

SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries

Curtis P Van Tassell¹, Timothy P L Smith², Lakshmi K Matukumalli^{1,3}, Jeremy F Taylor⁴, Robert D Schnabel⁴, Cynthia Taylor Lawley⁵, Christian D Haudenschild⁵, Stephen S Moore⁶, Wesley C Warren⁷ & Tad S Sonstegard¹

High-density single-nucleotide polymorphism (SNP) arrays have revolutionized the ability of genome-wide association studies to detect genomic regions harboring sequence variants that affect complex traits. Extensive numbers of validated SNPs with known allele frequencies are essential to construct genotyping assays with broad utility. We describe an economical, efficient, single-step method for SNP discovery, validation and characterization that uses deep sequencing of reduced representation libraries (RRLs) from specified target populations. Using nearly 50 million sequences generated on an Illumina Genome Analyzer from DNA of 66 cattle representing three populations, we identified 62,042 putative SNPs and predicted their allele frequencies. Genotype data for these 66 individuals validated 92% of 23,357 selected genome-wide SNPs, with a genotypic and sequence allele frequency correlation of $r = 0.67$. This approach for simultaneous *de novo* discovery of high-quality SNPs and population characterization of allele frequencies may be applied to any species with at least a partially sequenced genome.

The technology to multiplex thousands of SNPs into high-density assays has permitted genome-wide association studies for complex traits in the human^{1–5}. The ability to construct high-density SNP assays relies on the availability of a large number of SNPs with known genomic coordinates and minor allele frequencies (MAFs). Human SNP resources had been developed as part of the Human HapMap Project^{6,7} by the capillary-based Sanger resequencing of a few individuals⁸, with subsequent SNP validation and MAF estimation in panels of subject DNA representing disparate populations^{6,7}. Although successful, this approach relied on the low-coverage sequencing of a few individual samples and on the presence of a high-quality draft genome sequence. However, labor and reagent costs make this approach impractical for the development of high-density SNP assays for most species, which typically have draft genome sequences generated from the DNA

of a single inbred individual to facilitate the assembly of sequence contigs^{9,10}. Although these assemblies permit comparative genome studies¹¹, they do not allow the simultaneous identification of SNP resources for genetic studies. Moreover, the cost of complete genome resequencing for SNP discovery remains prohibitive, even considering the recent innovations in sequencing technologies.

The use of RRLs for SNP discovery was first described using Sanger sequencing¹²: pools of DNA from multiple individuals had been reduced in complexity by the size selection of fragments produced by complete restriction endonuclease digestion. This has the advantages of reducing the fraction of the genome present in the RRL by one to two orders of magnitude and ensuring that independently constructed libraries contain nearly identical fragment populations. Although Sanger sequencing of RRLs and the alignment of reads to each other or to a reference genome sequence will reliably identify SNPs, the depth of coverage necessary to simultaneously estimate MAFs is cost-prohibitive, particularly if the libraries contain a large number of fragments. Recently, RRLs produced by direct selection of target fragments from microarray hybridization has been described¹³. Although the number of unique fragments captured may not be sufficient to identify large numbers of genome-wide distributed SNPs, the approach has great utility for identification of SNPs within, for example, coding sequences. Also recently, next-generation sequencing of RRLs has been used for SNP discovery in plants¹⁴, but genomic complexity had been reduced by constructing cDNA libraries from tissue-specific transcriptomes. Owing to a low-fold sequence coverage, stringent filtering was required to identify 7,016 putative SNPs with an 85% validation rate, and validation and MAF estimation still required independent analyses¹⁴.

Development of high-density SNP assays in many species requires the development of efficient and cost-effective methods for identification of large numbers of randomly distributed SNPs and estimation of MAFs within target populations. To address this need, we designed an efficient and cost-effective approach for

¹Bovine Functional Genomics Laboratory, United States Department of Agriculture, Agricultural Research Service, 10300 Baltimore Avenue, Beltsville, Maryland 20705, USA. ²US Meat Animal Research Center, United States Department of Agriculture, Agricultural Research Service, P.O. Box 166, Clay Center, Nebraska 68933, USA.

