ORIGINAL ARTICLE

Use of bioinformatic SNP predictions in differentially expressed genes to find SNPs associated with *Salmonella* colonization in swine

J.J. Uthe[1], L. Qu[1,2], O. Couture[1], S.M.D. Bearson[3], A.M. O'Connor[4], J.D. McKean[4], Y.R. Torres[1], J.C.M. Dekkers[1,5], D. Nettleton[2,5] & C.K. Tuggle[1,5]

1 Department of Animal Science, Iowa State University, Ames, IA, USA
2 Department of Statistics, Iowa State University, Ames, IA, USA
3 USDA, ARS, National Animal Disease Center, Ames, IA, USA
4 Department of Veterinary Diagnostic and Production Animal Medicine, Iowa State University, Ames, IA, USA
5 Center for Integrated Animal Genomics, Iowa State University, Ames, IA, USA

Summary

Asymptomatic *Salmonella*-carrier pigs present a major problem in preharvest food safety, with a recent survey indicating >50% of swine herds in the United States have *Salmonella*-positive animals. *Salmonella*-carrier pigs serve as a reservoir for contamination of neighbouring pigs, abattoir pens and pork products. In addition, fresh produce as well as water can be contaminated with *Salmonella* from manure used as fertilizer. Control of *Salmonella* at the farm level could be through genetic improvement of porcine disease resistance, a potentially powerful method of addressing preharvest pork safety. In this research, we integrate gene expression profiling data and sequence alignment-based prediction of single nucleotide polymorphisms (SNPs) to successfully identify SNPs in functional candidate genes to test for the associations with swine response to *Salmonella*. A list of 2527 genes that were differentially regulated in porcine whole blood in response to infection with *Salmonella enterica* serovar Typhimurium were selected. In those genes, SNPs were predicted using ANEXdb alignments based on stringent clustering of all publically available porcine cDNA and expressed sequence tag (EST) sequences. A set of 30 mostly non-synonymous SNPs were selected for genotype analysis of four independent populations (n = 750) with *Salmonella* faecal shedding or tissue colonization phenotypes. Nine SNPs segregated with minor allele frequency ≥15% in at least two populations. Statistical analysis revealed SNPs associated with *Salmonella* shedding, such as haptoglobin (HP, p = 0.001, q = 0.01), neutrophil cytosolic factor 2 (NCF2 #2, p = 0.04, q = 0.21) and phosphogluconate dehydrogenase (p = 0.066, q = 0.21). These associations may be useful in identifying and selecting pigs with improved resistance to this bacterium.

Introduction

Salmonella enterica is one of the major causes of human food-borne gastroenteritis with an estimated 1.4 million cases and about 500 deaths in the United States annually (Mead *et al.* 1999). Pork is estimated to be one of the top five foods causing non-typhoid *Salmonella* illnesses in humans (Hoffmann *et al.* 2007). The US Food Safety Inspection Service indicated that about 7% of market hog carcasses

sampled between 1998 and 2000 contained *Salmonella* (Rigney *et al.* 2004). Furthermore, a survey carried out by the Collaboration of Animal Health and Food Safety Epidemiology showed in 2005 that approximately 58% of pig farm sites sampled in five states of the United States were positive for *Salmonella*. Most recently, a study conducted on 24 hog-farms in Iowa during 2006–09 found that 71.4% of farms had at least one *Salmonella*-positive faecal sample (Wang *et al.* 2010). *Salmonella*-colonized pigs are usually asymptomatic carriers of the bacterium and can shed *Salmonella* upon exposure to stress, thereby contaminating their pen-mates, shipping equipment and lairage or holding pens of the abattoirs (Hurd *et al.* 2001). *Salmonella* contamination of the food chain extends beyond just contaminated meat and processed pork products. Fresh produce can be contaminated using manure as fertilizer from farms with *Salmonella*-carrier swine (Guan & Holley 2003). Thus, reduction in on-farm *Salmonella* prevalence is an essential step to improve preharvest food safety.

Genetic improvement of disease resistance in animals is a potentially effective and sustainable method for addressing preharvest food safety issues. Genetic control of the porcine immune response to pathogens has been previously reported (Galina-Pantoja *et al.* 2009; Reiner *et al.* 2008; Clapperton *et al.* 2009). Differences in swine immune response have been found in porcine lines of high (HIR) and low immune response created by selection based on antibody and cellular assays (Wilkie & Mallard 1999). In general, HIR line animals demonstrated improved immune functions; however, they also developed autoimmune issues (Wilkie & Mallard 1999; Crawley *et al.* 2005). Several quantitative trail loci (QTL) related with immune response have been described in swine. Those include QTL for humoral as well as stress-induced innate immune response, leucocyte counts, cytokine concentration, mitogen-induced proliferation and levels of prevaccination antibody to *Esherichia coli* (Edfors-Lilja *et al.* 1998, 2000; Reiner *et al.* 2008; Wimmers *et al.* 2009; Lu *et al.* 2011). Furthermore, several specific regions in different porcine chromosomes have been associated with *S. Choleraesuis* burden in spleen and liver at 7 days postexperimental inoculation (Galina-Pantoja *et al.* 2009). It has also been suggested that several immune traits including levels of acute phase protein alpha-1, acid glycoprotein and subset CD11R1+ of peripheral blood mononuclear leucocytes can predict health status of swine (Clapperton *et al.* 2009). Even though these reports show that immune response in pigs is genetically controlled, few specific genetic variants

underlying variation in porcine immune response to *Salmonella* have been identified. In this research, we aimed to identify genetic polymorphisms in genes that are differentially regulated in pigs that persistently shed the bacteria versus low-shedding pigs, as such polymorphisms can be useful in predicting the effectiveness of immune responses. Sequence clustering and alignments were performed on all publicly available porcine cDNA sequences (Couture *et al.* 2009) to predict single nucleotide polymorphisms (SNPs) in the candidate genes selected from pig transcriptomic response data. Using a set of the most promising SNPs identified, 750 pigs from four different populations were genotyped and association of SNPs with *Salmonella* shedding or tissue colonization-related traits was identified.

