On-line Methods and Supplementary information

1. Sequence and assembly

1.1 From Duroc 2-14 DNA to Sscrofa11.1 assembly

1.1.1 Sample, sequencing and assembly

DNA was extracted from Duroc 2-14 cultured fibroblast cells passage 16-18 using the Qiagen Blood & Cell Culture DNA Maxi Kit, producing 139.15 µg DNA from three extractions. The high molecular weight DNA from this extraction was sequenced by Pacific Biosciences (PacBio) using their long read sequencing technology. Libraries for SMRT sequencing were prepared and sequenced as described previously (Pendleton *et al.*, 2015) using P6-C4 chemistry on the RSII using 213 SMRT cells. Initial read statistics are detailed in supplementary table ST1.

Supplementary Table ST1: Pacific Biosciences read statistics

	TJTabasco (Duroc 2-14)	MARC1423004
Chemistry	P6/C4	P5/C3 and P6/C4
Number of reads	12,328,735	32,960,338
Total length of reads (bp)	175,934,815,397	186,973,885,772
Mean read length (bp)	14,270	6,144
Read N50 (bp)	19,786	9,277

Contigs were assembled using the Falcon v0.4.0 assembly pipeline following the standard protocol. Quiver v. 2.3.0 (Chin *et al.*, 2013) was used to correct the primary and alternative contigs. Only the primary pseudo-haplotype contigs were used in the assembly.

1.1.2 Contig quality assessment and contig splitting

Paired-end Illumina reads from the same individual (http://www.ebi.ac.uk/ena/data/view/PRJEB9115) were mapped to the 3,206 haploid contigs and assessed for structural abnormalities using the methods described previously (Warr *et al.*, 2015). Briefly, 1,000 bp windows across the contigs were assessed for levels of abnormal mapping including high GC-normalized coverage, improper pairing and unexpected insert sizes. Additionally BAC end sequences (BES) (CHORI-242 library) (Humphray *et al.*, 2007) and fosmids (WTSI_1005 library: https://www.ncbi.nlm.nih.gov/clone/library/genomic/234/) (ENA accession:HE000001 – HE565349) (Skinner *et al.*, 2016) from the same individual (i.e. Duroc 2-14) were mapped to the contigs and regions with multiple occurrences of incorrect orientation were examined manually in the Integrative Genomics Viewer (IGV) (Robinson *et al.*, 2011). For 28 contigs where there was consistent evidence of structural disagreement between the contigs and the Illumina reads, BAC ends and fosmids, the contigs were split or trimmed.

1.1.3 Scaffolding

In order to establish an initial scaffold the contigs were mapped to Sscrofa10.2 using Nucmer (v3.23) (Kurtz *et al.*, 2004). The positioning of the contigs was determined by using the longest ascending subset of mapping locations using the show-coords tool from Mummer with the –g flag. Contigs with a %IDY below 95% were excluded. Contigs that mapped to regions substantially larger (>180%) or smaller (<10%) than the contig size were excluded. These tolerances were intentionally lenient due to the inflated gap sizes in the Sscrofa10.2 assembly (e.g. including 50 kb between scaffolds as required by the NCBI submission system in 2011) and highly fragmented nature of certain regions of Sscrofa10.2. Adjacent contigs were merged into a single fasta entry with Ns representing gaps between them. Gaps were estimated from the distance between the mapping locations against Sscrofa10.2, with an upper limit of 50 kb. Several of the remaining contigs were placed by identifying their longest alignment position, if this alignment was more than 50% the length of the contig and overlapped with a gap with a IDY>90% they were placed in the gaps with 25 bp gaps either side. 346 contigs covering 2.3 Gb were included in the initial chromosomal scaffolds.

1.1.4 Gap filling

PBJelly (English *et al.*, 2012) was used with the 65X raw PacBio reads to fill the gaps in the scaffolds. Default parameters were used for all stages except the assembly stage where max wiggle (-w) was set to 100 kb and max trim (-t) was set to 1,000 bp. These parameters were changed to account for the extremely inaccurate gap sizes and missing sequence in Sscrofa10.2 that will have influenced the estimated gap sizes, to allow heavily overlapping contigs to be closed and to allow potentially low-quality sequence at the end of contigs to be excluded. Following initial gap filling, PBJelly was rerun on the fasta output from the first round, with the unused contigs from the Falcon output added to the fasta to allow extension of the scaffolds. These contigs had been excluded initially to reduce secondary mapping positions. PBJelly is able to add contigs to the end of scaffolds, but not place whole contigs in gaps, so the initial mapping of contigs to scaffolds was examined to find if any of the contigs that had been excluded in this stage due to overlap with existing contigs might fill the gaps. Contigs were placed on a case-by-case basis if there was evidence of overlap with placed sequence on both sides of the gap, if the initial contig quality control was good, and if placement was well supported by BAC end mapping. Additionally, BACs for which the end sequences mapped to adjacent contigs providing evidence for scaffolding these adjacent contigs and for which finished quality sequence was publically available, were aligned and the gap filled and placed following the same restrictions as the unplaced contigs. On completion of these gap-filling procedures 108 gaps remained. Estimation of the size of the remaining gaps was based on BAC end mapping, using the known median insert size of the CHORI-242 library (see https://bacpacresources.org). Any gaps estimated to be <100 bp were sized at 100 bp and unspanned gaps were sized at 50 kb.

1.1.5 Targeted BAC sequencing to fill gaps

Five BACs from the CHORI-242 library were selected for further sequencing (CH242-188M9 (SSC16); CH242-323K10 (SSC18); CH242-284F8 (SSC18); CH242-61K12 (SSC1); CH242-168C15 (SSC12)) based on BAC ends mapping either side of gaps. The BAC clones were obtained from BACPAC (https://bacpacresources.org) and DNA was extracted using the Epicentre BACMAX DNA purification kit following manufacturer’s instructions. The BAC DNA was sequenced using Oxford Nanopore Technologies’ MinION sequencer using a barcoded 2D library following the discontinued protocol SQK-LSK208 on an R9 flow cell using MinKNOW v1.0.5. Sequences were assembled using Canu (Koren *et al.*, 2017) with default settings and each produced a single contig. The BAC vector sequences were removed from the contigs, the contigs were mapped to the assembly initially with Nucmer to confirm they mapped to the expected locations, with exact positions for placement determined by BWA-MEM (Li, 2013). All five contigs mapped to the expected positions and were placed to close the targeted gaps, leaving 103 gaps in the final Sscrofa11 assembly and closing chromosomes 16 and 18.

1.1.6 Polishing

Error correction was done using Arrow from the GenomicConsensus suite (https://github.com/PacificBiosciences/GenomicConsensus) using the original 65X PacBio coverage. This was followed by Pilon (Walker *et al.*, 2014) with fixlist restricted to "bases", but otherwise using default parameters and paired-end Illumina short read data that provided 50x genome coverage.

1.2 From MARC1423004 DNA to assembly USMARCv1.0

1.2.1 Sample, sequencing and assembly

DNA was isolated from barrow MARC1423004 using a salt extraction method. Briefly, frozen lung tissue was crushed into powder, scraped into a 15 mL tube, and suspended in 4 mL digestion buffer (10 mM NH_4Cl, 400 mM NaCl, 50 mM Na_2EDTA, pH 8.0). Digestion was initiated with 100 µL 20% SDS and 70 µL trypsin (5 mg/ml). This initial digestion was allowed to proceed at room temperature (approximately 22ºC) for one hour, and then 200 µL of 20% SDS and 50 µL of Proteinase K (50 mg/mL) were added. The digestion was incubated at 55ºC in a shaking water bath overnight (16 hours). Another 100 µL of Proteinase K were added and incubation extended for another 1.5 hours, until no remaining tissue pieces could be observed in the solution, and then 10 µL of RNase (10 U/µL) were added followed by additional incubation for one hour. 1.25 mL 5M NaCl was added, mixed by inversion, and the tube was centrifuged at 3200 x g at 4ºC. The supernatant was transferred to a fresh 15 mL tube, and DNA precipitated by addition of 2.5 volumes of 95% ethanol. The precipitate was removed using a hooked Pasteur pipet, dipped twice in separate tubes of 70% ethanol on ice, and allowed to briefly dry in air on the hook. The DNA was then eluted from the hook by placing it under 250 µL TE buffer (10 mM Tris-HCl, 0.1 mM EDTA) until the pellet slipped off into the buffer. The hook was then removed, and the DNA was allowed to dissolve into the buffer for several days at 4ºC until it appeared to be completely dissolved. The high molecular weight DNA from this extraction was sequenced by Pacific Biosciences (PacBio) using their long read sequencing technology. Libraries for SMRT sequencing were prepared and sequenced as described previously (Pendleton *et al.*, 2015) using P5/C3 and P6-C4 chemistry on the RSII. A total of 199 P5/C3 cells and 127 P6/C4 cells were produced. Initial read statistics are detailed in supplementary table ST1. Contigs were assembled using Celera Assembler v8.3rc2 (Berlin et al., 2015) using the command:

```
wgs-8.3/Linux-amd64/bin/PBcR -s pacbio.spec -fastq
filtered_subreads.fastq genomeSize=3000000000 -sensitive -l swine
sgeName=swine "sge=-p -500 -A swinenewsens"  useGrid=1 scriptOnGrid=1

and spec file:
merSize = 16

ovlMemory = 32
ovlStoreMemory = 32000
ovlThreads = 32
threads = 32
ovlConcurrency = 1
cnsConcurrency = 8
merylThreads = 32
merylMemory = 32000
frgCorrThreads = 16
frgCorrBatchSize = 100000
ovlCorrBatchSize = 100000

useGrid=1
scriptOnGrid=1
ovlCorrOnGrid=1
frgCorrOnGrid=1

sge = -A assembly
sgeScript = -pe threads 1
sgeConsensus = -pe threads 8
sgeOverlap = -pe threads 4 -l mem=2GB
gridEngineMhap = -pe threads 15 -l mem=2GB
sgeCorrection = -pe threads 15 -l mem=2GB
sgeOverlapCorrection  = -pe threads 1 -l mem=16GB
sgeFragmentCorrection=-pe threads 2 -l mem=2GB
sgeOverlapCorrection=-pe threads 1 -l mem=4GB

asmOvlErrorRate=0.1
asmUtgErrorRate=0.06
asmCgwErrorRate=0.1
asmCnsErrorRate=0.1
asmOBT=1
asmObtErrorRate=0.08
```

```
      asmObtErrorLimit=4.5

      batOptions=-RS -NS -CS
      utgGraphErrorRate=0.055
      utgGraphErrorLimit=4
      utgMergeErrorRate=0.055
      utgGraphErrorLimit=4

      ovlHashBits=24
      ovlHashLoad=0.80

      ovlHashBlockLength      =300000000
      ovlRefBlockLength       =0
      ovlRefBlockSize         =2000000
```

This initial assembly was 2.67 Gbp in 16,441 contigs and an N50 of 2.8 Mbp. Quiver from SMRTportal v. 2.3.0 (Chin *et al.*, 2013) was used to correct the assembly.

1.2.3 Scaffolding

The lung tissue from the pig was sent to Dovetail Genomics (Santa Cruz) for scaffolding by Chicago and HiRise as described (Putnam *et al.*, 2016). This process identified 270 putative misjoins in the contigs and output scaffolds 13,039 scaffolds (294 > 50 kb). The total length was 2.66 Gbp and scaffold N50 was 36.5 Mbp. The dovetail scaffolds were gap-filled where a single contig spanned the gap, correcting false breaks made by HiRise. The resulting assembly was used for reference-guided scaffolding based on the Sscrofa11.1 reference. In case of conflicts, with the exception of cross-chromosome joins, the USDA assembly was unchanged.

1.1.4 Gap filling

PBJelly (English *et al.*, 2012) was used with the 65X raw PacBio reads to fill the gaps in the scaffolds. Default parameters were used for all steps.

1.2.5 Polishing

Gap filling was followed by Pilon (Walker *et al.*, 2014) with fixlist restricted to “bases”, but otherwise using default parameters and paired-end Illumina short read data that provided 50x genome coverage. The final assembly of 2.8 Gbp has a scaffold N50 of 131.5 Mbp and a contig N50 of 6.4 Mbp (Table 1).

1.3 Anchoring the assemblies to chromosomes

1.3.1 Chromosome Preparation

Heparinized blood samples were cultured for 72 h in PB MAX Karyotyping medium (Invitrogen) at 37°C, 5% CO_2. Cell division was arrested by adding colcemid at a concentration of 10.0 µg/ml (Gibco) for 30 min prior to hypotonic treatment with 75 mM KCl and fixation to glass slides using 3:1 methanol:acetic acid.

1.3.2 Preparation and Selection of BAC clones for FISH

BAC clones with inserts of approximately 150 kb in size were selected for position using the Sscrofa10.2 NCBI database (www.ncbi.nim.nih.gov) and ordered from the PigE-BAC library (ARK-Genomics) (Anderson *et al.*, 2000) and the CHORI-242 Porcine BAC library (BACPAC, https://bacpacresources.org/). BAC clone DNA was isolated using the Qiagen Miniprep Kit (Qiagen) prior to amplification and direct labelling by nick translation. Probes were labelled with Texas Red-12-dUTP (Invitrogen) and FITC- Fluorescein-12-UTP (Roche) prior to purification using the Qiagen Nucleotide Removal Kit (Qiagen).

1.3.3 Fluorescence *in situ* hybridisation

Metaphase preparations were fixed to slides and dehydrated through an ethanol series (2 min each in 2×SSC, 70%, 85% and 100% ethanol at RT). Probes were diluted in a formamide buffer (Cytocell) with Porcine Hybloc (Insight Biotech) and applied to the metaphase preparations on a 37°C hotplate before sealing with rubber cement. Probe and target DNA were simultaneously denatured for 2 mins on a 75°C hotplate prior to hybridisation in a humidified chamber at 37°C for 16 h. Slides were washed post hybridisation in 0.4x SSC at 72°C for 2 mins followed by 2x SSC/0.05% Tween 20 at RT for 30 secs, and then counterstained using VECTASHIELD anti-fade medium with DAPI (Vector Labs). Images were captured using an Olympus BX61 epifluorescence microscope with cooled CCD camera and SmartCapture (Digital Scientific UK) system.

The Sscrofa11.1 and USMARCv1.0 assemblies were searched using BLAST with sequences derived from the BAC clones which had been used as probes for the FISH analyses. For most BAC clones these sequences were BAC end sequences (Humphray *et* *al.*, 2007), but in some cases these sequences were incomplete or complete BAC clone sequences (Groenen *et al.*, 2012; Skinner *et al.*, 2016). The links between the genome sequence and the BAC clones used in cytogenetic analyses by fluorescent *in situ* hybridization are summarised in Supplementary Table ST2.

The fluorescent *in situ* hybridization data indicate that the following chromosomal scaffolds in the USMARCv1.0 are inverted relative to the conventional cytogenetic orientation of the corresponding chromosomes: SSC1, SSC6, SSC7, SSC8, SSC9, SSC10, SSC11, SSC13, SSC14, SSC15, and SSC16. Whilst the USMARCv1.0 assembly of SSC16 appears overall to be in the reverse orientation with respect to the cytogenetic orientation and the Sscrofa11.1 assembly of this chromosome it also appears to harbour sequences at the start of the scaffold that perhaps belong at the other end of the scaffold.

The fluorescent *in situ* hybridization results also indicate areas where future assemblies might be improved. For example, the Sscrofa11.1 unplaced scaffolds contig 1206 and contig1914 may contain sequences that could be added to end of the long arms of SSC1 and SSC7 respectively. Examples of the primary fluorescent in situ hybridisation data are provided in Supplementary Figures SF1a, SF1b.

Supplementary Table ST2: Fluorescent *in situ* hybridisation results using named BAC clones as probes plus sequence matches for sequences derived from these BAC clones.

