

Lessons learned from the initial sequencing of the pig genome: comparative analysis of an 8 Mb region of pig chromosome 17

Elizabeth A Hart*, Mario Caccamo*, Jennifer L Harrow*, Sean J Humphray*, James GR Gilbert*, Steve Trevanion*, Tim Hubbard*, Jane Rogers* and Max F Rothschild†

Addresses: *Wellcome Trust Sanger Institute, Wellcome Tust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. †Centre for Integrated Animal Genomics, Kildee Hall, Iowa State University, Ames, IA 50011, USA.

Correspondence: Elizabeth A Hart. Email: eah@sanger.ac.uk

Published: 17 August 2007

Genome Biology 2007, 8:R168 (doi:10.1186/gb-2007-8-8-r168)

The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2007/8/8/R168>

Received: 1 March 2007

Revised: 6 July 2007

Accepted: 17 August 2007

© 2007 Hart et al.; licensee BioMed Central Ltd.

This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: We describe here the sequencing, annotation and comparative analysis of an 8 Mb region of pig chromosome 17, which provides a useful test region to assess coverage and quality for the pig genome sequencing project. We report our findings comparing the annotation of draft sequence assembled at different depths of coverage.

Results: Within this region we annotated 71 loci, of which 53 are orthologous to human known coding genes. When compared to the syntenic regions in human (20q13.13-q13.33) and mouse (chromosome 2, 167.5 Mb-178.3 Mb), this region was found to be highly conserved with respect to gene order. The most notable difference between the three species is the presence of a large expansion of zinc finger coding genes and pseudogenes on mouse chromosome 2 between *Edn3* and *Phactr3* that is absent from pig and human. All of our annotation has been made publicly available in the Vertebrate Genome Annotation browser, VEGA. We assessed the impact of coverage on sequence assembly across this region and found, as expected, that increased sequence depth resulted in fewer, longer contigs. One-third of our annotated loci could not be fully re-aligned back to the low coverage version of the sequence, principally because the transcripts are fragmented over several contigs.

Conclusion: We have demonstrated the considerable advantages of sequencing at increased read depths and discuss the implications that lower coverage sequence may have on subsequent comparative and functional studies, particularly those involving complex loci such as *GNAS*.

Background

The pig (*Sus scrofa*) occupies a unique position amongst mammalian species as a model organism of biomedical importance and commercial value worldwide. A member of

the artiodactyls (cloven-hoofed mammals), it is evolutionarily distinct from the primates and rodents. At 2.7 Gb, the pig genome is similar in size to that of human and is composed of 18 autosomes, plus X and Y sex chromosomes. Extensive

conservation exists between the pig and human genome sequence, making pig an important model for the study of human health and particularly for understanding complex traits such as obesity and cardiovascular disease. Alongside other recently sequenced mammalian species of biological significance, such as cow (sequenced to 7× coverage) and dog (sequenced to 7.5× coverage), the pig will be the next mammal to have its entire genome sequenced.

The Swine Genome Sequencing Consortium [1,2] has secured first phase funding from the USDA and many other institutions to achieve draft 4× sequence depth across the genome. The sequencing, being undertaken at the Wellcome Trust Sanger Institute, utilizes a bacterial artificial chromosome (BAC) by BAC strategy through a minimal tilepath provided by the integrated, highly contiguous, physical map of the pig genome [3,4]. Additional funding has been made available for increased sequencing on chromosomes 4, 7, 14 and the sequences of these chromosomes are now available from the PreENSEMBL website [5]. To test the usefulness of our approach to sequencing the pig genome and to obtain information for a quantitative trait locus (QTL) of interest, the *S. scrofa* physical map was used to identify a tilepath of 69 overlapping BACs across an 8 Mb region of SSC17 syntenic to human chromosome 20 (20q13.13-q13.33) and mouse chromosome 2 (167.5 Mb-178.3 Mb). For this study, the BACs were sequenced to a depth of 7.5× coverage and manually finished to High Throughput Genomic sequence (HTGS) Phase 3 standard. The high quality of the sequence enabled manual annotation to be performed using the same pipeline and standards as the GENCODE project [6].

Interest in pig chromosome 17 amongst researchers in the field of animal genomics has arisen following the identification of QTL on this chromosome that affect carcass composition and meat quality [7,8]. For medical scientists, the significance of this region lies in the presence of loci such as *PCK1* and *MC3R*, which have been linked to diabetes and obesity in mammals [9,10]. Furthermore, loci in the vicinity of 20q13.2 have been found significantly amplified in a number of human breast and gastric cancers [11,12]. Manual annotation of genomic sequence remains the most reliable method of accurately defining the exon and intron boundaries of genes and identifying alternatively spliced variants. However, this process can only be performed on high quality, finished, genomic sequence. Automatic gene annotation can be performed on draft genomic sequence, but the overall outcome is dependent on a reliable assembly, which in turn relies on the overall depth of sequencing. We address the anomalies that can arise in lower quality sequence here by comparing the assembly and annotation of draft pig genomic sequence generated using three different depths of read coverage. Complex genomic regions, in particular, benefit from increased sequence depth to provide a reliable platform for meaningful annotation. On pig chromosome 17, one such region is the *GNAS* complex locus, which encodes the stimulatory G-pro-

tein α subunit, a key component of the signal transduction pathway that links interactions of receptor ligands with the activation of adenylyl cyclase. This locus is subject to a complex pattern of imprinting in human, pig and mouse, with transcripts expressed maternally, paternally and biallelically utilising alternative promoters and alternative splicing [13-17].

We compare our annotation of pig chromosome 17 with that for the syntenic regions of human chromosome 20 (20q13.13-q13.33) and mouse chromosome 2 (167.5 Mb-178.3 Mb). Both of these chromosomes have been manually annotated by the HAVANA team [18] at the Wellcome Trust Sanger Institute and the data are publicly available via the VEGA browser [19]. The identification of similarities and differences between species across syntenic regions provides a wealth of information that can relate to chromosome structure, evolution and gene function. In this instance, our annotation and comparative analysis of this region of pig chromosome 17 will be of value to researchers in the fields of agronomics, genomics and biomedical sciences.

Results and discussion

Sequence clone tilepath identification

The region reported is in two contigs of finished BACs linked by one overlapping, unfinished BAC [EMBL:CU207400]. A minimal BAC tilepath was selected by assessing shared fingerprint bands in the contact of positional information derived from BAC end sequence alignments to the human genome.

Annotation of finished BAC sequence

This 8 Mb region of pig chromosome 17 is represented by 69 BACs derived from either a CHORI-242 library or a Male Large White × Meishan F1 PigE BAC library. Within this region we identified and annotated 71 loci. Of these, we identified 53 loci that are orthologous to known human coding (CDS) genes, 7 novel transcripts, 5 putative novel transcripts and 6 processed pseudogenes. A brief description of each locus and its position within the region is summarized in Table 1 and a feature map of the overall region, including the BAC tiling path, is illustrated in Figure 1. All of these data are publicly available via the VEGA website. In Table 2, the number and type of loci within this region of pig chromosome 17 are compared to the syntenic regions of human and mouse. All three species contain very similar numbers of known coding genes but differ in the number of novel transcripts and putative loci. Specifically in mouse, the number of novel CDS and unprocessed pseudogene loci differ considerably from pig. We have divided this region of pig chromosome 17 into three sections to undertake comparisons with the syntenic regions of human chromosome 20 and mouse chromosome 2 in turn.