Materials and methods

Porcine populations

The four porcine populations with qualitative or quantitative *Salmonella* shedding or tissue colonization phenotypes that were used in this study are as follows:

Field population

A recent survey of *Salmonella* incidence in swine farms in Iowa, USA, resulted in the collection of tissues and *Salmonella* faecal shedding data for >7000 animals (Wang *et al.* 2010). Faecal samples were collected on the farm 7 days before transport and marketing and were tested for *Salmonella* presence as described later. From this population, we obtained tissue samples and isolated genomic DNA from 168 faecal-positive pigs and 237 faecal-negative control pigs. These controls were selected from the cohort of animals collected on that farm/day where a positive case was found; where possible we used a sample that had been collected near where the faecal-positive sample was collected. For genotyping, we used these 405 field population pigs.

IAH-Compton population

This experimental challenge population contains a total of 228 pigs from four specific sire families of commercial pigs as described in the study by (van Diemen *et al.* 2002). Briefly, first generation offspring from two boars that were potentially susceptible and resistant as well as two unknown boars were experimentally infected with *S. enterica* serovar Choleraesuis (*S. Choleraesuis*), followed by the enumeration of the bacteria in liver and spleen 7 days

postinfection. The phenotypic data and DNA samples from this population were a gift from Pig Improvement Company. Even though the entire 228-pigs population was genotyped, only offspring (a total of 163 pigs) were used for statistical association analysis.

National Animal Disease Center (NADC)-40 population

This experimental challenge population contains 40 mixed breed pigs that were experimentally inoculated with *S. enterica* serovar Typhimurium (*S.* Typhimurium); details on this experiment were reported by Uthe *et al.* (2009). Briefly, at days 2, 7, 14 and 20 postinoculation (pi), quantitative *Salmonella* faecal shedding data were collected. At day 21 pi, ileo-caecal lymph nodes (ICLN) were collected, and *Salmonella* presence in this tissue was determined as described by Uthe *et al.* (2009).

NADC-77 population

Seventy-seven mixed breed pigs in this population were intranasally challenged at 7 weeks of age with 1×10^9 CFU of *S.* Typhimurium χ4232 as previously described by Uthe *et al.* (2009). At days 1, 2, 7, 14 and 21 pi, *Salmonella* faecal shedding was quantified.

All procedures involving animals in the NADC-40 and NADC-77 populations were lawful and approved by the USDA, ARS, NADC Animal Care and Use Committee.

Sample collection, *Salmonella* bacteriology and DNA isolation

For the field population, individual faecal samples (20–30 g) were collected into labelled plastic bags at the same time as animals were tattooed. Matched belly flap samples were collected at the abattoir from dressed carcasses using a unique slap tattoo number. Belly flap samples were placed in plastic bags and frozen for later DNA preparation. To identify pigs shedding *Salmonella*, qualitative bacteriology was performed as follows: 10 g samples of swine faeces were assayed in duplicates using *Salmonella* enrichment and selective media as previously described (Wang *et al.* 2010). Positive isolates were confirmed by serogroup antiserum agglutination assays.

To isolate DNA from belly flaps (field population) and liver tissues (NADC-40 population), about 20 mg of tissues samples was digested with proteinase K (Invitrogen, Carlsbad, CA, USA). DNA from tissue lysates was extracted using Wizard SV genomic DNA purification system (Promega, Madison, WI, USA) according to the manufacturer's protocol.

For the NADC-77 population, DNA was isolated from the pigs' whole blood preserved in Pax-tubes (PreAnalytiX; Qiagen, Valencia, CA, USA) using the DNeasy Blood and tissue kit (Qiagen), according to the manufacturers instructions. DNA was quantified by Nanodrop (Thermo Scientific, Wilmington, DE, USA), and DNA quality was checked by A260/A280 ratios and agarose gel electrophoresis.

Selection of functional candidate genes and prediction of putative SNPs

As candidates for SNP analysis, genes were selected that our Affymetrix microarray data indicated were differentially expressed in porcine whole blood 2 days after inoculation with *S. enterica* serovar Typhimurium (Tinghua Huang *et al.*, submitted manuscript). Based on the analysis of facal shedding, two classes of pigs were identified in this experimental *S.* Typhimurium challenge using the NADC-40 population: low-shedder pigs stopped shedding the bacteria after day 7 of the experiment and high-shedder pigs continued shedding the bacteria until the end of the 21-day study (Uthe *et al.* 2009). Genes that were differentially regulated between the two shedding classes with $q < 0.1$ and fold change ≥ 1.5 were selected for prediction of SNPs using the ANEXdb database (http://www.anexdb.org; Couture *et al.* 2009). Predicted SNPs in ANEXdb are created from a stringent clustering of similar sequences (using TGICL) and alignment (using CAP3) of all available porcine cDNA sequences submitted to NCBI by February 2009 (see Couture *et al.* 2009). Briefly, the assembly was designed to separate close gene family members as well as alternative splicing by first requiring at most 30-bp unmatched overhangs with at least 30-bp overlapping (being automatically adjusted for long sequences), having at least 94% identity in the overlapping area. Large clusters were then re-clustered using full-length clones, which increases the stringency of the programme, needing an almost perfect match without overhangs (Pertea *et al.* 2003). CAP3 was used on each cluster individually. To extract the SNPs for the sequences of interest, the target sequence for the porcine Affymetrix probe sets was used with BLAST to find the highest scoring match between the Affymetrix target sequence and the AnexDB consensus sequence. The SNPs in the ANEXdb were predicted by comparing the base at each location on the consensus sequence with all available cDNA sequences that overlapped that particular base. If there was a base difference at that location, the estimated frequency of the SNP

was calculated by the number of sequences with that SNP compared with the total number of sequences that overlapped that particular location. In addition to the frequency, the total number of sequences with and without the SNP were also recorded and stored in the database. To select a SNP for further analysis, a minimum of three counts of the minor allele that was different from the allele on the consensus sequence reported by the database was required.

