Collaborative Cross mice and their power to map host susceptibility to *Aspergillus fumigatus* infection

Caroline Durrant,[1] Hanna Tayem,[2] Binnaz Yalcin,[1] James Cleak,[1] Leo Goodstadt,[1] Fernando Pardo-Manuel de Villena,[3] Richard Mott,[1] and Fuad A. Iraqi[2,4]

[1]*Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, United Kingdom;* [2]*Department of Clinical Microbiology and Immunology, Sackler Faculty of Medicine, Tel Aviv University, Ramat Aviv, Tel Aviv 69978, Israel;* [3]*Department of Genetics, University of North Carolina, Chapel Hill, North Carolina 27599-72462, USA*

The Collaborative Cross (CC) is a genetic reference panel of recombinant inbred lines of mice, designed for the dissection of complex traits and gene networks. Each line is independently descended from eight genetically diverse founder strains such that the genomes of the CC lines, once fully inbred, are fine-grained homozygous mosaics of the founder haplotypes. We present an analysis of 120 CC lines, from a cohort of the CC bred at Tel Aviv University in collaboration with the University of Oxford, which at the time of this study were between the sixth and 12th generations of inbreeding and substantially homozygous at 170,000 SNPs. We show how CC genomes decompose into mosaics, and we identify loci that carry a deficiency or excess of a founder, many being deficient for the wild-derived strains WSB/EiJ and PWK/PhJ. We phenotyped 371 mice from 66 CC lines for a susceptibility to *Aspergillus fumigatus* infection. The survival time after infection varied significantly between CC lines. Quantitative trait locus (QTL) mapping identified genome-wide significant QTLs on chromosomes 2, 3, 8, 10 (two QTLs), 15, and 18. Simulations show that QTL mapping resolution (the median distance between the QTL peak and true location) varied between 0.47 and 1.18 Mb. Most of the QTLs involved contrasts between wild-derived founder strains and therefore would not segregate between classical inbred strains. Use of variation data from the genomes of the CC founder strains refined these QTLs further and suggested several candidate genes. These results support the use of the CC for dissecting complex traits.

[Supplemental material is available for this article.]

Laboratory mice are important models for many infectious diseases, and inbred strains of mice often show differences in susceptibility to infection, revealing the host mechanisms that perceive and clear pathogens. Genetic mapping of host–pathogen interactions in mice has identified many loci conferring resistance to various infections (Gervais et al. 1984; Scalzo et al. 1990; Marshall and Lemieux 1992; Stevenson et al. 1993; Malo and Skamene 1994; Kemp et al. 1997; Iraqi et al. 2000, 2001; Hernandez-Valladares et al. 2004) and has illuminated the corresponding mechanisms in humans (Vidal et al. 1993; Skamene 1994).

However, classical laboratory strains of mice originated from a small sample of founders (Beck et al. 2000), with shared ancestry largely contributed by *Mus mus domesticus* (Frazer et al. 2007; Yang et al. 2007). In contrast, wild-derived inbred strains carry genetic variation from other subspecies, accumulated over about 1 million yr (Guénet and Bonhomme 2003). Wild mice are constantly attacked by pathogens and are under strong selective pressure. Their genetic differences might reveal evolutionarily important mechanisms of resistance and susceptibility.

An advantage of working with genetic reference panels of inbred, or nearly inbred, mice is that it is possible to replicate experiments on the same genetic background—hence increasing the heritable fraction of variance—but economically, as each line only needs to be genotyped once. Several genetic reference panels of mice exist (Peirce et al. 2004; Williams et al. 2004; Grubb et al. 2009; Bennett et al. 2010). The latest is the Collaborative Cross (CC), a large panel of recombinant inbred lines derived from a genetically diverse set of eight inbred mouse strains (A/J, C57BL/6J, 129S1/SvImJ, NOD/LtJ, NZO/HiLtJ, CAST/EiJ, PWK/PhJ, and WSB/EiJ). Three founders of the CC (CAST/EiJ, PWK/PhJ, and WSB/EiJ) are wild-derived, and the CC has more recombination and genetic variation compared with that of other reference panels (Churchill et al. 2004; Chesler et al. 2008). Simulations indicate it should produce quantitative trait locus (QTL) mapping resolution in the megabase range (Valdar et al. 2005). The CC also has fewer rare variants segregating than in human populations, in the sense that the minor allele frequency of a variant segregating in the CC should not fall much below $1/8 = 12.5\%$. Furthermore, the genomes of the CC founder strains have been sequenced (http://www.sanger.ac.uk/resources/mouse/genomes/), so it is possible to reconstruct the genome of each CC line as a fixed mosaic of the founder chromosomes. It is then possible to test the association of each imputed sequence variant with the phenotype, e.g., using merge analysis (Yalcin et al. 2005).

There are three cohorts of CC mice under construction (Chesler et al. 2008; Iraqi et al. 2008; Morahan et al. 2008). Here we focus on a panel of 120 CC lines at Tel Aviv University (TAU), Israel. These lines are at a sufficiently advanced stage of breeding—between the sixth and 12th generation of brother–sister matings, which should confer over 80% homozygosity (Broman 2005)—that it is now timely to examine their genomes and evaluate their efficacy for mapping QTLs.

As an illustration of the power of the CC, in this study we dissect the genetic response of the CC to infection by *Aspergillus fumigatus* (Af), for which a mouse model for the infection in humans is well established (Smith 1972). Invasive disseminated aspergillosis is a serious disease in humans, inflicting severe damage to the kidneys, liver, spleen, brain, heart, and other organs. It is caused by infection by *Aspergillus*, a fungus common in soil, plant debris, and indoor air environments (Latgé 1999): Af, *Aspergillus flavus*, and *Aspergillus niger* are the most important infective species (Soubani and Chandrasekar 2002). Humans inhale at least several hundred Af airborne conidia (spores) per day, which can produce a wide range of allergic and invasive clinical manifestations depending on the host's immune status. The more serious form of the disease is invasive pulmonary aspergillosis, which is most common in individuals with defective immune systems; users of immunosuppressive therapies, such as those used to prevent rejection following organ transplantation; and individuals at late-stage human immunodeficiency virus infection. Survival rates in humans are ~50% (Nivoix et al. 2008).

We analyze the genomes of the TAU CC mice and show how they can be used to map QTL for susceptibility to Af at high precision and how using sequence variation data significantly refines the search for candidate genes. We find that the variation attributable to wild-derived mice is responsible for most of the QTLs mapped. To our knowledge this is the first report mapping susceptibility loci for invasive aspergillosis in immune-competent mice. Interestingly, our QTLs differ from a previous study in immune-compromised mice (Zaas et al. 2008).

Results

Structure of the genomes of the CC lines

We genotyped one mouse from each of 120 lines at 170,935 informative SNPs. We reconstructed the genomes of each line in terms of the founder strains using the HAPPY package (Mott et al. 2000). This provides a probabilistic reconstruction of the genome mosaic, taking into account that the genomes are not completely inbred and that there will be some genotyping error. Figure 1 shows typical reconstructions for the autosomes of two lines and indicates regions of residual heterozygosity. The shade of gray indicates the certainty that the founder strain is known: In general, over most of the genome there is a sharp well-defined mosaic reconstruction (reconstructions for all lines across the genome are available from http://mus.well.ox.ac.uk/CC/). Regions of ambiguity are either caused by residual heterozygosity or places where some founder strains have identical haplotypes. On average, 74% of loci within each line were predicted to be homozygous in terms of ancestral haplotype reconstruction: This figure underestimates the true level of inbreeding because loci where several ancestral strains share the same haplotype tend not to be called as homozygous (Supplemental Table S1). For the purposes of QTL mapping, on the basis of this analysis it seemed reasonable to ignore the residual heterozygosity within each line and treat all animals from a line as being genetically identical to the genotyped exemplar.