Table 1**List of manually annotated pig loci**

Locus name	Locus description	Start coordinate	End coordinate
PTPNI	Tyrosine phosphatase 1B	192295	261948
C17H20orf175	Orthologue of human <i>C20orf175</i>	263263	296984
<i>CH242-7P5.3</i>	Novel transcript	291165	303697
PARDB6	Par-6 partitioning defective 6 homolog beta (<i>Caenorhabditis elegans</i>)	370305	389059
BCAS4	Breast carcinoma amplified sequence 4	414609	479471
ADNP	Activity-dependent neuroprotector	492879	525950
DPMI	Dolichly-phosphate mannosyltransferase polypeptide 1, catalytic subunit	529091	552289
MOCS3	Molybdenum cofactor synthesis 3	552563	554931
KCNIG1	Potassium voltage-gated channel, subfamily G, member 1	600472	620055
<i>CH242-27718.2</i>	Novel transcript	862318	889344
<i>CH242-27718.1</i>	Putative novel transcript	891646	892957
NFATC2	Nuclear factor of activated T-cells	918501	1065177
<i>CH242-27718.4</i>	Putative novel transcript	1034571	1035756
ATP9A	Atpase, class II, type 9A	1111110	1247324
SALL4	Sal-like 4 (<i>Drosophila</i>)	1257555	1278123
<i>CH242-209L2.2</i>	Putative novel transcript	1323104	1324011
<i>CH242-209L2.1</i>	Pseudogene similar to part of human protein regulator of cytokinesis 1 (<i>PRC1</i>)	1323229	1323658
<i>CH242-511J12.1</i>	Ribosomal protein L27a (<i>RPL27A</i>) pseudogene	1496762	1497205
ZFP64	Zinc finger protein 64 homolog (mouse)	1577847	1666097
<i>CH242-300K12.1</i>	Novel transcript	1689884	1709546
TSHZ2	Teashirt family zinc finger 2	2317086	2773025
ZNF217	Zinc finger protein 217	2813463	2839285
<i>CR974566.1</i>	Thioltransferase (<i>GLRX1</i>) pseudogene	2937465	2937783
<i>CH242-271L5.2</i>	Novel transcript	3057211	3066632
<i>CH242-271L5.1</i>	Putative novel transcript	3077982	3079132
BCAS1	Breast carcinoma amplified sequence 1	3128687	3247224
CYP24A1	25-Hydroxyvitamin D3-24-hydroxylase	3318523	3339440
PFDN4	Prefoldin 4	3365055	3377364
DOK5	Docking protein 5	3600039	3751783
<i>CR956648.2</i>	Novel transcript	3756050	3764628
<i>CR956648.3</i>	Pseudogene similar to human <i>C11orf10</i>	3817744	3817975
CBLN4	Cerebellin precursor	4884433	4892814
<i>CR956393.1</i>	Ribosomal protein L27 (<i>RPL27</i>) pseudogene	4992534	4992878
MC3R	Melanocortin 3 receptor	5077795	5078875
C17H20orf108	Orthologue of human <i>C20orf108</i>	5151070	5162634
STK6	Serine/threonine kinase 6	5164100	5183118
CSTFI	Cleavage stimulation factor, 3' pre-RNA, subunit 1, 50 kda	5181641	5193333
C17H20orf32	Orthologue of human <i>C20orf32</i>	5200443	5240295
<i>CR956640.5</i>	Putative novel transcript	5224576	5225992
C17H20orf43	Orthologue of human <i>C20orf43</i>	5250245	5294040
C17H20orf105	Orthologue of human <i>C20orf105</i>	5271092	5277853
C17H20orf106	Orthologue of human <i>C20orf106</i>	5296497	5298326
TFAP2C	Transcription factor AP-2 gamma (activating enhancer binding protein 2 gamma)	5374025	5384555
<i>CH242-255C19.2</i>	Novel transcript	5409258	5410949
<i>CH242-266P8.1</i>	Ribosomal protein L27 (<i>RPL27</i>) pseudogene	5690286	5690695
BMP7	Bone morphogenetic protein 7 (osteogenic protein 1)	5794879	5886410
SPO11	SPO11 meiotic protein covalently bound to DSB-like (<i>Saccharomyces cerevisiae</i>)	5940695	5955823
RAE1	RAE1 RNA export 1 homolog (<i>Schizosaccharomyces pombe</i>)	5961819	5977901

Table 1 (Continued)**List of manually annotated pig loci**

RNPC1	RNA-binding region (RNPI, RRM) containing I	5993104	6007945
CTCFL	CCCTC-binding factor (zinc finger protein-like)	6070196	6102129
<i>CH242-37G9.1</i>	Novel transcript	6114046	6116136
PCK1	Phosphoenolpyruvate carboxykinase I (soluble)	6140516	6146484
ZBPI	Z-DNA binding protein I	6182270	6192447
TMEPAI	Transmembrane, prostate androgen induced RNA	6205610	6260081
C17H20orf85	Orthologue of human <i>C20orf85</i>	6580341	6590326
C17H20orf86	Orthologue of human <i>C20orf86</i>	6632471	6641977
PPP4R1L	Protein phosphatase 4, regulatory subunit I-like	6644116	6665957
RAB22A	RAB22A, member RAS oncogene family	6721985	6779521
VAPB	VAMP (vesicle-associated membrane protein)-associated protein B and C	6800573	6850820
STX16	Syntaxin 16	6911191	6939474
NPEPL1	Aminopeptidase-like 1	6946539	6962106
GNAS	GNAS complex locus	7056486	7123907
THIL	Th1-like (<i>Drosophila</i>)	7199960	7212384
CTSZ	Cathepsin Z	7212382	7220265
TUBB1	Tubulin, beta family 1	7232312	7239278
ATP5E	ATP synthase, H ⁺ transporting, mitochondrial F1 complex, epsilon subunit	7241534	7245591
C17H20orf45	Orthologue of human <i>C20orf45</i>	7245938	7255973
C17H20orf174	Orthologue of human <i>C20orf174</i>	7384250	7452090
EDN3	Endothelin 3	7496706	7519565
PHACTR3	Phosphatase and actin regulator 3	7697213	7882258
SYCP2	Synaptonemal complex protein 2	7895310	7972154

The locus name, description and relative co-ordinates within the 8 Mb region are given. Locus names denoted in bold indicate that the locus is orthologous to a known human locus.

