

Building *in vitro* tools for livestock genomics: chromosomal variation within the porcine PK15 cell line

M. Johnsson¹, J.M. Hickey² and M. K. Jungnickel^{2*}

¹ Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Box 7023, 750 07 Uppsala, Sweden; ² The Roslin Institute and Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Midlothian, EH25 9RG, Scotland, UK; *melissa.jungnickel@roslin.ed.ac.uk

Abstract

In this work, we called variants and used read coverage in combination with within-sample allele frequency to detect potential aneuploidy in two samples of an immortalised Porcine Kidney Epithelial (PK15) cell line. Porcine cell culture and cell lines are powerful tools for functional genomics and *in vitro* phenotypic testing of candidate causal variants. However, to be utilised for any type of genomic or variant interrogation assay, the genome sequence and structure of cell lines must be realised. Read coverage and within-sample allele frequencies suggest that several chromosomes are fully or partially aneuploid in PK15, including potential triploidy of chromosome 4 and tetraploidy of chromosome 17. We compared two PK15 cultured cells samples: a new American Type Culture Collection (ATCC) sample and one that has been utilised within the laboratory for an extended period (>10 years). The older PK15 sample showed evidence of additional structural variation and potentially clonal variation. We discuss implications for livestock functional genomics.

Introduction

Cell lines are an essential tool for functional genomics and functional studies of potential causative genetic variants. Large-scale epigenomic projects like ENCODE rely heavily on cell lines as models of different cell types, and functional work on causative genes in livestock make use of the relatively few livestock cell lines that exist (e.g. Walker et al., 2018).

Most CRISPR screens are conducted *in vitro*. This approach requires an exquisitely precise and detailed knowledge of the genome sequence of target cell lines. It is well established that such sequence details are essential prerequisites to design guide RNAs and direct Cas9/dCas9 function in CRISPR manipulation and perturbation assays. Less well known is the contribution of genome structure and cell line aneuploidy to CRISPR-genome interrogation. Aneuploidy, in which cells have an abnormal number of chromosomes, has been shown to have a devastating impact on cell phenotype and gene expression (Sheltzer et al., 2012). To our knowledge, an impact of aneuploidy on CRISPR gene manipulation has not been previously considered.

For this project, we used short read sequencing to characterise the genome of PK15, a classic pig cell line established in 1955 or 1956 from the kidney of an adult pig (Harris, 1959; Ruddle, 1961). We called variants and used read coverage in combination with within-sample allele frequency to detect potential aneuploidy in two samples of the PK15 cell line.

Materials & Methods

Samples and sequencing. We extracted DNA from two *in vitro* cultured PK15 cell lines. One PK15 cell line was ordered fresh from ATCC (product name: PK(15) CCL-33) and the other were PK15 cells that have been cultured at the Roslin Institute for > 10 years. Whole-genome

resequencing was performed on the Illumina platform by Genewiz (Essex) for the ATCC sample and Edinburgh Genomics (Edinburgh) for the Roslin sample.

Bioinformatics. We mapped reads to the Sscrofa11.1 reference genome (Warr et al., 2020) with BWA MEM version 0.7.12 (Li, 2013) after trimming adapter and low-quality ends with Trimmomatic version 0.36. We removed duplicate alignments with Picard version 2.9.0 and called variants with GATK 3.5 (McKenna et al., 2010) using the HaplotypeCaller and the GVCF workflow. We performed base quality score recalibration by first calling a preliminary set of variants, then using it as input to the GATK base quality score recalibration tool, and then calling a final set of variants using the recalibrated alignments. We filtered single nucleotide variants using the standard GATK hard filters. In order to estimate the allele frequencies within each sample, we extracted the allelic depth for each biallelic variant with bcftools version 1.10.2. In order to detect large-scale copy number variation and aneuploidy, we divided the pig genome chromosomes 1-18, X and Y into 10 kbp windows, and counted the number of reads mapping in each window with BEDTools version 2.26.0.

Results

The sequencing resulted in approximately 30X (ATCC sample) and 40X (Roslin sample) coverage of the pig genome. Variant calling detected 9,028,359 single nucleotide variants. Out of those, 3,565,539 were called as homozygous alternate in both samples, that is, they represent positions where the PK15 cell lines is likely to be fixed for a different allele than the Sscrofa11.1 reference genome.

Read coverage per chromosome suggest that several chromosomes are aneuploid in PK15. Figure 1 shows the average read coverage for a 10 kbp window on chromosomes 1-18, X and Y. In both samples, chromosomes 4, 9 and 17 have high read coverage. There are also differences between samples, such as high read coverage on chromosome 12 in the Roslin sample, not seen in the ATCC sample. When it comes to the sex chromosomes, the ATCC sample appears to be XX, consistent with the literature (Ruddle, 1961), while the Roslin sample appears to have lost one X chromosome and be X0.

