

# Efficient mapping of mendelian traits in dogs through genome-wide association

Elinor K Karlsson<sup>1,2</sup>, Izabella Baranowska<sup>3</sup>, Claire M Wade<sup>1,4</sup>, Nicolette H C Salmon Hillbertz<sup>3</sup>, Michael C Zody<sup>1</sup>, Nathan Anderson<sup>1</sup>, Tara M Biagi<sup>1</sup>, Nick Patterson<sup>1</sup>, Gerli Rosengren Pielberg<sup>5</sup>, Edward J Kulbokas III<sup>1</sup>, Kenine E Comstock<sup>6</sup>, Evan T Keller<sup>6</sup>, Jill P Mesirov<sup>1,2</sup>, Henrik von Euler<sup>7</sup>, Olle Kämppe<sup>8</sup>, Åke Hedhammar<sup>7</sup>, Eric S Lander<sup>1,9–11</sup>, Göran Andersson<sup>3</sup>, Leif Andersson<sup>3,5</sup> & Kerstin Lindblad-Toh<sup>1,5</sup>

With several hundred genetic diseases and an advantageous genome structure, dogs are ideal for mapping genes that cause disease. Here we report the development of a genotyping array with ~27,000 SNPs and show that genome-wide association mapping of mendelian traits in dog breeds can be achieved with only ~20 dogs. Specifically, we map two traits with mendelian inheritance: the major white spotting (*S*) locus and the hair ridge in Rhodesian ridgebacks. For both traits, we map the loci to discrete regions of <1 Mb. Fine-mapping of the *S* locus in two breeds refines the localization to a region of ~100 kb contained within the pigmentation-related gene *MITF*. Complete sequencing of the white and solid haplotypes identifies candidate regulatory mutations in the melanocyte-specific promoter of *MITF*. Our results show that genome-wide association mapping within dog breeds, followed by fine-mapping across multiple breeds, will be highly efficient and generally applicable to trait mapping, providing insights into canine and human health.

The genome of the modern purebred dog bears unmistakable evidence of two tight but widely spaced population bottlenecks: the first occurred at domestication and the second at breed creation. The bottleneck at breed creation and subsequent inbreeding, which produced high rates of inherited diseases in dog breeds, is evident in the genome structure. Within a single breed, linkage disequilibrium (LD) is extensive and haplotype blocks are long (500 kb to 1 Mb), with variation between breeds reflecting differences in historical populations<sup>1,2</sup>. Comparison of many different breeds reveals the short LD and shared haplotype blocks of the much older domestic dog population<sup>1</sup>. A high prevalence of the same disease in two different breeds, especially two related breeds, suggests that the same underlying risk factors were inherited from the ancestral population.

The genetic structure of the dog population suggests that it should be possible to map traits efficiently by a two-stage mapping strategy that uses both the long LD within breeds and the shorter LD across breeds<sup>3</sup>. In the first stage, genome-wide mapping within a single breed would use a relatively sparse marker set and a few dogs to identify a region of association of ~1 Mb. Simulations have suggested that a genome-wide map of ~15,000 SNPs will suffice to define a locus for

a recessive trait using 20 affected dogs and 20 controls<sup>3</sup>. In the second stage, the region of association would be narrowed to a few hundred kilobases by performing fine-mapping with a dense set of SNPs in multiple breeds.

Here we describe the development and general characteristics of a microarray chip containing ~27,000 SNPs and its application to genome-wide association mapping. Using this approach and only ~10 affected and ~10 control dogs, we successfully map two mendelian trait loci, thereby confirming our power predictions. For one of these traits, we reduce the association to a discrete region of ~100 kb by fine-mapping in two dog breeds. For both traits, we identify genes of biological relevance and putative mutations.

## RESULTS

### Development of a high-throughput genotyping array

We developed and validated an SNP array containing ~27,000 markers of high accuracy and relatively even spacing across the dog genome. From the 2.5 million SNPs in the genetic map<sup>1,3</sup>, we used a hierarchical scoring system to select a relatively evenly distributed set of SNPs on the basis of breed representation and technical performance

<sup>1</sup>Broad Institute of Harvard and Massachusetts Institute of Technology (MIT), 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. <sup>2</sup>Bioinformatics Program, Boston University, 44 Cummington Street, Boston, Massachusetts 02215, USA. <sup>3</sup>Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Biomedical centre, Box 597, SE-751 24 Uppsala, Sweden. <sup>4</sup>Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. <sup>5</sup>Department of Medical Biochemistry and Microbiology, Uppsala University, Box 597, SE-751 24 Uppsala, Sweden. <sup>6</sup>Department of Urology, University of Michigan, 1500 East Medical Center Drive, Ann Arbor, Michigan 48109, USA. <sup>7</sup>Department of Clinical Sciences, Swedish University of Agricultural Sciences, SE-750 07 Uppsala, Sweden. <sup>8</sup>Department of Medical Sciences, University Hospital, Uppsala University, SE-751 85 Uppsala, Sweden. <sup>9</sup>Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02142, USA. <sup>10</sup>Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. <sup>11</sup>Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02115, USA. Correspondence should be addressed to E.K.K. (elinor@broad.mit.edu), L.A. (leif.andersson@imbim.uu.se) or K.L.-T. (kersli@broad.mit.edu).

criteria. The SNP spacing averages  $87 \text{ kb} \pm 103 \text{ kb}$ ; 97% of 1-Mb bins across autosomes contain at least five SNPs and 100% contain at least two SNPs (**Supplementary Fig. 1a** online). Coverage of chromosome X is less dense, probably owing to the lower density of polymorphisms and the higher repeat content, with only 42% of 1-Mb bins containing five or more SNPs and 88% containing at least one SNP.

