Predicting deleterious sequence variants in the pig

M. Johnsson[1,2], R. Ros-Freixedes[1], C.Y. Chen[3], A.J. Mileham[4], G. Gorjanc[1], D.J. de Koning[2] & J.M. Hickey[1]

[1] The Roslin Institute and Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Easter Bush, Scotland, UK
[2] Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, SE-750 07, Uppsala, Sweden
[3] Genus Plc, 100 Bluegrass Commons Blvd., Suite 2200, Hendersonville, TN, USA
[4] Genus Plc, 1525 River Road, DeForest, WI, USA

Summary

In this paper, we present the prediction of deleterious variants in noncoding regions and protein-coding genes based on whole-genome sequence data of 673 pigs. Predicted sets are enriched for rare derived variants. The density of predicted deleterious variants across the genome correlates negatively with a proxy for recombination rate. The variants will be used for assessing the value of deleterious variants for increasing the accuracy of genomic selection, and for population genomics of and breeding to reduce deleterious load in pigs.

Keywords: pig, deleterious load, sequencing, population genetics

Introduction

Deleterious variation is a ubiquitous feature of genomes. Damaging mutations happen faster than they can be removed by natural selection, so populations suffer a deleterious load compared to a theoretical mutation-free condition (Haldane 1937). The load is exaggerated in livestock populations, due to population bottlenecks, inbreeding and lack of natural selection (Moyers et al. 2017).

Reducing deleterious load is potentially a way to increase livestock performance. There are two potential approaches. First, deleterious sequence variants could be used as prior information to improve prediction accuracy in genomic selection models (MacLeod et al. 2016, Westhues et al. 2017). Second, gene editing or designed matings could be used to reduce the frequency of individual deleterious variants. Both approaches would need sequence-level predictions of deleterious variants.

The aim of this paper is to predict deleterious sequence variants in pigs, genome-wide, to create deleterious variant sets for genomic selection and population genomics. First, we predict deleterious variants in a whole-genome dataset from a pig breeding population. Then, we assess the derived allele frequency spectra and genomic distribution of predicted deleterious variants.

Methods

Sequence data

The whole-genome dataset consists of variable coverage, whole-genome sequence from 673 individuals of the PIC Genus breeding programme. 58 individuals were sequenced at 30X on the Illumina HiSeq X platform, 598 at between 1X and 5X on the Illumina HiSeq 4000

platform, and the other 16 at intermediate coverage. We performed adapter trimming with Trimmomatic (Bolger et al. 2014), alignment with bwa (Li 2013), alignment deduplication with Picard (http://broadinstitute.github.io/picard/index.html), and variant calling with the GATK HaplotypeCaller (McKenna et al. 2010). We calculated allele frequencies of autosomal variants using VCFTools (Danecek et al. 2011).

Genome alignment and annotation

We created a reference-guided multiple sequence alignment of the pig and 10 other species. We used the pig (Sscrofa11.1), cow (UMD3.1.1), bison (UMD1.0), sheep (Oar_v3.1), goat (ARS1), horse (EquCab2.0), cat (Felis_catus_8.0), dog (CanFam3.1), mouse (GRCm38.p5), human (GRCh38.p10), and chicken (Galgal5) genomes. The multiple sequence alignment was created by pairwise alignment with Lastz (Harris 2007), processing with axtChain and chainNet (Kent et al. 2003), and integration using Multiz (Blanchette et al. 2004).

We estimated a phylogeny from the alignment using RAxML (Stamatakis 2014), using the chicken as an outgroup. We reconstructed ancestral states with the empirical Bayes method of the phangorn R package (Schliep 2010). We called an ancestral allele if the posterior difference between the best and second best base was greater than 0.5.

We classified variants as protein-coding synonymous, nonsynonymous or non-protein coding using the NCBI *Sus scrofa* annotation release 106 and the Variant Effect Predictor (McLaren et al. 2016). As a proxy for recombination rate, we counted exact matches of the CCCCACCCC motif (Tortereau et al. 2012) in 1 megabase bins of the pig reference genome.

