THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :

Institut National Polytechnique de Toulouse (Toulouse INP)

Discipline ou spécialité :

Pathologie, Toxicologie, Génétique et Nutrition

Présentée et soutenue par :

Mme MARIA EUGENIA MARTI MARIMON

le vendredi 9 novembre 2018

Titre :

3D genome conformation and gene expression in fetal pig muscle at late gestation

---

Ecole doctorale :

Sciences Ecologiques, Vétérinaires, Agronomiques et Bioingénieries (SEVAB)

Unité de recherche :

Génétique, Physiologie et Systèmes d'Elevage (GenPhySE)

Directeur(s) de Thèse :

MME MARTINE YERLE

M. SYLVAIN FOISSAC

Rapporteurs :

M. FREDERIC BANTIGNIES, CNRS

Mme CAROLE CHARLIER, UNIVERSITE DE LIEGE

Membre(s) du jury :

M. OLIVIER GADAL, CNRS TOULOUSE, Président

Mme MARTINE YERLE, INRA TOULOUSE, Membre

M. SYLVAIN FOISSAC, INRA TOULOUSE, Membre

Acknowledgments

I would like in first place to express my gratitude to the members of the jury for the time they spent evaluating this work. Thanks to Dr. Carole Charlier and Dr. Frédéric Bantignies for agreeing to be the reviewers of this thesis and for their kind and helpful feedback on my thesis. I would like to thank also to Dr. Olivier Gadal and Dr. Roderic Guigo for accepting to be part of the thesis committee.

Je voudrais remercier tout spécialement mes trois encadrants. Merci Martine d'avoir accepté d'encadrer cette thèse et de t'être rendue disponible quand j'en avais besoin malgré toutes tes responsabilités. Merci aussi d'avoir veillé à ce que cette thèse ait bien été menée à terme. Sylvain, un grand merci pour toute ton aide, tes conseils et ta patience. Oui, il a fallu l'être avec cette thésarde débutante en bioinfo dont tu as hérité. Mais surtout, Sylvain, merci pour ton optimisme et pour avoir su me remonter le moral dans les moments difficiles. Enfin, Yvette, je te remercie beaucoup pour tous ces échanges dans nos bureaux qui ont fait naitre des questions et des idées qui ont permis d'améliorer mon travail, ainsi que pour ces petits coups de pouce pour les manips (j'ai adoré travailler avec toi au labo).

D'autres personnes ont énormément contribué à mon travail et je leur en serai toujours reconnaissante. Tout particulièrement Nathalie V., merci à toi pour ton aide inestimable dans toutes les analyses biostatisiques, sans lesquelles ce travail n'aurait pas été possible. Malgré ta charge de travail, tu as toujours su trouver le temps pour m'aider. Je voudrais aussi remercier Pierre Neuvial, responsable du projet SCALES qui a servi à financer une bonne partie du séquençage permettant d'améliorer la qualité de mes données. Merci également Laurence L. pour nos échanges et pour ton implication dans mon premier article, j'ai vraiment apprécié de collaborer avec toi. Sarah, merci pour tous ces petits coups de main et pour ta gentillesse. Hervé, tu as été présent dans la plupart des moments clés de mon parcours à l'INRA depuis mon arrivée. Tu as été la toute première personne rencontrée, et qui de plus parlait l'espagnol!). C'est en partie grâce à toi que je me suis inscrite au master, que j'ai prolongé mon CDD avec Alain V. (merci à toi aussi), et que ce projet a vu le jour (je me souviens encore de cette réunion improvisée au patio avec Sylvain où nous avons établi les bases de cette thèse). Sans votre aide, je ne serais pas là aujourd'hui en train d'écrire ces lignes. Merci bien sûr à tous les membres de mon équipe Cytogène pour votre accueil chaleureux et votre soutien. Enfin, merci à Florence M., David R., Matthias Z., Diane E., Noémie T., Lisa B., Katia F., Nicolas M. et tous ceux que j'ai pu oublier pour leurs petites mais précieuses contributions.

La vie au labo n'aurait pas été la même sans « mes filles adorées », Manu, Lisa, Laure, Nathalie M. et Sonia, ainsi que sans toi Valentin (une petite pensée pour toi, même si tu as quitté le labo il y a longtemps). Yvette et Sylvain, je reviens à vous car plus qu'encadrants, vous êtes devenus de chers amis sur le plan personnel. Merci à tous pour tous ces moments partagés non seulement au travail mais aussi en soirées, vacances, etc., je vous porterai toujours dans mon cœur. Sachez que ces petites pauses-cigarette au patio avec vous les filles ainsi qu'avec Pitou, Yann, David, Sylvain, Juliette… vont

énormément me manquer. Enfin un grand merci en général à tous les membres du ex-LGC qui depuis mon arrivée m'ont fait me sentir comme chez moi et qui, pour une étrangère comme moi, sont devenus une grande famille.

A tous mes amis, spécialement aux « Toulousanos », et à ma famille un très grand merci; aux premiers pour me donner l'envie de rester en France, et pour me distraire et/ou me supporter quand la thèse devenait difficile; et à la deuxième, pour son aide inconditionnelle. "Gràcies mamà i papà, per tota la vostra ajuda i per creure sempre en mi: Si és que tinc la sort de tindre els millors pares del món! Ana i Dani gràcies també a vosaltres per estar sempre a l'escolta encara que la distància no ens permeta parlar més sovint".

Enfin, en dernier, mais le plus important, Stéphane, je ne pourrais pas trouver tous les mots nécessaires pour te remercier mais je vais essayer... Depuis le début tu as été toujours à mes côtés, pour les bons mais aussi pour les mauvais moments. Tu as souffert pour moi plus que moi-même dans mes plus grandes périodes de stress et tu t'es réjoui et senti fier de moi à chaque petit succès que j'ai pu avoir. Sans jamais lâcher et toujours prêt à m'aider et à me faire sourire, je ne peux pas imaginer un meilleur compagnon dans ma vie.

Table of contents

List of figures

List of tables

1 Abstract

In swine breeding industry, sows have been selected for decades on their prolificacy in order to maximize meat production. However, this selection is associated with a higher mortality of newborns. In this context, the skeletal fetal muscle is essential for the piglet's survival, as it is necessary for motor functions and thermoregulation. Besides, the three-dimensional structure of the genome has been proven to play an important role in gene expression regulation. Thus, in this project, we have focused our interest on the 3D genome conformation and gene expression in porcine muscle nuclei at late gestation. We have initially developed an original approach in which we combined transcriptome data with information of nuclear locations (assessed by 3D DNA FISH) of a subset of genes, in order to build gene co-expression networks. This study has revealed interesting nuclear associations involving *IGF2*, *DLK1* and *MYH3* genes, and highlighted a network of muscle-specific interrelated genes involved in the development and maturity of fetal muscle. Then, we assessed the global 3D genome conformation in muscle nuclei at 90 days and 110 days of gestation by using the High-throughput Chromosome Conformation Capture (Hi-C) method. This study has allowed identifying thousands of genomic regions showing significant differences in 3D conformation between the two gestational ages. Interestingly, some of these genomic regions involve the telomeric regions of several chromosomes that seem to be preferentially clustered at 90 days. More important, the observed changes in genome structure are significantly associated with variations in gene expression between the 90th and the 110th days of gestation.

2 General introduction

Pig breeding is one of the most important divisions in the French feed industry, being the French swine sector the third producer in EU, and the pork, the most consumed meat in France. In order to reach such levels of production, farmers have developed over the last fourty years breed selection programs based on cross-breeding plans designed to select phenotypic traits of interest. In that context, sows have been selected for their prolificacy. Unfortunately, the increasing number of piglets per litter has been correlated with an increase of newborns mortality. A key factor in this issue is the piglet's maturity, defined as the stage of full development leading to survival at birth. Indeed, developmental problems occurring at late gestation can lead to maturity defects during the perinatal period and consequently, to death. Therefore, it is important to understand the biological processes taking place in late gestation. For instance, the skeletal muscle represents the first reserve of glycogen in piglets, which is used during the first 24h after birth for piglet's thermoregulation.

Many studies have been performed in muscle tissue to identify key genes or molecular processes involved in muscle development and maturity. Nevertheless, it remains unclear how these genes or processes are regultated, not only in muscle tissue but generally in all kind of tissues. In fact, all cell types of a living organism have the same genetic material yet, they are morphologically and functionally different from each other. It is known that cell-type specific genes are responsible of phenotypic differences observed between cells, as it is also known that modulations of expression levels of a given gene can explain differences observed in a specific cell type under different conditions. Although in many cases the mechanisms of gene regulation have been well described, for many others some questions remain open: Which are all the factors and mechanisms responsible of gene expression regulation? How these mechanisms of gene regulation work? Some features present in the genome sequence itself such as promoters or enhancers, have been identified as key elements involved in gene expression regulation. Others, such as transcription factors or non-coding RNAs, associate to specific DNA sequences in order to regulate gene expression. Also, epigenetic modifications on histones or DNA have been show to play as well an important role in this regulation. All this knowledge has provided many clues to unravel the mechanisms involved in gene expression regulation but it remains insufficient to explain all phenotypic differences observed between cells. In the last fifteen years, numerous studies have emerged addressing the question of the 3D genome organization role in the regulation of gene expression. It has been proven that it exists an intimate relation between chromatin structure and gene expression, as will be presented in more detail in the third chapter of the "Bibliographic review".

The main objective of this thesis has been to explore any change on the 3D genome structure occurring in fetal porcine muscle between two developmental ages that could explain phenotypic differences observed at the level of gene expression at 90 and 110 days of gestation. For that purpose two different strategies were used.

Our fisrt approach combined the inference of gene co-expression networks with nuclear location information of a small set of genes. The expression data of a previous transcriptome study performed

by microarray analysis on muscle samples from 90 days and 110 days fetuses (Voillet et al., 2014) was used to build the networks. Concretely, we used the expression values of differentially expressed genes, identified in this previous study as genes having a potential role in piglet's maturity. Then, the 3D nuclear proximity between some pairs of genes was tested by 3D DNA Fluorencence *in situ* Hybridation (FISH) because either they appeared connected in the networks, they were identified as key genes in muscle development, or both. The resulting information of these 3D DNA FISH assays was used each time to infer a new gene co-expression network by reinforcing the edges between genes when they were found co-localized in the nucleus, or by preventing connexions between genes found distant in the nuclear space. This integration of gene expression and nuclear co-localization proved to be relevant as it revealed clusters of genes, around our target genes, related to muscle development.

In this first approach, we analyzed the nuclear proximity (distance) of a small number of genes in few (~ 60 - 100) nuclei. On our second approach, we sought to extend the scale of the analysis in order to explore the genome-wide structure of the DNA in a large population of muscle cells. We used the High-throughput Chromosome Conformation Capture (Hi-C) molecular approach, coupled to DNA sequencing and bioinformatics data analysis. This enabled to identify all genomic regions that were in nuclear proximity. Hi-C assays were performed on muscle samples from two gestational ages (90 and 110 days, 3 fetuses per condition). Large genomic regions, the so-called "A and B compartments", which are functionally different, were also identified. Although these compartments were highly conserved, we identifyied some genomic regions switching of compartment type between the two conditions. At a smaller scale, topologically associated domains (TADs), were also identified in both conditions. The differential analysis revealed global differences in the 3D chromatin structure between the two gestational ages. More precicely, it allowed identifying genomic regions that were proximal at 90 days of gestation but distant at 110 days and vice versa. Finally, we explored whether the differential of genome organization between the two gestational ages was associated with a differential in gene expression previously reported in the muscle transcriptome study. Small althought significant differences in gene expression were associated with those genomic regions showing a differential conformation between the two gestational ages.

The present manuscript is divided into four sections. The first one is a bibliographic review about the pig breeding context, which exposes the issue of neonatal mortality, the concept of maturity, and the role of muscle development and maturity in survival at birth. This is followed by a review about pig genome sequencing and annotation and the main transcriptome studies performed on fetal muscle. Lastly, a detailed review about the nuclear architecture will be presented to uncover: first, the general principles of the 3D genome organization then, the different approaches that allow to study this aspect and final, some specificities observed in pig. The methods used in this project will be presented in the second section. The third and fourth sections show respectively the results obtained with the two strategies mentioned before, used to study the 3D genome structure during fetal development and the integration of gene expression. In these two last sections, the results are presented and then discussed. The main results of this thesis are summarized in a final conclusion, followed by the presentation of the perspectives.

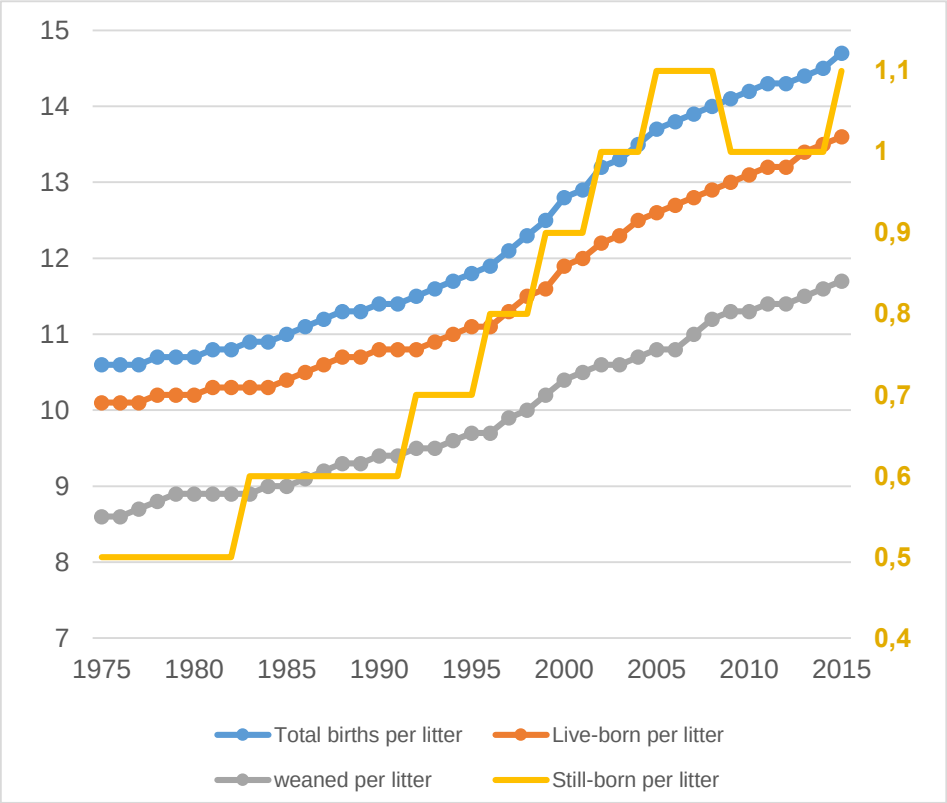


Figure 1. Evolution of average number of piglets per litter in France from 1975 to 2015. *The data used to build this graph were collected and treated by the GTTT (Technical Management of Sow Herds) of the French Porc and Pig Institut (IFIP).*

3 Bibliographic review

3.1 Chapter 1. Breeding context

3.1.1 Early mortality: a major breeding issue in pig farming

3.1.1.1 Background

Pig sector is an important economic motor of the livestock industry in France. In 2016, around 8.9% (24.3 million pigs) of the global production in the European Union (EU) was obtained in France (data obtained from the National Establishment of Agricultural and Sea products, FranceAgriMer, 2017). This makes the French swine sector into the third producer in EU after Germany and Spain. Actually, pork is the most consumed meat in France, before poultry and cattle, which makes France the fourth consumer (FranceAgriMer, 2017). To ensure this demand of meat, farmers have had to find strategies in order to increase their production.

3.1.1.2 Selection towards prolificacy

Beyond the breeding conditions (feeding, bedding, health, etc.), genetics and breed selection programs are among the most important aspects handled by farmers to increase their production. In this context, cross-breeding plans are used to combine different genotypes and select the best animals regarding genealogy, reproductive performance, growth, carcass type, and meat quality. For instance, pig males have been selected to improve feed conversion efficiency and carcass quality criteria, and sows from Large White (LW) line have been prolificness-enhanced to increase the number of live-born piglets per litter.

The selection towards increasing the prolificacy and meet production, has been unfortunately associated to an increment of perinatal mortality. Figure 1 shows data collected from the French Porc and Pig Institute (IFIP). A shift is observed between the years 1975 and 2015, with an increase in: (i) progeny (4.1 more piglets per litter), (ii) premature mortality (0.6 more still-born per litter) and (iii) postnatal mortality (1.5 of piglets died before reaching the weaning age in 1975 while 1.9 perished before weaning in 2015). The last, presenting the highest rate around the first 48-72 hours (corresponding to the perinatal mortality). In brief, the incidence of mortality has considerably raised in the last forty years, especially during the last fifteen years because of the application of novel selection practices for genetic improvement. This early mortality generates not only important economic losses (10-20% of total operating costs) for the swine industry but also raises ethical questions about animal welfare.

Early mortality is not a phenomenon restricted to the swine industry, other species in the agronomic sector suffer from the same losses. In the sheep industry, the mortality rate before the sixtieth day after birth is 13.6% and, before 48 hours of lambs' life, mortality represents more than 50% (data

obtained from the Sheep Breeding Institute, Idele, 2016). Although less pronounced than in the pig and sheep sectors, the proportion of perinatal mortality in the cattle industry is 5.2% (Perrin et al., 2011).

To finish with this overview of perinatal mortality, I would like to underline that humans are unfortunately not exempt from this problem, despite all the advances in medicine in the last years. In 2016, 5.6 million deaths were registered in children under five years old. Neonatal deaths (the first 28 days of life) accounted for 46% of all under-five deaths. Although the majority of them were attributable to neonatal infections, intrapartum-related events and congenital abnormalities, almost 35% of the neonatal mortality was due to preterm birth complications (data obtained from UNICEF, "Levels and Trends in Child Mortality Report 2017"), the last, especially regarding immaturity problems of newborns (hypothermia, hypoglycemia, respiratory distress, etc.).

3.1.1.3 Critical factors of piglets mortality

Many factors are responsible of pig losses, the most commons, the ones affecting the perinatal period and involved in stillbirths (prenatal stage), and deaths during the first 72 hours after birth (neonatal stage). Fetal losses can be explained by maternal effects (uterus anatomy, placenta development, and number of embryos). Neonatal deaths happening during farrowing can also be explained by maternal effects (farrowing issues, intrauterine hypoxia and hyperthermia caused by acidosis). Those happening in early breastfeeding can be due to maternal effects, breeding conditions or effects specific to the piglets. The piglet's weakness (malnutrition by low-quality colostrum and/or milk production from the sow) and maternal crushing are the most common causes of pre-weaned mortality. Other factors are important for piglet survival like the maternal skills/abilities (resource management (relationship with its appetite and body condition), its dairy milk production, the farrowing efficiency, the weight and size of the piglets and the maternal behavior). And last, but not less important, the vitality of the piglet defined as the piglet characteristics that will influence its survival and growth during the breastfeeding stage (Canario, 2006). This thesis is focused on this last item which is developed in the coming section.

3.1.2 Maturity and survival

3.1.2.1 Critical factors for piglets survival

Some studies have been performed to investigate and understand which conditions during pig fetal development alter the genetic merit for piglet survival. It has been observed that postnatal performance in pigs is mainly affected by the placental development, the size and weight of fetuses, and the levels of cortisone and glycogen. For instance, litters with high estimated breeding values for piglet survival present smaller and more regular placenta, smaller fetuses, higher cortisol concentrations, higher concentrations of glycogen in liver and skeletal muscle (*longissimus dorsi*) and higher percentages of carcass fat (Leenhouwers et al., 2002a). Similarly, intrauterine growth retardation (IUGR) have been associated with variation in birth weight within litters, pre-weaning survival and postnatal growth. Actually, IUGR is often produced due to high ovulation leading to high fetuses surviving to 30 days gestation. This is in detriment of a proper placenta development, especially limiting

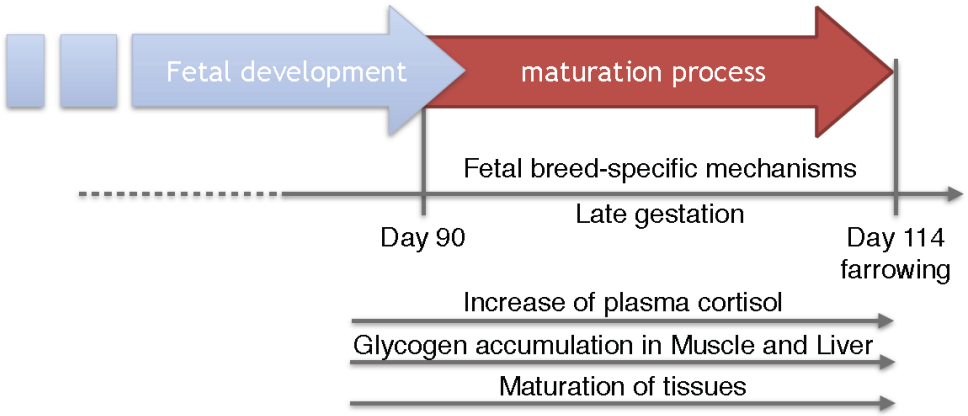


Figure 2. Specific mechanisms during pig maturation process in late gestation. (Voillet, 2016).

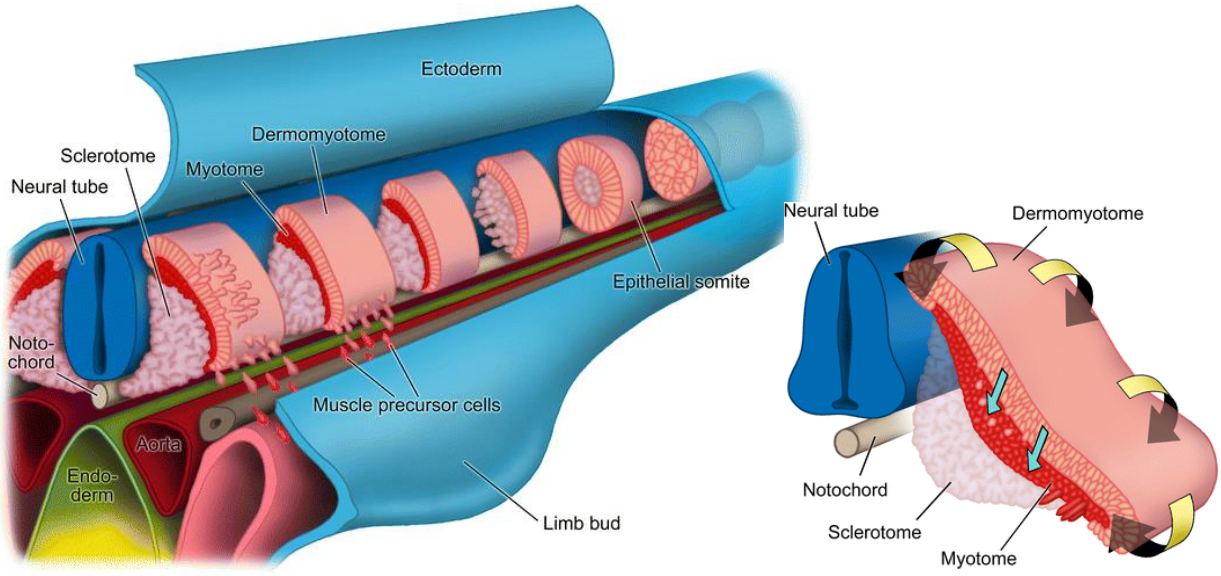


Figure 3. Primary trunk muscle embryonic development. *Skeletal striated muscle derives from the myotome, the middle layer of the somite segments. Myogenesis is initiated by delamination of cells of the dermomyotome that differentiate into skeletal muscle of the myotome* (Yusuf and Brand-Saberi, 2012)**.**

the availability of nutrients to the embryo during myogenesis (Foxcroft et al., 2006). These aspects support the idea that by selecting hyperprolific females to increase the number of pigs born, piglet survival is strongly impacted. Therefore, this strategy of selection should be critically evaluated in the context of pork production, as well as selection should be optimized to obtain slightly smaller but stronger piglets in terms of piglet survival.

3.1.2.2 Piglet's maturity

The ability of piglets to cope with hazards during birth or within the first days of life is closely linked to the fetal physiological maturity (van der Lende et al., 2001). A state of full development, due to a successful maturation process, promotes early survival after birth (Leenhouwers et al., 2002b, 2002a). Concretely, the maturity is described by the weight of birth, the body composition, the levels of metabolites, the ability to thermoregulate, the immune response and behavioral aspects (Canario, 2006; Foxcroft et al., 2006; Leenhouwers et al., 2002a). The fetal maturation process in pigs involves biological processes occurring between the 90th day and the term of gestation (around the 114th day) (Leenhouwers et al., 2002a). During this period, the most important events happening over the maturation process are the ones described in the previous section: an increase of plasma cortisol, the glycogen accumulation in muscle and liver and the maturation of tissues (Figure 2, (Voillet, 2016)).

Experimental results have also shown that breed-specific mechanisms could influence the physiological processes at the end of development and during the maturation process. Indeed, there are examples of breeds having different performances for piglet survival. For instance, the survival rate differs between the LW European breed and the Meishan (MS), a Chinese domestic breed. The LW breed which has been highly selected, presents a high incidence of mortality, while the primitive breed MS exhibits a strong potential for survival (Herpin et al., 1993). This disparity between extreme breeds in terms of maturity can be explained by breed-specific particularities happening during the muscle development and maturation, due to the fact that a proper functioning of this tissue is essential for piglet postnatal performance as mentioned before. The role of muscle maturity in survival at birth will be discussed in the following section, by focusing attention on the skeletal muscle.

3.1.3 The role of muscle maturity in survival at birth

There are three types of muscle in vertebrates, the skeletal muscle ("voluntary muscle" responsible of the skeletal movement), the smooth muscle ("involuntary muscle", in organs, blood vessels, skin, etc.) and the cardiac muscle (also "involuntary" but more similar in structure to the skeletal muscle). In this thesis, the interest is focused on the skeletal muscle, concretely on the *longuissimus dorsi* which is located in the trunk and it extends from the thoracic region to the sacrolumbar.

3.1.3.1 Myogenesis: the fetal skeletal muscle development

Skeletal muscle in the trunk of vertebrate embryo derives from the somites, segments of the paraxial mesoderm germ layer which is formed in the primitive blastopore during gastrulation (Figure 3). Somites are located at both sides of the neural tube and notocorde, and they are composed

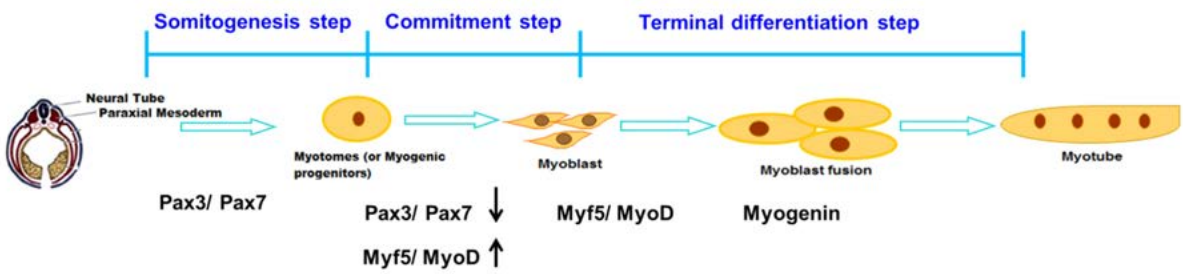


Figure 4. Myogenesis during embryonic development. *Myogenic progenitors from dermomyotome proliferate until the expression of myogenic regulatory factors (MRFs) that will determine cell commitment and differentiation through the sequential expression of different MRFs* (Jin et al., 2016)**.**

by three structures: the sclerotome (ventral compartment that originates vertebrae, ribs and cartilage), the myotome (middle layer originated from the dermomyotome, gives rise to skeletal striated muscles), and the dermomyotome (dorsal compartment originates dermis and hypodermis). After demomyotome formation, myogenesis is initiated by delamination of cells from the inward-curled borders (lips) of the dermomyotome. These detached cells move under the dermomyotome to generate the primary myotome and rapidly differentiate into skeletal muscle of the myotome. The dorsomedial portion of the myotome gives rise to the intrinsic back muscles (Buckingham, 2006; Chal and Pourquié, 2017; Yusuf and Brand-Saberi, 2012).

Cells in the dermomyotome express the Pax3 and Pax7 transcription factors, they are myogenic proliferating precursors in somites and do not express myogenic regulatory factors (MRFs) or muscle proteins (Figure 4). This is the so-called proliferation step. The determination step happens during the myotome formation, when myogenic precursors retreat from the cell-cycle. Then, these cells start to express Myf5 (myogenic factor 5), MyoD (MyoD1, myogenic determination factor), MRF4 (Myf6, myogenic factor 6) and to downregulate Pax3 (Paired box 3), becoming committed myoblasts (Buckingham, 2006; Chal and Pourquié, 2017; Yusuf and Brand-Saberi, 2012). In early development, myoblasts can either proliferate or differentiate. The differentiation step begins when myoblasts start expressing myogenin (Myf4/MYOG), MyoD and MRF4 (Buckingham, 2006). At this stage, the differentiating myoblasts are often named myocytes, which express specialized cytoskeletal proteins: Myh7 and Myh3 myosin heavy chains (MyHC), α-actine (Actc1), desmin, the Notch ligand jagged 2 and metabolic enzymes. Myocytes elongate and align to span the entire somite length and this process is controlled by Wnt11 signaling. Then they fuse leading to the formation of multinucleated myotubes which latter mature into myofibers. Myogenesis separated into two phases: an early embryonic or primary phase and a latter fetal or secondary phase. The first one results in the formation of primary myofibers (muscle cell polynucleated) (expressing slow MyHC and myosine light chain 1, MyLC1). During the second phase, myogenic precursors fuse among themselves or to the primary fibers and give rise to secondary myofibers expressing β-enolase, Nfix or MyLC3. Then these fibers also start to express fast MyHC isoforms (Chal and Pourquié, 2017).

Myofibers are filled of myofibrils which are bundles of protein filaments and responsible of muscle contraction. The process of myofibrils formation is called myofibrillogenesis. Myofibrils are composed of a repetitive contractile modules called sarcomeres and they are surrounded by the sarcolemma, a specialized plasma membrane for neural signal transduction by depolarization upon neural excitation. The filaments in a sarcomere are composed of actin and myosin (Chal and Pourquié, 2017).

It exists three main types of myofibers classified depending on their MyHC isoforms and metabolism. These are the slow-twitch oxidative (oxidative metabolism), the fast-twith oxydative (oxido-glycolytic meatabolism) and the fast-twitch glycolytic (glycolitic metabolism) fibers (Picard et al., 2002). There are eight isoforms of MyHC: four adult (I, IIa, IIx and IIb), 3 developmental (embryonic, fetal and α-cardiac), and one extraocular isoform (Perruchot et al., 2012). Oxidative slow-twitch fibers express slow MyHC (type I, Myh7), whereas glycolytic fast-twitch fibers express fast

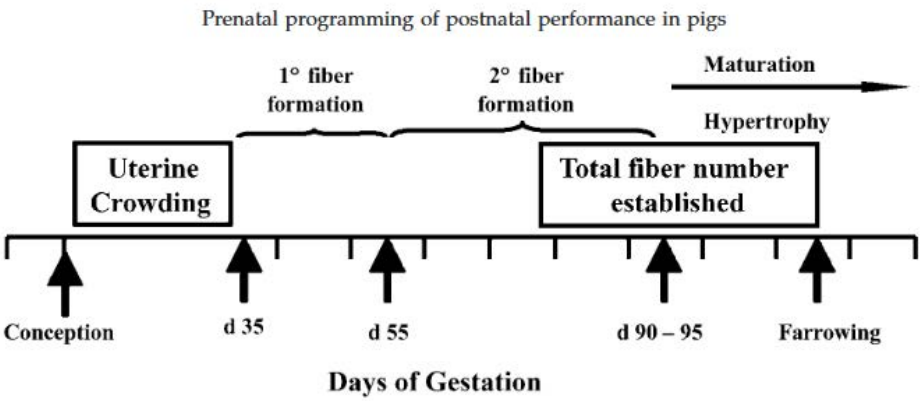


Figure 5. Schematic representation of the time-course of muscle fiber development in pig. *The muscle mass will be determined by the total number of fibers (TNF), established after the first and second wave of fiber formation, and by the prenatal and postnatal hypertrophy of those fibers* (Foxcroft et al., 2006)**.**

MyHC (types IIa (Myh2), IIb (Myh4) and IIx (Myh1)). Being the embryonic (Myh3) and slow MyHC the first to be expressed in early myogenesis phase, then fetal and neonatal fibers express perinatal MyHC (Myh8), finally the fast isoforms start to be expressed during late fetal myogenesis (Chal and Pourquié, 2017).