³Bioinformatics and Computational Biology, George Mason University, 10900 University Blvd., Manassas, Virginia 20110, USA. ⁴Division of Animal Sciences, 920 East Campus Drive, University of Missouri, Columbia, Missouri 65211, USA. ⁵Illumina, Inc., 25861 Industrial Blvd., Hayward, California 94545, USA. ⁶Department of Agricultural, Food and Nutritional Science, University of Alberta, 410 AgFor Centre, Edmonton, Alberta T6G 2P5, Canada. ⁷Genome Sequencing Center, Washington University School of Medicine, Campus Box 8501, 4444 Forest Park Avenue, St. Louis, Missouri 63108, USA. Correspondence should be addressed to C.P.V.T. (curt.vantassell@ars.usda.gov).

Table 1 | Summary of DNA sequence production and filtering

Run	Total sequences	Edits			
		CC present ^a (%)	N content ^b (%)	Pass quality control ^c (%)	Overall ^d (%)
1	8,366,699	7,463,431 (89.2)	7,463,431 (100.0)	7,267,282 (97.4)	86.9
2	7,139,161	6,366,225 (89.2)	6,289,525 (98.8)	6,172,427 (98.1)	86.5
3	11,182,655	10,216,432 (91.4)	10,060,676 (98.5)	9,382,671 (93.3)	83.9
4	16,477,806	14,683,651 (89.1)	14,567,407 (99.2)	10,715,303 (73.6)	65.0
5	19,808,937	17,695,540 (89.3)	17,518,045 (99.0)	15,659,391 (89.4)	79.1
Total	71,815,443	64,143,272 (89.3)	63,595,878 (99.1)	49,492,755 ^e (77.8)	68.9

^aSequences remaining after filtering those without a CC tag at the restriction site. ^bSequences remaining after sequentially filtering those with missing base calls. ^cSequences remaining after sequentially filtering those that failed quality control (see Methods). ^dOverall percentage of sequences retained after filtering. ^eAn additional run failed, but 295,681 useable sequences were made available by Illumina. Those data were included only in the total count that passed quality control.

simultaneous SNP discovery, validation and characterization using ‘next-generation’ sequencing technology. Our approach is based on deep sequencing of libraries of reduced complexity constructed from pooled DNA samples that represent populations of interest, to efficiently develop highly reliable SNP sets for species with limited available genome sequence. We used the draft bovine sequence assembly (Btau3.1) based on combined ‘shotgun’ and targeted large-insert clone sequencing from an inbred Hereford cow and her sire, respectively (<http://www.hgsc.bcm.tmc.edu/projects/bovine>), to compare, *in silico*, the characteristics of fragments produced by complete restriction digests of the entire genome. We compared candidate enzymes for library preparation based upon their predicted fragment population: (i) repetitive element content, (ii) number of unique fragments by size range, and (iii) genome coverage and distribution. Based on the characteristics of the fragment population produced by the selected enzyme, we optimized the read depth for a next-generation sequencing technology to simultaneously identify a large number of high-quality bovine SNP and estimate their MAFs. Our objective was to maximize the number of whole-genome detected SNPs while minimizing the false discovery rate and to simultaneously estimate MAFs.

RESULTS

Sequencing platform

We generated sequences from the bovine RRLs using the sequence-by-synthesis method on the Clonal Single Molecule Array (CSMA) platform¹⁵ from Illumina. At the time of this study, the technology produced a read length of only 25 bp, which was insufficient to enable the concurrent design of genotyping assays on existing SNP analysis platforms. We overcame this limitation by augmenting the short-read sequences with sequences from the draft bovine genome assembly flanking each SNP. Efficiency of the Illumina sequencing platform is maximized when input fragments are 70–200 bp (unpublished data), so we targeted this size range when evaluating potential restriction enzymes for RRL production.

