BMC Genomics

Assessing breed integrity of Göttingen Minipigs

Christian Reimer[1,2]*, Ngoc-Thuy Ha[1,2], Ahmad Reza Sharifi[1,2], Johannes Geibel[1,2], Lars Friis Mikkelsen[3], Martin Schlather[2,4], Steffen Weigend[2,5] and Henner Simianer[1,2]

Abstract

Background: Göttingen Minipigs (GMP) is the smallest commercially available minipig breed under a controlled breeding scheme and is globally bred in five isolated colonies. The genetic isolation harbors the risk of stratification which might compromise the identity of the breed and its usability as an animal model for biomedical and human disease. We conducted whole genome re-sequencing of two DNA-pools per colony to assess genomic differentiation within and between colonies. We added publicly available samples from 13 various pig breeds and discovered overall about 32 M loci, ~ 16 M. thereof variable in GMPs. Individual samples were virtually pooled breed-wise. F_{ST} between virtual and DNA pools, a phylogenetic tree, principal component analysis (PCA) and evaluation of functional SNP classes were conducted. An F-test was performed to reveal significantly differentiated allele frequencies between colonies. Variation within a colony was quantified as expected heterozygosity.

Results: Phylogeny and PCA showed that the GMP is easily discriminable from all other breads, but that there is also differentiation between the GMP colonies. Dependent on the contrast between GMP colonies, 4 to 8% of all loci had significantly different allele frequencies. Functional annotation revealed that functionally non-neutral loci are less prone to differentiation. Annotation of highly differentiated loci revealed a couple of deleterious mutations in genes with putative effects in the GMPs .

Conclusion: Differentiation and annotation results suggest that the underlying mechanisms are rather drift events than directed selection and limited to neutral genome regions. Animal exchange seems not yet necessary. The Relliehausen colony appears to be the genetically most unique GMP sub-population and could be a valuable resource if animal exchange is required to maintain uniformity of the GMP.

Keywords: Göttingen Minipigs, Conservation, Population structure

Background

The Göttingen Minipig (GMP) is an animal model with growing importance [1]. Created in the 1960's by crossing Minnesota Minipigs, Vietnamese Potbellied Pigs and the German Landrace, the breed has been under a fully documented closed breeding scheme ever since. The first colony was founded at the research farm of the University of Göttingen in Friedland, Germany, and later resettled to the Relliehausen research farm. Due to the growing customer interest of using GMPs, this facility

could not satisfy the demand anymore and therefor collaboration with Ellegaard Göttingen Minipigs A/S in Dalmose, Denmark, including a larger colony, was established in 1992. In 2003, animals from this colony were brought to Marshall BioResources, North Rose, New York, USA as the basis of a North American GMP breeding colony. In 2009, a second barrier colony was established in Dalmose, based on breeders from the first barrier colony to increase the production. Finally, animals from Dalmose were brought to Oriental Yeast Co., Ltd. in Nisshin, Japan, in 2013 to establish a barrier colony in Japan. After the initial animal transfer, all colonies remained under closed breeding without any genetic exchange, albeit being under a common controlled breeding scheme, coordinated by the animal breeding and genetics group at the University of

* Correspondence: creimer@gwdg.de
[1]Department of Animal Sciences, Animal Breeding and Genetics Group, University of Göttingen, Albrecht-Thaer-Weg 3, 37017 Göttingen, Germany
[2]Center for Integrated Breeding Research, University of Göttingen, Albrecht-Thaer-Weg 3, 37075 Göttingen, Germany
Full list of author information is available at the end of the article

Göttingen, Germany. Today all GMPs under that management are bred for two main traits, number of piglets born alive and an index comprising body weight measures at different ages. It is inherent in breeding practice that there is also a non-documented co-selection on tame behavior and against appearing malformations such as unwanted pigmentation [2]. Different intensities of conservational breeding techniques, such as adjustment of selective pressures and balancing selective pressures between the sexes, is used to account for the different herd sizes, eventually results in all colonies having comparable effective population sizes.

Managing the GMP in independent colonies in closed barriers is beneficial from a safety point of view. Additionally, a production unit close to the main sales market minimizes negative effects on animal welfare through long transports and prevents import complications and unnecessary bureaucracy. On the other hand, splitting a population reduces the effective population size of each sub-population, which increases the risk of genetic drift or the manifestation of recessive disorders [3]. Two concepts to counter these risks are purging of deleterious variants [4] or maintenance of genetic diversity [5]. Lacy [6] argues that drift is the most important factor in loss of genetic variance when effective population sizes are low, as in the case of the GMP [7], and the only effective measure to mitigate adverse effects would be animal exchange.

In this study we try to assess whether the genetic management was able to maintain the uniformity of the GMP breed, or if the isolated production units are already genetically diversified such that an exchange of breeders is inevitable. This was done by re-sequencing two representative DNA pools from each unit: candidates were sampled for low average relationship within a pool, but elevated relationship towards the remaining colony, allowing an assessment of the diversity within and between units.

Results
Sampling of the optimally representative candidates for pooling based on relationship measures resulted in candidate sets which exhibited lower inner-set mean relationship coefficients (a) compared to the mean relationship of the candidate set with the remaining colony mates (b). Both, the absolute level of relationship

and the difference between a and b were lowest for Relliehausen (RE) and highest for North Rose (NR), while Dalmose barrier 2 (DA2) and 3 and Nisshin (NI) were at the same level and exhibited similar difference between a and b (Table 1).

Variant calling discovered 67′056′755 raw biallelic SNPs. After the variant quality score recalibration and filtering, 32′615′461 (incl. 945′565 SNPs on chr. X) total non-monomorphic SNPs, thereof 4′121′427 (incl. 263′498 on chr. X) novel SNPs not documented in dbSNP were retained. Discarding loci, monomorphic in the minipigs and X-chromosomal loci for the analysis within minipigs left a data set containing 16′498′773 autosomal SNPs.
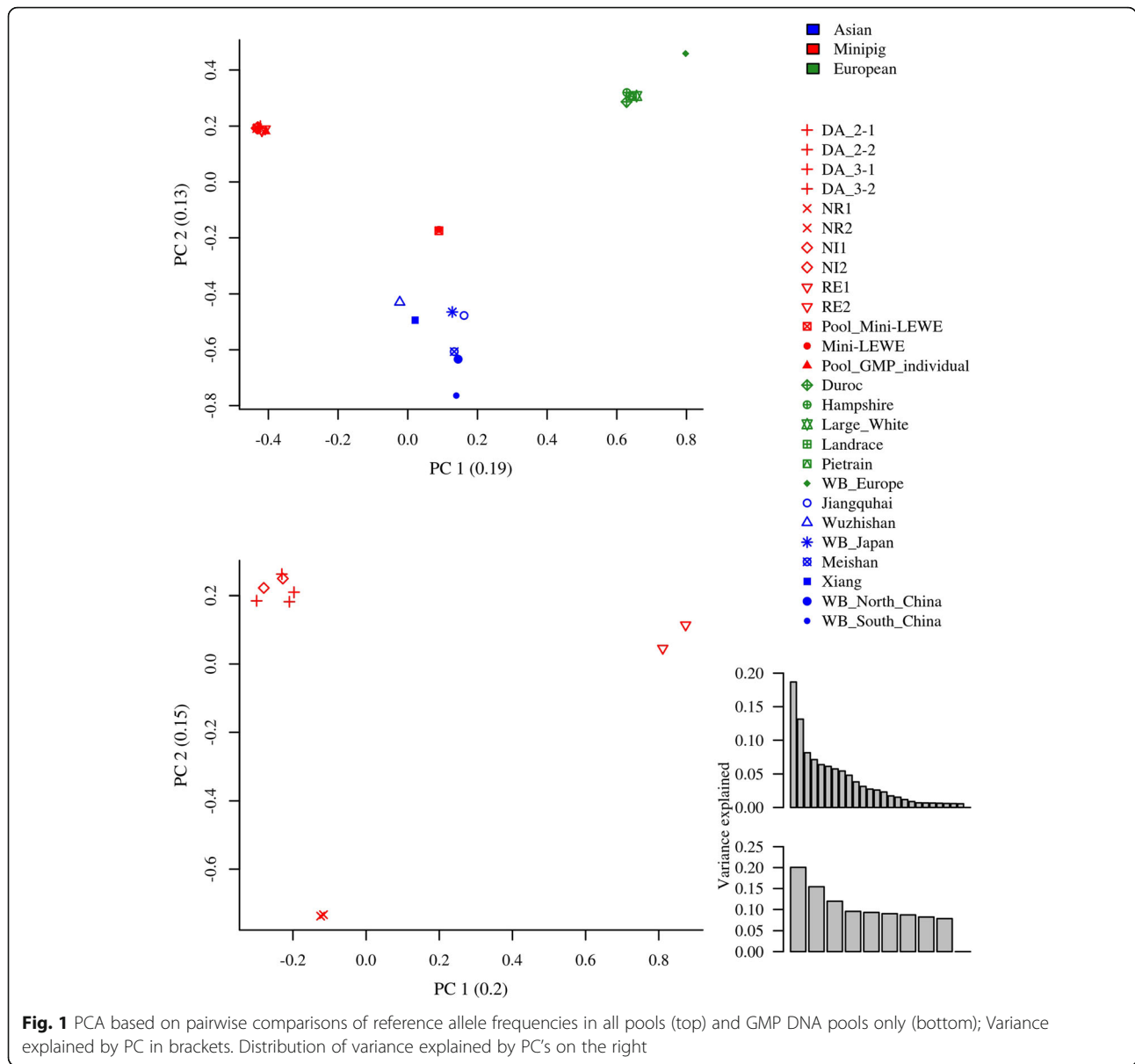
Principal component analysis
Principal component analysis (PCA) on the reference allele frequencies shows clear separation of the GMPs, Asian and European breeds, with the Mini-LEWE clustering close to the Asian breeds, by the first component (Fig. 1). The second component separates the Asian breeds from GMPs and Europeans. In the context of multiple breeds, no structure can be observed within the GMPs. Separate analysis shows three sub-groups within the GMPs consisting of the RE colony, NR colony and a composite group of DA2, DA3 and NI.

