Rapid Identification of Major-Effect Genes Using the Collaborative Cross

Ramesh Ram,*,† Munish Mehta,*,† Lois Balmer,*,† Daniel M. Gatti,‡ and Grant Morahan*,†,1
*Centre for Diabetes Research, Harry Perkins Institute of Medical Research, Nedlands 6009, WA, Australia, †Centre of Medical Research, University of Western Australia, Nedlands 6009, WA, Australia, and ‡The Jackson Laboratory, Bar Harbor, ME 04609
ORCID IDs: 0000-0002-4827-4778 (R.R.); 0000-0002-8562-7325 (G.M.)

ABSTRACT The Collaborative Cross (CC) was designed to facilitate rapid gene mapping and consists of hundreds of recombinant inbred lines descended from eight diverse inbred founder strains. A decade in production, it can now be applied to mapping projects. Here, we provide a proof of principle for rapid identification of major-effect genes using the CC. To do so, we chose coat color traits since the location and identity of many relevant genes are known. We ascertained in 110 CC lines six different coat phenotypes: albino, agouti, black, cinnamon, and chocolate coat colors and the white-belly trait. We developed a pipeline employing modifications of existing mapping tools suitable for analyzing the complex genetic architecture of the CC. Together with analysis of the founders' genome sequences, mapping was successfully achieved with sufficient resolution to identify the causative genes for five traits. Anticipating the application of the CC to complex traits, we also developed strategies to detect interacting genes, testing joint effects of three loci. Our results illustrate the power of the CC and provide confidence that this resource can be applied to complex traits for detection of both qualitative and quantitative trait loci.

THE Collaborative Cross (CC) project has been in progress for a decade (Churchill *et al.* 2004; Chesler *et al.* 2008; Iraqi *et al.* 2008; Morahan *et al.* 2008; Collaborative Cross Consortium 2012). The CC began from 56 nonreciprocal crosses of eight parental strains: A/J, C57BL/6J, 129S1SvImJ, NOD/LtJ, NZO/HILtJ, CAST/EiJ, PWK/PhJ, and WSB/EiJ. (For convenience, these strains are referred to below as A/J, C57BL/6J, 129S1, NOD, NZO, CAST, PWK and WSB.) Whole-genome sequencing showed that >85% of common species genetic variability was encompassed within these founder strains (Yalcin *et al.* 2011). Our breeding program generated over 900 lines (Morahan *et al.* 2008), with over 100 CC strains currently at inbreeding generation 15 or beyond.

The CC strains display a vast amount of variation in obvious attributes such as coat color, behavior, body weight, growth size, etc. (Collaborative Cross Consortium 2012). Over 38M SNPs and Indels have been identified among the CC founder strains, ensuring genetic diversity within the CC (Munger *et al.* 2014). A major advantage of the CC over conventional genetic approaches is that only one round of genotyping is required, and these data can be used whenever a new trait is characterized. Many of the CC strains have been genotyped using the MegaMUGA Illumina array, which provides a dense coverage genome-wide by typing 77,808 SNP markers. The founder haplotypes at each genomic interval can then be imputed using these genotypes (Mott *et al.* 2000; Yalcin *et al.* 2005; Zhang *et al.* 2014; Collaborative Cross Consortium 2012; also see *Materials and Methods*).

Application of these genetic data to analyze phenotypes of interest allows rapid detection of relevant loci. There are several factors that control the reliability of gene mapping with the CC. These include the number of lines tested for a trait of interest; the founder haplotype diversity present per locus among these strains; the effect of covariant factors on the desired trait of interest; the multigenic nature of the trait; the effect size of the gene on the trait of interest; and the presence of phenocopies. In the case of a monogenic trait, a group of CC lines sharing a common trait will share the same founder haplotype(s) at the causative genetic locus. In a polygenic trait, there will be some inconsistencies in the sharing of founder alleles and hence a linear

doi: 10.1534/genetics.114.163014
Manuscript received February 15, 2014; accepted for publication June 10, 2014
Available freely online through the author-supported open access option.
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.163014/-/DC1.
SNP genotypes for 110 strains are accessible online at http:///www.geniad.com/SNPBrowse.html. CC gene mapping can be done from http://www.sysgen.org/GeneMiner.
[1]Corresponding author: Centre for Diabetes Research, Harry Perkins Institute of Medical Research, QQ Block, QEII Medical Centre, 6 Verdun St., Nedlands, WA 6009, Australia. E-mail: grant.morahan@uwa.edu.au

mixed model can be used to evaluate the maximum-likelihood estimate (derived LOD score) for each genomic position with a suitable significance threshold to differentiate signal from noise. Recently, Bayesian Networks based analysis methods have also been proposed to map polygenic traits (Scutari *et al.* 2014). In the case of a categorical trait, we show below that an analysis using logistic regression or even Fisher's exact test is appropriate, especially in the case of small sample sizes.

The power of the CC was formally calculated by Valdar *et al.* (2006). They determined that 500 CC strains provided 67% power to detect a QTL with a 5% additive effect; power rose to ~100% when the QTL effect size exceeded 10%. Unfortunately, it seems unlikely that there will be 500 CC strains available for testing; most groups may be able to test fewer than 100 strains. Therefore, we sought empirical evidence for mapping genes using this lower number. In this report, we validated the utility of this reasonable number of CC strains for rapid mapping of genes mediating specific phenotypes. For this proof-of-principle exercise, we analyzed several coat color phenotypes, as this approach offered the advantage of easily ascertained phenotypes whose genetics have been well established (*cf*. Silvers 1979). In addition, we present a step-by-step guide that may be useful to researchers using the CC for the first time.

Materials and Methods

CC strains

The CC strains used in this study were bred by Geniad and housed in a specific pathogen-free facility at the Animal Resources Centre (Murdoch, WA, Australia) as described (Morahan *et al.* 2008). The Australian Code for the Care and Use of Animals for Scientific Purposes was followed, and the mice were maintained with appropriate ethics approvals. CC mice and data were kindly provided by Geniad. Genotypes for a further 25 CC strains produced at the other two CC colonies were obtained from a publicly available database (http://csbio.unc.edu/CCstatus/index.py?run=AvailableLines).

Quality control and preprocessing

First we obtained genotypes for the eight founders (eight replicates each) on the MegaMUGA genotyping platform from the University of North Carolina CC web site (http://csbio.unc.edu/CCstatus/index.py?run=GeneseekMM). We took consensus calls for each of eight replicates for each founder type. Among the 77,000 SNPs, some 69,245 SNPs were robustly homozygous in these inbred founder lines. Hence we extracted these 69,245 SNPs. For each strain, SNPs with a missing call were removed. PedPhase v3 (Li and Li 2009) was applied to determine the phase of the raw genotypes and to correct any genotyping errors.

Haplotype reconstruction

The phased and cleaned genotypes were separated into two sets of genotypes per strain, namely homozygous genotypes of allele 1 and homozygous genotypes of allele 2 for the genome to be treated as haploid (inbred). These data were used in HAPPY (Mott *et al.* 2000) in conjunction with 69,245 homozygous genotypes of the eight founder strains. We use the method "hdesign" in HAPPY to estimate the founder haplotype having the maximum-likelihood probability for genotype sets of allele 1 and 2 separately. A consensus of the resulting haplotype assignment was taken as the final call. In the regions where the genomes were heterozygous, the haplotype calls for alleles 1 and 2 differed. These data were recoded as 0, 1, and 0.5 for each of eight founder alleles at each marker, where 0 refers to nonfounder haplotype; 1, homozygous founder haplotype; and 0.5, heterozygous founder haplotype.

Candidate gene mapping

A step-by-step guide is presented in Figure 1, with a more detailed description in Supporting Information, File S1. The guide illustrates the steps involved in preprocessing genotyped SNPs, phasing, haplotype estimation, determining consensus haplotype code, and verification followed by qualitative/quantitative mapping methods using haplotype data. Most users will not need to concern themselves with the haplotype imputation steps. A detailed description of the mapping pipeline is provided in the Supporting Information.

Briefly, coat color traits were coded as cases and controls. A logistic regression model was fitted for the trait at each locus using the recoded eight variable haplotype data set (with 7 degrees of freedom). A one-way ANOVA chi-square test was used to estimate the *P*-value of association. In the case of the multinominal analysis, the coat colors were treated as qualitative values from 1 to 5. A false discovery rate (FDR) (Benjamini and Yekutieli 2001) correction method was used to define the genome-wide significant linkage peaks. Peaks were deemed significant after applying an FDR *P*-value correction, with an FDR of $P < 0.001$, while FDR $P < 0.01$ values were treated as suggestive. The founder strain(s) contributing to each trait were determined by deriving coefficients (log odds ratio) of the fit from the logistic/multinominal regression model and using plotting tools implemented in the DOQTL R package (Gatti *et al.* 2014). Then a list of putative genes at each locus was obtained by comparing founder alleles. From this list, identity of the candidate gene was arrived at by its relevance to the tissue studied (*e.g.*, skin and hair follicle).

