# Sequence variability, constraint and selection in the *CD163* gene in pigs

Martin Johnsson[1,2], Roger Ros-Freixedes[1], Gregor Gorjanc[1], Matt A. Campbell[3], Sudhir Naswa[3], Kimberly Kelly[3], Jonathon Lightner[3], Steve Rounsley[3], John M. Hickey[1]*

[1] The Roslin Institute and Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Midlothian, EH25 9RG, Scotland, United Kingdom.

[2] Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Box 7023, 750 07 Uppsala, Sweden.

[3] Genus plc, 1525 River Road, DeForest, WI 53532, USA.

* Corresponding author: john.hickey@roslin.ed.ac.uk

## Abstract

**Background**

In this paper, we investigate sequence variability, evolutionary constraint, and selection on the *CD163* gene in pigs. The pig *CD163* gene is required for infection by porcine reproductive and respiratory syndrome virus (PRRSV), a serious pathogen with major impact on pig production.

**Results**

We used targeted pooled sequencing of the exons of *CD163* to detect sequence variants in 35,000 pigs of diverse genetic backgrounds and search for potential knock-out variants. We then used whole genome sequence data from three pig lines to calculate a variant intolerance score, which measures the tolerance of genes to protein coding variation, a selection test on

protein coding variation over evolutionary time, and haplotype diversity statistics to detect recent selective sweeps during breeding.

**Conclusions**

We performed a deep survey of sequence variation in the *CD163* gene in domestic pigs. We found no potential knock-out variants. *CD163* was moderately intolerant to variation, and showed evidence of positive selection in the lineage leading up to the pig, but no evidence of selective sweeps during breeding.

Key words: targeted sequencing, variant intolerance, positive selection, selective sweep, PRRSV

## Introduction

In this paper, we investigate sequence variability, evolutionary constraint, and selection on the *CD163* gene in pigs. The pig *CD163* gene is required for infection by porcine reproductive and respiratory syndrome virus (PRRSV) [1], a serious pathogen with major impact on pig production [2]. PRRSV-resistant genome-edited pigs with modified *CD163* have been developed, either by knocking out the gene completely or targeting only the fifth scavenger receptor cysteine-rich (SRCR) domain, which is the one that the virus exploits [3–6].

The physiological functions of *CD163* include clearing of haemoglobin from blood plasma [7], adhesion of nucleated red blood cells to macrophages during red blood cell differentiation [8], and immune signalling [9–11]. When red blood cells rupture and haemoglobin is released into the blood stream, haemoglobin is bound by haptoglobin, and the haptoglobin—haemoglobin complex is taken up by macrophages using *CD163* receptors on

48 their surface [7]. Since its natural function is receptor-mediated endocytosis, *CD163* is a target
49 for pathogens entering cells. At least one other virus, the simian hemorragic fever virus [12],
50 has independently evolved to target *CD163*.

51 Given that genome editing of *CD163* has led to PRRVS-resistant pigs, we wanted to
52 see if natural loss-of-function variants for the *CD163* gene could be identified in elite pigs, in
53 order to investigate the opportunity to develop PPRSV resistance from within existing breeding
54 programs. The aims of this paper were to survey *CD163* sequence variation for such naturally
55 occurring knock-out variants, and to put *CD163* variability in the genomic context of genomic
56 variant intolerance and selection. We used targeted pooled sequencing of the exons of *CD163*
57 to detect sequence variants in 35,000 pigs of diverse genetic backgrounds. We then used whole
58 genome sequence data from three pig lines to put the *CD163* results in the context of the whole
59 genome. We used three complementary population genetic analyses: a variant intolerance
60 score, which measures the tolerance of genes to protein coding variation; a selection test on
61 protein coding variation over evolutionary time; and haplotype diversity statistics to detect
62 recent selective sweeps during breeding.

63 In summary, our results show that there were no potential knock-out variants in *CD163*
64 in these pigs. Furthermore, *CD163* was moderately intolerant to variation, and showed
65 evidence of positive selection in the lineage leading up to the pig, but no evidence of selective
66 sweeps during breeding.

67

## 68 Materials and Methods

69 We used targeted *CD163* exon sequence data from pools of 35,000 pigs of diverse
70 origins, and whole genome sequence data from three lines of pigs from the PIC breeding
71 programme. We used three complementary population genetic analyses: 1) residual variant

intolerance score based on segregating exonic SNPs; 2) a gene-based selection test incorporating synonymous and nonsynonymous divergence between the pig and cattle reference genomes; and 3) a selective sweep statistic based on haplotype diversity in imputed whole-genome sequence data from one of the lines.

## Data

We used targeted exon sequencing of the *CD163* gene from 35,000 pigs. The sample included nine lines from the breeding programme of the Pig Improvement Company (PIC). The DNA samples were previously collected in 2011-2016 as part of the operation of the breeding programme.

To put the targeted sequence data in a genomic context, we used whole-genome sequence data from three lines of pigs of the PIC breeding programme. These lines were also sampled in the targeted exome sequencing. We used 1146 individuals from line 1, sequenced at variable coverages. 84 of them were sequenced at 30X coverage, 11 at 10X coverage, 45 at 5X coverage, 561 at 2X coverage, and the remaining 445 at 1X coverage. The individuals and their sequence coverages were chosen with the AlphaSeqOpt algorithm [13,14], with the addition of sires that contributed a large proportion of the genotyped progeny in line 1 that were genotyped as part of the routine breeding activities of PIC. We used 408 individuals from line 2, and 638 individuals from line 3, all of them sequenced at 2X coverage. These individuals were sires that contributed a large proportion of the genotyped progeny in lines 2 and 3 that were genotyped as part of the routine breeding activities of PIC.