Comparative analysis: PTPN1 to CYP24A1

This region is well conserved between human, pig and mouse with respect to gene order. In human, this region (20q13.2) is of considerable interest because it is susceptible to amplification in a number of cancer lines, as shown by comparative genomic hybridization experiments [11,12,20]. In particular, *PTPN1*, *BCAS4*, *ZNF217* and *CYP24A1* have been found at increased copy numbers in human breast, ovarian, pancreatic and gastric cancer cell lines [12,21-23]. One noticeable difference between pig, mouse and human is the apparent absence of a *BCAS4* counterpart in mouse. *BCAS4* encodes a 203

Table 2**Comparison of loci type and number in pig, human and mouse**

Locus type	Pig	Human	Mouse
Known coding	53	54	52
Novel CDS	-	-	51
Novel transcript	7	15	22
Putative	5	24	12
Processed pseudogene	6	20	22
Unprocessed pseudogene	-	1	31
Expressed pseudogene	-	1	1
Total	71	115	191

amino acid protein of unknown function that shares homology with the cappuccino(*CNO*) locus in human, mouse and other mammalian species. We performed a BLASTP analysis to investigate whether a putative orthologue of *BCAS4* could be found elsewhere in the mouse genome, using the predicted pig and human *Bcas4* protein sequences to search ENSEMBL mouse (NCBI m36 assembly). However, the only homologous locus we identified in mouse was the *CNO* locus on chromosome 5. The relationship between *Bcas4* and *Cno* homologues can be visualized using TREEFAM [24] [TREEFAM:TF326629]. In human, additional alternative splice variants of *BCAS4* have been identified, with one potentially encoding a longer polypeptide of 211 amino acids. In human and mouse, five and seven novel transcripts or putative loci, respectively, lie between the *ZFP64* and *TSHZ2* loci. None of these appear to be conserved between the three species, and in pig only one novel transcript locus, *CH242-300K12.1*, was identified between *ZFP64* and *TSHZ2*.

Comparative analysis: PFDN4 to VAPB

Comparison of this region in pig, human and mouse reveals that it is highly conserved with respect to gene order and orientation. One notable difference between the three species in this region is the absence of porcine and murine counterparts of the human *C20orf107* locus. In human, the *C20orf107* locus lies immediately downstream of the *C20orf106* locus.

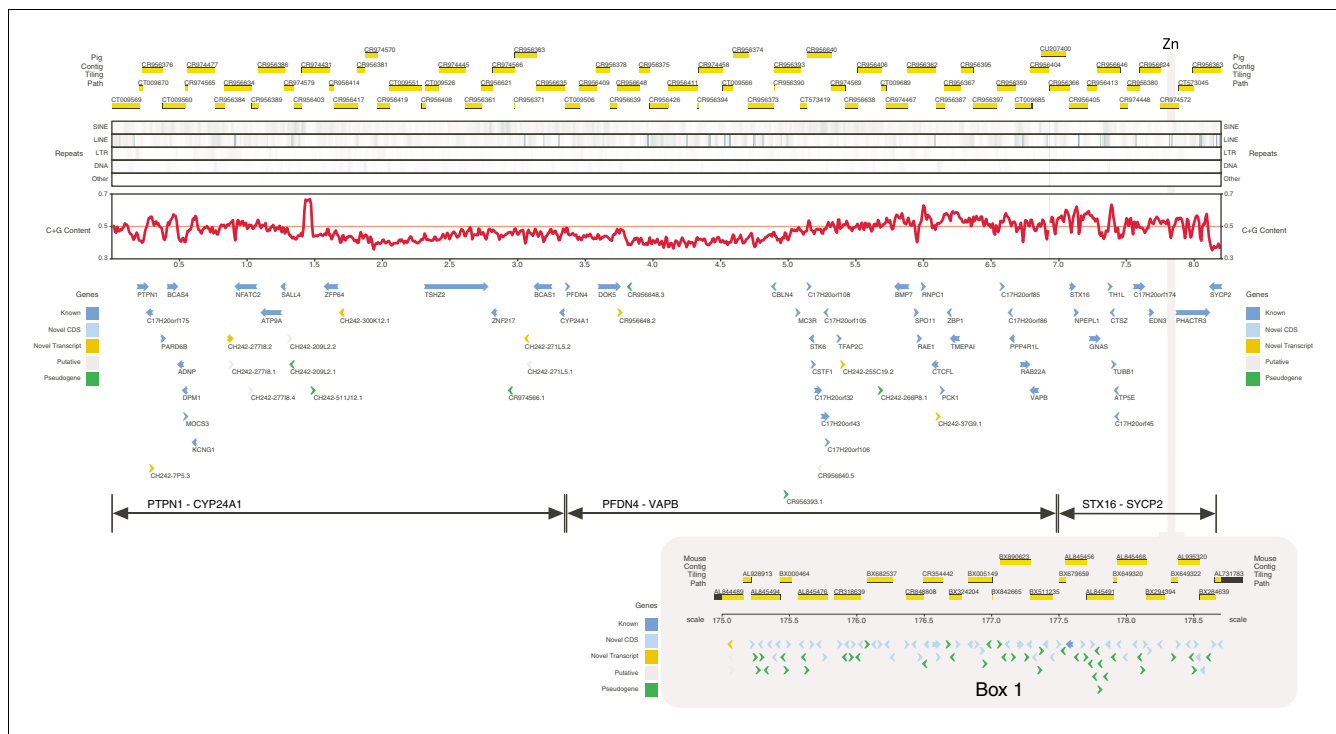


Figure 1
 Feature map of the 8 Mb region of pig chromosome 17. Each locus is depicted according to type, orientation and position. The tiling path of the sequenced BACs is shown along the top. Below this, the distribution of repeats and C + G content is shown. Box 1 illustrates the zinc-finger locus expansion that has occurred in mouse between *EDN3* and *PHACTR3*. The three regions described in the comparative analyses, *PTPN1-CYP24A1*, *PFDN4-VAPB* and *STX16-SYCP2*, are defined using double-headed arrows.

Both loci encode proteins of 171 amino acids and share 87% amino acid identity and 92% similarity. The function of these two proteins in human is unknown, although INTERPRO analysis predicts two transmembrane helices within these putative paralogues. The pig homologue of *C20orf106* encodes a protein of 170 amino acids that shares 63% identity and 78% similarity with both human *C20orf106* and *C20orf107* proteins and contains these two putative transmembrane helices. To further investigate the presence of *C20orf106* and *C20orf107* orthologues in other species, we compared this region across multiple organisms using ENSEMBL AlignSliceView [25]. Interestingly, it appears that the presence of both *C20orf106* and *C20orf107* loci is specific to primates: human, chimp and macaque all contain both *C20orf106* and *C20orf107* as neighboring loci whereas ENSEMBL non-primate species - for example, cow, rat and dog - appear to have only one or other of the two paralogues in the syntenic location. In the absence of additional species and a more detailed analysis it is not possible to draw definite conclusions regarding the evolutionary distribution of *C20orf106* and *C20orf107*. However, these observations suggest that the absence of *C20orf107* from this region in pig and mouse is not specific to these species.

Comparative analysis: *STX16* to *SYCP2*

The most striking difference between pig, human and mouse within this sub-region is the presence of a large cluster of zinc finger loci in mouse, between *Edn3* and *Phactr3*, that is completely absent from pig and human. This mouse-specific expansion is over 3.2 Mb in length and contains one known coding gene, 51 genes with a novel CDS and 30 unprocessed pseudogenes, all predicted to contain C2H2 Zinc finger type and KRAB box domains. These motifs have been found to confer DNA binding ability and behave as transcriptional repressor domains in a number of proteins [26]. Given that the full extent of duplication within this region of the mouse genome is still being resolved, there is potential for the total number of loci to be even greater.

