Functional & Integrative Genomics
https://doi.org/10.1007/s10142-019-00657-4

ORIGINAL ARTICLE

Computational detection and experimental validation of segmental duplications and associated copy number variations in water buffalo (*Bubalus bubalis*)

Shuli Liu[1,2] · Xiaolong Kang[1,3] · Claudia R. Catacchio[4] · Mei Liu[1,5] · Lingzhao Fang[1,6] · Steven G. Schroeder[1] · Wenli Li[7] · Benjamin D. Rosen[1] · Daniela Iamartino[8,9] · Leopoldo Iannuzzi[10] · Tad S. Sonstegard[11] · Curtis P. Van Tassell[1] · Mario Ventura[4] · Wai Yee Low[12] · John L. Williams[12] · Derek M. Bickhart[7] · George E. Liu[1]

Received: 16 August 2018 / Revised: 13 December 2018 / Accepted: 9 January 2019

Abstract

Duplicated sequences are an important source of gene evolution and structural variation within mammalian genomes. Using a read depth approach based on next-generation sequencing, we performed a genome-wide analysis of segmental duplications (SDs) and associated copy number variations (CNVs) in the water buffalo (*Bubalus bubalis*). By aligning short reads of Olimpia (the reference water buffalo) to the UMD3.1 cattle genome, we identified 1,038 segmental duplications comprising 44.6 Mb (equivalent to ~1.73% of the cattle genome) of the autosomal and X chromosomal sequence in the buffalo genome. We experimentally validated 70.3% (71/101) of these duplications using fluorescent *in situ* hybridization. We also detected a total of 1,344 CNV regions across 14 additional water buffaloes, amounting to 59.8 Mb of variable sequence or the equivalent of 2.2% of the cattle genome. The CNV regions overlap 1,245 genes that are significantly enriched for specific biological functions including immune response, oxygen transport, sensory system and signal transduction. Additionally, we performed array Comparative Genomic Hybridization (aCGH) experiments using the 14 water buffaloes as test samples and Olimpia as the reference. Using a linear regression model, a high Pearson correlation ($r = 0.781$) was observed between the $\log_2$ ratios between copy number estimates and the $\log_2$ ratios of aCGH probes. We further designed Quantitative PCR assays to confirm CNV regions within or near annotated genes and found 74.2% agreement with our CNV predictions. These results confirm sub-chromosome-scale structural rearrangements present in the cattle and water buffalo. The information on genome variation that will be of value for evolutionary and phenotypic studies, and may be useful for selective breeding of both species.

Shuli Liu, Xiaolong Kang, Claudia R. Catacchio and Mei Liu contributed equally to this work.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10142-019-00657-4) contains supplementary material, which is available to authorized users.

✉ Derek M. Bickhart
Derek.Bickhart@ARS.USDA.GOV

✉ George E. Liu
George.Liu@ARS.USDA.GOV

[1] USDA-ARS, Animal Genomics and Improvement Laboratory, Beltsville, Maryland 20705, USA

[2] College of Animal Science and Technology, China Agricultural University, Beijing 100193, China

[3] College of Agriculture, Ningxia University, Yinchuan 750021, China

[4] Department of Biology, University of Bari, 70126 Bari, Italy

[5] College of Animal Science and Technology, Shaanxi Key Laboratory of Agricultural Molecular Biology, Northwest A&F University, Yangling 712100, Shaanxi, China

[6] Department of Animal and Avian Sciences, University of Maryland, College Park, Maryland 20742, USA

[7] The Cell Wall Utilization and Biology Laboratory, US Dairy Forage Research Center, USDA, ARS, Madison WI 53706, USA

[8] AIA-LGS, Associazione Italiana Allevatori - Laboratorio Genetica e Servizi, Via Bergamo 292, 26100 (CR) Cremona, Italy

[9] Parco Tecnologico Padano, Via Einstein, Polo Universitario, 26900 Lodi, Italy

[10] Laboratory of Animal Cytogenetics and Gene Mapping, Nationa Research Council (CNR), ISPAAM, Via Argine 1085, 80147 Naples, Italy

[11] Recombinetics, 1246 University Ave W, St Paul, MN 55104, USA

[12] Davies Research Centre, School of Animal and Veterinary Sciences, University of Adelaide, Roseworthy, SA 5371, Australia

Published online: 07 February 2019

Keywords Segmental duplication · Copy number variation · *Bubalus bubalis* · Fluorescent *in situ* hybridization · Array Comparative Genomic Hybridization · Quantitative PCR

Introduction

Water buffalo were domesticated about 5,000 years ago and are the most important farm animal resource in developing tropical and subtropical countries, contributing greatly to the local economy of rural areas (Michelizzi et al. 2010). Two types of water buffalo are recognized, the river and the swamp buffalo. River buffalo have been selectively bred as dairy animals while swamp buffalo are typically used as draft animals (Zhang et al. 2007).

The use of molecular genetic approaches has increased the genetic gain in animal selection programs. After the release of the first *de novo* assembly of an Italian Mediterranean river buffalo (UMD_CASPUR_WB_2.0) (Williams et al. 2017), and with the availability of a 90K single nucleotide polymorphism (SNP) chip for buffalo (Iamartino et al. 2017), SNPs have been used in a range of studies of water buffalo (Colli et al. 2018; Whitacre et al. 2017). The current reference assembly is fragmented into 366,983 scaffolds with a low scaffold N50 of ~1.4Mb, which make it unsuitable for direct detection of large genetic variants that may contribute to the genetic diversity of the species (Williams et al. 2017). One such genetic variant, the copy number variation (CNV), which consists of duplications and deletions of DNA sequence ranging from 50bp to several megabases, impacts a large percentage of genomic sequence and potentially has a greater functional effect than SNPs (Henrichsen et al. 2009). CNVs have been extensively reported in human (Sudmant et al. 2015), primates (Gokcumen and Lee 2009), mouse (Henrichsen et al. 2009), zebrafish (Brown et al. 2012), dog (Nicholas et al. 2009), and livestock, including chicken (Yi et al. 2014), pig (Jiang et al. 2014), horse (Doan et al. 2012), sheep (Fontanesi et al. 2011), goat (Fontanesi et al. 2010) and cattle (Bickhart et al. 2016; Zhou et al. 2016; Liu et al. 2010). However, there have only been two surveys of CNVs in water buffalo, including a recent survey from our group (Li et al. 2018; Zhang et al. 2014).

Comparative Genomic Hybridization (CGH) arrays, SNP genotyping arrays, and high throughput sequencing (HTS) have been used for genome-wide CNV screens. However, the major limitation of CGH (comparative genomic hybridization) and SNP arrays is that they are indirect screens, providing no information on the actual structure of the variation detected (Pinto et al. 2011; Bickhart and Liu 2014; Li and Olivier 2013). Additionally, the resolution of CGH or SNP arrays is limited by the probe density of the array and certain genetic variants such as balanced rearrangements (e.g. inversions), and novel DNA sequence cannot be detected using these approaches. The decreasing cost of DNA sequencing has enabled CNV to be detected at a high effective resolution and sensitivity. Numerous methods have been developed for CNV detection using next-generation short-read sequencing, including read pair (RP), read depth (RD), split read (SR), sequence assembly (SA), and hybrid algorithms (combinatorial detection; e.g. Genome Strip) (Snyder et al. 2010; Mills et al. 2011; Handsaker et al. 2015). Among these, RD methods are highly sensitive in discovering duplications and are capable of determining exact copy number (CN) values for each genetic locus in an individual (Sudmant et al. 2010). The mrFAST/mrsFAST and whole genome shotgun sequence detection (WSSD) method (Sudmant et al. 2010; Alkan et al. 2009; Hach et al. 2010) can be used to construct personalized CNV maps in or near segmental duplication (SD) regions, by reporting all mapping locations for sequence reads, whereas other RD methods only take one mapping location per read into consideration. When a read is mapped to multiple best tied locations, a random locus is often selected for further downstream analyses. Due to the higher frequency of CNVs in or near duplication regions in the genome (Cheng et al. 2005; Bickhart et al. 2012), mrFAST and mrsFAST are especially efficient in detecting CNV within or near duplication- and repeat-rich regions. CNV detection methods that are locus-specific, including fluorescence *in situ* hybridization (FISH) and quantitative polymerase chain reaction (qPCR), can be used to detect large CNVs and often used to experimentally validate the CNVs predicted by genome-wide methods (Doan et al. 2012; Bickhart et al. 2012). Notably, CNV detection in most livestock have been limited to only one or two methods, and lack rigorous experimental validation.