Results

Genotyping and imputation of founder haplotypes

The coat phenotypes of the CC strains tested here are listed in Table S1. Genotypes were determined from CC breeders at inbreeding generation N16 and beyond. The raw genotype reads were subject to quality control, and the SNPs were positioned with reference to the mm9/build37 assembly. Residual heterozygosity per strain was calculated to be <10% (Table S2).

The founder haplotypes were reconstructed using data for 77,000 SNPs genome-wide (see *Materials and Methods*). Phasing was performed with PedPhase 3 (Li and Li 2009), and then for each marker the most likely founder haplotype was returned using HAPPY (Mott *et al.* 2000). The assigned haplotype call

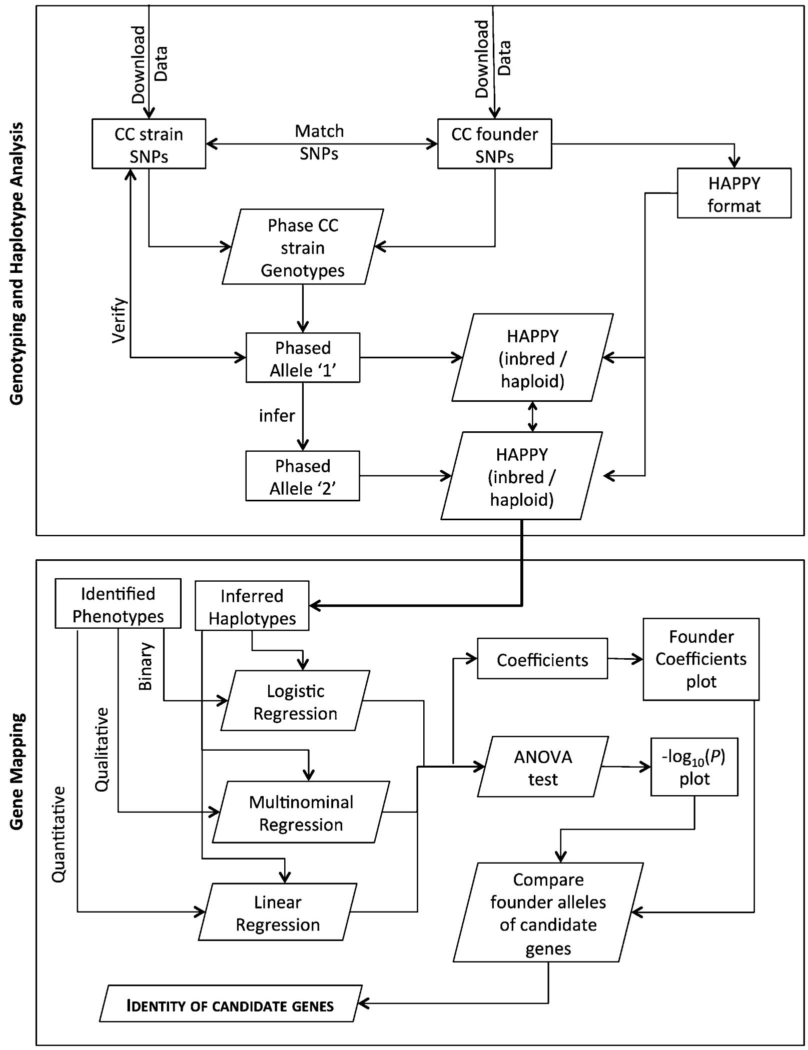

Figure 1 Overview of analytic pipeline. The methods are divided into two parts: (Top) genotyping and haplotyping analysis illustrates steps involved in the transfer from MegaMUGA genotypes to eight founder haplotypes and (Bottom) gene mapping illustrates the steps involved in testing identified phenotype values against genome-wide haplotype information, followed by identification of candidate causal genes.

was then used to reconstruct allele calls for each marker, and this data set was compared against the raw genotyping data for purposes of confirmation. Matching was over 97% for all strains.

An NxMxK weight matrix (where $N = 118$ strains, $M = 8$ founders, $K = 77{,}000$ SNPs) was used to summarize the genotype data. The eight founder weights were assigned based on reconstructed haplotypes as either homozygous weight = 1, heterozygous weight = 0.5 (split between the two founder alleles), or 0 otherwise. Kinship between the CC lines was calculated using raw genotypes and was generally found to be <60% (Table S3). Figure S1 shows the genome-wide correlation in the reconstructed haplotypes of the CC lines. No two CC lines had kinship >80%, demonstrating the genetic diversity of the CC population.

Extraction of nonsynonymous SNPs and common variants

There were ~69,000 SNPs on the MegaMUGA that were homozygous in the eight founders. We obtained founder genotypes for 170,000 SNPs at common variants typed in the JAX Mouse Diversity Genotyping Array (Yang *et al.* 2009). A further 85,000 nonsynonymous (ns) variants from the Sanger Mouse genome sequence project (Yalcin *et al.* 2011) were extracted by parsing query to their web interface. For these Diversity Array and nsSNPs, we imputed genotypes for each CC strain based on the haplotype calls (Yalcin *et al.* 2005). This yielded a genome-wide set of ~329,141 SNPs that could be used for SNP-wise association analyses.

Mapping strategy

An overview of the mapping strategy (including the haplotype inference steps described above) is shown in Figure 1. For the experiments below, we performed a logistic regression fit for the eight founder alleles at each locus (using R-GLM). We also tested the traits using Fisher's exact test (8×2 contingency table, with eight CC founders, two phenotypic values) per SNP (see Supporting Information). We found that Fisher's exact test was just as effective as the logistic regression model in finding QTL positions. However, its utility was limited for more complex studies since it cannot handle covariates.

Proof of principle: mapping the albino locus

Of 110 genotyped strains, 30 were albino. The phenotype was encoded as a binomial value (1, albino; 0, colored). Mapping

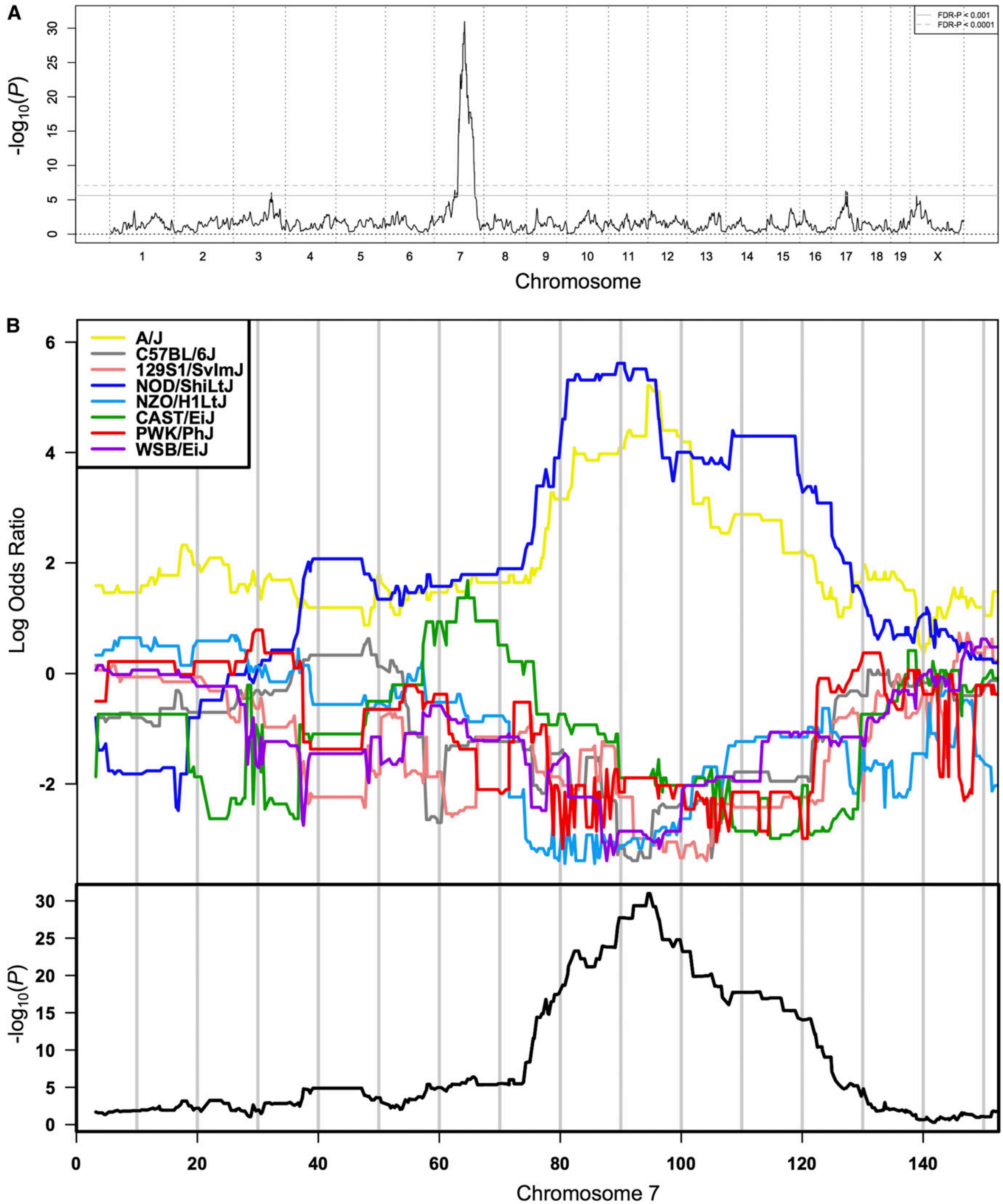

Figure 2 Mapping the albino trait. (A) Genome-wide scan comparing albino *vs.* colored CC strains. The *x*-axis shows the chromosomal position and the *y*-axis shows the $-\log_{10}(P)$ values; the *P*-values were derived from linkage haplotype data. The two threshold lines drawn represent 99.99% (adjusted $P < 0.0001$) confidence and 99.9% (adjusted $P < 0.0001$) confidence. (B) Founder coefficient plot for the chromosome carrying the peak locus. (Top) The plot of the calculated log-odds ratio of eight founder alleles over the chromosome where the founders are color coded. (Bottom) The $-\log_{10}(P)$ values at this chromosome.

