GENE-SWitCH

The regulatory GENomE of SWine and CHicken: functional annotation during development

Deliverable D1.5
Raw sequences of Capture Hi-C delivered to ENA and FAANG

Deliverable leader: INRAE

Author: Hervé ACLOQUE (INRAE)

Version: 1.0

Due date of deliverable (as in DoA):	M24
Actual submission date:	30/06/2022 (M36)

Dissemination level:

PU Public	**x**
CO Confidential, only for members of the consortium (including the Commission Services)	

Research and Innovation Action, SFS-30-2018-2019-2020 Agri-Aqua Labs
Duration of the project: 01 July 2019 – 30 June 2023, 48 months

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 817998

Table of contents

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 817998

1 Summary

Objective:

The aims of WP1 are i) to establish a collection of cryopreserved tissues representing three developmental stages (early organogenesis, late organogenesis and new-borns) of pig and chicken, and ii) to characterize these samples by a set of molecular assays (RNA-seq, ChIP-seq, ATAC-seq, RRBS, WGBS, Capture Hi-C) required for the deep functional annotation of these genomes.. Within these molecular assays, Capture Hi-C (high throughput chromosome conformation capture) is a technology designed to explore the genome-wide contacts of genomic regions of interest within a cell population.

As part of the annotation effort of the pig and chicken genomes, D1.5 aims to generated an exhaustive list genome-wide promoter contacts inliver and skeletal muscle).This information will also be used for the eQTL study in WP4 (T4.3).

Methods:

Tasks 1.1 and 1.2 of WP1 consisted in carrying out a collection of samples corresponding to the (7) selected tissues and (3) developmental stages, with associated metadata describing accurately the samples and the sampling process (described in D1.1 and D1.2). Regarding Capture Hi-C, for each species and each developmental stage, 2 tissues (liver and skeletal muscle) and 2 biological replicates (1 male and 1 female) were sampled. Starting from this tissues, the required single cell suspensions of pig samples were prepared by INRAE in Task 1.1. The required single cell suspensions of chicken samples were prepared by UEDIN and sent to INRAE in Task 1.2. The protocol used to prepare single cell suspensions has been uploaded to the FAANG Data Coordination Center (DCC) and is publicly available in the FAANG Data Portal (https://data.faang.org/api/fire_api/experiments/INRAE_SOP_GENESWITCH_CaptureHiC_Cell_Dissociation_20210324.pdf).

Two custom sets of capture probes (SureSelect Agilent) were designed to capture about 10,000 promoters in pig and chicken. Then, 24 Capture Hi-C libraries were produced using the Arima-HiC+ kit (Arima) and the SureSelect XT HS kit (Agilent) starting from snapfrozen single-cell suspension from liver and skeletal muscle.

Main results:

Twenty-four single cell suspensions (liver and muscle at the 3 developmental stages on 2 biological replicates for the 2 species) were processed for Capture Hi-C assays and the

resulting libraries were sent on dry ice to the Get-PlaGe platform in Toulouse (INRAE, France) for sequencing. The sequencing was designed to provide ~150M paired-end reads (2x150bp) per sample. Sequence data have been completed in February 2021 (pig samples) and in April and June 2022 (chicken samples). Fastq files and metadata have been uploaded to the FAANG Data Portal and are publicly available with the following accession numbers: PRJEB44486 (pig) and PRJEB53986 (chicken)

Teams involved:

INRAE, UEDIN

2 Introduction

The three-dimensional architecture of chromatin creates interactions and proximities between different regions of the genome, and plays a crucial role in regulating gene expression and cell phenotypes. Capture Hi-C (high throughput chromosome conformation capture) is one of the "Hi-C" technologies in use and is designed to characterize the genome-wide contacts of genomic regions of interest (e.g. gene promoters).

In GENE-SWitCH we decided to explore genome-wide promoter contacts in 2 tissues, liver and skeletal muscle from 2 biological replicates and for 3 developmental stages of pig and chicken (12 samples per species). The exhaustive list of promoter contacts will contribute to the functional annotation at these tissues. Moreover, it will provide complementary information to WP4 for the enhanced analysis of the pig eQTL study (T4.3) and for the study and epistatic interactions influencing growth in the chicken (T.4.4) by facilitating the identification of genetic variants that may affect such promoter contacts.