The recent release of a water buffalo draft reference genome has accelerated genomic studies and the application of genetic selection in this species (Williams et al. 2017). However, the draft assembly is highly fragmented and not as thoroughly annotated as the cattle reference genome. This study used a comparative alignment of the buffalo DNA sequences with the completed reference genome of *Bos taurus*, to systematically detect CNV in the water buffalo genome. The CNV identified were validated using a CGH-based whole-genome approach, followed by FISH and qPCR confirmation for selected CNVs. By assessing the CNV distribution of water buffalo at a genome-wide level, we provide information for studies into highly duplicated regions in the water buffalo genome e.g. to uncover duplicated genes that may be associated with agriculturally important traits.

Results

Genome-wide identification of segmental duplications

We retrieved the whole genome Illumina HTS reads of the water buffalo (Olimpia) whose sequence was recently assembled and released as the draft reference genome (Williams et al. 2017). We mapped the reads to the UMD3.1 cattle genome assembly (Zimin et al. 2009). We then detected segmental duplications (SDs) using a sliding window approach, based on a previously published mrsFAST-WSSD method (≥ 1kb in length, ≥ 90% sequence identity) (Alkan et al. 2009). Due to the short lengths of chrUn contigs (i.e. unplaced contigs) and the ambiguous mapping of the chrUn sequence reads, we excluded events mapping to chrUn contigs in the subsequent analyses. We discovered 1,038 SDs in the autosomes and X chromosome of Olimpia, spanning ~44.6 Mb (~1.73%) of the cattle genome. This is comparable with the previously predicted extent of CNVs in cattle (49.2 Mb, excluding SDs in cattle chrUn contigs) (Liu et al. 2009). The buffalo SDs ranged in size from 1270bp to 750,223bp, with an average size of ~43 kb (standard deviation = 60.5 kb) (Table S1).

FISH validation of the predicted segmental duplication

To confirm the SDs detected, we experimentally validated a subset of the largest (≥ 20 kb) duplicated regions by FISH. A total of 121 cattle BAC clones corresponding to the WSSD duplicated regions were used as probes which were hybridized against the buccal epithelial cells of Olimpia (Table S2). Twenty probes failed to generate hybridization signals. We observed multiple signals for 70.3% (71/101) of the remaining probes. As expected, the majority of duplicated sequences were intra-chromosomal (52/71), while inter-chromosomal duplications showed signals on multiple non-homologous chromosomes, accounting for less than 27% of the regions tested. These data suggested that tandem intra-chromosomal duplications predominate in the water buffalo genome, which is similar to other mammalian species (Nicholas et al. 2009; Liu et al. 2009). FISH results confirmed that the following genes are duplicated: peptidase inhibitor 3 (*PI3*), olfactory receptor (OR) genes and pregnancy-associated glycoprotein (PAG) gene families (Fig. 1).

CNV discovery and dataset statistics

To study CNVs that might be polymorphic or fixed, we aligned short reads from 14 additional water buffaloes to the UMD3.1 cattle genome using the mrsFAST short read aligner, and called CNVs using the WSSD read depth approach. Based on sequence RD against the reference genome, we detected CNVs for the 14 individuals in the autosomes and the X chromosome. The number of duplications ranged from 839 (ITWB2) to 900 (ITWB7), and the number of deletions varied from 0 (ITWB8) to 273 (ITWB6). While our method had sufficient power to detect duplications, variation in RD across the autosomes, measured in standard deviations (STDEVs), limited our discovery to extreme deletion events (Bickhart et al. 2012). The CNVs from all individuals (including Olimpia) were merged if overlaps were 1bp or greater. In total, we detected 1,344 unique CNV regions (CNVRs) (1,041 gains, 279 losses and 24 both), amounting to 59.8 Mbp or 2.2% of the total bases in the cattle genome. A full list of CNVRs are listed in Table S3. A representative overview of the CNV landscape mapped onto cattle chromosome 5 is shown in Fig. 2 and other individual chromosome plots in Figs. S1-S6.

Genes overlapping with copy number variation

Using BioMart, in the Ensembl database (Ensembl Genes 79), we obtained the IDs for the genes that were located within, or overlapped, with the detected CNVRs. We identified a total of 1,245 genes and 47.4% of the CNVRs encompass 1 or more genes (Table S3). Using the MrsFAST WSSD algorithm, we assigned a CN estimate to each gene. Gene regions outside the predicted CNVRs were found to have a median CN estimate of 2.05, suggesting that CNV detection and CN assignment were concordant. Genes within CNVRs were found to be highly variable in CN among individuals (minimum value: 0; maximum: 299; median: 5.43; average: 7.54) (Table S4). To test the hypothesis that particular gene classes were over-represented in duplicated regions, we assigned PANTHER terms to all genes that overlapped duplications (Mi et al. 2017). We observed statistically significant enrichments in genes that participate in immune response, oxygen transport, sensory system and signalling transduction, which is consistent with similar analyses of duplications in other organisms (false discovery rate [FDR] < 0.05, Table S5). Of the top 25 most copy number variable genes, most had functions related to the immune response, such as interferons, melanoma antigen family and PAG gene family (Table 1). One CNV impacted gene family is *PI3*, which encodes the trappin/elafin anti-microbial/immune system modulator protein (Belaaouaj et al. 1998; Fujishima et al. 2008), had a high CN value (average CN: 6.5) in the water buffalo (Fig. 3a, Table S4). Another highly CNV impacted gene, UL16 binding protein 3 (*ULBP3*) (average CN: 8.3), encodes one of several related ligands of the KLRK1/NKG2D receptor, which is involved in the regulation of both innate and adaptive immune responses (Vivier et al. 2002) (Fig. 3b, Table S4).

Validation with aCGH analysis

To confirm individual CNVs, we performed aCGH experiments using the 14 additional water buffalo samples with

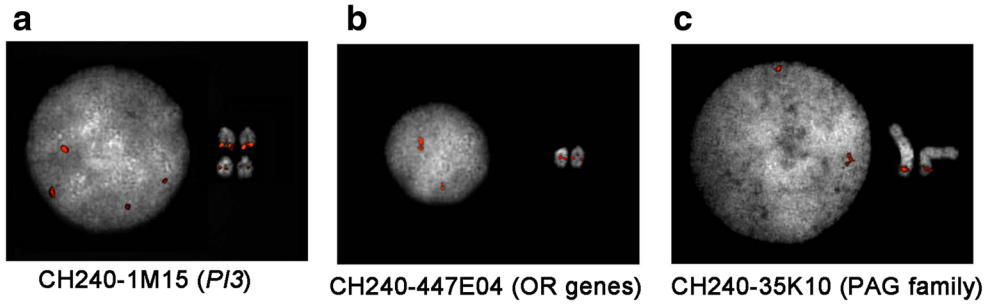

Fig. 1. Validation of segmental duplications by FISH analysis. **a** Example of an interchromosomal duplication detected with clone CH240-1M15, covering the *PI3* gene. (**b**, **c**) Two representative examples of tandem intrachromosomal duplications detected with clones CH240-447E04 and CH240-35K10, covering the OR genes and the PAG gene family