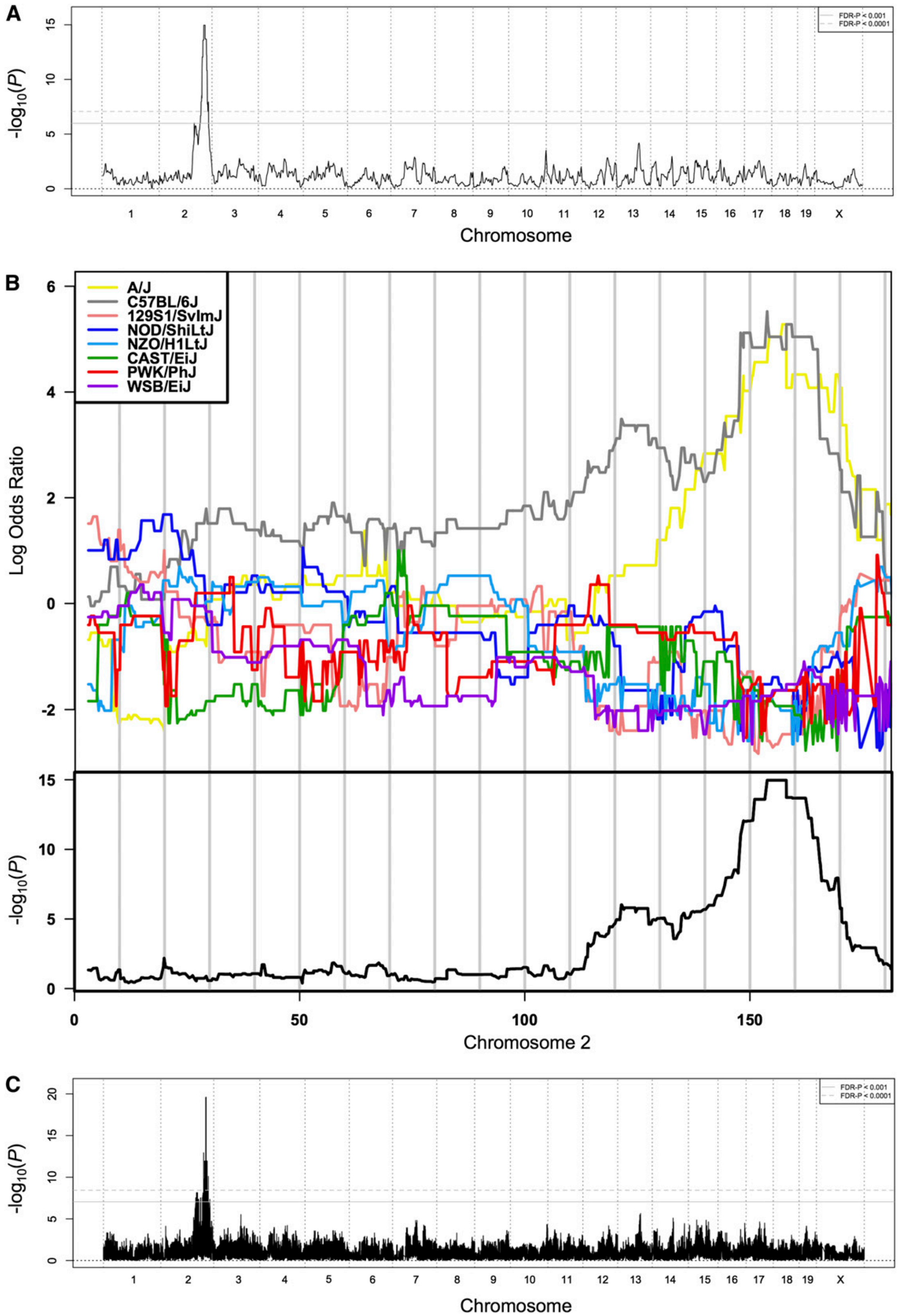

Figure 3 Mapping the agouti trait. (A) Genome-wide scan comparing agouti *vs.* non-agouti CC strains. Other details are as for Figure 2. (B) Founder coefficient plot for the chromosome carrying the peak locus. Details are as for Figure 2. (C) SNP-wise genome-wide scan. The *P*-values were derived from SNP-genotype data. Other details are as for Figure 2.

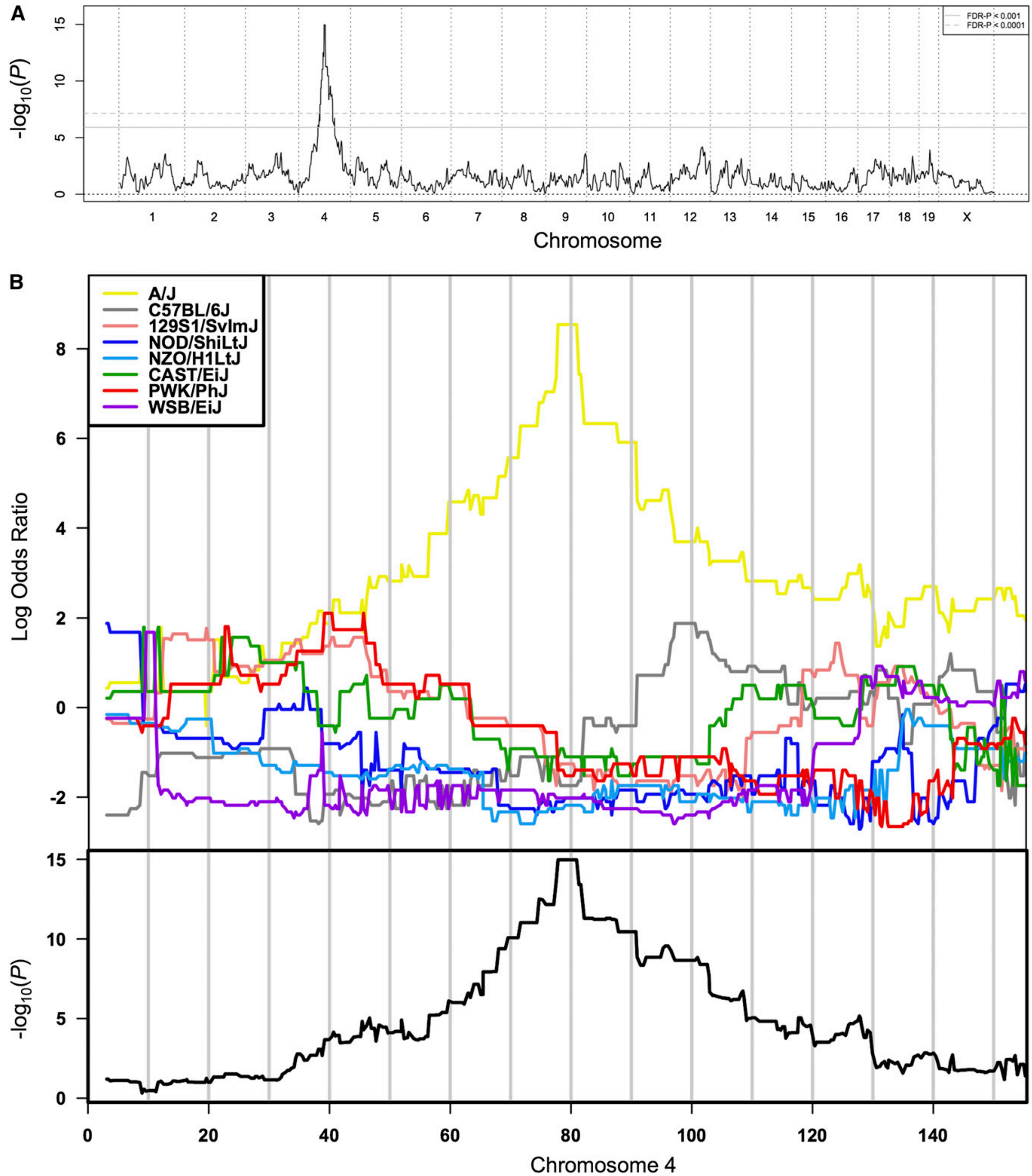

Figure 4 Mapping the cinnamon coat trait. (A) Genome scan comparing cinnamon *vs.* other colored CC strains. Other details are as for Figure 2. (B) Founder coefficient plot for the chromosomes carrying the peak locus. Details are as for Figure 2.

