Identification of low-confidence regions in the pig reference genome (Sscrofa10.2)

Citation for published version:
Warr, A, Robert, C, Hume, D, Archibald, AL, Deeb, N & Watson, M 2015, 'Identification of low-confidence regions in the pig reference genome (Sscrofa10.2)' Frontiers in genetics, vol. 6, 338. DOI: 10.3389/fgene.2015.00338

Digital Object Identifier (DOI):
10.3389/fgene.2015.00338

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Frontiers in genetics

frontiers
in Genetics

Identification of low-confidence regions in the pig reference genome (Sscrofa10.2)

Amanda Warr[1], Christelle Robert[1], David Hume[1], Alan L. Archibald[1], Nader Deeb[2], Mick Watson[1]*

[1]Edinburgh Genomics, The Roslin Institute, United Kingdom, [2]Genus plc, USA

Submitted to Journal:
Frontiers in Genetics

Specialty Section:
Livestock Genomics

ISSN:
1664-8021

Article type:
Original Research Article

Received on:
11 Aug 2015

Accepted on:
12 Nov 2015

Provisional PDF published on:
12 Nov 2015

Frontiers website link:
www.frontiersin.org

Citation:
Warr A, Robert C, Hume D, Archibald AL, Deeb N and Watson M(2015) Identification of low-confidence regions in the pig reference genome (Sscrofa10.2). *Front. Genet.* 6:338. doi:10.3389/fgene.2015.00338

Identification of low-confidence regions in the pig reference genome (Sscrofa10.2)

Amanda Warr[1], Christelle Robert[1], David Hume[1], Alan Archibald[1], Nader Deeb[2] and Mick Watson[1*]

1. The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Edinburgh, UK
2. Genus plc., Hendersonville, TN, USA

Correspondence:
Dr. Mick Watson
The Roslin Institute and Royal (Dick) School of Veterinary Studies
University of Edinburgh
Easter Bush
Edinburgh,EH25 9RG, UK
mick.watson@roslin.ed.ac.uk

Abstract

Many applications of high throughput sequencing rely on the availability of an accurate reference genome. Variant calling often produces large data sets that cannot be realistically validated and which may contain large numbers of false-positives. Errors in the reference assembly increase the number of false-positives. While resources are available to aid in the filtering of variants from human data, for other species these do not yet exist and strict filtering techniques must be employed which are more likely to exclude true-positives. This work assesses the accuracy of the pig reference genome (Sscrofa10.2) using whole genome sequencing reads from the Duroc sow whose genome the assembly was based on. Indicators of structural variation including high regional coverage, unexpected insert sizes, improper pairing and homozygous variants were used to identify low quality (LQ) regions of the assembly. Low coverage (LC) regions were also identified and analyzed separately. The LQ regions covered 13.85% of the genome, the LC regions covered 26.6% of the genome and combined (LQLC) they covered 33.07% of the genome. Over half of dbSNP variants were located in the LQLC regions. Of copy number variable regions (CNVRs) identified in a previous study, 86.3% were located in the LQLC regions. The regions were also enriched for gene predictions from RNA-seq data with 42.98% falling in the LQLC regions. Excluding variants in the LQ, LC or LQLC from future analyses will help reduce the number of false-positive variant calls. Researchers using WGS data should be aware that the current pig reference genome does not give an accurate representation of the copy number of alleles in the original Duroc sow's genome.

Keywords: Missassembly, Copy Number Variable Regions, Structural Variation, Draft Assemblies, False Positives

Introduction

Contemporary genetics research benefits from genomics tools and resources, including DNA sequencing and SNP (single nucleotide polymorphism) chips, which facilitate detailed quantitative molecular characterization of genetic variation at the population and individual level. A high quality reference genome sequence for the species of interest is an invaluable asset for the discovery of molecular genetic variants. Most reference genome sequences for species with large, complex genomes are incomplete representations of the genome sequence of a single individual or a small number of individuals. Given the extent of insertion/deletion (indel) polymorphisms and copy number variation (CNV) within species such individual reference genomes do not contain all the sequences present in the species of interest. Thus, there are two major flaws in the current single linear model for reference genomes as a framework for discovery and analysis of genetic variation: 1) errors and gaps in the reference genome assemblies most of which are incomplete drafts; and 2) using a haploid genome of one individual to represent the genome(s) of a species. In this paper we focus solely on the former.

Studies that employ variant calling from sequencing data to find variation in the genome produce large variant call sets (Belkadi et al., 2015;Robert et al., 2014;Bianco et al., 2015;Gudbjartsson et al., 2015). Most of these calls will be either false-positive, not relevant to the phenotype under investigation or benign (MacArthur et al., 2012). Failure to detect true variants (i.e. false-negatives) will also occur either as a result of insufficient sequence depth or gaps in the reference genome (real or technical).Filtering these datasets reduces the number of variants to a level which can be validated, however in the process researchers risk discarding the variants they are looking for.

Many applications of high throughput sequencing rely heavily on the accuracy of the available reference genome for the species. Errors in the reference genome increase the number of false-positive variant calls in data, resulting in a need for more stringent filters which may increase the risk of removing true-positives. Shortcomings in the reference genome will also increase the risk of missing true variants (i.e. false negatives). The human genome is more accurate than that of many other species and more resources are available to aid in the filtering of false-positive variants. Many reference genomes have a draft status and gaps and misassemblies are not uncommon (Kelley and Salzberg, 2010). Identifying misassembled regions in the reference genomes of non-human species and excluding them from analysis will help to reduce false-positives in variant calling data.