To finish with this section, hereafter a brief introduction about the adult muscle stem cells. They are the so-called satellite cells and are located between the basal lamina and the sarcolemma of each myofiber. They origined during embryogenesis from myogenic progenitors of the central dermomyotome expressing Pax7. The Pax7 progenitors pool is maintained by the Notch signaling. Satellite cells have a limited ability to replicate and will remain as quiescent $Pax7^{+}$ satellite cells in adult muscle. They are required for skeletal muscle regeneration, growth and maintenance through adulhood (Chal and Pourquié, 2017) (see (Crist et al., 2012) for more details).

3.1.3.2 Peculiarities of pig skeletal myogenesis and muscle metabolism

In pigs, and more generally in livestock animals, muscle fiber characteristics and ontogenesis influence the quality of meat. As discussed before, during myogenesis, two succesive waves of myoblasts are responsible of the myofiber ontogenesis and will lead to the formation of primary and secondary myofibers. In larger species as bovines, sheeps, pigs, but also in humans, it exists a third generation of myofibers (Picard et al., 2002; Rehfeldt et al., 2000). These tertiary myofibers appear during fetal life except for pigs, in which the third generation appears during the early postnatal period. Therefore, in the pig gestational timeline, the first wave of myoblast generation arrives around the 35^{th} day of fetal life, the second around the 55^{th}day, and the third between birth and the first 15^{th}days after birth (Picard et al., 2002). The total number of muscle fibers (TNF) and the myofibers size are important parameters playing a key role in meat quality and they have been influenced by lean meat growth selection (Rehfeldt et al., 2000). In pigs, the TNF is fixed around the 90^{th} day of gestation suggesting that the third generation of fibers (postnatal) is not quantitatively important. Primary and secondary fibers are under genetic and epigenetic (environmental) control respectively. The genetic aspect is explained by differences between breeds and the epigenetic one is mainly explained by maternal effects (maternal nutrition and offspring) (Picard et al., 2002; Rehfeldt et al., 2000). Regarding the maternal effects, intrauterin growth retardation has been observed in some fetuses of hyperprolific sows presenting high rates of conceptuses surviving to 30 days of gestation, resulting in detrimental effects of placental development. This limites the availability of nutrients to the embryo during the myogenesis and is translated into a decrease of the number of muscle fibers at 90 days of gestation (Foxcroft et al., 2006). Muscle mass is determinated not only by the TNF but also by the size of those fibers. Increases in muscle mass due to the fiber size are subjected to the prenatal and postnatal fiber hypertrophy (Figure 5). The hypertrophy depends on the accumulation of myonuclei (satellite cell proliferation) and muscle specific proteins.

Porcine muscle shows a unique distribution of fibers consisting in clusters of slow type I fibers surrounded by fast type II fibers. Primary and secondary myofibers express type I MyHC (embryonic

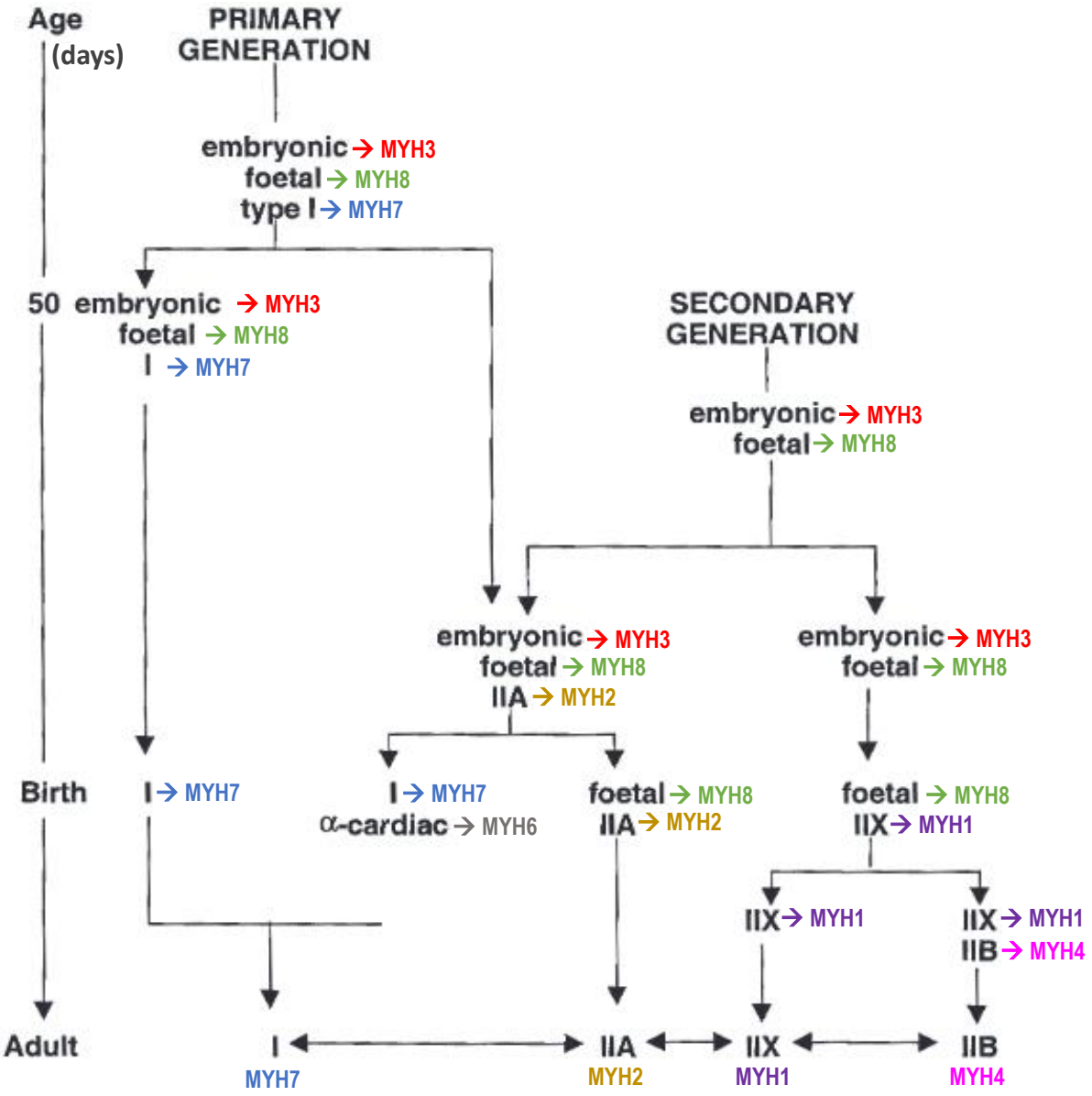


Figure 6. Schematic evolution of fiber type differentiation. *Myosin heavy chain (MYHC) isoform transitions in developing skeletal muscle of pig. Original figure created by* (Picard et al., 2002) *and modified to add the MYHC isoform genes in color.*

and fetal) but, in secondary fibers, type I MyHC is not expressed until late gestation (Figure 6). Fast type II MyHCs are mostly expressed after birth with exception of type IIa which expression increases from the last third of gestation. Another characteristic of porcine muscle is that the α-cardiac MyHC (Myh6) is also detected in early postnatal development (Picard et al., 2002). It seems that the fiber type composition could differ between different breeds. For instance, it was observed that MS pigs exhibited a decrease in the expression of the fastest isoform compared with LW pigs (Lefaucheur et al., 2004).

The oxidative metabolism represents the principal source of energy during fetal porcine life. At birth all muscles are oxidative, and glycolytic metabolism increases during the first postnatal weeks (from 0 to 15 days after birth in pigs). Globally, contractile and metabolic muscle fibers differentiate during the two first postnatal weeks, meaning that the main events occur soon after birth whereas they occur during fetal life in human, bovine and ovine (Picard et al., 2002). The carbohydrates metabolism is related to viability in perinatal period. Muscle glycogen reserves are the first source of energy for heat production used for piglets' thermoregulation during the first hours of life (Leenhouwers et al., 2002a; van der Lende et al., 2001).

3.1.3.3 Muscle and maturity

Low birth piglets born from hyperprolific sows are generally more immature, they present low number of secondary fibers, and exhibit lower postnatal growth performance and lean percentage than their mature littermates. To compensate, they tend to develop extremely large muscle fibers (giant fibers) to increase muscle mass solely through muscle fiber hypertrophy. This is associated with problems in fibers capacity to adapt to activity-induced demands, stress susceptibility and meat quality in modern meat-type pigs. Larger fibers present less mitochondria, and probably energy and oxygen supply are limited due to reduced capillarity density. Nuclear control of cellular processes may also be impaired because these kind of fibers present a low nuclear/cytoplasm ratio. Moreover, larger fibers belong to the white fast type, correlated with pale, soft, exudative meat conditions and their metabolism contributes to a fast pH decline which cannot be removed (Foxcroft et al., 2006; Rehfeldt and Kuhn, 2006; Rehfeldt et al., 2000). Mature piglets exhibiting a strong potential for survival show high concentrations of glycogen in *longissimus dorsi* muscle and liver, stimulated by an increase in cortisol concentrations. This may allow piglets to have a higher ability to maintain glucose levels during and after farrowing and to maintain body temperature in situation of late colostrum intake (Leenhouwers et al., 2002a).

The maturity process of the fetal muscle occurs during the last third of gestation, concretely between the 90th day and the perinatal period (Figures 2 and 5). Muscle studies regarding this gestational period, and performed in extreme breeds in terms of maturity, are particularly interesting to reveal biological processes involved in piglets survival. In this context, numerous transcriptome studies have been performed in porcine muscle tissue in the last decades. This approach is valuable because it allows assessing the gene expression profile easily in a particular tissue or condition (age, genotype, etc.). In the next chapter, a review of the different transcriptome approaches and studies will be presented focusing on studies performed in fetal muscle pig.

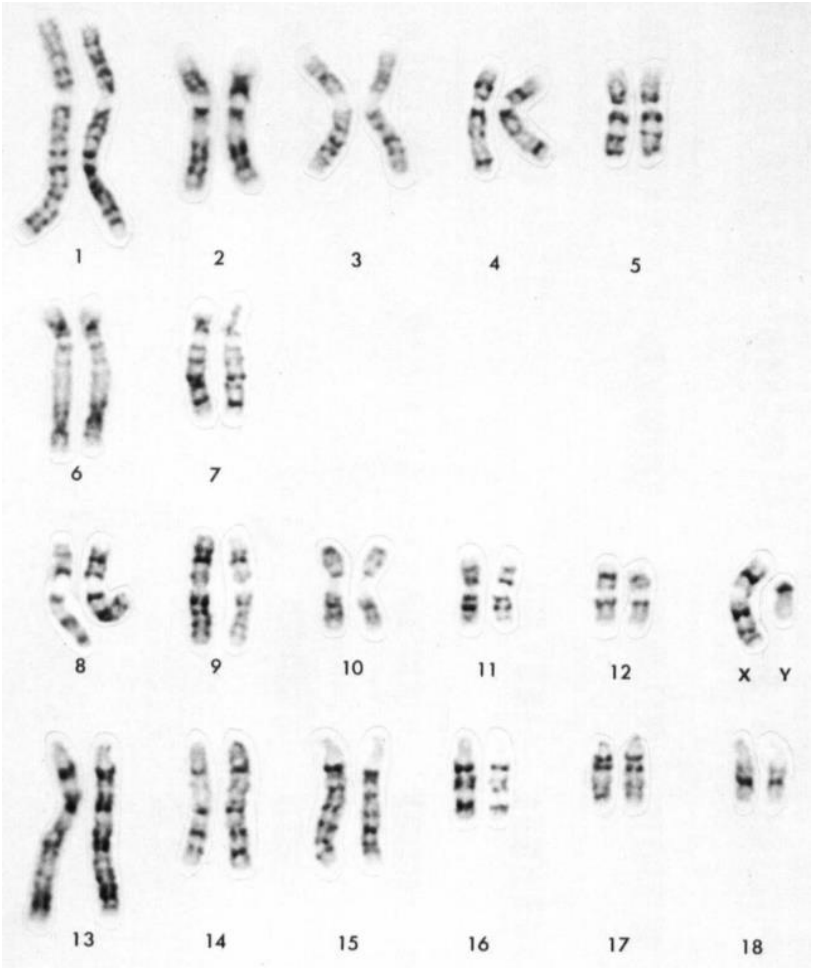


Figure 7. Representative GTG-banded male pig karyotype (Gustavsson, 1988)***.***

3.2 Chapter 2. Muscle transcriptome studies

The majority of phenotypes are complex and quantitative in nature. Understanding the rules that govern the transition from genotype to phenotype requires a comprehensive knowledge of the genome sequence information. Numerous projects of the Encyclopedia of DNA Elements (ENCODE) have been addressed in humans and classical model species. However, transcriptome complexity differs significantly between species (Barbosa-Morais et al., 2012), and little information is available for non-model species such as livestock animals compared with model species. Before addressing the subject of muscle transcriptome studies in pig, a brief description about specific features of the porcine genome sequence and annotation will be presented.

3.2.1 Functional Annotation of porcine genome

3.2.1.1 Main efforts in pig genome sequencing and annotation

The porcine genome is organized in 38 chromosomes (2n): 18 pairs of chromosomes and 2 sex chromosomes. The first five chromosomes are sub-metacentric as shown in Figure 7, chromosomes 6 and 7 are sub-telocentric, chromosomes 8 to 12 are metacentric, and the remaining six are telocentric (Gustavsson, 1988).

A prerequisite for mapping functional elements is a reference genome assembly. Contrary to the human or mouse genomes, which first drafts of their reference sequences were published in 2001 and 2002 respectively, first pig reference genome assembly was published in 2012 (Groenen et al., 2012), after more than 9 years of efforts since the Swine Genome Sequencing Consortium (SGSC) was created in 2003 (Schook et al., 2005). This pig whole genome *de novo* sequencing and assembly (Sscrofa10.2) was produced after the generation of genetic and physical maps (microsatellite linkage and whole-genome radiation hybrid maps). The SGSC adopted then the strategy of shotgun Sanger sequencing of bacterial artificial chromosome (BAC) clone end sequences (Humphray et al., 2007), and complemented latter with Illumina next-generation sequencing. For more details see: (Archibald et al., 2010; Chen et al., 2007; Groenen et al., 2012).

The current pig genome assembly (Sscrofa11.1) was produced and released in December 2016, and produced by the SGSC. Sequence data were largely obtained at 65x genome coverage in whole genome shotgun (WGS) Pacific Biosciences long reads. Sanger and Oxford Nanopore sequence data from a few BAC clones were used to fill gaps and, for final error correction, Illumina HiSeq2500 WGS paired-end and mate pair reads were used. Sscrofa11 replaces the previous assembly, Sscrofa10.2, which was largely established from the same Duroc DNA source. Sscrofa11.1 genome version is estimated to be ~2,500 Mb, with 41.97% of GC content. The total number of scaffolds is 706, with 583 unplaced scaffolds. It contains 1,118 contigs, and the N50 length for the contigs is 48,231,277. The final assembly is available in the public databases (GenBank/EMBL) under the accession number GCA_000003025.6. The primary source of the Sscrofa11.1 assembly is in the NCBI site https://www.ncbi.nlm.nih.gov/assembly/GCF_000003025.6/ (WGS in GenBank accession number:

AEMK00000000.2). Genome annotation for this genome version was available in July 2017 Ensembl v90. Sus scrofa genome contains 22,452 coding genes, 3,250 non coding genes, 178 pseudogenes and 49,488 gene transcripts.

Today, a few percentage of the human and mouse genomes (~0.37% and ~0.14% respectively) is found in unplaced scaffolds, while this percentage is higher in pig genome (~2.66%). Genome annotation is also poorer in pig than in model animals. In general, the annotation of genome sequence in domesticated and farmed species is limited to gene models using RNA expression and DNA variation data, which is insufficient to characterize the complexity of the transcriptomes in domesticated animals. These aspects highlight the difficulties that scientist must confront when working with species other than model ones.

In an effort to improve the annotation of newly assembled genomes of domesticated and non-model organisms, the Consortium of Functional Annotation of Animal Genomes (FAANG) was recently created (www.faang.org, (Andersson et al., 2015; Tuggle et al., 2016)). The aim of this Consortium is to produce comprehensive maps of functional elements based on common standardized protocols and procedures. Studies performed in the FAANG context have been mainly focalized on chicken, pig, cattle, and sheep, at neonatal and mature stages. Studied tissues include: skeletal muscle, adipose, liver and tissues collected from reproductive, immune and nervous systems. The main assays are based on RNA sequencing, chromatin accessibility and architecture, and histone marks. In this context, a French pilot project (FrAgENCODE) of the French National Institute of Agronomic Research (INRA) has been developed to asses the expression profiles, chromatin accessibility and structure in several tissues of four different farm species. This will be presented latter in more details.

3.2.2 Transcriptome technologies and approaches

3.2.2.1 DNA microarray and RNA-seq

The full range RNA molecules expressed by an organism comprise messenger RNA (mRNA), transfer RNA (tRNA), ribosomal RNA (rRNA), and regulatory RNAs (miRNA, RNAi, siRNA, piRNA, lncRNA, snRNA, snoRNA or circRNA), being the most abundant the rRNA, the tRNA and the mRNA. The transcriptome is defined as the full range of RNA molecules expressed by an organism, tissue, or cell type, in a particular condition, and is generally referred to the messenger RNA (mRNA) but it can also refer to other RNA types. Therefore, a transcriptome analysis allows determining expressed (active) and non-expressed (inactive) genes in a population of cells.

The two most commonly used transcriptomic techniques are DNA microarrays and RNA sequencing (RNA-seq). Microarrays, also known as DNA chip or biochip, are used since early 80s: (a) to measure thousands of genes at the same time, (b) for gene expression profiling, (c) to genotype multiple regions of a genome (d) for single nucleotide polymorphisms (SNP) or alternative splicing detection, etc. On the microarray, specific DNA sequences called “probes” or “oligos”, are used to hybridize anti-sense RNA or complementary DNA (cDNA, synthesized from a single stranded RNA). Probe-target is quantified by the detection of a fluorophore. RNA-seq, also named whole transcriptome

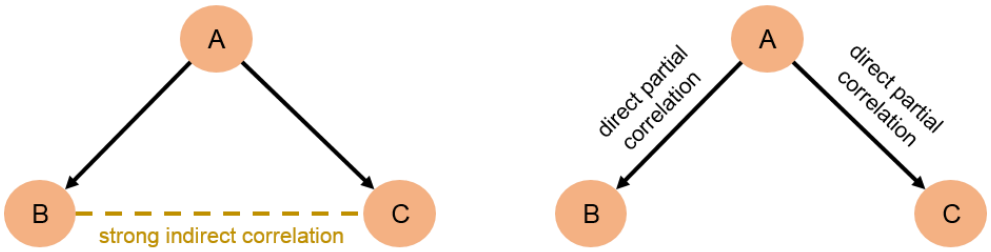


Figure 8. Schematic illustration of pairwise correlations and partial correlation assumptions. *Circles represent nodes (genes) and black arrows represent an observed correlation between nodes. Left: Computing pairwise correlations can lead to misconceptions. In the example, when two genes "B" and "C" are regulated by a common gene "A", the coefficient between the expression of "B" and the expression of "C" is strong as a consequence (dotted line). Right: By computing partial correlations there is no undesirable effects of strong indirect correlations.*

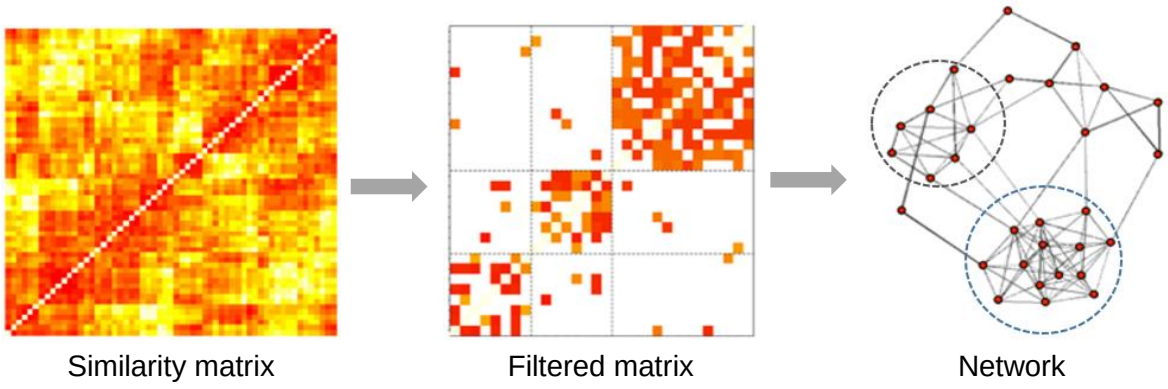


Figure 9. Basic steps of network inference. *First, pairwise similarities are computed (correlations in the simplest case). Second, the smallest (or less significant) similarities are filtered (using a threshold chosen either heuristically, or other more sophisticated methods). Third, the network is built from the remaining similarities.*

sequencing (WTSS), is a more recent technology, first used in 2008 (Lister et al., 2008) and based on next-generation sequencing (NGS). This technology is useful not only to obtain gene expression profiling but to detect alternative gene spliced transcripts, post-transcriptional modifications, or SNPs. It needs to prepare cDNA from isolated RNA before sequencing. RNA can be enriched for a specific type (i.e. by using 3' polyadenylated (poly (A)) tails to include only mRNA). After sequencing, transcriptome assembly and annotation are necessary before analyzing data. One common way for analyzing transcriptome data is by constructing gene co-expression networks. In the section below is presented a brief review about the methods and characteristic of these networks.

3.2.2.2 Co-expression networks

Gene co-expression networks are mathematical representations to model relations between genes behaving in a similar way across tissues and experimental conditions. In these kind of networks, each vertex (node) corresponds to a gene, and pairs of genes are connected by an edge when a significant co-expression relationship exists between the pairs. The first step to infer a co-expression network is to calculate pairwise similarities between pairs of genes (often by computing Pearson correlations for "relevance networks") (Zhang and Horvath, 2005). Although this approach can be useful to have a first look at relationships between co-expressed genes, it can also lead to misconceptions because Pearson correlations are sensitive to unwanted indirect effects, such as the effect of a common strong correlation with another gene (Figure 8). To account for the effect of all expression data and obtain a measure closer to direct interactions between genes, it is thus advised to use more sophisticated methods, such as Graphical Gaussian Models (GGM) (Edwards, 1995). GGM base the definition of the network on the measure of a *partial correlation*, i.e., a correlation between two gene expressions knowing the expression of all the other genes. This method was found more efficient, for instance, to group genes with a common function (Villa-Vialaneix et al., 2013). After computing pairwise similarities, those less significant in the similarity matrix are filtered by fixing a threshold to discard the less significant ones, then, the network is built from the remaining pairwise similarities between genes (Figure 9).

Once the network inferred, many network characteristics can be used to extract information about the most important nodes, or group of nodes, which will be helpful for interpreting the biological meaning of co-expressed genes. This is the so-called process of "Network mining". Network features can be classified as global characteristics of the network (i.e. density, transitivity), or as individual characteristics of a node (i.e. degree, betweenness). The individual characteristics are particularly interesting to extract the most important nodes, or genes in the case of gene co-expression networks. The degree of a node is the number of edges afferent to this node and the betweenness of the node is the number of shortest paths between pairs of nodes in the network that pass through that node. High-degree genes are connected to many other genes while high-betweenness genes are central and more likely to disconnect the network if removed. Finally, a clustering of the nodes can be performed to partition the network into groups of densely connected genes (sharing more edges than with other groups). These groups are called clusters or communities, and they are often used to find enriched biological processes or molecular functions by using Gene Ontology (GO) approaches.

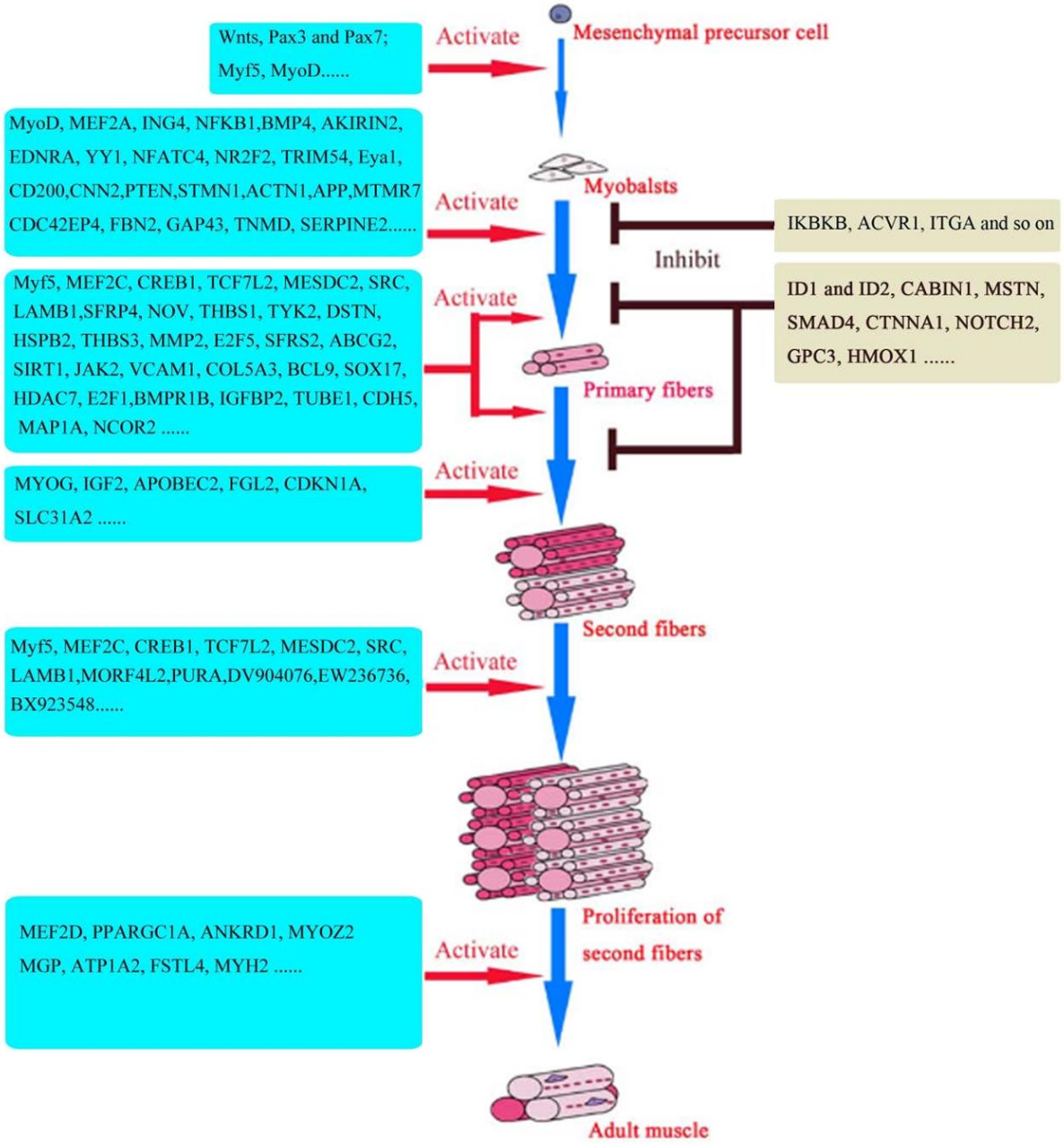


Figure 10. The probable roles of differentially expressed (DE) genes in the molecular regulation of myogenesis. *Red dots indicate the promoting roles of these genes in myogenesis and blank striping indicate the repressing roles* (Zhao et al., 2011).

3.2.3 Muscle transcriptome studies in pigs

Many transcriptome studies have been performed in pig to elucidate the mechanisms that govern porcine skeletal muscle development and maturity, but few of them address the gestational period when the skeletal muscle maturation process takes place (90th day of gestation and the end of gestation) (Voillet et al., 2014; Zhao et al., 2015). Some of these studies include, or are centered on the period before the maturation process (Cagnazzo et al., 2006; Tang et al., 2015a; Zhao et al., 2011, 2015), and some others include or mainly concern the study of the postnatal period (Ayuso et al., 2015; Óvilo et al., 2014; Sodhi et al., 2014; Xu et al., 2012; Zhao et al., 2011, 2015). Generally, most of them are based on comparisons between transcriptomes of two extreme breeds. Often, between highly selected breeds (selected for lean meet), and non-selected breeds. Four of these studies are particularly interesting as they explain phenotypic traits observed in breeds highly selected for muscle growth, and characterized by a high incidence of perinatal mortality. Firstly, some authors observed in Duroc (DU, high intramuscular fat) and Pietrain (PT, low intramuscular fat) breeds that myogenesis is more intense in late PT fetuses than in DU ones, and genes related to energy metabolism are expressed at a higher level in PT than in DU prenatal pigs (Cagnazzo et al., 2006). Then, a similar study performed in Lantang (LT, obese) and Landrace (LR, lean) breeds (Zhao et al., 2011) revealed that some differentially expressed genes might contribute in later myogenesis and more muscle fibers in LR than in LT. Another study, focused on the maturation process period in the MS (strong potential for survival) and the LW (high incidence of mortality) breeds, reported that: (a) genes involved in muscle development were enriched at 90 days of gestation, while those involved in metabolic functions were enriched at 110 days, (b) it exists a delay of gene expression in LW fetuses at 110 days of gestation which concerns globally genes involved in muscle development and metabolic functions (Voillet et al., 2014). Lastly, another study performed in the Tongcheng breed (TC, slow growth) and the American version of the LW, the Yorkshire breed (YK, fast growth, low back fat and high lean meat), revealed a higher number of myoblasts (myogenic progenitor cells) in early TC embryos than in YK embryos (Tang et al., 2015a). These results suggest that pig breeds characterized by low back fat and high lean meet composition as LW, LR, PT or YK, present a delay in expression of genes involved in muscle development and maturity.

To finish with this overview of the porcine muscle transcriptomes, the study of Zhao et al. (2011) performed at several time points of the prenatal and postnatal periods, nicely illustrates the whole process of myogenesis by indicating the main expressed genes for each step (Figure 10).

3.3 Chapter 3. Nuclear architecture

Genome sequence alone is not sufficient to explain cell type diversity and the overall coordination of nuclear activity in a particular tissue. Even though cis- and trans-acting regulatory sequences are among the most studied regulatory elements, they are not the only determinants of gene expression. For instance, epigenetic mechanisms such as histone and DNA modifications can also be responsible for tissue-specific expression of genes (Rothbart and Strahl, 2014). Nevertheless, numerous studies have demonstrated that the genome organization in the nucleus acts as an additional level of gene expression regulation (Osborne et al., 2004; Rieder et al., 2014; Schoenfelder et al., 2010; Zhao et al., 2006).

In the present section, the main generalities about nuclear architecture, more specifically about genome organization, will be presented. Additionally, an overview of the principal experimental methodologies and applications to study this matter will be presented, together with a description of current studies about genome organization performed in pig.

3.3.1 Higher order genome organization

3.3.1.1 Generalities

Genome organization extremely differs among biological organisms. The most important distinction regarding genome structure is the one found between prokaryotes and eukaryotes organisms. Prokaryotes lack of nuclear membrane, genome has relatively small size, is often circular, it generally contains only one chromosome, and may have additional DNA molecules (plasmids). In contrast, in eukaryotes the genome is located inside a nuclear membrane, and it contains larger and multiple linear DNA molecules (except for mitochondrial and chloroplast circular DNAs) which are condensed into chromosomes by association with histone proteins. Eukaryotic genome is also more complex with longer genes, and around only 1.22% of coding sequence (for protein-coding exons in human (ENCODE Project Consortium, 2012)) while prokaryotes has up to 90%.

The fundamental units of the genome are the chromosomes, which are made of chromatin in eukaryotic cells (DNA compacted by association to histone proteins). The higher level of DNA compaction is found in mitotic cells, with the metaphase chromosomes. Chromatin is subdivided into euchromatin, correlated to “open” and transcribed chromatin (R-bands of metaphasic chromosomes), and heterochromatin, more condensed chromatin (G-bands of metaphasic chromosomes) enriched into inactive and silenced chromatin regions. In interphase nuclei, the distribution of the chromatin is not random and is constrained by the presence of several nuclear structures such as, proteinaceous nuclear bodies (PML bodies, Cajal bodies or Polycomb bodies), nucleolus, nuclear lamina, nuclear pores, transcription factories (TFs) or splicing speckles (Schneider and Grosschedl, 2007).