was performed using a logistic regression model (LRM) fit over the reconstructed haplotype matrix. The resulting genome-wide distribution of P (ANOVA chi-squared) is shown in Figure 2A, together with FDR thresholds. The position of the peak SNP was at 93 Mb on chromosome 7. Applying a $-1\ -\log_{10}(P)$ drop restricted the locus interval to between 91 and 96 Mb. The coefficients (log odds ratio) of the fit from the LRM for the chromosome 7 region, together with the corresponding ANOVA test $-\log_{10}(P)$ values are shown in Figure 2B. This analysis clearly showed that haplotypes of

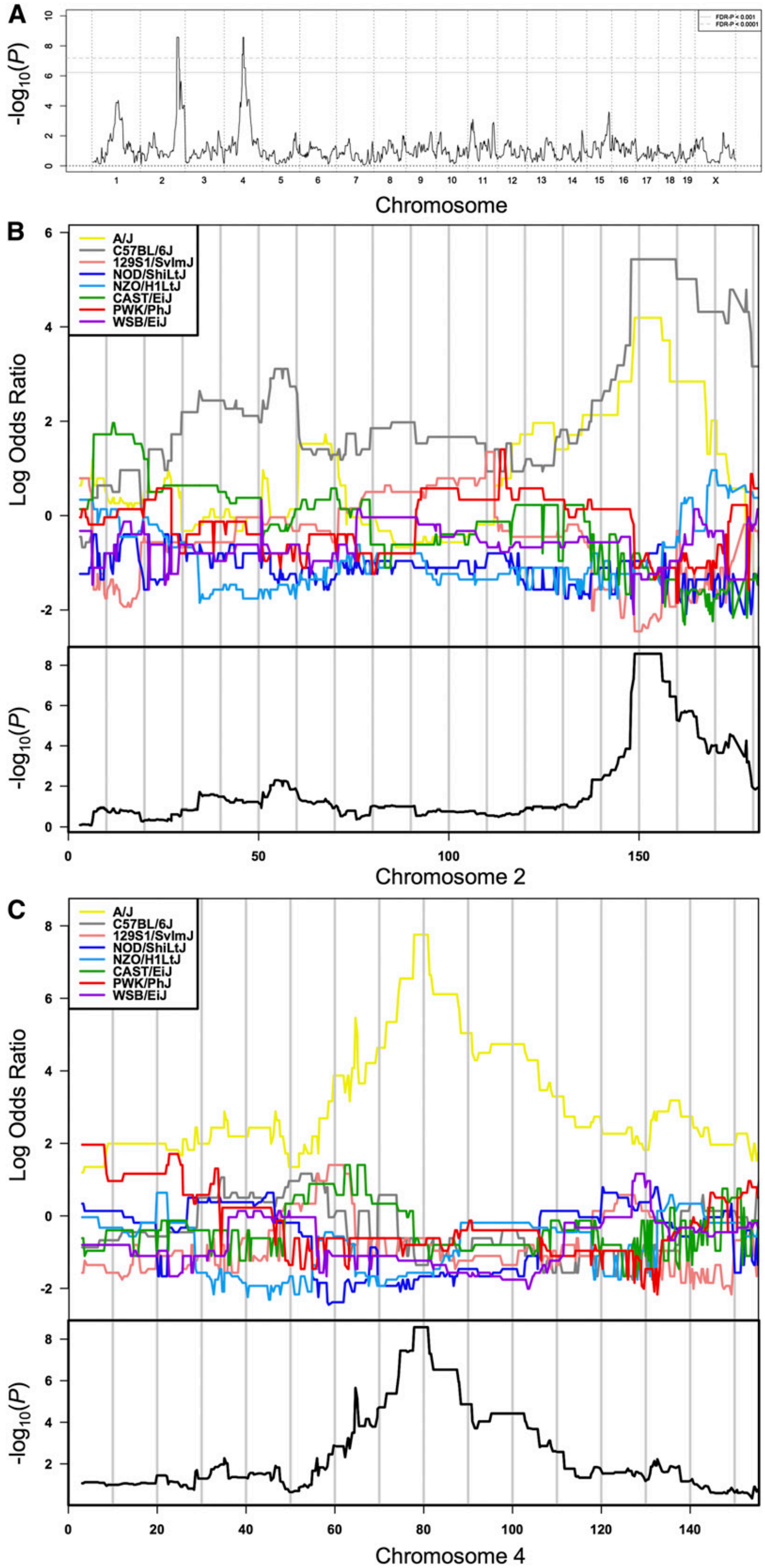

Figure 5 Mapping the chocolate coat trait. (A) Genome-wide scan comparing chocolate *vs.* other colored CC strains. Other details are as for Figure 2. (B & C) Founder coefficient plots for the chromosome carrying the peak loci on chromosome 2 and 4.

the two albino founders (NOD and A/J) contributed to the phenotype.

The catalog of 329,141 genome-wide SNPs (derived as described above) was assessed as an exercise in identifying the causative gene. Within the target region, there were only 9 genes (and 10 missense SNPs) in which the reference allele was present only in the colored group and the variant allele was present only in the albino group. Examining these

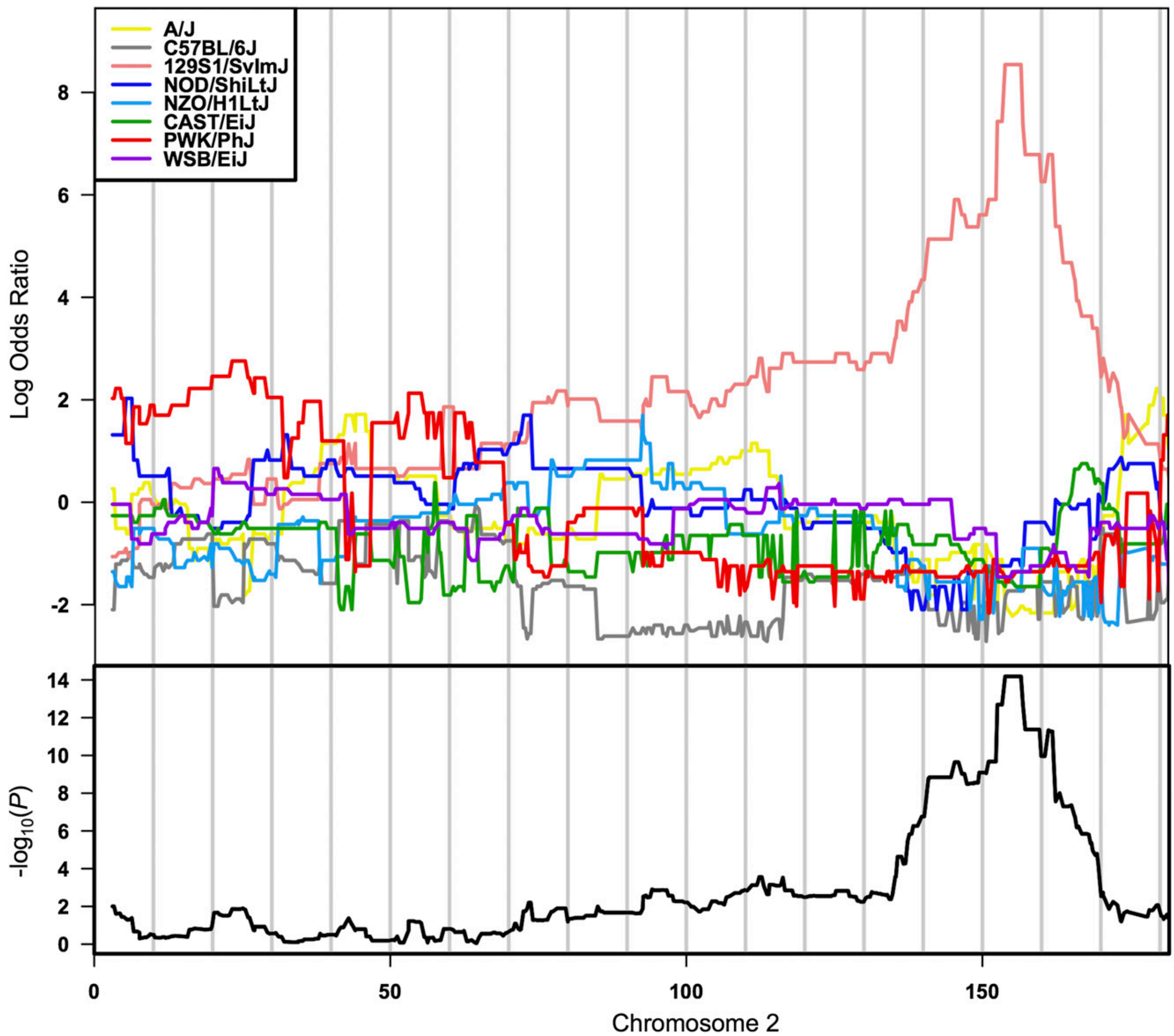

Figure 6 Mapping the white-belly trait. Founder coefficient plot for the chromosome carrying the peak locus after comparing genotypes of white-bellied *vs.* other colored CC strains.