Whole genome sequencing (WGS) produces fairly consistent coverage across the genome (Belkadi et al., 2015), however, the PCR step in the Illumina library preparation pipeline is known to introduce bias, particularly in regions of high or low GC content (Kozarewa et al., 2009). Modifications have been introduced to protocols to reduce this bias (Aird et al., 2011), however, sequencing depth and quality in GC-rich and -poor regions remain unreliable when using protocols involving a PCR step. Previous work has shown that copy number variation can be accurately detected in WGS data by looking for areas of excessively high or low read counts following adjustment for GC content (Yoon et al., 2009;Zhang and Backstrom, 2014). To identify misassemblies in the chicken genome, a previous study used a pool of multiple birds to account for true variation between individuals, treating regions where all individuals show low read counts as false tandem duplications (Zhang and Backstrom, 2014).

In this paper, we look to identify low-confidence regions in the reference genome assembly Sscrofa10.2 using whole genome sequencing reads from T. J. Tabasco (Duroc 2-14), the Duroc sow whose DNA was used in the assembly (Groenen et al., 2012). The assembly was constructed using a BAC-by-BAC method, covers 18 autosomes and 2 allosomes (with the Y chromosome constructed separately from the DNA of male pigs), and contains many gaps and sequences on unplaced scaffolds. Ideally, an individual’s sequencing reads mapped to that individual’s own assembled genome would show no true structural variation and any areas of structural variation could be considered a misassembly. But the reference genome is a haploid representation and cannot reflect areas of true heterozygous structural variation accurately. However, a conservative approach would treat variant calls in these areas as low-confidence until further verified. Regions with no structural variation between the sequencing reads and the reference genome can be considered high-confidence.

In addition to using read count to detect potential duplications or collapses, we use other indicators to identify different kinds of structural variation such as inversions, deletions and insertions as has been done previously to identify potentially disease causing structural variation in human genomes (Tuzun et al., 2005). Illumina paired-end sequencing generates read-pairs from the same DNA fragment that are a known distance apart (usually following a normal distribution), and in a known orientation with respect to the reference genome. Therefore, when read pairs are mapped to a reference, if they are not in the expected orientation, or are an abnormal distance apart, this may also be an indication of errors in the assembly.

Finally, when mapping reads from the same animal to the reference genome created from that animal, there should be no homozygous variant calls.

In this work, regions with abnormally high or low coverage, with high proportions of reads with unexpected insert sizes or a high proportion of reads which were improperly paired were identified. In addition, SNP and indel calling was carried out. Regions were considered low quality (LQ) if they had high coverage, a high proportion of unexpected insert sizes or improperly paired reads or if they were in proximity to a homozygous variant. LQ regions are the most likely to represent misassemblies in the genome. Regions which had low coverage (LC) were analyzed separately; these regions may not be misassembled but have poor coverage and may therefore be unreliable for accurate variant calling. Both regions were also analyzed together in a combined dataset (LQLC).

Following identification of regions of the reference which may be unreliable, publicly available data sets were downloaded and overlap with the regions calculated. The data sets downloaded were the coding region, dbSNP variants, copy number variable regions (CNVRs) identified by Paudel et al. (2013) using a method that assesses read depth, and gene predictions based on data obtained using RNA-seq methods. These data sets allowed for identification of the proportion of the coding region overlapping the unreliable regions, and to assess how commonly used methods of SNP and indel calling, CNVR calling and RNA-seq may have been affected by unreliable regions of the genome assembly. We would expect the coding region to be under represented in the unreliable regions because these regions are generally more complex, which should make assembly more accurate. If the unreliable regions are enriched for calls in these datasets, it may suggest that analysis of these regions produces a higher level of false-positives than the rest of the genome.

Methods

Sample, sequencing and alignment

Eight sets of paired-end, whole-genome, Illumina sequencing reads from a single sample from T.J. Tobasco, the sow whose genome was used to construct Sscrofa10.2, were used in this study (These can be accessed from http://www.ebi.ac.uk/ena/data/view/ERP010190). BWA (v0.6.2: Li and Durbin, 2009) was used to align the reads to the Sscrofa10.2 reference assembly using default parameters. The reads were mapped to both the chromosomes and the unplaced scaffolds from the assembly. Any reads which mapped to chromosome Y were excluded as the sequences were from a female pig; consequently, we are unable to comment on the quality of the assembly of chromosome Y.

Identifying regions with abnormal coverage

SAMtools was used to filter the data to remove reads with a mapping quality less than 2 or which were improperly paired. BEDtools (v2.16.2: Quinlan and Hall, 2010) bamtobed was used to extract the chromosome, start positions and the end positions of whole sequencing fragments. BEDtools GenomeCov was then used to find per-base fragment coverage across the genome. BEDtools MakeWindows was used to make windows of 1000bp across the whole genome. Gap data was downloaded from the UCSC table browser (Karolchik et al., 2004) and BEDtools intersect was used to remove windows intersecting gap data. The median coverage for each 1000 base window across the genome was calculated. GC content is known to have a significant effect on coverage in sequencing methods that involve a PCR stage (Kozarewa et al., 2009). Coverage was normalized by GC content as described by Yoon et al. (2009). Briefly, the read coverage in each 1 Kb window (w) was adjusted by a multiplying factor f, with f equal to the ratio of the overall median across all windows divided by the median of all windows with the same GC percentage as that of the window w. Using the median instead of the mean prevented these values from being inflated by extreme outliers, as described by Zhang and Backstrom (2014). Any window with a normalized coverage over 55 or under 27 (2 std. from the mean; 41) was defined as having an abnormal coverage.

