550. Extensive functional genomics information from early developmental time points for pig and chicken

H. Acloque[1], P.W. Harrison[2], W. Lakhal[3], F. Martin[2], A.L. Archibald[4], M. Beinat[1], M. Davey[4], S. Djebali[5], S. Foissac[6], S. Guizard[4], C. Guyomar[6], R. Kuo[4], C. Kurylo[6], O. Madsen[7], K. Miedzinska[4], M. Mongellaz[1], J. Smith[4], J. Smith[4], A. Sokolov[2], J. de Vos[7], E. Giuffra[1] and M. Watson[4*]

[1]Paris-Saclay University, INRAE, AgroParisTech, GABI, 78350 Jouy-en-Josas, France; [2]EMBL – European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, CB10 1SD, United Kingdom; [3]Epigenetics R&D, Diagenode S.A., Liège Science Park, Ru du Bois Saint-Jean 3, 4102 Liège, Belgium; [4]The Roslin Institute and R(D)SVS, University of Edinburgh, Easter Bush, EH25 9RG, United Kingdom; [5]IRSD, Université de Toulouse, INSERM, INRA, ENVT, UPS, 31024 Toulouse, France; France; [6]GenPhySE, Université de Toulouse, INRAE, ENVT, 31326 Castanet Tolosan, France; [7]Animal Breeding and Genomics, Wageningen University and Research, 6700 AH Wageningen, the Netherlands; mick.watson@roslin.ed.ac.uk

Abstract

The global Functional Annotation of Farm Animal Genomes initiative (FAANG) aims to improve animal breeding by improved genomic prediction via integration of functional genomics information. The GENE-SWitCH project has produced extensive functional genomics information for a variety of important tissues at early embryonic timepoints for both chickens and pigs. These datasets will be integrated to produce both tissue and time-point specific transcript, gene, and regulatory annotation for both species. In this paper, we describe the aims of the project, and the initial release of both raw and processed data.

Introduction

Genome annotation is often presented in a static, linear, fashion, as a set of genes, transcripts and regulatory regions which exist on the genome. However, we know that gene, transcript, and regulatory regions differ in both a time- and tissue- specific manner, hence current genome annotation data sets may miss important temporal and tissue-specific differences in the expression and use of functional genomic regions, and these regions may underlie key traits in livestock species. Addition of functional genomics information is therefore expected to improve animal breeding through improved genomic prediction (Andersson *et al.* 2015, Clark *et al.* 2020).

'The regulatory GENomE of SWine and CHicken: functional annotation during development (GENE-SWitCH)' project contributes to the global FAANG initiative and to EuroFAANG (https://eurofaang.eu), a synergy of five Horizon 2020 projects that share the common goal to discover links between genotype to phenotype in farmed animals and meet global FAANG objectives. A core activity of GENE-SWitCH is to generate extensive functional genomics information for key tissues during early development of both chickens and pigs, with the overall aim of improving genomic selection via more accurate genomic prediction using models that incorporate knowledge of functional regions of the genome.

Tissue- and time-point specific genome annotations are generated by standardized approaches (sampling, functional assays and analyses) to identify genes (coding and non-coding), transcripts, and regulatory sequences.

This paper constitutes a 'marker' paper of GENE-SWitCH in the context of the Toronto Statement on Pre-Publication Sharing (Toronto International Data Release Workshop Authors *et al.* 2009) in which we describe our plans for the first use of these data for genome-wide analyses and publication.

R.F. Veerkamp and Y. de Haas (eds)

DOI: 10.3920/978-90-8686-940-4_550, © H. Acloque *et al.* 2022

Materials & methods

Animals, tissues, time-points. Post-mortem samples were harvested from embryos (chicken: E8, E15; pig: 30, 70 days post-fertilisation) to capture early and late organogenesis; and from new born/hatched piglets and chicks. Large White pigs, that are one of the most widely used commercial pigs, were used. The chickens were from the Roslin broiler line.

Seven different tissues (liver, small intestine, kidney, lung, hindlimb muscle, brain, skin) were chosen to represent the three layers of embryonic and fetal development (endoderm, mesoderm, ectoderm) and the physiological functions of most relevance for the further integration in genomic prediction models.