Differentiation and distance measures
As measures of differentiation and genetic distance between the different breeds, F_{ST} and the Reynolds genetic distance (D_R) were estimated. When applied to the variable set of large breeds and minipig pools, both measures provided a similar picture of three strongly differentiated groups (Table 2). In principle, these three groups were the minipigs, the European breeds and the Asian breeds, respectively, with the exception, that the Mini-LEWE pool clustered with the Asian group. Comparing F_{ST} against D_R, the latter showed generally higher estimates, relatively inflated at moderate levels of differentiation/ distance (Fig. 2), but provided in general a very similar picture. Therefore, only F_{ST} was used for later purposes, such as the functional annotation. Focusing on the differentiation within the three groups, the GMP exhibited the lowest average differentiation (F_{ST}: 0.07; D_R: 0.11; see also Additional File 1: Supplementary Table 1), the European (F_{ST}: 0.16; D_R: 0.25) the second

Table 1 Mean coefficients of relationship within a colony sample and between the sample and the remaining stock, including number of successfully extracted probes

	RE	DA2	DA3	NR	NI
Relationship within sample	0.357	0.396	0.396	0.403	0.387
Relationship Sample/ Remaining Stock	0.376	0.404	0.402	0.410	0.397
Successful DNA extractions	28	24	23	28	24

Fig. 1 PCA based on pairwise comparisons of reference allele frequencies in all pools (top) and GMP DNA pools only (bottom); Variance explained by PC in brackets. Distribution of variance explained by PC's on the right

lowest and the Asian the highest (F_{ST}: 0.27; D_R: 0.37) (Table 2). The average differentiation to other groups was higher than the differentiation within the groups, clearly so for minipigs and for European breeds, but not
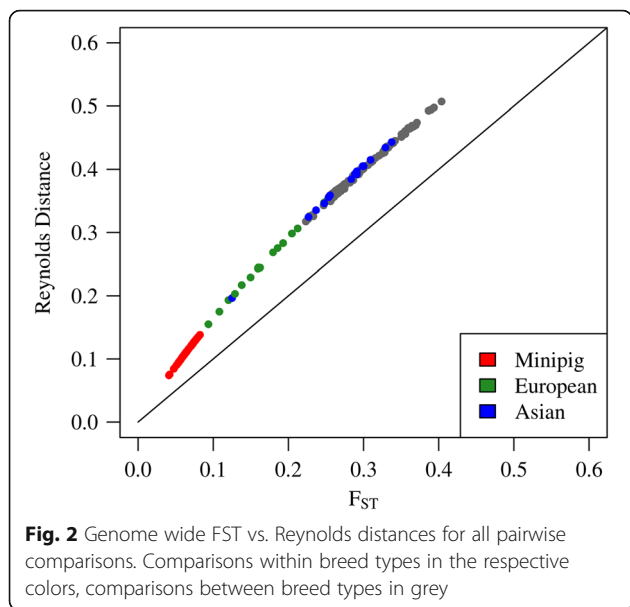
Table 2 Mean F_{ST} and D_R between European and Asian breeds and GMP

		GMP	Asian	European
	GMP	0.066	0.262	0.307
FST	Asian		0.271	0.327
	European			0.163
	GMP	0.113	0.364	0.408
D_R	Asian		0.372	0.428
	European			0.246

that clearly for the Asian breeds. The latter exposed an even lower average D_R (0.36 vs 0.37) and F_{ST} (0.26 vs. 0.27) to the GMP than within the group of Asian breeds.

Phylogeny

The UPGMA tree (Fig. 3) produced from F_{ST} values calculated with genome wide SNP data shows a clear clustering of the GMPs from the other breeds. The next level clusters contain (in that order) Xiang, Meishan, South Chinese wild boars, the Mini-LEWE and the North Chinese wild boars. The European breeds form their own cluster. Throughout 100 resamplings, the GMP cluster, the Mini-LEWE cluster and European cluster are rediscovered in every iteration, while the nodes connected to the Asian breeds seem unstable with

Reimer *et al. BMC Genomics*   (2020) 21:308

Page 4 of 13

Fig. 2 Genome wide FST vs. Reynolds distances for all pairwise comparisons. Comparisons within breed types in the respective colors, comparisons between breed types in grey

resampling probability between 18 and 87%, with the exception of the Japanese wild boar, that behaves like an outgroup sample. Even though, the European and the GMP clusters are distinct, the order within the clusters is variable. The node support within the European cluster spans from 56 to 72%, and between 18 to 86% in the GMP cluster. The most stable structure with 86% contains the RE pools and the least stable (18%) contains the DA and NI herds.

Stratification within the GMP

The genetic differences within the GMP were determined by comparing pools in terms of allele frequency differences, such as oppositely fixed alleles, extreme F_{ST} values between colonies, differences in the average expected heterozygosity within pools by a variation based approach employing an F-test statistic. Resulting loci detected by the aforementioned statistics were functionally annotated and imbalances between the various classes were checked for potential biases towards differentiated loci.