## Targeted sequencing of *CD163*

We used a hierarchical pooling strategy to be able to sequence *CD163* exons in many individuals cost-effectively. We pooled 96 DNA samples into one combined DNA sample and

97 constructed a shotgun sequencing library using the ThruPLEX Tag-seq kit from Rubicon
98 Genomics. This kit incorporates unique molecular identifiers that allow for a consensus
99 sequence to be generated from reads originating from the same molecule thus reducing the
100 impact of sequencing errors. 24 such barcoded libraries were combined and used as input into
101 a sequence capture reaction using baits designed against the exons of *CD163* gene (Arbor
102 Biosciences, Ann Arbor, MI). The product of the library capture was then used to generate
103 2x150bp reads on an Illumina MiSeq sequencer. This pooling scheme allowed us to sequence
104 up to 2304 samples per sequencing run. In total, 35,808 animals were sampled using this
105 scheme.

106 We aligned reads with bwa mem (v 0.7.15-r1140) [15] against the 10.2 version of the
107 pig genome with an added 33 kbp contig representing the *CD163* genomic region, which was
108 missing from this version of the reference genome. The coding sequence of *CD163* on this
109 contig is identical to the sequence in the version 11.1 of the pig reference genome. We used
110 Connor (https://github.com/umich-brcf-bioinf/Connor) to call consensus sequences from reads
111 with the same unique molecular identifier. We called variants from these consensus alignments
112 using the LoFreq variant caller [16]. We used snpEff [17] to classify the variants as
113 synonymous, nonsynonymous and stop gain variants.

114

115 ## Validation of potential knock-out variants

116 The potential stop gain variants detected in the pooled targeted sequencing data were
117 validated by sequencing of individual animals. We went back to the pools where the variants
118 were detected and sequenced amplicons of the appropriate exons from all the samples making
119 up the pool with individual barcodes on the MiSeq. None of the potential stop gain variants
120 were validated by individual sequencing.

121

## Whole-genome sequence data processing

We aligned reads to the pig genome (Sscrofa11.1) with bwa mem [15], removed duplicates with Picard (https://broadinstitute.github.io/picard/index.html), and called variants with the GATK HaplotypeCaller [18]. We filtered and processed variant call format files with VCFtools [19].

We used the Variant Effect Predictor [20] to find the protein-coding SNPs, and classifying them into synonymous and nonsynonymous SNPs. We used the Ensembl gene annotation [21], version 90. We downloaded variants in *CD163* from the Ensembl variation database.

## Residual variant intolerance score

The residual variant intolerance score [22] measures gene-level tolerance to mutations by counting segregating variants. To calculate the residual variant intolerance, we counted the number of nonsynonymous variants and the total number of variants in each gene, and calculated the studentised residual of the regression between them. We included variants that segregated in at least one of the three lines.

We applied residual variant intolerance score both at the level of the gene, and at the level of the protein domain [23], using protein domains found by identifying Pfam profiles in Ensembl protein sequences with PfamScan [24]. All gene-level analyses were performed on the principal transcript as designated with APPRIS annotation [25].

## Selection analysis in linage leading up to the pig

SnIPRE [26] uses a Poisson model to measure gene-level selection based on between-species divergence and within-species polymorphism. We calculated the divergence between the pig and cattle (UMD 3.1.1) reference genomes using the Nei-Gojobori method [27] which

147 finds the number of potential synonymous and nonsynonymous substitutions between two
148 codons. We aligned the reference genomes using Lastz [28], and refined the alignments using
149 the chain/net method [29]. We excluded all codons that were not fully aligned between
150 genomes, that is, any codon containing an alignment gap or a missing base in any of the
151 genomes. We ran the empirical Bayes implementation of SnIPRE, using the lme4 R package
152 [30].
153

154 ## Selective sweep analysis by haplotype diversity

155 We estimated haplotype diversity at CD163, at random 100 control genes of similar
156 length, and at 11 homologs of genes that are stably expressed in humans [31]. The control
157 genes were selected at random from genes of similar genomic length as CD163 (at most 10%
158 difference).
159 We imputed genome-wide sequence data to 65,000 pigs from line 1, using SNP chip
160 genotypes from 60K or a 15K SNP chip and the line 1 sequence data described above. We
161 extracted mapped read counts supporting each allele from low coverage samples, as outlined
162 in [Ros-Freixedes et al, in prep], and used multilocus hybrid peeling [32], as implemented in
163 AlphaPeel, to phase and impute all individuals to full sequence data in windows around the
164 selected genes.
165 We extracted all variants falling within exons and introns of the gene in the reference
166 genome and identified haplotypes carried by each individual in each gene, including only
167 genotyped individuals. For variants encompassing each gene, including introns, strings of
168 phased alleles were compared to define haplotypes carried by each individual in each parental
169 chromosome. Strings of alleles that were identical (with a mismatch threshold) between two
170 individuals were considered the same haplotype, and strings with multiple mismatches were
171 considered as different haplotypes. A maximum of two allele mismatches were allowed before

two strings were considered different haplotypes to account for sequencing or phasing errors. We then calculated haplotype homozygosities based on the pooled frequency of the two most common haplotypes ($H_{12}$) [33].

### Gene Ontology enrichment

We downloaded Gene Ontology Biological Process terms for Ensembl genes from BioMart [34], and ranked enriched biological processes by the p-value from Fisher's Exact test.

## Results

### *CD163* sequence variants identified

We used a hierarchical pooling strategy to sequence the exons of *CD163* from over 35,000 pigs from nine lines, and *CD163* variants from whole-genome sequencing of 1146, 638, and 408 pigs from three of the same lines.