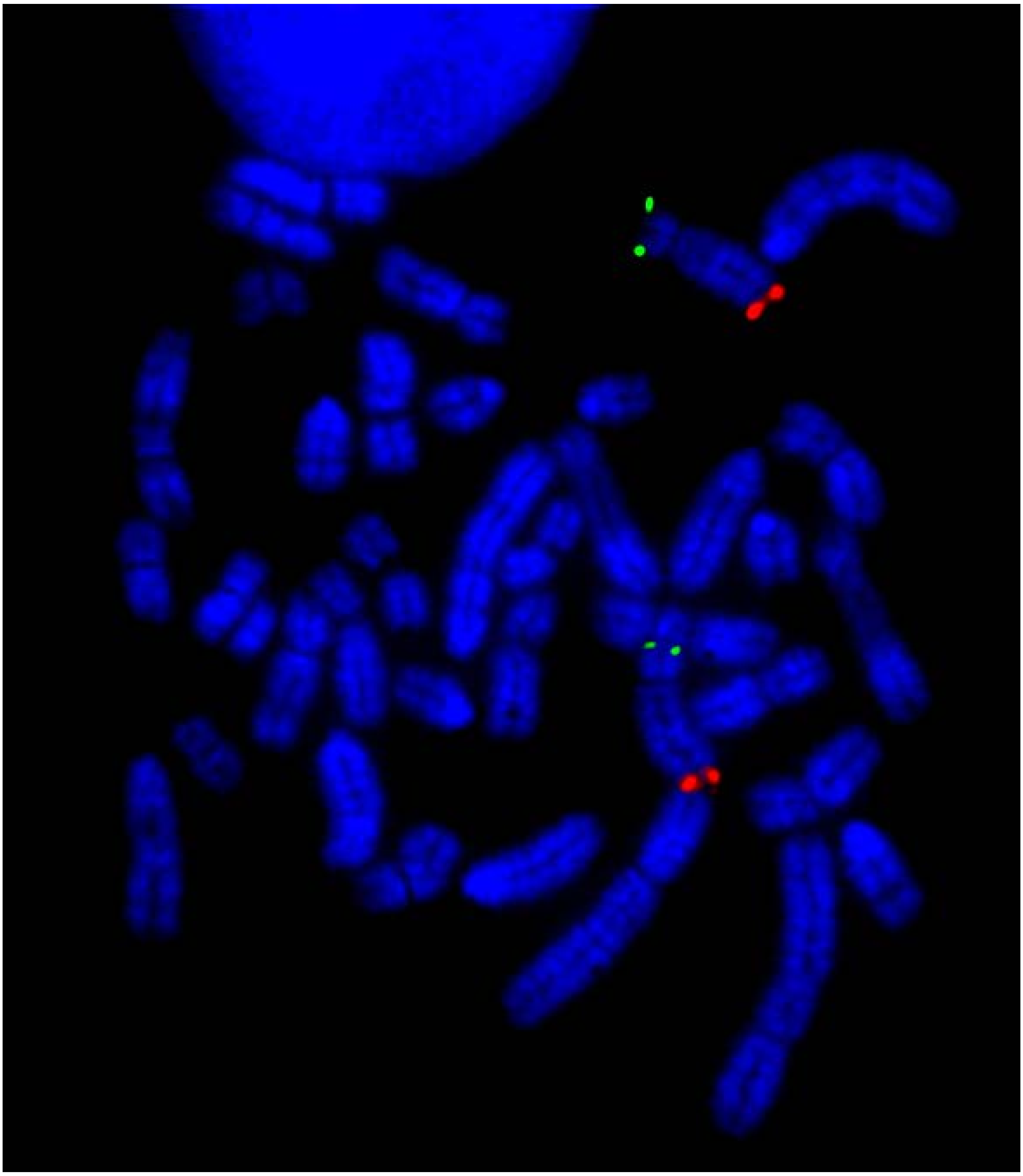
Chr	BAC Name	BES	FISH	Sscrofa11.1 coordinates	USMARCv1.0 coordinates
1	PigE-232G23	CT070230.1; CT218278.1	1p	1:615,021-619,597	1:280,453,704-280,458,272
1	CH242-248F13	FP340244.3	1p	1:1,470,202-1,660,001	1:279,368,385-279,558,294
1	CH242-151E10	CT239299.1; CT245986.1	1q	unplaced scaffold: Contig1206	1: 6,156,768-6,336,737
2	PigE-117G14	CT074446.1; CT074447.1	2p	2:19,406-161,226	2:537,026-678,808
2	PigE-8G19	CT260033.1; CT260032.1	2p	2:552,031-671,098	2:29,620-146,529
2	CH242-188K23	CU929880	2 cen	2:52,747,463-52,933,130	2:51,728,649-51,908,148
2	CH242-230M23	CT144824.1; CT258059.1	2 cen	2:53,300,582-53,472,497	no match
2	CH242-441A1	CT364255.1;CT364256.1	2 cen	2:53,458,574-53,652,606	2:52,095,251-52,095,932
2	CH242-294F6	CT378635.1; CT378634.1	2q	2:151,178,736-151,402,963	2:145,456,152-145,678,427
3	PigE-168G22	CT094069.1; CT094070.1	3p	3:301,813-509,346	3:218,358-425,025
3	CH242-315N8	CT359002.1; CT359003.1	3q	3:122,720,374-122,869,530	no match
4	PigE-262E12	CT082779.1; CT193441.1	4p	4:37,383-223,717	4: 96,811- 97,511
4	PigE-131J18	CT116562.1; CT171811.1	4p	4:449,934-626,677	4:322,853-499,367
4	PigE-85G21	CT070098.1; CT190031.1	4q	4:130,625,653-130,748,215	4:130,215,908-130,338,404
5	CH242-288F8	CT132004.1; CT211915.1	5p	5:170,319-344,353	5:188,019-362,653
5	PigE-178M22	CT139068.1; CT155898.1	5p	5:175,168-311,462	5:192,886-329113
5	CH242-133F9	CT166002.1; CT166003.1	5p	5:438,296-633,458	5:456,924-652,458
5	PigE-127K14	CT057696.1; CT057697.1	5p	5:1,003,455-1,129,329	5:1,024,261-1,148,699
5	PigE-74P10	CT188857.1; CT188858.1	5p	5:3,739,938-3,883,755	5:103,338,585-103,481,984
5	PigE-99L23	CT079916.1; CT106700.1	5p Mid	5:31,980,969-32,114,628	no match
5	CH242-63B20	FP102738	5q	5:104,304,289-104,489,770	no match
6	PigE-238J17	CT220438.1; CT220439.1	6p	6:2,333,972-2,522,065	6:162,952,836-163,141,204
6	PigE-199E24	CT272854.1; CT272853.1	6 below cen	6:62,771,286-62,952,647	6:104,969,580-105,152,317
6	CH242-510F2	CT396711.1; CT442620.1	6q	6:170,248,061-170,454,571	6:162,654-369,119
7	PigE-52L22	CT054562.1; CT063652.1	7p	7:188,339-317,255	7:125,463,765-125,463,765
7	PigE-246A1	CT203984.1; CT070741.1	7 cen	7:24,628,314-24,671,828	no match
7	PigE-230H8	CT120917.1	7q below cen	7:46,704,415-46,704,995	7:395,704-396,284
7	PigE-75E21	CT188956.1; CT261917.1	7q below cen	7:46,901,592-47,032,091	7:68,406-199,212
7	CH242-103I13	CU695123.2	7q	Unplaced scaffold: Contig1914	7:7,614,911-7,838,927

Chr	BAC Name	BES	FISH	Sscrofa11.1 coordinates	USMARCv1.0 coordinates
8	PigE-134L21	CT126839.1; CT172501.1	8p	8:570,904-705,341	8:280,369,080-280,502,409
8	PigE-2N1	CT229915.1; CT229916.1	8p	8:819,717-958,131	8:137,599,822-137,737,820
8	PigE-118B21	CT048761.1; CT091504.1	8q	8:138,491,413-138,647,394	8:322,914-478,869
9	CH242-65G4	CU695192.2	9p	9:320,582-511,079	9:137,686,630-137,874,917
9	PigE-126O17	CT170583.1; CT057320.1	9p	9:443,462-603,022	9:137,594,779-137,754,110
9	PigE-242D8	CT123266.1; CT123265.1	9 mid	9:67,752,381-67,910,109	9:71,096,887-71,254,731
9	CH242-411M8	CT362997.1; CT468791.1	9q	9:139,180,446-139,338,710	9:168,756-327,007
10	CH242-451I23	CT369304.1; CT459538.1	10p	Unplaced scaffold: Contig2471	10:71,863,534-72,028,842
10	CH242-36D16	CT345373.1; CT186999.1	10q	10:55,422,866-55,600,351	10:15,300,371-15,480,359
10	CH242-517L16	FP325295.2	10q	10:55,609,778-55,800,022	10:15,098,969-15,290,916
11	PigE-199B10	CT272693.1	11p	11:135,233-297,713	11:79,101,520-79,264,254
11	PigE-232N19	CT193346.1	11p	11:290,540-291,222	11:79,108,017-79,108,697
11	PigE-211E21	CT044498.1; CT044499.1	11p	11:1,584,043-1,743,425	11:77,663,220-77,822.434
11	CH242-239O11	CT146353.1; CT286242.1	11q	11:78,888,491-79,057,526	11:827,483-996,382
12	PigE-253K5	CT081057.1; CT204391.1	12p	12:324,614-524,015	12:3,288-206,400
12	PigE-124G15	CT056668.1; CT092177.1	12q	12:60,846,540-60,990,610	12:58,746,918-58,890,342
13	PigE-197C11	CT271598.1; CT271599.1	13p	13:556,804-694,010	13:204,579,401-204,716,338
13	PigE-179J15	CT124924.1; CT124925.1	13q	13:205,856,740-206,006,912	13:3,005,553-3,154,893
14	PigE-137C12	FP340551.3	14p	14:17,423-156,591	14:140,940,126-140,804,938
14	PigE-167E18	CT089616.1; CT089617.1	14q	14:141,407,495-141,435,234	14:98,899-125,652
15	PigE-90C11	CT190903.1; CT190904.1	15p	15:3,442,144-3,596,666	15:139,733,189-139,886,921
15	PigE-108N22	CT073138.1; CT046453.1	15 mid	15:56,903,229-57,028,679	no match
15	CH242-170N3	FP236135.2	15q	15:139,616,279-139,784,756	15:3,511,408-3,588,855
16	PigE-90L22	CT191132.1; CT113297.1	16p	16:109,696-235,547	16:87,402-212,531
16	PigE-124C22	CT056551.1; CT056550.1	16p	16:117,329-308,428	16:94,873-287,243
16	CH242-4G9	CT041970.1; CT041969.1	16p	16:141,557-324,802	16:118,753-303,587
16	PigE-173H6	CT123878.1; CT123877.1	16p	16:167,106-299,570	16:144,276-278,432
16	PigE-149F10	CT088298.1; CT153977.1	16p	16:596,671-782,524	16:78,918,129-79,108,868
16	CH242-42L16	CT347302.1; CT347303.1	16q	16:79,097,179-79,303,695	16:878,687-1,085,418
17	CH242-70L7	CT077340.1; CT077341.1	17p	17:545,995-673,770	17:464,378-592,438
17	PigE-190G24	CT126644.1; CT096362.1	17p	17:515,422-707,787	17:433,829-626,496
17	CH242-243H19	CT321876.1; CT321877.1	17q	17:61,760-582-61,937,945	17:62,450,941-62,628,249

Chr	BAC Name	BES	FISH	Sscrofa11.1 coordinates	USMARCv1.0 coordinates
18	PigE-253N22	CT081116.1; CT204433.1	18p	18:1,616,389-1,751,286	18:1,565,719-1,700,920
18	PigE-202I11	CT042866.1; CT254626.1	18q	18:55,539,630-55,700,409	18:55,320,418-55,481,057
X	CH242-447L20	CT377508.1; CT467360.1	Xp	X:505,086-692,549	no match
X	CH242-156O11	FP074895.7	Xp + Yp	X:6,337,709 6,584,993	X:7,588,110-7,597,109
X	CH242-19N1	CU856094.8	Xp	X:6,705,194-6,834,183	X:7,588,110-7,715,932
X	CH242-305A15	CU861979.13	Xq	X:125,384,028-125,529,813	X:126,150,718-126,296,945
Y	CH242-156O11	FP074895.7	Xp + Yp	Y:4,744,231-4,791,971	Y:32,909,634-32,923,401

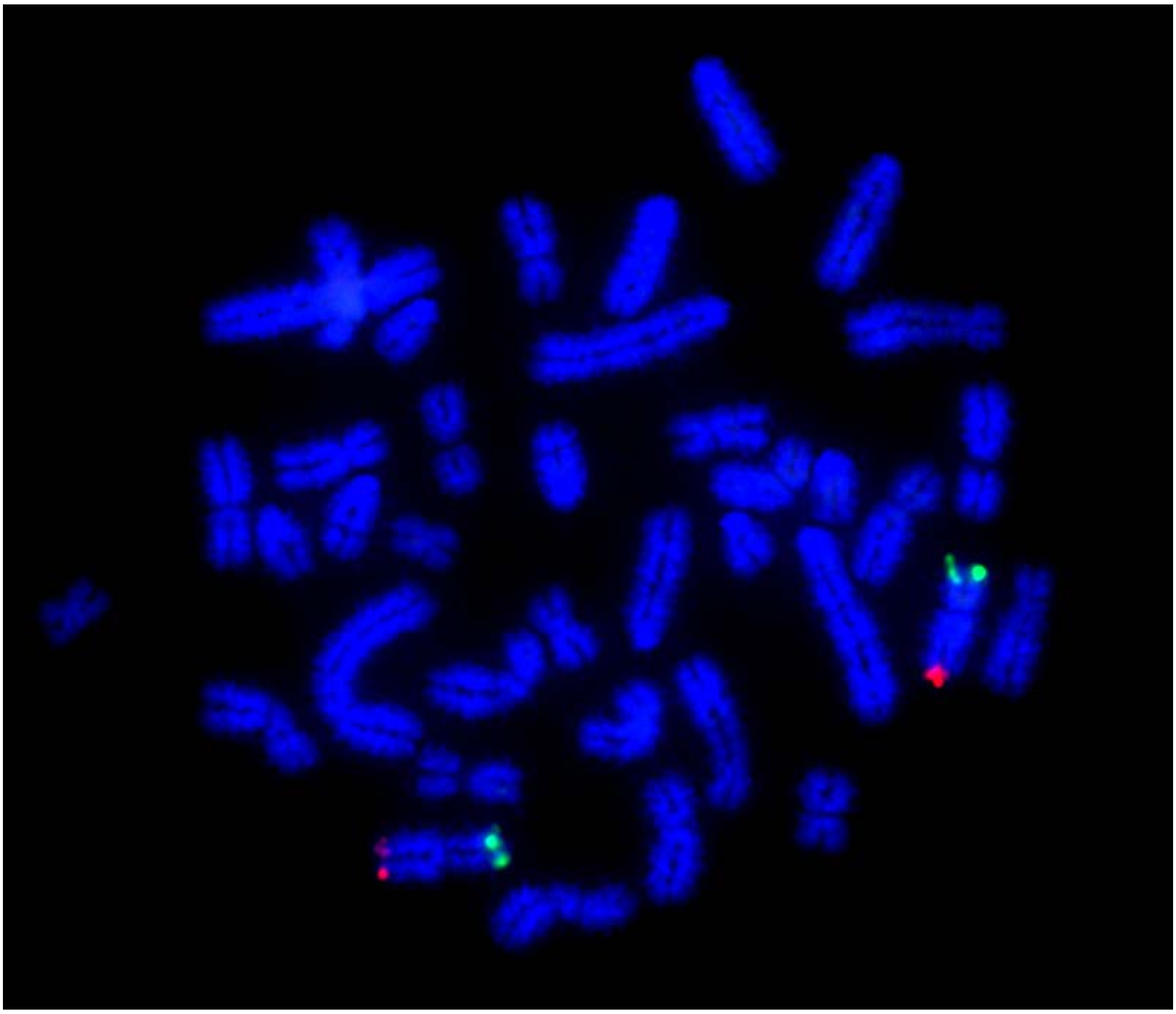
Supplementary Figure SF1a: Fluorescent *in situ* hybridisation assignments

a. SSC6 – p-telomeric end labelled with PigE-238J17, q-telomeric end labelled with CH242-510F2



Supplementary Figure SF1b: Fluorescent *in situ* hybridisation assignments

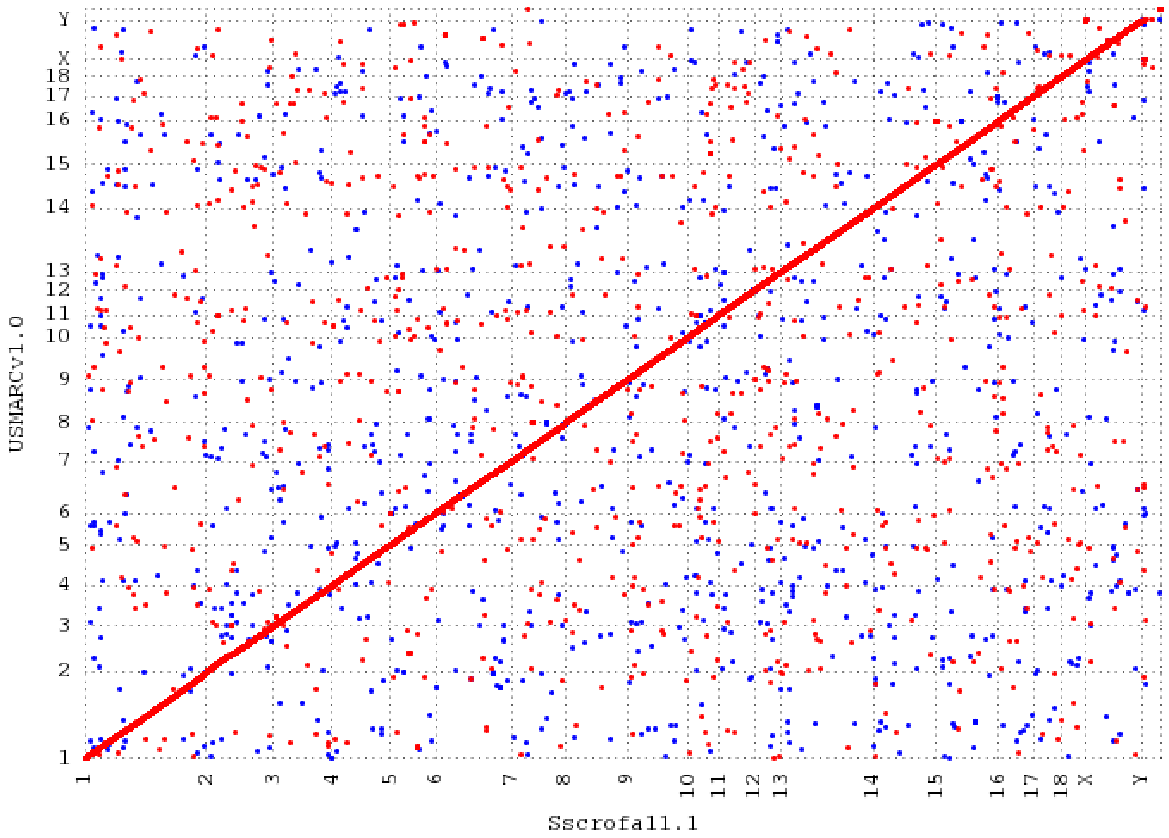
b. SSCX – p-telomeric end labelled with CH242-19N1, q-telomeric end labelled with CH242-305A15