Significance test of pool allele frequencies between and within colonies

The F-test compared the variation between the two pools with the variation between one of the pools against one foreign pool and could, in contrast to F_{ST}, add probabilistic evidence on differentiation between pools. On average, the NI colony had the lowest proportion of significantly ($p = 0.05$, Bonferroni corrected) differentiated loci, overall 4.7%, followed by DA3 and 2 with 5.3 and 5.4%, respectively, and RE with 5.7%. With 6.9%, NR had the highest proportion of significantly differentiated loci (Table 3; Supplementary Table 2). Focusing on the colonies separately (Fig. 4), only RE had comparable amounts of differentiated loci with all others. From the perspective of DA, NI, and NR, the level of differentiation to RE was clearly highest throughout all
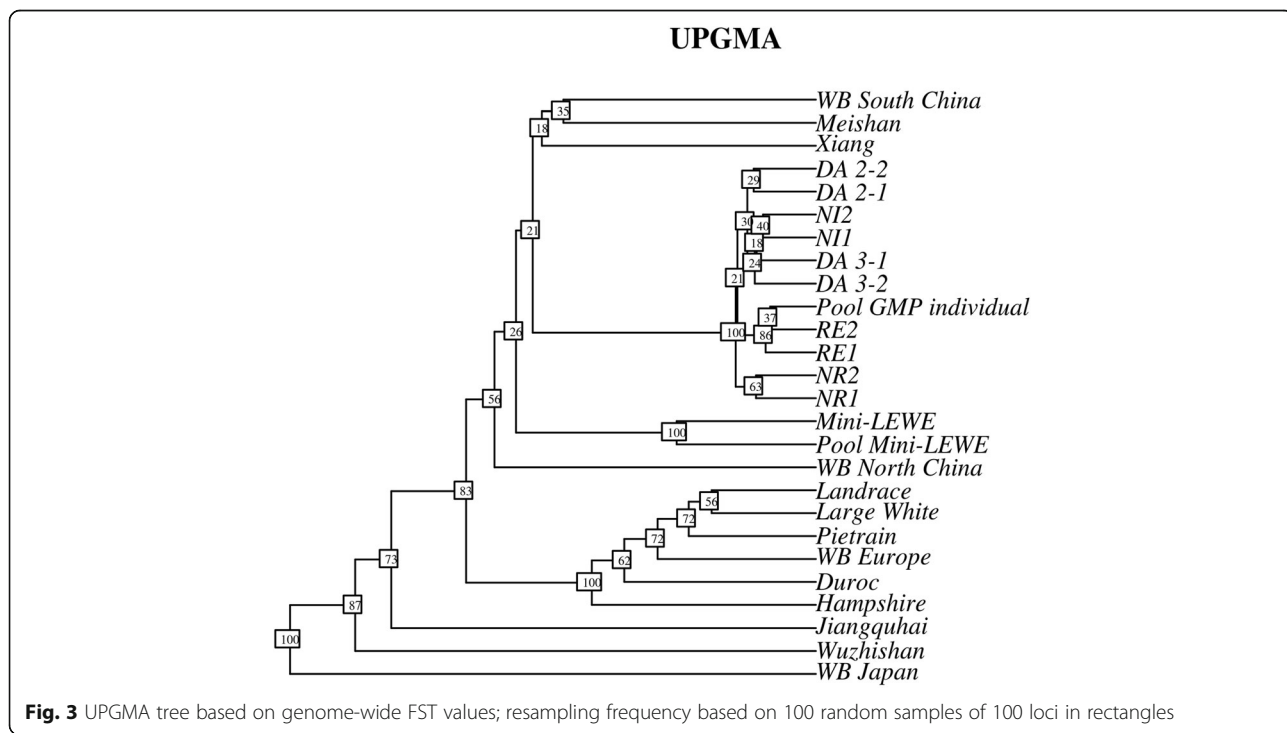

Fig. 3 UPGMA tree based on genome-wide FST values; resampling frequency based on 100 random samples of 100 loci in rectangles

Table 3 Proportion of SNP significantly different between colony and remote pool in F-test at 5%

	RE_1	RE_2	DA2_1	DA2_2	DA3_1	DA3_2	NR_1	NR_2	NI_1	NI_2
RE	0.00	0.00	0.06	0.06	0.06	0.06	0.06	0.05	0.06	0.06
DA2	0.07	0.07	0.00	0.00	0.05	0.05	0.05	0.05	0.05	0.05
DA3	0.07	0.07	0.05	0.05	0.00	0.00	0.05	0.05	0.04	0.04
NR	0.08	0.08	0.06	0.06	0.07	0.06	0.00	0.00	0.06	0.07
NI	0.07	0.07	0.04	0.04	0.04	0.04	0.04	0.04	0.00	0.00

comparisons, while the number of evaluated loci was relatively even throughout the comparisons. This means, e.g. RE pools could be significantly different from NR at a locus, since Variation within NR is low, and the average allele frequency in RE is far while using the more variable RE as a basis, the NR frequency is ranging within the variation of RE. Mostly both tested pools of a colony showed a similar amount of differentiation with the exception of NR versus the two NI pools. The highest proportion of differentiated loci was found, when NR was tested against the two RE pools.

Expected heterozygosity
Expected heterozygosity, as measure of variation within a pool, revealed that all pools exhibit similar levels of expected heterozygosity (Table 4) with RE and NI being highest and NR being lowest. When estimated for single pools, expected heterozygosity was between 0.271 and

0.283 for DA2 and DA3 and NR and between 0.280 and 0.285 in NI and RE. Estimated from the virtual union of both pools per colony, the values were about 0.01 higher, following the same trends (Table 5).

Fixed alleles and private polymorphisms
Table 6 depicts the correlation of allele frequencies of loci that had complete recordings and where each colony was fixed for either the reference or the alternative allele (Supplementary Table 3). Only 506 loci fulfilled this criterion. The correlations between the colonies based on these loci ranged between − 0.10 and + 0.14, with the highest being between RE and NI.

On the other hand, RE held by far the largest number of still variable loci while the other pools were fixed at one allele. Out of the 1′203′000 loci fulfilling the criterion of being variable in one colony while all others were fixed, 555′591 belonged to RE (Table 7). NR (192′896)
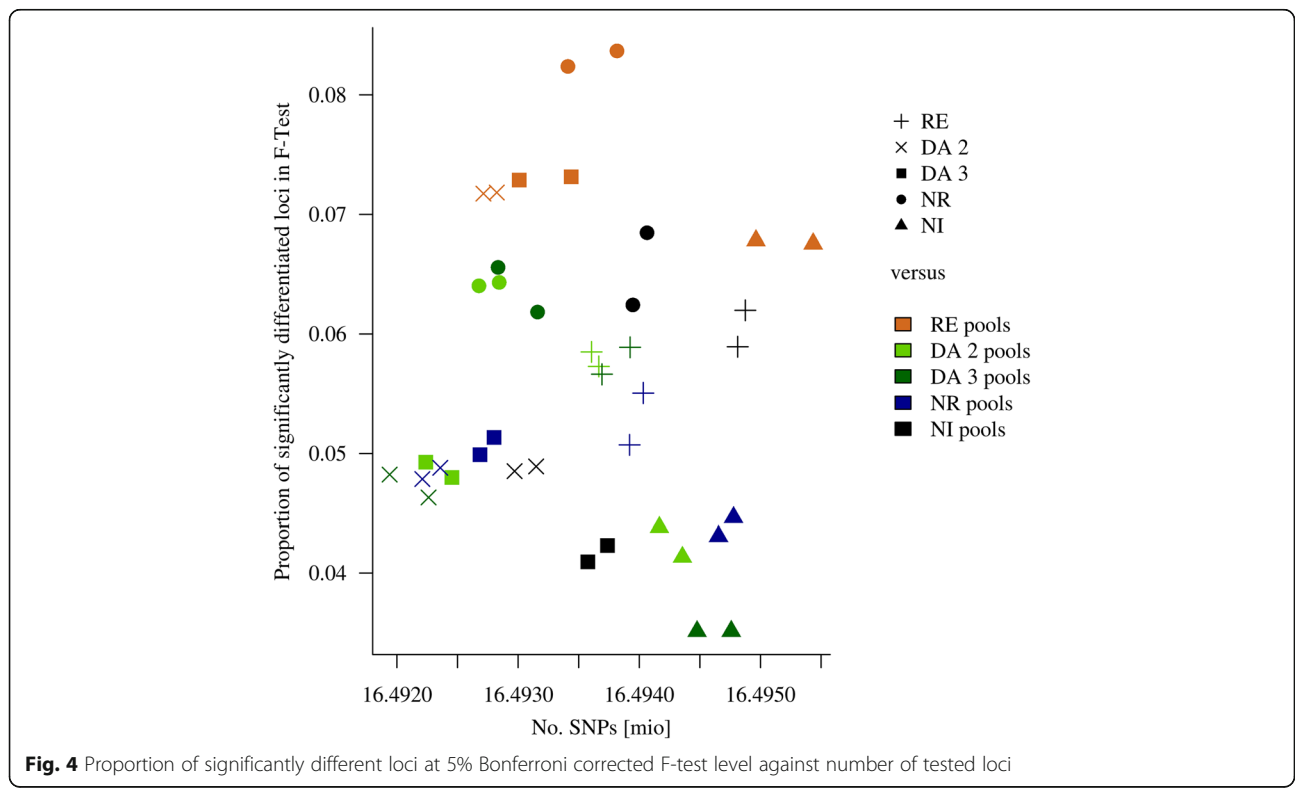