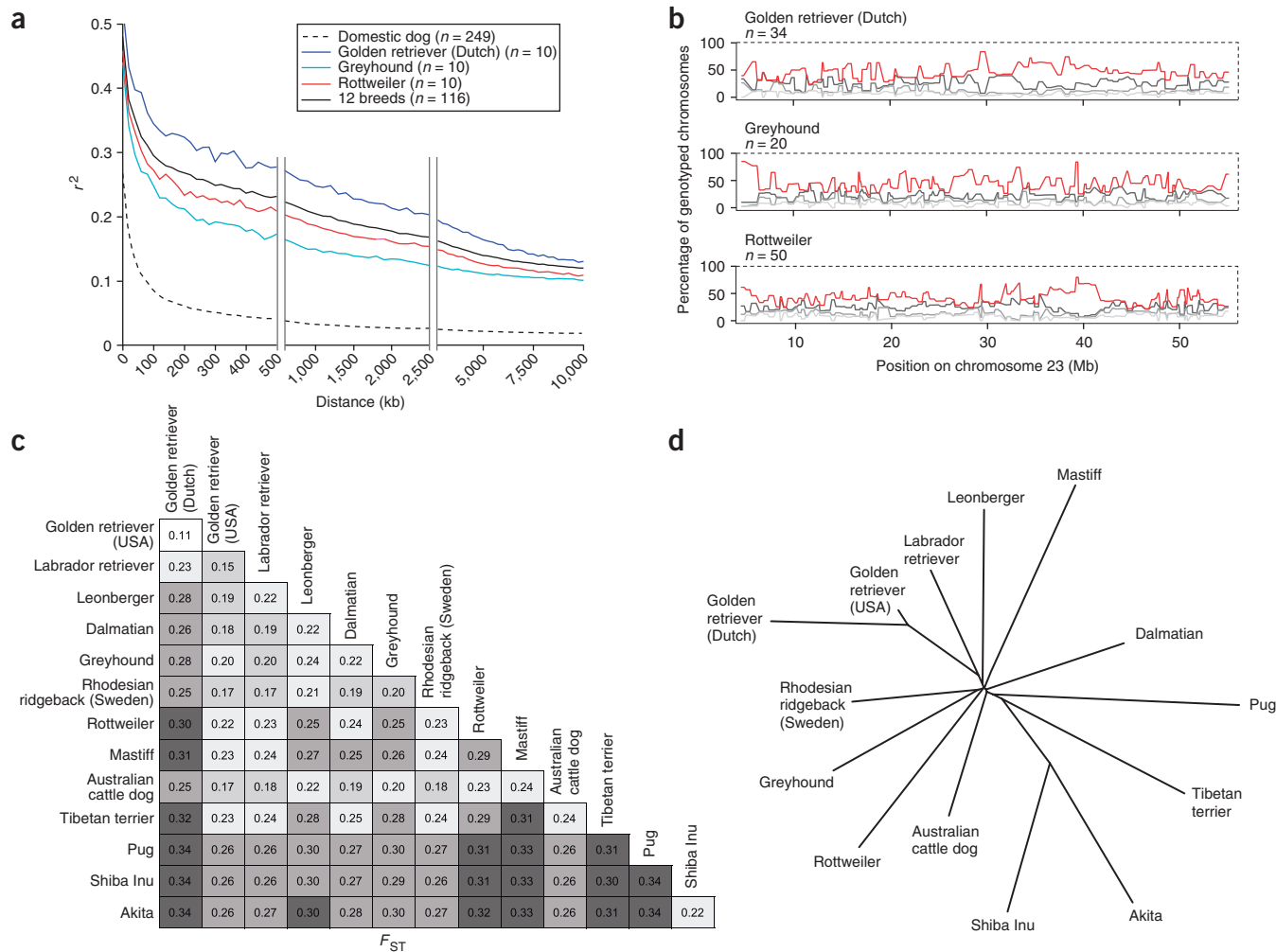
We genotyped a diverse collection of 252 samples, encompassing 21 diverse breeds and two wolves (**Supplementary Table 1a** online), and found that the array was informative for all breeds and both wolves. Specifically, 95.1% of SNPs had a minor allele frequency (MAF) of  $>5\%$  across all dogs analyzed. In any given dog, genotypes were reliably called for  $92.5 \pm 5.6\%$  of the SNPs, of which,  $27.0 \pm 3.8\%$  were called heterozygous (**Supplementary Table 2a** online). The percentages for the two wolves (a Chinese wolf and an Indian wolf) were in the range seen in dogs. In each breed with more than five samples ( $n = 14$ ), most SNPs ( $71.0 \pm 4.0\%$ ) were found to be polymorphic, ranging from 65.2% in the pug to 78.8% in the Australian cattle dog (**Supplementary Table 2b** and **Supplementary Fig. 2a** online). Five

dogs in a breed captured 83% of polymorphic SNPs, whereas ten dogs detected 93% of the variation (**Supplementary Fig. 1b**).

To test the accuracy of the array, a subset of the SNPs was independently genotyped by mass spectrometric genotyping. The overall validation rate was 99.1% ( $n = 1,161$ ) for all SNPs and 99.7% ( $n = 703$ ) for those SNPs with a call rate of  $>90\%$ . Thus, the SNPs with the highest call rates are also the most accurate.

### Genome-wide haplotype structure and breed relationships

By analyzing haplotype structure across the whole genome in 250 dogs, we confirmed that our early observations, based on sampling from ten genomic regions, were valid on a genome-wide scale. In particular, haplotype blocks within breeds are long and typically contain just 3–4 common haplotypes<sup>1,3</sup>. We found that LD measured by the square of the correlation coefficient ( $r^2$ ) is biphasic within breeds, initially dropping sharply but leveling out at  $\sim 100 \text{ kb}$  and remaining above background for 5–15 Mb (**Fig. 1a** and **Supplementary Fig. 2b** online). By contrast, LD across the 250 dogs drops quickly



**Figure 1** Genome structure in dog breeds determined using a genome-wide 27,000 SNP array. **(a)** The short LD in the ancient domestic dog population and the biphasic LD in breeds is measured as  $r^2$  over distance across all dogs (broken line) and within a breed (unbroken lines). Dutch golden retrievers (purple line) have shorter LD as compared with the USA population (**Supplementary Fig. 2b**). **(b)** Most breeds are less than 200 years old, leading to long haplotype blocks with 3–4 common haplotypes that vary in relative frequency, as shown for chromosome 23 in three different breeds. Although the most common haplotype (red line) may predominate, it is rarely fixed. **(c)** Population differentiation measured as  $F_{ST}$  demonstrates that stringent breeding practices and geography have created populations that are roughly twice as diverged as human populations. **(d)** A phylogenetic tree shows that most breeds are derived from a common ancestral population with the possible exception of the Akita and Shiba Inu (both Asian breeds). Branch length corresponds to  $F_{ST}$ .

**Table 1** Regions of complete homozygosity within a breed

	Homozygous blocks <sup>a</sup>	
	No. of regions	% of genome
> 100 kb	2,255	25
> 250 kb	686	14
> 500 kb	166	5.7
> 750 kb	53	2.6
> 1 Mb	23	1.4
> 2 Mb	1.4	0.2
> 5 Mb	0.1	0.0

<sup>a</sup>Average across seven breeds ( $n = 10$  dogs per breed) for autosomal chromosomes.

to background levels. Variability in the extent of LD reflects differences in population history. The Shiba Inu, a breed nearly wiped out by the Second World War, has the longest LD, whereas breeds with large populations, such as the greyhound, have the shortest LD. The average haplotype block size in breeds, defined by the four-gamete rule, is  $\sim 550$  kb. Although a common haplotype occasionally predominates, long regions of limited diversity are rare (Fig. 1b). Within each breed, there are  $\sim 166$  homozygous regions longer than 0.5 Mb ( $\sim 6\%$  of the genome) and only  $\sim 1.4$  longer than 2 Mb (Table 1).

Genetic differentiation between dog breeds is high, reflecting the tight bottlenecks at breed creation. Between breeds,  $F_{ST}$  (a measure of population differentiation<sup>4</sup>) varies from 0.15 to 0.34, which is much higher than in human populations (Fig. 1c). Even between the Dutch and American populations of golden retrievers,  $F_{ST}$  is 0.11, which is roughly equivalent to the  $F_{ST}$  value between European and East Asian human populations<sup>5</sup>. An  $F_{ST}$  phylogeny suggests that most breeds derive from a common ancestral population, but two Asian breeds, the Shiba Inu and Akita, are possibly more distantly related (Fig. 1d). Although a distinct lineage for the Spitz-type Asian breeds supports one of four breed clusters identified in a previous study based on 96 microsatellites<sup>6</sup>, we found no evidence for further subdivision into multiple breed clusters with the genome-wide set of  $\sim 27,000$  SNPs. The long branch lengths in the tree reflect tight breed-creation bottlenecks. A principal component analysis<sup>7</sup> of Dutch and American golden retrievers showed the distinct population stratification underlying the high  $F_{ST}$  (Supplementary Fig. 2c online).

