Ackerson IV, L., Olagunju, T., Kapoor, M., Durante, A., Stegemiller, M., Shira, K., Kuhn, K., Eory, L., McKay, S., Huang, W., Murdoch, B., Archibald, A., Smith, T., Tuggle, C., Rosen, B., Clark, E. (2026) Telomere to telomere (T2T) genomes for the pig, *Proceedings 13th World Congress on Genetics Applied to Livestock Production* 23812: https://doi.org/10.31274/wcgalp.23812
©2026 Ackerson IV, et al. CC BY-NC-ND

Telomere to telomere (T2T) genomes for the pig

L.K. Ackerson IV[1*], T.A. Olagunju[2], M. Kapoor[3], A. Durante[4], M.R. Stegemiller[2], K.A. Shira[2], K.L. Kuhn[5], L. Eory[6], S.D. McKay[7] W. Huang[1], B.M. Murdoch[2], A.L. Archibald[6], T.PL. Smith[5], C.K. Tuggle[3], B.D. Rosen[8], and E.L Clark[6,9]

[1]Department of Animal Science, Michigan State University, East Lansing, MI, USA; [2]Department of Animal, Veterinary and Food Sciences, University of Idaho, Moscow, ID, USA; [3]Department of Animal Science, Iowa State University, Ames, IA, USA; [4]GenPhySE, Université de Toulouse, INRAE, INP, ENVT, Castanet-Tolosan, France; [5]USDA Meat Animal Research Center, Clay Center, NE, USA; [6]The Roslin Institute, University of Edinburgh, Edinburgh, UK; [7]Division of Animal Sciences, University of Missouri, Columbia, MO, USA; [8]USDA Animal Genomics and Improvement Laboratory, Beltsville, MD; [9]European Molecular Biology Laboratory's European Bioinformatics Institute, Hinxton, Cambridge, UK; [*]Corresponding author: ackers24@msu.edu

Introduction

The current reference genome for the pig (*Sus scrofa*) was released to the public in 2017 and facilitated significant progress within both the swine industry and the livestock genomics research community. This reference, *Sscrofa11.1* (GCA_000003025.6), utilized long-read sequencing to achieve substantial improvement over its predecessor *Sscrofa10.2*. Contig N50 increased from 69.5 kb to 48.2 Mb (Warr *et al*., 2020). However, this reference assembly contained over a thousand fragmented contigs and a large number of gaps across the scaffolded chromosomal sequences. These imperfections frequently cause issues in downstream analyses such as sequence read alignment, variant calling, and annotation (Aganezov *et al*., 2022). In addition, the pseudo-haploid reference was derived from a single Duroc sow with Y chromosome from a different source and no haplotype resolution. These constraints pose certain complexities for ongoing research utilizing this reference assembly.

Recent advancements in long-read sequencing technologies offer a means to improve the current single linear swine reference genome. The primary reason for the prevalence of gaps in reference genomes is repetitive sequences, particularly those which span a length longer than the utilized reads in the assembly (Nurk *et al*., 2022). Gaps are primarily located in highly repetitive and heterochromatic regions including centromeres, telomeres, ribosomal DNA arrays, and transposable elements. As sequences are assembled via unique overlapping regions, short reads from long repetitive regions are indistinguishable and difficult to pinpoint in regard to their location in the genome. However, accurate long-read sequences are capable of spanning the entire region of repeats to include unique neighboring sequences, that allow the production of highly contiguous assemblies and reduces the number of gaps and unplaced sequences. Contemporary long-read technologies such as PacBio HiFi and Oxford Nanopore (ONT) routinely produce reads with 18-25 kb of length and 99.9% accuracy, and 50-150 kb of length and 95% accuracy, respectively (Wang *et al*., 2025; Koren *et al*., 2024). Recent improvements in the ONT platform have yielded higher accuracies and longer reads, reaching the scale of megabases for an individual read with the Ultra-Long (UL) technology, greatly improving our ability to span and assemble through complex repetitive regions of genomes.

