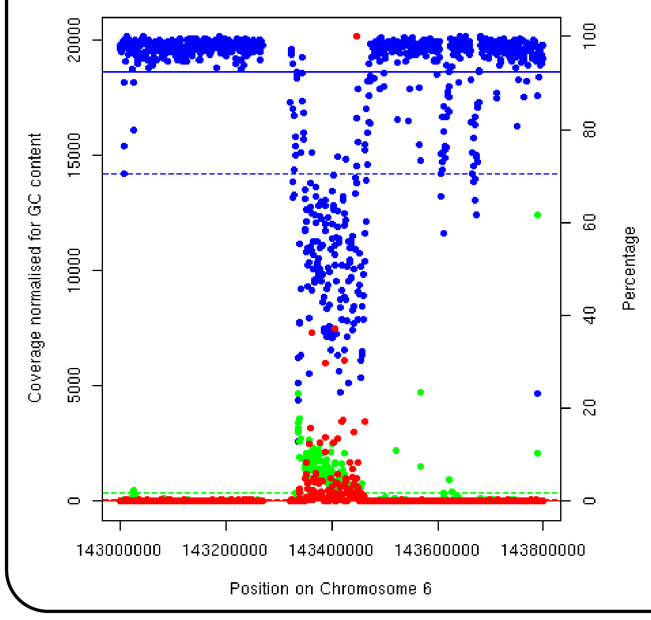
High-Quality, Highly Contiguous Re-Assembly of the Pig Genome

A Warr[1], R Hall[2], K Kim[2], S Koren[3], A Phillippy[3], DM Bickhart[4], BD Rosen[5], SG Schroeder[6], D Hume[1], R Talbot[1], W Chow[7], K Howe[7], D Nonneman[8], GA Rohrer[8], M Watson[1], TPL Smith[8], AL Archibald[1]

[1] The Roslin Institute and R(D)SVS, University of Edinburgh, Easter Bush, Edinburgh, UK **[2]** Pacific Biosciences, Menlo Park, CA, USA **[3]** National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA **[4]** Dairy Forage Research Center, USDA-ARS, Madison, WI, USA **[5]** ARS, USDA, Beltsville, MD, USA **[6]** Animal Genomics and Improvement Laboratory, USDA-ARS, Beltsville, MD, USA **[7]** The Wellcome Trust Sanger Institute, Hinxton, UK **[8]** USDA, ARS, USMARC, Clay Center, NE, USA

Introduction

Errors in reference genomes increase the number of **false-positives** in downstream analysis. Recently we have shown that the pig reference genome Sscrofa10.2 has misassemblies and poor mappabillity regions covering **up to a third of the genome [1]**. A new, high-quality reference genome is needed.