Tissues and developmental stages were chosen to complement those being characterised in complementary USDA-funded chicken and pig projects, minimising duplication, whilst including some element of replication. Links to the data and metadata for all samples are available at https://data.faang.org/projects/GENE-SWitCH

Functional genomic assays. The samples have been processed to yield chromatin, DNA and RNA for the following assays: Promoter Capture Hi-C, ATAC-seq, methylation profiling by whole genome and reduced-representation bisulphite sequencing, long read RNA sequencing using Pacific BioSciences (PacBio) Iso-Seq, RNA-seq (poly-A+), small RNA-seq, and chromatin immunoprecipitation sequencing (ChIP-seq) for sequences bound by modified histones (H3K4me1, H3K4me3, H3K27Ac and H3K27me3) and CTCF. Analyses of the sequence data from these assays will allow us to identify open chromatin indicative of active elements of the genome and key regulatory elements including promoters and enhancers as well as a more comprehensive view of the chicken and porcine transcriptomes and their dynamics during development.

Pipelines of analysis. As a basis for bioinformatics analysis of the GENE-SWitCH data, the nf-core pipelines have been adopted (Ewels *et al.* 2020). In the case of RNA-Seq, a GENE-SWitCH specific pipeline has been developed, TAGADA, which has additional features, in particular annotation of long non-coding RNAs. For DNA methylation analysis an extension of the nf-core/methylseq pipeline (GSM), including downstream analysis, was developed. In the case of IsoSeq data, an nf-core pipeline did not exist, therefore an IsoSeq pipeline has been developed and will be submitted to nf-core. All GENE-SWitCH bioinformatics pipelines are available on the FAANG github page: https://github.com/orgs/FAANG/.

Results

The data production phase of the project is almost complete with the generation of ChIP-seq data due for completion in Spring 2022. The sequence data have been deposited in the public domain sequence databases in accordance with the FAANG Data Sharing guidelines (https://www.faang.org/data-share-principle). These data are indexed on the GENE-SWitCH project pages in the FAANG Data Portal (https://data.faang.org/projects/GENE-SWitCH). Only those assays which have been completed are listed below.

RNA-Seq data. RNA-Seq reads have been used to identify and quantify expression of both known and novel transcripts and genes. The vast majority of the known reference genes (73.1% for pig and 81.5% for chicken) and transcripts (72.0% for pig and 78.8% for chicken) are expressed (TPM≥0.1) in at least one sample. In terms of novel genes, we predict an additional 2,103 for pig and 1,642 for chicken. Meanwhile, there are two to three times as many nc transcripts in the novel gene annotation as in the reference gene annotation (+92% for pig and +192% for chicken). From the novel annotation spliced transcripts, 8.9 and 22.1% extend a reference transcript by adding at least one exon in 5' or 3' for pig and chicken respectively, and 34.2 and 41.8% represent completely novel spliced transcripts. Around 15,000 transcripts are predicted as lncRNAs in the two species, of which the majority (around 70%) are intergenic.

The hierarchical clustering based on reference gene expression for pig can be seen in Figure 1. Samples cluster first by tissue for cerebellum, kidney and liver while for lung, ileum, muscle and skin, this is the developmental stage which is the main criterion for clustering, especially for the earliest stage.

IsoSeq data. IsoSeq is a method which uses PacBio sequencing to produce high quality, full length transcript sequences. In the GENE-SWitCH project, samples were pooled across individuals to produce one IsoSeq dataset per tissue, time-point and species (42 datasets). At the point of submission, IsoSeq data is only available for chicken, and IsoSeq has previously been applied to chicken tissues by Kuo *et al.* (2017) providing a useful comparator. An initial annotation for chicken using the IsoSeq data has been produced using standard IsoSeq filtering and clustering techniques, chimera removal, and removal of low confidence annotations. The number of detected transcripts is four times higher than in the Ensembl v105 annotation, and 2.5 times higher than in Kuo *et al.* (2017). The number of detected genes is similar to Ensembl v105 (Ensembl 24,356, GENE-SWitCH 24,152), however, the GENE-SWitCH IsoSeq annotation predicts 6,865 potential new genes.

Small RNA-Seq data. Between 30 and 120 million reads were sequenced per sample. Out of those, a median of 57.6% for pig and 45.0% for chicken could be mapped to mirBase mature miRNAs. From the remaining reads, a median of 39.0% reads for pig and 53.5% reads for chicken could be mapped to mirBase hairpin miRNAs. Similar to RNA-Seq, hierarchical clustering based on miRBase mature miRNA normalized expression led to a clustering by tissue (results not shown). As for miRNA discovery, 6,012 and 9,905 miRNAs were detected for pig and chicken, of which 4,873 (81.1%) and 8,124 (82.0%) were predicted as novel.