In silico restriction digestions

We first performed *in silico* digestions of Btau3.1 with candidate restriction enzymes to identify those predicted to yield 0.5–1 million fragments within the 70–200 bp size range. We estimated the proportion of repetitive sequence in the 25 bp at each fragment end (<http://www.repeatmasker.org>); plots of the predicted number of fragments and of repetitive element content against fragment sizes revealed the presence of repetitive elements

as ‘spikes’ in the plots at a particular fragment size (**Supplementary Fig. 1** online). We derived the optimal number of target fragments by modeling based on: (i) the target number of SNPs to be identified (> 50,000), (ii) estimated SNP density (1 SNP per 500 bp for the bovine genome), (iii) the depth of library sequencing as determined by the number of budgeted machine runs (five CSMA flow cells), and (iv) the number of independent chromosomes sampled per RRL (20–30-fold average depth of coverage per fragment end). We selected *HaeIII* because it produced the desired estimated number of fragments in the size range of 70–130 bp (872,835 predicted fragments), allowed avoidance of apparent repetitive elements in the fragments at 134 and 164 bp, provided a substantial reduction in repetitive elements (~25% compared to ~45% in the whole genome), and completely digested bovine genomic DNA *in vitro*. Furthermore, the resulting fragments had blunt ends, which avoided the need to enzymatically blunt the ends during library construction.

After *in silico* analyses and enzyme selection, we prepared 3 pooled DNA samples: the first DNA sample was from 15 cattle representing different Holstein lineages (HOL), the most popular US dairy breed; the second sample consisted of DNA from 35 Angus bulls (ANG), the most common beef breed in the US; and the third sample included at least two bulls each from Charolais, Gelbvieh, Hereford, Limousin, Red Angus and Simmental, the six next most common beef breeds (BEEF). To prepare the libraries, we digested these samples with *HaeIII* and isolated fragments in the 70–130 bp size range (avoiding a band at 91–92 bp of fragments harboring repetitive elements) from a preparative polyacrylamide gel (**Supplementary Fig. 2** online). The number of independent chromosomes (30, 70 or 32) represented in each pool ensured that, on average, each sequence read would sample a different chromosome, assuming tenfold sequence coverage of fragments within each pool. We predicted that by including 132 bovine chromosomes from 8 different breeds we could capture all common SNPs within the surveyed fragments and detect a portion of breed-specific or low-MAF SNPs in Holstein and Angus.

DNA sequencing and filtering

We independently sequenced the 3 RRLs (HOL, ANG and BEEF), generating over 71 million reads (**Table 1**). The yield of sequence per flow cell channel was variable, with ANG producing the fewest useful sequences per run. Consequently, we assigned a larger number of channels to this library than to HOL or BEEF. The most commonly observed sequence (3,710,175 occurrences) was an artifact corresponding to the Illumina adaptor oligonucleotide

Table 2 | Summary of SNP discovered, validated and characterized when genotyped in the 66 animals used for SNP discovery

Position in tag (bp)	Putative SNP ^a	Assayed SNP ^b	Called SNPs ^c (%)	Average MAFs for all called SNPs	Polymorphic SNPs (%)	Average MAFs for polymorphic SNPs
3	3,442	1,360	1,291 (94.9)	0.27	1,268 (98.2)	0.27
4	2,816	1,120	1,051 (93.8)	0.25	1,021 (97.1)	0.26
5	2,867	1,125	1,080 (96.0)	0.26	1,027 (95.1)	0.27
6	2,718	1,056	1,008 (95.5)	0.26	969 (96.1)	0.27
7	2,593	966	925 (95.8)	0.26	871 (94.2)	0.28
8	2,620	1,042	994 (95.4)	0.26	955 (96.1)	0.27
9	2,477	979	928 (94.8)	0.26	873 (94.1)	0.27
10	2,412	945	910 (96.3)	0.26	865 (95.1)	0.27
11	2,554	971	916 (94.3)	0.26	881 (96.2)	0.27
12	2,534	1,043	1,000 (95.9)	0.26	960 (96.0)	0.27
13	2,480	979	935 (95.5)	0.26	892 (95.4)	0.27
14	2,572	1,031	999 (96.9)	0.26	948 (94.9)	0.28
15	2,517	1,023	971 (94.9)	0.26	923 (95.1)	0.27
16	2,621	1,055	1,012 (95.9)	0.26	954 (94.3)	0.27
17	2,634	976	940 (96.3)	0.25	896 (95.3)	0.27
18	2,611	1,019	969 (95.1)	0.25	921 (95.0)	0.27
19	2,518	1,001	946 (94.5)	0.26	896 (94.7)	0.28
20	2,570	1,020	962 (94.3)	0.26	905 (94.1)	0.28
21	2,530	971	916 (94.3)	0.25	845 (92.2)	0.27
22	2,543	1,042	986 (94.6)	0.25	910 (92.3)	0.27
23	2,612	1,044	991 (94.9)	0.25	878 (88.6)	0.28
24	2,871	1,183	1,112 (94.0)	0.23	894 (80.4)	0.2
25	3,930	1,649	1,515 (91.9)	0.16	911 (60.1)	0.27
Total	62,042	24,600	23,357 (94.9)	0.25	21,463 (91.9)	0.27