Targeted sequencing of exons identified 140 single nucleotide variants in *CD163*. Whole genome sequencing in three lines identified 15 single nucleotide variants in *CD163*, two of which were nonsynonymous, the rest synonymous, and no potential knock-out variants. Table 1 shows the numbers of synonymous and nonsynonymous single nucleotide variants found in each dataset, and the overlap between them. Figure 1 shows the locations of variants in the *CD163* protein sequence and their frequencies in targeted and whole-genome sequencing.

The targeted sequencing also identified 14 potential knock-out variants. We followed up on these variants by performing individual sequencing of the animals constituting the pool from which the potential knock-out variant was identified. In this way, we ruled out all the

196 potential knock-out variants as false positives, likely caused by polymerase errors during
197 amplification before incorporation of unique molecular identifiers.

198 The *CD163* variants detected in whole genome sequence data of the three pig lines were
199 mostly concordant with the targeted sequencing. The discordant SNPs found in the whole
200 genome sequence data were rare. There is one nonsynonymous shared variant, K851R, which
201 occurs at high minor allele frequency. Out of the sequence variants detected, 10 of the variants,
202 most of them at higher frequency, were already present in the Ensembl variation database.
203

204 ## Residual variant intolerance score

205 *CD163* was not among the most variant intolerant genes, as measured by a residual
206 variant intolerance score. Figure 2 shows the distributions of gene-level and protein domain-
207 level residual variant intolerance scores with the position of *CD163* and its five variable
208 domains. *CD163* ranked as number 894 out of 17,982 variable autosomal genes. The five
209 variable SRCR domains of *CD163* ranked as 1037 (domain 9), 2686 (domain 7), number 8125
210 (domains 2 and 6), and 14,147 (domain 8) out of 19,930 variable protein domains, as measured
211 by residual variant intolerance score applied to protein domains identified with the Pfam
212 database.

213 We used the bottom 2% of the genome-wide residual variant intolerance distribution to
214 highlight 358 variant intolerant genes. They were enriched for basal cellular processes such as
215 microtubule-based movement, cell adhesion, and calcium ion transport (Figure 3).
216

217 ## Selection in the lineage leading up to the pig

218 *CD163* showed evidence of positive selection in the lineage leading up to the pig, as
219 estimated by the SnIPRE model. Figure 4 shows the selection estimates from the SnIPRE
220 model, highlighting positively and negatively selected genes and the position of *CD163*. We

found a total of 1125 putatively selected genes, 778 positively selected genes, and 347 putatively negatively selected. Positively selected genes in the lineage leading up to pig were enriched for cell surface receptor signalling, proteolysis, protein phosphorylation and terms related to lipids (Figure 3).

### Selective sweep analysis

We investigated haplotype diversity in *CD163* in one of the lines using imputed whole genome sequence data. *CD163* showed no evidence of recent selective sweep. We calculated the selective sweep test statistic $H_{12}$, and compared it to 100 randomly selected control genes of similar length. Figure 5 shows $H_{12}$ at *CD163*, the 100 control genes, a set of homologs of genes that are stably expressed in humans, and randomly selected genes labelled as intolerant by their residual variant intolerance score.

## Discussion

In this paper, we investigated sequence variability, evolutionary constraint, and selection on the *CD163* gene in pigs. We identified synonymous and nonsynonymous variants, but no potential knock-out variants in the gene. We found that *CD163* is relatively tolerant to variation, shows evidence of positive selection in the lineage leading up to the pig, and no evidence of selective sweeps during breeding. In the light of these results, we will discuss (i) variant intolerance scores; (ii) selection on *CD163* in the lineage leading up to the pig; (iii) the lack of evidence of selective sweeps; and (iv) technical aspects of the targeted exome sequencing method.

## Residual variant intolerance score

Variant intolerance scores measure the lack or excess of common nonsynonymous variants in a gene [22]. A low variant intolerance score for a gene indicates that its sequence is constrained, and correlates with gene essentiality [35]. The intermediate variant intolerance scores of *CD163* suggest that it is moderately constrained in the pig.

The 2% extreme tail of the variant intolerance distribution was enriched for genes related to microtubule-based movement. This is consistent with an enrichment of microtubule-genes in human essential genes [35].

## Selection in the lineage leading up to the pig

The SnIPRE model is a generalized mixed linear model that estimates the selection effect on each gene with the number of fixed nonsynonymous substitutions compared to an outgroup species [26]. A positive selection estimate means that when comparing the pig to the cow, there has been significantly more nonsynonymous fixed substitutions than expected under neutrality. The synonymous sites are assumed to be under negligible selection. The positive selection effect for *CD163* suggests that its sequence is quite flexible, and has rapidly evolved in the lineage leading up to the pig. Rapid evolution of *CD163* is consistent with its known role in infection. The estimated positive selection on other cell surface genes, including the T cell surface proteins *CD3*, *CD5* and *CD8* and immunoglobulin E receptor *FCER1A*, is consistent with previous observations of rapid immune gene evolution in pigs [36].