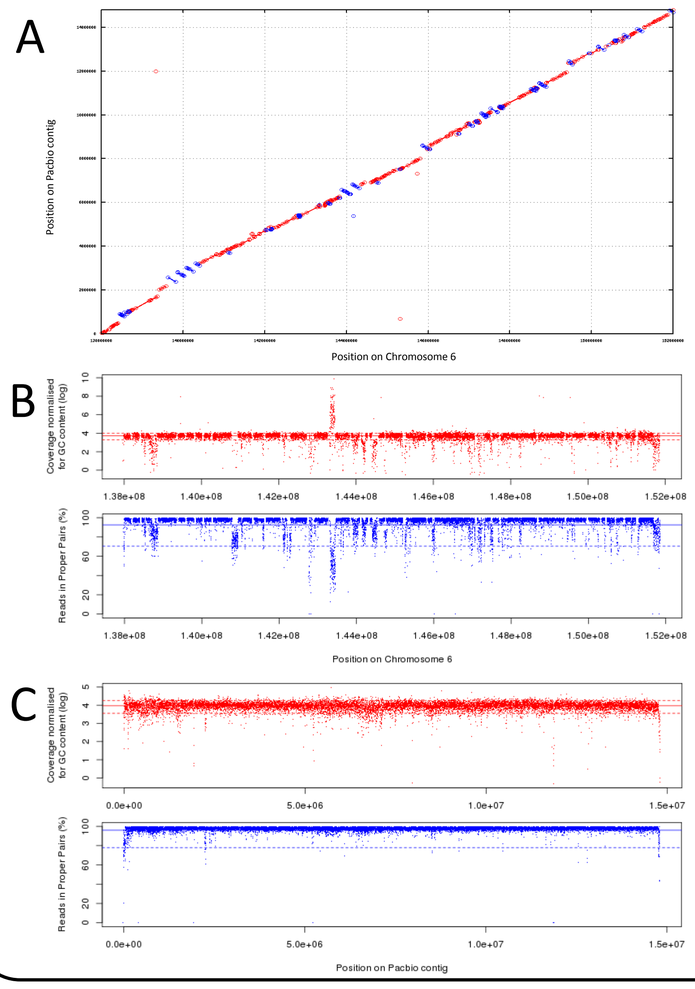


Left: Example of misassembled region in Sscrofa10.2. Data points represent the percentage properly paired **(blue)**, the GC-normalised coverage **(red)** and percentage larger than expected insert sizes **(green)** of 1000bp windows. Solid lines represent whole genome mean and dashed lines represent 2 std. from the mean.

Methods

- 65X **PacBio** C4 Sequencing of TJ Tabasco- **same sow** as Sscrofa10.2 assembly
- Error correction with **Quiver** (65X PacBio raw reads)
- **Falcon** Assembly
- Mapping to Sscrofa10.2 with **Nucmer** to order and orient contigs
- Gap Filling with **PBJelly** (65X PacBio raw reads)
- Error Correction with **Arrow** (65X PacBio raw reads) and **Pilon** (40X PE Illumina reads)
- Mapping of large unplaced contigs, sequenced BACs from original assembly, and BAC ends to assembly with **Nucmer** and **BWA mem**
- Placing of large unplaced contigs, sequenced BACs
- **MinION** sequencing of 5 BACs with ends spanning gaps, assembled with canu and mapped using **BWA mem.** All 5 successfully placed in expected position

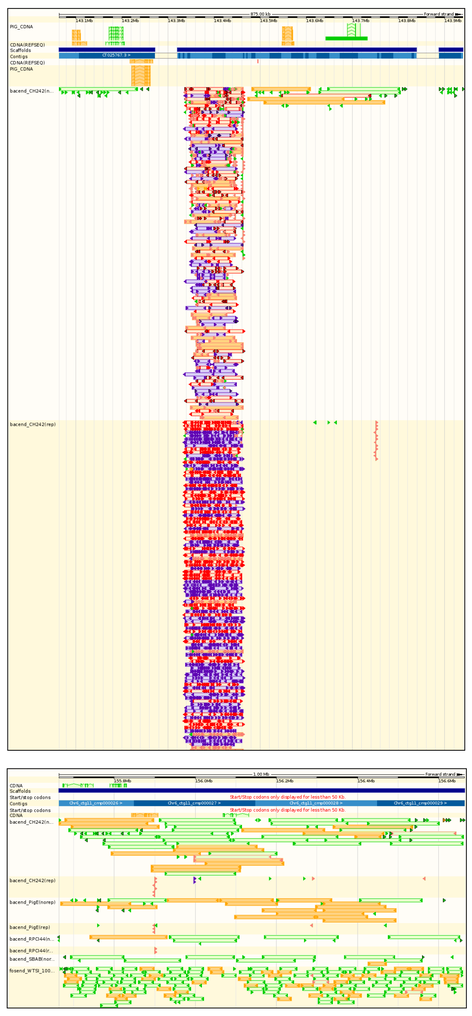
Corrected Misassemlies



Left : A: Alignment of region on Chr 6 between Sscrofa10.2 (x-axis) and new assembly (y-axis).
B, C: Coverage **(red)** and % properly paired **(blue)** for regions shown in **A** of paired end Illumina data for Sscrofa10.2 and new assembly, respectively

Improved consistency of coverage and improved % properly paired suggest the differences between the assemblies are corrections of misassemblies in Sscrofa10.2.