Regarding this deliverable, the tasks for UEDIN in T1.2 and INRAE in T1.1 included preparing single cell suspensions from liver and skeletal muscle from the chicken and pig samples respectively. The tasks for INRAE in T1.6 included (1) taking receipt (from UEDIN) of single cell suspension from chicken samples, (2) designing two sets of capture probes covering a wide-range of chicken and pig promoters respectively, (3) preparing 24 libraries of Capture Hi-C, (4) sending the libraries to GeT-PlaGe (INRAE) for sequencing, (5) checking for raw sequences quality and (6) uploading data and metadata to the FAANG DCC and ENA.

3 Results

3.1 Result 1: single-cell suspension

We successfully sampled all the tissues and performed single cell suspensions to perform the Capture Hi-C assay described in the DoA and corresponding to Tasks 1.6. The protocol used to prepare single cell suspensions has been uploaded to the FAANG DCC and is publicly available:
(https://data.faang.org/api/fire_api/experiments/INRAE_SOP_GENESWITCH_CaptureHiC_Cell_Dissociation_20210324.pdf)

3.2 Result 2 Probes' design

Two sets of capture probes (SureSelect Agilent) were designed to capture ~10k promoters in pig and chicken.

For each species, INRAE designed first a list of promoters (1kb length for each promoter, bed file) based on gene expression datasets (RNA-seq), chromatin accessibility datasets (ATAC-

seq) and available gene annotations. This (bed format) file was sent to the technical support of Arima Genomics (www.arimagenomics.com), that provided a new bed file including the restriction fragments that would be generated by the enzymes used in the Arima-HiC+ kit on the promoter regions.

This new promoter bed file was then sent to the technical support of Agilent (www.agilent.com) to provide a list of probes (120bp each) covering promoters ((1 to 3 probes per promoter). As the size of the design is limited (the Tier 3 from Agilent Sure Select XT kitscan cover up to 6Mb) it was impossible to cover all the promoters. Therefore, we selected only the probes (and promoters) that displayed the best score regarding capture efficiency.

The validated design is then processed by Agilent and is available online at:

https://earray.chem.agilent.com/suredesign/index.htm

Design name	Design ID	Species	date	number of probes	number of promoters	Reference genome
INRA_HiC_V1_XTHS	3225281	Pig	29/07/2019	31,520	10,238	Sscrofa11.1
Hi_c_poulet_V1	S3386943	Chicken	21/01/2022	33,654	12,652	Galgal6

3.3 Result 3: Production of Capture Hi-C libraries

Twenty-four libraries were produced using the Arima-HiC+ kit (Arima) and the SureSelect XT HS kit (Agilent) starting from snapfrozen single-cell suspension from liver and skeletal muscle.

Pig libraries have been performed on November 2020. The protocol used has been uploaded to the FAANG DCC and is publicly available:

https://data.faang.org/api/fire_api/experiments/INRAE_SOP_GENESWITCH_WP1_PROMOTER_CAPTURE_HI-C_20210325.pdf

Chicken libraries have been performed in March 2022. The protocol used the Arima Library Prep kit from Arima Genomics instead of the Kapa Hyper prep PCR-free, following the recommendations of Arima Tech support.

The protocol has been uploaded to the FAANG Data Portal and is publicly available:

https://data.faang.org/api/fire_api/experiments/INRAE_SOP_GENESWITCH_WP1_PROMOTER_CAPTURE_HI-C_CHICKEN_20220625.pdf

In both the pig and chicken protocols we performed the following Quality Control (QC) steps: i) estimation of the input amount, ii) measurement of digestion and biotinylation efficiency, iii) quality of DNA fragmentation, iv) quantification of DNA with a Qubit fluorometer after fragmentation, v) library production and Capture, and vi) libraries' profiles on a BioAnalyzer. When the libraries met all the quality criteria, they were sent for sequencing at the GeT-PlaGe (INRAE) facility in Toulouse (France).

3.4 Result 4: Sequencing of Capture Hi-C libraries

Capture Hi-C libraries were sequenced on a Novaseq6000 instrument from Illumina.

Sequencing of the pig libraries was achieved in February 2021. Quality control was carried out by INRAE to confirm that sequencing data met the quality criteria. An average of 165M paired-end reads (2x150bp) per sample was achieved.

