8. G. W. Mayr, T. Radenberg, U. Thiel, G. Vogel, L. R. Stephens, *Carbohydr. Res.* **234**, 247 (1992).
9. R. F. Irvine, M. J. Schell, *Nat. Rev. Mol. Cell Biol.* **2**, 327 (2001).
10. H. Streb, R. F. Irvine, M. J. Berridge, I. Schulz, *Nature* **306**, 67 (1983).
11. H. R. Luo *et al.*, *Biochemistry* **41**, 2509 (2002).
12. E. Dubois *et al.*, *J. Biol. Chem.* **277**, 23755 (2002).
13. W. Ye, N. Ali, M. E. Bembenek, S. B. Shears, E. M. Lafer, *J. Biol. Chem.* **270**, 1564 (1995).
14. B. Fleischer *et al.*, *J. Biol. Chem.* **269**, 17826 (1994).
15. A. Saiardi, A. C. Resnick, A. M. Snowman, B. Wendland, S. H. Snyder, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 1911 (2005).
16. S. J. York, B. N. Armbruster, P. Greenwell, T. D. Petes, J. D. York, *J. Biol. Chem.* **280**, 4264 (2005).
17. H. R. Luo *et al.*, *Cell* **114**, 559 (2003).
18. S. Huang, D. A. Jeffery, M. D. Anthony, E. K. O'Shea, *Mol. Cell. Biol.* **21**, 6695 (2001).
19. A. Saiardi, H. Erdjument-Bromage, A. M. Snowman, P. Tempst, S. H. Snyder, *Curr. Biol.* **9**, 1323 (1999).
20. A. M. Seeds, R. J. Bastidas, J. D. York, *J. Biol. Chem.* **280**, 27654 (2005).
21. S. Mulugu *et al.*, *Science* **316**, 106 (2007).
22. C. Auesukaree, H. Tochio, M. Shirakawa, Y. Kaneko, S. Harashima, *J. Biol. Chem.* **280**, 25127 (2005).
23. S. Huang, E. K. O'Shea, *Genetics* **169**, 1859 (2005).
24. J. S. Flick, J. Thorner, *Genetics* **148**, 33 (1998).
25. A. Saiardi, R. Bhandari, A. C. Resnick, A. M. Snowman, S. H. Snyder, *Science* **306**, 2101 (2004).
26. M. J. Schell *et al.*, *FEBS Lett.* **461**, 169 (1999).
27. We thank Y. Liu for preparation of strains; P. Fridy for reagents; B. Stern, H. Kim, C. Leimkuhler, and D. Schwarz for comments on the manuscript; members of the J.D.Y. laboratory for unpublished data, helpful discussions, and comments on the manuscript; and D. Kahne and his laboratory members for access to equipment. This work was supported by NIH R01 GM051377 (E.K.O.), DK070272 (J.D.Y.), and HL055672 (J.D.Y.); the David and Lucile Packard Foundation (E.K.O.); and the Howard Hughes Medical Institute (E.K.O. and J.D.Y.). The authors have no conflicting financial interest.

Supporting Online Material
www.sciencemag.org/cgi/content/full/316/5821/109/DC1
Materials and Methods
SOM Text
Figs. S1 to S5
Tables S1 and S2
References

19 December 2006; accepted 22 February 2007
10.1126/science.1139080

A Single *IGF1* Allele Is a Major Determinant of Small Size in Dogs

Nathan B. Sutter,[1] Carlos D. Bustamante,[2] Kevin Chase,[3] Melissa M. Gray,[4] Keyan Zhao,[5] Lan Zhu,[2] Badri Padhukasahasram,[2] Eric Karlins,[1] Sean Davis,[1] Paul G. Jones,[6] Pascale Quignon,[1] Gary S. Johnson,[7] Heidi G. Parker,[1] Neale Fretwell,[6] Dana S. Mosher,[1] Dennis F. Lawler,[8] Ebenezer Satyaraj,[8] Magnus Nordborg,[5] K. Gordon Lark,[3] Robert K. Wayne,[4] Elaine A. Ostrander[1]*

The domestic dog exhibits greater diversity in body size than any other terrestrial vertebrate. We used a strategy that exploits the breed structure of dogs to investigate the genetic basis of size. First, through a genome-wide scan, we identified a major quantitative trait locus (QTL) on chromosome 15 influencing size variation within a single breed. Second, we examined genetic variation in the 15-megabase interval surrounding the QTL in small and giant breeds and found marked evidence for a selective sweep spanning a single gene (*IGF1*), encoding insulin-like growth factor 1. A single *IGF1* single-nucleotide polymorphism haplotype is common to all small breeds and nearly absent from giant breeds, suggesting that the same causal sequence variant is a major contributor to body size in all small dogs.

Size variation in the domestic dog is extreme and surpasses that of all other living and extinct species in the dog family, Canidae (*1*, *2*). However, the genetic origin of this diversity is obscure. Explanations include increased recombination or mutation rates (*3*, *4*), a unique role of short repeat loci near genes (*3*), expansion of specific short interspersed nuclear elements (*5*), regulatory gene variation (*6*, *7*), or a readily altered developmental program (*1*, *6*). The domestic dog descended from the gray wolf at least 15,000 years ago (*8*–*10*), but the vast majority of dog breeds originated over the past few hundred years (*11*). Understanding the genetic basis for the rapid generation of extreme size variability in the dog would provide critical tests of alternative genetic mechanisms and insight into how evolutionary diversification in size could occur rapidly during adaptive radiations (*12*).

To investigate the genetic basis for size variation in dogs and understand how change in size might occur rapidly in dogs and other canids, we first initiated sequence-based marker discovery across a 15–megabase (Mb) interval on chromosome 15 in the Portuguese water dog (PWD), a breed that is allowed large variation in skeletal size by the American Kennel Club (*13*). Previously, based on 92 radiographic skeletal measurements for size and shape, we found that two QTL (FH2017 at 37.9 Mb and FH2295 at 43.5 Mb) within this region were strongly associated with body size in 463 PWDs from a well-characterized extended pedigree (*13*, *14*). We discovered 302 single-nucleotide polymorphisms (SNPs) and 34 insertion/deletion polymorphisms by sequencing 338 polymerase chain reaction (PCR) amplicons in four large and four small PWDs and in nine dogs from small and giant breeds (<9 and >30 kg average breed mass, respectively). We then measured the association between 116 SNPs and skeletal size in a sample of 463 PWDs and identified a single peak within 300 kb of the insulin-like growth factor 1 gene (*IGF1*) (Fig. 1A), confirming the FH2295 QTL. *IGF1* is an excellent candidate gene known to influence body size in both mice and humans (*15*–*17*).