## Selective sweep analysis

Selective sweeps occur when the fixation of one or more beneficial variants affects the allele frequencies at linked sites [37]. This signal of recent selection can be detected from population genetics data. When the beneficial variant is already present in the population as

269 standing variation, selection may give rise to a so called soft sweep, which may be more
270 difficult to detect than a sweep arising from a new mutation [38]. Since selection on standing
271 variation is the expectation in animal breeding, we used a statistic designed to detect soft
272 sweeps [33]. The lack of a selective sweep at *CD163* suggest that this gene has not been a
273 target of strong recent selection during pig breeding. However, selective sweep analysis cannot
274 rule the possibility that *CD163* variants could have small effects on some quantitative trait that
275 may have been subjected to subtle allele frequency shifts by selection.
276

277 ### Technical aspects of the targeted exon sequencing

278 Targeted exome sequencing of pooled samples is a feasible way to cost efficiently
279 sequence a gene in many individuals. However, as our unsuccessful validation of potential stop
280 gain variants show, this method suffers from low frequency false positives, likely due to
281 polymerase errors before incorporation of unique molecular identifiers. The targeted exon
282 sequencing and whole genome sequencing are in good agreement about higher frequency
283 variants, but since the targeted sequencing sampled a wider span of pig diversity, the rare
284 variant calls may represent genuine rare variants or sequencing errors. However, with the depth
285 of sequencing and validation, we are confident that there are no natural knock-out variants in
286 *CD163* in these pigs.
287

288 ## Conclusions

289 We performed a deep survey of sequence variation in the *CD163* gene in domestic pigs.
290 We found no potential knock-out variants. *CD163* was moderately intolerant to variation, and
291 showed evidence of positive selection in the lineage leading up to the pig, but no evidence of
292 selective sweeps during breeding.

**Funding**

The authors acknowledge the financial support from the BBSRC ISPG to The Roslin Institute BBS/E/D/30002275, from Grant Nos. BB/N015339/1, BB/L020467/1, BB/M009254/1, from Genus PLC, Innovate UK, and from the Swedish Research Council Formas Dnr 2016-01386.

**Author's contributions**

SR, MJ, JMH, JL, MAC, and GG conceived the study. SN and KK performed experiments. MJ, RRF, SN and SR analysed data. MJ, JMH, RRF and SR wrote the paper. All authors read and approved the final manuscript.

**Acknowledgements**

This work has made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF) (http://www.ecdf.ed.ac.uk).

## References

1. Calvert JG, Slade DE, Shields SL, Jolie R, Mannan RM, Ankenbauer RG, et al. CD163 expression confers susceptibility to porcine reproductive and respiratory syndrome viruses. J. Virol. Am Soc Microbiol; 2007;81:7371–9.

2. Holtkamp DJ, Kliebenstein JB, Neumann EJ, Zimmerman JJ, Rotto HF, Yoder TK, et al. Assessment of the economic impact of porcine reproductive and respiratory syndrome virus on United States pork producers. J. Swine Heal. Prod. 2013;21:72–84.

3. Wells KD, Bardot R, Whitworth KM, Trible BR, Fang Y, Mileham A, et al. Replacement of porcine CD163 scavenger receptor cysteine-rich domain 5 with a CD163-Like homolog confers resistance of pigs to genotype 1 but not genotype 2 porcine reproductive and

317 respiratory syndrome virus. J. Virol. Am Soc Microbiol; 2017;91:e01521-16.

318 4. Whitworth KM, Rowland RRR, Ewen CL, Trible BR, Kerrigan MA, Cino-Ozuna
319 AG, et al. Gene-edited pigs are protected from porcine reproductive and respiratory syndrome
320 virus. Nat. Biotechnol. Nature Research; 2016;34:20–2.

321 5. Whitworth KM, Lee K, Benne JA, Beaton BP, Spate LD, Murphy SL, et al. Use of
322 the CRISPR/Cas9 system to produce genetically engineered pigs from in vitro-derived oocytes
323 and embryos. Biol. Reprod. Oxford University Press; 2014;91:71–8.

324 6. Yang H, Zhang J, Zhang X, Shi J, Pan Y, Zhou R, et al. CD163 knockout pigs are
325 fully resistant to highly pathogenic porcine reproductive and respiratory syndrome virus.
326 Antiviral Res. Elsevier; 2018;

327 7. Kristiansen M, Graversen JH, Jacobsen C, Sonne O, Hoffman H-J, Law SKA, et al.
328 Identification of the haemoglobin scavenger receptor. Nature. Nature Publishing Group;
329 2001;409:198–201.

330 8. Fabriek BO, Polfliet MMJ, Vloet RPM, van der Schors RC, Ligtenberg AJM,
331 Weaver LK, et al. The macrophage CD163 surface glycoprotein is an erythroblast adhesion
332 receptor. Blood. Am Soc Hematology; 2007;109:5223–9.

333 9. Bover LC, Cardó-Vila M, Kuniyasu A, Sun J, Rangel R, Takeya M, et al. A
334 previously unrecognized protein-protein interaction between TWEAK and CD163: potential
335 biological implications. J. Immunol. Am Assoc Immnol; 2007;178:8183–94.

336 10. Van den Heuvel MM, Tensen CP, van As JH, Van den Berg TK, Fluitsma DM,
337 Dijkstra CD, et al. Regulation of CD 163 on human macrophages: cross-linking of CD163
338 induces signaling and activation. J. Leukoc. Biol. Soc Leukocyte Biology; 1999;66:858–66.

339 11. Philippidis P, Mason JC, Evans BJ, Nadra I, Taylor KM, Haskard DO, et al.
340 Hemoglobin scavenger receptor CD163 mediates interleukin-10 release and heme oxygenase-
341 1 synthesis. Circ. Res. Am Heart Assoc; 2004;94:119–26.

12. Caì Y, Postnikova EN, Bernbaum JG, Yú S, Mazur S, Deiuliis NM, et al. Simian hemorrhagic fever virus cell entry is dependent on CD163 and uses a clathrin-mediated endocytosis-like pathway. J. Virol. Am Soc Microbiol; 2015;89:844–56.

13. Gonen S, Ros-Freixedes R, Battagin M, Gorjanc G, Hickey JM. A method for the allocation of sequencing resources in genotyped livestock populations. Genet. Sel. Evol. BioMed Central; 2017;49:47.

14. Ros-Freixedes R, Gonen S, Gorjanc G, Hickey JM. A method for allocating low-coverage sequencing resources by targeting haplotypes rather than individuals. Genet. Sel. Evol. BioMed Central; 2017;49:78.