DNA genotyping

DNA genotyping was performed using MassARRAY technology (MassARRAY Compact System; Sequenom Inc., San Diego, CA, USA). Selection of SNPs for multiplex reaction, primer design (Table S1) and genotyping was made using the iPLEX reagent kit and the SEQUENOM TYPER 3.1 software according to the manufacturer's instructions (Sequenom, http://www.sequenom.com). Assays that had a minimum of 80% of genotyping calls and minor allele frequency (MAF) of at least 15% in at least one of the four populations were subsequently statistically analysed. Markers such as NCF1#2 and DDRGK1 that were insufficiently genotyped in Sequenom analysis were completely genotyped using tetra primer amplification refractory mutation system (ARMS)-PCR technology, with primers designed using software at http://cedar.genetics.soton.ac.uk/public_html/primer1.html (Ye *et al.* 2001). PCRs containing 30 ng of DNA template, 10 pmol of each inner primer and 1 pmol of each outer primer were prepared in 10-μl volume using GoTaq Hot Start Green Master Mix (Promega). Touchdown PCR conditions were as follows: 2 min at 95°C followed by 35 cycles of 1 min at 95°C, 1 min at 75°C for the first cycle, decreasing by 1°C each cycle until the temperature of 60°C was reached and continuing at that annealing temperature for the rest of the cycles and then 1 min at 72°C and final extension 10 min at 72°C. PCR products were detected by agarose gel electrophoresis (Table 1).

Statistical analysis

Statistical association analysis of genotype and *Salmonella* shedding and/or tissue colonization phenotype was performed in the statistical computing environment R, with all code available upon request.

NADC-40, NADC-77 and the Compton populations

Because of small sample sizes and/or non-normal distributions, permutation-based hypothesis testing was used to assess the significance of the association of each SNP with each phenotype in each population separately. Within each population, the test statistic used for permutation testing was carefully chosen based on the nature of the phenotype and the structure of block factors unique to each population, as described in the subsections later. Regardless of the test statistic selected, permutation testing was carried out for each population, phenotype and SNP as follows. First, the test statistic was computed for the original data. Then, SNP genotypes were randomly shuffled 19 999 times within the levels of the available block factor(s). The test statistic was then recomputed for each of these 19 999 data sets. The resulting 20 000 test statistics provided a reference distribution under the null hypothesis of no association. A permutation p-value was computed as the proportion of these 20 000 statistics that were as extreme as or more extreme than the statistic observed for the original data.

Details specific to each population are as follows:

IAH-Compton population

Two log-transformed count traits – spleen counts (splC) and liver counts (livC) – were analysed separately for each SNP. Each combination of group, sire and dam was treated as a distinct level of a block factor. The test statistic was chosen to be the weighted average of the within-SNP-genotype pairwise Euclidean distances, with weights being within-SNP-genotype sample size (see the MRPP (multire-

Table 1 Tetra primer amplification refractory mutation system-PCR primers and their products

Gene name and single nucleotide polymorphisms	Outer primers, 5'–3'	Inner, allele specific primers, 5'–3'	PCR products (bp)
NCF1#2 NM_001113220.1:c392G>A(P:S131N)	f: CAGGGCACCCTCAGTGAATACTGCAGCG r: CCGTGTTGGAATCCTGCGTCCTCTGCTT	f: GACGACCTCAAGCTCCCCGCGGAAAG r: CCGACTACCCTCTCCCGCCACTCACTTGT	G allele: 181; A allele: 234; Outer product: 359
DDRGK1 AK231807:c.90G>T(P:frame3:V5L)	f: GCCTACATCACAACCACAGCAACGGGGG r: GATCATGCTCTGAGAGTCCCCGCCAGCC	f: CTGGAAGCCACCGTCATGGTGGCTCTCT r: GAGCAGAGCCGCCACCACCAAATACACTAC	T allele: 324; G allele: 446; Outer product: 716

sponse permutation procedure) statistic proposed by Mielke & Berry 2007).

NADC-40 population

The phenotypes for this population included a univariate binary ICLN trait and a four-dimensional (four time points pi) shedding count trait. Sow was the only block factor. For the binary ICLN trait, a logistic regression model was used with block factor and SNP as a fixed factor of interest. The asymptotic p-value for testing for a SNP effect was computed as described for the field population. However, because the sample size was too small for the asymptotic approximation to work well, the asymptotic p-value was used as the test statistic in the permutation test to provide a finite sample exact p-value.

The four-dimensional shedding count trait was transformed by taking natural logarithms after adding one to each count to avoid infinite logarithms. The test statistic for this four-dimensional log-transformed trait was the MRPP statistic as in the IAH-Compton population, except that the Euclidean distance was computed in the four-dimensional space.

NADC-77 population

The phenotype for this population was a five-dimensional (five time points pi) shedding count trait. The test statistic for this five-dimensional trait was the MRPP statistic as in the IAH-Compton population, except that the Euclidean distance was computed in the five-dimensional space.

Field population

The binary *Salmonella* shedding status trait was used as the response variable. A logistic regression model was used, with farm-and-visit combination as a block factor and SNP as a fixed factor of interest. Owing to the complete separation of the binary trait with either block or the SNP of interest, not all maximum likelihood estimates of parameters were finite. Hence, Firth's penalized likelihood method was used for parameter estimation (Firth 1993). The modified (Rao's) score statistic was computed for the SNP of interest, based on Firth's modified score equation (Heinze & Schemper 2002). Because the number of explanatory variables was very large, the asymptotic p-value was then computed from the score statistic based on a chi-square distribution with degrees of freedom equal to the number of SNP levels minus one.

Error control for multiple tests

Because of multiple hypothesis tests performed in the analyses, we controlled SNP false discovery rates (FDR) from the set of p-values for each trait in each population separately, using a procedure similar to that of Storey & Tibshirani (2003). The default smoother method used by Storey & Tibshirani (2003) was not stable because we had only dozens of p-values. Thus, we chose to set the lambda parameter discussed by Storey & Tibshirani (2003) to be the observed p-value such that the difference between the p-value itself and the empirical distribution function evaluated at this p-value was maximum. When the number of p-values is large, this modification results in more conservative FDR control, i.e., it controls FDR by providing a wider safety margin. Note that this FDR estimator also provides a natural estimate of the proportion of false positives (PFP). Hence, even though the multiple testing procedure was applied to each phenotype-by-population combination separately, the procedure also provides a reasonable estimate of PFP when all hypotheses in the current study are considered together.