Haplotype analysis of 20 SNPs spanning *IGF1* further supported a role for the locus in determining body size. We observed that 889 of the 926 (96%) PWD chromosomes carry one of just two haplotypes, termed B and I. Dogs homozygous for haplotype B have a smaller median skeletal size [Fig. 1B; $P < 3.27 \times 10^{-7}$, analysis of variance (ANOVA)] and mass (fig. S1) than dogs homozygous for I and a lower level of IGF1 protein in blood serum (Fig. 1C; $P < 9.34 \times 10^{-4}$, ANOVA). In PWDs, 15% of the variance in skeletal size is explained by the *IGF1* haplotype. Linkage disequilibrium around *IGF1* in PWDs is too extensive to allow fine mapping, presumably because of the breed's recent origin and small population size (*18*, *19*). However, if a mutation at *IGF1* in general underlies genetic differences in size among dog breeds, comparison of breeds of different sizes that have distinct genealogical histories may allow fine mapping of the mutation. Moreover, because size has been the target of strong selection by dog breeders, we would expect to find a signature of selection surrounding the QTL in breeds of extreme small or giant size.

To test these predictions, we surveyed genetic variation for the same 116 SNPs in 526 dogs from 23 small (<9 kg) and 20 giant (>30 kg) breeds. To obtain an empirical distribution of our association mapping test statistics, we also surveyed variation in 83 SNPs with no known association to body size on canine chromosomes 1, 2, 3, 34, and 37. These data were analyzed first to determine if intense artificial selection on body size has resulted in a "selective sweep" (*20*), reducing variability and increasing allele frequency divergence near *IGF1*. We found a marked reduction in marker heterozygosity and

[1]National Human Genome Research Institute, Building 50, Room 5349, 50 South Drive MSC 8000, Bethesda, MD 20892–8000, USA. [2]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14850, USA. [3]Department of Biology, University of Utah, Salt Lake City, UT 84112, USA. [4]Department of Ecology and Environmental Biology, University of California, Los Angeles, CA 90095, USA. [5]Department of Molecular and Computational Biology, University of Southern California, Los Angeles, CA 90089, USA. [6]The WALTHAM Centre for Pet Nutrition, Waltham on the Wolds, Leicestershire, LE14 4RT, UK. [7]Department of Veterinary Pathobiology, University of Missouri, Columbia, MO 65211, USA. [8]Nestle Research Center (NRC-STL), St. Louis, MO 63164, USA.

*To whom correspondence should be addressed. E-mail: eostrand@mail.nih.gov

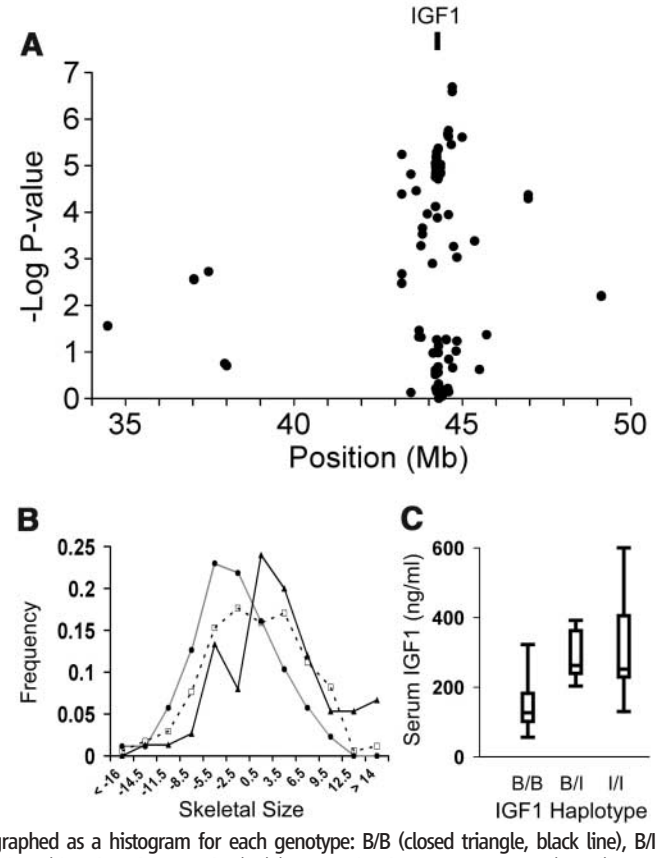


Fig. 1. Relationships of skeletal size, SNP markers, *IGF1* haplotype, and serum levels of the IGF1 protein in PWDs. (**A**) A mixed-model test for association between size and genotype. The association of three genotype categories (A_1A_1, A_1A_2, and A_2A_2) with skeletal size measurements was calculated with the use of all pairwise coefficients of consanguinity for 376 dogs. Each point represents a single SNP position on canine chromosome 15 and negative log *P* value for the association statistic. (**B**) PWD *IGF1* haplotypes and mean skeletal size. Haplotypes were inferred for 20 markers spanning the *IGF1* gene (chromosome 15: 44,212,792 to 44,278,140, CanFam1). Out of the 720 chromosomes with successful inference, 96% carry one of just two haplotypes, B and I, identical to haplotypes inferred for small and giant dogs, respectively (Fig. 3). Data are graphed as a histogram for each genotype: B/B (closed triangle, black line), B/I (open square, dashed line), and I/I (closed circle, gray line). (**C**) Serum levels of IGF1 protein (ng/ml) as a function of haplotype. Serum levels of IGF1 protein were assayed in 31 PWDs carrying haplotypes B and I. Box plots show the median (center line in box), first and third quartile (box ends), and maximum and minimum values (whiskers) obtained for each category: homozygous B/B (n = 15), heterozygous B/I (n = 7), and homozygous I/I (n = 9).