The genome-wide contribution of each founder strain to a CC line was close to the expected of 1/8 = 12.5% except for six lines
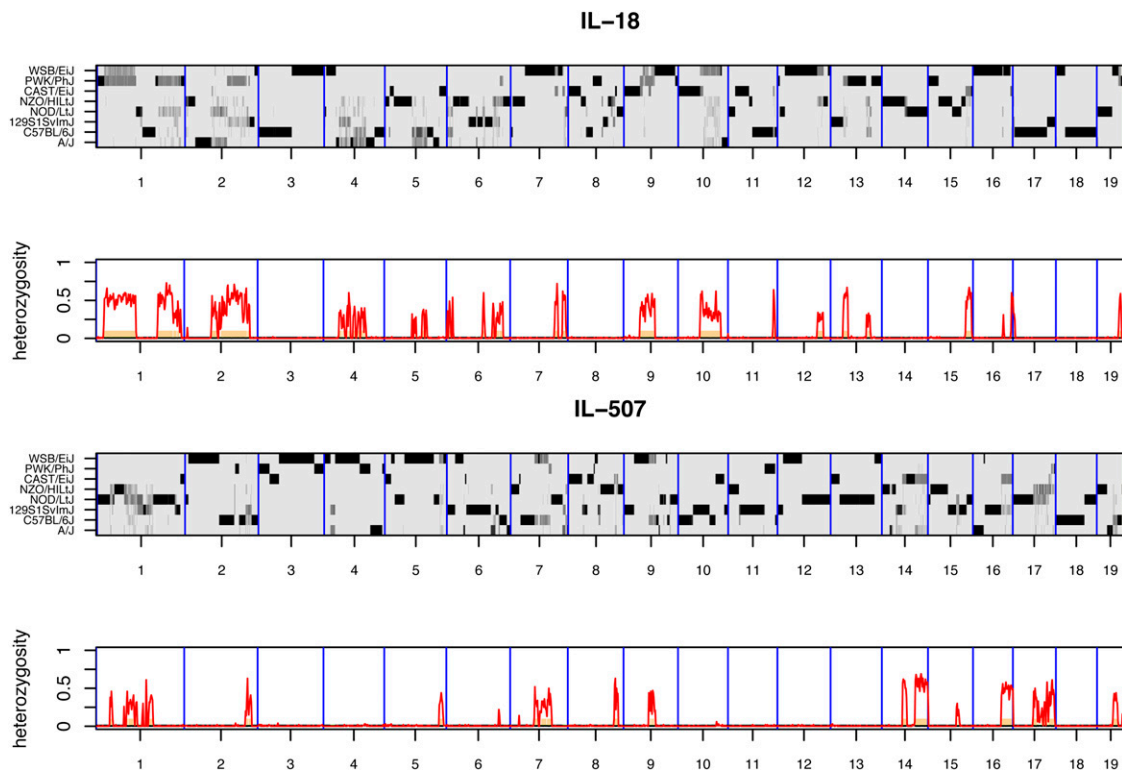

Figure 1. Reconstructions of the genomes of representative CC lines IL-18 and IL-507 from the hidden Markov model (HMM) implemented by HAPPY. The *x*-axis shows the 19 autosomes. Each reconstruction is represented by two panels. The *top* panel *y*-axis shows the eight CC founders, and the probability of descent from a founder at a locus is represented by the shade of gray, with white = 0 and black = 1. Regions where a single haplotype predominates appear as dark horizontal bands; loci with residual heterozygosity or where the founder haplotypes are indistinguishable are gray. The *lower* panel indicates local heterozygosity (red).

where one or more founder was entirely absent, which we attributed to breeding errors (Supplemental Table S1). We examined locus-specific variation in the contribution of the founders and identified 19 loci on 7 chromosomes (at FDR < 1%) where there was a genome-wide excess or deficiency of a founder strain. These are listed in Table 1 and Supplemental Figure S1. The deficient loci frequently lack the wild-derived strains WSB/EiJ (*M. m. domesticus*) or PWK/PhJ (*Mus mus musculus*; Fisher exact test *P* < 0.0022), suggesting that incompatibilities between subspecies may have caused these effects. Gene Ontology enrichment analysis of the genes at these loci failed to identify any over-represented classes of genes.

Susceptibility to infection by Af strain Af293

We first established there were different responses to Af293 infection caused by genetic variation in the host, by challenging a total of 50 females from four immune-competent inbred strains: BALB/cJ, DBA/2J, C3H/HeJ, and C57BL/6J. Post-mortem colony forming units (CFUs) testing confirmed that all mice were infected with a high fungal load, indicating mortality was due to *Af293* infection, but with variable survival times. BALB/cJ mice were highly resistant, with a mean survival of 23.2 (SE 1.93) d, DBA/2J and C3H/HeJ were highly susceptible to the infection, with a mean survival of 5.8 (SE 0.2) d and 7.0 (SE 0.49) d, respectively. C57BL/6J response was intermediate, with a mean survival of 12.7 (SE 20.8) d (Fig. 2). These differences were statistically significant ($P < 5.76 \times 10^{-16}$), and the broad-sense heritability (the explained randomness attributed to between-strain variation) was 0.88.

Susceptible mice tended to lose more weight and showed an increased body temperature relative to resistant animals. Post-mortem fungal load analysis indicated that the spleen, kidney, and liver, but not the lung or brain, were the target organs for systemic Af293 infection (data not shown). Susceptible mice showed fatigue, confusion, and limited motor activity leading ultimately to paralysis, which could be due to Af invasion of the brain, as is the case in humans (Latgé 1999, Soubani and Chandrasekar 2002) and in animal models (Bowman et al. 2001). Resistant mice were normally active during the course of infection, with fewer neurological or other deleterious symptoms (data not shown).

Table 1. Genomic loci in the CC lines with an excess (+) or deficiency (−) of a founder strain

Chromosome	Location (Mb)	Founder strain	Excess/deficiency
1	92.63–93.75	PWK/PhJ	−
2	20.65–52.65	PWK/PhJ	−
2	70.33–71.75	WSB/EiJ	+
2	77.35–92.70	WSB/EiJ	+
2	106.58–117.21	WSB/EiJ	+
2	106.58–125.67	PWK/PhJ	−
2	135.76–139.19	PWK/PhJ	−
3	29.85–35.11	C57BL/6J	+
3	37.48–38.65	C57BL/6J	+
3	42.18–43.98	C57BL/6J	+
3	65.08–66.27	PWK/PhJ	-
3	116.95–130.39	PWK/PhJ	-
4	46.63–53.05	NZO/HiLtJ	+
4	111.04–114.18	A/J	+
6	76.97–78.13	129S1/SvImJ	+
6	79.55–82.93	129S1/SvImJ	+
7	81.81–82.94	PWK/PhJ	−
7	84.22–87.25	PWK/PhJ	−
11	7.00–8.18	NZO/HiLtJ	+

Heritable differences between CC lines

Next, we measured responses to the Af293 challenge in 371 CC mice across 66 lines. There was a broad spectrum of survival times of 4–28 d post-infection (Supplemental Table S2), with highly significant differences between CC lines (Weibull survival regression analysis $P = 1.17 \times 10^{-41}$). Post-mortem testing confirmed that all mice, including those who survived to day 28, were infected at high fungal loads. The broad-sense heritability (the explained randomness attributed to between-line variation in the survival analysis) was 0.78. Figure 2 shows mean survival times of the four inbred lines and all 66 CC lines and their SE; full details for all the 66 lines are in Supplemental Table S2. Fourteen CC lines (21.2%) succumbed between 4 and 10 d, 21 lines (31.8%) between 10 and 20 d, and 14 lines (21.2%) between 20 and 27 d post infection. Seventeen resistant CC lines (25.8%) survived the challenge and were terminated at day 28 post-infection. Representative survival curves for 11 CC lines are presented in Figure 3A. A plot of the empirical cumulative distribution $\log(\log(D(t)))$ of deaths versus the log of time $\log(t)$ across all lines (Fig. 3B) shows an approximately linear relationship, consistent with a Weibull model.

Effects of generation, sex, and batch

The CC mice were tested at different generations of inbreeding, but a test for differences in survival times between generations was not significant ($P = 0.65$). In total, 205 males and 166 females were compared. There was also no overall significant effect of sex on survival ($P = 0.43$). Since males and females have different mean weights, this suggests that bodyweight at challenge did not affect survival. Phenotyping was carried out in eight batches, and a test for differences in the survival times between batches was significant ($P = 0.0032$). However, when QTLs were mapped with batch as a covariate, it did not materially affect the identified QTLs, so the mice were treated as a single population in the QTL analysis presented here (data not shown).