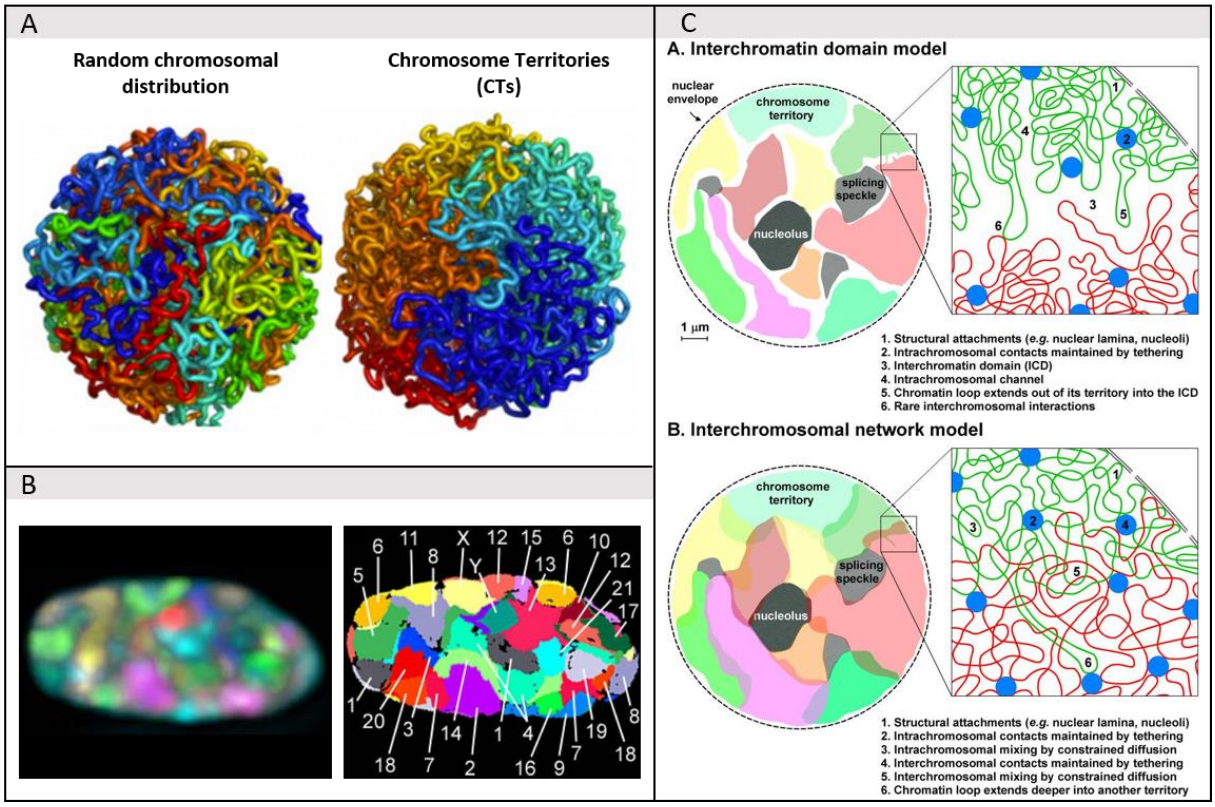


Figure 11. Chromosome territories (CTs). *(A) Random chromosomal distribution vs. the CTs model* (Lieberman-Aiden et al., 2009)*. (B) Representation and classification of chromosomes in a human fibroblast nucleus. Left: CTs were targeted by 3D FISH using seven different fluorochromes to construct the CT painting probes, and DAPI to counterstain the DNA. Right: False color representation of all CTs* (Bolzer et al., 2005)*. (C) Models of chromatin organization in mammalian nuclei. Up: In the CT-IC model, also called interchromatin domain (ICD) model, chromatin from different chromosomes is separated by an ICD compartment rich in nuclear machinery but poor in chromatin. Rare chromatin loops extending from CTs may invade the ICD space. Down: In the ICN model, chromatin from different chromosomes is allowed to expand into the surrounding territories; the presence of adjacent chromosomes, the nuclear membrane, and larger nuclear compartments restrict the amount of intermingling.* (Branco and Pombo, 2006).

3.3.1.2 Chromosome territories

In situ hybridization techniques have allowed visualizing individual chromosomes in the interphase nuclei. This permitted to accept the theory that chromosomes occupy discrete territories in the nucleus, the so-called chromosome territories (CTs), against the theory of global intermingling of interphase chromosomes (Figure 11A-B) (Bolzer et al., 2005). CTs are the basic principle of nuclear organization in animals, plants and yeast (Cavalli and Misteli, 2013; Cremer and Cremer, 2010). Although chromosomes occupy discrete regions in the nuclear volume, they are not necessarily completely separated one from each other. Different models are proposed, the most popular ones: the chromosome territory-interchromatin compartment (CT-IC) model, and the interchromatin network (ICN) model (Branco and Pombo, 2006) (Figure 11C). The CT-IC model postulates that two spatial compartments are present in the nucleus, one formed by the CTs, and the other one called the interchromatin-compartment (IC) and defined as a DNA-free space, rich in soluble nuclear machinery such as TFs or splicing speckles. The ICN model establishes that CTs are not separated by a DNA-free compartment, but chromatin expands into the surrounding CTs allowing a certain degree of intermingling at the interfaces of neighboring chromosomes with the presence of nuclear machinery in intermingling regions. In the CT-IC model, *trans*-chromosomal interactions could occur via extended chromatin loops, while in the ICN model, regions of intermingling would be more likely to produce *trans*-chromosomal interactions. A more recent study argues against these two models (Nagano et al., 2013). Firstly, local dissociations from CTs (necessary for extended loop formation in the CT-IC model) were not observed. Secondly, the observed preferential location of *trans*-chromosomal interactions associating some pairs of chromosomes, and the lack of contacts between other chromosome pairs, argue against the idea of domains completely immersed in other territories. These results do not exclude CT intermingling, but propose an intermediate model that includes preferential regions of intermingling altogether with DNA-free interface regions. It is not excluded that other eukaryotic organisms show a different chromosomal conformation. For instance, the CTs of the yeast *S. cerevisiae* are spatially less well defined and intermix to a much greater extent than those of higher eukaryotes. This is possibly due to yeast genome specificities (more decondensed chromatin, lack of large heterochromatin domains and smaller genome size) (Cavalli and Misteli, 2013).

Aside from the special distribution of chromosomes in CTs, it has been observed that p and q arms of metacentric chromosomes are also quite separated entities (Bickmore, 2013). Moreover, a special localization of centromeres has been observed in yeast, fly, mouse and human (Li et al., 2017). Indeed, the centromeres tend to cluster and are positioned at the periphery of the nucleolus during interphase, and this process is thought to play a role in determining the overall genome architecture. Finally, a specific phenomenon of homologous chromosomes pairing called transvection has been observed in *D. melanogaster* and other dipteran insects. This pairing can influence gene expression by forming interactions between regulatory elements on homologous chromosomes (Li et al., 2017).

Other elements influencing genome organization are gene density, active and repressive domains, and specialized nuclear structures. Regarding gene density, it has been observed in human, rodents, cattle and birds, that gene-rich chromosomes tend to be located towards the center of the nucleus

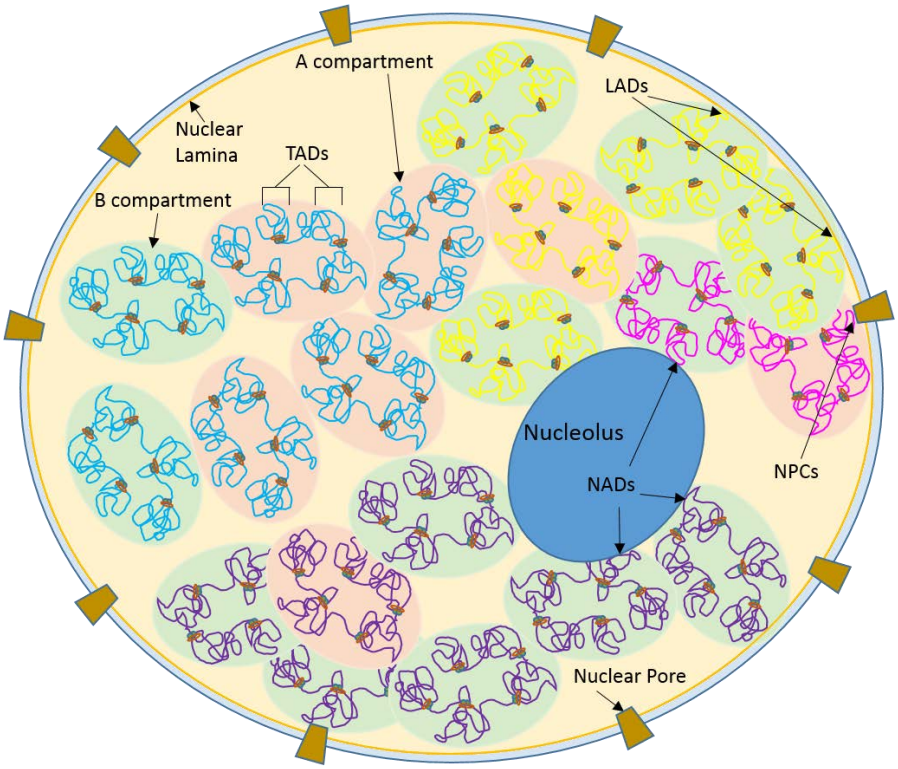


Figure 12. Nuclear architecture and genome organization. *This picture illustrates the principal components of the nucleus involved in the nuclear organization (nuclear lamina, nuclear pore and nucleolus), and the main genome domains (CTs, LADs, NADs, NPCs, A/B compartments and TADs). CTs are represented by different chromatin colors (blue, yellow, pink and purple).*

and gene-poor chromosomes toward the nuclear periphery (Cremer and Cremer, 2010; Gibcus and Dekker, 2013). Moreover, a polarized nuclear organization within chromosomes was observed, with gene-poor regions located towards the nuclear periphery compared with gene-rich regions from the same chromosome (Bickmore, 2013). Correlation between this non-random radial localization and gene activity has been observed in selected cases however, the nuclear periphery is not entire restrictive to transcription (Deng and Blobel, 2014).

3.3.1.3 NPCs, LADs, NADs, TFs and PcG domains

The main structures responsible of genome organization at the nuclear periphery are the nuclear lamina (NL) and the nuclear pores (NP) (Figure 12). NL is a protein network that covers the inner part of the nuclear envelope, and many studies have reported associations of the chromatin to this structure (Holwerda and de Laat, 2012). These are the so-called lamina-associated domains (LADs), reported before in human, fly and mouse (Li et al., 2017). They are characterized by heterochromatic regions, low gene density, transcriptional inactivity and depletion for transcription marks such as RNA polymerase II (RNAPII) and histone marks. LADs are large domains spanning (0.1-10 Mb) representing almost half the genome in a given cell population but not all LADs can physically be associated with the NL in each cell (Bickmore, 2013; Gibcus and Dekker, 2013). The differences in genome organization between a cell population and a single cell will be discussed later. Whereas NL associates with heterochromatin (inactive domains), NP are in some cases enriched for associations with euchromatin and active genes. These are the so-called nuclear pore complexes (NPCs) (Deng and Blobel, 2014; Gibcus and Dekker, 2013). For instance, in yeast active genes reside proximal to nuclear pores while in mammals, active genes did not exhibit such positioning preferences. Nevertheless, nuclear envelope is not the only organizer of genome structure, other nuclear bodies such as nucleolus, Polycomb bodies or TFs play an important role in genome organization.

Nucleoli are subnuclear structures specialized in ribosome biogenesis and enriched in RNA polymerase I (RNAPI) responsible of 45S rDNA transcription (Pombo and Dillon, 2015). In human, over 2000 clustered rRNA copies dispersed over five chromosomes (in human) are recruited together and transcribed on the surface of the fibrillar center within the nucleolus (Mercer and Mattick, 2013). Moreover, actively transcribed RNAPIII-dependent genes can also be found at the nucleoli, and some groups of RNAPII-dependent genes such as olfactory receptors have been also identified at the nucleoli; however these RNAPII-dependent genes are silent (Gibcus and Dekker, 2013). All these loci that associate at or near nucleoli are described as nucleolus-associated domains (NADs) (Figure 12). Hence, nucleoli are genome organizing structures bringing together actively transcribed RNAPI and RNAPII-dependent genes, as well as silenced repressive loci surrounding the sites of ribosomal synthesis (Deng and Blobel, 2014; Gibcus and Dekker, 2013). Nucleoli are a highly specialized example of RNAPI transcription factories (TFs) responsible of rDNA transcription but it is not the only one. TFs are defined as large nuclear assemblies containing a range of transcription factors and machinery constituents along with additional accessory proteins for RNA processing and splicing (Mercer and Mattick, 2013). RNAPII is associated to the transcription of most protein-coding genes. RNAPIII is responsible for the synthesis of 5S rRNA and tRNA and is also associated with clusters of 5S rRNA and tRNA transcripts (Rieder et al., 2012). RNPII and RNAPIII TFs are distributed through the nucleoplasm in foci, are more

abundant than RNAPI TFs, but contain far fewer polymerases (Pombo and Dillon, 2015). The number of TFs, and of polymerase molecules appears to depend on the cell type and species, for example, in HeLa cells there are about 8000 RNAPII factories and 2000 RNAPIII factories, each containing approximately 6 to 8 active enzymes (Pombo and Dillon, 2015; Rieder et al., 2012). In terms of nuclear structure, there is some evidence that TFs can lead to the clustering of co-regulated genes. Indeed it exists some cases where transcriptionally-related genes are transcribed in specialized TFs. A well-known example of this are the TFs enriched with the Klf1 transcription factor that mediates preferential co-associations with Klf1-responsive globin genes in erythroid cells (Schoenfelder et al., 2010).

Another structure playing a role in genome organization are the Polycomb bodies, identified in fly and mammals. They are composed by the Polycomb group (PcG) proteins, a collection of transcriptional regulatory factors mainly involved in gene silencing. PcG transcriptional repression occurs by imposing post-transcriptional modifications on histones and inducing chromatin condensation, which in turns restrain RNAPII elongation. More recently, PcG proteins have also been identified as coactivators of gene expression by regulating local topological interactions (Aranda et al., 2015). Regarding the regulation of chromatin structure, it was observed in mouse that most PcG-associated genes are contained within a loop flanked by CTCF/cohesin sites. These genes are included in the so-called chromatin structures PcG domains that average 112 Kb, and include repressive histone methylation marks (Dowen et al., 2014). Some of the best characterized PcG domains are the *Hox* gene clusters (Vieux-Rochas et al., 2015). PcG shapes intra-TADs (topologically associated domains) interactions and might help to stabilize and consolidate TADs of transcriptionally inactive regions of the genome (Aranda et al., 2015). A detailed view about the TADs will be discussed latter.

3.3.1.4 A and B compartments

Besides CTs, subchromosomal compartments within CTs have been identified. They are made-up of groups of multi-Mb chromosomal domains (median size ~3 Mb in mice, (Dixon et al., 2012)), mostly located in the same chromosome but can also be on different chromosomes (Gibcus and Dekker, 2013). Those are the A and B compartments, first described by Lieberman-Aiden et al. in human cells (Lieberman-Aiden et al., 2009). The A compartments are defined as transcriptionally permissive, euchromatic regions, which are gene-rich and DNase I hypersensitive areas, also referred as open compartments. Inversely, B compartments are considered as transcriptionally inert regions enriched for features of heterochromatin and nuclear lamina associations, which are gene-poor, DNase I insensitive, and are also referred as closed compartments (Bonora et al., 2014; Gibcus and Dekker, 2013). Different strategies have been adopted to define A and B compartments: by using High throughput Chromosome Conformation Capture (Hi-C) data, DNA methylation microarray data, DNase I hypersensitive sequencing, single cell whole-genome bisulfite sequencing, and Assay for Transposase Accessible Chromatin with high-throughput sequencing (ATAC-seq) (Fortin and Hansen, 2015).

Hi-C data allow to identify genomic compartments or domains at different scale levels depending on data resolution. Lieberman-Aiden et al. first defined the A and B compartments by analyzing low resolution matrices of a human lymphoblastoid cell line (Lieberman-Aiden et al., 2009).

Table 1. Specific characteristics of each A and B subcompartment according to Rao et al. 2014.

		A and B subcompartments					
		A1	**A2**	**B1**	**B2**	**B3**	**B4**
activating chromatin marks	H3K36me3	+	+	-	-	-	+
	H3K79me2	+	+		-	-	
	H3K27ac	+	+		-	-	
	H3K4me1	+	+		-	-	
silencing chromatin marks	H3K9me3		+		-	-	+
	H3K27me3			+			
	H4K20me3						+
gene features	density	+	+				
	expression	+	+				
genomic domains	LADs	-	-		+	+	
	NADs	-	-		+	-	

Later, they performed new experiments in the same cell line that allow obtaining high resolution matrices (1 Kb), and six nuclear subcompartments were identified (Table1) (Rao et al., 2014). These subcompartments were associated with distinct patterns of histone modifications and named: A1, A2, B1, B2, B3 and B4. A1 and A2 are gene dense, have highly expressed genes, harbor activating chromatin marks (H3K36me3, H3K79me2, H3K27ac and H3K4me1) and are depleted at LADs and NADs. A2 is more strongly associated with the presence of H3K9me3, has lower GC content, and contains longer genes than A1. B1 correlates positively with H3K27me3 and negatively with H3K36me3, suggestive of facultative heterochromatin. B2 and B3 tend to lack all of the above-noted marks. B2 includes 62% of pericentromeric heterochromatin and is enriched at LADs and NADs, while B3 is only enriched at LADs but strongly depleted at NADs. Finally, B4 is only present in a region highly enriched with members of the KRAB-ZNF superfamily genes, which exhibit a highly distinctive chromatin pattern, with strong enrichment for activating chromatin marks (H3K36me3) and heterochromatin-associated marks (H3K9me3 and H4K20me3).

It remains unclear whether these A and B compartments are stable or if they change in specific conditions. A recent study showed that changes in gene expression were associated with switches between compartments in 36% of the genome during mammalian development (Dixon et al., 2015). The A and B compartments are further subdivided into Topologically Associated Domains (TADs), which are further partitioned into smaller substructures and contact domains (Rao et al., 2014; Zhan et al., 2017). Last studies have been focused on the description of these smaller domains (TADs), and little work is available about descriptions of A and B compartments behavior in different conditions.

3.3.1.5 Topologically associated domains

Decreasing in the genome organization scale, domains smaller than A and B compartments were first identified by Nora et al. in mice active and inactive X chromosomes, and were named topologically associated domains (TADs) (Nora et al., 2012). TADs are contiguous genomic regions that range approximately 1 Mb size (Dixon et al., 2012; Nora et al., 2012). They are defined as chromatin domains enriched in highly-self interacting regions, with a frequency of intra-domain interactions higher than inter-domain interactions. These domains are highly conserved between cell types and across species, including human, mouse, fly, bacteria, yeast and plants (Björkegren and Baranello, 2018; Dixon et al., 2012), and genes located within the same TAD tend to have coordinated dynamics of expression during differentiation. Hence, TADs may play a role in coordinating the activity of groups of neighboring genes (Gibcus and Dekker, 2013).

A very characteristic feature of TADs is that their boundaries are enriched in DNA-binding proteins such as the CCCTC-binding factor (CTCF) in mouse, human and fly cells (Figure 13A) (Dixon et al., 2012; Li et al., 2017). This could suggest that CTCF might be involved in the establishment of TAD boundaries. However, in *Drosophila*, CTCF does not seem to have a role in loops formation (Björkegren and Baranello, 2018). Moreover, in human and mouse, only 15% of CTCF binding sites are located within boundary regions while the other 85% are present inside TADs (Dixon et al., 2012) indicating that CTCF alone is insufficient to separate different TADs (Ong and Corces, 2014). The role of CTCF at these sites will be addressed later.

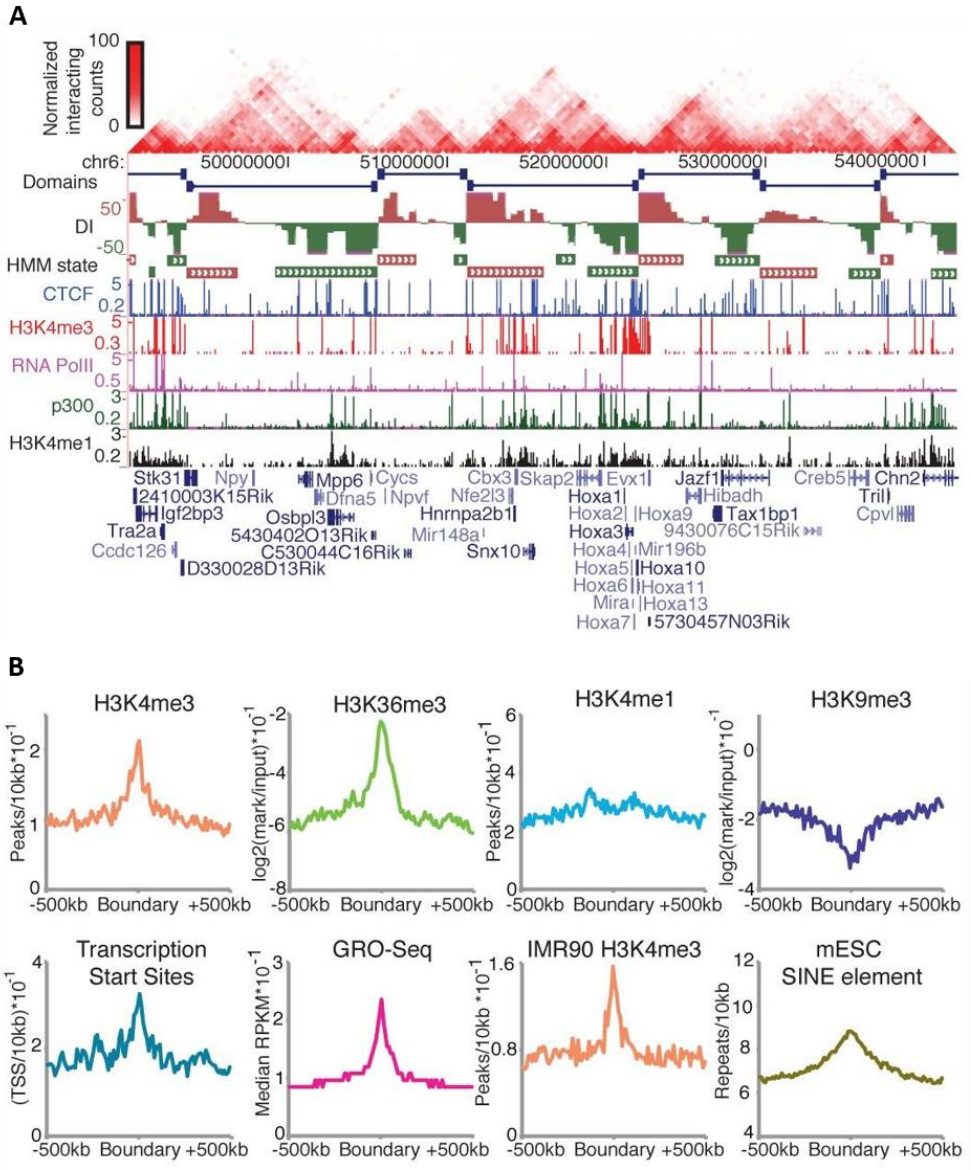


Figure 13. Topological domains and boundaries regions. *Images obtained from a study performed in mouse embryonic stem cells (mESC)* (Dixon et al., 2012). *(A) TADs obtained from Hi-C data (red triangles), overlayed on ChIP-seq data. (B) Density of peaks for different histone marks, transcription start sites, genome-wide nuclear run-on sequencing (GRO-Seq: short transcripts generated by engaged RNA polymerase) and SINE elements around TAD boundaries.*

CTCF is not the only element found enriched at TAD boundaries. Cohesin, a conserved ring-shaped protein complex, is often found co-localized with CTCF and enriched at TAD boundaries (Uusküla-Reimand et al., 2016). In fact, both CTCF and cohesin have been observed to be involved in gene expression regulation by shaping chromatin through loops formation (Björkegren and Baranello, 2018; Hnisz et al., 2016a; Rao et al., 2014). Chromatin marks associated with active transcription (more concretely with active promoters) of nearby genes, such as H3K4me3 and H3K36me3, have also been found enriched at boundaries. In contrast, non-promoter associated marks, such as H3K4me1 (associated with enhancers) and H3K9me3 (associated with heterochromatin), have been found not enriched or specifically depleted at boundary regions (Figure 13A and B). Likewise, transcription start sites (TSS) and repeat classes such as Short Interspersed Nuclear Element (SINE), are enriched at boundaries regions (Figure 13B), and "housekeeping genes" have been found strongly enriched near TAD boundaries (Cournac et al., 2016; Dixon et al., 2012). SINE-repetitive elements preferentially co-localize in the nuclear space and are enriched in transcription factors in human, mouse and fly, which may explain the global conservation of genome folding (Cournac et al., 2016). Besides, some non-coding RNAs may be involved in genome folding and gene expression regulation, such as the non-coding RNAs derived from Long Interspersed Nuclear Elements (LINEs) (Nozawa and Gilbert, 2014), or the intergenic long non-coding RNAs (lincRNAs). Indeed, lincRNAs show a preferential location at TAD boundaries, and are enriched in enhancer-like signatures, suggesting a regulation of proximal gene expression by modulating local chromosomal architecture (Tan et al., 2017). The fact that DNA associating proteins, transcriptional histone marks, TSS, repetitive elements, and lncRNAs, are preferentially enriched at TAD boundaries, together with coding genes localized near boundaries, strongly points to a potential role of boundary regions in the regulation of gene expression.

Smaller domains than the TADs (1 Mb) described by Dixon et al. have been observed in other studies (Rao et al., 2014; Sexton et al., 2012; Zhan et al., 2017) employing higher resolution maps. These domains range in size from 40 Kb to 3 Mb (median size 185 Kb), and are described as "contact domains" or sub-TAD structures (Rao et al., 2014). They were probably not observed by Dixon et al. because detecting smaller structures (sub-TADs) requires higher resolutions than the ones used in their study as discussed in Rao et al. (2014). As mentioned before, CTCF and cohesin are enriched at TAD boundaries but they also bind pervasively within TADs and are involved in the formation of sub-TAD structures, which are strongly associated with active regulatory sequences (Phillips-Cremins et al., 2013). It is not clear whether these subdomains are different from TADs or if they simply represent a further level of hierarchical organization (Björkegren and Baranello, 2018).

The 3D genome organization offers a hierarchical complexity (including from higher to lower scales: CTs, LADs and NADs, A and B compartments and subcompartments, TADs and sub-TADs) which is achieved with the presence of chromatin loops observed at the highest resolutions. The role of CTCF and cohesin, among other factors, in the mechanisms of loop formation has been quite well studied and will be detailed in the following section.

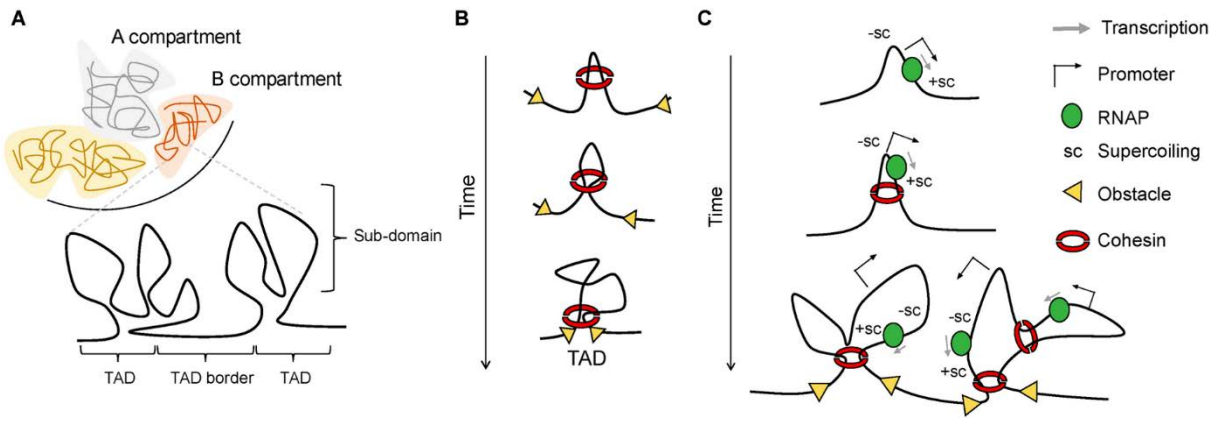


Figure 14. Mechanisms of loop formation. *The loop extrusion model here illustrated includes the cooperation of cohesin, CTCF, RNAP, and supercoiling. (A) Higher order organization of chromatin into A and B compartments, formed by topological domains (TADs), that are in turn subdivided into smaller domains. (B) Loop extrusion model, which main actors are cohesin and CTCF (yellow arrow), the last acting as a blocker. (C) RNAP introduces positive supercoiling that may be responsible of cohesin progression* (Björkegren and Baranello, 2018)**.**

3.3.2 Chromatin loops and gene-gene interactions

3.3.2.1 CTCF and cohesin functions

CTCF is an architectural protein conserved in most animals, and it contains a highly conserved DNA-binding domain (Kim et al., 2007). Around the 55,000-65,000 sites in mammalian genomes, approximately 5,000 are ultraconserved among species and tissues, whereas 30-60% of CTCF sites show cell-type specific distribution (Ong and Corces, 2014). This CTCF target selectivity can be explained by differential methylation in specific CpG dinucleotides at the CTCF recognition sequence (Wang et al., 2012). Classically, CTCF was initially associated to the roles of chromatin barrier (function to prevent repressive heterochromatin from spreading into a neighboring domain) and enhancer activity blocker (by association with insulators, sequences that block the action of enhancers on promoters). However, recent studies argue against these two proposed functions. Indeed, there is little evidence to support a generalized functional role for CTCF in separating domains with different epigenetic marks, and CTCF could participate in both, enhancer blocker and enhancer facilitator functions (Ong and Corces, 2014). New functions associated to CTCF are related to its ability to: (i) bring together distant sequences such as enhancer-promoters or distant gene segments, (ii) control transcriptional events such as RNAPII pausing and alternative mRNA splicing, (iii) stabilize interactions required for the formation of TAD borders together with the cooperation of other architectural proteins (Ong and Corces, 2014).

Cohesin is essential to establish sister chromatid cohesion during the S phase of the cell cycle, and maintaining it through G2 and mitosis, by forming a ring structure loaded onto DNA during G1. A large number of cohesin-binding sites co-localizes with binding of CTCF, and it is been suggested that both proteins are primary involved in promoting promoter-enhancer interactions by forming chromatin loops. But, they could also have some involvement in delineating boundaries between TADs (Pombo and Dillon, 2015).

3.3.2.2 Insulated neighborhoods (CTCF/cohesin-mediated loops)

One hypothesis to explain the mechanism of loop formation mediated by CTCF and cohesin is the loop extrusion model. Björkegren et al. proposed cohesin as a loop extruding factor, in a way that DNA could pass through the ring and the extrusion would stop when the ring meets an obstacle. This obstacle could be a DNA site occupied by CTCF on each side of the growing loop (Figure 14B). In addition, they proposed that RNAP may be involved in this mechanism, which could also contribute to the formation of TADs structure (Figure 14A-C). Indeed RNAPII has been detected in loop structures included within CTCF-mediated in chromatin contact domains, suggesting that CTCF-anchor regions are the foci for transcriptional activity (Tang et al., 2015b). Positive supercoiling introduced by RNAP, could “push” cohesin along the double helix, providing an impulse for the extrusion of the loop (Björkegren and Baranello, 2018).