Fig. 4 Proportion of significantly different loci at 5% Bonferroni corrected F-test level against number of tested loci

Table 4 Expected Heterozygosity within pools

	RE_1	RE_2	DA2_1	DA2_2	DA3_1	DA3_2	NR_1	NR_2	NI_1	NI_2
H_{exp}	0.284	0.285	0.278	0.277	0.283	0.276	0.272	0.271	0.280	0.283
SD	0.188	0.185	0.189	0.189	0.189	0.190	0.193	0.193	0.189	0.185
Nloci [mio.]	16.497	16.497	16.496	16.496	16.496	16.497	16.497	16.497	16.497	16.498
NNA [mio.]	1'831	1'455	2'556	2'754	2'493	2'167	2'116	2'204	1'380	1'154

and NI (156′502) with about 80′000 loci carried more than the DA units (163′853 and 134′158).

Annotation

Functional annotation of loci significant in F-test, showing oppositely fixed alleles and exhibiting extreme F-test values revealed that most loci were in intergenic or intronic regions (compare Table 8, i.e. F-test: 65% intron and 19% intergenic) followed by ∼ 13% upstream and downstream variants. Exonic variants were present to an extent of less than 1%. Potential protein changing variants like start or stop codons were barely present at a 5% significance level in the F-test and nearly absent among loci with oppositely fixed alleles. Compared to the unselected background, intergenic, intron, up– and downstream variants were slightly higher represented in both, the 5% F-test level and for the oppositely fixed loci, while exonic variants were in majority less frequent, especially for stop codons and in oppositely fixed loci only five such stop codons were present.

Annotating SNPs in different levels of F_{ST} supported these findings. Start and stop codon changes could only be found at lower F_{ST} levels, while synonymous and missense mutations showed a decline in frequency towards high F_{ST} values, while up- and downstream, intron and intergenic variants were unaffected or increased in frequency (Fig. 5).

In the highest F_{ST} class with values > 0.9 there were 34 missense variants in all pairwise comparisons (Supplementary Table 4), 10 of them had a predictedly deleterious function (Table 9). While some annotations pointed to artefacts or novel genes, 3 of them were located in genes known by name, among them *Carbohydrate Sulfotransferase 12* (*CHST12*, DA2 vs DA3 and DA2 vs NI), *Arrestin Beta 1* (*ARRB1*, DA2 vs NR and NR vs. RE), *Autoimmune Regulator* (AIRE; DA2 vs. NR) and

Calcium/Calmodulin Dependent Protein Kinase II Beta (*CAMK2B*, DA3 vs NR). RVIS scores for those genes were not extreme, with − 0.84 (among 11.18% most intolerant genes) *ARRB1*, − 0.62 (17.16%) for *CAMK2B*, − 0.57 (19.01%) for AIRE and 0.27 (70.58%) for CHST12.

Annotation of loci with variability in only one colony, while they were fixed in all other colonies, resembled the fractions of functional classes already known from the F-test and F_{ST} annotations (Table 10), but due to the higher number of private polymorphic loci in RE, the absolute numbers of loci annotated to potentially protein changing classes, such as missense mutations, was therefore higher in RE (2′380) than in all other colonies. DA3 colony carried the lowest number of missense mutations (538). Still, every colony carried at least one stop codon gain or loss, RE even 15.

Discussion

The aim of our study was to determine whether the integrity of the breed Göttingen Minipigs was compromised by the current production and the genetic management system that relies on genetic isolation of production units. First, the classification of the GMP samples within the context of various pig breeds representing worldwide porcine genetic variation was evaluated with phylogenetic and population genetic methods. Secondly, genetic identity of the breed was assessed by multiple approaches describing variability within and differentiation between the separated barrier colonies.

Discriminability of Göttingen Minipigs from other pig breeds

Our PCA, genetic distance and phylogenetic results show clearly distinct groups of European pigs, Asian pigs and Göttingen Minipigs. The distance between the

Table 5 Expected Heterozygosity estimated from the virtual union of both unit pools

	RE	DA2	DA3	NR	NI
H_{exp}	0.298	0.292	0.294	0.285	0.295
SD	0.175	0.175	0.178	0.181	0.176
Nloci [M]	16.498	16.498	16.498	16.499	16.499
NNA	260	452	346	441	231

Table 6 Correlation between genotypes for loci that were completely fixed within each unit

	RE	DA2	DA3	NR	NI
RE	1.00	−0.10	0.02	0.05	0.14
DA2	−0.10	1.00	−0.07	−0.02	0.01
DA3	0.02	−0.07	1.00	0.02	0.12
NR	0.05	−0.02	0.02	1.00	0.11
NI	0.14	0.01	0.12	0.11	1.00

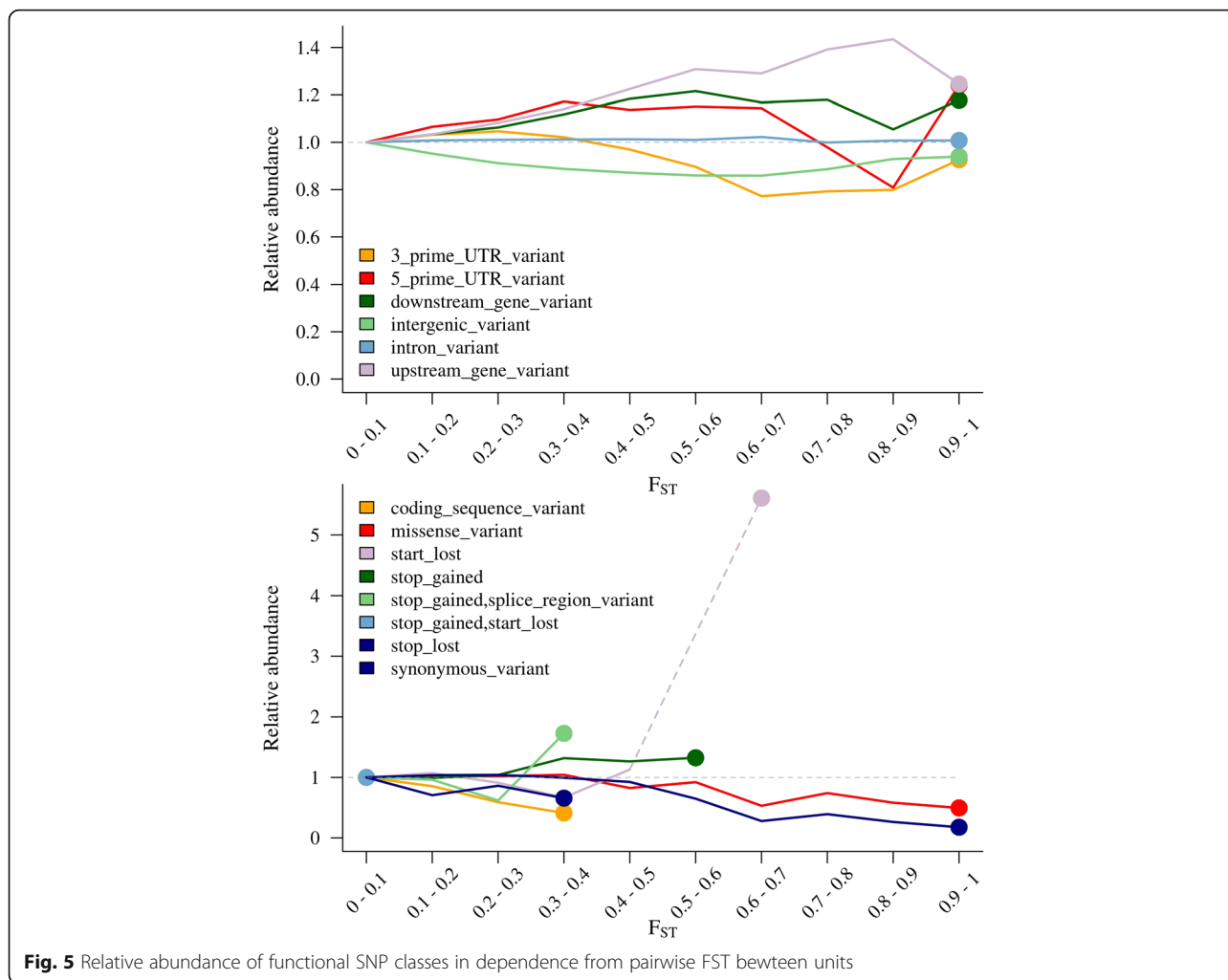**Table 7** Number of private polymorphism; left: completely recorded loci; right: missing information (NA) allowed