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

16. Wilm A, Aw PPK, Bertrand D, Yeo GHT, Ong SH, Wong CH, et al. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res. Oxford University Press; 2012;40:11189–201.

17. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). Taylor & Francis; 2012;6:80–92.

18. 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. Cold Spring Harbor Lab; 2010;20:1297–303.

19. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. Oxford University Press; 2011;27:2156–8.

20. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The

367 ensembl variant effect predictor. Genome Biol. BioMed Central; 2016;17:122.

368 21. Aken BL, Achuthan P, Akanni W, Amode MR, Bernsdorff F, Bhai J, et al. Ensembl

369 2017. Nucleic Acids Res. Oxford University Press; 2016;45:D635–42.

370 22. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic intolerance to

371 functional variation and the interpretation of personal genomes. PLoS Genet. Public Library of

372 Science; 2013;9:e1003709.

373 23. Gussow AB, Petrovski S, Wang Q, Allen AS, Goldstein DB. The intolerance to

374 functional genetic variation of protein domains predicts the localization of pathogenic

375 mutations within genes. Genome Biol. BioMed Central; 2016;17:9.

376 24. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam

377 protein families database: towards a more sustainable future. Nucleic Acids Res. Oxford

378 University Press; 2016;44:D279–85.

379 25. Rodriguez JM, Maietta P, Ezkurdia I, Pietrelli A, Wesselink J-J, Lopez G, et al.

380 APPRIS: annotation of principal and alternative splice isoforms. Nucleic Acids Res. Oxford

381 University Press; 2012;41:D110–7.

382 26. Eilertson KE, Booth JG, Bustamante CD. SnIPRE: selection inference using a

383 Poisson random effects model. PLoS Comput. Biol. Public Library of Science;

384 2012;8:e1002806.

385 27. Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and

386 nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 1986;3:418–26.

387 28. Harris RS. Improved pairwise alignment of genomic DNA. The Pennsylvania State

388 University; 2007.

389 29. Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. Evolution's cauldron:

390 duplication, deletion, and rearrangement in the mouse and human genomes. Proc. Natl. Acad.

391 Sci. National Acad Sciences; 2003;100:11484–9.

30. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. arXiv Prepr. arXiv1406.5823. 2014;

31. Eisenberg E, Levanon EY. Human housekeeping genes, revisited. Trends Genet. Elsevier; 2013;29:569–74.

32. Whalen A, Ros-Freixedes R, Wilson DL, Gorjanc G, Hickey JM. Hybrid peeling for fast and accurate calling, phasing, and imputation with sequence data of any coverage in pedigrees. bioRxiv. Cold Spring Harbor Laboratory; 2017;228999.

33. Garud NR, Messer PW, Buzbas EO, Petrov DA. Recent selective sweeps in North American Drosophila melanogaster show signatures of soft sweeps. PLoS Genet. Public Library of Science; 2015;11:e1005004.

34. Smedley D, Haider S, Durinck S, Pandini L, Provero P, Allen J, et al. The BioMart community portal: an innovative alternative to large, centralized data repositories. Nucleic Acids Res. Oxford University Press; 2015;43:W589–98.

35. Bartha I, di Iulio J, Venter JC, Telenti A. Human gene essentiality. Nat. Rev. Genet. Nature Publishing Group; 2017;nrg-2017.

36. Groenen MAM, Archibald AL, Uenishi H, Tuggle CK, Takeuchi Y, Rothschild MF, et al. Analyses of pig genomes provide insight into porcine demography and evolution. Nature. Nature Research; 2012;491:393–8.

37. Smith JM, Haigh J. The hitch-hiking effect of a favourable gene. Genet. Res. (Camb). Cambridge University Press; 1974;23:23–35.

