Copy number variation of individual cattle genomes using next-generation sequencing

Derek M. Bickhart, Yali Hou, Steven G. Schroeder, et al.

Genome Res. 2012 22: 778-790 originally published online February 2, 2012
Access the most recent version at doi:10.1101/gr.133967.111

Resource

Copy number variation of individual cattle genomes using next-generation sequencing

Derek M. Bickhart,[1,8] Yali Hou,[1,2,8] Steven G. Schroeder,[1] Can Alkan,[3,9] Maria Francesca Cardone,[4] Lakshmi K. Matukumalli,[1] Jiuzhou Song,[2] Robert D. Schnabel,[5] Mario Ventura,[3,4] Jeremy F. Taylor,[5] Jose Fernando Garcia,[6] Curtis P. Van Tassell,[1] Tad S. Sonstegard,[1] Evan E. Eichler,[3,7] and George E. Liu[1,10]

[1]*USDA-ARS, ANRI, Bovine Functional Genomics Laboratory, Beltsville, Maryland 20705, USA;* [2]*Department of Animal and Avian Sciences, University of Maryland, College Park, Maryland 20742, USA;* [3]*Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington 98195, USA;* [4]*Department of Genetics and Microbiology, University of Bari, Bari 70126, Italy;* [5]*Division of Animal Sciences, University of Missouri, Columbia, Missouri 65211, USA;* [6]*UNESP-Univ Estadual Paulista, Rua Clóvis Pestana, 793, Araçatuba, SP, Brazil;* [7]*Howard Hughes Medical Institute, Seattle, Washington 98195, USA*

Copy number variations (CNVs) affect a wide range of phenotypic traits; however, CNVs in or near segmental duplication regions are often intractable. Using a read depth approach based on next-generation sequencing, we examined genome-wide copy number differences among five taurine (three Angus, one Holstein, and one Hereford) and one indicine (Nelore) cattle. Within mapped chromosomal sequence, we identified 1265 CNV regions comprising ~55.6-Mbp sequence—476 of which (~38%) have not previously been reported. We validated this sequence-based CNV call set with array comparative genomic hybridization (aCGH), quantitative PCR (qPCR), and fluorescent in situ hybridization (FISH), achieving a validation rate of 82% and a false positive rate of 8%. We further estimated absolute copy numbers for genomic segments and annotated genes in each individual. Surveys of the top 25 most variable genes revealed that the Nelore individual had the lowest copy numbers in 13 cases (~52%, χ^2 test; *P*-value <0.05). In contrast, genes related to pathogen- and parasite-resistance, such as *CATHL4* and *ULBP17*, were highly duplicated in the Nelore individual relative to the taurine cattle, while genes involved in lipid transport and metabolism, including *APOL3* and *FABP2*, were highly duplicated in the beef breeds. These CNV regions also harbor genes like *BPIFA2A (BSP30A)* and *WC1*, suggesting that some CNVs may be associated with breed-specific differences in adaptation, health, and production traits. By providing the first individualized cattle CNV and segmental duplication maps and genome-wide gene copy number estimates, we enable future CNV studies into highly duplicated regions in the cattle genome.

[Supplemental material is available for this article.]

Copy number variations (CNVs) are gains and losses of genomic sequence >50 bp between two individuals of a species (Mills et al. 2011). Substantial progress has been made in understanding CNVs in mammals, especially in humans (Redon et al. 2006; Conrad et al. 2009; Altshuler et al. 2010; Mills et al. 2011) and rodents (Graubert et al. 2007; Guryev et al. 2008; She et al. 2008; Yalcin et al. 2011). While single nucleotide polymorphisms (SNPs) are more frequent, CNVs impact a higher percentage of genomic sequence and have potentially greater effects, including the changing of gene structure and dosage, altering gene regulation and exposing recessive alleles (Zhang et al. 2009). In particular, segmental duplications (SDs) were shown to be one of the catalysts and hotspots for CNV formation (Sharp et al. 2005; Alkan et al. 2009; Marques-Bonet et al. 2009). Several common CNVs have been shown to be important in both normal phenotypic variability and disease susceptibility in human (Aitman et al. 2006; Fellermann et al. 2006; Le Marechal et al. 2006; Fanciulli et al. 2007; Yang et al. 2007; Stankiewicz and Lupski 2010). Although analyses of a subset of CNVs provided evidence of linkage disequilibrium with flanking SNPs (McCarroll et al. 2008), a significant portion of CNVs fell in genomic regions not well-covered by SNP arrays, such as SDs, and thus were not genotyped (Locke et al. 2006; Estivill and Armengol 2007; Campbell et al. 2011). Combining CNV and SNP data in human genome-wide association studies has associated CNVs with diseases such as intellectual disability, autism, schizophrenia, neuroblastoma, Crohn's disease, and severe early-onset obesity (de Vries et al. 2005; Sharp et al. 2006; Sebat et al. 2007; Cook and Scherer 2008; Bochukova et al. 2009; Diskin et al. 2009; Glessner et al. 2009; Shi et al. 2009; Stefansson et al. 2009).

Comparative genomic hybridization (CGH) and SNP arrays are routinely used for CNV screens, and their performances have been extensively reviewed (Lai et al. 2005; LaFramboise 2009; Winchester et al. 2009; Pinto et al. 2011). Although these platforms offer some detection power in SD regions, they are often affected by low probe density and cross-hybridization of repetitive sequence. In addition, only a relative copy number (CN) increase or decrease is reported with respect to the reference individual in array comparative genomic hybridization (aCGH). This poses a particular problem in the detection of CNVs in SD regions, as the test individual's CN may differ from that of the reference by

[8]**These authors contributed equally to this work.**
[9]**Present address: Department of Computer Engineering, Bilkent University, Ankara 06800, Turkey.**
[10]**Corresponding author.**
E-mail George.Liu@ars.usda.gov.
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.133967.111.

22:778–790 © 2012 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/12; www.genome.org

a smaller proportion than is detectable using array-based calling criteria.

The advent of next-generation sequencing (NGS) and complementary analysis programs has provided better approaches to systematically identify CNVs at a genome-wide level. These sequence-based approaches, which are becoming more popular due to the ongoing developments and cost decreases in NGS, allow CNV reconstruction at a higher effective resolution and sensitivity. Different methods to detect CNVs using sequence data were presented in the 1000 Genomes Project pilot studies (Sudmant et al. 2010; Mills et al. 2011) and have been previously reviewed (Snyder et al. 2010). Read depth (RD) methods used to analyze the 1000 Genomes Project data contributed high-resolution CNV calls with the capability of determining exact CN values for each genetic locus in an individual (Sudmant et al. 2010). Specifically, mrFAST/mrsFAST and whole-genome shotgun sequence detection (WSSD) (Alkan et al. 2009; Hach et al. 2010; Sudmant et al. 2010) are able to construct personalized CNV maps in or near SD regions by reporting all mapping locations for sequence reads, whereas other RD methods consider only one mapping location per read. Since CNVs are often found in or near duplicated regions in the genome (Cheng et al. 2005; Marques-Bonet et al. 2009), mrFAST and mrsFAST are more appropriate for detecting CNV in duplication- and repeat-rich regions.

Recently, interest in CNV detection has extended into domesticated animals (Chen et al. 2009b; Fontanesi et al. 2009; Nicholas et al. 2009; Bae et al. 2010; Fadista et al. 2010; Liu et al. 2010; Ramayo-Caldas et al. 2010; Fontanesi et al. 2011; Kijas et al. 2011). For example, in ridgeback dogs, duplication of *FGF3*, *FGF4*, *FGF19*, and *ORAOV1* causes hair ridge and predisposition to dermoid sinus (Hillbertz et al. 2007). The "wrinkled" skin phenotype and a periodic fever syndrome in Chinese Shar-Pei dogs are caused by a duplication upstream of *HAS2* (Olsson et al. 2011). The white coat color in pigs and sheep is caused by a duplication involving *KIT* and *ASIP*, respectively (Moller et al. 1996; Norris and Whan 2008). The chicken peacomb phenotype was linked to a duplication near the first intron of *SOX5* (Wright et al. 2009). Similarly, partial deletion of the bovine gene *ED1* causes anhidrotic ectodermal dysplasia in cattle (Drogemuller et al. 2001). Given the heritability of CNVs and their higher rates of mutation, it is possible that CNVs may be associated with or affect animal health and production traits under recent selection. *Bos taurus indicus* are better adapted to warm climates and demonstrate superior resistance to tick infestation than *Bos taurus taurus* breeds (Porto Neto et al. 2011). Likewise, beef and dairy cattle breeds display distinct patterns in selected metabolic pathways related to muscling, marbling, and milk composition traits. It is possible that CNVs may be associated with these agriculturally important traits.

The availability of two alternative cattle reference genomes (Btau_4.0 and UMD3.0) (The Bovine Genome Sequencing and Analysis Consortium 2009; Zimin et al. 2009) has opened new avenues of cattle genome research. Using the Btau_4.0 assembly, we previously applied an approach combining MegaBlast and WSSD to detect cattle SD and discovered 94.4 Mbp of duplicated sequence in the reference genome (Liu et al. 2009). Our earlier array-based studies in cattle have also uncovered significant differences in CNV frequency among breeds, as well as several genes associated with CNVs like *ULBP* and *PGR* (Liu et al. 2010; Hou et al. 2011). These studies confirm that CNVs are common, associated with SDs, and often occur in gene-rich regions in cattle. Here, we describe the first use of NGS data to detect CNVs in the cattle genome. Using mrsFAST and WSSD, we also analyzed genome-wide gene copy number estimates in order to explore their potential functional and evolutionary contributions to breed-specific traits. By providing the first individualized bovine CNV and SD maps and genome-wide gene copy number estimates, we enable future CNV studies into highly duplicated regions in the cattle genome.