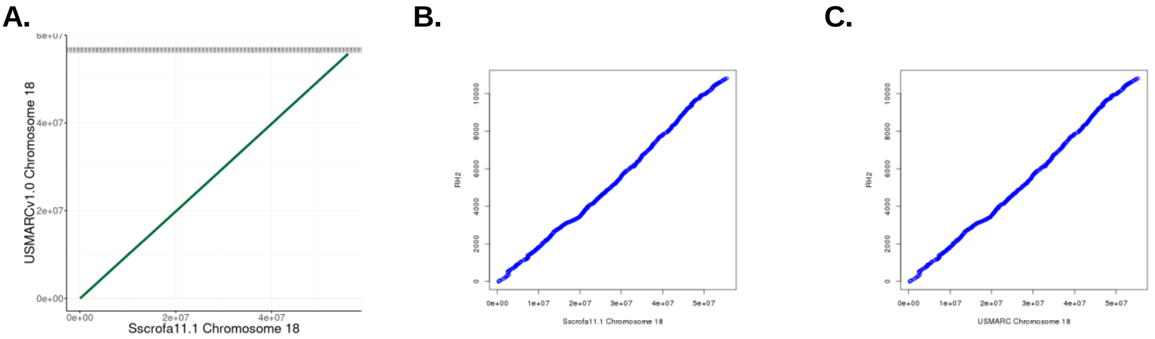
1.4 Quality Assessment of Sscrofa11.1 and USMARCv1.0 assemblies

1.4.1 Order and orientation

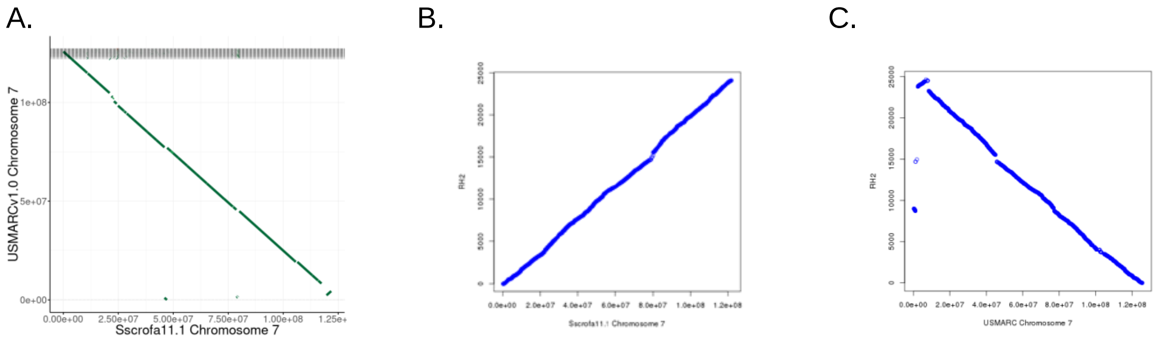
In addition to assigning and orienting the scaffolds on chromosomes as described above, order and orientation within chromosome assemblies was checked by alignment to the radiation hybrid map (Servin *et al.*, 2012) and alignments amongst the assemblies (Sscrofa10.2, Sscrofa11.1 and USMARCv1.0). The overall alignments indicate that the new assemblies (Sscrofa11.1, USMARCv1.0) are essentially co-linear with each other and with the radiation hybrid map (Figure 1, Supplementary Figure SF2). At the level of individual chromosomes, order and orientation within chromosome 18, for example, is consistent between Sscrofa11.1 and USMARCv1.0 and both SSC18 chromosome assemblies are supported by the radiation hybrid map (Supplementary Figure SF3). However, although the alignments of other chromosomes with the radiation map also support the overall co-linearity of the sequence and radiation hybrid maps, there are some differences in local order and orientation between the Sscrofa11.1 and USMARCv1.0 as illustrated in Supplementary Figures SF4 and SF5 for SSC7 and SSC8 respectively.



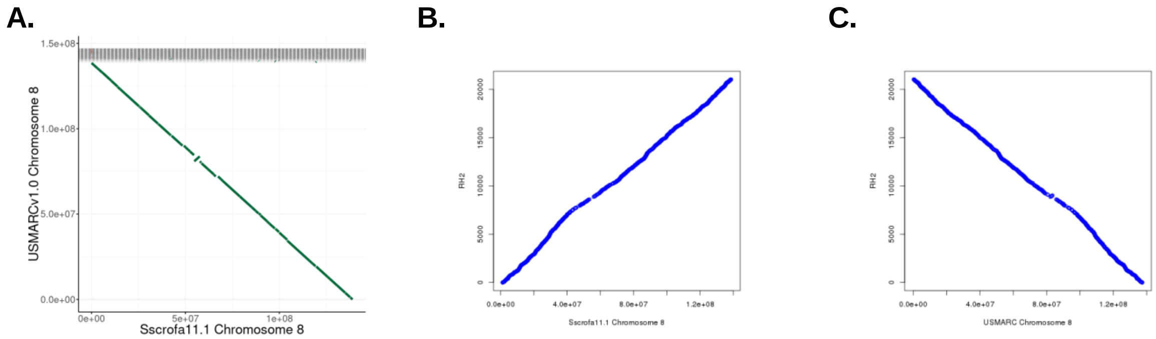
Supplementary Figure SF2: Alignment of Sscrofa11.1 and USMARCv1.0 assemblies after correcting inversions of USMARCv1.0 chromosome scaffolds



Supplementary Figure SF3: Order and orientation of SSC18 assemblies: A. alignment of Sscrofa11.1 and USMARCv1.0 assemblies of SSC18; B. alignment of Sscrofa11.1 and radiation hybrid map (RH2); C. alignment of USMARCv1.0 and radiation hybrid map (RH2).



Supplementary Figure SF4: Order and orientation of SSC7 assemblies: **A.** alignment of Sscrofa11.1 and USMARCv1.0 assemblies of SSC7; **B.** alignment of Sscrofa11.1 and radiation hybrid map (RH2); **C.** alignment of USMARCv1.0 and radiation hybrid map (RH2).



Supplementary Figure SF5: Order and orientation of SSC8 assemblies: **A.** alignment of Sscrofa11.1 and USMARCv1.0 assemblies of SSC8; **B.** alignment of Sscrofa11.1 and radiation hybrid map (RH2); **C.** alignment of USMARCv1.0 and radiation hybrid map (RH2).

The matches shown in the grey zone at the top of each plot of the Sscrofa11.1 versus USMARCv1.0 alignments probably represent a mix of repetitive sequences and matches to the unplaced scaffolds in the USMARCv1.0 assembly.

Whether the differences between Sscrofa11.1 and USMARCv1.0 in order and orientation within chromosomes represent assembly errors or real chromosomal differences will require further research. The sequence present at the telomeric end of the long arm of chromosome 7 (after correcting the orientation of the USMARCv1.0 SSC7 assembly) is missing from the Sscrofa11.1 SSC7 assembly, and currently located on a 3.8 Mbp unplaced scaffold (AEMK02000452.1) that harbours several genes including DIO3, CKB and NUDT14 whose orthologues map to human chromosome 14 as would be predicted from the pig-human comparative map (Meyers *et al.*, 2005). This omission will be corrected in an updated assembly in future.

1.4.2 BUSCO and Cogent analyses

The assembly was assessed for completeness using BUSCO (Simão *et al.*, 2015) (Table ST3) and Cogent (https://github.com/Magdoll/Cogent), and assessed for structural accuracy by checking consistency between markers from radiation hybrid maps (Servin *et al.*, 2012) and the assembly. PacBio transcriptome (Iso-Seq) data consisting of high-quality isoform sequences from 7 tissues (diaphragm, hypothalamus, liver, skeletal muscle (*longissimus dorsi*), small intestine, spleen and thymus) from the pig whose DNA was used as the source for the USMARCv1.0 assembly were pooled together for Cogent analysis. Cogent is a tool that identifies gene families and reconstructs the coding genome using full-length, high-quality (HQ) transcriptome data without a reference genome. Cogent partitioned 276,196 HQ isoform sequences into 30,628 gene families, of which had at least 2 distinct transcript isoforms. Cogent then performed reconstruction on the 18,708 partitions. For each partition, Cogent attempts to reconstruct coding 'contigs' that represent the ordered concatenation of transcribed exons as supported by the isoform sequences. The reconstructed contigs were then mapped back to Sscrofa11.1 and contigs that could not be mapped or map to more than one position are individually examined.

Supplementary Table ST3: BUSCO statistics, BUSCOv2 (OrthoDBv9)

	Sscrofa10.2	Sscrofa11.1	USMARCv1.0
Complete BUSCOs	80.9%	93.8%	93.1%
Complete and single-copy BUSCOs	80.2%	93.3%	92.6%
Complete and duplicated BUSCOs	0.7%	0.5%	0.5%
Fragmented BUSCOs	8.2%	3.5%	3.5%
Missing BUSCOs	10.9%	2.7%	3.4%
Total BUSCO groups searched	4,104	4,104	4,104

1.4.3 Assemblytics

A comparison of pig genome assemblies was undertaken using the Assemblytics tools (Nattestad and Schatz, 2016) (http://assemblytics.com). The comparisons are listed in Table ST4.

Supplementary Table ST4a: Assemblytics comparisons

Reference		**Sscrofa10.2 (GCF_000003025.5)**
Query	**Assembly accession**	
Sscrofa11.1	GCA_000003025.6	http://qb.cshl.edu/assemblytics/analysis.php?code=i0H3KuHhWjKO5Tn7nsXg
USMARCv1.0	GCA_002844635.1	http://qb.cshl.edu/assemblytics/analysis.php?code=faROmPzOlMp1q5IdToO8
Reference		**Sscrofa11.1**
Query	**Assembly accession**	
Sscrofa11.1	GCA_000003025.6	N/A
USMARCv1.0	GCA_002844635.1	http://assemblytics.com/analysis.php?code=4rscWrlT7paorSvTMI7L
Bamei	GCA_001700235.1	http://assemblytics.com/analysis.php?code=gpCq8VWG4aWrocrlCWww
Berkshire	GCA_001700575.1	http://assemblytics.com/analysis.php?code=dvVxU3qkCNUR3rWpm2FI
Hampshire	GCA_001700165.1	http://assemblytics.com/analysis.php?code=V6jWeDYKywLu4Av40Ikh
Jinhua	GCA_001700295.1	http://qb.cshl.edu/assemblytics/analysis.php?code=UxtEbFk065DWQBpYz0sV
Landrace	GCA_001700215.1	http://qb.cshl.edu/assemblytics/analysis.php?code=7V7QGUCXrNAtFcGL6DMT
LargeWhite	GCA_001700135.1	http://qb.cshl.edu/assemblytics/analysis.php?code=UymCHs1NirQkdMFFbM1e
Meishan	GCA_001700195.1	http://qb.cshl.edu/assemblytics/analysis.php?code=toDVmO7nus0BbyMCGKSc
Pietrain	GCA_001700255.1	http://qb.cshl.edu/assemblytics/analysis.php?code=TIlXYB2uQYgWbf5YqNXk
Rongchang	GCA_001700155.1	http://qb.cshl.edu/assemblytics/analysis.php?code=HzggG8kBPJ6uKWWEvZOV
Tibetan	GCA_000472085.2	http://qb.cshl.edu/assemblytics/analysis.php?code=o9WtyIF6wTnGsEeAiizn
Wuzhishan	GCA_000325925.2	http://qb.cshl.edu/assemblytics/analysis.php?code=UbH3avfeoW19DjJmVC8C

Supplementary Table ST4b: Assemblytics comparisons

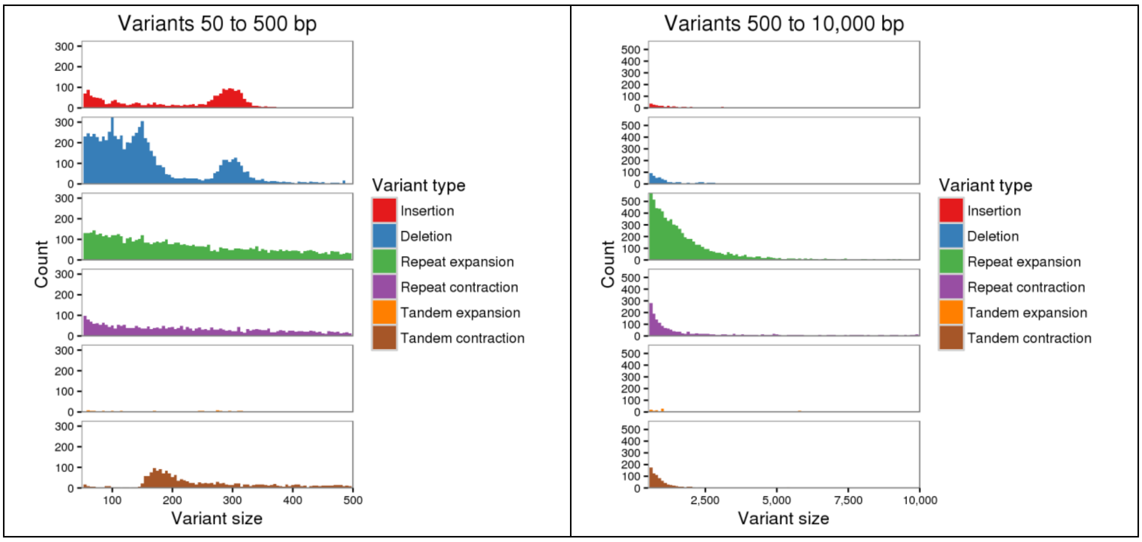
Reference		USMARCv1.10
Query	**Assembly accession**	
Sscrofa11.1	GCA_000003025.6	http://assemblytics.com/analysis.php?code=4rscWrlT7paorSvTMI7L
USMARCv1.0	GCA_002844635.1	N/A
Bamei	GCA_001700235.1	http://assemblytics.com/analysis.php?code=A1doW581DPkQKXlwfbtB
Berkshire	GCA_001700575.1	http://qb.cshl.edu/assemblytics/analysis.php?code=5dCXFbth2110zsguw58t
Hampshire	GCA_001700165.1	http://qb.cshl.edu/assemblytics/analysis.php?code=Xe5ENqAjsxeNcrK7TaRp
Jinhua	GCA_001700295.1	http://qb.cshl.edu/assemblytics/analysis.php?code=nqEihnLJRPsjNswVxV9J
Landrace	GCA_001700215.1	http://qb.cshl.edu/assemblytics/analysis.php?code=tfrtkAXiy148TUsb8HIJ
LargeWhite	GCA_001700135.1	http://qb.cshl.edu/assemblytics/analysis.php?code=lZM3EFMBzo9KyytQMSWH
Meishan	GCA_001700195.1	http://qb.cshl.edu/assemblytics/analysis.php?code=K9qeCrVxr9znPtFanHd3
Pietrain	GCA_001700255.1	http://qb.cshl.edu/assemblytics/analysis.php?code=U1n9D7z7DtRvbWjqEdTH
Rongchang	GCA_001700155.1	http://qb.cshl.edu/assemblytics/analysis.php?code=nEk3faE5s8YYckjNuvN7
Tibetan	GCA_000472085.2	http://qb.cshl.edu/assemblytics/analysis.php?code=NqjCZ7wvt6D0vm7Ai4tN
Wuzhishan	GCA_000325925.2	http://qb.cshl.edu/assemblytics/analysis.php?code=mEqp9WaGi9eceSY4Vid6

Supplementary Table ST4c: Assembly statistics* for pig genome assemblies subject to Assemblytics ananlyses