QTL mapping

We used the QTL mapping methodology that we previously established and validated for the MAGIC panel of recombinant inbred lines of *Arabidopsis thaliana* (Kover et al. 2009), adapted for the analysis of survival traits. The genome was divided into 8533 intervals (loci). The probability distribution of descent from the eight founders at each interval was calculated from the HAPPY HMM (Mott et al. 2000) and used to test for association between founder haplotype at each locus and the median survival time from each line. Because there are eight known haplotypes segregating at a locus, association is primarily tested at the level of differences between founder haplotypes rather than between SNP alleles, with merge analysis subsequently used to test imputed sequence variants (Yalcin et al. 2005). The genome scan is shown in Figure 4. Confidence intervals (CIs) were estimated for each QTL separately by simulating a QTL of similar magnitude and strain effects close to the observed QTL peak. Genome-wide thresholds for significance were computed from permutations of the phenotypes. In 95% of permuted genome scans, the global maximum logP was <5.63 (E < 0.05), in 90% it was <5.30 (E < 0.1), and in 50% it was <4.0 (E < 0.5).

Seven QTLs for survival time were mapped at genome-wide E < 0.5 (overall FDR 7%). We identified two distinct significant QTLs on chr 10 with logP of 5.43 (E < 0.10) and 5.78 (E < 0.05). Five additional significant QTL were mapped on chromosomes 2, 3, 8
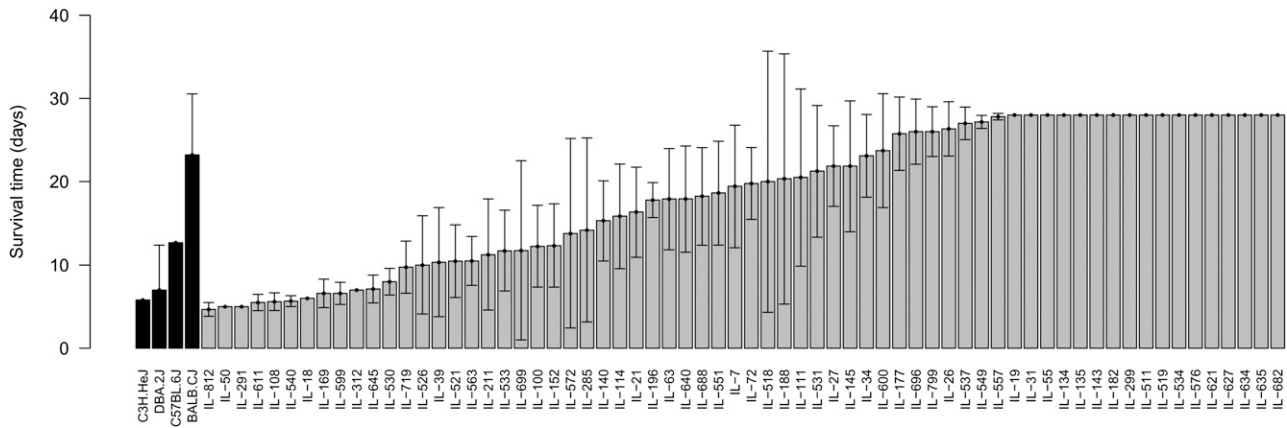
Figure 2. Representative mean survival time in days post-infection with *Aspergillus fumigatus*, of standard inbred lines (black) and of CC lines (gray), with standard errors indicated.

(Fig. 5), 15, and 18, with logP of 4.08 (E < 0.5), 4.24 (E < 0.5), 6.24 (E < 0.05), 5.72 (E < 0.05), and 4.48 (E < 0.5). The three E < 0.5 QTLs are only suggestive. The QTLs and their effect sizes (randomness explained by the QTL in the survival analysis) and CIs are tabulated in Table 2. QTLs are named with the prefix *Asprl* (Aspergillosis-resistant locus). The simulated distribution of CI widths for *Asprl*1 is shown in Figure 5, which is similar to the other QTLs. Typically the 50% width is narrow, so the median distance between the true QTL location and the peak logP is <1 Mb in most QTLs. However, the distribution is heavy-tailed, with broader 90% and 95% CIs.

At each QTL we estimated the effects on survival time due to each of the founder strains, relative to WSB/EiJ, shown in Figure 6. The QTLs are caused by markedly distinct patterns of contrasts between founders, principally involving the wild-derived founders. *Asprl*1 involves a contrast between WSB/EiJ and all other strains; *Asprl*2, between CAST/EiJ and the others; *Asprl*3, predominantly between WSB/EiJ and the others; *Asprl*4, between CAST/EiJ and NZO/HiLtJ versus the others; and *Asprl*7, PWK/PhJ versus the others. Thus most of the QTLs would not segregate in a study involving only classical strains.

Association analysis of sequence variants and candidate genes

We used merge analysis (Yalcin et al. 2005) to impute and test the association of sequence variants segregating between the CC founders within the QTLs. This takes advantage of the ancestry of the CC to infer the alleles of each CC line based on its genome mosaic and sequence variation data in the founder strains. Where a QTL is caused by a single diallelic variant, we expect to have a high chance of testing a very tightly linked tagging SNP with the identical strain distribution pattern in the founders as the causal variant. We also expect the merge analysis of such a SNP to produce higher logP-values than does the eight-way haplotype test in the interval containing a causal variant, due to the reduction in the dimension of the test. If this is not observed, one possibility other than a false positive is that the QTL is caused by a combination of linked variants. It is also possible that an unknown and therefore untested sequence variant (e.g., an indel or copy number variant) that is not tagged by a known SNP could be responsible, although this is unlikely since most indels have similar SDP to a neighboring SNP (Yalcin et al. 2004). The merge analysis for one QTL, *Asprl*1 on chr 8, is shown in Figure 7.

We found increased merge analysis logP for the QTLs on chromosomes 2, 3, 15, and 18 and the first QTL on chr 10, along with an additional region on chr 10 previously unidentified because the eight-way haplotype test did not reach genome-wide significance. Supplemental Table S3 lists the genes under each QTL with the significant merge SNPs nearby. As expected, those QTLs (*Asprl*1, -2, -3, -4, -7) with contrasts involving one wild-derived strain versus the other seven strains were confirmed by merge analysis, where SNPs in which the minor allele is private to the wild-derived strain distribution pattern (SDP) were the most significant. In these QTLs, we can exclude the great majority of variants from being causal. However, although the fraction of SNPs with the correct SDP is small, these are evenly distributed across the QTL, with logP values that track those of the eight-way haplotype logP. Consequently, many genes under the QTL will contain or be close to a variant with that SDP. This situation contrasts with that of a QTL where the SDP of the best SNP has a minor allele frequency closer to 50%, and the spatial decay in linkage disequilibrium concentrates SNPs with the correct SDP around a few genes.

We classified the sequence variants under the QTLs according to whether their merge logP was greater than the corresponding eight-way haplotype logP and by their relationship to the genome annotation (genic [subdivided into coding, UTR, upstream, downstream, or intron], or intergenic [subdivided into repetitive or nonrepetitive]), and calculated the enrichment of variants with high merge logP values in each category (Supplemental Table S4). Interestingly, the genic variants under the 95% CIs for *Asprl*1 and to some extent, *Asprl*4 and *Asprl*6 are enriched for high merge logP values. Under the 95% CI of *Asprl*1, ~12% of coding variants had merge logP higher than the corresponding haplotype test, compared to 4% for intergenic variants (Fisher Exact test $P < 3 \times 10^{-14}$) (Supplemental Table S3).