Results and Discussion

CNV discovery and data set statistics

We obtained Illumina NGS data for four taurine (three unrelated Angus, one Holstein) and one indicine (Nelore) cattle (Supplemental Table S1). Additionally, we simulated NGS reads using Sanger sequence reads of the sequenced cow, L1 Dominette 01449, a Hereford cow of European descent, and named its result as DTTRACE. The amount of sequence data for each animal varied from 4× (Hereford and Holstein) to nearly 20× (Angus and Nelore) coverage, allowing sufficient power to detect CNVs >20 kbp in length (Table 1). Since two of our animals (Holstein and Hereford) were sequenced primarily as single sequence reads, and we aimed to provide absolute genome-wide gene copy number estimates in this study, we used an RD detection method similar to that previously described (Alkan et al. 2009). See Methods for full details of mrsFAST alignment and WSSD CNV discovery parameters. Based on sequence RD against the reference genome (Alkan et al. 2009; Sudmant et al. 2010), we detected a total of 1265 unique CNV regions (CNVRs) across all analyzed individual animals (average length = 49.1 kbp), amounting to 55.6 Mbp of variable sequence or 2.1% of the cattle genome (Fig. 1). A full list of CNV calls can be found in Supplemental Table S2. As expected, the "uncharacterized chromosome" (chrUn), which consists of sequence that cannot be uniquely mapped to the genome, contains much variable polymorphic sequence (Liu et al. 2009). Our analysis indicated that 36.7 Mbp of chrUn (944 regions) may be copy number variable between individuals. Due to the shorter

Table 1. Sequence data sets and window estimates

Animal abbreviation	Breed	Platform	Number of reads[a]	Raw X coverage	Autosome reads per 5-kbp window	Autosome reads STDEV	Duplications	Deletions	Variable nucleotides
BINE	*B. t. indicus* Nelore	Illumina GAIIx	1,294,595,641	19	1034.65	285.79	803	64	35.3 Mb
BTAN1	*B. t. taurus* Angus	Illumina GAIIx	1,177,885,036	17	1328.19	328.08	793	4	40.3 Mb
BTAN2	*B. t. taurus* Angus	Illumina GAIIx	1,318,356,916	19	1285.88	340.24	801	5	40.6 Mb
BTAN3	*B. t. taurus* Angus	Illumina GAIIx	1,219,531,192	18	1019.7	283.89	798	7	40.5 Mb
BTHO	*B. t. taurus* Holstein	Illumina GAIIx	287,255,229	4	237.38	57.76	751	3	37.7 Mb
Dominette (DTTRACE)	*B. t. taurus* Hereford	Sanger	307,909,731	4	340.25	95.07	668	0	36.4 Mb

[a]Total number of reads after filtering for quality scores and sectioning the reads into nonoverlapping 36-bp fragments.

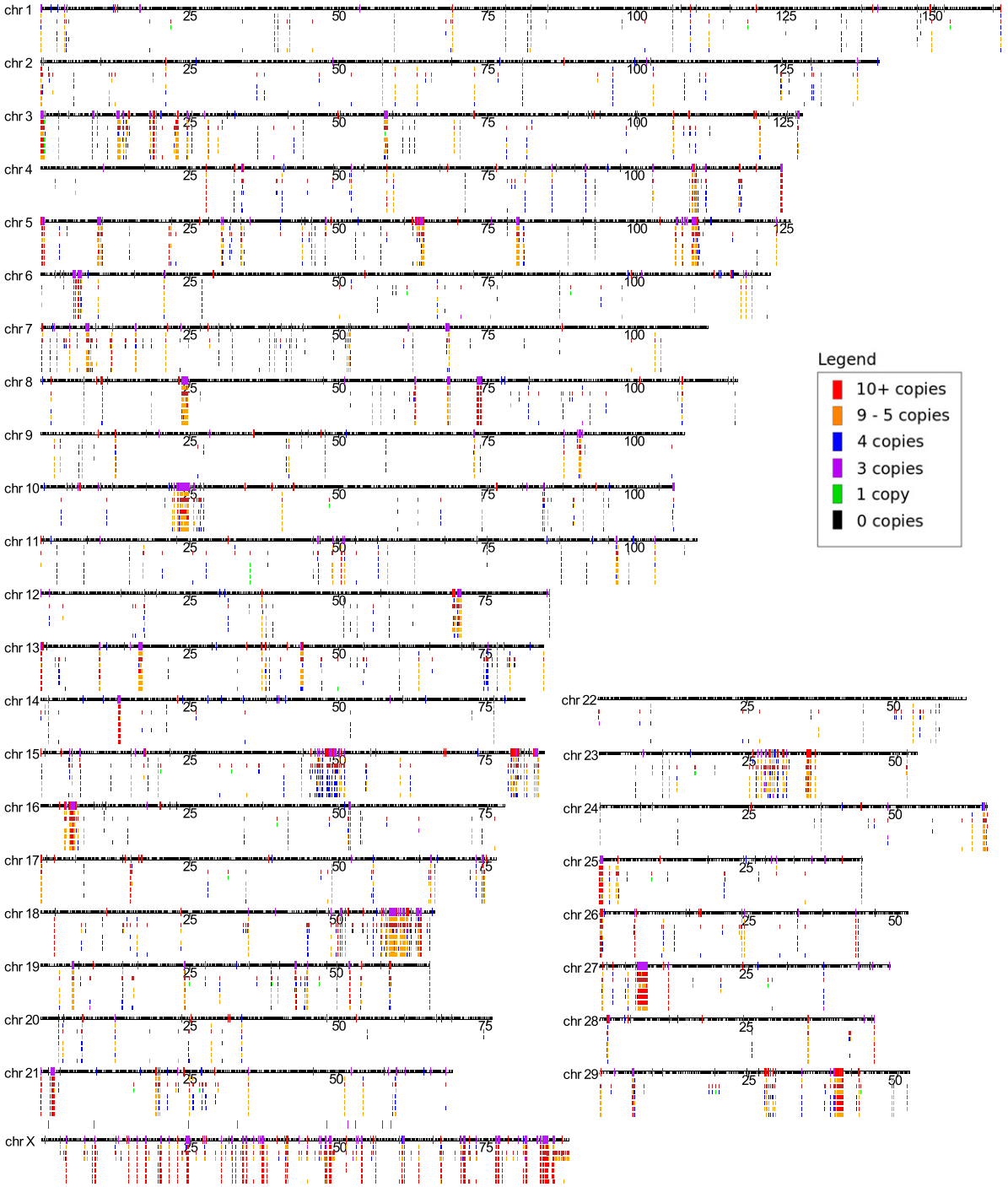

Figure 1. Individualized cattle CNV map. The Btau_4.0 assembly is represented as black bars with assembly gaps indicated by white boxes on the chromosomes. Larger bars intersecting the chromosomes represent the previously discovered WSSD (red), WGAC (blue), and WSSD/WGAC joint-prediction (purple) regions. Tracks underneath the chromosomes represent the CNV data sets (in order from *top* to *bottom*) for DTTRACE, merged CNVRs from all data sets, BINE, BTAN1, BTAN2, BTAN3, and BTHO. The colors for each bar in the animal data set tracks represent the average estimated CN for each CNV as shown in the legend. The merged CNVR track does not have CN information and is uniformly colored brown.

lengths of the chrUn contigs and the ambiguous mapping of chrUn sequence reads, candidate CNVRs on chrUn require further investigation. While our method had sufficient power to detect duplications, variance in RD across the autosomes—measured in standard deviations (STDEVs)—limited our discovery to only the extreme deletion events (Table 1). In the following analyses, we focused on further characterization of the high-confidence CNVRs (mostly duplications) from Btau_4.0 known chromosomes.

We constructed duplication maps for each of the six genomes and estimated the absolute copy number of each 1-kbp genomic

interval and over 9000 annotated cattle RefSeq genes. We compared the extent of overlap among duplicated sequences and reclassified duplicated sequences as shared or specific to an individual based on the predicted copy numbers in the analysis of these genomes. A significant proportion of the CNVs were shared among all three Angus (BTAN) individuals (35.7 Mbp out of 45.3 Mbp shared; 78.8%) with BTAN2 having the fewest unique CNVs (Supplemental Fig. S1). Apart from the *RNASE1* (pancreatic ribonuclease) gene, where BTAN10 (BTAN3) has a significant duplication compared to all other animals (CN of five vs. two for each animal), most of the CNVs were shared among the Angus individuals, with few novel CNV events among them. To simplify discussion, we used BTAN2 as a representative Angus bull in our comparisons of CNVs across the breeds due to its lower content of unique CNVs among the Angus individuals. A total of 23.4 Mbp of large SDs is shared among the four cattle breeds (36% of the intervals and 43% by base pair) (Supplemental Fig. S1).

We found the greatest CNV diversity for the indicine Nelore individual, BINE, with a total of 245 CNVs corresponding to 5.9 Mbp of sequence (Supplemental Fig. S1). Pairwise comparisons of BINE to the taurine individuals also yielded consistently lower shared CNV nucleotide space, suggesting that CNV differences between subspecies are greater than across the breeds within a subspecies. DTTRACE and BTAN2 shared a significant proportion of CNV space (33.9 Mbp out of 55.6 Mbp shared; 61%) and CNVRs (574 BTAN2 CNVs overlapped with 581 DTTRACE CNVs; 71% and 87%, respectively), likely due to the close relationship between the Hereford and Angus breeds, both of which are used for beef production. The second largest shared space was found in a three-way comparison of BTAN2, BTHO, and DTTRACE (6.6 Mbp), indicating that more CNVs are shared among taurine breeds than between the taurine and indicine subspecies. This result is consistent with previously reported breed phylogenies based on SNP analyses (Decker et al. 2009). Additionally, BINE had significantly more small CNV events (185 CNVs under 15-kbp in length) than the other animals (94–148 CNVs under 15-kbp in length) (Supplemental Fig. S2).

Comparison with published cattle SD results

Using the sequenced Hereford cow Dominette, we compared its CNV call set (DTTRACE) with our published cattle SD to validate our method (Liu et al. 2009). We trimmed Dominette's Sanger sequence reads to simulate 36-bp sequence reads similar to those found in several of our Illumina GAIIx sequenced libraries. Comparisons of DTTRACE against autosomal WSSD calls from the previous study (Liu et al. 2009) revealed 67% (25.8 Mbp/38.5 Mbp DTTRACE CNV total length) overlap of base pairs (Supplemental Fig. S3) and a similar amount of duplicated bases across the placed chromosomes (Supplemental Table S4).

We next assessed the ability of our method to accurately predict the sizes of published duplicated sequences. We first extracted and filtered 1020 duplication intervals from duplicated sequence identified by WGAC (whole-genome assembly comparison) and WSSD from the SD study (Liu et al. 2009), requiring intervals >20 kbp in length with <80% of their sequence occupied by common repeats. We then compared the 332 remaining SD intervals with similarly filtered CNVs from DTTRACE (Fig. 2A). We found a strong correlation (r = 0.904) between these two data sets. This confirmed that large candidate CNV (length >20 kbp) calls can be made with high confidence as previously reported (Alkan et al. 2009). Correlations of estimated CNV sizes between the two studies had better agreement when CNV size predictions were below 100 kbp (~63% of shared CNVs). By contrast, the original SD study predicted larger CNV sizes when lengths were above 200 kbp (~37% of shared CNVs). We suspect that merging the duplication predictions from the two different methodologies used in our earlier SD study (WGAC and WSSD) increased the size of SD predictions with respect to our new prediction, overinflating the predicted sizes for the SDs >200 kbp.