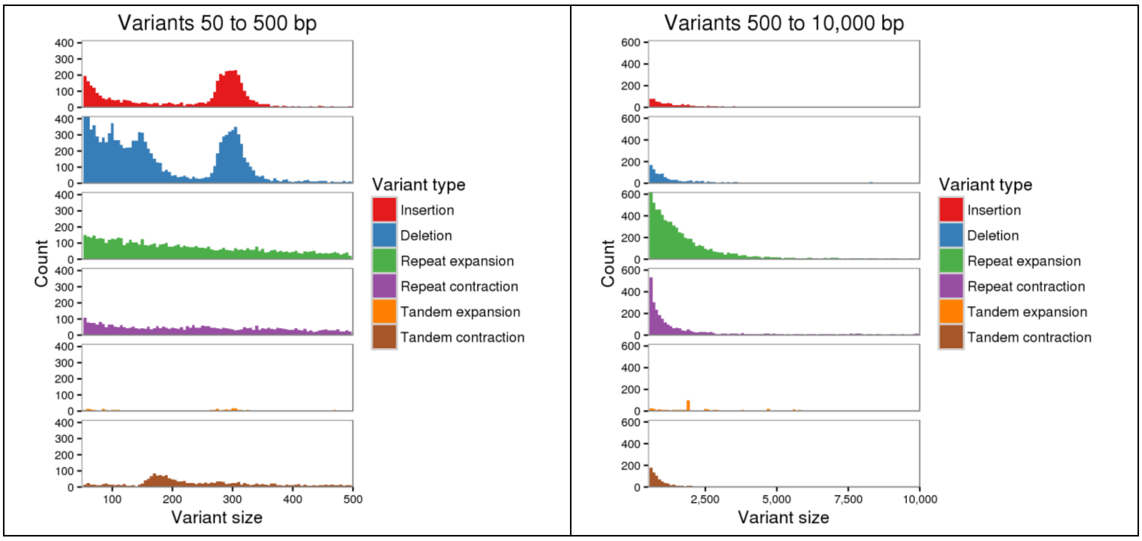
Assembly	Accession	Total (bp)(ungapped)	Scaffolds	Scaffold N50	Contigs	Contig N50
Sscrofa11.1	GCA_000003025.6	2,472,047,747	706	88,231,837	1,118	48,231,277
USMARCv1.0	GCA_002844635.1	2,623,130,238	14,818	131,458,098	14,818	6,372,407
Bamei	GCA_001700235.1	2,433,636,520	129,335	1,529,027	187,466	70,893
Berkshire	GCA_001700575.1	2,414,739,650	94,468	1,655,397	137,661	94,651
Hampshire	GCA_001700165.1	2,418,011,428	82,206	1,550,023	122,452	102,417
Jinhua	GCA_001700295.1	2,433,032,022	115,554	1,478,908	158,796	95,227
Landrace	GCA_001700215.1	2,420,570,845	94,659	1,407,841	141,909	88,142
LargeWhite	GCA_001700135.1	2,430,896,979	102,342	2,441,555	150,742	88,831
Meishan	GCA_001700195.1	2,438,814,343	133,833	1,248,180	201,146	63,263
Pietrain	GCA_001700255.1	2,415,062,022	88,436	1,663,542	139,497	80,611
Rongchang	GCA_001700155.1	2,429,730,895	120,246	2,325,000	173,508	79,093
Tibetan	GCA_000472085.2	2,379,878,366	72,068	861,885	148,234	57,199
Wuzhishan	GCA_000325925.2	2,453,484,489	137,577	5,853,977	272,163	31,939

* source NCBI Assembly

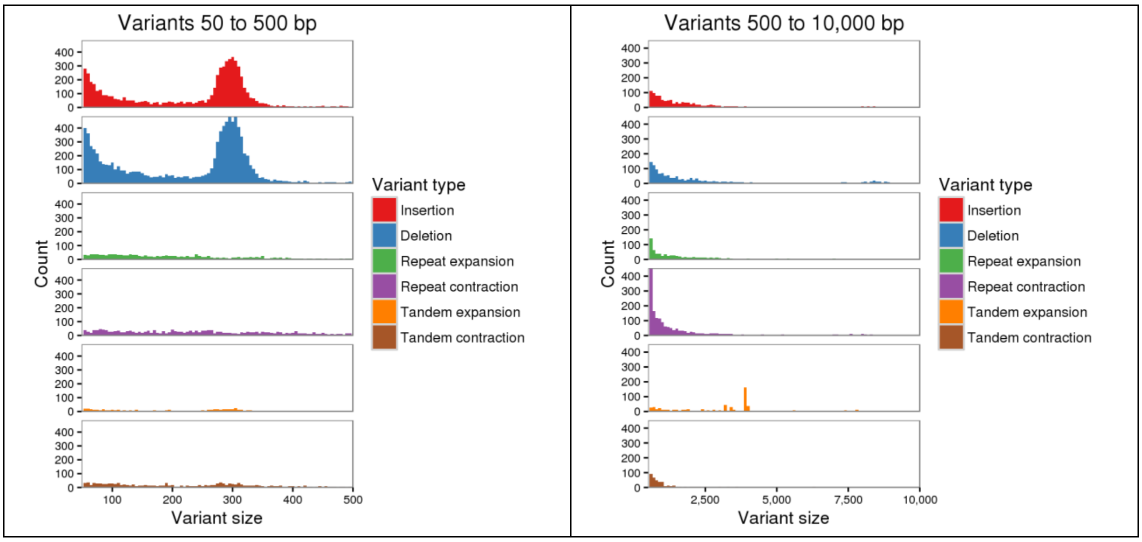
In all the pairwise comparisons amongst the former Sscrofa10.2 assembly and the new Sscrofa11.1 an USMARCv1.0 assemblies there is a peak of insertions and deletion with sizes of about 300 bp (Supplementary Figures SF6a-c). We assume that these correspond to SINE elements. Despite the fact that the Sscrofa10.2 and Sscrofa11.1 assemblies are representations of the same pig genome, there are many more differences between these assemblies than between the Sscrofa11.1 and USMARCv1.0 assemblies. We conclude that many of the differences between the Sscrofa11.1 assembly and the earlier Sscrofa10.2 assemblies represent improvements in the former. Some of the differences may indicate local differences in terms of which of the two haploid genomes has been captured in the assembly. The differences between the Sscrofa11.1 and USMARCv1.0 will represent a mix of true structural differences and assembly errors that will require further research to resolve.



Supplementary Figure SF6a: Assemblytics comparison of Sscrofa11.1 (query) against the Sscrofa10.2 (reference) i). (left hand panel) variants from 50 to 500 bp; ii). (right hand panel) variants from 500 to 10,000 bp.



Supplementary Figure SF6b: Assemblytics comparison of USMARCv1.0 (query) against the Sscrofa10.2 (reference) i). (left hand panel) variants from 50 to 500 bp; ii). (right hand panel) variants from 500 to 10,000 bp.



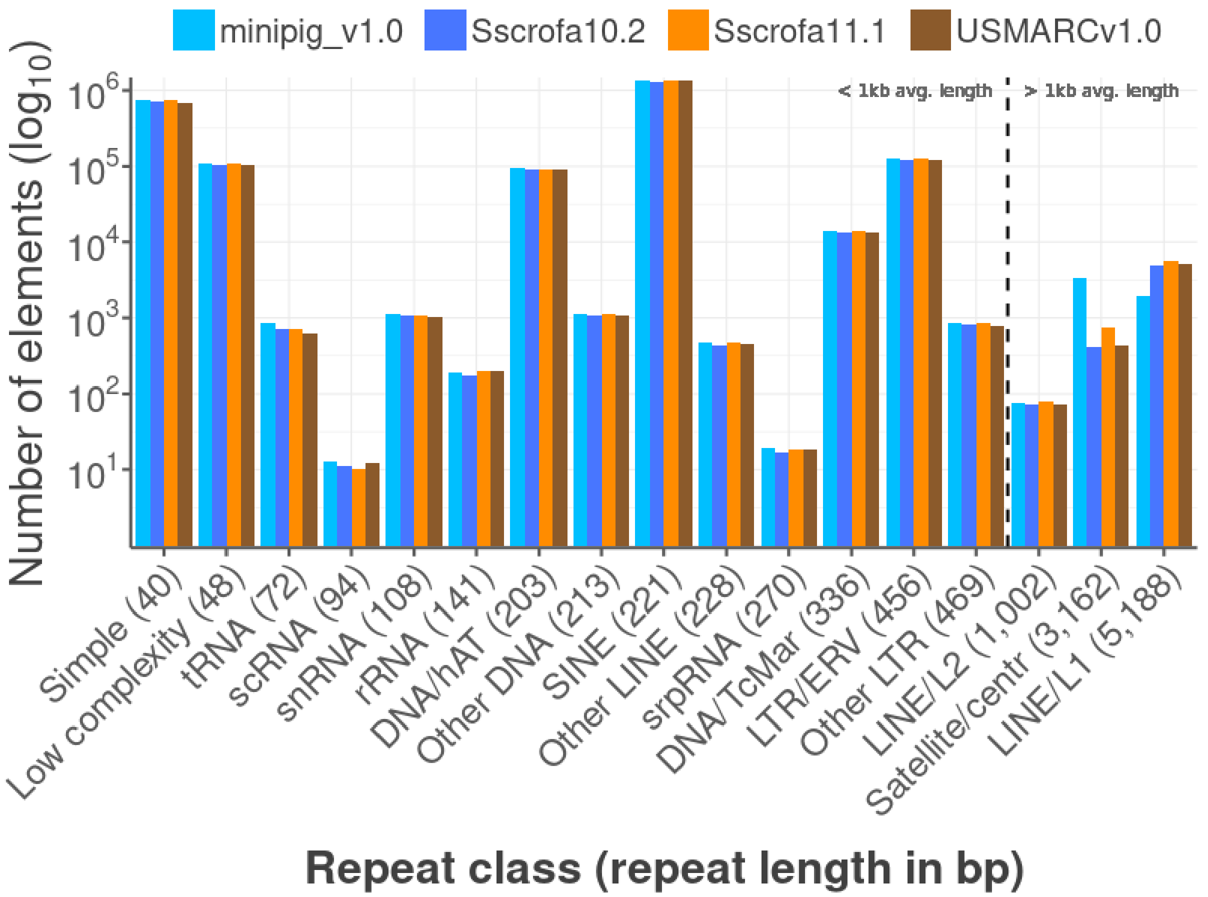
Supplementary Figure SF6c: Assemblytics comparison of USMARCv1.0 (query) against the Sscrofa11.1 (reference) i). (left hand panel) variants from 50 to 500 bp; ii). (right hand panel) variants from 500 to 10,000 bp.

2. Analyses

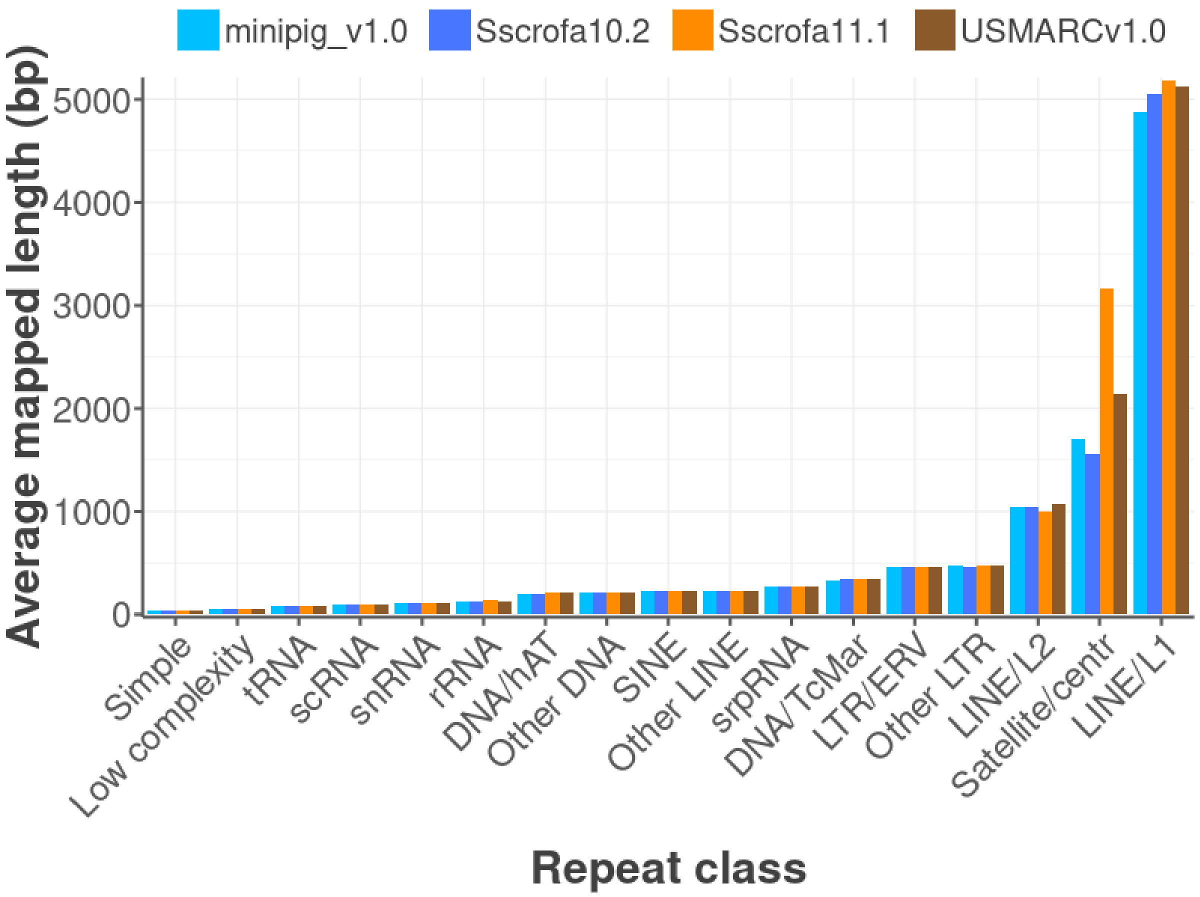
2.1 Repeat analysis

Repeats were identified using RepeatMasker (v.4.0.7) (Smit et al. 2013) with a combined repeat database including Dfam (v.20170127) (Hubley *et al.*, 2016) and RepBase (v.20170127) (Bao, Kojima and Kohany, 2015) on the minipig_v1.0, Sscrofa10.2, Sscrofa11.1 and USMARCv1.0 assemblies. RepeatMasker was run with "sensitive" (-s) setting using sus scrofa as the query species (-- species "sus scrofa"). Repeats which showed greater than 40% sequence divergence or were shorter than 70% of the expected sequence length were filtered out from subsequent analyses. The presence of potentially novel repeats was assessed by RepeatMasker using the novel repeat library generated by RepeatModeler (v.1.0.11) (Smit and Hubley, 2008).

The numbers of the different repeat classes and the average mapped lengths of the repetitive elements identified in these four pig genome assemblies are summarised in Supplementary Figures SF7 and SF8 respectively.



Supplementary Figure SF7: Counts of repetitive elements in four pig assemblies. Counts are given for repeat classes for which percent divergence was less than 40% and mapped length was above 70% relative to the RepBase database entries.



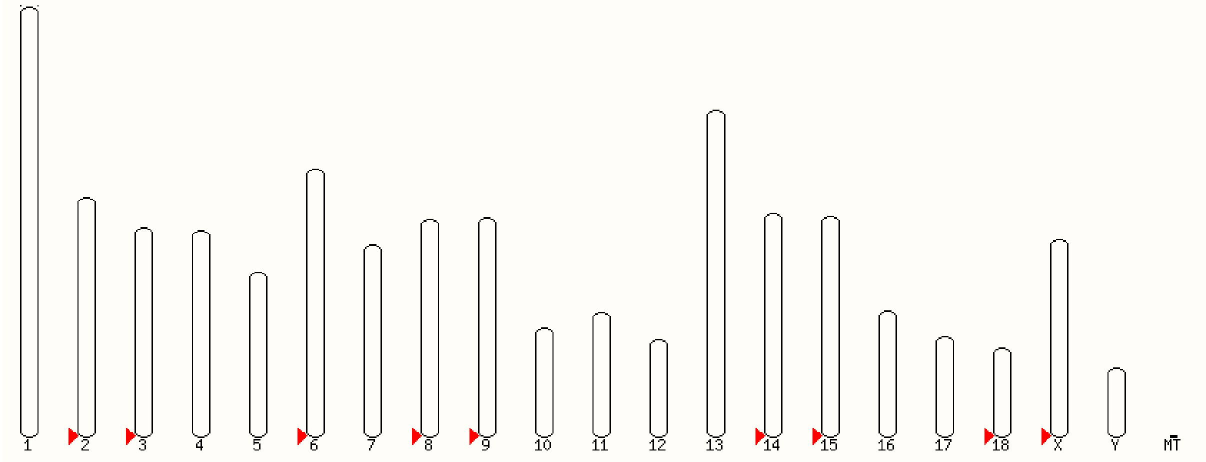
Supplementary Figure SF8: Average mapped length of repetitive elements in four pig genomes.

2.1.1 Telomeres

Telomeres were identified by running Tandem Repeat Finder (TRF) (Benson, 1999) with default parameters apart from Mismatch (5) and Minscore (40). The identified repeat sequences were then searched for the occurrence of five identical, consecutive units of the TTAGGG vertebrate motif or its reverse complement and total occurrences of this motif was counted within the tandem repeat. Regions which contained at least 200 identical hexamer units, were >2kb of length and had a hexamer density of >0.5 were retained as potential telomeres (Supplementary Table ST5; Supplementary Figure SF9). As chromosomes SSC1-SSC12 inclusive are metacentric we would have expected to identify telomeric sequences on the short arms of these chromosomes.