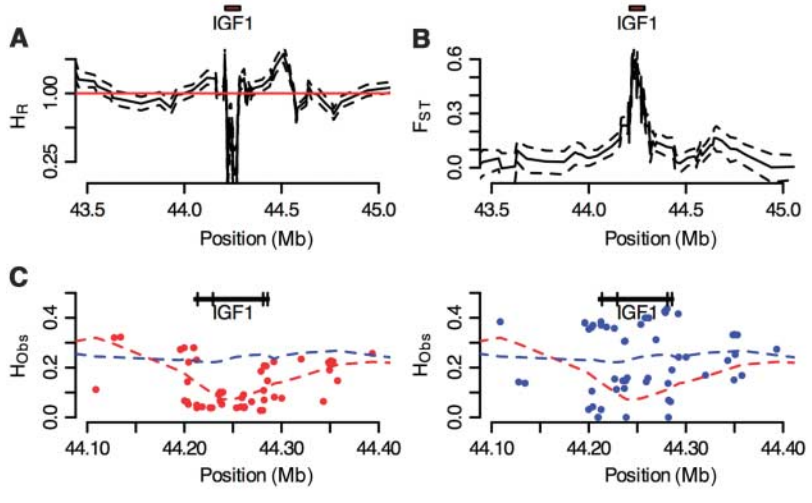


Fig. 2. Signatures of recent selection on the *IGF1* locus across 22 small and giant dog breeds. (**A**) Heterozygosity ratio (H_R) for small versus giant dogs. (**B**) Genetic differentiation (F_{ST}) for small versus giant dogs. For both (A) and (B), a sliding 10-SNP window across *IGF1* was used. Dashed lines delimit the 95% confidence intervals based on nonparametric bootstrap resampling. The *IGF1* gene interval is indicated above the graphs as a red box drawn to scale. (**C**) Observed heterozygosity (H_{Obs}) of SNPs near *IGF1* typed in small breeds (<9 kg) and giant breeds (>30 kg). Small breeds have a reduction in observed heterozygosity compared with that of giant breeds. Red and blue points are average observed heterozygosity in small and giant breeds, respectively. Dashed lines are locally weighted scatterplot smoothing (LOWESS) best fit to the data. The *IGF1* gene is shown as a black bar with exons indicated by vertical lines.

increased genetic differentiation between small and giant dogs centered on *IGF1* (Fig. 2). Specifically, near *IGF1*, average heterozygosity in small dogs is only 25% of that in large dogs, genetic differentiation (F_{ST}, where ST represents subpopulation) peaks significantly at 0.6, and overall heterozygosity is sharply reduced (Fig. 2B) (figs. S2 to S5). Together, these results suggest that a narrow and precisely defined genomic region holds the variant (or variants) responsible for small size in a disparate set of small dog breeds.

We next tested for association between each SNP and average breed size (Fig. 3A). The null hypothesis of no association between body size and marker frequency across breeds is rejected (Bonferroni-correct P value < 0.05) for 25 contiguous SNPs defining an 84-kb interval spanning the same region that shows evidence of a selective sweep (chromosome 15 base pairs 44,199,850 to 44,284,186) (Figs. 2 and 3A). The Mann-Whitney U statistic provides a uniform distribution of P values for 83 genomic control markers (fig. S6). Similarly, P values from Fisher's exact test of association across individuals were smaller than 10^{-100} in the 84-kb interval; although these P values are clearly biased by confounding population structure (fig. S6), as evidenced by the 83 genomic control markers [for which the minimum P value was 10^{-20} (fig. S7)], the result is significant.

Analysis of specific breed haplotypes shows that a 20-SNP haplotype spanning *IGF1* is shared by all 14 sampled small dog breeds (Fig. 3, B and C) and is identical to haplotype B in small PWDs. This haplotype was observed in only three of the nine giant breeds because most giant dogs carry one or both of two distinct haplotypes: F and I. SNP 5, located at base pair position 44,228,468 (Fig. 3B), is the best candidate for being proximate to the causative mutation for the following reasons: (i) It distinguishes haplotypes A, B, and C, associated with small body size, from haplotypes D to L, which are common in large breeds; (ii) an ancestral recombination graph suggests an absence of recombination between SNPs 4 and 5 (fig. S8); and (iii) marker analysis in the golden jackal and gray wolf indicates that the SNP 5 A allele of small breeds is the derived condition (fig. S9) (table S1). To further assess the association between body size and the SNP 5 A allele, we genotyped six tagging SNPs that distinguish all major *IGF1* haplotypes in a set of 3241 dogs from 143 breeds (Fig. 4) (table S2). The frequency of the SNP 5 A allele is strongly negatively correlated with breed average mass across this large sample of breeds (Fig. 4, Spearman's rank correlation coefficient $\rho = -0.773$; $P < 2.2 \times 10^{-16}$; likelihood ratio test = 2882.3, $\chi^2_{df=1} < 2 \times 10^{-16}$, logistic regression of allele frequency on body size). A strong negative correlation remains when the 22 breeds used to discover SNP 5 are removed from the analysis ($\rho = -0.729$; $P < 2.2 \times 10^{-16}$,

Spearman's rank correlation). Exceptions, such as the large Rottweiler or small whippet breeds, may carry compensatory mutations at other size QTL or recombinants that could aid fine mapping at *IGF1*. Our results show that a single *IGF1* haplotype is common to a large sample of small dogs and strongly imply that the same causal variant (or variants) is a major influence on the phenotype of diminished body size.

The *IGF1* gene is a strong genetic determinant of body size across mammals; mice genetically deficient in IGF1 are just 60% normal birth weight (*15*), and a human with a homozygous partial deletion of the gene was born 3.9 SD below normal length (*16*, *17*). IGF1 binds the type 1 IGF receptor, a tyrosine kinase signal transducer. This interaction promotes cell growth and organismal longevity (*21*) and induces cellular differentiation (*22*). Serum levels of IGF1 protein (*23*) have been found to correlate with body size in toy, miniature, and standard poodles (*24*). These studies did not compare *IGF1* genetic variation with differences in serum IGF1 protein concentrations; we observed that PWDs carrying the B haplotype of the *IGF1* gene have significantly lower serum levels of IGF1 (Fig. 1C).