9 genes in the GXD gene expression database (Smith *et al.* 2014) showed that only the Tyrosinase (*Tyr*) gene had significant expression in skin and hair follicle; the G allele of the *Tyr* missense SNP rs31191169 encodes an amino acid change (Cys to Ser) that is predicted by PROVEAN (Choi *et al.* 2012) to have a damaging effect on the protein (Protein seq. ID: NP_035791). The albino trait is known to be due to tyrosinase deficiency (Russell and Russell 1948), and mutations in *Tyr* have been functionally validated as causing albino coat color (Tanaka *et al.* 1990).

Thus, in a few simple steps we could rapidly map and identify the causative gene and variant for this example trait. This demonstrated the power of the CC for rapid gene identification.

Analyzing the agouti trait

Next, we compared 64 pigmented strains. Fifteen of these had black coats while the rest were agouti. A genome scan was conducted using the same methods as above. As shown in Figure 3A, the peak SNP was at 154 Mb of chromosome 2; the $-\log_{10}(P)-1$ confidence interval was between 153.8 and 158.0 Mb. The B6 and A/J founder strains clearly showed allelic differentiation at this locus (Figure 3B). A SNP-wise analysis of 329,141 SNPs revealed 23 significantly associated SNPs in the candidate region (Figure 3C). Among these, there were 11 nsSNPs in seven genes, but none of these were expressed in skin or hair follicle. A query of the Sanger database yielded a total of two SNPs overlapping the *agouti* gene with appropriate allelic distribution between the strains. However, neither of these SNPs was nonsynonymous. Thus, although we could rapidly identify associated SNPs, this low-level approach could not detect the genetic variant responsible for the agouti trait. This is perhaps not surprising since the molecular basis of the non-agouti trait in C57BL/6J strains is the insertion of a retrotransposon into

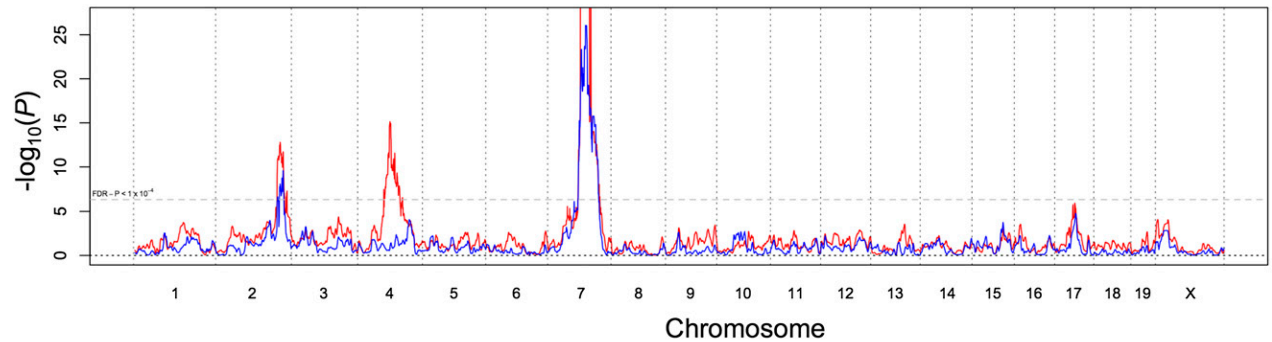

Figure 7 Modeling coat color as a complex trait. Genome-wide scan comparing all CC strains with different coat colors considered as individual traits. The *P*-values were derived from linkage haplotype data. The red lines were derived from multinominal analysis of coat color traits; blue lines were derived from analysis of coat color traits given a quantitative value. The threshold line represents 99.99% (adjusted $P < 0.0001$) confidence that applies to both analyses.

an intron of the *agouti* gene (Bultman *et al.* 1994). [Note that although A/J is albino, it too carries a non-agouti allele (Bultman *et al.* 1994).]

Analyzing the cinnamon coat trait

Cinnamon (or brown agouti) is a coat color dilution trait that is not exhibited by any of the CC founder strains. However, 15 of the 64 pigmented CC strains showed this trait, so we investigated their genetics. The linkage plot is shown in Figure 4A, and the coefficients of the fit for chromosome 4 are shown in Figure 4B. The peak was on chromosome 4, with a confidence threshold between 78 and 81 Mb. The peak was defined by A/J founder alleles; all strains with the cinnamon trait had the A/J haplotype at the locus. In this region, there was only one missense SNP whose alleles showed the appropriate strain distribution pattern: *rs28091500*, located in *Tyrp1*. The A allele was present in the strains with cinnamon coats. This allele encodes the amino acid substitution C110Y, predicted by PROVEAN (Choi *et al.* 2012) to be deleterious. *Tyrp1* encodes tyrosinase-related protein, which has been shown to cause the brown color dilution trait (Bennett *et al.* 1990).

Analyzing the chocolate coat trait

Chocolate may be considered as a darker shade of brown than cinnamon. It is another color dilution trait that is not evident in the CC founder strains. We compared the 64 pigmented strains, of which 9 had chocolate-colored coats. Two significant peaks were seen (Figure 5A): between 79.5 and 80.5 Mb on chromosome 4 and between 149 and 156 Mb of chromosome 2. The coefficients are summarized in Figure 5, B and C. The chocolate and cinnamon coat mice shared the same chromosome 4 gene/allele (*i.e.*, *Tyrp1*). However, all the chocolate coat mice had either a C57BL/6 or an A/J allele at the *agouti* locus compared to the cinnamon mice, suggesting the non-agouti allele at chromosome 2 interacts with *Tyrp1* to produce the chocolate brown coat. Hence, analysis of CC data could rapidly generate a model in which these genes interact to produce the trait of interest.

White-belly gene mapping

Some CC strains have paler fur in the belly area. This trait was also apparent in the 129S1 founder strain. We compared 64 pigmented strains of which 14 displayed a white belly. There was only a single linkage peak. This was on chromosome 2 and overlapped the region harboring the *agouti* (*a*) gene, as shown in Figure 6. Only the 129S1 haplotype contributed to the allelic differentiation. This strain bears an *agouti* mutation (A^w) that is known to induce hypo-pigmentation in the belly area (Dickie 1969).

Modeling coat color as a complex trait

To extend the utility of the CC to mapping genes for complex traits, we tested whether loci could be mapped robustly in a three-gene system. To do so, we modeled coat color as a complex trait, considering all five coat traits displayed by our CC strains. Two analytical methods were used. First, modeling was done with the traits distributed as multinominal categories, and multinominal logistic regression analysis was performed using R-Multinom fit and the *P*-value was obtained from an ANOVA chi-square test. In the second method, coat color was naively assigned a number on a scale from zero (white) through cinnamon, agouti, and chocolate to black (100%) and analyzed using a linear model; the *P*-value was obtained by an ANOVA *F*-test. The results are shown in Figure 7, together with a conservative FDR threshold. Both methods could readily detect linkage to the agouti and albino loci. The multinominal method also correctly identified the contribution of the third locus (*Tyrp*). This example shows that the level of complexity found in a three-gene interaction system could be successfully analyzed using our panel of CC strains and suggests a simple method for accurately mapping the genes of interest.

Reliability of gene mapping using a smaller sample of CC strains

We envision that researchers will prefer to ascertain phenotypes in a smaller set of strains, using these data to map key genes, and validate these in a second, smaller set of CC strains selected to maximize mapping power. To enable

Table 1 Empirical testing of likelihood of successful gene mapping using 50 strains

	No. of loci			
No. of trials	Significant	Suggestive	NS	FP
27	3	0	0	0
86	2	1	0	2
65	2	0	1	1
133	1	2	0	1
292	1	1	1	2
282	1	0	2	0
70	0	3	0	0
100	0	2	1	0
72	0	1	2	0
23	0	0	3	0
Total: 1150				

Permutation analyses were performed using the multinominal QTL scan described in Figure 7. From the set of 110 CC strains, 50 were selected at random for each of 1150 analyses. In each scan, a corrected threshold of $P < 0.001$ was considered as significant, while $P < 0.01$ was considered as suggestive. FP, false positive. NS, no significant linkage observed.

such a scenario, it is important to evaluate the reliability of mapping in a set of strains smaller than the 110 used above. Therefore, we evaluated linkage in >1000 randomly selected sets of 50 strains. Of 1150 permutations, 27 showed genome-wide significance at all three genes with no significant false positives in any of the 27 permutations (Table 1). A total of 885 scans (77% of the total) resulted in at least one of three test loci being detected with genome-wide significance, while 316 scans (27% of the total) resulted in at least suggestive significance at all three test loci. Only 6 scans (<1%) resulted in false positives at the genome-wide significance level.