Supplementary Table ST5: Predicted telomere locations in the Sscrofa11.1 assembly. Number of exact matches of the vertebrate TTAGGG repeat sequence was used to identify candidate telomeres.

Chr	Start	End	Number of hexamers	Region length (kb)	Strand	Hexamer content
2	151,924,806	151,935,981	1,609	11.2	+	86.4%
3	132,840,959	132,848,913	1,046	8.0	+	78.9%
6	170,835,933	170,843,587	957	7.7	+	75.0%
8	138,963,948	138,966,197	208	2.2	+	55.5%
9	139,499,115	139,512,083	1,836	13.0	+	84.9%
14	141,745,369	141,755,446	1,201	10.1	+	71.5%
15	140,408,314	140,412,713	595	4.4	+	81.2%
18	55,971,782	55,982,971	1,571	11.2	+	84.2%
X	125,929,106	125,939,592	1,329	10.5	+	76.0%



Supplementary Figure SF9: Predicted locations of telomeres in the Sscrfoa11.1 assembly

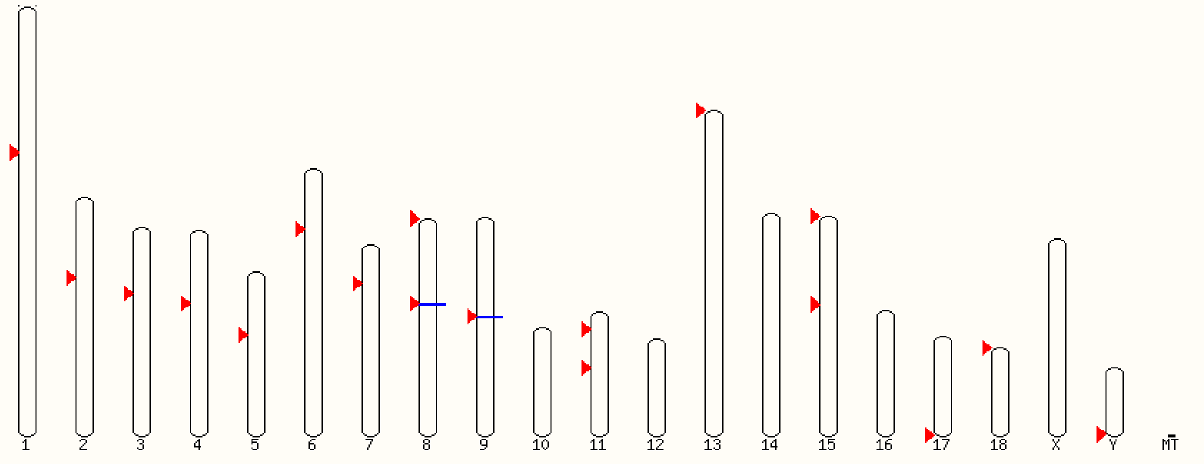
2.1.2 Centromeres

Centromeres were predicted using the following strategy. First, the RepeatMasker output, both default and novel, was searched for centromeric repeat occurrences. Second, the assemblies were searched for known, experimentally verified, centromere specific repeats (Miller, Hindkjær and Thomsen, 1993) (Riquet et al., 1996) in the Sscrofa11.1 genome. Then the three sets of repeat annotations were merged together with BEDTools (Quinlan and Hall, 2010) (median and mean length: 786 bp and 5775 bp, respectively) and putative centromeric regions closer than 500 bp were collapsed into longer super-regions. Regions which were

>5kb were retained as potential centromeric sites (Supplementary Table ST6; Supplementary Figure SF10).

Supplementary Table ST6: Predicted centromere locations in the Sscrofa11.1 assembly

Chr	Start	End	Repeat content (bp)	Region length (bp)	Repeat content
1	92,615,481	92,672,216	46,164	56,735	81.4%
1	92,760,768	92,881,119	110,990	120,351	92.2%
1	93,266,464	93,430,514	80,940	16,4050	49.3%
2	50,550,173	50,777,308	198,336	227,135	87.3%
3	41,776,737	41,860,603	35,376	83,866	42.2%
4	46,443,460	46,472,085	28,625	28,625	100.0%
5	39,774,025	39,828,563	54,538	54,538	100.0%
5	39,878,566	40,207,105	328,539	328,539	100.0%
6	38,712,705	38,886,534	163,335	173,829	94.0%
7	24,578,125	24,606,761	28,636	28,636	100.0%
8	144	20,905	20,761	20,761	100.0%
8	54,585,508	54,685,241	21,099	99,733	21.2%
9	63,144,551	63,503,859	356,770	359,308	99.3%
11	11,220,831	11,222,126	1,295	1,295	100.0%
11	35,726,738	35,728,355	1,617	1,617	100.0%
11	35,804,210	35,809,503	5,293	5,293	100.0%
11	35,870,705	35,878,206	7,501	7,501	100.0%
13	34	152,474	150,375	152,440	98.6%
15	1,649	36,105	10,369	34,456	30.1%
15	56,407,100	56,427,869	9,798	20,769	47.2%
17	63,189,675	63,361,433	171,758	171,758	100.0%
18	619	17,212	16,593	16,593	100.0%
Y	42,496,777	42,515,903	17,954	19,126	93.9%



Supplementary Figure SF10: Predicted centromere locations in the Sscrofa11.1 assembly.

2.2 Transcriptome data used for building gene models

Two new sources of transcriptome sequence data were generated for use in building gene models as described below – Annotation (Ensembl). First, long read transcript data (Iso-Seq) were generated on the Pacific Bioscience RSII platform. Second, short read Illumina RNA-Seq data.

2.2.1 Iso-Seq

The following tissues were harvested from MARC1423004 at age 48 days: brain (BioSamples: SAMN05952594), diaphragm (SAMN05952614), hypothalamus (SAMN05952595), liver (SAMN05952612), small intestine (SAMN05952615), skeletal muscle – *longissimus dorsi* (SAMN05952593), spleen (SAMN05952596), pituitary (SAMN05952626) and thymus (SAMN05952613).

Total RNA from each of these tissues was extracted using Trizol reagent (ThermoFisher Scientific) and the provided protocol. Briefly, approximately 100 mg of tissue was ground in a mortar and pestle cooled with liquid nitrogen, and the powder was transferred to a tube with 1 ml of Trizol reagent added and mixed by vortexing. After 5 minutes at room temperature, 0.2 mL of chloroform was added and the mixture was shaken for 15 seconds and left to stand another 3 minutes at room temperature. The tube was centrifuged at 12,000 x g for 15 minutes at 4°C. The RNA was precipitated from the aqueous phase with 0.5 mL of isopropanol. The RNA was further purified with extended DNase I digestion to remove potential DNA contamination. The RNA quality was assessed with a Fragment Analyzer (Advanced Analytical Technologies Inc., IA). Only RNA samples of RQN above 7.0 were used for library construction. PacBio IsoSeq libraries were constructed per the PacBio IsoSeq protocol. Briefly, starting with 3 µg of total RNA, cDNA was synthesized by using SMARTer PCR cDNA Synthesis Kit (Clontech, CA) according to the IsoSeq protocol (Pacific Biosciences, CA). Then the cDNA was amplified using KAPA HiFi DNA Polymerase (KAPA Biotechnologies) for 10 or 12 cycles followed by purification and size selection into 4 fractions: 0.8-2 kb, 2-3 kb, 3-5 kb and >5 kb. The fragment size distribution was validated on a Fragment Analyzer (Advanced Analytical Technologies Inc, IA) and quantified on a DS-11

FX fluorometer (DeNovix, DE). After a second round of large scale PCR amplification and end repair, SMRT bell adapters were separately ligated to the cDNA fragments. Each size fraction was sequenced on 4 or 5 SMRT Cells v3 using P6-C4 chemistry and 6 hour movies on a PacBio RS II sequencer (Pacific Bioscience, CA). Short read RNA-Seq libraries were also prepared for all nine tissue using TruSeq stranded mRNA LT kits and supplied protocol (Illumina, CA), and sequenced on a NextSeq500 platform using v2 sequencing chemistry to generate 2 x 75 bp paired-end reads.

2.2.1.1 Error-correction and redundancy reduction of PacBio Iso-Seq full-length cDNA reads

The Read of Insert (ROI) were determined by using *consensustools.sh* in the SMRT-Analysis pipeline v2.0, with reads which were shorter than 300 bp and whose predicted accuracy was lower than 75% removed. Full-length, non-concatemer (FLNC) reads were identified by running the classify.py command. The cDNA primer sequences as well as the poly(A) tails were trimmed prior to further analysis. Paired-end Illumina RNA-Seq reads from each tissue sample were trimmed to remove the adaptor sequences and low-quality bases using Trimmomatic (v0.32) (Bolger, Lohse and Usadel, 2014) with explicit option settings: *ILLUMINACLIP:adapters.fa: 2:30:10:1:true LEADING:3 TRAILING:3 SLIDINGWINDOW: 4:20 LEADING:3 TRAILING:3 MINLEN:25*, and overlapping paired-end reads were merged using the PEAR software (v0.9.6) (Zhang *et al.*, 2014). Subsequently, the merged and unmerged RNA-Seq reads from the same tissue samples were *in silico* normalized in a mode for single-end reads by using a Trinity (v2.1.1) (Grabherr *et al.*, 2011) utility*, insilico_read_normalization.pl*, with the following settings*: --max_cov 50 --max_pct_stdev 100 --single*. Errors in the full-length, non-concatemer reads were corrected with the preprocessed RNA-Seq reads from the same tissue samples by using proovread (v2.12) (Hackl *et al.*, 2014). Untrimmed sequences with at least some regions of high accuracy in the *.trimmed.fq* files were extracted based on sequence IDs in *.untrimmed.fa* files to balance off the contiguity and accuracy of the final reads.

2.2.2 RNA-Seq

In addition to the Illumina short read RNA-seq data generated from MARC1423004 and used to correct the Iso-Seq data (see above), Illumina short read RNA-seq data (PRJEB19386) were also generated from a range of tissues from four juvenile Duroc pigs (two male, two female) and used for annotation as described below. Extensive metadata with links to the protocols for sample collection and processing are linked to the BioSample entries under the Study Accession PRJEB19386. The tissues sampled are listed in Supplementary Table ST7. Sequencing libraries were prepared using a ribodepletion TruSeq stranded RNA protocol and 150 bp paired end sequences generated on the Illumina HiSeq 2500 platform in rapid mode.

Supplementary Table ST7: Tissue samples characterised by Illumina short read RNA-Seq analyses

Tissue	**BioSample accession**	**alias**	**Animal**	**Sex**
alveolar macrophages	SAMEA103886124	SUS_RI_DUR21-30	Duroc 21	female
alveolar macrophages	SAMEA103886168	SUS_RI_Pig 21_DUR_30	Duroc 21	female
alveolar macrophages	SAMEA103886137	SUS_RI_DUR22-60	Duroc 22	male
alveolar macrophages	SAMEA103886112	SUS_RI_Pig 22_DUR_60	Duroc 22	male
amygdala	SAMEA103886173	SUS_RI_R-Dur_23-08	Duroc 23	female
amygdala	SAMEA103886162	SUS_RI_Dur_24-C-S0	Duroc 24	male
brain, frontal lobe	SAMEA103886139	SUS_RI_R-Dur_23-01	Duroc 23	female
brain, frontal lobe	SAMEA103886156	SUS_RI_R-Dur_24-41	Duroc 24	male
brain stem	SAMEA103886128	SUS_RI_R-Dur_23-05	Duroc 23	female
brain stem	SAMEA103886129	SUS_RI_R-Dur_24-45	Duroc 24	male
caecum	SAMEA103886133	SUS_RI_DUR21-19	Duroc 21	female
caecum	SAMEA103886120	SUS_RI_DUR22-48	Duroc 22	male
caecum	SAMEA103886151	SUS_RI_Pig 22_DUR_48	Duroc 22	male
cerebellum	SAMEA103886116	SUS_RI_R-Dur_23-09	Duroc 23	female
cerebellum	SAMEA103886131	SUS_RI_R-Dur_24-49	Duroc 24	male
colon	SAMEA103886132	SUS_RI_Dur_23-21	Duroc 23	female
colon	SAMEA103886147	SUS_RI_Dur_24-61	Duroc 24	male
corpus callosum	SAMEA103886154	SUS_RI_R-Dur_23-10	Duroc 23	female
corpus callosum	SAMEA103886167	SUS_RI_R-Dur_24-50	Duroc 24	male
duodenum	SAMEA103886155	SUS_RI_Dur_23-22	Duroc 23	female

Tissue	BioSample accession	alias	Animal	Sex
duodenum	SAMEA103886176	SUS_RI_Dur_24-62	Duroc 24	male
epididymis	SAMEA103886140	SUS_RI_DUR22-58	Duroc 22	male
hippocampus	SAMEA103886122	SUS_RI_Dur_23-B-S0	Duroc 23	female
hippocampus	SAMEA103886114	SUS_RI_R-Dur_24-51	Duroc 24	male
ileum	SAMEA103886163	SUS_RI_Dur_23-23	Duroc 23	female
ileum	SAMEA103886121	SUS_RI_Dur_24-63	Duroc 24	male
kidney cortex	SAMEA103886174	SUS_RI_DUR21-09	Duroc 21	female
kidney cortex	SAMEA103886153	SUS_RI_DUR22-39	Duroc 22	male
heart, left ventricle	SAMEA103886169	SUS_RI_DUR21-12	Duroc 21	female
heart, left ventricle	SAMEA103886172	SUS_RI_DUR22-43	Duroc 22	male
lymph node, mesenteric	SAMEA103886127	SUS_RI_DUR21-22	Duroc 21	female
lymph node, mesenteric	SAMEA103886115	SUS_RI_DUR22-51	Duroc 22	male
medulla oblongata	SAMEA103886135	SUS_RI_R-Dur_23-06	Duroc 23	female
medulla oblongata	SAMEA103886142	SUS_RI_R-Dur_24-46	Duroc 24	male
occipital lobe	SAMEA103886158	SUS_RI_R-Dur_23-02	Duroc 23	female
occipital lobe	SAMEA103886177	SUS_RI_R-Dur_24-42	Duroc 24	male
omentum	SAMEA103886145	SUS_RI_DUR21-65	Duroc 21	female
omentum	SAMEA103886146	SUS_RI_DUR22-73	Duroc 22	male
penis	SAMEA103886166	SUS_RI_DUR22-59	Duroc 22	male
pituitary gland	SAMEA103886152	SUS_RI_Dur_23-14	Duroc 23	female
pituitary gland	SAMEA103886150	SUS_RI_Dur_24-54	Duroc 24	male
pituitary gland	SAMEA103886149	SUS_RI_DUR21-06	Duroc 21	female
pons	SAMEA103886159	SUS_RI_R-Dur_23-07	Duroc 23	female
pons	SAMEA103886164	SUS_RI_R-Dur_24-47	Duroc 24	male
skeletal muscle	SAMEA103886171	SUS_RI_DUR21-24	Duroc 21	female
skeletal muscle	SAMEA103886118	SUS_RI_DUR22-75	Duroc 22	male
spleen	SAMEA103886157	SUS_RI_DUR21-25	Duroc 21	female
spleen	SAMEA103886170	SUS_RI_DUR22-55	Duroc 22	male
stomach	SAMEA103886111	SUS_RI_Dur_23-24	Duroc 23	female
stomach	SAMEA103886134	SUS_RI_Dur_24-64	Duroc 24	male
thalamus	SAMEA103886136	SUS_RI_R-Dur_23-13	Duroc 23	female
thalamus	SAMEA103886160	SUS_RI_R-Dur_24-53	Duroc 24	male
tonsils	SAMEA103886125	SUS_RI_DUR22-56	Duroc 22	male
uterus	SAMEA103886126	SUS_RI_DUR21-27	Duroc 21	female

2.3 SNP chip variants

2.3.1 SNP chip probes mapped to assemblies

The probes from four commercial SNP chips were mapped to the Sscrofa10.2, Sscrofa11.1 and USMARCv1.0 assemblies using BWA MEM (Li and Durbin, 2009) and a wrapper script (https://github.com/njdbickhart/perl_toolchain/blob/master/assembly_scripts/alignAndOrderSnpProbes.pl).

- Illumina PorcineSNP60 ((Ramos *et al.*, 2009), https://emea.illumina.com/products/by-type/microarray-kits/porcine-snp60.html)
- Affymetrix Axiom™ Porcine Genotyping Array (https://www.thermofisher.com/order/catalog/product/550588)
- Gene Seek Genomic Profiler Porcine – HD beadChip (http://genomics.neogen.com/uk/ggp-porcine)
- Gene Seek Genomic Profiler Porcine v2– LD Chip (http://genomics.neogen.com/uk/ggp-porcine)

Probe sequence was derived from the marker manifest files that are available on the provider websites. In order to retain marker manifest coordinate information, each probe marker name was annotated with the chromosome and position of the marker's variant site from the manifest file. All mapping coordinates were tabulated into a single file, and were sorted by the chromosome and position of the manifest marker site. In order to derive and compare relative marker rank order, a custom Perl script (https://github.com/njdbickhart/perl_toolchain/blob/master/assembly_scripts/pigGenomeSNPSortRankOrder.pl) was used to sort and number markers based on their mapping locations in each assembly.