Long-read sequencing technologies have facilitated the production of complete, telomere-to-telomere (T2T) assemblies in humans, sheep, goats, and many more (Nurk *et al*., 2022; Olagunju *et al*., 2025; Wu *et al*., 2024). Beyond the key advantage of illuminating novel and previously unresolved regions of the genome, studies leveraging these new T2T assemblies

have consistently shown improvements in downstream analyses such as short-read mapping and variant calling (Aganezov *et al*., 2022). Here we present our newly developed haplotype-resolved T2T genome assemblies from an F1 male fetus from a mating between a Hampshire boar and a composite Yorkshire x Landrace sow (HxYL). Together with other T2T genome assemblies, these assemblies will further advance swine genome research and genetic improvement.

Materials & Methods

A trio-based approach utilizing a commercially relevant Hampshire x (Yorkshire x Landrace) cross was used to generate two T2T genome assemblies, one for each parental haplotype (Koren *et al*., 2018). An average sized male gestational day 72 fetus of the cross was used to generate long-read sequencing data, including a combination of highly accurate HiFi reads for bicep femoris, lung, and brain frontal cortex, and ultra-long (UL) reads for liver, lung and spleen. Short-read Illumina sequencing was performed on samples from the parents, and Omni-C was generated from fetal fibroblast cells; both of which served to aid haplotype resolution of the diploid assembly.

The parental haplotypes were assembled several times using Verkko and HiFiasm with varying assembler parameters, sequence data allocations, and read filtering criteria to fully evaluate the optimal output assembly with the software and sequence data (Rautiainen *et al*., 2023; Antipov *et al*., 2025; Cheng *et al*., 2021). In the end, the assembly produced with Verkko v2.2 using quality filtered HiFi, Duplex, and ONT-UL reads produced the best assembly in terms of contiguity and T2T statistics. Two sets of haplotype-resolved assemblies were then deduced from this diploid assembly, the first using parental short-read data in accordance with the trio-binning approach, and the second using spatial Omni-C data. Importantly, both of these assemblies have the same core Verkko v2.2 diploid assembly, the haplotype resolution being the only differing aspect – albeit they were nearly identical.

Each assembly was then moved to a manual curation step, that serves to resolve and detangle ambiguous regions and gaps in the assembly that the assembler (Verkko) failed to confidently scaffold through. These regions were detangled in a variety of ways, primarily using logic and alignment information from mapping the ONT reads to the graph and evaluating read support through every possible path in the ambiguous regions. When needed, the Omni-C assembly was used to offer additional evidence towards a solution of the gaps in the trio-binning assembly graph. To resolve the rDNA sequences in the assembly on Chromosomes 8 and 10, we generated a representative morph of each haplotype's rDNA array using assembly sequences in the region and the software *Ribotin* (Rautiainen, 2024). Once all four (two haplotypes and two chromosomes) representative morphs were produced, copy number for each morph was estimated using k-mer counting and normalization via *CONKORD* (Borcherm, 2025*)*. The morph and its respective CNV estimation were then leveraged together to produce a representative rDNA sequence, each of which was aligned to their corresponding region in the genome and anchored into the assembly.

The diploid assembly was evaluated for the presence of adaptor or foreign sequences, and if present they were excised from the assembly. The mitochondrial contig was assembled, circularized, and oriented to the correct starting position. Once the assembly was structurally sound, meaning that solutions to tangles and gaps that we could resolve with very high confidence were applied, the trio-binning assembly was moved forward to the polishing stage. This stage entailed aligning the reads used to build the assembly back to the final product to evaluate accuracy and fix any errors that were introduced due to low coverage, manual intervention, or a variety of other reasons. This step primarily serves to improve the base level accuracy of the assembly. Briefly, we used Canu (Koren *et al*., 2017) to bin the reads, BWA-MEM (Li and Durbin, 2009) and Winnowmap (Jain *et al*., 2020) to align the Illumina and HiFi reads, respectively, DeepVariant (Poplin *et al*., 2018) to call the variants, Merfin (Formenti *et*

al., 2022) & BCFtools (Danecek *et al*., 2021) to filter and apply the variants to the assembly, and Meryl and Merqury (Rhie *et al*., 2020) to evaluate quality value (QV). The polished assembly was then re-oriented at the chromosome level such that the sequences span *p*-arm to *q*-arm in accordance with standard convention; this was validated using centromeric signals and alignments to the existing reference genome. The final assembly then went through rigorous quality evaluation to assess assembly statistics relevant to completeness, contiguity, and correctness.