Predictions of deleterious variants

We used GERP++ to calculate constraint scores across the multiple genome alignment (Davydov et al. 2010). We consider derived alleles at positions with a GERP score > 1 as potentially deleterious.

We also predicted deleterious variants in protein coding genes using residual variant intolerance scores (RVIS) (Petrovski et al. 2013). We calculated intolerance scores for all genes, and for protein domains derived from Pfam (Finn et al. 2016). We consider nonsynonymous variants in genes and domains at the bottom 5% of the respective intolerance score distributions as potentially deleterious.

Results

The whole-genome sequence data of 673 pigs produced five sets of predicted deleterious variants: constrained noncoding variants (1,062,820), constrained protein-coding variants (21,080), protein-coding variants in intolerant genes (6219), and protein-coding variants in intolerant protein domains (3808). The overlaps between protein-coding variant sets are shown in Figure 1. There is limited overlap between the different prediction sets, with only 538 variants shared between constrained variants and the two intolerance variant sets.

The predicted deleterious variant sets have skewed derived allele frequency spectra, with excess low frequency variants. The constrained noncoding variant set has excess low frequency variants compared to variants that are not constrained (Figure 2a). The coding variant sets have excess low frequency variants compared to the whole set of nonsynonymous variants (Figure 2b). Nonsynonymous variants cause substitutions of amino acids to proteins, and are therefore already likely to be enriched for deleterious variants. However, the excess of low frequency variants is greater in the predicted deleterious variant sets, in particular the constrained protein-coding variants.

There is a negative relationship between the local density of deleterious variants and the local recombination rate. Figure 3a shows the ratio of noncoding deleterious variants to tolerated variants in 1 megabase bins along the chromosome 1. There is a depletion of deleterious variants near the ends of chromosomes, where recombination rate is higher (Tortereau et al. 2012). Figure 3b shows the local relationship between the ratio of constrained noncoding variants to tolerated variants, and the count of a sequence motif associated with high recombination rate, along the whole genome. The Pearson correlation is -0.26.

Discussion

We predicted deleterious sequence variants based on sequence constraint over evolutionary timescales (GERP++), and based on identifying genes and protein domains intolerant to variation (RVIS). We assessed the allele frequency spectra, and the relationship with a proxy for recombination rate, and recovered known features of deleterious variants.

Skewed allele frequency spectra, with excess low frequency variants compared to control variant sets, are consistent with the idea that purifying selection decreases the frequency of deleterious variants (Cooper et al. 2010). This suggests that the predicted variant sets are enriched for deleterious variants.

There is a negative relationship between noncoding constrained variants and recombination rate. This negative correlation is consistent with more efficient selection against deleterious variants in high recombination regions (Hussin et al. 2015). This is a known feature of deleterious variant distributions across genomes, and again, suggests that the constrained noncoding variant set is enriched for deleterious variants.

The prediction methods rely on different principles, and result in variant sets of different size and allele frequency skew, and with limited overlap. The variant sets based on sequence constraint (GERP++) show a greater excess of low frequency variants, but also contain more variants, many of which are likely to be false positives. Protein-coding mutations have larger deleterious effects than noncoding mutations (Kryukov et al. 2005), which may be a reason to favour protein-coding variant sets. It remains to be seen how the different variant sets perform in terms of genomic prediction accuracy.

In summary, we predicted potentially deleterious sequence variants in the pig, and showed that they are enriched for rare derived variants, and enriched in genomic regions of low recombination rate. The variant sets will be used for assessing the value of deleterious variants for increasing the accuracy of genomic selection, and for population genomics of and breeding to reduce deleterious load in pigs.