	Without NA	with NA
RE	555′591	555′765
DA2	163′853	163′935
DA3	134′158	134′265
NR	192′896	192′974
NI	156′502	156′579

European and Asian breeds reflects the current scientific consensus that domestication happened independently in Europe and Asia about 9000 years ago [8]. The European breeds appear generally closer to each other, which might be explained through the different domestication processes in both centers: while the European breeds emerged more or less directly from relatively uniform wild boar strains [9], the Asian domestication history is characterized by complex human driven dispersal of domesticated pigs in the South East Asian archipelagos, sometimes interrupted by feral states, before pigs eventually reached the Asian mainland [10]. This might explain why the European group clusters closely together in the UPGMA tree with higher resampling support than the Asian group. The tree, based on genome wide SNPs, clusters together, on one hand Xiang, Meishan and the South Chinese wild boars and on the other hand Jiangquhai and the North Chinese wild boars, interrupted by the Mini-LEWE. This is in contradiction to Ai et al. [11] who found that Meishan clustered together with the North Chinese wild boars, and could also support the low resampling probabilities found among the Asian breeds. The Mini-LEWE, a composite miniature breed, developed by crossing Vietnamese Potbellied Pigs, Saddlebacks and German Landrace, is in our study represented by a DNA pool of 10 females and a virtual pool made up from two sequenced individuals. Although it appears that individual sequences are not fully comparable to pools, since mixing of individual sequences and pool sequences leads to clustering of the respective sample types (results not shown). Still, the virtual and the DNA Mini-LEWE pools are clearly identified as one breed. Therefore, the virtual pooling seems to be a suitable measure to make different types of data comparable. In the case of the GMP, both types were mixed and both analyses, PCA and the UPGMA tree show that it is easily discriminable from all other breeds. The phylogenetic tree supports a GMP clade with 100% resampling support, which is located among the Asian pig breeds. This can be explained by the cross-breeding history in which Vietnamese Potbellied Pigs, Minnesota Minipigs and German Landrace were involved [12]. An earlier study [7] estimated that about 70% of the GMP genome are of Asian origin. In the PCA, the first component explains the difference between the three breed groups as the main source of variation, accounting for 19% of the genetic variability, while the second component discriminates the GMP and European breeds from the Asian breeds. Following the interpretation of Kim et al. [13], the average F_{ST} between the three groups, ranging between 0.26 and 0.33, suggests that still a major part of the total variability can be assigned to differences among individuals. Anyway, albeit using

Table 8 Relative amount [%] of significantly differentiated and oppositely fixed loci per functional class and relative abundance of loci in differentiated classes in comparison to all background loci

	Relative amount of loci per class		Relative abundance compared to background	
	5% bonf	Opp. Fixed	5% bonf	Opp. fixed
3_prime_UTR_variant	1.0790	1.1905	0.9906	1.0913
5_prime_UTR_variant	0.2728	0.3345	0.9919	1.2232
coding_sequence_variant	0.0002	0.0000	0.9558	0.0000
downstream_gene_variant	6.4463	7.1600	1.0199	1.1224
intergenic_variant	19.9130	19.4910	0.9892	0.9698
intron_variant	65.4566	64.3371	1.0019	0.9848
missense_variant	0.2416	0.3683	1.0033	1.5229
start_lost	0.0009	0.0000	1.2111	0.0000
stop_gained	0.0020	0.0047	0.9747	2.1566
stop_gained,splice_region_variant	0.0001	0.0000	0.5296	0.0000
stop_gained,start_lost	0.0000	0.0000	5.2287	0.0000
stop_lost	0.0005	0.0000	1.0325	0.0000
synonymous_variant	0.5686	0.4842	1.0205	0.8975
upstream_gene_variant	6.0184	6.6297	0.9945	1.0989

Fig. 5 Relative abundance of functional SNP classes in dependence from pairwise FST bewteen units

microsatellite data, they encountered similar estimates for F_{ST} values in a set of breeds comparable to this study. Therefore, we conclude that the GMP is still a distinct breed that can be easily distinguished from other breeds.

Variation and differentiation within and between the GMP pools

While it seems particularly easy to distinguish the GMP from other pig breeds, it is more difficult, but relevant from a breeders' point of view, to determine, if genetic

Table 9 Annotation of deleterious missense variants with pairwise FST of 1

Chr	Pos [bp]	Pool 1	Pool 2	Ensembl ID	RS ID	SIFT	Gene name
3	1′741′574	DA2	DA3	ENSSSCG00000007567	rs1108271931	0	CHST12
3	1′741′574	DA2	NI	ENSSSCG00000007567	rs1108271931	0	CHST12
9	9′545′629	DA3	NR	ENSSSCG00000014852	rs793754709	0	ARRB1
9	9′545′629	NR	RE	ENSSSCG00000014852	rs793754709	0	ARRB1
10	2′699′680	DA2	DA3	ENSSSCG00000042899	–	0.03	–
10	2′699′680	DA2	NR	ENSSSCG00000042899	–	0.03	–
13	207′155′372	DA2	NR	ENSSSCG00000025294	rs343158168	0.03	AIRE
14	132′394′490	DA2	DA3	ENSSSCG00000046807	–	0.04	–
14	132′394′490	DA2	NR	ENSSSCG00000046807	–	0.04	–
18	50′955′382	DA3	NR	ENSSSCG00000039535	rs713007753	0.01	CAMK2B

Table 10 Relative amount (in per cent) of private polymorphism loci per functional class (absolute number of loci in brackets)