The removal of multimappers prior to coverage analysis may cause the detection of low coverage regions in certain sequence contexts in the genome that are more likely to contain multimappers (e.g. repetitive regions). Multimapped reads were extracted from the original bam file and read counts for these were calculated using Bedtools Coverage and the same 1000bp windows used in the above coverage analysis; additionally raw read counts for each window were calculated in the same way from the original bam file. The percentage of reads in each window which were multimapped was calculated. Windows with >50% multimapped reads are likely to have been identified as low coverage due to the removal of these reads before coverage analysis. The regions with >50% multimappers were merged and intersect with the LC regions was calculated using Bedtools.

Identifying regions with abnormal insert sizes

The mean and standard deviation of the insert sizes was calculated using Picard InsertSizeMetrics (v1.113: http://sourceforge.net/p/picard/wiki/Main_Page/). Insert sizes were considered abnormal if they were more than 2 standard deviations from the mean (427 bp). The merged BAM file was filtered for abnormally large (above 588 bp) and small (below 266 bp) insert sizes. BEDtools coverage was used to find the read count of the abnormal reads and the original BAM file using 1000 base windows with 200 overlap created with BEDtools MakeWindows. These data were used to calculate the percentage of abnormal reads in each window. A high proportion of small insert sizes was defined as a window with over 9.47% small insert sizes (2 std. deviations above the mean of 4.22%) and a high proportion of large insert sizes was defined as a window with over 1.86% large insert sizes (2 std. above the mean of 0.12%).

Identifying regions with a low proportion of properly paired reads

The mapped reads were filtered using SAMtools for the SAM flag 0x2, removing reads which were flagged as improperly paired. The percentage of properly paired reads was calculated as described for insert sizes. Any window with fewer than 70.59% (2std. below the mean of 92.5%) properly paired reads was considered abnormal.

Variant calling

SNPs and indels were called using SAMtools mpileup, BCFtools and vcfutils varFilter. The resultant vcf file was filtered for homozygous variants, indicative of errors in the reference genome or sequencing errors. In order to include the entire regions covered by reads overlapping each variant, the regions spanning from 100 bases before to 100 bases after each variant were considered low quality.

Merging

BEDtools was used to merge the regions identified by the above parameters into LQ, LC and LQLC regions. BEDtools intersect was used to find regions of each group which overlapped with the coding region (regions downloaded from UCSC table browser (Karolchik et al., 2004)). Sanger's gEVAL website (http://geval.sanger.ac.uk/index.html) was used to inspect BAC and fosmid end alignments in the identified regions.

Assessing effect of identified regions on public data

Known variant data were downloaded from dbSNP (Sherry et al., 2001) and the number of variants overlapping the abnormal regions were calculated. To assess the potential effect of these regions on WGS resequencing studies in pigs, the regions identified as CNVRs in Paudel et al. (2013) were downloaded and the number of regions overlapping the abnormal regions from the current study were calculated. Gene predictions based on RNA-seq data were downloaded from Ensembl (Cunningham et al., 2015) and the number of bases overlapping the identified regions calculated.

Results

Alignment

582,271,856 reads mapped to the reference and 94.66% of these were properly paired (551,173,366 reads).

Abnormal regions

The effect of GC content on median coverage was as expected, with both high and low GC content regions having poor median coverage (Figure 1).

While the coverage following GC normalization did follow a normal distribution, several extreme outliers inflated the mean and standard deviation. R (R Development Core Team, 2009) was used to find the mean and standard deviation of the majority of the data by overlaying a normal distribution on the data (Figure 1). Using this method, we determined the mean coverage to be 41X and the standard deviation to be 7.

Regions identified by the parameters measured are summarized in table 1. In total, 2.6% of the genome had abnormally high coverage, and 26.6% of the genome abnormally low coverage. Regions with a high percentage of fragment pairs with abnormally low and high insert sizes cover 3.99% and 1.52% of the genome respectively. Regions with a low percentage of properly paired reads cover 4.95% of the genome. One of the largest regions identified (77.8Kb) has abnormal coverage, insert sizes and read orientation (Figure 2A), and this is not uncommon, further examples are shown in figures 2B and 2C.

There were a total of 62,463 regions with >50% multimappers and of these 99.3% overlapped with the LC regions. 66% of the regions identified as LC overlapped with the multimapped regions. The remaining LC regions had an unremarkable distribution of GC contents (data not shown) and the majority (81%) had 0 multimappers. The median read count per window for the whole genome was 264 and the median read count per window for the LC regions excluding those with >50% multimappers was 161.
We identified a total of 583,093 homozygous variants. Following merging, there were 245,972 regions identified as abnormal due to proximity to these variants covering 63,085,828 bases (2.25% of the genome).

Merged regions

After merging the regions with abnormal insert sizes, abnormal read orientation and homozygous variant calls, we were left with 409,905 regions identified as being LQ, covering 13.85% of the genome.