Also, several differences in methods likely contributed to the discrepancies between the previous SD study and our new data set. First, different alignment programs were used in these two studies. The original SD study used MegaBlast to align full-length Sanger reads to a lightly masked Btau_4.0 assembly. Our method used

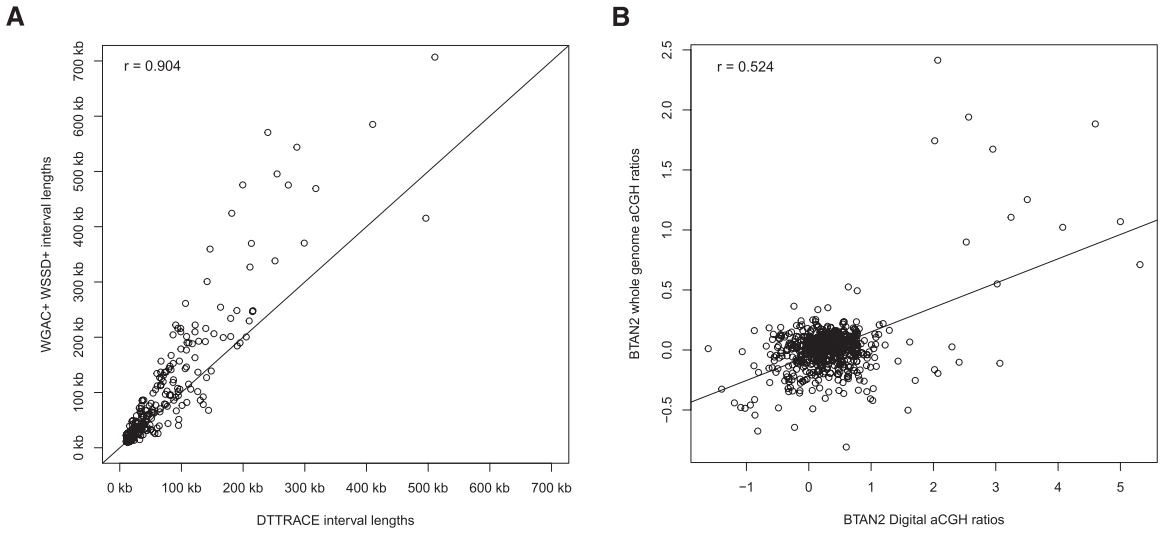

Figure 2. Correlation between computational predictions and experimental validations. (*A*) A good agreement of lengths (r = 0.904) exists between previously discovered WSSD+, WGAC+, and predicted DTTRACE duplications. (*B*) Calculated digital aCGH probe values (BTAN2_ngs) were compared with probe $\log_2$ ratios from a whole-genome aCGH (BTAN2_whole). Digital aCGH values were estimated using a $\log_2$ ratio of the 1-kbp CN windows from BTAN2 divided by CN estimates from DTTRACE. A moderate correlation (r = 0.524) was found for aCGH probe values and digital aCGH values within CNV intervals >20 kbp that had fewer than 80% of their lengths occupied by common repeats.

mrsFAST (Hach et al. 2010) to align trimmed Sanger reads to a thoroughly masked Btau_4.0 assembly. Only common repeats with <10% sequence divergence and bovine-specific repeat sequences were masked in the SD study (Liu et al. 2009). By contrast, we heavily masked Btau_4.0 with RepeatMasker, Tandem Repeats Finder, and WindowMasker and further extended 36 bp in both directions to remove edge effects for mapping short reads (see Methods). Although the smaller (<5 kbp) CNVs that did not overlap with the previous SD results could represent true duplications, we could not rule out the potential for mapping differences resulting from the use of short reads and different alignment programs. Therefore, we considered these calls separately (as "artifacts") and removed them from our current call sets to focus on higher confidence CNVs.

Experimental validation

We performed extensive experimental validation to confirm individual copy number variants, including aCGH, quantitative PCR (qPCR), and fluorescent in situ hybridization (FISH). We performed four aCGH experiments using BTAN1, BTAN2, BTHO, and BINE as test samples and Dominette as the reference for all experiments. To test the potential for our method to generate false positive results, we sought to confirm diploid regions (CN = 2) by our method with the array results. Since BINE shows the greatest CNV diversity and the lowest overlap with DTTRACE, it was the most divergent sample available to test the variability within the predicted two-copy regions. We selected all 1-kbp genomic regions and excluded all windows that intersected with known CNVs. Based on their RD values, we determined that these invariant diploid regions had an average CN of 1.967 and a STDEV of ~0.215. To make the CN estimates comparable to the aCGH results, we created log ratios of BINE CN estimates with DTTRACE CN estimates similar to the previously described digital aCGH approach (Sudmant et al. 2010). We then matched and compared the digital aCGH values with the actual aCGH probe log ratios based on their proximity to each other. We defined each intersection as being congruent if both values were within two STDEVs of the baseline of 0 (indicating a two-copy state in both individuals). Using this model, we found that only 8% of the matched ratios were significantly divergent from the expected baseline. This divergence percentage (~8%) suggests that our false discovery rate within the unique regions was low even within a comparison of different subspecies (*Bos taurus indicus* vs. *Bos taurus taurus*). Using the same digital aCGH approach, we compared our RD predicted CNV intervals with aCGH results. Based on predicted CN values within filtered CNVs (>20 kbp that contained <80% common repeat content), we generated digital aCGH values and compared them to aCGH probe $\log_2$ ratios using a linear regression model (Fig. 2B). Pearson's correlation values (r) ranged from 0.429 to 0.524 among the taurine individuals (BTAN1, BTAN2, and BTHO in Supplemental Fig. S4, Fig. 2B, and Supplemental Fig. S5, respectively). The lowest correlation was found for BINE (r = 0.264) (Supplemental Fig. S6). Discrepancies between the digital and experimental aCGH may be partly explained by the diminishing ability of aCGH to determine absolute differences between highly duplicated segments (Locke et al. 2003). Additionally, discrepancies for BINE are probably related to the sequence divergence between the indicine and taurine subspecies that could have influenced probe hybridization and sequence alignment. Indeed, a brief survey of SNPs within our sequence alignment files indicated that BINE had over twice the number of SNPs (897,124 SNPs) than the Angus individuals (average: ~400,000 SNPs) and nearly nine times the amount of SNPs than BTHO (113,775 SNPs; data not shown). Under our settings, mrsFAST allowed for up to two mismatches in a sequence read during alignment, which may have influenced BINE predicted RD in regions of the genome that were divergent from the taurine reference Btau_4.0 assembly. Another limitation of mrsFAST is its omission of gapped alignments in order to save computational time (Hach et al. 2010). However, these limitations were less likely to influence our CNV calling, as we only focused on the larger CNVs (>10 kbp). Until a *Bos taurus indicus* reference genome assembly is available, our CNV calls for the BINE individual can only be based on the Btau_4.0 assembly. Even so, CNVs predicted in BINE were likely to contain fewer false positive CNVs in non-variable, two-copy regions as demonstrated by the low 8% false discovery rate (FDR), suggesting our calls likely represent true variation in the BINE genome.

Quantitative PCR assays were designed to confirm previously unreported CNVs as well as CNVs within or near annotated genes. We chose to investigate 12 predicted CNVRs in different animals, using two distinct primer sets per locus (see Supplemental Table S4). Our qPCR analysis used a modified ddCT method to determine relative CN as described previously (Hou et al. 2011). The only exception was that we used absolute CN estimates from the DTTRACE data set at each qPCR locus that was not diploid in Dominette (our control sample). Using this correction, we found that 82% of our qPCR results (46 confirmed/56 total) agreed with our CNV predictions in these regions (see Supplemental Table S5). If we counted CNVs as confirmed only when both qPCR primer sets were positive, we had 70% agreement (23 confirmed/33 total) under these stringent criteria. The discrepancies between the qPCR and WSSD results may represent small CNV events missed in the WSSD calls. Likewise, instances where SNPs or small indels existed among individuals may have caused the qPCR assay to report a different value from the WSSD analysis.

We also attempted to validate CNV calls using FISH; however, the lack of cell lines from the same individuals forced us to use cell lines from related animals from the same breeds. We tested CNV predictions using a FISH analysis on interphase nuclei within three cell lines derived from Hereford, Angus, and Holstein animals, respectively, using 51 large insert BAC clone probes that corresponded to predicted CNVRs. Out of 51 tests, 28 probe locations (~55%) showed a detectable change in CN among the cell lines (see Supplemental Table S6; Supplemental Fig. S7), showing variable signal numbers either among three cells (8) or between haplotypes (20). The results of all FISH experiments are available online at http://bfgl.anri.barc.usda.gov/cattleCNV/. All 28 duplication signals were tandemly clustered. Similar to the mouse and dog genomes (She et al. 2008; Nicholas et al. 2009), these data reinforce that tandem distributions of CNV are predominant in the cattle genome (Fig. 1). Discrepancies between our predicted CNVRs and the FISH results are most likely due to the difference in animals used in each analysis. In addition, we also note that BAC-FISH has a limited ability to detect tandem duplications and duplications <40 kbp in length.

Other causes, such as the draft status of the cattle genome assembly, make it difficult to determine CNVs on the sex chromosomes. For example, chrX presented a challenge as all five studied cattle were bulls, and chrY currently exists only in putative contigs. Even so, nearly 80% of the CNVRs (69/87) on chrX identified in the taurine individuals agree with previously predicted SD regions (Liu et al. 2009). Several predicted CNVs in BINE were detected near the end of chrX (Fig. 1), which contains the

pseudoautosomal region (PAR) (Sonstegard et al. 1997). It is possible that reads from the PAR of chrY were instead mapped to chrX, thereby inflating RD values within that region of the chromosome.

Copy number polymorphic genes

Using the cattle RefSeq gene annotations, we identified copy number polymorphic genes and then assigned a CN estimate to each gene. A total of 413 out of 9571 (4.3%) RefSeq gene transcripts overlapped putative CNVRs, while 9158 (95.7%) did not. Gene transcripts outside of predicted CNVRs were found to have a median CN estimate of 2.01, again reconfirming a low false positive rate in predicted nonvariable regions. The overlapping gene transcripts were found to be highly variable in CN among individuals (minimum value: 0.10; maximum: 242; median: 3.77; average: 5.80). Fifteen of the 25 most variable genes had functions related to the immune response and host defense, such as the interferon, defensin, and GIMAP (GTPase, IMAP) families (Table 2). High CN genes often belong to multiple-member gene families, and it is likely that our CN estimates are actually representative of real duplications of these paralogs. For complete lists of CNV gene intersections, see Supplemental Table S7. Of the 413 overlapping genes, 195 were found to be completely overlapped by CNV intervals. Genes covered by CNVs varied on an individual basis, with BINE having the most genes intersecting with predicted CNVs (313 genes), BTHO having the least (239 genes), and 178 genes being shared by all studied animals. Of the 178 shared genes, multiple members were previously identified in other studies, such as the olfactory receptors and the beta-defensin gene families (Liu et al. 2010; Hou et al. 2011). Other previously identified CNV-gene overlaps were detected, such as the BSP30 gene family (Liu et al. 2009). This BPI-like protein gene family has been through several ruminant-specific amplifications, and each gene family member encodes a highly transcribed, salivary, anti-microbial peptide (Haigh et al. 2008). *BSP30A* and *BSP30B* are two specific members of this gene family (Wheeler et al. 2007) found to have CNVR overlaps (Fig. 3A). All five sequenced animals had predicted CNVs within the region (average CN: 7–11), with notably smaller CN values predicted upstream of the *BSP30A* locus in Dominette (average CN: 4). These CNV predictions were confirmed by whole-genome aCGH, with all animals showing consistently higher CN counts than Dominette near the *BSP30A* locus. *BSP30A* and *BSP30B* comprise 15%–30% of the total protein content of bovine saliva, making them an important first defense enzyme against orally ingested parasites/bacteria (Haigh et al. 2008). Duplications in this region may be a ruminant-specific response to evolutionary pressures from soil-based parasites and bacteria encountered while grazing.