Results

Selection of functional candidate genes for SNP analysis

Global gene expression profiling of pigs persistently shedding *S.* Typhimurium versus low-shedder pigs identified 2527 annotated differentially regulated genes ($q < 0.1$, fold change in response to infection ≥ 1.5; Figure 1). Analysis of these genes by the ANEXdb database (http://www.anexdb.org; Couture *et al.* 2009) identified 4346 SNPs in 330 differentially expressed genes that included upregulated as well as downregulated genes in the persistent-shedder and low-shedder categories (Figure 1). To focus on the SNPs most likely to affect protein function, only SNPs located within a predicted open reading frame were selected, resulting in 1236 SNPs in 167 genes. Further, polymorphisms causing a non-synonymous amino acid substitution were chosen, resulting in 65 SNPs in 57 genes. In addition, another 11 SNPs were identified within these differentially expressed genes from predictions in Pig EST Data Explorer (PEDE), a database of cDNA sequences from Western and Chinese breeds (Uenishi *et al.* 2007), or the compre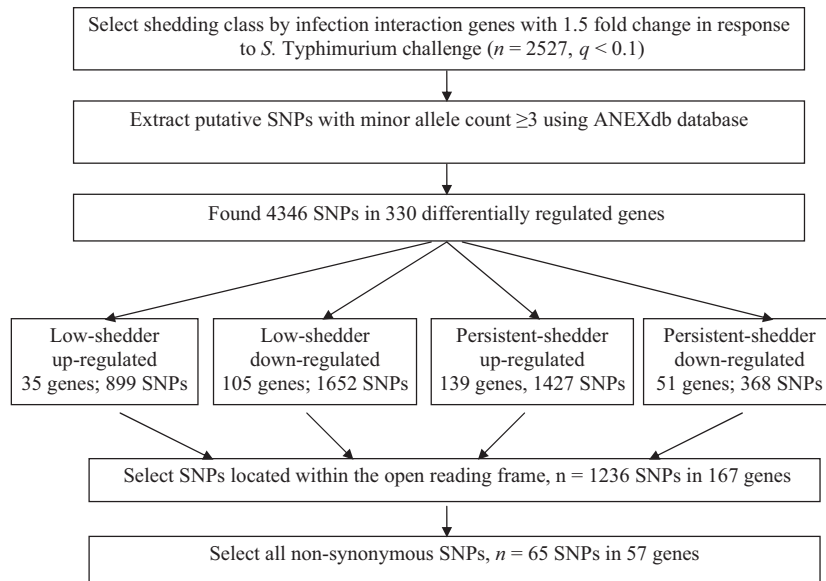hensive DFCI (Dana Farber Cancer Institute) Pig Gene Index (http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/gimain.pl?gudb=pig).

```
┌─────────────────────────────────────────────────────────────────────┐
│   Select shedding class by infection interaction genes with 1.5 fold change in response │
│                 to S. Typhimurium challenge (n = 2527, q < 0.1)                         │
└─────────────────────────────────────────────────────────────────────┘
                                      │
┌─────────────────────────────────────────────────────────────────────┐
│        Extract putative SNPs with minor allele count ≥3 using ANEXdb database           │
└─────────────────────────────────────────────────────────────────────┘
                                      │
┌─────────────────────────────────────────────────────────────────────┐
│              Found 4346 SNPs in 330 differentially regulated genes                      │
└─────────────────────────────────────────────────────────────────────┘
```

| Low-shedder up-regulated 35 genes; 899 SNPs | Low-shedder down-regulated 105 genes; 1652 SNPs | Persistent-shedder up-regulated 139 genes, 1427 SNPs | Persistent-shedder down-regulated 51 genes; 368 SNPs |

```
┌─────────────────────────────────────────────────────────────────────┐
│   Select SNPs located within the open reading frame, n = 1236 SNPs in 167 genes         │
└─────────────────────────────────────────────────────────────────────┘
                                      │
┌─────────────────────────────────────────────────────────────────────┐
│         Select all non-synonymous SNPs, n = 65 SNPs in 57 genes                         │
└─────────────────────────────────────────────────────────────────────┘
```

Figure 1 Strategy for gene selection and putative single nucleotide polymorphisms (SNP) identification from ANEXdb for 65 SNPs. An additional 11 putative SNPs were selected from the Pig EST Data Explorer or DFCI databases.

Genotyping and association analysis of SNPs in functionally relevant genes

Based on the gene selection criteria described earlier, we identified a set of 76 SNPs in 66 genes, and 30 SNPs in 27 genes were able to be multiplexed for genotyping by Sequenom software. Tetra-ARMS PCR manual genotyping was employed to complete genotyping where <80% of alleles were typed by Sequenom; final genotyping data were obtained for 28 SNPs. Thirteen SNPs in 12 genes segregated in the populations tested, another 13 SNPs were monomorphic, and two SNPs came from a misalignment of two closely related genes and were removed from further analysis. Of these segregating SNPs, nine were predicted by ANEXdb alignments (of 21 predictions tested), while two were predicted from PEDE alignments (of five predictions tested) and two from DFCI alignments (of two predictions tested). Thus, this paper reports the analysis results of 13 segregating SNPs in 12 genes. With MAF of 15% or greater, eleven SNPs in ten genes segregated in at least one of the four populations; while nine SNPs in eight genes segregated in two or more populations (Table 2).

Statistical analysis assessed the associations of the SNPs with *Salmonella* shedding or tissue colonization phenotypes (Table 3). In the field population, a SNP in the haptoglobin (HP) gene was associated (p = 0.001, estimated FDR or *q* = 0.01) with *Salmo-*

nella shedding status (Figure 2). In the NADC-77 population, SNP #2 in the neutrophil cytosolic factor 2 (NCF2) gene was potentially associated (p = 0.04, *q* = 0.21) with quantitative faecal bacterial shedding over the time course of infection (Figure 3). A SNP in the phosphogluconate dehydrogenase (PGD) gene also showed suggestive evidence of association (p = 0.07. *q* = 0.21) with bacterial shedding in the NADC-77 population (data not shown).