38. Hermisson J, Pennings PS. Soft sweeps. Genetics. Genetics Soc America; 2005;169:2335–52.

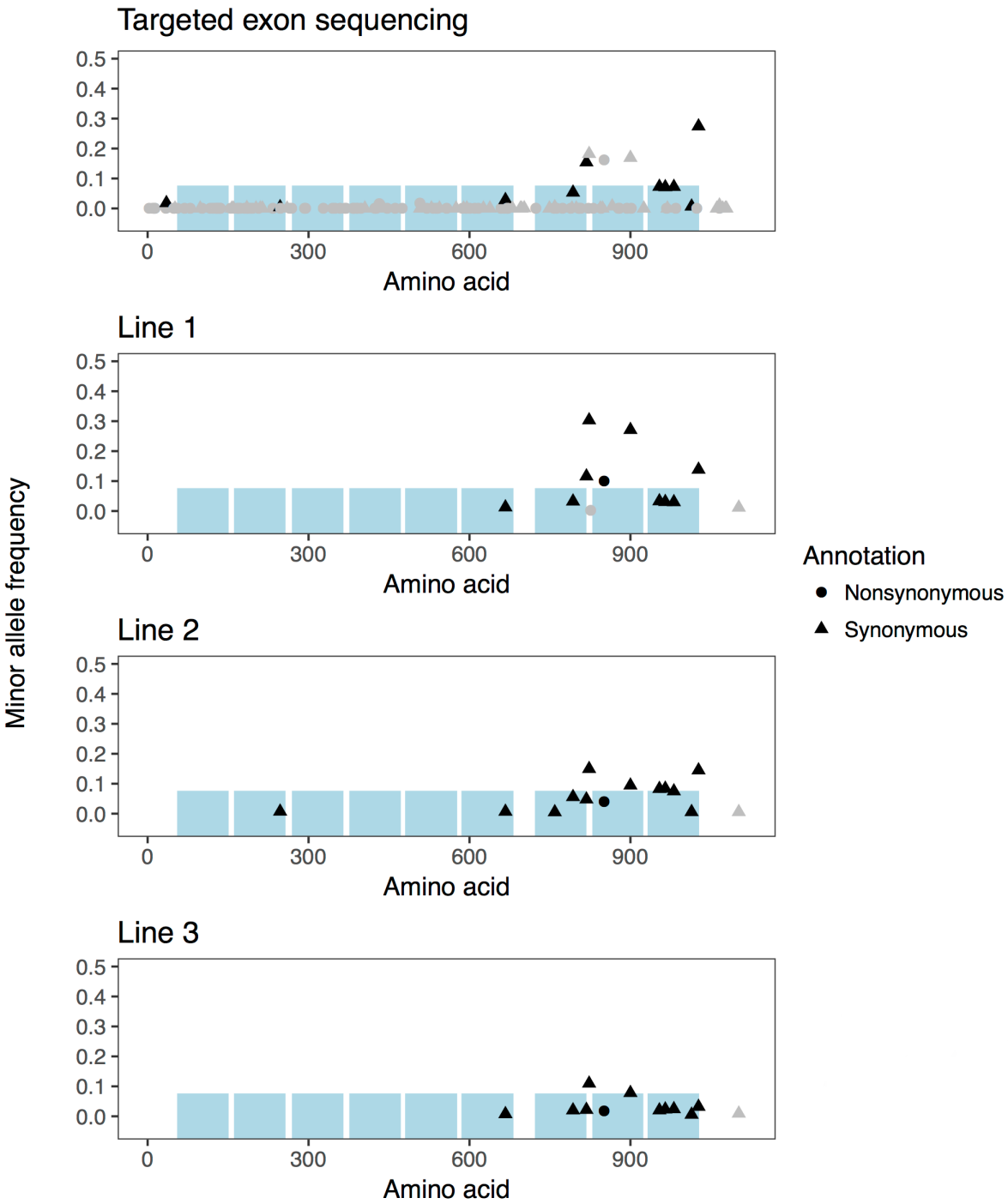
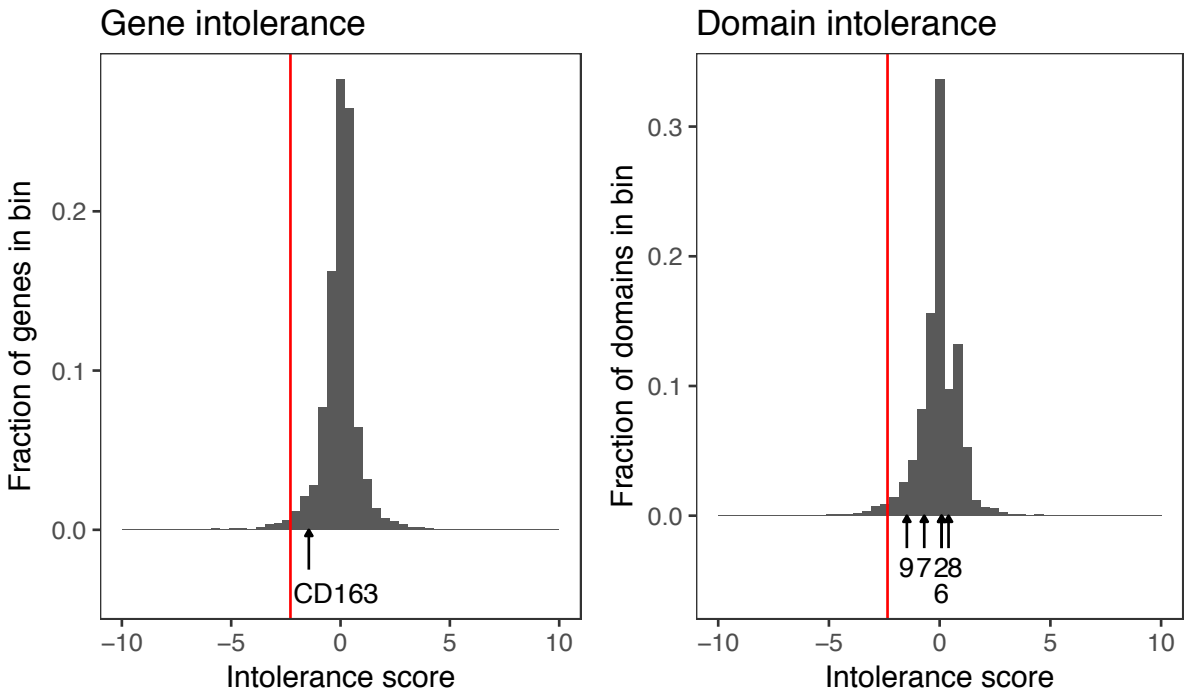



*Figure 1: Protein-coding SNPs in CD163 in targeted exon sequencing and whole genome sequencing of three lines. Minor allele frequency showing that discordant variants are rare. Grey and black coloured points indicate replication. Grey dots in targeted sequencing are variants not present in the Ensembl variation database. Grey dots in the whole genome sequenced lines are variants not replicated by the targeted exon sequencing. The blue boxes represent SRCR domains.*




424

425 *Figure 2: Residual variant intolerance score distributions with CD163 highlighted. The red*

426 *line is the 2% threshold, and arrows indicate the position of CD163 and five of its SRCR*