Olimpia as the reference. We compared the RD predicted CNV intervals with the aCGH results. To make the CN estimates comparable with the aCGH results, we calculated $\log_2$ ratios between CN estimates for the 14 buffaloes and that of Olimpia using a digital aCGH approach (Sudmant et al. 2010). Based on the predicted CN values within filtered CNVs (> 20kb that contained < 80% common repeat content), we generated $\log_2$ ratios between CN estimates and compared them with $\log_2$ ratios of the aCGH probes using a linear regression model (Sudmant et al. 2010). Within the CNV regions, we observed a high correlation (Pearson $r = 0.781$) between $\log_2$ ratios of CN estimates and aCGH $\log_2$ ratios (Fig. 4). The computational prediction and aCGH validation of Olimpia and another three randomly selected individuals in two CNV regions are presented in Fig. 5. These two regions cover the *PI3*, *PAG3* and *PAG6* immune related genes. The duplication of these three genes for Olimpia were confirmed in FISH analysis (Fig.1, Table S2).

qPCR analysis

We designed quantitative PCR assays to test 11 predicted CNVRs within or near annotated genes including *PI3*, *PAG6* and other randomly chosen genes. We randomly selected 6 individuals to investigate the 11 CNV regions and designed one primer set for each locus (Table S6). The basic transcription factor 3 (*BTF3*) gene was chosen as the control with the assumption that there were two copies of DNA segment in this region. The validation rates of the 11 loci in the 6 samples varied from 63.6% to 90.9% with an average validation rate of 74.2% (Table S7). We selected four of the CNVRs that were validated in all 6 individuals to compare the WGS-predicted CN values with the qPCR estimates (Fig. 6). We observed a high correlation between qPCR-derived CN and WGS-derived CN, supporting the reproducibility of our genome wide CNV detection methods.

Discussion

This study carried out a systematic investigation of the genome-wide CNV landscape of water buffalo. We identified 1,344 CNVRs in 15 water buffaloes (including Olimpia, the reference animal for the buffalo genome assembly) and validated this sequence-based CNV set using aCGH, qPCR and FISH. Agreement in CNV assignment was found among all four methods. Two previous studies have focused on the discovery of CNVs in the water buffalo genome. One study used the NimbleGen 3×720K CGH array, and found more than half of CNVRs discovered in buffalo were shared with cattle (Zhang et al. 2014). However, this CGH array approach has several inherent drawbacks, including hybridization noise,

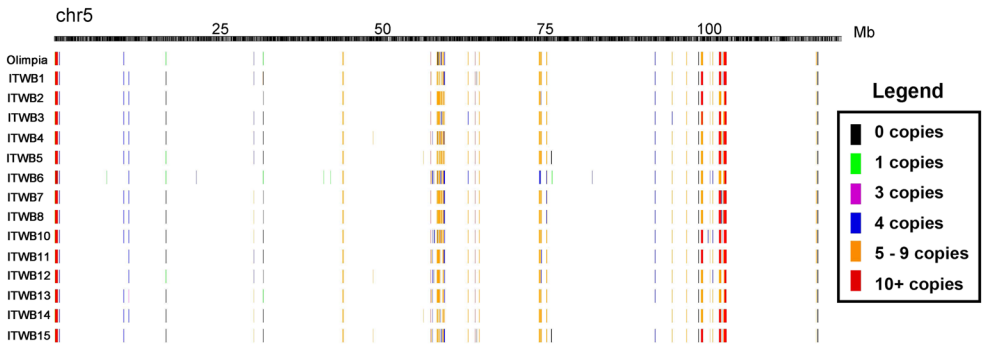

Fig. 2. CNV map of water buffalo chromosome 5. The UMD3.1 assembly is represented as black bars with assembly gaps indicated by white boxes on the chromosomes. Tracks underneath the chromosomes represent the SDs for Olimpia, and CNV data sets for 14 additional buffaloes. The colors for each bar in the animal data set tracks represent the average estimated CN for each CNV as shown in the legend

Table 1. Top 25 genes with copy number variations genes in the individual buffalo with the locations given on the UMD 3.1 bovine reference genome

Ensembl ID	Symbol	Location	Size (bp)	CNV overlap percentage (%)	Average CN
ENSBTAG00000045940	*MGC148328*	chrX:25444909-25445676	768	100	220.8
ENSBTAG00000034626	*FBXO16*	chr8:10095870-10133299	37,430	100	27.3
ENSBTAG00000035319	*MAD2L1*	chr6:6013172-6020467	7,296	100	14.2
ENSBTAG00000045554	*IFNW1*	chr8:22762551-22763138	588	100	14.2
ENSBTAG00000014534	*EEF1A1*	chr9:13233554-13236949	3,396	100	12.5
ENSBTAG00000046398	*PAG19*	chr29:39261026-39270024	8,999	100	12.4
ENSBTAG00000008291	*PROCR*	chr13:65052810-65106553	53,744	86.63	12.3
ENSBTAG00000035745	*MAGEB4*	chrX:118671361-118675384	4,024	100	12.0
ENSBTAG00000036277	*PAG4*	chr29:38793756-38802922	9,167	100	12.0
ENSBTAG00000046366	*CLDN34*	chrX:143692535-143693164	630	100	12.0
ENSBTAG00000006304	*PAG6*	chr29:40068356-40077715	9,360	100	11.9
ENSBTAG00000047892	*MGC157408*	chr29:39457513-39466695	9,183	100	11.6
ENSBTAG00000034064	*MGC133764*	chrX:118633280-118636998	3,719	100	11.3
ENSBTAG00000046638	*IFNAH*	chr8:23073254-23073823	570	100	11.2
ENSBTAG00000048133	*PAG15*	chr29:38659682-38668543	8,862	100	11.1
ENSBTAG00000037784	*MGC157405*	chr29:39731524-39740647	9,124	100	11.1
ENSBTAG00000022348	*PAG3*	chr29:39998682-40007570	8,889	100	10.8
ENSBTAG00000045674	*IFNB1*	chr8:23231884-23232444	561	100	10.6
ENSBTAG00000036172	*PAG20*	chr29:39004039-39013165	9127	100	10.5
ENSBTAG00000047141	*PAG7*	chr29:38518914-38624602	105,689	100	10.1
ENSBTAG00000037908	*PAG1*	chr29:39180301-39189626	9,326	100	9.8
ENSBTAG00000026102	*PAG9*	chr29:39863946-39872971	9,026	100	9.8
ENSBTAG00000038497	*CT47B1*	chrX:4355493-4387239	31,747	100	9.6
ENSBTAG00000040340	*PAG21*	chr29:39049999-39059075	9,077	100	9.6
ENSBTAG00000033993	*PRP14*	chr23:34693485-34706026	12,542	100	9.4

