



This is the peer reviewed version of the following article: Bovo, S., A. Ribani, M. Muñoz, E. Alves, J. P. Araujo, R. Bozzi, and R. Charneca et al. 2020. "Genome-Wide Detection Of Copy Number Variants In European Autochthonous And Commercial Pig Breeds By Whole-Genome Sequencing Of DNA Pools Identified Breed-Characterising Copy Number States". *Animal Genetics* 51 (4): 541-556. doi:10.1111/age.12954, which has been published in final form at <https://doi.org/10.1111/age.12954>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions <http://www.wileyauthors.com/self-archiving>.

Document downloaded from:



1 **Genome-wide detection of copy number variants in European autochthonous and commercial**  
2 **pig breeds by whole genome sequencing of DNA pools identified breed-characterising copy**  
3 **number states**

4

5 Samuele Bovo<sup>1</sup>, Anisa Ribani<sup>1</sup>, Maria Muñoz<sup>2</sup>, Estefania Alves<sup>2</sup>, Jose P. Araujo<sup>3</sup>, Riccardo Bozzi<sup>4</sup>,  
6 Rui Charneca<sup>5</sup>, Federica Di Palma<sup>6</sup>, Graham Etherington<sup>6</sup>, Ana I. Fernandez<sup>2</sup>, Fabián García<sup>2</sup>, Juan  
7 García-Casco<sup>2</sup>, Danijel Karolyi<sup>7</sup>, Maurizio Gallo<sup>8</sup>, Kristina Gvozdanović<sup>9</sup>, José Manuel Martins<sup>5</sup>,  
8 Marie-José Mercat<sup>10</sup>, Yolanda Núñez<sup>2</sup>, Raquel Quintanilla<sup>11</sup>, Čedomir Radović<sup>12</sup>, Violeta Razmaite<sup>13</sup>,  
9 Juliette Riquet<sup>14</sup>, Radomir Savić<sup>15</sup>, Giuseppina Schiavo<sup>1</sup>, Martin Škrlep<sup>16</sup>, Graziano Usai<sup>17</sup>, Valerio  
10 J. Utzeri<sup>1</sup>, Christoph Zimmer<sup>18</sup>, Cristina Ovilo<sup>2</sup>, Luca Fontanesi<sup>1</sup>

11

12 <sup>1</sup> Department of Agricultural and Food Sciences, Division of Animal Sciences, University of  
13 Bologna, Viale Fanin 46, 40127 Bologna, Italy.

14 <sup>2</sup> Departamento Mejora Genética Animal, INIA, Crta. de la Coruña, km. 7,5, 28040, Madrid, Spain.

15 <sup>3</sup> Centro de Investigação de Montanha (CIMO), Instituto Politécnico de Viana do Castelo, Escola  
16 Superior Agrária, Refóios do Lima, 4990-706 Ponte de Lima Portugal.

17 <sup>4</sup> DAGRI – Animal Science Section, Università di Firenze, Via delle Cascine 5, 50144 Firenze, Italy.

18 <sup>5</sup> MED – Mediterranean Institute for Agriculture, Environment and Development & Universidade de  
19 Évora, Pólo da Mitra, Apartado 94, 7006-554 Évora, Portugal.

20 <sup>6</sup> Earlham Institute, Norwich Research Park, Colney Lane, Norwich, NR47UZ, United Kingdom

21 <sup>7</sup> Department of Animal Science, Faculty of Agriculture, University of Zagreb, Svetošimunska c. 25,  
22 10000 Zagreb, Croatia.

23 <sup>8</sup> Associazione Nazionale Allevatori Suini (ANAS), Via Nizza 53, 00198 Roma, Italy.

24 <sup>9</sup> Faculty of Agrobiotechnical Sciences Osijek, University of Osijek, Vladimira Preloga 1, 31000,  
25 Osijek, Croatia.

26 <sup>10</sup> IFIP Institut du porc, La Motte au Vicomte, BP 35104, 35651, Le Rheu Cedex, France.

27 <sup>11</sup> Programa de Genética y Mejora Animal, IRTA, Torre Marimon, 08140 Caldes de Montbui,  
28 Barcelona, Spain.

29 <sup>12</sup> Department of Pig Breeding and Genetics, Institute for Animal Husbandry, 11080 Belgrade-  
30 Zemun, Serbia.

31 <sup>13</sup> Animal Science Institute, Lithuanian University of Health Sciences, Baisogala, Lithuania.

32 <sup>14</sup> GenPhySE, Université de Toulouse, INRA, Chemin de Borde-Rouge 24, Auzeville Tolosane,  
33 31326 Castanet Tolosan, France.

34 <sup>15</sup> Faculty of Agriculture, University of Belgrade, Nemanjina 6, 11080, Belgrade- Zemun, Serbia.

35 <sup>16</sup> Kmetijski Inštitut Slovenije, Hacquetova 17, SI-1000 Ljubljana.

36 <sup>17</sup> AGRIS SARDEGNA, Loc. Bonassai, 07100 Sassari, Italy.

37 <sup>18</sup> Bäuerliche Erzeugergemeinschaft Schwäbisch Hall, Schwäbisch Hall, Germany.

38

39 \* Corresponding author

40 E-mail addresses:

41 LF: [luca.fontanesi@unibo.it](mailto:luca.fontanesi@unibo.it)

42

43 **Short title:** CNV in European pig breeds

## 44 **Summary**

45 In this study we identified copy number variants (CNVs) in 19 European autochthonous pig breeds  
46 and in two commercial breeds (Italian Large White and Italian Duroc) that represent important genetic  
47 resources for this species. The genome of 725 pigs was sequenced using a breed specific DNA pooling  
48 approach (30-35 animals per pool) obtaining an average depth per pool of 42×. This approach  
49 maximized CNV discovery as well as the related copy number states characterizing, on average, the  
50 analysed breeds. By mining more than 17.5 billion reads, we identified a total of 9592 CNVs (~683  
51 CNVs per breed) and 3710 CNV regions (CNVRs; 1.15% of the reference pig genome), with an  
52 average of 77 CNVRs per breed that was considered as private. A few CNVRs were analysed in more  
53 details, together with other information derived from sequencing data. For example, the CNVR  
54 encompassing the *KIT* gene was associated with coat colour phenotypes in the analysed breeds,  
55 confirming the role of the multiple copies in determining breed specific coat colours. The CNVR  
56 covering the *MSRB3* gene was associated with ear size in most breeds. The CNVRs affecting the  
57 *ELOV6* and *ZNF622* genes were private features observed in the Lithuanian Indigenous Wattle and  
58 in the Turopolje pig breeds, respectively. Overall, genome variability here unravelled can explain part  
59 of the genetic diversity among breeds and might contribute to explain their origin, history and  
60 adaptation to a variety of production systems.

61

62 **Keywords:** CNV; *ELOV6*; Genetic resource; *KIT*; *MSRB3*; Next generation sequencing; *Sus scrofa*;  
63 *ZNF622*.

## 64 **Introduction**

65         Livestock genomes have been shaped by natural and artificial selection, leading to the  
66 accumulation of a broad range of phenotypic and genetic variability that have largely contributed to  
67 differentiate populations and constitute modern breeds. As a result, livestock populations and breeds  
68 represent a reservoir of genetic diversity, harbouring genetic variants that span from single nucleotide  
69 polymorphisms (SNPs) to more complex structural variants, some of which with small to large  
70 phenotypic effects on a variety of exterior and economically relevant traits (Andersen *et al.* 2011).  
71 Copy number variants (CNVs) are a type of structural variants in the form of large DNA segments,  
72 usually more than 1kb of length, which are present in a variable copy number within a species as  
73 compared to its reference genome (Feuk *et al.* 2006).

74         CNVs represent an important source of genetic variability, by influencing phenotypes through  
75 a variety of molecular mechanisms such as gene dosage effect, disruption or alteration of coding and  
76 regulatory regions among several other modifications (Redon *et al.* 2006, Zhang *et al.* 2006, Bickhart  
77 & Liu 2014). Detection of CNVs is technically challenging when applied on genome-wide scale and  
78 different technologies have been applied to this aim. Among them, the most commonly used are array  
79 comparative genome hybridization (aCGH), high density SNP chip and high-throughput sequencing  
80 (HTS) platforms (Winchester *et al.* 2009; Alkan *et al.* 2011; Pirooznia *et al.* 2015; Pollard *et al.* 2018).  
81 However, due to the decreased cost of HTS analyses and the advantage that this approach has to  
82 obtain more precise information on CNVs, whole genome resequencing is becoming a standard  
83 approach to discover and characterize CNVs in complex genomes.

84         Genetic diversity described by CNVs and CNV regions (CNVRs; i.e. CNVs present in different  
85 individuals in the same or overlapping genome regions) has been extensively studied in livestock,  
86 including, for example, cattle (Fadista *et al.* 2010; Bickhart *et al.* 2012), sheep (Fontanesi *et al.* 2011;  
87 Yang *et al.* 2018), goats (Fontanesi *et al.* 2010b; Liu *et al.* 2019), rabbits (Fontanesi *et al.* 2012) and  
88 chickens (Yi *et al.* 2014), among other species. Several studies investigating CNVs and CNVRs have  
89 been also reported in pigs, including also an interspecies survey within the genus *Sus* (Paudel *et al.*

90 2015). Studies have been focused on the main commercial European breeds (i.e. Duroc, Landrace,  
91 Large White, Hampshire, Yorkshire, Piétrain) (e.g. Fadista *et al.* 2008; Li *et al.* 2012; Chen *et al.*  
92 2012; Fowler *et al.* 2013; Wang *et al.* 2014, 2015a, c, 2019b; Jiang *et al.* 2014; Wiedmann *et al.* 2015;  
93 Revay *et al.* 2015; Long *et al.* 2016; Revilla *et al.* 2017; Stafuzza *et al.* 2019) and Asian breeds  
94 (Meishan, Erhualian) (Wang *et al.* 2012, 2014, 2015b, c; Li *et al.* 2012; Chen *et al.* 2012; Jiang *et al.*  
95 2014). Other studies screened commercial pig populations in the attempt to capture part of the missing  
96 heritability (expected to be explained by CNVs) on economically important traits, including number  
97 of piglets born alive (Stafuzza *et al.* 2019), fertility (Revay *et al.* 2015), meat quality traits (Wang *et*  
98 *al.* 2015c), fatty acid composition and growth traits (Revilla *et al.* 2017), fat deposition (Fowler *et al.*  
99 2013; Schiavo *et al.* 2014), among other traits.

100 Although the modern pig industry relies on few commercial pig breeds, autochthonous pig  
101 populations subsist in many different regions, mainly associated with local and traditional niche  
102 markets (Čandek-Potokar and Nieto 2019). These breeds represent genetic resources adapted to local  
103 agro-climatic and environmental conditions. Up to date, the genome architecture of CNVs has been  
104 studied mainly in Asian autochthonous populations/breeds (Li *et al.* 2012; Wang *et al.* 2014, 2015b,  
105 2019a; Jiang *et al.* 2014; Dong *et al.* 2015; Xie *et al.* 2016). European autochthonous pig breeds have  
106 been mainly investigated by exploring their genetic variability using SNP data (e.g. Ovílo *et al.* 2002;  
107 Tomás *et al.* 2011; Wilkinson *et al.* 2013; Silió *et al.* 2016; Yang *et al.* 2017; Muñoz *et al.* 2018,  
108 2019; Schiavo *et al.* 2018, 2019, 2020a, b; Ribani *et al.* 2019). A few studies, using SNP arrays,  
109 analysed CNVs in European autochthonous pig breeds (e.g. Iberian, Swallow-Bellied Mangalitsa)  
110 (Ramayo-Caldas *et al.* 2010; Fernández *et al.* 2014; Molnár *et al.* 2014).

111 Results of CNV studies in pigs showed a limited degree of agreement in terms of CNVRs  
112 number and size ranges. Even if part of these discrepancies may be attributed to breed-specific  
113 genome features, the remaining discrepancies may derive from the different technologies and  
114 algorithms used to unravel CNVs, which mainly used aCGH and SNP arrays. Few other studies  
115 analysed CNVs and CNVRs in the pig genome using HTS platforms (e.g. Rubin *et al.* 2012; Jiang *et*

116 *al.* 2014; Paudel *et al.* 2015; Wang *et al.* 2015c, 2019b; Long *et al.* 2016; Revilla *et al.* 2017; Keel *et*  
117 *al.* 2019).