In some cases, CN varied widely among individuals especially within highly duplicated gene families (Table 2). For example, the transcription factor gene *SUHW2* (zinc finger protein 280B) had a large CN variance among animals (STDEV = 12.6). Dominette had the fewest copies of *SUHW2* at 2.6; BTHO and BINE: 22.2 and 22.9, respectively; and the three Angus bulls each had more than 30 copies of *SUHW2*. BINE was predicted to have the lowest CN values for 13 of the top 25 variable genes (P-value $<7 \times 10^{-5}$). Other CN variable genes of note include *CATHL4*, *KRTAP9-2*, *LAP*, *TAP*, and several *PAG* genes. *CATHL4*, an antimicrobial peptide coding gene, was found to have a higher copy number in BINE than in the taurine animals. Indolicidin, the protein product of *CATHL4*, can induce autophagic cell death in the parasite *Leishmana donovani*, the causative agent of the parasitic disease Leishmaniasis (Bera et al. 2003). This is particularly interesting as the indicine breeds are known for their increased parasite resistance compared to taurine breeds (Berman 2011). Knockouts of *CATH*-family genes in mice have revealed that the antimicrobial peptide products of

Table 2. Top 25 copy number variable genes in sampled individual cattle

Gene ID	RefSeq accession	Gene size (bp)	Covered percentage	BINE	BTAN1	BTAN2	BTAN3	BTHO	DTTRACE
MGC134093	NM_001077070	8620	100	48.2	199.6	176.9	103.8	242.0	215.0
MGC134093[a]	NM_001077070	8653	100	42.4	170.7	149.5	87.8	210.8	182.3
LOC780876	NM_001079796	7960	100	12.2	37.9	33.6	19.1	53.9	47.8
SUHW2	NM_001077935	8466	100	22.9	31.7	34.5	35.6	22.2	2.6
IFNB3	NM_001114297	639	100	38.0	46.9	25.4	23.2	27.5	48.2
IFNB1	NM_174350	637	100	33.0	35.4	22.6	21.7	17.6	30.6
FBXO16	NM_001078119	32,822	100	14.6	29.0	26.1	28.5	22.6	24.1
BNBD-4	NM_174775	1921	100	11.2	16.5	25.2	16.5	11.4	14.0
DEFB1	NM_175703	15,258	100	9.8	15.5	24.3	15.7	11.0	14.4
KRTAP9-2	NM_001105020	1464	100	4.5	9.0	19.0	13.5	14.4	13.0
LOC100126815	NM_001111069	8197	100	20.1	6.8	8.7	9.3	7.8	8.2
DEFB5	NM_001130761	1783	100	9.7	14.2	22.4	14.2	8.9	12.7
GIMAP1	NM_001083677	7827	97.06	5.6	2.1	11.9	13.1	5.7	1.9
ISG12(A)	NM_001038050	8891	89.07	3.6	12.3	6.6	13.6	4.9	2.9
LAP	NM_203435	1789	100	10.1	15.2	21.2	17.0	10.1	12.5
TAP	NM_174776	1819	100	7.4	13.3	18.7	15.2	8.3	11.6
BNBD10	NM_001115084	1677	100	10.4	15.5	21.3	18.4	11.1	13.8
PAG6	NM_176617	9330	100	7.2	16.2	17.5	18.0	12.4	11.3
BNBD10	NM_001115084	1678	100	11.4	15.0	20.3	18.3	11.0	12.9
TUBA1B	NM_001114856	3600	100	17.3	8.5	11.6	14.6	9.9	7.6
CATHL4	NM_174827	1374	100	13.5	6.6	7.5	11.6	4.0	6.7
PAG21	NM_176630	9079	100	5.3	13.0	13.5	13.9	9.5	9.6
GIMAP7	NM_001080257	8418	100	6.5	3.6	11.0	10.1	6.1	2.7
GIMAP4	NM_001046060	7149	100	5.7	3.8	10.4	10.4	6.0	2.8
PAG15	NM_176624	8862	100	4.8	12.0	12.9	13.0	8.9	9.1
PAG1B	NM_174411	9331	100	5.2	12.0	13.1	13.5	8.9	8.9

[a]Parts of MGC134093 are mapped to Btau_4.0 multiple times.

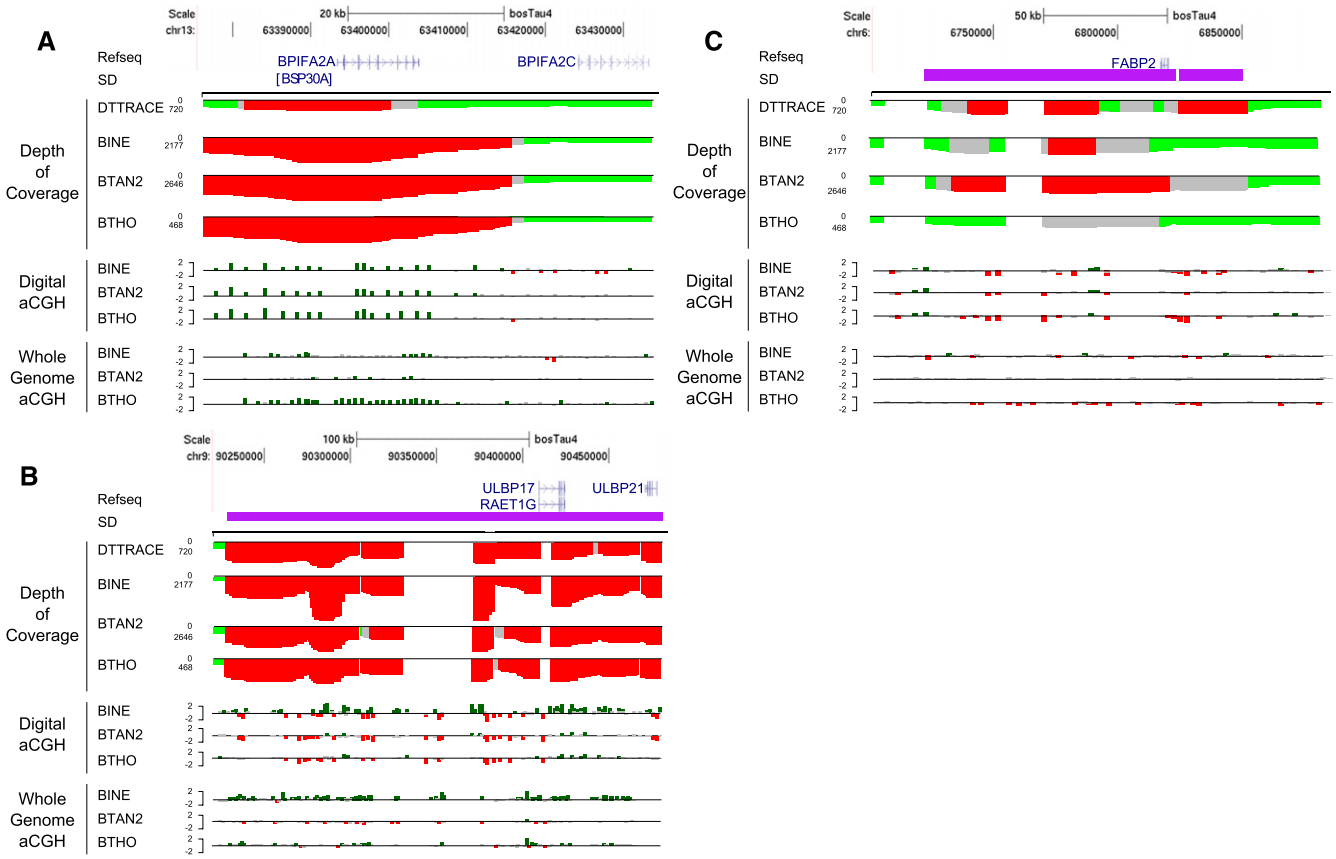

Figure 3. Computational predictions and aCGH validations of segmental duplication copy number differences for six cattle genomes. Depth-of-coverage tracks for DTTRACE, BINE, BTAN2, and BTHO are *below* a UCSC track for each investigated gene region. Regions colored in red on the plot indicate excessive read depth (> mean + 4 × STDEV), whereas gray regions indicate intermediate read depth (> mean + 3 × STDEV). Normal read depth values are colored green (mean ± 2 × STDEV). Digital aCGH tracks show the $\log_2$ ratio of the copy number of each listed animal compared to DTTRACE, with high values listed in green (>0.5); low values: red (<−0.5); and nominal values: gray ($0.5 > x > -0.5$). Whole-genome CGH array experiments, using Dominette as a reference sample in all cases, are listed *below* the digital aCGH experiments. Color schemes for the aCGH plots are the same as for the digital aCGH. Previously detected segmental duplications (SDs) are shown *below* the UCSC plot, if present in the region. (*A*) CNVs intersecting the *BPIFA2A (BSP30A)* locus (chr13:63364661-63487495). A duplication of this region was predicted for all animals and was confirmed by whole-genome aCGH. (*B*) In the *ULBP17* locus (chr9:90209622-90499803), BINE was predicted to have a higher copy number than DTTRACE across the region from both read depth and aCGH experiments. (C) The promoter region of *FABP2* (chr6:6701747-6888288) was a predicted duplication in Dominette (Hereford; beef breed), BTAN2 (Angus; beef), and BINE (Nelore; dual-purpose) but not in BTHO (Holstein; milk).

these genes influence Leishmaniasis lesion development and tissue colonization of the parasite (Kulkarni et al. 2011). Additionally, a study on the resistance of indicine cattle to Leishmaniasis found antibody production with no onset of disease symptoms despite contact with the sandfly carriers of *L. donovani* (Alam et al. 2011). Two other beta-defensin-like molecules, Lingual (*LAP*) and Tracheal (*TAP*) antimicrobial peptides, were found to have significant CNV among individuals. Both the *LAP* and *TAP* peptides are activated in response to bacterial infection and are expressed in a wide range of epithelial tissues, including mammary epithelial cells (Isobe et al. 2009; Lopez-Meza et al. 2009). CN changes of these alleles may influence host resistance to mastitis, which is an important economic trait in the breeding of dairy cattle. Additionally, BINE was found to have the lowest CN (4.5) of the *KRTAP9-2* gene, a member of the keratin-associated protein (KAP) gene family. Since cattle skin is the infestation site for ticks, collagens, keratins, and their associated proteins have been suggested to play a role in tick resistance. Wang and colleagues compared gene expression patterns in response to tick infestation using tick-resistant and -susceptible cattle skin samples (Wang et al. 2007). They reported that, in susceptible skins, *KRTAP9-2* showed more active transcription before infestation and a more dramatic reduction in transcription following infestation. If CN increases at the *KRTAP9-2* locus influence the gene expression in vivo, this would be a significant allele in the determination of tick resistance in cattle. The Interferon tau (*IFN-tau-c1)* and pregnancy-associated glycoprotein (*PAG)* loci were also highly variable in CN among individual animals. These gene families are involved in reproduction, with *IFN-tau-c1* influencing the maternal immune system's recognition of conceptus (Walsh et al. 2011) and the *PAG* genes serving as important secretory products of trophoblasts (Xie et al. 1997). Several *PAG* paralogs were identified as highly variable in CN, including three members of modern *PAG* groups (*PAG15, PAG6*, and *PAG21*) (Telugu et al. 2009). Given the detected inter-individual variability, the *PAG* family expansion may represent important differences in fertility and reproduction among the surveyed individuals.