Discussion

Field and experimental data indicate that some pigs colonized with *Salmonella* become chronic carriers, while others can clear the bacteria (Hurd *et al.* 2001). However, the underlying causes for such differences in response to *Salmonella* at the genetic level are not known. Recent work has characterized the porcine transcriptional response to *Salmonella* (Wang *et al.* 2007, 2008) and has demonstrated that differences between swine in their phenotypic response to this bacterium can be detected at the gene expression level (Tinghua Huang *et al.*, submitted manuscript). We successfully used this novel global gene expression information to select genes that may contribute to host control of *Salmonella*, identified polymorphisms in those genes using our expressed sequence database, and analysed these variants for their association with *Salmonella*-related traits in four porcine populations. An essential part of genetic

Table 2 Segregating single nucleotide polymorphisms (SNPs) genotyped across the four porcine populations

SNP	Nomenclature	NADC-40 population, n = 40		NADC-77 population, n = 77		Field population, n = 405		Compton population, n = 228	
		Allele frequency, %	% pigs genotyped	Allele frequency, %	% pigs genotyped	Allele frequency, %	% pigs genotyped	Allele frequency, %	% pigs genotyped
DDRGK1	AK231807:c.90G>T(P:frame3:V5L)	T-26; G-74	98	T-36; G-64	100	T-24; G-76	93	T-19; G-81	91
ELF3	AK233286.1:c.818T>C(P:frame2:S229P)	C-15; T-85	100	C-16; T-84	100	C-12; T-88	95	C-32; T-68	99
GLS2	AK233325.1:c.1990A>G(P:frame3:R587K)	G-19; A-81	98	G-28; A-72	100	G-23; A-77	97	G-18; A-82	99
HLA-A	AK231632.1:c152C>T(P:frame2:H36Y)	T-9; C-91	95	T-6; C-94	83	T-12; C-88	75	T-49; C-51	63
HP #2	NM_214000.1:c471C>T	T-12; C-88	98	T-39; C-61	95	T-16; C-84	88	T-36; C-64	99
ITIH1#2	NM_213924.1:c227C>T	T-13; C-88	30	Monomorphic	–	T-8; C-92	12	T-13; C-87	34
NCF1#1	NM_0011113220.1:c237G>T(P:R79S)	T-49; G-51	98	G-29; T-71	100	T-50; G-50	90	G-25; T-75	95
NCF1#2	NM_0011113220.1:c392G>A(P:S131N)	A-41; G-59	100	G-30; A-70	99	A-44; G-56	93	G-25; A-75	96
NCF2 #2	TC222915:c1108G>A[a]	A-38; G-63	100	A-45; G-55	96	A-41; G-59	93	A-45; G-55	99
PGD	ITC.15044.548:c742A>G(P:frame3:H214R)[b]	G-42; A-58	93	G-24; A-76	100	A-50; G-50	88	G-11; A-89	98
RNH1 #1	ITC.15044.937:c130A>G (P:frame1:S44G)[b]	G-41; A-59	85	A-48; G-52	66	A-45; G-55	81	A-23; G-77	83
SERPINB1	XM_001926744:c845C>G(P:frame3:D263E)	G-10; C-90	100	G-5; C-95	99	G-4; C-96	95	G-10; C-90	99
SULT1A1 #2	NM_213765.1:c441G>A(P:A112T)	A-6; G-94	100	A-4; G-96	100	A-7; G-93	94	A-20; G-80	94

[a]Sequence from DFCI database: http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/gimain.pl?gudb=pig.
[b]Sequence from ANEXdb database: http://www.anexdb.org.
NCF2, neutrophil cytosolic factor 2; HP, haptoglobin; NADC, National Animal Disease Center; PGD, phosphogluconate dehydrogenase.

Table 3 Association of single nucleotide polymorphisms (SNPs) with phenotypes in the four populations

SNP	NADC-40 population, n = 40		NADC-77 population, n = 77	Field population, n = 405	Compton population, n = 163	
	Fecal counts p-value	Ileo-caecal lymph node counts p-value	Fecal counts p-value	Shedding status p-value	Spleen counts p-value	Liver counts p-value
DDRGK1	0.455	0.301	0.866	0.691	0.779	0.460
ELF3	0.826	0.454	0.744	0.272	0.724	0.877
GLS2	0.637	0.257	0.077	0.448	0.139	0.279
HLA-A	Not analyzed	Not analyzed	Not analyzed	Not analyzed	Not analyzed	Not analyzed
HP #2	0.664	0.644	0.559	0.001*	0.314	0.703
ITIH1#2	Not analyzed	Not analyzed	Not analyzed	Not analyzed	Not analyzed	Not analyzed
NCF1#1	0.253	0.268	0.609	0.740	0.254	0.139
NCF1#2	0.567	0.927	0.442	0.902	0.280	0.299
NCF1 #1+2	0.741	0.383	0.490	0.672	0.461	0.425
NCF2 #2	0.970	0.174	0.044*	0.366	0.630	0.434
NCF2 #1+2	0.970	0.196	0.156	0.150	0.401	0.492
PGD	0.734	0.350	0.066*	0.050	Not analyzed	Not analyzed
RNH1	0.647	0.538	0.125	0.707	0.279	0.458
SERPINB1	Not analyzed	Not analyzed	Not analyzed	Not analyzed	Not analyzed	Not analyzed
SULT1A1 #2	Not analyzed	Not analyzed	Not analyzed	Not analyzed	0.940	0.760

SNPs with least 80% of genotyping calls and minor allele frequency ≥15% statistically analyzed. Asterisk denote a statistically significant association with estimated false discovery rates or $q \leq 0.21$.
NCF2, neutrophil cytosolic factor 2; HP, haptoglobin; NADC, National Animal Disease Center; PGD, phosphogluconate dehydrogenase.

association analysis is creating natural or experimental challenge populations with appropriate phenotypic records, which in large food animals is a very difficult process, involving substantial expense and logistic work. In this research, we were able to use samples and phenotypic data from the largest published experimental *Salmonella* challenge population in pigs, the IAH-Compton population (van Diemen *et al.* 2002). Also, we created two additional experimental challenge populations, NADC-40 and NADC-77. Furthermore, by sampling pigs from 24 farms across Iowa, we created a unique data set representing naturally occurring *Salmonella* colonization status in pigs. Thus, in total for genotyping and association analysis, we used 750 pigs from four populations with a variety of *Salmonella* exposures including *S.* Typhimurium (NADC-40, NADC-77), *S.* Choleraesuis (IAH-Compton) and mixed *Salmonella* spp. (field population).