Finally, to identify possible causative variants, we sequenced the exons of *IGF1* in a panel of nine small and giant dogs and found only one variation in coding sequence, a synonymous SNP in exon 3 [chromosome 15 base pair position 44,226,324, *Canis familiaris* genome assembly 1 (CanFam1)]. Extensive resequencing within introns and flanking genomic sequence was also undertaken (table S3). Several additional SNPs (table S4) and an antisense oriented retrotransposon (table S5) unique to small breeds were identified. Alleles of a dinucleotide CA_n microsatellite in the *IGF1* promoter were also significantly associated with body size in the PWDs ($P < 1.4 \times 10^{-6}$, ANOVA) and the small and giant breeds ($P < 2.2 \times 10^{-14}$, chi–square test; table S6). All of these variations were in strong linkage disequilibrium and therefore a causative variant could not be definitively identified by this approach. Given the difficulty of developing inbred dog lines segregating small size, future studies will focus on using knock-in mice to explore the effect of these variants on phenotypes.

Our findings suggest that a single *IGF1* haplotype substantially contributes to size variation in the domestic dog. Because our sample includes small breeds that are distantly related (*25*) and reproductively isolated, and because the extent of haplotype sharing at *IGF1* is relatively small, the sequence variant or variants probably predate the common origin of the breeds and likely evolved early in the history of dogs. The early appearance of this allele may have facilitated the rapid genesis of size diversity in the domestic dog. The first archaeological record of dogs, beginning about 12,000 to 15,000 years ago (*9*, *26*), shows that size diversity was present early in the history of domestication. For example, dog remains from eastern Russia dated to 14,000 to 15,000 years ago are similar in size and conformation to great Danes, whereas slightly younger dog remains from the Middle East and Europe (10,000 to 12,000 years ago) are similar in size to small terriers (*9*, *26*, *27*). The early and widespread appearance of small size suggests that an ancesral small dog *IGF1* haplotype was readily spread over a large geographic area by trade and human migration and was maintained in local gene pools by selection. Such early selection on dogs may have been manifest as intentional artificial selection exercised by early humans or as an adaptive trait for coexistence with humans in the more crowded confines of developing villages and cities (*28*).

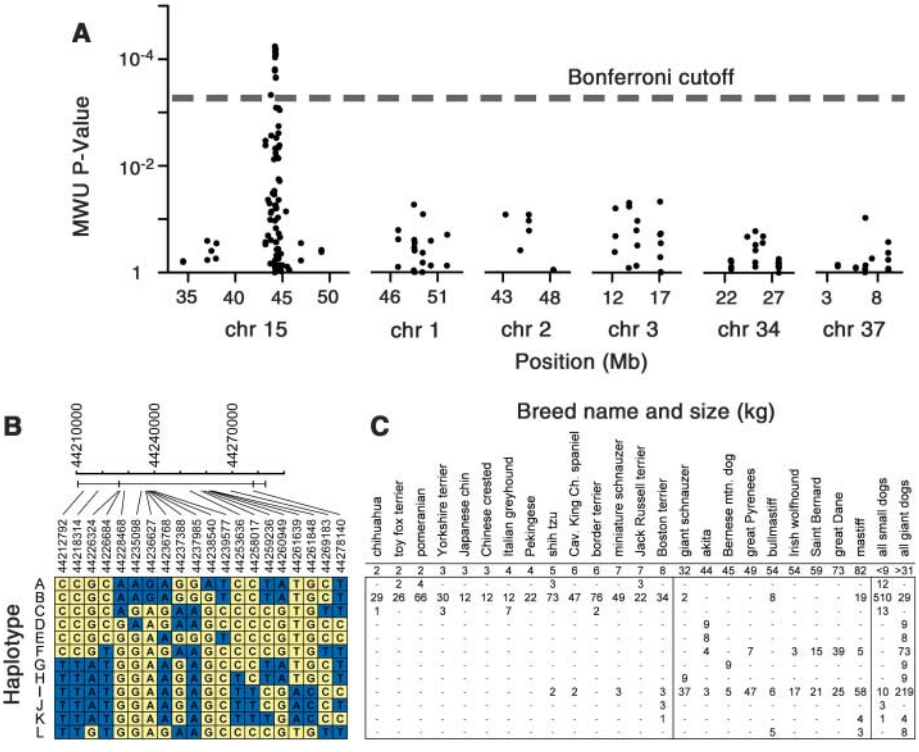


Fig. 3. Evidence of association and *IGF1* haplotypes for 14 small and 9 giant breeds. (**A**) Mann-Whitney U (MWU) *P* values for tests of association between individual SNPs and body size (small versus giant) for 116 SNPs on chromosome 15 and 83 SNPs on five control chromosomes. The dashed line indicates Bonferroni correction for multiple tests. Only breeds with data for at least 10 chromosomes were included (14 small and 9 giant breeds). (**B**) Haplotypes for the 20 markers spanning the small breed sweep interval near *IGF1*. The haplotypes were inferred independently in each breed. For each individual, fractional chromosome counts were summed for all haplotypes with at least 5% probability according to the haplotype inference software program PHASE. Chromosome sums for each breed were rounded to integer values; several breeds have odd numbers of chromosomes due to rounding error. Only inferred haplotypes carried by at least three dog chromosomes total (i.e., >0.5% frequency overall) are shown. Sequence reads collected from golden jackal (*Canis aureus*) were used to determine the ancestral allele for each SNP. The haplotypes are rows labeled A to L, and marker alleles are colored yellow for ancestral state (matching the nucleotide observed in the golden jackal) and blue for derived state. SNP positions within *IGF1* are shown at the top with *IGF1* introns (horizontal line) and exons (vertical bars) indicated. (**C**) Breed name and the average size of adult males in kilograms are provided. Small breeds less than 9 kg and giant breeds greater than 30 kg are grouped for totals shown at the far right.

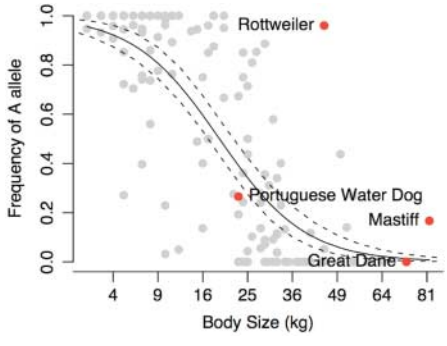


Fig. 4. Association of body size and frequency of the SNP 5 A allele. Binomial regression of allele frequency on square root of mean breed mass. Dashed lines indicate the 95% confidence interval on the predicted equation line as estimated from nonparametric bootstrap resampling. Between 5 and 109 (median = 22) dogs were genotyped for each of 143 breeds. The PWD is highlighted in red along with three giant breeds that have larger breed average masses than is predicted by their SNP 5 allele frequency.