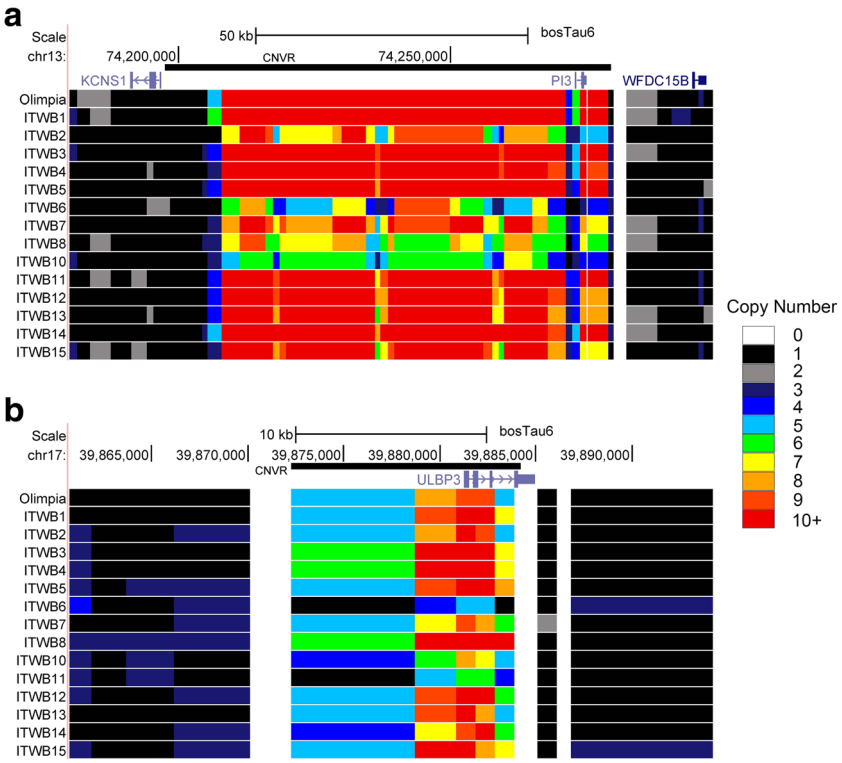

Fig. 3. Genes with copy number variations in individual water buffaloes. **a** Copy number values for each animal were plotted within the *PI3* locus (chr13:74180018-74298194) using the color scheme depicted in the legend. Heatmap boxes represent 1-kbp sliding, nonoverlapping windows in the region. **b** Copy number values within the *ULBP3* locus (chr17:39860701-39894230). The duplications of these two loci were confirmed using qPCR

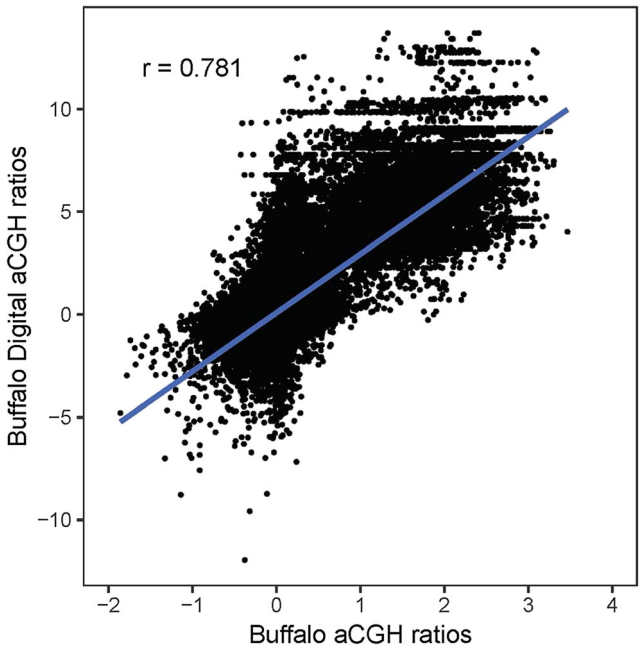

Fig. 4. Correlation between digital aCGH values (log_2 ratios between CN estimates) and whole genome aCGH (log_2 ratios of hybridization probes). Digital aCGH values were estimated using a log_2 ratio of the 1-kbp CN windows from each water buffalo individual divided by CN estimates from Olimpia. A high correlation (r = 0.781) was found for aCGH probe values and digital aCGH values within CNV intervals > 20 kb that had fewer than 80% of their lengths occupied by common repeats

low resolution, and that novel and rare variations are not detected (Snijders et al. 2001). Using a comparative alignment and selective filtering approach, our previous study on buffalo CNVs focused on the comparative analysis of genome features shared between buffalo and cattle (Li et al. 2018). The study identified large deletions and smaller variations in the gene regulatory regions which may impact on gene expression (Li et al. 2018). Our earlier study used the R package, cn.mops (Klambauer et al. n.d.) and JaRMS (Oldeschulte et al. 2017) to assess CNVs. These two methods tend to discover more deletion events than duplication events. In contrast, the mrsFAST-WSSD method has the tendency to detect more duplications, especially in the repetitive segmental duplication regions. We compared the CNV regions identified in these two studies, and found 49.9% (670/1344) CNVRs were also discovered using the methods of cn.mops and JaRMS, covering 25.9% (11.6 Mb) of all the variable sequences.

We found that the water buffalo studied shared several high copy number regions with cattle (Bickhart et al. 2012). For example, position of 25 Mb on chromosome 10 is enriched with 22 CNV regions, covering T-cell receptor alpha variable (TRAV) gene family members (Fig. S2). The copy number of the CNV cluster varied from 3 to 10 (average: 5.5). TRAV genes encode variable domains of the T-cell receptor alpha chain. T-cell receptors recognize foreign antigens and bind to major histocompatibility complex (MHC) molecules, which in turn are encoded by the genes located on chromosome 23 (Fig. S4), expressed on the surface of antigen presenting cells (Nikolich-Zugich et al. 2004). The expansion of the TRAV gene CN in buffalo could be partially due to the requirement for a substantial immune-regulatory T-cell population in this species to combat a wide range of pathogens

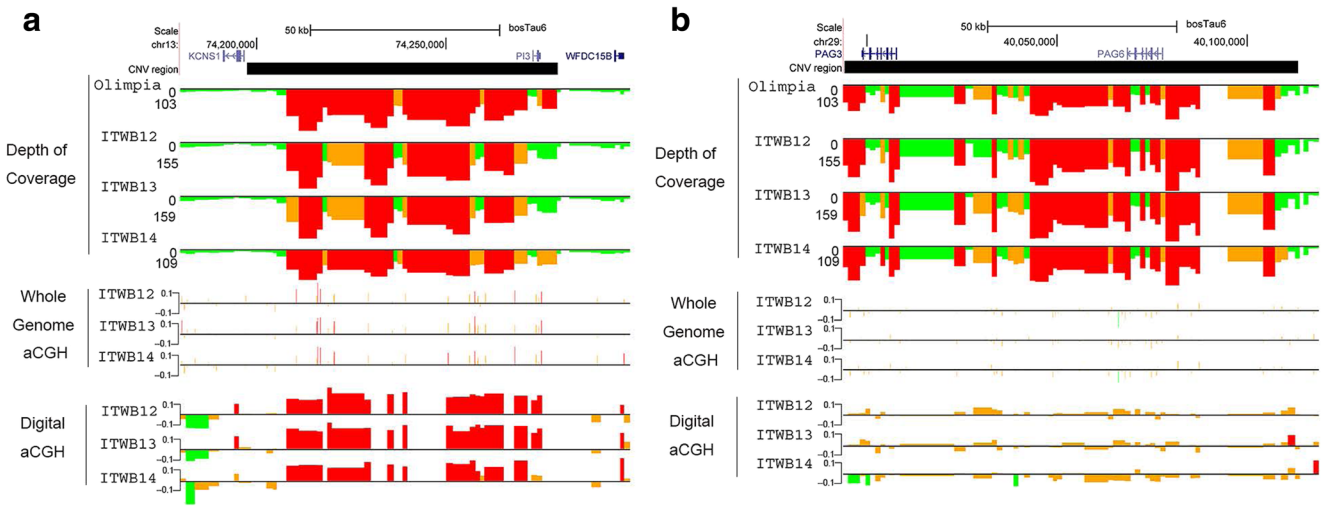

Fig. 5. Computational predictions and aCGH validations of segmental duplication copy number differences. Depth-of-coverage tracks for Olimpia, ITWB12, ITWB13 and ITWB14 are shown below a UCSC track for each investigated gene region. Regions colored in red on the plot indicate excessive read depth (> mean + 1.5 × STDEV), whereas orange regions indicate intermediate read depth (> mean + 1 × STDEV). Normal read depth values are colored green (within mean ± 1 × STDEV). Digital aCGH tracks show the log_2 ratio of the copy number of each listed animal compared to Olimpia, with high value listed in green (> 0.3); low values: red (< -0.3) and nominal values: orange (0.3 ≥ x ≥ -0.3). Whole-genome CGH array experiments, using Olimpia 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. The CNVRs are shown below the UCSC plot. **a** CNVs intersecting the *PI3* locus (chr13:74180018-74298194). A duplication of this region was predicted for all animals and was confirmed by whole-genome aCGH. **b** CNVs intersecting the *PAG3* and *PAG6* locus (Bovine chr29:39994004-40119007)