### Genome-wide association mapping

To demonstrate the effectiveness of gene mapping in dogs, we used genome-wide association to map two recessive traits: the ridgeless phenotype in Rhodesian ridgebacks and white coat color in boxers (Fig. 2).

In the Rhodesian ridgeback breed, a characteristic dorsal ridge of inverted hair growth is inherited as an autosomal dominant trait over the normal ridgeless phenotype<sup>8</sup>. The *Ridged* allele predisposes dogs to dermoid sinuses (closed neural tube defects similar to dermal sinuses in humans), suggesting a mutation affecting secondary neurulation<sup>9</sup>. By genotyping 9 ridgeless Rhodesian ridgebacks and 12 ridged controls, we mapped the *ridgeless* allele to a 750-kb region on chromosome 18 (Fig. 2a;  $\chi^2$ -test, nominal  $P$  value ( $P_{\text{raw}}$ ) =  $9.6 \times 10^{-8}$  and  $P$  value corrected for genome-wide search ( $P_{\text{genome}}$ ) =  $1.4 \times 10^{-3}$  on the basis of 100,000 permutations; software package PLINK<sup>10</sup>). This association is 100-fold stronger than that for any other region in the genome (the next highest being  $P_{\text{genome}} = 0.2$ ). Using the Haploview program, we identified a haplotype defined by three SNPs across 750 kb that is homozygous in all but one *Ridged* dog and absent from

the *ridgeless* dogs ( $P_{\text{raw}} = 1.3 \times 10^{-7}$ ; chromosome-wide significance,  $P_{\text{chr}} < 1 \times 10^{-4}$ ; 25,000 permutations; Fig. 2c). This region contains five genes, including three fibroblast growth factor genes (*FGF3*, *FGF4* and *FGF19*). In chick embryos, *FGF3* and *FGF4* are both expressed in the primitive streak during neurulation and later in parts of the neural ectoderm<sup>11,12</sup>. In an accompanying paper<sup>13</sup>, we report that the *Ridged* mutation is a 133-kb duplication that includes all three FGF genes.

We next mapped the locus responsible for the absence of skin and coat pigmentation in white boxers, a semi-dominantly inherited trait in which heterozygous dogs appear part solid, part white (termed 'flash'; Supplementary Fig. 3b online). White boxers suffer increased rates of deafness, reminiscent of the human auditory-pigmentary disorders Waardenburg and Tietz syndromes<sup>14,15</sup>. Breeding studies in the 1950s designated the white coat variant as the extreme-white or  $s^w$  allele of the major white spotting (*S*) locus<sup>16</sup>. Other alleles assigned to this locus are Irish spotting ( $s^i$ ), seen in Basenji (Supplementary Fig. 3f) and Bernese mountain dogs, and piebald spotting ( $s^p$ ), seen in beagles, fox terriers (Supplementary Fig. 3e) and English springer spaniels. Previous research has excluded several candidate genes<sup>17,18</sup>.

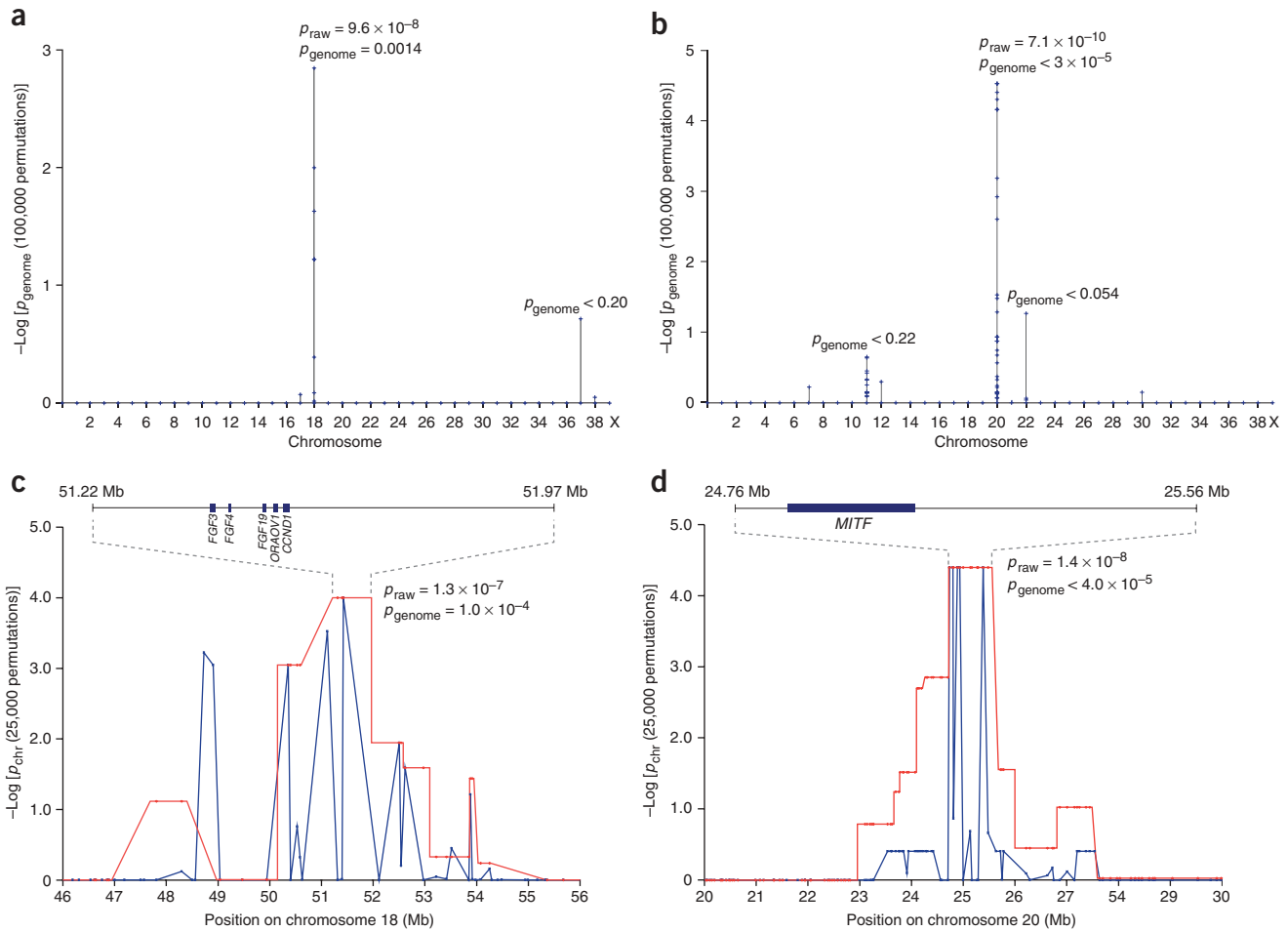
By genotyping ten white ( $s^w/s^w$ ; Supplementary Fig. 3a) and nine solid (*S/S*; Supplementary Fig. 3c) boxers, we mapped  $s^w$  to an associated region of less than 1 Mb containing only one gene: *microphthalmia-associated transcription factor* (*MITF*). The most strongly associated SNP ( $P_{\text{raw}} = 7.1 \times 10^{-10}$ ,  $P_{\text{genome}} = 3 \times 10^{-5}$ ) lies within a haplotype of 800-kb defined by 11 SNPs ( $P_{\text{raw}} = 1.4 \times 10^{-8}$ ,  $P_{\text{chr}} = 4.0 \times 10^{-5}$ ) that is homozygous in all white boxers and absent from solid dogs (Fig. 2b,d). The predominant haplotype in solid boxers has a frequency of 78%, and several minor haplotypes are also present. The sequenced boxer, with intermediate 'flash' pigmentation, is heterozygous for the white haplotype and the predominant solid haplotype. The association is 1,000-fold stronger than any other region in the genome.