In contrast to the significant differences between pig and mouse between the *EDN3* and *PHACTR3* loci, the rest of this sub-region remains highly conserved across the three species, including the *GNAS* locus, one of the most complex loci to be found in mammalian genomes. A comparison of the *GNAS* transcripts annotated in pig, human and mouse can be viewed directly in VEGA using Pig MultiContigView [27], as is shown in Figure 2. To generate this simultaneous view of *GNAS* transcripts in all three species, pig *GNAS* should be viewed in VEGA ContigView. 'Homo_sapiens chromosome 20' should then be chosen from the 'View alongside' menu and

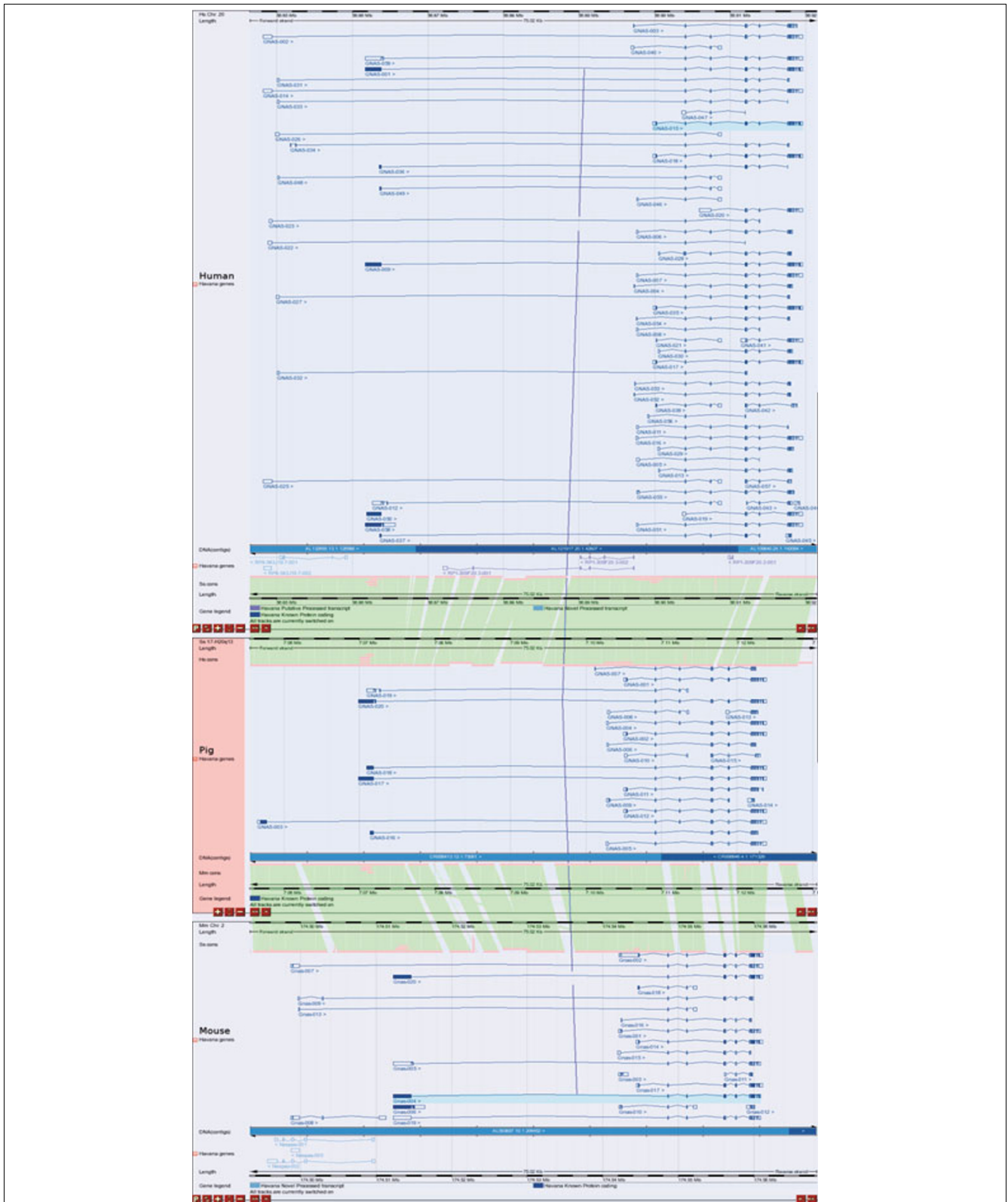


Figure 2
 Comparison of *GNAS* transcripts in human, pig and mouse. A screenshot taken from VEGA Pig MultiContigView, comparing *GNAS* transcripts annotated in human (top panel), pig (middle panel) and mouse (bottom panel). The vertical blue lines joining loci in VEGA MultiContigView represent orthologous relationships between loci across species.

'Mus_musculus:2' added from the 'Comparative' drop-down menu. *GNAS* has been well studied in human and mouse and encodes four proteins - *Gsa*, *Nesp Xla α* and *Alex* - that have been well-characterized in both species. Of these *GNAS* products, the most well-conserved are the alternatively spliced variants of *Gsa*, the alpha-stimulatory subunit of GTP-binding protein, which is biallelically expressed in human and mouse. The best known of these *Gsa* isoforms is 394 amino acids long in all three species. In pig and human these *Gsa* proteins are 100% identical with respect to primary structure, while the mouse orthologue differs by the substitution of just one amino acid. Paternally expressed, the large variant of G-protein α subunit known as *Xla α* utilizes a large, upstream first exon compared to the *Gsa* variants [14,28]. The pig *Xla α* homologue is predicted to be 1,005 amino acids long and shares 78% identity and 82% similarity with the human and 65% identity and 70% similarity with the mouse *Xla α* proteins, which are 1,037 and 1,133 amino acids long, respectively. The capacity to encode the most unusual of the *GNAS* products, *Alex*, is also conserved in pig. *Alex* is translated in a different reading frame to *Xla α* and has been described in rat and human [16,29]. The pig *Alex* protein is predicted to be 564 amino acids long while the human and mouse *Alex* proteins are 625 amino acids and 725 amino acids long, respectively. This difference in length is partly due to divergence within a proline-rich and leucine-rich stretch of amino acids that lie between residues 298 and 398 in porcine *Alex*. Alignment of these predicted pig, mouse and human *Alex* proteins reveals they are less conserved than the other *GNAS*-encoded proteins: pig *Alex* protein shares approximately 61% identity and 70% similarity with human *Alex* protein and 44% identity and 51% similarity with mouse *Alex* protein. Finally, expressed exclusively from maternal alleles in human and mouse, the *NESP55* transcript encodes neuroendocrine secretory protein 55. Pig *Nesp55* shares 82% identity and 89% similarity with human *Nesp55* (68% identity and 80% similarity with mouse *Nesp55*). At the mouse and human *GNAS* loci, maternally imprinted *NESP55* antisense transcripts have been identified [30-32], unofficially known as *Nespas* and *SANG*, respectively. However, we have been unable to identify a pig *GNAS* antisense transcript. Pig has diverged sufficiently from human and mouse such that the exons of these antisense transcripts are not conserved. In human, *GNAS* appears to be the only locus that is imprinted within this region, 20q13.32: the two genes, *TH1* and *CTSZ*, which lie downstream of *GNAS*, have been found to be biallelically expressed [33].