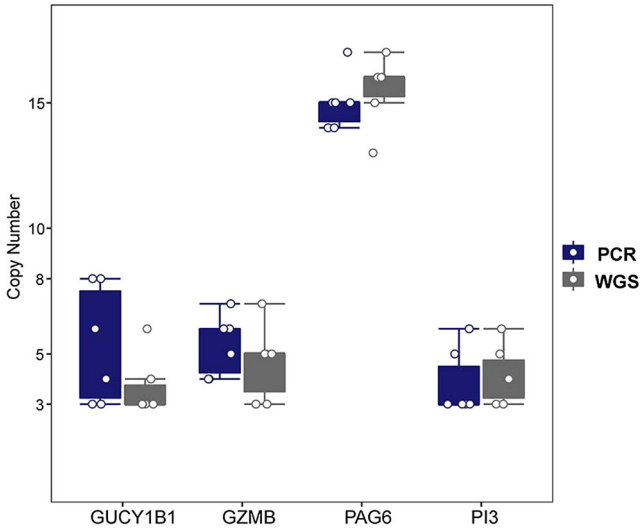

Fig. 6. The CN observed from qPCR experiments and estimated from WGS for four genes (*GUCY1B1, GZMB, PAG6* and *PI3*)

(Connelley et al. 2014). Another example of CN-divergent gene families is the PAG gene family members on chromosome 29 (Fig. S5). In the water buffalo genome, we found nine CNVRs, ranging from 38.3 Mb to 42.5 Mb on chromosome 29, of which the copy number varied from 5 to 26 (average: 11.4). PAG genes are abundantly expressed in the placenta of species within the Cetartiodactyla order where they play a role in the sequestration of fetal antigens at the placenta-uterine interface (Wallace et al. 2015). We found a large cluster of CNVs covering the PAG genes in all of the sequenced buffalo individuals, indicating the pervasiveness of the duplication events. Similarly, the PAG genes have been shown to be duplicated in eight diverse cattle breeds (Bickhart et al. 2016). As we proposed before (Bickhart et al. 2016), it is possible that duplications of the genes identified in this and other studies are indicative of subfunctionalization, neofunctionalization, or overdominance effects on structurally polymorphic PAG gene alleles. The other examples of shared CNVRs are the ~47 – 52 Mb region of chromosome 15 associated with OR genes, ~25 – 30 Mb region of chromosome 23 associated with cattle MHC (*BoLA)* family members and ~5 – 6 Mb region of chromosome 27 associated with β-Defensin (*DEFB*) family members.

We detected gene with largest CN differences between the water buffaloes studied here and 75 cattle individuals from eight breeds/subspecies (Bickhart et al. 2016) (i.e. with differences of the average CN values > 4). They included several cell cycle-related genes (*MIS18BP1*, *MAD2L1* and *CNTLN*), several genes related to immune function (*DEFB1*, *DEFB5*, *DEFB7* and *NRIH4*), the skin disease related genes (melanomas, like *PRAME* and *TNFRSF10*), as well as neuron system (*FZD3*) (Fig. 7a). One mitotic associated gene, mitotic arrest deficient 2 like 1 (*MAD2L1*), showed higher CN values in cattle (average CN: 35.1) than in water buffalo (average CN: 14.2) (Fig. 7b). The *MAD2L1* encoded protein is identified as a vital mediator of the chromosomal control pathway (Kato et al. 2011). It has been reported that the copy number loss of mitotic arrest deficient genes may be related to human fetal loss (Nath et al. 2012). A previous study identified frizzled class receptor 3 (*FZD3*) as one of the most stratified genes for taurine and indicine animals (Bickhart et al. 2016). Interestingly, the CN values of this gene in water buffaloes (average CN: 15.5) were two times more than that in cattle (average CN: 4.5), suggesting they could have been under different selection pressures in the two species (Fig. 7c). *FZD3* contributes to axonal growth in the central nervous system (Wang et al. 2002). The difference in CN values between cattle and water buffaloes may be partially driven by the differences occurring during domestication or from natural selection. We also discovered other functionally important genes stratified in CN values in these two species although the CN differences were less than 4. For example, the *PI3* gene has more copies (average CN: 6.5) in water buffalo than in cattle (average CN: 3), and their CN distributions seldom overlapped (Fig. 7d). The *PI3* gene is implicated in resistance to fungal and bacterial pathogens, so copy number variability in water buffalo may indicate structurally polymorphic alleles in this species that confer different resistance to these pathogens. These shared CNV clusters and CN differential genes in two species warrant further investigation to understand whether the CN affects phenotypes, particularly those related to economically important traits.

One limitation of this study is that the UMD3.1 cattle reference genome was used as a basis for CNV detection in water buffaloes. We chose to align the buffalo sequence data to the cattle reference genome because the highly fragmented water buffalo draft reference genome, which contains a large number of smaller scaffolds that are difficult to analyse using a window-based CNV detection method. It is possible that the structural differences between the cattle and water buffalo genomes detected may have been the result of using a different species as reference. Alignment of the water buffalo sequences with the bovine UMD3.1 reference genome identified 6.6 point mutations per 1000bp. The alignment method, mrsFAST, allows two mismatches per aligned read, which represents a 96% identity cutoff for each 50bp sequence. As the water buffalo diverged 12.3Myr ago from its last common ancestor with the cattle, 15% of buffalo sequences do not have a match with the cattle genome (Williams et al. 2017). These sequences represent either buffalo specific DNA or sequence absent in the UMD3.1 cattle reference genome. This missing buffalo specific genomic sequence is likely to have resulted in a loss of buffalo specific CNVs (i.e. false negatives). Other regions, which are divergent from the cattle assembly, may also influence the prediction of buffalo CNVs. However, in most of the unique and gene-rich genomic regions, buffalo sequence was highly comparable with sequence in the cattle