Minimum number of strains required for analysis of uncommon traits

In our characterization of CC strains, we have observed some traits that are exhibited by only a small number of strains. To determine the minimum number of strains required for reliable mapping of an unusual trait, we used the chocolate coat color as a model. All 501 combinations of between two and eight of the nine chocolate strains were tested to determine what the minimum number of strains would be required for successful mapping of uncommon traits, with comparison to all other colored strains. The comparison group was all other non-albino strains. As shown in Table 2, both loci that contribute to the trait achieved better signals than background using at least six strains, while genome-wide significance was achieved using at least seven strains.

Discussion

The purpose of this study was to provide the proof of principle for applying the CC resource for rapid mapping and identification of genes responsible for traits of interest. Although it was originally planned to produce 1000 CC strains, a combination of factors including poor breeding performance and insufficient funding precluded a resource of this magnitude. Therefore, it was important to establish whether a smaller panel of CC strains would be sufficient to support robust gene mapping in view of the published power estimates calculated for 500 CC strains (Valdar *et al.* 2006).

Our results showed that a panel of ~100 CC strains supported rapid mapping of each of five coat color traits. A sixth trait (white head blaze) was also assigned to the *Kitl* gene (Zsebo *et al.*, 1990) (not shown because this had been demonstrated in analyses of the "pre-CC" by Aylor *et al.* 2011). In addition to gene mapping, this CC panel was also able to support not only identification of the causative gene, but also the genetic variants responsible for determining the albino, chocolate, and cinnamon coat traits.

Mapping of genes for dichotomous traits in the CC is therefore likely to be a very powerful application of this resource. Pilot studies in a screen of only 50 CC strains could identify those with phenotypes at the extremes of the range. A dichotomous test of the extreme phenotype strains should reveal likely candidates for major-effect genes. More complex traits may also be successfully analyzed, as demonstrated with the multinominal analysis of five coat colors. We also demonstrated that major-effect genes could be readily mapped using LRM analyses of CC data.

Table 2 Evaluation of the minimum number of strains required to map interacting major-effect loci

		$-\log_{10}(P)$				
		Chromosome 2: *a* locus		Chromosome 4: *Tyrp1* locus		Other
No. of test strains (*n*)	No. of combinations (9C*n*)	Minimum	Maximum	Minimum	Maximum	Maximum
2	36	1.67	1.94	1.44	1.67	1.94
3	84	2.65	2.97	2.36	2.65	2.65
4	126	3.57	3.94	3.24	3.57	3.57
5	126	4.45	4.84	4.08	4.45	4.45
6	84	5.27	5.69	4.88	5.27	4.02
7	36	6.05	6.05	5.64	6.05	4.48
8	9	6.8	6.8	6.37	6.8	3.76

Genome scans of all combinations of the nine test (chocolate) strains were compared against all other non-albino strains. Results summarize the minimum and maximum $-\log_{10}(P)$ scores determined at the *a* and *Tyrp1* loci. The maximum scores for any other loci (*i.e.*, false positives) are shown for comparison. (Note that the *Tyrp1* scores are generally lower than those for *a* because the cinnamon strains shared the same founder haplotypes at this locus.)

We investigated how few strains were needed for reliable mapping of genes of interest using the CC resource. Our results suggest positive identification of least one of three loci at genomic significance in every 3 of 4 random scans of mapping using a subset of 50 CC strains, while all but 23 scans (*i.e.*, 98%) resulted in detection of one or more of the test loci (*a*, *Tyrp1*, and *Tyr*) with at least suggestive significance. Furthermore, there was a very low rate of false positives (<1%). This work supports a two-stage strategy for mapping using CC strains: an initial scan of phenotypes in 50 strains is likely to detect loci that can be validated in a second stage using CC strains selected to maximize mapping power. Finally, our modeling to determine how few strains were needed to map an uncommon trait showed that as few as 6 strains may be sufficient to obtain suggestive true positives at the candidate loci. These results provide the basis for future investigations using the CC.

The plot of the log-odds of each founder allele calculated at each locus is an accurate way of representing and interpreting the founder haplotype bearing the causative allele. A follow-up SNP-based analysis using a catalog of well-annotated variants would help to narrow down the locus interval and to identify the likely causative gene. With the application of cluster computing, analyses could be expanded to utilize the millions of variants identified from sequencing the founders' genomes (Yalcin *et al.* 2011). Another useful resource for investigating candidate SNPs is the ECCO database (Nguyen *et al.* 2014), which enables researchers to interrogate sequence variation of functional elements for each of 19 tissues/cell types. ECCO catalogs sequence variation in ~300,000 functional elements (*e.g.*, promoters, enhancers, and CTCF-binding sites) active across 17 inbred mouse strains, including the CC founders. Thus, candidate SNPs can be evaluated for effects on *cis*-acting regulatory elements.

This proof-of-principle study tested monogenic traits for which single genes exerted large effects. We demonstrated the suitability of the CC for efficient mapping of major-effect genes and defining the underlying causative genetic variants. Obviously, more complex traits, affected by factors such as epistasis and plieotropy, will be more challenging. Nevertheless, the results presented here showing the rapid and robust identification of genes for qualitative categorical traits provide confidence that future studies of quantitative phenotypes with complex genetic architectures will also benefit from the power of the CC.

Acknowledgments

This work was supported by Discovery Project Grant DP110102067 from the Australian Research Council; by Program Grant 1037321 and Project Grant 1069173 from the National Health and Medical Research Council of Australia; and by the Diabetes Research Foundation of Western Australia. R.R. is supported by the Sunsuper Ride to Conquer Cancer in association with the Harry Perkins Institute of Medical Research. D.M.G. was supported by National Institutes of Health grants P50 GM076468 and R01 GM070683.

Literature Cited

Aylor, D. L., W. Valdar, W. Foulds-Mathes, R. J. Buus, R. A. Verdugo *et al.*, 2011 Genetic analysis of complex traits in the emerging Collaborative Cross. Genome Res. 21: 1213–1222.

Benjamini, Y., and D. Yekutieli, 2001 The control of the false discovery rate in multiple testing under dependency. Ann. Stat. 29: 1165–1188.

Bennett, D. C., D. Huszar, P. J. Laipis, R. Jaenisch, and I. J. Jackson, 1990 Phenotypic rescue of mutant brown melanocytes by a retrovirus carrying a wild-type tyrosinase-related protein gene. Development 110: 471–475.

Bultman, S. J., M. L. Klebig, E. J. Michaud, H. O. Sweet, M. T. Davisson *et al.*, 1994 Molecular analysis of reverse mutations from nonagouti (a) to black-and-tan (a(t)) and white-bellied agouti (Aw) reveals alternative forms of agouti transcripts. Genes Dev. 8: 481–490.

Chesler, E. J., D. R. Miller, L. R. Branstetter, L. D. Galloway, B. L. Jackson *et al.*, 2008 The Collaborative Cross at Oak Ridge National Laboratory: developing a powerful resource for systems genetics. Mamm. Genome 19: 382–389.

Choi, Y., G. E. Sims, S. Murphy, J. R. Miller, and A. P. Chan, 2012 Predicting the functional effect of amino acid substitutions and indels. PLoS ONE 7: e46688.

Churchill, G. A., D. C. Airey, H. Allayee, J. M. Angel, A. D. Attie *et al.*, 2004 The Collaborative Cross: a community resource for the genetic analysis of complex traits. Nat. Genet. 36: 1133–1137.

Collaborative Cross Consortium, 2012 The genome architecture of the Collaborative Cross mouse genetic reference population. Genetics 190: 389–401.

Dickie, M. M., 1969 A white-bellied agouti. Mouse News Lett. 40: 29.

Gatti, D. M., K. L. Svenson, A. Shabalin, L. Wu, W. Valdar *et al.*, 2014 Quantitative Trait Locus Mapping Methods for Diversity Outbred Mice. G3 4: 1623–1633.

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

Li, X., and J. Li, 2009 An almost linear time algorithm for a general haplotype solution on tree pedigrees with no recombination and its extensions. J. Bioinform. Comput. Biol. 7: 521–545.

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

Mott, R., C. J. Talbot, M. G. Turri, A. C. Collins, and J. Flint, 2000 A new method for fine-mapping quantitative trait loci in outbred animal stocks. Proc. Natl. Acad. Sci. USA 97: 12649–12654.

Munger, S. C., N. Raghupathy, K. Choi, A. K. Simons, D. M. Gatti *et al.*, 2014 RNA-Seq Alignment to Individualized Genomes Improves Transcript Abundance Estimates in Multiparent Populations. Genetics 198: 59–73

Nguyen, C., A. Baton, and G. Morahan, 2014 Comparison of sequence variants in transcriptomic control regions across 17 mouse genomes. Database. DOI: doi: 10.1093/database/bau020.