Acknowledgements
The authors acknowledge the financial support of BBSRC ISPG to The Roslin Institute BB/P013759/1, BBSRC Grant No. BB/N004736/1, Genus plc, and The Swedish Research Council Formas Dnr 2016-01386. This work has made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF) (http://www.ecdf.ed.ac.uk).

References

Blanchette, M., W. J. Kent, C. Riemer, L. Elnitski, A. F. A. Smit, K. M. Roskin, R. Baertsch, K. Rosenbloom, H. Clawson, and E. D. Green. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res. 14:708–715.

Bolger, A. M., M. Lohse, and B. Usadel. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120.

Cooper, G. M., D. L. Goode, S. B. Ng, A. Sidow, M. J. Bamshad, J. Shendure, and D. A. Nickerson. 2010. Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. Nat. Methods 7:250–251.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, and S. T. Sherry. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–2158.

Davydov, E. V, D. L. Goode, M. Sirota, G. M. Cooper, A. Sidow, and S. Batzoglou. 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput. Biol. 6:e1001025.

Finn, R. D., P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, and A. Sangrador-Vegas. 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 44:D279–D285.

Haldane, J. B. S. 1937. The effect of variation of fitness. Am. Nat. 71:337–349.

Harris, R. S. 2007. Improved pairwise alignment of genomic DNA. The Pennsylvania State University.

Hussin, J. G., A. Hodgkinson, Y. Idaghdour, J.-C. Grenier, J.-P. Goulet, E. Gbeha, E. Hip-Ki, and P. Awadalla. 2015. Recombination affects accumulation of damaging and disease-associated mutations in human populations. Nat. Genet. 47:400–404.

Kent, W. J., R. Baertsch, A. Hinrichs, W. Miller, and D. Haussler. 2003. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc. Natl. Acad. Sci. 100:11484–11489.

Kryukov, G. V, S. Schmidt, and S. Sunyaev. 2005. Small fitness effect of mutations in highly conserved non-coding regions. Hum. Mol. Genet. 14:2221–2229.

Li, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv Prepr. arXiv1303.3997.

MacLeod, I. M., P. J. Bowman, C. J. Vander Jagt, M. Haile-Mariam, K. E. Kemper, A. J. Chamberlain, C. Schrooten, B. J. Hayes, and M. E. Goddard. 2016. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genomics 17:144.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, and M. Daly. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–1303.

McLaren, W., L. Gil, S. E. Hunt, H. S. Riat, G. R. S. Ritchie, A. Thormann, P. Flicek, and F. Cunningham. 2016. The ensembl variant effect predictor. Genome Biol. 17:122.

Moyers, B. T., P. L. Morrell, and J. K. McKay. 2017. Genetic costs of domestication and improvement. bioRxiv:122093.

Petrovski, S., Q. Wang, E. L. Heinzen, A. S. Allen, and D. B. Goldstein. 2013. Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genet. 9:e1003709.

Schliep, K. P. 2010. phangorn: phylogenetic analysis in R. Bioinformatics:btq706.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

Tortereau, F., B. Servin, L. Frantz, H.-J. Megens, D. Milan, G. Rohrer, R. Wiedmann, J. Beever, A. L. Archibald, and L. B. Schook. 2012. A high density recombination map of the pig reveals a correlation between sex-specific recombination and GC content. BMC Genomics 13:586.

Westhues, M., T. A. Schrag, C. Heuer, G. Thaller, F. H. Utz, W. Schipprack, A. Thiemann, F. Seifert, A. Ehret, and A. Schlereth. 2017. Omics-Based Hybrid Prediction In Maize. bioRxiv:134668.