This mechanism of loops formation likely involves CTCF dimerization due to the convergent orientation of the two CTCF motif present at the loop anchors (Rao et al., 2014), suggesting that CTCF may participate to the stabilization and maintain of the loop. Rao et al. observed that the vast majority

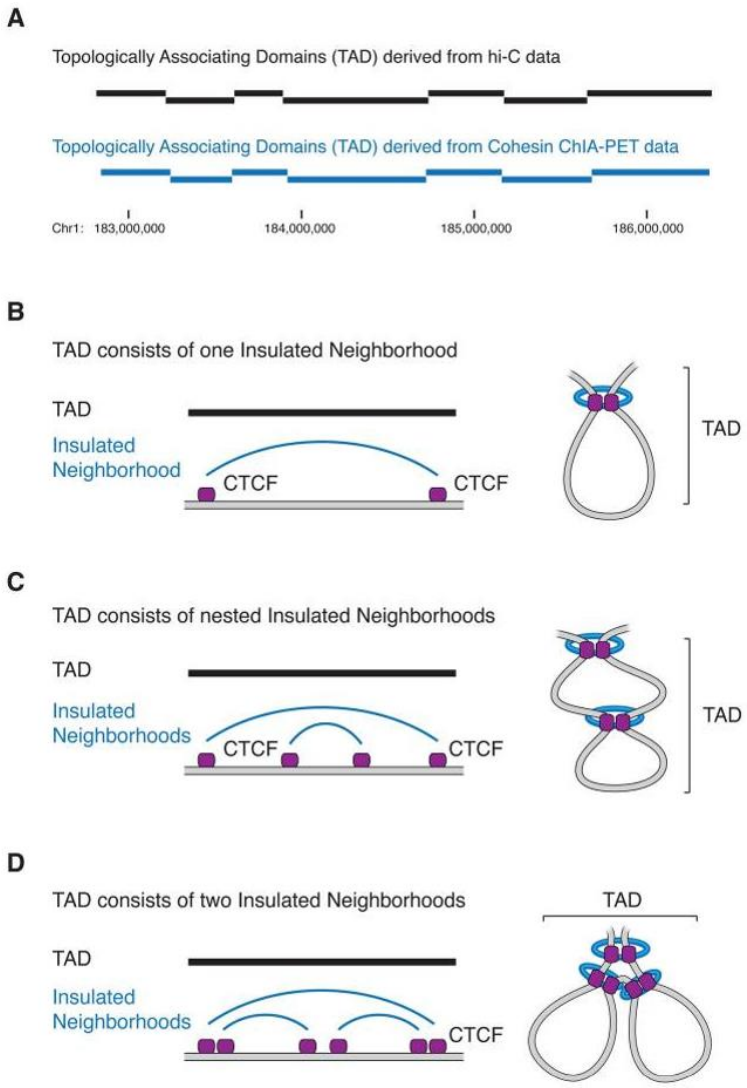


Figure 15. Models of loop domains to constitute TAD structure and sub-structure. *(A) Identification of TADs (black and blue bars) by Hi-C and cohesin ChIA-PET. (B-D) Models of a TAD that consist of (B) an insulated neighborhood, (C-D) nested insulated neighborhoods (sub-TADs within a TAD)* (Hnisz et al., 2016a).

of detected loops bound CTCF and two proteins of the cohesin complex (RAD21 and SMC3) at the loop anchor region. Two-thirds of loops contain a single CTCF-binding motif, but 92% of motif pairs are facing one another. Many of the loops detected by Rao et al. demarcate sub-TADs (contact domains, 185 Kb median size), suggesting that CTCF delimits structural and regulatory domains, and that the two anchor sequences of the loop are located at these domain boundaries. These are referred as "loop domains". Figure 15 illustrates possible models of loop domain to form TAD structures and sub-structures (Hnisz et al., 2016a). Finally, they also observed that loops frequently have a promoter at one anchor locus and an enhancer at the other one. Enhancers are defined as segments of DNA occupied by multiple transcription factors that recruit co-activators and RNAPII to target genes, which are generally located far away from the gene promoter (Hnisz et al., 2016a). Genes whose promoters are associated to a loop are higher expressed than those that do not associate, and cell type-specific loops are associated with changes in expression (Rao et al., 2014).

CTCF-CTCF loops have been called "insulated neighborhoods", defined as chromatin loops formed by a CTCF-CTCF homodimer, co-bound with cohesin, and containing at least one gene. The median of an insulated neighborhood is ~190 Kb and contains three genes (Figure 16B). The majority of enhancer-gene interactions occur within these loops, which are necessary for normal gene activation and repression (Figure 16C). Perturbation of their loop anchors (i.e. deletion of CTCF binding sites) leads to local gene dysregulation (Figure 16D). Insulated neighborhood boundaries serve either to constrain the activity of enhancers, or maintain repression of genes within the neighborhood (Figure 16E).

Above, a detailed view about CTCF-CTCF mediated loops is presented. However, as mentioned before, not all the loops involve a CTCF dimer (two-thirds of loops contain a single CTCF-binding motif), and some loops are detected without the presence of CTCF (Rao et al., 2014). This suggest that other mechanisms of loop formation exist and may be involved in different functions than those explained by the insulator-mediated looping. For instance, intrachromosomal looping may be required for: (i) efficient recycling of RNAPII after transcription termination (Figure 17A); (ii) bringing distant enhancers in contact with promoters without CTCF dimerization (Figure 17B); (iii) polycomb-dependent repression (Figure 17C) (Cavalli and Misteli, 2013).

3.3.2.3 Gene-gene interactions

After this description of genomic compartments/regions from the highest order of organization (CTs) to the smallest one (chromatin loops), some well-described examples of gene-gene interactions will be presented in this section (see (Hou and Corces, 2012) for review).

The β-globin locus has been probably one of the most studied example of association between genes and regulatory sequences. 3D DNA and RNA Fluorescence In Situ Hybridization (FISH) combined with 3C assays revealed that distal genomic regions co-localize in mouse erythroid nuclei (Osborne et al., 2004; Tolhuis et al., 2002). Concretely, the locus control region (LCR), located 40-60 Kb

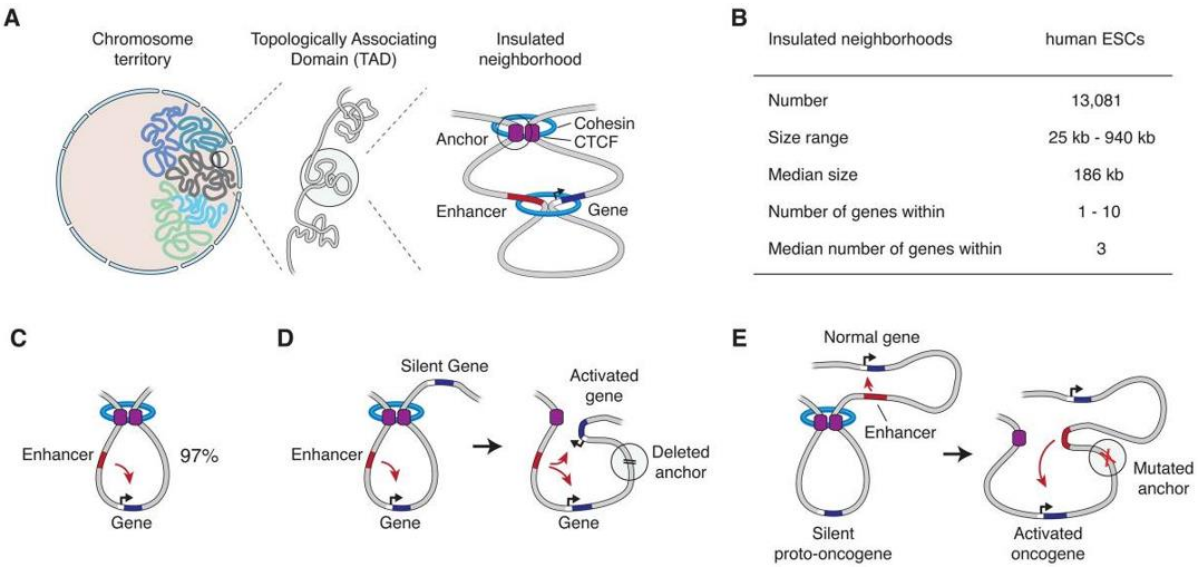


Figure 16. Insulated neighborhood functions. *(A) Hierarchy of chromosome structures: CTs, TADs, and insulated neighborhoods. Loop anchor establishes by CTCF dimerization and cohesion binding. (B) Features of insulated neighborhoods in human embryonic stem cells (ESCs). (C) 90% of enhancer-gene interactions occur within insulated neighborhoods in human ESCs. (D) Deletion of insulated neighborhood anchors leads to gene misregulation. (E) Mutations of insulated neighborhood anchors in tumor cells lead to oncogene activation* (Hnisz et al., 2016a).

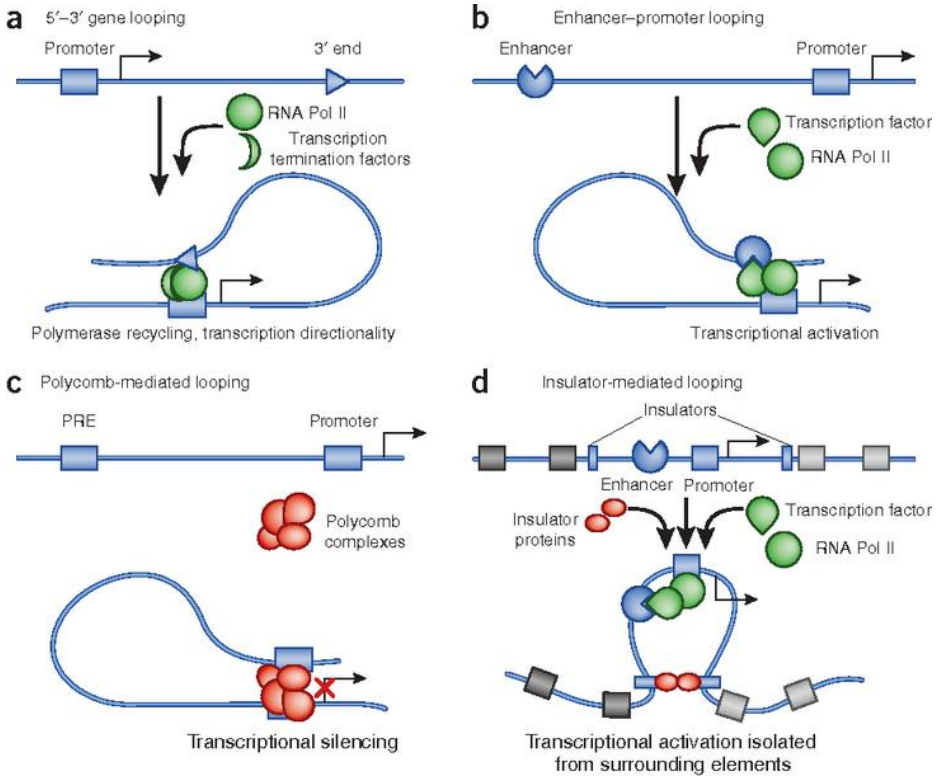


Figure 17. Transcription regulatory chromatin loops. *(A) Intragenic loops joining the 5' and 3' end of genes may allow recycling of RNAPII. (B) Enhancer-promoter loops mediated by sequence-specific transcription factors (possibly assisted by ncRNAs, or CTCF and cohesin). (C) Loops between polycomb-bound regions and promoters, prevent RNAPII recruitment. (D) Insulator-mediated loops, mediated by insulator proteins such as CTCF* (Cavalli and Misteli, 2013).

away from the active genes, come in close spatial proximity with these globin genes. The nuclear co-localization was observed not only between erythroid-specific genes but between highly transcribed genes that co-localize in the same transcription factory (TF) at high frequencies, and movement into or out of these factories results in activation or abatement of transcription (Osborne et al., 2004).

Another example in mouse olfactory neurons, is the nuclear aggregation of silent olfactory receptor (OR) genes from different chromosomes (Clowney et al., 2012). In this study, a spatial segregation between active and silent OR alleles was observed by combining DNA and RNA FISH with immunofluorescence against heterochromatic marks and transcriptional active marks. Inactive alleles were found associated with heterochromatic foci, while active alleles were found associated with Pol II in euchromatic territories. This phenomenon might explain the monoallelic nature of OR expression.

Allele specific regulation of imprinted genes through long-range chromosomal interactions has been studied in detail in the human and mouse Igf2/H19 loci (see (Hou and Corces, 2012) for review). In this locus, *Igf2* is expressed from the paternal allele and *H19* from the maternal allele. The imprinting control region (ICR) which is methylated on the paternal but not on the maternal alleles, is responsible for the regulation of this allele-specific expression (Macdonald, 2012). This ICR contains CTCF-binding sites and, when the ICRs are non-methylated, CTCF can bind to these sites. 3C assays showed that the CTCF-loop structure formed in the maternal allele prevents the accessibility of the enhancer to *Igf2* (Hou and Corces, 2012), avoiding *Igf2* expression. Additionally, *trans*-interactions between the *H19* ICR and other imprinted loci have been observed using 4C assays (Zhao et al., 2006). In this study, perturbations of the CTCF recognition site or CTCF binding lead to a loss of interactions and miss-regulation of imprinted genes.

Co-expressed genes have been found co-localized in TFs, this is the case of erythroid-specific and highly transcribed genes, mentioned at the beginning of this section (Osborne et al., 2004), but this is not the only example. For instance, it was observed by 3D FISH that upon differentiation of human multipotent stem cells, co-expressed genes associated with either the same splicing speckle or with the same TF (Rieder et al., 2014). In another study which combined 3D RNA FISH, Immuno-FISH and 4C-derived assays, it was observed that mouse globin genes interact with many other transcribed genes, and Klf1-regulated genes preferentially co-associate with specialized TFs enriched with the transcription factor Klf1 (Schoenfelder et al., 2010). This idea of gene associations driven by specialized factors has been also formulated in a study performed in estrogen-treated human breast adenocarcinoma cells (MCF-7) (Fullwood et al., 2009). In this study, the ChIA-PET assay was used to determine interacting chromatin regions associated with the estrogen receptor alpha (ER-α), and it was proposed that ER-α form chromatin looping structures around target genes for coordinated transcriptional regulation of these genes.

Another example of co-expressed genes interacting in the nucleus was observed in Human Umbilical Vein Endothelial Cells (HUVECs) when upon TNFα (a major proinflamatory cytokine) stimulation, TNFα-induced genes were hierarchically transcribed when engaged also hierarchically in chromosomal interactions (Fanucchi et al., 2013). This is an elegant illustration of the dynamic aspect of genome organization, which will be discussed in the following section.

All the studies presented here, are examples of long-range interactions between active genes. However, there are some cases where silent genes form repressive interactions such as polycomb-repressed Hox genes in *Drosophila* and mammalian cells (see (Hou and Corces, 2012) for review). For instance, it was observed that Hox genes only co-localize in polycomb bodies in tissues where these genes are repressed (Bantignies and Cavalli, 2011).

3.3.3 Dynamic organization of the genome

The position of genes in the nucleus is not fixed, for instance genes can move in and out of TFs, resulting in activation or abetment of their transcription (Osborne et al., 2004). The ability of chromosome large domains (such as A and B compartments) to move in a given cell is limited because of their several megabases in size. At scales of several hundreds of kilobases, chromatin is considerably more dynamic. The mobility and movements of gene loci have been studied by live cell imaging, showing that loci have a constrained radius of diffusion of ~0.5 μm. This volume corresponds to the TADs scale (1 Mb), suggesting that interactions between any two loci located within a TAD are sufficiently dynamic to have an opportunity to engage in long-range interactions (Gibcus and Dekker, 2013). Imaging of relative positions of individual genes or subnuclear compartments by 3D FISH in fixed cells, has shown that locations can change at different stages of gene activation and/or cell differentiation (Schneider and Grosschedl, 2007). Nevertheless, other studies show that genes can change expression without altering nuclear location (Hakim et al., 2009; Kocanova et al., 2010).

Sometimes changes in chromatin location can happen at great scale. Striking changes in chromosome positioning are rare, but have been reported to occur within minutes (Pombo and Dillon, 2015). Another example of chromatin dynamics is observed in LADs, which can be classified in constitutive (cLADs) and facultative (fLADs) LADs. It was observed in mouse that cLADs are maintained across a wide range of cell types and across species (between human and mouse), contrary to fLADs which are rather cell-type specific (Meuleman et al., 2013). Moreover, *in vivo* analyses in a human fibrosarcoma cell line show that some LADs relocate to the periphery of the nucleolus after mitosis (Kind et al., 2013). TAD reorganization has also been observed in the regulation of the *Hox* gene clusters, one of the best characterized PcG domains. *Hox* genes form large H3K27me3-marked (inactive) TADs, located within an A compartment. In mouse ES cells, when transcription is activated, specific *Hox* genes progressively segregate into an active TAD and this process is accompanied by a switch in histone modifications (Aranda et al., 2015).

Little is known about how DNA moves or is relocated in some of these examples of genome dynamics. One possibility is that active polymerase can function as a motor that pulls in its template. Alternatively, other molecular motors such as actin and myosin could be involved in the relocalization of the DNA template (Schneider and Grosschedl, 2007). Moreover, it has been proposed that the nucleoskeleton (a dense, filamentous structure containing many proteins: lamins, titin, actin, myosins, DNA binding proteins and the general transcription machinery) may direct the traction of genes to nuclear bodies such as TFs (Mercer and Mattick, 2013). Recent studies argue in favor to this hypothesis by suggesting that an actin-based nucleoskeleton would be involved in gene regulation and genome

organization (Xie and Percipalle, 2017). For instance, some authors suggested that nuclear actin is required for rapid long-range movement of U2 genes towards Cajal bodies in HeLa cells (Dundr et al., 2007). Other authors observed that after transcriptional activation, the migration of a chromosomal locus from nuclear periphery to the interior was perturbed in actin or myosin mutants in a rat cell line (Chuang et al., 2006). More recently, it was observed in budding yeast that both cytoskeletal and nuclear actin drive local chromosomal movements, such as telomeres dynamics (Spichal et al., 2016). In addition, chromatin modifications such as post-transcriptional histone modifications could directly affect the structure of the chromatin and may also have a role in the positioning of chromosomes (Schneider and Grosschedl, 2007). Possible mechanisms explaining the role of post-transcriptional modified histones in chromatin (during cell differentiation in mammals) or telomeres (in yeast) anchoring to the nuclear periphery, have been recently reviewed (Harr et al., 2016).

3.3.4 Single cell genome organization

Due to this capacity of individual loci to diffuse in the nuclear space, is quite difficult that two cells exhibit exactly the same genome organization at the same time. Most of the recent studies performed to uncover the global organization of the genome have been performed on cell populations. Therefore, the average interaction maps generated using population-based methods are an ensemble of many different genome landscapes (Cavalli and Misteli, 2013) and do not show the reality of what is happening in an individual cell. Single-cell approaches have permitted to fill this gap (Nagano et al., 2013; Stevens et al., 2017). For instance, Nagano et al. showed that it exists cell-to-cell variability in chromosome structure, and that intradomain contacts are more robust (generally conserved) at the single-cell level that interdomain contacts, which are highly variable between individual cells. They also observed that some domains are more likely to present *trans*-chromosomal contacts at the surface of CTs than others (Nagano et al., 2013). A recent study highlighted that the structure of TADs and loops vary substantially from cell-to-cell. However, A/B compartments, LADs, active enhancers and promoters are quite stable among all cells in a population. This suggests that the last could drive chromosome and genome folding (Stevens et al., 2017). Thus, data coming from cell population-based methods, need to be interpreted carefully and if possible combined with single-cell data. All chromatin contacts occurring in a cell population, cannot be present simultaneously in an individual cell, cell-to-cell variability and other physical constrains will prevent this to happen.

3.3.5 3D genome architecture and disease

The structural integrity of the 3D genome topology is crucial for the correct functioning of an organism. In the normal human population, approximately 5% of the genome is structurally variable, including deletions, duplications (copy number variants, CNVs), inversions, and translocations (Lupiáñez et al., 2015). Chromosomal rearrangements (CRs), more concretely the breakpoints, occur in evolutionary “fragile” genomic regions characterized by the presence of high chromatin contacts (Berthelot et al., 2015). Balanced rearrangements such as inversions, or CNVs limited to non-coding DNA, have the potential to disrupt the integrity of the genome, leading to alteration of gene expression

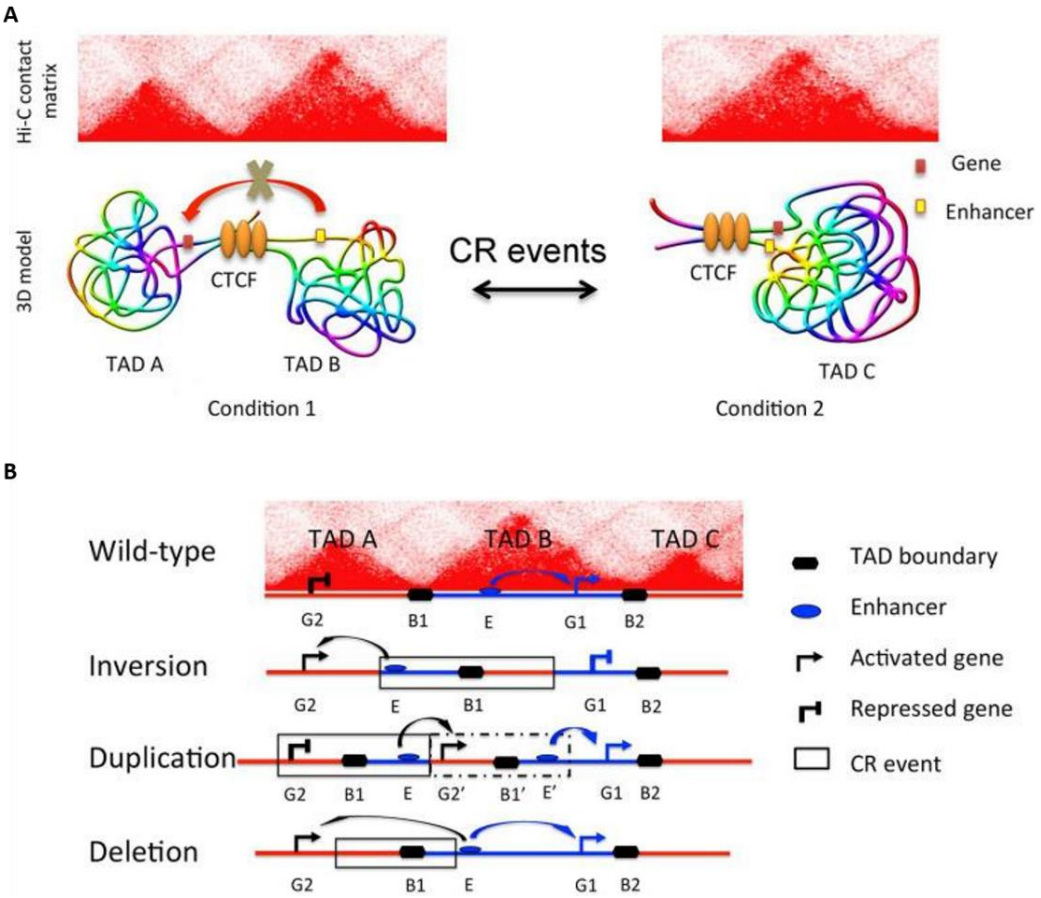


Figure 18. Chromosomal rearrangement (CR) events affect TAD structures. *(A) Model of chromatin TAD variation caused by CR. Every red triangle represents one TAD. The loss of a TAD boundary due to a CR event may cause a phenomenon of TAD fusion, bringing in proximity a promoter and an enhancer initially separated by the TAD boundary. (B) Model for pathogenicity of CR events that alter gene expression through 3D chromosome structure* (Li et al., 2016).

levels and patterns. Only a fraction of enhancers contact the nearest promoter, whereas most skip one or more genes (de Laat and Duboule, 2013), and these contacts are restricted by the TADs structure. When CR events affect the structure of TADs, more concretely the TAD boundaries, they alter the interactions between enhancers and promoters, leading to abnormal expression of genes (Figure 18) (Li et al., 2016). CR associated disease, such as developmental diseases and cancer can potentially be caused by chromosomal 3D structure alterations when a TAD boundary is deleted or a novo TAD boundary is created. A recent study shows that CR events cause polydactyly diseases through altering CTCF-associated TAD boundary domains (Lupiáñez et al., 2015). Global and more specific alterations in the 3D genome organization have been described in cancer. For instance, higher order chromosomal changes were detected between breast cancer cells and a normal epithelial cell line (Barutcu et al., 2015). In this study, a decrease on the interaction frequencies in breast cancer cells was observed in small gene-rich chromosomes, associated with a higher occurrence of open compartments of these chromosomes. Moreover, telomeric and subtelomeric regions displayed more frequent intra-chromosomal interactions in epithelial cells than in cancer cells. Another study in breast cancer cells, allowed detecting differentially interacting loci enriched for cancer proliferation and estrogen-related genes after hormone stimulation. These loci were correlated with higher estrogen receptor α-binding events and gene expression, suggesting a role of estrogen hormone on genome reorganization (Mourad et al., 2014). Besides these global changes, more precise disruptions in genome topology may explain pathological processes. For instance, a disruption in the insulated neighborhood structures may be also involved in cancer processes. An aberrant activation of proto-oncogenes by enhancers, normally located outside the neighborhoods, might be due to a loss of an insulated neighborhood boundary (Figure 16E). This was observed in acute lymphoblastic leukemia, where T-cells contain recurrent microdelections that eliminate boundary sites of insulted neighborhoods containing proto-oncogenes (Hnisz et al., 2016b).

3.3.6 3D genome architecture approaches

Initially, genome organization has been studied by microscopy, particularly with fluorescence *in* situ hybridization (FISH). This approach has permitted to obtain valuable information about the nuclear organization although uncompleted. These assays allow analyzing specific aspects of genome folding, but do not permit to uncover global aspects of chromatin structure and genome topology. Since 2002, the development of Chromosome Conformation Capture (3C)-based technologies, has led to a kind of revolution in the domain of genome topology. The reason is that these technologies have the potential to quantify almost all frequency contacts between distal DNA segments in a cell population.

3.3.6.1 Population-based methods (3C, 4C, 5C, Capture-C, Hi-C, ChIA-PET)

All 3C-derived methods are based on the same principle. First, chromatin is fixed, often using formaldehyde agent to create covalent bounds between DNA fragments bridged by proteins. Second, DNA is digested by using a restriction enzyme. Third, sticky ends are religated under diluted conditions (to promote intramolecular ligations between cross-linked fragments), which creates “hybrid”

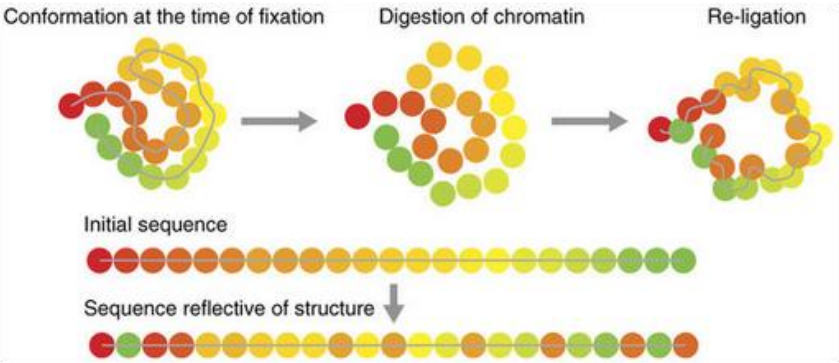


Figure 19. Common principle in 3C-based techniques. *The chromatin fiber is initially digested into short restriction fragments (represented by beads), after which a ligation reaction is performed to create large DNA concatemers in which the order of the fragments reflects the three-dimensional structure of the chromatin at the time of fixation.* (Davies et al., 2017).

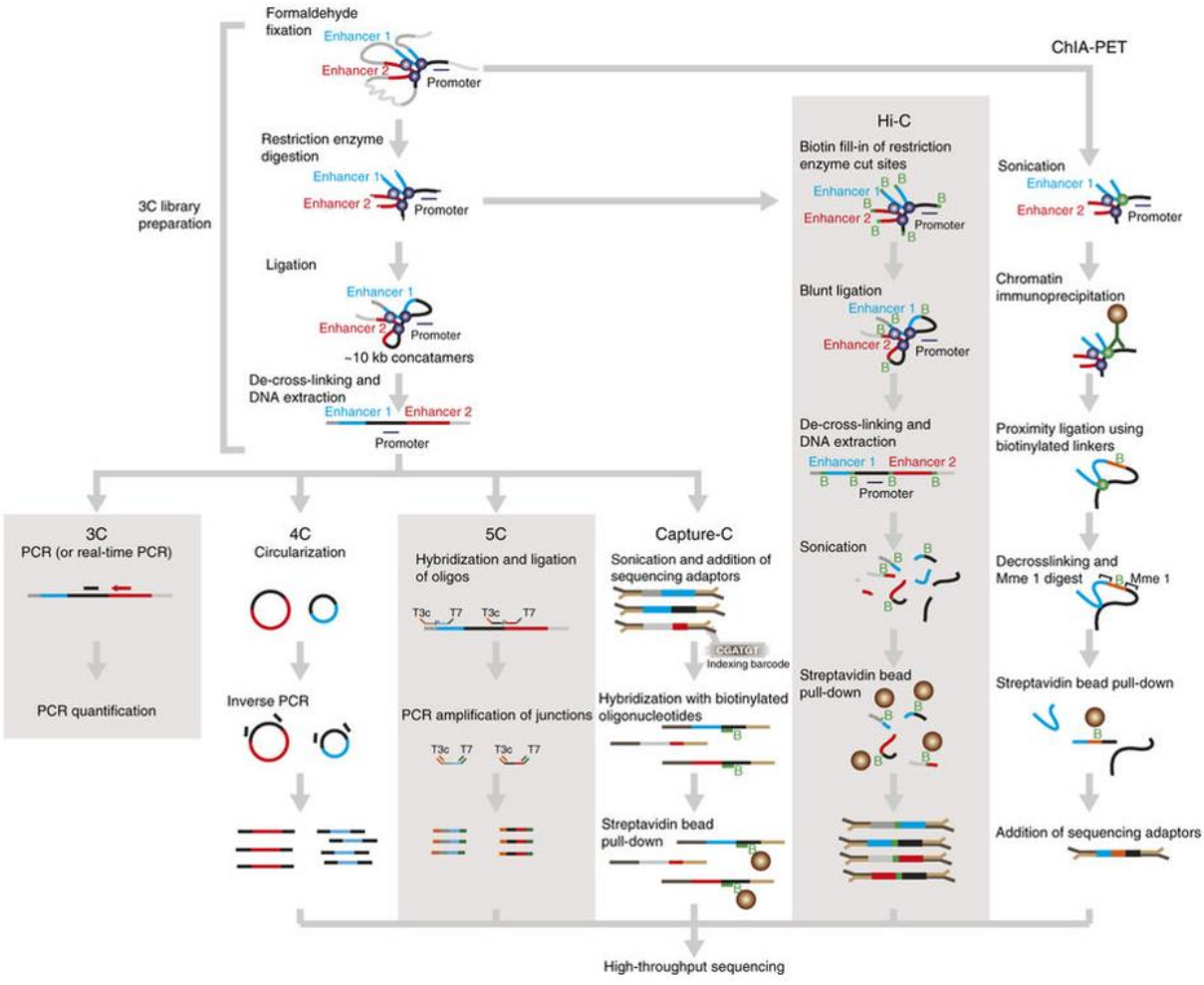


Figure 20. Overview of the different 3C-based technologies. *Cross-linking, digestion and ligation steps are common to all 3C-based methods* (Davies et al., 2017).

fragments. Only ~5% of the restriction fragments in a 3C library ligate back to their original partner (Figure 19) (Davies et al., 2017). The final step consists in reversion of cross-linking, and DNA extraction containing the hybrid fragments (Figure 20). The idea is that DNA fragments being far away on the linear sequence of the genome, but co-localizing in space, can be ligated to each other (Figure 19). In the initial method (termed 3C), Dekker et al. performed PCR amplification using primers designed near and towards ligation junctions, followed by gel electrophoresis (Dekker et al., 2002) (Figure 20). Nowadays, to quantify interaction frequencies, electrophoresis has been replaced by real-time PCR. 3C method allows uniquely detecting contacts between small numbers of fixed restriction fragments for suspected interactions, rather than identifying new interactions. Actually, 3C is known as a "one vs. one" strategy, because it serves to obtain pairwise interaction frequencies between a known pair of loci (de Wit and de Laat, 2012).

An evolution of the 3C method was the development of circular chromosome conformation capture (4C), also termed as a "one vs. all" approach (Zhao et al., 2006). This method allows identifying all potential partners for any specific loci of interest in the genome, through an additional step of circularization after decross-linking of 3C fragments (Figure 20). Small circularized fragments are generally generated with a second digestion by using a different restriction enzyme. Then, an inverse PCR is performed by using primers designed close to the first restriction enzyme site of the target locus, and oriented towards the unknown sequence to amplify any interacting partners. Initially, microarray was used to identify interacting partners, but this has now been replaced by high-throughput sequencing. An improvement of the 4C method is the recently developed protocol described as Unique Molecular Identifier (UMI)-4C (Schwartzman et al., 2016). Without circularization step, 3C fragments are sonicated and sequencing adapters ligated uniquely to one end of each 3C sonicated fragment. Hybrid fragments are then amplified by using a universal adaptor primer and a primer in the target sequence. UMI-4C also allows multiplexing by using different sequencing adapters.

The first jump from interrogating only one target sequence to many at a time (genome-scale assays), came with the development of the chromosome conformation capture carbon copy (5C) technology (Dostie et al., 2006). 5C can be described as a "many vs. many" method, because it allows the simultaneous detection of millions of interactions by using a mix of primers in a single assay. This approach allows to define functional contacts for all the genes in a locus simultaneously. 5C primer pairs anneal to either site of all the ligation junctions in the sequence of interest. Forward and reverse primers contain universal tails (usually T7 and T3 respectively) joined at the ends of 5C primers, a middle specific sequence complementary to the target locus, followed by half of the restriction enzyme site (Figure 20). Reverse primers have in addition a phosphate group at the 5' end of the half restriction enzyme site. Only primers annealed next to each other at the ligation junction can be ligated with Taq ligase by means of the phosphate group. All ligated 5C primers are simultaneously amplified using a pair of universal primers that anneal on the T7 and T3 sequences. In brief, this method allows detecting contact events concerning different regions within a particular locus.