Russell, L. B., and L. W. Russell, 1948 A study of the physiological genetics of coat color in the mouse by means of the dopa reaction in frozen sections of skin. Genetics 33: 237–262.

Scutari, M., P. Howell, D. J. Balding, and I. Mackay, 2014 Multiple Quantitative Trait Analysis Using Bayesian Networks. Genetics 198: 129–137.

Silvers, W. K., 1979 *The Coat Colors of Mice: A Model for Mammalian Gene Action and Interaction*. Springer Verlag, Berlin.

Smith, C. M., J. H. Finger, T. F. Hayamizu, I. J. McCright, J. Xu *et al.*, 2014 The mouse gene expression database (GXD): 2014 update. Nucleic Acids Res. 42: D818–D824.

Tanaka, S., H. Yamamoto, S. Takeuchi, and T. Takeuchi, 1990 Melanization in albino mice transformed by introducing cloned mouse tyrosinase gene. Development 108: 223–227.

Valdar, W., J. Flint, and R. Mott, 2006 Simulating the collaborative cross: power of quantitative trait loci detection and mapping resolution in large sets of recombinant inbred strains of mice. Genetics 172: 1783–1797.

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

Yalcin, B., K. Wong, A. Agam, M. Goodson, T. M. Keane *et al.*, 2011 Sequence-based characterization of structural variation in the mouse genome. Nature 477: 326–329.

Yang, H., Y. Ding, L. N. Hutchins, J. Szatiewicz, T. A. Bell *et al.*, 2009 A customized and versatile high-density genotyping array for the mouse. Nat. Methods 6: 663–666.

Zhang, Z., W. Wang, and W. Valdar, 2014 Bayesian Modeling of Haplotype Effects in Multiparent Populations. Genetics 198: 139–156.

Zsebo, K. M., D. A. Williams, E. N. Geissler, V. C. Broudy, F. H. Martin *et al.*, 1990 Stem cell factor is encoded at the *Sl* locus of the mouse and is the ligand for the c-kit tyrosine kinase receptor. Cell 63: 213–224.

Communicating editor: J. B. Holland

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.163014/-/DC1

Rapid Identification of Major-Effect Genes Using the Collaborative Cross

Ramesh Ram, Munish Mehta, Lois Balmer, Daniel M. Gatti, and Grant Morahan

DOI: 10.1534/genetics.114.163014

Table S1 List of CC lines genotyped and their coat color traits. The coat colors were coded numerically as follows: 0: Albino, 2.5: cinnamon, 5: agouti, 7.5: chocolate, 10: black.

Sample.id	Color	Sample.id	Color	Sample.id	Color
BUTTSWORTH	0	LOU	2.5	RAE2	5
CIS	0	MAF2	2.5	ROGAN	5
CIS2	0	MAK	2.5	SAT	5
CUE	0	MERCURI	2.5	SEH	5
DAM	0	NUY	2.5	SOLDIER	5
DIF	0	PAT	2.5	STUCKY	5
DONNELL	0	REM	2.5	TAS	5
DUB	0	BACKWOOD	5	VUX2	5
DUH	0	BAX2	5	WAD	5
FER2	0	BEM	5	XAM5	5
FEW	0	BOLSEN	5	XAR	5
GIT	0	BOM	5	YID	5
HAX2	0	CAESAR	5	YOX	5
JEUNE	0	CAMERON	5	ZOE	5
JUNIOR	0	DAVIS	5	DEB	7.5
LEY	0	DELTA	5	DOF	7.5
MILLER	0	DOCTOR	5	MARTINI	7.5
MOK	0	DOD	5	PUB	7.5
NUK	0	ERIC	5	TOP	7.5
POR	0	FIM	5	VAK2	7.5
REV	0	FUD	5	WOB2	7.5
RUB	0	FUF	5	YAT	7.5
SOZ	0	GALASUPREME	5	ZAT	7.5
TUY	0	GET	5	BOON	10
VIT	0	GIG	5	DET3	10
VOY	0	HAZ	5	GAV	10
VUN	0	HIP	5	GEK2	10
WAB2	0	MEY	5	HOE	10
WOLLOMAI	0	MIW	5	JAFFA	10
XED	0	MOP	5	JUD	10
BEW	2.5	PEF	5	MEE	10
BIS	2.5	PEF2	5	ROSE	10
CIV2	2.5	PER2	5	TOFU	10
ELROD	2.5	PIPING	5	YIL	10
FIV	2.5	POG	5	ZAQ	10
FLINTOFF	2.5	POH	5	ZIF2	10
HAE	2.5	POT	5		

Table S2 List of CC lines genotyped and their percent residual heterozygosity.

Sample.id	%HET	Sample.id	%HET	Sample.id	%HET
BACKWOOD	4.0	GAV	2.2	PUB	1.6
BAX2	18.8	GEK2	4.9	RAE2	13.2
BEM	2.5	GET	4.5	REM	18.4
BEW	5.6	GIG	0.1	REV	2.4
BIS	7.1	GIT	6.0	ROGAN	2.3
BOLSEN	6.0	HAE	15.9	RONCHI	6.5
BOM	2.5	HAX2	6.4	ROSE	1.2
BOON	2.3	HAZ	8.4	RUB	11.0
BUTTSWORTH	3.3	HIP	3.7	SAT	1.7
CAESAR	4.4	HOE	2.8	SEH	5.9
CAMERON	8.2	JAFFA	6.7	SOLDIER	9.2
CIS	4.3	JEUNE	9.9	SOZ	7.6
CIS2	5.4	JUD	0.0	STUCKY	2.5
CIV2	8.3	JUNIOR	4.2	TAS	0.4
CUE	10.7	LEY	3.8	TOFU	4.9
DAM	14.2	LOU	6.1	TOP	5.9
DAVIS	9.0	MAF2	30.5	TUY	12.7
DEB	14.0	MAK	8.6	VAK2	7.5
DELTA	5.8	MARTINI	4.5	VIT	0.8
DET3	7.5	MEE	5.3	VOY	9.8
DIF	7.9	MERCURI	4.9	VUN	7.7
DOCTOR	1.3	MEY	8.4	VUX2	2.4
DOD	4.0	MILLER	4.5	WAB2	3.7
DOF	16.7	MIW	9.4	WAD	2.3
DONNELL	3.3	MOK	3.4	WOB2	5.7
DUB	11.8	MOP	4.5	WOLLOMAI	7.1
DUH	8.2	NUK	2.4	XAM5	25.4
ELROD	6.9	NUY	22.8	XAR	15.7
ERIC	16.9	PAT	0.9	XED	14.7
FER2	7.0	PEF	2.2	YAT	7.7
FEW	0.4	PEF2	12.5	YID	2.0
FIM	3.3	PER2	2.0	YIL	10.2
FIV	3.0	PIPING	0.6	YOX	1.2
FLINTOFF	26.0	POG	8.7	ZAQ	19.5
FUD	8.5	POH	1.6	ZAT	14.3
FUF	7.2	POR	7.7	ZIF2	2.8
GALASUPREME	1.6	POT	2.3	ZOE	6.4

Table S3 Genome-wide SNP based IBS (Identity by Descent) Kinship matrix for the CC strains.

Available for download as an Excel file at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.163014/-/DC1