	RE	DA2	DA3	NR	NI
3_prime_UTR_variant	1.2176 (6317)	1.1844 (1790)	1.2121 (1516)	1.2375 (2211)	1.1192 (1634)
5_prime_UTR_variant	0.378 (1961)	0.3474 (525)	0.3622 (453)	0.4125 (737)	0.3199 (467)
coding_sequence_variant	0.0002 (1)	NA (NA)	NA (NA)	0.0006 (1)	NA (NA)
downstream_gene_variant	7.346 (38113)	7.1572 (10817)	7.1015 (8882)	7.2189 (12898)	6.6622 (9727)
intergenic_variant	39.5045 (204959)	39.6937 (59991)	41.8575 (52352)	39.7633 (71045)	40.8015 (59571)
intron_variant	43.3518 (224920)	43.5875 (65876)	41.7496 (52217)	43.179 (77148)	43.6521 (63733)
missense_variant	0.4587 (2380)	0.4268 (645)	0.4302 (538)	0.525 (938)	0.389 (568)
start_lost	0.001 (5)	0.0013 (2)	NA (NA)	NA (NA)	0.0007 (1)
stop_gained	0.0021 (11)	0.0053 (8)	0.0048 (6)	0.0022 (4)	0.0048 (7)
stop_gained,splice_region_variant	NA (NA)	NA (NA)	NA (NA)	NA (NA)	NA (NA)
stop_gained,start_lost	NA (NA)	NA (NA)	NA (NA)	NA (NA)	NA (NA)
stop_lost	0.0008 (4)	0.0007 (1)	NA (NA)	0.0011 (2)	NA (NA)
synonymous_variant	0.6557 (3402)	0.6213 (939)	0.5525 (691)	0.6571 (1174)	0.5082 (742)
upstream_gene_variant	7.0837 (36752)	6.9746 (10541)	6.7297 (8417)	7.0029 (12512)	6.5424 (9552)

isolation of the five breeding colonies has led to differentiated subpopulations. Applying PCA on the 10 GMP pools, we were able to see a trend to three subgroups consisting of NR, RE, and a cluster comprising NI and DA, respectively. The presence of a certain level of stratification is expected and has been observed beforehand, i.e. in studies in dogs [14] or sheep [15]. In the latter study, several breeds with heterogeneous breeding background split into sub clusters, for example dependent on their origin (American Suffolk vs British Suffolk, $F_{ST} \sim 0.058–0.064$; African vs American Dorpers, $F_{ST} = 0.053$) or phenotypic differences (Australian Poll Dorset vs American Dorset, $F_{ST} = 0.082$), while New Zealand and American Texel appeared indistinguishable ($F_{ST} = 0.025$). Studies comparing clearly distinct pig and cattle breeds, respectively, found F_{ST} values between 0.06 and 0.40 [16, 17]. F_{ST} of ~ 0.1 was found between relatively similar breeds, for example Large White and Landrace, while values higher 0.3 indicated major differentiation, such as between Nellore and Holstein cattle or Asian and European pig breeds. These values matched our findings between the European, Asian and GMP groups. Within the GMP, even two randomly composed pools from the same unit had a minimum differentiation of about 0.05. Between the aforementioned clusters F_{ST} was about 0.06 to 0.8 and therefore between the differentiation observed between the sheep breeds from separate origins and clearly distinct breeds. We explain this by genetic drift and slight differences in the actual breeding management, since the three clusters are confounded with the three partners in GMP breeding, even though all follow the same general breeding goal. We also assume that ascertainment bias affecting the variant discovery procedure might have an elevating effect on

differentiation since incorporation of various pig breeds in the discovery sample might allow for calling more variants and especially more heterogeneous variants than calling in a GMP only sample [18].

Comparison of our results with the F_{ST} levels found in the aforementioned studies implies that our colonies are at the edge of splitting into sub-populations, and we did not expect all genomic regions between all pairwise combinations of the five units being similarly differentiated, when focusing on individual loci. The F-test (Table 3) identified about 4 to 8% of the genome to be objected by differentiation which is similar to the range found in a comparable study by Amaral et al. [19]. We hypothesize that genetic differentiation should be attributed to drift rather than to selection, if it affects neutral loci relatively more than loci with putative harmful consequences on protein translation, such as stop codon gains or deleterious missense mutations. This was supported by an under-representation of detrimental variation among highly differentiated loci. The 10 loci representing deleterious missense mutations with maximum F_{ST} were partly located in genes with known function. Although the RVIS values from the genetic intolerance analysis suggest that the respective genes are relatively tolerant to changes in functional variation, there should be further research on the real functionality, since some genes may have influence on important features of GMPs. While high differentiation in one candidate, *CAMK2B*, seems even vital to maintain cellular functions [20, 21], for example, *ARRB1* has found to be involved in feed conversion in pigs [22], *AIRE* is an important pathway gene of auto-immune response [23], and *CHST12* is differentially expressed in Oocytes after nuclear transfer in mice [24]. Nine of these SNPs are identified when the DA2 and DA3

subpopulation are involved in a pairwise comparison, three of them are between DA2 and 3 and five are between a DA unit and NR, indicating that there is detectable differentiation between the DA units and NR. When focusing on the subset of loci where all units are fixed for either allele, it is interesting, that the correlations between all units at such loci range around zero, indicating no clear pattern of shared fixed loci among colonies.

When we looked at expected heterozygosity as a measure of variability, RE and NI exhibit just slightly higher values than the other colonies, but it is notable, that RE holds about as many private polymorphisms as all other units together, making it an indispensable resource of genetic variability. We explain this with the consequent implementation of the mating scheme based on the optimum genetic contribution concept [25] in RE.

Not only is the preservation of a common genetic identity for all colonies of the GMP important [26], but also the risk of inbreeding depression and loss of variation due to drift is increased in artificially reduced subpopulations [6]. To counter this in future, two strategies appear feasible: First, the exchange of genetic material, e.g. via artificial insemination, and second, selection of a most diverse set of breeders as basis for future breeding. The first strategy, in which semen from RE, the major reservoir of remaining variability, would be used to inseminate breeders in the other units, as it is commonly used by dog breeders, would harbor various risks of spreading diseases and disorders between units. The second option of selecting a most diverse set of breeders from the respective unit also has the potential to increase heterozygosity, as can be observed in NI whose founding population was established in that very way. It can be taken as an example of the Bulmer effect [27] that genetic variation in the relatively large colony of DA3 wasn't lost while selection and assortative mating was conducted, and could be largely recovered when the NI founders were chosen for maximum diversity.

Conclusions
Our study based on assessment of differentiation found that the GMP as a breed is easily discriminable from other pig breeds. The finding of genetic distances and differentiation between the isolated breeding colonies of the GMP being minor in comparison to such between distinct breeds is taken as evidence of a successful conservation breeding program, even though some indications of stratification were detected. Functional annotation revealed that loci with functional impact are less differentiated than neutral loci, which implies that the force differentiating the colonies appears to be rather random drift than selective pressure. Selection pressures, through the current breeding activities, might on the other hand have ensured similarity of all colonies in

regions which are expected to underlie traits important for the use of the GMPs as animal models. The detection of putatively deleterious highly differentiated variations in candidate genes with functions important for GMPs suggest that research needs to be undertaken to confirm their functionality in our animals.

Albeit animal exchange seems not yet necessary, the RE subpopulation harbors the highest amount of genetic variation while not being especially differentiated from all other colonies.

Methods
Samples
A joint pedigree was created from the pedigrees of all colonies in the five separated barrier facilities (Research Farm Relliehausen: Relliehausen (RE); Ellegaard Göttingen Minipigs A/S: Dalmose barrier 2 and 3 (DA2, DA3); Marshall BioResources: North Rose (NR); Oriental Yeast Co., Ltd.: Nisshin (NI)). Numerator relationship matrices were constructed with Wrights coefficient of relationship [28] for each colony and all animals alive within a colony in November 2015. A set of 30 individuals was selected for blood sampling with the following procedure in each colony, respectively: all candidates available for blood sampling consisting of only non-pregnant, healthy sows without genetic disorders were identified. A subset of 30 animals was randomly sampled from this list and the relationship within the set (a) and between the animals in the set and all remaining animals in the colony (b) were calculated. Both values were combined in an index $I = 0.8*a - 0.2*b$, to minimize relationship within the samples while maximizing relationship with the sample and the remaining colony. This sampling was repeated up to 25´000 times and restarted every time a new index value went below the previously recorded one. The procedure was stopped after 25´000 rounds without improvement.