For prediction and high-throughput extraction of putative and SNPs in differentially expressed genes, we used ANEXdb, a new database with the most recent and largest number of porcine consensus sequences and predicted SNPs, which were created from all publicly available ESTs (Couture *et al.* 2009). Although genome-wide high-throughput genotyping technologies are available to type large

numbers of SNPs, we chose to assess association of a highly refined list of candidate genes. Owing to the relatively small number of animals, genome-wide genotyping would not have provided enough power for association analysis. By selecting a relatively small number of putative SNPs with a minor allele incidence of three or greater and without any preliminary genotyping, we were able to confirm segregation of 43% of ANEXdb-predicted SNPs and 40% of the PEDE-predicted SNPs in our populations. Many of the predicted SNPs in the ANEXdb database are derived from sequence data of a broad population of pig breeds, including diverse non-commercial breeds such as Chinese Meishan and European wild boar which we did not survey in our populations with *Salmonella* colonization phenotypes. Thus, our data validate the usefulness of ANEXdb-predicted SNPs tested at a minor allele incidence level of n = 3. In addition, ANEXdb was useful in confirming SNP predictions in differentially regulated porcine genes that had no annotation at NCBI. In this study, we provide evidences for three associations of 13 segregating SNPs; however, we did not identify SNPs with association across populations. Owing to the different phenotypes measured across populations, associations with a specific phenotype may be difficult to detect in multiple populations.
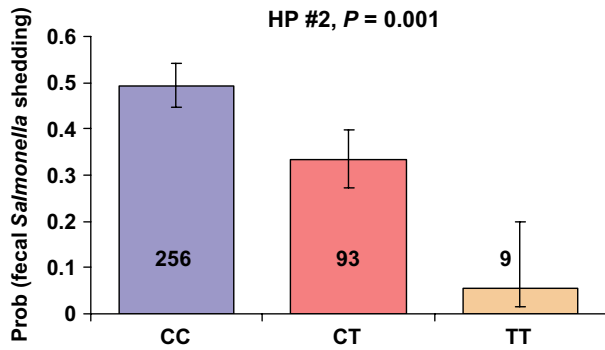
Figure 2 Association of haptoglobin #2 (p = 0.001, *q* = 0.01) with *Salmonella* faecal shedding in the field population. Number of pigs having the particular genotype is indicated on the bar graphs. *Y*-axis represents the fitted probability of positive faecal shedding (see *Materials and methods*).

The segregating SNPs in this study belonged to genes from several differential expression categories, as targeted by the gene selection process. In particular, polymorphisms were identified in genes that were upregulated in persistent-shedders as well as in low-shedders (HP, NCF1, SULT1A1, PGD and SER-PINB1); upregulated only in persistent shedders with no change in expression in low shedders (GLS2, DDRGK1 and RNH1); and upregulated in low shedders but downregulated in persistent shedders (ELF3, HLA-A and ITIH1). The SNPs associating with *Salmonella* shedding or tissue colonization were in genes involved in a variety of cellular functions from immunity to metabolism. The HP gene has a SNP associated with *Salmonella* shedding status in the field population and encodes a multifunctional acute phase protein involved in haemoglobin metabolism and inflammation (Huntoon *et al.* 2008). Expression of HP is induced upon different inflammatory stimuli, preventing loss of iron, protecting cells from oxidative damage; stimulating T-cell responses and development of adaptive immunity (Huntoon *et al.* 2008). In addition, HP is one of the genes that are non-tolerized by iterated lipopolysaccharide (LPS) stimulation (Foster *et al.* 2007). Repeated treatment of a murine macrophage cell line with LPS did not repress but instead further induced the expression of HP, indicating that this gene can directly contribute to antibacterial defence function of the host. In calves, levels of serum HP were correlated with diarrhoea, morbidity and rectal temperatures after mixed infection with *S.* Dublin, *S.* Enteriditis and *S.* Heidelberg (Deignan *et al.* 2000). Swine levels of serum HP serve as a marker of clinical or sub-clinical disease status and correlate with age and weight gain
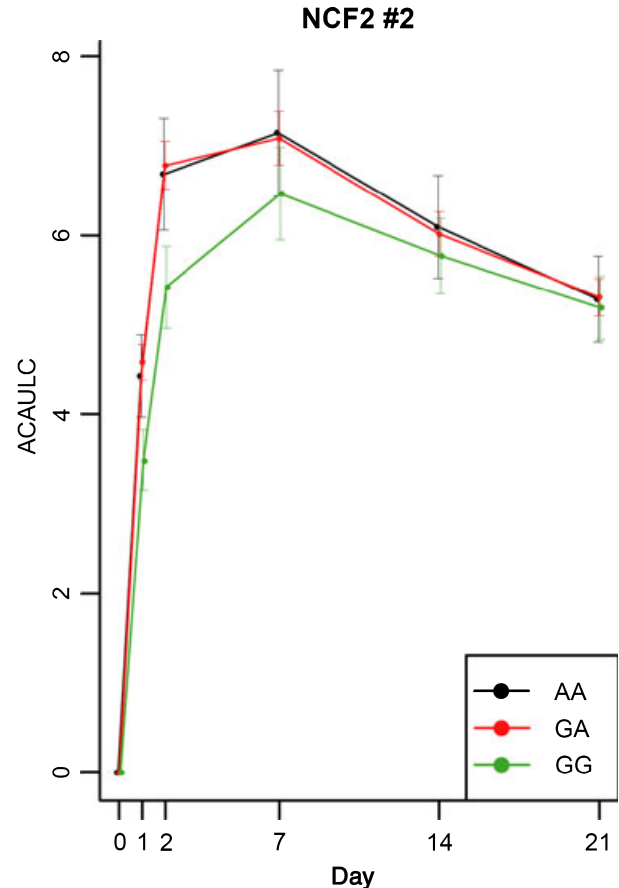

Figure 3 Association of single nucleotide polymorphisms #2 in the neutrophil cytosolic factor 2 gene with *Salmonella* faecal shedding in the National Animal Disease Center-77 population, p = 0.04, *q* = 0.21. There were n = 12 pigs of A/A genotype, n = 42 pigs of G/A genotype and n = 20 pigs of G/G genotype. *Y*-axis represents the day-averaged cumulative area under the log curve of shedding counts, which indicates the average shedding per day from day 0 to each day of interest, on the natural log scale.