In total, 337,276 regions were identified as being LQLC and the regions covered a total of 928,664,896 bases (33.07% of the genome). If the multimapped regions are excluded from the LC regions and the remaining LC regions are merged with the LQ regions these cover 17.3% of the genome.
The coding region data downloaded from UCSC table browser covered 587,219,382 bases (excluding chromosome Y) and of these 81,566,904 (13.89%) intersected with the LQ regions.

Of the coding region, 154,875,678 bases (26.37%) intersected with the LQLC regions.

Impact on public data

The proportion of variants from real data sets from Paudel et al. (2013) and dbSNP (Sherry et al., 2001) that fall in the abnormal regions are summarized in table 2.

Paudel et al. (2013) identified 61,761 multi-copy regions (MCR), and from these identified 3,118 CNVRs. Of the CNVRs 1,081 (34.66%) lie in the LQ regions and 2,692 (86.3%) lie in the LQLC regions identified here.

The data downloaded from dbSNP (Release 104. Accessed: 05/05/2015) contain 52,634,111 known variants. In total, 19,121,760 (36.33%) dbSNP variants were located in the LQ regions, 15,483,445 (29.42%) dbSNP variants were located in the LC regions and 27,009,232 (51.3%) dbSNP variants were located in the LQLC regions.

The gene predictions based on RNA-sequencing data covered 41,788,900 bases, 26.69% of these bases were in the LQ region (11,155,280), 27.19% were in the LC region (11,360,980) and 42.98% were in the LQLC regions (17,959,798).

Discussion

This work emphasizes the importance of accuracy in reference genomes in variant discovery research. Previous work by Zhang and Backstrom (2014) used sequencing reads from multiple chickens to detect misassemblies in the chicken genome. Here we used data from the same individual used to construct the pig reference assembly. We are therefore able to assess the assembly without introducing potential true variation that may be present by chance in multiple individuals; however, regions of the genome may have been incorrectly identified as low-quality due to true structural variation at heterozygous sites.

Regions of Sscrofa10.2 identified in this study were enriched for variants from dbSNP. The fact that the regions identified were enriched for variants in dbSNP, with the LQLC regions containing over half of the dbSNP variants, supports the assertion that these regions are enriched for false-positives; dbSNP contains large numbers of SNPs that are not validated and are potentially false-positives (Mitchell et al., 2004;Musumeci et al., 2010).

In the CNVR study by Paudel et al. (2013), 61,761 MCRs were identified and the authors state that the majority of these were common in all individuals sequenced; in this study 60,281 regions were identified as having high coverage and it is likely that there is overlap between these results. Studies looking for copy number gains may benefit from excluding the LQ regions from analysis. From the MCRs, 3,118 CNVRs were identified. The authors estimated that of these 2,664 (85.43%) were likely to be neutral or nearly neutral as they were common between different groups or were in non-genic regions, which is very similar to the number of CNVRs in the data that overlap the LQLC regions in the current study (2,692; 86.3%). CNVRs are called from sequencing data by comparison of read counts for a region with the average across the genome; it is likely that there are many false tandem repeats or collapsed repetitive regions in the assembly that would cause false copy number loss or gain calls. While regions identified as CNVRs are potentially variable regions between populations, breeds and individuals, calls based solely on comparison with the reference will give false-positives and false estimates of the copy numbers in true variable regions. Paudel et al. (2013) used copy number comparisons between individuals from different populations to identify MCRs that were variable between groups, which likely removed the majority of the false-positives. Other studies have used array-based methods to detect CNVRs in the pig genome (Chen et al., 2012;Wang et al., 2012) and of the regions identified in these studies, almost all of them fall in the LQLC regions (data not shown). This suggests these regions truly are enriched for CNVRs; however, enrichment of the unreliable regions for CNVRs may also suggest unreliable assembly around large duplications. In studies using whole genome resequencing, often small sample sizes are used and too much confidence may be given to the reference. It would be advisable in studies using Sscrofa10.2, and references of other species that may contain similar inaccuracies, not to call CNVRs based solely on comparison with the reference, but from regional variation in read count between individuals as has been done previously for genomes which lack a reference following co-assembly (Nijkamp et al., 2012) and when comparing sequences from cancer cells to healthy cells (Chiang et al., 2009;Koboldt et al., 2012). Similarly, researchers using other techniques that rely on counting reads mapped to the reference genome such as ChiP-seq and RNA-seq should be aware that these errors may cause inaccurate calling or expression estimates. In RNA-seq, read counts are used to estimate expression levels; unexpected CNV between the reference and the sample sequence could cause over- or under-exaggerated read counts, potentially resulting in false-positives or false-negatives. RNA-seq is prone to off-target mapping (Mortazavi et al., 2008), particularly at higher depth (Tarazona et al., 2011); true peaks can often be distinguished from off-target mapping using an expression threshold. However, misrepresentation of the copy number of a region in the reference assembly may exaggerate off-target peaks above the threshold and cause false-positives, exaggerate true peaks causing inaccurate expression estimates, or reduce true peaks causing false-negatives or underestimation of expression. The regions identified here were enriched for RNA-seq gene predictions, more so than the annotated coding region, which may suggest an increased false-positive rate in these regions from this method.