Comparison of draft sequence assemblies

The manual annotation produced in this project is not only useful for comparative analyses but also can be used as a reference set to judge the influence of sequence coverage on gene annotation. For the purpose of this study, our 8 Mb region of pig chromosome 17 was sequenced to a depth of 7.5 \times coverage and manually finished to GenBank HTGS Phase 3 standard to produce sequence with a predicted error rate of less than 1 in

100,000 bases. However, the international pig genome sequencing project currently has funding to generate in the first phase of sequencing only draft sequence at 3-4 \times coverage overall (with the exception of chromosomes 4, 7 and 14, which will be sequenced to an improved draft using sequence targeted to close gaps). To assess the impact of sequencing coverage on contig size and gene integrity, we automatically assembled sequence reads obtained from 384-well plates of shotgun sequencing to represent differing amounts of coverage across the region: 2.5 \times , 5 \times and 7.5 \times (see Materials and methods for details).

We chose to count only contigs greater than 2 kb in our analysis, thus excluding short bacterial contaminants and single pass reads. When we assembled reads at a depth of 2.5 \times coverage, the mean number of contigs obtained per clone was 27 and the average total contig length was 138 kb. If coverage is increased to 5 \times , the mean number of contigs obtained per clone decreases to 13 and the average total contig length increases to 179 kb. When we increased the level of coverage further to 7.5 \times the mean number of contigs obtained per clone is reduced to 5 and the average total contig length achieved is 184 kb. Therefore, increasing the read coverage for each BAC clone results in fewer, longer contigs per clone. These results are illustrated in Figure 3, where the difference in contig number obtained after automatic assembly of reads at a level of either 5 \times and 7.5 \times coverage is represented using dot-plots for two different BAC clones: CH242-247L10 [EMBL:CR956646] and CH242-155M9 [EMBL:CR956640]. CH242-247L10 contains the 3' end of the *GNAS* complex locus and the downstream *TH1L*, *CTSZ*, *TUBB*, *ATP5E*, *C17H20orf45* loci. At a level of 5 \times coverage, CH242-247L10 is assembled into 10 contigs longer than 2 kb, with the 50 kb region containing the 3' end of *GNAS* and its immediate downstream region (defined by a black rectangle) dispersed over 4 contigs. However, increasing the level of coverage to 7.5 \times reduced the total number of contigs longer than 2 kb to three, such that the *GNAS* downstream region is now contained within a single contig. A manual finishing step is still required to link these 3 contigs, but the assembly is much improved in comparison. In Figure 3b, CH242-155M9 contains the pig *C20orf106* gene. As mentioned previously, pig lacks the paralogous locus, *C20orf107*, which lies immediately downstream of *C20orf106* in human. At a depth of 5 \times coverage, CH242-155M9 is assembled into six contigs longer than 2 kb, with the region immediately downstream of *C20orf106* (defined by a black rectangle) divided between three of these contigs. Using this assembly, it may not be easily ascertained whether the *C20orf107* gene is absent in pig or falls within a gap in the assembly. Increasing the coverage to 7.5 \times decreases the total number of contigs to three (again, a manual finishing step would be required to link these three contigs) and we can be more confident that the *C20orf107* locus is absent in pig and does not simply fall within a gap in the assembly.

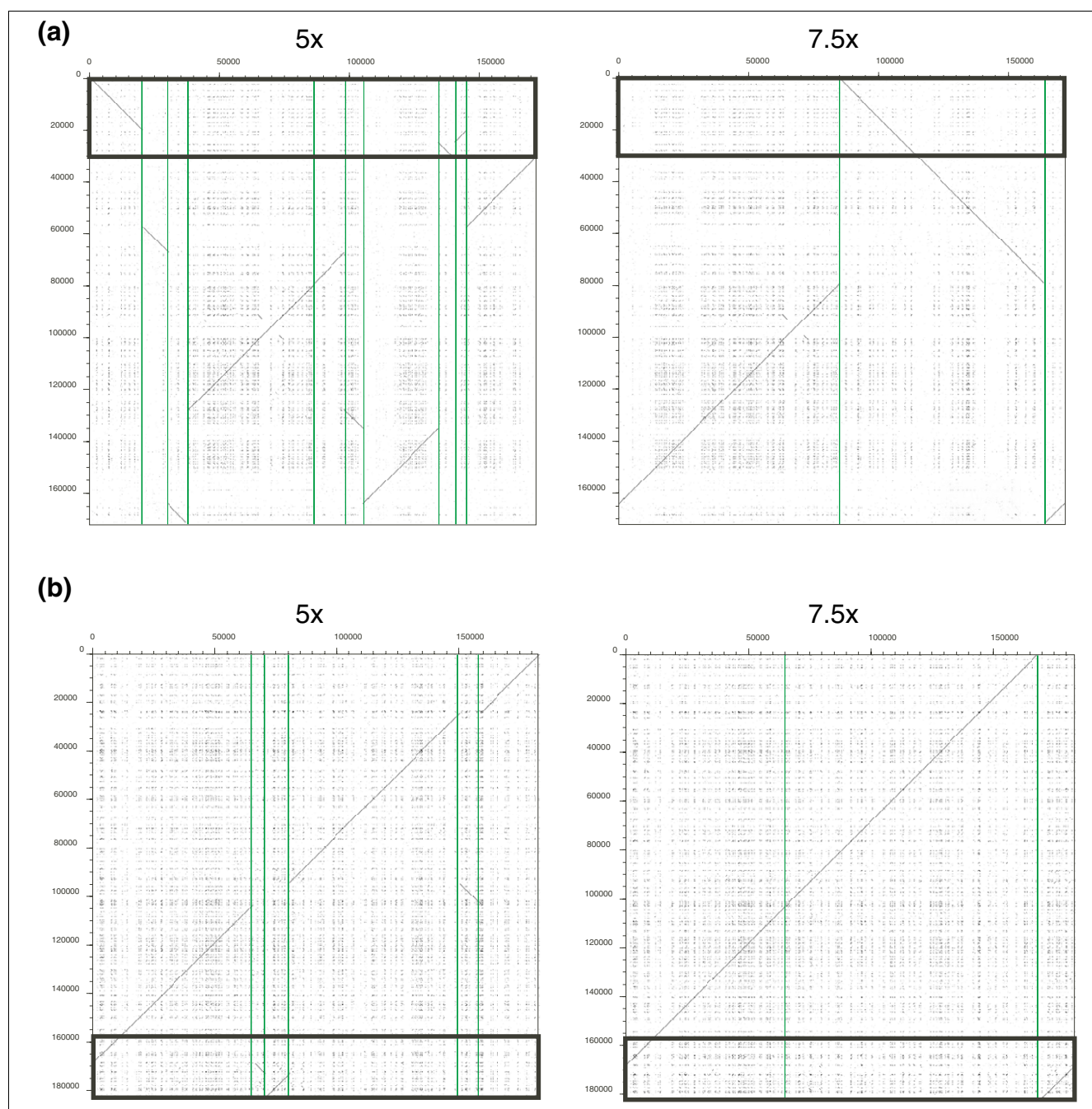


Figure 3
Comparison of 5 \times and 7.5 \times coverage assemblies. Dot-plots of finished BAC sequence against either 5 \times or 7.5 \times assembled sequence for BACS (a) CH242-247L10 and (b) CH242-155M9. Individual contigs, represented on the x-axis, are separated by vertical green lines. In (a) the black rectangle depicted on the graphs represents the *GNAS* downstream region. In (b) the black rectangle depicted on the graphs defines the vicinity of the pig *C20orf106* locus.