*MITF* is an important developmental gene with a complex regulation implicated in pigmentary and auditory disorders in humans and mice<sup>19–21</sup>. *MITF* is thus an ideal candidate locus for  $s^w$ , which affects both pigmentation and hearing.

### Fine-mapping the coat color locus

To map the mutation more finely, we studied a second breed, bull terriers, in which  $s^w$  segregates (Supplementary Fig. 3d). We genotyped 127 dogs (23 solid, 13 flash and 25 white boxers, and 16 solid, 16 flash and 34 white bull terriers; Supplementary Table 1c) for 115 SNPs across 4.6 Mb, including 69 SNPs within the associated region of 800 kb ( $11 \pm 14$  kb average spacing) and 46 SNPs in 3.8 Mb of flanking sequence ( $86 \pm 58$  kb average spacing). In the white boxers, homozygosity extends for 736 kb; thus, the additional SNPs do not narrow the region. The genotypes of the white bull terriers, however, define two regions of homozygosity (43 kb and 203 kb) interrupted by a region of 30 kb that has three common haplotypes (frequencies 0.83, 0.10 and 0.05).

We first mapped the locus in boxers ( $\chi^2 = 92$ ) and bull terriers ( $\chi^2 = 104$ ) separately to confirm independent association and then combined the two data sets to identify a narrower region of strong association ( $\chi^2 = 194$ ; Fig. 3a). Haplotype analysis revealed a 102-kb region (24.847–24.949 Mb) that includes two distinct blocks with perfect genotype-phenotype correlation in both breeds: a block of seven SNPs (29–48 kb) at the melanocyte-specific promoter 1M and exons 2–6 and a block of six SNPs (87–95 kb) at exon 1B (Fig. 3b and Supplementary Fig. 4b online). In addition, a region downstream of exon 6 (5–19 kb) is identical in all dogs and thus cannot be excluded as a site of the  $s^w$  mutation. We note that a single isolated



**Figure 2** Genome-wide association mapping of two mendelian-inherited traits. **(a)** The recessive allele *ridgeless* in Rhodesian ridgebacks was mapped with 9 ridgeless and 12 ridged dogs. **(b)** The extreme white ( $s^w$ ) coat color allele was mapped with nine white boxers and ten solid boxers. For both traits, a single locus with strong genome-wide significance was identified. Significance of association was calculated with the software package PLINK over 100,000 permutations. **(c,d)** Significant association with long-range, breed-specific haplotypes is evident for the *ridgeless* phenotype **(c)**; 750 kb, three SNPs) and the white coat color **(d)**; 800 kb, 11 SNPs). Chromosome-wide association for SNPs (blue) and blocks (red) defined by the four-gamete rule was calculated by using Haploview<sup>33</sup> with 25,000 permutations.

SNP at 25.4 Mb shows association; this observation probably reflects coincidental allele sharing between distinct haplotypes (**Supplementary Figs. 3a and 4a**).

### Mutation screening of fine-mapped regions

To identify candidate mutations, we produced complete finished sequence from BAC clones representing the solid and white haplotypes. Across the associated 102-kb region, we identified 124 polymorphisms. Notably, all occur in noncoding sequence, implying that the  $s^w$  allele encodes a regulatory mutation. We examined these polymorphisms in a larger collection of white, solid and flash bull terriers and boxers and in control solid dogs of other breeds (**Supplementary Tables 1d and 3a** online). Of the 124 polymorphisms, 78 were not concordant with the coat color phenotype (the white allele either was not homozygous in white dogs or was present in solid dogs), leaving 46 candidates. Although any of these polymorphisms could represent the  $s^w$  mutation, we focused particularly on polymorphisms located in or near segments of genomic sequence showing strong cross-species conservation (see Methods). Only three polymorphisms fitted this description, and all are

located immediately upstream of the transcriptional start site of the melanocyte-specific ( $M$ ) promoter of *MITF*: a short interspersed nuclear element (SINE) insertion in the white haplotype (3-kb upstream), a length polymorphism in the  $M$  promoter (<100-bp upstream) and a single base change at an unconserved position close to conserved elements (~1,100-bp upstream). The  $M$  promoter of *MITF* is a critical regulator of melanocyte development, survival and migration<sup>22,23</sup>.

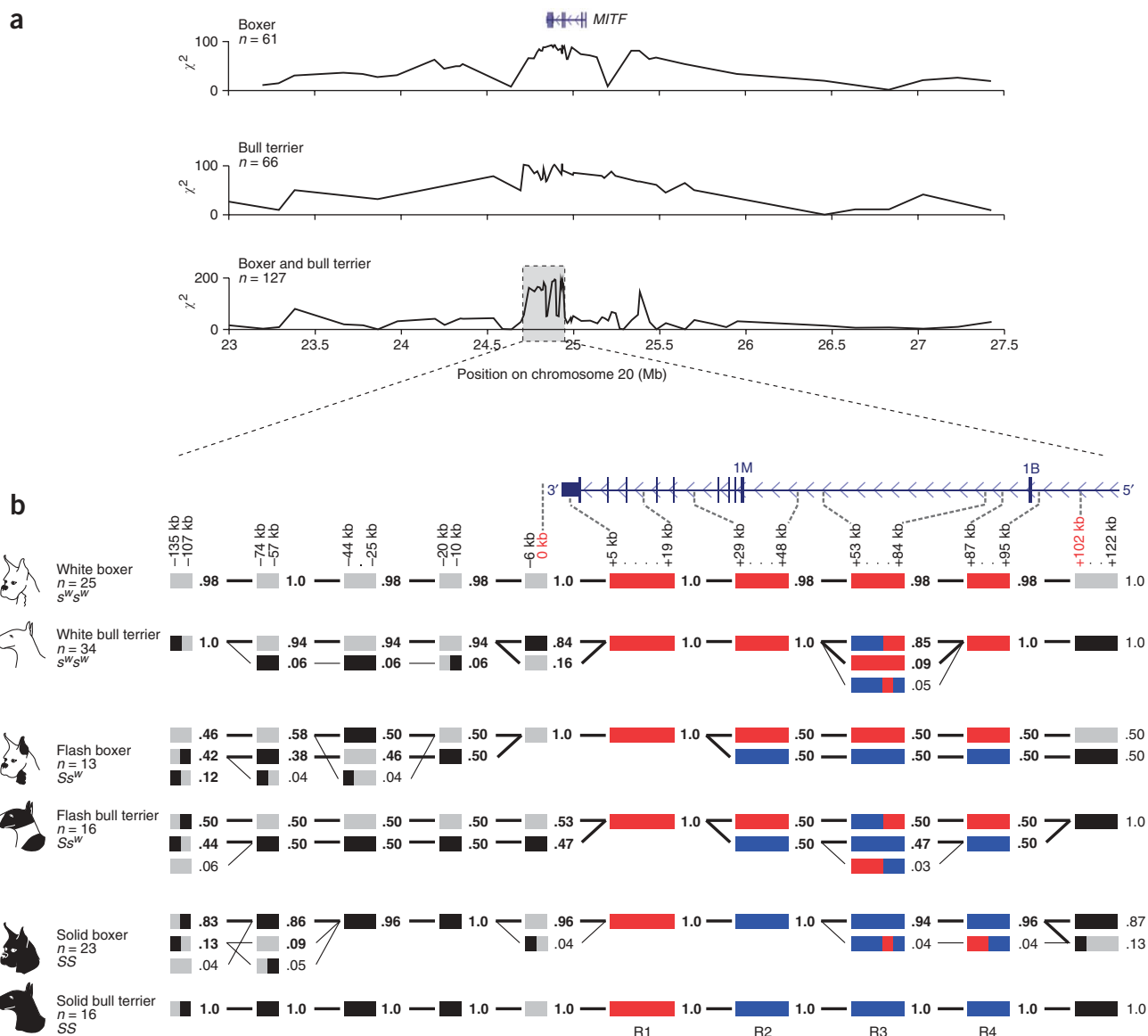
The SINE-Cf element is inserted 3,026-bp upstream of the transcriptional start site for the  $M$  transcript and 229-bp downstream of three clustered lymphoid-enhancing factor 1 (LEF1) binding motifs in ~20 bases of sequence unique to the dog genome (**Fig. 4a**). These LEF1 sites are located in sequence that is not present in human or mouse, but are probably functional because there are three additional LEF1 sites located closer to the  $M$  promoter (228 bp upstream) that are conserved across human, mouse and dog and have been shown to facilitate *MITF* self-activation in human cells<sup>24</sup>. All white boxers ( $n = 14$ ) and all white bull terriers ( $n = 13$ ) tested were homozygous for the SINE insertion, whereas the flash boxers ( $n = 20$ ) and flash bull terriers ( $n = 10$ ) were all heterozygous. None of the 80 solid dogs