A large amount of the genome showed low coverage. While these regions may suggest errors in the reference genome, such as false tandem duplications (Zhang and Backstrom, 2014), they do so with less confidence than the other parameters measured. The study by Paudel et al. (2013) reported a considerable number of copy number losses and subsequently excluded these from further analysis as they were likely enriched for false-positives; the fact that this excess of low coverage regions has been encountered by other researchers may suggest that the problem is with the quality of the genome assembly or region mappability rather than the quality of the data used in the current study. Regions with low coverage were analyzed separately as low coverage may be an indicator of the quality of the sequencing data, PCR bias or poor mappability and not necessarily inaccuracy in the reference. The majority of the LC regions were explained by their large proportion of multimappers; the regions were identified as low coverage because multimappers were excluded from the coverage analysis. These regions may not be misassembled, but rather of poor mappability due to, for example, low complexity or repetitive sequences. Of the LC regions which were not explained by multimappers there was no evidence of extreme GC content causing the reduced coverage and the majority contained no multimappers; the low coverage in these regions likely relates to misassembled areas in the reference genome, or potentially heterozygous structural variants in the individual. Where the low coverage is explained by poor mappability, it may still be advisable to exclude these regions from SNP and indel analyses as this is likely to yield low quality variants with a high rate of false-positives. Studies requiring identification of only the highest quality variants would reduce computational burden and false-positive rate by excluding the LC regions. In studies more concerned with finding variants relating to a specific phenotype, if LC regions are included, variants identified in them may be treated as low-confidence, but not necessarily excluded entirely. The percentage of dbSNP variants in the LC region is not as high as in the LQ region, however fewer variants may be called in poor mappability regions due to the common practice of filtering out low mapping quality reads before proceeding to variant calling, reducing depth and subsequently the chances of calling a variant in these regions. The proportion of the genome identified here as LQ is likely to be an over-estimation of the proportion that is misassembled. The individual may have true, heterozygous structural variation that cause some of these regions to appear misassembled and this analysis has been intentionally strict to allow downstream bioinformatic analysis to focus on only the highest confidence regions of the genome by excluding LQLC regions. The number of variants identified in studies employing variant calling is often extreme and strict filtering techniques are employed to reduce the number to a more tractable level for validation (MacArthur et al., 2012; Ai et al., 2015). Excluding regions which are likely to be enriched for false-positives may significantly reduce computational burden and increase accuracy. Strict filtering after variant calling may cause the loss of variants of interest and it is desirable to reduce the initial number of variant calls as much as possible to reduce the need for excessive filtering. While variants of interest may lie in the low-confidence regions identified here, the excess of false-positives in the region make it unlikely that they will be easily identified. However, discovery of variants outside of these regions will benefit from the reduced number of false-positives in the dataset. Many variant callers and filtration methods will consider depth and mapping quality and are likely to exclude a number of false-positive variants from these regions by default; however, computational burden would be decreased by excluding unreliable regions, which will be particularly relevant with large datasets. . Other methods that use regional read count data need to be aware that Sscrofa10.2 does not accurately represent the copy number of alleles in the original Duroc sow's genome. Clearly in studies searching for CNVRs, excluding the LQLC regions, which are potentially enriched for true CNVRs, is not an option. In such studies it would be beneficial to compare individuals in a study with one another rather than with the reference, as is done in somatic variant calling comparisons between healthy cells and cancer cells (Roberts et al., 2013), to filter out variation that is common in all individuals, or to exclude the LQ regions only. The degree to which misassemblies will affect research results depends on a number of factors including the tools used, the type of misassembly and the type of analysis; for example, the incorrect order of contigs will negatively affect read-pair mapping and collapsed duplications may cause incorrect calling of SNPs - though SNP callers may accurately filter many of these. Similar inaccuracies to those found here are likely to be present in the reference genomes of other non-human species. With the price of sequencing continuing to fall, the number of large-scale sequencing studies on species with draft genomes will undoubtedly increase; awareness of inaccuracies in these references will decrease computational burden and increase accuracy. Identifying regions that are inaccurate and producing new, more accurate assemblies will greatly increase the power of whole-genome resequencing studies in non-human species.

Availability of Data

The regions identified in this study have been made available as three bed files: LQ regions, LC regions and LQLC regions. BED files are available to download from http://www.ark-genomics.org/outputs/identification-low-confidence-regions-pig-reference-genome-sscrofa102

Conflicts of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding

This work was funded by grants from the Roslin Foundation, Genus Plc and by a BioSciences KTN CASE studentship. The work was enabled by funding from the Biotechnology and Biological Sciences Research Council including Institute Strategic Programme and National Capability grants (BBSRC; BBS/E/D/20310000, BB/J004243/1). DNA sequencing was provided by Edinburgh Genomics (http://genomics.ed.ac.uk). Edinburgh Genomics is partly supported through core grants from the National Environmental Research Council (NERC R8/H10/56), Medical Research Council (MRC MR/K001744/1) and The Biotechnology and Biological Sciences Research Council (BBSRC BB/J004243/1).

