UNIVERSITÀ DEGLI STUDI DI MILANO
Ph.D Course in Veterinary and Animal Science
Class XXXII

GENOMIC VARIATION IN LIVESTOCK USING DENSE SNP CHIP DATA

Dott.ssa ERICA GORLA
Tutor Prof. Alessandro Bagnato
Co-Tutor Dott.ssa Maria G. Strillacci

Academic Year 2018-2019

To my Family and Friends

INDEX

ABSTRACT & RESEARCH AIM

- **Part I** describes two possible approaches to investigate Mexican chicken genetic variation, using selective sweeps and Copy Number Variants (CNV). CNVs are genomic polymorphisms that influence phenotypic expression and are an important source of genetic variation in populations.

  The aim of the first study here presented was to characterize the genetic variability of the Mexican chicken's population and to disclose any underlying population structure. A total of 213 chickens were sampled in different rural production units located in 25 states of México. Genotypes were obtained using the Affymetrix Axiom® 600K Chicken Genotyping Array. The Identity by Descent (IBD) and the Principal Components Analysis (PCA) were performed by SVS software on pruned SNPs. Analyses done with ADMIXTURE identified three ancestors and determined, for each individual, the proportion of the genetic contribution from each of the three ancestors. The results of the Neighbor-Joining (NJ) analysis were consistent with those obtained by the PCA. All methods used in this study did not allow a classification of Mexican chicken in distinct genetic groups. A total of 3,059 Run of homozygosity (ROH) were identified and, being mainly short in length (< 4 Mb), these regions are indicative of a low inbreeding level in the population. Finally, findings from the ROH analysis indicated the presence of natural selective pressure in the population of Mexican chicken.

  In the second study we used CNVs to investigate genetic variability in the Mexican Creole chicken and to relate this variation to the available gene annotation. The Hidden Markov Model of the PennCNV software detected a total of 1,924 CNVs in the chicken genome of 256 individuals. Input data were LOGR

Ratio and B allele frequency obtained with the Axiom® Genome-Wide Chicken Genotyping Array (Affymetrix). The mapped CNVs comprised 1,538 gains and 386 losses resulting, at population level, in 1,216 CNV regions (CNVRs), of which 959 gains, 226 losses and 31 complexes (i.e. containing both losses and gains). The CNVRs covered a total of 47 Mb of the whole genome sequence length, corresponding to 5.12 % of the chicken galGal4 autosome assembly. This study allowed a deep insight into the structural variation in the genome of unselected Mexican chicken population, which up to now has not been genetically characterized. The genomic study disclosed that the population, even if presenting extreme morphological variation, couldn't be organized in differentiated genetic subpopulations. Finally, this study provided a chicken CNV map based on the 600K SNP chip array, jointly with a genome-wide gene copy number estimates in a native, unselected for more than 500 years, chicken population.

Genetic variation can be caused by adaptive evolutionary changes and by artificial selection. The genetic makeup of populations is the result of a long-term process of selection and adaptation to specific environments and ecosystems.

The two studies here presented indicate that the Mexican chicken clearly appear to be a unique Creole chicken population that was not subjected to a specific directional selection. Results provide a genetic knowledge that can be used as a basis for the genetic management of a unique genetic resource. Industry is likely envisaging to use the female native populations mating them with selected males to increase the productivity and the economic revenue of family farming agriculture, which is a large reality of United States of México.

- **Part II** describes a CNV scan and a population analysis of turkey populations coming from different countries.

  The domesticated turkey was brought to Europe in late 1500 by Spanish conquerors from Central America, likely from Mexico.

  The evolution of the Mexican turkey population occurred as such independently for more than 500 years from the European ones and the commercial hybrids.

  This study investigates the genomic diversity of several turkey populations using CNVs as source of variation.

  A total of 116 individuals from 6 Italian breeds (Colle Euganei, Bronzato Comune Italiano, Parma e Piacenza, Brianzolo, Nero d'Italia and Ermellinato di Rovigo), 7 Narragansett, 38 commercial hybrids and 31 Mexican turkeys, were processed with the Affymetrix 600K SNP turkey array. The CNV calling was performed with the HMM of PennCNV software. CNV were summarized into CNV regions (CNVRs) at population level using BEDTools. Variability among populations has been addressed by hierarchical clustering (pvclust R package) and by principal component analysis (PCA). A total of 2,987 CNV were identified covering 4.65% of the autosomes of the Turkey_5.0/melGal5 assembly. The CNVRs including at least 2 individuals were 362, 189 gains, 116 losses and 57 complexes. Among these regions the 51% contain genes. This study is the first CNV mapping of turkey population using 600K SNP chip. CNVs clustered the individuals according to population and their geographical origin. CNVs are also known to be indicators of adaptation, as some researches are suggesting investigating different species. The outcomes of this are likely reflecting the human action on domestication of domesticated turkey after its introduction into Europe and the directional selection occurring in the last 40 years to produce a fast-growing heavy bird.

- **Part III** describes the CNV mapping in the Valdostana Red Pied (VRP) cattle breed, an autochthonous Italian dual-purpose cattle population reared in the Alps, and the comparison with the CNV maps detected in previous studies in the Italian Brown Swiss (IBS) and in the Mexican Holstein (HOL). Many studies have focused on identifying CNVs within and between human and livestock populations alike, but only few have explored population-genetic properties in cattle based on CNVs derived from a high-density SNP array.

  In this study in cattle we report a high-resolution CNV scan, using the Illumina 777k BovineHD Beadchip, for VRP, a population that did not undergo strong selection for production traits. After stringent quality control and filtering, CNVs were called across 108 bulls using the PennCNV software. A total of 6,784 CNVs were identified, summarized to 1,723 CNV regions (CNVRs) on 29 autosomes covering a total of ~59 Mb of the UMD3.1 assembly. Among the mapped CNVRs, there were 812 losses, 832 gains and 79 complexes. A total of 171 CNVRs were common to VRP, IBS and HOL. Between VRP and IBS, 474 regions overlapped, while only 313 were in common between VRP and HOL, indicating a more similar genetic structure among populations with common origins, i.e. the Alps. The clustering and admixture analyses showed a clear separation of the three breeds into three distinct clusters. In order to describe the distribution of CNVs within and among breeds we used the pair V_{ST} statistic. We considered only the CNVRs shared by more than 5 individuals within breed. We identified unique and highly differentiated CNVs (n=33), some of which could be due to specific breed selection and adaptation. Genes and QTL within these regions were also characterized adding evidence to the relationship between CNV and adaptation

GENERAL INTRODUCTION

In the last decade thanks to the availability of new technologies such as high-density SNP chips genotyping and short read sequencing and their cost reduction, the possibility to obtain a great deal of previously inaccessible genomic information has opened up. A large part of this genetic information can be used to analyse genetic variability among populations of different species, including livestock populations (Franzer et al., 2007; Zhang et al., 2011; Vignal et al., 2002).

Different indicators can be employed to investigate genetic variability along the genome. In this work we used high-density SNP data focussing on Copy Number Variants (CNV) and Runs of Homozigosytiy (ROH).

CNV are a class of genomic variation known to be related to gene expression deletion may be due to loss of deleterious genes during a species evolution (Hull et al., 2017), while duplication is driven by directional selection (Perry et al., 2007).

ROH are directly related to mating strategies. Long ROH are, in fact, an indicator of recent inbreeding, i.e. mating of related individuals in the last generations (Kirin et al., 2010). Shorter ROH are, on the other hand indicating of ancient mating occurrence among related individuals: recombination events across several generation allow, in fact, to break long DNA tract in homozygosity (Pemberton et al., 2012).

ROH can also be used to identify the genomic regions that are under directional selection according the selection strategies of the populations (Purfield et al, 2010).

Genetic diversity in genomic era

During evolution, natural and human-imposed selection, affected genomic structure of livestock populations. The differences in the genome structure affect phenotypic expression, driving the extreme variability that can be disclosed between native low producing breeds and highly selected one or hybrids (Xu et al., 2014; Fleming et al., 2017). Generally, the native populations are said to be very well adapted to harsh environmental conditions, while selected populations to outperform in artificially controlled environments, as the intensive farming ones (Thornton et al., 2009).

In last decades the artificial selection, also based on genomic information for the last years, was employed to improve performances for productive traits in cattle and chicken, driving a quick change in the genome (Hayes et al., 2009; Meuwissen et al, 2001).

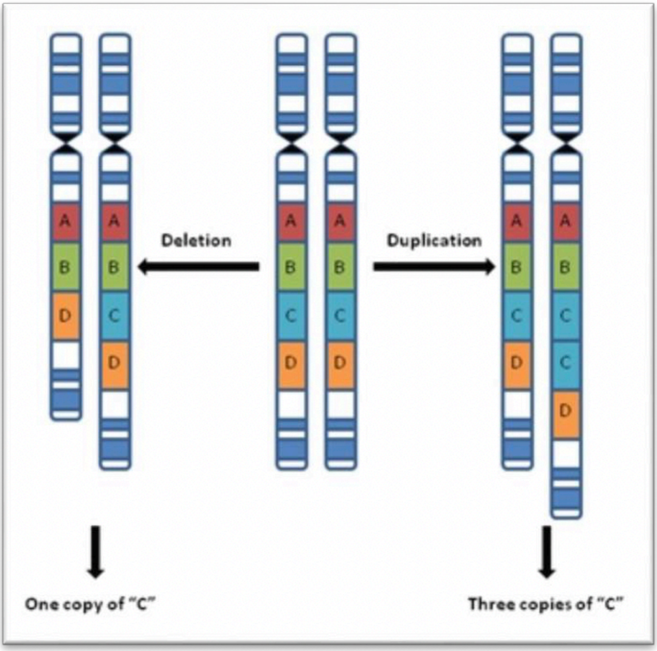
On the other hand, natural selection and adaptation to environment are capable to modify the phenotypic characteristics of individuals over time, and thus of populations, as well as their genomic structure (Hoffman et al., 2000).

The natural selection for adaptive and survival traits as well as the artificial selection for productive traits, may lead to the presence of genomic signatures as a response to selective pressure (Fleming et al, 2016).

Recently one of the research efforts in livestock is addressed to identify strategies to preserve population biodiversity and maintain genetic diversity (Herrero-Medrano et al., 2013). The very recent and fast development of genomic technologies led to the development of many indicators that can be useful to preserve genetic diversity in conjunction with improvement of livestock performances. Among these indicators the most used in livestock in the recent past are SNP markers, a neutral indicator of genomic variation. The ROH can be determined from information on SNP genotype and are now widely suggested to monitor genomic inbreeding in the population.

More recently CNVs are becoming a marker studied in several species. CNVs are an interesting class of non-neutral markers as a large proportion of them is overlapping annotated genes.

Copy number variants



Copy Number Variants are defined as genomic structural variations (duplications or deletions) ranging from at least 50 bp to several mega base (Mb), that can be distributed over the whole genome and that has been found in all species (Mills et al., 2011).

These structural variations affect a larger portion of genome in respect to Single Nucleotide Polymorphism (SNPs), and this result in a significant influence of CNVs on phenotypic variation

(Mills et al., 2011).

CNVs can also impact the phenotype of individuals, altering the allele through different mechanisms, i.e. changing the coding sequence of a gene creating paralogs that can alter gene functions or altering the expression level of a gene, altering the genes dosage (Iskow et al., 2012).

This may lead to phenotypic variation also in selected populations for commercial traits, as well as in disease susceptibility, describing up to the 30% of the genetic variation in gene expression (Stranger et al., 2007; Henrichsen et al., 2009).

The evidence of a direct effect of CNVs in determining complex disease expression in human, e.g autism and schizophrenia, as well as in livestock species has been recently widely studied (Zhang et al., 2009; Norris et al., 2008; Pinto et al., 2010; Sebat et al., 2007). Additionally, differential selection for CNVs has been reported to generate genomic diversity in adaptation to specific environments (Chain et al. 2014; Iskow et al. 2012). Therefore, studies in human and mice, confirm the idea that CNVs could be exposed to selection pressure during the evolution (Zahng et al., 2009).

The first comprehensive human CNV map was edited by Iafrate et al. (2004), and Redon et al. (2006), and since then several studies based on CNV mapping were done in many species, including some livestock species, such as chicken (Gorla et al., 2017; Drobik-Czwarno et al., 2018), cattle (Bagnato et al., 2015; Prinsten et al., 2016), pigs (Ramayo-Caldas, et al., 2010; Schiavo et al., 2014) goat (Liu et al., 2019), and sheep (Liu, et al. 2013; Zhu et al., 2016), using SNP chip. Fewer studies have investigated intra-breeds genetic diversity in cattle (Bickhart et al., 2016) and chicken (Strillacci et al., 2017). The use of CNVs as markers to investigate population genetic diversity among population and to explore population structure is gradually becoming an emerging research topic for livestock animal, even though up to now it has been focused mainly in cattle (Xu et al., 2016; Strillacci et al., 2018).

Techniques and Software for CNVs detection

The techniques currently available for the identification of CNVs are several:

-) fluorescence in situ hybridization technique (FISH). This technique is a type of hybridization that uses probes whose presence can be highlighted by marking with fluorochromes. The principle on which it is based is that for which any DNA sequence is capable of binding itself to its complementary sequence. The probes hybridized and marked with fluorochromes are directly visualized under the microscope. FISH allows to identify CNVs as visible microscopic alterations (Wain et al., 2009).

-) comparative genomic hybridization array (aCGH). This technique requires DNA and control samples to be labelled respectively with Cyclochrome (green) and Cy5 (red), and then hybridized together on a specific Microarray (long oligonucleotides or BAC clones). Both the total red and green fluorescence intensity for each sample is measured, as is the ratio between the intensities of the two fluorochromes. These intensities are then processed with specific software to identify CNVs

-) Next Generation Sequencing (NGS). The NGS technique allows to detect more types of structural variation with a single sequencing trial. The CNV detection methods based on this technique can be classified into five main different strategies:

Paired-end mapping (PEM), Split read (SR) -based methods, Read depth (RD) -based approach, Assembly (AS) -based approach and a combined RD-PEM approach (Zhao et al. 2013).

-) SNP genotyping array (SNP chip). The SNP genotyping array is a hybridization-based technique that allows the identification of hundreds of thousands of structural variants (SNPs) with a high degree of resolution. An SNP array consists of a set of DNA probes (specific for the amplification of each SNP) fixed to the solid surface of the chip. The principle on which this technique is based is given by the specificity of hybridization between complementary nucleotide sequences. The last two techniques are the most used and most reliable for genome wide detection of CNVs.

-) quantitative PCR (qPCR).

It is a method based on a simple modification of PCR, that allow the quantification of target DNA, using fluorescent or intercalating dyes to detect PCR product as it accumulates during PCR cycles. In addition to being used to quantify DNA, (mitochondrial DNA and cDNA), qPCR can be used in the validation of CNVs. (Wain et al, 2009)

A wide range of algorithms is currently available for the identification of CNVs, starting from the data obtained from different genotyping techniques. To identify CNVs from the data obtained with the SNP chip, i.e. Log R Ratio (LRR) and B allele frequency (BAF) two of the most commonly used and reliable algorithms are the HMM of PennCNV (Wang et al., 2007) and CNAM of SVS8 by Golden Helix (Golden Helix Inc., Bozeman, MT, USA).

PennCNV is one of the most used software for CNV identification and use the Hidden Marckov Model (HMM) for the CNV detection. It incorporates multiple factors, including the log R ratio (LRR), B Allele Frequency (BAF), the marker distance, and the population frequency of the B allele (PFB).

-) The BAF, a normalized measure of fluorescence intensity of each allele, allows defining if a CNV is present in the homozygous or heterozygous form.

-) The LRR, normalized measurement of the total allelic intensity signal of a given SNP, allows to attribute the CNV state, i.e. duplication or deletion state (defined also as gain or loss) in a given chromosomal region. PennCNV integrates a computational approach by applying a regression model to the GC content to reduce waviness. Copy number variations were also detected using the Hidden Marckov Model parameter file.

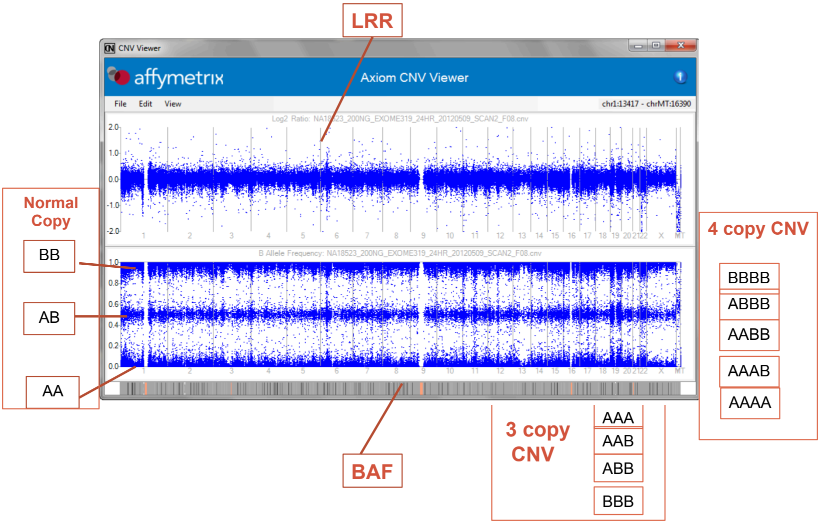
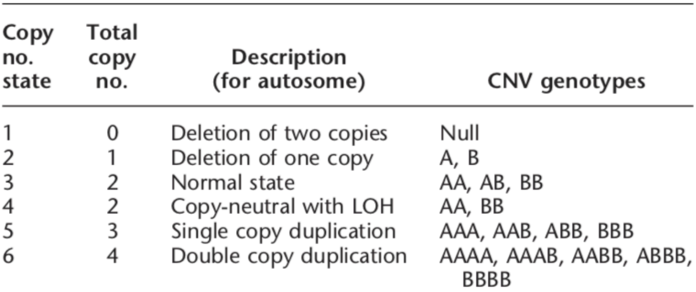

LRR
CNV Viewer
affymetrix
Axiom CNV Viewer
File Edit View
chr1:13417 - chrMT:16390
Log2 Ratio: NA18523_200NG_EXOME319_24HR_20120509_SCAN2_F08.cnv
2.0
1.0
0.0
-1.0
-2.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X MT
B Allele Frequency: NA18523_200NG_EXOME319_24HR_20120509_SCAN2_F08.cnv
1.0
0.8
0.6
0.4
0.2
0.0
Normal Copy
BB
AB
AA
4 copy CNV
BBBB
ABBB
AABB
AAAB
AAAA
3 copy CNV
AAA
AAB
ABB
BBB
BAF

Figure 1. An illustration of Log R Ratio (LRR) and B Allele Freq (BAF) values for the chromosome 1 of an individual

A normal chromosome region has LRR values cantered around zero and has three BAF genotype clusters, as represented as AA, AB, and BB genotypes in boxes, and with LRR values centred around zero. Therefore, the increased copy number for a CNV region can be detected based on an increased number of peaks in the BAF distribution, as well as increased LRR values.



Copy no. state	Total copy no.	Description (for autosome)	CNV genotypes
1	0	Deletion of two copies	Null
2	1	Deletion of one copy	A, B
3	2	Normal state	AA, AB, BB
4	2	Copy-neutral with LOH	AA, BB
5	3	Single copy duplication	AAA, AAB, ABB, BBB
6	4	Double copy duplication	AAAA, AAAB, AABB, ABBB, BBBB

Table 1. Description of hidden states, copy numbers and their genotypes for each possible detection state from PennCNV

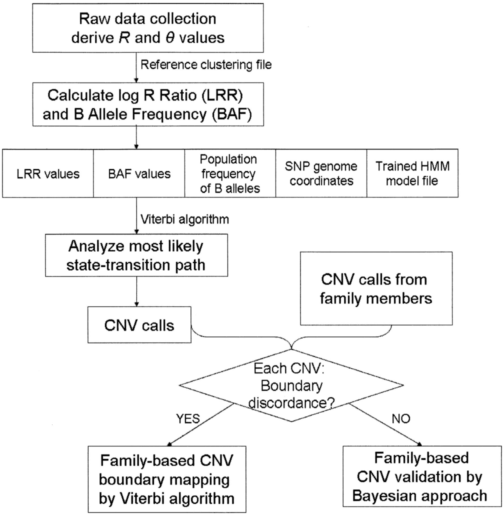

Raw data collection derive R and θ values
Reference clustering file
Calculate log R Ratio (LRR) and B Allele Frequency (BAF)
LRR values
BAF values
Population frequency of B alleles
SNP genome coordinates
Trained HMM model file
Viterbi algorithm
Analyze most likely state-transition path
CNV calls from family members
CNV calls
Each CNV: Boundary discordance?
YES
NO
Family-based CNV boundary mapping by Viterbi algorithm
Family-based CNV validation by Bayesian approach

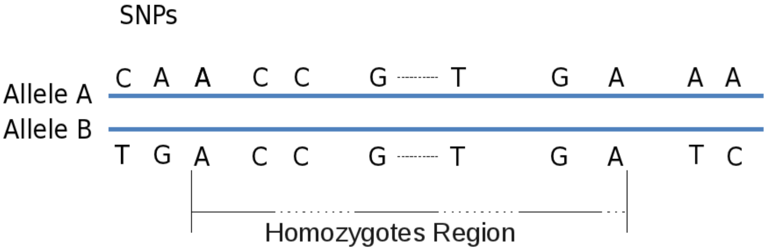
Figure 2. A flowchart outlining the procedure for CNV calling from genotyping data (Wang et al., 2007)

The SVS 8.4 by Golden Helix® (Golden Helix Inc., Bozeman, MT, USA) use a different algorithm, the Copy Number Analysis Module (CNAM) using only LRR as input information. The CNAM is able to process raw intensity data, to detect copy number variations. It identifies the CNV boundaries at a single probe level.

The pipeline in SVS8 performs an accurate Quality Assurance on LRR, using quality filters such as the derivative log ratio spread (DLRS), Genomic waves detection in log ratio data and Principal component analysis (PCA), as in Diskin et al., (2008). This step is fundamental to reduce the false positive calling of CNV. CNV detection can be then performed using two segmentation algorithms: the univariate method, used mainly for the detection of rare and/or large CNV, which considers only one sample at the time. The multivariate method, that uses all samples at the same time and is recommended to detect small, common CNV.

The basic principle of the CNAM is conceived to identify the CNV in the genome where a given sample's mean LRR value is different from the population average reference value. When the mean LRR is around zero the sample has the same number of copies as the reference. Otherwise, when the LRR segment mean is above zero usually there is a copy number gain, and when the LRR segment mean is below zero, there is a copy number loss. The CNAM is able then to detect with a specific methodology when change respect to the population reference are to be considered true gain or loss (Golden Helix Inc., Bozeman, MT, USA).

Runs of homozygosity



Runs of homozygosity (ROH) are defined as continuous homozygous segments in the DNA sequence that are common among individuals and populations (Gibson et al., 2006) and may be used to define individual autozygosity (McQuillan et al., 2008). This process occurs when parents share common ancestor and pass shared DNA fragments to their offspring, which inhered chromosomal segments that are identical by descent (IBD) from both parents (Wright 1922). Those homozygous segments can form ROH in the progeny genome (Broman & Weber 1999). Generally, in livestock it is accepted that ROH with a length of ~10 Mb are a consequence of recent inbreeding (maximum five generations ago), while shorter ROH (~1 Mb) can considered a consequence of ancient positive selection effect (50 generations ago) (Purfield et al., 2012). In fact, recombination events may break long chromosome into shorter segments reducing their size along the selection process.

Since 2006 the use of high-density SNP array to identified ROH was explored first in human (Gibson et al., 2006; McQuillan et al., 2008; Kirin et al., 2010; Nothnagel et al., 2010) and then livestock species: first in cattle (Ferencakovic et al., 2013a, Ferencakovic et al., 2013b; Kim et al., 2013), swine (Bosse et al., 2012; Herrero-Medrano et al. 2013), sheep (Beynon et al. 2015; Muchadeyi et al. 2015), goat (Guangul 2014) and chicken (Strillacci et al., 2018).

Two major methods can be used to define ROH: observational genotype-counting algorithms (Purcell et al. 2007) and model-based algorithms (Pemberton et al, 2012).

The first approach consists in scanning using an algorithm each chromosome by moving a fixed size window along the whole length of the genome searching stretches of consecutive homozygous SNPs (Purcell et al. 2007).

The software mainly used to ROH detection are: PLINK (Purcell et al. 2007), SVS (Golden Helix SNP & Variation Suite v.7.6.8), GERMLINE (Gusev et al. 2009), BEAGLE (Browning & Browning 2010).

PLINK v1.9 software (Purcell et al. 2007), for example, uses the first approach, by considering a given SNP to be potentially in an ROH and calculating the proportion of completely homozygous windows that encompass the given SNP. If this proportion is higher than a defined threshold, the SNP is considered as being in a ROH.

GERMLINE (Gusev et al. 2009) and SVS Golden Helix 8.4 software (SVS) (Golden Helix Inc., Bozeman, MT, USA) on the other hand, are examples of haplotype-matching algorithms for calculation of identity- by-descent (IBD) and can also be used to identify ROH, as a special case of IBD within an individual.

Finally, BEAGLE (Browning and Browning 2010) is based on a Model-based approaches, which use Hidden Markov Models (HMM) to account for background levels of LD.

A strong limitation for the studies based on ROH, is the lack of a common criteria for their definition across population and studies, that is not only determined by the ROH length but also function of parameters used in their detection as number of missing genotypes of heterozygous markers allowed in a run. This lack of consensus makes it difficult to compare studies as also commented by authors comparing different algorithms for their detection (Howrigan et al. 2011; Ku et al. 2011). It is then always useful to consider the minimum length of ROH, the density of the SNP chip used, the minimum number of SNPs allowed in a ROH as suggested by Peripolli et al. (2017) in a recent review of ROH studies in livestock.

Reference

- Bagnato, A., Strillacci, M.G., Pellegrino, L., Schiavini, F., Frigo, E., Rossoni, A., Fontanesi, L., Maltecca, C., Prinsen, R.T. and Dolezal, M.A., (2015). Identification and validation of copy number variants in Italian Brown Swiss dairy cattle using Illumina Bovine SNP50 Beadchip®. *Ital. J. Anim. Sci. 14*(3), p.3900 https://doi.org/10.4081/ijas.2015.3900.
- Beynon S.E., Slavov G.T., Farré M., Sunduimijid B., Waddams K.·, Davies B.· et al. (2015) Population structure and history of the Welsh sheep breeds determined by whole genome genotyping. BMC Genet. 2015 Jun 20;16:65. doi: 10.1186/s12863-015-0216-x.
- Bickhart D.M., Xu L., Hutchison J.L., Cole J.B., Null D.J., Schroeder S.G., et al. (2016) Diversity and population-genetic properties of copy number variations and multicopy genes in cattle. DNA Res. 23(3):253-262. doi: 10.1093/dnares/dsw013.
- Bosse M., Megens H.J., Madsen O., Paudel Y., Frantz L.A., Schook L.B., et al. (2012) Regions of homozygosity in the porcine

genome: consequence of demography and the recombination landscape. PLoS Genetics 8, e1003100. doi: 10.1371/journal.pgen.1003100

- Broman K. & Weber J.L. (1999) Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain. Am J Hum Genet. 1999 Dec;65(6):1493-500.
- Browning S.R., Browning B.L. (2010) High-resolution detection of identity by descent in unrelated individuals. Am J Hum Genet, 86(4):526-539. doi: 10.1016/j.ajhg.2010.02.021.
- Chain F.J.J., Feulner P.G.D., Panchal M., Eizaguirre C., Samonte I.E., Kalbe M., et al. (2014) Extensive Copy-Number Variation of Young Genes across Stickleback Populations. PLoS Genet 10(12): e1004830. https://doi.org/10.1371/journal.pgen.1004830.
- Diskin S.J., Li M., Hou C., Yang S., Glessner J., Hakonarson, H., et al. (2008) Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. Nucleic. Acids. Res. 36:e126. doi: 10.1093/nar/gkn556.
- Drobik-Czwarno W., Wolc, A., Fulton J. E., & Dekkers J. C. (2018). Detection of copy number variations in brown and white layers based on genotyping panels with different densities. Genetics Selection Evolution, 50(1), 54.
- Ferenčaković M., Sölkner J., Curik I. (2013) Estimating autozygosity from high-throughput information: effects of SNP density and genotyping errors. Genet Sel Evol. Oct 29;45:42. doi: 10.1186/1297-9686-45-42.
- Fleming, D. S., J. E. Koltes, A. D. Markey, C. J Schmidt, C. Ashwell, M. F. Rothschild, M. E. Persia, J. M. Reecy, and S. J. Lamont. (2016). Genomic analysis of Ugandan and Rwandan chicken ecotypes using a 600 k genotyping array. BMC Genomics. 17:407. doi: 10.1186/s12864-016-2711-5.

- Fleming D.S., S. Weigend, H. Simianer, A. Weigend, M. F. Rothschild, C. J. Schmidt, C. M. Ashwell, M. E. Persia, J. M. Reecy and S. J. Lamont. (2017). Genomic Comparison of Indigenous African and Northern European Chickens Reveals Putative Mechanisms of Stress Tolerance Related to Environmental Selection Pressure. G3 (Bethesda). 7(5): 1525–1537. doi: 10.1534/g3.117.041228.
- Frazer K.A., Ballinger D.G., Cox D.R., et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature.*;449(7164):851–861. doi: 10.1038/nature06258
- Gibson J., Newton E.M. & Collins A. (2006) Extended tracts of homozygosity in outbred human populations. Human Molecular Genetics 15, 789–95. doi: 10.1093/hmg/ddi493.
- Gorla E., Cozzi M.C., Román-Ponce S.I., López F.R., Vega-Murillo V.E., Cerolini S., Bagnato A., Strillacci M.G. (2017) Genomic variability in Mexican chicken population using copy number variants. BMC Genet. Jul 3;18(1):61. doi: 10.1186/s12863-017-0524-4.
- Guangul S.A. (2014). Design of community based breeding programs for two indigenous goat breeds of Ethiopia. Doctoral thesis, University of Natural Resources and Life Sciences, Vienna.
- Gusev A., Lowe J.K., Stoffel M., Daly M.J., Altshuler D., Breslow J.L., et al. (2009) Whole population, genome-wide mapping of hidden relatedness. Genome Res. 19(2):318-326 doi: 10.1101/gr.081398.108.
- Hayes, B.J., P.J. Bowman, A.J. Chamberlain, and M.E. Goddard. (2009) Invited review: Genomic selection in dairy cattle: Progress and challenges. J. Dairy Sci. 92:433–443. doi: 10.3168/jds.2008-1646.
- Henrichsen C.N., Vinckenbosch N., Zollner S., Chaignat E., Pradervand S., Schutz F., et al. (2009) Segmental copy number

variation shapes tissue transcriptomes. Nat. Genet.; 41:424–429. doi: 10.1038/ng.345.

- Herrero-Medrano J.M., Megens H.-J., Groenen M.A.M., Ramis G., Bosse M., Perez-Enciso M. & Crooijmans R.P.M.A. (2013) Conservation genomic analysis of domestic and wild pig populations from the Iberian Peninsula. BMC Genetics 14, 106. doi: 10.1186/1471-2156-14-106
- Hoffmann A. A., M. J. Hercus. (2000) Environmental Stress as an Evolutionary Force. BioScience, 50, 3, March 2000, Pages 217–226. https://doi.org/10.1641/0006-3568(2000)050[0217:ESAAEF]2.3.CO;2
- How.rigan D.P., Simonson M.A. & Keller M.C. (2011) Detecting autozygosity through runs of homozygosity: a comparison of three autozygosity detection algorithms. BMC Genomics 12, 460. DOI: 10.1186/1471-2164-12-460
- Hull, R.M., Cruz, C., Jack, C.V., Houseley. J. (2017) Environmental change drives accelerated adaptation through stimulated copy number variation. PLoS Biol. 15(6):e2001333. doi: 10.1371/journal.pbio.2001333.
- Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W., Lee, C., (2004). Detection of large-scale variation in the human genome. Nat Genet. 36(9): 949-51. Doi: 10.1038/ng1416.
- Kim E.S., Cole J.B., Huson H., Wiggans G.R., Van Tassell C.P., Crooker B.A., et al. (2013) Effect of artificial selection on runs of homozygosity in U.S. Holstein cattle. PLoS One 8, e80813. doi: 10.1371/journal.pone.0080813.
- Kirin M., McQuillan R., Franklin C., Campbell H., McKeigue P.M. & Wilson J.F. (2010) Genomic runs of homozygosity record population history and consanguinity. PLoS One 5, e13996. doi: 10.1371/journal.pone.0013996.