Both haplotype assemblies have been deposited to NCBI with accession numbers GCA_054392225.1 (paternal) and GCA_054392235.1 (maternal).

Results & Discussion

The final haplotype-resolved assemblies of the HxYL hybrid are of high quality with every chromosome (2n=38) having both telomeres correctly placed. The paternal Hampshire (*H*) haplotype has 16 chromosomes achieving T2T at the contig level, including chromosomes 1-13, 17, 18, and Y; and the 3 remaining chromosomes (14, 15, and 16) achieving T2T at the scaffold level. The maternal Yorkshire x Landrace (*YL*) haplotype has 17 chromosomes achieving T2T at the contig level, including chromosomes 1-14, 17, 18, and X; with the remaining 2 chromosomes (15 and 16) being T2T scaffolds. The final diploid assembly contains a total of 9 unresolved gaps, 5 on the paternal and 4 on the maternal haplotypes. All gaps are located on acrocentric chromosomes 14, 15, and 16. The *H* and *YL* haplotypes have QV scores of 56.71 and 56.76, respectively. Contiguity is extremely high; the *H* and *YL* haplotype assemblies have contig N50 values of 147.07 Mb and 145.42 Mb, respectively. To assess completeness of these assemblies, we determined BUSCO scores using the *cetartiodactyla_odb10* lineage dataset which yielded scores of 95.7% and 98.3% for the *H* and *YL* haplotype assemblies, respectively. We note that the paternal assembly has a lower completeness score as a result of the X chromosome exclusion. Collectively, these assemblies identified 232 Mb of novel sequence or previously unassembled regions (PURs), compared to Sscrofa11.1.

One major advantage of T2T and long-read based assemblies is the ability to identify and genotype large structural variants. By aligning the *H* and *YL* haplotypes, we identified a 52.9 Mb inversion near the q arm of the paternal (*H*) chromosome 15 relative to the maternal and *Sscrofa11.1* reference chromosome 15 (Figure 1), which has been confirmed via fibroblast optical mapping data.

Conclusions

Here we present two nearly gapless, high quality T2T haplotype-resolved assemblies for the Hampshire and Yorkshire x Landrace breeds. Each assembly greatly improves upon the current reference genome in terms of contiguity, while also achieving true, stand-alone haplotype resolution as opposed to a mixture between two haplotypes in pseudo-haploid assemblies. Our analysis demonstrates that these assemblies can be used to identify large scale structure variants effectively. Additional assemblies, including T2T or non-T2T but highly contiguous assemblies will be useful towards building a comprehensive pangenome to further advance swine genome research.

Acknowledgements

We acknowledge support from the Roslin Foundation, BBSRC Roslin Institute ISP Pump-Priming Awards 2018 & 2023, BBSRC Awards BB/W018772/1, BB/W019108/1 and BBS/E/RL/230001B, USDA-NIFA: 2024-67015-42253, 2022-67015-37055, and NRSP8 Swine Genome Coordination Effort, and USDA-ARS Project #3040-31000-104-00D.