Finally, we combined the genes associated by merge analysis with published gene function data to identify genes supported by both sources of information. The most significant QTL is *Asprl*1 on chr 8, containing the promising candidate interferon regulatory factor 2 (*Irf2*), which is involved in the host response to infectious diseases (Harada et al. 1989; Masumi et al. 2009). Merge analysis identified an associated SNP near *Irf2* (logP = 5.63). The a priori candidates under *Asprl*3 on chr 15 include the lysosomal-associated protein transmembrane 4B (*Laptm4b*) (Kawai et al. 2001) and
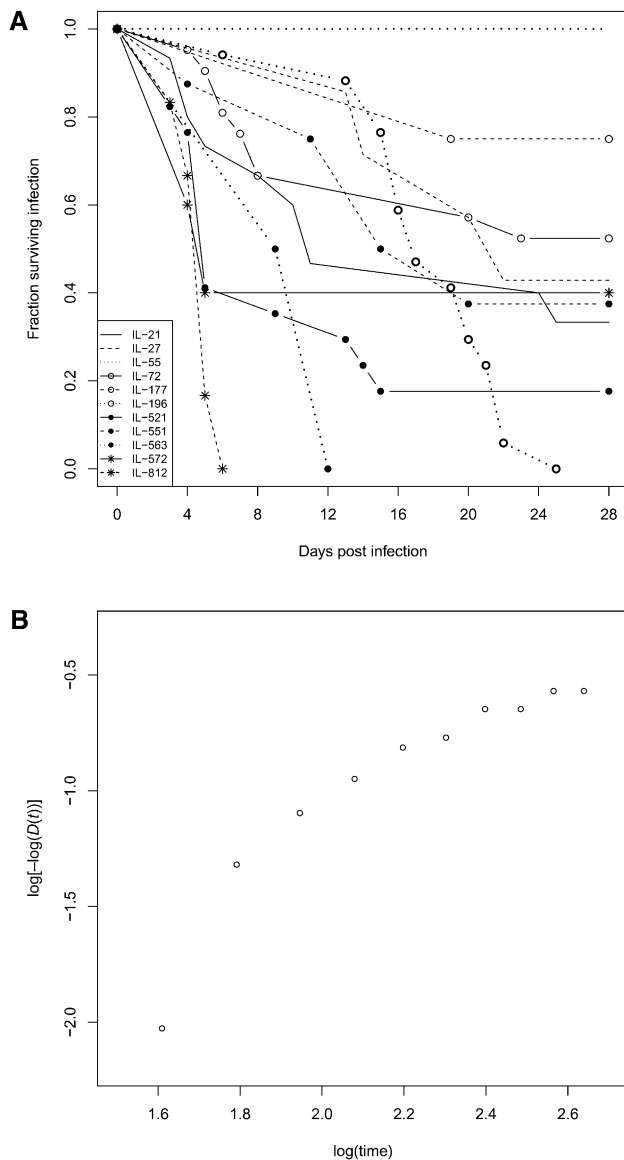
Figure 3. (*A*) Survival curves of selected CC lines following infection with *Aspergillus fumigatus*. (*B*) log[−log(*D*(*t*))] versus log(time) for all CC animals.

the heat-responsive protein 12 (*Hrsp12*) (Samuel et al. 1997). These are supported by merge analysis, which identified a run of sequence variants at logP = 6.18 with the SDP WSB/EiJ versus others, encompassing *Laptm4b* and *Hrsp12* as well as *Matn2, Rp130, BC030476*, and *Pop1*. For a full discussion of candidate genes under all QTLs, see Supplemental File S1.

Discussion

This report, and that by Aylor et al. (2011), demonstrates the utility of the CC in the analysis of complex traits in mouse models of human disease, using 66 partially inbred CC lines to fine-map QTLs for a complex disease resistance trait. Several hundred inbred CC lines will be made available to the research community over the next few years, and using more lines will improve mapping resolution and power. Nonetheless, we have shown that a modest

number of lines is useful if there is sufficient replication (three to five mice in this study) within each line. The CIs of the QTLs, in combination with merge analysis of sequence variants, were small enough to identify candidate genes, although further confirmation work is required. Many of these candidates are involved in innate and adaptive immune responses, including T and B cells and cytokine-related genes.

This is the first study to use immune-competent mouse strains to dissect the genetic response to infection by *Aspergillus* and is the first use of the CC for infectious disease. It is noteworthy that we did not detect any QTLs in the mouse MHC locus on chr 17, and that our QTLs do not replicate a previous study (Zaas et al. 2008) of susceptibility to Af293 in immune-compromised mice. That study compared 10 classical strains and identified a locus on chr 17 containing plasminogen (*Plg*). Three of the strains (129/SvJ, C57BL/6J, A/J) are among the CC founders (if 129/SvJ is equated to 129S1/SvImJ). The *Plg* QTL involved a contrast between 129 and C57BL/6J versus A/J. Therefore, in principle, the QTL should segregate and be detectable in our study, especially as we phenotyped more mice (371 compared with 100). However we found no evidence for a QTL anywhere on chr 17 (logP < 2 throughout). A likely explanation is that we used immune-competent mice, while Zaas et al. (2008) immune-compromised the mice before challenge with Af293. It is possible that immunosuppression alters the host response mechanism and the pathways activated during the infection. Furthermore in the study by Zaas et al. (2008), infection was via inhalation, where the main defense mechanism is by alveolar macrophages in the lung (Dubourdeau et al. 2006). The lungs of the mice in our study were clear from *Aspergillus*, suggesting a different defense mechanism.

Our results confirm the important contribution of the wild-derived strains to the CC. First, these strains are slightly underrepresented at some loci, and because the CC is descended from inbred strains, we know that all alleles in the CC are viable on a suitable genetic background. Therefore one explanation for underrepresentation is that there are incompatible combinations of alleles at different loci (epistasis) between the subspecies of *M. musculus*. Second, the majority of the QTLs we mapped were attributed to differences between the subspecies of *M. musculus*. This is to be expected, given the large number of variants that segregate between them, and it means we expect to find many novel QTLs for other traits in the CC. The most significant QTL we identified, *Aspr1*1, is caused by a contrast between the wild-derived strain WSB/EiJ versus the others. Unexpectedly, the very significant enrichment of coding sequence variants with high logP values under *Aspr1*1 suggests there may not be a single causal variant at this QTL but rather an accumulation of multiple causal alleles with identical strain distribution patterns arising on the WSB/EiJ background, due to selection and historical isolation. In general, at a QTL caused by a single founder, only sequence variants that follow the same strain distribution pattern can be causal for the QTL. For some QTLs no sequence variants were identified by merge analysis, suggesting either that the catalog of variants is incomplete or that the QTL effects are irreducibly based on differences between haplotypes rather than single variants. Therefore completing the sequencing in the founders, beyond those loci that can be assembled from short-read sequence data, will be of great utility.

The task of going from a QTL to a causal gene responsible for susceptibility to a given infectious disease would not have been possible few years ago. By combining new reagents like the CC with DNA variation data, the identification of mechanisms of resistance to pathogens in mice is now within reach.
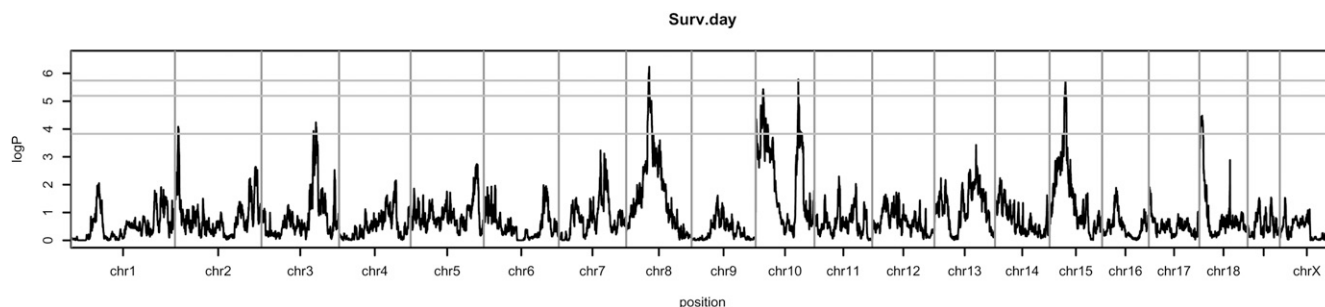
Figure 4. Genome scan of susceptibility to *Aspergillus fumigatus* in 66 CC lines. The *x*-axis is genome location; the *y*-axis is the logP of the test of association between locus and survival time. Genome-wide thresholds of association at E < 0.5, E < 0.1, and E < 0.05 expectation levels are indicated by the horizontal gray lines at logP = 3.97, 5.26, and 5.64, respectively (i.e., the threshold *P* means that in a fraction *P* of permutations the genome-wide maximum logP exceeded the threshold).