Sequencing of the chicken libraries was achieved in June 2022. Quality control was carried out by INRAE to confirm that sequencing data meet the quality criteria. An average of 444M paired-end reads (2x150bp) per sample was achieved.

Sequencing statistics:

Pig

Sample ID	# Reads	Yield (Mbases)	Mean Quality Score	% Bases >= 30
SSC_INRAE_GS_WP1_CHiC_liver_FT_30dpf_37	158M	47,400	35,95	93,91
SSC_INRAE_GS_WP1_CHiC_liver_FT_30dpf_40	160M	48,000	35,93	93,82
SSC_INRAE_GS_WP1_CHiC_liver_FT_70dpf_1	200	60,000	35,94	93,86
SSC_INRAE_GS_WP1_CHiC_liver_FT_70dpf_4	180M	54,000	35,95	93,93
SSC_INRAE_GS_WP1_CHiC_liver_NB_M_2	156M	46,800	35,9	93,68
SSC_INRAE_GS_WP1_CHiC_liver_NB_F_3	233M	69,900	35,91	93,76
SSC_INRAE_GS_WP1_CHiC_hindlimb_muscle_FT_30dpf_37	155M	46,500	35,93	93,83
SSC_INRAE_GS_WP1_CHiC_hindlimb_muscle_FT_30dpf_39	140M	42,000	35,9	93,66
SSC_INRAE_GS_WP1_CHiC_hindlimb_muscle_FT_70dpf_1	134M	40,200	35,93	93,81
SSC_INRAE_GS_WP1_CHiC_hindlimb_muscle_FT_70dpf_4	170M	51,000	37,76	93,35
SSC_INRAE_GS_WP1_CHiC_gluteus_medius_NB_M_2	149M	44,700	35,91	93,71
SSC_INRAE_GS_WP1_CHiC_gluteus_medius_NB_F_4	140M	42,000	35,95	93,93
mean	165M	49,500		

chicken

Sample ID	# Reads	Yield (Mbases)	Mean Quality Score	% Bases >= 30
GGA_INRAE_GS_WP1_CHiC_Liver_E8_1	412M	123,670	35,95	93,91
GGA_INRAE_GS_WP1_CHiC_Muscle_E8_1	394M	118,090	35,93	93,82
GGA_INRAE_GS_WP1_CHiC_Liver_E8_2	321M	96,458	35,94	93,86
GGA_INRAE_GS_WP1_CHiC_Muscle_E8_2	657M	196,974	35,95	93,93
GGA_INRAE_GS_WP1_CHiC_Liver_E15_1	755M	226,500	35,9	93,68
GGA_INRAE_GS_WP1_CHiC_Muscle_E15_1	390M	117,000	35,91	93,76
GGA_INRAE_GS_WP1_CHiC_Liver_E15_2	481M	144,300	35,93	93,83
GGA_INRAE_GS_WP1_CHiC_Muscle_E15_2	332M	99,600	35,9	93,66
GGA_INRAE_GS_WP1_CHiC_Liver_HC_1	304M	91,200	35,93	93,81
GGA_INRAE_GS_WP1_CHiC_Muscle_HC_1	332M	99,600	35,76	93,35
GGA_INRAE_GS_WP1_CHiC_Liver_HC_2	435M	130,500	35,91	93,71
GGA_INRAE_GS_WP1_CHiC_Muscle_HC_2	512M	153,600	35,95	93,93
mean	444M	133,000		

3.5 Result 5: Submission of data to FAANG DCC and ENA

Fastq files and metadata have been uploaded to the FAANG DCC and European Nucleotide Archive in April 2021 (pig) and June 2022 (chicken) and are publicly available with the following accession numbers: PRJEB44486 (pig) and PRJEB53986 (chicken).

4 Deviations and delays

Delays in the production and sequencing of chicken libraries were due to:

- delays in the sampling of chicken tissues (T1.2; 9 months delay);
- delays in providing the annotations necessary for the design of the probes (T2.1; 9 months delay);
- unexpected and COVID related 6-month delay for the design and the production of probes (chicken) (Arima Genomics and Agilent technical supports and manufacturing);
- resequencing of 11 libraries was necessary in order to increase the number of valid reads (2 months delay).

5 Acknowledgements

We would like to acknowledge Arima Genomics and Agilent for their technical support and follow-up.