A more recent 3C-based technology is the Capture-C, which can generate genome-wide interaction profiles from hundreds of viewpoints in a single assay. This method was developed to

analyze hundreds of *cis*-regulatory landscapes (Hughes et al., 2014). In Capture-C, 3C fragments are sonicated, and sequencing adapters with indexed barcodes are added (Figure 20). This unique random ends allow identifying and removing PCR duplicates in a subsequent data analysis. The library is then enriched for fragments of interest using biotinylated probes designed for each viewpoint, before amplification and sequencing. An improvement of this method is the so-called next-generation (NG) Capture-C (Davies et al., 2016), which uses a different method of probes design that increases the efficiency of the oligonucleotide capture process.

The method that has permitted to interrogate interaction frequencies between all parts of the genome is the High throughput chromosome conformation capture (Hi-C), which is referred as an "all vs. all" method (Lieberman-Aiden et al., 2009). In Hi-C, the protocol for creating the 3C template is slightly modified (Figure 20). Before ligation, the digested ends are filled-in with a biotynilated nucleotide, followed by a blunt-end ligation. Then DNA is purified and sheared. Ligation junctions are enriched by a streptavidin bead pull-down, and sequenced by high-throughput sequencing. Hi-C has been found very appropriate for determining megabase-scale contacts and large-scale chromatin structure, such as A and B compartments and TADs. An improvement of this method is the so-called *in situ* or in-nucleus Hi-C. In this protocol the ligation step is performed within preserved nuclei, instead of performing in-solution diluted cross-linked chromatin ligation (Nagano et al., 2013). This allows capturing chromatin interactions more consistently and reducing experimental noise and bias, compared to the in-solution method (Nagano et al., 2015). A variant of the Hi-C method is the capture Hi-C, which combines capture-C and Hi-C libraries to enrich in target sequences and to exclude further uninformative background. In addition to determining interaction frequencies and chromatin structure, Hi-C has been found useful for other applications, such as in *de novo* assembly (Burton et al., 2013), and in metagenome analysis (Marbouty and Koszul, 2015).

Another "all vs. all" method is the chromatin interaction analysis by paired-end tag (ChIA-PET) sequencing (Fullwood et al., 2009). This method combines chromatin immunoprecipitation (ChIP) with 3C, an offers the possibility to analyze all chromatin interactions between sites bound by a specific protein. After cross-linking and sonication, ligation junctions between DNA sites are pulled-down with an antibody against the protein of interest. Then, DNA sequences tethered together and to the target protein are connected through proximity ligation with DNA linkers. These linkers are biotynilated and contain MmeI restriction sites. MmeI is able to recognize these sites and cut DNA few bases away of the restriction site, allowing short fragments to be extracted with a streptavidin bead pull-down, and then identified by paired end (PE) sequencing.

A little mention about the ChIP-seq (chromatin immunoprecipitation sequencing) approach will be done in this section. Even though this is not an approach to study DNA-DNA interactions, ChIP-seq has allowed to uncover important aspects of genome organization and function. ChIP-seq is based on cross-linking of DNA-protein interactions, and enrichment of DNA sequences associated with the target protein, by using specific antibodies, followed by DNA sequencing. It was first developed to study genome sequences associated to histone modifications (Barski et al., 2007). Later, it was assessed to study regions associated with structural proteins such as CTCF and cohesin (Rao et al., 2014), and other

DNA-binding proteins. The integration of ChIP-seq and Hi-C data has permitted to study mechanisms of loop formation and TAD structure, and to define active and repressed chromatin domains.

3.3.6.1.1 Hi-C resolution

An important parameter that will determine the scale level of study of the 3D genome organization is the resolution. For instance, in a Hi-C approach which has the potential to capture all genomic regions in proximity, high resolutions will permit to detect "small" chromatin structures such as loops. However, lower resolutions for the same Hi-C experiment will not allow to identify such structures, but could permit identifying larger domains such as A and B compartments or TADs. The resolution depends on many factors, the most important concerns the choice of the restriction enzyme and the sequencing depth. The most limiting, the restriction enzyme, because contacts between DNA sequences can be detected only at restriction enzyme cut sites (Davies et al., 2017; Han et al., 2018). This means that contacts within two restriction enzyme sites will not be identified. For instance a four-cutter enzyme will produce smaller fragments than a six-cutter enzyme (256 bp vs. 4096 bp), increasing in 16-fold the resolution of the library. This is a global approximation considering that resolution will not be the same all over the genome, because the distribution of restriction sites is not uniform. Therefore, an increase of resolution can be achieved by substituting restriction enzymes by other enzymes. This is the case of Hi-C variants using DNase and MNaseI (termed micro-C), which can cut at any site along the genome and have the potential to generate single base pair resolution. Theoretically, if 1 bp fragment size is achieved, the resolution is no longer dependent on the restriction enzyme but is determined by the sequencing depth. However, the sequencing depth is intimately linked to the restriction enzyme choice and to the genome size. Indeed, the sequencing requirements of the 3C libraries are related to the square of the number of restriction fragments in the genome (Davies et al., 2017). Moreover, because of the quadratic nature of "all vs. all" data, an increase in resolution by 10-fold requires a 100-fold increase in depth (de Wit and de Laat, 2012). For instance, the 1 Mb resolution achieved by performing Hi-C experiments in mammals, was based on 10 million PE reads (Lieberman-Aiden et al., 2009). In this case, an increase in resolution from 1 Mb to 100 Kb, would need 1 billion PE reads instead of 10 million. Likewise, generating contact profiles with a resolution from 40 Kb to 1 Kb in the human genome, requires from hundreds of millions to multiple billion PE reads. Therefore, the cost and computational resources are far too expensive for most laboratories (Han et al., 2018).

3C library complexity is another critical factor for resolution, and it is mainly affected by the initial number of cells used, the digestion and ligation efficiency and the cumulative loss of material from each step before sequencing (Davies et al., 2017). When library complexity and/or sequencing are insufficient to explore contacts at the level of individual restriction fragments, the resolution will be determined by an appropriate bin size. Binning improves the signal strength and reduce biases, the inconvenient is that: (i) the profile becomes skewed by density of restriction enzyme sites, (ii) the original signal is smoothed, hiding the quality of underlying data, (iii) the resolution decrease (Davies et al., 2017).

3.3.6.1.2 Limits and biases of 3C-based methods

Regarding the limits and biases of the 3C-based methods, 3C is limited to the detection of spatial relationships between known DNA sequences and it can detect contacts only in a limited range (not exceeding a few hundred of kilobases) (Han et al., 2018). 4C allows very long-range contacts to be detected, however, the amplification of GC-rich fragments by inverse PCR is inefficient, resulting in biases in the interaction profile. In addition, it is not possible to differentiate between PCR duplicates and unique ligation junctions (Davies et al., 2017). In contrast, UMI-4C allows removing PCR duplicates during data analysis (Han et al., 2018). In 5C, the resolution is determined by the spacing between neighboring probes on the linear chromosome template (de Wit and de Laat, 2012). It can never reach the resolution of 4C, Hi-C, and Capture-C, as not every end of each restriction fragment will allow the design of a 5C probe. It can also miss weak, long-range contacts, which are detectable by 4C, Hi-C and Capture-C. Moreover, differences in the hybridization efficiency of the probes can cause bias, and it is only possible to determine contacts between forward and reverse probes. As in 4C, the levels of PCR duplication cannot be determined. Although Capture-C can be used to detect hundreds of informative interactions, individual interactions themselves do not have the depth of data of a good 4C experiment for the same region. NG Capture-C is then a better alternative in terms of sensivity and resolution, which allows weak *cis* long-range and *trans* interactions to be quantified (Davies et al., 2017; Han et al., 2018). Regarding Hi-C, the number of contacts determined from any individual restriction fragment is around 100-fold lower than in 4C and Capture-C, even in the recent Hi-C data sets at 1 Kb resolution (Rao et al., 2014). That makes Hi-C a relative insensitive method to determine fine-scale (<40 Kb) interactions between regulatory elements present within TADs (Davies et al., 2017). Even though Hi-C has relative low biases, it is still systematically affected by the distance between restriction sites, the G+C content and the presence of repetitive regions. But, several methods have been developed to attempt to correct these biases (Davies et al., 2017). Compared to the high levels of enrichment of NG Capture-C, the two-fold increase in resolution in capture Hi-C seems negligible and has been balanced against the more extended protocol (Hi-C) and extra loses in library complexity due to decreases on cell numbers in each step (Davies et al., 2017). Regarding the ChIA-PET method, a limitation is the low library complexity due to the relative low levels of enrichment of ChIP, which implies that the number of reads used to identify individual interactions is usually low (Davies et al., 2017).

In conclusion, the Hi-C method is unique in its ability to determine genome-wide interaction profiles, and to define whole genome large domains to globally determine basic rules of genome organization. However, to define the details of small-scale interactions that dictate regulation of individual genes, 4C or NG Capture-C are more appropriate because need less requirements in terms of sequencing depth than high resolution Hi-C experiments (Davies et al., 2017).

3.3.6.2 Single-cell methods (single-cell Hi-C, 3D DNA and RNA-FISH)

As mentioned before, the 3D genome organization is not static but dynamic, which makes almost impossible that two cells exhibit exactly the same genome organization. The techniques described above, give an overview of all possible chromatin interactions in a cell population. However,

Table 2. Comparison of super-resolution microscopy techniques (Sydor et al., 2015)**.**

	Structured Illumination Microscopy (SIM)	Stimulated Emission Depletion Microscopy (STED)	Photo-Activated Localization Microscopy (PALM) and Stochastic Optical Reconstruction Microscopy (STORM)
Principle	Uses interference-generated light patterns to create a Moire effect from which higher-resolution information can be extracted	Reduces the effective excitation volume with a depletion laser	Stochastically activates a subset of photoswitchable probes at a time, and then determines the centroid position of each point spread function
Microscopy type	Wide-field	Laser scanning confocal	Wide-field
xy Resolution	100–130 nm	20–70 nm	10–30 nm
Axial resolution	~300 nm	40–150 nm	10–75 nm
Probes	Common photostable organic dyes and fluorescent proteins	Particular photostable organic dyes and fluorescent proteins	PALM: photoswitchable fluorescent proteins STORM: photoswitchable organic dyes
Temporal resolution	Milliseconds to seconds	Milliseconds to seconds	Seconds to minutes
Photodamage	Low to moderate	Moderate to high	Moderate
Photobleaching	Moderate to high	Moderate to high	High for single fluorophores, low overall
Live imaging?	Yes	Yes	Yes
Post-image processing required?	Yes	No	Yes
Maximum number of simultaneous colors	4	3	PALM: 2 STORM: 3
Considerations	Straightforward multicolor experiments and sample preparation. Reconstruction algorithm may cause artifacts	Best temporal resolution at the highest spatial resolution; however maximal in-plane can be at the expense of axial resolution	Highest spatial resolution; however sensitive to labeling density. Crosstalk between fluorophores maybe an issue

a single cell will not be able to present all of them in a given moment due to physical constrains. To study this aspect, a variant of the *in situ* Hi-C method termed "single-cell Hi-C" has been developed (Nagano et al., 2013). Basically the protocol is the same, except that before reverse the crosslinks and purify the biotinilated Hi-C ligation junctions, individual nuclei are selected under the microscope and placed into individual tubes, which allows creating single-cell libraries. Although this approach gives an idea of the global chromatin structure in an individual cell, several single-cell Hi-C experiments need to be performed to identify conserved and variable regions among cells.

Before the appearance of the 3C-based methods, classical genome architecture studies were mainly performed by 3D Fluorescence In Situ Hybridization (FISH). This method allows labelling specific loci (DNA FISH), whole chromosomes (chromosome painting), nascent RNAs (RNA FISH), and protein complexes such as transcription factories (immuno-FISH). Performing these approaches in 3D-preserved interphase nuclei has permitted to uncover important aspects of the 3D nuclear organization (Bickmore, 2013; Chaumeil et al., 2013). FISH experiments can be done by using direct or indirect detection. For direct detection, probes are labelled by incorporation of a nucleotide associated to a fluorophore. For indirect detection, probes are labelled with biotin, and revealed with avidin/streptavidin associated to fluorophore. Chromosomes can be labelled by using paint probes, a mix of several probes that cover a large portion of the whole chromosome. DNA probes are generally constructed with bacterial artificial chromosomes (BACs) containing the gene of interest, and then hybridize on the corresponding gene sequence after DNA denaturation. RNA probes are constructed with PCR products from the amplification of the target gene, or synthetically (RNA FISH oligo probes, 40-50 nucleotides), and hybridization occurs without DNA denaturation. Protein complexes labelled by immuno-FISH use generally primary antibodies that recognize the protein of interest, and are revealed with secondary antibodies labelled with a fluorophore. For 3D FISH, the nuclear integrity must be preserved during all the process, including a fixation step, and soft conditions of permeabilization. All these methods permit multiple labelling to visualize several loci in a single experiment. Nuclei are generally analyzed by confocal microscopy, and 3D distances between loci, RNAs, or protein complexes, can be measured by using specific software. However, the smallest measured distance will depend on the resolution, which is determined by microscopy characteristics and parameters. Classical confocal microscopy has a relative low resolution (200 nm at the best in the *x* and *y* axes), which means that two loci located at less than 200 nm cannot be separated. Other technologies have been developed since 1994, termed super-resolution fluorescence microscopy, or nanoscopy, achieving up to 10-fold improvement in resolution (Table 2). These instruments allow obtaining up to 20-40 nm resolution in the focal plane (*x* and *y*), and 50-80 nm in the depth direction (*z* plane). More recently, a new technology offering ultra-high resolution have been developed, achieving up to 10-20 nm of isotropic resolution (*x*, *y* and *z*) (Huang et al., 2016). These new technologies have been widely used for imaging proteinaceous nanostructures such as bacteriophages, PLM bodies, nuclear pore complex or centriole. However, imaging sequence-specific chromatin loci remains more challenging even though some progress have been done (Cremer et al., 2017; Sieben et al., 2018; Sydor et al., 2015).

Transcription activator-like effectors (TALEs) conjugated with fluorescent proteins, or clustered regularly interspersed short palindromic repeat (CRISPR)/CRISPR-associated protein 9

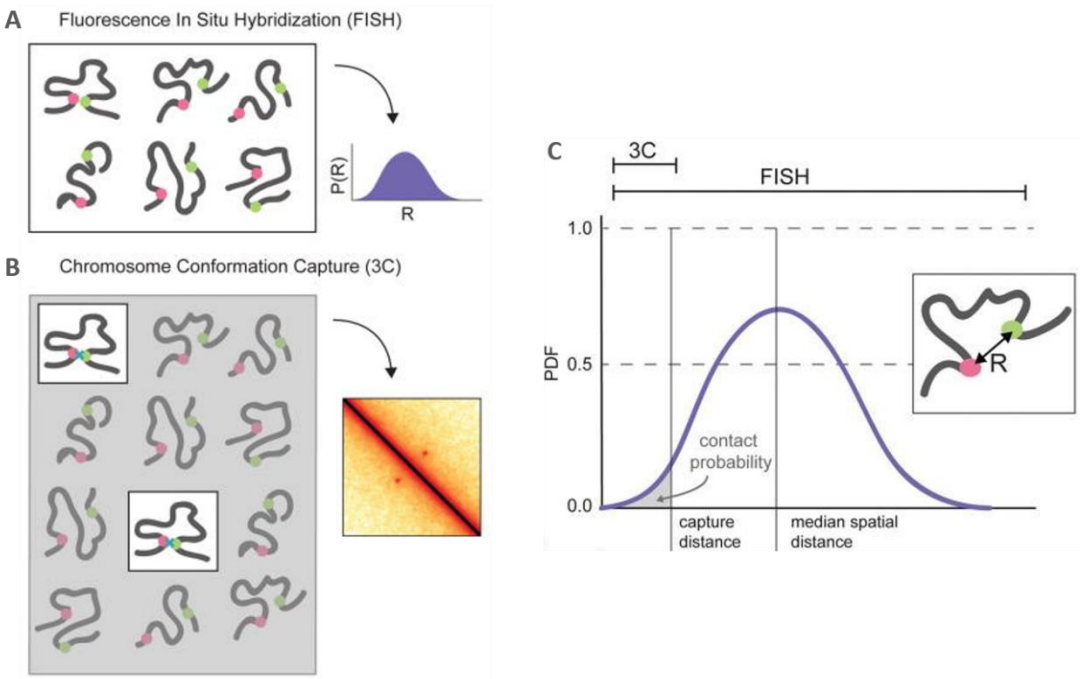


Figure 21. Illustrated relationship between 3C and FISH. *(A) FISH obtains information for all cells in a population to build up a full distribution of pairwise distances between labeled loci. (B) 3C-based approaches (including 4C, 5C, Hi-C) capture contacts from the small fraction of cells where two loci are within the capture radius. (C) Illustration of a PDF pairwise spatial distance, R, between two loci for a large population of cells. Theoretically, FISH can measure the full pairwise spatial distance distribution. 3C captures contacts that occur at distances less than the capture distance* (Fudenberg and Imakaev, 2017).

(Cas9) system, have been used for fluorescence labelling of specific loci in live cells. Concretely, CRISPR/Cas9 system has permitted multicolor labeling and measuring of 3D distances between different loci (Ma et al., 2015; Qin et al., 2017). A version of the CRISPR/Cas9 system uses nuclease-inactive Cas9 (dCas9), to label three orthogonal Cas9 variants by fusion to green, red or blue fluorescent proteins (GFP, RFP and BFP respectively). Then, single-guide RNAs (sgRNAs) are designed to target the specific loci. These two constructions (fluorescent dCas9, and sgRNAs containing the target sequences) are each subcloned into a different plasmid vector. Then, both plasmids are co-transfected in the cells. The dCas9-fluorescent protein and the sgRNAs will be then expressed into the cells and will associate to form a complex able to target the loci of interest by fluorescence labelling.

3.3.6.3 Comparison between FISH and 3C-based methods

FISH and 3C-based methods, are both used to detect spatial chromatin interactions, and many 3C-based studies used FISH to validate some of the detected interactions. However, both techniques differ in many aspects and must be interpreted with caution. FISH are low-throughput assays, because they are restricted to viewpoints corresponding to regions targeted by specific probes. In contrast, 3C-based methods offer a genome-wide view of genome organization. FISH has limited spatial resolution but distances are measured directly in individual cells. 3C methods extrapolate physical proximity by considering that ligation frequency is inversely proportional to the real spatial separation between to loci. Moreover, they provide average frequencies across all cells in a population. Some studies have proposed an equivalence between direct 3D measured distances and interaction frequencies. For instance, Wang et al. observed that Hi-C contact frequency was inversely proportional to the fourth power of the mean spatial distance (Wang et al., 2016). Nevertheless, these kind of comparisons must be taken with caution. Indeed, it has been demonstrated that contact frequency is distinct to average spatial distance (Fudenberg and Imakaev, 2017), and that data coming from FISH and 3C-type experiments are not always concordant (Williamson et al., 2014). Comparing the simplest case of 3C and FISH, in which each method probes the relationship between a pair of loci, large shifts can be observed due to the nature of each approach. Measuring 3D spatial distances between a pair of loci in several cells, allows the measurement of the probability density function (PDF) between a pair of loci (Figure 21). While FISH makes possible measuring distances at any location of the two loci (except for the limits of microscopy resolution) (Figure 21A), 3C only capture contacts when the loci are closer than the contact radius. Such small “distances” detected by 3C correspond to very rare contacts in the cell population (Figure 21B-C). Thus, to compare 3C and FISH, it would be necessary imaging much more cells than conventionally done in FISH studies (around 100 nuclei or even less), and overcome the resolution limits of microscopy to obtain the full spatial distribution of two loci.

Other intrinsic factors of both methods could hinder the reconciliation of both approaches. In FISH, the probe size and the chromatin movement during denaturation and hybridization could affect the distances (Fudenberg and Imakaev, 2017). FISH may be no appropriate to capture weak or transient interactions. In 3C, formaldehyde preferentially crosslinks with lysine, tryptophan, and cysteine side chains in proteins, which could bias interaction frequencies (Williamson et al., 2014). It seems that FISH and 3C-based methods, such as Hi-C or 5C, may agree or may be more comparable when comparing large-scale domains such as CTs, A and B compartments or TADs. However at higher resolution, 3C

interaction frequency may not always simply reflect physical distances (Wang et al., 2016; Williamson et al., 2014). Finally, the dynamics of chromatin structure and cell-to-cell variation, is not appreciable by 3C-based methods. It cannot be determined whether interactions between multiple loci occur simultaneously or sequentially and/or whether they are mutually exclusive, which is possible by FISH (de Wit and de Laat, 2012). In conclusion, visual and molecular approaches should be complementary to each other and models of 3D genome organization should be extrapolated from data validated by independent methods.

3.3.7 3D Pig genome organization

3.3.7.1 Assessed by 3D DNA FISH

Few studies have been performed in pig regarding the nuclear architecture and genome organization. The majority of them are based on FISH assays to assess different aspects of nuclear bodies and gene-gene associations related to gene expression profiles. For instance, it was observed during in vitro adipogenesis that some up-regulated genes are relocated more internally, found on loops and projections of chromatin away from CTs, associated (often in clusters) to splicing speckles, and their CTs are decondensed (Szczerbal and Bridger, 2010; Szczerbal et al., 2009). However, another study on the same in vitro adipogenesis system showed that the relationship between transcription activity and gene positioning exists only for some genes but not all (Kociucka et al., 2012). A more recent study about nuclear substructures changes during differentiation of porcine mesenchymal stem cells (MSCs) into adipocytes has been performed by the same research group (Stachecka et al., 2018). After differentiation, they observed changes in nuclear size and shape (smaller and less spherical nuclei), as well as a preferential location in nuclear interior of nucleoli, and a clustering of telomeres. In differentiated cells, they also observed that chromocenters (a densely staining aggregation of heterochromatic regions) were more diffused than in MSCs, but no change in speckles and PML bodies' number were detected.

Other studies have been performed in neutrophils before and after activation by lipopolysaccharide (LPS) stimulation to mimic bacterial infection (Yerle-Bouissou et al., 2009). In both conditions, it was observed that centromeres associate to form chromocenters (preferentially between chromosomes with the same morphology), but after activation, some of these chromocenters disperse. Telomeres were observed to form clusters but no difference was observed upon LPS activation. They presented a more internal position than chromocenters, which were found significantly closer to nuclear border after activation. It was observed as well, that some chromosomes decondense upon LPS activation. Similar studies were performed in macrophages and neutrophils before and after LPS activation (Solinhac et al., 2011). In this study, it was observed that some up-regulated genes change their position with respect to CTs upon activation by increasing the distances to CTs edges, while down-regulated genes did not change their position.

3.3.7.2 Assessed by population-based methods

As mention in the second chapter, current efforts to improve functional annotation of livestock species have been made thanks to the creation of the FAANG Consortium (Andersson et al., 2015). The core assays of this Consortium are mainly focused on three aspects: identification of transcribed elements, study of the chromatin accessibility, histone modification marks, and genome organization. Although most of the studies are in progress, many FAANG contributors have already produced relevant data for these three aspects of the functional annotation (Tuggle et al., 2016). For instance, RNA-seq data have been generated in liver, muscle and T cells for chicken, cattle, pig and goat by the French National Institute of Agronomic Research (INRA), and by the University of California-Davis (UC Davis). ChIP-seq data for histone marks and CTCF protein have been also generated for chicken, cattle and pig by the UC Davis group.

Regarding genome organization, Hi-C assays have been performed on liver samples of pig (LW), chicken, and goat, from adult animals (two males and two females) by the INRA contributors to the FAANG Consortium (FR-AgENCODE project). Chromatin accessibility as well as transcriptome profiles have been also assessed by ATAC-seq and RNA-seq assays respectively on the same samples. These data have been integrated and are issue of a recent publication (Foissac et al., 2018). This study has permitted to: (1) extend the catalog of protein-coding and non-coding transcripts; (2) reveal differentially expressed transcripts with unknown function (including new lncRNAs in syntenic regions); (3) detect differentially accessible ATAC-seq peaks mapped to putative regulatory regions and enriched with predicted transcription factor binding sites; (4) show a consistency with results from gene expression (RNA-seq) and chromatin accessibility (ATAC-seq) in topological A and B compartments of the genome (Hi-C).

3.4 Chapter 4: Objective and strategy of this thesis

The main objective of this thesis has been to establish the relationship between genome organization and gene expression in muscle tissue during late fetal development.Two main approaches were developped for this purpose:

3.4.1 Combining 3D DNA FISH and gene expression for network inference

First, a single-cell approach was used to determine by 3D DNA FISH specific gene associations in the nuclear space for a small selected group of genes. Initially, we performed a study mainly focused on three target genes (*IGF2*, *DLK1* and *MEG3*). These genes correspond to two imprinted loci of particular interest in the agronomic context: *IGF2* for being a key element in fetal growth and development, involved in pig muscle growth and fat deposition (Nezer et al., 1999; St-Pierre et al., 2012; Van Laere et al., 2003) and *DLK1 for* being related to the control of muscle development and regeneration (Waddell et al., 2010).This preliminary study allowed us to detect by 3D DNA FISH *trans*-

interactions between these three genes in fetal liver and muscle tissues, and to reveal that these interactions involve the expressed alleles in muscle cells (Lahbib-Mansais et al., 2016). To extend this study, we further analyzed other nuclear associations between these three initial genes and four new genes (*MEST* and *DCN* imprinted genes, and *MYH3* and *RPL32* non-imprinted), *MYH3* being also a gene of major interest because of its important role in fetal muscle development (Schiaffino et al., 2015; Voillet et al., 2018).

Beyond these, a transcriptome study previously performed in our laboratory on muscle tissue, revealed differentially expressed genes (DEGs) associated with two fetal gestational ages (90 and 110 days of gestation) and four genotypes (Large White (LW), Meishan (MS), and the two reciprocal crossbreeds) (Voillet et al., 2014). The expression data and the information about DEGs observed in this study, together with the information of nuclear gene associations obtained by 3D DNA FISH, were combined to develop a new iterative method of gene co-expression network inference. This approach has enabled to obtain a robust gene co-expression network that spotlights significant biological processes related to foetal muscle development through the combination of spatial gene association and gene expression data. This study has recently been issue of a publication in the Scientific Reports journal (Marti-Marimon et al., 2018) and is further detailed after the “Materials and methods” section.

3.4.2 Global genome organization assessed by Hi-C and gene expression analysis

Second, after this initial single-cell approach focussed on a reduced number of genes associations analyzed on a few number of cells, a population-based approach was used to explore globally the changes occurring at the level of chromatin structure for a large cell population of muscle tissue during late fetal development. For that purpose we first assessed the 3D genome organization in pig fetal muscle at the 90th and the 110th day of gestation, by determining all interacting regions in the genome. To do this, Hi-C assays were performed by adapting the FAANG experimental protocol of Hi-C to fetal muscle tissue. As being a contributor of the FAANG Consortium, Hi-C data pipeline implementation was done in collaboration with the INRA research group in charge of Hi-C data production and analysis. Then, we further explored whether significant differences in the 3D genome organization exist between the two gestational ages. Finally, we combined the Hi-C and transcriptome data to investigate whether changes in genome organization are linked to changes in gene expression. This approach is further detailed on the “Global genome organization assessed by Hi-C and gene expression” section.

4 Materials and methods

4.1 Ethics statement

All tissues sampled for the 3D DNA FISH experiments were collected on pigs bred for the project (ANR-09-GENM-005-01, 2010–2015). The ethical committee of the Midi-Pyrénées Regional Council approved the experimental design (authorization MP/01/01/01/11). Tissues sampled for the Hi-C and ChIP-seq experiments were collected on pigs bred and financed by the AAP INRA of the Animal Genetics Division, 2014. The experimental design was approved and authorized by the ethical committee (No. 84) in animal experimentation of the French Ministry of National Education, Higher Education, and Scientific Research (authorization No. 02015021016014354).

For both samplings, the experiment authorization number for the experimental farm GenESI (Genetics, testing and innovative systems experimental unit) is A 17 661. All the fetuses used in this study were males and were obtained by caesarean after euthanasia of sows and fetuses. The procedures performed in this study and the treatment of animals complied with European Union legislation (Directive 2010/63/EU) and French legislation in the Midi-Pyrénées Region of France (Decree 2001-464).

4.2 Gene co-expression network approach

4.2.1 Transcriptome data

4.2.1.1 Microarray data description

Expression data were obtained from a previous transcriptome study of skeletal muscle in pig for two fetal gestational ages (90 and 110 days of gestation) associated with four fetal genotypes (two extreme breeds for mortality at birth –Large White (LW) and Meishan (MS)– and two reciprocal crosses –MSxLW and LWxMS). The final dataset consisted of 44,368 probes for 61 samples under eight different conditions (four genotypes at two gestational ages). A precise description of the experimental design and data collection can be found in (Voillet et al., 2014). Normalized expression data (log2 transformed) and sample information are available in NCBI (GEO accession number GSE56301).

4.2.1.2 Microarray data pre-processing

Missing values were imputed with k NN (R package “impute” function, with $k = 3$). Gene annotation was updated (nblast/NCBI July 2017, Sscrofa10.2) and the 40,847 annotated probes were found to correspond to 13,855 unique genes. For each gene, the probe with the highest average correlation with the other probes associated with the same gene, was selected to serve as a representative in further statistical analyses.

4.2.2 Network inference and analysis

4.2.2.1 Network inference

Networks were inferred using Gaussian Graphical Models (GGMs (Edwards, 1995)) from n = 61 samples. From expression data, GGMs build a graph (or network) in which vertices are genes and edges represent the conditional dependency structure between those genes. GGMs are based on the estimation of partial correlations (*i.e.*, correlations between two gene expressions when the expression of all the other genes is known). They were preferred over relevance networks (Butte and Kohane, 2000) because they improve measurements of direct relations between gene expressions by accounting for the effect of all expression data, and because they were found to be more efficient for grouping together genes with a common function in a previous study (Villa-Vialaneix et al., 2013).

Since the number of samples was smaller than the number of genes used for network inference, the models were fitted with a sparse penalty (Meinshausen and Bühlmann, 2006) to address the issues of high-dimensional data and edge selection. In addition, as many examples have shown that co-expressed genes occasionally tend to interact preferentially or consolidate in specialized foci of the nuclear environment (Osborne et al., 2004; Rieder et al., 2014; Schoenfelder et al., 2010; Zhao et al., 2006), when *a priori* information about nuclear gene co-localization is available, the latter was included in the model using the approach described in (Villa-Vialaneix et al., 2014). The details of the method and of the tuning of the different parameters are given in Appendix 1 “Description of the model used for network inference”.

4.2.2.2 Practical implementation of network inference

The starting point of the analysis was the inference of a network with no a priori information about co-localization. Since network inference based on partial correlation can only be performed with a limited number of genes (because of the number of samples) and since the number of unique genes (p = 13,855) was too high compared to the number of samples (n = 61), we applied two restrictions to the original list. First, we restricted the list to genes that were reported as differentially expressed (DEG) (Voillet et al., 2014). Secondly, among these DEGs, only those that had an absolute value for their correlation with either *IGF2*, *DLK1* or *MEG3* larger than 0.84 were kept. This final list contained 359 genes, provided in the Appendix 2 “Gene description and cluster allocation”.

4.2.2.3 Network inference interactions and 3D FISH validations

Based on network inference results or on genes found to be connected in the IGN of (Varrault et al., 2006), 3D DNA FISH experiments were performed to check whether pairs of genes of interest were co-localized in the 3D nuclear space. These experiments were conducted in an iterative manner with network inference. More precisely, network inference was performed with the following *a priori* conditions: (1) Network 0: was inferred with no *a priori* information, as a baseline for comparison; (2) Network 1: was inferred using a priori information from the triple association found in (Lahbib-Mansais et al., 2016) by giving the three pairs *IGF2-DLK1*, *IGF2-MEG3* and *DLK1-MEG3* as known co-localized genes. Network 1 was then used to propose candidate pairs of genes for testing by 3D DNA

FISH for Network 2 (*IGF2-RPL32*) and Network 3 (*DLK1-MYH3*); (3) Network 2: in addition to the initial three pairs, Network 2 was inferred using *a priori* information provided by the results of the new 3D DNA FISH experiments by giving the pairs *IGF2-MEST*, *DLK1-MEST*, *MEG3-MEST*, *MEG3-DCN*, *DLK1-DCN*, and *RPL32-IGF2* as known to be co-localized and *IGF2-DCN* as known not to be co-localized; (4) Network 3: in addition to the 10 previous pairs, Network 3 was inferred using *a priori* information provided by the results of new 3D DNA FISH experiments by giving the additional pairs *IGF2-MYH3*, *DLK1-MYH3, MEG3-MYH3* and *MEST-MYH3* as known co-localized genes.

All simulations were performed with the free statistical software R (https://cran.r-project.org). The inference was performed using our own scripts (available at https://github.com/tuxette/internet3D) and the graphs were displayed and analyzed using the R package igraph (Csárdi and Nepusz, 2006).