tested, including boxers ( $n = 15$ ), bull terriers ( $n = 6$ ) and 59 dogs from 9 solid breeds, had the SINE element.

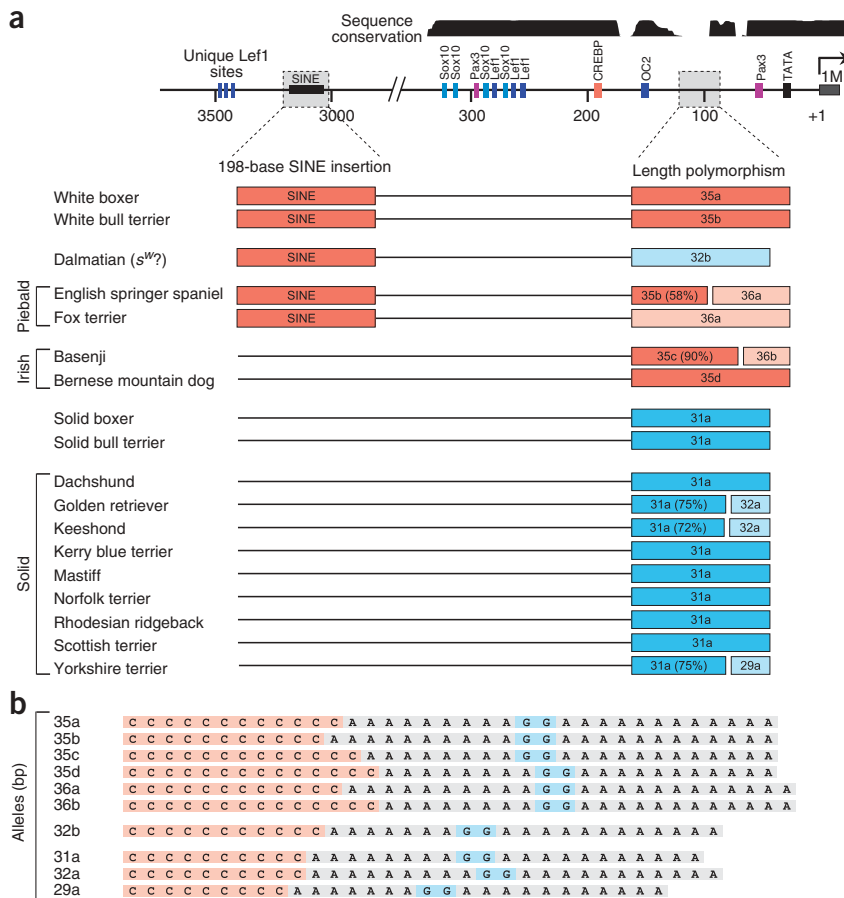
The second polymorphism is a set of short insertion-deletions (indels) located 60–95-bp upstream of the TATA box of the *M* promoter between the OC2- and PAX3-binding sites<sup>25,26</sup>, within a canine-specific 20-bp insertion (Fig. 4b). The flanking sequence is highly conserved across mammals. At this site, the white boxers ( $n = 10$ ) and bull terriers ( $n = 6$ ) tested had alleles of 35 bp, 4-bp longer than the allele in solid boxers ( $n = 4$ ) and bull terriers ( $n = 10$ ). The third polymorphism is a single base polymorphism

at a position that is variable among mammals and, thus, unlikely to affect function.

There is also a 12-bp deletion that is orthologous to exon B, a promoter used in a transcript of unknown function seen in humans and mice<sup>21,27</sup> (Supplementary Fig. 4d online). This deletion, however, is unlikely to be related to  $s^w$  for two reasons: transcript B seems to be specific to the Euarchontoglires clade (which includes human and mouse; Supplementary Fig. 4e), and the deletion does not correlate perfectly with the coat color phenotype (it was found in 4 of 23 Rhodesian ridgebacks screened; Supplementary Table 3b).



**Figure 3** Fine-mapping of coat color in boxers and bull terriers. **(a)** Broad association in boxers ( $\max \chi^2 = 92$ ) and bull terriers ( $\max \chi^2 = 104$ ) results in a smaller, highly associated region after combining the two breeds ( $\max \chi^2 = 194$ ). Coincidental allele sharing between the long, breed-specific white boxer and white bull terrier haplotypes produces an isolated single peak at 25.4 Mb, but the SNP shows only partial correlation with phenotype (Supplementary Fig. 4a). **(b)** The 102-kb region of association contains two blocks of perfect correlation of  $s^w$  to one haplotype (R2 and R4). The white boxer allele is shown in red, and the alternative allele, when present, in blue. Also in the 102-kb region are a block with no apparent polymorphism that cannot be definitively excluded (R1) and an intermediate, uncorrelated region that does not show perfect genotype-phenotype correlation and thus is unlikely to contain the causative mutation (R3). Outside the associated region, the two alleles for each SNP are shown in light and dark gray. The position of each SNP relative to the start of the 102-kb region is shown on top. Frequency is shown to the right of each haplotype, and common haplotypes (>5%) are in bold. Haplotypes were inferred with Haploview<sup>33</sup>. Dogs used for fine-mapping are listed in Supplementary Table 1c.