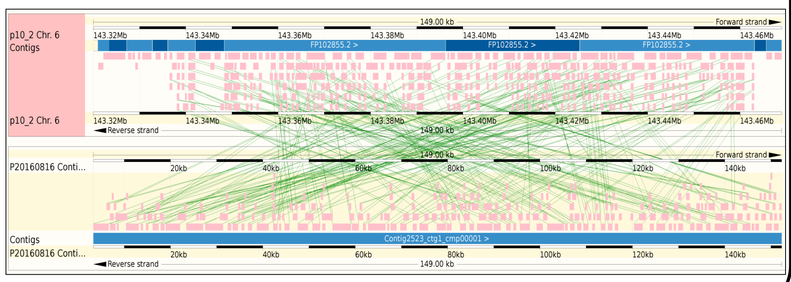
Corrected Misassemlies



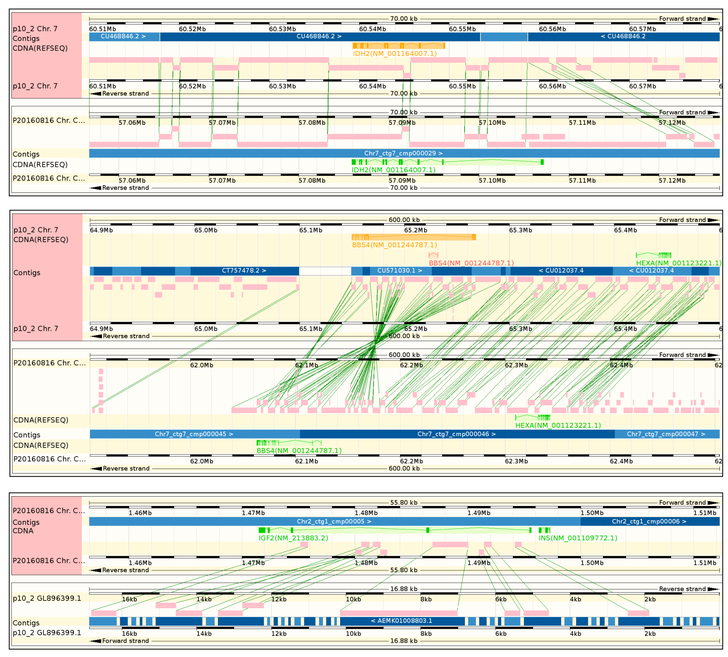
Left (top): Misassembled region in Sscrofa10.2 viewed on the gEVAL **[2]** browser showing poor BAC end mapping.
Left (bottom) Same region of the new assembly viewed on gEVAL browser showing well mapped BAC and fosmid libraries
Below: gEVAL key

	Individual	Paired
	Mapped 1 or 2 times in no_rep	Paired, correct direction, correct distance.
	N/A	Paired end, wrong direction (<>, >>, <<)
	N/A	Paired end, wrong distance (library insert size)
	Mapped more than twice	One of the hits mapped more than twice.
	N/A	Same end, artificial pairing.
	Discrepancy between ends fpc location and the fpc contig its currently mapped to	N/A
	Highlighted clone (read)	Highlighted clone (pair)
	Black border, Paired end in close vicinity outside current window	N/A

Right: gEVAL visualisation of rearrangements in a section of the above region between Sscrofa 10.2 (top) and new assembly (bottom)



Corrected Genes



A, B: gEVAL visualisation of rearrangements that complete genes and correct the order and orientation of exons from Sscrofa10.2 (top) to the new assembly (bottom)
C: gEVAL visualisation of placement of genes missing from Sscrofa 10.2 in the new assembly (top) which were previously lost in a highly fragmented unplaced scaffold (bottom)

Assembly Statistics

Contig N50: 48Mbp
Contig L50:15
Scaffold N50: 88Mbp
Scaffold L50:9
Length: 2.46Gb
Spanned Gaps: 79
Unspanned Gaps: 24

Conclusion

We have produced a much improved pig reference assembly, **Sscrofa11, which is now available on NCBI (Acc. GCA_000003025.5)**. This assembly will allow for more accurate analysis of the pig genome.

References:
1. Warr et al. (2015) Identification of Low-Confidence Regions in the Pig Reference Genome (Sscrofa10.2). Frontiers in Genetics doi:10.3389/fgene.2015.00338
2. Chow et al. (2016) gEVAL — a web-based browser for evaluating genome assemblies. Bioinformatics doi:10.1093/bioinformatics/btw159

amanda.warr@roslin.ed.ac.uk @amandatron89