^aPutative SNPs identified from analysis of sequences across all three RRL. ^bSNPs successfully designed and synthesized oligonucleotides on the Illumina iSelect platform. ^cSNPs successfully genotyped on the Illumina iSelect platform.

sequence. Because we prepared the libraries using *HaeIII*, properly ligated fragments included a two-nucleotide CC tag at the restriction site. We discarded the 7,672,171 sequences that did not begin with this tag (including adaptor sequences), leaving ~64 million reads for analysis. After removing 547,394 sequences containing at least one unknown base and 14,103,123 sequences with low-quality scores, 49,492,755 reads (1.3 billion bp) corresponding to 5,022,143 unique sequences remained for SNP discovery. The most frequently occurring sequence (713,735 occurrences) was a perfect match to more than 100 locations in Btau3.1 and represented a contaminating repetitive element. We discarded reads that exactly matched more than one position in Btau3.1, and those that uniquely mapped but yielded more than one putative SNP when clustered. To improve alignment to the genome sequence, we prefixed to the complete sequence, including the two CC nucleotides, two additional nucleotides (GG) inferred from the requisite presence of the *HaeIII* restriction site.

SNP discovery and validation

We identified a total of 62,042 putative SNPs within fragments that mapped uniquely to Btau3.1, in which only a single variable position was detected and each alternate allele was represented at least twice (see Methods). Putative SNP were approximately uniformly distributed across autosomes, although there was regional variation (**Supplementary Fig. 3** online). We used tag and flanking sequence derived from Btau3.1 to design 24,600 Infinium assays for the Illumina iSelect genotyping platform, and successfully genotyped 23,357 of these SNP loci. We individually genotyped the 66 cattle comprising the three RRLs to assess the concordance of

sequencing-derived allele frequencies estimated in the SNP discovery with those from the individuals represented in the RRLs. The genotypes indicated a modest rate of false SNP discovery, with 21,463 or 92% of assayed SNP being polymorphic (**Table 2**). False discovery rate varied with location of the predicted SNP within the sequence read. The percentage of monomorphic loci was 1.8–7.8% for predicted SNPs located in positions 3–22 of the sequence reads, but was 11.4–39.9% for SNPs within the last three positions of the read (**Table 2**). The average MAF across all 23,357 genotyped loci was 0.250, but for the 21,463 polymorphic SNPs average MAF was 0.272. Considering only the putative SNPs identified from sequence differences in read positions 3–22, these average MAFs were 0.258 and 0.271, respectively. These results demonstrate that the approach was successful in identifying common SNPs segregating within the three screened populations. For comparison, we designed an additional 6,200 Infinium assays for the Illumina iSelect genotyping platform by sampling from the 118,249 SNPs produced by the Baylor College of Medicine (<ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Btaurus/snp/Btau20050310/README>). These SNPs had been discovered using single-pass Sanger sequences aligned to the reference genome and had not been experimentally validated. A total of 5,809 loci were successfully genotyped and 90% were polymorphic, but these SNPs had substantially lower average MAFs than the SNPs discovered by our strategy (**Supplementary Table 1** online).