(Chung *et al.* 2008). Differences in HP mRNA expression in whole blood were found at 2 days postinoculation with *S.* Typhimurium, with about 2- and 1.3-fold induction in low-shedder and persistent-shedder pigs, respectively (Tinghua Huang *et al.*, submitted manuscript). Other polymorphisms in the porcine HP sequence have been identified by other researchers and were associated with HP baseline serum levels (Diack *et al.* 2008) as well as with serum concentration of complement factor 3c, following immunization against *Mycoplasma hyopneumoniae*, Aujeszky's disease and porcine reproductive and respiratory syndrome virus (Wimmers *et al.* 2009). We report a SNP at position 471 of HP sequence NM_214000.1, with pigs of C/C genotype having the

highest probability of being *Salmonella* shedders compared with C/T and T/T pigs (Figure 2).

A SNP (TC222915:c1108G>A) in NCF2 was potentially associated with *Salmonella* shedding over the time course of infection in the NADC-77 population. The NCF2 gene encodes a cytosolic subunit of neutrophil nicotinamide adenine dinucleotide phosphate (NADPH) oxidase, the enzyme for the production of superoxide anions exerting microbicidal effect in the neutrophil phagosomes. Oxidative burst capacity of neutrophils has been shown to be involved in swine and mice resistance to *Salmonella* (van Diemen *et al.* 2002; Sancho-Shimizu & Malo 2006), suggesting an involvement of NCF2 in this process. Our research indicates that pigs with the G/G genotype at NCF2 TC222915:c1108G>A have significantly reduced the numbers of the bacterium in faeces at 1 and 2 days postinoculation with *S.* Typhimurium.

A SNP in the PGD gene, which causes an amino acid substitution from histidine to arginine at position 214, was marginally associated with *Salmonella* shedding in the NADC-77 population. The enzyme PGD or phosphogluconate dehydrogenase catalyses the oxidation of glucose-6-phosphate to 6-phosphogluconolactone, the first step in the pentose phosphate pathway resulting in the production of pentose sugars and NADPH involved in respiratory burst of neutrophils and macrophages (Mason *et al.* 2007). The activity of NADPH is well known to be involved in host response to *Salmonella* (van Diemen *et al.* 2002), suggesting an involvement of PGD in the swine response to this bacterium. Our microarray analysis revealed that expression of PGD was induced fivefold ($q = 0.001$) in blood of persistent-shedder pigs and twofold ($q = 0.1$) in low-shedder pigs (Tinghua Huang *et al.*, submitted manuscript). Pigs with homozygous G/G genotype of the SNP in PGD had fewer bacteria at day 1 postinoculation and to note, all seven pigs with G/G genotype fell within the low-shedder category.

This research demonstrates a novel approach in the selection of candidate genes for association analyses of genetic polymorphisms by integrating gene expression profiling data and bioinformatic analysis tools. We also recognize that there are genes that could control swine resistance to *Salmonella* that are not differentially expressed in response to this bacterium and that our selection method would not find such genes. In our SNP selection, we specifically targeted non-synonymous SNPs in the genes differentially regulated between pigs belonging to different faecal shedding classes. For the first time, we have associated specific SNPs in the HP and NCF2 genes

and provided suggestive evidence for the association of PGD with *Salmonella* shedding in pigs. These associated SNPs can now be evaluated as potential candidates for selection of pigs with improved ability to control *Salmonella* and shed fewer bacteria. Validation of these associated markers will need to be made as additional and larger populations with *Salmonella*-related phenotypes are developed.

References

Chung H.Y., Park H.J., Yoon S.H., Baek M.G., Kim J.Y., Kim M.J. (2008) Detection of SNPs in porcine haptoglobin and apolipoprotein genes. *Biochem. Genet.*, **46**, 744–754.

Clapperton M., Diack A.B., Matika O., Glass E.J., Gladney C.D., Mellencamp M.A., Hoste A., Bishop S.C. (2009) Traits associated with innate and adaptive immunity in pigs: heritability and associations with performance under different health status conditions. *Genet. Sel. Evol.*, **41**, 54.

Couture O., Callenberg K., Koul N., Pandit S., Younes R., Hu Z., Dekkers J., Reecy J., Honavar V., Tuggle C. (2009) ANEXdb: an integrated animal ANnotation and microarray EXpression database. *Mamm. Genome*, **20**, 768–777.

Crawley A.M., Mallard B., Wilkie B.N. (2005) Genetic selection for high and low immune response in pigs: effects on immunoglobulin isotype expression. *Vet. Immunol. Immunopathol.*, **108**, 71–76.

Deignan T., Alwan A., Kelly J., McNair J., Warren T., O'Farrelly C. (2000) Serum haptoglobin: an objective indicator of experimentally-induced *Salmonella* infection in calves. *Res. Vet. Sci.*, **69**, 153–158.

Diack A.B., Gladney C.D., Mellencamp M.A., Stear M.J., Eckersall P.D. (2008) Associations between polymor-

phisms in the porcine haptoglobin gene and baseline levels of serum haptoglobin. *Dev. Biol. (Basel)*, **132**, 255–259.

van Diemen P.M., Kreukniet M.B., Galina L., Bumstead N., Wallis T.S. (2002) Characterisation of a resource population of pigs screened for resistance to salmonellosis. *Vet. Immunol. Immunopathol.*, **88**, 183–196.