We also detected CN differences for several interesting immune function-related genes in BINE as well as for several genes related to lipid metabolism and transport in the taurine individuals. Previous studies have identified *ULBP17* as a potential

highly duplicated gene in cattle (Liu et al. 2010; Hou et al. 2011). To our knowledge, this is the first study to assign estimates of CN to the region on an individual basis (Fig. 3B). While all animals were predicted to have a CNV overlapping the entire *ULBP17* locus, BINE had the highest predicted CN for the gene by a factor of two (14 CN compared to ~6–7 CN for taurine breeds). Another interesting CNV-gene intersection within the BINE individual was found for the last exons of the aldehyde oxidase 1 (*AOX1*) gene (Fig. 4A; Supplemental Fig. S8). This CNV is BINE-specific and was confirmed by whole-genome array and qPCR analysis. Aldehyde oxidase produces hydrogen peroxide and is often implicated in drug metabolism and detoxification (Garattini et al. 2009).

Promising breed-differential CNVs associated with lipid transport and metabolism were found in the taurine cattle, though their potential impacts on beef and milk production remain elusive. An interesting CNV was identified directly upstream of the *FABP2* locus in BTAN1, BTAN2, BTAN3, and Dominette simulation (Fig. 3C). *FABP2* encodes a small, fatty acid-binding protein expressed in the proximal portion of the intestines and typically binds bent-conformation fatty acids for transport across cell membranes (Glatz and van der Vusse 1996). Animals with this CNV include the taurine beef breeds (BTAN1, BTAN2, BTAN3: Angus; Dominette: Hereford) and BINE (Nelore, also beef). The human Ala54Thr allele of *FABP2* has been shown to have lower binding efficiency and a weaker promoter in vitro and is likely to contribute to the development of insulin resistance and lower lipid oxidation rates (Formanack and Baier 2004). We hypothesize that CNVs upstream of *FABP2* may be associated with feed efficiency and lipid uptake, i.e., variation of the *FABP2* locus could potentially increase its expression in the intestines and, therefore, increase fatty acid sequestration from feed in beef breeds.

Another lipid metabolism associated gene, apolipoprotein L, 3 (*APOL3*), was identified as a CNV in all animals (Fig. 4B; Supplemental Fig. S9). *APOL3* is expressed in all tissues in humans but has a higher expression in the prostate and placenta and is involved in the transport of cholesterol (Page et al. 2001). PANTHER molecular function analysis (Thomas et al. 2003; Mi et al. 2010) revealed an enrichment of lipid transporter activity proteins in

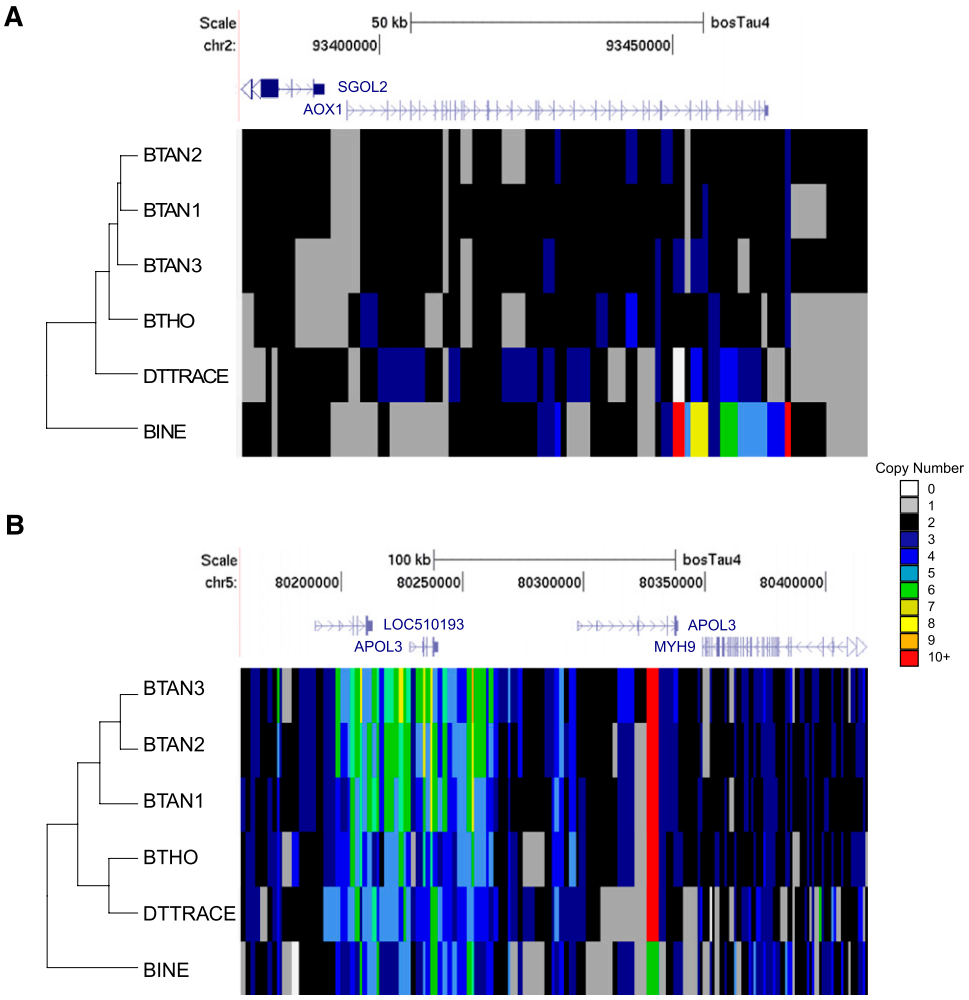

Figure 4. Cluster analysis of copy number variable genes in individual cattle. (*A*) Copy number values for each animal were plotted within the *AOX1* locus (chr2:93376314-93484307) using the color scheme depicted in the legend. Heatmap boxes represent 1-kbp sliding, nonoverlapping windows in the region. The dendrogram indicates the hierarchical ordering of animals based on a Pearson's hierarchical clustering of the CN values within the region. Within *AOX1*, the last exons are predicted to have a higher CN in BINE than in any other animal. This observation was confirmed using aCGH and qPCR. (*B*) A heatmap of *APOL3* reveals significantly higher CN in the three Angus animals (BTAN3, BTAN2, and BTAN1) for the first *APOL3* transcript (NM_001100297) than in the other breeds (chr5:80158821-80417344).

BTAN1, BTAN2, BTAN3, BTHO, and Dominette (P-value < 5.0×10^{-3}) compared to BINE, suggesting that the taurine breeds may have a larger number of CNV-associated genes in lipid metabolism and transport functions (see Supplemental Table S8). A network analysis also revealed enrichment of genes involved in lipid metabolism, supporting the PANTHER analysis results (see Supplemental Table S9).

Heatmap analyses

To provide an evolutionary perspective to our analyses, we created heatmaps with Pearson's correlation-based hierarchical clustering using the CN values for regions within selected gene loci. *AOX1* was a confirmed difference between BINE and the taurine cattle and is particularly evident within a CN heatmap analysis (Fig. 4A). Hierarchical clustering of animals based on CN content within the *AOX1* locus mirrored the evolutionary history of the studied cattle breeds based on SNP genotyping (Decker et al. 2009), with BINE as the clear outlier and the Angus individuals in a branching clade. BTHO and Dominette were arranged sequentially next to Angus individuals. A hierarchical clustering of CN values near the *APOL3* loci again grouped the Angus animals and set BINE as the outlier; however, the BTHO and DTTRACE lineages split into a separate clade (Fig. 4B). Therefore, heatmap analyses of breed-specific CNVs within the sequenced individuals generate cluster trees consistent with the generally accepted breed history (Decker et al. 2009). We predict that future sequence-based CNV studies using a larger sample size and outgroups will find CNV selective sweeps within the cattle breeds.

While both taurine and indicine individuals had CNVs that intersected immune function-related genes, the distinct resistance traits of zebu cattle (Berman 2011) lend great importance to the study of *Bos taurus indicus*-specific immune function gene CNVs. In that regard, we have identified several CNVs that may represent variations between BINE and the remaining *B. t. taurus* animals. *ULBP17* and *CATHL4* appear to have been recently duplicated in BINE, suggesting that these gene expansions are in response to increased viral and bacterial/helminthic pathogens, respectively. It is also interesting to note the lower copy numbers of *KRTAP9-2* and other genes in BINE.

Overlap with SD and other genomic features

We next sought to categorize the overlap of CNVs with other genomic features, and we found significant overlap with the previously identified SD regions (Liu et al. 2009), human disease gene orthologs, and cattle quantitative trait loci (QTL). We overlapped CNVRs with SDs and found that 64.5% (35.9 Mbp) of CNVRs overlapped with SDs. We tested the significance of this result by generating 1000 random, simulated CNVRs and checked their proximity to known SD regions (Supplemental Fig. S13). Only 3.0% of the simulated CNVs directly overlapped with SDs, compared to the 46.7% of our observed CNVRs, suggesting a 15.6-fold enrichment. Increasing the flanking regions of the SDs by 100 kbp on both sides increased the number of overlapping CNVs (6.6% simulated and 56.5% of observed CNVRs). We noted that the tandem cluster pattern of cattle CNVs is a dramatic contrast when compared to a preponderance of interspersed duplications of human CNVs. A strong correlation of CNVs and SDs in cattle further supports the hypothesis that their formation mechanisms are mainly due to nonallelic homologous recombination (NAHR) (Liu et al. 2010). Several CNVRs that spanned QTL and human orthologous Online Mendelian Inheritance in Man annotations were identified. For instance, multiple CNVRs directly overlapped with QTL for marbling (intramuscular fat content), carcass weight, milk yield, and clinical mastitis (see Supplemental Table S10). Out of 1265 total CNVRs, 211 (16.7%) overlapped genes associated with human diseases, including intellectual disability, autism, schizophrenia, and Crohn's disease. Other overlapping QTL were involved in many production and reproduction traits, such as marbling score, calving ease, gestation length, pregnancy rate, and inseminations per conception. Such regions warrant future study to determine the extent that CNVs may contribute to QTL.

Conclusions and future directions

Our study presents the first description of sequenced-based CNV within cattle genomes based on analyzing the genomic sequence of six individuals. By considering all possible map locations for a read in an efficient manner, we have been able to leverage the dynamic range of NGS reads to accurately predict absolute copy number of some of the most structurally complex regions of the cattle genome for the first time. We identified a total of 1265 unique CNVRs in five *Bos taurus taurus* (three Angus, one Hereford, and one Holstein) and one *Bos taurus indicus* (Nelore) individuals. We found the patterns of the bovine SDs vary greatly, with only 40% of the duplications being shared. We also confirmed that the most extreme CNV corresponds to genes embedded within SDs (a 15.6-fold enrichment), and most of these differences involve tandem changes in copy as opposed to duplicative transpositions to new locations. These results provide a prelude to a 1000 Cattle Genomes Project, which could lead to a deep catalog of cattle CNVs by population-scale genome sequencing.