- Ku C.S., Naidoo N., Teo S.M. & Pawitan Y. (2011) Regions of homozygosity and their impact on complex diseases and traits. Hum Genet. 2011 Jan;129(1):1-15. doi: 10.1007/s00439-010-0920-6.
- Liu, J., Zhang, L., Xu, L., Ren, H., Lu, J., Zhang, X., et al. (2013). Analysis of copy number variations in the sheep genome using 50K SNP BeadChip array. BMC genomics, 14(1), 229. doi:10.1186/1471-2164-14-229.
- Liu M., Zhou Y., Rosen B.D., Van Tassell C.P, Stella A., Tosser-Klopp G., et al. (2019). Diversity of copy number variation in the worldwide goat population. Heredity 122, 636–646 doi:10.1038/s41437-018-0150-6.
- McQuillan R., Leutenegger A. L., Abdel-Rahman R., Franklin C. S., Pericic M., Barac-Lauc L.,et al. (2008) Runs of homozygosity in European populations. American J. Hum. Genet. 83, 359–72. doi: 10.1016/j.ajhg.2008.08.007.
- Meuwissen T.H.E., Hayes B. J., Goddard M. E. (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics, 2001, 157: 1819–1829.
- Mills R.E., Walter K., Stewart C., Handsaker R.E., Chen K., Alkan C., et al. (2011) Mapping copy number variation by population-scale genome sequencing. Nature. 470: 59-65. doi:10.1038/nature09708.
- Iskow R.C., Gokcumen O. & Lee C. (2012) Exploring the role of copy number variants in human adaptation. *Trends in genetics TIG* 28, 245–257, https://doi.org/10.1016/j.tig.2012.03.002.
- Muchadeyi F.C., Malesa M.T., Soma P. & Dzomba E.F. (2015) Runs of homozygosity in Swakara pelt producing sheep: implications on sub-vital performance. Proc. Assoc. Advmt. Anim. Breed. Genet. (2015) 21: 310-313.
- Norris B.J., Whan V.A. (2008) A gene duplication affecting

expression of the ovine ASIP gene is responsible for white and black sheep. Genome Res 18: 1282–1293. doi: 10.1101/gr.072090.107.

- Nothnagel M., Lu T.T., Kayser M. & Krawczak M. (2010) Genomic and geographic distribution of SNP defined runs of homozygosity in Europeans. Hum Mol Gen 19, 2927–35. doi: 10.1093/hmg/ddq198.
- Pemberton T.J., Absher D., Feldman M.W., Myers R.M., Rosenberg N.A., Li J.Z. (2012) Genomic patterns of homozygosity in worldwide human populations. Am J Hum Genet.91(2):275–92. doi: 10.1016/j.ajhg.2012.06.014.
- Peripolli, E., Munari, D.P., Silva, M.V.G.B., Lima, A.L.F., Irgang, R., & Baldi, F. (2017). Runs of homozygosity: current knowledge and applications in livestock. Anim Genet. 2017 Jun;48(3):255-271. doi: 10.1111/age.12526.
- Perry, G.H., Dominy, N.J., Claw, K.G., Lee, A.S., Fiegler, H., Redon, R., et al. (2007). Diet and the evolution of human amylase gene copy number variation. Nat. Genet. 39(10):1256-60. DOI: 10.1038/ng2123.
- Pinto D., Pagnamenta A.T., Klei L., Anney R., Merico D., et al. (2010) Functional impact of global rare copy number variation in autism spectrum disorders. Nature 466: 368–372. doi: 10.1038/nature09146
- Prinsen R.T.M.M., Strillacci M.G., Schiavini F., Santus E., Rossoni A., Maurer V., Bieber A., et al. (2016) A genome-wide scan of copy number variants using high-density SNPs in Brown Swiss dairy cattle. Liv Sci;191:153–160, http://dx.doi.org/10.1016/j.livsci.2016.08.006.
- Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M.A., Bender D., et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum

Genet. 81(3):559-575. doi:10.1086/519795.

- Purfield, D.C., Berry, D.P., McParland, S., and Bradley, D.G. (2012) Runs of homozygosity and population history in cattle. BMC genetics, 13(1), 70. doi: 10.1186/1471-2156-13-70.
- Ramayo-Caldas Y.,Castelló A., Pena R.N., Alves E., Mercadé A., Souza C.A., et al. (2010) Copy number variation in the porcine genome inferred from a 60 k SNP BeadChip. *BMC genomics* 11, 1 doi: 10.1186/1471-2164-11-593.
- Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M. H., Carson, A.R., Chen, W. (2006) Global variation in copy number in the human genome. Nature 444:444-454.
- Schiavo G., Dolezal M.A., Scotti E., Bertolini F., Calò D.G., Galimberti G., et al. (2014) Copy number variants in Italian Large White pigs detected using high-density single nucleotide polymorphisms and their association with back fat thickness. Anim Genet. 45:745–749. doi: 10.1111/age.12180.
- Sebat J., Lakshmi B., Malhotra D., Troge J., Lese-Martin C., et al. (2007) Strong Association of De Novo Copy Number Mutations with Autism. Science 316: 445–449. doi:10.1126/science.1138659.
- Strillacci M.G., Cozzi M.C., Gorla E., Mosca F., Schiavini F., Román-Ponce S.I., et al. (2017) Genomic and genetic variability of six chicken populations using single nucleotide polymorphism and copy number variants as markers. Animal.11(5):737–45. doi: 10.1017/S1751731116002135.
- Strillacci M.G., Gorla E., Cozzi M.C., Vevey M., Genova F., Scienski K., et al. (2018) A copy number variant scan in the autochthonous Valdostana Red Pied cattle breed and comparison with specialized dairy populations. PLoS ONE 13(9):e0204669. https://doi.org/10.1371/journal.pone.0204669

- Stranger B.E., Forrest M.S., Dunning M., Ingle C.E., Beazley C., et al. (2007) Relative Impact of Nucleotide and Copy Number Variation on Gene Expression Phenotypes. Science 315: 848–853. DOI:10.1126/science.1136678
- Thornton, P. K., van de Steeg J., Notenbaert, A. and Herrero M. (2009) The impacts of climate change on livestock and livestock systems in developing countries: a review of what we know and what we need to know. Agric. Syst. 101: 113–127. https://doi.org/10.1016/j.agsy.2009.05.002.
- Vignal, A., Milan, D., SanCristobal, M., & Eggen, A. (2002). A review on SNP and other types of molecular markers and their use in animal genetics. *Genet Sel Evol. 34*(3), 275. doi: 10.1186/1297-9686-34-3-275.
- Wain L. V., Armour J. A. L., Tobin M. D. (2009). Genomic copy number variation, human health, and disease. Lancet. 374:340–50. doi: 10.1016/S0140-6736(09)60249-X.
- Wang K., Li M., Hadley D., Liu R., Glessner J., Grant S.F. et al. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res., 17(11), 1665-1674 doi: 10.1101/gr.6861907.
- Wright S. (1922). Coefficients of inbreeding and relationship. American Naturalist 56, 330–8.
- Xu L., Bickhart D.M., Cole J.B., Schroeder S.G., Song J., Van Tassell C.P, et al. (2015) Genomic Signatures Reveal New Evidences for Selection of Important Traits in Domestic Cattle. *Mol Biol Evol. 32(3):711-25*. doi: 10.1093/molbev/msu333.
- Xu L., Hou Y., Bickhart D.M., Yang Z., Hay el H.A., Song J., et al. (2016) Population-genetic properties of differentiated copy number variations in cattle. Sci. Rep. 2016;6:23161. doi: 10.1038/srep23161.

- Zhang F., Gu W., Hurles,M.E., & Lupski, J.R. (2009). Copy number variation in human health, disease, and evolution. *Annual review of genomics and human genetics*, *10*, 451-481. doi: 10.1146/annurev.genom.9.081307.164217
- Zhang, J., Chiodini R., Badr A., & Zhang, G. (2011). The impact of next-generation sequencing on genomics. *J Genet Genomics 20; 38(3): 95–109. doi: 10.1016/j.jgg.2011.02.003.*
- Zhao M., Wang Q., Wang Q., Jia P., Zhao Z., (2013). Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC Bioinformatics 14 Suppl. 11. doi: 10.1186/1471-2105-14-S11-S1.
- Zhu, C., Fan, H., Yuan, Z., Hu, S., Ma, X., Xuan, J., et al. (2016) Genome-wide detection of CNVs in Chinese indigenous sheep with different types of tails using ovine high-density 600K SNP arrays. Sci. Rep., *6*, 27822. doi: 10.1038/srep27822.

PART I

Investigation of genomic variability in Mexican chicken populations

Local chicken populations are considered an important genetic resource; they are able to adapt successfully during the years in areas with peculiar environmental characteristics, with limited support (Hall and Bradley, 1995) by farmers in terms of health management, feed supply and recovery facilities.

Often these populations are not of economic interest for intensive farming practices and so there is a lack of knowledge about their phenotype and genetic variation, with consequent possible loss of information on the genes that favouring their adaptability to local harsh environments (Mahammi et al., 2016).

In Mexico the poultry population, despite showing large morphological variability, is not classified into breeds but considered a unique backyard population generally classified as "creole" chicken (*Gallus gallus domesticus*), resulting from undefined crosses among different reeds imported into Mexico from Spanish conquerors (Segura-Correra et al., 2004; Rodriguez et al., 1996).

Creole chickens show a wide range of variable biotypes, with different morphological features and characterized by high feed conversion, low growth rate, low egg production, and small egg size under semi-intensive or harsh environmental conditions (Segura-Correa et al., 2004, 2005). The Mexican chicken population has been, de facto, under natural adaptive selection for more than 500 years, making it a very interesting case for studying genetic variation related to resilience in harsh environments.

In the past only a few studies tried to characterize phenotypes and performance of Mexican creole chickens and up to now, no molecular characterisation studies related to genetic variability and phylogenetic analysis of this population had been realized using dense panels of SNPs. The Mexican population is, in fact, a genetic resource that could express genes lost in the industrial selection process, mainly targeted to increase meat and egg productions.

In the absence of population characterization data and documentation of their origin, DNA polymorphism can provide a valuable and one of the most reliable indicators of genetic diversity within and between a given set of populations (Ceccobelli et al. 2013).

We here used different genetic information to investigate genetic variability of a sample of Mexican chickens, in order to reveal any underlying population structure and breed differentiation. We use a dense SNP panel to detect and analyse CNVs and ROH to identify genetic variation and evidences of selection signatures. We also identified the genes that according to the mapped genomic variation, appear to be under positive selection for adaptive variants in an outbred population.

Reference

- Ceccobelli S., Lorenzo P.D., Lancioni H., Castellini C., Ibáñez L.V., Sabbioni A., et al. (2013) Phylogeny, genetic relationships and population structure of five Italian local chicken breeds. *Italian J AnimSci*. 12(3):e66. https://doi.org/10.4081/ijas.2013.e66.
- Hall, S.J., and Bradley D.G. (1995). Conserving livestock breed biodiversity. *Trends. Ecol. Evol*. 10: 267–270 https://doi.org/10.1016/0169-5347(95)90005-5.
- Mahammi F.Z., Gaouar S.B. S., Laloë D., Faugeras R., Tabet-Aoul N., Rognon X., et al. (2016). A molecular analysis of the patterns of genetic diversity in local chickens from western Algeria in comparison with commercial lines and wild jungle fowls. *J. Anim. Breed. Genet.* 133 59–70. doi: 10.1111/jbg.12151.
- Rodriguez J.C., Allaway C.E., Wassink G.J., Segura J.C., Rivera T. (1996) Estudio de la Avicultura de traspatio en el municipio de Dzununcàn. *Yucatàn Vet Mex*. 27(3):215–9.
- Segura-Correa J.C., Sarmiento-Franco L., Magaña-Monforte J.G., Santos-Ricalde R. (2004). Productive performance of Creole chickens and their crosses raised under semi-intensive management conditions in Yucatan, Mexico. *Br. Poult. Sci.* 45(3):342–5. doi: 10.1080/00071660410001730833.
- Segura-Correa J.C., Juarez-Caratachea A., Sarmiento-Franco L.,

Santos-Ricalde R. (2005). Growth of Creole chickens raised under tropical conditions of Mexico. *Trop Anim Health Prod*. 37(4):327–32. doi: 10.1007/s11250-005-3863-5.

I) Looking at genetic structure and selection signatures of the Mexican chicken population using Single Nucleotide Polymorphism markers

M.G. Strillacci, V.E. Vega-Murillo, S.I. Román-Ponce, F.J. Ruiz López, M.C. Cozzi, **E. Gorla**, S. Cerolini, F. Bertolini, L. Fontanesi, A. Bagnato (2018).

Poultry Science
97:791–802.

doi/10.3382/ps/pex374/4767756

Introduction

The knowledge of the genetic variation within and across populations is essential in the process of identification of local genetic resources (i.e. individuals of local poultry breeds) to be maintained in animal genetics conservation efforts (Cavalchini et al., 2007). Microsatellites markers have been widely used to analyse genetic variability in the chicken population (Strillacci et al., 2009; Al-Qamashoui et al., 2014; Ceccobelli et al., 2015). Recently, the availability of high throughput genomic information, i.e. sequencing data and high-density Single Nucleotide Polymorphism (SNP) arrays, has opened the possibility to investigate the genetic structure of populations using a very large number of markers and to highlight genomic regions where events related to selection pressure occur (Fleming et al. 2016; Strillacci et al., 2017). Chicken can be easily utilized for the study of the signatures of selection under artificial breeding conditions, thanks to their relatively fast reproduction time (Brown et al., 2003). Theoretically, functional genes under selection are exposed to a change in allele frequency that can be identified analysing the characteristic DNA pattern that derives, known as selection signature (Fan et al., 2014). In other words, selection signatures are, particular patterns of DNA that can be identified in regions of the genome that include a mutation, that is, or have been, under selection in the population (Qanbari and Simianer, 2014). Whenever in positive selection for a particular allele, these regions are expected to exhibit larger homozygosity than expected under Hardy Weinberg equilibrium. Many measures can be utilized to estimate genetic variability pattern along the genome using marker data; among them Runs of Homozygosity (ROH) are contiguous lengths of homozygous genotypes that develop as a result of parental transmission of identical haplotypes (Gibson et al., 2006). Long ROH (~10 Mb) are a consequence of recent inbreeding (up to five generations ago), whereas shorter ROH (~1 Mb) can be related to a more distant ancestral positive selection effect (up to 50 generations ago), because of recombination events that break long chromosome into segments (Mastrangelo et al., 2016) have reduced their size along the reproductive events. Recently the availability of sequencing and high-throughput SNP datasets has permitted to release chromosome-wide molecular diversity and population structure studies (Nimmakayala et al. 2014). Furthermore, it is possible to disclose traces of positive selection and identify possible candidate genes associated with selection (Fan et al., 2014).

Local chicken populations are considered an important genetic resource, derived after thousands of years of successful adaptation in areas with peculiar environmental characteristics, with limited veterinary and management support (Hall and Bradley, 1995). Phenotypic traits variability is little known in backyard poultry population, as well as those genes that cause their adaptability to local environments. It is also not clear if the geographical origin of that local chicken population is one of the causes of their genetic differentiation, making them so various (Mahammi et al., 2016). In México, poultry population is not classified in breeds, but there is a diffusion of the Creole chicken (*Gallus gallus domesticus*), coming from European chickens brought to México by the Spanish conquerors during the 16th century. They originate form undefined crosses among different breeds for almost 500 years. Because of that, Creole chickens include a wide range of variable biotypes, having different morphological features and characterized by high feed conversion, low growth rate, low egg production and small egg size under semi-intensive or scavenging conditions (Segura-Correa et al., 2004, 2005). The Mexican population is, de facto, under natural adaptive selection for more than five centuries making it a very interesting one to disclose genetic variation related to resilience in harsh environments. The Mexican population is in fact a genetic resource that can express genes lost in the industrial selection process, targeted to increase meat and egg productions.

As recently well disclosed by Fleming et al. (2016, 2017) studying genetic variation in African native populations, the existence of proprietary genetic variation in native breeds related to specific environmental conditions (e.g. hot and humid climates or heat waves) is the basic knowledge for its introgression in F1 individuals, crossing for native females populations (natural selection occurring in population) and artificially selected males (artificial selection). To our knowledge, there have been some attempt to characterize phenotype and performance of Mexican creole chickens but up to now, no molecular characterisation studies related to genetic variability and phylogenetic analysis of this population have yet been realized using dense panels of SNPs, except the recent study of Gorla et al. (2017) who used Copy Number Variants to dissect genetic variability in the Mexican population.

The aim of this study was to describe the genetic variability of Mexican chickens to reveal any underlying population structure using a dense SNP panel and to identify selection signatures using ROH, characterizing the inbreeding level of this chicken population and disclosing the genes under positive selection for adaptive variants in an outbred population.

Materials and methods

Sampling and genotyping

In the present study, a total number of 213 chickens feathers were sampled in different rural production units located in 25 states of México (Aguascalientes, Baja California Sur, Campeche, Chiapas, Chihuahua, Coahuila, Colima, México City, Durango, Estado de México, Guanajuato, Guerrero, Hidalgo, Jalisco, Morelos, Nayarit, Nuevo León, Oaxaca, Querétaro, Tabasco, Tamaulipas, Tlaxcala, Veracruz, Yucatain and Zacatecas) by INIFAP (Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias). This samples collection is part of institucional Project "Identificación de los recursos genéticos pecuarios para su evaluación, conservación y utilización sustentable en México. Aves y cerdos. SIGI NUMBER 10551832012" coordinated with the activities of the Centro Nacional of Recursos Genéticos (**CNRG**) at Tepatitlán, Jalisco (México) engaged in promoting strategic research to solve the most important problems of productivity, competitiveness, equity and sustainability at the forest, agricultural and livestock sectors in México (http://www.inifap.gob.mx/SitePages/centros/cnrg.aspx). The samples are owned by the CNRG who control their access and reuse. Original owners of individuals have donated the samples to CNRG who gave consent for re-use for research purposes. The study did not require any ethical approval according to national rules, according to EU regulation, as it does not foresee sampling from alive animals. The University of Milan permit for the use of collected samples in existing bio-banks was released with n. OPBA-56-2016.

DNA extraction from feathers and genotyping were performed at GeneSeek (Lincoln, NE) using a commercial kit and the

Affymetrix Axiom® 600K Chicken Genotyping Array, containing 580,954 SNPs, distributed across the genome with an average spacing of 1.7 Kb, respectively. The galGal4 chicken assembly was used in this study as reference genome. Only markers positioned on chromosome 1 to 28 were used in this study.

A quality control of raw intensity files using the standard protocol in the Affymetrix Power Tools package (www.affimetrix.com) was performed in order to guarantee a high quality of genotyping data. Samples with Dish Quality Control (**DQC**: the closer the value is to one, the better the signal separates from the background) <0.82 and with quality control (QC) call rates <97% were excluded from downstream analysis. The quality verified samples were used for subsequent SNP analyses using dedicated software.

Morphological chicken characterization

Morphological characteristics of collected Mexican Creole individuals are extremely variable in terms of feathers colours, shapes (i.e. naked neck/breast or not, fighting characteristics), comb, and size. The measurement of morphological characteristics of birds was done according to the FAO Guidelines (2012), which is the recognised standard. Measures were taken at sampling and recorded in a database. The STANDARD procedure of SAS 9.4 (2013) was used to create a dataset with standardized values (mean = zero; standard deviation = 1) for the following four quantitative variables: i) body length - length between the tip of the rostrum maxillare (beak) and that of the cauda (tail, without feathers); the bird’s body should be completely drawn throughout its length; ii) wingspan: length in cm between tips of right and left wings after both are stretched out in full; iii) breast circumference:

circumference of the chest (taken at the tip of the pectus, hind breast); iv) length of the shank: (length in cm of the shank from the hock joint to the spur of either leg). Subsequently, a FASTCLUS procedure of SAS 9.4 (2013) was used to perform a disjoint cluster analysis on the basis of distances computed from one or more quantitative variables. The observations were divided into clusters such that every observation belongs to one and only one cluster. Four clusters were generated by the program with a Cubic Clustering Criterion (CCC) of 15.74. Following the FASTCLUS the CANDISC procedure was run on the four body measures as variables using the clusters previously created as classes. The CANDISC procedure performs a multivariate one-way analysis of variance and provides four multivariate tests under the hypothesis that the class means vectors are equal.

Genetic characterization

Different approaches and software were used in order to disclose the genetic structure of Mexican chickens:

a) using SVS Golden Helix 8.4 software (SVS) (Golden Helix Inc., Bozeman, MT, USA) the Identity by Descent (IBD) estimation and the Principal Components Analysis (PCA) were performed. The IBD is a measure of the relatedness of the pair of individuals and indicates how many alleles at any marker in each of two individuals came from the same ancestral chromosomes. The estimation of the IBD between all pairs of samples was done after the application of LD pruning option. Relationship-based pruning was performed and one member of each pair of animals with an observed genomic relatedness greater than 0.25 was removed from further analyses. The PCA, of pairwise individual genetic distances, was performed based on allele frequencies of pruned SNPs. To visualize the individual samples relatedness graphically in multi-dimensions the rgl R package (https://CRAN.R-project.org/package=rgl) was used.

b) The ADMIXTURE v. 1.3.0 software was employed to infer the most probable number of ancestral populations based on the SNP genotype data (Alexander et al., 2009). ADMIXTURE was run from K = 2 to K = 6, and the optimal number of clusters (K-value) was determined as the one having the lowest cross-validation error. Each inferred chicken population structure was visualized using R script suggested in the ADMIXTURE procedure.

c) Wright's statistics, including observed heterozygosity (H_O), expected heterozygosity (H_E), and inbreeding estimates (F_{IS}) were calculated with SVS.

d) Neighbor-Joining (NJ) tree, constructed based on the allele sharing distances (DAs) as the genetic distance between not-related individuals, was created and graphically represented using PEAS (Xu et al., 2010) and FigTree version 1.4.2 (http://tree.bio.ed.ac.uk/software/figtree) software, respectively.

e) The Arlequin v.3.5.2.2 software (Excoffier and Lisher, 2010) was used to perform an Analysis of MOlecular VAriance (AMOVA) a tool to check how the genetic diversity is distributed among individuals within groups, whose structure is quantified by F_{ST}.

f) ROH analysis was performed for each individual (complete SNP dataset = 471,730), using the SVS software. The ROH was defined by: i) a minimum of 1500 kb in size and 50 homozygous SNPs; ii) one heterozygous SNP was permitted in ROH, so that the length of the ROH was not disrupted by an occasional heterozygote; iii) five missing SNPs were allowed in the ROH; iv) maximum gap between SNPs of 100 Kb was predefined in order to assure that the SNP density did not affect the ROH. According to the nomenclature reported by other authors (Curik et al. 2014), ROH were grouped into five classes of length: <2Mb, 2-4Mb, 4-8Mb, 8-16Mb and >16Mb. Number, total length and the average of ROH length were calculated across individuals within chicken population. In addition, the percentage of the total genome length affected by ROH was also estimated.

Results

Morphological chicken characterization

The analysis of morphological measures (body length, wingspan, breast circumference, length of the shank) to cluster individuals separated the population into four different groups. While the clusters 1 (Cl_1), 3 (Cl_3), and 4 (Cl_4) were composed by 36, 74, and 72 animals, respectively, only one individual belonged to cluster 2. This latter individual, as such as cluster 2, were eliminated from subsequent analyses. In Figure 1 the scatter plot of the canonical variables one *vs.* two based on morphological measures is shown. The distinction among the three clusters was clearly displayed on the canonical variable plotted as x-axis representing 99% of the total variance.

The Table S1 shows that the individuals in the three clusters exhibit different sizes with CL_1 being in general the smaller individuals, CL_4 the intermediate, CL_3 the cluster with birds of larger dimension. This appears also from Figure 1 where CV_1 clearly differentiate the three clusters with CL_4, the green one, and intermediate respect the other two.

Genetic characterization

SNPs with Minor Allele Frequency (MAF) value ≤ 0.01, HWE value 0.00001, SNPs non on autosomal chromosomes from 1 to 28 and SNPs having a call rate <97% were excluded, reducing to 471,730 markers the number of SNPs used for the statistical analysis. SNPs passing the QC were pruned for LD using a threshold of $r^2 = 0.5$. LD trimming resulted in another 207,245 SNPs pruned from the dataset, ensuing in a final set of 264,485 SNPs used in the downstream analysis. Of the 213 animals sampled, 31 showed an IBD value greater than 0.25 with at least one other individual of the population, and then were subsequently removed leaving 182 animals for the population structure analyses. The remaining population is thus holding individuals with IBD less than 0.25 IBD value as maximum value. Out of the 16,471 IBD values only 337 were comprised in the interval 0.125 ≤ IBD < 0.25, 561 in the interval 0.0625 ≤ IBD < 0.125, 646 between 0.0625 ≤ IBD < 0.03125, while the remaining all less than 0.03125. According to this distribution we considered all individuals unrelated. The Heat map created using the IBD estimates values is showed in Figure 1S in Additional File 1.

The program ADMIXTURE was run for K values from one to six (Figure 2A). The lowest cross validation error was found at K = 3, that represent the number of ancestors in the Mexican populations (Figure 2B). A number of K greater than three does not produce a larger number of ancestor's contribution in the living population, as it is visible in Figure 2A.

The Figure 2C is a graphical representation of the 182 individuals grouped according to the proportion of the three ancestors' contribution. One individual result to derive entirely from ancestor 1, while seven derives entirely from ancestor 2 and 11 from ancestor 3. A total of 25 individuals showed to derive from two ancestors while the largest proportion of the sample, 138 individuals, showed a genetic composition that derived from all the three identified ancestors. Apparently, there is no clear relationship between the morphological clustering and the ancestor's composition. Table S2 shows the bird count according to the ancestor's composition classes and the morphological clusterization. The largest part of individuals pertaining to ancestor 1 class (i.e. 57%) showed morphological characteristics of birds classified as cluster 3, while individuals pertaining to ancestor 2 (i.e. 50%) and 3 (i.e. 49%) showed characteristics of animals classified as cluster 4.

The results of the NJ analysis depicted in Figure 2D are consistent with those obtained by the PCA.

It is possible to note that the major part of samples is grouped according to the ancestor's composition, but individual differences based on DAs did not allow a clear division of birds in well separated clusters (Figure 2D).

The results of the PCA agreed well with the findings outlined above, as showed on Figure 3A e 3B. In both PCA analyses of Figure 3 there is no clustering of individuals neither for morphological cluster than for the ancestor classification, as points are mixed in all distributions depicted.

The Table 1 reports the results for the AMOVA analysis. The analysis account of individual classification was based on morphological clustering (Cl_1; Cl_3 and Cl_4). We considered three hypotheses: Hypothesis 1) Cl_1 + Cl_3 *vs* Cl_4; Hypothesis 2) Cl_1 + Cl_4 *vs* Cl_3; Hypothesis 3) Cl_3 + Cl_4 *vs* Cl_1. All the hypotheses indicate that the most part of variability is observed within clusters, 99.72 % (Hypothesis 1), 99.56% (Hypothesis 2) and 99.49% (Hypothesis 3), with a much smaller amount of the variance component occurring among groups 0% (Hypothesis 1), 0.18% (Hypothesis 2) and 0.22% (Hypothesis 3) (Table 1). The AMOVA confirmed the results obtained with the PCA. In other words, the genetic variation of the Mexican population appears to be mostly related to the individual genetic variability rather than to the genetic diversity expressed by the clustering classification obtained on the basis of the morphological characteristics.

Run of homozygosity (ROH) analysis

The SVS software identified a total of 3,059 runs across Mexican chicken population. (Supplementary Table S3). Six individuals did not show any ROH in any of the 28 chromosomes. Likewise, the chromosomes 16 and 25 showed no evidence of ROH in all genotyped individuals. Results revealed that there were marked differences in terms of number and length of ROH across individuals.

The ROH have been defined with 305 and 6,629 SNPs as minimum and maximum number of SNPs. The average number of SNPs falling into a ROH was consistent among ROH length category, ranging from 824 (ROH <2Mb) to 3,977 (ROH >8-16 Mb) SNPs.

The identified ROH are mainly short in length; in fact, the ROH of 2-4 Mb and <2 Mb are the most frequent classes of length identified (i.e. 84%). Instead, no ROH were found within the >16Mb length class (Table 2).

The number of ROH per individual ranged from one to 115, with a mean number of ROH for sample of 17.38 (Figure 3). The Figure 3 also shows the relationship between number and averaged total length of ROH for each individual (mainly ranged from 1.7 to 2.8 Mb). Only two samples showed a very high number of ROH (i.e 110 and 115 ROH). The average size of ROH of these two individuals is nevertheless similar to other subject. ROH larger than three Mb were found in 38 individuals representing 21% of the total sample and showing a count range of ROH from one to 65. The amount of the genome covered by ROH per individual ranged (as mean values) from 1,563,036 bp to 4,387,646 bp (Figure 4).

The relative frequencies of ROH (calculated as number of ROH per class on total number of ROH) within each chromosome and by length classes (Table 2) were also calculated. The ROH of 8-16 Mb size were found in longer chromosomes, the total number of ROH appeared to be proportional to their lengths and were distribution appeared homogeneous across them.

The genomic regions most commonly associated with ROH have been identified by selecting the top 1% of the SNPs most frequently observed in the ROH (Top 1% ROH). Figure 5 shows the incidence of ROH segments across the genome and as appear, the genomic distribution of ROH segments was clearly non-uniform across chromosomes. A total of 11 regions were identified with frequencies of ROH segments exceeding 1% of the whole population (Top 1% ROH) in the first eight chromosomes, excluding chr 6.

After downloading the list of chicken autosome galGal4 genes (GCA_000002315.2) from Ensembl database (http://www.ensembl.org), the annotation of gene mapping within the Top 1% ROH is reported in Table 3.

Discussion

Patterns of high-density SNPs variation were used in this study to detect genetic variability in a chicken population collected in several states of México. All findings provided in this research using several statistic approaches, confirmed and highlighted a not structural classification of individuals in well differentiate subpopulations, even if ADMIXTURE statistic identified three possible ancestors to define the predominant genetic background in our population.

The effective number of polymorphic SNPs (considered as the number of SNP in which at least one heterozygous individual was identified) represents the 99.9% of the total loci. The moderately high values of H_O (0.319) and H_E (0.348) reflect the high percentage of polymorphic SNP; the low F_{is} value (0.084) are indicative of a low level of inbreeding in the population and of the relatively high number of birds in heterozygous state. In other native populations where the heterozygosity and F_{is} was recently calculated using a high-density SNP chip (Strillacci et al., 2017), the H_O varies from 0.21 to 0.34, the H_E from 0.17 to 0.32 and the F_{is} from -0.19 to 0.094. These populations nevertheless are very well characterized in different breeds, thus showing more homogeneity within the same group of individuals. Using a 60K SNP chip (Johansson and Nelson, 2015) found an F_{is} value of -0.09 and 0.17 in two local chicken populations indicating that farmers do not increase inbreeding excessively. Our results thus show that in outbreed Creole population as the Mexican one, the genetic variability appear larger respect to local populations defined in breeds.

As expected, the results of AMOVA showed that most of the genetic variation occurred within populations in all the three hypotheses here considered and confirm the absence of a genetic structure in the Mexican chicken population. The slightly negative value for the variance in fact, as obtained in hypothesis

1 (i.e. -0.47), can occur in absence of genetic structure, and is a quite common occurrence in AMOVA, as the real parameter value has to be considered zero. The negative or slightly positive values of among groups variance and ΦCT for all hypothesis (Excoffier, 2007), thus confirm the absence of a hierarchical genetic structure in the Mexican poultry population. These findings also confirm the results from Gorla et al. (2017) who, using a different approach and a different class of genetic markers, did not disclosed a genomic structure in the Mexican chicken population. We did not consider the analysis by ancestor as the classes are extremely numerous and unbalanced among them (see Table 1).

It is to be recalled that the Mexican poultry population is a Creole unique genetic pool that have not been selected for target traits for more than 500 years. As consequences of its adaptation to the environmental conditions and production, some genomic region may be fixed in individuals as a result of positive selection.

These results here obtained for ROH are in concordance with those identified in previous studies (Gibson et al., 2006) where short ROH with high frequencies were identified in outbred individuals, as well as the intermediate sizes runs. ROH greater than 10 Mb, generally identified in individuals belonging to populations with high levels of background relatedness, have been also identified in 2%–26% of individuals pertaining to outbred populations (Pemberton et al., 2012), and in a proportion of 14% in our birds. These findings may reflect a recent parental relatedness or be the result of a recombination lack that allows uncommonly long ancestral genomic segments to persist in the population (Pemberton et al., 2012). Additionally, findings from the ROH analysis indicated that natural selection affected allele frequencies in specific regions of the Mexican chicken genome (Figure 5).

Among the annotated genes in the ROH regions, in fact, some are worth mentioning because their functions could play important roles in the historical genetic dynamic occurred to the Mexican chicken population.