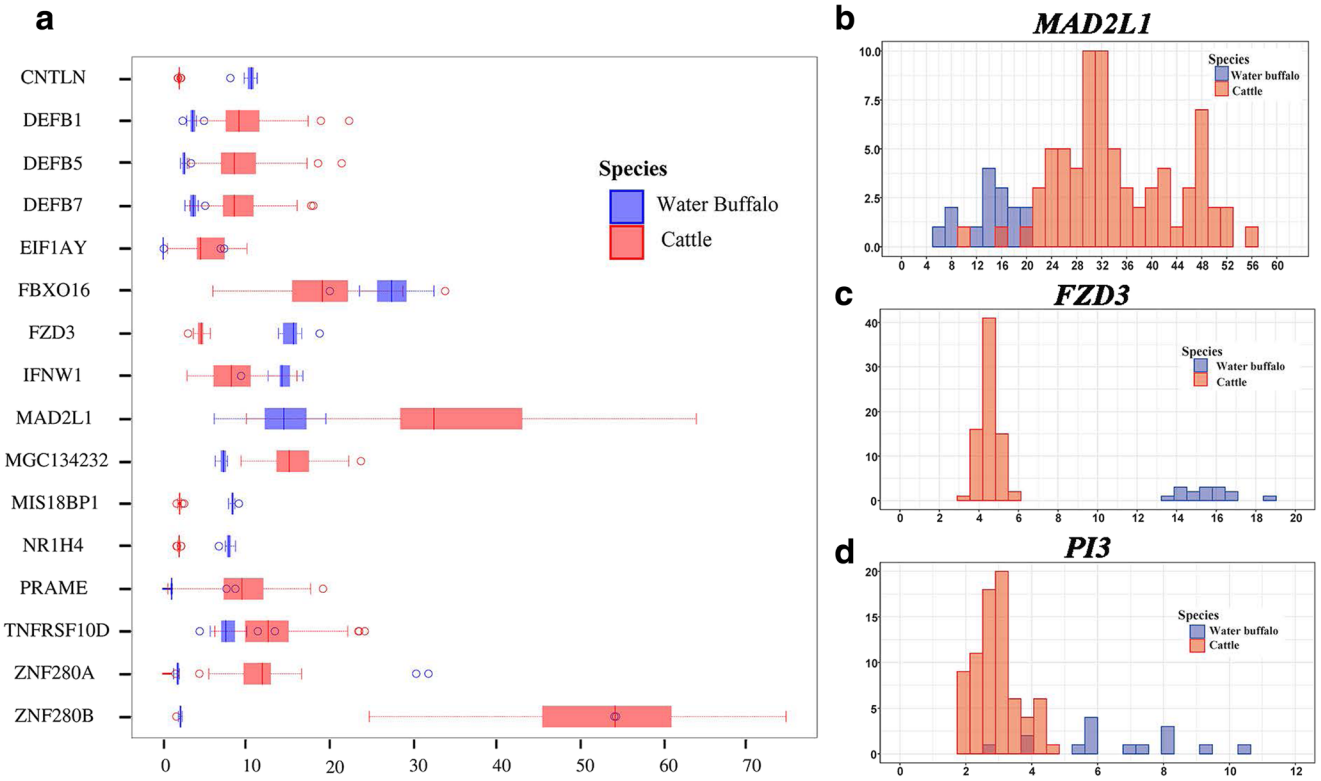

Fig. 7. Copy number different genes between cattle and water buffaloes. **a** The boxplot of CN values in 16 genes showing the highest CN differences (> 4) between cattle and water buffaloes (legend insets denote group colors). Histograms showing the distribution of CNs among the unrelated individuals in each group are plotted for *MAD2L1* (**b**), *FZD3* (**c**), and *PI3* (**d**). X-axis values indicate copy number and Y-axis values indicate 3 sample count

genome. For CNV detection based on the CGH array, the probes of 50-60 bp in length usually allowed one mismatch to target the specific chromosomal region, which is more stringent than mrsFAST alignment. Discrepancies between the digital and experimental aCGH may be partly explained by the differences in mapping/hybridization efficiency between mrsFAST and aCGH probes. For FISH, the hybridization of cattle BAC clones to Olimpia chromosomes is less likely to be affected by small variation in the sequence, thus FISH can be used to distinguish single copy signal vs. duplication signal, accommodating sequence divergence up to 20% in probe hybridization regions.

Future directions

It is important to note that performance of CNV detection using short read sequence data relies heavily on accurate mapping of reads. In highly repetitive regions, misalignment may lead to a high rate of false positive CNV calls. The latest long read sequencing technologies offer new opportunities in CNV detection by providing high confidence breakpoint analysis (Sedlazeck et al. 2018). In addition, more long reads can be more confidently anchored to repetitive sequences that often mediate the formation of SVs (Lucas Lledo and Caceres 2013).

Long read sequences and improved mapping technologies will lead to more precise CNV detection. These methods will facilitate further investigation of the structural organization of copy number variable regions in water buffaloes through population-level sequencing. With more confident detection of CNVs using long-read technology, the long-term goal is to explore the association of CNVs with important economic traits and incorporate them into selection programmes.

Methods

Data retrieval and sequence alignment Illumina sequence data from an inbred, female Italian Mediterranean buffalo (Olimpia) were retrieved from NCBI BioProject PRJNA207334 submitted by the International Water Buffalo Genome Consortium (Iamartino et al. 2017). We retrieved sequence data for the 14 additional water buffaloes (ITWB1 to ITWB8 and ITWB10 to ITWB15, paired-end reads of 100 bp, ~ 10X coverage, Table S8) from NCBI bioproject PRJNA350833 submitted by a previous study (Whitacre et al. 2017). As the UMD_CASPUR_WB_2.0 water buffalo draft genome assembly is highly fragmented, our analysis was based on the Cattle UMD3.1 assembly (Zimin et al. 2009). We masked repeats of the cattle assembly using RepeatMasker (version open-3-3-0) (using the -s option and

cattle RepBase libraries), Tandem Repeats Finder (version 3.21), and WindowMasker. We then aligned the buffalo reads to the masked UMD3.1 using mrsFAST (version 2.5.0.4) (Hach et al. 2010), allowing up to two mismatches (i.e., 48/50, ~96% sequence identity).

CNV calling using Read depth method We then processed aligned reads within sliding windows using the WSSD pipeline as previously described (Bickhart et al. 2012). Reads were counted and the GC bias was corrected using Locally Weighted Scatter-plot Smoother (LOESS). We called the CNVs based on the read depth in three different sizes and types of windows. The procedure and criteria for the CNV calling were similar to that of the previous study (Alkan et al. 2009). We estimated the CN within 1-kb non-overlapping windows across all placed chromosomes. The non-overlapping estimates of CN served as a good approximation of CN within non-masked, non-gapped regions of the genome.

Validation of water buffalo CNVs using aCGH Agilent whole genome high-density CGH arrays containing ~974,016 oligonucleotide probes were designed and fabricated on a single slide to provide an evenly distributed coverage on UMD3.1 with an average interval of ~3.1 kb between probes. We performed standard genomic DNA labelling (Cy3 for samples and Cy5 for references), hybridizations, array scanning, spatial correction, and data normalization as previously described (Liu et al. 2010; Bickhart et al. 2012).

qPCR validation We designed primers using a custom script that incorporated Primer3 and Exonerate to identify unique binding sites for primer design (Bickhart et al. 2012; Untergasser et al. 2012). Only the following Primer3 setting 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 Table S6. We conducted qPCR experiments using SYBR green chemistry in triplicates, each with a reaction volume of 25 μl, as previously described (Hou et al. 2011). PCRs were run on a BioRad MyIQ or iQ5 thermocycler. We chose an intron-exon junction of *BTF3* as a reference location for all qPCR experiments with the assumption that there were two copies of the DNA segment in this region. We performed analysis of resultant crossing cycle thresholds (CT) using the relative comparative CT method and normalized against the control gene. Finally, a value of 3 or above was considered as gain and a value of 1 or below was considered as loss.

FISH validation We selected one hundred twenty-one cattle BAC clones from the bovine BAC library (CHORI-240 at http://bacpacresources.org/bovine240.htm) for experimental validation by FISH (Liu et al. 2010; Bickhart et al. 2012). These clones contain large (≥ 20 kb) regions where copy number variations were predicted in Olimpia. We performed FISH experiments as previously described (Liu et al. 2009; Snijders et al. 2001). We prepare both interphase and metaphase nuclei using the buccal epithelial cells of Olimpia. We examined Metaphase nuclei to identify the chromosomal origins of FISH signals. Interphase nuclei analysis allowed us to evaluate the occurrence of tandem duplications.