The ubiquitous occurrence of the *IGF1* B haplotype in a diverse panel of small breeds clearly does not support unorthodox explanations of phenotypic diversity in the dog such as elevated mutation or recombination rates. Rather, we show that a single *IGF1* allele is a major determinant of small size in dogs and that intense artificial selection has left a signature in the proximity of *IGF1* that can readily be found by genomic scans of breeds sharing a common phenotype. The ability to identify a gene contributing to morphology without doing a genetic cross, but instead by using centuries of dog breeding, highlights the contribution that the study of canine genetics can make to an understanding of mammalian morphogenesis. These results provide a precedent for future studies aimed at identifying the genetic basis for complex traits such as behavior and skeletal morphology in dogs and other species with small populations that have experienced strong artificial or natural selection.

References and Notes

1. R. K. Wayne, *Evolution* **40**, 243 (1986).
2. R. K. Wayne, *J. Morphol.* **187**, 301 (1986).
3. J. W. Fondon 3rd, H. R. Garner, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 18058 (2004).
4. C. Webber, C. P. Ponting, *Genome Res.* **15**, 1787 (2005).
5. W. Wang, E. F. Kirkness, *Genome Res.* **15**, 1798 (2005).
6. R. K. Wayne, *J. Zool.* **210**, 381 (1986).
7. P. Saetre *et al.*, *Brain Res. Mol. Brain Res.* **126**, 198 (2004).
8. P. Savolainen, Y. P. Zhang, J. Luo, J. Lundeberg, T. Leitner, *Science* **298**, 1610 (2002).
9. S. J. Olsen, *Origins of the Domestic Dog* (Univ. of Arizona Press, Tucson, AZ, 1985).
10. C. Vila *et al.*, *Science* **276**, 1687 (1997).
11. J. Sampson, M. M. Binns, in *The Dog and Its Genome*, E. A. Ostrander, K. Lindblad-Toh, U. Giger, Eds. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 2006), vol. 44, pp. 19–30.
12. B. Van Valkenburgh, X. Wang, J. Damuth, *Science* **306**, 101 (2004).
13. K. Chase *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 9930 (2002).
14. K. Chase, D. R. Carrier, F. R. Adler, E. A. Ostrander, K. G. Lark, *Genome Res.* **15**, 1820 (2005).
15. J. Baker, J. P. Liu, E. J. Robertson, A. Efstratiadis, *Cell* **75**, 73 (1993).
16. K. A. Woods, C. Camacho-Hubner, D. Barter, A. J. Clark, M. O. Savage, *Acta Paediatr. Suppl.* **423**, 39 (1997).
17. K. A. Woods, C. Camacho-Hubner, M. O. Savage, A. J. Clark, *N. Engl. J. Med.* **335**, 1363 (1996).
18. N. B. Sutter *et al.*, *Genome Res.* **14**, 2388 (2004).
19. K. Lindblad-Toh *et al.*, *Nature* **438**, 803 (2005).
20. J. P. Pollinger *et al.*, *Genome Res.* **15**, 1809 (2005).
21. R. Kooijman, *Cytokine Growth Factor Rev.* **17**, 305 (2006).
22. P. Cohen, *Horm. Res.* **65**, 3 (2006).
23. R. P. Favier, J. A. Mol, H. S. Kooistra, A. Rijnberk *J. Endocrinol.* **170**, 479 (2001).
24. J. E. Eigenmann, D. F. Patterson, E. R. Froesch, *Acta Endocrinol. (Copenh.)* **106**, 448 (1984).
25. H. G. Parker *et al.*, *Science* **304**, 1160 (2004).
26. H. Epstein, *The Origin of the Domestic Animals of Africa* (Africana Publishing, New York, 1971).
27. M. V. Sablin, G. A. Khlopachev, *Curr. Anthropol.* **43**, 795 (2002).
28. E. Tchernov, L. K. Horwitz, *J. Anthropol. Archaeol.* **10**, 54 (1991).

29. We thank the hundreds of dog owners who contributed samples; the AKC Canine Health Foundation; S. Hoogstraten-Miller and I. Ginty for assistance at dog shows; P. Cruz for assistance with automated PCR primer designs; S. Kim for analytical assistance, and R. Pelker for assistance with blood serum assays of IGF1. Funded by the National Human Genome Research Institute (E.A.O., N.B.S., E.K., S.D., P.Q., H.G.P., and D.S.M.), the NSF (R.K.W.), NIH grant no. 5 T32 HG002536 (M.M.G.), NSF grant 0516310 (C.D.B. and L.Z.), NSF grant DBI 0606461 (B.P.), NIH grant P50 HG002790 (K.Z. and M.N.), and the National Institute of General Medical Sciences 063056, the Judith Chiara Charitable Trust, and the Nestle Purina Company (K.G.L.).

Supporting Online Material

www.sciencemag.org/cgi/content/full/316/5821/112/DC1
Materials and Methods
Figs. S1 to S9
Tables S1 to S6
References

1 November 2006; accepted 8 March 2007
10.1126/science.1137045

Binding of the Human Prp31 Nop Domain to a Composite RNA-Protein Platform in U4 snRNP

Sunbin Liu,[1]* Ping Li,[1,2]* Olexandr Dybkov,[1] Stephanie Nottrott,[1] Klaus Hartmuth,[1] Reinhard Lührmann,[1]† Teresa Carlomagno,[2]† Markus C. Wahl[3]†

Although highly homologous, the spliceosomal hPrp31 and the nucleolar Nop56 and Nop58 (Nop56/58) proteins recognize different ribonucleoprotein (RNP) particles. hPrp31 interacts with complexes containing the 15.5K protein and U4 or U4atac small nuclear RNA (snRNA), whereas Nop56/58 associate with 15.5K–box C/D small nucleolar RNA complexes. We present structural and biochemical analyses of hPrp31-15.5K-U4 snRNA complexes that show how the conserved Nop domain in hPrp31 maintains high RNP binding selectivity despite relaxed RNA sequence requirements. The Nop domain is a genuine RNP binding module, exhibiting RNA and protein binding surfaces. Yeast two-hybrid analyses suggest a link between retinitis pigmentosa and an aberrant hPrp31-hPrp6 interaction that blocks U4/U6-U5 tri-snRNP formation.