References

Ai, H., Fang, X., Yang, B., Huang, Z., Chen, H., Mao, L., et al. (2015). Adaptation and possible ancient interspecies introgression in pigs identified by whole-genome sequencing. *Nat Genet*. 10.1038/ng.3199

Aird, D., Ross, M.G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., et al. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology* 12**,** R18-R18. 10.1186/gb-2011-12-2-r18

Belkadi, A., Bolze, A., Itan, Y., Cobat, A., Vincent, Q.B., Antipenko, A., et al. (2015). Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proceedings of the National Academy of Sciences*. 10.1073/pnas.1418631112

Bianco, E., Nevado, B., Ramos-Onsins, S.E., and Pérez-Enciso, M. (2015). A Deep Catalog of Autosomal Single Nucleotide Variation in the Pig. *PLoS ONE* 10**,** e0118867. 10.1371/journal.pone.0118867

Chen, C., Qiao, R., Wei, R., Guo, Y., Ai, H., Ma, J., et al. (2012). A comprehensive survey of copy number variation in 18 diverse pig populations and identification of candidate copy number variable genes associated with complex traits. *BMC Genomics* 13**,** 733-733. 10.1186/1471-2164-13-733

Chiang, D.Y., Getz, G., Jaffe, D.B., O'kelly, M.J.T., Zhao, X., Carter, S.L., et al. (2009). High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Meth* 6**,** 99-103. 10.1038/nmeth.1276

Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., et al. (2015). Ensembl 2015. *Nucleic Acids Research* 43**,** D662-D669. 10.1093/nar/gku1010

Groenen, M.A., Archibald, A.L., Uenishi, H., Tuggle, C.K., Takeuchi, Y., Rothschild, M.F., et al. (2012). Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* 491**,** 393-398. 10.1038/nature11622

Gudbjartsson, D.F., Helgason, H., Gudjonsson, S.A., Zink, F., Oddson, A., Gylfason, A., et al. (2015). Large-scale whole-genome sequencing of the Icelandic population. *Nat Genet* advance online publication. 10.1038/ng.3247

Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., et al. (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Research* 32**,** D493-D496. 10.1093/nar/gkh103

Kelley, D.R., and Salzberg, S.L. (2010). Detection and correction of false segmental duplications caused by genome mis-assembly. *Genome Biology* 11**,** R28-R28. 10.1186/gb-2010-11-3-r28

Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., Mclellan, M.D., Lin, L., et al. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research* 22**,** 568-576. 10.1101/gr.129684.111

Kozarewa, I., Ning, Z., Quail, M.A., Sanders, M.J., Berriman, M., and Turner, D.J. (2009). Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of GC-biased genomes. *Nature methods* 6**,** 291-295. 10.1038/nmeth.1311

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

Macarthur, D.G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., et al. (2012). A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. *Science* 335**,** 823-828. 10.1126/science.1215040

Mitchell, A.A., Zwick, M.E., Chakravarti, A., and Cutler, D.J. (2004). Discrepancies in dbSNP confirmation rates and allele frequency distributions from varying genotyping error rates and patterns. *Bioinformatics* 20**,** 1022-1032. 10.1093/bioinformatics/bth034

Mortazavi, A., Williams, B.A., Mccue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Meth* 5**,** 621-628. http://www.nature.com/nmeth/journal/v5/n7/suppinfo/nmeth.1226_S1.html

Musumeci, L., Arthur, J.W., Cheung, F.S.G., Hoque, A., Lippman, S., and Reichardt, J.K.V. (2010). Single Nucleotide Differences (SNDs) in the dbSNP Database May Lead to Errors in Genotyping and Haplotyping Studies. *Human mutation* 31**,** 67-73. 10.1002/humu.21137

Nijkamp, J.F., Van Den Broek, M.A., Geertman, J.-M.A., Reinders, M.J.T., Daran, J.-M.G., and De Ridder, D. (2012). De novo detection of copy number variation by co-assembly. *Bioinformatics* 28**,** 3195-3202. 10.1093/bioinformatics/bts601

Paudel, Y., Madsen, O., Megens, H.-J., Frantz, L.a.F., Bosse, M., Bastiaansen, J.W.M., et al. (2013). Evolutionary dynamics of copy number variation in pig genomes in the context of adaptation and domestication. *BMC Genomics* 14**,** 449-449. 10.1186/1471-2164-14-449

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26**,** 841-842. 10.1093/bioinformatics/btq033

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing* [Online]. Vienna, Austria : the R Foundation for Statistical Computing. Available: http://www.R-project.org/ [Accessed 01/05/2015].

Robert, C., Fuentes-Utrilla, P., Troup, K., Loecherbach, J., Turner, F., Talbot, R., et al. (2014). Design and development of exome capture sequencing for the domestic pig (Sus scrofa). *Bmc Genomics* 15**,** 9. 10.1186/1471-2164-15-550

Roberts, N.D., Kortschak, R.D., Parker, W.T., Schreiber, A.W., Branford, S., Scott, H.S., et al. (2013). A comparative analysis of algorithms for somatic SNV detection in cancer. *Bioinformatics* 29**,** 2223-2230. 10.1093/bioinformatics/btt375

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., et al. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research* 29**,** 308-311.

Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A., and Conesa, A. (2011). Differential expression in RNA-seq: A matter of depth. *Genome Research* 21**,** 2213-2223. 10.1101/gr.124321.111

Tuzun, E., Sharp, A.J., Bailey, J.A., Kaul, R., Morrison, V.A., Pertz, L.M., et al. (2005). Fine-scale structural variation of the human genome. *Nat Genet* 37**,** 727-732. http://www.nature.com/ng/journal/v37/n7/suppinfo/ng1562_S1.html

Wang, J., Jiang, J., Fu, W., Jiang, L., Ding, X., Liu, J.-F., et al. (2012). A genome-wide detection of copy number variations using SNP genotyping arrays in swine. *BMC Genomics* 13**,** 273-273. 10.1186/1471-2164-13-273

Yoon, S., Xuan, Z., Makarov, V., Ye, K., and Sebat, J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res* 19**,** 1586-1592. 10.1101/gr.092981.109

Zhang, Q., and Backstrom, N. (2014). Assembly errors cause false tandem duplicate regions in the chicken (Gallus gallus) genome sequence. *Chromosoma* 123**,** 165-168. 10.1007/s00412-013-0443-8

Table 1: Table summarizing the regions identified by different parameters measured

	No. of features	**Mean feature size**	**Percentage of genome**
High coverage	60,281	1,202	2.6%
Small insert	82,097	1,363	3.99%
Large insert	31,833	1,343	1.52%
Improperly paired	77,785	1,786	4.95%
Homozygous variants	245,972	256	2.25%
Low Quality (LQ)	409,905	949	13.85%
Low coverage (LC)	119,251	6,275	26.6%

Total (LQLC)	337,276	2,753	33.07%

Table 2: Table summarizing the proportion of called variants in real data that fall in the abnormal regions identified in the current study

	Total	**Low Quality (LQ)**	**Low Coverage (LC)**	**Combined (LQLC)**
% of genome	-	13.85%	26.6%	33.07%
% of coding region	-	13.89%	17.72%	26.37%
dbSNP variants[a]	52,634,111	19,121,760 (36.33%)	15,483,445 (29.42%)	27,009,232 (51.3%)
CNVRs[b]	3,118	1,081 (34.66%)	1,706 (54.71%)	2,692 (86.3%)
RNA-seq genes[c] **(intersecting bases)**	41,788,900	11,155,280 (26.69%)	11,360,980 (27.19%)	17,959,798 (42.98%)

[a] Data from dbSNP database (Sherry, et al., 2001)

[b] Data from Paudel, et al. (2013)

[c] Data from Ensembl (Cunningham, et al., 2015)

Figure 1: Plot showing median coverage of windows against percentage of GC content (left). Histogram showing the distribution of window coverage, red line represents a normal distribution (right).

Figure 2: Plots showing examples of abnormal regions for multiple parameters on Chromosomes 6 (A), 12 (B) and X (C) (top). GC normalized coverage (red) uses the left Y axis. Percentage of properly paired reads (blue) and percentage of high insert sizes (green) use the right Y axis. Means are represented by solid lines and 2 std from the mean are represented by dashed lines. Bottom shows same regions viewed on the gEVAL browser with poorly mapped fosmids (top) and bac ends (bottom) shown in red.