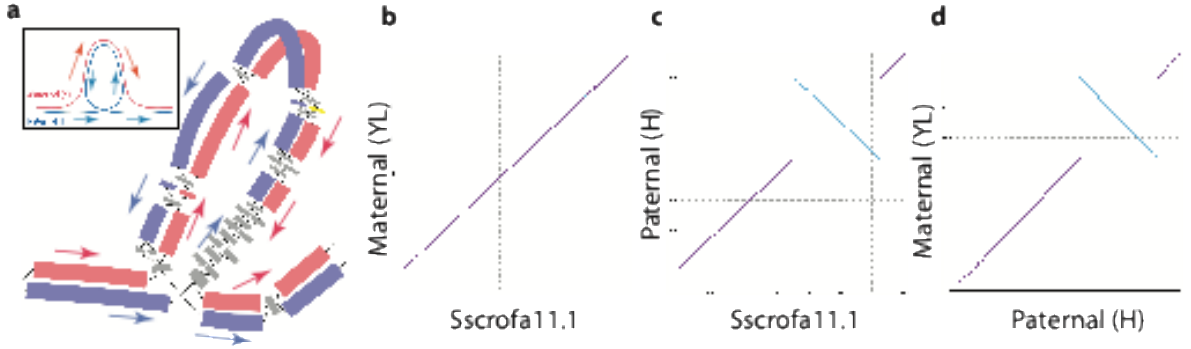

Figure 1. Identification of a large-scale inversion on the paternal chromosome 15. (**a**) Illustration of the inversion and a plot of the inversion and surrounding nodes of the assembly graph. Red nodes are unambiguously maternally derived, and blue paternal. Grey nodes are either homozygous between the two haplotypes or unresolved. (**b**) Dot plot between the maternal haplotype and the Sscrofa11.1 reference. (**c**) Dot plot between the paternal haplotype and the Sscrofa11.1 reference. (**d**) Dot plot between the maternal and paternal haplotypes.

References

Aganezov S., Yan S.M., Soto D.C., *et al*. (2022) Science 376(6588):eabl3533. https://doi.org/10.1126/science.abl3533

Antipov D., Rautiainen M., Nurk S., *et al*. (2025) Genome Res. 35(7):1583–1594. https://doi.org/10.1101/gr.280383.124

Borcherm M., CONKORD copy number variation detection software. Available at: https://github.com/borcherm/CONKORD, visited on: 03-12-2025.

Cheng H., Concepcion G.T., Feng X., *et al*. (2021) Nat. Methods 18(2):170–175. https://doi.org/10.1038/s41592-020-01056-5

Danecek P., Bonfield J.K., Liddle J., *et al*. (2021) Twelve years of SAMtools and BCFtools. GigaScience 10(2):giab008. https://doi.org/10.1093/gigascience/giab008

Formenti G., Rhie A., Walenz B.P., *et al*. (2022) Nat. Methods 19(6):696-704. https://doi.org/10.1038/s41592-022-01440-3

Jain C., Rhie A., Zhang H., *et al*. (2020) Bioinformatics 36(Supplement_1):i111–i118. https://doi.org/10.1093/bioinformatics/btaa435

Koren S., Bao Z., Guarracino A., *et al*. (2024) Genome Res. 34(11):1919–1930. https://doi.org/10.1101/gr.279334.124

Koren S., Rhie A., Walenz B.P., *et al*. (2018) Nat. Biotechnol. 36(12):1174–1182. https://doi.org/10.1038/nbt.4277

Koren S., Walenz B.P., Berlin K., *et al*. (2017) Genome Res. 27(5):722–736. https://doi.org/10.1101/gr.215087.116

Li H., and Durbin R. (2009) Bioinformatics 25(14):1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Nurk S., Koren S., Rhie A., *et al*. (2022) Science 376(6588):44–53. https://doi.org/10.1126/science.abj6987

Olagunju T.A., Pospelova M., Schwartz J.C., *et al*. (2025) bioRxiv:2025.09.30.679298. https://doi.org/10.1101/2025.09.30.679298

Poplin R., Chang P.-C., Alexander D., *et al*. (2018) Nat. Biotechnol. 36(10):983–987. https://doi.org/10.1038/nbt.4235

Rautiainen M., (2024) Bioinformatics 40(3):btae124-btae124. https://doi.org/10.1093/bioinformatics/btae124

Rautiainen M., Nurk S., Walenz B.P., *et al*. (2023) Nat. Biotechnol. 41(10):1474–1482. https://doi.org/10.1038/s41587-023-01662-6

Rhie A., Walenz B.P., Koren S., Phillippy A.M., *et al*. (2020) Genome Biol. 21(1):245. doi.org/10.1186/s13059-020-02134-9

Wang B., Jia P., Gao S., *et al*. (2025) Genomics Proteomics Bioinformatics 23(1):qzaf003. https://doi.org/10.1093/gpbjnl/qzaf003

Warr A., Affara N., Aken B., *et al*. (2020) GigaScience 9(6):giaa051. https://doi.org/10.1093/gigascience/giaa051

Wu H., Luo L.-Y., Zhang Y.-H., *et al*. (2024) Nature Communications 15:10041. https://doi.org/10.1038/s41467-024-54188-z