We evaluated the accuracy of estimation of allele frequencies from the sequence data by comparison to frequencies determined from the genotype data (**Fig. 1**). In our simulation of one pool of individuals from a single population in which the distribution of

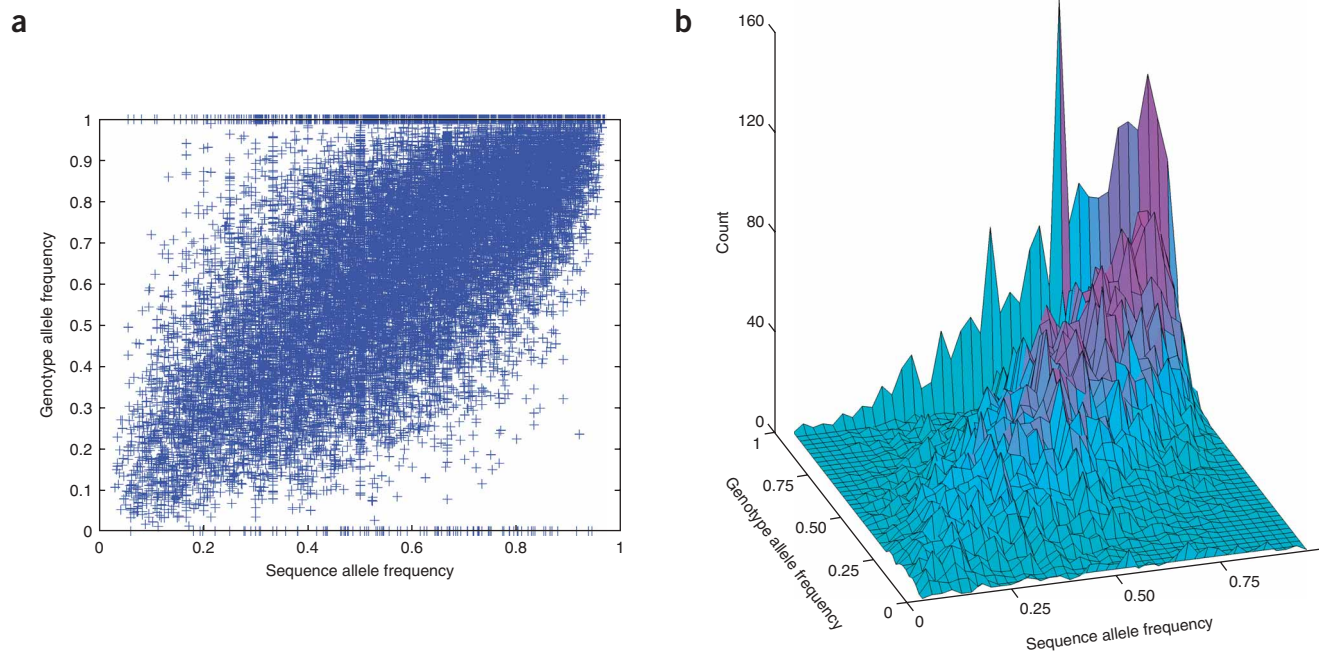


Figure 1 | Distribution of sequence and genotype derived allele frequencies ($r = 0.67$) in the SNP discovery populations. **(a)** Scatter plot of 21,463 SNPs. **(b)** Surface plot.

SNP allele frequencies followed a beta distribution, the expected correlation of sequence and genotype-determined allele frequencies was very high (> 0.9). However, our modeling did not account for population stratification (correlation performed across populations), sequencing errors or the selection of SNP to be genotyped based on sequence MAFs. In view of these confounding factors, the correlation between genotype and sequence derived allele frequencies was remarkably high ($r = 0.67$). Overall, sequence derived estimates of MAF from a 22-fold coverage appeared satisfactory; however, we anticipate improved results from increasing both the sequence depth and enhancing prediction accuracy in the target population by increasing the number of individuals represented in each RRL.

DISCUSSION

Our results indicate that next-generation sequencing technologies offer an opportunity for deep sequencing of RRLs to simultaneously identify, validate and characterize large numbers of high-quality SNPs in target populations. Our total reagent costs to identify 62,042 putative SNPs and to simultaneously estimate MAF was approximately \$0.48 per SNP. Further, we estimated our false SNP discovery rate to be 8% by genotyping 23,357 SNPs in the discovery samples. To place this in context, we estimate that the cost of discovery of 118,249 putative bovine SNPs by Sanger sequencing and alignment to a draft assembly was \$2.95 per SNP, assuming that the 348,958 bovine sample survey Sanger reads each cost \$1. While these SNPs have about a 10% false discovery rate (**Supplementary Table 1**) and do not have MAF estimated in the discovery process, they are further limited in utility by the fact that they map to only 56,634 unique sequence reads. The value of simultaneous estimation of MAF in the target population is clear, with the RRL (Sanger)-derived putative SNP having 8.1% (10%) monomorphic loci and 61.5% (47.4%) SNPs with MAFs ≥ 0.2 (**Supplementary**

Table 1). Our strategy was designed to minimize false positive SNP discovery and to identify a large number of common SNPs for the development of a high-density genome-wide association assay and to simultaneously predict MAFs, and this strategy almost certainly resulted in the discarding of significant numbers of real SNPs. We considered this to be an acceptable compromise.