118 In this study, we provide a detailed survey of CNVs and CNVRs in the pig genome by whole  
119 genome resequencing of DNA pools constituted from 21 European pig breeds: 19 autochthonous  
120 breeds belonging to nine different countries and two Italian commercial breeds. These breeds, some  
121 of them untapped, stem from different production systems and breeding programmes in Europe.  
122 Therefore, dissection of their genome architecture at the level of CNVs could provide new insights  
123 into their histories, origin, potential selection signatures and adaptation to different local agro-climatic  
124 and environmental conditions.

125

## 126 **Materials and methods**

### 127 **Animals**

128 Blood samples were collected from a total of 30 or 35 animals from each of the 21 pig breeds  
129 included in the study, distributed in nine European countries (from West to East and then North; Fig.  
130 1): Portugal (Alentejana and Bísara); Spain (Majorcan Black); France (Basque and Gascon); Italy  
131 (autochthonous: Apulo-Calabrese, Casertana, Cinta Senese, Mora Romagnola, Nero Siciliano and  
132 Sarda; and commercial breeds: Italian Large White and Italian Duroc); Slovenia (Krškopolje pig,  
133 hereafter indicated as Krškopolje); Croatia (Black Slavonian and Turopolje); Serbia (Moravka and  
134 Swallow-Bellied Mangalitsa); Germany (Schwäbisch-Hällisches Schwein); and Lithuania  
135 (Lithuanian indigenous wattle and Lithuanian White old type). Selection of individuals for sampling  
136 was performed by avoiding highly related animals (no full- or half-sibs), balancing between sexes,  
137 and prioritizing adult individuals or at least animals with adult morphology. All animals were  
138 registered to their respective Herd Books and presented standard breed characteristics. Details on the  
139 analysed animals and investigated breeds, including geographical distribution and phenotypic  
140 description, are reported in Table S1.

141

142 **DNA samples and sequencing**

143 Genomic DNA was extracted from 8–15 mL of peripheral blood for each pig, collected in  
144 Vacutainer tubes containing 10% 0.5 M EDTA (ethylenediaminetetraacetic acid, disodium dihydrate  
145 salt) at pH 8.0. The extraction was performed using either a standardized phenol-chloroform  
146 (Sambrook *et al.* 1989) or the NucleoSpin® Tissue commercial kit (Macherey-Nagel, Düren,  
147 Germany). A total of 21 DNA pools were constructed, including in each pool 30 or 35 individual  
148 DNA samples pooled at equimolar concentration (Table S2).

149 A sequencing library was generated for each DNA pool by using the Truseq® Nano DNA HT  
150 Sample preparation Kit (Illumina, CA, USA), following the manufacturer's recommendations.  
151 Briefly, DNA was randomly sheared to obtain 350 bp fragments which were end polished, A-tailed,  
152 and ligated with the full-length adapter for Illumina sequencing with further PCR amplification. PCR  
153 products were purified (AMPure XP system) and libraries were analysed for size distribution by  
154 Agilent 2100 Bioanalyzer and quantified using real-time PCR. The qualified libraries were then fed  
155 into an Illumina Hi-Seq sequencer for paired-end sequencing, obtaining 150 bp length reads.

156

157 **Quality controls and sequence alignment**

158 Obtained reads underwent several cleaning and filtering steps including removal of (i) adapters,  
159 (ii) reads containing more than 10% unknown bases (N) and (iii) reads containing low quality bases  
160 ( $Q \leq 5$ ) over 50% of the total sequenced bases. FASTQ files were sub-sequentially inspected with  
161 FASTQC v.0.11.7 (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) that highlighted  
162 very high-quality reads.

163 Reads were mapped on the latest version of the *Sus scrofa* reference genome (Sscrofa11.1) with  
164 BWA tool 0.7.17 (Li & Durbin, 2009) (function: MEM) and the parameters for paired-end data.  
165 Picard v.2.1.1 (<https://broadinstitute.github.io/picard/>) was used to remove duplicated reads. Whole  
166 genome sequencing data statistics are reported in Table S2.

167



## 168 **Detection of CNVs and CNVRs from sequencing data**

169 The cn.Mops v.1.32 tool (Klambauer *et al.* 2012) was used to identify autosomal CNVs.  
170 cn.Mops was run with default parameters except for the window size that was lowered to 750 bp.  
171 Since three consecutive genome windows positive for copy number are required by cn.Mops to assert  
172 the presence of a CNV, the minimum size of a detected CNV was 2250 bp. The 750 bp window size  
173 allowed us to detect short CNVs (CNV  $\geq$  3 kbp with default parameters) with a length fitting the  
174 definition of CNV (usually more than 1 kbp). Smaller window sizes were tested resulting in longer  
175 computational times without any specific indication on their reliability. CNVs identified in the  
176 different breeds were merged into CNVRs with Bedtools v.2.17.0 (Quinlan & Hall 2010) (function:  
177 merge) whenever overlapping genome windows, constituting the different CNVs, were encountered.

178 CNVRs were then compared with previous studies. The comparison was carried out remapping  
179 CNVRs on the Sscrofa11.1 using the NCBI genome remapping tool  
180 (<https://www.ncbi.nlm.nih.gov/genome/tools/remap>) looking for CNVRs sharing at least one  
181 nucleotide, as proposed by Keel *et al.* (2019).

182

## 183 **Cluster analysis of breeds based on CNVRs**

184 Pig breeds were clustered based on the read count ratio of each genome window covered by a  
185 CNVR. This ratio was defined as  $\frac{RC}{RC_g}$ , where  $RC$  and  $RC_g$  indicate the exact number and the average  
186 number of reads in a genome window for a specific pig breed, respectively. Hierarchical clustering  
187 was computed in R v.3.6 (R Core Team, 2018) (function: hclust) using the Ward.D2 distance (we  
188 excluded genome windows presenting a ratio  $\geq$  50 in at least one pig breed).

189

## 190 **Genomic analysis of repeated elements in CNVs/CNVRs and flanking regions**

191 The GFF file reporting the location of repeated elements interspersed in the *S. scrofa* genome  
192 was downloaded from the UCSC Genome Browser (<https://genome.ucsc.edu/>). For CNVs/CNVRs

193 and the related 1-kb flanking regions, we counted the number of bases overlapping each repeated  
194 element (Bedtools; function: intersect), assessing their enrichment via Fisher's exact test as  
195 implemented in Python 2.6 (Scipy library; function: stats.fisher\_exact; alternative hypothesis:  
196 greater). We considered statistically enriched classes of repeated elements presenting a  $P < 0.05$ ,  
197 Bonferroni corrected.

198

## 199 **Annotation of CNVRs**

200 Annotated genes overlapping the identified CNVRs were retrieved from the Sscrofa11.1  
201 NCBI's GFF file by using Bedtools (function: intersect). Functional analysis was carried out with  
202 PANTHER (Mi *et al.* 2019) via Fisher's exact test. Analyses were run over a subset of the Gene  
203 Ontology – Biological Process resource (PANTHER GO-slim v.14.1; release 2019-03-12; no. = 2004  
204 biological processes) and the Reactome database (Reactome v.65; release 2019-03-12; no. = 1569  
205 pathways). We made use of pig specific gene annotations. We considered statistically enriched terms  
206 presenting a  $P < 0.05$ , FDR corrected.

207 The presence of QTLs in CNVRs was evaluated and tested via Fisher's exact test. QTLs were  
208 downloaded from the Pig Quantitative Trait Locus Database (Pig QTLdb; release 39) (Hu *et al.* 2019)  
209 and checked. Distribution of QTL size pointed out a fraction of long QTLs (> 2 Mbp) probably due  
210 lack of resolution derived by the information retrieved from several QTL studies. These QTLs were  
211 discarded. We noted that for a given QTL class (i.e. trait) several DNA markers, defining the QTL in  
212 different breeds, were close to each other. Thus, QTLs that were less than 500 kbp of distance were  
213 merged with Bedtools (function: merge) to obtain QTL regions. The final dataset presented a total of  
214 295 traits and 1978 QTL regions. For each trait, the fraction of CNVR nucleotides overlapping QTLs  
215 was retrieved with Bedtools (function: intersect). Fisher's exact test was run in Python, retrieving  
216 statistically enriched traits presenting a  $P < 0.05$ , Bonferroni corrected.

217

## 218 **Results**

## 219 **Sequenced reads and genome wide identification of CNVs**

220 About 17.5 billion reads were produced from the sequencing of the 21 pig DNA pools. On  
221 average, each DNA pool presented about 417.7 million of mapped reads spanning 98.5% of the *S.*  
222 *scrofa* reference genome, with an average read depth of about 42×. Summary statistics of sequencing  
223 data are reported in Table S2.

224 Using cn.Mops we identified a total of 9592 CNVs (14344 events) across the 21 analysed  
225 breeds. On average, each pig breed had 683 CNVs (median = 601; min. = 209, Sarda; max. = 1440  
226 Turopolje) covering 0.18% (s.d. = 0.09%) of the reference genome, with the smallest fraction in Sarda  
227 (0.04%) and the largest coverage in Turopolje (0.40%), reflecting the lowest and highest number of  
228 CNVs, respectively (Table 1). For each pig breed, CNVs were divided in losses (copy number < 2,  
229 as inferred by cn.Mops) and gains (copy number > 2, as inferred by cn.Mops) that represented the  
230 most frequent copy number (CN) state characterizing the animals analysed in the pools. On the whole,  
231 we identified a total of 3492 losses, 5012 gains and 638 showing a mix of copy number loss and gain.  
232 The losses/gains ratio was around 0.79. Stratified by chromosome, this value ranged from 0.57 to  
233 1.02, for SSC12 and SSC1, respectively (Table S3). Considering the CNVs detected in each breed,  
234 the number of losses and gains strongly correlated ( $r = 0.93$ ). CNV length ranged from 2250 to  
235 560250 bp. The longest CNV (560250 Mbp) was detected on SSC8 in the Italian Large White and  
236 Lithuanian White Old Type pig breeds (Table 1). The number of CNVs and the chromosome length  
237 had a medium-high Pearson's correlation coefficient ( $r = 0.69$ ;  $P < 0.05$ ).

238

## 239 **Identification of CNVRs**

240 CNVs were merged across breeds resulting in a total of 3710 CNVRs (Table S4). The  
241 distribution of CNVRs along each chromosome is presented in Fig. 2. SSC1, SSC2 and SSC3 had the  
242 largest number of detected CNVRs (no. = 359, no. = 361 and no. = 307, respectively; Table 2). The  
243 number of CNVRs and the chromosome length highly correlated ( $r = 0.87$ ;  $P < 0.05$ ). Positive  
244 correlation ( $r = 0.92$ ,  $P < 0.05$ ) was observed also between the number of CNVRs and their total

245 length. On average, each pig breed had 586 CNVRs (min. = 180 in Sarda; max. = 1257 in Turopolje;  
246 Table S5). Among the 3710 CNVRs, 1615 (43.5%) were breed specific (and indicated as private  
247 CNVRs; Table S5). Size of CNVRs ranged from 2250 bp up to 560250 bp (the same of CNVs), with  
248 an average length of 7038 bp and a median value of 3750 bp (Table 2). Distribution of CNVR size  
249 showed a decrease in CNVR counts while increasing their size. CNVRs occupied a total of 26.1 Mbp,  
250 equal to 1.15% of the Sscrofa11.1 reference genome. Among the CNVRs, based on the copy number  
251 state (i.e. the number of copies; CN state) provided by cn.MOPS, 1305 (35.2%) had only copy number  
252 gains (duplication), 1323 (35.6%) had only copy number losses (deletion), and 1082 (29.2%) showed  
253 a mix of copy number losses and gains from different pig breeds.

254         The 3710 detected CNVRs encompassed a total of 34821 genome windows. After filtering, the  
255 read count ratio of each genome window was used to cluster pig breeds (Fig. 3), which grouped breeds  
256 in agreement to their main specific phenotypes or their geographic origin. A first group encompassed  
257 breeds that have a coat colour with white background or white patterns (Lithuanian Indigenous  
258 Wattle, Italian Large White, Krškopolje, Bísara and Lithuanian White Old Type). This may be due to  
259 the strong signals of genome windows encompassing the *KIT* gene, that accounts for ~15% of the  
260 total positive windows for CNVs. The two reddish brown coloured breeds (Mora Romagnola and  
261 Italian Duroc) were on the same branch. Three autochthonous Italian breeds (Casertana, Nero  
262 Siciliano and Sarda) constituted a cluster whereas one Portuguese and one Spanish breed (Alentejana  
263 and Majorcan Black, respectively) constituted another cluster. The Turopolje pig breed was the only  
264 one that clustered apart from all other breeds.