Figure 1.JPEG

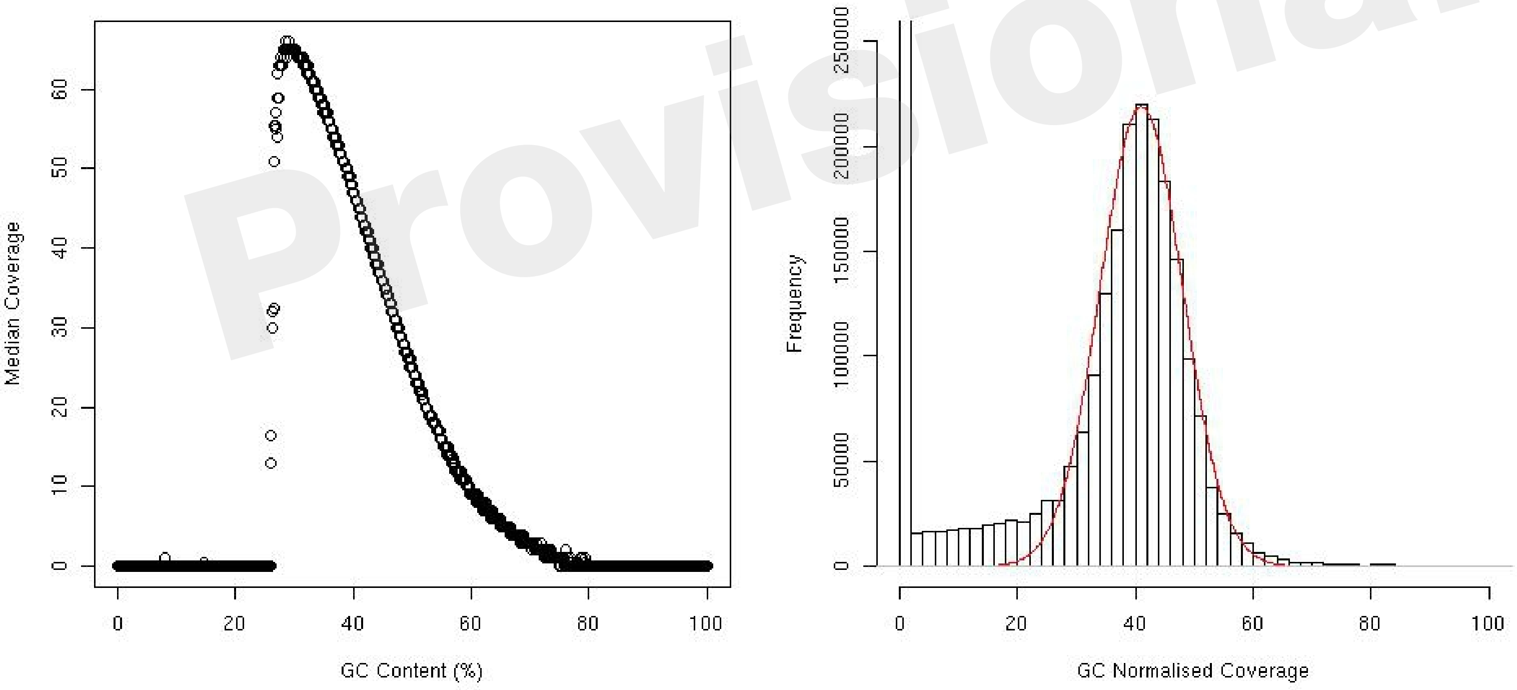
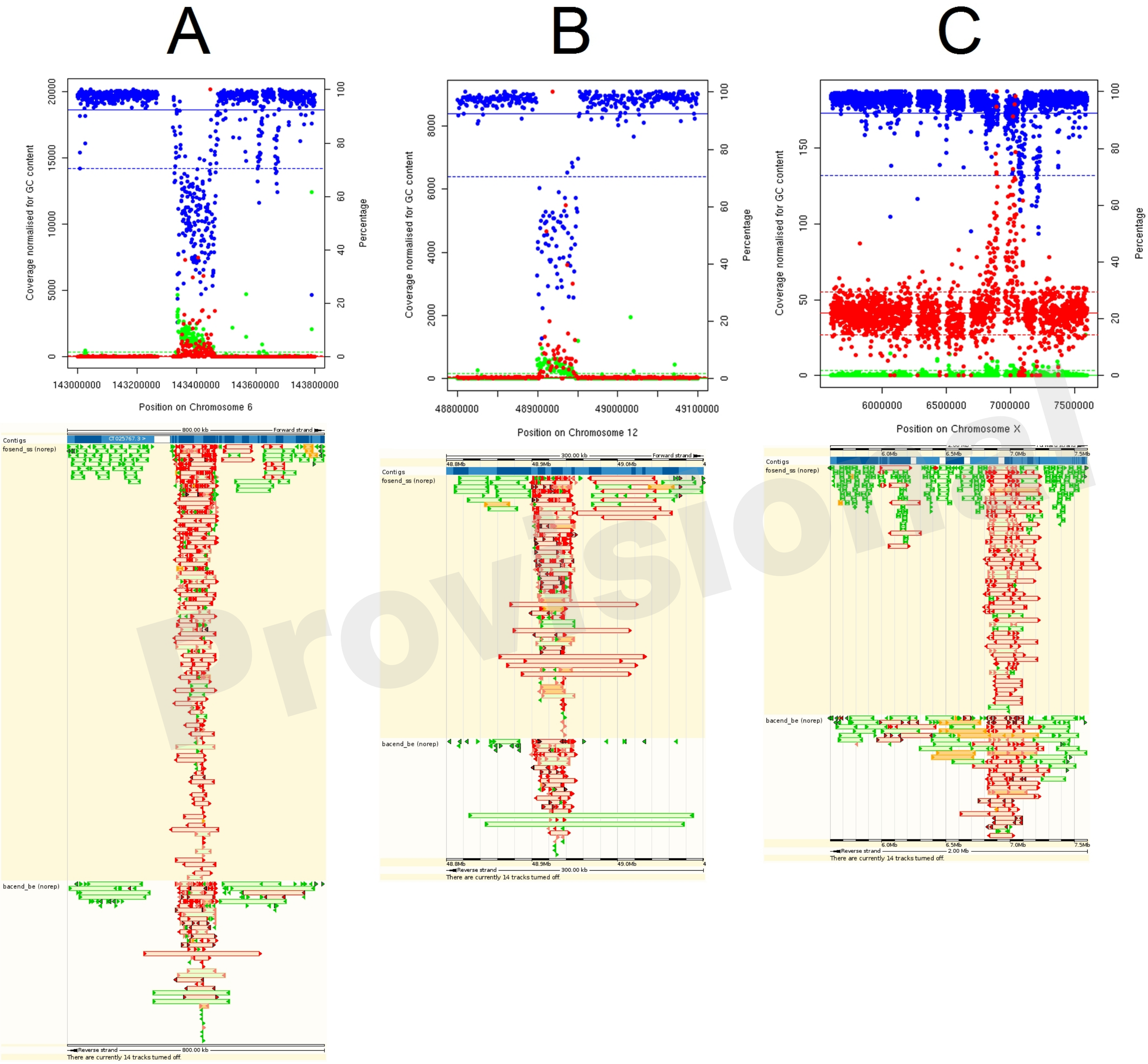


Figure 2.JPEG


A
B
C
Coverage normalised for GC content
Percentage
Position on Chromosome 6
Position on Chromosome 12
Position on Chromosome X
800.00 kb
Forward strand
Contigs
CT025767.3 >
fosend_ss (norep)
bacend_be (norep)
Reverse strand
There are currently 14 tracks turned off.
300.00 kb
48.8Mb
48.9Mb
49.0Mb
2.00 Mb
6.0Mb
6.5Mb
7.0Mb
7.5Mb