For species with at least a draft genome sequence, SNP discovery can be optimized by identifying an appropriate restriction enzyme using *in silico* digests of the sequence. Pooling of DNA samples from a large number of distantly related individuals, combined with deep sequencing of a small fraction of the genome using RRLs, results in the identification of a large number of common SNPs and a substantial number of rare SNPs. Most importantly, the approach simultaneously generates estimates of allele frequencies in the discovery populations and can even identify SNPs with fixed allelic differences between groups. This result suggests that the approach may have considerable utility for the discovery of SNPs that are associated with disease by forming discovery populations from affected and unaffected individuals. SNPs with large differences in allele frequency between the discovery populations are strongly associated with the disease phenotype and are likely to be in linkage disequilibrium with disease risk loci. Owing to high efficiency and low cost, the approach also allows access to advanced genomic technologies to scientists from developing countries working with less well-characterized species or regionally adapted strains.

In the absence of a genome sequence, the approach can be implemented using longer sequence read technologies, or a combined approach of longer reads to generate flanking sequence and of higher-density short reads for SNP discovery. However, a limitation in the application of this method in the absence of a genome sequence is the inability to order SNPs within the sequence assembly. This challenge may be overcome by genotyping linkage

mapping populations or radiation hybrid panels¹⁶, or by using comparative genomic information to infer likely or relative genome position (if closely related genomes have more advanced information). Nevertheless, use of these alternate approaches likely will result in some loss of efficiency compared to the strategy demonstrated here. Finally, genotyping using short-read sequencing technologies combined with sample indexing to track the source of alleles may soon be a feasible alternative to multiple library construction. In summary, our approach provides an efficient and economical means of producing data for the design of high-density SNP genotyping platforms for species with draft sequence assemblies and provides a framework for methods in species that lack genome sequence.

The strategy of using next-generation sequencing in conjunction with reduced representation has completely changed the landscape of high-density SNP assay development, particularly for non-human genome applications. It is no longer necessary to conduct distinct projects to identify putative SNPs and then validate and characterize their allele frequencies in target populations. The scale and efficiency of SNP discovery achieved in this study suggests that it is now practical to deep-sequence RRLs to identify the SNP with utility for high-density genome-wide association studies in targeted populations.

METHODS

Restriction digestion. We performed *in silico* digests of the cattle (*Bos taurus*) genome sequence assembly (Btau3.1) for 84 commercially available Type II DNA restriction enzymes, assuming complete digestion. We considered methylation sensitivity for each enzyme and its isoschizmers when selecting the enzymes to test. Characteristics we evaluated included: (i) number of fragments predicted in the target size range (70–200 bp), (ii) proportion of repetitive nucleotides in the terminal 25 base pairs of each fragment end, and (iii) proportion of unknown (N) base pairs in the terminal 25 base pairs. The identification of repetitive elements was based on annotation from the UCSC cattle genome browser (<http://genome.ucsc.edu/cgi-bin/hgGateway?&org=Cow&>). We first chose *DraI* (TTT[^]AAA) based upon a considerable reduction in repetitive element content within the fragments in the size range 70–130 bp and an almost uniform size distribution of predicted digested fragments. However, we observed little enzyme activity when we tested several cattle DNA samples *in vitro* despite activity in control genomic DNA samples from porcine and ovine individuals (data not shown). As a result, we selected for this project a second enzyme, *HaeIII* (GG[^]CC), with excellent *in silico* digestion characteristics.