265

## 266 **Repeated elements within and flanking CNVs and CNVRs**

267         Highly repetitive sequences were investigated for their co-occurrence with CNVs and CNVRs  
268 (Table S6). The following classes of repeated elements were statistically over-represented within  
269 CNVs: long interspersed nuclear elements (LINE), long terminal repeats (LTR), satellites, rolling-  
270 circle (RC/Helitron) and pseudogenes (tRNAs, snRNAs, srpRNAs, and rRNAs). Additionally, CNV

271 flanking regions (1-kbp per side) were enriched for the following classes: short interspersed nuclear  
272 elements (SINE), simple repeat and low complexity. CNVRs differed for the absence of RC elements  
273 and the absence of SINE and srpRNAs in the 1-kbp flanking regions. However, SINE were over-  
274 represented when the flanking region size was extended to 10-kbp.

275

#### 276 **QTLs in CNVRs**

277 A total of 1978 QTL regions, associated to 554 phenotypic traits, were retrieved from the pig  
278 QTL database. CNVRs overlapped a total of 336 QTL regions representing 295 phenotypic traits.  
279 Enrichment analysis identified 126 traits (~ 43%) significantly over-represented ( $P < 0.05$ , Bonferroni  
280 corrected). These traits spanned different classes, including meat quality, body shape and  
281 conformation, reproduction, disease susceptibility, haematological and metabolism related traits  
282 (Table S7).

283

#### 284 **Functional annotation of CNVRs and detailed analysis of selected genes**

285 A total of 1571 genes overlapped the identified CNVRs, including 1296 protein coding genes,  
286 261 lncRNAs, 3 miRNAs and 11 tRNAs. The number of overlapped genes correlated with the number  
287 of CNVRs ( $r = 0.99$ ). A total of 993 protein-coding genes were annotated by PANTHER and used  
288 for functional enrichment over the GO slim Biological process resource. A total of 17 terms were  
289 over-represented (Table S8), encompassing different biological processes such as sensory perception,  
290 nervous system process, fatty acid metabolic process, gene expression and biological adhesion. Over  
291 the Reactome database, PANTHER over-represented the olfactory signalling pathway and the related  
292 mechanism of transduction mediated by G protein-coupled receptors (Table S8). Analysis of genes  
293 located in private CNVRs did not identify any over-represented process/pathway.

294 The *v-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog (KIT)* and the  
295 *methionine sulfoxide reductase B3 (MSRB3)* genes were two important genes presenting variable  
296 copies among breeds. CNVs affecting the *KIT* gene are responsible for different coat colour

297 phenotypes (Johansson Moller *et al.* 1996; Marklund *et al.* 1998; Johansson *et al.* 2005; Rubin *et al.*  
298 2012) whereas variable copies of the *MSRB3* have been associated with ear size in pigs (Chen *et al.*  
299 2018).

300 The detailed analysis of the *KIT* gene indicated the presence of the four duplicated regions  
301 (DUP1-4; Fig. 4a) previously described by Rubin *et al.* (2012). Structural variants as well as the  
302 presence of the splice mutation at the first base in intron 17 (g.41486012G>A, rs345599765) are all  
303 required for manifesting a solid white coat colour (Marklund *et al.* 1998). Using sequence data, we  
304 estimated the allele frequencies of this SNP (Fig. 4a; Table S9) to complement CNV results. Pools  
305 from colored pigs did not show any CNV and the splice mutation. White pigs (Italian Large White  
306 and Lithuanian White Old Type) had DUP1-4 and the splice mutation (allele A). However, allele  
307 frequencies were divergent (Table S9) suggesting a different structure of the CNV (different gene  
308 copies with the the A or G nucleotides). The Sarda (not fixed for any coat colour and including many  
309 spotted animals) and Lithuanian Indigenous Wattle breeds presented DUP1, did not have DUP2-4  
310 and had allele A (the only two other breeds having the splice mutation). Bísara, another spotted breed,  
311 had also DUP2-3. The piebald breed Basque and the belted breed Cinta Senese had DUP2-4, whereas  
312 the other two belted breeds (Krškopolje and Schwäbisch-Hällisches Schwein) had only DUP2 and  
313 DUP4.

314 The detailed analysis of the *MSRB3* gene region revealed the presence of the 38.4-kbp  
315 duplication (SSC5:29826981-29865653; Fig. 4b) previously described by Chen *et al.* (2018). Copy  
316 number gains encompassing the *MSRB3* exons 6 and 7 have been associated with large ear size in  
317 Chinese pig breeds and with half-floppy ears in Landrace pigs (Chen *et al.* 2018). Alentejana, Cinta  
318 Senese, Mora Romagnola, Italian Duroc and Italian Large White that are breeds characterized by  
319 small/medium ear size, had a normal copy number state (that means no gain of copies). The remaining  
320 pig breeds showed variable copy number which seems to be correlated to ear size (Fig. 4c).  
321 Regression analysis between the average CN state and the ear size (coded as follows: small = 1,  
322 medium = 1.5, medium/large = 1.75 and large = 2) resulted in a positive association ( $P = 0.0001$ ).

323 However, other breeds characterized by small ears (i.e. Nero Siciliano and Sarda) had variable copy  
324 numbers. Variability in ear size was also analysed by estimating the allele frequency of two SNP in  
325 the 5' flanking region (g.29695369C>T; rs340841870) and in the 3'-untranslated region  
326 (g.29862412C>T; rs326411202) of the *MSRB3* gene, that Zhang *et al.* (2015) reported to be  
327 associated with ear size. These SNP positions are not included in the CNVR of this gene. For each  
328 SNP, the regression analysis pointed out a significant association between allele frequencies and ear  
329 size ( $P < 0.0001$ ). Additionally, allele frequencies of these two SNPs (Table S9) correlated with the  
330 average CN state ( $|r| > 0.8$ ;  $P < 0.0001$ ; Fig. 4d).

331 We further explored genomic regions harbouring private information on CNVRs. Among them,  
332 we identified two interesting examples. The first one, characterizing Lithuanian Indigenous Wattle  
333 pigs, encompassed the intron 10 of the *ELOVL fatty acid elongase 6 (ELOVL6)* gene (Fig. 5a).  
334 Variants in this gene has been associated with fatty acid composition in pigs (Corominas *et al.* 2013).  
335 The second one, characterizing Turopolje pigs, was the *Zinc finger protein 622 (ZNF622)* gene, a  
336 regulator of early embryonic development (Hasegawa *et al.* 2015). The CNV affecting this gene was  
337 quite complex. Copy number gains were in the correspondence of the exonic regions but also included  
338 the complete intron 1, intron 2 and intron 5. Most of introns 3 and 4 were not affected by CN gains  
339 (only small and contiguous intronic segments to the exonic regions were included in the CN gains)  
340 (Fig. 5b). The regions with CN gains were clearly evidenced in all breeds except Turopolje, which  
341 did not have any copy number and, in part, in Krškopolje and Italian Duroc, that had CN higher than  
342 that of Turopolje but lower than that of all other breeds (Fig. 5b).

343

#### 344 **Comparison with other studies**

345 The positions of CNVRs we detected were compared with the CNVRs reported by previous  
346 studies, which analysed different pig breeds and other species of the *Sus* genus using whole genome  
347 sequencing. A total of five datasets, which investigated Asian pig breeds, commercial and European  
348 pig breeds, and five species of the genus *Sus*, were considered for this comparison (Table S4). The

349 overlap ranged from about 10 to 25% (Table S10). Overall, a total of 595 CNVRs detected in our  
350 work (16%) overlapped with CNVRs reported by the considered studies (Table S4).

351

## 352 **Discussion**

353 In this study we carried out a genome-wide CNV/CNVR analysis in 19 European  
354 autochthonous and two Italian commercial pig breeds. Breeds were analysed by using a whole  
355 genome sequencing strategy from breed specific DNA pools to maximize CNV discovery. CNVs  
356 were detected via cn.MOPS, a tool that implements a Bayesian approach that models depth of  
357 coverage across samples by decomposing its variability in a part coming from copy numbers and the  
358 remaining part due to noise, in order to reduce false discoveries (Klambauer *et al.* 2012). Other  
359 software based on different assumptions have been also developed and used for CNV detection from  
360 HTS datasets. However, there is no consensus in the literature on the strategy and methodology that  
361 might be applied for this purpose.

362 As our study was based on DNA pools from a large number of populations, we maximized the  
363 power of cn.MOPS in reducing the false discovery rate, as this tool is specifically designed to deal  
364 with multiple samples.

365 Even if this design could not precisely define the exact number of copy gains or losses for all  
366 animals in the sequenced pools, the obtained results made it possible to capture within breed averaged  
367 states. This was supported by the agreement among the different coat colour phenotypes and the  
368 expected CN states, rightly detected at the *KIT* locus which indirectly confirmed and validated CNV  
369 calls from cn.MOPS. This approach demonstrated that CNVs detected using whole genome  
370 sequencing can be useful to identify breed specific features (including in this definition the most  
371 frequent breed features) and describe genetic diversity across pig breeds, complementing SNP based  
372 studies.

373 We confirmed a high correspondence between CNV data detected from sequenced DNA pools  
374 and SNP information using Pearson's correlation calculated considering the fraction of the pig



375 genome covered by CNVRs detected for each breed (Table 1) and SNP based diversity measured on  
376 the same animals genotyped with the GeneSeek® GGP Porcine HD Genomic Profiler (Muñoz *et al.*  
377 2019). Among these SNP averaged variability parameters, correlation with the above mentioned  
378 CNVR parameter was highly negative with both the minor allele frequency (MAF;  $r = -0.90$ ) and  
379 expected heterozygosity ( $r = -0.90$ ), whereas highly positive correlation with the Fixation Index ( $F_{ST}$ ;  
380  $r = 0.96$ ) values. These correlations mean that when within breed variability was low, it increased the  
381 possibility to identify losses/gains at variable CN state and that the fraction of the genome covered  
382 by CNVRs detected in DNA pools is a good indicator of the diversity among breeds.

383 With few differences, these breeds were clustered resembling the relationships that we already  
384 reported using array SNP datasets obtained from individually genotyped pigs and SNPs detected from  
385 whole genome sequencing (Muñoz *et al.* 2018, 2019; Bovo *et al.*, in preparation). Geographical and  
386 some major morphological features (i.e. coat colour) mainly determined breed clusters obtained from  
387 CN states. Turopolje, the breed that accounted for the largest number of CNVs (with the largest  
388 fraction of the genome covered by CNVRs), was clustered apart, as also reported with SNP data  
389 (Muñoz *et al.* 2018, 2019; Bovo *et al.*, in preparation).

390 Some CNVRs were considered as breed specific or identified in a few breeds, suggesting that  
391 this variability might contribute to determine several phenotypic characteristics that distinguish  
392 autochthonous and commercial European breeds. In addition, considering the whole patterns of  
393 CNVRs that we detected, a quite high frequency of these events was classified as mixed CNVs  
394 (including both gains and losses). This indicates that despite breeds share genome regions affected  
395 by CNV, the single breed carries a gain or a loss specific for the breed itself.

396 In the current study, an average of 77 CNVRs (~16% of all breed reported CNVRs) was  
397 considered as private for each analysed breed, highlighting the power of the DNA pooling strategy in  
398 capturing distinctive breed features. However, as the sequencing depth is not so high, for a given  
399 private CNVR we cannot completely exclude the possibility that few animals of the other investigated  
400 breeds could carry the same alleles in these regions. The remaining CNVRs were shared among two

401 or more breeds, indicating that admixture and crossbreeding events or a common origin might have  
402 contributed to spread this variability. However, further studies are needed to clarify their allelic status  
403 or their common origin, as in our first survey we did not characterize into detail the precise breakdown  
404 positions and structure of all identified CNVRs.

405 CNVRs we detected overlapped genes involved in different biological processes including  
406 nervous system and sensory perception such as olfactory signalling. Brain functions control several  
407 behaviours, including feeding, habitat selection, reproduction and social interaction that strongly  
408 depend on the genetics architecture of an individual (Bendesky & Bargmann 2011). Several studies  
409 in mammals including pigs reported CNVs in genes involved in the olfactory signalling pathway,  
410 linking gene variability to food foraging and mate recognition abilities (Paudel *et al.* 2015; Keel *et*  
411 *al.* 2019). In addition, considering the overlapping of CNVRs and QTL regions, the main traits  
412 associated with changes in CN state were meat quality, body shape and conformation, reproduction  
413 and metabolism. Variability in chromosome regions harbouring functionally relevant genes or QTL  
414 may reflect the adaptation of these breeds to different production systems and environments.