On chr1 within the ROH_1 (at 41.38-43.21 Mb) lies the KITLG (KIT ligant) gene that has a role in controlling the migration, survival and proliferation of melanocytes; also, rare mutations in the mouse homolog of the KITLG gene are known to affect coat colour (Sulem et al., 2007). Additionally, Metzger et al. (2015) highlighted the role of this gene in the horse reproduction efficiency, claiming its general effect in all livestock populations. The AICDA (activation-induced cytidine deaminase (AID)) gene mapping within the ROH_2, encodes for a DNA editing protein that plays an essential role in some events of immunoglobulin (Ig) diversification: somatic hypermutation, class switch recombination and Ig gene conversion (Carãtao et al., 2013). These processes generate the vast diversity of antibodies required to challenge a nearly infinite number of antigens that immune systems encounter (Keim et al., 2013). In the same ROH_2 the VWF (von Willebrand factor) gene and the FGB (Fibrinogen beta chain), the FGG (Fibrinogen gamma chain) and the FGA (Fibrinogen alpha chain) genes located within ROH_8, are four of the eight hemostatic genes resulted down regulated in studies based on RNA-Seq analysis on breast muscle of chickens affected by “Wooden Breast disease” (Mutryn et al., 2015). The ROH_9 on chr5 (2.60-3.95 Mb) harbours the BDNF (brain derived neurotrophic factor) gene, which is considered important for the heat stress response in chicken (Lamont et al., 2014). Furthermore, previous findings indicating that the BDNF gene prevents the death of cultured chick retinal ganglion cells, and as reported by Herzog et al. (1994) the tightly controlled expression of the BDNF gene might be important in the coordinated development of the visual system in chicks.

The same ROH_9 includes the LGR4 (leucine rich repeat containing G protein-coupled receptor 4) gene that in human is associated with low bone mineral density (Styrkarsdottir et al., 2013).

Within the ROH_8 and ROH_10 map genes that are closely linked to immune system (Table 3). More precisely, within the first region map two duplicated genes, the TLR2A (toll-like receptor 2 family member A) and the TLR2B (toll-like receptor 2 family member B), both orthologs of the single TLR2 of mammals. These genes mediate innate immune responses via recognition of pathogen-associated molecular patterns (PAMPs) such as dsRNA of some viruses, or lipopolysaccharide of Gram-negative bacteria (Downing et al., 2010). Miyagi et al. (2007) demonstrated that regulation of basal levels of particular STATs including STAT1 and STAT4 and their receptor association, contributes to innate production of the IFN-γ of NK cells. Also, the STAT4 gene encodes a transcription factor involved in the signalling pathways of several cytokines, including interleukin-12 and interleukin-23 (IL-12) (Martinez et al. 2008).

A recent work by Fleming et al. (2017) has mapped ROH in several indigenous African and European populations. The authors do not report the list of genomic position of the 4167 consensus ROH mapped in their populations, so it is not possible to compare the overlapping with our results. Nevertheless, the number of ROH mapped is comparable with the one found in this study. Finally, among the three ROH that Fleming et al. (2017) are reporting in detail, no one is overlapping those found in this study.

Gene ontology (GO) and pathway analyses for genes included into the Top ROH (Supplementary Table S4) were performed using GenCLiP2.0, an online server for functional clustering of genes (http://ci.smu.edu.cn/GenCLiP2.0/analysis.php?random=new) accounting for false discovery rate. The GO analysis revealed that they are clustered into a 10 group of genes that were involved in a variety of cellular functions such as sex differentiation, reproductive system development, regulation of response to stress, programmed cell death, tissue and organ development, and so on. KEGG Pathway analysis showed the involvement of several signal pathways, but only five were significant after FDR correction (in Supplementary Table S5, as the Q-values).

The Literature Mining Gene Network tool (provided by GenCLiP2.0), that searches for genes linked to keywords based on up-to-date literature profiling, revealed that the twenty-two genes included within Top 1% ROH have been associated mainly with the keywords ‘‘stress’’, “muscle”, “immune response” and “reproduction”, as reported in Figure 6. Edges in Network correspond to literature that associate two genes with each other, while the relative edge-labels indicate the number of related articles. To further examine the Top 1% ROH content, quantitative trait loci (QTL) that overlapped with these genomic regions were identified by downloading the QTL list from the animal QTL database (http://www.animalgenome.org/cgi-bin/QTLdb/GG/index). We filtered out the QTL that are larger than 5 Mb and only QTL overlapping for at least 50% with the ROH were considered. As reported in Table 3, the most represented QTL are those associated to body conformation and structure (i.e. breast muscle percentage and weight, tibia bone mineral density, body weight, abdominal fat weight and, muscle fibre density and diameter). The same holds for QTLs with a size comprised between 5 and 10 Mb.

The study indicates that the Mexican population is well adapted to the diverse farming conditions that can be found in the United States of México. The population clearly appear to be a unique Creole chicken population that was not subjected to a specific breeding strategy to improve performance but shows selection sweeps due to the occurring natural selection for more than 500 years. As the population was maintained mainly as a backyard population, possibly the farmers have reproduced the more productive, more fertile, and more resistant individuals regardless to plumage colour or morphological characteristics. The adaptation of the population to environmental conditions, its resilience to various challenges, makes it very interesting as native genetic resource to be used in family non-intensive farming, in order to raise their income.

In some states of México where with a very important poultry and swine intensive farming, there is no specific financial support for poultry family farming by Mexican program "Sin hambre" (no hungry - http://sinhambre.gob.mx/). In these states the goal of is to improve sanitary conditions in intensive farming to favour the exportation to USA and Europe. Nevertheless, the "Sin hambre" project helps greatly local family to farm chicken and increase their revenue, providing a commercial channel for the egg production. This can be easily supported with the local Mexican chicken population adapted to local environmental conditions. A strategy sometime used at present is to cross the local Creole population with highly productive breeds as the, e.g. Rhode Island or to provide farmers directly with F1 hybrids. This practice nevertheless requires a very careful management of the local well adapted population to avoid the loss of the genetic variability that guarantee the resilience of the individuals in very harsh environments.

The study provide a genetic knowledge that can be used as a basis for the genetic management of a unique and very large Creole population, especially in the view of using it in production of hybrids to increase the productivity and economic revenue of family farming agriculture, which is a large reality of United States of México.

Tables

Table 1. Hierarchical AMOVA analysis among the clusters obtained based on allele frequencies of pruned SNPs.

Hypotheses	Variance component (%)			Fixation indexes[a]					
	Among groups	Among clusters within groups	Within clusters	ΦCT	P-value[b]	ΦSC	P-value[b]	ΦST	P-value[b]
Cl_1+Cl_3 *vs* Cl_4	-0.47[c]	0.75	99.72	-0.005	1.000±0.000	0.007	0.000±0.000***	0.003	0.000±0.000***
Cl_1+Cl_4 *vs* Cl_3	0.18	0.26	99.56	0.002	0.658±0.011	0.003	0.043±0.001*	0.004	0.001±0.001*
Cl_3+Cl_4 *vs* Cl_1	0.22	0.29	99.49	0.002	0.332±0.016	0.003	0.004±0.002*	0.005	0.000±0.000***

[a] ΦCT = variation among groups divided by total variation, ΦSC = variation among sub-groups divided by the sum of variation among sub-groups within groups and variation within sub-groups, ΦST = the sum of variation groups divided by total variation.

[b] ns = P > 0.05, * = P < 0.05, *** = P < 0.001.

[c] Negative values are presented, but we can consider this value effectively equal to zero.

Table 2. Numbers of ROH per chromosome according to ROH classes of length.

Classes of ROH						
Chr	<2 Mb (*)	2-4 Mb (*)	4-8 Mb (*)	8-16 Mb (*)	>16 Mb (*)	Total
1	247 (0.34)	350 (0.48)	123 (0.17)	11 (0.02)	0 (0)	731
2	190 (0.35)	255 (0.47)	83 (0.15)	11 (0.02)	0 (0)	539
3	147 (0.38)	182 (0.47)	56 (0.14)	5 (0.01)	0 (0)	390
4	110 (0.34)	162 (0.5)	45 (0.14)	8 (0.02)	0 (0)	325
5	97 (0.39)	104 (0.42)	46 (0.18)	3 (0.01)	0 (0)	250
6	51 (0.42)	46 (0.38)	23 (0.19)	1 (0.01)	0 (0)	121
7	41 (0.32)	71 (0.56)	14 (0.11)	1 (0.01)	0 (0)	127
8	39 (0.41)	50 (0.53)	6 (0.06)	0 (0)	0 (0)	95
9	23 (0.34)	42 (0.62)	3 (0.04)	0 (0)	0 (0)	68
10	27 (0.37)	36 (0.49)	10 (0.14)	0 (0)	0 (0)	73
11	27 (0.44)	30 (0.48)	5 (0.08)	0 (0)	0 (0)	62
12	20 (0.48)	20 (0.48)	2 (0.05)	0 (0)	0 (0)	42
13	16 (0.43)	19 (0.51)	2 (0.05)	0 (0)	0 (0)	37
14	15 (0.52)	14 (0.48)	0 (0)	0 (0)	0 (0)	29
15	6 (0.22)	18 (0.67)	3 (0.11)	0 (0)	0 (0)	27
16	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0
17	10 (0.36)	18 (0.64)	0 (0)	0 (0)	0 (0)	28
18	9 (0.53)	8 (0.47)	0 (0)	0 (0)	0 (0)	17
19	12 (0.52)	10 (0.43)	1 (0.04)	0 (0)	0 (0)	23
20	21 (0.49)	19 (0.44)	3 (0.07)	0 (0)	0 (0)	43
21	3 (0.38)	5 (0.63)	0 (0)	0 (0)	0 (0)	8
22	1 (0.33)	2 (0.67)	0 (0)	0 (0)	0 (0)	3
23	3 (0.6)	2 (0.4)	0 (0)	0 (0)	0 (0)	5
24	6 (0.67)	3 (0.33)	0 (0)	0 (0)	0 (0)	9
25	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0
26	0 (0)	1 (1)	0 (0)	0 (0)	0 (0)	1
27	1 (0.25)	3 (0.75)	0 (0)	0 (0)	0 (0)	4
28	2 (1)	0 (0)	0 (0)	0 (0)	0 (0)	2
Total	1,124 (0.36)*	1,470 (0.48)*	425 (0.14)*	40 (0.02)*	0 (0)*	3,059

(*) Proportion calculated as number of ROH per class over the total number of ROH

Table 3. The eleven Top 1% ROH identified on Mexican chicken autosomes by SVS.

ROH_ID	Chr	Start	End	Length	Genes*	QTL (http://www.animalgenome.org/cgi-bin/QTL_IDdb/GG/index)
ROH_1	1	41,387,392	43,210,859	1,823,467	***NTS, KITLG, DUSP6***	Femur bending strength QTL_ID (6758); Yolk weight QTL_ID (24938, 24939, 24940); Breast muscle percentage QTL_ID (95427); Growth (post-challenge) QTL_ID (65829)
ROH_2	1	73,476,359	75,663,705	2,187,346	*CCND2, NDUFA9,* ***NTF3***, ***VWF***, *TEAD4, RHNO1,* ***FKBP4***, *FOXM1, NANOG,* ***AICDA***, *PHC1*	
ROH_3	1	146,817,860	147,817,564	999,704		
ROH_4	2	51,510,051	54,136,958	2,626,907	*PSMA2, STK17A,* ***EGFR***, *SEC61G*	Body weight (70 days) QTL_ID (12390); Body weight (56 days) QTL_ID (12391); Body weight (70 days) QTL_ID (12392)
ROH_5	2	70,951,155	71,838,862	887,707	*ENS-3, mir6545*	Body weight (42 days) QTL_ID (6899); Abdominal fat weight QTL_ID (6900); Breast muscle weight QTL_ID (6968)
ROH_6	2	86,255,347	87,498,657	1,243,310	*IRX1, IRX2*	Egg shell color QTL_ID (1914); Body weight (42 days) QTL_ID (6901); Breast muscle percentage QTL_ID (12569)
ROH_7	3	68,723,915	70,012,473	1,288,558		
ROH_8	4	18,018,373	20,496,038	2,477,665	*IDS, TLR2A, TLR2B, TRIM2, MND1, SFRP2,* ***FGB, FGA, FGG,*** *NPY2R, CTSO, mir7469*	Abdominal fat weight QTL_ID (19531, 19535, 19538); Body weight (40 days) QTL_ID (6659); Egg shell color QTL_ID (3348); Muscle fiber density QTL_ID (19534, 19537); Muscle fiber diameter QTL_ID (19533,19536,19340); Residual feed intake QTL_ID (7057); Subcutaneous fat thickness QTL_ID (19532, 19539); Yolk weight QTL_ID (3349)

ROH_9	5	2,126,161	4,221,327	2,095,166	*PRMT3, ANO5, SLC17A6, GAS2, SVIP, ANO3, FIBIN, LIN7C,* ***BDNF,*** *KIF18A, METTL15, BBOX1, LGR4,* ***SLC5A1,*** *mir1775, mir1760*	Body weight (28 days) QTL_ID (95415, 195416)
ROH_10	7	6,661,093	8,199,140	1,538,047	*COL18A1, SLC19A1,* ***COL6A1, COL6A2****, FTCD, LSS****, S100B****, ITGB2, ADARB1,* ***GLS****,* ***STAT1, STAT4***	Shank weight QTL_ID (9161); Breast muscle weight QTL_ID (6982)
ROH_11	8	9,141,018	11,122,757	1,981,739	***PLA2G4A, PTGS2,*** *C8H1ORF27, AMY1AP, AMY1A, SLC30A7,* ***CRK,*** *CDC14A, DBT, SASS6, MFSD14A, SLC35A3, HOXA3, PALMD, mir6561, mir1610*	Thigh meat-to-bone ratio QTL_ID (6721); Abdominal fat percentage QTL_ID (2183); Body weight (day of first egg) QTL_ID (14465); Tibia bone mineral density QTL_ID (24365)

*Genes in bold are those included in Networks

Figures

Figure 1. PCA based on morphological features (Body length, Wingspan, Breast circumference, Length of the shank): Cl_1: blue, Cl_3: red, and Cl_4: green). Canonical variable 1: CV_1; Canonical variable 2: CV_2

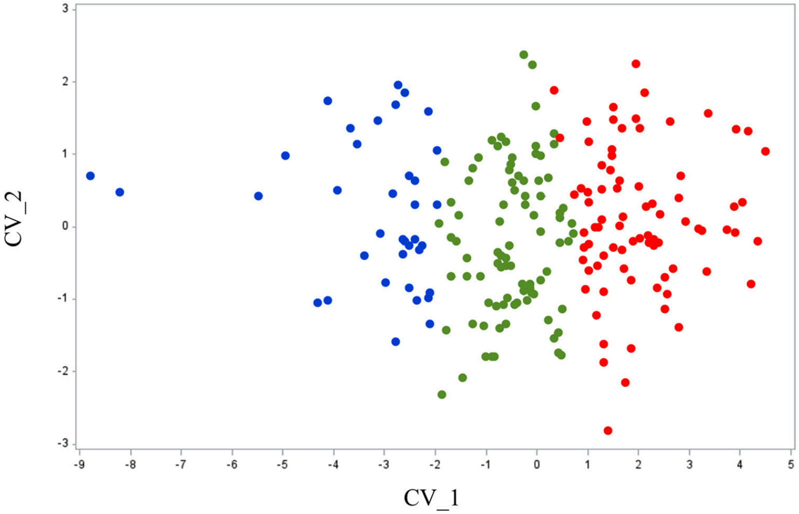


Figure 2. Graphical representation of Mexican chicken population genetic structure. A) ADMIXTURE k=2-K=6 barplots; B) Optimal number of clusters according to cross-validation error; C) Count of individuals based on 3 ancestors' composition; D) NJ tree: classification of individuals according to allele sharing distances. Individuals were labelled according to the morphological cluster (i.e. from 1 to 4) they belong and their individual (e.g. CL1012 = morphological cluster 1, individual 012) and the ancestors' composition from ADMIXTURE.

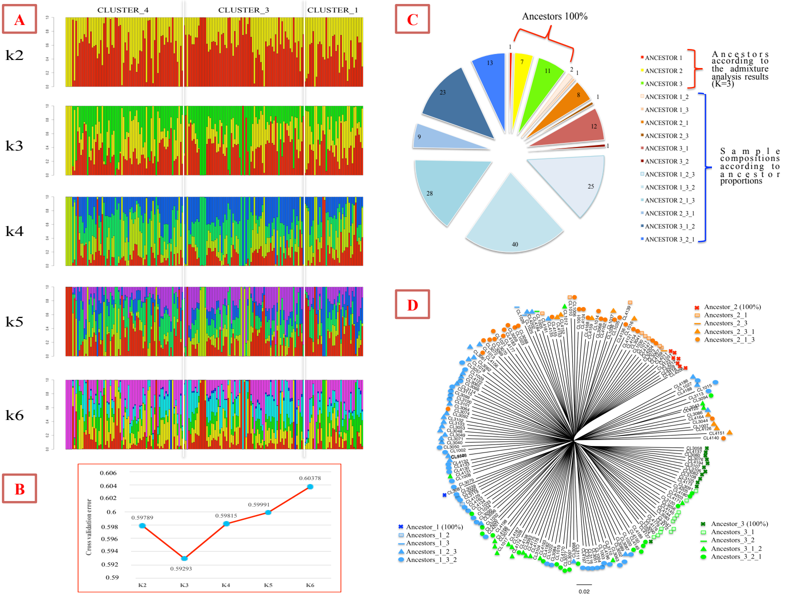
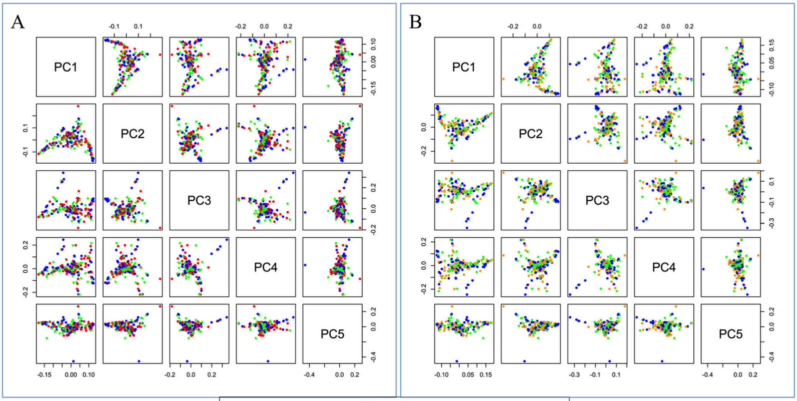


Figure 3. A) PCA based on allele frequencies of SNPs (individuals were coloured according to the three morphological clusters: Cl_1: blue, Cl_3: red, and Cl_4: green); B) PCA based on allele frequencies of SNPs (individuals were coloured according to the individual ancestor's composition: ancestor_1: blue, ancestor_2: orange, and ancestor_3:green).



Eigenvalues				
PC1 = 4.349	PC2= 1.612	PC3 = 1.273	PC4 = 1.062	PC5 = 1.025

Figure 4. Relationship between number and averaged length of ROH in each individual.

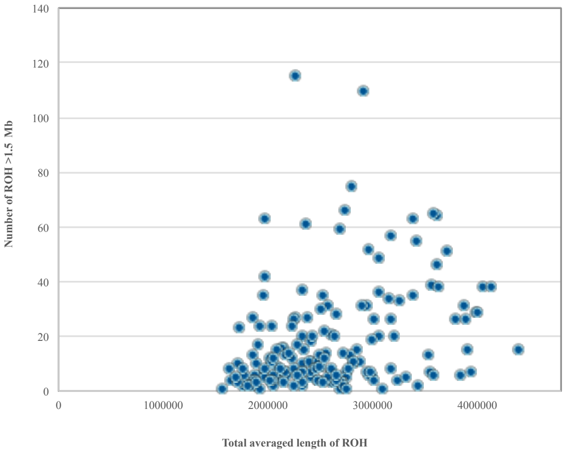


Figure 5. SNPs incidence in ROH identified by SVS. Red line indicates the adopted threshold: Top 1% of the observations.

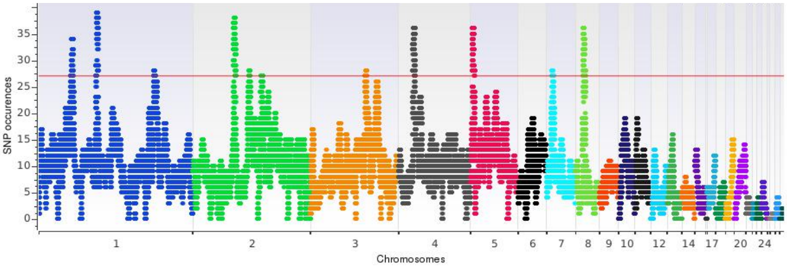
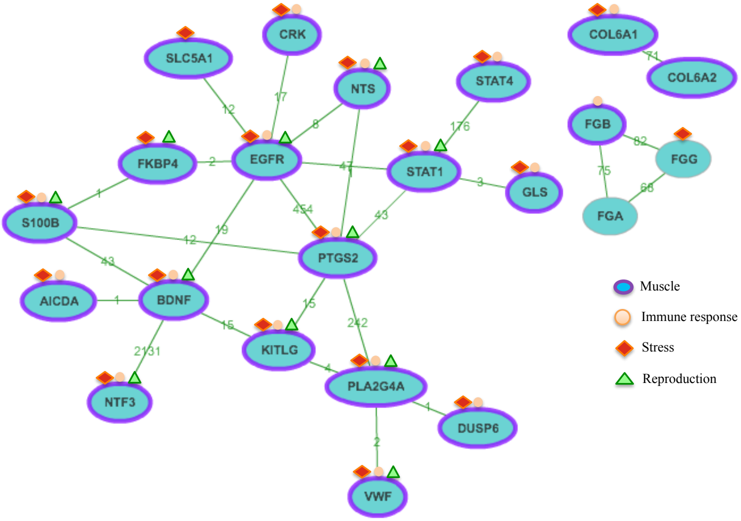


Figure 6. Network of genes included in the Top 1% Mexican chicken ROH.



Supporting information

All supplementary files are available at:

https://doi.org/10.3382/ps/pex374/4767756

Reference

- Alexander, D. H., J. Novembre, and K. Lange. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19:1655-1664.
- Al-Qamashoui, B., H. Simianer, I. Kadim, and S. Weigend. 2014. Assessment of genetic diversity and conservation priority of Omani local chickens using microsatellite markers. Trop. Anim. Health Prod. 46:747-752.
- Brown, W.R.A., S. J. Hubbard, C. Tickle, and S. A. Wilson. 2003. The chicken as a model for large-scale analysis of vertebrate gene function. Nat. Rev. Genet. 4:87–98.
- Caratão, N., C. S. Cortesão, P. H. Reis, R. F. Freitas, C. M. Jacob, A. C. Pastorino, M. Carneiro-Sampaio, and V. M. Barreto. 2013. A novel activation-induced cytidine deaminase (AID) mutation in Brazilian patients with hyper-IgM type 2 syndrome. Clin. Immunol. 148:279-286.
- Cavalchini, L. G., S. P. Marelli, M. G. Strillacci, M.C. Cozzi, M. Polli, and M. Longeri. 2007. Heterozygosity analysis of Bionda Piemontese and Bianca di Saluzzo chicken breeds by microsatellites markers: a preliminary study. J. Anim. Sci. 6:63-5.
- Ceccobelli, S., P. Di Lorenzo, H. Lancioni, L. V. Monteagudo Ibáñez, M. Tejedor, C. Castellini, V. Landi, A. Martínez Martínez, J. D. Delgado Bermejo, J. L. Vega Pla, J. M. Leon Jurado, M. García, G. Attard, A. Grimal, S. Stojanovic, K. Kume, F. Panella, S. Weigend, and E. Lasagna. 2015. Genetic diversity and phylogeographic structure of sixteen Mediterranean chicken breeds assessed with microsatellites and mitochondrial DNA. Livest. Sci. 175:27–36.
- Curik, I., M. Ferenčaković, and J. Sölkner. 2014. Inbreeding and runs of homozygosity: A possible solution to an old problem.

Livest. Sci. 166:26-34.

- Downing, T., A. T. Lloyd, C. O'Farrelly, and D. G. Bradley. 2010. The differential evolutionary dynamics of avian cytokine and TLR gene classes. J. Immunol. 184:6993-7000.
- Excoffier, L. 2007. Analysis of Population Subdivision. In: Balding, David J.; Bishop, Martin; Cannings, Chris (eds.) Handbook of Statistical Genetics 980-1020. Chichester: John Wiley & Sons.
- Excoffier, L., and H. E. L. Lischer. 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol. Ecol. Resour. 10:564–567.
- Fan. H., Y. Wu, X. Qi, J. Zhang, J. Li, X. Gao, L. Zhang, J. Li, and H. Gao H. 2014. Genome-wide detection of selective signatures in Simmental cattle. J. Appl. Genet. 55:343-351.
- FAO, 2012. Phenotypic characterization of animal genetic resources. FAO Animal Production and Health Guidelines No. 11. Rome (http://www.fao.org/docrep/015/i2686e/i2686e00.pdf) (accessed 01.02.2017).
- Fleming, D. S., J. E. Koltes, A. D. Markey, C. J Schmidt, C. Ashwell, M. F. Rothschild, M. E. Persia, J. M. Reecy, and S. J. Lamont. 2016. Genomic analysis of Ugandan and Rwandan chicken ecotypes using a 600 k genotyping array. BMC Genomics. 17:407.
- Fleming D. S., S. Weigend, H. Simianer, A. Weigend A, M. F. Rothschild, C. J. Schmidt, C. Ashwell C, M. E. Persia, J. M. Reecy, S. J. Lamont. 2017. Genomic Comparison of Indigenous African and Northern European Chickens Reveals Putative Mechanisms of Stress Tolerance Related to Environmental Selection Pressure. G3 (Bethesda). 7(5):1525-1537.

- Gibson, J., N. Morton, and A. Collins. 2006. Extended tracts of homozygosity in outbred human populations. Hum. Mol. Genet. 15:789–795.
- Gorla, E., M. C. Cozzi, S. I. Roman-Ponce, F. J. Ruiz-Lopez, V. E. Vega-Murillo, S. Cerolini, A. Bagnato, and M. G. Strillacci. 2017. Genomic variability in Mexican chicken population using Copy Number Variants. BMC Genetics 18:61.
- Hall, S.J., and D. G. Bradley. 1995. Conserving livestock breed biodiversity. Trends. Ecol. Evol. 10: 267–270
- Herzog, K. H., K. Bailey, and Y. A. Barde. 1994. Expression of the BDNF gene in the developing visual system of the chick. Development 120:1643-1649.
- Johansson, A. M., and R. M. Nelson. 2015. Characterization of genetic diversity and gene mapping in two Swedish local chicken breeds. Front. Genet. 6:44.
- Keim, C., D. Kazadi, G. Rothschild, and U. Basu. 2013. Regulation of AID, the B-cell genome mutator. Genes Dev. 27:1-17.
- Lamont, S. J., D. J. Coble, A. Bjorkquist, M. F. Rothschild, M. Persia, C. Ashwell, and C. Schmidt. 2014. Genomics of heat stress in chickens. Proceedings, 10th World Congress of Genetics Applied to Livestock Production. Vancouver, BC, Canada, August 17-22, 2014.
- Mahammi, F. Z., S. B. Gaouar, D. Laloë, R. Faugeras, N. Tabet-Aoul, X. Rognon, M. Tixier-Boichard, and N. Saidi-Mehtar. 2016. A molecular analysis of the patterns of genetic diversity in local chickens from western Algeria in comparison with commercial lines and wild jungle fowls. Journal of Animal Breeding and Genetics. J. Anim. Breed. Genet. 133:59-70.
- Martínez, A., J. Varadé, A. Márquez, M. C. Cénit, L. Espino, N. Perdigones, J. L. Santiago, M. Fernández-Arquero, H. de la

Calle, R. Arroyo, J. L. Mendoza, B. Fernández-Gutiérrez, E. G. de la Concha, E. Urcelay. 2008. Association of the STAT4 gene with increased susceptibility for some immune-mediated diseases. Arthritis Rheum. 58:2598-2602.

- Mastrangelo, S., M. Tolone, R. Di Gerlando, L. Fontanesi, M. T. Sardina, and B. Portolano. 2016. Genomic inbreeding estimation in small populations: evaluation of runs of homozygosity in three local dairy cattle breeds. Animal. 10:746-754.
- Metzger, J., M. Karwath, R. Tonda, S. Beltran, L. Águeda, M. Gut, I. G. Gut, and O. Distl. 2015. Runs of homozygosity reveal signatures of positive selection for reproduction traits in breed and non-breed horses. BMC Genomics 16:764.
- Miyagi, T., M. P. Gil, X. Wang, J. Louten, W. M. Chu, and C. A. Biron. 2007. High basal STAT4 balanced by STAT1 induction to control type 1 interferon effects in natural killer cells. J. Exp. Med. 204:2383–2396.
- Mutryn, M. F., E. M. Brannick, W. Fu, W. R. Lee, and B. Abasht. 2015. Characterization of a novel chicken muscle disorder through differential gene expression and pathway analysis using RNA-sequencing. BMC Genomics 16:399.
- Nimmakayala, P., A. Levi, L. Abburi, V. L. Abburi, Y. R. Tomason, T. Saminathan, V. G. Vajja, S. Malkaram, R. Reddy, T. C. Wehner, S. E. Mitchell, U. K. Reddy. 2014. Single nucleotide polymorphisms generated by genotyping by sequencing to characterize genome-wide diversity, linkage disequilibrium, and selective sweeps in cultivated watermelon. BMC Genomics. 15:767.
- Pemberton, T. J., D. Absher, M. W. Feldman, R. M. Myers, N. A. Rosenberg, and J. Z. Li. 2012. Genomic patterns of homozygosity in worldwide human populations. Am. J. Hum.

Genet. 91:275-292.

- Qanbari, S., and H. Simianer. 2014. Mapping signatures of positive selection in the genome of livestock. Livest. Sci. 166:133–143.
- Segura-Correa, J. C., L. Sarmiento-Franco, J. G. Magaña-Monforte, and R. Santos-Ricalde. 2004. Productive performance of Creole chickens and their crosses raised under semi- intensive management conditions in Yucatan, Mexico, Br. Poult. Sci. 45:342- 345.
- Segura-Correa, J. C., A. Juarez-Caratachea, L. Sarmiento-Franco, and R. Santos-Ricalde. 2005. Growth of Creole chickens raised under tropical conditions of Mexico. Trop. Anim. Health. Prod. 37:327-332.
- SAS Institute. 2013. SAS 9.4 language reference concepts. Cary, NC: SAS Institute.
- Styrkarsdottir, U., G. Thorleifsson, P. Sulem, D. F. Gudbjartsson, A. Sigurdsson, A. Jonasdottir, A. Oddsson, A. Helgason, O. T. Magnusson, G. Bragi Walters, M. L. Frigge, H. T. Helgadottir, H. Johannsdottir, K. Bergsteinsdottir, M. H. Ogmundsdottir, J. R. Center, T. V. Nguyen, J. A. Eisman, C. Christiansen, E. Steingrimsson, J. G. Jonasson, L. Tryggvadottir, G. I. Eyjolfsson, A. Theodors, T. Jonsson, T. Ingvarsson, I. Olafsson, T. Rafnar, A. Kong, G. Sigurdsson, G. Masson, U. Thorsteinsdottir, and K. Stefansson. 2013. Nonsense mutation in the LGR4 gene is associated with several human diseases and other traits. Nature. 497:517-520.
- Strillacci, M. G., S. P. Marelli, M. C. Cozzi, E. Colombo, M. Polli, M. Gualtieri, A. Cristalli, P. Pignattelli, M. Longeri, and L. Guidobono Cavalchini. 2009. Italian autochthonous chicken breeds conservation: evaluation of biodiversity in Valdarnese

Bianca breed (Gallus gallus domesticus). Avian. Biol. Res. 2:229-233.

- Strillacci, M. G., M. C. Cozzi, F. Schiavini, E. Gorla, S. Cerolini, S. I. Román-Ponce, F. J. Ruiz López, and A. Bagnato. 2017. Genomic and genetic variability of six Italian chicken populations using SNP and CNV as markers. Animal. 11:737-745.
- Sulem, P., D. F. Gudbjartsson, S. N. Stacey, A. Helgason, T. Rafnar, K. P. Magnusson, A. Manolescu, A. Karason, A. Palsson, G. Thorleifsson, M. Jakobsdottir, S. Steinberg, S. Pálsson, F. Jonasson, B. Sigurgeirsson, K. Thorisdottir, R. Ragnarsson, K. R. Benediktsdottir, K. K. Aben, L. A. Kiemeney, J. H. Olafsson, J. Gulcher, A. Kong, U. Thorsteinsdottir, and K. Stefansson. 2007. Genetic determinants of hair, eye and skin pigmentation in Europeans. Nature Genet. 39:1443-1452.
- Xu, S., S. Guputa, and L. Jin. 2010. PEAS V1.0: A Package for Elementary Analysis of SNP Data. Mol. Ecol. Resour. 10:1085-1088.