427 *domains.*

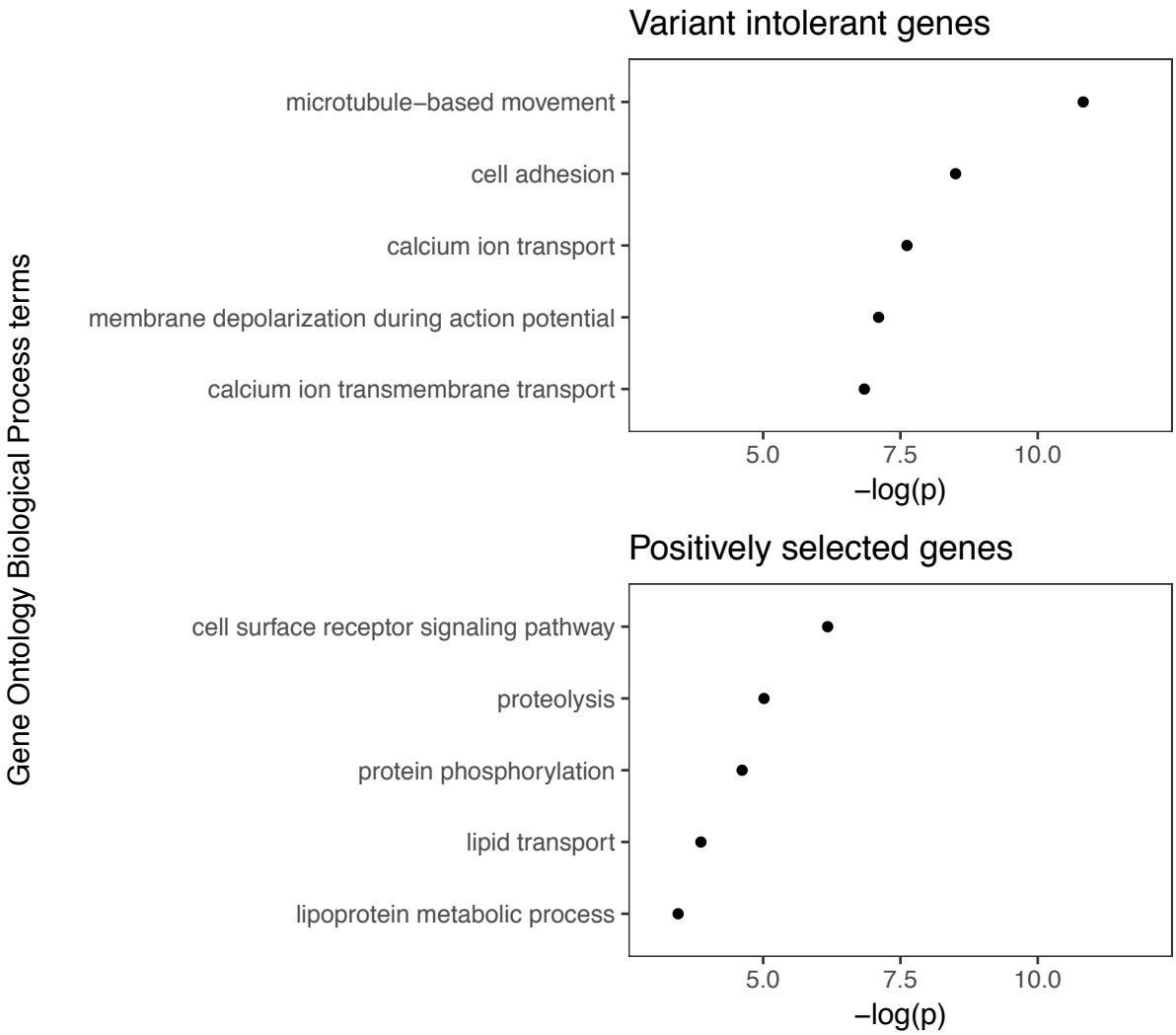
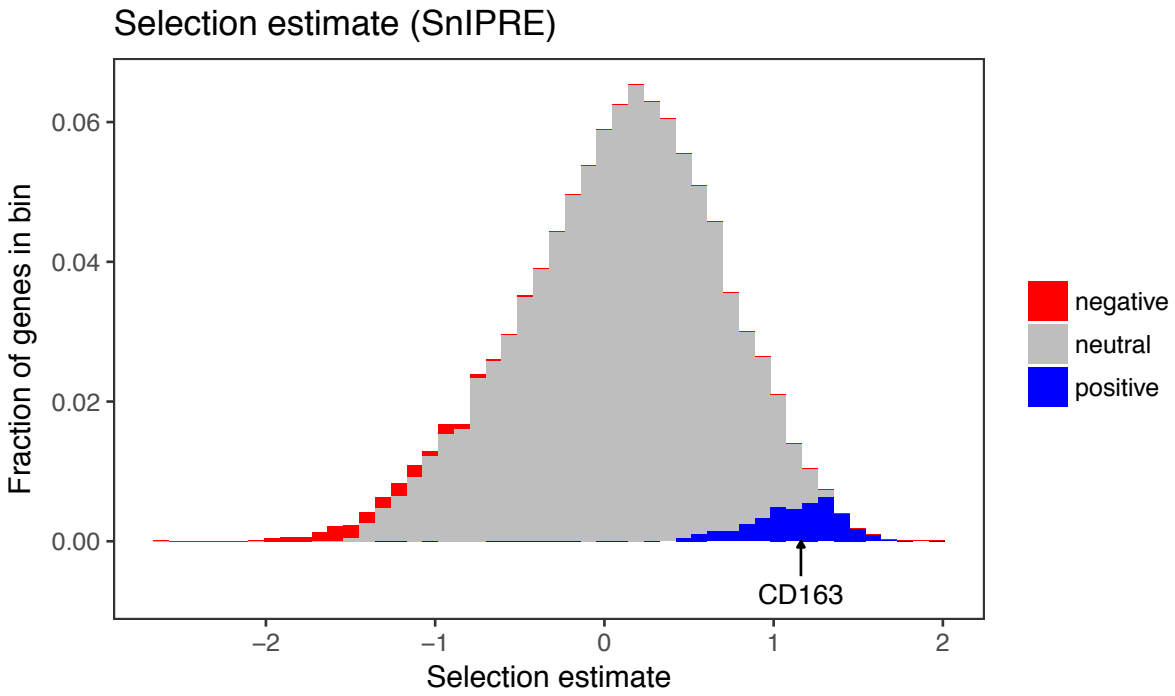