Figures

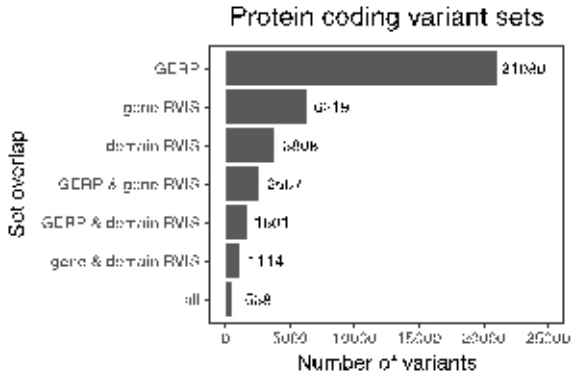

Figure 1. Overlap between sets of predicted protein-coding deleterious variants.

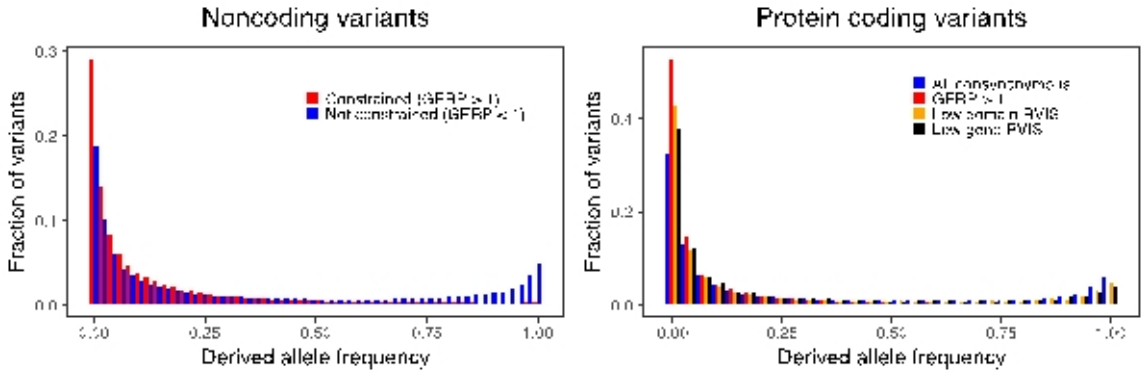

Figure 2. Derived allele frequency spectra of a) noncoding variants, comparing constrained and tolerated variants; and b) protein-coding variants, comparing predicted deleterious sets to all nonsynonymous variants.

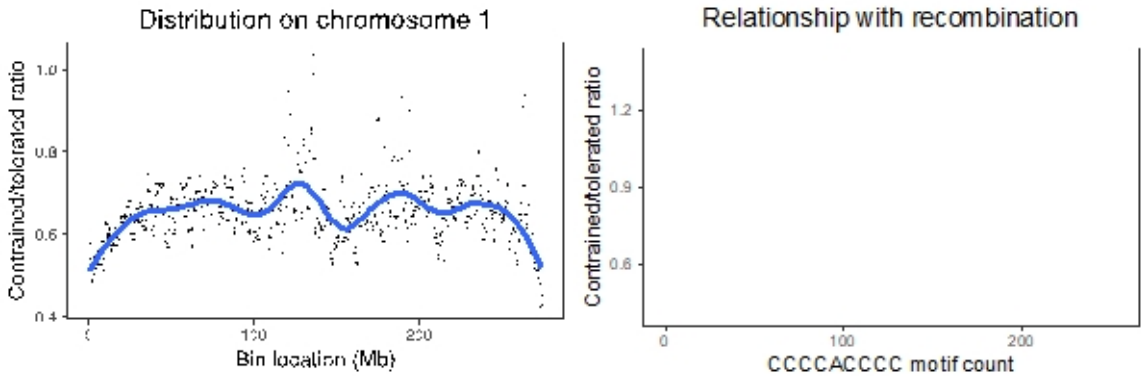

Figure 3. a) Ratio of constrained noncoding (GERP score > 1) to tolerated variants in 1 megabase bins on pig chromosome 1. b) Relationship between constrained to tolerated ratio and count of a recombination associated sequence motif in 1 megabase bins along the pig genome.