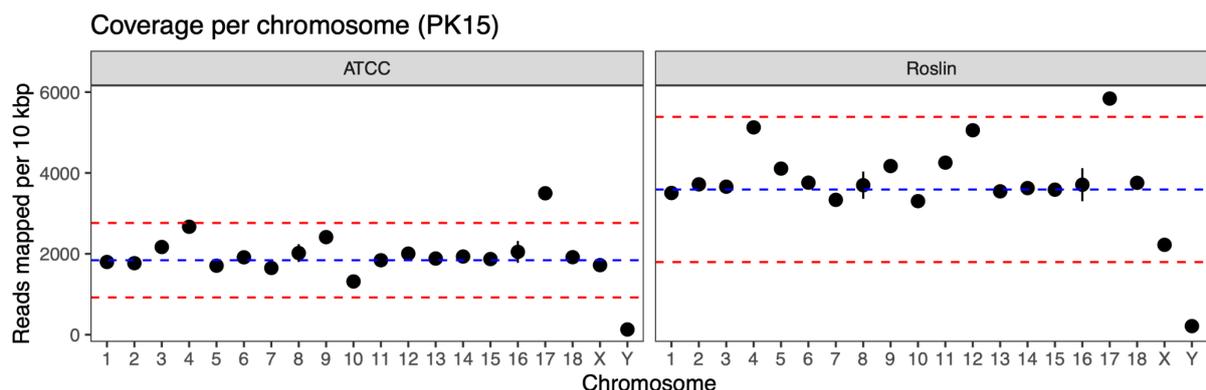


Figure 1. Average read coverage for 10 kbp windows on chromosomes 1-18, X and Y in the PK15 genome. Panels show the two samples of the cell line. Horizontal lines indicate the median coverage for a 10 kbp window (blue) as well as 0.5 x and 1.5 x the median (in red).

Read coverage within chromosomes also suggest that there are large-scale structural variants within chromosomes. Figure 2 shows the read coverage in 10 kbp windows within four

selected chromosomes. Chromosome 1, which shows little evidence of gross aneuploidy at the whole chromosome level (Figure 1), still potentially has a duplication within the Roslin sample, at around 200 Mbp. Chromosome 4 appears mostly triploid, but with several regions of higher and lower coverage especially in the Roslin sample. Chromosome 17 appears to be tetraploid in the ATCC sample and triploid in the Roslin sample.

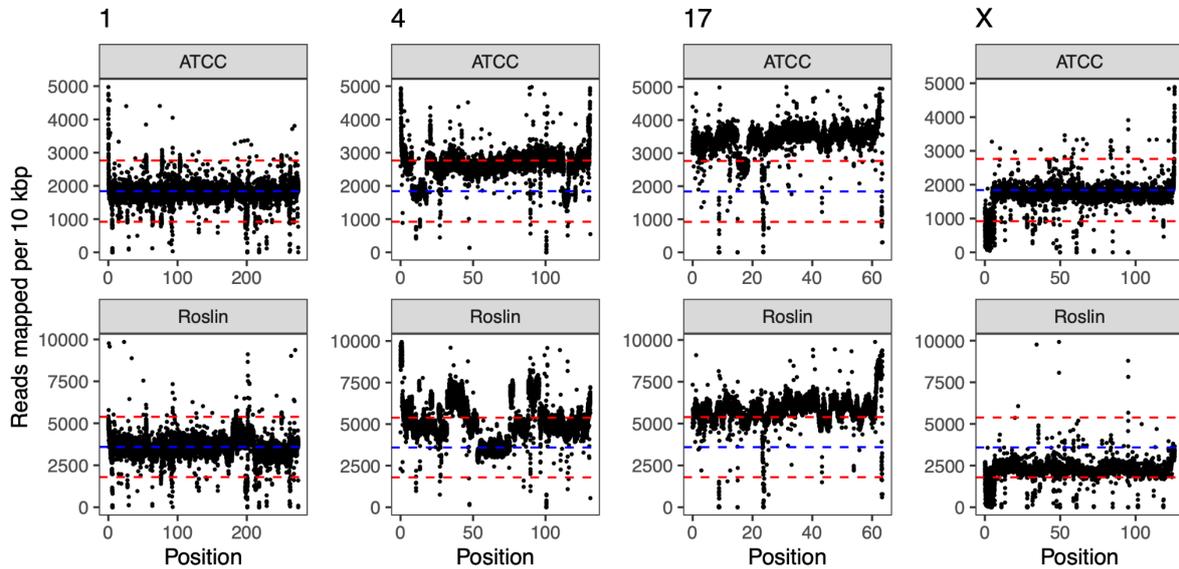


Figure 2. Read coverage in 10 kbp windows on chromosomes 1, 4, 17 and X in the PK15 genome. Panels show the two sample of the cell line. Horizontal lines indicate the median coverage for a 10 kbp window (blue) as well as 0.5 x and 1.5 x the median (in red).