II) Genomic variability in Mexican chicken population using copy number variants

E. Gorla, M. C. Cozzi, S. I. Román-Ponce, F. J. Ruiz López, V. E. Vega Murillo, S. Cerolini, A. Bagnato, M.G. Strillacci (2017).

BMC Genetics
2017, 18:6.

doi/10.1186/s12863-017-0524-4

Background

Copy Number Variants (CNV) are genomic structural variations distributed over the whole genome in all species and refers to genomic segments of at least 50 bp in size [1], for which copy number differences have been observed in comparison to reference genome assemblies (insertions, deletions and more complex changes) [2-3]. Sequencing of the chicken genome, released in 2004 [4], has facilitated the use of molecular markers for breed/ecotype characterization. Structural variation has been recognized as an important mediator of gene and genome evolution within populations. In the last decades, microsatellite markers have been often used to perform phylogenetic analysis and studies on genetic variability in chicken populations [5-6-7]. Although numerous studies investigating genetic variation have focused on SNPs, there is a growing evidence for the substantial role of structural DNA polymorphism in phenotypic diversity [8]. It has been shown that CNVs are ubiquitous in the genome and can contribute substantially to phenotypic variability and disease susceptibility in humans [8-9] and animals [10-11]. The underlying assumption is that CNVs are changing the gene structure and dosage and altering the gene regulation [8-12-13]. Even if CNVs are less frequent than SNPs in terms of absolute numbers, CNVs cover a larger proportion of the genome and have, therefore, a large potential effect on phenotypic variability [14]. Compared with humans and other model organisms, there is limited research on the extent and impact of CNVs in the chicken genome.

In Mexico the poultry population, even if it shows large morphological variability, is not divided into breeds or strains and, possibly, can be considered as a unique widespread Creole chicken population (*Gallus gallus domesticus*), as the result of undefined crosses among different breeds imported into Mexico from Europe for almost 500 years [15-16]. Creole chickens include, in fact a wide variety of biotypes with different colors of plumage and morphological features that are widely distributed in the country [17]. In the absence of comprehensive breed characterization data and documentation of the origin of breeding populations, DNA polymorphism provides the most reliable estimates of genetic diversity within and between a given set of populations [18].

Several studies have been developed in the recent past to detect CNV in poultry using low-density 60K SNP chips [19] or aCGH [20-21-22]. The major limits of these studies reside in the density of the spots of the used arrays and the limited sample size. It has been already suggested by Jia et al. [23] that the use of the 600K SNP array can improve the efficiency of the CNV detection in the poultry species. The whole genome sequence data can improve the detection of small CNVs but, even if desirable and employed by some authors [24-25], is still economically too demanding to be realized over a large number of samples.

The aim of this study was to map the CNV in the Mexican chicken population with an unprecedented resolution using high density SNP chip (i.e. 600K Affymetrix SNP chip) on a large number of individuals (i.e. 256) and to characterize the genetic variability of the Mexican Creole chicken's population using CNV as genomic markers.

Methods

Sampling and genotyping

In this study a collection of 265 individuals of the Mexican chicken population, from different farms across 26 states of

United States of Mexico, was previously sampled by Instituto Nacional de Investigaciones Forestales, Agricola y Pecuarias (INIFAP) within the institutional activities of the Centro Nacional de Recursos Geneticos at Tepatiplan, Jalisco. As mentioned hereinbefore, a classification of the Mexican population in breeds does not exist. For this reason, the birds have been considered as a unique Creole population and sampled in several states of Mexico.

Samples were processed and genotyped within the framework of a previous project of INIFAP using the 600K Affymetrix Axiom® Chicken Genotyping Array, containing 580,954 SNPs distributed across the genome, with an average spacing of about 1.8 kb and data made available for the present study. A commercial service provider performed the genotyping and the DNA extraction from feathers. The galGal4 chicken assembly was used in this study as reference genome.

Quality assurance of CNV raw data and CNV detection

The CNV detection was performed on a total of 471,730 SNPs on the first 28 chicken autosomes.

The Axiom® Analysis Suite software (Affymetrix) was used to perform raw intensity data Quality Control and run the genotyping algorithms. Default quality control settings were applied to filter for low quality samples before running the genotyping analysis, to exclude the ones with call rates < 97% and Dish Quality Control <0.82. The Axiom® CNV summary software tool was used to generate input files for CNV prediction analysis.

The CNV detection was performed with PennCNV software [26] using Log R Ratio and the B allele frequency [27] obtained with the Axiom® CNV Summary Tool software. The individual-based

CNV calling was performed using the default parameters of the Hidden Markov Model (HMM): standard deviation of LRR <0.30, BAF drift as 0.01 and waviness factor at 0.05 and a minimum of 3 SNP was required to define a CNV. The distribution of CNV per individual spanned from 0 CNV to more than 100. Up to 79 CNV the distribution was continuous, while a step to more than 100 CNV was detected in 9 birds. To avoid the introduction of possible false positive and a bias in the CNV interpretation they were then filtered out as the number of CNVs detected appeared to be outlier respect to the CNV distribution, leaving 256 samples for further analyses. It is worth to mention that Zhang et al. [19] have performed a validation of the CNV called by PennCNV, using the CNVPartition program obtaining an overlapping of results of 99%. Additionally, recent studies in cattle [28] have used two software to map CNV based on different algorithms: the HMM of PennCNV, based on the CNV identification on B allele frequency and Log R ratio, and the CNAM of SVS (Golden Helix) basing the identification only on Log R ratio. These studies provide an additional empirical evidence of the results provided by Xu et al. [29] that in their study concluded that using multiple CNV calling algorithms might also increase the false positive rate.

In addition to detect the outliers as hereinbefore described, in order to minimize the false positive callings, the PennCNV was run using different ".hmm" files (agre.hmm, affygw6.hmm, hh550.hmm), which is known that may affect substantially the false positive as well as the false negative rate. The online PennCNV manual (www.openbioinformatics.org) in fact instruct the user that the agre.hmm file produces an excess of false positive calls respect to the default affygw6.hmm file, which has been criticized to produce a low number of CNV calls (i.e. excess of false negative) respect to other calling software and algorithms. Additionally, we used the hh550.hmm file in the calling process, which is based on a chip with the closest number of SNPs respect to the SNP chip used here. To reduce the false calling rate, we have then considered valid only the CNV calls obtained both with the agre.hmm and the hh550.hmm files. The number of CNV calls resulted using the affygw6.hnm files were negligible respect to other two files, but anyhow present in the consensus here obtained. The hmm file supplied to the HMM of PennCNV, (www.openbioinformatics.org), provides to the algorithm the expected signal intensity values for different states of CNV and the expected probability for the transition in different copy number state. As described in the PennCNV user manual, however, the transition probability is a function of the distance between neighboring markers. This makes the choice of a correct hmm file, in respect to the density of markers, a critical step in the mapping of CNV to control false positive and negative calls.

CNVR annotation

After downloading the list of chicken autosome galGal4 genes (GCA_000002315.2) from Ensembl database (Release 88) (http://www.ensembl.org), the gene annotation was performed using the software Bedtools, *intersect* command [30], identifying the genes fully included in, or partially overlapping, the defined CNVRs. Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis were performed using the Panther database (http://pantherdb.org).

Clustering analysis using CNVRs.

A clustering analysis was performed considering CNVRs found in this study [31]. A scoring matrix of the CNVRs was constructed, attributing the “0” or “1” values to indicate the absence or the presence of a CNV in a specific CNVR. A hierarchical agglomerative clustering was then applied to the scoring matrix using the pvclust function of the pvclust R package [32]. Multiscale bootstrap resampling (no. 10,000 bootstraps) was used to obtain the Approximately Unbiased P-value (AU), in order to determine the robustness of branches. The Unweighted Pair Group Method with Arithmetic mean (UPGMA) was the Agglomerative method chosen.

Results

CNV and CNVR detection

The Table 1 reports the descriptive statistics of identified CNVs and CNVRs. The HMM of the PennCNV software detected a total of 1,924 CNVs; among these, 386 were deletions (i.e. loss state) and 1,538 were duplication (i.e. gain state), with a deletions/duplications CNV ratio of 0.25, calculated as the total number of losses divided by the total number of gains.

The CNVs overlapping among samples were summarized across all individuals into 1,216 CNVRs (959 gains, 226 losses and 31 complex), covering a total of 47 Mb of sequence length, corresponding to 5.12 % of 28 autosomes in the galGal4 assembly (Additional File 1: Sheet 1).

In Figure 1 the CNVRs map, divided in gain, loss and complex on each chromosome is shown.

In Table 2 the number of CNVRs found is reported, together with the state and the proportion of coverage by chromosome. The coverage proportion is smaller than 10% for all chromosomes, except for 16, 18, 24, 27 ones.

CNVRs were classified as singleton if detected in only one individual. Among the identified CNVRs, 1,009 (82.9%) were present in singleton, 127 (10.4%) in two individuals, 30 (2.4%) in three individuals, 11 (0.9%) in four individuals, and 39 (3.2%) in five or more individuals. For every state (i.e. gain, loss, complex) CNVRs were divided according to their length into four classes: <1 kb, 1-10 kb, 10–100 kb, >100 kb; Figure 2 reports the CNVRs count for each class of CNVRs length.

The majority of the 1,065 CNVRs identified in this study had a length comprised between 10 kb and 100 kb, of which 471 comprised between 1 kb and 10 kb and 594 comprised between 10 kb and 100 kb. A total of 39 CNVRs had a length lower than 1 kb while 112 CNVRs showed a size longer than 1 Mb (Figure 3). The highest number of gain and complex CNVRs are those with a length of 10–100 kb, while the loss CNVRs were present at largest frequency within a length of 1–10 kb (Figure 3).

The regions mapping in a large number of individuals were: the CNVR on chromosome 1 at 42.96-43.13 Mb, identified in 61 samples; the CNVR on chromosome 12 at 1.12-1.22 Mb, identified in 56 samples; the CNVR on chromosome 16 at 1,253-533,589 bp, identified in 53 samples; the CNVR on chromosome 1 at 193.13 – 193.24 Mb, identified in 52 samples.

The Figure 3 shows the sample count for every CNVR state according to the previously defined 4 CNVR length classes (as shown in Figure 2). The sample count classes were defined as: 1 (singleton), 2–5, 6–20 and > 20.

The gain CNVRs (Figure 3.A) have a sample count distribution with most of the regions falling into the 10–100 kb class. The loss CNVRs (Figure 3.B) have a sample count distribution with most of the regions falling into the 1–10 kb class. Class 1 mostly represents the gain regions. Furthermore, class 1 is the most frequent in all length classes. The highest length and sample classes mainly belong to the gain regions. In the complex region

(Figure 3.C) the class mostly represented is the 10-100 kb one. More precisely, the most represented sample class is the 2–5 class falling mainly within the 10–100 kb length class. Furthermore, class 2-5 is the most frequent. Lastly, all the sample classes are distributed mostly within the 1–10 and 10-100 length classes.

CNVR annotation

The intersection analysis performed between the chicken gene database (Ensembl galGal4) and our CNVRs allowed the identification (within or overlapping the consensus CNVRs) of 1,543 Ensemble genes ID, corresponding to 1,068 genes with an official gene ID. Out of the 1,216 CNVRs identified in this study, 783 (64.4%) encompassed one or more genes, while 433 (35.6%) did not involve any gene. More specifically, among these genes, 1,028 (96.25%) were protein-coding genes, 34 (3.1%) were miRNAs and 6 (0.56%) were small nuclear RNAs (Additional File 1: Sheet 4).

The Panther database provided the annotation information, according to GO terms and KEGG pathways, for only 865 chicken genes. The Additional File 1 reports the annotation output including only terms resulted statistically significant after Bonferroni correction (p-value < 0.05): 27 classified as Cellular Component, 11 as Molecular Function, and 28 as Biological Process. The significant GO terms were mainly involved in muscle contraction, sensory perception of sound, response to stimulus, cellular component morphogenesis and movement, and cell communication (Additional File 1: Sheet 5). Instead, the KEGG pathway analysis indicated that these genes are involved in 166 pathways, but none of which was significant after Bonferroni correction.

Clustering analysis using CNVRs.

The Figure 4 shows the cluster-tree built for the chicken Mexican Creole population based on CNVRs similarities.

In the plot (Figure 4), the branch length is not directly proportional to the genetic distance estimated among samples. The Approximately Unbiased P-value (AU-P in red) and Bootstrap Probability value (BP-P in green), indicative of how strongly the cluster is supported by the data, were shown for each node, as well as the Edge numbers (in light grey). As can be read from Figure 4 mostly all AU-P and BP-P values are zero, showing no difference among branch in which individuals are clustered in: there is no cluster with both AU-P and BP-P values greater than 0.

Discussion

CNV and CNVR detection

The use of a high-density SNP chip allows to disclose smaller CNVs compared to studies performed in the recent past that were based on a 60K SNP chip [19] or on aCGH [20-21-22]. The average probe distance in the SNP chip used here is in fact more or less 1,8 kb (galGal4) allowing the identification of short CNVs. The smaller CNV (i.e. 92 bp.) that was detected in this study (Table 1), according to the criteria of minimum 3 SNPs to map a CNV, overlaps with the one mapped by Yi et al. [24] using a sequencing approach.

Chromosome 16 is the only one with a very large proportion of length covered by CNV, i.e. 81% (Table 2). This may be due to the small length of the autosome and to the presence of the Major Histocompatibility Complex (MHC), which is known to be affected by variation in genome copy number as reported by

Fulton et al. [33]. The region is resulting in this study as a complex CNVR but having the majority of individual CNVs (46 over 52) to be gain variant (45 heterozygous duplications, 1 homozygous duplication). The existence of such a CNV is possibly due to the importance that the MHC has in immune resistance. As it is known by literature in fact, the high number of polymorphic sites, closely associated with resistance against infection diseases (e.g. Marek's disease, avian Influenza, Rous sarcoma disease, avian leukosis, infectious bursal disease, avian infectious bronchitis, *Salmonella enteritidis*, *E. coli* and other bacterial diseases), characterizes this complex [34-35].

The large proportion of singleton CNVRs has been previously reported in chicken populations also by Yi et al. [24], Han et al. [22] and Strillacci et al. [36], finding a total fraction of 68.8%, 76.5% and 75%, respectively. Our findings confirm their results and showed that also in the Mexican chicken population the segregation of CNVs exists among individuals.

Comparison with previous chicken CNV studies

In order to perform a comparison with previous studies mapping CNVs in chicken, we migrated autosomal CNVRs coordinates from galGal3 to galGal4 for the CNVRs identified by Tian et al. [21], by Crooijmans et al. [20] and by Han et al. [22] using the UCSC liftOver tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver). In total 201 out of 308 (65%) autosomal CNVRs detected by Tian et al. [21], 837 out of 1,504 (56%) mapped by Croijmans et al. [20] and 134 out of 264 (50.75%) identified in Han et al. [22] were converted successfully.

The comparison among the CNVRs found in this study and those found in other 7 studies [19-20-21-22-24-25-36] is reported in Table 3 and in the Additional File 1: Sheet 2 showing the number of CNVRs overlapping among the studies.

The 1,216 CNVRs detected in this study overlap with 617 mapped by others confirming that a proportion of 51% of them were found by independent methods and in other populations (Additional File 1: Sheet 2).

As reported in Table 3, the proportion of overlapping CNVRs between this study and each of the other 7 studies ranged from 2.38% to 35.19%. Independently from the breeds included in all studies, the CNVRs detection is mainly influenced by the sample size and by the algorithm and the technology used to CNVs mapping (i.e. aCGH *vs.* SNP or whole genome sequence). The largest overlap rates occurred in fact when the comparison is done with studies using in their analyses a large sample of individuals [24- 25]. On the contrary, a low overlap occurred when the comparison was performed with studies that employed a low number of samples, when CNVs were detected with different technical methods (i.e. aCGH or whole genome sequencing) and calling algorithms.

No CNVR is simultaneously common to this and to all the 7 other studies here considered. The Additional File 1: Sheet 3 reports the list of CNVRs simultaneously shared by our study and at least 3 other ones among the 7 here considered, and the annotated genes found in the regions. As shown, the CNVR common among 7 studies are 4 and are located on chromosome 1 at 42.96-43.13 Mb, chromosome 5 at 2.6-3.9 Mb, chromosome 8 at 15.45-15.47 Mb and chromosome 9 at 3.42- 3.49 Mb.

In particular the CNVR on chromosome 1 is common to 7 studies and includes the *KITLG* (*KIT ligant*), a pigmentation candidate gene that has a role in controlling the migration, survival and proliferation of melanocytes. Rare mutations in the mouse homolog of *KITLG* are known to affect coat color [37].

Additionally, Metzger et al. [38] highlighted the importance of this gene in the reproduction efficiency in horses claiming its general effect in all livestock populations.

The CNVR on chromosome 5 (2.60-3.95 Mb) (Additional File 1: Sheet 3) harbors the *BDNF* (*brain derived neurotrophic factor*) gene, which seems to be involved in chicken heat stress response. In fact, Lamont et al. [39] reports that early thermal conditioning allows increased transcription of the *BDNF* gene in response to heat stress later in the bird's life. Furthermore, previous findings indicate that *BDNF* prevents the death of cultured chick retinal ganglion cells and, as reported by Herzog et al. [40], the tightly controlled expression of the *BDNF* gene might be important in the coordinated development of the visual system in chicks. Also, in the same CNVR on chromosome 5 lies the *LGR4* (*leucine rich repeat containing G protein-coupled receptor 4*) gene that in human is associated with low bone mineral density [41]

In the region on chromosome 8 no genes were annotated, while in the region on chromosome 9 the *IMP4 (U3 small nucleolar ribonucleoprotein)* and the *VPS8 (Vacuolar Protein Sorting-Associated Protein 8 Homolog)* genes are annotated, but there are no studies that associate these genes to specific traits.

CNVR annotation

Additionally, quantitative trait loci (QTL) from chicken QTLdb (http://cn.animalgenome.org/cgi-bin/QTLdb/GG/index) were downloaded in order to examine their overlapping with the identified CNVRs. Because the confidence intervals of some QTL were too large, we considered QTL less than 5 Mb of length. A total of 656 CNVRs overlapped with 918 QTL, corresponding to

172 different traits that included mainly: body weight, body size, carcass traits, fatness traits, Marek's disease-related traits, and eggshell (Additional File 1: Sheet 6).

Some of the genes identified in our CNVR have already been associated with functional traits in chickens in previous studies. The region identified on chromosome 4 at 80.75-81.02 Mb contains the gene *SORCS2* (*sortilin related VPS10 domain containing receptor 2*) associated with aggressive behavior traits in males [42]. The region on chromosome 1 at 130.82-130.89 Mb includes the gene *OCA2* (*oculocutaneous albinism II*)*.* This gene had highly significant effects on body weight in weeks 11–12 in chicken, as reported by Gu et al. [43] and is also involved in pigmentation [44]. The CNVRs on chromosome 1 at 65.63- 65.98 Mb and at 66.02- 66.03 Mb harbor *SOX5* (*SRY-box 5*) gene, which is involved in chicken the Pea-comb expression. In fact, Pea-comb is caused by a duplication located near conserved non-coding sequences in intron 1 of the gene [45]. Three regions on chromosome 1 at 146.55-146.59 Mb, at 147.08-147.13 Mb and at 147.78-147.80 Mb harbor the *glypican 6 (GPC6)* gene, *glypican 5* (*GPC5*) gene, which are located within the QTL for bodyweight identified in previous studies [46-47].

The CNVR on chromosome 18 (5.00-5.02 Mb) includes the *FASN* (*fatty acid synthase*) gene that has been identified as one of the genes that control fat deposition in chickens (i.e. fat bandwidth, abdominal fat percentage and abdominal fat weight) [48].

Finally, some genes included in 10 different CNVRs found in this study are classified into the pathway for salmonella infection (http://www.genome.jp/dbget-bin/www_bget?gga05132).

These genes are: *IFNG* (*interferon gamma*) (chromosome 1 at 34.95-35.16 Mb), *DYNC2H1* (*dynein cytoplasmic 2 heavy chain 1*) (chromosome 1 at 182.31-182.3 Mb), *WASF1* (*WAS protein family* *member 1*) (chromosome 3 at 66.86-66.87 Mb), *ARPC2* (*actin related protein 2/3 complex subunit 2*) (chromosome 7 at 22.60-22.70 Mb), *TJP1* (*tight junction protein 1*) (chromosome 10 at 6.08- 6.11 Mb), *DYNC1LI2* (*dynein cytoplasmic 1 light intermediate chain 2*) (chromosome 11 at 11.42- 11.51 Mb), *FLNB* (*filamin B*) (chromosome12 at 8.87-8.87 Mb), *RAB7A* (*member RAS oncogene family*) (chromosome 12 at 9.15- 9.15 Mb), *ARPC1B* (*actin related protein 2/3 complex subunit 1B*) (chromosome 14 at 4.38- 4.38 Mb), *PLEKHM2* (*pleckstrin homology and RUN domain containing M2*) (chr21 at 4.21- 4.22 Mb).

Clustering analysis using CNVRs

The results of this study suggest that there is not a clear division in classifiable subpopulations based on the CNVR characterization and, thus, that the Mexican Creole chicken population can be considered a unique genetic mix. These results are different to the ones recently found by Strillacci et al. [36] using the same approach in Italian well defined chicken breeds clearly clustered by CNVRs classification and by Tian et al. [21] and Wang et al. [49] in chicken and pigs respectively, showing additional evidence of the usefulness of CNV as markers for differentiating individuals. To provide a validation of the approach here used to cluster individuals of the Mexican population with CNVs we performed a PCA and a hierarchical clustering using the SNP genotypes: no clustering was obtained, and the population resulted as for CNVs a unique genetic mix (Additional File 2: Figure S1).

Conclusion

This study is the first CNV genomic analysis on a large sample of individuals of the Mexican chicken population based on high-density SNP chips. It provides insights into the genetic and genomic architecture of the Mexican Creole chicken population, providing valuable genomic source of structural variation to enrich the chicken CNV map, helping future CNV association studies for important traits in chickens. The major result resides in the disclosure of the genetic homogeneity of the Mexican chicken population. This result allows to consider all individuals of population as a unique genetic mix deriving from the introduction of chicken in the American continent, following the colonization from Europe. According to our results the CNV variation in the population does not allow to disclose breeding strategy addressed to specific selection criteria. The same method, we used here based on the CNV, was able to dissect properly different Italian breeds in a previous study [36]. The results of this study, thus, suggest that there is not a clear division in classifiable subpopulations based on the CNVR characterization and that the Mexican Creole chicken population can be considered a unique mix of genetics. Most of the 1,216 CNVRs detected were novel variants disclosed thanks to the HD SNP chips here used, which enrich the current poultry CNV database. This mapping is having a particular value because it is based on a unique poultry population, that we assumed to own a larger genetic variability respect to selected commercial population, as reproduction is based on an outbreeding mating system by more than 500 years. Finally, we detected 1,543 Ensemble genes ID overlapping with CNVRs, including genes involved in well-known phenotypes such as KITLG and *OCA2* on chromosome 1, *SORCS2* on chromosome 4, *FASN* on chromosome 18. Also, some genes included in 10 different CNVRs found in this study, belong to the pathway for salmonella infection. The MHC region on chromosome 16, which has great interest for disease resistance, lies on a region that is in common among the CNVRs of four studies.

Tables

Table 1. Descriptive statistics for Copy Number Variants (CNVs) and Copy Number Variants Regions (CNVRs) identified in the Mexican chicken population

Type	No.	Length	Min length	Max length	Mean length	Median length	Total Coverage
CNVs							
Loss	386	12,575,609	92	574,231	32,579	6,038	1.37%
Gain	1,538	74,022,420	138	1,345,291	42,129	22,810	8.05%
All	1,924	86,598,029	92	1,345,291	45,009	19,273	9.42%
CNVRs							
Loss	226	3,920,955	92	279,420	17,349.36	4,950	0.43%
Gain	959	38,550,088	138	1,345,291	40,198.21	15,414	4.19%
Complex	31	4,580,519	3,501	607,435	147,758.7	60,250	0.50%
All	1,216	47,051,562	92	1,345,291	38,693.72	13,897.5	5.12%

Table 2. Number and proportion of genome covered (Coverage %) by Gain, Loss and Complex Copy Number Variants Regions per chromosome (CHR).

CHR	Gain (*)	Loss (*)	Complex (*)	Total	Coverage (%)
1	186 (3.94)	46 (0.38)	6 (0.29)	238	4.61
2	140 (4.78)	31 (0.38)	2 (0.14)	173	5.29
3	101 (3.02)	18 (0.11)	0 (0)	119	3.13
4	58 (3.40)	20 (0.36)	0 (0)	78	3.75
5	58 (6.43)	8 (0.15)	0 (0)	66	6.58
6	41 (3.61)	9 (0.15)	1 (0.15)	51	3.91
7	36 (4.03)	2 (0.02)	1 (0.46)	39	4.51
8	32 (4.55)	1 (0.30)	1 (0.68)	34	5.53
9	25 (3.22)	8 (0.23)	0 (0)	33	3.45
10	32 (5.06)	9 (0.79)	2 (1.11)	43	6.96
11	17 (2.64)	7 (0.78)	1 (0.19)	25	3.61
12	26 (2.73)	4 (0.16)	0 (0)	30	2.89
13	30 (3.88)	8 (1.05)	1 (0.52)	39	5.45
14	32 (7.72)	7 (2.05)	1 (0.20)	40	9.97
15	18 (1.90)	3 (0.12)	1 (0.31)	22	2.33
16	0 (0)	0 (0)	1 (81.60)	1	81.60
17	8 (2.28)	5 (0.97)	0 (0)	13	3.26
18	12 (3.54)	7 (2.06)	2 (5.03)	21	10.63
19	22 (8.32)	4 (0.23)	1 (0.91)	27	9.46
20	17 (3.57)	3 (0.26)	2 (0.39)	22	4.22
21	9 (1.60)	5 (0.30)	0 (0)	14	1.90
22	8 (4.31)	2 (0.74)	1 (0.62)	11	5.67
23	9 (4.78)	5 (0.95)	1 (0.73)	15	6.46
24	12 (9.91)	2 (0.24)	0 (0)	14	10.14
25	3 (2.41)	3 (1.13)	2 (2.39)	8	6.48
26	6 (2.27)	5 (2.11)	1 (1.46)	12	5.84
27	11 (6.04)	4 (3.66)	1 (10.74)	16	20.45
28	10 (3.36)	0 (0)	2 (2.24)	12	5.61
Total	959	226	31	1,216	

* Coverage of CNVR by chromosome and state (gain/loss/complex) relatively to each chromosome length.

Table 3. Comparison between CNVRs detected in this study and in other 4 published ones.

Study	Method	Samples	Breeds	CNVR	Length overlap (Mb)	Common CNVR	Overlap (%)
Crooijmans et al. [20]	aCGH	64	7	837*	4.49	92	7.57
Tian et al. [21]	aCGH	22	11	201*	0.969	29	2.38
Zhang et al. [19]	SNP chip (60K)	475	11	438	19.903	80	6.58
Han et al. [22]	aCGH	10	4	134*	1.311	29	2.38
Yi et al. [24]	Sequencing	12	12	8,487	10.424	428	35.19
Yan et al. [25]	Sequencing	6	2	5,009	2.933	256	21.05
Strillacci et al. [36]	SNP chip (600K)	96	6	564	3.855	109	8.96
This Study	SNP chip (600K)	256	1	1,216	47.05		

* This value refers to the number of CNVRs after the shifting to

Figures

Figure 1. Physical distribution of the Copy Number Variants Regions (CNVRs) according to states (gain, loss and complex).

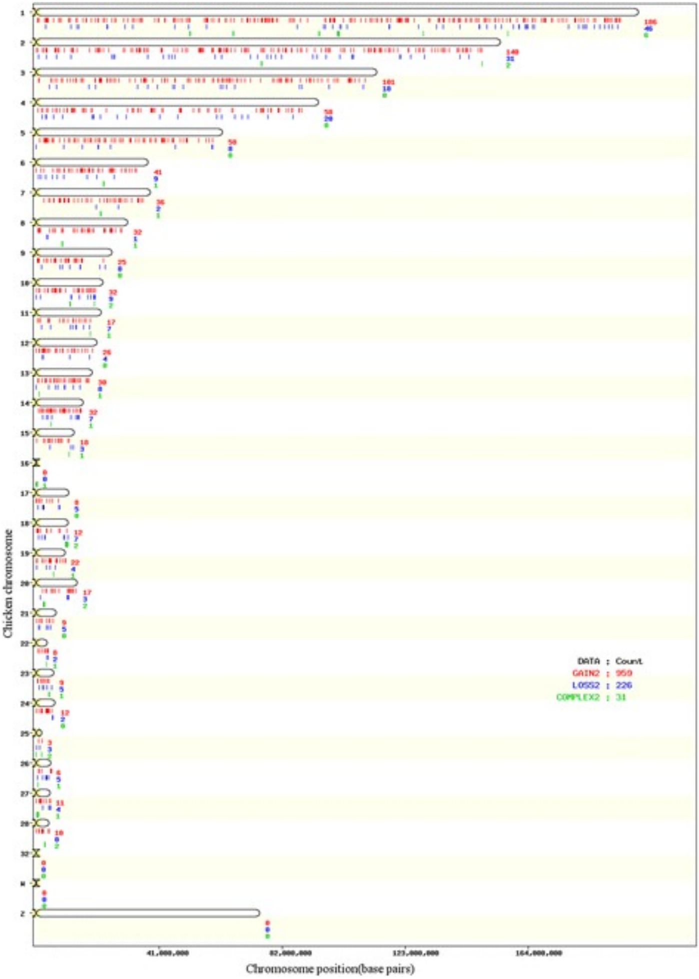


Figure 2. Distribution of CNVRs lengths identified with PennCNV

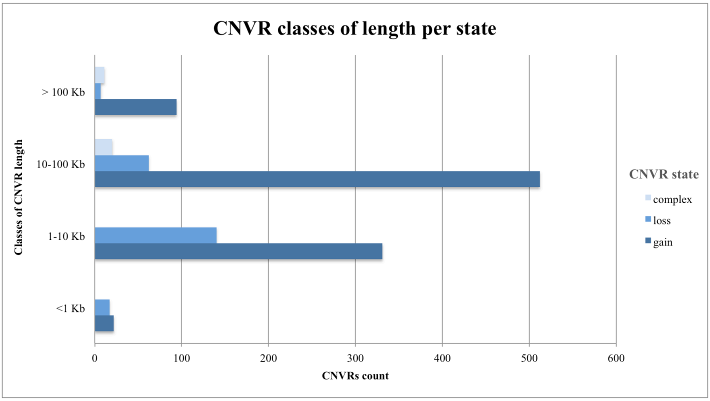


Figure 3. Sample count per classes of samples (1 singleton; 2-5; 6-20; >20) in each class of CNVR length (<1; 1-10; 10-100; >100 kb), according to the different CNVRs states.