**Figure 4** Alleles by breed for the two candidate mutations. **(a)** Two candidate mutations are found within a region 3.5-kb upstream of the *M* promoter of the *MITF* gene. Solid dogs in all breeds lack the SINE insertion and have a short (29–32-bp) allele in the *M* promoter. White boxers and bull terriers and piebald ( $s^p$ ) breeds have both the SINE insertion and a longer promoter allele (35–36 bp), whereas Irish spotted ( $s^i$ ) dogs lack the SINE element but have a longer variant at the promoter. Dalmatians ( $s^{w?}$ ) carry the SINE element and a private short allele, suggesting a unique mutation. **(b)** Alleles observed for the length polymorphism in the *M* promoter of *MITF* contain a cytosine repeat (red) and two adenine repeats (grey) separated by two guanines (blue).

### Other alleles at the *S* locus

We also examined the two most likely candidate variants in 16 different breeds reported to have specific *S*-locus phenotypes (Fig. 4a). The breeds included three carrying white ( $s^w$ ) alleles, two fixed for piebald ( $s^p$ ) alleles, two fixed for Irish spotting ( $s^i$ ) alleles and nine fixed for solid (*S*) alleles. Pigmentation phenotypes in dogs range from solid to all white, and pigment disappears last from regions of highest embryonic melanoblast density<sup>28</sup>; this phenomenon is consistent with regulatory mutations that variably affect expression of *MITF* from the *M* promoter (*MITF-M*).

For both variants, the allele found in the white boxers and bull terriers was not seen in solid dogs. The SINE insertion was found in all white ( $s^w$ ) and piebald ( $s^p$ ) breeds, but not in the Irish spotting ( $s^i$ ) or solid (*S*) breeds. The length polymorphism is long (35–36 bp) in the white, piebald and Irish spotted breeds and short (29–32 bp) in the solid dogs. The sequence variability in the long variant (six alleles in six breeds) as compared with the short variant (four alleles in 12 breeds) might reflect reduced selective pressure on the mutated sequence or similar mutations arising many times. Dalmatians, which are reported to be white ( $s^w$ ) with black spots caused by a second locus<sup>16</sup>, are fixed for a private 32-bp allele.

### Selection at the coat color locus

In dog breeds that have been bred to fixation for one of the white spotting phenotypes, we would expect to see genetic evidence of strong recent selection in the form of extensive homozygosity around the *S* locus. To test this prediction, we genotyped the full set of 115 fine-mapping SNPs in Basenjis ( $s^i$ ), Bernese mountain dogs ( $s^i$ ), beagles ( $s^p$ ), English springer spaniels ( $s^p$ ) and Dalmatians. In two selected breeds (24 Basenjis and 25 Dalmatians), we indeed found extensive homozygosity of a single haplotype (660 kb and 560 kb, respectively). Several other breeds (21 beagles, four English springer spaniels and six Bernese mountain dogs) showed only short-range homozygosity (21 kb, 49 kb and 96 kb, respectively), comparable to that seen in the solid ridgebacks (54 kb). With the exception of beagles (a breed with very variable pigmentation<sup>16</sup>), the region of homozygosity in all of the breeds overlaps the *M* promoter and includes the two most likely candidate mutations, consistent with selection at this locus.

### DISCUSSION

The unique history of the domestic dog has produced over 400 genetically distinct breed populations and a genome structure particularly advantageous to gene mapping<sup>1</sup>. Here we have shown that genome-wide association mapping with only ~27,000 SNPs and ~20 dogs identifies a single discrete region of the genome for each of two recessive traits. The mapping is unambiguous: the genome-wide *P* values are 100-fold to 1,000-fold stronger for the associated regions than for any other region in the genome. In addition, the sample is only half as large as our original projection

of ~40 dogs<sup>3</sup>. In studies to be reported elsewhere, we have also mapped a dominantly inherited trait, primary hyperparathyroidism in Keeshonden, with only ~30 affected and ~40 control dogs, as predicted. If our estimates continue to hold true, it should be possible to map risk factors for genes that confer a 3–5-fold increase in risk for a trait with only 100–300 affected and 100–300 control dogs. We consider that this strategy has strong potential for the mapping of complex traits.

Our results have important implications for the design of genetic mapping studies in dog. First, genotype data for 13 diverse breeds clearly show that LD is bimodal: within breeds it extends over long distances owing to recent breed-creation bottlenecks, but across breeds it drops off more rapidly than in human populations. This finding confirms observations based on a few genomic regions<sup>1,2</sup>. Although the precise extent of LD varies on the basis of breed history, average LD extends > 5 Mb in all breeds studied. Genome-wide LD mapping should thus be effective in all breeds.

Second, for genome-wide LD mapping, it is most effective to study unrelated affected and control dogs within a breed. By contrast, family-based linkage designs will yield much larger linked regions owing to limited recombination within a pedigree. With unrelated

dogs, associated regions will then reflect the haplotype block size in dog breeds, ~0.5–1 Mb, and should be small enough for efficient fine-mapping.

Third, dog breeds, despite their recent common origins, are very distinct populations. The analysis of population differentiation, calculated as the genome-wide  $F_{ST}$  value between populations, suggests that typical breeds are 2–3 times as diverged as human population groups. Therefore, it is not advisable to combine multiple breeds for genome-wide association analysis. In addition,  $F_{ST}$  values show that American and European golden retrievers are roughly as diverged as European and Asian human populations, suggesting that affected and control dogs should be geographically matched to minimize population stratification.

Fourth, after initial LD mapping, it should be possible to perform fine-mapping across multiple dog breeds to obtain a smaller associated region of 100 kb or less that reflects the ancestral haplotype block size before breed creation. In boxers and bull terriers, two closely related breeds, white dogs share a 34-kb region containing the candidate mutations. The dorsal ridge mutation described in a companion paper<sup>1</sup> is shared between two seemingly unrelated breeds. Given the recent origins of breeds and the reported high degree of ancestral haplotype sharing<sup>1,3</sup>, many disease-causing mutations are likely to be carried on ancestral haplotypes of 10–100 kb that are shared between breeds. Using multiple breeds to define precisely the associated haplotype will limit the number of candidate mutations, a particularly important step for identifying regulatory mutations where ascribing function is more difficult and time consuming.

Last, our canine SNP array has sufficient marker density to identify a block of association of 0.5–1 Mb and shows similar polymorphism frequencies across the breeds tested. It should thus be useful for dog genetic studies in general.

Our results also suggest that the genetic analysis may help to pinpoint genes that underwent strong selection during the creation of dog breeds. Specific genetic variants under strong selection should lie within large blocks that are homozygous within the breed. The *MITF* locus provides a good example: in certain breeds bred for coat color (such as Dalmatian and Basenji), the locus shows extensive homozygosity (>0.5 Mb), consistent with a single fixed haplotype that underwent recent strong selection. Although extensive blocks of homozygosity may provide clues to loci that have undergone strong selection in breeds, interpreting such data will require careful characterization of the background noise caused by random drift. Within a typical breed, there are ~160 homozygous regions of >0.5 Mb, corresponding to ~6% of the genome (Table 1), many of which are probably due to random drift. By looking for overlapping regions of homozygosity in multiple breeds that share the same phenotype, it may be possible to decrease the noise and to identify selected loci accurately. Extensive homozygosity, however, may not always mark selected loci. Some breeds clearly under selection for white spotting phenotypes, such as the Bernese mountain dog, show only short-range homozygosity at *MITF* (although they have consistent genotypes at the two candidate variants; Fig. 4a).