Most eukaryotic pre-mRNAs contain introns that are removed before translation by a multi-megadalton ribonucleoprotein (RNP) enzyme, the spliceosome (*1*–*3*). A spliceosome is assembled anew on each intron from small nuclear (sn) RNPs and non-snRNP splice factors (*4*, *5*). The RNP network of the spliceosome is extensively restructured during its maturation (*2*, *6*, *7*), reflected by changing RNA interactions. The U6 snRNA is delivered to the pre-mRNA in a repressed state, in which catalytically important regions are base-paired to the U4 snRNA (*8*, *9*). During spliceosome activation, the U4-U6 interaction is disrupted, U4 snRNA is released, and U6 snRNA forms short duplexes with U2 snRNA and the pre-mRNA substrate (*6*). Understanding this catalytic activation of the spliceosome requires detailed structural information on the snRNPs.

As for other complex RNPs (*10*), the U4/U6 di-snRNP is built in a hierarchical manner. A U4 5′ stem loop (U4 5′-SL) between two base-paired stems of U4/U6 serves as a binding site for the highly conserved U4/U6-15.5K protein (*11*). 15.5K binds to and stabilizes a kink turn (K turn) in the U4 5′-SL (*12*) and is required for subsequent recruitment of the human (h) Prp31 protein to the U4/U6 di-snRNP (*13*). hPrp31 does not interact with either the 15.5K or the RNA alone (*13*, *14*), but it is not known whether 15.5K merely prestructures the RNA for subsequent binding of hPrp31 or whether 15.5K provides part of the hPrp31 binding site. hPrp31 is essential for pre-mRNA splicing (*15*) and is a component of both major and minor spliceosomes. In the latter, the U4 snRNA is replaced by the U4atac snRNA (Fig. 1A). Nevertheless, both snRNAs bind 15.5K, and both primary RNPs incorporate hPrp31 in a strictly hierarchical manner (*13*, *16*).

The 15.5K protein also binds to a K turn in box C/D small nucleolar (sno) RNAs (*17*, *18*), but subsequently Nop56 and Nop58 (Nop56/58; Nop5p in archaea) are recruited to the snoRNPs (Fig. 1A) (*17*, *19*). Stem II of the snRNAs and snoRNAs (Fig. 1A) encompasses crucial identity elements for secondary protein binding. In the box C/D snoRNAs, stem II is longer by one base

[1]Abteilung Zelluläre Biochemie, Max-Planck-Institut für Biophysikalische Chemie, Am Faßberg 11, D-37077 Göttingen, Germany. [2]AG Flüssig-NMR Spektroskopie, Max-Planck-Institut für Biophysikalische Chemie, Am Faßberg 11, D-37077 Göttingen, Germany. [3]AG Makromolekulare Röntgenkristallographie, Max-Planck-Institut für Biophysikalische Chemie, Am Faßberg 11, D-37077 Göttingen, Germany.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: Reinhard.Luehrmann@mpi-bpc.mpg.de (R.L.); taco@nmr.mpibpc.mpg.de (T.C.); mwahl@gwdg.de (M.C.W.)

Supporting Online Material

Materials and methods

Sample and Data Collection

Whole blood was collected from purebred dogs with written consent from dog owners. Wild canid genomic DNA samples were also typed (*1*). This process was reviewed and approved by the animal care and use committees at the National Human Genome Research Institute, University of Utah, and the University of Missouri. Genomic DNA was extracted from blood by a standard phenol-chloroform protocol. Portuguese water dog samples were whole genome amplified (repli-G kit, Qiagen) prior to SNPlex genotyping but un-amplified DNA was used for sequence based marker discovery.

SNPs and insertion/deletion polymorphisms (Table. S4) were discovered by sequencing PCR amplicons (Table. S3) from dog genomic DNA. Sequencing reactions (Applied Biosystems) were bi-directional from exonuclease/shrimp alkaline phosphatase cleaned PCR amplicons by standard methods. Sequence data were collected on an ABI 3730xl and aligned and genotyped using phred/phrap and consed. SNP genotyping utilized the SNPlex platform (Applied Biosystems) following the manufacturer's protocol with 40-200 ng genomic DNA (small and giant breeds) or 80-200 ng whole genome amplified genomic DNA (Portuguese water dog) from each sample.

Serum levels of *IGF1* in Portuguese water dogs were measured by ELISA following standard methods.

Mixed model for Portuguese water dog fine-mapping

A mixed model was applied for fine mapping within the Portuguese water dog population since the shared ancestry within the breed could lead to spurious associations. To reduce the affect of this cryptic relatedness between dogs, we applied the mixed model analysis of Yu et al (*2*) using:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

where **Y** is the vector of the skeletal size trait; **α** is a vector of fixed effect, the SNP effect we are testing; **u** is a vector of random effect reflecting the polygenetic background; and **X** and **Z** are known incidence matrices relating the observations to fixed and random effects, respectively. The essential idea is that relatedness is incorporated into the model. The variance in the model can be expressed as:

$$Var\begin{bmatrix}\mathbf{u}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{K}\sigma_u^2 & 0\\ 0 & \mathbf{I}\sigma_e^2\end{bmatrix}$$

where **K** is the consanguinity matrix estimated from the known pedigree, which reflects the genetic background correlations between individuals.