4.2.2.4 Network mining and clustering

Nodes of importance to the network structure were obtained by computing the degree and the betweenness centrality measurement for every node. Node clustering was performed by applying the Louvain algorithm (Blondel et al., 2008), which performs fast approximate optimization of the modularity (Clauset et al., 2004). All clusterings were found to be significant using the permutation test described in (Montastier et al., 2015) by generating 500 random networks with the same degree distribution (all clusterings were found to have a modularity larger than that obtained on the 500 random networks, p-value < 0.002). Clusters were compared using two methods: first, pairwise contingency tables between clusters were computed. Second, the normalized mutual information (NMI (Danon et al., 2005)) between pairs of clusterings was obtained. The NMI is a number between 0 and 1 measuring the similarity between two clusterings and is maximum (equal to 1) when the two clusterings are identical.

4.2.3 Functional analysis of the networks

4.2.3.1 Gene Ontology (Webgestalt)

Functional enrichment analysis based on GO was performed using the web tool Webgestalt (WEB-based GEne SeT AnaLysis Toolkit, http://www.webgestalt.org/option.php) updated on January 27, 2017 (Wang et al., 2013; Zhang et al., 2005). The web tool uses the Fisher exact test and controls for the number of false positives among the declared significant GOs terms. The False Discovery Rate procedure was used ((Benjamini and Hochberg, 1995), FDR < 5%). The analysis was performed using the Overrepresentation Enrichment Analysis (ORA) method, selecting non-redundant Biological Processes (BPs).

4.2.3.2 Ingenuity Pathway Analysis

The final network was analysed through the use of Ingenuity Pathway Analysis version 01-12 (updated on March 31st, 2018). Ingenuity Pathway Analysis (IPA, Ingenuity Systems; QIAGEN, Inc., Valencia, CA, USA, https://analysis.ingenuity.com/pa) contains a large bibliographic database (Ingenuity Pathways Knowledge Base) with various molecular relationships already identified between two genes (protein-protein interaction, ligand-receptor regulation, enzymatic modification,

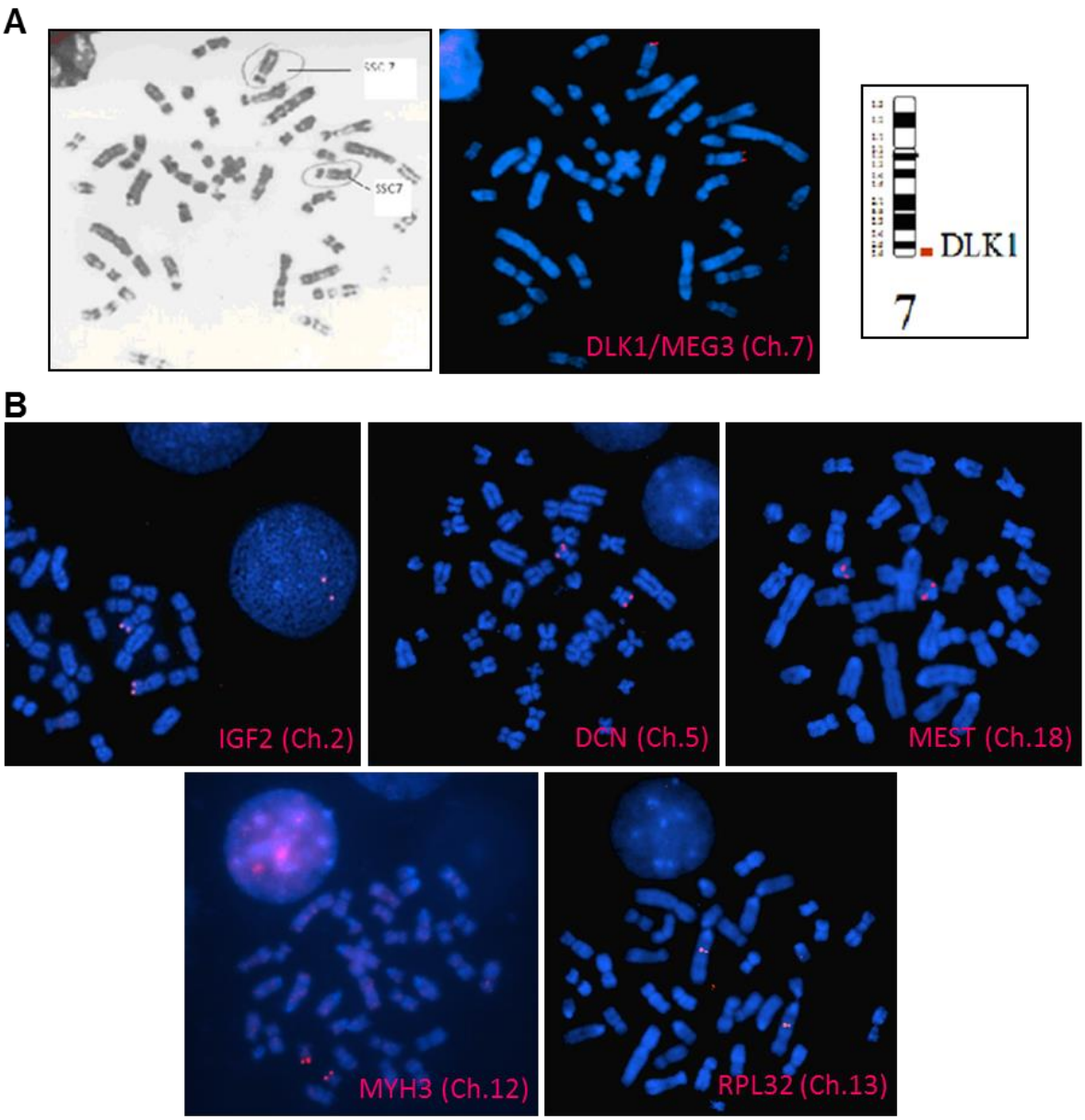


Figure 22. Verification of BAC probes specificity and location by 2D DNA FISH on porcine methaphases. *DNA probes were labeled with alexa fluor-568 (red), chromosomes with DAPI (blue) and images were obtained by fluorescence microscopy wide field. Each metaphase was shown before with G-banding for chromosome identification. (A) Example for the DLK1/MEG3 locus. The comparison of the two images allows the identification of chromosomes bearing spots. (B) The same procedure was applied for the other genes.*

transcriptional expression regulation, etc.). The obtained network is a graphic representation of the molecular relationships between molecules. All edges are supported by at least one reference from the literature, or from canonical information stored in the Ingenuity Pathways Knowledge Base. The obtained networks were improved for representation using Path Designer. Nodes are displayed using various shapes that represent the functional class of the gene product. The Functional Analysis identified the biological functions, the canonical pathways and the upstream regulators that were the most relevant to the dataset. Molecules from the dataset that were associated with biological functions, canonical pathways or upstream regulators in the Ingenuity Knowledge Base were considered for the analysis. Fisher's exact test was used to calculate a right-tailed *p*-value determining the probability that each function and pathway assigned to that dataset is due to chance alone. The networks proposed by IPA were cleaned (some nodes/genes were discarded) in order to keep only the genes necessary to connect the co-expressed genes. The three first networks were merged and regulation information was added to highlight transcription factors that could explain unexpected gene co-expression and nuclear co-localization (e.g. *MYH3* and *IGF2*; see Appendix 3 "Biological network reconstructed following Ingenuity data analyses").

4.2.4 Gene-gene nuclear associations

4.2.4.1 3D DNA FISH in interphase nuclei

Tissue preparation: Foetal muscle tissue was obtained from the *Longuissimus dorsi* muscle of 90-day gestation ♀MSxLW♂ pig and prepared as described in (Lahbib-Mansais et al., 2016) with slight modifications. In addition, muscle sample from the LW breed at 90-day gestation was also used to test some gene associations. When needed, stored muscle fibre packets were permeabilised for 8 min in cytoskeleton extraction buffer (100 mM NaCl, 300 mM sucrose, 3 mM MgCl2, 10 mM PIPES pH 6.8) containing 0.5% Triton X 100 and then fixed in cold 4% paraformaldehyde for 5 min. After washing in cold PBS, muscle packets were manually dilacerated directly on Superfrost glass slides (CML, Nemours, France) to isolate individual fibres, and air-dried before adding DNA probes for *in situ* hybridization.

DNA probes construction: Bacterial artificial clones (BACs) containing genes were isolated from porcine BAC libraries (available at the Biological Resources Center-GADIE, INRA, Jouy-en-Josas, France http://abridge.inra.fr/) using specific primers designed with Primer3 software (http://primer3.sourceforge.net/) (Appendix 4 "Information about BACs used as probes for 3D DNA FISH experiments"). For multiple label experiments, approximately 120 ng of each BAC DNA were random priming labelled (Bioprime DNA labelling kit, Invitrogen, Cergy Pontoise, France) directly by incorporation of dUTP Alexa Fluor (488 or 568) or indirectly with Biotin 6 dUTP detected by immuno-FISH. Chromosomal localizations of all BAC probes were controlled by 2D DNA FISH on porcine metaphases (Figure 22) prepared from lymphocytes according to standard protocols (Yerle et al., 1994).

IGF2 was previously localized on SSC2p17, *DLK1/MEG3* on SSC7q26 and *ZAR1* on SSC8q11-12 (Lahbib-Mansais et al., 2016). In this study, additional genes were localized on pig

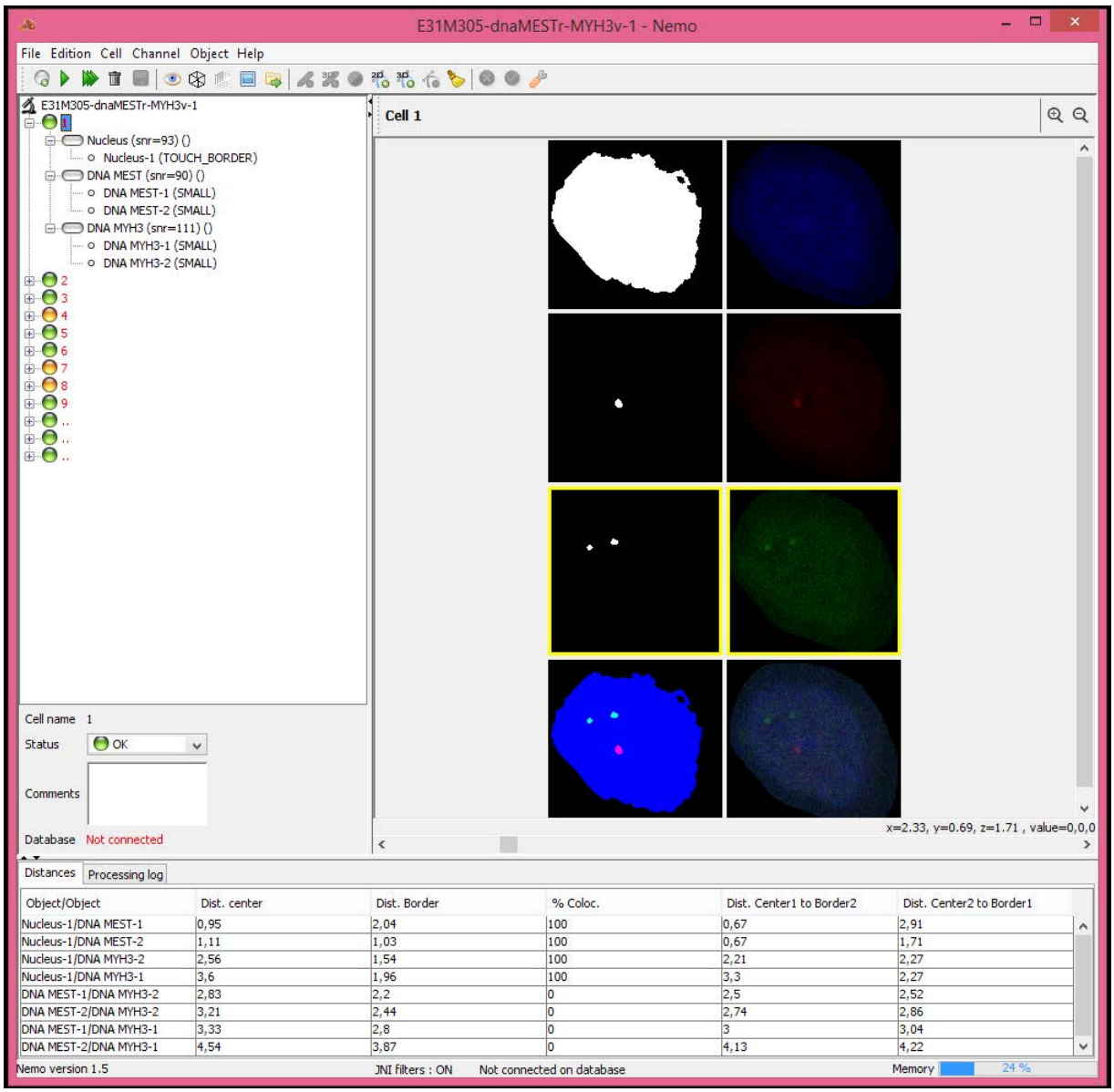


Figure 23. Illustrative exemple of a NEMO analysis window. *Nuclei are initially segmented and numbered from the raw images of a confocal field. Left panel: list of all segmented nuclei of a confocal field. For each nuclei, the deteted objects for each channel are listed (blue: nucleus; red: the alleles of MEST; and green: the two alleles of MYH3). Right panel: decomposition of each channel and final merge. The detected objects are visualized in the left column, and the raw images in the right column. Lower panel: distance measures and percentage of co-localization between each object. The center-to-center 3D distances (Dist. center) were chosen for further analyses.*

metaphases: *MYH3* on SSC12q, *MEST* on SSC18, *RPL32* on SSC13q24-33, *DCN* on SSC5qter, and *PRLR* on SSC16.

3D DNA Fluorescence in situ hybridization: 3D DNA FISH experiments were conducted using specific probes to label each gene with a different colour as described in (Lahbib-Mansais et al., 2016) with slight modifications. Probes were resuspended in hybridization buffer (50% formamide, 10% dextran sulphate, 2 mg/ml BSA, 2× SSC) at a final concentration of 110 ng/µl. Nuclear DNA and probes were simultaneously denatured at 74°C for 7 min and then incubated overnight at 37°C in a wet atmosphere (DAKO hybridizer). Washes were then performed with gentle agitation, first twice in 2× SSC at room temperature (RT) for 8 min, then twice for 3 min in 2× SSC, 50% formamide pH 7.0 at 40°C, and finally twice for 15 min in 2× SSC, then in PBS at RT. When a biotin-labelled probe was used, biotins were detected by incubating the slides with streptavidin-Alexa 568 or 488 at a final concentration of 5 µg/ml for 1 hour at RT.

4.2.4.2 Confocal microscopy and image analysis

Image captures: 3D acquisitions were performed at the T.R.I. Genotoul (Toulouse Réseau Imagerie, http://trigenotoul.com/en) imaging core facility in Toulouse (France). Image stacks were captured at different depths with a Leica TCSSP2 confocal microscope (Leica Instruments, Heidelberg, Germany) equipped with an oil immersion objective (plan achromatic 63× N.A. = 1.4). The Z-stacks (around 60 confocal planes per capture) were acquired at 1024 × 1024 pixels per frame using a 8-bit pixel depth for each channel at a constant voxel size of 0.077 × 0.077 × 0.284 µm.

Image analyses: Images were analysed with a specific software for measuring the 3D distances between signals (genes) (NEMO (Iannuccelli et al., 2010)) (Figure 23) as described in (Lahbib-Mansais et al., 2016). Euclidean distances were computed with respect to the x, y and z resolutions. Given the resolution on the z axis, at least three pixels corresponding to 0.852 µm (0.284 x 3) were required for a high resolution between two separate signals; consequently, 1µm was chosen as the upper cut-off for associated signals.

Gene-gene associations: In all 3D DNA FISH experiments, nuclei were only analysed when 4 signals (corresponding to the 2 alleles of each gene) were present. "Associated" signals were considered to be those separated by a distance (d) ≤ 1 µm, and were divided into two different classes: "close" signals (0.5 < d ≤ 1 µm), and "co-localized" signals (d ≤ 0.5 µm). The great majority of associations concerned uniquely one allele from each gene. To establish the threshold for distinguishing between associated and non-associated genes, two 3D DNA FISH experiments were performed as negative controls: first, between two genes (*ZAR1* and *PRLR*) located on different chromosomes and expressed at a very low level in muscle cells (Voillet et al., 2014), second, between *IGF2* (highly expressed) and *ZAR1* (low expression) (Lahbib-Mansais et al., 2016). In both cases, the two genes were found to be associated in only 8% of the analysed nuclei. Considering this value as a sporadic association between loci not expected to be associated, a 10% value was arbitrarily chosen to distinguish between associated and non-associated genes.

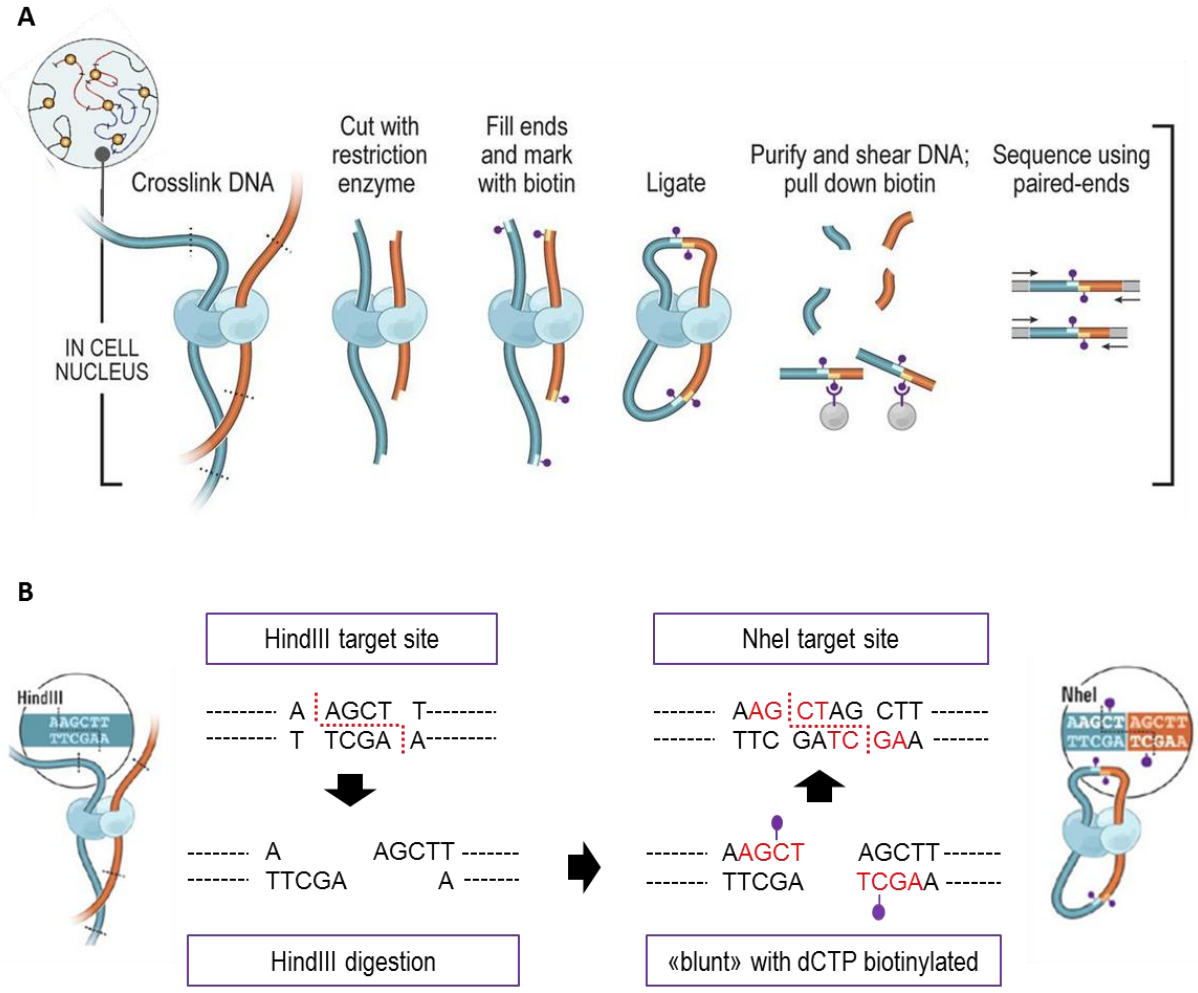


Figure 24. Hi-C experimental procedure. *(A) Illustration of the main steps performed in the Hi-C assays* (Rao et al., 2014). *(B) Detailed view of the digestion, ends filling and ligation steps. Figure adapted from* (Lieberman-Aiden et al., 2009).

4.3 Nuclear architecture and gene expression approach

4.3.1 Transcriptome data

4.3.1.1 Microarray data description

Expression data were obtained from the previous transcriptome study of skeletal muscle in pig during development (Voillet et al., 2014). The dataset consists of 44,368 probes for 17 samples (LW animals) at two different gestational stages (8 samples for the 90 days gestational age and 9 samples for the 110 days). A precise description of the experimental design and data collection can be found in (Voillet et al., 2014). Normalized expression data (log2 transformed) and sample information are available in NCBI (GEO accession number GSE56301).

4.3.1.2 Microarray probes alignment and annotation

Since the microarray was originally designed on a former version of the pig genome, the probes were aligned to the Sscrofa11.1 assembly version by using BLAT (v.35x1) with the parameters -minIdentity=95 -mask=lower. Alignments were obtained and processed by keeping unique best hits only with a minimum score of 30, and in case of several "blocks" in the alignment of a given probe -across two exons for instance- the longest block (with a minimum length of 20) was kept. The 42,885 resulting probes were then annotated depending on their mapping position relatively to the annotated genes of the Ensembl v90 annotation. Probes that overlapped an annotated gene from the Ensembl annotation -either within the entire genic region or on an annotated exon- were assigned to the corresponding gene ID. A total of 38,043 probes could be assigned to a gene, from which, 30,594 correspond to probes mapped to exonic regions. The total of distinct genes targeted with probes mapped to genes was 13,530 and those targeted with probes mapped to exons were 12,465.

4.3.2 High-throughput chromosome conformation capture (Hi-C)

The experimental FAANG protocol, based on the in situ Hi-C protocol used in (Rao et al., 2014) (available in http://ftp.faang.ebi.ac.uk/ftp/protocols/assays/INRA_SOP_Hi-C_HA_v1_20160610.pdf), was slightly modified in terms to adapt the Hi-C experiments and libraries to the fetal muscle tissue. A detailed description of the experimental procedure (Figure 24A) is presented below.

4.3.2.1 Hi-C experiments

Muscle nuclei isolation and crosslink: Longissimus dorsi muscle samples from three 90- and three 110-day post coitum (p.c.) fetus (3 males at 90 days, 2 males and 1 female at 110 days) of a European Large White (LW) breed (F1 ♂LW x LW♀), were frozen in isopentane cooled with liquid nitrogen and stored at -80 ^{0}C until needed. For each sample, around 1.5 g of frozen stored fetal muscle was thawed at Room Temperature (RT) and dissected with scalpel blades to obtain a homogenate of mashed muscle. Dissected tissue was washed in phosphate-buffered saline (PBS) to remove blood. Nuclei were disaggregated by rubbing (pipetting up-down many times), filtered through a cell strainer

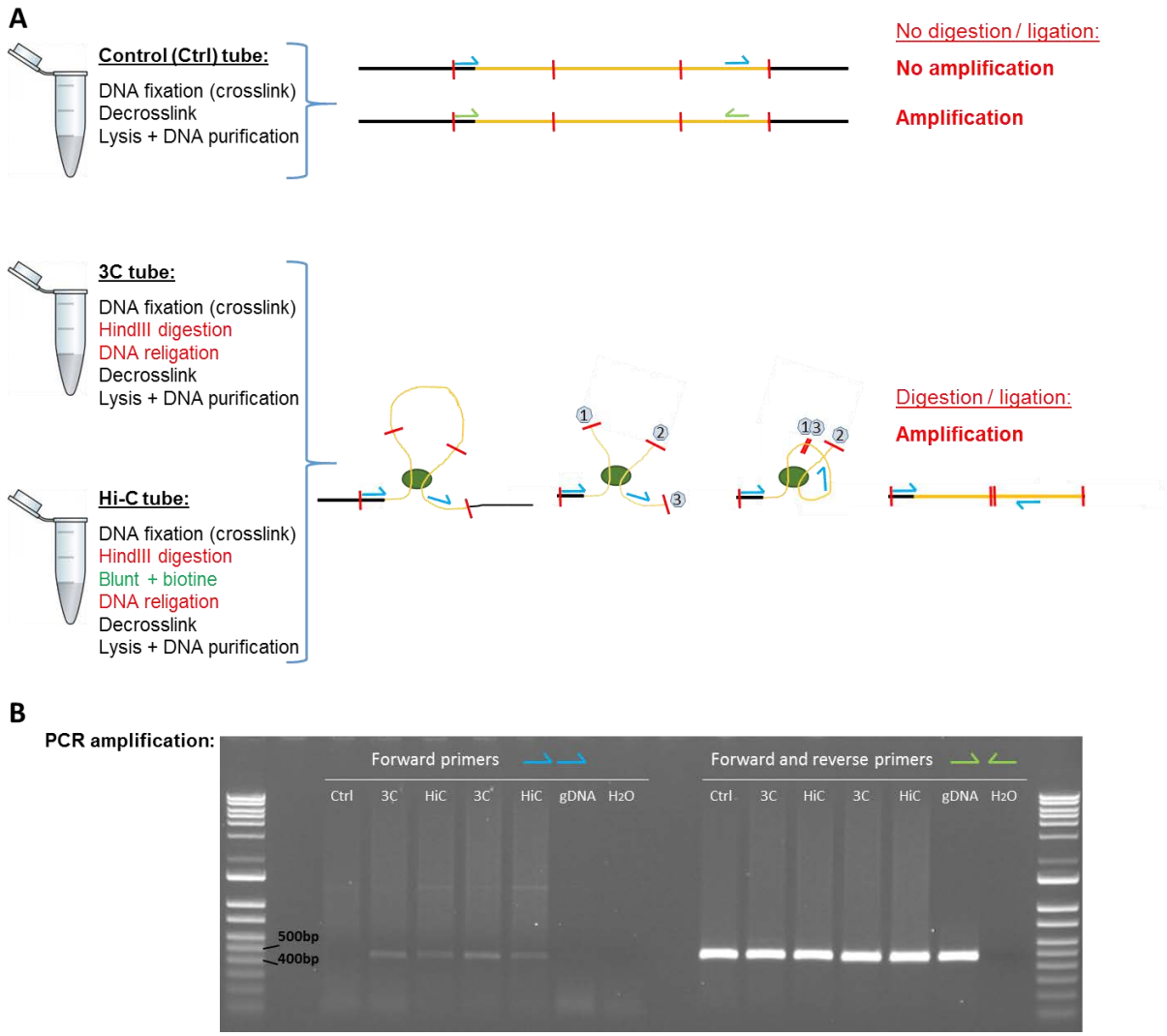


Figure 25. PCR quality control of Hi-C products. *A couple of forward primers (blue arrows) and a couple of forward and reverse primers (green arrows) were used to amplify the same genomic region in the Control, 3C and Hi-C tubes (left side of the A panel), as well as in genomic DNA (gDNA) and in absence of DNA (H2O). (A) In the Control tube, where the steps of digestion and ligation were not performed (similar to gDNA), the region was amplified with the forward and reverse primers but no amplification was observed with the couple of forward primers as expected. In both 3C and Hi-C tubes, digestion and religation were performed allowing in some religation events switching the sense of one DNA fragment thus permitting the amplification with the couple of forward primers. (B) Migration of the PCR products in a 1% agarose gel.*

(70 µm) and centrifuged at 1200g 5 min to get a high yield of cells. Pellet was resuspended in 3 ml Dulbecco's Modified Eagle's Medium (DMEM) with glutamax (1% formaldehyde) and incubated 10 min at RT. To quench fixation, 0.125 M final glycine was added 5 min at RT, then cooled 5 min on ice. After 5 min centrifugation at 1200 g, pellet was washed in ice-cold PBS (with protease inhibitors). An aliquot of cell suspension was stained with 4',6-diamidino-2-phénylindole (DAPI) and phalloidin to check nuclei quality and integrity (see Appendix 5 "Quality check of nuclear integrity in Hi-C experimental steps").

Nuclei permeabilization and endonuclease-based DNA fragmentation: For each Hi-C experiment, 3 pellets were prepared with around 5 million cells per tube (named: Hi-C, 3C and control). Pellet in control tube was resuspended in 200 µl of water and kept at 4 °C. Tubes Hi-C and 3C were resuspended in 0.05 % Sodium Dodecyl Sulfate (SDS) and incubated 10 min at 62 °C. To quench the SDS, 0.1 % final Triton X-100 was added for 15 min at 37 °C, then 100U of HindIII in 25 µl of 10X NEbuffer 2.0 were added to digest overnight at 37 °C on the wheel. Fifty µl of water were added to the 3C tube.

Biotinylation, ligation and decrosslink: To fill overhangs with marked dNTPs and obtain blunt ends 50 µl of fill-in master mix (200 nM dATP, dGTP, dTTP, biotin-14-dCTP, 50U Klenow) were added to the Hi-C tube only, and both tubes (Hi-C and 3C) were incubated at 37 °C, 1 hour on the wheel. Then they were incubated at 62 °C 20 min to inactivate the enzyme and 900 µl of ligation mix was added to each tube (1.3X T4 DNA ligase reaction buffer, 1.1% Triton X-100, 130 ng/ml BSA, 2000U T4 DNA ligase) and incubated 1 hour at RT and then overnight at 4°C on the wheel. Proteins were degraded by incubating the three tubes (Hi-C, 3C and control) with 50 µl of Proteinase K (20 mg/ml) and 120 µl of 10% SDS at 55 °C for 30 min, then with 130 µl of NaCl 5M at 68°C overnight.

DNA purification and enrichment of biotinylated DNA: DNA was precipitated with 1.6 volume of 100% ethanol and 0.1 volume of 3M sodium acetate (pH5.2) at -80°C 15 min, then centrifuged at 4°C (15400rpm, 10 min), the pellet was resuspended in 70% ethanol, centrifuged at 4 °C (15400rpm, 5 min) and dissolved in nuclease-free water (20 min at 37°C). To desalt and purify DNA, 1.8 volume of CleanPCR magnetic beads were added and incubated 5 min and after washing twice for 30 seconds with 80% ethanol and letting dry for 3 min, the beads were resuspended in TE (10:1, Tris 10 mM pH8.0, EDTA 0.1 mM) buffer solution for control and 3C tubes and in TE (10:0.1) for Hi-C tubes.

Removing non-ligated biotinylated DNA: To remove non-ligated biotinylated DNA, 28 µl of T4 DNA polymerase mix reaction (714 ng/ml bovine serum albumin (BSA), 5.3X NE Buffer 2, 357 nM dATP, 357 nM dGTP and 30U T4 DNA polymerase) were added to the Hi-C tubes and incubated at 12 °C for 90 min. The reaction was stopped by adding 2 µl of 0.5 M Ethylenediaminetetraacetic acid (EDTA) and heating 20 min at 75 °C. Then DNA was purified with magnetic beads as explained before and resuspended in TE (10:0.1).

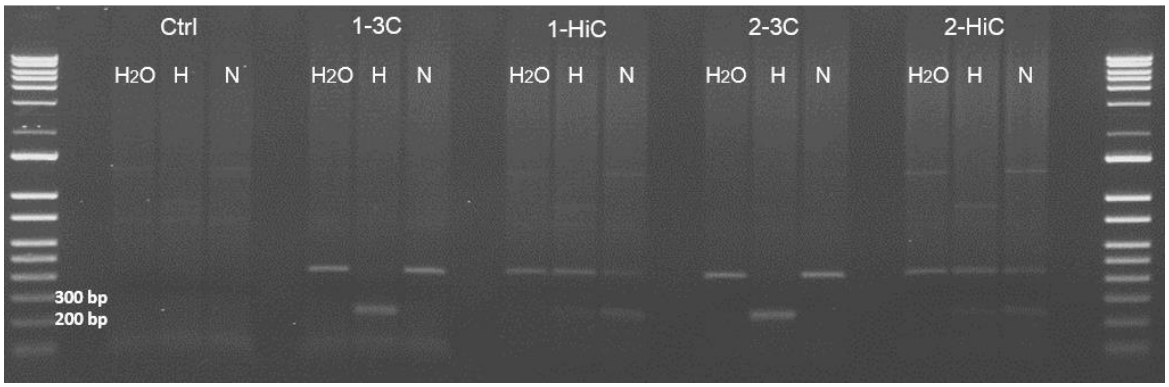


Figure 26. Digestion quality control of the PCR products. *PCR products from the 3C and Hi-C tubes amplified with a couple of forward primers were digested with HindIII (H) and NheI (N) restriction enzymes and the digestion products were migrated in a 1% agarose gel. The digestions were also performed in absence of DNA (Ctrl), and absence of enzymes (H2O) as negative controls. 3C tubes, where no filling ends with biotin incorporation was performed, were digested only with HindIII (smaller band ~210 bp). Hi-C tubes, where biotin incorporation was performed with the resulting formation of a NheI target site, were mainly digested with NheI (~210 bp) and slightly digested with HindIII.*

4.3.2.2 Quality control of Hi-C experiment

The quality control is based on the following principle: when DNA is digested with HindIII, after filling and ligation of the digested ends, the HindIII target site disappears and a new site, which is recognized by the NheI restriction enzyme, is created instead (Figure 24B). To check the efficiency of the Hi-C assays, PCR are performed around one HindIII restriction site with two forward primers (Fwd1: 5' TCTGGGCAGGTCACTCATT 3'; Fwd2: 5' TCTCGGGATGCTGAGTGTTT 3'; product size = 425 bp). A reverse primer combined with Fwd1 was used as a control (Rv1: 5' AAACACTCAGCATCCCGAGA 3'; product size = 465 bp). In Hi-C and 3C assays, some religation events allow switching the sense of one DNA fragment and PCR amplification with these primers is possible (Figure 25). Then the PCR amplification products from the couple of forward primers are digested either with HindIII or NheI (product sizes = 201 + 215 bp). For 3C experiments, HindIII should cleave the PCR products while NheI should not. For Hi-C experiments, NheI should cleave most of the PCR products while HindIII should cleave only a small fraction (Figure 26).