415 The impact of this type of variability on exterior characteristics of the pigs has been already  
416 demonstrated for the CNVs in the *KIT* gene region affecting coat colours and patterns, which  
417 characterize the *Dominant white* phenotype (Rubin *et al.* 2012). Other evidences came for the CNVs  
418 in the *MS3B3* gene region, involved in ear size as mainly reported in Chinese breeds (Chen *et al.*  
419 2018). These CNVRs were also detected in our study with some interesting new information for some  
420 of the analysed breeds.

421 The complexity of the *Dominant white KIT* locus has been explained by the presence of six  
422 main allele groups (in addition to a few other potential variants; Fontanesi & Russo 2013): (i) a  
423 recessive wild-type allele *i* (that is carried by wild boar and coloured pigs), (ii) the *Patch* allele  $I^P$   
424 (determining spotted patterns), (iii) the *Belt* allele  $I^{Be}$  (determining the belted phenotype), (iv) the  
425 *Roan/Gray* allele  $I^{Rn}$  or  $I^d$  (causing the grey-roan phenotype), (v) the dominant white alleles *I*,  
426 comprising several forms (e.g.  $I^1$ ,  $I^2$  and  $I^3$ ) and causing the white solid phenotype that mainly

427 characterize Large White and Landrace breeds and (vi) the  $I^L$  allele, a null and lethal allele (Johansson  
428 Moller *et al.* 1996; Marklund *et al.* 1998; Johansson *et al.* 2005; Rubin *et al.* 2012). Variants in this  
429 chromosome region are mainly associated with a 450-kbp duplication encompassing the entire *KIT*  
430 gene (DUP1; the only CN of the  $I^P$  allele), including also another 4.3-kbp duplication (DUP2) located  
431 ~100 kbp upstream of *KIT* gene, and a 23-kbp duplication (DUP3) ~100 kbp downstream from *KIT*,  
432 which in turn resulted to contain another 4.3-kbp duplication (DUP4; Rubin *et al.* 2012). The *I* alleles  
433 presented variable copy numbers of DUP1/2/3/4, whereas DUP2/3/4 were identified in pigs with the  
434  $I^{Be}$  allele (Rubin *et al.* 2012). Moreover, a recent whole genome resequencing study uncovered new  
435 *KIT* alleles conferring different coat colour phenotypes (Wu *et al.* 2019). The CN state states that we  
436 identified in our study encompassed all four duplicated regions, describing for the first time the  
437 structure of the *KIT* gene in several autochthonous pig breeds (Fig. 4a).

438 In addition, analysis of sequencing data let us to estimate the frequency of the splice mutation  
439 g.41486012G>A (rs345599765) that distinguish the CN state of the  $I^P$  from the *I Dominant white*  
440 allele series (Marklund *et al.* 1998). As expected, all breeds that did not show any duplicated regions  
441 are characterized by solid coat colours and did not have the splice mutation. They are considered to  
442 carry only the *i* wild-type at the *Dominant white* locus. Sarda, which is a breed not fixed for any coat  
443 colours and that includes also white and white spotted pigs, showed the presence of DUP1, with some  
444 faint signs at the DUP4 position (with a low frequency of the splice mutation). Several alleles at the  
445 *KIT* gene might be present in this breed, including  $I^P$ , *I* variants and  $I^{Be}$  forms. A similar pattern was  
446 observed in the Lithuanian Indigenous Wattle breed, which includes mainly spotted pigs. According  
447 to the CN state observed in this breed,  $I^P$  might be the most frequent allele, even if other and  $I^{Be}$  and  
448 *I* forms (including also DUP4) might be present. A more marked copy number pattern was evidenced  
449 for the Bísara breed (which has mainly heterogeneous coats: grey or black and white or spotted) that  
450 reported DUP1 copy number status similar to Sarda and Lithuanian Indigenous Wattle) in addition to  
451 DUP2-3 (without signals indicating the presence of DUP4).

452 The analysis of the *KIT* gene region in breeds characterized by a belted phenotype, even if not  
453 homogeneous, indicated that more alleles at this locus might produce belted pigs even if with some  
454 different phenotypic effects. Cinta Senese and Basque had equal CN state at DUP2-3 but differed in  
455 DUP4 (higher in Basque and lower in Cinta Senese). Cinta Senese is a classical belted breed whereas  
456 Basque pigs are usually black and white with heterogenous patterns but usually with black head and  
457 rump. Other breeds having white belts of varying size and shape (Krškopolje and Schwäbisch-  
458 Hällisches Schwein) showed only DUP2 and DUP4. The connection between the two breeds might  
459 be derived by ancestral origins (not clearly defined), that preserved the same structure at the *Dominant*  
460 *white* locus. Wu *et al.* (2019) observed that the presence of DUP2 together with DUP4 can produce  
461 a belted phenotype in Duroc × (Landrace × Large White) hybrid pigs. The presence of multiple alleles  
462 conferring a belted phenotype is also confirmed by the results of the analysis of the rs328592739 SNP  
463 in the *KIT* gene that was associated with the belted pattern in Cinta Senese pigs (Fontanesi *et al.* 2016)  
464 but not in Krškopolje and Schwäbisch-Hällisches Schwein pigs (Ogorevc *et al.* 2017).

465 White breeds (Italian Large White and Lithuanian White Old Type) had a classical copy number  
466 pattern in DUP1-4 and the splice mutation already described for completely white pigs carrying *I*  
467 alleles (Fontanesi *et al.* 2010a). Heterogeneity on the presence of the splice mutation suggested that  
468 *Dominant white* alleles having different G/A ratios at this position. In Lithuanian White Old Type,  
469 gene copies at this position carried G only in 1 out of 5 copies (as estimated from its 0.20 frequency).  
470 In Italian Large White, about 2 out of 3 gene copies carried the G nucleotide (G = 0.68), suggesting  
471 that the CNV structure in this breed might be determined by different *Dominant white* alleles than  
472 those frequently present in the Lithuanian White Old Type breed.

473 Interesting copy number patterns were also observed in the region of SSC5 encompassing the  
474 last exons of the *MSRB3* gene (Fig. 4b), which is associated with ear size (Chen *et al.* 2018). These  
475 authors proposed that large ear size is due to the increased CN state in this region, which affects the  
476 expression of the nearby miR-584-5p that in turn inhibits the expression of its target gene *MSRB3*.  
477 Our CNV analysis for the *MSRB3* gene across autochthonous European pig breeds indicated, with

478 the exception of some breeds, a significant correlation between ear size and the average CN state  
479 (Fig. 4c). The latter also correlated with allele frequencies estimated for the rs340841870 and  
480 rs326411202 SNPs (outside this CNVR), which suggested the presence of linkage between these two  
481 types of variants: allele C at both positions is associated with a normal copy state whereas the  
482 alternative allele at both sides (T) is associated with the presence of 5 or 6 copies (of the linked  
483 multiple copy region), as estimated from the sequencing data in the CNVR. Even if pigs of the studied  
484 breeds were in general described to have breed-specific traits, heterogeneity for ear size has been  
485 already reported in some breeds which might not actually have fixed ear size shape (Schiavo *et al.*  
486 2019). Therefore, correlation between CN state and ear size might not precisely estimated by the  
487 DNA pooling approach (Fig. 4b). It is also worth mentioning that ear size and position have been  
488 already shown to be under polygenic control with a few major genes affecting these traits (e.g. Wei  
489 *et al.* 2007; Ma *et al.* 2009; Ren *et al.* 2011). Thus, other genomic regions and polymorphisms could  
490 be responsible for the ear size phenotype in some of the analysed breeds.

491 The CNV in the *ELOVL6* gene might interesting to explain economically relevant traits,  
492 considering the role of this gene in affecting fatty acid composition in pigs (Corominas *et al.* 2013).  
493 Other studies reported that variability in this gene or variability in its expression level might explain,  
494 at least in part, differences of intramuscular fat accumulation and lipid metabolism among breeds,  
495 which are relevant for meat quality, considering also genotype-feeding interactions to design  
496 appropriate fatty-acid diets in pigs to maximize this aspect (e.g. Benítez *et al.* 2016; Muñoz *et al.*  
497 2018; Revilla *et al.* 2018). Association studies and functional analysis of the CNV in this gene are  
498 needed to understand if this variability could be involved in affecting meat quality traits in pigs.  
499 Targeted analyses are also needed to detect with more precision if this variability segregates within  
500 the analysed breeds as well as in other breeds in which meat quality parameters are important factors  
501 determining the quality of their products.

502 Detailed analyses of CN states of some chromosome regions can also identify (or suppose) the  
503 occurrence of other or more complex mutational events that might not be properly considered as

504 derived by CNVs. The case of the *ZNF622* gene that reported three distinct copy number gains  
505 (mainly in the correspondence of exonic regions) might raise a few hypotheses on the occurrence of  
506 this strange pattern. The three divided copy number gains might be due to the presence of a  
507 pseudogene derived by the *ZNF622* gene (inserted somewhere into the genome) or that the  
508 duplication of the gene subsequently underwent other mutational events that eliminated most of the  
509 sequence of introns 3 and 4 (Fig. 5b). Other studies are needed to clarify these hypotheses. After a  
510 preliminary analysis, CN states reported in the correspondence of this gene appeared to produce a  
511 private condition in the Turopolje breed that did not have any copy number gain (common in all other  
512 breeds). Inspection of the clustering analysis for the CN at this gene in all breeds, indicated that two  
513 other breeds (Krškopolje and Italian Duroc) might not have fixed copy number gains, mainly in the  
514 correspondence of the annotated exons of the *ZNF622* gene.

515 On the whole, our survey on European pig breeds reported that CNVRs occupy 26.1 Mbp,  
516 representing 1.15% of the reference genome size. Compared to other whole genome sequencing based  
517 studies, this genome fraction is similar to what was reported by Paudel *et al.* (2015) and Keel *et al.*  
518 (2019) (17.83 and 22.9 Mbp, respectively). Other two studies (Paudel *et al.* 2013; Jiang *et al.* 2014)  
519 identified larger fractions of the pig genome covered by CNVRs (39.2 and 102.8 Mbp, respectively).  
520 Although this divergence could be attributed in part to the algorithms used to detect CNVs and the  
521 sequencing approaches (single pigs vs pools of individuals), it might be also due to differences among  
522 the studied pig populations. Distribution of CNVR sizes showed a decrease in CNVR counts while  
523 increasing their size, as also described by Jiang *et al.* (2014). Differences among breeds were also  
524 clearly shown in our study, as detailed above. Some of the CNVRs we detected in our study  
525 overlapped with CNV events reported by the other whole genome sequencing mentioned studies (on  
526 average, ~13% of overlap), pointing out that they could exist also in other breeds that we did not  
527 survey. However, they represent just fraction a small fraction, strengthening the evidence that CNV  
528 are breed-specific genome features. Additional studies are needed to obtain a global overview of  
529 CNVs segregating in the *Sus scrofa* species, by comparing more breeds and populations.

530 As CNVs mutate about 2-3 times faster than SNPs, some of the CNVRs that we detected across  
531 several breeds could eventually also be derived from recurrent mutational events through nonallelic  
532 homologous recombination, potentially driven by the presence of repeated regions within or in  
533 flanking positions (Liu *et al.* 2012). Analyses of CNVRs and their flanking regions identified  
534 enrichments of different classes of repeated elements, confirming what other studies reported this  
535 species (e.g. Paudel *et al.* 2013; Wang *et al.* 2015b). This further suggest that these sequence features  
536 might contribute to chromosome instability and mutational mechanisms promoting these structural  
537 changes also in *Sus scrofa*.

538 Our study investigated CNVs in the porcine genome over a large number of pig breeds that  
539 represent important European genetic resources for this species. This variability can explain part of  
540 the genetic diversity among breeds and might contribute to explain their origin, history and adaptation  
541 to a variety of production systems. Further studies are needed to better understand how CNVs could  
542 be considered in defining conservation programmes of these autochthonous genetic resources.