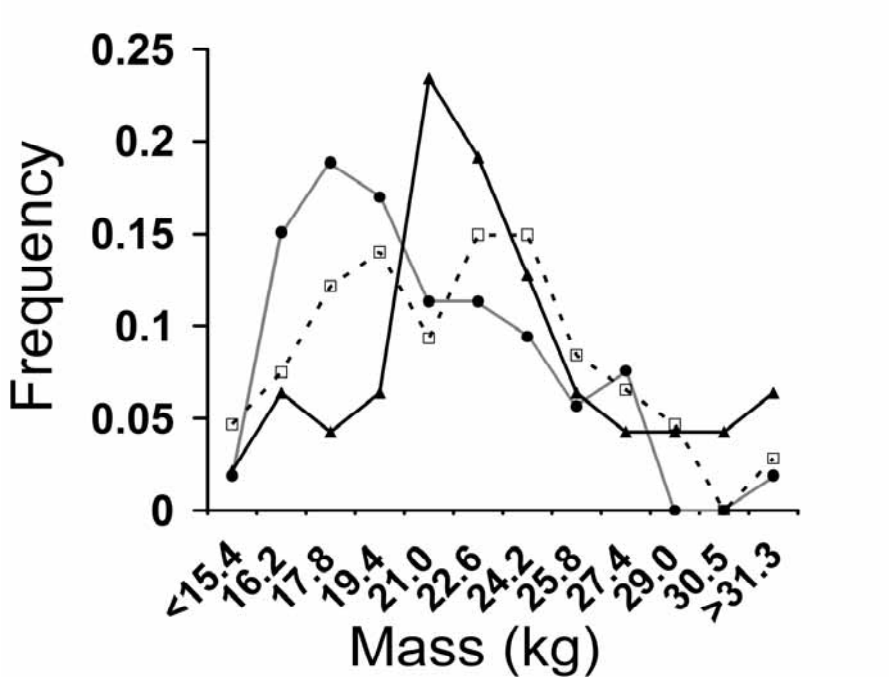
Mann-Whitney U test for association

When testing for association across structured populations such as dog breeds, there is a large inflation of nominal p-values in Fisher's exact test that is caused by the relatedness between samples within populations (see Fig. S6). Because dogs from different breeds are only very distantly related, a reasonable strategy is to only remove cryptic relatedness within breeds by collapsing the information obtained from dogs within the same breed into an allele frequency distribution. For each breed, we first calculated the relative frequency of the minor allele at a marker and then conducted a Mann-Whitney U test comparing the frequency in small dog breeds with the frequency in giant dog breeds. The test rejects the null hypothesis of no association if there is a large difference in the median allele frequency across small breeds as compared to the median frequency in large breeds.

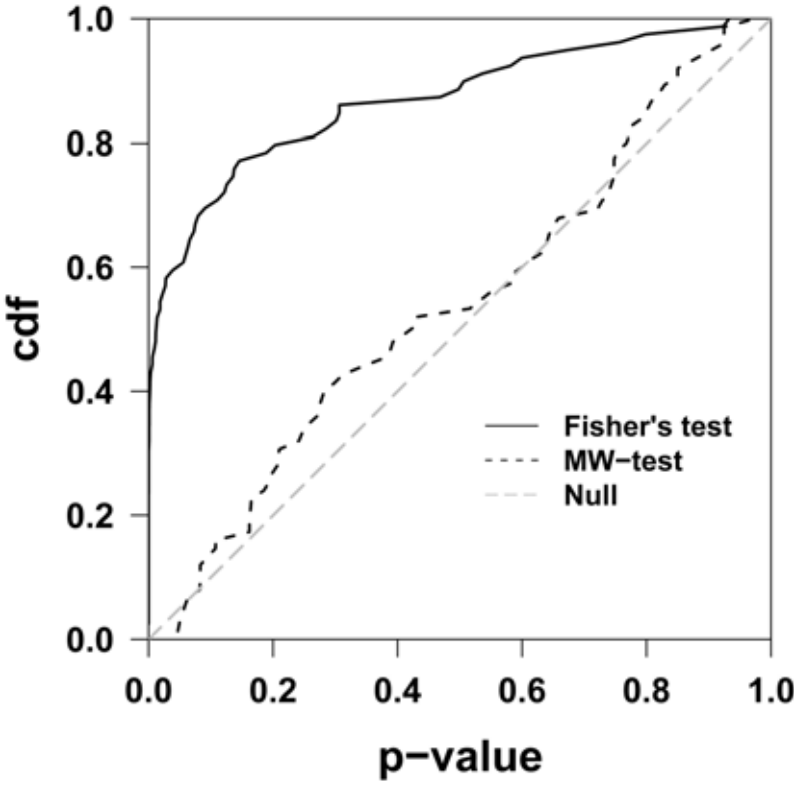
Estimation of the ancestral recombination graph

An ancestral recombination graph was reconstructed for a 1.2 Mb interval (chr15:43.7-44.9 Mb) that includes the *IGF1*core region from 1052 sequences of all small and giant dog breeds and is rooted with data from the golden jackal (*Canis aureus*) using the software SHRUB (*3*) [http://www.cs.ucdavis.edu/~yssong/lu.html]. Given a set of sequences and the ancestral sequence, SHRUB uses efficient branch and bound methods

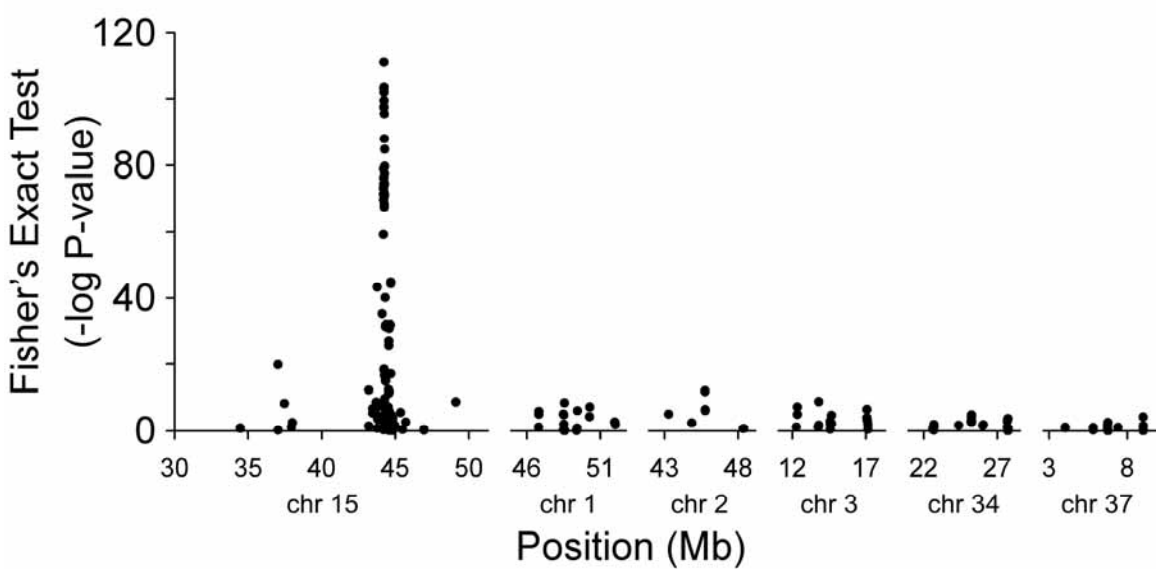
to compute the minimum number of recombination events necessary to explain the data and generates ARGs consistent with the data.