A Spearman's rank order (rho) value was calculated for each assembly (alternative hypothesis: rho is equal to zero; $p < 2.2 \times 10^{-16}$) (Supplementary Table ST9). This rank order comparison was estimated by ordering all of the SNP probes from all chips by their listed manifest coordinates against their relative order in each assembly (with chromosomes ordered by karyotype). Any unmapped markers in an assembly were penalized by giving the marker a "-1" rank in the assembly ranking order. The methods are similar to what those used to assess the relative order of the ARS1 Goat assembly RH map vs the scaffold order ((Bickhart *et al.*, 2017) see Supplementary Note 1).

Supplementary Table ST8: SNP chip markers mapped to pig genome assemblies

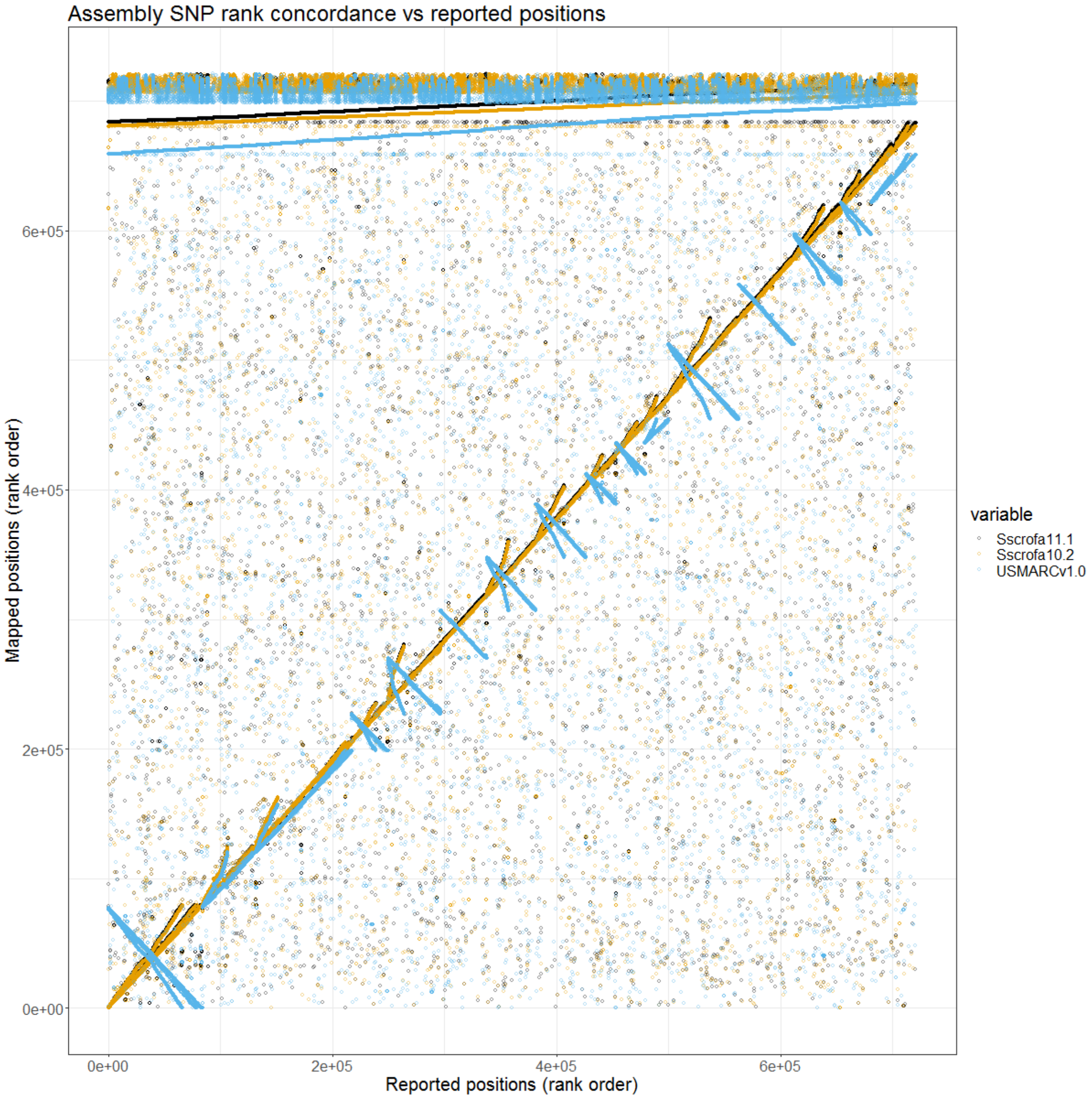
Assembly	Mapped / unmapped	AxiomHD	PorcineSNP60	GGP LD	80K
Sscrofa10.2	mapped	633,705	59,590	50,530	68,046
	unmapped	24,987	1,975	385	470
Sscrofa11.1	mapped	628,280	61,299	50,586	68,270
	ummapped	30,412	266	329	246
USMARCv1.0	mapped	618,771	60,692	50,042	67,604
	unmapped	39,921	873	873	912

Supplementary Table ST9: Spearman’s rank order

Assembly	Rho
Sscrofa10.2	0.88464
Sscrofa11.1	0.88890
USMARCv1.0	0.81260

In order to examine general linear order of placed markers on each assembly, the marker rank order (y axis; used above in the Spearman's rank order test) was plotted against the rank order of the probe rank order on the manifest file (x axis) (Supplementary Figure SF11).

Supplementary Figure SF11: Assembly SNP rank concordance versus reported chromosomal location



The analyses reveal some interesting artifacts that suggest that the SNP manifest coordinates for the porcine 60K SNP chip are still derived from an obsolete (Sscrofa9) reference in contrast to all other manifests (Sscrofa10.2). Also, it confirms that several of the USMARCv1.0 chromosome scaffolds are inverted with respect to the canonical orientation of pig chromosomes. Such inversions are due to the agnostic nature of genome assembly and post-assembly polishing programs. Unless these are corrected post-hoc by manual curation, they result in artefactual inversions of the entire chromosome. However, such inversions do not generally impact downstream analysis that does not involve the relative order/orientation of whole chromosomes. The large band of points at the top of the plot corresponds to marker mappings on the unplaced contigs of each assembly. These unplaced contigs often correspond to assemblies of alternative haplotypes in heterozygous regions of the reference animal (Koren *et al.*, 2018). Marker placement on these segments suggests that these variants are tracking different haplotypes in the population, which is the desired intent of genetic markers used in Genomic Selection.

3. Annotation (Ensembl)

3.1 Repeat Finding

After loading into a database, the Sscrofa11.1. genomic sequence was screened for sequence patterns, including repeats using RepeatMasker (Smit et al., 2013-5) (version 4.0.5) with parameters '-nolow –species "sus scrofa" –engine "crossmatch"', dustmasker (Camacho *et al.*, 2009) and TRF (Benson, 1999). Both executions of RepeatMasker and dustmasker combined masked 45.04% of the assembly.

3.2 Raw computes

Transcription start sites (TSS) were predicted using Eponine-scan (Down and Hubbard, 2002). CpG islands [Micklem, G., unpublished] longer than 400 bases and tRNAs (Lowe and Eddy, 1996) were also predicted. The results of Eponine-scan, CpG and tRNAscan are for display purposes only and are not used subsequently in the gene annotation process.

Genscan (Burge and Karlin, 1997) was run across the repeat-masked sequence and the results were used as input for UniProt (Goujon *et al.*, 2010), UniGene (Sayers *et al.*, 2010) and Vertebrate RNA (www.ebi.ac.uk/ena/) alignments by BLAST+ (Camacho *et al.*, 2009). Passing only Genscan results to BLAST is an effective way of reducing the search space and therefore the computational resources required. The resulting alignments to the Sscrofa11.1 assembly included 5,680,769 UniProt, 4,801,230 UniGene and 4,414,040 Vertebrate RNA sequences.

3.3 Generation of gene models

Various sources of transcript and protein data were investigated and used to generate gene models using a variety of techniques and are outlined here. The number of gene models generated are summarised in Table ST10.

Table ST10: Gene model generation overview

Pipeline	**Source**	**Number of models**
Species specific cDNAs	RefSeq, ENA	45,589
PacBio Iso-Seq	USDA MARC	326,217
RNA-Seq	The Roslin Institute	572,419
Olfactory receptors	Human and mouse Ensembl Release 89	1,212
IG/TR genes	IMGT®	1,803
Protein-to-genome	Subset of UniProt vertebrate proteins	509,769

3.3.1 cDNA alignments

Pig cDNAs were downloaded from ENA and RefSeq, and aligned to the Sscrofa11.1 assembly using Exonerate (Slater and Birney, 2005). A minimal sequence length of 60 bp was used and a cut-off of 97% identity and 90% coverage were required for an alignment to be processed further. The cDNAs are mainly used for display purposes, but can be used to add untranslated regions (UTRs) to the protein coding transcript models if they have matching introns.

Table ST11: Species specific cDNAs aligned against Sscrofa11.1

Species	Initial mRNA sequences	Sequences aligned
Pig	45,571	45,526

3.3.2 PacBio Iso-Seq transcript data

PacBio Iso-Seq data are high coverage long read transcriptomic data that allows for correction for the high error rate in raw PacBio reads. The consensus sequences representing nine tissues (brain, diaphragm, hypothalamus, liver, skeletal muscle (*longissimus dorsi*), pituitary, small intestine, spleen, and thymus were downloaded from the short read archive (SRA: PRJNA351265) after correction using Illumina short reads from the same tissue type. The sequences were aligned to the genome using Exonerate (Slater and Birney, 2005) using a cut-off of 95% identity and 90% coverage. All the Iso-Seq data sets had 3' capping and were used for adding UTRs to homology-based protein coding models. All Iso-Seq data sets were used as lincRNA candidates for our lincRNA prediction pipeline.

Table ST12: PacBio Iso-Seq sequences aligned against Sscrofa11.1

Tissue sample	Initial Iso-Seq sequences	Aligned sequences
Liver	588,957	491,796
Thymus	567,700	374,515
Hypothalamus	414,021	256,930
Brain	398,629	354,494
Skeletal muscle (*l. dorsi*)	410,420	361,494
Diaphragm	459,911	391,813
Spleen	674,053	449,425
Pituitary	411,562	252,707
Small intestine	494,538	406,144

3.3.3 Protein-to-genome alignment

Protein sequences were downloaded from UniProt and aligned to the Sscrofa11.1 assembly in a splice aware manner using GenBlast (She *et al.*, 2011). The set of proteins aligned to the genome was a subset of UniProt proteins used to provide a broad targeted coverage of the pig genome. The set consisted of the following:

- Pig PE level 1, 2, 3
- Human PE level 1, 2, 3
- Mouse PE level 1, 2, 3
- Other mammals PE level 1, 2, 3
- Other vertebrates PE level 1, 2, 3

Note: PE level = protein existence levelA cut-off of 50 percent coverage and identity and an e-value of e-1 were used for GenBlast (She *et al.*, 2011) with the exon repair option turned on. The top 5 transcript models built by GenBlast for each protein passing the cut-offs were kept. This process produced 509,769 transcript models in total.

3.3.4 RNA-seq pipeline

RNA-Seq data downloaded from ENA PRJEB19386 were used in the annotation. These RNA-Seq data consisted of 150 bp paired end reads from libraries prepared using a stranded library protocol from ribo-depleted total RNA from Duroc pigs. The dataset comprised RNA-Seq data from 28 tissue and cell samples: alveolar macrophages, amygdala, brain stem, caecum, cerebellum, colon, corpus callosum, duodenum, epididymis, frontal lobe (brain), hippocampus, ileum, kidney cortex, left ventricle (heart), mesenteric lymph node, medulla oblongata, occipital lobe, omentum, penis, pituitary gland, pons, skeletal muscle, spleen, stomach, thalamus, tonsil, uterus (Supplementary Table ST7). A merged file containing reads from all tissues was also created. The merged data was less likely to suffer from model fragmentation due to read depth. The available reads were aligned to the Sscrofa11.1 assembly using BWA. A 50 percent allowed mismatch criteria was applied to identify potential splice junctions. Initial rough exon/intron boundaries were generated via the BWA alignments and then refined by mapping the reads in a splice-aware manner using Exonerate (Slater and Birney, 2005). The split reads and the processed BWA alignments were combined to produce 1,060,366 transcript models in total. The predicted open reading frames were compared to UniProt proteins using NCBI BLAST. Models with poorly scoring or no BLAST alignments were split into a separate class and considered as potential lincRNAs.

Supplementary Table ST13: Tissue-specific values for initial read counts along with the percent of mapped and properly paired reads. The final column shows the count of potential transcript models build per tissue.

Tissue name	Total reads	Mapped	Properly paired	Transcript models
Alveolar macrophages	508,512,918	92.69%	64.31%	34,867
Amygdala	170,434,766	93.73%	64.43%	38,118
Brain stem	124,538,342	93.33%	61.60%	35,791
Caecum	444,611,528	92.40%	71.78%	40,716
Cerebellum	158,560,324	94.09%	64.42%	36,132
Colon	168,263,230	90.80%	61.79%	34,520
Corpus callosum	148,039,874	93.75%	62.68%	37,474
Duodenum	346,909,970	91.94%	62.08%	40,112
Epididymis	186,743,514	92.74%	69.27%	37,377
Frontal lobe	119212918	94.30%	59.99%	35,119
Hippocampus	164,637,176	94.72%	62.38%	36,403
Ileum	166,645,682	91.96%	69.83%	36,661
Kidney cortex	258,616,430	95.30%	86.38%	35,544
Left ventricle	265,075,268	95.33%	86.11%	33,125
Mesenteric lymph node	448,893,104	93.24%	69.37%	40,250
Medulla oblongata	141,361,800	93.18%	58.96%	42,716
Occipital lobe	13,3884,172	94.23%	64.25%	35,390
Omentum	179,713,086	93.70%	84.61%	27,570
Penis	179,834,564	93.15%	71.84%	37,121
Pituitary	164,402,132	93.64%	61.23%	35,482
Pituitary gland	131,196,396	95.15%	86.44%	33,800
Pons	134,913,426	93.80%	61.90%	35,974
Skeletal muscle	206,977,278	92.09%	81.55%	32,011
Spleen	194,924,210	94.26%	83.52%	35,130
Stomach	141,172,602	92.49%	70.72%	33,326
Thalamus	149,227,654	93.84%	53.67%	36,047
Tonsil	320,766,440	94.22%	74.78%	38,154
Uterus	90,381,988	94.56%	59.49%	31,178

3.3.5 IG and TR genes

All pig, cow, sheep, human and mouse IG/TR V, C and J segment protein sequences were downloaded from IMGT® (Lefranc *et al.*, 2015) and aligned against the Sscrofa11.1 assembly using Exonerate (Slater and Birney, 2005) using '—max-intron 50000' and only the models with 95% coverage and 80% identity were kept. We generated 1,803 gene models. For positions where there were overlapping transcript models, the transcript model with the highest combined alignment coverage and percent identity was kept as the representative model for the locus.

3.3.6 Olfactory receptor genes

We used the manually curated human and mouse set (Ensembl release 89) and pig olfactory receptor sequences (Nguyen *et al.*, 2012). The sequences were aligned against the genome with Exonerate (Slater and Birney, 2005) and only the models with high similarity (95% coverage, 95% identity) were kept, yielding 1,212 gene models.

3.3.7 Selenocysteine proteins

Known selenocysteine proteins were aligned against the Sscrofa11.1 assembly using Exonerate (Slater and Birney, 2005). The models generated were checked for the presence of selenocysteines in the same positions as the known proteins. We generated 103 models.

3.3.8 Filtering the models

The filtering phase decided the subset of protein-coding transcript models, generated from the model-building pipelines, that would comprise the final protein-coding gene set in the GeneBuild. Models were filtered based on information such as what pipeline they were generated using, how closely related the data are to the target species (i.e. pig) and how good the alignment coverage and percent identity to the original data are. Models were filtered using the LayerAnnotation and GeneBuilder modules. The Apollo software (Lewis *et al.*, 2002) was used to visualise the results of the filtering.

3.3.9 Collapsing the transcript set

The LayerAnnotation module was used to define a hierarchy of input data sets, from most preferred to least preferred. The output of this pipeline included all transcript models form the highest ranked input set. Models from the lower ranked input sets are included only if their exons do not overlap a model from an input set higher in the hierarchy. Note that models cannot exist in more than one layer. For UniProt proteins, models were also separated into clades. To help selection during the layering process. Each UniProt protein was in one clade only, for example mammal proteins were present in the mammal clade and were not present in the vertebrate clade to avoid aligning the proteins multiple times.