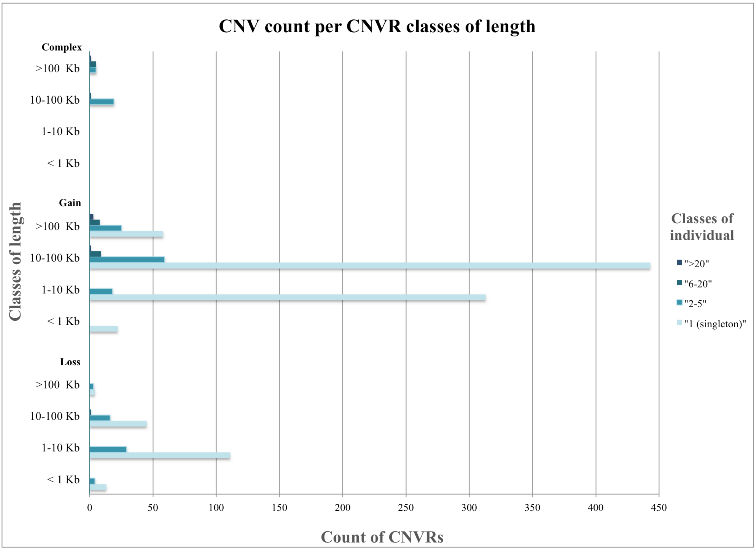
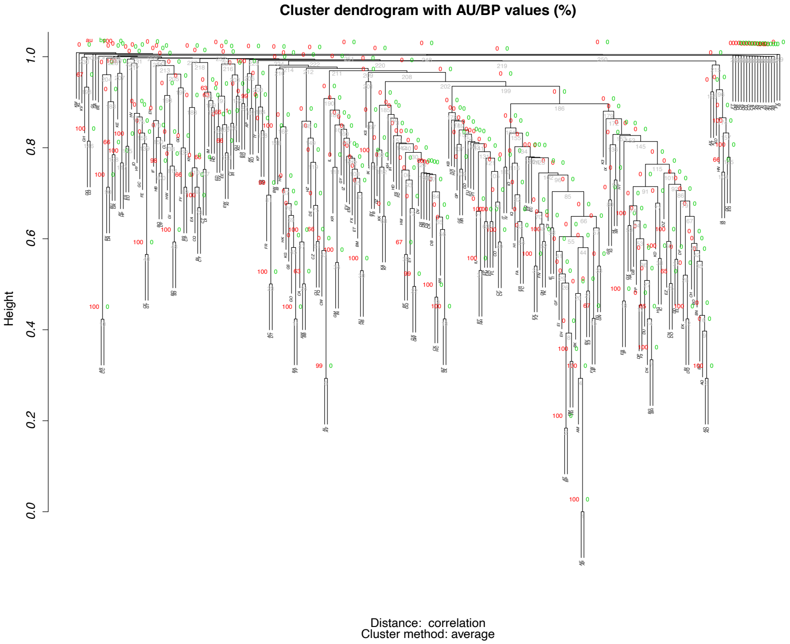


Figure 4. Cluster dendrogram with AU/BP values (%)



Supporting information

All supplementary files are available at:

https://doi.org/10.1186/s12863-017-0524-4

References

1. Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, et al. Mapping copy number variation by population-scale genome sequencing. Nature. 2011; 470: 59-65.
2. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nat Rev Genet. 2006; 7: 85–97.
3. Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, et al. Copy number variation: new insights in genome diversity. Genome Res. 2006; 16: 949–961.
4. Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, et al. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature. 2004; 432: 695-716.
5. Al-Qamashoui B, Simianer H, Kadim I, Weigend S. Assessment of genetic diversity and conservation priority of Omani local chickens using microsatellite markers. Trop Anim Health Prod. 2014; 46: 747-752.
6. Strillacci MG, Marelli SP, Cozzi MC, Colombo E, Polli M, Gualtieri M, et al. Italian autochthonous chicken breeds conservation: evaluation of biodiversity in Valdarnese Bianca breed (Gallus gallus domesticus). Avian Biol Res. 2009; 2: 229-233.
7. Ceccobelli S, Di Lorenzo P, Lancioni H, Monteagudo Ibáñez LV, Tejedor M, Castellini C, et al. Genetic diversity and phylogeographic structure of sixteen Mediterranean chicken breeds assessed with microsatellites and mitochondrial DNA. Livest Sci. 2015; 175: 27-36
8. Zhang F, Gu W, Hurles ME, Lupski JR: Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet. 2009; 10: 451–481.
9. McCarroll SA, Altshuler DM. Copy-number variation and association studies of human disease. Nat Genet. 2007; 39:

37–42.

10. Wang X, Nahashon S, Feaster TK, Bohannon-Stewart A, Adefope N. An initial map of chromosomal segmental copy number variations in the chicken. BMC Genomics. 2010; 11:351.
11. Yalcin B, Wong K, Agam A, Goodson M, Keane TM, Gan X, et al. Sequence-based characterization of structural variation in the mouse genome. Nature. 2011; 477: 326–329.
12. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number in the human genome. Nature. 2006; 444: 444–454.
13. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, et al. Origins and functional impact of copy number variation in the human genome. Nature. 2010; 464:704–712.
14. Yang Z, Zhuan B, Yan Y, Jiang S, Wang T. Integrated analyses of copy number variations and gene differential expression in lung squamous-cell carcinoma. Biol Res. 2015; 48(1):47.
15. Segura-Correa JC, Sarmiento-Franco L, Magaña-Monforte JG, Santos-Ricalde R. Productive performance of Creole chickens and their crosses raised under semi-intensive management conditions in Yucatan, Mexico, Br Poult Sci. 2004; 45(3): 342-345.
16. Rodriguez JC, Allaway CE, Wassink, GJ, Segura JC, Rivera T. Estudio de la Avicultura de traspatio en el municipio de Dzununcàn, Yucatàn. Vet Mex. 1996; 27(3): 215-219.
17. Segura-Correa JC, Juarez-Caratachea A, Sarmiento-Franco L, Santos-Ricalde R. Growth of Creole chickens raised under tropical conditions of Mexico. Trop Anim Health Prod. 2005; 37(4): 327-332.
18. Ceccobelli S, Lorenzo PD, Lancioni H, Castellini C, Ibáñez LV, Sabbioni A, et al. Phylogeny, genetic relationships and population structure of five Italian local chicken breeds.

Italian J AnimSci. 2013; 12(3):e66.

19. Zhang H, Du ZQ, Dong JQ, Wang HX, Shi HY, Wang N et al. Detection of genome-wide copy number variations in two chicken lines divergently selected for abdominal fat content. BMC Genomics. 2014; 15:517.
20. Crooijmans RP, Fife MS, Fitzgerald TW, Strickland S, Cheng HH, Kaiser P, Redon R, Groenen MA. Large scale variation in DNA copy number in chicken breeds. BMC Genomics. 2013;14:398.
21. Tian M, Wang Y, Gu X, Feng C, Fang S, Hu X, Li N. Copy number variants in locally raised Chinese chicken genomes determined using array comparative genomic hybridization. BMC Genomics. 2013; 14:262.
22. Han R, Yang P, Tian Y, Wang D, Zhang Z, Wang L, Li Z, Jiang R, Kang X. Identification and functional characterization of copy number variations in diverse chicken breeds. BMC Genomics. 2014; 15: 934.
23. Jia X, Chen S, Zhou H, Li D, Liu W, Yang N. Copy number variations identified in the chicken using a 60K SNP BeadChip. Anim Genet. 2012; 44:276-84.
24. Yi G, Qu L, Liu J, Yan Y, Xu G, Yang N. Genome-wide patterns of copy number variation in the diversified chicken genomes using next-generation sequencing. BMC Genomics. 2014; 15(1): 962.
25. Yan Y, Yang N, Cheng HH, Song J, Qu L. Genome-wide identification of copy number variations between two chicken lines that differ in genetic resistance to Marek's disease. BMC Genomics. 2015; 16: 843.
26. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant S, Hakonarson H, Bucan M. PennCNV: an integrated hidden Markov model designed for high- resolution copy number variation detection in whole-genome SNP genotyping data.

Genome Res. 2007; 17 (11):1665–1674.

27. Peiffer DA, Le JM, Steemers F.J., Chang W, Jenniges T, Garcia F, et al. High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. Genome Res. 2006; 16: 1136–1148.
28. Durán Aguilar M, Román Ponce SI, Ruiz López FJ, González Padilla E, Vásquez Peláez CG, Bagnato A, Strillacci MG. Genome-wide association study for milk somatic cell score in holstein cattle using copy number variation as markers. J Anim Breed Genet. 2017; 134(1), 49-59.
29. Xu L, Hou, Y, Bickhart DM, Song J, Liu GE. Comparative analysis of CNV calling algorithms: literature survey and a case study using bovine high-density SNP data. Microarrays. 2013; 2(3): 171-185.
30. Quinlan, AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26: 841–842.
31. Gazave E, Darré F, Morcillo-Suarez C, Petit-Marty N, Carreño A, Marigorta UM, Ryder OA, Blancher A, Rocchi M, Bosch E, Baker C. Copy number variation analysis in the great apes reveals species-specific patterns of structural variation. Genome Res. 2011;21(10):1626-39.
32. Suzuki R, Shimodaira H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics. 2006; 12: 1540-1542.
33. Fulton JE, McCarron AM, Lund AR, Pinegar KN, Wolc A, Chazara O et al. A High-Density SNP Panel Reveals Extensive Diversity, Frequent Recombination and Multiple Recombination Hotspots Within the Chicken Major Histocompatibility Complex B Region Between BG2 and CD1A1. Genet Sel Evol. 2016; 48: 1.
34. Wang X, Byers S. Copy number variation in chickens: a

review and future prospects. Microarrays. 2014; 3: 24–38.

35. Garcia-Camacho L, Schat KA, Brooks Jr. R, Bounous DI. Early cell-mediated immune responses to Marek's disease virus in two chicken lines with defined major histocompatibility complex antigens. Vet Immunol Immunopathol. 2003; 95(3): 145–153.
36. Strillacci MG, Cozzi MC, Gorla E, Mosca F, Schiavini F, Román-Ponce SI et al. Genomic and genetic variability of six chicken populations using single nucleotide polymorphism and copy number variants as markers. Animal. 2017; 11(5): 737–745.
37. Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, Magnusson KP, et al. Genetic determinants of hair, eye and skin pigmentation in Europeans. Nature Genet. 2007; 39(12): 1443-52.
38. Metzger J, Karwath M, Tonda R, Beltran S, Águeda L, Gut M, Gut IG, Distl O. Runs of homozygosity reveal signatures of positive selection for reproduction traits in breed and non-breed horses. BMC Genomics. 2015; 16(1):764.
39. Lamont SJ, Coble DJ, Bjorkquist A, Rothschild MF, Persia M, Ashwell C, et al. Genomics of heat stress in chickens. Proceedings, 10th World Congress of Genetics Applied to Livestock Production. Vancouver, BC, Canada. August 17-22, 2014.
40. Herzog KH, Bailey K, Barde YA. Expression of the BDNF gene in the developing visual system of the chick. Development. 1994; 120(6): 1643-1649.
41. Styrkarsdottir U, Thorleifsson G, Sulem P, Gudbjartsson DF, Sigurdsson A, Jonasdottir A, et al. Nonsense mutation in the LGR4 gene is associated with several human diseases and other traits. Nature. 2013; 497(7450): 517-520.
42. Li Z, Zheng M, Abdalla BA, Zhang Z, Xu Z, Ye Q, et al. Genome-

wide association study of aggressive behaviour in chicken. Sci Rep. 2016; 6:30981.

43. Gu X, Feng C, Ma L, Song C, Wang Y, Da Y, Li H, et al. Genome-wide association study of body weight in chicken F2 resource population. PLoS One. 2011;6(7): e21872.
44. Zhang J, Liu F, Cao J, Liu X. Skin Transcriptome Profiles Associated with Skin Color in Chickens. PloS One. 2015;10(6): e0127301.
45. Wright D, Boije H, Meadows JR, Bed'Hom B, Gourichon D, et al. Copy number variation in intron 1 of SOX5 causes the Pea-comb phenotype in chickens. PLoS Genet. 2009; 5(6): e1000512.
46. Sewalem A, Morrice DM, Law A, Windsor D, Haley CS, Ikeobi CO, et al. Mapping of quantitative trait loci for body weight at three, six, and nine weeks of age in a broiler layer cross Poultry Sci. 2002; 81(12):1775-81.
47. Carlborg Ö, Hocking PM, Burt DW, Haley CS. Simultaneous mapping of epistatic QTL in chickens reveals clusters of QTL pairs with similar genetic effects on growth. Genet Res. 2004; 83(03):197-209.
48. D'Andre HC, Paul W, Shen X, Jia X, Zhang R, Sun L, Zhang X. c. J Anim Sci Biotechnol. 2013; 4(1): 43.
49. Wang Z, Chen Q, Yang Y, Liao R, Zhao J, Zhang Z, et al. Genetic diversity and population structure of six Chinese indigenous pig breeds in the Taihu Lake region revealed by sequencing data. Anim Genet. 2015; 46 (6): 697–701.

Online database

- Thomas lab at the University of Southern California. http://pantherdb.org. December 28, 2016.
- University of California, Santa Cruz. https://genome.ucsc.edu/cgi-bin/hgLiftOver. December 13, 2016.

- NAGRP - Bioinformatics Coordination Program. http://cn.animalgenome.org/cgi-bin/QTLdb/GG/index. December 30, 2016.
- Kanehisa, M.; "Post-genome Informatics". http://www.genome.jp/dbget-bin/www_bget?gga05132. January 1, 2017.

PART II

CNV mapping and population structure in turkey populations

Turkey (*Meleagris gallopavo*) domestication process began about 2000 years ago in ancient North America (i.e., the combined North and Central American sub-continents) (Thornton and Emery, 2015). The wild form of turkey was divided into seven subspecies (Howard and Moore, 1984) located in different geographic areas and having morphological and plumage differences.

The Mexican turkey is supposed to be the first ancestor of domestic turkeys (Crawford, R.D., 1990). Turkeys from central America underwent two main migratory processes. In the 16th century turkeys have been introduced into Europe and spread quickly across European countries (Schorger, 1966). In 17th Century French, Dutch and English colonists brought them back into North America, where they crossed them with local wild eastern subspecies (*Meleagris gallopavo silvestris*) (Crawford, 1984, 1990).

Since then, Turkeys experimented a massive expansion and became the second worldwide source of poultry meat, in particular in developing countries. In last 40 years, turkey stock almost tripled, average meat production per bird doubled, and selection pressure for economically important traits, such as egg production, meat quality and body weight were enhanced, showing an intensive selection process on turkey populations. (FAO)

Recently scientific studies focusing on turkey genetics developed rapidly, thanks to the availability of a reference whole genome sequence (Dalloul et al., 2010).

The following study is the first to use high density SNP chip to create CNV map in the Turkey species (*Meleagris Gallopavo*) in several autochthonous populations: the Mexican turkey, the Narragansett, 6 Italian breeds and a commercial hybrid, and to identify annotated genes harboured in the mapped CNVRs.

Reference

- Crawford R.D. (1990). Origin and history of poultry species. Poultry breeding and genetics, 1-41.
- Crawford R.D. 1984. Chapter 47. Turkey. In: Mason, I.L., editor. Evolution of Domesticated Animals. Longman Inc., New York.
- Food and agriculture organization statistical division (FAOSTAT) of the United Nations. http://faostat.fao.org/
- Dalloul R.A., Long J.A., Zimin A.V., Aslam L., Beal K., Ann Blomberg L., Bouffard P. et al. (2010) Multi-Platform Next-Generation Sequencing of the Domestic Turkey (Meleagris gallopavo): Genome Assembly and Analysis. PLoS Biol, 8(9):e1000475. doi: 10.1371/journal.pbio.1000475
- Howard R., and Moore A. (1991). A complete checklist of the birds of the world (No. Ed. 2). Academic Press Ltd.
- Schorger A.W. (1966). The wild turkey: its history and domestication (No. QL696. G2 S3)
- Thornton E.K. and Emery K.F. (2015). The Uncertain Origins of Mesoamerican Turkey Domestication. J Archaeol Method Th doi: 10.1007/s10816-015-9269-4

III) Copy Number Variation Mapping and Genomic Variation of autochthonous and commercial turkey populations.

Maria G. Strillacci, **Erica Gorla**, Angel Ríos-Utrera, Vicente E. Vega-Murillo, Moises Montaño-Bermudez, Adriana Garcia-Ruiz, Silvia Cerolini, Sergio I. Román-Ponce, Alessandro Bagnato

Frontiers in Genetics
2019, 10:982.

doi: 10.3389/fgene.2019.00982

Introduction

The domestication of the wild turkey appears to occur in Mexico between 200 B.C and 700 A.D. (Crawford, 1992). The domesticated turkey has been introduced in Europe from Mexico and central America starting in late 15th century (Schorger, 1966) by the Spanish conquerors. The diffusion of the turkey population in the European territory was very fast, close to 50 km per year as indicated by Crawford (1992). The rapid diffusion in Europe was possibly facilitated because of their farming, as turkey was appreciated for its meat (Schorger, 1996). Then, since 15th century, the populations of European and Mexican turkey evolved independently for more than 500 years.

At present in Europe there is a clear differentiation in several turkey breeds, indicating that farmers and breeders have selected the turkey populations according to a directional goal for more than five centuries. Additionally, in the last 40 years, companies developed a structured breeding plan to produce commercial hybrids selected to maximize meat production[1].

In this study six Italian autochthonous breeds (Colle Euganei (CoEu); Bronzato Comune Italiano (BrCI); Parma e Piacenza (PrPc); Brianzolo (BR); Nero d'Italia (NI) Ermellinato di Rovigo (ErRo)), the Narragansett, the Mexican turkey and a hybrid population were considered to disclose genome structural variations in a wide dataset of individuals from differently evolved populations.

The selection operated by farmers in the past 5 centuries in the Italian populations determined the appearance of strong variation in plumage colors, in body size and weight, differentiating the populations in breeds (Cavalchini, 1983). This

[1] https://www.coe.int/t/e/legal_affairs/legal_co-operation/biological_safety_and_use_of_animals/farming/Rec%20Turkeys.asp

differentiation was possibly also facilitated by the geopolitical structure of Italy in middle ages, structured in a large number of small states with very limited exchange of goods and populations, making each turkey population genetically isolated from the others. Plumage of these breeds spans from totally black (Nero d'Italia) to white with black streaks (Ermellinato di Rovigo), while it is generally bronze like or with bronze reflection in all the other Italian populations. Body size is also showing a considerable difference among the Italian breeds with male weight spanning from 4.5 to 6.5 Kg in the Brianzolo and reaching 12 kg in the Ermellinato di Rovigo (Table 1). Due to the fact that local farming occurred for centuries, it is expected that genetic bottleneck occurred in the Italian populations. The Mexican turkey population has historically been farmed as a backyard population without any directional selection for centuries, with a plumage very variable in its color and a weight close to 6 kg in males. In fact, in this population, there is no any structured selection program, while its genetic peculiarity is a strong argument in favor of its conservation (Utrera et al., 2016). In the farming system birds are free to migrate facilitating the exchange of genetics across the country, favoring the genetic variability occurring in the population thus contributing to its morphological homogeneity irrespectively from the geographical location. The Narragansett breed (NARR) originated in Rhode Island and was recognized as a breed at the end of the 19th century. The NARR was originally developed in Rhode Island by colonies returning to America from Europe in 16th century, bringing back turkeys of the Norfolk Black breed and crossing them with the native American ones (Ekarius, 2007).

In the last 40 years the intensive selection in turkey produced a fast-growing meat bird, a commercial hybrid (HYB). The selection for heavy turkey started presumably in north America and preferred the white pigmentation to other plumage colors (Christman and Hawes, 1999; Ekarius 2007). Birds are selected according to a strong directional mating system to improve weight at slaughter and feed conversion efficiency. The hybrid population here used is a common commercial line of selected heavy turkey (white plumage) that reach in males a weight of 20 kg or more.

Even though the directional selection occurring in European populations for more than 500 years determined that breeds differentiated in morphology and in performances, the European and central American populations share a common genetic background, because their common ancestral origin. This holds true also for commercial turkey line where, nevertheless, the intense directional selection performed in the last 40 years, affected dramatically the physiology, the adult weight, the growth rate, the behavior and the bird's sociality respect to the wild type (EU directive, 2001).

The Copy Number Variants (CNVs) are genomic structural variants recognized to have an active role in gene regulation (Redon et al., 2006; Gamazon and Stranger, 2015) and capable to identify genomic variation among populations. Their use in identifying genomic variation among populations is particularly relevant as several authors found a large proportion (up to 60% in chicken) of mapped CNVs Regions (CNVRs) harboring annotated genes related to expressed phenotypes caused by the specific evolution occurred in the populations (Gorla et al., 2017; Strillacci et al., 2018; Drobik-Czwarno et al., 2018).

The goal of this study is to produce the first CNV map in the Turkey species (*Meleagris gallopavo*) using high density SNP chip information in several populations: the Mexican turkey, the Narragansett, 6 Italian breeds and a commercial hybrid, and to produce a GO analysis of annotated genes in the mapped CNVRs. The strong directional selection occurring in high producing hybrids, the one occurred in the differentiation of the Narragansett and the Italian Turkey breeds, and the adaptive selection in the Mexican turkey population is then discussed according to the genes harbored in the CNVRs. The second goal of this study is to identify the existing variability among the breeds and populations using the mapped CNV, since knowledge of their genomic variation can be used to interpret the phenotypic variability.

Materials and methods

Sampling and SNP chip processing

A total of 115 biological samples from individuals belonging to six Italian breeds (Colle Euganei: CoEu – 22; Bronzato Comune Italiano: BrCI – 5; Parma e Piacenza: PrPc – 15; Brianzolo: BR – 32; Nero d'Italia: NI – 31; Ermellinato di Rovigo: ErRo - 10), 7 Narragansett turkeys (NARR), 38 commercial hybrids (HYB), 30 Mexican turkeys (MEX) were available from previous collections or deriving from other research projects and part of the University of Milan repository of animal samples. The University of Milan permit for the use of collected samples in existing bio-banks was released with n. OPBA-56-2016. The Mexican sample collection is part of the institutional Project "Identificación de los recursos genéticos pecuarios para su evaluación, conservación y utilización sustentable en México. Aves y cerdos. SIGI NUMBER 10551832012" coordinated with the activities of the Centro Nacional of Recursos Genéticos (CNRG) at Tepatitlán, Jalisco (México)[2]. Original owners of sampled individuals gave consent for re-use for research purposes. The study did not require any ethical approval according to national rules, according to EU regulation, as it does not foresee sampling from alive animals.

The samples of the Italian breeds belong to individuals originally collected in different areas of North Italy (Veneto, Lombardia and Emilia Romagna), in nine small farms dedicated to the breeding of one or two breeds each. The MEX individuals were originally sampled across twelve different States of Mexico, characterized by various climatic and geographical environments. The individuals belong to backyards small groups, spread over many small farms. These birds, at best of our

[2] http://www.inifap.gob.mx/SitePages/centros/cnrg.aspx

knowledge, did not undergo any selection by the owners, who let them reproduce according to a natural occurring random mating as they are raised as a backyard population. The Narragansett individuals were originally sampled from two family farms in North Italy A brief description of each turkey population including a picture, the sampling geographical area, the plumage color, the adult body weight and the fertility performance are reported in **Table 1**. The commercial hybrid comes from a unique farm in the Lombardia region in north Italy from the same batch of birds.

DNA extraction from feathers (Mexican samples) and blood (all others) samples were performed using ZR Genomic DNA™ Tissue MiniPrep (Zymo, Irvine, CA, U.S.A.) according to the procedures relative to different tissue. DNA was quantified using NanoQuant Infinite®m200 (Tecan, Männedorf, Switzerland) and diluted to 50 ng/µl. Samples were processed on the Axiom® Turkey Genotyping Array (Affimetrix), containing 634,067 SNPs. The Turkey_5.0 (GCA_000146605.1) genome assembly was used in this study as reference genome.

A quality control of raw intensity files using the standard protocol in the Affymetrix Power Tools package (www.affimetrix.com) was performed in order to guarantee a high quality of obtained data. Default quality control settings, according to the manual (www.affimetrix.com) were applied to filter for low quality samples, i.e. genotyping call rate <98% and Dish Quality Control <0.82.

CNVs detection and subsequent analysis

The Log R Ratio (LRR) and the B allele frequency (BAF) values were obtained using the Axiom® CNV Summary Tool software. Outlier samples for LRR were identified using the SVS 8.4 software (SVS) (Golden Helix Inc., Bozeman, MT, USA) through: i) the overall distribution of Derivative Log Ratio Spread (DLRS) values; ii) screened according to GC content, which is correlated to a long-range waviness of LRR values by the wave detection factor algorithm as in Diskin et al., (2008).

The CNV detection was performed on the data of birds passing quality controls on 30 autosomes, using two different calling algorithms: i) the Copy Number Analysis Module (CNAM) of SVS[3]; ii) the Hidden Markov Model (HMM) of PennCNV software[4]. In order to reduce the false positive calls a consensus map of CNV obtained by the two algorithms was produced.

The CNV calling performed with SVS has been obtained using the univariate analysis based on LRR values, with the following options: univariate outlier removal, a limit of not more than 100 segments per 10,000 markers with a minimum of 3 marker per segment, and 2,000 permutations per pair with a p-value cut off of 0.005, according to the SVS 8.4 user manual.

The PennCNV calling (Wang et al., 2007) was based on LRR and BAF values using the default parameters: standard deviation of LRR <0.30, BAF drift as 0.01 and waviness factor at 0.05 and a minimum of 3 SNP was required to define a CNV. In addition, as to reduce the false calling rate function of the hmm parameter file proper of PennCNV, the CNV call was obtained using three different "hmm" files (agre.hmm, affygw6.hmm, hh550.hmm). The online PennCNV manual describes that the agre.hmm file produces an excess of false positive calls respect to the default affygw6.hmm file (both specific for Affymetrix SNP array), which instead is known to produce a low number of CNV calls (i.e. excess of false negative) respect to other calling software and algorithms. The hh550.hmm file (specifically developed for Illumina SNP arrays) has been considered in the calling process, because is based on a SNPs chip density closest to the one used in this study.

After the four CNVs detections (i.e. one for each hmm file and the one from SVS8.4), the outputs were compared, at individual level and within each population, using the -intersectBed command of Bedtools software (Quinlan and Hall, 2010). For each individual, the consensus_CNVs were defined as the length of the DNA tract full overlapping across at least two detections. CNVs were classified in loss (0 and 1 from the PennCNV output) and in gain

[3] http://goldenhelix.com

[4] http://penncnv.openbioinformatics.org/en/latest/

(3 and 4 from PennCNV output) and were constant across the different callings.

CNV regions (CNVRs) at population level were obtained by merging consensus_CNVs that overlap by at least 1 bp using the -megeBed command of Bedtools (Quinlan and Hall 2010) in at least two birds. The identified CNVRs were classified as "breed_CNVRs" and "shared_CNVRs", when occurred in only one breed (i.e. BR, BrCl, CoEu, ErRo, NI and PrPc) or population (i.e. NARR MEX and HYB), or in at least two ones, respectively. CNVRs were classified within population in gain (all consensus_CNVs gain), loss (all consensus_CNV loss) and complex (consensus_CNVs both gain and loss). Singleton CNVs were considered also to be singleton CNVRs.

Genes were annotated within the CNVRs using the NCBI Turkey_5.0 gene dataset (annotation Release 102) and the Bedtools "-intersectBed" command was used to catalogue these genes to the corresponding regions. Gene Ontology terms (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analyses were performed using the DAVID Bioinformatic Database[5]. Only LOC genes catalogued in NCBI Database as protein genes were considered.

Different approaches were used to disclose population structure and diversification of all turkey population. In order to provide the required input for different analyses two different matrices were built using CNV data: i) the first matrix (matrix_1) was built by assigning a value of "1" (presence of CNV), or "0" (normal state) to each sample-CNV for each CNVR, without considering the CNV state; ii) the second matrix (matrix_2) was built assigning the sample-CNV genotypes: "0" homozygous deletion, "1" heterozygous deletion, "2" normal state (absence of CNV in that region), "3" heterozygous duplication and "4" homozygous duplication. For details see Strillacci et al., (2018).

The Past software (Hammer et al., 2001) was employed to perform and visualize two principal component analyses (PCAs), the first based on the matrix_1 as input data, while the second

[5] https://david.ncifcrf.gov/tools.jsp

based on matrix_2. In addition, two 3D PCAs were performed with the "rgl" package of R[6] on PCAs results. The pvclust R package was utilized using the same matrixes to carry out two Hierarchical Clustering Analyses (HCA) applying 10,000 bootstraps (Suzuki and Shimodaira, 2006).

The STRUCTURE Software v.2.3.4 (Pritchard et al., 2002; Falush et al., 2003) was used to represent the population structure of the populations studied, on the basis of matrix_1. We used the STRUCTURE admixture model without the LocPrior option and setting 5,000 as burning period and 10,000 as iterations, performing five replicates for each K value from 2 to 20 and assuming nine different populations. Structure Harvester software (Dent and vonHoldt, 2012) was used to obtain the best K values, on the basis of STRUCTURE results, providing the DeltaK values according to the heuristic method reported by Evanno et al. (2005). The STRUCTURE PLOT software (Ramasamy et al., 2014) was employed to graphically visualize each cluster assignment of the K obtained.

Results

CNVs and CNVRs maps

A total of 13 samples (5 NI, 2 PrPc, 4 MEX, 1 ErRo and 1 BR) were excluded during quality assurance: three because of high DLRS values, seven because wave factor values, and three for their exceptionally high number of called CNVs. Consequently, the final CNV dataset used for genomic variation analyses comprised a total of 177 turkeys.

The total number of CNVs called was 2,987 (**Supplementary Table 1**) and varied in terms of number and size among the individuals of each population, as reported in **Table 2**. CNVs ranged from 819 bp to 453.5 kb in size with an average length of 115.2 kb, covering a total length of about 41 Mb (4.65%) of the turkey genome (chromosomes 1-30). The BrCl and the HYB shown shorter average CNVs respect to other populations, while

[6] https://CRAN.R-project.org/package=rgl

in the MEX one the longest average CNVs length was found. The HYB birds are also the most homogeneous for the average length of CNVs (**Figure 1A**). The MEX breed is the one with the largest number of CNVs per individual (i.e. 28) while the HYB is the one with the lowest (i.e. 10).

Duplications were higher than deletions in the majority of populations except for BrCI, CoEu and NARR breeds, where the ratio gain/loss (losses are the sum of the total copy numbers 0 and 1; gains are the sum of the total copy numbers 3 and 4) are inverted as showed in **Figure 1B**. The gain/loss ratio is similar in HYB, MEX, NI, and PrPc populations (about 65% vs. 35%), instead the proportion of duplication and deletion are differently represented in the other populations. The CNVRs including at least 2 individuals were 362 counting 189 gains, 116 losses and 57 complexes and their distribution on the chromosomes is shown in **Figure 1C**.

Statistics of CNVRs for each population are reported in **Table 3**. A total of 1,659 CNVRs (OverAll) were obtained across all populations with 412 loss, 1,190 Gain and 57 Complex.

Details on CNVRs are reported in the **Supplementary Table S2** for those including at least two individuals and detected across breeds, i.e. shared_CNVRs. The 1,297 singleton CNVRs, representing 64% of all detected ones, are listed in the **Supplementary Table S3.** The **Supplementary Figure S1** is showing the distribution of singleton among breeds/populations and the distribution of loss and gains across all populations and by breed/population. The largest proportion of CNVRs resulted to be gain, i.e. 77% across all breeds/populations, with a proportion of singletons of 64%. This result is consistent with the proportion of singleton identified in chickens by others (Gorla et al., 2017; Yi et al., 2014).

The Venn Diagram (Heberle et al., 2015) shown in **Figure 2 A** represents the amount of CNVRs shared among the populations, grouping them as ITA (all Italian breeds), NARR (the Narragansett), MEX (the Mexican turkey population) and HYB (the commercial cross). The reason of this grouping resides in the type of evolution of the populations: the Italian breeds are all highly selected for breed standard phenotypes and possibly highly inbred; the Mexican population has been under an outbreeding mating system, with no directional selection undertaken for centuries; the NARR is a cross between the wild American turkey and the US domestic Bronze turkey; the HYB is a commercial population obtained by a strong directionally selection for heavy body weight. Three CNVRs resulted common to all populations and a large proportion of ITA CNVRs are shared with MEX and HYB, 65 and 42 CNVRs respectively.

In **Table 4** the details of the thirty-two CNVRs detected in at least ten samples and the genes laying in the same regions are reported. Among those, the three regions in common to all turkey populations, as shown also in **Figure 2A** are located on chr3 at 92,889,953 – 92,936,492 (CNVR_1126, gain), on chr4 at 26,993 – 164,704 (CNVR_1240, gain) and on chr4 at 68,446,449 – 68,522,752 (CNVR_1371, complex). In the CNVR_1371, the one also found in the largest number of individuals from all breeds, is annotated the *CD8A* gene that is related to immune and inflammatory response (Li et al., 1999). In the other two common regions, CNVR_1126 and the CNVR_1240, the *FK1L* and the *TLR2A* gene are annotated. respectively involved in immune and inflammatory response and in feather keratin multigene family with implication in feather evolution (Li et al., 2013; Velová et al., 2018).

Other two regions are shared by a large number of individuals of ITA breeds and have been detected on chr4 at 63,830,569 – 63,854,111 in CNVR_1357 (62 birds from ITA breeds) and CNVR_1358 (65 birds from ITA breeds). These two regions are both a loss, are very close on the genome being 13,382 bp apart and have been detected in almost the same samples of the same ITA breeds. No genes are annotated within these two CNVRs. Ten CNVRs in **Table 4** are common to ITA and MEX, 5 common to ITA and HYB and only 1 in common between ITA and NARR. Among these regions 9 of them including genes (CNVR_163, CNVR_1243, CNVR_1246, CNVR_1598, CNVR_488, CNVR_644, CNVR_987, CNVR_1025). There are no regions shared only among HYB, NARR and MEX.

The Venn Diagram in **Figure 2B** shows in detail the distribution of CNVRs among the six Italian breeds. It is worthy of mention that the gene *CD8A* is in a CNVR common to all the Italian breeds (in the red circle).