Library construction. We prepared three pools of DNA containing 16 BEEF, 15 HOL and 35 ANG cow DNA samples, and digested 5.5 µg of DNA from each pool with *HaeIII* (New England Biolabs) as suggested by the manufacturer, using 50 units of enzyme in a total volume of 150 µl. The reaction proceeded overnight at 37 °C to ensure complete digestion. We fractionated the resulting fragments on a 6% non-denaturing polyacrylamide gel (16 cm × 14 cm, 1.5 mm thick) and stained with Syber Gold (Molecular Probes) to reveal the predicted, distinctive bands of repetitive elements at 70 and 130 bp, as well as an unanticipated band at ~92 bp. We presumed this to be the result of an undescribed bovine repetitive element containing *HaeIII* sites,

which we confirmed by *in silico* digestion of the contigs represented in the unassembled draft genome sequence. We excised the digestion products between 70 and 130 bp, taking care to avoid the band at 92 bp (**Supplementary Fig. 2**).

To shear the gel pieces, we nested a 0.5-ml microcentrifuge tube (with a hole in the bottom formed with an 18-gauge needle) containing the gel slices inside a 2-ml microcentrifuge tube and centrifuged the sample at 16,873g for 2–3 min. We added 150 µl of buffer (8 mM Tris pH 8.0, 0.08 mM EDTA, 1.25 M ammonium acetate) to the sheared gel pieces, vortexed the sample, and eluted DNA at 4 °C overnight, followed by a 15-min incubation at 65 °C. To purify the eluate, we placed the slurry in a SNAP column (Invitrogen) and centrifuged the sample at 16,873g for 2 min. We precipitated the DNA by adding ethanol in the presence of glycogen carrier to the eluent, and washed the pellet twice with 500 µl cold 70% ethanol, dried it and resuspended the DNA in 11 µl of TE (10 mM Tris pH 8.0, 0.1 mM EDTA).

To add an adenine overhang to the 3' ends of each strand, we treated the gel-isolated blunt-ended fragments produced by *HaeIII* digestion with Klenow fragment of DNA polymerase lacking exonuclease activity in the presence of dATP in a 50 µl reaction volume using the buffer and enzyme provided by Illumina, Inc., at 37 °C for 20 min followed by inactivation of the enzyme at 70 °C for 5 min. We precipitated the modified fragments with ethanol in the presence of glycogen carrier, dried them and resuspended them in 10 µl of TE. We added adapters required for sequencing on the Illumina CSMA sequencing-by-synthesis platform by ligation in a 50 µl final volume using adapters and buffers supplied in a proprietary kit (Illumina, Inc.). Ligation proceeded for 15 min at 22 °C to produce fragments with 65 bases added from the adaptor primers.

We separated the ligation products on a 6% nondenaturing polyacrylamide gel as previously described, and excised the ligation products of 130–190 bp from the gel, eluted and precipitated them as previously described, and resuspended the final fragment pellet in 15 µl of TE. We amplified the fragments in duplicate by 18 cycles of PCR in 50-µl reactions using 7 µl of the fragment eluate as template in a reaction containing amplification primers, buffer and thermostable polymerase from the Illumina Genomic DNA Sample Preparation kit. The thermocycling parameters were 30 s at 98 °C, followed by 18 cycles of 98 °C 10 s, 65 °C 30 s, 72 °C 30 s, and a final elongation step of 72 °C for 5 min. We purified the amplification products using a Qiaquick PCR Purification kit (Qiagen) as directed by the manufacturer, and eluted them from the column in 50 µl of the provided buffer. We quantified the final products by absorbance spectroscopy, indicating total yields of 1.5–2 µg. We sent the PCR products to Illumina for sequencing (**Supplementary Methods** online).

SNP discovery and mapping. After screening to require that the first two sequenced bases were the expected terminal sequence (CC) derived from *HaeIII* digestion, we retained the remaining 23 bases (tags) for analysis. We screened the data for overall quality by requiring an average base quality score of at least 25 across the entire tag for all reads. The methods used to align sequence tags are available in **Supplementary Methods**. To enhance the likelihood that aligned tag pairs differing by a single nucleotide corresponded to SNPs, we imposed several constraints. First, we required a minimum nucleotide quality score of 27 at the SNP

position. Second, we required that for each tag pair, one of the two alternate tag allele sequences (extended to include the inferred presence of the restriction site bases) uniquely map to Btau3.1. Furthermore, we required that the tag sequence containing the alternate allele have no exact match to Btau3.1 because such a match would indicate two possible genomic positions for the SNP. To reduce SNP associated with repetitive elements or gene families, we discarded tags that contained more than one SNP. Finally, we required a minimum threshold of two copies of each allele (including the Btau3.1 aligned sequence) to call a SNP. We calculated sequence depth as the total number of reads corresponding to a tag, excluding the Btau3.1 copy. We calculated MAFs from the allele counts within the sequence copies from the region pooled across all three libraries.

