covariates and all the QTLs called for a given phenotype. Including the genetic random component would result in underestimation of most of the effect sizes because

part of the variance would have been attributed to it. Thus, the QTL effect sizes reported are probably overestimates.

Number of genes mapping to a QTL. The number of genes mapping to each QTL confidence interval was calculated using Ensembl protein-coding genes and genes coding for microRNAs (downloaded from BioMart[48]).

Overlap with rat genome database (RGD) QTLs and with QTLs detected in the mouse heterogeneous stock. The calculation of the overlap between RGD and rat heterogeneous stock QTLs as well as between mouse and rat heterogeneous stock QTLs is given in the **Supplementary Note**.

Pathway analysis for the QTLs detected in the rat and mouse heterogeneous stocks. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway terms were retrieved using the R KEGG.db package. We used INRICH[49] to find enrichment of pathways in the mouse and rat phenotypic QTLs (as defined by the 90% confidence interval) called at a low significance threshold (20th percentile of the extreme value distribution). We report the empirical and corrected *P* values from INRICH.

37. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
38. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
39. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069–2070 (2010).
40. Kent, W.J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
41. Fiddy, S., Cattermole, D., Xie, D., Duan, X.Y. & Mott, R. An integrated system for genetic analysis. *BMC Bioinformatics* **7**, 210 (2006).
42. 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).
43. Hinrichs, A.S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34**, D590–D598 (2006).
44. R Core Development Team. *A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, 2004).
45. Magrane, M. & Consortium, U. UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)* **2011**, bar009 (2011).
46. Söding, J., Biegert, A. & Lupas, A.N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **33**, W244–W248 (2005).
47. Eswar, N., Eramian, D., Webb, B., Shen, M.Y. & Sali, A. Protein structure modeling with MODELLER. *Methods Mol. Biol.* **426**, 145–159 (2008).
48. Kasprzyk, A. BioMart: driving a paradigm change in biological data management. *Database (Oxford)* **2011**, bar049 (2011).
49. Lee, P.H., O'Dushlaine, C., Thomas, B. & Purcell, S.M. INRICH: interval-based enrichment analysis for genome-wide association studies. *Bioinformatics* **28**, 1797–1799 (2012).

doi:10.1038/ng.2644