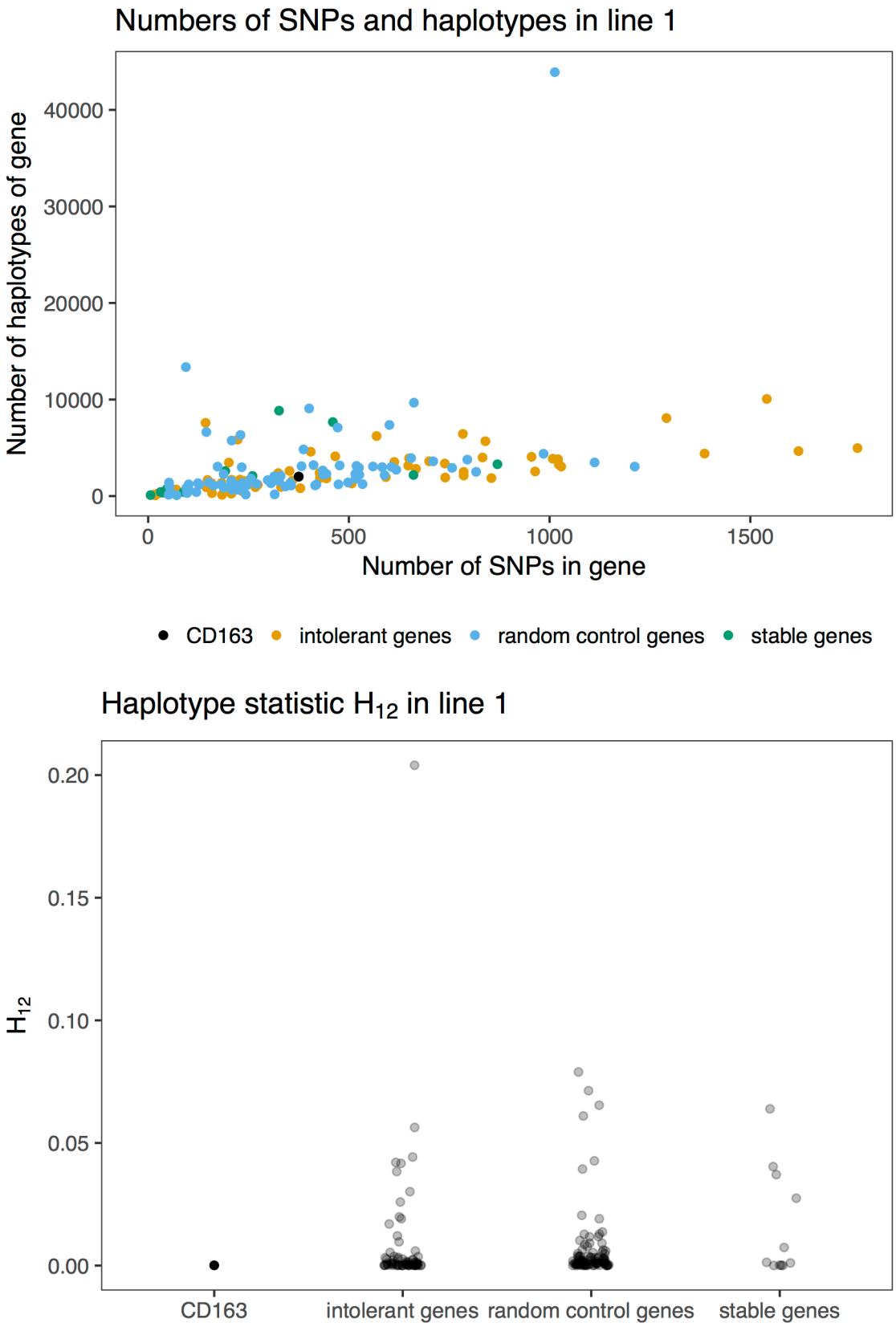

*Figure 3: The five most enriched Gene Ontology Biological Process terms in variant intolerant genes and positively selected genes, with the negative logarithm of the p-value of Fisher's exact test.*




433

434 *Figure 4: SnIPRE selection estimates. The arrow indicates the position of CD163, which is one*

435 *of the potentially positively selected (blue) genes.*




436

437 *Figure 5: Number of haplotypes, number of SNPs in gene, and selective sweep statistic $H_{12}$ of*

438 *CD163, 100 control genes of similar length, 100 intolerant genes with low residual variant*

439 *intolerance score, and 11 control genes that are homologs of human genes with stable*

440 *expression across many tissues.*

441 *Table 1: Number of pairwise shared SNPs between variant sets.*

| | *Synonymous* | | | | *Nonsynonymous* | | | |
|---|---|---|---|---|---|---|---|---|
| | *Targeted* | *Line 1* | *Line 2* | *Line 3* | *Targeted* | *Line 1* | *Line 2* | *Line 3* |
| *Targeted* | 49 | 9 | 12 | 10 | 91 | 1 | 1 | 1 |
| *Line 1* | | 10 | 10 | 11 | | 2 | 1 | 1 |
| *Line 2* | | | 13 | 10 | | | 1 | 1 |
| *Line 3* | | | | 11 | | | | 1 |

442