Additional methods. The methods used for assay design for SNP validation are available in **Supplementary Methods**.

Accession codes. Single Nucleotide Polymorphism database (dbSNP): ss86273438–ss86341791.

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

ACKNOWLEDGMENTS

J.F.T. and R.D.S. were supported by National Research Initiative grants 2005-35205-15448, 2005-35604-15615, 2006-35205-16701 and 2006-35616-16697 from the US Department of Agriculture Cooperative State Research, Education and Extension Service. C.P.V.T., T.S.S., and L.K.M. were supported by National Research Initiative grant 2006-35205-16888 from the US Department of Agriculture Cooperative State Research, Education, and Extension Service and by Projects 1265-31000-081D and 1265-31000-090-00D from the United States Department of Agriculture Agricultural Research Service. T.P.L.S. was supported by Project 5438-31000-073D from the US Department of Agriculture Agricultural Research Service. L.K.M. was also supported by National Research Initiative grant 2006-35205-17878 from the US Department of Agriculture Cooperative State Research, Education and Extension Service. We gratefully acknowledge the early prepublication access under the Fort Lauderdale conventions to the draft bovine genome sequence provided by the Baylor College of Medicine Human Genome Sequencing Center and the Bovine Genome Sequencing Project Consortium.

AUTHOR CONTRIBUTIONS

C.P.V.T. and L.K.M. developed and implemented the SNP discovery algorithm; J.F.T. and C.P.V.T. performed SNP discovery modeling; C.P.V.T., T.S.S. and L.K.M. performed *in silico* genome analysis; W.C.W. suggested the reduced representation strategy; T.P.L.S. constructed the RRLs; J.F.T., R.D.S., T.P.L.S. and T.S.S. identified

cows for DNA pools; T.S.S. managed the DNA collection; C.T.L. genotyped the discovery animals and managed the assay synthesis; C.D.H. sequenced the RRLs; L.K.M., T.S.S., R.D.S., S.S.M. and T.P.L.S. conducted pilot validations; and C.P.V.T., J.F.T., T.S.S. and T.P.L.S. coordinated manuscript writing and editing.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at <http://www.nature.com/naturemethods/>.

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

1. Klein, R.J. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385–389 (2005).
2. Libioulle, C. *et al.* Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS Genet.* **3**, e58 (2007).
3. Sladek, R. *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881–885 (2007).
4. Zanke, B.W. *et al.* Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat. Genet.* **39**, 989–994 (2007).
5. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
6. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
7. The International HapMap Consortium. The international HapMap project. *Nature* **426**, 789–796 (2003).
8. Nickerson, D.A., Tobe, V.O. & Taylor, S.L. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res.* **25**, 2745–2751 (1997).
9. International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695–716 (2004).
10. Lindblad-Toh, K. *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803–819 (2005).
11. O'Brien, S.J. *et al.* The promise of comparative genomics in mammals. *Science* **286**, 458–481 (1999).
12. Altshuler, D. *et al.* An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**, 513–516 (2000).
13. Albert, T.J. *et al.* Direct selection of human genomic loci by microarray hybridization. *Nat. Methods* **4**, 903–905 (2007).
14. Barbazuk, W.B., Emrich, S.J., Chen, H.D., Li, L. & Schnable, P.S. SNP discovery via 454 transcriptome sequencing. *Plant J.* **51**, 910–918 (2007).
15. Johnson, D.S., Mortazavi, A., Myers, R.M. & Wold, B. Genome-wide mapping of *in vivo* protein-DNA interactions. *Science* **316**, 1497–1502 (2007).
16. McKay, S.D. *et al.* Construction of bovine whole-genome radiation hybrid and linkage maps using high-throughput genotyping. *Anim. Genet.* **38**, 120–125 (2007).