Methods

Animals

All animal work was carried out at the small animal facility at The Sackler Faculty of Medicine, TAU, Israel. The Institutional Animal Care and Use Committee of TAU approved all experimental protocols. Mice were housed on hardwood chip bedding in open-top cages and were given tap water and rodent chow ad libitum.

Aspergillus *challenge*

Af strain 293 (Af293) was provided by Prof. Nir Osherov (Department of Clinical Microbiology and Immunology, Sackler Faculty of Medicine, Tel-Aviv University, Israel). Af293 was obtained by growth on Sabouraud dextrose agar (SDA; Difco Laboratories) containing chloramphenicol (Sigma Chemicals). Conidia were harvested in 0.2% (v/v) Tween 80 (Sigma Chemicals), resuspended in triple-distilled water, and counted with a hemocytometer. Conidia numbers were confirmed by colony counts of SDA plate dilutions after 48-h incubation at 37°C.

Mice were inoculated intravenously (IV) via the lateral tail vein using a 27-gauge syringe. Each mouse was challenged with 10^7 conidia of freshly harvested Af293 with a final volume of 0.2 mL in saline medium. This dose should be sufficient to ensure severe infection occurs regardless of the weight or sex of the animal. Negative control CC mice were injected IV with saline to confirm that saline alone would not shorten the animals' lifespan. The survival time of each mouse was recorded for a maximum of 28 d post-infection, when all surviving mice were sacrificed and tested post-mortem for Af293 load by plating extracts from different tissues for growth on SDA containing chloramphenicol and counting CFUs.

Classical inbred mouse lines

Ten immune-competent 8-wk-old female mice from each of the strains BALB/cJ, C57BL/6J, DBA/2J, and C3H/HeJ were purchased from Harlan, Israel, and challenged with Af293.

CC lines

Full details of the development of the TAU CC lines are described by Iraqi et al. (2008). A total of 371 of 8- to 10-wk-old male and female mice, from 66 lines of the TAU CC at inbreeding generation between six and 12, were used in this study (i.e., five to six mice per line were phenotyped). The mice were challenged with Af293 in eight equally sized batches.

Genotyping

We genotyped one mouse from each of the 120 CC lines generated at TAU, of which 85 were genotyped at WTCHG (Oxford UK) on the Mouse Diversity Array (Yang et al. 2009) and 39 mice at UNC (Chapel Hill, USA), with a forerunner of this array. Sixty-six of these 120 were used for Af challenge. We removed SNPs with heterozygous or missing genotypes in the 8 CC founders, or that were not in common between the arrays, leaving 170,935 SNPs. The SNPs were mapped onto build 37 of the mouse genome.

Data analysis

Data analysis was performed using the statistical software R (R Development Core Team 2009), including the R package HAPPY. HBREM (Mott et al. 2000). Additional R code used in the analyses along with the genotype and phenotype data is available from http://mus.well.ox.ac.uk/CC/.

Reconstruction of CC ancestral genome mosaics

The methodology used to reconstruct the CC genomes as mosaics of the eight CC founder genomes is based on the hidden Markov model (HMM) HAPPY (Mott et al. 2000). This was originally developed for heterogeneous stock (HS) mice but was later adapted to work with recombinant inbred lines and validated in the MAGIC genetic reference panel of the plant *A. thaliana* (Kover et al. 2009). The HMM can be run in two modes: either assuming a diploid heterogeneous genome (where each chromosome is an independent haplotype mosaic) or assuming an inbred, homozygous genome. Because the CC lines were not completely inbred at the time of the experiment, we used the former. The extent of observable historical recombination in a chromosome of a recombinant inbred line
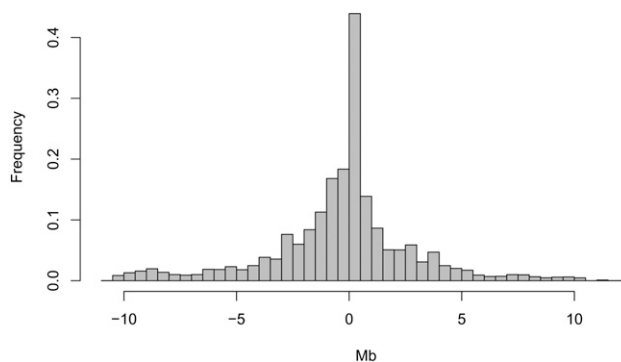

Figure 5. Simulation-based empirical distribution of mapping resolution for QTL *Asprl*1. The *x*-axis is the distance of the maximum logP in the region from the simulated locus. The *y*-axis is the fraction of 100,000 simulations. Note that the *x*-axis labeling is offset to the *left* side of each bar, so the central highest peak is in reality centered at zero.

Table 2. Positions of QTLs associated with susceptibility to *A. fumigatus* infection in 66 CC lines

QTL	Chr	logP	H²	Sig. level	Location (Mb)	Width (Mb; no. of genes)	Location (Mb)	Width (Mb; no. of genes)	Location (Mb)	Width (Mb; no. of genes)
					50% CI		90% CI		95% CI	
*Asprl*1	8	6.24	0.162	0.05	49.39–51.89	2.50 (5)	44.25–55.51	11.26 (64)	42.29–58.96	16.67 (87)
*Asprl*2	10	5.78	0.154	0.05	96.41–97.35	0.94 (6)	94.35–99.01	4.66 (29)	93.65–99.58	5.93 (39)
*Asprl*3	15	5.72	0.153	0.05	33.45–35.26	1.81 (14)	30.97–37.52	6.55 (56)	29.29–38.50	9.21 (65)
*Asprl*4	10	5.43	0.148	0.10	16.72–18.62	1.90 (14)	12.03–23.04	11.01 (74)	9.97–24.57	14.60 (121)
*Asprl*5	18	4.48	0.130	0.50	8.59–10.80	2.21 (18)	5.26–13.58	8.32 (60)	4.38–14.54	10.16 (66)
*Asprl*6	3	4.24	0.125	0.50	115.41–118.78	3.37 (30)	109.57–123.56	13.99 (81)	108.46–125.52	17.06 (100)
*Asprl*7	2	4.08	0.122	0.50	9.31–10.51	1.20 (64)	6.35–12.72	6.37 (105)	5.49–14.31	8.82 (139)

The chromosome (Chr), negative log *P*-value (logP), randomness explained by the locus (H²), and genome-wide significance level reached (Sig. level) are given, as well as the position and length of the empirical 50%, 90%, and 95% confidence intervals. The numbers of genes (known protein-coding, pseudogenes, and RNA genes) in these intervals are indicated in parentheses.

depends on g, the number of effective generations of breeding, before the genome becomes so inbred that few meioses are detectable. For the CC, there are three generations of crossing before inbreeding, during which all recombinants should be detectable. We approximate the effect of the 10–20 further generations of inbreeding as four effective generations, thus setting $g = 7$. Computational experiments (data not shown) confirmed that the precise choice of g is not critical to the resulting haplotype mosaic. To allow for genotyping error, we configured the HMM to allow a small probability of 0.001 that any founder was consistent with any SNP allele.

We reconstructed the genome mosaic of each CC line in terms of the eight CC founders across the 170,935 genotypes to compute probabilities of descent from founders for each of the SNP intervals, which we reduced to 8533 intervals by averaging the matrices in groups of $n = 20$ consecutive SNPs. This reduction made analyses faster and reduced further the effects of genotyping error. Mean heterozygosity was computed across each window of 20 SNPs.