Among the 362 CNVRs a total of 140 mapped only in one specific population, the breed_CNVRs, as reported in **Supplementary Table S4**. The mapped genes in any species and the corresponding references for each association studies, the associated phenotypes and the organism involved are also indicated.

The largest number of breed_CNVRs occurred in the MEX turkey population with 45 regions followed by the NI with 33. The lowest number of breed_CNVRs was found in the BrCI and in the NARR with 1 and 4 breed_CNVRs respectively. The number of genes annotated in the breed_CNVRs was 26 and 21 in MEX and NI, while the number of genes in breed_CNVR in other other populations was between 1 and 8. The gene *IMMPL2* is harbored by 2 breed_CNVRs, one in the BI (CNVR_69) and one in the NI (CNVR_70). The two regions are very close even if they do not overlap.

The results of the GO TERM and KEGG pathway analyses obtained using DAVID considering the genes found in the 362 shared_CNVRs are reported in the **Supplementary Table S5** into clustered and not clustered groups of genes.

The **Supplementary Table S6** contains the information generated from the KEGG and GO Term analysis using DAVID from breed_CNVRs. The information was obtained using *Meleagris gallopavo* as background species and integrated and confirmed using the *Gallus gallus* as background, in case of absence of complete information for the *Meleagris gallopavo* species.

Genetic Variability across turkey populations

Two clustering analysis were performed based on two different matrixes (matrix_1 and matrix_2) described hereinbefore. Both the cluster dendrograms, **Figure 3A** based on matrix_1 and **Figure 4A** based on matrix_2, showed distinct clades grouping animals belong to the same populations. It is interesting to note that MEX and NARR always clustered very close. Also, Italian breeds and the Hybrid group form well distinct clusters according to their origin.

In all the PCAs graphs in **Figure 3B 3C**, **4B** and **4C** the clustering results show two main clades: NARR, MEX and HYB were grouping closer, while the ITA breeds clustered in a separate one.

The STRUCTURE software was employed to infer population structure and gene flow of the individuals of the 9 populations studied. We calculate a number of K from K=2 to K=20 to identify the true number of possible clusters (subpopulation) in which is possible to divide the populations. The estimated likelihood (LnP (D)) values were used to find the ΔK to distinguish the break in slope of the distribution of LnP (D) values at the true K. The analyses identify K=13 the best likely K value, suggesting that the population could be divided into 13 genetic groups.

Even though K=5 show the second higher value (**Supplementary Figure S2)** it is not possible to well differentiate the populations as in K =13. In fact, for K=2 to K=12 it is not possible to assign each population to a clear distinct cluster, while for K=14 to K=20 the high level of admixture in each of the population result in not significant successive clustering.

Discussions

The results from this study are likely reflecting the human action on turkey populations, i.e. its migration to Europe and then back to America, and the directional selection occurring in the last 40 years to produce a fast-growing heavy bird.

The study considers three main groups of birds that reproduce and adapt according to different constrains and environmental conditions. The MEX population developed in a natural environment, with no (or very little) intervention by humans in mating and with no (or very little) supplement of feed and harsh rearing conditions. The Italian populations are the result of a phenotypic selection operated by individual farmers in their small group of individuals and operated to obtain birds that best perform in the semi-extensive farming system (backyard with recovery availability and feeding supplement) that characterized the middle ages poultry system of Italy and Europe. The HYB population, in the last 40 years, has been heavily directionally selected, through a very well-structured genetic improvement and breeding plans to improve weight and growing performances and to best perform in an artificially controlled environment with unlimited feed supplement.

Our study is the first CNV mapping in a worldwide turkey sampling, from populations collected across different continents, and disclosed similarities and variation in CNVs and CNVRs across the populations studied. Because of the diversity in their breeding history and actual farming environmental conditions the MEX, ITA and HYB populations provide an interesting model to investigate CNV variation, and their relation to gene expression and rearing conditions. The CNV, in fact, are widely recognized to be a non-neutral genomic structural variation related to positive and directional selection. The CNV has been recently successfully used in poultry to differentiate breeds and populations with different genetic background (Gorla et al., 2017; Strillacci et al., 2017; Sohrabi et al, 2018), as well as in other species (Xu et al. 2016; Strillacci et al. 2018). Interestingly in chicken Sohrabi et al. (2018) discuss long-term adaptation of animals to rural and hard rearing conditions in relation to a specific expressed trait linked to a CNV identified in the Creeper indigenous chicken local population that is adapted to the harsh environmental condition of southeastern Iran. Additionally, a recent study on a eukaryotic model (Hull et al., 2017) showed that environmental changes are accelerating adaptation through the stimulation of copy number variation and that this is not a random effect but has a cause effect relationship. Perry et al. (2007) also demonstrated that directional selection due to starch diet (i.e. environmental factor) is increasing specific copies of the genes involved in starch metabolism producing as such CNV gains. The CNV difference among populations is here shown in particular by the variation in the number of CNV per bird that is the lowest in the HYB (10 on average) and the largest in the MEX (28 CNV) and by the CNV length that in the HYB is much less variable than in the other two group (ITA and MEX) of birds. These findings support the hypothesis that the variability in CNV (size and number), as in the MEX vs. the HYB, is possibly related to the different breeding and selection underwent in these populations and to the environmental conditions where they were farmed: MEX very harsh rearing one, HYB controlled artificial environment and ad libitum feeding. The same holds true for the ITA vs. MEX and HYB.

Most of the genes found do not show previous associations with any specific function or pathway in turkey, since associations studies in turkey are only a few, but most of those genes have been previously studied and linked to functions in other species such as chicken, pig, bovine, birds, mice, zebrafish and human, as reported in **Supplementary Table 4**.

Thirty-two regions were detected in at least 10 individuals, and 14 of them include 29 genes, that are known to be involved in different traits in different species (**Table 4**), such as immune response (*TLR2A and CD8A),* feather evolution (*FK1L*), feed efficiency (*PRKG1* and *LMAN1),* growth traits (*TCF15, FAM110A*) and residual feed intake (*TACC1, PLEKHA2, TM2D2, ADAM9, IDO2, C24H8orf4, ZMAT4*), as reported in **Table 4**.

There are three CNVRs in common among all the populations; one of them harbors the *CD8A* gene, which is known to have a role in the host immune and inflammatory response in chicken (Li et al., 1999). The polymorphism of the *CD8A* gene has been studied in 5 lines of turkey populations by Li et al. (1999) who found a loss of this gene in one half of the turkey of a studied line. This loss can be related to the CNVR_1371 found in this study where 34 CNVRs were loss and 34 gain. All the ErRo resulted to have a loss, CoEu had 12 loss (over 13 birds), BrCI 4 loss (over 5 birds), while other populations have a more balanced representation between loss and gain CNVs.

The *TLR2A* gene has been shown to be involved in the bird's evolution with a strong driving of TLR due to positive selection

(Velova et al., 2018). It is interesting to note that our results show that CNVR_1240 include the *TLR2A* gene with only normal and gain state. Even if the question of the adaptive value of the TLR genetic variation is still unresolved the results found here are supporting the hypothesis that positive selection is driving the evolution of the gene towards duplication of copies as proposed recently by Velova et al., (2018).

Other genes in the CNVRs here found (**Supplementary Table S4**) are associated with immunity and inflammatory response in mice (*TCF7, ARHGEF5*), chicken (*VMO1, GUCY1A2, NBN*), bovine (*NEK11*) and in all species (*PARP15*) as reported in previous studies (Velova et al., 2018; Zhu et al., 2015; Wang et al., 2009; Lim and Song 2015; Saelao et al., 2018; Jang et al., 2015; Strillacci et al., 2014; Daugherty et al., 2014). Among the genes reported in the **Supplementary Table S4**, the *IMMP2L* gene lies in CNVR_69 which is common to NI and BR. This gene was associated with fertility in mice (Bharadwaja et al., 2014) and with collective behavior in zebrafish (Tang et al., 2018). The presence of this gene in a gain CNVR may have some link with the typical collective behavior of the turkey.

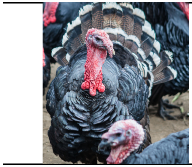
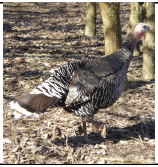
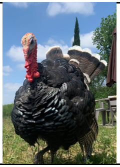
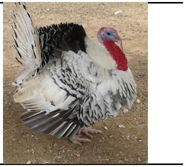
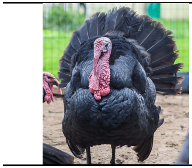
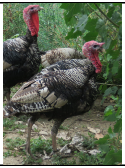
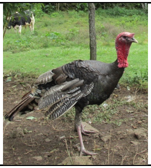
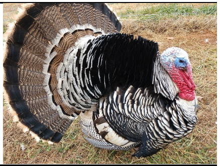
Conclusions

This study represents the first CNV mapping using high density SNP chip on turkey. It provides a first insights into the genomic architecture of the turkey population, laying the groundwork for future structural variation investigation in turkey species. In this study we have focused on the CNV, a structural variation linked to phenotypic expression regulation, in order to identify similarities across populations of the structural genome covered by this large variation.

The turkey populations are a unique resource to identify evolutionary process affecting the structural genome since it is possible to access to populations under positive selection only and, on the other extreme, under heavy artificial selection. The most complete isolation of the MEX turkey population and the European ones together to the HYB provide a unique model to disclose the effect of the adaptation to environment and directional artificial selection performed by humans on the structural genome.

Tables

Table 1. Population name, sampling area, weight (kg) and plumage color of the turkey populations considered in the study.

Brianzolo (BR)*	Bronzato Comune Italiano (BrCI)*	Colle Euganei (CoEu)**	Ermellinato di Rovigo (ErRo)**
			
Origin Area: North Italy (Lombardia) **Weight (Kg):** F:2.1-3.2; M: 4.5-6.4 **N. eggs/year:** 47 **Fertility:** 77-78% **Plumage:** Black, bronzed, reticulated gray (common), bronzed with white wings. **Description:** Early and disease-resistant bird. Rural breeding, numerical consistency extremely small.	**Origin Area**: North-East Italy (Veneto) **Weight (Kg):** F:3-3.5; M: 6-7 **N. eggs/year:** 70-100 **Fertility:** 92-93% **Plumage:** brilliant black with intense bronze reflections. **Description:** Rustic breed with a strong hatching attitude. Breeding in local areas.	**Origin Area**: North-East Italy (Veneto) **Weight (Kg):** F:3; M: 5 **N. eggs/year:** N/A **Fertility:** N/A **Plumage:** bronzed with metallic reflections. **Description:** Rustic breed with a strong hatching attitude. Local breeding, numerical consistency extremely small.	**Origin Area**: North-East Italy (Veneto) **Weight (Kg):** F:4-6; M: 10-12 **N. eggs/year:** 70-80 **Fertility:** 86-92% **Plumage:** white with black streaks. **Description:** Rustic breed with slow growing excellent grazers. Breeding in local areas.
Nero Italiano (NI)*	**Parma e Piacenza (PrPc)****	**Mexican (MEX)*****	**Narragansett (NARR)**; §**
			
Origin Area: North Italy (Lombardia) **Weight (Kg):** F:2.1-3-9; M: 4.9-7.1 **N. eggs/year:** 41 **Fertility:** 84-85% **Plumage:** Black. **Description:** Rustic breed with a strong hatching attitude Breeding in local areas.	**Origin Area**: North Italy (Emilia Romagna) **Weight (Kg):** F:6.5; M: 12 **N. eggs/year:** N/A **Fertility:** N/A **Plumage:** Steel gray with white streaks. **Description:** Local breeding, numerical consistency extremely small.	**Origin Area**: Mexico **Weight (Kg):** F: 3.2; M: 5.7 Kg **N. eggs/year:** N/A **Fertility:** N/A **Plumage:** Different colors. **Description:** Backyard birds. Unselected extremely variable in term of phenotype and production.	**Origin Area**: Rhode Island (USA) **Weight (Kg):** F: 8.2; M: 15 Kg **N. eggs/year:** N/A **Fertility:** N/A **Plumage:** Steel gray color. **Description:** Breeding in Europe and locally in Italy.

*Data from https://www.pollitaliani.it/portfolio-articoli/razze/: **Data from: http://www.agraria.org/tacchini/neroitalia.htm; *** Data from: Utrera et al., (2016);
§Picture from: https://commons.wikimedia.org/wiki/File:Narragansett_Turkey,_male.jpg

Table 2. Summary of CNVs identified in each population.

Breed	N. of samples	N. CNVs	CNV per sample Min-Max (average)	Loss (0/1)*	Gain (3/4)*	Min length	Max length	Mean length	Coverage	Total Coverage (%)
BR	31	412	4-34 (13)	185	227	1,221	214,517	15,715	6,474,485	0.73
BrCl	5	63	6-24 (12)	38	25	1,271	25,586	7,357	463,483	0.05
CoEu	22	354	8-37 (16)	191	163	1,096	184,966	11,762	4,163,692	0.47
ErRo	9	135	8-30 (10)	53	82	1,221	362,781	11,569	1,561,859	0.18
NI	26	567	6-69 (22)	192	375	1,096	283,259	12,436	7,038,934	0.8
PrPc	13	232	7-42 (18)	85	147	1,328	230,199	16,307	3,783,129	0.43
NARR	7	96	10-22 (14)	51	45	1,301	83,743	13,105	1,258,113	0.14
MEX	26	734	12-49 (28)	245	489	819	453,485	16,979	12,462,363	1.41
HYB	38	394	4-20 (10)	128	266	1,070	62,316	9,964	3,935,744	0.45
Total	**177**	**2,987**	**4-69 (17)**	**1,168**	**1,819**	**819**	**453,485**	**115,194**	**41,141,802**	**4.65**

*0=homozygous deletion, 1=heterozygous deletion, 3=heterozygous duplication, and 4=homozygous duplication

Table 3. Summary of CNVRs identified for each turkey's population.

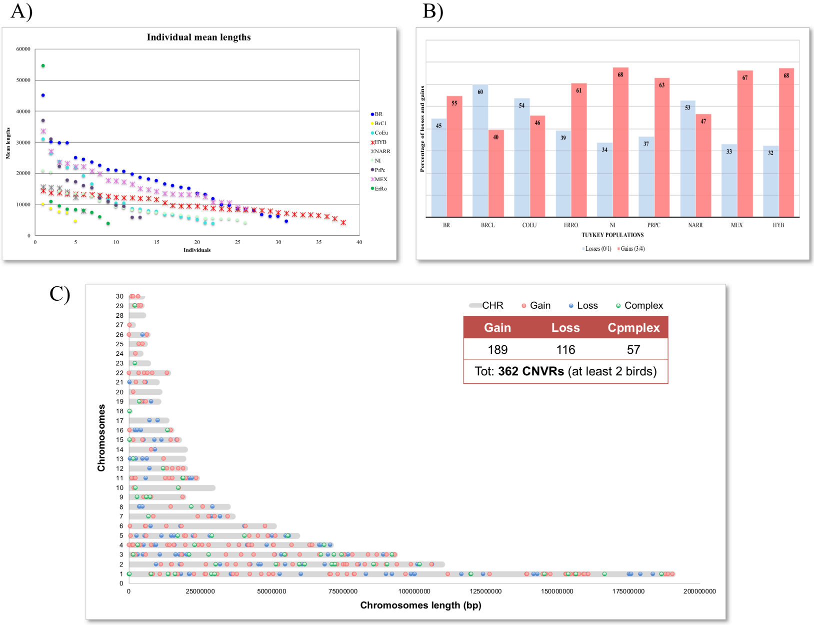
Breed	CNVR	Loss	Gain	Complex	Min length	Max length	Mean length	Coverage	Total Coverage (%)
BR	223	53	168	2	1,221	214,517	12,293	2,741,386	0.31
BrCl	47	24	23	0	1,383	25,586	7,063	331,977	0.04
CoEu	195	56	138	1	1,096	186,030	10,542	2,055,612	0.23
ErRo	108	79	29	0	1,221	362,781	12,634	1,364,494	0.15
NI	358	58	293	7	1,096	283,259	15,186	5,436,564	0.62
PrPc	186	59	126	1	1,328	230,199	14,029	2,609,445	0.30
NARR	77	39	38	0	1,301	83,743	11,494	885,013	0.10
MEX	575	185	385	5	843	453,485	15,864	9,122,023	1.03
HYB	243	59	181	3	1,070	62,316	8,830	2,145,688	0.24
OverAll	**1,659**	**412**	**1190**	**57**	**843**	**453,485**	**13,612**	**22,581,871**	**2.55**

Table 4**.** List of the CNVRs mapped in at least 10 birds with chromosome, start bp, end bp, CNVR length and CNVR state. For each of the CNVRs the count of birds for each population (ITA, NARR, MEX HYB) is reported together with their total. The genes annotated in each region are listed with the trait of interest and the reference.

N_CNVR	Chr	CNVR	CNVR	CNVR			ITA				NARR	MEX	HYB	Total Samples	CNVR state	Genes	Trait by gene: (species)	References
		start	end	length	BR	BrCl	CoEu	ErRo	NI	PrPc								
CNVR_113	1	46402671	46430314	27643						1			17	18	gain			
CNVR_126	1	52847470	52853786	6316					9				4	13	loss			
CNVR_163	1	76320966	76430128	109162	6				8	7		2		23	gain	*OVSTL, TCRb1*	OVSTL: eggshell calcified layer (quail)	Mann and Mann, 2015
CNVR_206	1	98886764	98931838	45074						2			17	19	loss			
CNVR_210	1	99904908	99927304	22396	9		1			1			1	12	loss			
CNVR_307	1	145466178	145680695	214517	9				1				1	11	complex			
CNVR_757	2	30461083	30521978	60895			9					4		13	complex			
CNVR_780	2	42604981	42606860	1879	9	1		2				1		13	loss			
CNVR_809	2	57899261	57923296	24035					1		5	4		10	complex			
CNVR_843	2	72167387	72173022	5635	10		4		4					18	loss			
CNVR_920	2	101084671	101088748	4077		1		2	6		1		4	14	loss			
CNVR_1088	3	20396386	20399251	2865	10		18			11				39	loss			
CNVR_1152	3	54655570	54693060	37490					1			2	10	13	complex	*OPN5L1*		
CNVR_1226	3	92889953	92936492	46539	1	1	1			1	1	6	4	15	gain	*FK1L*		
CNVR_1240	4	26993	164704	137711	2		1		3	2	2	5	3	18	gain	*TLR2A*	host immune response (Birds)	Velová et al., 2018
CNVR_1243	4	1581791	1620844	39053						2			9	11	gain	*GRIA2*		
CNVR_1246	4	3011587	3071312	59725	5			2		1		4		12	complex	*FSTL5*		
CNVR_1259	4	8948522	8954649	6127	12				5					17	loss			
CNVR_1357	4	63830569	63837531	6962	21		19		24	1				65	loss			
CNVR_1358	4	63850913	63854111	3198	19		19		23	1				62	loss			
CNVR_1371	4	68446449	68522752	76303	4	4	13	7	13	2	2	7	16	68	complex	*CD8A*	host immune and inflammatory response (Poultry)	Yi et al., 2014
CNVR_1408	5	15840153	15842835	2682	3		10		1			4		18	loss			
CNVR_1586	7	28038559	28062433	23874	2		1	1	2	2		5	2	15	gain			
CNVR_1598	8	3846585	3850061	3476		1	8	1		2		2		14	loss	*PRKG1*	feeding efficiency (bovine)	Taye et al., 2017
CNVR_465	11	1004126	1053713	49587	1		2	1	2			6		12	gain	HNRNPL		
CNVR_488	11	18985991	19015763	29772	6		1	1	3		1			12	complex	*LMAN1, CPLX4*	LMAN1: feed efficiency and feeding behavior (pig)	Reyer et al., 2017
CNVR_644	16	4206442	4209316	2874	12		11					1		24	loss	*GRIN2A*		
CNVR_970	21	5878926	5903943	25017		5	3	2						10	loss			
CNVR_987	22	5386977	5429908	42931	2		1	1	2				4	10	gain	*SLC52A3, RSPO4, SRXN1*		
																TCF15	growth (bovine)	Paredes-Sánchez et al., 2015
																FAM110A	growth (human, bovine)	Espigolan et al., (2015)
																ANGPT4	birth weight (human)	Turan et al., (2012)
																SCRT2	self- reported helping behavior (human)	Primes and Fieder (2018)
CNVR_1003	24	2359444	2545474	186030	1		12							13	gain	*TACC1, PLEKHA2, TM2D2, ADAM9, IDO2, C24H8orf, ZMAT4*	(all genes) residual feed intake (bovine)	Hardie et al., (2017)
CNVR_1024	26	6388747	6431016	42269			1		2			19		22	gain			
CNVR_1025	27	157671	192870	35199			1		5			10		16	gain	*VPS45, NUPL2*		

Figures

Figure 1. Graphical representation of identified CNVRs. A) Distribution of Individual mean length for each population; B) Percentage of losses and gains CNVRs in each population; C) Map of CNVRs in the autosomes according with states.



Gain	Loss	Cpmplex
189	116	57
Tot: **362 CNVRs** (at least 2 birds)		

Figure 2. Venn diagrams of CNVRs identified: A) in turkeys grouped according to ITA-breeds; NARR; MEX and HYB; B) in the six Italian turkey breeds.

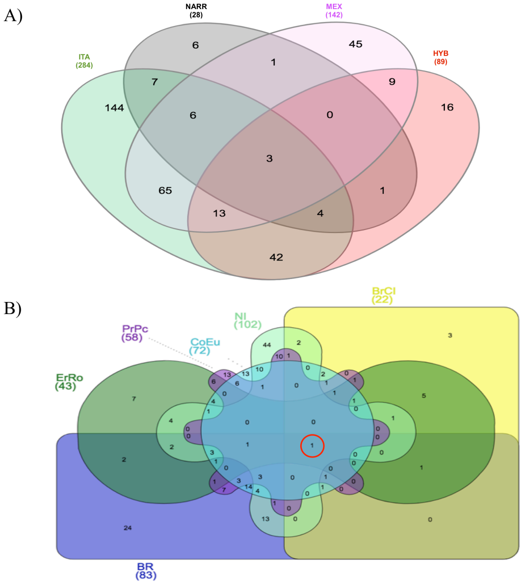


Figure 3. Hierarchical clustering and PCAs based on CNVRs (CNV encoded as in matrix_1). A), B) and C) are the dendrogram, the PCA-2D and the PCA-3D, respectively.

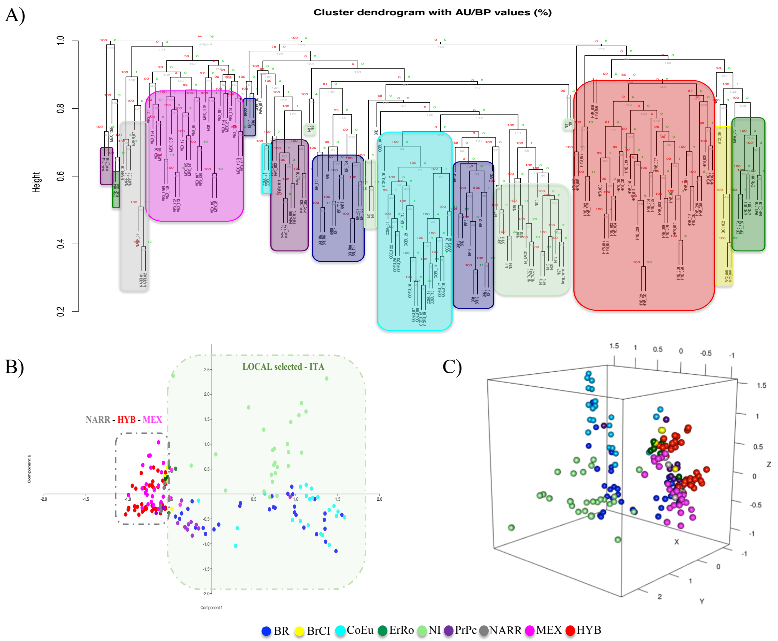
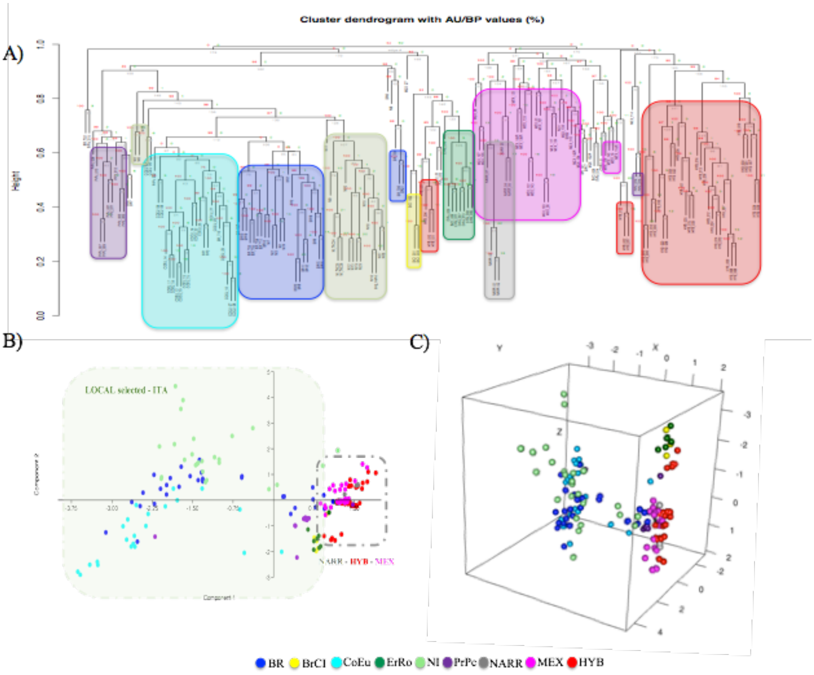


Figure 4. Hierarchical clustering and PCAs based on CNVRs (CNV encoded as in matrix_2). A), B) and C) are the dendrogram, the PCA-2D and the PCA-3D, respectively.



Supporting information

All supplementary files are available at:

https://doi.org/10.3389/fgene.2019.00982

References

- Bharadwaj, M.S., Zhou, Y., Molina, A.J., Criswell, T., Lu, B. (2014). Examination of bioenergetic function in the inner mitochondrial membrane peptidase 2-like (Immp2l) mutant mice. Redox Biol. 2:1008-15. doi: 10.1016/j.redox.2014.08.006.
- Boije, H., Harun-Or-Rashid, M., Lee, Y.J., Imsland, F., Bruneau, N., Vieaud, A., et al. (2012). Sonic Hedgehog-signalling patterns the developing chicken comb as revealed by exploration of the pea-comb mutation. PLoS One. 7(12):e50890. doi: 10.1371/journal.pone.0050890.
- Cavalchini L.G. (1983). IL TACCHINO allevamento, incubazione, patologia. 1° Edition, 304 pages. Edagricole.
- Condro, M.C., White, S.A. (2014). Recent advances in the genetics of vocal learning. Comp. Cogn. Behav. Rev. 9:75-98. doi: 10.3819/ccbr.2014.90003.
- Crawford, R.D. (1992). Introduction to Europe and diffusion of domesticated turkey from the America. Arch. Zootec. 41:307-314.
- Daugherty, M.D., Young, J.M., Kerns, J.A., Malik, H.S. (2014). Rapid evolution of PARP genes suggests a broad role for ADP-ribosylation in host-virus conflicts. PLoS Genet. 10(5):e1004403. doi: 10.1371/journal.pgen.1004403.
- Day, A.E., Quilter, C.R., Sargent, C.A., Mileham, A.J. (2002). Characterization of the porcine sperm adhesion molecule gene SPAM1–expression analysis, genomic structure, and chromosomal mapping. Anim Genet. 33(3):211-4.
- Dent, E.A. and vonHoldt, B.M. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Cons. Genet. Res. 4(2): 359–361. doi: 10.1007/s12686-011-9548-7.

- Diskin, S.J., Li, M., Hou, C., Yang, S., Glessner, J., Hakonarson, H., Bucan, M., Maris, J.M., Wang, K. (2008) Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. Nucleic. Acids. Res. 36:e126. doi: 10.1093/nar/gkn556.
- Drobik-Czwarno, W., Wolc, A., Fulton, J., Dekkers, J.C. (2018) Detection of copy number variations in brown and white layers based on genotyping panels with different densities. Genet. Sel. Evol. 50(1):54. doi: 10.1186/s12711-018-0428-4.
- Ekarius, Carol. Storey's Illustrated Guide to Poultry Breeds by Carol Ekarius. Storey Publishing, LLC; 1 edition (May 30, 2007), ISBN-978-1580176675.
- Espigolan R., Baldi F., Boligon A.A., Souza F.R., Fernandes Júnior G.A., Gordo D.G. et al. (2015). Associations between single nucleotide polymorphisms and carcass traits in Nellore cattle using high-density panels. Genet. Mol. Res.14(3):11133-44. doi: 10.4238/2015
- Evanno, G., Regnaut, S., Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. Mol. Ecol. 14(8):2611–2620.
- Fang, M., Du H., Hu, Y., Zhou, X., Ouyang, H., Zhang, W., et al. (2011). Identification and characterization of the pig ABIN-1 gene and investigation of its association with reproduction traits. J. Genet. 90(1):e10-20.
- Falush, D., Stephens, M., Pritchard, J.K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics. 164(4):1567-87.
- Gorla, E., Cozzi, M.C., Román-Ponce, S.I., Ruiz López, F.J., Vega-Murillo, V.E., Cerolini, S., et al. (2017). Genomic variability in Mexican chicken population using copy number variants. BMC Genetics. 18(1):61. doi: 10.1186/s12863-017-0524-4.

- Gamazon, E.R., Stranger, B.E. (2015). The impact of human copy number variation on gene expression. Brief. Funct. Genomics. Sep;14(5):352-357.
- Hammer, Ø., Harper, D.A.T., Ryan, P.D. (2001). PAST: Paleontological statistics software package for education and data analysis. Palaeontologia Electronica (2001); 4(1):9. http://palaeo-electronica.org/2001_1/past/issue1_01.htm.
- Hardie, L.C., VandeHaar, M.J., Tempelman, R.J., Weigel, K.A., Armentano, L.E., Wiggans G.R., et al. (2017). The genetic and biological basis of feed efficiency in mid-lactation Holstein dairy cows. J. Dairy. Sci. 100(11):9061-9075. doi: 10.3168/jds.2017-12604.
- Heberle, H., Meirelles, G.V., da Silva, F.R., Telles, G.P., Minghim, R. (2015). InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. BMC Bioinformatics. 16:169. doi: 10.1186/s12859-015-0611-3.
- Hull, R.M., Cruz, C., Jack, C.V., Houseley. J. (2017) Environmental change drives accelerated adaptation through stimulated copy number variation. PLoS Biol. 15(6):e2001333. doi: 10.1371/journal.pbio.2001333.
- Jang, H.J., Lee, H.J., Kang, K.S., Song, K.D., Kim, T.H., Song, C.S., et al. (2015). Molecular responses to the influenza A virus in chicken trachea-derived cells. Poult. Sci. 94(6):1190-1201. doi: 10.3168/jds.2017-12604.
- Letunic, I. and Bprk, P. (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 44(W1):W242-5 doi: 10.1093/nar/gkw290.
- Li, Y.I., Kong, L., Ponting, C.P., Haerty, W. (2013). Rapid evolution of Beta-keratin genes contribute to phenotypic differences that distinguish turtles and birds from other

reptiles. Genome Biol. Evol. 5(5):923-33. doi: 10.1093/gbe/evt060.

- Li, Z., Nestor, K.E., Saif, Y.M,. Fan Z., Luhtala M., Vainio O. (1999). Cross-reactive anti-chicken CD4 and CD8 monoclonal antibodies suggest polymorphism of the turkey CD8alpha molecule. Poult. Sci. 78(11):1526-31.
- Lim, W. and Song, G. (2015). Differential expression of vitelline membrane outer layer protein 1: hormonal regulation of expression in the oviduct and in ovarian carcinomas from laying hens. Mol. Cell Endocrinol. 399:250-8. doi: 10.1016/j.mce.2014.10.015.
- Lovell, P.V., Carleton, J.B. and Mello, C.V. (2013). Genomics analysis of potassium channel genes in songbirds reveals molecular specializations of brain circuits for the maintenance and production of learned vocalizations. BMC Genomics. 14:470. doi: 10.1186/1471-2164-14-47.
- Mann, K. and Mann, M. (2015). Proteomic analysis of quail calcified eggshell matrix: a comparison to chicken and turkey eggshell proteomes. Proteome Sci. 13:22. doi: 10.1186/s12953-015-0078-1.
- Paredes-Sánchez, F.A., Sifuentes-Rincón, A.M., Cabrera, A.S., Pérez, C.A.G., Bracamonte, G.M.P., Morales, P.A. (2015). Associations of SNPs located at candidate genes to bovine growth traits, prioritized with an interaction networks construction approach. BMC Genetics. 16:91. doi: 10.1186/s12863-015-0247-3.
- Pelz, L., Purfürst, B., Rathjen, F.G. (2017). The cell adhesion molecule BT-IgSF is essential for a functional blood–testis barrier and male fertility in mice. J. Biol. Chem. 292(52):21490-21503. doi: 10.1074/jbc.RA117.000113.
- Perry, G.H., Dominy, N.J., Claw, K.G., Lee, A.S., Fiegler, H., Redon,

R., et al. (2007). Diet and the evolution of human amylase gene copy number variation. Nat. Genet. 39(10):1256-60. DOI: 10.1038/ng2123.