Within-sample allele frequencies (Figure 3) for heterozygous sites also suggest aneuploidy. We find examples of chromosome that are likely diploid, triploid, tetraploid and haploid:

- For a diploid chromosome, one would expect a frequency of $1/2$. Considering chromosome 1, the ATCC sample appears mostly diploid. In the Roslin sample, however, there are two distinct modes of allele frequency, suggesting that some part of the cell line has an aneuploid karyotype.
- For a triploid chromosome, allele frequencies should instead be $1/3$ and $2/3$. In both samples, chromosome 4 shows two allele frequency modes, suggesting triploidy.
- For a tetraploid chromosome resulting from a duplication of both homologous chromosomes, the expected allele frequency is $1/2$. On chromosome 17, the ATCC has allele frequency centred on 0.5 , while the Roslin sample shows two modes. As they both have high read coverage (Figure 2), we hypothesise that chromosome 17 is tetraploid in the ATCC sample and mostly triploid in the Roslin sample.
- For a haploid chromosome, only frequencies 0 and 1 are possible. In the Roslin sample, the X chromosome has very few heterozygous sites, suggesting that it is haploid.

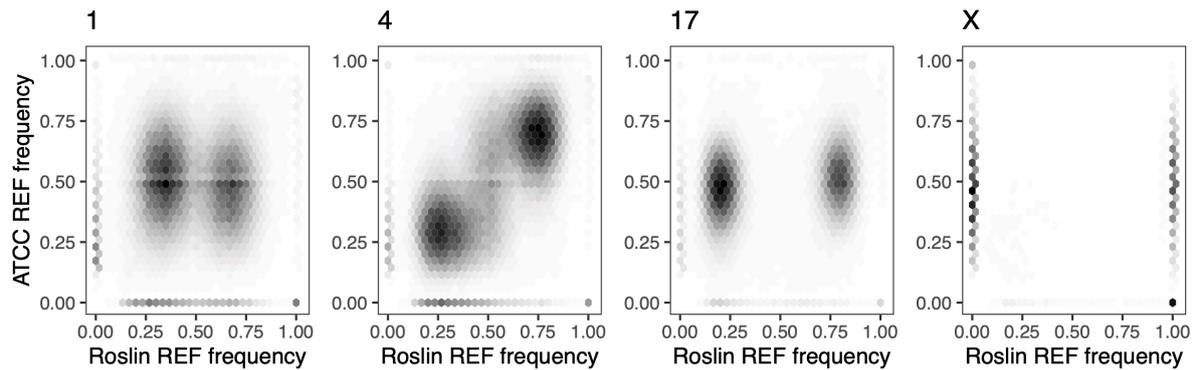


Figure 3. Within-sample allele frequency for the reference allele, calculated from allelic read counts, for single nucleotide variants that are heterozygous in at least one of the samples on chromosomes 1, 4, 17 and X.

Discussion

Our results suggest that the PK15 cell line has aneuploidies and complex structural variants in its genome. While such genome anomalies have not impacted over-expression studies that have been popular for the last 10-15 years (see review by (Prelich, 2012)), this cell line aneuploidy and clonal variation will directly impact CRISPR genome editing/genome manipulation studies; a genome editing assay with 20% homology-directed repair efficiency would generate repair rates of 4% in diploid and 0.8% in triploid genome sites respectively.

We are currently generating a more thorough bioinformatic to detect structural variants from short reads and segment the genome by ploidy. However, to fully describe the genome structure of a cell line, both long read sequencing and genome assembly is required. While currently too time consuming and expensive for routine cell line characterization, a method to examine cell line aneuploidy in more detail should be considered for key, routinely genome-manipulated and edited cell lines.

In summary, we suggest that genome aneuploidy should be considered when planning genome editing assays in *in vitro* cell lines.

References

- Harris, M. (1959) *Cancer Res* 19, 1020–1024.
- Li, H. (2013) arXiv:1303.3997. <https://arxiv.org/abs/1303.3997>
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, et al. (2010) *Genome Res* 20, 1297–1303. <https://doi.org/10.1101/gr.107524.110>
- Prelich, G. (2012) *Genetics* 190, 841–854. <https://doi.org/10.1534/genetics.111.136911>
- Ruddle, F.H. (1961) *Cancer Res* 21, 885.
- Sheltzer, J.M., Torres, E.M., Dunham, M.J., Amon, A. (2012) *Proc Natl Acad Sci USA* 109, 12644–12649. <https://doi.org/10.1073/pnas.1209227109>
- Walker, L.R., Engle, T.B., Vu, H., Tosky, E.R., Nonneman, et al. (2018) *PLoS Genet* 14, e1007750. <https://doi.org/10.1371/journal.pgen.1007750>
- Warr, A., Affara, N., Aken, B., Beiki, H., Bickhart, D.M., et al. (2020) *GigaScience* 9, giaa051. <https://doi.org/10.1093/gigascience/giaa051>