543

#### 544 **Acknowledgements**

545 SB received a fellowship from the Europe-FAANG COST Action. This work has received  
546 funding from the University of Bologna RFO 2016-2019 programme, the Italian MIUR 2017  
547 *PigPhenomics* project, the Slovenian Agency of Research (grant P4-0133) and from the European  
548 Union's Horizon 2020 research and innovation programme under grant agreement No. 634476 for  
549 the project with acronym TREASURE. The content of this article reflects only the authors' view and  
550 the European Union Agency is not responsible for any use that may be made of the information it  
551 contains.

552

#### 553 **Availability of data**

554 Sequence data generated and analysed in the current study are available in the EMBL-EBI European  
555 Nucleotide Archive (ENA) repository (<http://www.ebi.ac.uk/ena>), under the study accession

556 PRJEB36830. CNVRs are available as Supplementary Table S4 and from the corresponding author  
557 on reasonable request.

558

559

560

## 561 **Competing interests**

562 The authors declare they do not have any competing interests.

563

## 564 **References**

565 Alkan C., Coe B.P. & Eichler E.E. (2011) Genome structural variation discovery and genotyping.  
566 *Nature Reviews Genetics* **12**, 363–76.

567 Andersen I.L., Nævdal E. & Bøe K.E. (2011) Maternal investment, sibling competition, and offspring  
568 survival with increasing litter size and parity in pigs (*Sus scrofa*). *Behavioral Ecology and*  
569 *Sociobiology* **65**, 1159–67.

570 Bendesky A. & Bargmann C.I. (2011) Genetic contributions to behavioural diversity at the gene-  
571 environment interface. *Nature Reviews Genetics* **12**, 809–20.

572 Benítez R., Núñez Y., Fernández A., Isabel B., Rodríguez C., Daza A., López-Bote C., Silió L. &  
573 Óvilo C. (2016) Adipose tissue transcriptional response of lipid metabolism genes in growing  
574 Iberian pigs fed oleic acid v. carbohydrate enriched diets. *Animal* **10**, 939–46.

575 Bickhart D.M., Hou Y., Schroeder S.G., Alkan C., Cardone M.F., Matukumalli L.K., Song J.,  
576 Schnabel R.D., Ventura M., Taylor J.F., Garcia J.F., Van Tassell C.P., Sonstegard T.S., Eichler  
577 E.E. & Liu G.E. (2012) Copy number variation of individual cattle genomes using next-  
578 generation sequencing. *Genome Research* **22**, 778–90.

579 Bickhart D.M. & Liu G.E. (2014) The challenges and importance of structural variation detection in  
580 livestock. *Frontiers in Genetics* **5**, 37.



- 581 Čandek-Potokar M. & Nieto Liñan R.M. (2019) European Local Pig Breeds - Diversity and  
582 Performance. A study of project TREASURE. IntechOpen, doi:10.5772/intechopen.83749.
- 583 Chen C., Liu C., Xiong X., Fang S., Yang H., Zhang Z., Ren J., Guo Y. & Huang L. (2018) Copy  
584 number variation in the *MSRB3* gene enlarges porcine ear size through a mechanism involving  
585 miR-584-5p. *Genetics Selection Evolution* **50**, 72.
- 586 Chen C., Qiao R., Wei R., Guo Y., Ai H., Ma J., Ren J. & Huang L. (2012) A comprehensive survey  
587 of copy number variation in 18 diverse pig populations and identification of candidate copy  
588 number variable genes associated with complex traits. *BMC Genomics* **13**, 733.
- 589 Corominas J., Ramayo-Caldas Y., Puig-Oliveras A., Pérez-Montarelo D., Noguera J.L., Folch J.M.  
590 & Ballester M. (2013) Polymorphism in the *ELOVL6* gene is associated with a major QTL  
591 effect on fatty acid composition in pigs. *PLoS One* **8**, e53687.
- 592 Dong K., Pu Y., Yao N., Shu G., Liu X., He X., Zhao Q., Guan W. & Ma Y. (2015) Copy number  
593 variation detection using SNP genotyping arrays in three Chinese pig breeds. *Animal Genetics*  
594 **46**, 101–9.
- 595 Fadista J., Nygaard M., Holm L.E., Thomsen B. & Bendixen C. (2008) A snapshot of CNVs in the  
596 pig genome. *PLoS One* **3**, e3916.
- 597 Fadista J., Thomsen B., Holm L.-E. & Bendixen C. (2010) Copy number variation in the bovine  
598 genome. *BMC Genomics* **11**, 284.
- 599 Fernández A.I., Barragán C., Fernández A., Rodríguez M.C. & Villanueva B. (2014) Copy number  
600 variants in a highly inbred Iberian porcine strain. *Animal Genetics* **45**, 357–66.
- 601 Feuk L., Carson A.R. & Scherer S.W. (2006) Structural variation in the human genome. *Nature*  
602 *Reviews Genetics* **7**, 85–97.
- 603 Fontanesi L., Beretti F., Martelli P.L., Colombo M., Dall'olio S., Occidente M., Portolano B., Casadio  
604 R., Matassino D. & Russo V. (2011) A first comparative map of copy number variations in the  
605 sheep genome. *Genomics* **97**, 158–65.

- 606 Fontanesi L., D'Alessandro E., Scotti E., Liotta L., Crovetto A., Chiofalo V. & Russo V. (2010a)  
607 Genetic heterogeneity and selection signature at the *KIT* gene in pigs showing different coat  
608 colours and patterns. *Animal Genetics* **41**, 478–92.
- 609 Fontanesi L., Martelli P.L., Beretti F., Riggio V., Dall'Olio S., Colombo M., Casadio R., Russo V. &  
610 Portolano B. (2010b) An initial comparative map of copy number variations in the goat (*Capra*  
611 *hircus*) genome. *BMC Genomics* **11**, 639.
- 612 Fontanesi L., Martelli P.L., Scotti E., Russo V., Rogel-Gaillard C., Casadio R. & Vernesi C. (2012)  
613 Exploring copy number variation in the rabbit (*Oryctolagus cuniculus*) genome by array  
614 comparative genome hybridization. *Genomics* **100**, 245–51.
- 615 Fontanesi L. & Russo V. (2013) Molecular genetics of coat colour in pigs. *Acta Agriculturae*  
616 *Slovenica* **4**, 16.
- 617 Fontanesi L., Scotti E., Gallo M., Nanni Costa L. & Dall'Olio S. (2016) Authentication of “mono-  
618 breed” pork products: Identification of a coat colour gene marker in Cinta Senese pigs useful  
619 to this purpose. *Livestock Science* **184**, 71–7.
- 620 Fowler K.E., Pong-Wong R., Bauer J., Clemente E.J., Reitter C.P., Affara N.A., Waite S., Walling  
621 G.A. & Griffin D.K. (2013) Genome wide analysis reveals single nucleotide polymorphisms  
622 associated with fatness and putative novel copy number variants in three pig breeds. *BMC*  
623 *Genomics* **14**, 784.
- 624 Hasegawa Y., Taylor D., Ovchinnikov D.A., Wolvetang E.J., de Torrenté L., Mar J.C. (2015)  
625 Variability of gene expression identifies transcriptional regulators of early human embryonic  
626 development. *PLoS Genetics* **11**, e1005428.
- 627 Hu Z.L., Park C.A. & Reecy J.M. (2019) Building a livestock genetic and genomic information  
628 knowledgebase through integrative developments of Animal QTLdb and CorrDB. *Nucleic*  
629 *Acids Research* **47**, D701–10.

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

633 Johansson Moller M., Chaudhary R., Hellmén E., Höyheim B., Chowdhary B. & Andersson L. (1996)  
634 Pigs with the dominant white coat color phenotype carry a duplication of the KIT gene encoding  
635 the mast/stem cell growth factor receptor. *Mammalian Genome* **7**, 822–30.

636 Johansson A., Pielberg G., Andersson L. & Edfors-Lilja I. (2005) Polymorphism at the porcine  
637 Dominant white/KIT locus influence coat colour and peripheral blood cell measures. *Animal*  
638 *Genetics* **36**, 288–96.

639 Keel B.N., Lindholm-Perry A.K. & Snelling W.M. (2016) Evolutionary and Functional Features of  
640 Copy Number Variation in the Cattle Genome. *Frontiers in Genetics* **7**, 207.

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

644 Li H. & Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform.  
645 *Bioinformatics* **25**, 1754–60.

646 Li Y., Mei S., Zhang X., Peng X., Liu G., Tao H., Wu H., Jiang S., Xiong Y. & Li F. (2012)  
647 Identification of genome-wide copy number variations among diverse pig breeds by array CGH.  
648 *BMC Genomics* **13**, 725.

649 Liu P., Carvalho C.M.B., Hastings P.J. & Lupski J.R. (2012) Mechanisms for recurrent and complex  
650 human genomic rearrangements. *Current Opinion in Genetics & Development* **22**, 211–20.

651 Liu M., Zhou Y., Rosen B.D., Van Tassell C.P., Stella A., Tosser-Klopp G., Rupp R., Palhière I.,  
652 Colli L., Sayre B., Crepaldi P., Fang L., Mészáros G., Chen H., Liu G.E. & ADAPTmap  
653 Consortium. (2019) Diversity of copy number variation in the worldwide goat population.  
654 *Heredity* **122**, 636–46.

655 Long Y., Su Y., Ai H., Zhang Z., Yang B., Ruan G., Xiao S., Liao X., Ren J., Huang L. & Ding N.  
656 (2016) A genome-wide association study of copy number variations with umbilical hernia in  
657 swine. *Animal Genetics* **47**, 298–305.

658 Ma J., Qi W., Ren D., Duan Y., Qiao R., Guo Y., Yang Z., Li L., Milan D., Ren J. & Huang L. (2009)  
659 A genome scan for quantitative trait loci affecting three ear traits in a White Duroc x Chinese  
660 Erhualian resource population. *Animal Genetics* **40**, 463–7.

661 Marklund S., Kijas J., Rodriguez-Martinez H., Rönnstrand L., Funa K., Moller M., Lange D., Edfors-  
662 Lilja I. & Andersson L. (1998) Molecular basis for the dominant white phenotype in the  
663 domestic pig. *Genome Research* **8**, 826–33.

664 Mi H., Muruganujan A., Ebert D., Huang X. & Thomas P.D. (2019) PANTHER version 14: more  
665 genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic  
666 Acids Research* **47**, D419–26.

667 Molnár J., Nagy T., Stéger V., Tóth G., Marincs F. & Barta E. (2014) Genome sequencing and  
668 analysis of Mangalica, a fatty local pig of Hungary. *BMC Genomics* **15**, 761.

669 Muñoz M., Bozzi R., García F., Núñez Y., Geraci C., Crovetto A., García-Casco J., Alves E., Škrlep  
670 M., Charneca R., Martins J.M., Quintanilla R., Tibau J., Kušec G., Djurkin-Kušec I., Mercat  
671 M.J., Riquet J., Estellé J., Zimmer C., Razmaite V., Araujo J.P., Radović Č., Savić R., Karolyi  
672 D., Gallo M., Čandek-Potokar M., Fontanesi L., Fernández A.I. & Óvilo C. (2018a) Diversity  
673 across major and candidate genes in European local pig breeds. *PLoS One* **13**, e0207475.

674 Muñoz M., Bozzi R., García-Casco J., Núñez Y., Ribani A., Franci O., García F., Škrlep M., Schiavo  
675 G., Bovo S., Utzeri V.J., Charneca R., Martins J.M., Quintanilla R., Tibau J., Margeta V.,  
676 Djurkin-Kušec I., Mercat M.J., Riquet J., Estellé J., Zimmer C., Razmaite V., Araujo J.P.,  
677 Radović Č., Savić R., Karolyi D., Gallo M., Čandek-Potokar M., Fernández A.I., Fontanesi L.  
678 & Óvilo C. (2019) Genomic diversity, linkage disequilibrium and selection signatures in  
679 European local pig breeds assessed with a high density SNP chip. *Scientific Reports* **9**, 13546.

680 Muñoz M., García-Casco J.M., Caraballo C., Fernández-Barroso M.Á., Sánchez-Esquiliche F.,  
681 Gómez F., Rodríguez M.D.C. & Silió L. (2018b) Identification of candidate genes and  
682 regulatory factors underlying intramuscular fat content through *longissimus dorsi* transcriptome  
683 analyses in heavy Iberian pigs. *Frontiers in Genetics* **9**, 608.