Using EXONERATE [34] in conjunction with a splice-aware model, we investigated whether our manual annotation performed on the finished BACs could be aligned back to the 2.5 \times , 5 \times and 7.5 \times assemblies. In total, 71 loci were annotated within the finished BACs. For each of these genes we selected the longest transcript and discarded any that spanned multiple finished clones, leaving us with 58 transcripts, which we

attempted to align back to the 2.5 \times , 5 \times and 7.5 \times assemblies. We counted only transcripts that could be fully re-aligned along their entire length. From the pool of 58 annotated transcripts we were able to fully re-align 54 to our 7.5 \times assembly, 39 to our 5 \times assembly and just 10 to our 2.5 \times assembly. This means that 33% of our annotated transcripts could not be fully re-aligned to the 5 \times assembly. Where re-alignment was

unsuccessful, the most common reason was that the transcript spanned multiple contigs. In other instances, however, re-alignment failure was linked to mis-assemblies and low quality regions.

These results indicate that the impact of low-coverage sequencing on the structure of the assembly is considerable. Reducing the number of sequence reads from a depth of 7.5× to 5× and 2.5× increases the number of contigs within the assembly, decreases the total length of contigs and is likely to introduce errors in sequence organization due to the presence of gaps in sequence coverage. As a result, annotation of gene loci will be less precise and large genes are likely to be incomplete or artificially re-arranged.

Conclusion

The generation and manual annotation of this 8 Mb region of pig chromosome 17 will provide a useful resource for researchers in the field of pig genomics, as well as scientists with a more general interest in mammalian comparative genomics. Importantly, we have also shown that increasing the sequence depth across this region of the pig genome has several material advantages with respect to coverage and quality.

We have identified 71 loci that lie between *PTPN1* at the centromeric end of pig chromosome 17 and *SYPC2* at the telomeric end. Comparison of this region with the 9.38 Mb and 10.8 Mb syntenic regions of human chromosome 20 and mouse chromosome 2, respectively, has revealed both striking similarities and differences between the three species. The most significant difference between pig, human and mouse is the presence of a 3.2 Mb expansion of zinc finger loci in mouse, absent in human and pig, which has occurred between *Edn3* and *Phactr3* and could represent an event of evolutionary significance in the mouse lineage. Additional differences between the three species include the existence of *C20orf107* in human that is absent from pig and mouse and the absence of the *BCAS4* locus from mouse that is conserved in human and pig. We detected 12 transcribed non-coding loci specific to pig that may warrant further investigation. Eight of these lay between *PTPN1* and *CYP24A1*, a region of interest subject to amplification in human cancer cell lines and associated with complex traits such as type 2 diabetes [35,36]. Furthermore, our annotation of the porcine orthologue of *GNAS* will contribute towards the characterization of this enigmatic complex locus. The predicted primary structures of the four putative pig *GNAS* products - *Gsa*, *Xlas*, *Alex* and *Nesp* - are comparable to their counterparts in human and mouse. Interestingly, imprinted regions on other pig chromosomes have been linked to a range of QTLs [37], which suggests the region encompassing the pig *GNAS* locus is worthy of further analysis.

In addition to providing locus information within the confines of the sequence, we have used this test region of pig chromosome 17 to demonstrate the value of genome sequencing at increased levels of coverage. The advent of large-scale sequencing projects in the last two decades has been accompanied by the formulation of mathematical models to quantitatively determine the strategic design of such projects. The models proposed by Lander and Waterman [38], which extended the earlier theories of Clarke and Carbon [39], have provided theoretical guidelines for standard fingerprint mapping and shotgun sequencing projects and have been developed by others [40,41] as the nature and scale of sequencing projects has evolved. These algorithms continue to be relevant, particularly to assess the design, quality and value of new sequencing technologies and their applications to projects such as re-sequencing [42,43], which themselves will bring new challenges to the field. In this study, we have not set out to perform a detailed quantitative investigation into the effect of sequence depth on sequence assembly. However, we have taken advantage of this test region of the pig genome to illustrate the impact of read coverage on the structure and contiguity of the pig genome assembly and, importantly, annotation. We have shown that increasing sequence coverage from 5× (which is above the overall target depth of the pig genome) to 7.5× greatly improves the assembly of sequence reads into contigs. Specifically, it results in fewer and longer contigs, which improves the reliability of the genome assembly overall. A high degree of confidence in the fidelity of the genome assembly is advantageous in complex regions - for example, *GNAS* - that may contain non-coding regulatory sequences. It is preferable that such regions are kept as intact as possible, but our analysis showed the region just downstream of the *GNAS* locus to be fragmented over four contigs using the 5× assembly. Assembly errors that occur in intergenic regions may not be immediately obvious, but can have implications for subsequent analyses of non-coding regions. Using the *C20orf106/C20orf107* loci in human as a second example, we showed that 5× coverage is insufficient to determine with confidence whether a pig orthologue of *C20orf107* is absent from the pig lineage or simply falls within a gap in our assembly. Clearly, it is important to eliminate doubts such as these for meaningful comparative analyses. Genome annotation, whether automated or manual, is highly dependent on the integrity of the genome assembly. While reduction of errors at the base level is pertinent to improving the quality of shotgun sequence [44], our pilot study has focused on the impact of sequence structure on the quality of the final product. In particular, we assessed the effect of read coverage on genome annotation. We found that we were unable to fully re-align one-third of our annotated transcripts back to the 5× assembly, indicating that multiple contigs, gaps and assembly errors caused by low coverage sequencing significantly affect the quality of genome annotation. The value of a genome is dependent on the quality of its annotation, which makes sequencing coverage an important consideration in project design. There is no doubt that the 3-4× sequencing of the pig

genome will provide researchers with another extremely valuable layer of information for mammalian comparative studies. However, the additional advantages that could be gained by additional investment should not be underestimated. Improving the level of sequencing coverage will undoubtedly provide a better platform for automated annotation and downstream analyses. Given the importance of pig as an agricultural species and a biomedical model, greater advances in many aspects of porcine and mammalian science might be made if further funding was made available to improve the overall coverage of the entire pig genome.

Materials and methods

Mapping and sequencing

A physical map of the porcine genome was constructed using the fingerprints and end sequences generated from over 264,000 BACs from 4 BAC libraries and ordering information derived from pig radiation hybrid markers and sequence homology to the human genome. The current assembly contains just 172 contigs and covers >98% of the genome.