DNA of two times ten animals per colony, randomly chosen from the available samples from the previously selected 30 candidates, was pooled using equimolar amounts of the individual DNA. 150 bp paired-end sequencing was done on an Illumina HiSeq 4000 with an aim coverage of 30X and an insertsize of about 420 bp. Publicly available data from 13 various pig breeds [29–31] as used in a previous study [32] were incorporated (see also Table 1). Raw data was aligned to the reference genome Sscrofa11.1 [33] with BWA 0.7.12 [34], sorting, merging of different libraries and marking duplicates were done with Picard tools 2.10.5 [35], base qualities were recalibrated with GATKs BQSR [36, 37] using the available SNPs from dbSNP as validation [38]. Biallelic SNPs were called with the Haplotype Caller from GATK 4.0.8.1 in gVCF-mode. SNPs were filtered with the VQSR tool of GATK that uses machine learning to assess the validity of a SNP. SNPs contained in the

Table 11 Additional porcine samples used in Reimer et al. (2014)

Breed	Number of Samples	Average Depth	Class	Subclass
Duroc	4	5.98	European	Domestic
Hampshire	2	6.49	European	Domestic
Jiangquhai	1	8.20	Asian	Domestic
Large White	14	6.46	European	Domestic
Landrace	5	6.36	European	Domestic
Meishan	4	6.83	Asian	Domestic
Pietrain	5	5.61	European	Domestic
Xiang	2	6.27	Asian	Domestic
European wild boar	6	6.44	European	Wild
Asian wild boar	5	6.27	Asian	Wild
Göttingen Minipigs external	1	12.76	Minipig	Göttingen
Göttingen Minipigs	10	13.01	Minipig	Göttingen
Mini-LEWE	2	13.93	Minipig	Berlin
Mini-LEWE pool	10	13.14	Minipig	Berlin
Wuzhishan	1	11.02	Asian	Domestic

Affymetrix Axiom_PigHD_v1 array were used to train the model incorporating the variant attributes Qualityby-Depth (QD), MappingQuality (MQ), MQRankSumTest, ReadPositionRankSumTest, FisherStrand (FS), StrandOddsRatio (SOR) and depth (DP). A truth sensitivity filter level of 99.0 was applied. Monomorphic loci, X-chromosomal loci and loci without records were removed.

Principal component analysis (PCA)

PCA was constructed from matrix X, which contained samples in rows and genetic loci in columns. Elements of X were the respective reference allele frequencies. Every locus was centered and scaled beforehand. Due to the nature of data from short-read sequencing, matrix X contained numerous missing values. To account for these, missing information was recoded in matrix N with the same dimensions as X, in which records and missing positions were recoded as 1 or 0, respectively. The adjusted covariance structure was subsequently modeled as $E = \frac{XX'}{NN'}$. Eigenvectors V and eigenvalues λ were achieved bei eigenvalue decomposition of E. Loadings L where the corrected eigenvalues where negative values were set to zero. Principal components were calculated by multiplying the eigenvectors with a diagonal-matrix containing square roots of L.

Fixation index and Reynolds distance

Fixation index (F_{ST}) and Reynolds distance (D_R) were estimated between breed pools. Therefore read information of individuals was virtually pooled by breed-wise summation of reads supporting the reference and the alternative allele, respectively.

Reference allele frequency in each breed k per locus was estimated as $p_k = \frac{R_{ref}}{R_{ref}+R_{alt}}$, with $R_{ref/alt}$ denoting the number of reads supporting either the reference or alternative allele, and F_{ST} calculated per locus as F_{ST}

$$= \frac{H_T - \bar{H}_S}{H_T} = \frac{\bar{p}*(1-\bar{p}) - \frac{p_1*(1-p_1) + p_2*(1-p_2)}{2}}{\bar{p}*(1-\bar{p})} \;, \text{ with } \bar{p}$$

$= \frac{p_1+p_2}{2}$. Reynolds distance was estimated as $D_R = \frac{1}{2} *$

$$\frac{\sum_{i=1}^{2}(p_{1i}-p_{2i})2}{1-\sum_{i=1}^{2}p_{1i}p_{p2i}}, \text{ where } i \text{ reflects the } i^{th} \text{ allele at a biallelic}$$

locus, namely the reference allele or the alternative allele, respectively [39]. Both measures were averaged over all pairwise complete loci to gain genome-wide values

Phylogeny

A phylogenetic tree was constructed from genome-wide F_{ST} values from all autosomal loci, using the clustering algorithm UPGMA as implemented in the package "phangorn" [40]. The resulting tree reliability was determined by comparison to 100 trees constructed from 100 randomly sampled loci each.

Test of allele frequency differences between pools

We employed an F-test based statistic to determine statistically significant variation patterns between pools for every locus (eq. 1).

$$F = \frac{V_I}{V_O}, \tag{1}$$

where V_I is the pooled variance within a unit, e.g. RE1 and RE2, estimated as $V_I = \frac{p_{RE1}+p_{RE2}}{2}*(1-\frac{p_{RE1}+p_{RE2}}{2})*\frac{2}{10}$, and where V_O represents the variance between the aforementioned unit and a remote pool, e.g. NI1 estimated as $\frac{p_{RE1}+p_{NI1}}{2}*(1-\frac{p_{RE1}+p_{NI1}}{2})*\frac{2}{10}$. The degrees of freedom where assumed to be nine, since every pool was made up from ten animals.

Heterozygosity, fixed alleles and private polymorphisms

Expected heterozygosity at locus i was estimated from original pools and the virtual pool for each colony as $H_{exp_i} = 2*p_i*(1-p_i)$, where p_i is the reference allele frequency. It was further assessed whether a single colony was fixed for one allele, while the others were fixed for the other allele. To assess variability remaining in only one colony, loci where all colonies apart from one were fixed were identified. This was done both for the subset of loci without missing information and for loci where single colonies had missing information.

Annotation

Loci identified in the aforementioned tests were functionally annotated with the Ensemble Genes database (version 98; Sscrofa11.1 [41]). Genetic intolerance of predicted deleterious missense mutations at highly differentiated loci was assessed by calculation of Residual Variation Intolerance Scores (RVIS [42]).

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12864-020-6590-4.

Additional file 1 Supplementary Table 1. Genome wide FST (upper triangle) and Reynolds distance (lower triangle).

Additional file 2 Supplementary Table 2. Number of SNPs significant at 5% Bonferroni corrected significance level in F-Test.

Additional file 3 Supplementary Table 3. Functional annotation of SNPs oppositely fixed between one unit and a distant pool from another unit.

Additional file 4 Supplementary Table 4. Functional annotation of SNPs dependent on the F_{ST} value between two colonies.

Abbreviations

DA: Colony of Dalmose, Denmark; DNA: Deoxyribonucleic acid; D_R: Reynolds genetic distance; F_{ST}: Fixation index; GMP: Göttingen Minipigs; KEGG: Kyoto Encyclopedia of Genes and Genomes; M: Million; Mini-LEWE: Minischwein Lehnitz-Wendefeld; NI: Colony of Nisshin, Japan; NR: Colony of North Rose, NY, USA; PCA: Principal component analysis; RE: Colony of Relliehausen, Germany; SNP: Single nucleotide polymorphism; UPGMA: Unweighted pair group method with arithmetic mean

Ethics approval

The blood samples were provided by the individual partner organisations and breeders of the Göttingen Minipigs as part of the scientific collaboration and taken under national legislation and individually approved:
Japan: Blood collections performed in our facility were reviewed/approved by Animal Welfare Committee of the Ina MP Breeding Center, Oriental Yeast Co., Ltd.
NY/USA: Blood collection performed at our site in the United States was done in accordance with procedures approved by the Institutional Animal Care and Use Committee (IACUC) at Marshall BioResources.
Denmark: The blood sampling was approved by the Danish Animal Experiment Inspectorate under licenses number 2012-15-2934-00615.
Germany: Blood sampling for this study was approved by the animal welfare body (Tierschutzbeauftragte/ Tierschutzkommission) of the University of Göttingen and is documented under reference E1–20.
All partner organisations have given written informed consent to use the animals in the study, as being owners of the individual animals.

Consent for publication

Not Applicable.

Competing interests

The authors declare that there are no competing interests.

Author details

[1]Department of Animal Sciences, Animal Breeding and Genetics Group, University of Göttingen, Albrecht-Thaer-Weg 3, 37017 Göttingen, Germany. [2]Center for Integrated Breeding Research, University of Göttingen, Albrecht-Thaer-Weg 3, 37075 Göttingen, Germany. [3]Ellegaard Göttingen Minipigs A/S, Soroe Landevej 302, 4261 Dalmose, Denmark. [4]Institute of Mathematics, University of Mannheim, B6 26, 68131 Mannheim, Germany.