It is important to note that only one individual per each breed was studied here except for the Angus breed. Any breed level inferences need to be further tested with a larger sample size. While our study included only one *Bos taurus indicus* individual, CNVs identified in BINE reveal promising areas for future research into indicine-specific CNVs of economic relevance. Given the sequence divergence of BINE from the taurine animals, it is likely that some CNV calls may have been missed and/or CN may have been underestimated by mapping BINE's reads to a taurine reference genome. Notwithstanding, our observations that BINE shows the greatest diversity in CNV (867 events) is consistent with it being more distantly related. The creation of a separate reference assembly for indicine cattle would facilitate resolving potential differences in chromosome and gene structure that may have been detected by our study.

We have detected 413 complete genes as copy number variable among six cattle. We report breed-specific copy number differences in a Nelore individual as excellent candidates for pathogen and parasite resistance (*CATHL4*, *ULBP7*, and *KRTAP9-2*). In addition, copy number differences were detected for several lipid metabolism and transport genes in the taurine individuals. This study also provides new CNVs and CN estimates across the cattle genome, enabling further research into highly duplicated gene families and chromosome segments. Although these genes warrant future investigation, the ability to use NGS to accurately predict their copy number provides the first step to make genotype and phenotype correlations in these complex areas of the genome. The next challenge will be further definition of the sequence content and structural organization of these dynamic and important regions of the cattle genome through population-level sequencing including trio (parents and offspring). The long-term

goal is to identify CNV-associated economic traits and incorporate them into animal genomic selection systems.

Methods

Sequencing and data acquisition

Based on the breed history and pedigree information, breeds and individuals were selected as the representatives of the modern cattle population. The chosen breeds and their origins and features are summarized in Supplemental Table S1, including three *Bos taurus taurus* breeds—Angus, Hereford, and Holstein—and one *Bos taurus indicus* zebu breed—Nelore. Genomic DNA samples were purified from semen or blood as described (Sonstegard et al. 2000). All DNA samples were analyzed by spectrophotometry and agarose gel electrophoresis.

Four taurine bulls (BTAN1, BTAN2, BTAN3, and BTHO) and one indicine bull (BINE) were sequenced using both single- and paired-end libraries on the Illumina GAIIx; however, most of the NGS data from BTHO (~57%) and some of the data from BINE (~24%) were in single-end read libraries. Since two of our animals (Holstein and Hereford) were sequenced primarily as single sequence reads and we aimed to provide absolute genome-wide gene copy number estimates in this study, we only used an RD detection method similar to that previously described (Alkan et al. 2009). We excluded sequence reads if they had a first base quality score of two. When needed, we trimmed longer reads into nonoverlapping 36 bp to reduce the read length heterogeneity prior to sequence alignment. For the sequenced Hereford cow, L1 Dominette 01449 (Dominette), we downloaded its Sanger sequence reads from the NCBI Trace Archives (ftp://ftp-trace.ncbi.nlm.nih.gov/; DTTRACE; also see Experimental Validation). After clipping identified vector sequences, we trimmed the remaining Sanger read sequence into nonoverlapping 36-bp fragments. As detected in previous simulations of increasing coverage (Alkan et al. 2009; Waszak et al. 2010; Mills et al. 2011), a genome coverage greater than fourfold is sufficient for the RD detections of CNVs.

Sequence alignment

Since most of genome annotations, including our earlier SD analysis, were based on the Btau_4.0 assembly, we used that genome assembly to align sequence reads. Repeats were masked using RepeatMasker (Smit et al. 1996) (using the -s option and cattle RepBase libraries), Tandem Repeats Finder (Benson 1999), and WindowMasker (Morgulis et al. 2006). Masked regions were further extended by 36 bp in both directions to reduce edge alignment effects (Sudmant et al. 2010). We then aligned ~5.3 billion 36-bp reads (~190 gigabases) to the masked Btau_4.0 using mrsFAST (Hach et al. 2010), allowing up to two mismatches (i.e., 34/36, ~94.4% sequence identity). Approximately 20% of the raw reads were mapped to the unmasked portion of the genome (~40%) with an average mapping count of 1.1 per read.

Read depth analysis

Aligned reads within sliding windows were then processed using the WSSD pipeline as previously described (Alkan et al. 2009). This approach uses three different sizes and types of windows to call CNVs, refine their breakpoints, and determine CNs within a particular region. Reads were first counted in overlapping, sliding 5-kbp windows of nonmasked, nongapped sequence. The GC bias of the Illumina GAIIx platform was corrected using LOESS smoothing toward a pattern of uniform coverage at all GC percentage bins as previously described (Alkan et al. 2009). Assuming that most of the genome is in a diploid state, the corrected RD mean and standard deviation (STDEV) were calculated for each individual (Table 1). Since five bulls were used, chrX was analyzed separately, and its RD value was not used to determine thresholds for CNV calling.

CNV calls were initially made using conservative criteria and were subsequently refined using higher resolution settings to determine breakpoints. Initial calls were selected if six out of seven or more sequential 5-kbp overlapping windows had RD values that varied significantly from the average (duplications > mean + 4 × STDEV; deletions < mean − 3 × STDEV). Calls were then refined using the GC% corrected RD means from 1-kbp overlapping windows, albeit with a less stringent cutoff value (duplications > mean + 3 × STDEV). Deletions were not refined in this fashion, given their less stringent calling criteria. Only CNV calls >10 kbp in length were kept in the final data set.

Finally, CN was estimated within 1-kbp nonoverlapping windows across all placed chromosomes. These nonoverlapping estimates of CN serve as a good approximation of CN within nonmasked, nongapped regions of the genome.

DTTRACE simulation and artifact removal

Removal of short-read mapping artifacts was performed using DTTRACE simulation as previously reported (Alkan et al. 2009). As described above, clipped and trimmed 36-bp sequence fragments from the Dominette trace data were aligned to the masked Btau_4.0 assembly using mrsFAST. SD intervals were predicted using the same parameters used with the real Illumina WGS read sets. We then compared these predictions to our published cattle duplications (Liu et al. 2009) and classified any intervals (or subintervals) as short-read mapping "artifacts" if they did not agree with the known duplication set. Such regions were subsequently removed from the SDs predicted in all six cattle genomes. A total of 5.2 Mbp of autosomal artifact regions was identified and removed from all CNV call sets.

Identification of cattle CNVs using aCGH

Whole-genome high-density CGH arrays manufactured by Nimblegen containing ~2,166,464 oligonucleotide probes (NCBI GEO accession no. GPL11314) were designed and fabricated on a single slide to provide an evenly distributed coverage on UMD3.0 with an average interval of ~1.2 kbp between probes. Standard genomic DNA labeling (Cy3 for samples and Cy5 for references), hybridizations, array scanning, spatial correction, and data normalization were performed as previously described (Liu et al. 2010). Since we aligned to Btau_4.0, probe coordinates were migrated from UMD3.0 to Btau_4.0 using liftOver (http://hgdownload.cse.ucsc.edu/admin/exe/). Approximately 95% (2,066,074/2,166,464) of the probes were successfully converted. Segmentation was performed using the segMNT v1.1 algorithm in NimbleScan ver2.5. We then tested a series of $\log_2$ ratio shifts (0.5 and 0.3) and affected neighboring probe counts (3 and 5) to evaluate their impact on the FDR in the self-self control hybridization. We chose the 0.5 $\log_2$ shift and three neighboring probe criteria (0.5_3) to call CNVs, under which no false-positive calls were found in self-self hybridization experiments. Since all test samples were from bulls (one X chromosome) and our reference was a cow (two X chromosomes), we shifted the chrX baselines to negative values. High-confidence calls were subsequently filtered and merged as previously described (Liu et al. 2010). Because of the strict filtering criteria, a substantial false-negative rate was expected.

qPCR validation

Primers were designed using a custom script that incorporated Primer3 (http://frodo.wi.mit.edu) and Exonerate (http://www.

ebi.ac.uk/~guy/exonerate/) to identify unique binding sites for primer design. Only the following Primer3 settings were changed from default values: The amplicon length was set to 150–250 bp, and the GC clamp value was set to 2. Primer information is shown in Supplemental Table S4. qPCR experiments were conducted using SYBR green chemistry in triplicate reactions, each with a reaction volume of 25 μl, as previously described (Hou et al. 2011). Reactions were amplified on a BioRad MyIQ or iQ5 thermocycler. An intron-exon junction of *BTF3* was chosen as a reference location for all qPCR experiments. Analysis of resultant crossing cycle thresholds (C_T) was performed using the relative comparative C_T method. Calibrations of C_T values were derived from amplification of reference and test primers on Dominette genomic DNA. The CN estimates for nonoverlapping 1-kbp windows in the DTTRACE data set were used as Dominette's expected copy numbers. The copy number for each test region was calculated as $2^{(1+ddCT)}$. Agreement of the estimated test copy number with the expected reference copy number was determined using simple heuristics that incorporated the difference of the DTTRACE predicted CN from a value of two into the estimated test copy number (Supplemental Figs. S10–S12). To reduce batch and platform effects, plates were designed to amplify the reference gene and Dominette in each experiment.

FISH validation

FISH experiments were performed as previously described (Ventura et al. 2003; Liu et al. 2009). Fifty-one cattle BAC clones (CHORI-240) were selected with large (≥20 kbp) copy number variable regions. Both interphase and metaphase nuclei were prepared using three cell lines from Coriell Cell Repositories (AG08501: Hereford male smooth muscle cell; AG08423: Angus female fibroblast; and AG10375: Holstein male fibroblast). A single BAC clone (297K6) was used as a control in each FISH experiment. Differentially labeled test and control BAC clones were cohybridized to one slide. To determine the copy number of the test BAC, we calculated the ratio between the number of signals of the test BAC and the number of signals of the control BAC. We counted 40–50 nuclei for each slide and reported their averages. Metaphase nuclei were examined to identify the chromosomal origins of FISH signals. More intense FISH signals, which localized to a single site, were subsequently examined by interphase nuclei. Interphase analyses were controlled for replication by comparing cells at both G_1 and G_2 stages of arrest.

Gene content

Gene content of cattle CNVRs was assessed using Ensembl genes (ftp://ftp.ensembl.org/pub/current_fasta/bos_taurus/pep/), the Glean consensus gene set, cattle RefSeq, and in silico mapped human RefSeq (the UCSC Genome Browser website at http://genome.ucsc.edu/). We obtained a total of 26,977 bovine peptides from Ensembl. In addition, using the PANTHER classification system, we tested the hypothesis that the PANTHER molecular function, biological process, and pathway terms were under- or overrepresented in CNVRs after Bonferroni correction (Liu et al. 2010). It is worth noting that a portion of the genes in the bovine genome have not been annotated or have been annotated with the designation "unknown function," which may cause an underestimation of the influence of CNVs on genes/genetic features.