Gene content We assessed gene content of cattle CNVRs using the BioMart Database (http://www.ensembl.org/biomart/martview/). Ensembl genes overlapping with CNVRs, completely or partially, were considered as copy number variable and selected for further analysis. To gain an insight into the functional enrichment of the genes with copy number variations, we tested the hypothesis that the PANTHER molecular function, biological process, and pathway terms were under- or overrepresented in CNVRs after false discovery rate (FDR) correction using the PANTHER classification system (Mi et al. 2017).

Comparison of the gene CN between cattle and water buffaloes We collected the CN values of all annotated genes in 75 cattle individuals (Bickhart et al. 2016), and compared them with the gene CN in the water buffalo. We focused on the common genes, shared but CN differential genes, as well as buffalo-specific CN variable genes. Sixteen genes with the highest average CN differences (> 4) in the two species are highlighted in the Fig. 7a.

Acknowledgements We thank Reuben Anderson and Alexandre Dimtchev for technical assistance. 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. The USDA is an equal opportunity provider and employer.

Author Contributions DMB and GEL conceived and designed the experiments. JLW, DI, LI, SGS, TSS, CPVT, CRC, and MV collected samples and/or generated HTS and FISH data. DMB, SL, XK, ML, and BDR performed computational and statistical analyses for HTS, aCGH and qPCR. SL, DMB and GEL wrote the paper. All authors read and approved the final manuscript.

Funding GEL was partially supported by appropriated project 1265-3200-083-00D from the USDA Agricultural Research Service (Beltsville Agricultural Research Center), AFRI grant number 2013-67015-20951 from the USDA National Institute of Food and Agriculture (NIFA) Animal Genome and Reproduction Programs, and BARD grant number US-4997-17 from the US-Israel Binational Agricultural Research and Development (BARD) Fund. WL and DMB were supported by appropriated project 5090-31000-024-00-D from the USDA Agriculture Research Service (Dairy Forage Research Center). WYL and JLW are funded by the JS Davies Bequest to the University of Adelaide.

Data Availability The aCGH raw data from the 14 water buffaloes have been submitted to the NCBI under GEO accession ID GSE118117. All 101 FISH results are posted on http://www.biologia.uniba.it/dl/buffalo_CNV/.

Compliance with ethical standards

Consent for publication Not applicable.

Competing interests The authors declare that they have no competing interests.

Abbreviations aCGH, array Comparative Genomic Hybridization; BoLA, Bovine Leucocyte Antigens; BTF3, basic transcription factor 3; CN, copy number; CNVRs, CNV regions; CNVs, copy number variations; CT, cycle thresholds; DEFB, β-Defensin; FCGR3A, Fc fragment of IgG receptor IIIa; FDR, false discovery rate; FZD3, frizzled class receptor 3; FISH, fluorescence in situ hybridization; HTS, high throughput sequencing; KLRK1, killer cell lectin like receptor K1; LOESS, Locally Weighted Scatter-plot Smoother; MAD2L1, mitotic arrest deficient 2 like 1; MHC, major histocompatibility complex; OR, olfactory receptor; PAG, pregnancy-associated glycoprotein; PI3, peptidase inhibitor 3; RD, read depth; RP, Read pair; SA, sequence assembly; SDs, segmental duplications; SNPs, single nucleosome polymorphisms; SR, split read; STDEVs, standard deviations; TRAV, T cell receptor alpha variable; ULBP3, UL16 binding protein 3; WSSD, whole genome shotgun sequence detection

References

Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C, Malig M, Mutlu O, Sahinalp SC, Gibbs RA, Eichler EE (2009) Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet 41(10):1061–1067

Belaaouaj A, McCarthy R, Baumann M, Gao Z, Ley TJ, Abraham SN, Shapiro SD (1998) Mice lacking neutrophil elastase reveal impaired host defense against gram negative bacterial sepsis. Nat Med 4(5): 615–618

Bickhart DM, Liu GE (2014) The challenges and importance of structural variation detection in livestock. Front Genet 5:37

Bickhart DM, Hou Y, Schroeder SG, Alkan C, Cardone MF, Matukumalli LK, Song J, Schnabel RD, Ventura M, Taylor JF, Garcia JF, van Tassell CP, Sonstegard TS, Eichler EE, Liu GE (2012) Copy number variation of individual cattle genomes using next-generation sequencing. Genome Res 22(4):778–790

Bickhart DM, Xu L, Hutchison JL, Cole JB, Null DJ, Schroeder SG, Song J, Garcia JF, Sonstegard TS, Van Tassell CP et al (2016) Diversity and population-genetic properties of copy number variations and multicopy genes in cattle. DNA research : an international journal for rapid publication of reports on genes and genomes 23(3): 253–262

Brown KH, Dobrinski KP, Lee AS, Gokcumen O, Mills RE, Shi X, Chong WW, Chen JY, Yoo P, David S et al (2012) Extensive genetic diversity and substructuring among zebrafish strains revealed through copy number variant analysis. Proc Natl Acad Sci U S A 109(2):529–534

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(7055):88–93

Colli L, Milanesi M, Vajana E, Iamartino D, Bomba L, Puglisi F, Del Corvo M, Nicolazzi EL, Ahmed SSE, Herrera JRV et al (2018) New Insights on Water Buffalo Genomic Diversity and Post-Domestication Migration Routes From Medium Density SNP Chip Data. Front Genet 9:53

Connelley TK, Degnan K, Longhi CW, Morrison WI (2014) Genomic analysis offers insights into the evolution of the bovine TRA/TRD locus. BMC Genomics 15:994

Doan R, Cohen N, Harrington J, Veazey K, Juras R, Cothran G, McCue ME, Skow L, Dindot SV (2012) Identification of copy number variants in horses. Genome Res 22(5):899–907

Fontanesi L, Martelli PL, Beretti F, Riggio V, Dall'Olio S, Colombo M, Casadio R, Russo V, Portolano B (2010) An initial comparative map of copy number variations in the goat (Capra hircus) genome. BMC Genomics 11:639

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(3):158–165

Fujishima S, Morisaki H, Ishizaka A, Kotake Y, Miyaki M, Yoh K, Sekine K, Sasaki J, Tasaka S, Hasegawa N, Kawai Y, Takeda J, Aikawa N (2008) Neutrophil elastase and systemic inflammatory response syndrome in the initiation and development of acute lung injury among critically ill patients. Biomedicine & pharmacotherapy = Biomedecine & pharmacotherapie 62(5):333–338

Gokcumen O, Lee C (2009) Copy number variants (CNVs) in primate species using array-based comparative genomic hybridization. Methods 49(1):18–25

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(8):576–577

Handsaker RE, Van DV, Berman JR, Genovese G, Kashin S, Boettger LM, SA MC (2015) Large multiallelic copy number variations in humans. Nat Genet 47(3):296–303

Henrichsen CN, Vinckenbosch N, Zollner S, Chaignat E, Pradervand S, Schutz F, Ruedi M, Kaessmann H, Reymond A (2009) Segmental copy number variation shapes tissue transcriptomes. Nat Genet 41(4):424–429

Hou Y, Liu GE, Bickhart DM, Cardone MF, Wang K, Kim ES, Matukumalli LK, Ventura M, Song J, VanRaden PM et al (2011) Genomic characteristics of cattle copy number variations. BMC Genomics 12:127

Iamartino D, Nicolazzi EL, Van Tassell CP, Reecy JM, Fritz-Waters ER, Koltes JE, Biffani S, Sonstegard TS, Schroeder SG, Ajmone-Marsan P et al (2017) Design and validation of a 90K SNP genotyping assay for the water buffalo (Bubalus bubalis). PLoS One 12(10):e0185220