[5]Institute of Farm Animal Genetics of the Friedrich-Loeffler-Institut, Höltystraße 10, 31535 Neustadt, Germany.

References

1. Swindle MM, Makin A, Herron AJ, Clubb FJ, Frazier KS. Swine as models in biomedical research and toxicology testing. Vet Pathol. 2012;49:344–56.
2. Simianer H, Köhn F. Genetic management of the Göttingen Minipig population. J Pharmacol Toxicol Methods. 2010;62:221–6.
3. Fitzpatrick JL, Evans JP. Reduced heterozygosity impairs sperm quality in endangered mammals. Biol Lett. 2009;5:320–3.
4. Hedrick PW. Purging inbreeding depression and the probability of extinction: full-sib mating. Heredity (Edinb). 1994;73:363–72.
5. Bosse M, Megens H-J, Madsen O, Crooijmans RPMA, Ryder OA, Austerlitz F, et al. Using genome-wide measures of coancestry to maintain diversity and fitness in endangered and domestic pig populations. Genome Res. 2015;25: 970–81.
6. Lacy RC. Loss of genetic diversity from managed populations: interacting effects of drift, mutation, immigration, selection, and population subdivision. Conserv Biol. 1987;1:143–58.
7. Gaerke C, Ytournel F, Sharifi a R, ECG P, Ludwig A, Simianer H. Footprints of recent selection and variability in breed composition in the Göttingen Minipig genome. Anim Genet. 2014:381–91.
8. Giuffra E, Kijas JMH, Amarger V, Carlborg O, Jeon J-T, Andersson L. The origin of the domestic pig: independent domestication and subsequent introgression. Genetics. 2000;154:1785–91.
9. Scandura M, Iacolina L, Crestanello B, Pecchioli E, Di Benedetto MF, Russo V, et al. Ancient vs. recent processes as factors shaping the genetic variation of the European wild boar: are the effects of the last glaciation still detectable? Mol Ecol. 2008;17:1745–62.
10. Larson G, Cucchi T, Fujita M, Matisoo-Smith E, Robins J, Anderson A, et al. Phylogeny and ancient DNA of Sus provides insights into neolithic expansion in island Southeast Asia and Oceania. Proc Natl Acad Sci U S A. 2007;104:4834–9.
11. Ai H, Fang X, Yang B, Huang Z, Chen H, Mao L, et al. Adaptation and possible ancient interspecies introgression in pigs identified by whole-genome sequencing. Nat Genet. 2015;47:217–25.
12. Glodek P, Oldigs B. Das Göttinger Miniaturschwein. Berlin and Hamburg: Parey; 1981.
13. Kim TH, Kim KS, Choi BH, Yoon DH, Jang GW, Lee KT, et al. Genetic structure of pig breeds from Korea and China using microsatellite loci analysis. J Anim Sci. 2005;83:2255.
14. Quignon P, Herbin L, Cadieu E, Kirkness EF, Hédan B, Mosher DS, et al. Canine population structure: assessment and impact of intra-breed stratification on SNP-based association studies. PLoS One. 2007;2:e1324.
15. Kijas JW, Townley D, Dalrymple BP, Heaton MP, Maddox JF, McGrath A, et al. A genome wide survey of SNP variation reveals the genetic structure of sheep breeds. PLoS One. 2009;4:e4668.
16. Ai H, Huang L, Ren J. Genetic diversity, Linkage Disequilibrium and Selection Signatures in Chinese and Western Pigs Revealed by Genome-Wide SNP Markers. PLoS One. 2013;8:e56001.
17. McKay SD, Schnabel RD, Murdoch BM, Matukumalli LK, Aerts J, Coppieters W, et al. An assessment of population structure in eight breeds of cattle using a whole genome SNP panel. BMC Genet. 2008;9:37.
18. Geibel J, Reimer CHUW, Weigend S, Weigend A, Pook T, Simianer H. How Array Design Affects SNP Ascertainment Bias. bioRxiv. 2019:833541.
19. Amaral AJ, Ferretti L, Megens H-J, Crooijmans RPMA, Nie H, Ramos-Onsins SE, et al. Genome-wide footprints of pig domestication and selection revealed through massive parallel sequencing of pooled DNA. PLoS One. 2011;6:e14782.
20. Verardo LL, Nascimento CS, Silva FF, Gasparino E, Toriyama E, Barbosa AR, et al. Identification and expression levels of pig miRNAs in skeletal muscle. Livest Sci. 2013;154:45–54.
21. Hudmon A, Schulman H. Structure-function of the multifunctional Ca2+/calmodulin-dependent protein kinase II. Biochem J. 2002;364:593–611.
22. Piórkowska K, Żukowski K, Tyra M, Szyndler-Nędza M, Szulc K, Skrzypczak E, et al. The Pituitary Transcriptional Response Related to Feed Conversion in Pigs. Genes (Basel). 2019;10:712.
23. Fierabracci A. Recent insights into the role and molecular mechanisms of the autoimmune regulator (AIRE) gene in autoimmunity. Autoimmun Rev. 2011;10:137–43.
24. Suzuki T, Minami N, Kono T, Imai H. Zygotically activated genes are suppressed in mouse nuclear transferred embryos. Cloning Stem Cells. 2006; 8:295–304.
25. Meuwissen TH. Maximizing the response of selection with a predefined rate of inbreeding. J Anim Sci. 1997;75:934.
26. Bollen P, Ellegaard L. The Göttingen Minipig in pharmacology and toxicology. Pharmacol Toxicol. 1997;80:3–4.
27. Bulmer MG. The effect of selection on genetic variability. Am Nat. 1971;105: 201–11.
28. Wright S. Coefficients of inbreeding and relationship. Am Nat. 1922;56:330–8.
29. Rubin C-J, Megens HJ, Martinez Barrio A, Maqbool K, Sayyab S, Schwochow D, et al. Strong signatures of selection in the domestic pig genome. Proc Natl Acad Sci U S A. 2012;109:19529–36.
30. Fang X, Mou Y, Huang Z, Li Y, Han L, Zhang Y, et al. The sequence and analysis of a Chinese pig genome. Gigascience. 2012;1:16.
31. Vamathevan JJ, Hall MD, Hasan S, Woollard PM, Xu M, Yang Y, et al. Minipig and beagle animal model genomes aid species selection in pharmaceutical discovery and development. Toxicol Appl Pharmacol. 2013;270:149–57.
32. Reimer C, Rubin C-J, Sharifi AR, Ha N-T, Weigend S, Waldmann K-H, et al. Analysis of porcine body size variation using re-sequencing data of miniature and large pigs. BMC Genomics. 2018;19:687.
33. Warr A, Affara N, Aken B, Beiki H, Bickhart DM, Billis K, et al. An improved pig reference genome sequence to enable pig genetics and genomics research. bioRxiv. 2019:668921.
34. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics. 2009;25:1754–60.
35. Picard. http://picard.sourceforge.net/. Accessed 2013-07-26. 2009.
36. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.
37. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. In: Current protocols in bioinformatics. Hoboken: John Wiley & Sons, Inc.; 2013. p. 11.10.1–11.10.33.
38. Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29: 308–11.
39. Eding H, Bennewitz J. Measuring genetic diversity in farm animals. In: Oldenbroek K, editor. Utilisation and conservation of farm animal genetic resources. 1st ed. Wageningen, The Netherlands: Wageningen Academic Publishers; 2007. p. 103–30.
40. Schliep KP. Phangorn: phylogenetic analysis in R. Bioinformatics. 2011;27: 592–3.
41. Hunt SE, McLaren W, Gil L, et al. Ensembl variation resources. Database. 2018;2018:1–12. https://doi.org/10.1093/database/bay119.
42. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genet. 2013;9.

Publisher's Note