Beyond the general lessons for genetic mapping in dogs, the specific results concerning the coat color and ridge phenotypes have interesting implications. Neither is caused by a mutation in protein-coding sequence: white coat color phenotype in boxers and bull terriers is due to variation in the *M* promoter of the *MITF* gene, whereas the ridge phenotype in Rhodesian ridgebacks is due to a genomic duplication of several FGF genes. We suspect that the creation of dog breeds will often have involved selection for subtle mutations affecting the level, timing or tissue-specific expression of key developmental genes.

Indeed, *Mitf*-null mutations in mouse cause severe phenotypes, including extensive depigmentation, hearing loss, and acute eye and bone disorders. The closest mouse model of the dog phenotype is the less severe black-eyed white *Mitf*<sup>mi-bw</sup> allele, which has an L1 insertion in intron 3 that abolishes *Mitf-M* expression and reduces expression of *Mitf-H* and *Mitf-A*. This mutation prevents melanocyte formation, making the mice both white and universally deaf<sup>29,30</sup>. The *s<sup>w</sup>* allele in boxers and bull terriers confers an even milder phenotype: only ~2% of white dogs have bilateral deafness<sup>31</sup>, suggesting that *MITF-M* expression sufficient for limited melanocyte migration persists in most dogs<sup>19,22</sup>. In addition, any patches of color have normal pigmentation, indicating that *MITF-M* is expressed in mature melanocytes<sup>19,22</sup>. Detailed studies of the *M* promoter of *MITF* will be required to understand the precise effects on gene regulation.

Regulatory mutations that disrupt the expression of *MITF-M* during crucial developmental time points would explain not only the white coat phenotype, but also other *S*-locus alleles. White spotting phenotypes in dogs span a continuum from full pigmentation to all white. As the proportion of white increases, pigmentation disappears last from regions of highest embryonic melanoblast density<sup>28</sup>, consistent with disruption of the *M* promoter, a regulator of melanocyte development, survival and migration. We propose that, for each white spotting allele, the combination of *MITF-M* regulatory mutations defines the extent of pigmentation. These mutations potentially include the SINE and length polymorphism identified, in addition to others absent from the boxer breed (which carries only the *S* and *s<sup>w</sup>* alleles). Spots in Dalmatians appear after birth and may result from a later round of melanoblast proliferation<sup>32</sup>.

Our work suggests that dog genetics will prove to be a powerful tool for elucidating mammalian genome function, including genetic factors underlying disease. Because dogs and humans have very similar gene repertoires and share much of their environment, it is likely that many of the same pathways will be involved in related traits and diseases. Our results clearly show that genetic association studies within breeds will facilitate identification of genes responsible for mendelian traits. The challenge ahead will be to extend this methodology to complex traits with direct relevance for human medicine.

## METHODS

**SNP array development and data sets.** To achieve fairly uniform genome coverage and utility in many breeds, we selected 64,039 SNPs from non-overlapping 25-kb bins in which SNPs located within *Sty1* fragments of 300–800 bp had been ranked on the basis of their location within the fragment, repetitiveness of sequence and the breed source. A 5- $\mu$ m array was generated by Affymetrix. Genome-wide genotype data from the canine Affymetrix GeneChip array were generated with the human 500K array protocol but with a smaller hybridization volume of 125  $\mu$ l owing to the smaller surface area of the canine array. Probe intensity data were processed by the Affymetrix BRLMM (Bayesian Robust Linear Model with Mahalanobis distance classifier) genotype calling method. A set of 26,625 high-performing SNPs ('27K set') that performed consistently well in the initial test of 92 arrays (at  $P < 0.25$ , the call rate was >90% and the heterozygous call rate was 2–80%) was selected for all further analysis. For detailed information on the arrays see <http://www.broad.mit.edu/mammals/dog/caninearray/>.

**Genome structure in breeds.** Using Haploview<sup>33</sup>, we calculated  $r^2$  versus distance for all SNPs with MAF > 5% and call rate > 75% and measured haplotype block size by using the four-gamete rule with a fourth haplotype frequency cutoff of 0.1. We excluded arrays with call rate < 70%. We assessed stratification between populations with the principal components analysis implemented in the software Eigensoft<sup>7,34</sup>. We measured population differentiation by using an  $F_{ST}$  estimator across the 27K set of array SNPs (see **Supplementary Methods** online for details), and subsequently calculated

the phylogenetic tree by using the Fitch-Margoliash method in PHYLIP<sup>35</sup>. Sample numbers are summarized in **Supplementary Table 1a**.

**Genome-wide association.** For genome-wide mapping, we performed a case-control association analysis on all SNPs with MAF > 0.05 and call rate > 75% by using the software package PLINK. We excluded arrays with call rate < 70%. We ascertained genome-wide significance through phenotype permutation testing ( $n = 100,000$ ). The most associated haplotype was identified with Haploview; blocks were defined by the four-gamete rule and chromosome-wide significance was calculated by permutation testing ( $n = 25,000$ ) for SNPs with MAF > 0.05 and call rate > 75%. Sample numbers are summarized in **Supplementary Table 1b**.

**Fine-mapping.** For fine-mapping and array validation, we generated SNP genotypes using the SEQUENOM MassARRAY platform. Using PLINK, we calculated SNP association for all SNPs with MAF > 0.1, call rate > 75% and good functionality (all three genotypes observed in a breed). We manually defined haplotype block boundaries at positions where genotypes provided evidence of a historical recombination and then measured haplotype frequencies in those blocks with Haploview. Sample numbers are summarized in **Supplementary Table 1c**.

**Identifying the candidate mutations for  $s^w$ .** We generated finished sequence data for one BAC from each chromosome of the sequenced boxer genome, identified by genotyping five SNPs known to differ between the two haplotypes. Using the program diffseq, we identified all 124 polymorphisms between the two BAC sequences in the 102-kb associated region. To identify candidate mutations, we resequenced boxers, bull terriers and solid dogs from multiple breeds and identified the 46 polymorphisms that showed complete correlation with phenotype. Out of these 46 variants, we identified three mutations that seemed most likely to be functional on the basis of cross-species conservation. We analyzed four species Dog/Human/Mouse/Rat Multiz conservation scores downloaded from the University California Santa Cruz (UCSC) dog genome browser<sup>36</sup>. For any region that aligned with the human genome, we also considered the 17-species alignments currently in the UCSC human genome browser. The 43 other polymorphisms that were considered less likely to be functional fell into three groups: 36 short polymorphisms (SNPs or 1-bp indels) in unconserved sequence (none had a conservation score of >0.4 within 5 bases or >0.75 within 50 bp); five longer indels (2–8 bp) occurring in unconserved, repetitive sequence (as annotated by RepeatMasker); and two polymorphisms (an SNP and a 5-bp indel) for which the white allele was the ancestral variant on the basis of 11 mammals in the USCS human genome browser. Sample numbers are summarized in **Supplementary Table 1d**, and the 124 polymorphisms are described in **Supplementary Table 3a**. The indel in exon B was assessed in a larger number of dogs ( $n = 115$ ) by fragment analysis, and the SINE insertion upstream of the *M* promoter was assessed by PCR, followed by size separation on an agarose gel.