684 Ogorevc J., Zorc M., Škrlep M., Bozzi R., Petig M., Fontanesi L., Čandek-Potokar M. & Dovč P.  
685 (2017) Is KIT locus polymorphism rs328592739 related to white belt phenotype in Krškopolje  
686 pig? *Agriculturae Conspectus Scientificus* **82**, 155–61.

687 Ovilo C., Clop A., Noguera J.L., Oliver M.A., Barragán C., Rodriguez C., Silió L., Toro M.A., Coll  
688 A., Folch J.M., Sánchez A., Babot D., Varona L. & Pérez-Enciso M. (2002) Quantitative trait  
689 locus mapping for meat quality traits in an Iberian x Landrace F2 pig population. *Journal of*  
690 *Animal Science* **80**, 2801–8.

691 Paudel Y., Madsen O., Megens H.J., Frantz L.A.F., Bosse M., Bastiaansen J.W.M., Crooijmans  
692 R.P.M.A. & Groenen M.A.M. (2013) Evolutionary dynamics of copy number variation in pig  
693 genomes in the context of adaptation and domestication. *BMC Genomics* **14**, 449.

694 Paudel Y., Madsen O., Megens H.J., Frantz L.A.F., Bosse M., Crooijmans R.P.M.A. & Groenen  
695 M.A.M. (2015) Copy number variation in the speciation of pigs: a possible prominent role for  
696 olfactory receptors. *BMC Genomics* **16**, 330.

697 Pirooznia M., Goes F.S. & Zandi P.P. (2015) Whole-genome CNV analysis: advances in  
698 computational approaches. *Frontiers in Genetics* **6**, 138.

699 Pollard M.O., Gurdasani D., Mentzer A.J., Porter T. & Sandhu M.S. (2018) Long reads: their purpose  
700 and place. *Human Molecular Genetics* **27**, R234–41.

701 Quinlan A.R. & Hall I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic  
702 features. *Bioinformatics* **26**, 841–2.

703 R Core Team. (2018) R: A language and environment for statistical computing. R Foundation for  
704 Statistical Computing, Vienna, Austria.

705 Ramayo-Caldas Y., Castelló A., Pena R.N., Alves E., Mercadé A., Souza C.A., Fernández A.I., Perez-  
706 Enciso M. & Folch J.M. (2010) Copy number variation in the porcine genome inferred from a  
707 60 k SNP BeadChip. *BMC Genomics* **11**, 593.

708 Redon R., Ishikawa S., Fitch K.R., Feuk L., Perry G.H., Andrews T.D., Fiegler H., Shaperro M.H.,  
709 Carson A.R., Chen W., Cho E.K., Dallaire S., Freeman J.L., Gonzalez J.R., Gratacos M., Huang  
710 J., Kalaitzopoulos D., Komura D., MacDonald J.R., Marshall C.R., Mei R., Montgomery L.,  
711 Nishimura K., Okamura K., Shen F., Somerville M.J., Tchinda J., Valsesia A., Woodwark C.,  
712 Yang F., Zhang J., Zerjal T., Zhang J., Armengol L., Conrad D.F., Estivill X., Tyler-Smith C.,  
713 Carter N.P., Aburatani H., Lee C., Jones K.W., Scherer S.W. & Hurles M.E. (2006) Global  
714 variation in copy number in the human genome. *Nature* **444**, 444–54.

715 Ren J., Duan Y., Qiao R., Yao F., Zhang Z., Yang B., Guo Y., Xiao S., Wei R., Ouyang Z., Ding N.,  
716 Ai H. & Huang L. (2011) A missense mutation in PPARD causes a major QTL effect on ear  
717 size in pigs. *PLoS Genetics* **7**, e1002043.

718 Revay T., Quach A.T., Maignel L., Sullivan B. & King W.A. (2015) Copy number variations in high  
719 and low fertility breeding boars. *BMC Genomics* **16**, 280.

720 Revilla M., Puig-Oliveras A., Castelló A., Crespo-Piazuelo D., Paludo E., Fernández A.I., Ballester  
721 M. & Folch J.M. (2017) A global analysis of CNVs in swine using whole genome sequence  
722 data and association analysis with fatty acid composition and growth traits. *PLoS One* **12**,  
723 e0177014.

724 Revilla M., Puig-Oliveras A., Crespo-Piazuelo D., Criado-Mesas L., Castelló A., Fernández A.I.,  
725 Ballester M. & Folch J.M. (2018) Expression analysis of candidate genes for fatty acid  
726 composition in adipose tissue and identification of regulatory regions. *Scientific Reports* **8**,  
727 2045.

728 Ribani A., Utzeri V.J., Geraci C., Tinarelli S., Djan M., Veličković N., Doneva R., Dall’Olio S., Costa  
729 L.N., Schiavo G., Bovo S., Usai G., Gallo M., Radović Č., Savić R., Karolyi D., Salajpal K.,  
730 Gvozdanić K., Djurkin- Kušec I., Škrlep M., Čandek- Potokar M., Oviló C. & Fontanesi L.

731 (2019) Signatures of de-domestication in autochthonous pig breeds and of domestication in  
732 wild boar populations from *MC1R* and *NR6A1* allele distribution. *Animal Genetics* **50**, 166–71.

733 Rubin C.J., Megens H.J., Martinez Barrio A., Maqbool K., Sayyab S., Schwochow D., Wang C.,  
734 Carlborg Ö., Jern P., Jørgensen C.B., Archibald A.L., Fredholm M., Groenen M.A.M. &  
735 Andersson L. (2012) Strong signatures of selection in the domestic pig genome. *Proceedings*  
736 *of the National Academy of Sciences of the United States of America* **109**, 19529–36.

737 Sambrook J., Fritsch E.F. & Maniatis T. (1989) *Molecular cloning: a laboratory manual*. 2<sup>nd</sup> Ed. Cold  
738 Spring Harbor Laboratory Press, Cold Spring Harbor, USA.

739 Schiavo G., Bertolini F., Galimberti G., Bovo S., Dall’Olio S., Nanni Costa L., Gallo M. & Fontanesi  
740 L. (2020a) A machine learning approach for the identification of population-informative  
741 markers from high-throughput genotyping data: application to several pig breeds. *Animal* **14**,  
742 223-32.

743 Schiavo G., Bovo S., Bertolini F., Tinarelli S., Dall’Olio S., Nanni Costa L., Gallo M. & Fontanesi L.  
744 (2020b) Comparative evaluation of genomic inbreeding parameters in seven commercial and  
745 autochthonous pig breeds. *Animal*, 1–11, doi:10.1017/S175173111900332X.

746 Schiavo G., Bertolini F., Utzeri V.J., Ribani A., Geraci C., Santoro L., Óvilo C., Fernández A.I., Gallo  
747 M. & Fontanesi L. (2018) Taking advantage from phenotype variability in a local animal  
748 genetic resource: identification of genomic regions associated with the hairless phenotype in  
749 Casertana pigs. *Animal Genetics* **49**, 321–5.

750 Schiavo G., Bovo S., Tinarelli S., Bertolini F., Dall’Olio S., Gallo M. & Fontanesi L. (2019) Genome-  
751 wide association analyses for several exterior traits in the autochthonous Casertana pig breed.  
752 *Livestock Science* **230**, 103842.

753 Schiavo G., Dolezal M.A., Scotti E., Bertolini F., Calò D.G., Galimberti G., Russo V. & Fontanesi  
754 L. (2014) Copy number variants in Italian Large White pigs detected using high-density single  
755 nucleotide polymorphisms and their association with back fat thickness. *Animal Genetics* **45**,  
756 745–9.

757 Silió L., Barragán C., Fernández A.I., García-Casco J. & Rodríguez M.C. (2016) Assessing effective  
758 population size, coancestry and inbreeding effects on litter size using the pedigree and SNP data  
759 in closed lines of the Iberian pig breed. *Journal of Animal Breeding and Genetics* **133**, 145–54.

760 Stafuzza N.B., Silva R.M. de O., Fragomeni B. de O., Masuda Y., Huang Y., Gray K. & Lourenco  
761 D.A.L. (2019) A genome-wide single nucleotide polymorphism and copy number variation  
762 analysis for number of piglets born alive. *BMC Genomics* **20**, 321.

763 Tomás A., Ramírez O., Casellas J., Muñoz G., Sánchez A., Barragán C., Arqué M., Riart I., Óvilo C.,  
764 Noguera J.L., Amills M. & Rodríguez C. (2011) Quantitative trait loci for fatness at growing  
765 and reproductive stages in Iberian × Meishan F(2) sows. *Animal Genetics* **42**, 548–51.

766 Wang J., Jiang J., Fu W., Jiang L., Ding X., Liu J.F. & Zhang Q. (2012) A genome-wide detection of  
767 copy number variations using SNP genotyping arrays in swine. *BMC Genomics* **13**, 273.

768 Wang J., Jiang J., Wang H., Kang H., Zhang Q. & Liu J.-F. (2014) Enhancing genome-wide copy  
769 number variation identification by high density array CGH using diverse resources of pig  
770 breeds. *PLoS One* **9**, e87571.

771 Wang J., Jiang J., Wang H., Kang H., Zhang Q. & Liu J.-F. (2015c) Improved detection and  
772 characterization of copy number variations among diverse pig breeds by array CGH. *G3* **5**,  
773 1253–61.

774 Wang Z., Sun H., Chen Q., Zhang X., Wang Q. & Pan Y. (2019a) A genome scan for selection  
775 signatures in Taihu pig breeds using next-generation sequencing. *Animal* **13**, 683–93.

776 Wang H., Wang C., Yang K., Liu J., Zhang Y., Wang Y., Xu X., Michal J.J., Jiang Z. & Liu B.  
777 (2015a) Genome wide distributions and functional characterization of copy number variations  
778 between Chinese and Western pigs. *PLoS One* **10**, e0131522.

779 Wang L., Xu L., Liu X., Zhang T., Li N., Hay E.H., Zhang Y., Yan H., Zhao K., Liu G.E., Zhang L.  
780 & Wang L. (2015b) Copy number variation-based genome wide association study reveals  
781 additional variants contributing to meat quality in Swine. *Scientific Reports* **5**, 12535.



782 Wang Y., Zhang T. & Wang C. (2019b) Detection and analysis of genome-wide copy number  
783 variation in the pig genome using an 80 K SNP Beadchip. *Journal of Animal Breeding and*  
784 *Genetics* doi: 10.1111/jbg.12435.

785 Wei W.H., Koning D.J. de, Penman J.C., Finlayson H.A., Archibald A.L. & Haley C.S. (2007) QTL  
786 modulating ear size and erectness in pigs. *Animal Genetics* **38**, 222–6.

787 Wiedmann R.T., Nonneman D.J. & Rohrer G.A. (2015) Genome-Wide Copy Number Variations  
788 Using SNP Genotyping in a Mixed Breed Swine Population. *PLoS One* **10**, e0133529.

789 Wilkinson S., Wilkinson Lu Z.H., Megens H.J., Archibald A.L., Haley C., Jackson I.J., Groenen  
790 M.A.M., Crooijmans R.P.M.A., Ogden R. & Wiener P. (2013) Signatures of diversifying  
791 selection in European pig breeds. *PLoS Genetics* **9**, e1003453.

792 Winchester L., Yau C. & Ragoussis J. (2009) Comparing CNV detection methods for SNP arrays.  
793 *Briefings in Functional Genomics & Proteomics* **8**, 353–66.

794 Wu Z., Deng Z., Huang M., Hou Y., Zhang H., Chen H. & Ren J. (2019) Whole-genome resequencing  
795 identifies KIT new alleles that affect coat color phenotypes in pigs. *Frontiers in Genetics* **10**,  
796 218.

797 Xie J., Li R., Li S., Ran X., Wang J., Jiang J. & Zhao P. (2016) Identification of copy number  
798 variations in Xiang and Kele pigs. *PLoS One* **11**, e0148565.

799 Yang B., Cui L., Perez-Enciso M., Traspov A., Crooijmans R.P.M.A., Zinovieva N., Schook L.B.,  
800 Archibald A., Gatphayak K., Knorr C., Triantafyllidis A., Alexandri P., Semiadi G., Hanotte  
801 O., Dias D., Dovč P., Uimari P., Iacolina L., Scandura M., Groenen M.A.M., Huang L. &  
802 Megens H.-J. (2017) Genome-wide SNP data unveils the globalization of domesticated pigs.  
803 *Genetics Selection Evolution* **49**, 71.