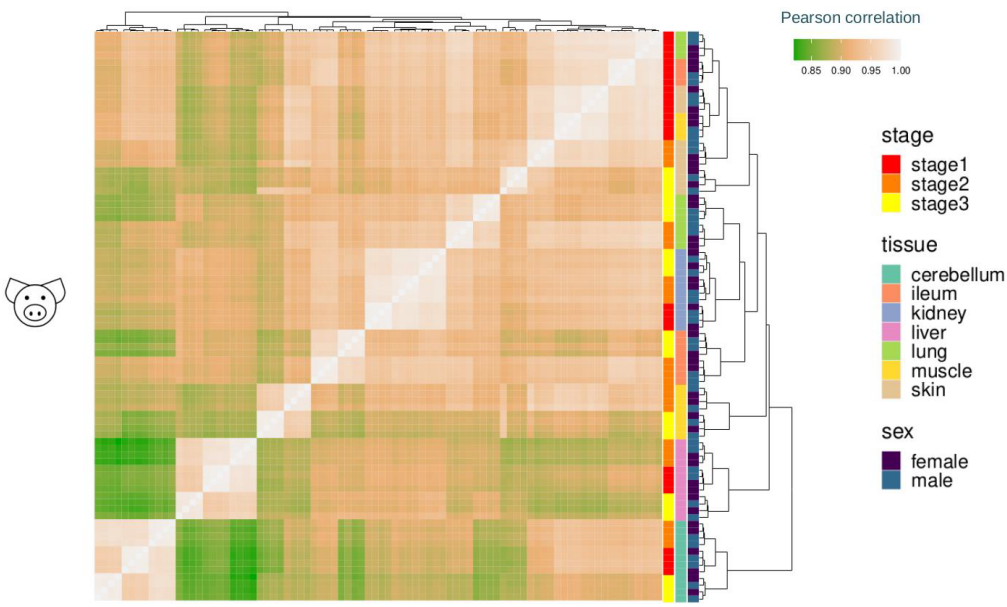

Figure 1. Hierarchical clustering of RNA-Seq data for pig. Hierarchical clustering of pig RNA-Seq data demonstrates that, as expected, samples largely cluster first by tissue, and then by developmental stage.

Bisulfite sequencing. Reduced-representation bisulfite sequencing was performed on 63 samples, and the remaining 21 samples were submitted for whole-genome bisulfite sequencing. Quality of all samples as well as mapping rates for both pig and chicken were very high. Clustering analysis of the samples provides insight into the development of the tissues, where we observe both tissue- and time-point- specific clustering in both species (results not shown).

Discussion

Functional genomics data is hypothesized to improve animal breeding by enabling more accurate genomic prediction, and a core aim of the GENE-SWitCH project is to produce extensive functional and regulatory genomics information for early developmental time-points in both pigs and chickens. Here, we have described an initial data release of both raw and processed data, as well as describing early results of how these data are likely to impact the genome annotation of these species. All data are available through the FAANG data portal.

We are using these multiple datasets: (1) to characterise the complexity of the coding and non-coding transcriptome and its dynamics during development in pigs and chickens; (2) to characterise the regulatory and epi- genomes of pigs and chickens; (3) to compare the regulatory landscape and its influence on gene expression and development in species separated by millions of years of evolutionary time (mammals, e.g. pigs and birds, e.g. chickens). In addition, to publishing the results of these analyses, enhanced annotation of the chicken and pig genomes will be released via the Ensembl Genome Browser (https://www.ensembl.org/index.html) initially as additional tracks in the Rapid Release site (https://rapid.ensembl.org/index.html). This enhanced annotation will facilitate the development of improved models for genomic prediction that incorporate knowledge of the likely functions of genome sequences as well as accounting for genetic variation.

Acknowledgements

This work is part of the GENE-SWitCH project that has received funding from the European Union's Horizon 2020 Research and Innovation Programme under the grant agreement no. 817998.

References

Clark, E.L., Archibald, A.L., Daetwyler, H.D. *et al.* From FAANG to fork: application of highly annotated genomes to improve farmed animal production. (2020) Genome Biology 21: 285. https://doi.org/10.1186/s13059-020-02197-8

Ewels, P.A., Peltzer, A., Fillinger, S. *et al.* (2020) The nf-core framework for community-curated bioinformatics pipelines. Nature Biotechnology 38:276–278. https://doi.org/10.1038/s41587-020-0439-x

Kuo, R.I., Tseng, E., Eory, L. *et al.* (2017) Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human. BMC Genomics 18:323.

Andersson, L., Archibald, A.L. *et al.* (2015) Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. Genome Biology 16:57.

Toronto International Data Release Workshop Authors, Birney E., Hudson T.J., Green E.D., Gunter C., *et al.* (2009) Nature 461(7261):168-170. https://doi.org/10.1038/461168a