The locus-specific fraction of CC lines carrying each of the founders was estimated by summing the HMM posterior probabilities at each interval across all lines. Genome-wide thresholds for significance were computed by permuting the identities of the founders separately within each line, then recomputing the locus-specific fractions and recording the genome-wide maximum and minimum fractions in the permuted data. This process was repeated 200 times to estimate the upper and lower thresholds exceeded in 10% of permutations.

Survival analysis

The probability that the survival time Y_i for an individual from CC line i exceeds x days is modeled by a Weibull distribution:

$$\Pr(Y_i > x) = \exp(-\mu_i x^\gamma),$$

where γ is a scale parameter and $\log \mu_i$ is a linear predictor of covariates

$$\log \mu_i = \mu + \lambda_i,$$

such that μ is the overall mean and λ_i the effect of line i. We also fitted models with sex as a covariate, but as this was nonsignificant, we omit these results (data not shown). The Weibull survival model is both a proportional hazard and an accelerated failure time model and is thus applicable to a wide range of survival analyses. Goodness of fit was assessed by plotting $\log\log[-\log(D(t))]$ against $\log(t)$, where $D(t)$ is the number of deaths occurring before time t, and the plot should be approximately linear. This model was fitted to survival data using the R functions survfit and survreg in the R survival package. Comparisons between models were tested using a likelihood ratio test, in the R ANOVA function. We also fitted the Cox proportional-hazards model (coxph) to the data as a comparison.

Estimating heritability

We compared the fit of a null model (0) to all 371 CC mice with no covariates and a genetic model (G) with CC line (i.e., genetic) effects. Heritability usually refers to the proportion of variation between individuals in a population that can be attributed to genetic factors. For normally distributed quantitative traits, both heritability (H²) and QTL effect sizes are estimated from the corresponding fraction of attributable variance in an ANOVA. For survival times, particularly when the observations are censored, heritability cannot be estimated in the same way, but several alternatives exist (Hielscher et al. 2010). Here we use the proportion of explained randomness $\rho^2 = 1 - \left(L_G/L_0\right)^{-2/E}$, where L_G, L_0 are the maximized likelihoods from the Weibull genetic and null models and E is the total number of deaths observed (O'Quigley et al. 2005). This generalizes ANOVA heritability in the sense that asymptotically $\rho^2 = H^2$ for quantitative traits with normal errors.

QTL survival analysis

In CC line i at SNP interval (locus) L, the HAPPY HMM probability of descent from founder strains s,t is denoted by $P_{Li}(s,t)$. The presence of a QTL at the locus L is tested using Weibull survival analysis, with the linear predictor

$$\log \mu_i = \mu + \sum_{st} P_{Li}(s,t)(\beta_s + \beta_t) = \mu + \sum_s X_{Lis}\beta_s,$$

where μ is the overall mean, β_s is the effect of founder strain haplotype s at locus L, and $X_{Lis} = \Sigma_t P_{Li}(s,t)$. This model assumes additive effects between the haplotypes at a locus; since the genomes are predominantly homozygous, there is no gain in considering dominance. By construction, $\sum_s X_{Lis} = 2$ (for a diploid organism), so the maximum likelihood estimates $\hat{\beta}_s$ are not independent and can be parameterized in several ways. Here they are expressed as differences from the effect for WSB/EiJ, with $\hat{\beta}_{WSB}$ being set to 0 and the effect of WSB/EiJ represented by the mean $\hat{\mu}$.

For QTL analysis, the median survival time for each CC line is used in the model fitting (thus each line is represented by one mouse). The parameters are estimated by maximum likelihood using the R survreg function, and the presence of a QTL is tested by comparing the log-likelihood for the model with that of a simpler null model in which all the $\hat{\beta}_s = 0$. Significance is reported as the
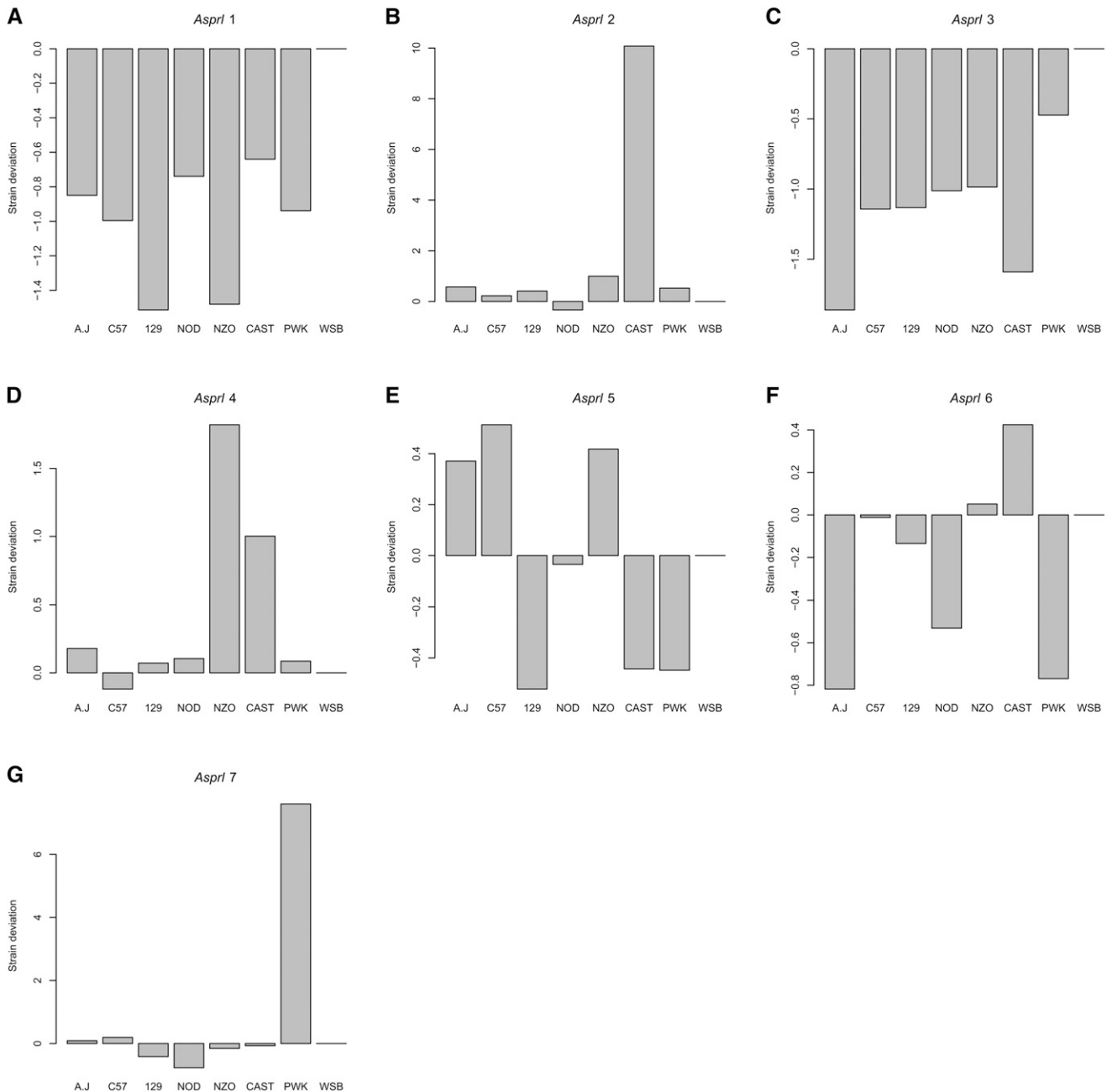
Figure 6. Estimated effects on survival time after *Aspergillus fumigatus* infection, for the eight CC founder strains for each of the *Asprl* QTLs. Effects are shown as deviations relative to WSB/EiJ, which is assigned the trait effect of 0. The *x*-axis of each plot shows the founder strains; the *y*-axis shows the estimated parameter $\hat{\beta}_s$ for the haplotype of the CC founder *s*.

logP, the negative $\log_{10}$ of the *P*-value of the likelihood ratio test of the null versus QTL model, using the R ANOVA function. Genome-wide significance is estimated by permutation, where the CC line labels are permuted between the phenotypes. QTL effect sizes are estimated as the proportion of the explained randomness attributable to the locus effects ($\Sigma_s X_{Lis} \beta_s$) at the QTL. Trait effects (plotted in Fig. 6) are the estimates $\hat{\beta}_s$ reported by the survreg function, relative to founder WSB/EiJ.