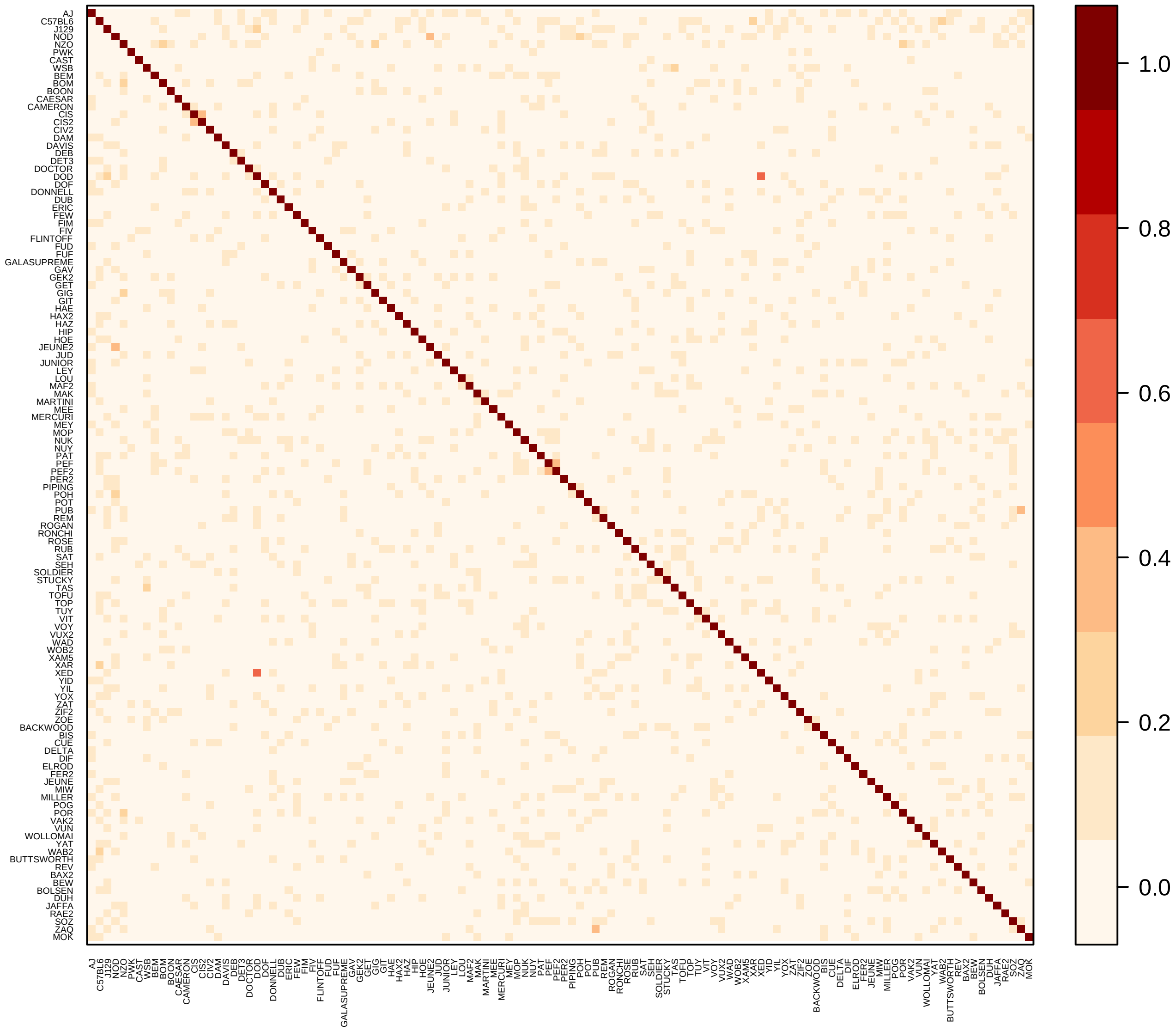

Figure S1 Haplotype correlation Matrix: Heat-map plot of Genome-wide haplotype based kinship matrix for CC strains.

File S1

SUPPLEMENTARY RESULTS and SUPPLEMENTARY METHODS

SUPPLEMENTARY RESULTS

Mapping of Agouti trait using Fisher's exact: We tested the agouti trait using Fisher's exact test (8 x 2 contingency table, with eight CC founders, two phenotypic values) per SNP. Figure S2 shows the results of this analysis. We found that Fisher's exact test was just as effective as the logistic regression model in finding QTL positions. However, its utility was limited for more complex studies since it cannot handle covariates.

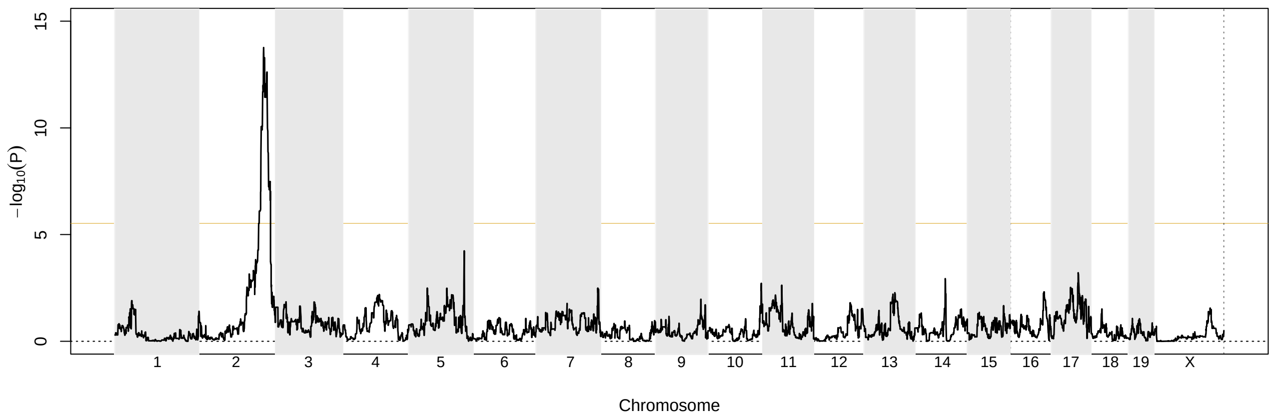

Figure S2 Plot of QTL mapping using Fisher's exact test for the agouti trait.

SUPPLEMENTARY METHODS

Haplotype reconstruction

1. For each Geniad strain:

    a. Extract all 69,245 homozygous non-monomorphic SNPs for the 8 cc founders.

    b. Extract the exact same 69,245 SNPs of the strain tested.

    c. Take out SNPs that are called ('N') during genotyping. If greater than 62,320 SNPs remain, then proceed to next step.

    d. Follow rules: A <-> T, G <-> C, to correct and match strand to that of the founder genotypes at all SNPs

    e. At each of 20 chromosomes: SNPs are incrementally parsed through PEDPHASE v3 to determine phased Allele 1 homozygous genotypes.

    f. Check is performed to ensure integrity of phased Allele 1 data against original genotyping data.

    g. Phased allele 2 genotypes are simply homozygous genotypes of the left over allele at each SNP.

    h. HAPPY format .'alleles' files are created for each chromosome with SNPs placed in order of increasing genetic distance. (cM) (refer to HAPPY manual)

    i. HAPPY format '.data' files are created for each chromosome, in which two set of genotypes: phased allele 1 homozygous genotype and phased allele 2 homozygous genotype are kept and labeled 'strain.name_A' and 'strain.name_B'.

    j. HAPPY objects are created for each pair of 20 '.alleles' and '.data' file with options set: phase = 'known', haploid = TRUE

    k. The founder weights for allele 1 genotype and allele 2 genotype are calculated individually using hdesign() method in HAPPY. This will return two sets of weights in the range of 0 -1.

    l. The founder that carries highest weight in each set is determined as say, founder_A and founder_B respectively. In a homozygous block, founder_A will be same as founder_B. Thus concensus is founder_A. In a heterozygous block founder_A and founder_B are not same. Thus concensus is 'founder_A or founder_B' (example: "AJ or B6").

    m. OPTIONAL: For added confidence: go back to step h, place SNPs in reverse order for each chromosome and adjust genetic distance (cM) accordingly, then repeat steps (i) to (l) with SNPs in reverse orientation. Then, sum the hdesign weights of a marker computed in forward and reverse orientation. This will reurn two summed sets of founder weight in the range 0-2. From this, founder_A and founder_B can be determined as before.

    n. Once founder haplotype for each non-missing SNP for the strain is determined, this is then transferred to the full SNP set containing 69,245 SNPs, where SNPs with missing genotype are assigned (imputed with) haplotype as that of the SNP with non-missing genotype immediately above the missing SNP. In case, should missing genotype occur at the start of a chromosome, haplotype imputed is that of the non-missing SNP immediately below.

  o. Thus haplotype is reconstructed at all 69,245 markers for the strain

2. For N strains: 2 D Matrix M = N x 69,245 Haplotype matrix is obtained by repeating step 1 for all N strains individually.
3. A 3 D matrix 'geno' is then created with dimensions (N_strains x 8_founders x 69,245_SNPs) and assigned '0' value throughout. Then for each strain/SNP pair of 2 D matrix M, a homozygous founder haplotype is assigned a value 1, and in the case of heterozygous pair of founder haplotypes, each founder in the pair is assigned a value 0.5; in the 'geno' matrix.
4. KINSHIP: N x N kinship matrix can be obtained by calculating correlation coefficient between reconstructed haplotypes of pairs of strain. Alternatively IBS matrix can be derived from using raw genotypes and utilized as kinship.

QTL mapping

A. Logistic regression based QTL (qualitative) mapping

1. Using R, following code is used to test each marker:
   a. fit1 = glm(as.factor(phenotype) ~ geno[,,1], family=binomial)
   b. fit2 = glm(as.factor(phenotype) ~ 1, family=binomial) # Null
   c. anova(fit1,fit2, test='Chisq') # anova test

B. Multi-nominal Logistic regression based QTL (qualitative) mapping

1. Using R, the following code is used to test each marker:
   a. library(nnet)
   b. fit1 = multinom(as.factor(phenotype) ~ geno[,,1])
   c. fit2 = multinom(as.factor(phenotype) ~ 1) # Null model
   d. anova(fit1,fit2,test='Chisq') # anova test

C Linear regression based QTL (quantitative) mapping

1. Using R, the following code is used to test each marker:
   a. fit1 = lm(phenotype ~ geno[,,1])
   b. fit2 = lm(phenotype ~ 1) # Null model
   c. anova(fit1,fit2,test='F') # anova F-test

D Linear mixed model based QTL (quantitative) mapping

1. Using R, the following code is used to test each marker:

    a. library(DOQTL)

    b. use scanone() method with appropriate options (ref. manual)