804 Yang L., Xu L., Zhou Y., Liu M., Wang L., Kijas J.W., Zhang H., Li L. & Liu G.E. (2018) Diversity  
805 of copy number variation in a worldwide population of sheep. *Genomics* **110**, 143–8.

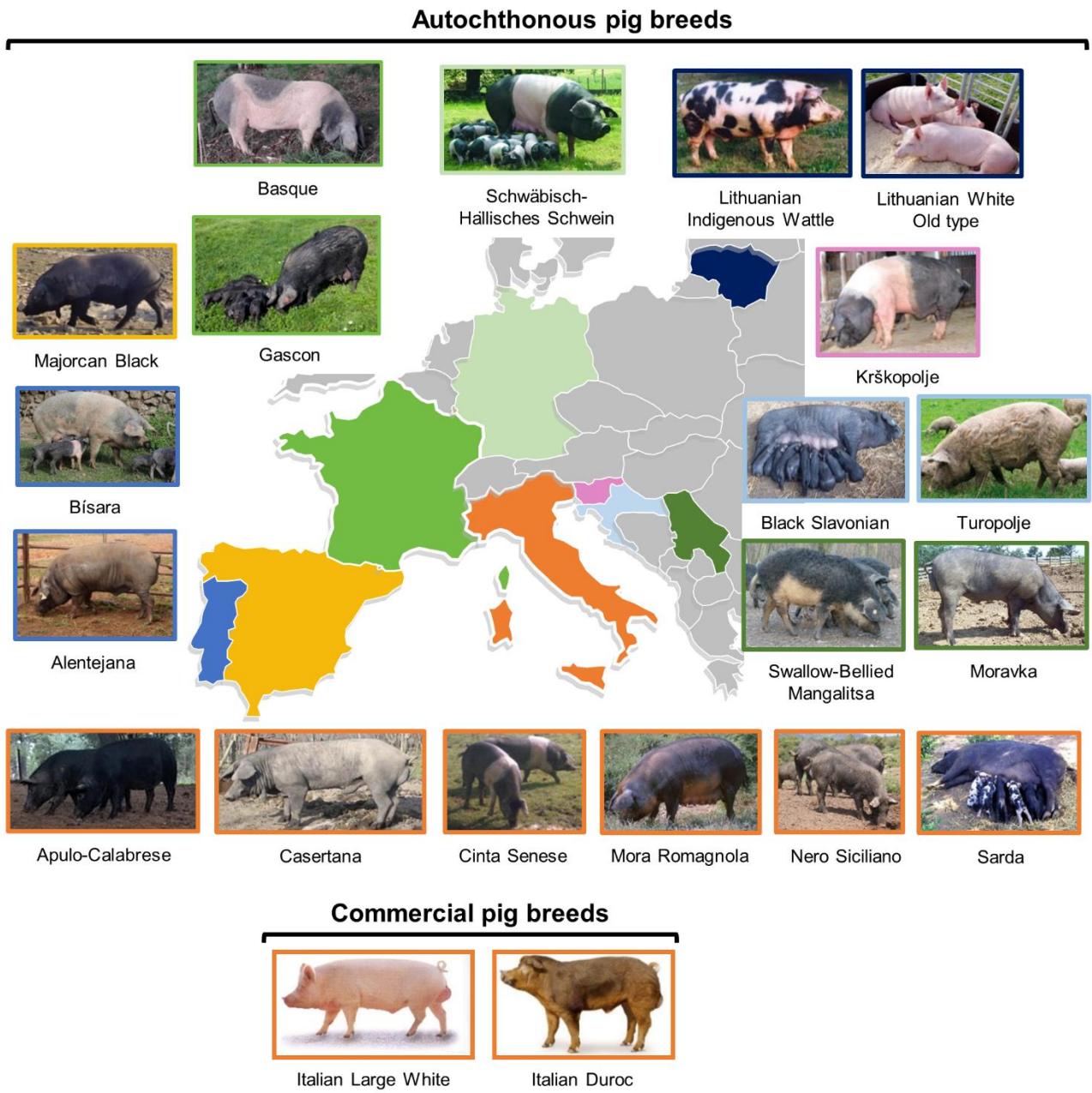
806 Zhang F., Gu W., Hurles M.E. & Lupski J.R. (2009) Copy number variation in human health, disease,  
807 and evolution. *Annual Review of Genomics and Human Genetics* **10**, 451-81.

808 Zhang Y., Liang J., Zhang L., Wang L., Liu X., Yan H., Zhao K., Shi H., Zhang T., Li N., Pu L. &  
809 Wang L (2015) Porcine *methionine sulfoxide reductase B3*: molecular cloning, tissue-specific  
810 expression profiles, and polymorphisms associated with ear size in *Sus scrofa*. *Journal of*  
811 *Animal Science and Biotechnology* **6**, 60.

812

813 **Figures**

814 **Fig. 1.** Phenotypes and geographical origin of the 21 analysed pig breeds.



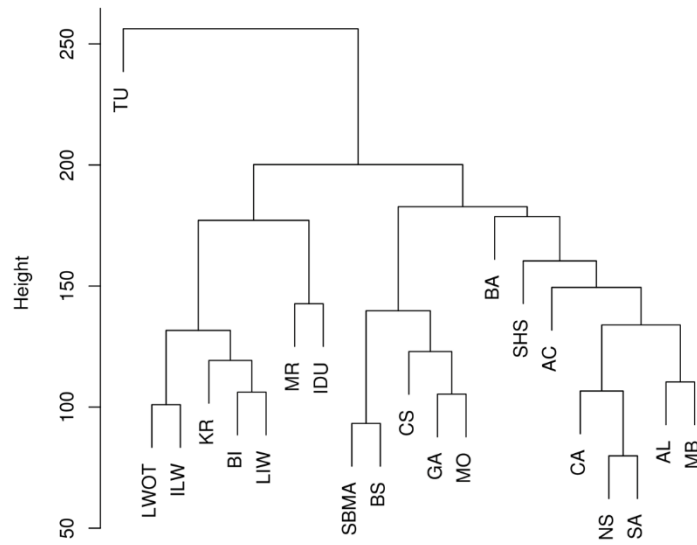
815

816 **Fig. 2.** Distribution of CNVRs along each autosomal chromosome.



817

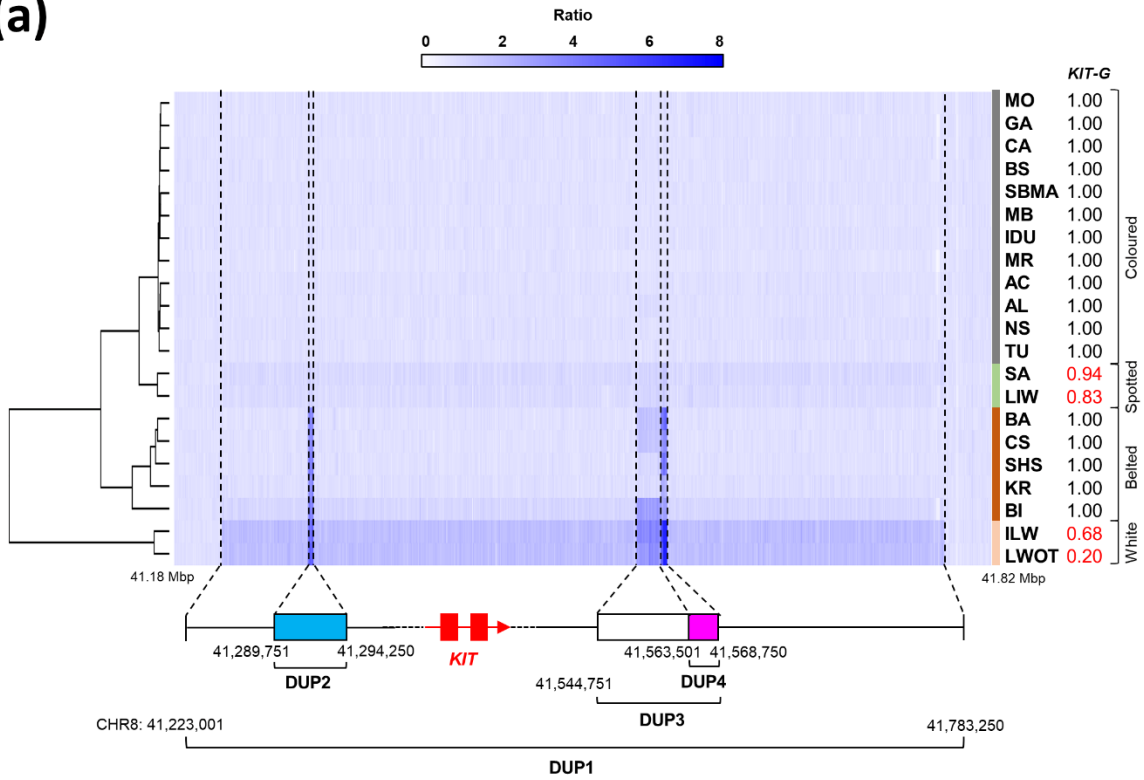
818 **Fig. 3.** Dendrogram representing the hierarchical clustering of the copy number state. Acronyms of  
819 the breed name are explained in Table 1.



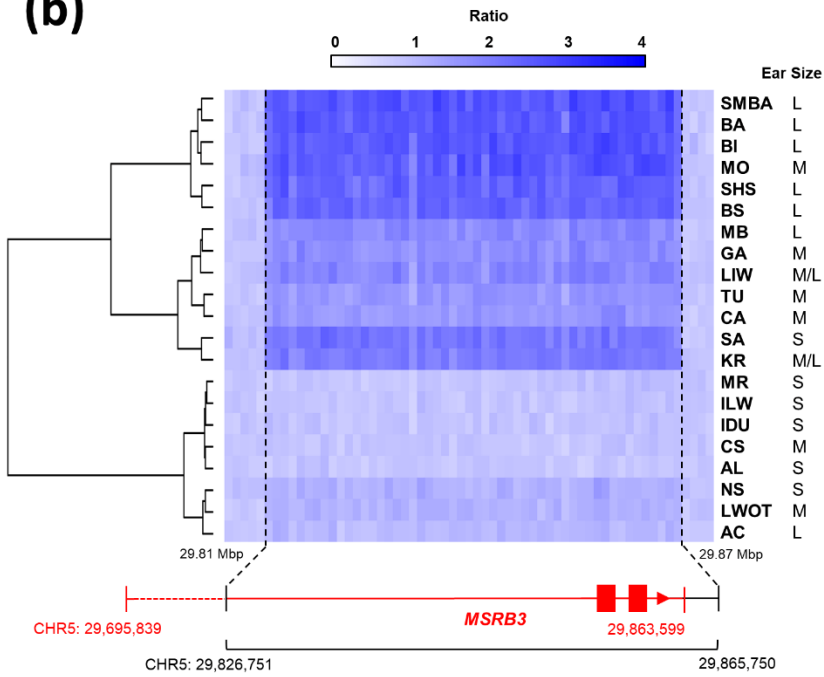
820

821 **Fig. 4. (a)** Heatmap of the read count ratios over the *KIT* gene. Coat colour reported in the  
822 correspondence of the breeds indicates breed main characteristics. SA (Sarda) has heterogeneous and  
823 not-fixed patterns. It was included among the spotted based on the frequency of this phenotype in the  
824 breed and according to the copy number (CN) state at this locus. Basque (BA) has spotted/belted  
825 heterogeneous patterns but was included among the belted breeds according to the CN state at this  
826 locus – (see text and Table S1 for details). KIT-G: frequency of the allele G of the single nucleotide  
827 polymorphism (SNP) rs345599765 (splice mutation of the intron 17; Marklund *et al.* 1998). **(b)** Heatmap  
828 of the read count ratios over the *MSRB3* gene. Ear size indicated in (b): L = large; M = medium; S =  
829 small (see text and Table S1 for details). The light-dark blue bar at the top of (a) and (b) indicates the  
830 CN ratio (1 = normal state without any gain or loss). For each breed, the read count ratio was  
831 computed in 750-bp consecutive genome windows. Acronyms of the breed name are explained in  
832 Table 1. **(c)** Average CN state of the *MSRB3* gene in relation to ear size. **(d)** Relationship between  
833 the average CN state of the *MSRB3* gene and the SNPs rs340841870 (green) and rs326411202 (blue).  
834 Pearson's correlation coefficient ( $r$ ) are reported.