- Pickrell, J. K. & Pritchard, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. 8(11):e1002967. doi: 10.1371/journal.pgen.1002967
- Pinto, D., Darvishi, K., Shi, X., Rajan, D., Rigler, D., Fitzgerald, T., et al. (2011). Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nat Biotechnol. 29(6):512-20. doi: 10.1038/nbt.1852.
- Primes, G. and Fieder, M. (2018). Real-life helping behaviours in North America: A genome-wide association approach. PLoS One. 13(1):e0190950. doi: 10.1371/journal.pone.0190950
- Pritchard, J.K., Stephens, M. and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics. 155(2):945-59.
- Ramasamy, R.K., Ramasamy, S., Bindroo, B.B., Naik, V.G. (2014). STRUCTURE PLOT: a program for drawing elegant STRUCTURE bar plots in user friendly interface. Springerplus. 3:431 doi: 10.1186/2193-1801-3-431
- Redon R., Ishikawa S., Fitch K.R., Feuk L., Perry G.H., Andrews T.D., et al. (2006). Global variation in copy number in the human genome. Nature. 444(7118):444-54. doi: 10.1038/nature05329.
- Reich, D., Thangaraj, K., Patterson, N., Price, A.L., Singh, L. (2009). Reconstructing Indian population history. Nature. 461(7263), 489-94. doi: 10.1038/nature08365.
- Reyer, H., Shirali, M., Ponsuksili, S., Murani, E., Varley, P.F., Jensen, J., et al. (2017). Exploring the genetics of feed efficiency and feeding behaviour traits in a pig line highly selected for

performance characteristics. Mol. Genet. Genomics. 292(5):1001-1011. doi: 10.1007/s00438-017-1325-1

- Quinlan, A.R., Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26(6):841-2. doi: 10.1093/bioinformatics/btq033
- Saelao, P., Wang, Y., Gallardo, R.A., Lamont, S.J., Dekkers, J.M., Kelly, T., et al. (2018). Novel insights into the host immune response of chicken Harderian gland tissue during Newcastle disease virus infection and heat treatment. BMC Vet. Res. 14(1):280. doi: 10.1186/s12917-018-1583-0
- Strillacci, M.G., Cozzi, M.C., Gorla, E., Mosca, F., Schiavini, F., Román-Ponce, S.I., et al. (2017). Genomic and genetic variability of six chicken populations using single nucleotide polymorphism and copy number variants as markers. Animal. 11(5):737-745. doi: 10.1017/S1751731116002135.
- Strillacci, M.G., Frigo E., Schiavini F., Samoré A.B., Canavesi F., Vevey M., et al. (2014). Genome-wide association study for somatic cell score in Valdostana Red Pied cattle breed using pooled DNA. BMC Genetics 15:106. doi: 10.1186/s12863-014-0106-7.
- Strillacci, M.G., Gorla, E., Cozzi, M.C., Vevey, M., Genova, F., Scienski, K., et al. (2018). A copy number variant scan in the autochthonous Valdostana Red Pied cattle breed and comparison with specialized dairy populations. PLoS One 13(9):e0204669. doi: 10.1371/journal.pone.0204669.
- Suzuki, R., Shimodaira, H. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics. 22(12):1540-2. doi: 10.1093/bioinformatics/btl117.
- Sohrabi, S.S., Mohammadabadi, M., Wu, D.D., Esmailizadeh, A. (2018). Detection of breed-specific copy number variations in

domestic chicken genome. Genome. 61(1):7-14. doi: 10.1139/gen-2017-0016.

- Tang, W., Zhang, G., Serluca, F., Li, J., Xiong, X., Coble, M., et al. (2018). Genetic architecture of collective behaviors in zebrafish. bioRxiv. 1:350314. doi: 10.1101/350314.
- Tardif, S., Akrofi, A.S., Dass, B., Hardy, D.M., MacDonald, C.C. (2010). Infertility with impaired zona pellucida adhesion of spermatozoa from mice lacking TauCstF-64. Biol Reprod. 83(3):464-72. doi: 10.1095/biolreprod.109.083238.
- Taye, M., Kim, J., Yoon, S.H., Lee, W., Hanotte, O., Dessie, T., et al. (2017). Whole genome scan reveals the genetic signature of African Ankole cattle breed and potential for higher quality beef. BMC Genetics. 18(1):11. doi: 10.1186/s12863-016-0467-1.
- Turan N., Ghalwash M.F., Katari S., Coutifaris C., Obradovic Z., Sapienza C. (2012). DNA methylation differences at growth related genes correlate with birth weight: a molecular signature linked to developmental origins of adult disease. BMC Med. Genomics. 5:10. doi: 10.1186/1755-8794-5-10.
- Utrera, A.R., Ponce, S.I.R., Izquierdo, A.V., Torres, E.C., Covarrubias, A.C., De La Cruz Colín, L. et al. (2016). Analysis of morphological variables in Mexican backyard turkeys (Meleagris gallopavo gallopavo). Revista Mexicana De Ciencias Pecuarias. 7(3): 377-389.
- Velová, H., Gutowska-Ding, M.W., Burt, D.W., Vinkler, M. (2018). Toll-Like Receptor Evolution in Birds: Gene Duplication, Pseudogenization, and Diversifying Selection. Mol. Biol. Evol. 35(9):2170–2184. doi: 10.1093/molbev/msy119.
- Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S.F., et al. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-

genome SNP genotyping data. Genome Res. 17(11):1665-74.

- Wang, Z., Kumamoto, Y., Wang, P., Gan, X., Lehmann, D., Smrcka, A.V., et al. (2009). Regulation of immature dendritic cell migration by RhoA guanine nucleotide exchange factor Arhgef5. J Biol Chem. 284(42):28599-606. doi: 10.1074/jbc.M109.047282.
- Xu, L., Hou, Y., Bickhart, D.M., Yang, Z., Hay el, H.A., Song, J., et al. (2016). Population-genetic properties of differentiated copy number variations in cattle. Sci. Rep. 6:23161. doi: 10.1038/srep23161.
- 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. doi: 10.1186/1471-2164-15-962.
- Zhu, Y., Wang, W., Wang, X. (2015) Roles of transcriptional factor 7 in production of inflammatory factors for lung diseases. J. Transl. Med. 13:273 doi: 10.1186/s12967-015-0617-7.

PART III

Analysis of population structure based on copy number variation in cattle specialized breed

CNV diversity in cattle breeds may reveal the genetic basis of their respective phenotypic differences and provide insights on their adaptation to environments: extensive farming or intensive farming systems.

We performed the first CNV mapping of Valdostana Red Pied cattle (VRP) breed based on high density SNP chip, comparing the CNVs identified in the VRP with those already available for the Mexican Holstein (HOL) and Italian Brown Swiss (IBS) cattle. The comparison aimed at disclosing a possible relationship between the proprietary genomic structure of each breed and their fitness to different farming systems.

We use different techniques, such as Principal Component Analysis; clustering analysis using the pvclust function of the pvclust R package (Suzuki & Shimodaira, 2006); the admixture model of STRUCTURE software v.2.3.4 (Pritchard et al., 2002; Falush et al., 2003) to investigate population structure and finally we investigate possible regions under selection using V_{ST} statistic as defined in Redon et al., (2006).

Reference

- Falush D., Stephens M., Pritchard J.K. (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics. 164:1567–1587. Doi: 10.1111/j.1471-8286.2007.01758.x.
- Pritchard J.K., Stephens M. and Donnelly P. (2000) Inference of population structure using multilocus genotype data. Genetics. 155, 945–959.
- Redon R., Ishikawa S., Fitch K.R., Feuk L., Perry G.H., Andrews T.D., Fiegler H., et al. (2006) Global variation in copy number in the human genome. Nature. 444(7118):444-54. doi: 10.1038/nature05329.
- Suzuki R., Shimodaira H. (2006) Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics.* 12: 1540-1542. doi: 10.1093/bioinformatics/btl117.

IV) A copy number variant scan in the autochthonous Valdostana Red Pied cattle breed and comparison with specialized dairy populations

Maria Giuseppina Strillacci, **Erica Gorla**, Maria Cristina Cozzi, Mario Vevey, Francesca Genova, Kathy Scienski, Maria Longeri, Alessandro Bagnato.

PLoS ONE
2018, 13(9): e0204669.

doi: 10.1371/journal.pone.0204669

Introduction

The use of genomic information in dairy cattle breeding has taken high priority in recent years, as genomic selection has been adopted to improve genetic gain for production traits such as milk production [1] and meat quality [2] in cattle breeding programs. In the last 50 years, artificial and natural selection has provoked changes within the cattle genome, causing relevant phenotypic and genetic variability and resulting in the adaptation to local environments [3].

Structural variations, as the Copy Number Variants (CNVs), are one of the major contributors to genetic diversity and phenotypic variation [4]. Liu et al., (2010) [5]. Underlined the importance of CNVs in disclosing genetic diversity among populations and in breeds evolution.

CNVs were defined as large-scale insertions and deletions, ranging from 50 bp to several megabases (Mb) [6]. Compared to SNPs, which are commonly used to detect the existing genetic variation in cattle, CNVs involve larger genomic regions and may have stronger effects on gene regulation and expression. These effects include the modification of gene dosage and structure, which in turn cause exposure of recessive alleles and the alteration of gene regulation [7,8]. Studies in several species have found that CNVs are sources of phenotypic variability as well as disease susceptibility, describing up to 30% of the genetic variation in gene expression [9,10].

CNVs have been mapped in several livestock species [11,12,13], although their use as markers to explain intra-breed genetic diversity has been explored in only a few species [14,3,15]. CNV properties used to explore the diversity and structure of cattle populations remains an issue of little investigation [16]. The study of genetic variation in local populations is a fundamental step in understanding the evolutionary processes that lead to the divergence and differentiation of breeds. Since the mid 20th century, the strong selective pressure to increase milk production in cattle has led to the specialization of breeds that were once dual-purpose in the past (i.e. Brown Swiss) to where their structure in terms of size and physiology has drastically changed.

The Valdostana Red Pied (VRP), farmed in the Aosta Valley located in the northwest Alps of Italy, is an autochthonous dual-purpose cattle breed that did not undergo any specialized intensive selection for neither milk nor meat. This population is bred for milk and meat, and possesses fairly considerable milk production considering the size of the animal (mature weight of 500 kg on average). It is a well-adapted breed to harsh environments as those that animals face during summer pasture in the Alps. Therefore, it is thought that the VRP's genetic background is a population that diverged less than specialized populations as the Brown, from the ancestral cattle populations of the Alps.

CNV diversity in cattle breeds may reveal the genetic basis of their respective phenotypic differences and provide insights on their adaptation to environments: extensive farming vs. intensive farming systems.

In this study we mapped the CNVs of 143 Valdostana Red Pied (VRP) bulls in order to identify structural variations in this breed's genome. Additionally, we compare the VRP's CNVs with those already identified in the Mexican Holstein (HOL) and Italian Brown Swiss (IBS) cattle to highlight genomic structure diversity possibly linked to differences in breed fitness. Breeds were chosen because of their selection histories. VRP remains a dual-purpose breed, HOL has been heavily selected for milk production and intensive farming and IBS, while a dairy cattle breed for not more than 20 years, was initially selected for dual-purpose characteristics.

Results

CNV and CNVR detection in VRP breed

The stringent quality control performed with SVS® allowed for the identification of 35 outlier individuals that were identified according to the Derivative Log Ratio Spread (DLRS) and genomic wave factor values. A total of 6,784 CNVs were detected with PennCNV software across the 29 autosomal chromosomes in a final dataset of 108 VRP bulls. Among these, 3,990 were deletions (i.e. loss states 0 and 1) and 2,794 were duplications (i.e. gain states 3 and 4), with a deletions/duplications CNV ratio of 1.42 calculated as the total number of losses divided by the number of gains. The CNV count ranged from 38 to 141 CNVs per sample, with an average of 62 CNVs. Additionally, the size of CNVs ranged from 31,558 to 103,139 bp, with an average size of 55,566 bp. Table 1 shows the descriptive statistics of the identified CNVs and CNV regions (CNVRs) at population level according to their state.

All the CNVs were merged into 1,723 unique CNVRs (832 gains, 812 losses and 79 complex) across all individuals, covering a total of 59.4 Mb of the genome, which corresponds to 2.36% of the bovine UMD3.1 assembly.

In Table S1 the complete list of CNVRs in the VPR is reposted. In Fig 1, the map displays gain, loss and complex CNVRs on each chromosome. Table S2 reports the number of CNVRs by state (gain, loss and complex) and the proportion of coverage by chromosome in the VPR. Although CNVRs were found on all autosomes, the number and the total size of CNVRs per chromosome were not correlated with their lengths.

The regions mapped in a large number of individuals were: chr12 at 72.42-74.59 Mb (n = 104 samples) and 70.49-72.12 Mb (n = 91 samples), chr5 at 117.28-117.64 Mb (n = 107 samples), and chr10 at 23.89-25.26 Mb (n = 76 samples). In some cases, subjects contribute with two or more adjacent CNVs to the location of these regions.

A classification based on CNVR length was performed for each state (i.e. gain, loss, complex) and the CNVRs have been divided into three classes of length: 1-10 kb, 10–100 kb, >100 kb (Fig 2). The majority of CNVRs identified in this study (n =1,043) have a length comprised between 10 kb and 100 kb. The class of length comprised between 10 and 100 kb harbors the highest number of gain, loss and complex CNVRs. In addition, 593 CNVRs have a length comprised between 1 and 10 kb, while only 87 CNVRs had a size longer than 1 Mb.

Additionally, each class of CNVRs length has been divided into four classes of CNV frequency per individual (1, 2-4, 5-15, ≥ 16). The frequency count is shown in Fig 3. Thus, for every state, CNVRs were defined as singleton regions (if defined by one single individual), rare regions (if determined by 2-4 individuals), moderately recurring (if determined by 5-15 individuals), or recurring regions if including at least 16 individuals (Fig 3). In general, among the identified CNVRs, 1,061 (58.9%) were singleton, 440 (25.5%) were rare regions and, 267 (15.5%) are CNVRs identified in more than 5 individuals. If we consider CNVR states, the occurrences of singleton and rare regions were the most frequent both in gain and loss regions as shown in Fig 3.

Annotation of Valdostana Red Pied CNVRs

A total of 882 Ensembl gene IDs (Ensembl UMD3.1), corresponding to 442 genes with an official ID, have been identified in the 1,723 CNVRs of the VPR. Five hundred and thirty-six regions (31.1%) encompassed one or more genes, while 1,187 (68.9%) did not involve any gene (Table S3).

The GO Term and KEGG pathway analysis was performed using the DAVID Classification database. After FDR (p-value < 0.05), terms resulting as statistically significant included 12 genes involved in heart development as "Biological Process," and 4 genes involved in glucoside activity as "Molecular Component." The complete list of Biological Process, Cellular Component, and Molecular Function is reported in Table S4.

Comparison of CNVs across populations

A comparison among VRP, HOL, and IBS cattle breeds was performed using CNVs called here and previously published, summarized in Table S5 and in the Venn diagram of Fig. S1. We observe that 171 CNVRs are shared among the three breeds, while 1,107, 1,800, and 1,161 unique CNVRs belong to the VRP, IBS and HOL, respectively. In particular, the CNVRs found in HOL overlap with 18.16% (313 CNVRs) of those found in VRP, while the identified regions in IBS overlap with 27.51% (474 CNVRs) of those found in VRP. Considering the lengths of the common 171 CNVRs, we can observe that those shared by VRP and IBS have an average length of 29.82 Mb (50.17% of the length of the CNVRs identified in this study), while the ones common to HOL and VRP, show an average length of 24.15 Mb (40.06% of the length of the CNVRs detected).

Principal component analysis

In the PCAs, the first two principal components explain 10.2% and 3.1% respectively of the total variability of data (PC1 and PC2) for Analysis 1. The same occurs for Analysis 2 where 10.5% and 2.3% of the total genetic variation is explained by PC1 and PC2. Both analyses clearly identified three clusters corresponding to the three breeds (Figs 4A and 4B). While VRP and IBS breeds appeared to be closer, a clear separation resulted between IBS and VRP in respect to HOL.

Clustering to infer population structure

The STRUCTURE software was employed to analyze the genetic structure of the 396 animals of IBS, VRP and HOL. The analysis identified the true number of clusters (subpopulation) in which it is possible to divide the considered pools of individuals. i.e. VRP, IBS and HOL. Both the analyses (Analyses 1 and 2) assumed a model with 12 clusters (K=12). Based on the heuristic test, the estimated likelihood (LnP (D)) values were used to obtain the ΔK values in order to distinguish the break in slope of the distribution of LnP (D) values at the true K. The analyses identify K=3 as the likely K value suggesting that the population should be divided into 3 genetic groups: the VPR, the IBS and the HOL. In both analyses at K=2, VRP and IBS were clearly assigned to a unique group distinct from HOL. At K=3, the three breeds resulted in a clear separation of three clusters and most of the individuals were assigned to a cluster according to the breed division. From K=4 to K=12, the high level of admixture in each of the breeds (in particular in the HOL) shows that the successive clustering is not significant (Figs 4A and 4B).

The cluster tree represented in Fig 5 was built using the CNVR differences identified in the three considered populations. Each node of the tree reports the AU-P and Bootstrap probability values and the edge number. As reported by [17] the AU-P value is considered more accurate than the BP-P value. Even if many AU-P values reported for every node of the tree are low, maybe due to the number of CNVRs considered in this analysis (171 regions share among the three breeds), the majority of individuals are grouped in three distinct clusters corresponding to the three populations (breed-cluster). To be noted that, IBS and VRP, although separated in different clusters, come from a common node.

Population Differentiated CNVs on V_{ST}

In order to test if the CNVs can be related with population-specific selection, we calculated the pairwise V_{ST} among every combination of the three breeds (HOL vs IBS, VRP vs HOL, and VRP vs IBS). The V_{ST} statistic defines values that range from 0 to 1; the high V_{ST} values (close to 1), similar to F_{ST}, suggest differentiation between populations, while low values (close to 0) are indicative of very similar populations.

To calculate the V_{ST} we used a total of 930 CNVs (only those identified in at least 5 individuals in each population), defined by 1,222 SNPs. The defined threshold, taking into account the pairwise of V_{ST} > mean + 2 standard deviations, identified a total of 33 CNVs (Fig 6): 8 for HOL vs IBS; 13 for VPR vs HOL;, 12 for VPR vs IBS. The genes and QTL annotated in these CNVs are reported in Table 2.

Discussion

Although recent studies on CNVs in cattle breeds using high-density SNP chips have been performed, limited knowledge regarding genetic variability and CNV characterization in local populations like the VRP is available. This study is the first CNV scan on the VRP using a high-density SNP chip, and provides valuable information of the structural genomic variation able to enrich the Bovine CNV map. A total of 6,784 CNVs were detected in the autosomes of 108 VRP bulls, and breed-specific region under selection were identified comparing CNVs mapped here and those available from previously published studies for IBS (n=164) [18] and HOL (n= 124) [19] populations. We observed a similar number of duplications (gain state) and deletions (loss state) in VRP and IBS, while the number of deletions (loss state) is superior to the number of duplications (gain state) in the HOL breed. The latter result was previously reported for the Holstein breed in several studies based on SNPs [20] and whole genome sequencing [21]. These results suggest the existence of high genetic variability among these breeds. When we assessed population structure, both principal component analyses revealed that the three cattle breeds form non-overlapping clusters, which is evident given that they are three separated populations, even though the second PCA shows a clearer separation among IBS and VRP. The same results are found by the hierarchical clustering, performed on a matrix based on presence or absence of a CNV in a CNVR, which also exhibits that the HOL, VRP and IBS samples are grouped in three distinct clusters. Also, both the admixture analyses revealed that at K=3 the three breeds result in three clearly separated clusters, and most of the individuals are assigned to a cluster according to their breed division. Very interestingly at K=2 IBS and VPR result a unique genetic population. Till 30 years ago in fact both VPR and IBS were sharing the same selection by breeders: milk, meat and adaptation to pasture. This latter characteristic is fundamental for breeds that during summer face the environmental challenge of pasturing in harsh mountain. This is still the ongoing selection objective for the VPR, while the IBS selection pushed in the last 30 years towards the specialization of the population as a dairy breed. Nowadays, in fact, the IBS is a specialized dairy breed with a large proportion of genes coming from the US Brown, historically selected for milk production. The results of this study show that IBS and VPR still are very close populations as the 30 years of strong directional selection in the IBS is still not sufficient to completely differentiate the two populations.

Regarding the HOL since 1950, Mexico has imported Holstein germplasm (mainly animals and semen) largely from the USA and Canada to increase the productivity of its dairy cattle populations [22]. The same occurred in Italian Holstein where more than 80% of the genetic origin is attributed to US bulls [23]. The HOL population here analyzed thus can be considered a representative sample of the genetic background that USA population has diffused all over the world in the last century after importation from the Holstein and Frisian regions of north Europe. The HOL population then has an origin mostly completely different than VPR and IBS. This result clear at K=2 where HOL population is clustered separately from VPR and IBS. Additionally, the HOL at K=3 is showing common CNV regions with the IBS and in a very minor extent to the VPR. We may speculate that this has occurred because the selection in the IBS to increase milk production has generated CNVs of common importance between HOL and IBS. Nevertheless, at K=3 IBS and VPR remain very well differentiated from the HOL and results to be 2 distinct populations.

The pairwise V_{ST} for the three comparisons (HOL vs IBS, VRP vs HOL, and VRP vs IBS) was estimated in order to identify CNVs under a population-specific selection. According to the V_{ST}, we identified a total of 33 CNVs that differing in frequencies in the above-mentioned comparisons, 8, 12 and 13, respectively, could be considered involved in breed selection. The high V_{ST} values in the comparison of VRP vs IBS, as shown in Table 2, are closer to zero in respect to the V_{ST} results obtained comparing HOL to the other two breeds, which are closer to one. This confirms the genetic similarity described above between the two populations and their difference from the HOL.

Among the 33 genomic regions, 21 CNVs encompass 22 genes, some of which have a well-known phenotype associated in cattle or in other species. The lysozyme gene (*LYZ*) (VPR vs IBS) on BTA 5, for example, encodes for the 1,4-beta-N-acetylmuramidase C. It belongs to a class of enzymes that lyse the cell walls of certain gram-positive bacteria and has also been described in other important functions including inactivation of certain viruses, enhancement of phagocytic activity for leukocytes and macrophages, and control of inflammation [23]. For the same breed comparison, the CNV on BTA 10 contains leucine-rich repeat containing 49 (*LRRC49*), which has been associated with subcutaneous fat and marbling score in the Canchim beef breed by [25]. In respect to the HOL vs IBS comparison, the CNV on BTA 23 overlaps BCL2 antagonist/killer 1 (*BAK1*). This gene plays a crucial role in inducing apoptosis, and [26] associates it with carcass measurements in beef cattle breeds. Also, sortilin related VPS10 domain containing receptor 2 (*SORCS*2) on BTA 6 has been associated with lipid metabolism in different mammal species and with back fat thickness in the Nellore beef breed by [27]. The CNV identified on BTA 13 overlaps with lipin 3 (*LPIN3*). This gene has been associated with both lipodystrophy in humans and with back fat thickness in cattle by [28]. Also, [29] defines this gene as a potential marker for hepatic metabolic adaptations to negative energy balance, as well as for altered physiological state occurring during the transition period in cattle, like adipose tissue lipolysis or hepatic fatty acid oxidation.

Finally, in the last comparison of VRP vs HOL, the possible candidate genes under selection are reelin (*RELN)*, gamma-aminobutyric acid type A receptor alpha2 subunit (*GABRA2*), and solute carrier family 9 member C2 (*SLC9C2*). The *RELN* gene, on BTA 4, is involved in the regulation of mammary gland morphogenesis [30]. These authors also report a down-regulation of *RELN* in lactating pregnant cows, showing an imbalance and possible lower availability of this protein affecting embryo differentiation and development. The *GABRA2* gene, located on BTA 6, is involved in stress response in the mouse species [31]. Lastly, the *SLC9C2* gene is located within a CNVR associated with a polyunsaturated fatty acid profile in intramuscular fat of the *Longissimus thoracis* muscle in a Nellore cattle population (Lemos, 2017. Online Thesis; http://hdl.handle.net/11449/150817).

In addition, *EPHB3*, *PRAME*, *TSPY*, *and ZNF280B* were identified as genes under selection and have also been reported in [16], who reported a comparison between Taurine (included Holstein and Brown Swiss cattle breeds) and two African multipurpose populations using V_{ST}. Furthermore, 12 QTLs overlapped with the significant CNVs resulting from the V_{ST} analysis, and some of these have already been linked to functional processes in cattle (Table 2).

In general, our analyses revealed distinctiveness among the IBS and VRP in respect to HOL, especially related to genes regulating the distribution of intramuscular lipids, which is indicating a difference in metabolism of individuals. In particular we may speculate that the use of resources in HOL is not addressed to fat deposition and in a more general context to body weight, differently than in the double purpose VPR breed, an in a minor extent in the IBS, a double purpose breed till few years ago.

Conclusions

In this project, we performed the first CNV mapping in an autochthonous cattle population, the Valdostana Red Pied breed, using high-density SNP genotypes. The study permitted to disclose a CNV map in a local population well adapted to a harsh environment., and to compare it with 2 cosmopolitan populations, the Holstein and the Brown Swiss. One of the major indications of this study is that the directional selection occurring in population is affecting the genome in term of CNVs. Particularly the comparison among a very selected and specialized population, the HOL, a population as the Italian Brown Swiss where a directional selection occurred only recently, and a population under a very limited selection pressure for milk and meat but maintained adapted to environment as the VPR, discloses differentiated CNVRs where genes and QTL related to their selection history are annotated.

Materials and Methods

Sampling and genotyping

The Associazione Nazionale Allevatori Bovini di Razza Valdostana (A.N.A.Bo.Ra.Va.) provided commercial semen doses of 143 bulls. No animals were involved directly in this study; consequently, no ethical approval was required. Genomic DNA was extracted from semen using the ZR Genomic DNA TM Tissue MiniPrep (Zymo, Irvine, CA, U.S.A.). DNA was quantified using

NanoQuant Infinite®m200 (Tecan, Männedorf, Switzerland) and diluted to 50 ng/μl as required in order to apply the Illumina Infinium protocol. DNA samples were genotyped using BovineHD Genotyping BeadChip Illumina (Illumina Inc., San Diego, USA) containing 777,962 polymorphic SNPs with a median <3 kb gap spacing.

CNV and CNVR detection in VRP breed

Intensity signals from all SNPs were clustered using the Illumina BeadStudio software V.2.0 (Illumina Inc.). Samples with a call rate below 98% were excluded. The signal intensity data of log-R ratio (LRR) and B allele frequency (BAF) were exported from the Illumina BeadStudio software on all the autosomes. As quality control, the overall distribution of derivative log ratio spread (DLRS) values was used in the SVS 8.4 software (Golden Helix Inc.) to identify and filter outlier samples [32]. In addition, individuals were also screened according to their GC content, which is correlated to a long range waviness of LogR ratio values and outlier samples, as detected by the SVS 8.4 wave detection factor algorithm [33], were edited. The PennCNV software (http://penncnv.openbioinformatics.org/en/latest/) was used for CNV calling in the VRP breed. PennCNV is based on a Hidden Markov Model (HMM) algorithm using as input the LRR and BAF data from the SNP arrays. Only samples with a standard deviation (SD) of LRR <0.30 and with default set of BAF drift as 0.01 were used to call CNV. Additionally, a minimum of three adjacent SNPs was required for the detection. The CNV regions (CNVRs) were defined as described by [34], using the BedTools software (-mergeBed command) [35], through merging overlapping CNVs by at least 1 bp. CNVRs were classified as "gain" if there was a duplication of the genome, "loss" if there was a deletion, or "complex" if the region comprised both gain and loss events.

Comparison of CNVs across populations

In this study, we used CNVs to study the population-genetic properties in cattle. In order to identify genomic diversity among the three populations (VPR, HOL, and IBS), we used the individual CNVs available from [18] and those identified in Italian bulls selected from [19]. CNV calling was performed following the same procedures as in our study, and only CNVs identified (within each breed) in at least five individuals were considered in this comparison. Based on CNV two different matrices (number of individuals by number of CNV) were built and applied for analysing population genetic properties. The first matrix was built by presence ("1") or absence ("0") of a CNV in a CNVR, without considering if CNVs were a gain or a loss (Analysis 1) as used in the studies of [13-15]. The second matrix was built according to the CNV genotypes: "0" homozygous deletion, "1" heterozygous deletion, "2" normal state (absence of CNV in that region), "3" heterozygous duplication and "4" homozygous duplication (Analysis 2) as applied in [36]. The use of two different approach to inform the matrices built was chosen to explore if the presence of the CNV in a CNVR is sufficient to discriminate genomic variation among individuals and if the availability of the CNV genotype is providing additional information. Different approaches and software were used in order to disclose population structure and diversification of the three breeds considered. The Past software [37] was employed to perform two different principal component analyses (PCAs) of pairwise individual genetic distances based on allele frequencies of CNVRs classified according to Analyses 1 and 2 (as above). The

STRUCTURE v2.3.4 software [38,39] was used to obtain a complete representation of the population structure of the considered breeds, using both the two matrices built as hereinbefore described. The Admixture model of STRUCTURE without the LocPrior option was used, with a 5,000 burning period and 10,000 iterations, performing five repeats for each K value from 2 to 12 and assuming three different populations. On the basis of STRUCTURE results, the best K values were calculated using the Structure Harvester software [40], which provides the DeltaK values according to the heuristic method reported by [41]. The Distruct software [42] was utilized to graphically visualize each cluster assignment for K of 2 to 12. A clustering analysis was then performed using the pvclust package of the software R [17], applying a hierarchical agglomerative clustering to the scoring matrix based on Analysis 1 (as default input for this application). In order to obtain the Approximately Unbiased P-value (AU) and identify the branches robustness, a multiscale bootstrap resampling (n=10,000 bootstraps) was used. For the hierarchical clustering method, we employed the Unweighted Pair Group Method with Arithmetic mean (UPGMA).

In order to identify novel and exclusive population-differentiated loci, the V_{ST} statistic (highly correlated with Wright's fixation index of F_{ST}) was used. As defined in [34], V_{ST} is calculated by considering $(V_T-V_S)/V_T$, where V_T is the variance in LRRs mean of SNPs (within defined CNVR) estimated among individuals of two populations and V_S is the average variance within each breed, weighted for breed size (in our case: VRP *vs* HOL, VPR *vs* IBS, and HOL *vs* IBS).

Annotation and Gene Ontology and Pathway Analysis

The full Ensembl UMD3.1 gene set for the autosomal chromosomes was downloaded from Ensemble Genome Browse database (release 90 - August 2017), using BioMart (http://www.ensembl.org/biomart). Gene ontology (GO) and KEGG pathways analyses were performed with the high classification stringency option and FDR correction, using the DAVID database (https://david.ncifcrf.gov). The analyses allowed the identification of molecular functions, biological processes, cellular components and pathways for the genes included in the consensus CNVRs. In addition, the National Animal Genome Research Program database (https://www.animalgenome.org) was utilized to catalogue bovine QTL overlapping in both VRP's CNVRs and within significant CNVs.