Network identification

In silico mapped human RefSeq genes in CNVRs were analyzed using Ingenuity Pathways Analysis (IPA) v9.0 (Ingenuity Systems) as previously described (Hou et al. 2011). The accessions of unique genes were imported into the software and subsequently mapped to their corresponding annotations in the Ingenuity Pathways Knowledge Base. The networks accommodating these unique genes (also called focus molecules) were identified in comparison with the comprehensive global networks developed by IPA. The molecule network was illustrated with an assigned relevance score, the number of focus molecules, as well as the top function of the networks. In the process of analysis, each network was set to have a maximum of 35 molecules by default, and only human was chosen for the species option (vs. human, mouse, and rat). We used all for the confidence level, including evidences of experimentally observed, predicted high or moderate confidence. The identified networks were further presented as a network graph showing the biological relationship among different molecules in which molecules were represented as nodes, distinguished by shapes based on the functional category, connected by distinct edges according to the interaction between molecules.

Heatmap hierarchical cluster analysis

Heatmaps were generated using the estimated CN windows for each animal. The gplots R package (http://cran.r-project.org/web/packages/gplots/index.html) was used to graph the CN values and generate hierarchical cluster dendrograms for each animal, using Pearson's correlation.

Cattle CNV distribution and association with segmental duplications and other features

Association between CNVs and SDs was tested by 1000 random simulations by selecting valid genomic segments from the length distribution of 1265 CNVs and determining if the segments overlapped at least one SD. Additional genomic features were obtained from public databases. Determination of the overlap between CNVRs and genomic features was performed as previously described (Liu et al. 2010).

Data access

All aCGH data have been submitted to the NCBI Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE31018.

Acknowledgments

We thank members of the Illumina Bovine HD SNP Consortium for sharing their samples. We also thank T. Brown, R. Anderson, and A. Dimtchev for technical assistance. E.E.E. is supported by NIH grants GM058815, HG002385 and is an investigator of the Howard Hughes Medical Institute. J.F.T. and C.P.V.T. were supported by Agriculture and Food Research Initiative competitive grant no. 2009-65205-05635 from the USDA National Institute of Food and Agriculture Animal Genome Program. G.E.L. was supported by NRI/AFRI grants nos. 2007-35205-17869 and 2011-67015-30183 from the USDA CSREES (now NIFA) and Project 1265-31000-098-00 from USDA-ARS. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the US Department of Agriculture.

Authors' contributions: G.E.L. and D.M.B. conceived and designed the experiments. D.M.B., Y.H., C.A., L.K.M., and J.S. performed in silico prediction and computational analyses. Y.H., D.M.B., M.F.C., and M.V. performed aCGH, qPCR, and FISH confirmation. S.G.S., J.F.G., R.D.S., J.F.T., T.S.S., and C.P.V.T. collected samples and sequence data. G.E.L., D.M.B., and E.E.E. wrote the paper.

References

Aitman TJ, Dong R, Vyse TJ, Norsworthy PJ, Johnson MD, Smith J, Mangion J, Roberton-Lowe C, Marshall AJ, Petretto E, et al. 2006. Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature* **439:** 851–855.

Alam MS, Ghosh D, Khan MGM, Islam MF, Mondal D, Itoh M, Islam MN, Haque R. 2011. Survey of domestic cattle for anti-Leishmania antibodies and Leishmania DNA in a visceral leishmaniasis endemic area of Bangladesh. *BMC Vet Res* **7:** 27. doi: 10.1186/1746-6148-7-27.

Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C, Malig M, Mutlu O, et al. 2009. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet* **41:** 1061–1067.

Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu FL, Bonnen PE, de Bakker PIW, Deloukas P, Gabriel SB, et al. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* **467:** 52–58.

Bae JS, Cheong HS, Kim LH, NamGung S, Park TJ, Chun JY, Kim JY, Pasaje CF, Lee JS, Shin HD. 2010. Identification of copy number variations and common deletion polymorphisms in cattle. *BMC Genomics* **11:** 232. doi: 10.1186/1471-2164-11-232.

Benson G. 1999. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res* **27:** 573–580.

Bera A, Shashi S, Nagaraj R, Vaidya T. 2003. Induction of autophagic cell death in *Leishmania donovani* by antimicrobial peptides. *Mol Biochem Parasitol* **127:** 23–35.

Berman, A. 2011. Invited review: Are adaptations present to support dairy cattle productivity in warm climates? *J Dairy Sci* **94:** 2147–2158.

Bochukova EG, Huang N, Keogh J, Henning E, Purmann C, Blaszczyk K, Saeed S, Hamilton-Shield J, Clayton-Smith J, O'Rahilly S, et al. 2009. Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature* **463:** 666–670.

The Bovine Genome Sequencing and Analysis Consortium. 2009. The Genome Sequence of Taurine Cattle: A window to ruminant biology and evolution. *Science* **324:** 522–528.

Campbell CD, Sampas N, Tsalenko A, Sudmant PH, Kidd JM, Malig M, Vu TH, Vives L, Tsang P, Bruhn L, et al. 2011. Population-genetic properties of differentiated human copy-number polymorphisms. *Am J Hum Genet* **88:** 317–332.

Chen WK, Swartz JD, Rush LJ, Alvarez CE. 2009b. Mapping DNA structural variation in dogs. *Genome Res* **19:** 500–509.

Cheng Z, Ventura M, She X, Khaitovich P, Graves T, Osoegawa K, Church D, deJong P, Wilson RK, Paabo S, et al. 2005. A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* **437:** 88–93.

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, et al. 2009. Origins and functional impact of copy number variation in the human genome. *Nature* **464:** 704–712.

Cook EH Jr, Scherer SW. 2008. Copy-number variations associated with neuropsychiatric conditions. *Nature* **455:** 919–923.

de Vries BBA, Pfundt R, Leisink M, Koolen DA, Vissers LELM, Janssen IM, van Reijmersdal S, Nillesen WM, Huys EHLP, De Leeuw N, et al. 2005. Diagnostic genome profiling in mental retardation. *Am J Hum Genet* **77:** 606–616.

Decker JE, Pires JC, Conant GC, McKay SD, Heaton MP, Chen K, Cooper A, Vilkki J, Seabury CM, Caetano AR, et al. 2009. Resolving the evolution of extant and extinct ruminants with high-throughput phylogenomics. *Proc Natl Acad Sci* **106:** 18644–18649.

Diskin SJ, Hou C, Glessner JT, Attiyeh EF, Laudenslager M, Bosse K, Cole K, Mosse YP, Wood A, Lynch JE, et al. 2009. Copy number variation at 1q21.1 associated with neuroblastoma. *Nature* **459:** 987–991.

Drogemuller C, Distl O, Leeb T. 2001. Partial deletion of the bovine ED1 gene causes anhidrotic ectodermal dysplasia in cattle. *Genome Res* **11:** 1699–1705.

Estivill X, Armengol L. 2007. Copy number variants and common disorders: Filling the gaps and exploring complexity in genome-wide association studies. *PLoS Genet* **3:** 1787–1799.

Fadista J, Thomsen B, Holm LE, Bendixen C. 2010. Copy number variation in the bovine genome. *BMC Genomics* **11:** 284. doi: 10.1186/1471-2164-11-284.

Fanciulli M, Norsworthy PJ, Petretto E, Dong R, Harper L, Kamesh L, Heward JM, Gough SC, de Smith A, Blakemore AI, et al. 2007. FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nat Genet* **39:** 721–723.

Fellermann K, Stange DE, Schaeffeler E, Schmalzl H, Wehkamp J, Bevins CL, Reinisch W, Teml A, Schwab M, Lichter P, et al. 2006. A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. *Am J Hum Genet* **79:** 439–448.

Fontanesi L, Beretti F, Riggio V, Gómez González E, Dall'olio S, Davoli R, Russo V, Portolano B. 2009. Copy number variation and missense mutations of the agouti signaling protein (ASIP) gene in goat breeds with different coat colors. *Cytogenet Genome Res* **126:** 333–347.

Fontanesi L, Beretti F, Martelli PL, Colombo M, Dall'olio S, Occidente M, Portolano B, Casadio R, Matassino D, Russo V. 2011. A first comparative map of copy number variations in the sheep genome. *Genomics* **97:** 158–165.

Formanack ML, Baier LJ. 2004. Variation in the FABP2 promoter affects gene expression: Implications for prior association studies. *Diabetologia* **47:** 349–351.

Garattini E, Fratelli M, Terao M. 2009. The mammalian aldehyde oxidase gene family. *Hum Genomics* **4:** 119–130.

Glatz JFC, van der Vusse GJ. 1996. Cellular fatty acid-binding proteins: Their function and physiological significance. *Prog Lipid Res* **35:** 243–282.

Glessner JT, Wang K, Cai G, Korvatska O, Kim CE, Wood S, Zhang H, Estes A, Brune CW, Bradfield JP, et al. 2009. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature* **459:** 569–573.

Graubert TA, Cahan P, Edwin D, Selzer RR, Richmond TA, Eis PS, Shannon WD, Li X, McLeod HL, Cheverud JM, et al. 2007. A high-resolution map of segmental DNA copy number variation in the mouse genome. *PLoS Genet* **3:** e3. doi: 10.1371/journal.pgen.0030003.

Guryev V, Saar K, Adamovic T, Verheul M, van Heesch SA, Cook S, Pravenec M, Aitman T, Jacob H, Shull JD, et al. 2008. Distribution and functional impact of DNA copy number variation in the rat. *Nat Genet* **40:** 538–545.

Hach F, Hormozdiari F, Alkan C, Hormozdiari F, Birol I, Eichler EE, Sahinalp SC. 2010. mrsFAST: A cache-oblivious algorithm for short-read mapping. *Nat Methods* **7:** 576–577.

Haigh B, Hood K, Broadhurst M, Medele S, Callaghan M, Smolenski G, Dines M, Wheeler T. 2008. The bovine salivary proteins BSP30a and BSP30b are independently expressed BPI-like proteins with anti-*Pseudomonas* activity. *Mol Immunol* **45:** 1944–1951.

Hillbertz NHCS, Isaksson M, Karlsson EK, Hellmen E, Pielberg GR, Savolainen P, Wade CM, Von Euler H, Gustafson U, Hedhammar K, et al. 2007. Duplication of FGF3, FGF4, FGF19, and ORAOV1 causes hair ridge and predisposition to dermoid sinus in Ridgeback dogs. *Nat Genet* **39:** 1318–1320.

Hou Y, Liu GE, Bickhart DM, Cardone MF, Wang K, Kim ES, Matukumalli LK, Ventura M, Song J, Vanradan PM, et al. 2011. Genomic characteristics of cattle copy number variations. *BMC Genomics* **12:** 127. doi: 10.1186/1471-2164-12-127.

Isobe N, Nakamura J, Nakano H, Yoshimura Y. 2009. Existence of functional lingual antimicrobial peptide in bovine milk. *J Dairy Sci* **92:** 2691–2695.

Kijas JW, Barendse W, Barris W, Harrison B, McCulloch R, McWilliam S, Whan V. 2011. Analysis of copy number variants in the cattle genome. *Gene* **482:** 73–77.