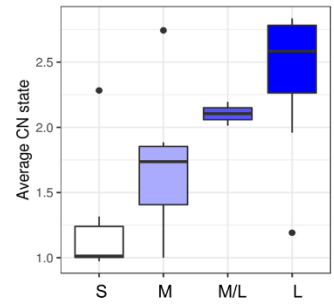
(a)



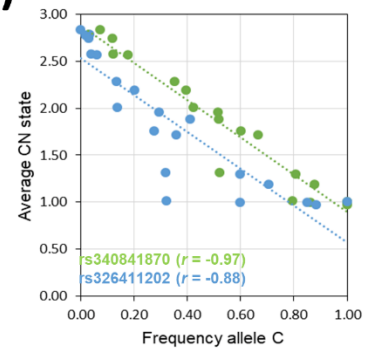
(b)



(c)



(d)



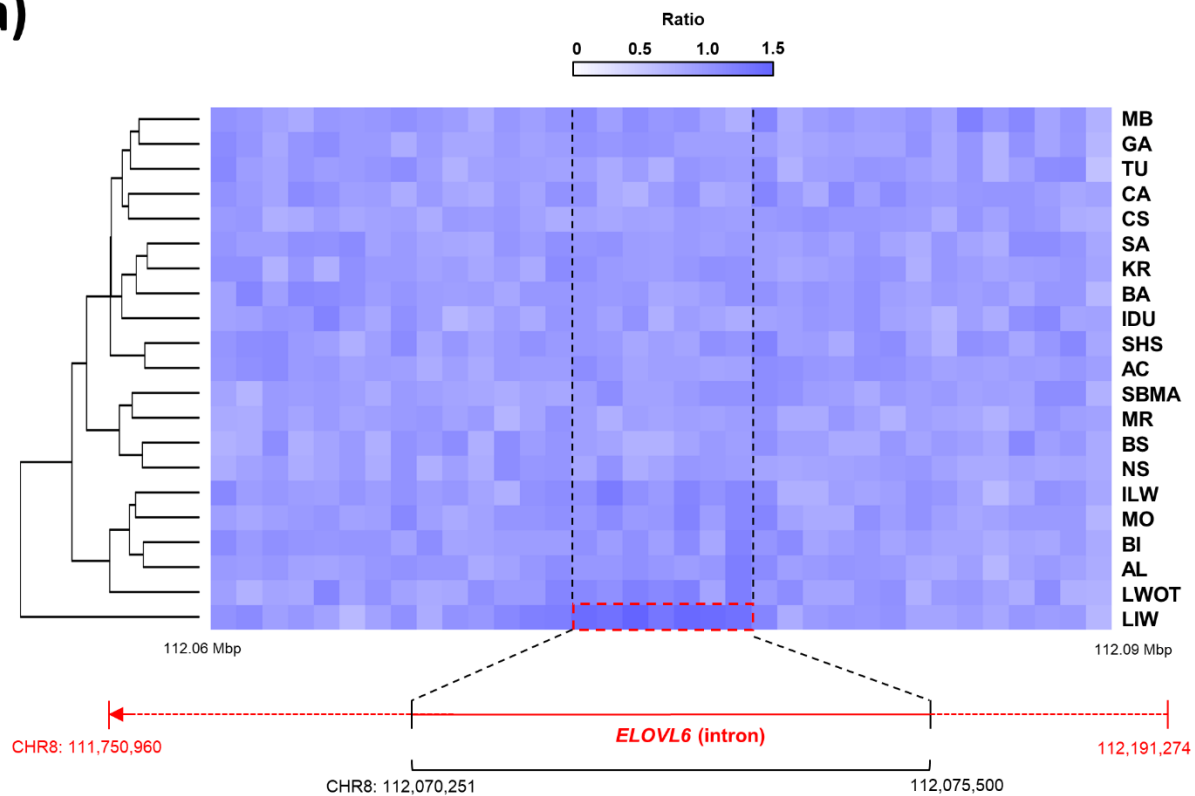
835

836

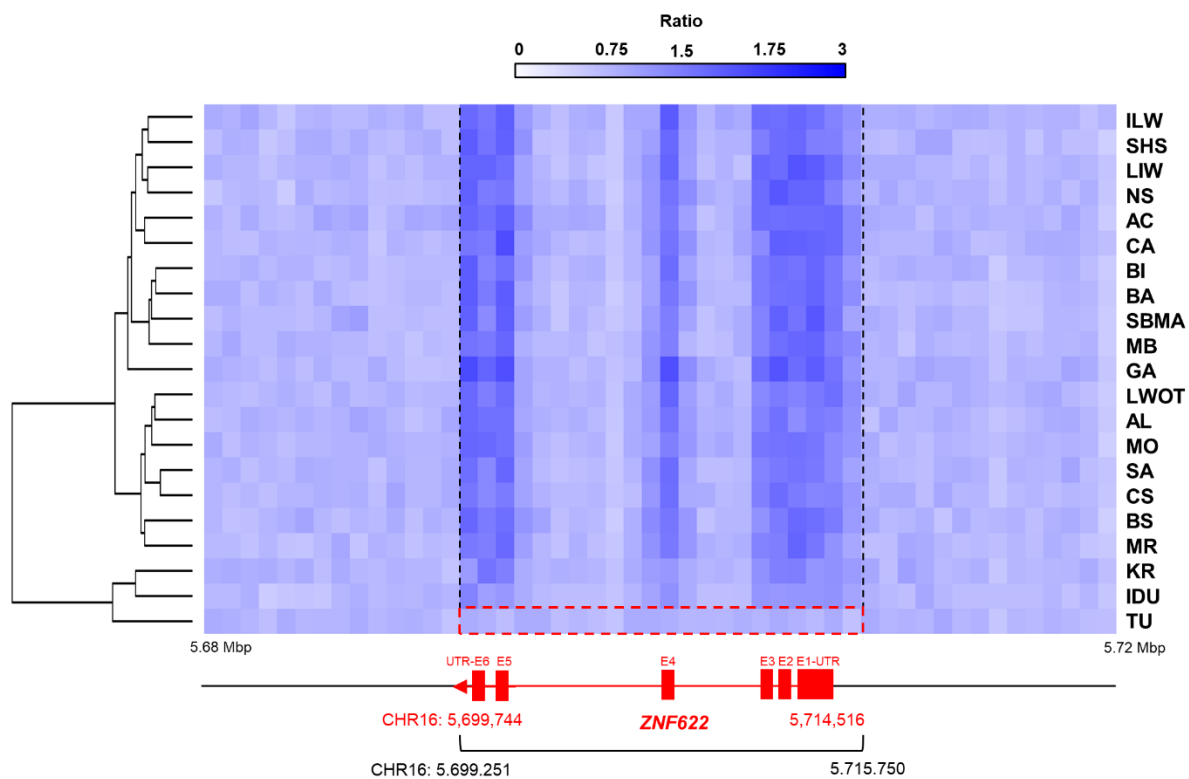
837 **Fig. 5.** Heatmap of the read count ratios over the *ELOVL6* (a) and *ZNF622* (b) genes. Exons below  
838 the heatmap for the *ZNF622* gene are numbered (E1-E6) according to the annotation in the  
839 Sscrofa11.1 genome version. Untranslated regions (UTR) are also reported. The light-dark blue bar  
840 at the top of (A) and (B) indicates the copy number (CN) ratio (1 = normal state without any gain or  
841 loss). For each breed, the read count ratio was computed in 750-bp consecutive genome windows.  
842 Acronyms of the breed name are explained in Table 1.



(a)



(b)



843

844

845 **Tables**846 **Table 1.** Summary of CNVs of the 21 analysed pig breeds. Data are stratified by breed.

<b>Breed</b>	<b>Short name</b>	<b>CNV<sup>1</sup></b>	<b>CNL<sup>2</sup></b>	<b>CNG<sup>3</sup></b>	<b>Length<sub>Min</sub><sup>4</sup></b>	<b>Length<sub>Max</sub><sup>5</sup></b>	<b>Length<sub>Median</sub><sup>6</sup></b>	<b>% length in CNV<sup>7</sup></b>
<b>Autochthonous</b>								
Alentejana	AL	601	345	256	2250	69750	3000	0.17
Apulo-Calabrese	AC	676	313	363	2250	142500	3000	0.18
Basque	BA	1122	626	496	2250	99750	3750	0.29
Bísara	BI	437	162	275	2250	63000	3000	0.11
Black Slavonian	BS	504	225	279	2250	142500	3000	0.13
Casertana	CA	596	272	324	2250	113250	3000	0.16
Cinta Senese	CS	662	352	310	2250	89250	3750	0.19
Gascon	GA	781	379	402	2250	126000	3000	0.20
Krškopolje	KR	510	152	358	2250	101250	3000	0.13
Lithuanian Indigenous Wattle	LIW	710	295	415	2250	90750	3750	0.19
Lithuanian White Old Type	LWOT	711	308	403	2250	560250	3750	0.21
Majorcan Black	MB	546	328	218	2250	101250	3750	0.15
Mora Romagnola	MR	1255	647	608	2250	137250	3000	0.34
Moravka	MO	391	159	232	2250	100500	3000	0.10
Nero Siciliano	NS	298	149	149	2250	42750	3000	0.07
Sarda	SA	209	72	137	2250	38250	3000	0.04
Schwäbisch-Hällisches Schwein	SHS	576	277	299	2250	147000	3000	0.15
Swallow-Bellied Mangalitsa	SBMA	757	433	324	2250	121500	3000	0.22
Turopolje	TU	1440	845	595	2250	99750	3750	0.40
<b>Commercial</b>								
Italian Duroc	IDU	1111	249	862	2250	116250	3000	0.28
Italian Large White	ILW	451	148	303	2250	560250	3000	0.14

847 <sup>1</sup> Total no. of copy number variants; <sup>2</sup> Total no. of copy number losses; <sup>3</sup> Total no. of copy number gains; <sup>4</sup> Minimum length (bp) of CNVs; <sup>5</sup> Maximum  
848 length (bp) of CNVs; <sup>6</sup> Median length (bp) of CNVs; <sup>7</sup> Percentage of the *S. scrofa* genome occupied by CNVs.

849 **Table 2.** Summary of CNVRs of the 21 analysed pig breeds stratified by chromosome.

<b>Chromosome</b>	<b>CNVR<sup>1</sup></b>	<b>Length<sub>Min</sub><sup>2</sup></b>	<b>Length<sub>Max</sub><sup>3</sup></b>	<b>Length<sub>Median</sub><sup>4</sup></b>	<b>% length in CNVR<sup>5</sup></b>
SSC1	359	2250	137250	3760	0.88
SSC2	361	2250	43500	3760	1.54
SSC3	162	2250	147750	3010	0.82
SSC4	227	2250	81000	3760	0.99
SSC5	215	2250	46500	3760	1.45
SSC6	302	2250	120750	3760	1.07
SSC7	167	2250	96750	3760	1.16
SSC8	244	2250	560250	3760	1.37
SSC9	259	2250	159000	3760	1.86
SSC10	114	2250	85500	3010	0.96
SSC11	159	2250	153750	3760	1.54
SSC12	110	2250	108000	3385	1.33
SSC13	307	2250	91500	3760	1.05
SSC14	212	2250	195750	3760	1.34
SSC15	196	2250	63000	3760	0.86
SSC16	138	2250	41250	3010	0.88
SSC17	132	2250	80250	3010	1.27
SSC18	46	2250	16500	3010	0.35

850 <sup>1</sup> Total no. of copy number variant regions; <sup>2</sup> Minimum length (bp) of CNVs; <sup>3</sup> Maximum length (bp)  
851 of CNVs; <sup>4</sup> Median length (bp) of CNVs; <sup>5</sup> Percentage of the chromosome occupied by CNVRs.  
852

853 **Supporting information**

854 **Table S1.** Details on the analysed animals and investigated breeds, including geographical  
855 distribution and phenotypic description.

856 **Table S2.** Summary statistics of whole-genome sequencing.

857 **Table S3.** Summary statistics of detected CNVs stratified by chromosome.

858 **Table S4.** CNVRs detected over all analysed breeds.

859 **Table S5.** Summary statistics of detected CNVRs, stratified by pig breed.

860 **Table S6.** Over-represented repeated element classes.

861 **Table S7.** Within CNVRs over-represented QTLs.

862 **Table S8.** Within CNVRs over-represented biological functions.

863 **Table S9.** Allele frequency of the single nucleotide polymorphisms at the *KIT* and *MSRB3* genes  
864 estimated from sequencing data.

865 **Table S10.** Summary statistics of CNVRs previously identified in other studies.