**URLs.** Information on the CanFam2.0 genome is available at <http://www.genome.ucsc.edu>. diffseq, <http://bioweb.pasteur.fr/docs/EMBOSS/diffseq.html>; PLINK, <http://pngu.mgh.harvard.edu/~purcell/plink/>

Note: Supplementary information is available on the Nature Genetics website.

#### ACKNOWLEDGMENTS

We thank the Genetic Analysis Platform at the Broad Institute of MIT and Harvard for performing the SNP array genotyping, and L. Gaffney for assistance with figures. The work was supported by the AKC/Canine Health Foundation (grant 373), the Foundation for Strategic Research, and the Donald and Jo Ann Petersen Endowed Research Fund of the University of Michigan Comprehensive Cancer Center.

Published online at <http://www.nature.com/naturegenetics>

Reprints and permissions information is available online at <http://npg.nature.com/reprintsandpermissions>

- Lindblad-Toh, K. *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803–819 (2005).
- Sutter, N.B. *et al.* Extensive and breed-specific linkage disequilibrium in *Canis familiaris*. *Genome Res.* **14**, 2388–2396 (2004).

- Wade, C.M., Karlsson, E.K., Mikkelsen, T.S., Zody, M.C. & Lindblad-Toh, K. The dog genome: sequence, evolution and haplotype structure. In *The Dog and Its Genome* (eds. Ostrander, E.A., Giger, U. & Lindblad-Toh, K.) 179–207 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 2006).
- Hartl, D.L. & Clark, A.G. *Principles of Population Genetics* (Sinauer Associates, Sunderland, MA, 2007).
- Keinan, A., Mullikin, J.C., Patterson, N. & Reich, D. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nat. Genet.* **39**, 1251–1255 (2007).
- Parker, H.G. *et al.* Genetic structure of the purebred domestic dog. *Science* **304**, 1160–1164 (2004).
- Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
- Hillbertz, N.H. & Andersson, G. Autosomal dominant mutation causing the dorsal ridge predisposes for dermoid sinus in Rhodesian ridgeback dogs. *J. Small Anim. Pract.* **47**, 184–188 (2006).
- Copp, A.J., Greene, N.D. & Murdoch, J.N. The genetic basis of mammalian neurulation. *Nat. Rev. Genet.* **4**, 784–793 (2003).
- Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
- Karabagli, H., Karabagli, P., Ladher, R.K. & Schoenwolf, G.C. Comparison of the expression patterns of several fibroblast growth factors during chick gastrulation and neurulation. *Anat. Embryol. (Berl.)* **205**, 365–370 (2002).
- Ladher, R.K., Wright, T.J., Moon, A.M., Mansour, S.L. & Schoenwolf, G.C. FGF8 initiates inner ear induction in chick and mouse. *Genes Dev.* **19**, 603–613 (2005).
- Salmon Hillbertz, N.H.C. *et al.* Duplication of *FGF3*, *FGF4*, *FGF19* and *ORAOV1* causes hair ridge and predisposition to dermoid sinus in Ridgeback dogs. *Nat. Genet.*, advance online publication 30 September 2007 (doi:10.1038/ng.2007.4).
- Dourmishev, A.L., Dourmishev, L.A., Schwartz, R.A. & Janniger, C.K. Waardenburg syndrome. *Int. J. Dermatol.* **38**, 656–663 (1999).
- Tietz, W. A syndrome of deaf-mutism associated with albinism showing dominant autosomal inheritance. *Am. J. Hum. Genet.* **15**, 259–264 (1963).
- Little, C.C. *The Inheritance of Coat Color in Dogs* (Comstock Publishing Associates, Ithaca, NY, 1957).
- Metallinos, D. & Rine, J. Exclusion of *EDNRB* and *KIT* as the basis for white spotting in Border Collies. *Genome Biol.* **1** research0004.1–research0004.4 (2000).
- van Hagen, M.A. *et al.* Analysis of the inheritance of white spotting and the evaluation of *KIT* and *EDNRB* as spotting loci in Dutch boxer dogs. *J. Hered.* **95**, 526–531 (2004).
- Smith, S.D., Kelley, P.M., Kenyon, J.B. & Hoover, D. Tietz syndrome (hypopigmentation/deafness) caused by mutation of *MITF*. *J. Med. Genet.* **37**, 446–448 (2000).
- Tassabehji, M., Newton, V.E. & Read, A.P. Waardenburg syndrome type 2 caused by mutations in the human microphthalmia (*MITF*) gene. *Nat. Genet.* **8**, 251–255 (1994).
- Steingrimsson, E., Copeland, N.G. & Jenkins, N.A. Melanocytes and the microphthalmia transcription factor network. *Annu. Rev. Genet.* **38**, 365–411 (2004).
- Widlund, H.R. & Fisher, D.E. Microphthalmia-associated transcription factor: a critical regulator of pigment cell development and survival. *Oncogene* **22**, 3035–3041 (2003).
- Levy, C., Khaled, M. & Fisher, D.E. MITF: master regulator of melanocyte development and melanoma oncogene. *Trends Mol. Med.* **12**, 406–414 (2006).
- Saito, H. *et al.* Melanocyte-specific microphthalmia-associated transcription factor isoform activates its own gene promoter through physical interaction with lymphoid-enhancing factor 1. *J. Biol. Chem.* **277**, 28787–28794 (2002).
- Jacquemin, P. *et al.* The transcription factor octnuc-2 controls the microphthalmia-associated transcription factor gene. *Biochem. Biophys. Res. Commun.* **285**, 1200–1205 (2001).
- Bondurand, N. *et al.* Interaction among *SOX10*, *PAX3* and *MITF*, three genes altered in Waardenburg syndrome. *Hum. Mol. Genet.* **9**, 1907–1917 (2000).
- Udono, T. *et al.* Structural organization of the human microphthalmia-associated transcription factor gene containing four alternative promoters. *Biochim. Biophys. Acta* **1491**, 205–219 (2000).
- Burns, M. & Fraser, M.N. *Genetics of the Dog: the Basis of Successful Breeding* (Oliver & Boyd, Edinburgh, London, 1966).
- Motohashi, H., Hozawa, K., Oshima, T., Takeuchi, T. & Takasaka, T. Dysgenesis of melanocytes and cochlear dysfunction in mutant microphthalmia (*mi*) mice. *Hear. Res.* **80**, 10–20 (1994).
- Yoshida, H., Kunisada, T., Kusakabe, M., Nishikawa, S. & Nishikawa, S.I. Distinct stages of melanocyte differentiation revealed by analysis of nonuniform pigmentation patterns. *Development* **122**, 1207–1214 (1996).
- Strain, G.M. Deafness prevalence and pigmentation and gender associations in dog breeds at risk. *Vet. J.* **167**, 23–32 (2004).
- Jordan, S.A. & Jackson, I.J. A late wave of melanoblast differentiation and rostrocaudal migration revealed in patch and rump-white embryos. *Mech. Dev.* **92**, 135–143 (2000).
- Barrett, J.C., Fry, B., Maller, J. & Daly, M.J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).
- Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
- Felsenstein, J. PHYLIP, phylogeny inference package (version 3.2). *Cladistics* **5**, 164–166 (1989).
- Karolchik, D. *et al.* The UCSC Genome Browser Database. *Nucleic Acids Res.* **31**, 51–54 (2003).