We also compared the fit of the above model to that of a mixed-effects Cox proportional hazards model as implemented in the function coxme() in R package kinship. This model is fitted

to all the data (not just the median from each line) by including a random effect γ_i for each line *i*. This did not change the detected QTLs appreciably (data not shown).

Estimation of CIs

We estimated the CI for each QTL by simulation. Accurate estimates of QTL mapping resolution should take into account local patterns of linkage disequilibrium. We devised a method that preserved the genotypes of the data, while simulating survival times caused by a QTL in the neighborhood of the observed QTL
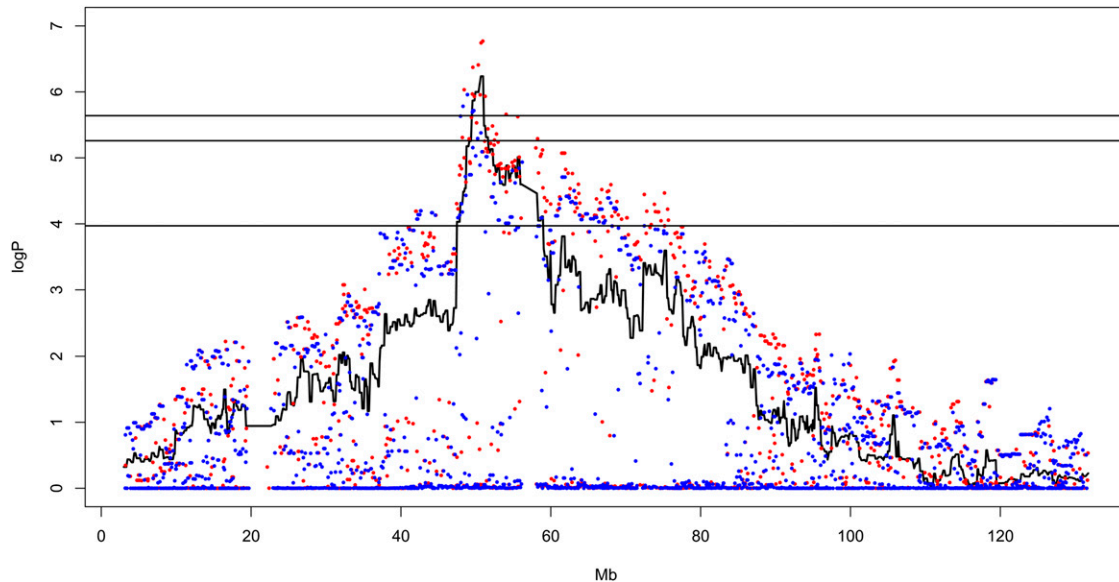
Figure 7. Merge analysis of sequence variants around the QTL *Asprl*1 on chr 8, for susceptibility to *Aspergillus fumigatus* infection in the CC. The *x*-axis is genome location; the *y*-axis is the logP of the test of association between locus and survival time. The continuous black line is the genome scan in Figure 4. The blue and red dots are the results of merge analysis tests of sequence variants segregating in the eight founders of the CC. For clarity, the great majority of variants with merge logP < 1 are not shown. Biallelic SNPs are in blue; variants with more than two alleles (caused when alleles are unknown in one or more founders, which are then treated as private alleles) are red. Image is taken from the genome scan viewer (http://mus.well.ox.ac.uk/gscan_flot/).

peak, and with a similar logP to that observed. We first extracted the parameter estimates $\hat{\beta}_s$ and residuals $\hat{r}_i$ of the fitted survival model at the QTL peak. Let $\hat{t}_i$ be a random permutation of $\hat{r}_i$. Then in a marker interval K within 5 Mb of the QTL peak L, we simulated a set of survival times Z_{iK} caused by a QTL at K by substituting the parameter estimates and permuted residuals:

$$Z_{iK} = \hat{t}_i \, \exp(\hat{\mu} + \sum_s X_{Kis}\hat{\beta}_s).$$

We then rescanned the region and found the interval with the highest logP. We simulated 1000 QTLs at each of 100 intervals K and estimated the *p*% CI from interval containing p% of the simulated local maxima.

Association analysis of sequence variation segregating between the CC founders

Except for a small number of de novo mutations arising during breeding, all sequence variants segregating in the CC should also segregate in the CC founders. Therefore we use the merge analysis methodology (Yalcin et al. 2005) to test which variants under a QTL peak were compatible with the pattern of action at the QTL. A variant with *A* alleles inside the locus *L* merges the eight CC founders into *A* < 8 groups according to whether they share the same allele at the variant (*A* = 2 in the case of SNPs). This merging is characterized by an 8x*A* merge matrix M_{sa} defined to be 1 when strain *s* carries allele *a*, and 0 otherwise. The effect of this merging is tested by comparing the fit of the QTL model above with one in which the Nx8 matrix X_{Lis} is replaced by the NxA matrix $Z_{ia} = \Sigma_s X_{Lis} M_{sa}$. We use the Perlegen SNP database (http://mouse.perlegen.com/mouse/download.html) to test sequence variants globally and the Sanger mouse genomes database (http://www.sanger.ac.uk/resources/mouse/genomes/) for individual genes.

Within the QTLs, we classified the sequence variants according to the genome annotation as repetitive, intergenic, upstream, downstream, UTR, intronic, or coding. We then classified variants according to whether their merge logP was greater

or less than the corresponding haplotype-based logP. The enrichment of variants with high logP values within each category was computed.

Data access

The data from this study are available at http://mus.well.ox.ac.uk/CC/.

References

Aylor DL, Valdar W, Foulds-Mathes W, Buus RJ, Verdugo RA, Baric RS, Ferris MT, Frelinger JA, Heise M, Frieman MB, et al. 2011. Genetic analysis of complex traits in the emerging Collaborative cross. *Genome Res* (in press). doi: 10.1101/gr.111310.110.

Beck JA, Lloyd S, Hafezparast M, Lennon-Pierce M, Eppig JT, Festing MF, Fisher EM. 2000. Genealogies of mouse inbred strains. *Nat Genet* **24:** 23–25.

Bennett BJ, Farber CR, Orozco L, Min Kang H, Ghazalpour A, Siemers N, Neubauer M, Neuhaus I, Yordanova R, Guan B, et al. 2010. A high-resolution association mapping panel for the dissection of complex traits in mice. *Genome Res* **20:** 281–290.

Bowman JC, Abruzzo GK, Anderson JW, Flattery AM, Gill CJ, Pikounis VB, Schmatz DM, Liberator PA, Douglas CM. 2001. Quantitative PCR assay to measure *Aspergillus fumigatus* burden in a murine model of disseminated aspergillosis: demonstration of efficacy of caspofungin acetate. *Antimicrob Agents Chemother* **45:** 3474–3481.

Broman KW. 2005. The genomes of recombinant inbred lines. *Genetics* **169:** 1133–1146.

Chesler EJ, Miller DR, Branstetter LR, Galloway LD, Jackson BL, Philip VM, Voy BH, Culiat CT, Threadgill DW, Williams RW, et al. 2008. The Collaborative Cross at Oak Ridge National Laboratory: developing a powerful resource for systems genetics. *Mamm Genome* **19:** 382–389.

Churchill GA, Airey DC, Allayee H, Angel JM, Attie AD, Beatty J, Beavis WD, Belknap JK, Bennett B, Berrettini W, et al. 2004. The Collaborative Cross, a community resource for the genetic analysis of complex traits. *Nat Genet* **36:** 1133–1137.

Dubourdeau M, Athman R, Balloy V, Huerre M, Chignard M, Philpott DJ, Latge J-P, Ibrahim-Granet O. 2006. *Aspergillus fumigatus* induces innate immune responses in alveolar macrophages through the MAPK pathway independently of TLR2 and TLR4. *J Immunol* **177:** 3994–4001.