Layer 1:

- Pig seleno-proteins
- Pig olfactory receptors with >= 90% coverage and 97% identity
- All vertebrate seleno-proteins with full RNA-seq support
- IG and TR genes

Layer 2:

- Pig cDNA models with >= 90% coverage and 97% identity
- Pig IsoSeq models with protein support >= 80% coverage and identity and full RNA-seq support
- RNA-seq models with >=95% coverage and identity
- Pig curated UniProt proteins from PE levels 1 & 2 with >=80% coverage and identity and full RNA-seq support
- Pig curated UniProt proteins from PE levels 3 with >=95% coverage and identity and full RNA-seq support
- All vertebrate curated UniProt proteins from PE levels 1 & 2 with >=95% coverage and identity and full RNA-seq support

Layer 3:

- RNA-seq models with >=80% coverage and identity

Layer 4:

- Pig curated UniProt proteins from PE levels 1 & 2 with >=50% coverage and identity
- Pig IsoSeq models with protein support >= 80% coverage and identity

Layer 5:

- Pig curated UniProt proteins from PE levels 3 with >=80% coverage and identity
- All vertebrate curated UniProt proteins from PE levels 1 & 2 with >=80% coverage and identity

Layer 6:

- RNA-seq models with >= 50% coverage and identity
- Pig IsoSeq models with protein support >= 50% coverage and identity
- Pig curated UniProt proteins from PE levels 3 with >=50% coverage and identity
- All vertebrate curated UniProt proteins from PE levels 1 & 2 with >=50% coverage and identity

Layer 7:

- Pig UniProt proteins from PE levels 1 & 2 & 3 with >=80% coverage and identity and full RNA-seq support
- All vertebrate UniProt proteins from PE levels 1 & 2 with >=80% coverage and identity and full RNA-seq support

Layer 8:

- Pig UniProt proteins from PE levels 1 & 2 & 3 with >=50% coverage and identity and full RNA-seq support
- All vertebrate UniProt proteins from PE levels 1 & 2 with >=50% coverage and identity and full RNA-seq support
- Pig IsoSeq models with protein support >= 50% coverage and identity which may have retained an intron

Layer 9:

- Pig UniProt proteins from PE levels 1 & 2 & 3 with >=80% coverage and identity
- All vertebrate UniProt proteins from PE levels 1 & 2 with >=80% coverage and identity

Layer 10:

- Pig UniProt proteins from PE levels 1 & 2 & 3 with >=50% coverage and identity
- All vertebrate UniProt proteins from PE levels 1 & 2 with >=50% coverage and identity

3.3.10 Addition of UTR to coding models

The set of coding models was extended into the untranslated regions (UTRs) using RNA-seq, cDNA and Iso-Seq sequences. The source of the UTRS was prioritised with UTR coming from cDNAs and Iso-Seq, then RNA-seq.

3.3.11 Generating multi-transcript genes

The steps described above generated a large set of potential transcript models, many of which overlapped one another. Redundant transcript models were collapsed and the remaining unique set of transcript models were clustered into multi-transcript gene where each transcript in a gene has at least one coding exon that overlaps a coding exon from another transcript within the same genes.

At this stage the gene set comprised 23,025 genes with 46,511 transcripts.

3.3.12 Pseudogenes

The Pseuodgene module was run to identify pseudogenes from within the set of gene models. A total of 178 genes were labelled as pseudogenes or processed pseudogenes.

3.3.13 Small ncRNAs

Small structured non-coding genes were added using annotations taken from RFAM (Griffiths-Jones *et al.*, 2003) and miRBase (Griffiths-Jones *et al.*, 2006). BLAST+ was run for these sequences and models built using the Infernal software suite (Eddy, 2002).

3.3.14 lincRNAs discovery

Using the transcriptomic data set, we tried to predict long intergenic non-coding RNAs (lincRNAs). We used the RNA-seq and Iso-Seq data which were filtered against the protein-coding gene set. Candidate lincRNAs that overlapped a protein-coding gene were discarded. The Pfam analysis of InterProScan was run against the filtered gene set. Candidate lincRNAs with a Pfam domain were also discarded.

3.3.15 Cross-referencing and stable identifiers

Before public release the transcripts and translations were given external references (cross-references to external databases). Stable identifiers were assigned to each gene, transcript, exon and translation. As earlier pig genome sequences have been annotated by Ensembl previously a comparison was made to the previous gene set and as many stable identifiers as possible were mapped between the two annotations.

3.3.16 Gene expression

The Illumina RNA-Seq data (Supplementary Table ST7) were also processed by the EBI Gene Expression Atlas (GXA) team (Papatheodorou *et al.*, 2018) (https://www.ebi.ac.uk/gxa/home) to generate a baseline gene expression atlas (Expression Atlas release 25, August 2017). These gene expression data can be visualised in the Ensembl genome browser from the gene page.

3.3.17 Comparison of Ensembl and NCBI annotation

The Sscrofa11.1 assembly was also annotated independently by the NCBI (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Sus_scrofa/106/). We have compared these two annotations (Supplementary Table ST14).

Supplementary Table ST14: Comparison of Ensembl and NCBI annotation of Sscofa11.1

Ensembl		**NCBI**			
		missing (relative location)			
		in common	(intragenic)	(intergenic)	other
Protein-coding	22,452	18,772	270	1,785	[*]1,625
Non-coding	3,250	811	1,158	1,281	
Pseudogenes	178	121	1	56	
NCBI		**Ensembl**			
		missing (relative location)			
		in common	(intragenic)	(intergenic)	other
Protein-coding	20,790	18,772	119	1,899	
Non-coding	6,460	811	541	3,730	[**]1,378
Pseudogenes	3,084	121	124	1,214	[*]1,625

[*] 1,625 genes annotated as protein-coding by Ensembl are annotated as pseudogenes by NCBI

[**] 1,378 genes annotated as non-coding by NCBI are annotated as protein-coding by Ensembl

3.3.18 Annotation of the USMARCv1.0 assembly

Annotation for USMARCv1.0 was carried out using the Ensembl pipeline and the same key steps as outlined for Sscrofa11.1. To help with the consistency of annotation, the same set of long and short read transcriptomic data were used in the annotation of USMARCv1.0. As the annotations were done two years apart there was some variance in terms of the underlying code base used to generate the annotations. We plan to update the Sscrofa11.1 annotation in future to take advantage of these upgrades, though the effect on the overall geneset is likely to be marginal due to the amount of high quality transcriptomic data available for the original annotation.

4. References

Anderson, S. I. *et al.* (2000) 'A large-fragment porcine genomic library resource in a BAC vector', *Mammalian Genome*, 11(9), pp. 811–814. doi: 10.1007/s003350010155.

Bao, W., Kojima, K. K. and Kohany, O. (2015) 'Repbase Update, a database of repetitive elements in eukaryotic genomes', *Mobile DNA*, 6(1). doi: 10.1186/s13100-015-0041-9.

Benson, G. (1999) 'Tandem repeats finder: A program to analyze DNA sequences', *Nucleic Acids Research*, 27(2), pp. 573–580. doi: 10.1093/nar/27.2.573.

Bickhart, D. M. *et al.* (2017) 'Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome', *Nature Genetics*, 49(4), pp. 643-650. doi: 10.1038/ng.3802.

Bolger, A. M., Lohse, M. and Usadel, B. (2014) 'Trimmomatic: A flexible trimmer for Illumina sequence data', *Bioinformatics*, 30(15), pp. 2114–2120. doi: 10.1093/bioinformatics/btu170.

Burge, C. and Karlin, S. (1997) 'Prediction of complete gene structures in human genomic DNA', *Journal of Molecular Biology*, 268(1), pp. 78–94. doi: 10.1006/jmbi.1997.0951.

Camacho, C. *et al.* (2009) 'BLAST+: Architecture and applications', *BMC Bioinformatics*, 10: 421. doi: 10.1186/1471-2105-10-421.

Chin, C. S. *et al.* (2013) 'Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data', *Nature Methods*, 10(6), pp. 563–569. doi: 10.1038/nmeth.2474.

Down, T. A. and Hubbard, T. J. P. (2002) 'Computational detection and location of transcription start sites in mammalian genomic DNA', *Genome Research*, 12(3), pp. 458–461. doi: 10.1101/gr.216102.

Eddy, S. R. (2002) 'A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure', *BMC Bioinformatics*, 3: 18. doi: 10.1186/1471-2105-3-18.

English, A. C. *et al.* (2012) 'Mind the Gap: Upgrading Genomes with Pacific Biosciences RS

Long-Read Sequencing Technology', *PLoS ONE*, 7(11): e47768. doi: 10.1371/journal.pone.0047768.

Goujon, M. *et al.* (2010) 'A new bioinformatics analysis tools framework at EMBL-EBI', *Nucleic Acids Research*, 38(SUPPL. 2): W695-699. doi: 10.1093/nar/gkq313.

Grabherr, M. G. *et al.* (2011) 'Full-length transcriptome assembly from RNA-Seq data without a reference genome', *Nature Biotechnology*, 29(7), pp. 644–652. doi: 10.1038/nbt.1883.

Griffiths-Jones, S. *et al.* (2003) 'Rfam: An RNA family database', *Nucleic Acids Research*, pp. 439–441. doi: 10.1093/nar/gkg006.

Griffiths-Jones, S. *et al.* (2006) 'miRBase: microRNA sequences, targets and gene nomenclature.', *Nucleic Acids Research*, 34(suppl_1), pp. D140–D144. doi: 10.1093/nar/gkj112.

Groenen, M. A. M. *et al.* (2012) 'Analyses of pig genomes provide insight into porcine demography and evolution', *Nature*, 491(7424), pp. 393–398. doi: 10.1038/nature11622.

Hackl, T. *et al.* (2014) 'Proovread: Large-scale high-accuracy PacBio correction through iterative short read consensus', *Bioinformatics*, 30(21), pp. 3004–3011. doi: 10.1093/bioinformatics/btu392.

Hubley, R. *et al.* (2016) 'The Dfam database of repetitive DNA families', *Nucleic Acids Research*, 44(D1), pp. D81–D89. doi: 10.1093/nar/gkv1272.

Humphray, S. J. *et al.* (2007) 'A high utility integrated map of the pig genome.', *Genome Biology*, 8(7), p. R139.

Koren, S. *et al.* (2017) 'Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation', *Genome Research*, 27(5), pp. 722–736. doi: 10.1101/gr.215087.116.

Koren, S. *et al.* (2018) 'De novo assembly of haplotype-resolved genomes with trio binning', *Nature Biotechnology*, 36, pp. 1174-1182. doi: 10.1038/nbt.4277.

Kurtz, S. *et al.* (2004) 'Versatile and open software for comparing large genomes.', *Genome*

Biology, 5(2), p. R12. doi: 10.1186/gb-2004-5-2-r12.

Lefranc, M. P. *et al.* (2015) ‘IMGT R, the international ImMunoGeneTics information system R 25 years on’, *Nucleic Acids Research*, 43(D1), pp. D413–D422. doi: 10.1093/nar/gku1056.

Lewis, S. E. *et al.* (2002) ‘Apollo: a sequence annotation editor.’, *Genome Biology*, 3(12), p. RESEARCH0082. doi: 10.1186/gb-2002-3-12-research0082.

Li, H. and Durbin, R. (2009) ‘Fast and accurate short read alignment with Burrows-Wheeler transform’, *Bioinformatics*, 25(14), pp. 1754-1760. doi: 10.1093/bioinformatics/btp324

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv:1303.3997v1 [q-bio.GN]*.

Lowe, T. M. and Eddy, S. R. (1996) ‘TRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence’, *Nucleic Acids Research*, 25(5), pp. 955–964. doi: 10.1093/nar/25.5.0955.

Meyers, S. N. *et al.* (2005) ‘Piggy-BACing the human genome: II. A high-resolution, physically anchored, comparative map of the porcine autosomes’, *Genomics*, 86(6), pp.739-752. doi: 10.1016/j.ygeno.2005.04.010.

Miller, J. R., Hindkjær, J. and Thomsen, P. D. (1993) ‘A chromosomal basis for the differential organization of a porcine centromere-specific repeat’, *Cytogenetic and Genome Research*, 62(1), pp. 37–41. doi: 10.1159/000133441.

Nattestad, M. and Schatz, M. C. (2016) ‘Assemblytics: A web analytics tool for the detection of variants from an assembly’, *Bioinformatics*, 32(19), pp. 3021-3023. doi: 10.1093/bioinformatics/btw369.

Nguyen, D. T. *et al.* (2012) ‘The complete swine olfactory subgenome: expansion of the olfactory gene repertoire in the pig genome’, *BMC Genomics*, 13(1), 584. doi: 10.1186/1471-2164-13-584.

Papatheodorou, I. *et al.* (2018) ‘Expression Atlas: Gene and protein expression across multiple studies and organisms’, *Nucleic Acids Research*, 46(D1), pp. D246-D251. doi: 10.1093/nar/gkx1158.

Pendleton, M. *et al.* (2015) ‘Assembly and diploid architecture of an individual human genome via single-molecule technologies’, *Nature Methods*, 12(8), pp. 780–786. doi: 10.1038/nmeth.3454.

Putnam, N. H. *et al.* (2016) ‘Chromosome-scale shotgun assembly using an in vitro method for long-range linkage’, *Genome Research*, 26(3), pp. 342–350. doi: 10.1101/gr.193474.115.

Quinlan, A. R. and Hall, I. M. (2010) ‘BEDTools: A flexible suite of utilities for comparing genomic features’, *Bioinformatics*, 26(6), pp. 841–842. doi: 10.1093/bioinformatics/btq033.

Ramos, A. M. *et al.* (2009) ‘Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology’, *PLoS ONE*, 4(8), e6524. doi: 10.1371/journal.pone.0006524.

Robinson, J. T. *et al.* (2011) ‘Integrative genomics viewer’, *Nature Biotechnology*, 29(1), pp. 24–26. doi: 10.1038/nbt.1754.

Sayers, E. W. *et al.* (2010) ‘Database resources of the National Center for Biotechnology Information.’, *Nucleic Acids Research*, 38(Database issue), pp. D5-16. doi: 10.1093/nar/gkp967.

Servin, B. *et al.* (2012) ‘High-resolution autosomal radiation hybrid maps of the pig genome and their contribution to the genome sequence assembly’, *BMC Genomics*, 13(1), 585. doi: 10.1186/1471-2164-13-585.

She, R. *et al.* (2011) ‘genBlastG: Using BLAST searches to build homologous gene models’, *Bioinformatics*, 27(15), pp. 2141–2143. doi: 10.1093/bioinformatics/btr342.

Simão, F. A. *et al.* (2015) ‘BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs’, *Bioinformatics*, 31(19), pp. 3210–3212. doi: 10.1093/bioinformatics/btv351.

Skinner, B. M. *et al.* (2016) ‘The pig X and Y Chromosomes: structure, sequence, and evolution’, *Genome Research*, 26(1), pp. 130–139. doi: 10.1101/gr.188839.114.

Slater, G. S. C. and Birney, E. (2005) ‘Automated generation of heuristics for biological sequence comparison', *BMC Bioinformatics*, 6, 31. doi: 10.1186/1471-2105-6-31.

Smit, A., Hubley, R & Green, P. (2013-2015). *RepeatMasker Open-4.0.* [Online]. Available: http://www.repeatmasker.org [Accessed 16/05/2016].