Sequence clones were sub-cloned into 4-6 kb inserts in pUC 19 and sequenced to up to 8-fold depth with Applied Biosystems (Foster City, CA, USA) Big Dye v3 chemistry. Sequence reads were assembled using PHRAP. Assembled clones were improved by one round of primer walking to extend sequence contigs and close gaps before the clones were examined and final gap closure and checking procedures were carried out. The integrity of the finished clones was assessed by reference to three restriction enzyme digests compared to virtual digestions performed on the sequence assembly before sequence accessions were declared finished and entered into EMBL/GenBank HTGS Phase 3.

Sequence annotation

Manual annotation was performed on the pig genomic sequence by the Wellcome Trust Sanger Institute Havana team as follows: The finished porcine sequence was analyzed using an automatic ENSEMBL pipeline [45] with modifications to aid the manual curation process. The G + C content of each clone sequence was analyzed and putative CpG islands were marked. Interspersed repeats were detected using RepeatMasker using the mammalian library along with porcine-specific repeats submitted to EMBL/NCBI/DDBJ and simple repeats using Tandem Repeats Finder [46]. The combination of the two repeat types was used to mask the sequence. The masked sequence was searched against vertebrate cDNAs and expressed sequence tags (ESTs) using WU-BLASTN and matches were cleaned up using EST2_GENOME. A protein database combining non-redundant data from SwissProt and TrEMBL was searched using WU-BLASTX. *Ab initio* gene structures were predicted using FGENESH and GENSCAN. Predicted gene structures were manually annotated according to GENCODE standards [6]. The gene categories are described on the VEGA website [19]:

'Known' genes are identical to known pig cDNAs or are orthologous to known human loci; 'Novel CDS' loci have an open reading frame (ORF), are identical to spliced ESTs or have some similarity to other genes and proteins; 'Novel transcript' is similar to novel CDS but no ORF can be determined unambiguously; 'Putative' genes are identical to spliced pig ESTs but do not contain an ORF; and 'Pseudogenes' are non-functional copies of known or novel loci.

Comparison of draft sequence assemblies

We calculated the three depths of coverage (2.5×, 5× and 7.5×) that were compared across this particular region as follows. We predicted an insert size of 173 kb for our BAC libraries and the average read length achieved during sequencing was 713 base pairs. Therefore, for this region, approximately 240 sequencing reads represent a depth of 1× coverage. For each BAC, approximately 600 passed reads were obtained from a 384-well plate after quality checking. Thus, one plate of 600 passed reads represents approximately 2.5× coverage for that clone; two plates constitute around 1,200 passed reads and is equivalent to up to 5× coverage; three plates constitute approximately 1,800 passed reads and is equivalent to up to 7.5× coverage. Using PHRAP, we automatically re-assembled 62 BAC clones from the 8 Mb region using one, two and three plates of passed reads to obtain the 2.5×, 5× and 7.5× assemblies, respectively. Assembled contigs that were shorter than 2 kb were discarded. The resulting assemblies for each clone were compared to each other directly with respect to contig number and length. Our manually annotated loci were re-aligned to each of the three assemblies using EXONERATE [34] in conjunction with a splice-aware model to avoid spurious hits. For each annotated locus we selected the longest transcript (where alternative variants had been annotated) but discarded transcripts that spanned multiple finished clones. Thus, we attempted to re-align a total of 58 transcripts to our 2.5×, 5× and 7.5× assemblies. Only transcripts that could be re-aligned entirely back to the assembly across their full length were counted as being successfully re-aligned. All of the sequence traces from this project have been deposited in the trace repository and are available from the ENSEMBL trace server [47].

Abbreviations

BAC, bacterial artificial chromosome; EST, expressed sequence tag; QTL = quantitative trait locus.

Authors' contributions

Manual annotation of finished pig BACs and subsequent comparative analysis was undertaken by EA Hart. The comparison of draft pig sequence assemblies was performed by M Caccamo.

Acknowledgements

The authors thank M Ramos, J Reecy and ZL Hu from Iowa State University for their assistance on this project. From the Sanger Institute, we wish to thank the Anacode team, the HAVANA team, Richard Clark, Paul Hunt, Carol Scott and the DNA sequencing division. This work was funded by the National Pork Board, Iowa Pork Producers Association, the Iowa Agriculture and Home Economics Experiment Station, State of Iowa and Hatch funding and the Wellcome Trust.