Frazer KA, Eskin E, Kang HM, Bogue MA, Hinds DA, Beilharz EJ, Gupta RV, Montgomery J, Morenzoni MM, Nilsen GB, et al. 2007. A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature* **448:** 1050–1053.

Gervais F, Stevenson M, Skamene E. 1984. Genetic control of resistance to *Listeria* monocytogenes: regulation of leukocyte inflammatory responses by the Hc locus. *J Immunol* **132:** 2078–2083.

Grubb SC, Maddatu TP, Bult CJ, Bogue MA. 2009. Mouse phenome database. *Nucleic Acids Research* **37:** D720–D730.

Guénet JL, Bonhomme F. 2003. Wild mice: an ever-increasing contribution to a popular mammalian model. *Trends Genet* **19:** 24–31.

Harada H, Fujita T, Miyamoto M, Kimura Y, Maruyama M, Furia A, Miyata T, Taniguchi T. 1989. Structurally similar but functionally distinct factors, IRF-1 and IRF-2, bind to the same regulatory elements of IFN and IFN-inducible genes. *Cell* **58:** 729–739.

Hernandez-Valladares M, Rihet P, ole-MoiYoi OK, Iraqi FA. 2004. Mapping of a new quantitative trait locus for resistance to malaria in mice by a comparative mapping approach with human chromosome 5q31-q33. *Immunogenetics* **56:** 115–117.

Hielscher T, Zucknick M, Werft W, Benner A. 2010. On the prognostic value of survival models with application to gene expression signatures. *Stat Med* **29:** 818–829.

Iraqi F, Clapcott SJ, Kumari P, Haley CS, Kemp SJ, Teale AJ. 2000. Fine mapping of trypanosomiasis resistance loci in murine advanced intercross lines. *Mamm Genome* **11:** 645–648.

Iraqi F, Sekikawa K, Rowlands J, Teale A. 2001. Susceptibility of tumour necrosis factor-alpha genetically deficient mice to *Trypanosoma congolense* infection. *Parasite Immunol* **23:** 445–451.

Iraqi F, Churchill G, Mott R. 2008. The Collaborative Cross, developing a resource for mammalian systems genetics: A status report of the Wellcome Trust cohort. *Mamm Genome* **19:** 379–381.

Kawai J, Shinagawa A, Shibata K, Yoshino M, Itoh M, Ishii Y, Arakawa T, Hara A, Fukunishi Y, Konno H, et al. 2001. Functional annotation of a full-length mouse cDNA collection. *Nature* **409:** 685–690.

Kemp SJ, Iraqi F, Darvasi A, Soller M, Teale AJ. 1997. Localization of genes controlling resistance to trypanosomiasis in mice. *Nat Genet* **16:** 194–196.

Kover PX, Valdar W, Trakalo J, Scarcelli N, Ehrenreich IM, Purugganan MD, Durrant C, Mott R. 2009. A multiparent advanced generation inter-cross to fine-map quantitative traits in *Arabidopsis thaliana*. *PLoS Genet* **5:** e1000551. doi: 10.1371/journal.pgen.1000551.

Latgé JP. 1999. *Aspergillus fumigatus* and *aspergillosis*. *Clin Microbiol Rev* **12:** 310–350.

Malo D, Skamene E. 1994. Genetic control of host resistance to infection. *Trends Genet* **10:** 365–371.

Marshall P, Lemieux C. 1992. The I-CeuI endonuclease recognizes a sequence of 19 base pairs and preferentially cleaves the coding strand of the *Chlamydomonas moewusii* chloroplast large subunit rRNA gene. *Nucleic Acids Res* **20:** 6401–6407.

Masumi A, Hamaguchi I, Kuramitsu M, Mizukami T, Takizawa K, Momose H, Naito S, Yamaguchi K. 2009. Interferon regulatory factor-2 induces megakaryopoiesis in mouse bone marrow hematopoietic cells. *FEBS Lett* **583:** 3493–3500.

Morahan G, Balmer L, Monley D. 2008. Establishment of "The Gene Mine": a resource for rapid identification of complex trait genes. *Mamm Genome* **19:** 390–393.

Mott R, Talbot CJ, Turri MG, Collins AC, Flint J. 2000. A method for fine mapping quantitative trait loci in outbred animal stocks. *Proc Natl Acad Sci* **97:** 12649–12654.

Nivoix Y, Velten M, Letscher-Bru V, Moghaddam A, Natarajan-Amé S, Fohrer C, Lioure B, Bilger K, Lutun P, Marcellin L, et al. 2008. Factors associated with overall and attributable mortality in invasive aspergillosis. *Clin Infect Dis* **47:** 1176–1184.

O'Quigley J, Xu R, Stare J. 2005. Explained randomness in proportional hazards models. *Stat Med* **24:** 479–489.

Peirce J, Lu L, Gu J, Silver L, Williams R. 2004. A new set of BXD recombinant inbred lines from advanced intercross populations in mice. *BMC Genet* **5:** 7. doi: 10.1186/1471-2156-5-7.

R Development Core Team. 2010. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Samuel SJ, Tzung SP, Cohen SA. 1997. Hrp12, a novel heat-responsive, tissue-specific, phosphorylated protein isolated from mouse liver. *Hepatology* **25:** 1213–1222.

Scalzo AA, Fitzgerald NA, Simmons A, La Vista AB, Shellam GR. 1990. Cmv-1, a genetic locus that controls murine cytomegalovirus replication in the spleen. *J Exp Med* **171:** 1469–1483.

Skamene E. 1994. The Bcg gene story. *Immunobiology* **191:** 451–460.

Smith GR. 1972. Experimental aspergillosis in mice: aspects of resistance. *J Hyg (Lond)* **70:** 741–754.

Soubani AO, Chandrasekar PH. 2002. The clinical spectrum of pulmonary aspergillosis. *Chest* **121:** 1988–1999.

Stevenson FK, Longhurst C, Chapman CJ, Ehrenstein M, Spellerberg MB, Hamblin TJ, Ravirajan CT, Latchman D, Isenberg D. 1993. Utilization of the VH4-21 gene segment by anti-DNA antibodies from patients with systemic lupus erythematosus. *J Autoimmun* **6:** 809–825.

Valdar W, Flint J, Mott R. 2005. Simulating the collaborative cross: power of QTL detection and mapping resolution in large sets of recombinant inbred strains of mice. *Genetics* **172:** 1783–1797.

Vidal SM, Malo D, Vogan K, Skamene E, Gros P. 1993. Natural resistance to infection with intracellular parasites: isolation of a candidate for Bcg. *Cell* **73:** 469–485.

Williams R, Bennett B, Lu L, Gu J, DeFries J, Carosone-Link PJ, Rikke B, Belknap J, Johnson T. 2004. Genetic structure of the LXS panel of recombinant inbred mouse strains: a powerful resource for complex trait analysis. *Mamm Genome* **15:** 637–647.

Yalcin B, Fullerton J, Miller S, Keays DA, Brady S, Bhomra A, Jefferson A, Volpi E, Copley RR, Flint J, et al. 2004. Unexpected complexity in the haplotypes of commonly used inbred strains of laboratory mice. *Proc Natl Acad Sci* **101:** 9734–9739.

Yalcin B, Flint J, Mott R. 2005. Using progenitor strain information to identify quantitative trait nucleotides in outbred mice. *Genetics* **171:** 673–681.

Yang H, Bell TA, Churchill GA, Pardo-Manuel de Villena F. 2007. On the subspecific origin of the laboratory mouse. *Nat Genet* **39:** 1100–1107.

Yang H, Ding Y, Hutchins LN, Szatkiewicz J, Bell TA, Paigen BJ, Graber JH, de Villena FP, Churchill GA. 2009. A customized and versatile high-density genotyping array for the mouse. *Nat Methods* **6:** 663–666.

Zaas AK, Liao G, Chien JW, Weinberg C, Shore D, Giles SS, Marr KA, Usuka J, Burch LH, Perera L, et al. 2008. Plasminogen alleles influence susceptibility to invasive aspergillosis. *PLoS Genet* **4:** e1000101. doi: 10.1371/journal.pgen.1000101.