Tables

Table 1. Descriptive statistics for CNVs and CNVRs detected in VRP breed

State*	No.	Mean Length	Min Length	Max Length	Total Coverage
CNVs					
0	1,434	59,322	1,245	581,425	3.39%
1	2,556	45,839	1,264	523,180	5.72%
3	2,779	56,924	1,030	1,052,912	6.00%
4	15	52,381	3,270	273,013	0.01%
All	6,784	59,322	1,245	581,425	15.10%
CNVRs					
Loss	812	29,827.30	1,263	494,272	0.53%
Gain	832	26,438.23	1,029	692,847	0.88%
Complex	79	167,388.85	1,714	2,170,361	0.96%
All	1,723	34,498.03	1,029	2,170,361	2.36%

*0=homozygous deletion, 1=heterozygous deletion, 3=heterozygous duplication, and 4=homozygous duplication

Table 2. List of CNVRs and gene and QTL annotation for pairwise V_{ST} > Mean + 2 S.D

CHR	CNV Start	CNV End	Length	V_{ST}	IND*	Genes	QTL**
VPR *vs* IBS							
1	83218713	83238102	19389	0.141	5	*EPHB3*	Conformation score QTL (106404, 106405), Average daily gain QTL (106246), Muscularity QTL (106247, 106248)
2	56375294	56403140	27846	0.132	5		
3	71477185	71486626	9441	0.165	11		
5	3434356	3439861	5505	0.133	6		
5	40181727	40209934	28207	0.141	6	*CNTN1*	
5	44705963	44718715	12752	0.14	5	*LYZ*	
9	71525299	71608476	83177	0.143	7		
10	17775153	17784123	8970	0.123	16	*LRRC49*	
13	43884430	43940108	55678	0.117	21	*AKR1C3*	
16	7901886	7948314	46428	0.11	12		
16	80271680	80284738	13058	0.157	7		
18	61894649	61918012	23363	0.246	37		
25	18666885	18674448	7563	0.128	11	*ERI2, REXO5, DCUN1D3*	
HOL *vs* IBS							
3	93310320	93315045	4725	0.615	7		Somatic cell score QTL (122082)
6	118543527	118545281	1754	0.587	5	*SORCS2*	
7	4226753	4238450	11697	0.591	7	*COPE*	
8	83242450	83261773	19323	0.769	5	*TSPY*	
13	70667271	70698983	31712	0.6	21	*LPIN3, EMILIN3*	
17	25056695	25119996	63301	0.874	97	*PRAME*	Average daily gain QTL (106236), Conformation score QTL (106238,

							106239)
17	51115979	51370688	254709	0.651	60		Conformation score QTL (106240)
23	7655804	7688981	33177	0.595	58	*BAK1, GGNBP1, ITPR3*	
				VRP *vs* HOL			
4	45062559	45072215	9656	0.618	6	*RELN*	
5	108810406	108866833	56427	0.358	6	*DCP1B*	
6	66451170	66465621	14451	0.358	5	*GABRA2*	
7	43487164	43498441	11277	0.462	67	*LOC788287*	Calving ease (maternal) QTL (106493)
8	105250028	105303832	53804	0.331	7	*COL27A1*	
10	23133923	23160598	26675	0.305	16		
15	1277543	1312041	34498	0.312	27		
16	56458959	56475433	16474	0.3	26	*SLC9C2*	
17	73004371	73023888	19517	0.453	7	*ZNF280B, ZNF280A*	
18	59154291	59182962	28671	0.301	5		Length of productive life QTL (123783)
24	61918390	62143246	224856	0.304	9	*BCL2, KDSR*	Body weight gain QTL (69320), Daughter pregnancy rate QTL (107040)
25	7380550	7388001	7451	0.307	6		Lean meat yield QTL (36946)
28	43916806	43924903	8097	0.534	7		

*IND = individuals per CNVR;

**https://www.animalgenome.org/cgi-bin/QTLdb/BT/index

Figures

Fig 1. Distribution of the CNVRs on the chromosomes according to their state (gain, loss and complex)

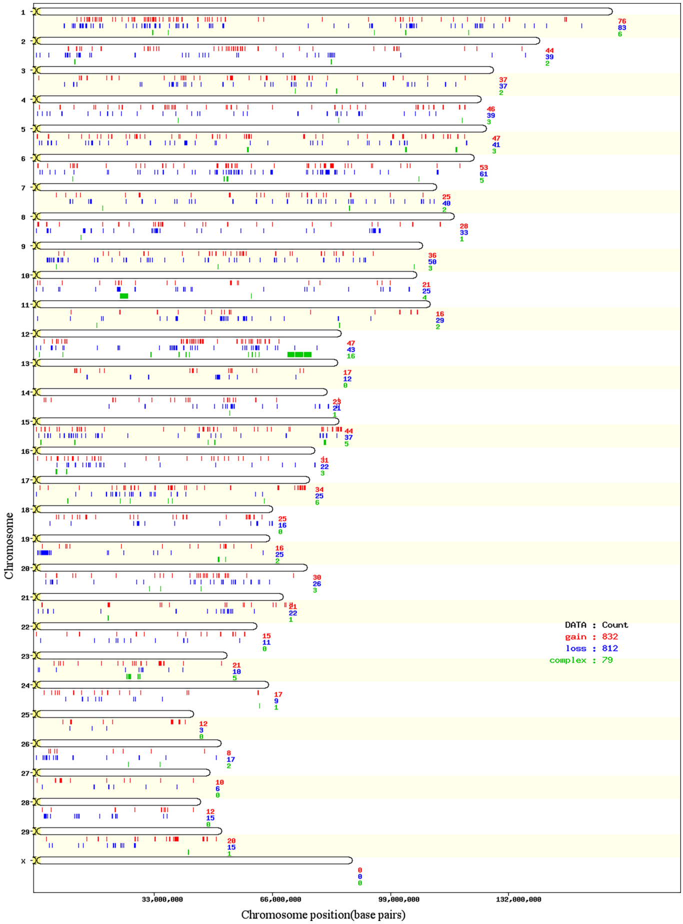


Fig 2. Distribution of CNVR lengths in VRP identified with PennCNV

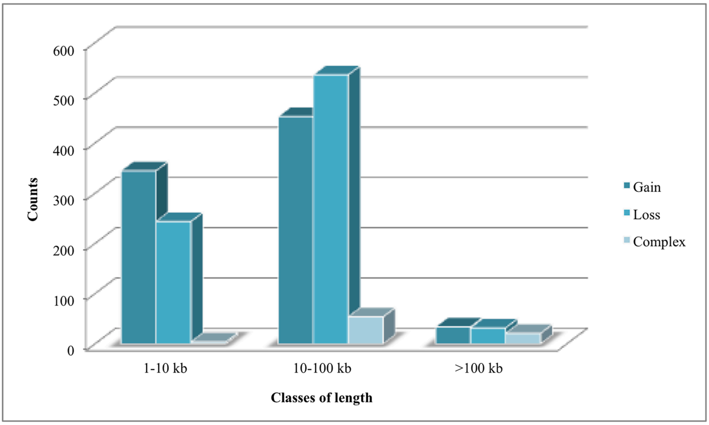


Fig 3. Sample count per individual class (1 singleton; 2-5; 5-15; >16) in each class of CNVR length (1-10; 10-100; >100 kb), according to CNVR states.

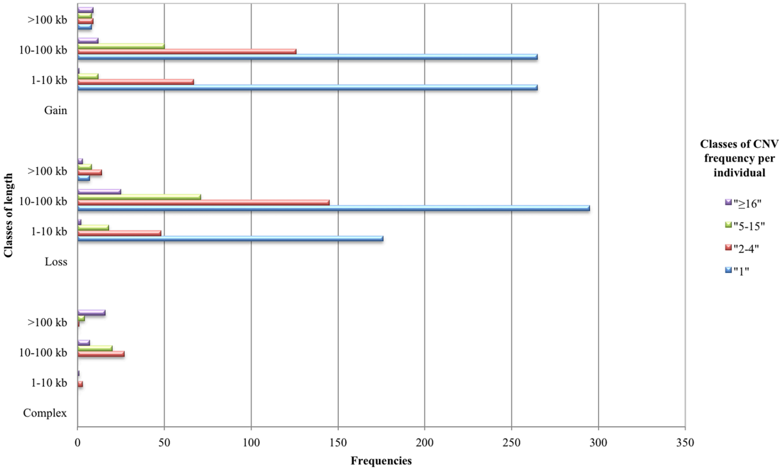


Fig 4. PCAs and population STRUCTURE analyses of three cattle breeds (VRP, IBS and HOL) based on CNVs. Twelve subpopulation clusters inferred by STRUCTURE are represented by different colors (K2-K12). A) Analyses run considering five CNV genotypes: (1) normal state, (2) homozygous deletion, (3) heterozygous deletion, (4) homozygous duplication, or (5) heterozygous duplication; B) Analyses run considering presence or absence of a CNV in a CNVR

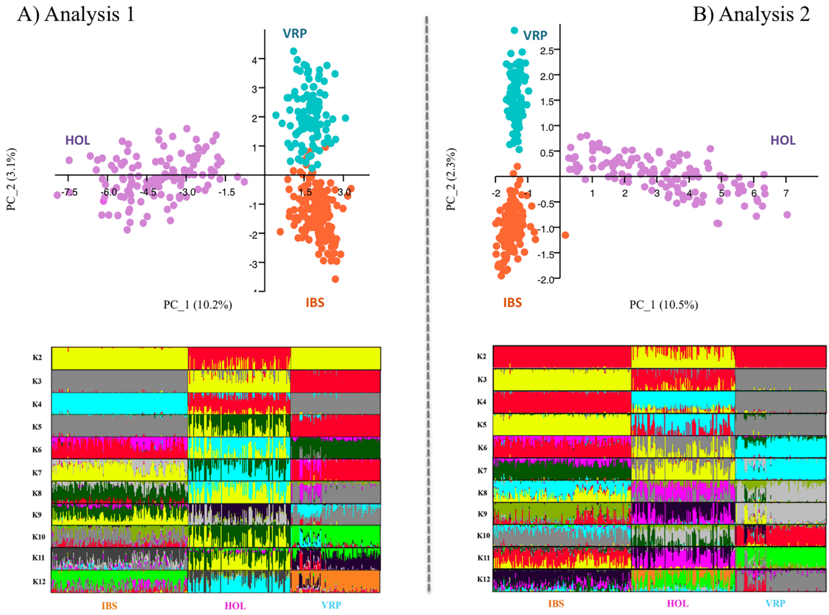


Fig 5. Dendrogram obtained from clustering analysis based on common CNVRs of VRP, IBS and HOL breeds

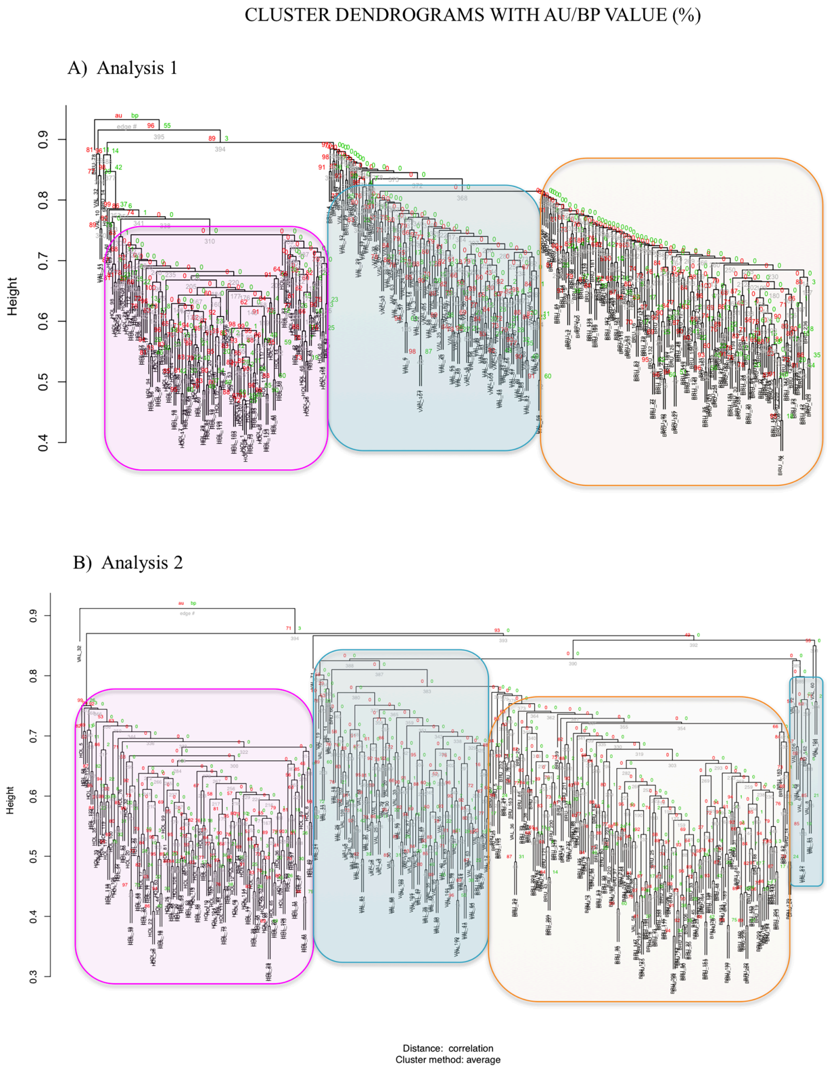
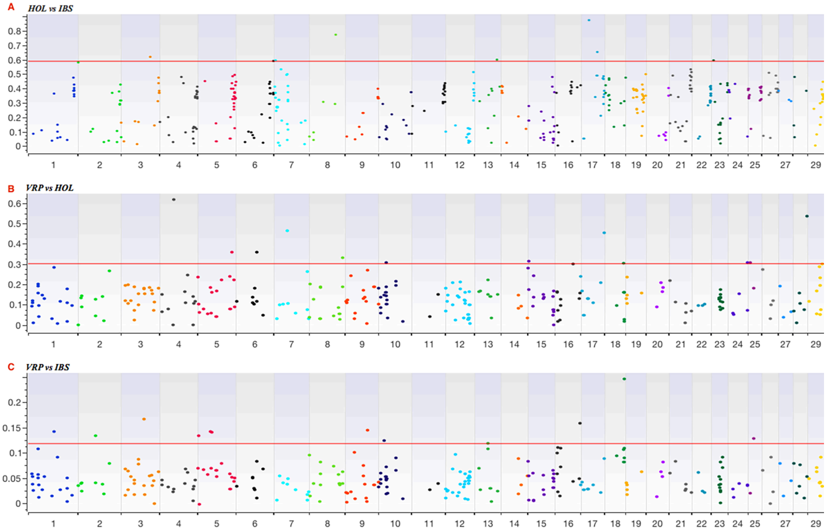


Fig 6. Genome wide V_{ST} value plots for CNVs in the combinations: A) HOL vs IBS; B) VRP vs HOL; C) VRP vs IBS



Supporting information

All supplementary files are available at:

https://doi.org/10.1371/journal.pone.0204669

References

[1] VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, et al. Invited review: Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 2009;92:16–24. doi: 10.3168/jds.2008-1514.

[2] Mannen H. Identification and utilization of genes associated with beef qualities. Anim. Sci. J. 2011;82:1–7. doi: 10.1111/j.1740-0929.2010.00845.x.

[3] Bickhart DM, Xu L, Hutchison JL, Cole JB, Null DJ, Schroeder SG, et al. Diversity and population-genetic properties of copy number variations and multicopy genes in cattle. DNA Res. 2016; 23(3):253-262. doi: 10.1093/dnares/dsw013.

[4] Pezer Z, Harr B, Teschke M, Babiker H, Tautz D. Divergence patterns of genic copy number variation in natural populations of the house mouse (Mus musculus domesticus) reveal three conserved genes with major population-specific expansions. Genome Res. 2015; 25:1- 11. doi: 10.1101/gr.187187.114.DC1.

[5] Liu GE, Hou Y, Zhu B, Cardone MF, Lu J, Cellamare A, et al. Analysis of Copy Number Variations among diverse cattle breeds. Genome Res. 2010;20(5):693–703. doi: 10.1101/gr.105403.110.

[6] Mills RE, Klaudia W, Stewart C, Handsaker RE, Chen K, Alkan C, et al. Mapping copy number variation by population-scale genome sequencing. Nature. 2011;470(7332):59-65. doi: 10.1038/nature09708.

[7] Zhang F, Gu W, Hurles ME, Lupski JR. Copy number variation in Human Health, Disease, and Evolution. Annu. Rev. Genomics Hum. Genet. 2009;10:451–481. doi: 10.1146/annurev.genom.9.081307.164217.

[8] Hou Y, Bickhart DM, Chung H, Hutchison JL, Norman HD, Connor EE, Liu GE. Analysis of copy number variations in Holstein cows identify potential mechanisms contributing to differences in residual feed intake. Funct. Integr. Genomics. 2012;12:717–723. doi: 10.1007/s10142-012-0295-y.

[9] Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science. 2007;315:848–853. doi: 10.1126/science.1136678.

[10] Henrichsen CN, Vinckenbosch N, Zollner S, Chaignat E., Pradervand S, Schutz F, et al. Segmental copy number variation shapes tissue transcriptomes. Nat. Genet. 2009; 41:424–429. doi: 10.1038/ng.345.

[11] Schiavo G, Dolezal MA, Scotti E, Bertolini F, Calò DG, Galimberti G, et al. Copy number variants in Italian Large White pigs detected using high-density single nucleotide polymorphisms and their association with back fat thickness. Anim Genet. 2014; 45:745–749. doi: 10.1111/age.12180.

[12] Bagnato A, Strillacci MG, Pellegrino L, Schiavini F. Frigo, E, Rossoni A, et al. Identification and validation of copy number variants in Italian Brown Swiss dairy cattle using Illumina Bovine SNP50 Beadchip. It J Anim Sci. 2015; 14:552–558. doi: 10.4081/ijas.2015.3900.

[13] Gorla E, Cozzi MC, Román-Ponce SI, Ruiz López FJ, Vega-Murillo VE, Cerolini S, et al. Genomic variability in Mexican chicken population using copy number variants. BMC Genet. 2017;18(1):61. doi: 10.1186/s12863-017-0524-4.

[14] Gazave E, Darré F, Morcillo-Suarez C, Petit-Marty N, Carreño A, Marigorta UM, et al. Copy number variation analysis in the great apes reveals species-specific patterns of structural

variation. Genome Res. 2011;21(10):1626-1639. doi: 10.1101/gr.117242.110.

[15] Strillacci MG, Cozzi MC, Gorla E, Mosca F, Schiavini F, Román-Ponce SI, et al. Genomic and genetic variability of six chicken populations using single nucleotide polymorphism and copy number variants as markers. Animal. 2017;11(5):737–45. doi: 10.1017/S1751731116002135.

[16] Xu L, Hou Y, Bickhart DM, Yang Z, Hay el HA, Song J, et al. (2016). Population-genetic properties of differentiated copy number variations in cattle. Sci. Rep. 2016;6:23161. doi: 10.1038/srep23161.

[17] Suzuki R, Shimodaira H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics. 2006;12:1540-1542. doi: 10.1093/bioinformatics/btl117.

[18] Durán Aguilar M, Román Ponce SI, Ruiz López FJ, González Padilla E, Vásquez Peláez CG, Bagnato A, et al. Genome-wide association study for milk somatic cell score in holstein cattle using copy number variation as markers. J Anim Breed Genet. 2017; 134(1):49-59. doi: 10.1111/jbg.12238.

[19] Prinsen RTMM, Strillacci MG, Schiavini F, Santus E, Rossoni A, Maurer V, Bieber A, et al. A genome-wide scan of copy number variants using high-density SNPs in Brown Swiss dairy cattle. Liv Sci; 2016;191:153–160, http://dx.doi.org/10.1016/j.livsci.2016.08.006.

[20] Jiang L, Jiang J, Yang J, Liu X, Wang J, Wang H, et al. Genome-wide detection of copy number variations using high-density SNPgenotyping platforms in Holsteins. BMC Genomics. 2013;14:131. doi: 10.1186/1471-2164-14-131.

[21] Gao Y, Jiang J, Yang S, Hou Y, Liu GE, Zhang S, et al. CNV discovery for milk composition traits in dairy cattle using

whole genome resequencing. BMC Genomics. 2017;18(1):265. doi: 10.1186/s12864-017-3636-3.

[22] Valencia M, Montaldo HH, Ruíz F. Interaction between genotype and geographic region for milk production in Mexican Holstein cattle. Arch. Zootec. 2008; 57(220):457-463.

[23] Malecca C, Canavesi F, Gandini G, Bagnato A. Pedigree analysis of Holstein dairy cattle population. Interbull Bulletin. 2002; 29:168:172. https://journal.interbull.org/index.php/ib/article/view/760/751

[24] Niyonsaba F, Ogawa H. Protective roles of the skin against infection: implication of naturally occurring human antimicrobial agents beta-defensins, cathelicidin LL-37 and lysozyme. J Dermatol Sci. 2005;40(3):157-68. doi: 10.1016/j.jdermsci.2005.07.009.

[25] Mokry FB, Higa RH, de Alvarenga Mudadu M, Oliveira de Lima A, Meirelles SL, Barbosa da Silva MV, et al. Genome-wide association study for backfat thickness in Canchim beef cattle using Random Forest approach. BMC Genet. 2013; 5:14-47. doi: 10.1186/1471-2156-14-47.

[26] Jiang Z, Michal JJ, Chen J, Daniels TF, Kunej T, Garcia MD, et al. Discovery of novel genetic networks associated with 19 economically important traits in beef cattle. Int J Biol Sci. 2009;5(6):528-542.

[27] Júnior GA, Costa RB, de Camargo GM, Carvalheiro R, Rosa GJ, Baldi F, et al. Genome scan for postmortem carcass traits in Nellore cattle. J Anim Sci. 2016;94(10):4087-4095. doi: 10.2527/jas.2016-0632.

[28] Lee SH, van der Werf J, Lee SH, Park EW, Gondro C, Yoon D, et al.. Genome wide QTL mapping to identify candidate

genes for carcass traits in Hanwoo (Korean Cattle). Genes & Genomics. 2012; 34(1):43-49. doi: 10.1007/s13258-011-0081-6.

[29] Loor JJ, Everts RE, Bionaz M, Dann HM, Morin DE, Oliveira R, et al. Nutrition-induced ketosis alters metabolic and signaling gene networks in liver of periparturient dairy cows. Physiol Genomics. 2007;32(1):105-16.

[30] Cerri RL, Thompson IM, Kim IH, Ealy AD, Hansen PJ, Staples CR, et al. Effects of lactation and pregnancy on gene expression of endometrium of Holstein cows at day 17 of the estrous cycle or pregnancy. J Dairy Sci. 2012; 95(10):5657-5675. doi: 10.3168/jds.2011-5114.

[31] Dai J, Wang X, Chen Y, Wang X, Zhu J, Lu L. Expression quantitative trait loci and genetic regulatory network analysis reveals that Gabra2 is involved in stress responses in the mouse. Stress. 2009;12(6):499-506. doi: 10.3109/10253890802666112.

[32] Pinto D, Darvishi K, Shi X, Rajan D, Rigler D, Fitzgerald T, et al. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nat. Biotechnol. 2011;29:512-520. doi: 10.1038/nbt.1852.

[33] Diskin SJ, Li M, Hou C, Yang S, Glessner J, Hakonarson H, et al. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. Nucleic Acids Res. 2008;36(19):e126. doi: 10.1093/nar/gkn556.

[34] Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, et al., Global variation in copy number in the human genome. Nature. 2006;444(7118):444-54. doi: 10.1038/nature05329.

[35] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–842. doi: 10.1093/bioinformatics/btq033.

[36] Lou H, Li S, Yang Y, Kang L, Zhang X, Jin W, et al. A Map of Copy Number Variations in Chinese Populations. PLoS ONE. 2011; 6(11):e27341. doi: 10.1371/journal.pone.0027341.

[37] Hammer Ø, Harper DAT, Ryan PD. PAST: Paleontological statistics software package for education and data analysis. Palaeontologia Electronica. 2001;4(1):9. http://palaeo-electronica.org/2001_1/past/issue1_01.htm.

[38] Pritchard JK, Stephens M and Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155, 945–959.

[39] Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics. 2003;164:1567–1587. DOI: 10.1111/j.1471-8286.2007.01758.x

[40] Dent EA, vonHoldt B M. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Cons Genet Res. 2012; 4(2): 359-361. doi: 10.1007/s12686-011-9548-7.

[41] Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. Mol Ecol. 2005;14(8):2611– 2620. doi:10.1111/j.1365-294X.2005.02553.x

[42] Rosenberg NA. Distruct: a program for the graphical display of population structure. Mol. Ecol. Notes. 2004;4:137–138. doi: 10.1046/j.1471-8286.2003.00566.

GENERAL DISCUSSION

In the studies here presented we focused on structural variations, SNP (Single Nucleotide Polymorphism) and CNV (Copy Number Variants), to disclose and characterize the genomic variation in populations and breeds of different species, cattle and poultry. The possibility to realize such type of studies resides in the recent availability of the reference genome, e.g. turkey, and high-density SNP chip that have been recently developed and released to the animal science community. The SNP genotypes are a class of neutral markers while CNV can contain in a large proportion (up to 60% in poultry) annotated genes related to expressed phenotypes.

Moreover, the study of the Run of Homozygosity (ROH) shown in the Part I, allows to deepen the genomic variability and genomic modifications occurred to the Mexican chicken population.

Findings from the ROH analysis indicated that natural selection affected allele frequencies in specific regions of the Mexican chicken genome and some of the annotated genes in the ROH regions could play important roles in the historical genetic dynamic of this population.

The use of avian species is particularly interesting as we considered populations that were separated by 500 years of evolution in different farming and mating systems. The Mexican chicken population resulted to be a mix of a limited number of genetic founders that were brought in Central America by Spanish conqueror in 16th century.

On the other hand, they brought the turkey from Central America to Europe where this bird found a rapid expansion. The turkey populations were selected in Europe and differentiated in several distinct breeds and in the last four decades in a highly selected hybrid for meat production. Instead, the Mexican turkey population maintained its own mating system as a backyard population.

The genetic comparison of unselected Mexican population of chicken and turkey populations with the selected ones, disclosed common and proprietary structural variation that we speculate is due to the different evolution and selection occurred.

The genetic variation existing in the backyard chicken population of Mexico was mapped using both SNP genotypes and CNV as markers. Results are suggesting that this creole population is a genetic mix derived by three ancestors (Part 1) supporting the evidence that a very limited number of birds funded the actual genetic mix.

The genetic variability resulted in Mexican population respect to the one identified in selected Italian native populations (Strillacci et al., 2016) suggest that the environmental context affected the structural evolution: in Mexico the population evolved expressing genes favourable to the harsh environment, while the Italian population were influenced by the selection operated by farmers for a higher production. Even if further studies are necessary to deeply investigate these findings, the genes harboured in the CNV show a differentiation according to the fitness to the environment of these populations: in a general context, in fact, CNV loss occurred where deleterious genes may be identified, while CNV gains include genes related to production traits. CNV gains, as showed by literature, are generally related to directional selection increasing the number of copies of a gene: e.g. the starch gene in humans and dogs, where has been shown an increase of number of copy of the "starch gene" when a nutritional diet based on starch is consumed vs a non-starch diet cohort. Recently, in fact, a study on a eukaryotic model (Hull et al., 2017) showed that environmental changes are accelerating adaptation through the stimulation of copy number variation and that this is not a random effect but has a cause effect relationship. Additionally, Perry et al. (2007) demonstrated that directional selection due to starch diet (i.e. environmental factor) is increasing specific copies of the genes involved in starch metabolism producing as such CNV gains.

In turkey the variation was here studied using CNV. The turkey model is particularly interesting as we mapped the genetic variation on the "original population", the Mexican one, and compared it to several populations that derived from that genetic pool.

Results are likely reflecting the human action on turkey populations, i.e. its migration to Europe and then back to America, and the directional selection occurring in the last 40 years to produce a fast-growing heavy bird.

The study considers three groups of birds that reproduce and evolve according to different constrains and environmental conditions. The Mexican population developed in a natural environment, with no (or very little) intervention by humans in mating and with no (or very little) supplement of feed. The Italian populations are the result of a phenotypic selection operated by individual farmers in their small group of individuals, to obtain birds that best perform in the semi-extensive farming system (backyard with recovery availability and feeding supplement) that characterized the middle ages poultry system of Italy and Europe. The Hybrid population, in the last 40 years, has been heavily directionally selected, through a very well-structured genetic improvement and breeding plans to improve weight and growing performances and to best perform in an artificially controlled environment with unlimited feed supplement.

This study is the first CNV mapping in a worldwide turkey sampling, from populations collected across different continents and disclosed similarities and variations in CNVs and CNVRs across the populations studied. Because of the diversity in their selection history and actual farming environmental conditions the Mexican, Italian and Hybrid populations provide an interesting model to investigate CNV variation.

These recent findings support the hypothesis that the variability in size of CNV and their number in the Mexican population respect to Hybrid, is possibly related to the different selection and breeding undergoing in these populations, and to the environmental conditions where they are farmed. The impact of the different selection performed on the CNV variability is here supported by the variation in the number of CNV per bird that is the lowest in the Hybrid (10 on average) while the largest in the Mexican (28 CNV) population. Additionally, the length of the CNVs in the Hybrid group is much less variable than in the other two groups (Italian and Mexican).

A first general evidence that can be drawn from the studies on avian specie is that the environmental effect (either environment itself or human intervention) on population evolution is likely affecting the genome structure in term of number of copies of genes: deleterious genes are lost while those related to directional selection are increased in number of copies.

A second evidence is that CNV, as SNP, can be efficiently used to identify genetic distinct clusters. While SNP are neutral markers and can show the long-term evolution, the CNV markers are involved in gene regulation and expression, thus allowing a functional interpretation of the results. The comparison of the results obtained with the SNP markers and CNV, in terms of genetic clustering of the populations, show comparable results.

In the third part of this thesis we presented a CNV mapping based on high density SNP chips in the Valdostana Red Pied, a double purpose breed, comparing it with the CNV detected in two specialized breeds, the Mexican Holstein and the Italian Brown Swiss.

In cattle many studies on CNV using high-density chip have been performed, but this is the first CNV scan in the VRP a local autochthonous population of northwest Italy. The VRP population selection occurring in the last decades was addressed to increase milk and meat production, maintaining the ability of the population to cope with harsh environments and summer pasture practice.

The approach we used in this study is somehow similar to the one in poultry for the comparison among populations: the VRP very well adapted to environmental conditions of the Alps vs the IBS originally a double purpose but strongly selected for milk in the last 30 years and the Holstein strongly selected and specialised for milk production.

The comparison of CNV regions across population using the V_{ST} statistic allows disclosing proprietary deletion or duplication related to the peculiar evolution of the population. The HOL is showing duplications harbouring genes related to production efficiency, while on the contrary the VRP CNV variation is more related to adaptive genes.

In general, the results obtained in the different species show the capability of CNV as markers to disclose genetic variation among populations not identified by SNP and the possibility to relate this variation to annotated genes. This appears to be true across different livestock species, i.e. poultry and cattle.

The results here obtained showed a clear differentiation among the populations analysed. In general populations and breeds evolving in harsh environments show CNV regions related to adaptive genes. While populations and breeds farmed in artificial controlled environmental conditions, i.e. intensive farming, show a genomic CNV evolution that can be related to the strong directional selection for production traits.

Our findings are a first overview on the comparison between selected and unselected populations using non-neutral markers. Further insights should be considered using genomic data that may include other layers of information as expression data and epigenetic marks.

Table 2. Summary of analysis performed in the four studies included in this thesis

	I paper	**II paper**	**III paper**	**IV paper**
POP	**CHICKEN** -Mexican creole chicken	**CHICKEN** -Mexican creole Chicken	**TURKEYS** -6 Italian breeds -Narragansett -Hybrid -Mexican Turkeys	**CATTLE** -Valdostana Red Pied, -Mexican Holstein, -Italian Brown Swiss
CNV detection	-Penn CNV	/	-Penn CNV -Golden Helix	Penn CNV
Structure Analysis	-Hierarchical clustering (pvclust R)	-PCA (Past software) -IBD -Wright's statistics -ADMIXTURE -AMOVA -ROH	-PCA (Past software) -Hierarchical clustering (pvclust R) - STRUCTURE software	-PCA (Past software) -STRUCTURE software -Hierarchical clustering (pvclust R) -V_{ST}
Gene Annotation and KEGG - Go Term Analysis	-Ensemble database -Panther database	-Ensemble Database	-NCBI Turkey_5.0 gene dataset -DAVID Database	-Ensemble database -DAVID Database

Reference

- Strillacci M.G., Cozzi M.C., Gorla E., Mosca F., Schiavini F., Román-Ponce, S. I., et al. (2017). Genomic and genetic variability of six chicken populations using single nucleotide polymorphism and copy number variants as markers. Animal.11(5):737-745. doi: 10.1017/S1751731116002135.
- Hull, R. M., Cruz, C., Jack, C. V., & Houseley, J. (2017). Environmental change drives accelerated adaptation through stimulated copy number variation. *PLoS biol. 15*(6), e2001333. doi: 10.1371/journal.pbio.2001333
- Perry, G. H., Dominy, N. J., Claw, K. G., Lee, A. S., Fiegler, H., Redon, R., et al. (2007). Diet and the evolution of human amylase gene copy number variation. *Nature genetics*, *39*(10), 1256. doi: 10.1038/ng2123.

Acknowledgments

I would like to express my thanks to Prof. Bagnato and Dr.ssa Strillacci for the support and teaching given to me in these years working together.

A big thank you to my family and friends who have always supported me and above all to my daughters who inspire me everyday