Jiang J, Wang J, Wang H, Zhang Y, Kang H, Feng X, Wang J, Yin Z, Bao W, Zhang Q, Liu JF (2014) Global copy number analyses by next generation sequencing provide insight into pig genome variation. BMC Genomics 15:593

Kato T, Daigo Y, Aragaki M, Ishikawa K, Sato M, Kondo S, Kaji M (2011) Overexpression of MAD2 predicts clinical outcome in primary lung cancer patients. Lung Cancer 74(1):124–131

Klambauer G, Schwarzbauer K, Mayr A, Clevert DA, Mitterecker A, Bodenhofer U, Hochreiter S cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. Nucleic Acids Res, 2012 40(9):e69

Li W, Olivier M (2013) Current analysis platforms and methods for detecting copy number variation. Physiol Genomics 45(1):1–16

Li W, Bickhart DM, Ramunno L, Iamartino D, Williams JL, Liu GE (2018) Comparative sequence alignment reveals River Buffalo genomic structural differences compared with cattle. Genomics. https://doi.org/10.1016/j.ygeno.2018.02.018

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

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(5):693–703

Lucas Lledo JI, Caceres M (2013) On the power and the systematic biases of the detection of chromosomal inversions by paired-end genome sequencing. PLoS One 8(4):e61292

Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, Thomas PD (2017) PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Res 45(D1):D183–D189

Michelizzi VN, Dodson MV, Pan Z, Amaral ME, Michal JJ, McLean DJ, Womack JE, Jiang Z (2010) Water buffalo genome science comes of age. Int J Biol Sci 6(4):333–349

Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK, Chinwalla A, Conrad DF, Fu Y, Grubert F, Hajirasouliha I, Hormozdiari F, Iakoucheva LM, Iqbal Z, Kang S, Kidd JM, Konkel MK, Korn J, Khurana E, Kural D, Lam HY, Leng J, Li R, Li Y, Lin CY, Luo R, Mu XJ, Nemesh J, Peckham HE, Rausch T, Scally A, Shi X, Stromberg MP, Stütz AM, Urban AE, Walker JA, Wu J, Zhang Y, Zhang ZD, Batzer MA, Ding L, Marth GT, McVean G, Sebat J, Snyder M, Wang J, Ye K, Eichler EE, Gerstein MB, Hurles ME, Lee C, McCarroll S, Korbel JO, 1000 Genomes Project (2011) Mapping copy number variation by population-scale genome sequencing. Nature 470(7332):59–65

Nath S, Moghe M, Chowdhury A, Godbole K, Godbole G, Doiphode M, Roychoudhury S (2012) Is germline transmission of MAD2 gene deletion associated with human fetal loss? Mol Hum Reprod 18(11): 554–562

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(3):491–499

Nikolich-Zugich J, Slifka MK, Messaoudi I (2004) The many important facets of T-cell repertoire diversity. Nat Rev Immunol 4(2):123–132

Oldeschulte DL, Halley YA, Wilson ML, Bhattarai EK, Brashear W, Hill J, Metz RP, Johnson CD, Rollins D, Peterson MJ, Bickhart DM, Decker JE, Sewell JF, Seabury CM (2017) Annotated draft genome assemblies for the Northern Bobwhite (colinus virginianus) and the scaled quail (callipepla squamata) reveal disparate estimates of modern genome diversity and historic effective population size. G3 7(9): 3047–3058

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(6):512–U576

Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC (2018) Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods 15(6):461–468

Snijders AM, Nowak N, Segraves R, Blackwood S, Brown N, Conroy J, Hamilton G, Hindle AK, Huey B, Kimura K, Law S, Myambo K, Palmer J, Ylstra B, Yue JP, Gray JW, Jain AN, Pinkel D, Albertson DG (2001) Assembly of microarrays for genome-wide measurement of DNA copy number. Nat Genet 29(3):263–264

Snyder M, Du J, Gerstein M (2010) Personal genome sequencing: current approaches and challenges. Genes Dev 24(5):423–431

Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, Sampas N, Bruhn L, Shendure J, Genomes P et al (2010) Diversity of human copy number variation and multicopy genes. Science 330(6004):641–646

Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, Coe BP, Baker C, Nordenfelt S, Bamshad M, Jorde LB, Posukh OL, Sahakyan H, Watkins WS, Yepiskoposyan L, Abdullah MS, Bravi CM, Capelli C, Hervig T, Wee JTS, Tyler-Smith C, van Driem G, Romero IG, Jha AR, Karachanak-Yankova S, Toncheva D, Comas D, Henn B, Kivisild T, Ruiz-Linares A, Sajantila A, Metspalu E, Parik J, Villems R, Starikovskaya EB, Ayodo G, Beall CM, di Rienzo A, Hammer MF, Khusainova R, Khusnutdinova E, Klitz W, Winkler C, Labuda D, Metspalu M, Tishkoff SA, Dryomov S, Sukernik R, Patterson N, Reich D, Eichler EE (2015) Global diversity, population stratification, and selection of human copy-number variation. Science 349(6253):aab3761

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG (2012) Primer3–new capabilities and interfaces. Nucleic Acids Res 40(15):e115

Vivier E, Tomasello E, Paul P (2002) Lymphocyte activation via NKG2D: towards a new paradigm in immune recognition? Curr Opin Immunol 14(3):306–311

Wallace RM, Pohler KG, Smith MF, Green JA (2015) Placental PAGs: gene origins, expression patterns, and use as markers of pregnancy. Reproduction 149(3):R115–R126

Wang Y, Thekdi N, Smallwood PM, Macke JP, Nathans J (2002) Frizzled-3 is required for the development of major fiber tracts in the rostral CNS. J Neurosci 22(19):8563–8573

Whitacre LK, Hoff JL, Schnabel RD, Albarella S, Ciotola F, Peretti V, Strozzi F, Ferrandi C, Ramunno L, Sonstegard TS, Williams JL, Taylor JF, Decker JE (2017) Elucidating the genetic basis of an oligogenic birth defect using whole genome sequence data in a non-model organism. Bubalus bubalis Scientific reports 7:39719

Williams JL, Iamartino D, Pruitt KD, Sonstegard T, Smith TPL, Low WY, Biagini T, Bomba L, Capomaccio S, Castiglioni B, Coletta A, Corrado F, Ferré F, Iannuzzi L, Lawley C, Macciotta N, McClure M, Mancini G, Matassino D, Mazza R, Milanesi M, Moioli B, Morandi N, Ramunno L, Peretti V, Pilla F, Ramelli P, Schroeder S, Strozzi F, Thibaud-Nissen F, Zicarelli L, Ajmone-Marsan P, Valentini A, Chillemi G, Zimin A (2017) Genome assembly and transcriptome resource for river buffalo, Bubalus bubalis (2n = 50). GigaScience 6(10):1–6

Yi G, Qu L, Liu J, Yan Y, Xu G, Yang N (2014) Genome-wide patterns of copy number variation in the diversified chicken genomes using next-generation sequencing. BMC Genomics 15:962

Zhang Y, Sun D, Yu Y, Zhang Y (2007) Genetic diversity and differentiation of Chinese domestic buffalo based on 30 microsatellite markers. Anim Genet 38(6):569–575

Zhang L, Jia S, Yang M, Xu Y, Li C, Sun J, Huang Y, Lan X, Lei C, Zhou Y, Zhang C, Zhao X, Chen H (2014) Detection of copy number variations and their effects in Chinese bulls. BMC Genomics 15:480

Zhou Y, Utsunomiya YT, Xu L, el HA H, Bickhart DM, Sonstegard TS, Van Tassell CP, Garcia JF, Liu GE (2016) Comparative analyses across cattle genders and breeds reveal the pitfalls caused by false positive and lineage-differential copy number variations. Sci Rep 6: 29219

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(4):R42