References

- Swine Genome Sequencing Consortium [http://www.piggenome.org/index.php]
- Wellcome Trust Sanger Institute Swine Genome Sequencing and Mapping [http://www.sanger.ac.uk/Projects/S_scrofa/]
- PreENSEMBL Pig Clone Map [http://pre.ensembl.org/Sus_scrofa_map/index.html]
- Humphray SJ, Scott C, Clark R, Marron B, Bender C, Camm N, Davis J, Jenks A, Noon A, Patel M, et al.: **A high utility integrated map of the pig genome.** *Genome Biol* 2007, **8**:R139.
- PreENSEMBL Pig Genome Sequence [http://pre.ensembl.org/Sus_scrofa/index.html]
- Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, Lagarde J, Gilbert JG, Storey R, Swarbreck D, et al.: **GENCODE: producing a reference annotation for ENCODE.** *Genome Biol* 2006, **7**(Suppl 1):S4.
- Malek M, Dekkers JCM, Lee HK, Baas TJ, Rothschild MF: **A molecular genome scan analysis to identify chromosomal regions influencing economic traits in the pig. I. Growth and body composition.** *Mamm Genome* 2001, **12**:630-636.
- Ramos AM, Helm J, Sherwood J, Rocha D, Rothschild MF: **Mapping of 21 genetic markers to a QTL region for meat quality on pig chromosome 17.** *Anim Genet* 2006, **37**:296-297.
- Beale EG, Hammer RE, Antoine B, Forest C: **Disregulated glyceroneogenesis: PCK1 as a candidate diabetes and obesity gene.** *Trends Endocrinol Metab* 2004, **15**:129-135.
- Butler AA: **The melanocortin system and energy balance.** *Pep-tides* 2006, **27**:281-290.
- Kallioniemi A, Kallioniemi O-P, Piper J, Tanner M, Stokke T, Chen L, Smith HS, Pinkel D, Gray JW, Waldmann : **Detection and mapping of amplified DNA sequences in breast cancer by comparative genomic hybridisation.** *Proc Natl Acad Sci USA* 1994, **91**:2156-2160.
- Yang SH, Seo MY, Jeong HJ, Jeung H-C, Shin J, Kim SC, Noh SH, Chung HC, Rha SY: **Gene copy number change events at chromosome 20 and their association with recurrence in gastric cancer patients.** *Clin Cancer Res* 2005, **11**:612-620.
- Kozasa T, Itoh H, Tsukamoto T, Kaziro Y: **Isolation and characterization of the human G(s) α gene.** *Proc Natl Acad Sci USA* 1988, **85**:2081-2085.
- Kehlenbach RH, Matthey J, Huttner WB: **XL alpha s is a new type of G protein.** *Nature* 1994, **372**:804-809.
- Ischia R, Lovisetti-Scamihorn P, Hogue-Angeletti R, Wolkersdorfer M, Winkler H, Fischer-Colbrie R: **Molecular cloning and characterization of NESP55, a novel chromogranin-like precursor of a peptide with 5-HT1B receptor antagonist activity.** *J Biol Chem* 1997, **272**:11657-11662.
- Klemke M, Kehlenbach RH, Huttner WB: **Two overlapping reading frames in a single exon encode interacting proteins - a novel way of gene usage.** *EMBO J* 2001, **20**:3849-3860.
- Thomsen H, Lee HK, Rothschild MF, Malek M, Dekkers JCM: **Characterization of quantitative trait loci for growth and meat quality in a cross between commercial breeds of swine.** *J Anim Sci* 2004, **82**:2213-2228.
- Wellcome Trust Sanger Institute HAVANA Team [http://www.sanger.ac.uk/HGP/havana/]
- Vertebrate Genome Annotation Browser [http://vega.sanger.ac.uk/]
- Tanner MM, Tirkkonen M, Kallioniemi A, Collins C, Stokke T, Karhu R, Kowbel D, Shadravan F, Hintz M, Kuo WL, et al.: **Increased copy number at 20q13 in breast cancer: defining the critical region and exclusion of candidate genes.** *Cancer Res* 1994, **54**:4257-4260.
- Bärlund M, Monni O, Weaver JD, Kauraniemi P, Sauter G, Heiskanen M, Kallioniemi O-P, Kallioniemi A: **Cloning of BCAS3 (17q23) and BCAS4 (20q13) genes that undergo amplification, overexpression, and fusion in breast cancer.** *Genes Chromosomes Cancer* 2002, **35**:311-317.
- Mahlamäki EH, Bärlund M, Tanner M, Gorunova L, Hoglund M, Karhu R, Kallioniemi A: **Frequent amplification of 8q24, 11q, 17q, and 20q-specific genes in pancreatic cancer.** *Genes Chromosomes Cancer* 2002, **35**:353-358.
- Collins C, Rommens JM, Kowbel D, Godfrey T, Tanner M, Hwang S, Polikoff D, Nonet G, Cochran J, Myambo K, et al.: **Positional cloning of ZNF217 and NABC1: genes amplified at 20q13.2 and overexpressed in breast carcinoma.** *Proc Natl Acad Sci USA* 1998, **95**:8703-8708.
- Tree Families (Treefam) Database [http://www.treefam.org]
- ENSEMBL AlignSliceView for the C20orf106 Locus [http://www.ensembl.org/Homo_sapiens/alignsliceview?c=20:54535127.5;w=30000;align=opt_align_259]
- Urrutia R: **KRAB-containing zinc-finger repressor proteins.** *Genome Biol* 2003, **4**:231.
- VEGA MultiContigView for the GNAS Locus [http://vega.sanger.ac.uk/Sus_scrofa/multicontigview?sl=hs;w=69422;c=17-H20q13:3A7265015.5%3A1;h=w;wl=69422;c1=20%3A56882955.5%3A1;flip=Mus_musculus:2]
- Abramowitz J, Grenet D, Birnbaumer M, Torres HN, Birnbaumer L: **XL α s, the extra-long form of the α -subunit of the Gs G protein, is significantly longer than suspected, and so is its companion Alex.** *Proc Natl Acad Sci USA* 2004, **101**:8366-8371.
- Freson K, Jaeken J, Van Helvoirt M, de Zegher F, Wittevrongel C, Thys C, Hoylaerts MF, Vermeylen J, Van Geet C: **Functional polymorphisms in the paternally expressed XL α s and its cofactor ALEX decrease their mutual interaction and enhance receptor-mediated cAMP formation.** *Hum Mol Genet* 2003, **12**:1121-1130.
- Wroe SF, Kelsey G, Skinner JA, Bodle D, Ball ST, Beechey CV, Peters J, Williamson CM: **An imprinted transcript, antisense to Nesp, adds complexity to the cluster of imprinted genes at the mouse Gnas locus.** *Proc Natl Acad Sci USA* 2000, **97**:3342-3346.
- Williamson CM, Skinner JA, Kelsey G, Peters J: **Alternative non-coding splice variants of Nespas, an imprinted gene antisense to Nesp in the Gnas imprinting cluster.** *Mamm Genome* 2002, **13**:74-79.
- Hayward BE, Bonthron DT: **An imprinted antisense transcript at the human GNAS1 locus.** *Hum Mol Genet* 2000, **9**:835-841.
- Bonthron DT, Hayward BE, Moran V, Strain L: **Characterization of THI and CTSZ, two non-imprinted genes downstream of GNAS1 in chromosome 20q13.** *Hum Genet* 2000, **107**:165-175.
- Slater G, Birney E: **Automated generation of heuristics for biological sequence comparison.** *BMC Bioinformatics* 2005, **6**:31.
- Klupa T, Malecki MT, Pezzolesi M, Ji L, Curtis S, Langefeld CD, Rich SS, Warram JH, Krolewski AS: **Further evidence for a susceptibility locus for type 2 diabetes on chromosome 20q13.1-q13.2.** *Diabetes* 2000, **49**:2212-2216.
- Bento JL, Palmer ND, Mychaleckyj JC, Lange LA, Langefeld CD, Rich SS, Freedman BI, Bowden DW: **Association of protein tyrosine phosphatase 1B gene polymorphisms with type 2 diabetes.** *Diabetes* 2004, **53**:3007-3012.
- De Koning D-J, Rattink AP, Harlizius B, Van Arendonk JAM, Brascamp EW, Groenen MAM: **Genome-wide scan for body composition in pigs reveals important role of imprinting.** *Proc Natl Acad Sci USA* 2000, **97**:7947-7950.
- Lander ES, Waterman MS: **Genomic mapping by fingerprinting random clones: a mathematical analysis.** *Genomics* 1988, **2**:231-239.
- Clarke L, Carbon J: **A colony bank containing synthetic ColEI hybrid plasmids representative of the entire E. coli genome.** *Cell* 1976, **9**:91-101.
- Wendl MC, Barbazuk WB: **Extension of Lander-Waterman theory for sequencing filtered DNA libraries.** *BMC Bioinformatics* 2005, **6**:245.
- Wendl MC: **A general coverage theory for shotgun DNA sequencing.** *J Comput Biol* 2006, **13**:1177-1196.
- Sundquist A, Ronaghi M, Tang H, Pevzner P, Batzoglu S: **Whole-genome sequencing and assembly with high-throughput, short-read technologies.** *PLoS ONE* 2007, **2**:e484.
- Bentley DR: **Whole-genome re-sequencing.** *Curr Opin Genet Dev* 2006, **16**:545-552.
- Tammi MT, Arner E, Kindlund E, Andersson B: **Correcting errors in shotgun sequences.** *Nucleic Acids Res* 2003, **31**:4663-4672.
- Potter SC, Clarke L, Curwen V, Keenan S, Mongin E, Searle SMJ, Stabenau A, Storey R, Clamp M: **The Ensembl analysis pipeline.**

Genome Res 2004, **14**:934-941.

46. Benson G: **Tandem repeats finder: A program to analyze DNA sequences.** *Nucleic Acids Res* 1999, **27**:573-580.
47. **ENSEMBL Trace Server** [<http://trace.ensembl.org/>]