4.3.2.3 Hi-C libraries production and sequencing

The whole process of Hi-C libraries production and sequencing was performed at the GeT-PlaGe (Génome & Transcriptome - Plateforme Génomique) (https://get.genotoul.fr/en/ in Toulouse, France).

DNA fragmentation and sizing: 1.4 µg of DNA from the Hi-C experiments were fragmented with a Covaris machine. Then, 0.55 volumes of CleanPCR magnetic beads were added to the fragmented DNA to select fragments < 600 bp (5 min incubation and keeping the supernatant), and 0.7 volumes of beads were added again (5 min incubation and removing supernatant) to remove fragments < 200 bp. Then beads were washed with 80% ethanol and DNA was recovered with Resuspension Buffer.

Biotinylated DNA purification: To purify biotinylated DNA, 1 volume of M-280 streptavidin magnetic Dynabeads was added and after 15 min incubation, the supernatant was removed and the beads were washed 4 times with beads wash buffer (Nextera Mate Pair Preparation Kit, Illumina) and twice with Resuspension buffer. From this point, all steps were performed while DNA remains attached to the beads.

End repair, 3' adenylation and adapters ligation: To repair DNA breaks, 60 µl of water and 40 µl of End Repair Mix 2 (TruSeqNano DNA library prep, Illumina) were added and incubated 30 min at 30°C, then beads were washed as explained before. To allow the adapters ligation, an 'A' nucleotide was added to the 3' ends by adding 17.5 µl of water and 12.5 µl of A-Tailing Mix (TruSeqNano DNA library prep, Illumina) and incubating 30 min at 37 °C and then 5 min at 70°C to inactivate the enzyme. To ligate the adapters to the DNA extremities, 2.5 µl of Resuspension Buffer, 2.5 µl of DNA Ligase Mix and 2.5 µl of DNA Adapter Index (TruSeqNano DNA library prep, Illumina) were added (10 min incubation at 30°C, then 5 µl of Stop ligation Buffer) and then beads were washed as before.

PCR enrichment and DNA purification: DNA was amplified by 12 PCR cycles (15 sec at 98 °C – 30 sec at 60 °C – 30 sec at 72 °C) by resuspending beads in 50 µl of PCR mix (25 µl Enhanced PCR

Table 3. Libraries size and concentration estimations of the libraries.

	library size estimated with FA (bp)	library concentration by qPCR (nM)
Rep1-90	619	10,74
Rep2-90	635	4,19
Rep3-90	547	15,01
Rep1-110	540	19,87
Rep2-110	570	19,86
Rep3-110	644	35,69

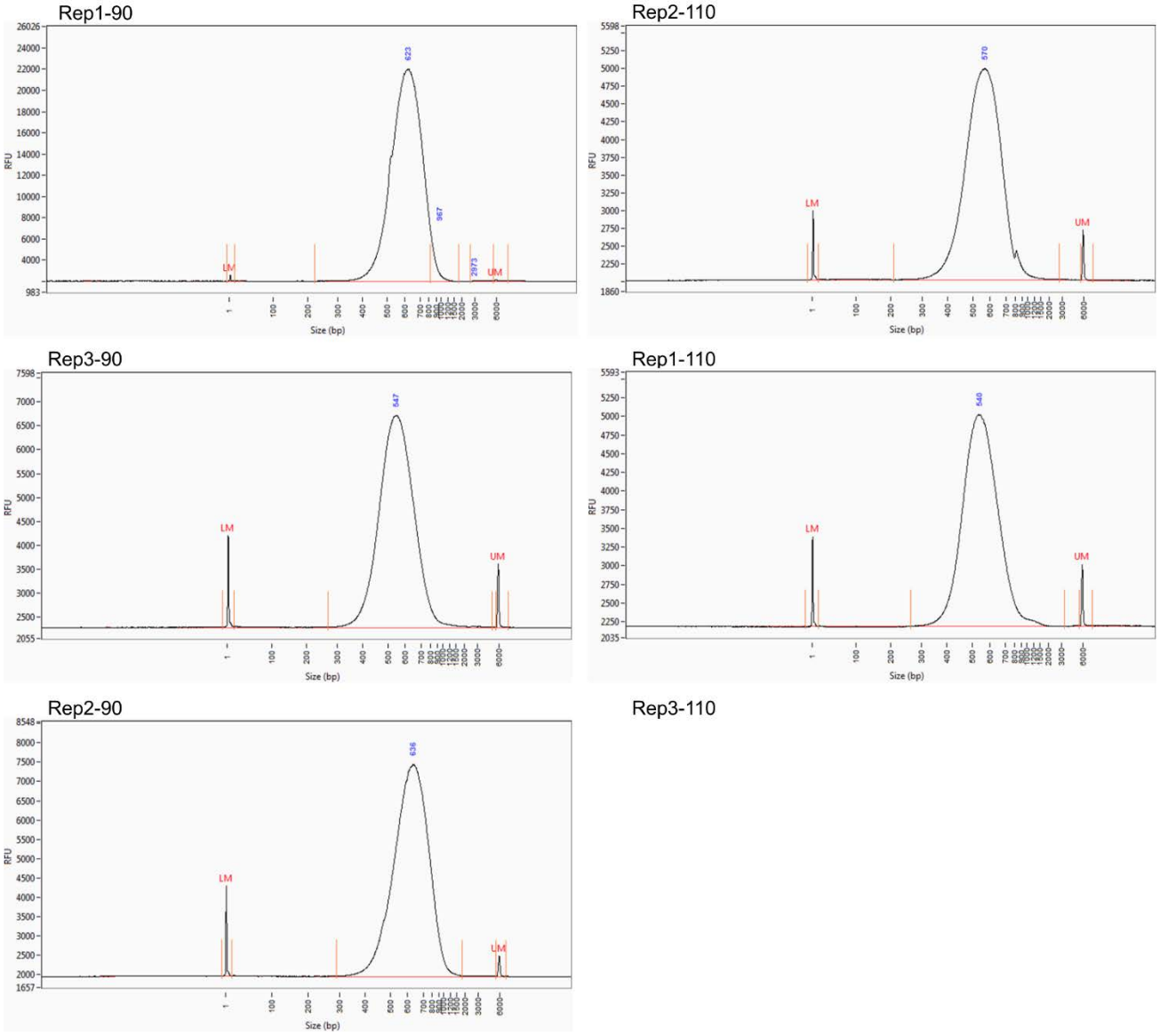


Figure 27. Fragment analyzer profiles of the Hi-C libraries. *Missing data for Rep3-110.*

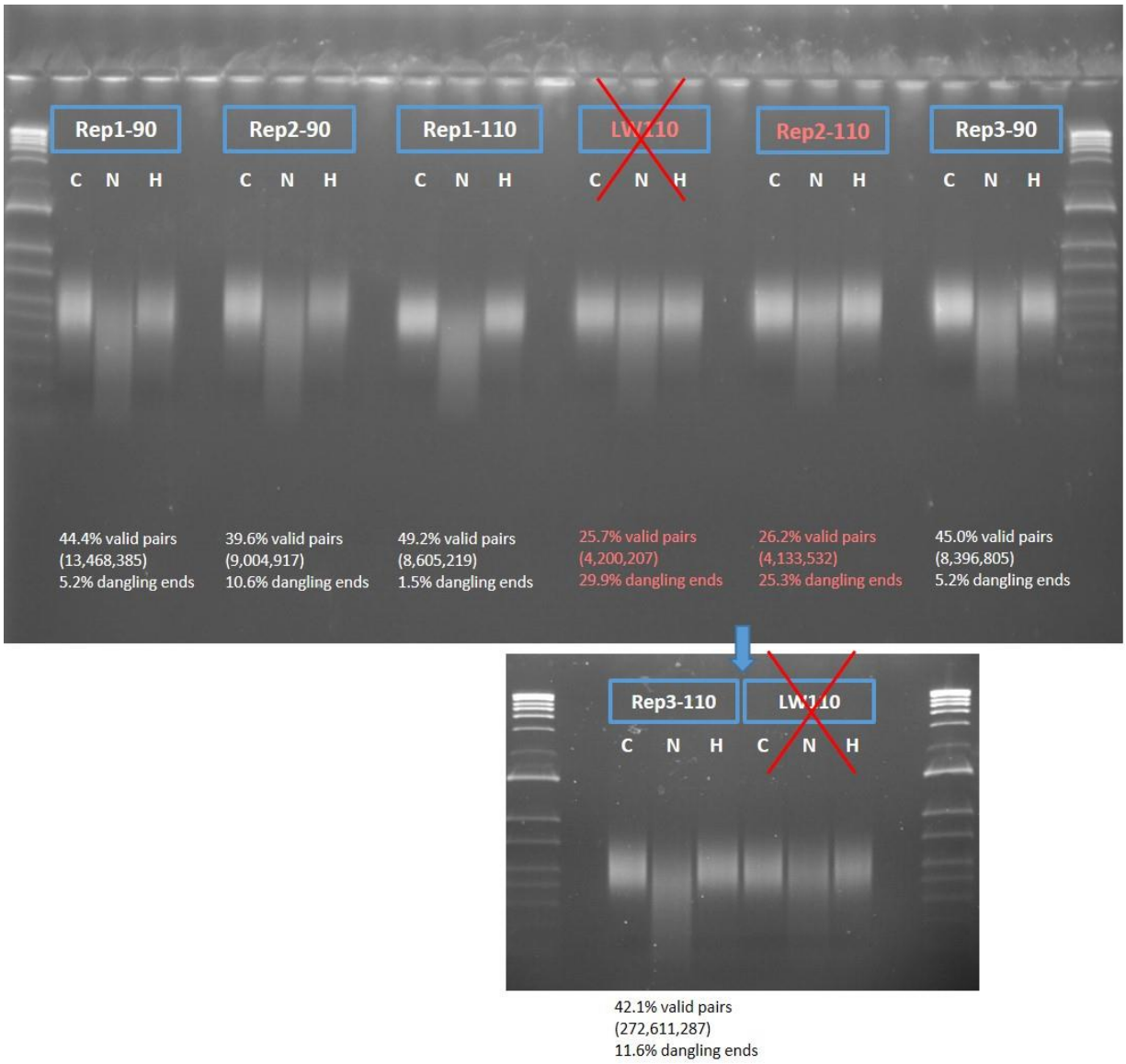


Figure 28. Digestion quality control of the Hi-C libraries. *All six initial Hi-C libraries were digested with HindIII (H) and NheI (N) restriction enzymes and run in a 1% agarose gel. After an initial low depth sequencing and Hi-C data processing (see below), two Hi-C libraries (110 days gestation) presented low percentages of valid pairs and high of dangling ends compared with the others. This corresponds with the less efficient digestion profile observed in these two libraries (upper panel). Two new Hi-C libraries were prepared in order to replace these two, but only one (Rep3-110) showed a good digestion profile (lower panel) and was used to replace the one presenting a less proportion of valid pairs.*

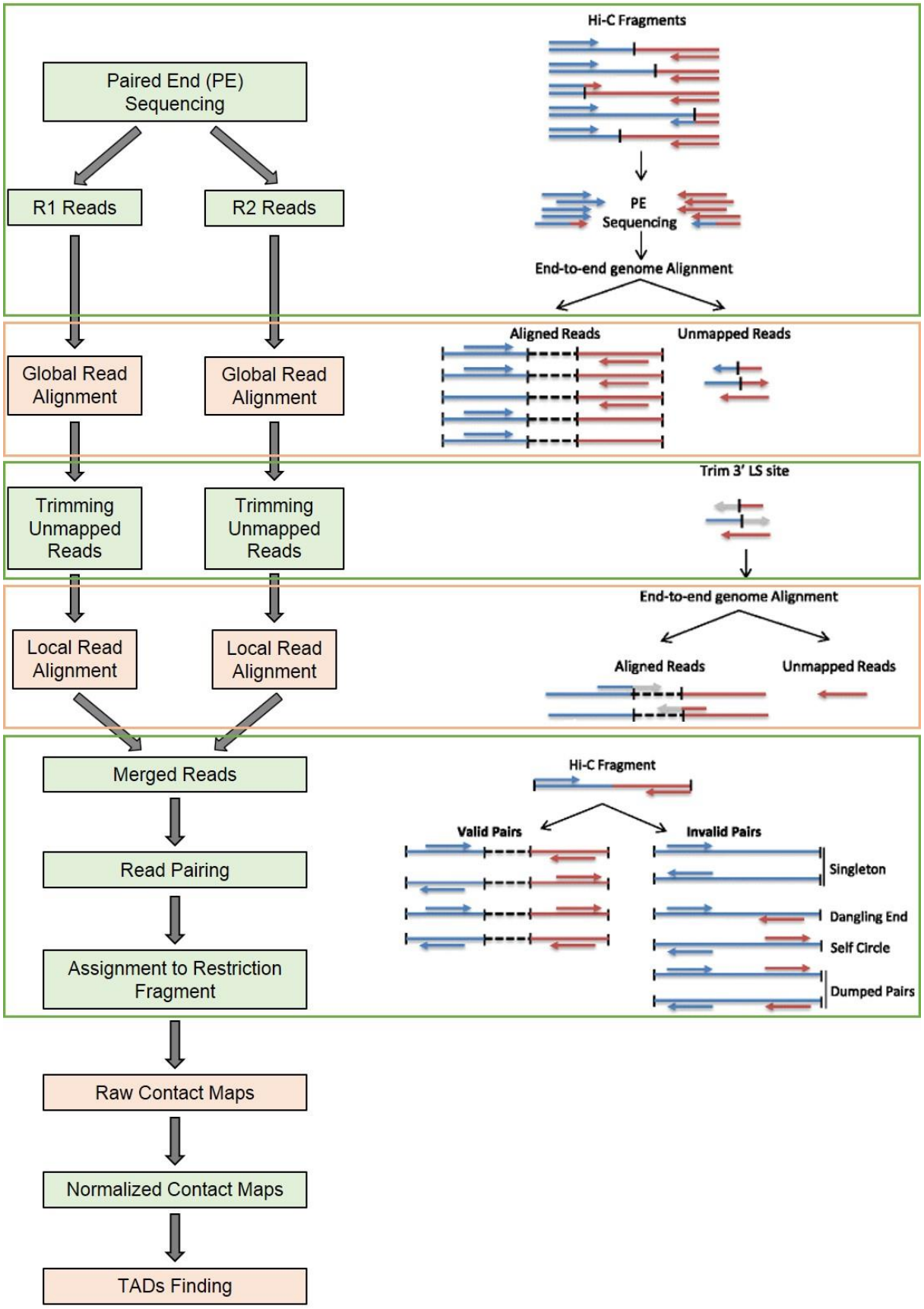


Figure 29. Hi-C pipeline workflow. *Figure created with images obtained from* (Servant et al., 2015).

mix, 5 µl PCR primer Cocktail and 20 µl water, TruSeqNano DNA library prep, Illumina). To recover DNA from the beads, 0.6 volume of CleanPCR magnetic beads were added and incubated 5 min, then washed twice with 80% ethanol, resuspended in 30 µl of Resupension Buffer and after placing in a magnetic rack, supernatant containing the libraries was recovered.

Size selection control and sequencing: Libraries size was then controlled with the Fragment Analyzer (FA) (Figure 27) and quantified by qPCR (Table 3). In addition an aliquot was digested by using the NheI and HindIII enzymes to verify if selected fragments are the ones containing the filled-in biotinylated religation sites as done in (Belton et al., 2012) (Figure 28) and they were sequenced in pool in one HiSeq3000 lane to validate their quality. For depth sequencing, the pool was paired end (PE) sequenced in 9 lines of a HiSeq3000 (reads size = 150 bases), producing from ~ 476 M to 685 M read pairs per library in total.

4.3.2.4 Hi-C data processing

The total 3,447,428,742 PE reads were processed with a bioinformatics pipeline developed in the context of the FAANG consortium (Andersson et al., 2015) (Foissac et al., 2018). This pipeline mainly combines existing software: HiC-Pro v2.9.0 for mapping the reads (with Bowtie 2 v2.3.3.1 (Langmead and Salzberg, 2012)) and obtaining the contact matrices (Servant et al., 2015), ICE to normalize the matrices (Imakaev et al., 2012), HiTC v1.18.1 for various tools (Servant et al., 2012) and Armatus v2.1 to call TADs (Filippova et al., 2014). Figure 29 shows the main steps of the pipeline described below: read cleaning, trimming, mapping and pairing, detection and filtering of valid interaction products, binning, contact map normalization, and TADs finding.

Mapping: A first smaller dataset was initially mapped to the *Sus scrofa* genome version 10.2, and latter mapped to *Sscrofa11.0*. Then after re-sequencing, the full dataset of reads was mapped to the *Sus scrofa* genome (version 11.1) in two steps. First, during what is called "global alignment" in the HiC-Pro software, both reads of each pair were mapped independently in single-end (SE) mode with bowtie2 using the full read sequences. The reason for this SE mode is that paired-end (PE) mappers typically use the genomic distance between potential positions of reads to assist the mapping process since the read-to-read distances are expected to fit the size distribution of the sequenced fragments. Here, this genomic distance is irrelevant since the goal of the Hi-C protocol is to capture long-range and *trans* ligation events, so reads from the same pair are initially mapped separately. Also, due to their chimeric nature, many reads might not directly map on the genomic sequence over their entire length. Indeed, Hi-C hybrid fragments are issue of digestion and religation events that bring together two interacting loci. Therefore, a single read might contain sequences from distant genomic regions when spanning the ligation junction, which hampers the mapping process. Consequently, around 30-40% of reads were not mapped during the global alignment. These chimeric reads need to be trimmed in order to remove the portion beyond the ligation junction, and then re-mapped during a "local alignment" step, still in SE mode. To do that HiC-Pro needs to be provided with the re-ligation sequence (AAGCTAGCTT) resulting from the HindIII digestion, fill-in, and ligation. Global and local alignments use `bowtie2-2.3.3.1-linux-x86_64` and `bowtie2` options: `--very-sensitive -L 30` (global) or `-L 20` (local);

`--score-min L,-1,-0.1` (global) or `L,-0.6,-0.2` (local); `--end-to-end`; and `–reorder`. From 46.2 to 73.7% of the previously unmapped reads could be successfully mapped and therefore retrieved after trimming. In the last step of the mapping process, read pairs were rebuilt whenever possible using the set of all SE-mapped reads (either from the global or local alignment) to generate the final alignment files in bam format. Singletons (reads for which the "mate" could not be mapped: ~14% of the initial material) were discarded, resulting in a total of 2,367,601,471 read pairs (68.7% of the initial material).

Detection of valid interactions: Mapped read pairs were classified into valid and invalid pairs. The valid pairs are those with reads in the expected mapping configuration, meaning that both reads map near a HindIII restriction site. More precisely, the cumulative distance between their 5'-end to the closest HindIII site downstream should fit within the range of the expected molecular size distribution of the library. The distribution of these genomic distances (read1-to-HindIIIsite1 + read2-to-HindIIIsite2) is estimated during the quality control step and used to define the threshold values of the accepted range: from 20 bp to 1 Kb. Invalid pairs with a fragment size outside of this range were therefore discarded by specifying the parameters -i 20 (MIN_INSERT_SIZE) -I 1000 (MAX_INSERT_SIZE). Read pairs classified as dangling end and self-cycle ligation were discarded. These are obtained when both reads of the read pair belong to the same restriction fragment of the genome. The self-circularized ligation products correspond to pairs with reads in opposite directions within the same restriction fragment, and the unligated dangling end products correspond to pairs with reads facing each other (Figure 29) (Imakaev et al., 2012). PCR duplicates (redundant pairs with both reads at the same positions: about 13.6% of the initial material) were also filtered out.

Contact map generation: Only valid pairs involving two different restriction fragments are used to build the contact maps. To build them, the genome is segmented into intervals of equal number of bases, called bins. Then, the number valid read pairs (number of contacts or counts) per bin pair is reported. In this study, the binning was generated at 500, 200 and 40 Kb (bin size), which define the resolution of the Hi-C matrices. Large bin sizes allow identifying higher order chromatin conformation features, while smaller bin sizes allow identifying local chromatin structures. Therefore, the smaller is the bin size, the higher will be the resolution. However, this parameter of resolution is intimately linked to the sequencing depth. For instance, at high resolutions, the genome is highly segmented, and the number of read counts per bin pair is low. Therefore, high resolution Hi-C matrices need higher sequencing depth in order to compensate this effect.

Intra-matrix normalization and display of contact maps: HiC-Pro uses for normalization a fast sparse-based implementation of the iterative correction and eigenvector decomposition (ICE) method (matrix balancing ICE normalization). Contrary to any parametric normalization approaches, this method does not explicitly model specific biases introduced by experimental procedures and by intrinsic properties of the genome (such as GC content, mappability or restriction site density), but treat them globally instead. ICE normalization assumes that the bias for detecting contacts between two regions can be represented as the product of the individual biases of these regions. Concretely, this method

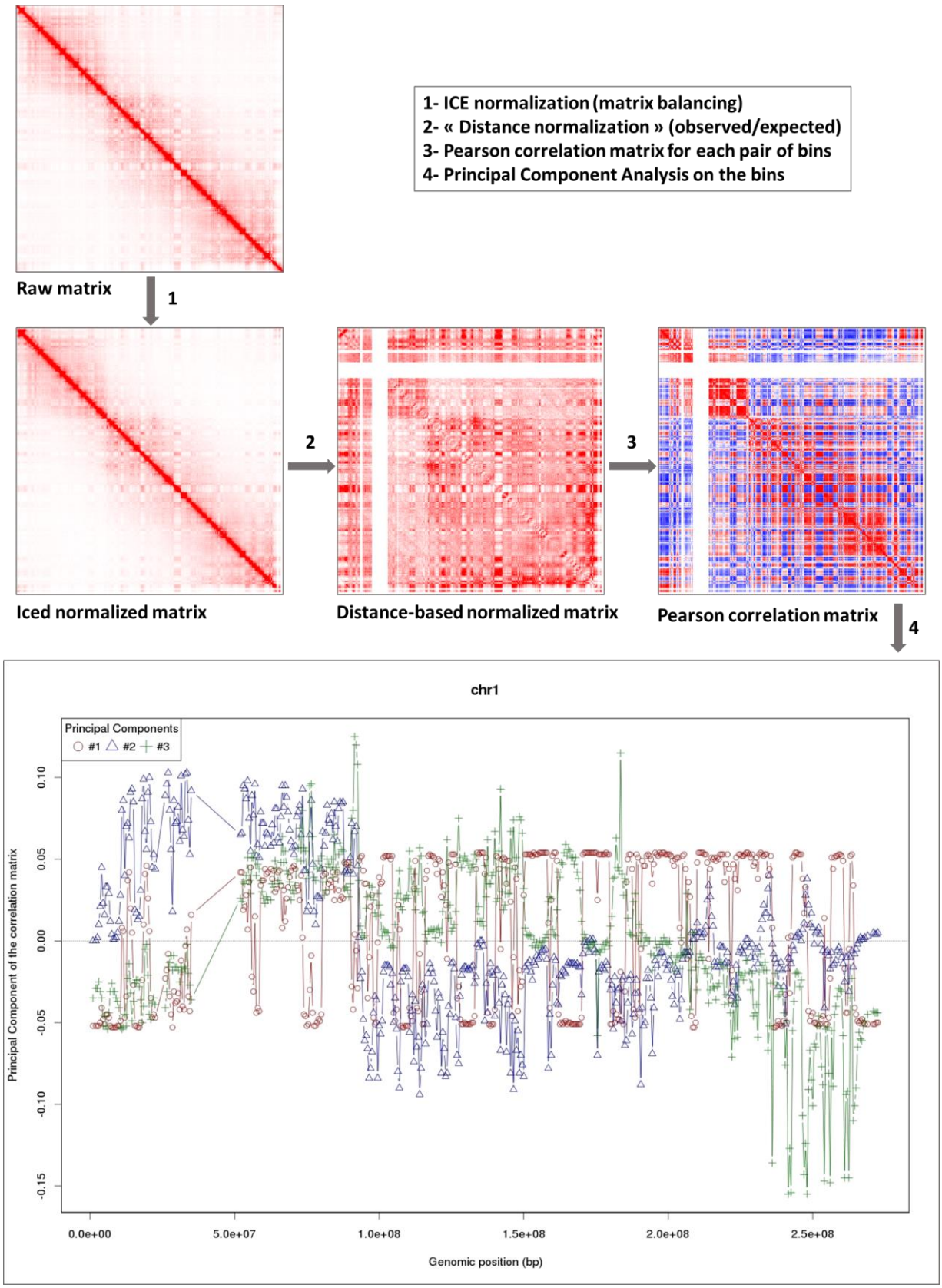


Figure 30. Method for predicting Hi-C A and B compartments. *Illustration of the A/B compartment calling workflow using the 90 days merged matrices of chromosome 1. Lower panel: the first three eigenvectors are shown along the chromosome to illustrate the relevance of PC#1 as the discriminative value to segregate bins between A and B compartments.*

assumes an equal exposure to contacts for each region with all the others (Imakaev et al., 2012; Servant et al., 2015). 2% of bins showing the lowest counts were filtered out by fixing the following parameter: `FILTER_LOW_COUNT_PERC = 0.02`. The maximum number of iterations was also fixed to 100 (`MAX_ITER = 100`).

The pipeline here presented uses **HiTC R** / Bioconductor package v1.18.1 (Servant et al., 2012) to visualize the normalized contact maps generated by HiC-Pro.

TADs finding: Only the longest 25 chromosomes/scaffolds were considered for TAD calling, with a specific focus on the 18 autosomes. 40-Kb resolution matrices (HiC-Pro output format) were extracted for each chromosome separately and converted to square matrices before running Armatus (http://www.cs.cmu.edu/~ckingsf/software/armatus/ (Filippova et al., 2014)) to generate TADs. Armatus uses a multiscale approach to find TADs at various size scales. This method uses a score function that encodes the quality of putative domains based on their local density of interactions. The algorithm used to identify topological domains in chromatin from interaction matrices uniquely requires a single specific parameter γ_{max}, which was fixed to 0.5. Armatus generates then TADs at different γ values (from 0 to 0.5) by incrementing this parameter in steps of 0.05. As γ decreases, the average size of the domains increase, conform to a hierarchical domain structure. Armatus offers additionally the possibility to obtain a consensus set of TADs, which have been showed to persist across multiple resolutions (Filippova et al., 2014). Therefore, consensus TADs were used in this study for subsequent analysis.

CTCF prediction: The position specific frequency matrix corresponding to the CTCF-binding motif was recovered from the JASPAR Transcription Factor Binding Sites (TFBS) catalogue (http://jaspar.genereg.net/, (Mathelier et al., 2016)). CTCF genomic occurrences were predicted by running FIMO v.4.11.1 (Grant et al., 2011) software with the JASPAR CTCF frequency matrix. Then, the density of CTCF predicted motifs with respect to TADs was obtained.

A/B compartments detection: A and B compartments were obtained for each chromosome after matrix balancing ICE normalization, followed by a distance-based normalization, using the method described in (Lieberman-Aiden et al., 2009) (Figure 30). ICE-normalized counts, Kij, were corrected for a distance effect with:

$$\widehat{K}ij = \frac{Kij - \overline{K}^d}{\sigma^d}$$

in which $\overline{K}ij$ is the distance-corrected count for the bins i and j, $\overline{K}^d$ is the average count over all pairs of bins at distance $d = d(i,j)$ and σ^d is the standard deviation of the counts over all pairs of bins at distance d. Then, Pearson correlations were computed between bins, by using interaction counts with all the other bins of the same chromosome, and a Principal Component Analysis (PCA) was performed on the correlation matrix. The overall process was performed similarly to the method implemented in the **R**/Bioconductor package **HiTC** (Servant et al., 2012). Boundaries between A and B compartments were identified according to the sign of the first PC (eigenvector). Since PCAs had to be performed on each chromosome separately, the average counts on the diagonal of the normalized matrix were used to

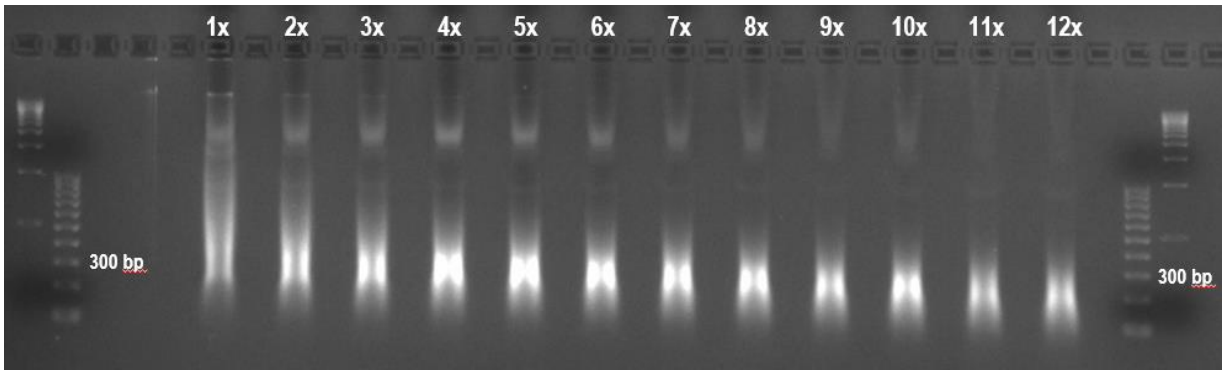


Figure 31. DNA sonication test*Erreur ! Signet non défini.. Before running the sonication on the real samples, several sonication cycles (from 1 to 12) were performed on fetal porcine muscle cells to determine the number of cycles necessary to obtain DNA fragments of 300 bp. Six cycles were observed to be appropriate for the obtention of 300 bp DNA fragments.*

identify which PC sign (+/-) should be assigned to A and B compartments for each chromosome. This allowed to automatically obtain a homogeneous assignment across chromosomes.

4.3.3 Chromatin Immunoprecipitation sequencing (ChIP-seq)

All ChIP-seq experiments and ChIP-seq libraries production were performed during a three months PhD mobility in the context of a collaboration with FAANG contributor members of the ABG (Animal Breeding and Genetics) group at Wageningen University, Netherlands. The sequencing of the ChIP-seq libraries was performed at the GeT-PlaGe (Génome & Transcriptome - Plateforme Génomique) (https://get.genotoul.fr/en/ in Toulouse, France).

4.3.3.1 ChIP-seq experiments

Muscle nuclei isolation and crosslink: Muscle samples from the same six animals used in Hi-C assays (3 fetuses at 90 days of gestation and 3 fetuses at 110 days) were used for ChIP-seq assays. For each sample, nuclei from around 0.8g of frozen stored fetal muscle were obtained as explained before for Hi-C experiments. Pellet was resuspended in 1/10 volumes of buffer A with formaldehyde (148 mM NaCl, 1.48 mM EDTA, 0.74 mM Egtazic Acid (EGTA), 74 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES), 11% formaldehyde) and incubated 10 min at RT. 1/10 volume of ice-cold 1.25 M glycine was added for 2 min at RT to quench fixation. Nuclei were centrifuged 5 min at 1600g, then the pellet was washed with ice-cold PBS (with protease inhibitors) and nuclei were centrifuged again 5 min (4 °C, 1600g) and resupended in 28 ml buffer C (150 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 50 mM HEPES) at 4 °C for 10 min. An aliquot of cell suspension was stained with DAPI and phalloidin to check nuclei quality and integrity. Nuclei were then centrifuged 5 min (4 °C, 1600g) and the pellet resuspended in 1X incubation buffer (0.75% SDS, 5% Triton X-100, 750 mM NaCl, 5 mM EDTA, 2.5 mM EGTA, 100 mM HEPES, 1/50 volume PIC reagent) so the final concentration was around 15 million cells/ml.