Smit, A.F.A. & Hubley, R. (2008-2015) *RepeatModeler Open-1.0*. 2008-2015 http://www.repeatmasker.org

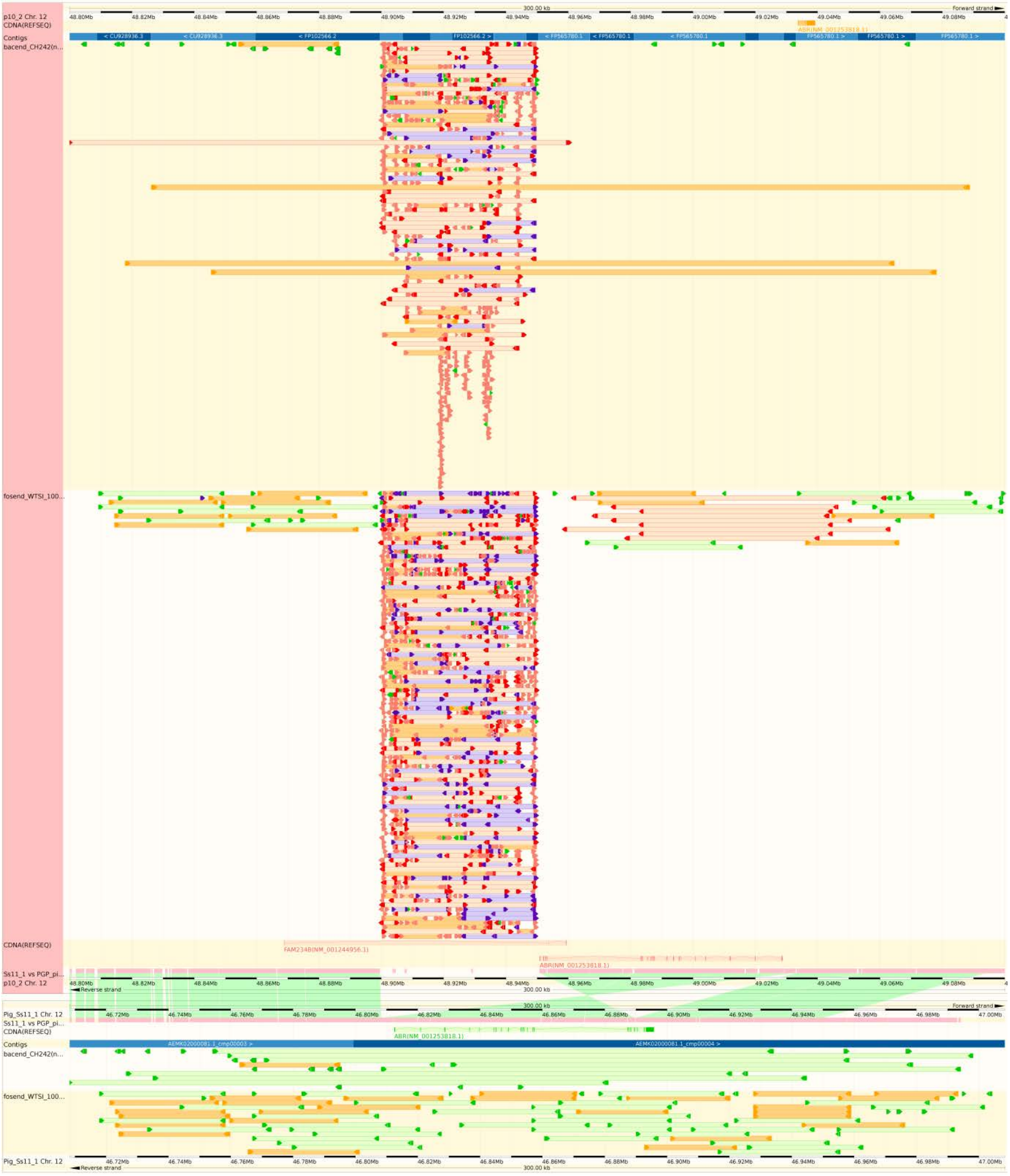
Walker, B. J. *et al.* (2014) 'Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement', *PLoS ONE*, 9(11), e112963. doi: 10.1371/journal.pone.0112963.

Warr, A. *et al.* (2015) 'Identification of Low-Confidence Regions in the Pig Reference Genome (Sscrofa 10.2)', *Frontiers in Genetics*, 6, 338. doi: 10.3389/fgene.2015.00338.

Zhang, J. *et al.* (2014) 'PEAR: A fast and accurate Illumina Paired-End reAd mergeR', *Bioinformatics*, 30(5), pp. 614–620. doi: 10.1093/bioinformatics/btt593.

5. Further supplementary figures

The following figures (Supplementary Figures SF12-16) illustrates improvements in the assemblies as discussed in the main paper text.

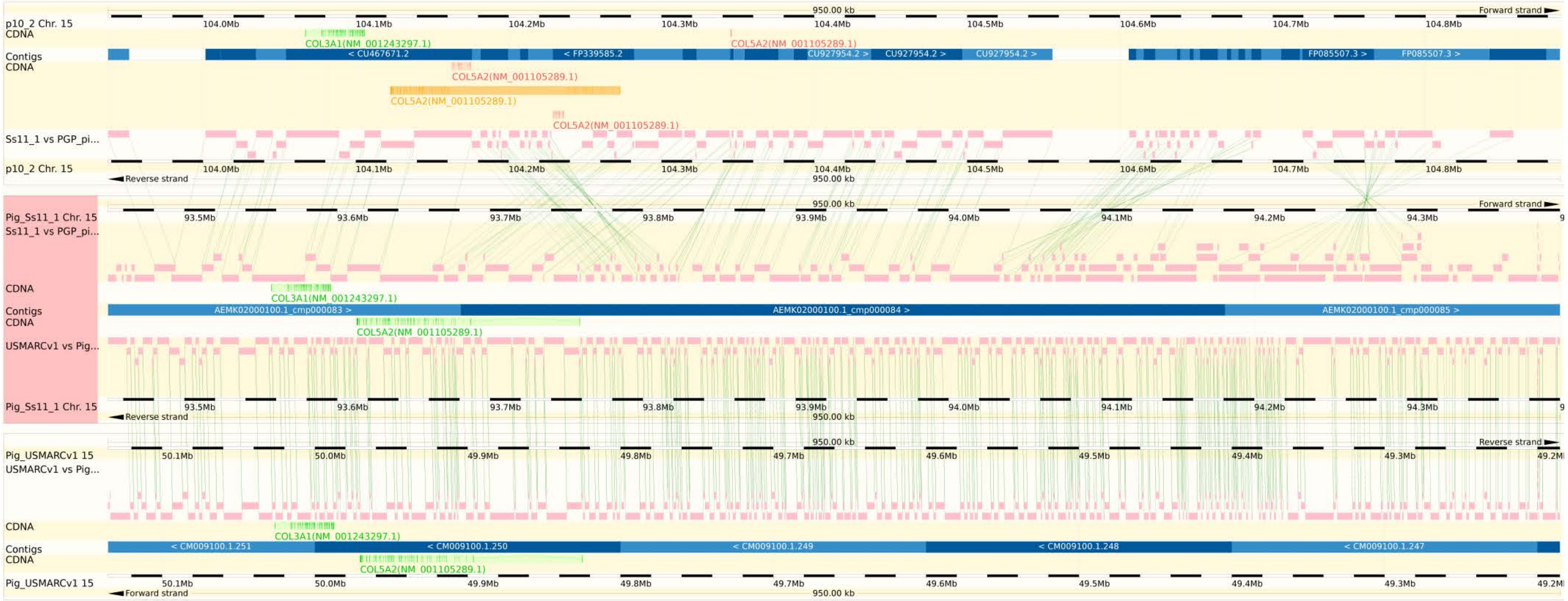


Supplementary Figure SF12: Illustration of improvement in local order and orientation and reduction in sequence redundancy

The alignment of isogenic CH242 BAC end and WTSI_1005 fosmid end sequences with the Sscrofa10.2 (upper panel with pink bar on left hand side) and Sscrofa11.1 (lower panel).

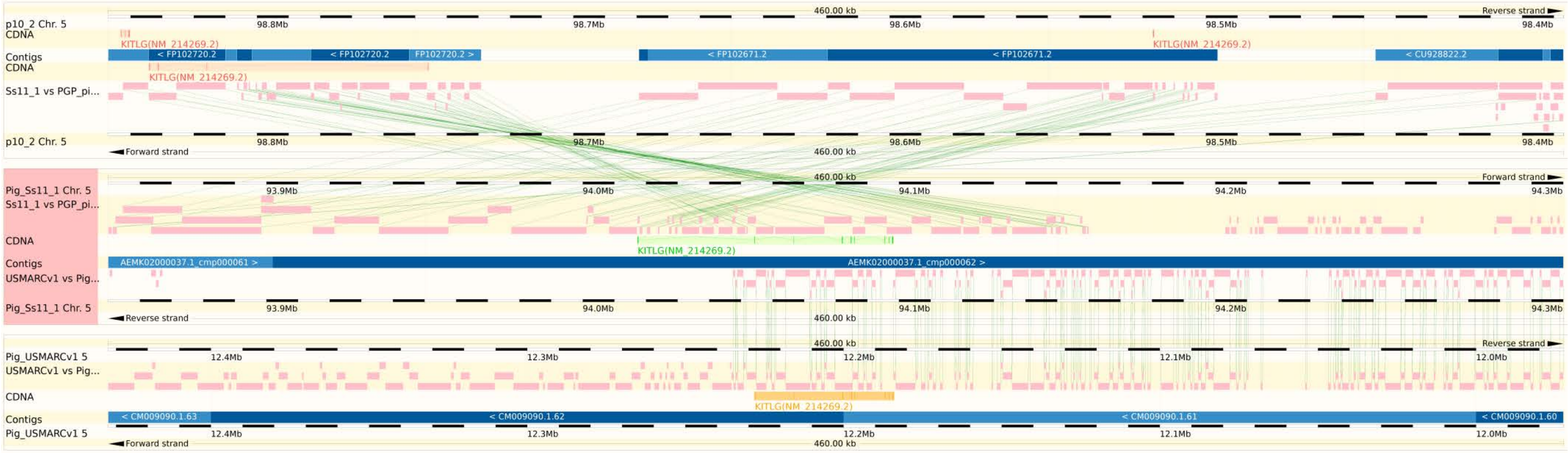
Red arrows indicate incorrect orientation of the paired end sequences, purple arrows are sequences which are present multiple times, green and orange arrows indicate the end sequences are correctly oriented. The distances between correctly oriented end sequences are as expected (green) or either greater or less than expected (orange) for the clone insert size for the fosmid or CH242 BAC libraries.

Supplementary Figure SF13: gEVAL comparison of Sscrofa10.2, Sscrofa11.1 and USMARCv1.0 at *COL3A1*, *COL5A2* loci.



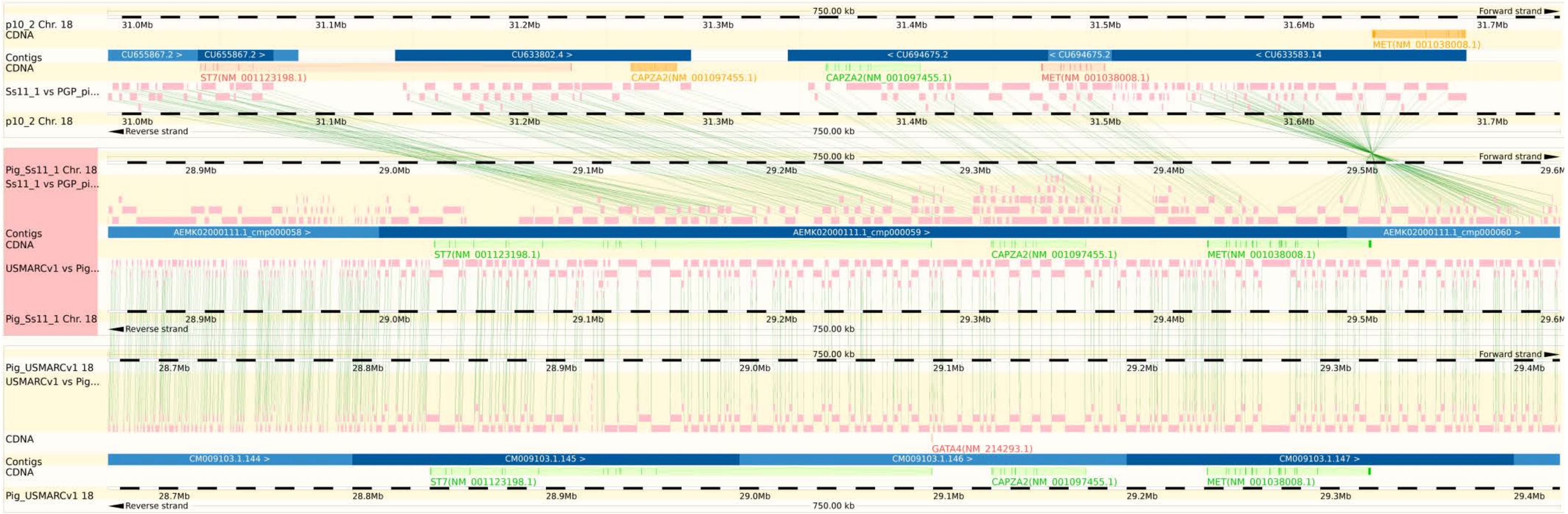
In the new assembly (Sscrofa11.1, middle row marked with pink vertical block) an improved gene model for COL5A2 can be annotated; in the previous assembly (Sscrofa10.2, upper row) the order and orientation of sequence contigs within BAC clone CH242-40P12 (ENA: FP339585.2) are not resolved. There is good agreement between the Sscrofa11.1 (middle row) and the USMARCv1.0 (lower row) although the USMARCv1.0 assembly of SSC15 is inverted relative to Sscrofa11.1.

Supplementary Figure SF14: gEVAL comparison of Sscrofa10.2, Sscrofa11.1 and USMARCv1.0 at the *KITLG* locus.



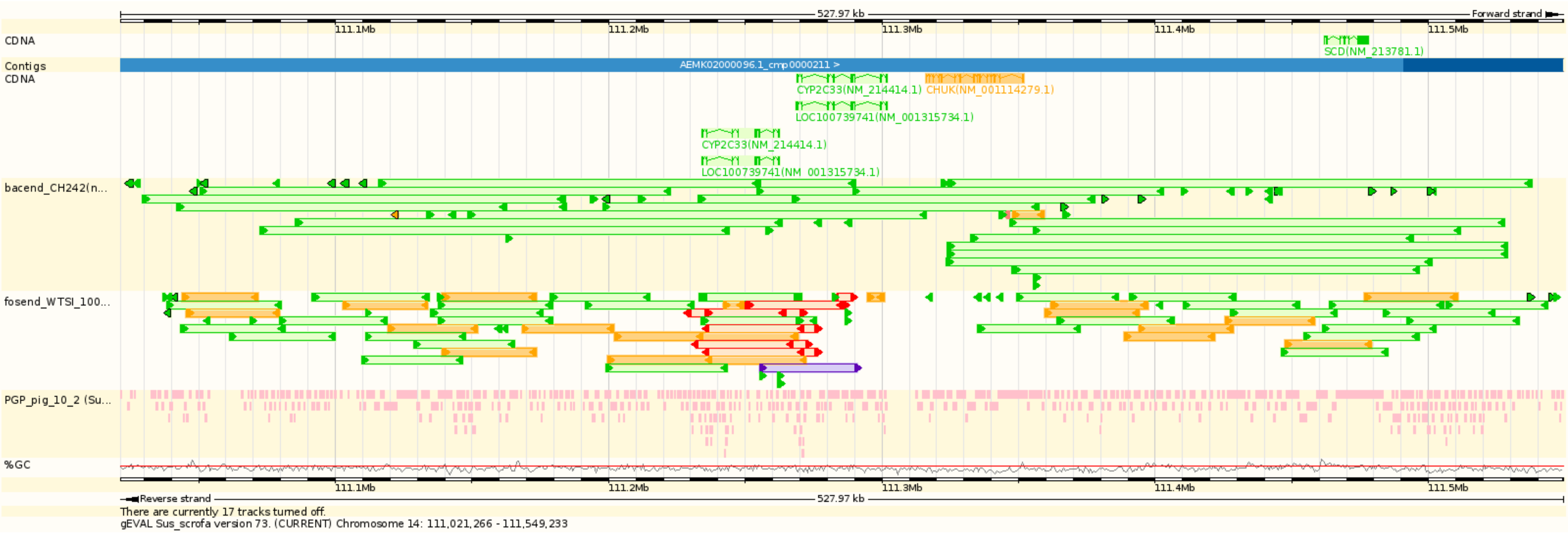
The new assembly (Sscrofa11.1, middle row with pink vertical block at left hand side) resolves the sequences encoding *KITLG* which were split across two small scaffolds in Sscrofa10.2 (upper row). Although there is good agreement between Sscrofa11.1 (middle row) and USMARCv1.0 (lower row) assemblies in the right hand half of the region on SSC5 above, there is additional sequence present in the Sscrofa11.1 assembly between *DUSP6* and *KITLG*, the gene model for *KITLG* appears incomplete in the USMARCv1.0 assembly. Again the USMARCv1.0 is inverted relative to Sscrofa11.1.

Supplementary Figure SF15: gEVAL comparison of Sscrofa10.2, Sscrofa11.1 and USMARCv1.0 across the *ST7*, *CAPZA2* and *MET* loci on SSC18



The new assembly (Sscrofa11.1, middle row with pink block at left hand side) resolves the coding sequences for i) *ST7* that were previously split across two small scaffolds; *CAPZA2* that was similarly split across two small scaffolds; and iii) the *MET* sequences that were previously split as a result of an error in the orientation of the sequence drawn from BAC clone CH242-385N7 (ENA: CU633583.14) with respect to the sequence from BAC clone CH242-150K23 (ENA: CU694675.2) that harbours parts of the *MET* locus. This error in the incorporation of the CH242-385N7 (ENA: CU633583.14) in the Sscrofa10.2 assembly (upper row) is particularly unfortunate as this BAC had been sequenced to finish quality. There is good agreement between the Sscrofa11.1 (middle row) and USMARCv1.0 (lower row) assemblies with both SSC18 assemblies also being in the same orientation.

Supplementary Figure SF16: Absence of *ERLIN1* gene, duplication of *CYP2C33*


527.97 kb
Forward strand
111.1Mb
111.2Mb
111.3Mb
111.4Mb
111.5Mb
CDNA
SCD(NM_213781.1)
Contigs
AEMK02000096.1_cmp0000211 >
CDNA
CYP2C33(NM_214414.1)
CHUK(NM_001114279.1)
LOC100739741(NM_001315734.1)
CYP2C33(NM_214414.1)
LOC100739741(NM_001315734.1)
bacend_CH242(n...
fosend_WTSI_100...
PGP_pig_10_2 (Su...
%GC
Reverse strand
There are currently 17 tracks turned off.
gEVAL Sus_scrofa version 73. (CURRENT) Chromosome 14: 111,021,266 - 111,549,233