Kulkarni MM, Barbi J, McMaster WR, Gallo RL, Satoskar AR, McGwire BS. 2011. Mammalian antimicrobial peptide influences control of cutaneous *Leishmania* infection. *Cell Microbiol* **13:** 913–923.

LaFramboise T. 2009. Single nucleotide polymorphism arrays: A decade of biological, computational, and technological advances. *Nucleic Acids Res* **37:** 4181–4193.

Lai WR, Johnson MD, Kucherlapati R, Park PJ. 2005. Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics* **21:** 3763–3770.

Le Marechal C, Masson E, Chen JM, Morel F, Ruszniewski P, Levy P, Ferec C. 2006. Hereditary pancreatitis caused by triplication of the trypsinogen locus. *Nat Genet* **38:** 1372–1374.

Liu GE, Ventura M, Cellamare A, Chen L, Cheng Z, Zhu B, Li C, Song J, Eichler EE. 2009. Analysis of recent segmental duplications in the bovine genome. *BMC Genomics* **10:** 571. doi: 10.1186/1471-2164-10-571.

Liu GE, Hou Y, Zhu B, Cardone MF, Jiang L, Cellamare A, Mitra A, Alexander LJ, Coutinho LL, Dell'Aquila ME, et al. 2010. Analysis of copy number variations among diverse cattle breeds. *Genome Res* **20:** 693–703.

Locke DP, Segraves R, Carbone L, Archidiacono N, Albertson DG, Pinkel D, Eichler EE. 2003. Large-scale variation among human and great ape genomes determined by array comparative genomic hybridization. *Genome Res* **13:** 347–357.

Locke DP, Sharp AJ, McCarroll SA, McGrath SD, Newman TL, Cheng Z, Schwartz S, Albertson DG, Pinkel D, Altshuler DM, et al. 2006. Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *Am J Hum Genet* **79:** 275–290.

Lopez-Meza JE, Gutierrez-Barroso A, Ochoa-Zarzosa A. 2009. Expression of tracheal antimicrobial peptide in bovine mammary epithelial cells. *Res Vet Sci* **87:** 59–63.

Marques-Bonet T, Girirajan S, Eichler EE. 2009. The origins and impact of primate segmental duplications. *Trends Genet* **25:** 443–454.

McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, Shapero MH, de Bakker PI, Maller JB, Kirby A, et al. 2008. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet* **40:** 1166–1174.

Mi H, Dong Q, Muruganujan A, Gaudet P, Lewis S, Thomas PD. 2010. PANTHER version 7: Improved phylogenetic trees, orthologs, and collaboration with the Gene Ontology Consortium. *Nucleic Acids Res* **38:** D204–D210.

Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK, et al. 2011. Mapping copy number variation by population-scale genome sequencing. *Nature* **470:** 59–65.

Moller MJ, Chaudhary R, Hellmen E, Hoyheim B, Chowdhary B, Andersson L. 1996. Pigs with the dominant white coat color phenotype carry a duplication of the KIT gene encoding the mast/stem cell growth factor receptor. *Mamm Genome* **7:** 822–830.

Morgulis A, Gertz EM, Schaffer AA, Agarwala R. 2006. WindowMasker: Window-based masker for sequenced genomes. *Bioinformatics* **22:** 134–141.

Nicholas TJ, Cheng Z, Ventura M, Mealey K, Eichler EE, Akey JM. 2009. The genomic architecture of segmental duplications and associated copy number variants in dogs. *Genome Res* **19:** 491–499.

Norris BJ, Whan VA. 2008. A gene duplication affecting expression of the ovine ASIP gene is responsible for white and black sheep. *Genome Res* **18:** 1282–1293.

Olsson M, Meadows JRS, Truve K, Pielberg GR, Puppo F, Mauceli E, Quilez J, Tonomura N, Zanna G, Docampo MJ, et al. 2011. A novel unstable duplication upstream of HAS2 predisposes to a breed-defining skin phenotype and a periodic fever syndrome in Chinese Shar-Pei dogs. *PLoS Genet* **7:** e1001332. doi: 10.1371/journal.pgen.1001332.

Page NM, Butlin DJ, Lomthaisong K, Lowry PJ. 2001. The human apolipoprotein L gene cluster: Identification, classification, and sites of distribution. *Genomics* **74:** 71–78.

Pinto D, Darvishi K, Shi XH, Rajan D, Rigler D, Fitzgerald T, Lionel AC, Thiruvahindrapuram B, MacDonald JR, Mills R, et al. 2011. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nat Biotechnol* **29:** 512–520.

Porto Neto LR, Jonsson NN, D'Occhio MJ, Barendse W. 2011. Molecular genetic approaches for identifying the basis of variation in resistance to tick infestation in cattle. *Vet Parasitol* **180:** 165–172.

Ramayo-Caldas Y, Castelló A, Pena RN, Alves E, Mercadé A, Souza CA, Fernández AI, Perez-Enciso M, Folch JM. 2010. Copy number variation in the porcine genome inferred from a 60 k SNP BeadChip. *BMC Genomics* **11:** 593. doi: 10.1186/1471-2164-11-593.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, et al. 2006. Global variation in copy number in the human genome. *Nature* **444:** 444–454.

Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J, et al. 2007. Strong association of de novo copy number mutations with autism. *Science* **316:** 445–449.

Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, Pertz LM, Clark RA, Schwartz S, Segraves R, et al. 2005. Segmental duplications and copy-number variation in the human genome. *Am J Hum Genet* **77:** 78–88.

Sharp AJ, Hansen S, Selzer RR, Cheng Z, Regan R, Hurst JA, Stewart H, Price SM, Blair E, Hennekam RC, et al. 2006. Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nat Genet* **38:** 1038–1042.

She X, Cheng Z, Zollner S, Church DM, Eichler EE. 2008. Mouse segmental duplication and copy number variation. *Nat Genet* **40:** 909–914.

Shi J, Levinson DF, Duan J, Sanders AR, Zheng Y, Pe'er I, Dudbridge F, Holmans PA, Whittemore AS, Mowry BJ, et al. 2009. Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature* **460:** 753–757.

Smit A, Hubley R, Green P. 1996. RepeatMasker Open-3.0. http://www.repeatmasker.org.

Snyder M, Du J, Gerstein M. 2010. Personal genome sequencing: Current approaches and challenges. *Genes Dev* **24:** 423–431.

Sonstegard TS, Lopez-Corrales NL, Kappes SM, Stone RT, Ambady S, Ponce de Leon FA, Beattie CW. 1997. An integrated genetic and physical map of the bovine X chromosome. *Mamm Genome* **8:** 16–20.

Sonstegard TS, Garrett WM, Ashwell MS, Bennett GL, Kappes SM, Van Tassell CP. 2000. Comparative map alignment of BTA27 and HSA4 and 8 to identify conserved segments of genome containing fat deposition QTL. *Mamm Genome* **11:** 682–688.

Stankiewicz P, Lupski JR. 2010. Structural variation in the human genome and its role in disease. *Annu Rev Med* **61:** 437–455.

Stefansson H, Ophoff RA, Steinberg S, Andreassen OA, Cichon S, Rujescu D, Werge T, Pietilainen OP, Mors O, Mortensen PB, et al. 2009. Common variants conferring risk of schizophrenia. *Nature* **460:** 744–747.

Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, Sampas N, Bruhn L, Shendure J, Eichler EE. 2010. Diversity of human copy number variation and multicopy genes. *Science* **330:** 641–646.

Telugu BPVL, Walker AM, Green JA. 2009. Characterization of the bovine pregnancy-associated glycoprotein gene family—analysis of gene sequences, regulatory regions within the promoter, and expression of selected genes. *BMC Genomics* **10:** 185. doi: 10.1186/1471-2164-10-185.

Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. 2003. PANTHER: A library of protein families and subfamilies indexed by function. *Genome Res* **13:** 2129–2141.

Ventura M, Mudge JM, Palumbo V, Burn S, Blennow E, Pierluigi M, Giorda R, Zuffardi O, Archidiacono N, Jackson MS, et al. 2003. Neocentromeres in 15q24-26 map to duplicons which flanked an ancestral centromere in 15q25. *Genome Res* **13:** 2059–2068.

Walsh SW, Williams EJ, Evans ACO. 2011. A review of the causes of poor fertility in high milk producing dairy cows. *Anim Reprod Sci* **123:** 127–138.

Wang YH, Reverter A, Kemp D, McWilliam SM, Ingham A, Davis CK, Moore RJ, Lehnert SA. 2007. Gene expression profiling of Hereford Shorthorn cattle following challenge with *Boophilus microplus* tick larvae. *Aust J Exp Agric* **47:** 1397–1407.

Waszak SM, Hasin Y, Zichner T, Olender T, Keydar I, Khen M, Statz AM, Schlattl A, Lancet D, Korbel JO. 2010. Systematic inference of copy-number genotypes from personal genome sequencing data reveals extensive olfactory receptor gene content diversity. *PLoS Comput Biol* **6:** e1000988. doi: 10.1371/journal.pcbi.1000988.

Wheeler TT, Hood KA, Maqbool NJ, McEwan JC, Bingle CD, Zhao S. 2007. Expansion of the Bactericidal/Permeability Increasing-like (BPI-like) protein locus in cattle. *BMC Genomics* **8:** 75. doi: 10.1186/1471-2164-8-75.

Winchester L, Yau C, Ragoussis J. 2009. Comparing CNV detection methods for SNP arrays. *Brief Funct Genomics Proteomics* **8:** 353–366.

Wright D, Boije H, Meadows JRS, Bed'hom B, Gourichon D, Vieaud A, Tixier-Boichard Ml, Rubin CJ, Imsland F, Hallböök F, et al. 2009. Copy number variation in intron 1 of SOX5 causes the Pea-comb phenotype in chickens. *PLoS Genet* **5:** e1000512. doi: 10.1371/journal.pgen.1000512.

Xie S, Green J, Bixby JB, Szafranska B, DeMartini JC, Hecht S, Roberts RM. 1997. The diversity and evolutionary relationships of the pregnancy-associated glycoproteins, an aspartic proteinase subfamily consisting of many trophoblast-expressed genes. *Proc Natl Acad Sci* **94:** 12809–12816.

Yalcin B, Wong K, Agam A, Goodson M, Keane TM, Gan XC, Nellaker C, Goodstadt L, Nicod J, Bhomra A, et al. 2011. Sequence-based characterization of structural variation in the mouse genome. *Nature* **477:** 326–329.

Yang Y, Chung EK, Wu YL, Savelli SL, Nagaraja HN, Zhou B, Hebert M, Jones KN, Shu Y, Kitzmiller K, et al. 2007. Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): Low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. *Am J Hum Genet* **80:** 1037–1054.

Zhang F, Gu W, Hurles ME, Lupski JR. 2009. Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet* **10:** 451–481.

Zimin AV, Delcher AL, Florea L, Kelley DR, Schatz MC, Puiu D, Hanrahan F, Pertea G, Van Tassell CP, Sonstegard TS, et al. 2009. A whole-genome assembly of the domestic cow, *Bos taurus*. *Genome Biol* **10:** R42. doi: 10.1186/gb-2009-10-4-r42.

Received October 31, 2011; accepted in revised form February 1, 2012.