Edfors-Lilja I., Wattrang E., Marklund L., Moller M., Andersson-Eklund L., Andersson L., Fossum C. (1998) Mapping quantitative trait loci for immune capacity in the pig. *J. Immunol.*, **161**, 829–835.

Edfors-Lilja I., Wattrang E., Andersson L., Fossum C. (2000) Mapping quantitative trait loci for stress induced alterations in porcine leukocyte numbers and functions. *Anim. Genet.*, **31**, 186–193.

Firth D. (1993) Bias reduction of maximum likelihood estimates. *Biometrika*, **80**, 27–38.

Foster S.L., Hargreaves D.C., Medzhitov R. (2007) Gene-specific control of inflammation by TLR-induced chromatin modifications. *Nature*, **447**, 972–978.

Galina-Pantoja L., Siggens K., van Schriek M.G.M., Heuven H.C.M. (2009) Mapping markers linked to porcine salmonellosis susceptibility. *Anim. Genet.*, **40**, 795–803.

Guan T.Y., Holley R.A. (2003) Pathogen survival in swine manure environments and transmission of human enteric illness – a review. *J. Environ. Qual.*, **32**, 383–392.

Heinze G., Schemper M. (2002) A solution to the problem of separation in logistic regression. *Stat. Med.*, **21**, 2409–2419.

Hoffmann S., Fischbeck P., Krupnick A., McWilliams M. (2007) Using expert elicitation to link foodborne illnesses in the United States to foods. *J. Food Prot.*, **70**, 1220–1229.

Huntoon K.M., Wang Y., Eppolito C.A., Barbour K.W., Berger F.G., Shrikant P.A., Baumann H. (2008) The acute phase protein haptoglobin regulates host immunity. *J. Leukoc. Biol.*, **84**, 170–181.

Hurd H.S., McKean J.D., Wesley I.V., Karriker L.A. (2001) The effect of lairage on *Salmonella* isolation from market swine. *J. Food Prot.*, **64**, 939–944.

Lu X., Gong Y.F., Liu J.F., Wang Z.P., Hu F., Qiu X.T., Luo Y.R., Zhang Q. (2011) Mapping quantitative trait loci for cytokines in the pig. *Anim. Genet.*, **42**, 1–5.

Mason P.J., Bautista J.M., Gilsanz F. (2007) G6PD deficiency: the genotype–phenotype association. *Blood Rev.*, **21**, 267–283.

Mead P.S., Slutsker L., Dietz V., McCaig L.F., Bresee J.S., Shapiro C., Griffin P.M., Tauxe R.V. (1999) Food-related illness and death in the United States. *Emerging Infect. Dis.*, **5**, 607–625.

Mielke P.W., Berry K.J. (2007) Permutation Methods: A Distance Function Approach. Springer, New York.

Pertea G., Huang X., Liang F., Antonescu V., Sultana R., Karamycheva S., Lee Y., White J., Cheung F., Parvizi B., Tsai J., Quackenbush J. (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics*, **19**, 651–652.

Reiner G., Fischer R., Hepp S., Berge T., Köhler F., Willems H. (2008) Quantitative trait loci for white blood cell numbers in swine. *Anim. Genet.*, **39**, 163–168.

Rigney C.P., Salamone B.P., Anandaraman N., Rose B.E., Umholtz R.L., Ferris K.E., Parham D.R., James W. (2004) *Salmonella* serotypes in selected classes of food animal carcasses and raw ground products, January 1998 through December 2000. *J. Am. Vet. Med. Assoc.*, **224**, 524–530.

Sancho-Shimizu V., Malo D. (2006) Sequencing, expression, and functional analyses support the candidacy of Ncf2 in susceptibility to *Salmonella* Typhimurium infection in wild-derived mice. *J. Immunol.*, **176**, 6954–6961.

Storey J.D., Tibshirani R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.

Uenishi H., Eguchi-Ogawa T., Shinkai H., Okumura N., Suzuki K., Toki D., Hamasima N., Awata T. (2007) PEDE (Pig EST Data Explorer) has been expanded into Pig Expression Data Explorer, including 10 147 porcine full-length cDNA sequences. *Nucleic Acids Res.*, **35**, D650–D653.

Uthe J.J., Wang Y., Qu L., Nettleton D., Tuggle C.K., Bearson S.M.D. (2009) Correlating blood immune parameters and a CCT7 genetic variant with the shedding of *Salmonella* enterica serovar Typhimurium in swine. *Vet. Microbiol.*, **135**, 384–388.

Wang Y., Qu L., Uthe J.J., Bearson S.M.D., Kuhar D., Lunney J.K., Couture O.P., Nettleton D., Dekkers J.C.M., Tuggle C.K. (2007) Global transcriptional response of porcine mesenteric lymph nodes to *Salmonella* enterica serovar Typhimurium. *Genomics*, **90**, 72–84.

Wang Y., Couture O.P., Qu L., Uthe J.J., Bearson S.M.D., Kuhar D., Lunney J.K., Nettleton D., Dekkers J.C.M., Tuggle C.K. (2008) Analysis of porcine transcriptional response to *Salmonella* enterica serovar Choleraesuis suggests novel targets of NFkappaB are activated in the mesenteric lymph node. *BMC Genomics*, **9**, 437.

Wang B., Wesley I.V., McKean J.D., O'Connor A.M. (2010) Sub-iliac lymph nodes at slaughter lack ability to predict *Salmonella* enterica prevalence for swine farms. *Foodborne Pathog. Dis.*, **7**, 795–800.

Wilkie B., Mallard B. (1999) Selection for high immune response: an alternative approach to animal health maintenance? *Vet. Immunol. Immunopathol.*, **72**, 231–235.

Wimmers K., Murani E., Schellander K., Ponsuksili S. (2009) QTL for traits related to humoral immune

response estimated from data of a porcine F2 resource population. *Int. J. Immunogenet.*, **36**, 141–151.

Ye S., Dhillon S., Ke X., Collins A.R., Day I.N. (2001) An efficient procedure for genotyping single nucleotide polymorphisms. *Nucleic Acids Res.*, **29**, E88–88.

Supporting information

Additional Supporting information may be found in a online version of this article:

Table S1 Design of Sequenom assays.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any Supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding for the article.