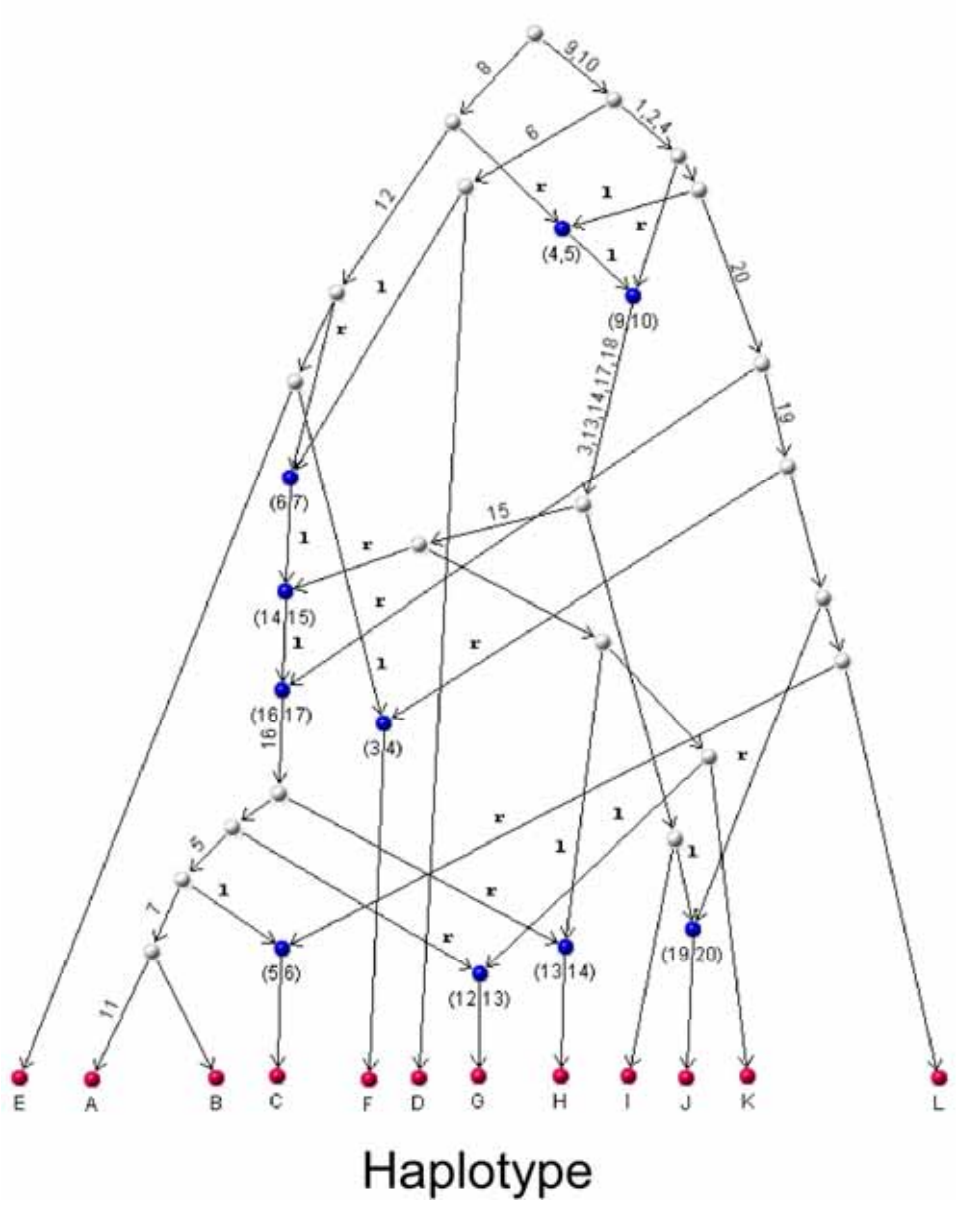
Supplementary Figure S1. Portuguese water dog *IGF1* haplotypes and mass. Haplotypes were inferred for 20 markers spanning the *IGF1* gene (cfa15:44,212,792-44,278,140, Canfam1). Out of the 720 chromosomes with successful inference, 96% carry one of just two haplotypes, "B" and "I", identical to haplotypes inferred for small and giant dogs, respectively (see Fig. 3). Data are graphed as a histogram for each genotype: I/I (closed triangle, solid line), B/I, (open square, dashed line) and B/B (closed circle, grayed line).



Supplementary Figure S6. Cumulative distribution function for Fisher's exact test and Mann-Whitney U statistic calculated from 83 genomic control SNPs genotyped in small and giant dogs.



Supplementary Figure S7. Fisher's exact test p-values for tests of association between individual SNPs and body size (small vs. giant) for 116 SNPs on chromosome 15 and 83 SNPs on five control chromosomes. Only breeds with data for at least ten chromosomes were included (14 small and 9 giant breeds). Note that, unlike p-values in Fig. 3A, these p-values clearly reflect confounding by population structure (see Material and Methods).



Supplementary Figure S8. An ancestral recombination graph that is consistent with the 12 haplotypes shown in Fig. 3B for the interval chr15:44,212,792 – 44,278,140. Red dots denote the 12 haplotypes, white dots denote coalescent events and blue dots indicate recombination vertices. The numbers in parentheses below recombination vertices denote breakpoint intervals, given as SNP positions reading from left to right in Fig 3b. Numbers along the edges in the graph indicate mutations. Recombination branches are labeled "l" or "r" to denote material to the left or right of recombination breakpoints.