DNA fragmentation: A preliminary test of DNA sonication was performed in fetal muscle nuclei to define the number of sonication cycles needed to obtain DNA fragments of approximately 300 bp (Figure 31). The suspension was then sonicated in a water bath at 4 °C using the Bioruptor Pico sonicator (6 cycles, 30 seconds on/30 seconds off in order to favor fragments of 300 bp), and after 5 min centrifugation (4 °C, 13000rpm) the supernatant was snap freeze and stored at -80 °C. An aliquot of supernatant was used for a decrosslink test by adding 2 µl of 10 mg/ml of Proteinase K (PK) at 65 °C for 1hour.

Chromatin Immunoprecipitation: When needed, 300 µl chromatin (around 4.5 million cells) was thawed and incubated overnight at 4 °C in final concentration 0.1% BSA, 1X PIC, 1X incubation buffer, 3 µl CTCF antibody). To recover CTCF binding fragments, 15 µl of protein A/G magnetic beads (50% slurry), previously washed and resuspended with incubation buffer (0.1% BSA), were added to the 300 µl of chromatin and incubated 90 min at 4 °C on the wheel. Then beads were washed at 4 °C for 5 min twice with wash buffer 1 (0.1% SDS, 0.1% sodium deoxycholate (DOC), 1% Triton, 150 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 20 mM HEPES), once with wash buffer 2 (same as wash buffer 1

but 500 mM NaCl), once with wash buffer 3 (250 mM LiCl, 0.5% DOC, 0.5% Nonidet P-40, 1 mM EDTA, 0.5 mM EGTA, 20 mM HEPES) and twice with wash buffer 4 (1 mM EDTA, 0.5 mM EGTA, 20 mM HEPES). After washing steps, 400 µl of fresh elution buffer (1% SDS, 0.1 M sodium bicarbonate) were added and incubated 20 min at RT.

Decrosslink and DNA purification: 0.2 M final NaCl and 100 ng/ml final PK were added to the supernatant and to 10% input sample (30 µl of sonicated DNA before immunoprecipitation in 370 µl of elution buffer) and were incubated at 65 °C overnight (shaking at 1000rpm). DNA was purified in a MinElute column (MinElute purification kit Qiagen) by adding 5 volumes of PB buffer and centrifugation 1 min at 13000rpm (removing flow through), then eluted in 20 µl EB buffer (1 min at 37 °C, then 1 min at 13000rpm). The elution product was the final ChIP sample.

4.3.3.2 ChIP-seq libraries production and sequencing

End repair, 3' adenylation and adapters ligation: To repair DNA breaks and allow the adapters ligation, 7 µl End Repair & A-Tailing buffer, and 3 µl of End Repair & A-Tailing Enzyme Mix (Kapa Hyper Prep Kit) were added to 50 µl of fragmented double-stranded DNA (5 ng) and incubated 30 min at 20°C, then 30 min at 65 °C to inactivate the enzyme. To ligate the adapters, the 60 µl End Repair & A-Tailing reaction product were incubated at 20 °C for 15 min in the thermo-shacker with 30 µl of ligation buffer and 10 µl of DNA ligase (Kapa Hyper Prep Kit), 5 µl of nuclease free water and 5 µl of final 28 nM NEXTflex-96™ DNA Barcodes. The adapter ligation reaction product was then purified by incubating with 0.8 volume of Agencourt AMPure XP reagent at RT for 15 min, then beads were washed twice with 80% ethanol, resuspended in 22.5 µl of elution buffer (10 mM Tris-HCl, pH 8.0) and incubated for 2 min to elute the DNA from the beads (keeping supernatant).

PCR enrichment, DNA purification, size selection and sequencing: Libraries were amplified by 10 PCR cycles (15 sec at 98 °C – 30 sec at 60 °C – 30 sec at 72 °C) by adding to the 20 µl of Adapter-ligated library 25 µl of PCR mix (25 µl of 2X KAPA HiFi Hotstart Ready Mix and 5 µl of 10X Library Amplification Primer Mix). PCR product was purified in a MinElute column as described before. A final step of size selection was performed by loading the 20 µl of amplified libraries in an E-Gel™ iBase Power System (2% agarose, program 2) and running for 16 min and 30 seconds to collect the 300 bp band. Afterwards, a qPCR with specific primers was performed as a quality control and libraries size was controlled on the Bioanalyzer. The 6 libraries and the 2 input DNAs were PE sequenced in one HiSeq 3000 lane.

4.3.3.3 ChIP-seq data analyses

Mapping: PE reads were mapped to the *Sus scrofa* genome version 11.1.90 (obtained from the NCBI and released in December 2016) with `bwa mem` (bwa v.0.7.12-r1039) and the option `-M`. Resulting sam files were converted to the bam format with `samtools view -bS` (samtools v.1.3.1 (Li et al., 2009)), sorted with samtools sort and indexed with samtools index. PCR duplicates were removed with `samtools rmdup`.

Peak calling: The peaks were called using macs2 callpeak (MACS2 v2.1.1.20160309 (Feng et al., 2012)) with the options `-f BAMPE, -g 2.4e9, --keep-dup all` and `–q 0.01`.

4.3.4 Differential analyses

A differential analysis was performed to extract bin pairs that are significantly differentially connected between the two conditions (90 and 110 days of gestation), at three different resolutions (500, 200 and 40 kb). A method similar to the one described in (Lun and Smyth, 2015), with some adaptations, was used to perform this task. More precisely, a 3-step approach was used which consisted in three steps:

1) *Filtering step*, in which low count bin pairs were removed from the dataset: this step is used to leverage the effect of multiple testing correction (and improve the testing procedure power) by removing low count bin pairs that have a very low chance to be found differentially expressed. We choose to use a fixed threshold ($\tau = 30$, which corresponds to a minimum of 5 reads per sample on average) based on the total number of reads, across the 6 samples, associated to a given bin pair, to filter out irrelevant bin pairs from the differential analysis.

2) *Normalization step (inter-matrices normalization)*, to make the different matrices comparable. As stressed in Lun and Smyth (2015), contrary to RNA-seq data, a normalization based on the (potentially corrected) library size is not sufficient for Hi-C data. Indeed, the complexity of the protocol generally generates additional biases that result in trended differences between libraries (as visible on MA plots). To correct such biases, we used the method proposed in (Ballman et al., 2004) and implemented in the Bioconductor **R** package **csaw** (Lun and Smyth, 2016) that performs a non-linear normalization based on a fast loess algorithm. Gene and sample specific offsets were computed and incorporated in the Generalized Linear Model (GLM) described in the next step, to correct trended differences.

   The efficiency of the normalization was controlled using PCA and MA plots on pseudo counts ($\log_2$ transformed counts), before and after normalization.

3) *Differential analysis step*: this step was performed using a Generalized Linear Model (GLM) based on the Negative Binomial (NB) distribution with a condition (two-level factor: 90/110 days) fixed effect. The model was estimated with the implementation of the **R** package **edgeR** (Robinson et al., 2010) and log ratio tests were used to assess the significativity of the condition effect on each bin pair proximity. *p*-values were corrected using (Benjamini and Hochberg, 1995) procedure to control the False Discovery Rate.

4.3.5 Gene ontology (GO) analysis

The GO functional analysis was performed among the human homologs of genes mapped to the differential bin pairs. The enrichment of ontological categories was tested with the hypergeometric test

implemented in the **R** package **GOstats** v2.32.0 (Falcon and Gentleman, 2007) using the following ontologies: biological processes (BP), molecular functions (MF) and cellular components (CC). Ensembl gene IDs were converted to entrez gene IDs via the R package org.Hs.eg.db v3.1.2, and mapped to gene ontology through the R package GO.db v3.0.0.

The GO terms associated with the biological process hierarchy are sorted by their p values corrected for multiple testing (Benjamini–Hochberg correction (Benjamini and Hochberg, 1995)).

4.3.6 Integrative analysis with expression data

Expression data was obtained from a previous transcriptome study in pig using microarray probes (Voillet et al., 2014). For each probe ID of that study the following information was available:

1) The corresponding sequence of the probe
2) The average expression value that was measured with the probe in fetal muscle samples from Large White pigs at 90 days of development (log-normalized value)
3) The same for samples at 110 days of development.
4) Log fold change (logFC) of these expression values at 110 vs. 90 days. As the reference time point was 90 days, a positive logFC involves a higher expression in 110-days pigs.

As the transcriptome characterization -including the microarray design- was performed using a former version of the pig genome (Sscrofa 10.2)- we first proceeded to remap the probe sequences on the more recent 11.1 genome version in order to anchor expression data on reliable genomic positions. Stringent filtering steps were applied to keep only high quality hits (unique best hits with more than 30 matches).

To compare the average expression in A vs. B compartment we simply computed the mean expression value of all the probes in each compartment using bedtools map and considered the resulting distribution in A vs. B compartments. This was done separately for 90 and 110 days, using expression values and compartments from the same condition. To investigate the dynamic of expression in compartment-switching regions we considered the logFC expression values of the probes and split them into compartment-switching categories using bedtools: no switch, A to B, B to A. The distribution of all logFC in each category was then compared. Boxplots and statistical tests were carried on in R.

The same approach was used to compare logFC expression values in the regions that were identified by the differential analysis of Hi-C data. Since a same genomic region (bin) can be simultaneously involved in both a bin pair with a positive logFC proximity value and a bin pair with a negative logFC proximity value (for instance, a chr1 region that is both significantly closer from a chr2 region at 90 days and significantly closer from chr3 at 110 days) we chose to discard such regions. We therefore considered the distribution of probe logFC expression values in three categories of bins based on the Hi-C differential analysis: bin that were not involved in any significantly different bin pair, bins that were only found in bin pairs with negative logFC and bins that were only found in pairs with positive logFC.

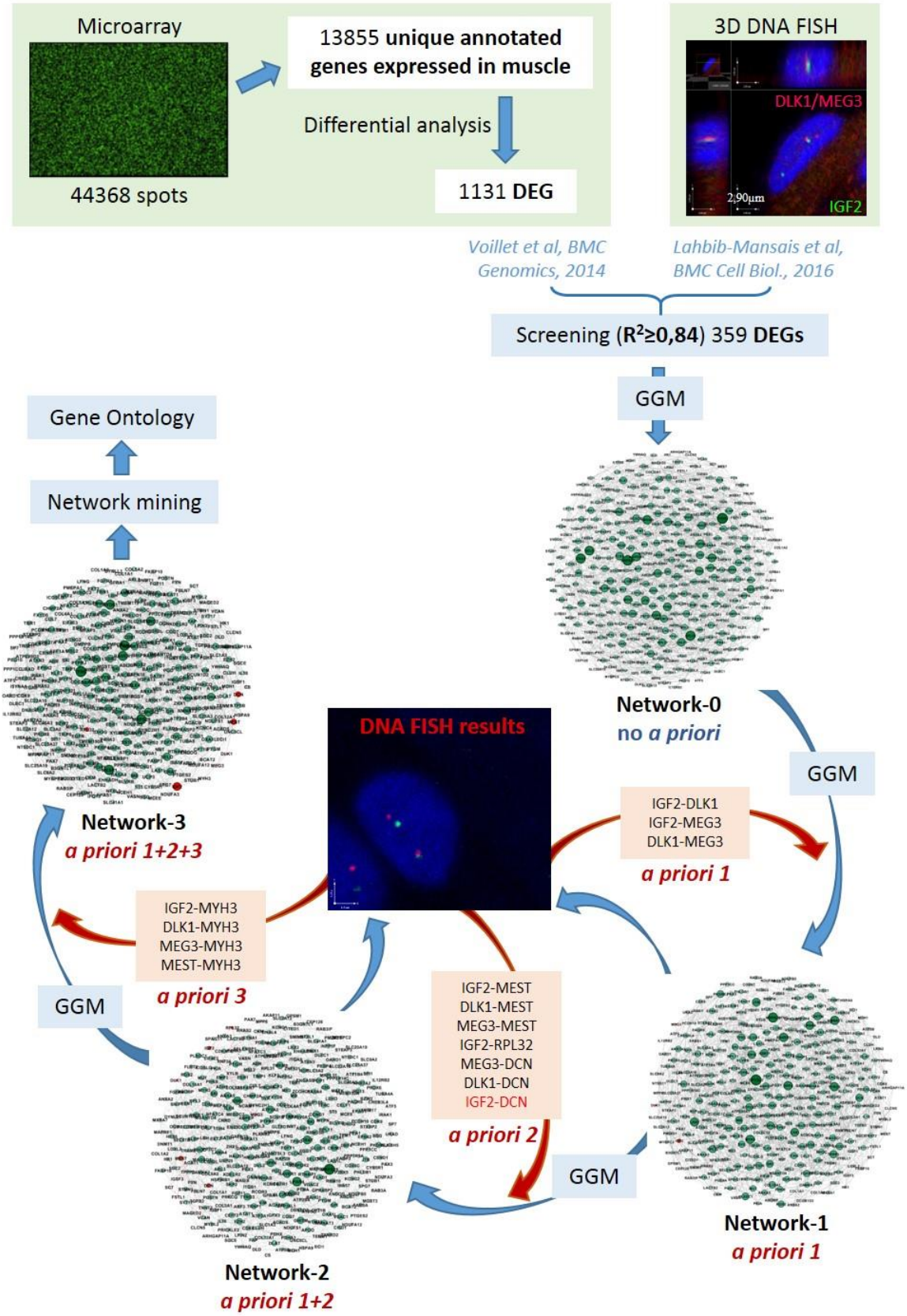


Figure 32. Experimental design. *Published data are represented in green squares (microarray data and 3D DNA FISH data), statistical methods are represented in blue (GGM: Gaussian Graphical Models) and new information about spatial localization used for network inference is represented in red.*

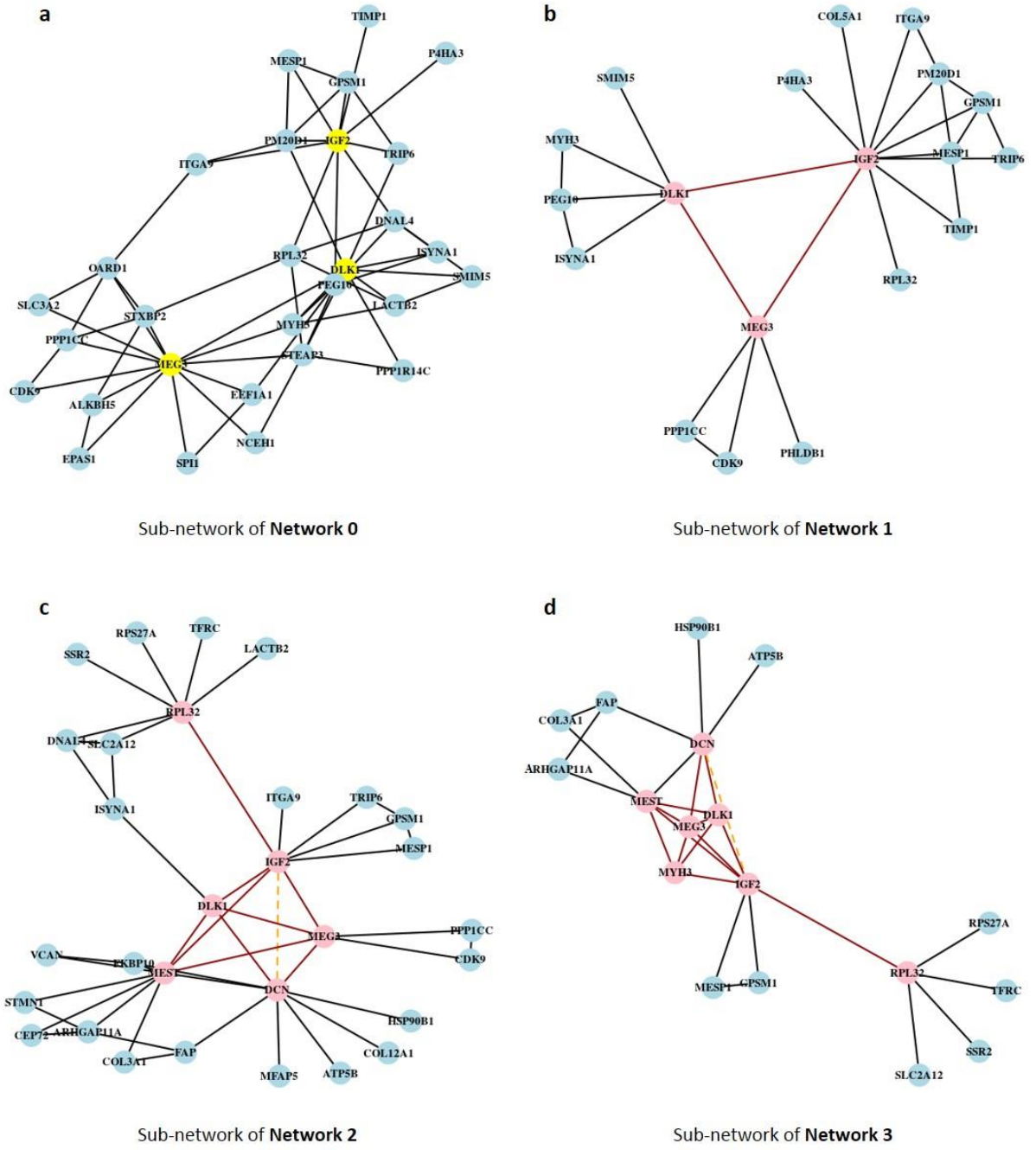


Figure 33. Analysis of gene associations. *Pink nodes represent target genes, red edges represent the known associations observed by 3D DNA FISH and the dotted orange edge represents the observed as not associated after 3D FISH validations. Because networks are very dense and contain many genes, a sub-network restricted to the target genes and their direct neighbors is extracted from each network, and presented in this figure. (a) Network 0 is inferred without a priori information, and restricted to the nodes corresponding to IGF2, DLK1 and MEG3 (in yellow). To infer Networks 1, 2 and 3, new a priori information of spatial localization is introduced for the following pairs of genes: (b) IGF2 DLK1, IGF2 MEG3 and DLK1 MEG3 for Network 1; (c) IGF2 MEST, (DLK1/MEG3) MEST, (DLK1/MEG3) DCN, RPL32 IGF2, IGF2 DCN for Network 2; (d) IGF2 MYH3, DLK1 MYH3, MEG3 MYH3 and MEST MYH3 for Network 3.*

5 Combining 3D DNA FISH and gene expression for network inference

The results presented in this study have been issue of a recent publication in the journal of Scientific Reports (Marti-Marimon et al., 2018). The whole article is provided at the end of this thesis (Apendix 21).

5.1 Results

5.1.1 Network inference iteration and 3D FISH validations

The whole process involving the data selection, the network inference and the 3D FISH validations is summarized in Figure 32. Network 0 was inferred with no *a priori* knowledge and contained 2,279 edges for 359 nodes (density: 3.55%). A sub-network extracted around the three target genes is shown in Figure 33a.

Network 1 was built based on the triple co-localization of *IGF2, DLK1 and MEG3* found in our previous study (Lahbib-Mansais et al., 2016). This *a priori* information was used to reinforce the existence of an edge between the pairs *IGF2-DLK1*, *IGF2-MEG3* and *DLK1-MEG3* in Network 1 (sub-network in Figure 33b), which contained 2,250 edges (density: 3.50%). In both graphs (Network 0 without *a priori* and Network 1 with *a priori*), we found a direct connection between the genes *IGF2* and *RPL32*. The *IGF2-RPL32* association was thus tested by 3D DNA FISH, because it involved one of our 3 initial target genes (*IGF2, DLK1* and *MEG3*), and because it was also found in the Imprinted Gene Network (IGN) of (Varrault et al., 2006). The 3D DNA FISH assay revealed that *IGF2* and *RPL32* were associated in 20% of the analyzed nuclei (Table 4, Figure 34a).

Additionally, we used 3D DNA FISH to analyze *MEST* and *DCN* associations with each of the three target genes, because they were also connected in the IGN (Table 4 and Figure 34b-e).

This new information about spatial co-localization in the nucleus was entered in our model as an *a priori* to build Network 2 (with 2,091 edges and 3.25% of density) (sub-network in Figure 33c). Specifically, in addition to the three pairs *IGF2-DLK1*, *IGF2-MEG3* and *DLK1-MEG3* given as associated in Network 1, we gave the following pairs of genes as known to be co-localized: *IGF2-MEST* (34% of analyzed nuclei presenting an association), (*DLK1/MEG3*)-*MEST* (in 34% of analyzed nuclei), (*DLK1/MEG3*)-*DCN* (in 15% of analyzed nuclei) and *RPL32-IGF2* (in 20% of analyzed nuclei). The pair *IGF2-DCN* was given as not co-localized (with 10% of nuclei presenting an association) (Table 4, Figure 34b-e). *DLK1* and *MEG3* are two imprinted genes located in the same cluster, and are both present in the same Bacterial Artificial Chromosome (BAC) used for the 3D DNA FISH experiments, because of their proximity on the genomic sequence (Appendix 4). Consequently, we considered

Table 4. Association percentages of tested gene pairs. *Associated signals (close + co localized) are considered as those separated by a 3D distance (d) ≤ 1 µm, and are divided into two different classes: "close" signals (0.5 < d ≤ 1 µm), and "co localized" signals (d ≤ 0.5 µm). * Genes imprinted in pig.*

Gene associations	Number of nuclei analysed	Percentage of nuclei with signals			
		Distant (d > 1 µm)	Close (0,5 < d ≤ 1 µm)	Co-localized (d < 0.5 µm)	Associated (d ≤ 1 µm)
MEST* - IGF2*	100	66	32	2	**34**
MEST* - (DLK1-MEG3)*	90	66	28	6	**34**
DCN - (DLK1-MEG3)*	73	85	15	0	**15**
RPL32 - IGF2*	80	80	16	4	**20**
DCN - IGF2*	98	90	7	3	**10**
IGF2* - MYH3	58	48	43	9	**52**
(DLK1-MEG3)* - MYH3	69	55	38	7	**45**
MEST* - MYH3	103	74	23	3	**26**
ZAR1 - IGF2*	61	92	8	0	**8**
ZAR1 - PRLR	63	92	8	0	**8**

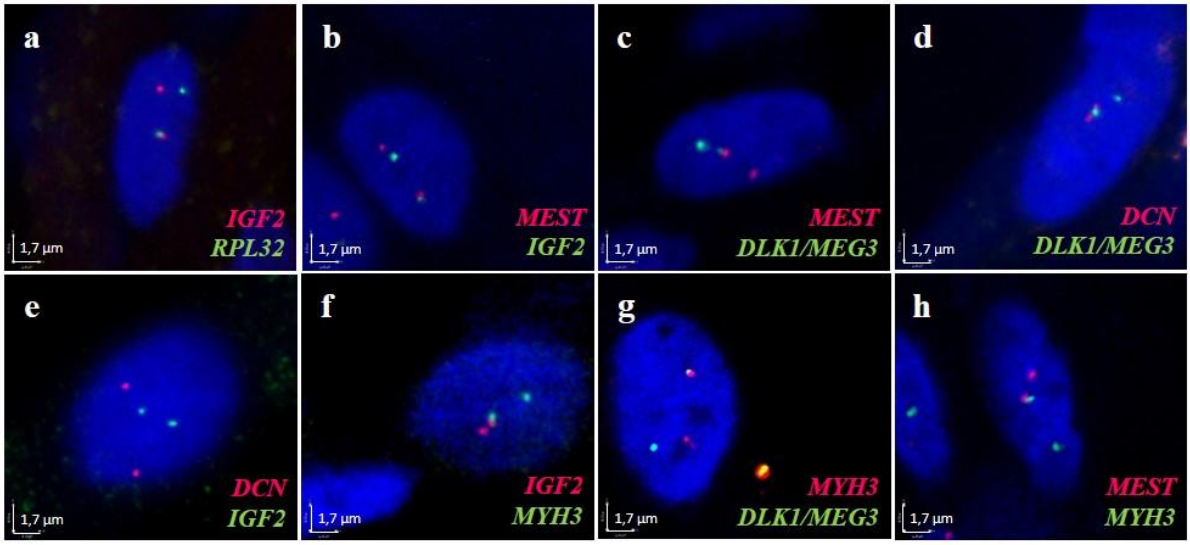


Figure 34. Analysis of gene associations by DNA FISH. *Extended focus of 3D image sections from confocal microscopy and overlay of the 3 channels (blue, red and green) were obtained with Volocity v6.0 software (Perkin Elmer). The four signals in the nuclei correspond to the two alleles of each gene. Nuclei are counterstained with DAPI (blue). In all experiments, the percentage of association between genes was higher than 10% except for (e). Scale = 1.7 µm.*

44

Table 5. Normalized mutual information (NMI) between pairs of clusterings. *NMI measure the similarity between two clusterings. The value is comprised between 0 and 1 and is equal to 1 when the two clusterings are identical.*

	Network 0	Network 1	Network 2	Network 3
Network 0	1	0.3893	0.3381	0.3244
Network 1	0.3893	1	0.4007	0.3923
Network 2	0.3381	0.4007	1	0.4152
Network 3	0.3244	0.3923	0.4152	1

DLK1/MEG3 as a simple locus for all 3D DNA FISH analyses, even though they are considered to be single genes for network inference.

To obtain the last network (Network 3), we used 3D DNA FISH to test for associations involving *MYH3* because it was found to be connected to *DLK1* and *MEG3* in Network 0 and to *DLK1* in Network 1. We found *MYH3* associated with i) *IGF2* in 52% of the analyzed nuclei, ii) *DLK1/MEG3* in 45% of the analyzed nuclei, and iii) *MEST* in 26% of the analyzed nuclei (Table 4, Figure 34f-h). Thus, in addition to the *a priori* information given in Networks 1 and 2, we gave the following new associations (*IGF2-MYH3*, *DLK1-MYH3*, *MEG3-MYH3* and *MEST-MYH3*) to infer Network 3 (2,091 edges, density = 3.25%) (Sub-network in Figure 33d).

5.1.2 Network mining (network structure with key genes)

For each network, two main numerical characteristics (degree and betweenness) were used to detect key genes with respect to the network structure. The degree of a node (in this case, of a gene) is the number of edges afferent to this gene. The betweenness of the node (gene) is the number of shortest paths between pairs of genes in the network that pass through that gene. High-degree genes are connected to many other genes while high-betweenness genes are central and more likely to disconnect the network if removed. We analyzed the evolution of the betweenness and degree from Network 0 to Network 3. Appendix 6 "Evolution of the betweenness and degree values of a subset of genes from Network 0 to Network 3" shows a subset of 25 genes selected as key genes for the network structure because they showed a high betweenness or a high degree value or both a high betweenness and a high degree, or because they were among genes whose associations tested positive with 3D DNA FISH. Most of the genes presenting the highest betweenness values in Network 0, still kept or increased this numerical characteristic in Network 3 after network inference iterations. However, important changes were observed in some genes. For instance, *AKR7A2, DLK1, EGFR*, *MEG3, MYH3 and RPL32*, showed more than a 40% decrease in betweenness accompanied by a decrease in degree (> 25%) when Network 3 was obtained. *DCN* showed a pronounced decrease in its degree while its betweenness was slightly modified. Interestingly, *MEST* and *IGF2* were found to have a mixed profile of betweenness and degree: in Network 3, we observed a 46% loss for *MEST* in gene connections, as compared to Network 0, while its betweenness increased by 160%. Similarly, a 30% loss of connections and a 426% gain in betweenness was observed for *IGF2*.

5.1.3 Network clustering

To analyze the evolution of the network structure from Network 0 to Network 3, clustering of the genes was performed on each network (for more details, see "Network mining and clustering" in "Materials and Methods" and Appendix 2 and 7 "Gene description and cluster allocation" and "Cluster parameters"). Four significant clusterings (p-value < 0.002) were obtained, one for each network. A total of nine clusters were obtained in Network 0, six in Network 1, eight in Network 2 and six in Network 3. Networks 0 and 3 were analyzed in depth to search for any correspondence between clusters (Appendix 8 "Pairwise contingency tables between clusterings"). Four clusters in Network 0 were found

Table 6. Comparison of GOBP in clusters 1 and 2 between Network 0 and Network 3. *GO terms enriched in one of the clusters as well as all GO terms associated to one of the three target genes at least (even if not significantly enriched). In bold, the smallest FDR value for a given GOBP term when the difference between the FDR of the two clusters is higher than one order of magnitude. Genes tested by 3D DNA FISH are in red bold.*

		Network 0 - Cluster 1		**Network 3 - Cluster 1**	
GO ID	GOBP Terms	Genes	FDR	Genes	FDR
43062	Extracellular structure	*POSTN, COL1A1, COL1A2, COL3A1, COL5A1, COL16A1, LAMA4, MFAP5*	5,76E-05	*POSTN, COL1A1, COL1A2,COL3A1, COL5A1, COL5A2, COL16A1,* ***DCN****, FAP, FBN1, ABI3 bp, ANXA2, LAMA4*	**1,14E-08**
71417	Cellular response to organonitrogen compound	*COL1A1, COL1A2,COL3A1, COL5A2, COL16A1, FYN, KLF3, ZFP36L1, HSP90B1*	**6,80E-04**	*COL1A1, COL1A2,COL3A1, COL5A2, COL16A1, DNMT1, FBN1,* ***IGF2****, HSP90B1*	1,16E-02
45995	Regulation of embryonic development	*COL5A1, COL5A2, FGFR1, LAMA4, LFNG*	2,24E-03	*COL5A1, COL5A2, FGFR1, LAMA4, LFNG*	1,16E-02
71559	Reponse to transforming growth factor beta	*POSTN, COL1A1, COL1A2, COL3A1, FYN, ZFP36L1*	**2,35E-03**	*POSTN, COL1A1, COL1A2,COL3A1, FBN1*	1,24E-01
44236	Multicellular organism metabolic process	*COL1A1, COL1A2,COL3A1, COL5A1, COL5A2*	2,35E-03	*COL1A1, COL1A2,COL3A1, COL5A1, COL5A2, FAP*	3,05E-03
43588	Skin development	*COL1A1, COL1A2, COL3A1, COL5A1, COL5A2, ZFP36L1*	**3,18E-03**	*COL1A1, COL1A2, COL3A1, COL5A1, COL5A2*	1,44E-01
1101	Reponse to acid chemical	*COL1A1, COL1A2,COL3A1, COL5A2, COL16A1, NFATC4*	1,17E-02	*COL1A1, COL1A2,COL3A1, COL5A2, COL16A1, DNMT1, NFATC4*	2,27E-02
1501	Skeletal system development	*POSTN, COL1A1, COL1A2, COL3A1, COL5A2, FGFR1, TMEM119*	1,43E-02	*POSTN, COL1A1, COL1A2, COL3A1, COL5A2, FBN1, FGFR1, ANXA2, TMEM119,* ***IGF2***	3,05E-03
		Network 0 - Cluster 2		**Network 3 - Cluster 2**	
72350	Tricarboxylic acid metabolic process	*CS, DLAT, DLD, NNT, MDH1, PDHA1*	3,02E-06	*CS, DLAT, DLD, NNT, MDH1, PDHA1*	2,11E-05
51186	Cofactor metabolic process	*COQ7, DLAT, DLD, NNT, HK1, ACACB, NMNAT3, ACAT1, MDH1, PDHA1, PDHX*	**2,97E-05**	*DLAT, DLD, IBA57, NNT, GPI, ACACB, NMNAT3, MDH1, PDHA1, FLAD1, MCEE*	1,34E-03
72524	Pyridine-containig compound metabolic process	*DLD, NNT, HK1, NMNAT3, MDH1, PDHA1, PDHX*	**1,00E-04**	*DLD, NNT, GPI, NMNAT3, MDH1, PDHA1*	1,11E-02
6631	Fatty acid metabolic process	*CPT1B, ECI1, DLAT, DLD, ACACB, ACADS, ACAT1, PDHA1, PTGES2, PDHX*	**1,00E-04**	*CPT1B, ECI1, DLAT, DLD, FABP3, ACACB, ACADS, PDHA1, ADIPOR2, PTGES2, MCEE*	1,17E-03
6091	Generation of precursor metabolites and energy	*CS, DLAT, DLD, NNT, HK1, MDH1, OXA1L, ATP5B, PDHA1, SLC25A3*	1,09E-04	*CS, DLAT, DLD, NNT, GPI, MDH1, NDUFA3, NDUFB5, NDUFS1, OXA1L, ATP5B, PDHA1, SLC25A3, CISD1, NDUFA12, PYGM*	**1,32E-07**
6090	Pyruvate metabolic process	*DLAT, DLD, HK1, PDHA1, PDHX*	5,42E-03	*DLAT, DLD, GPI, PDHA1, BSG*	2,32E-02
6790	Sulfur compound metabolic process	*VCAN,* ***DCN****, DLAT, DLD, ACACB, ACAT1, PDHA1, PDHX*	**7,47E-03**	*DLAT, DLD, IBA57, ACACB, PDHA1, MCEE*	4,79E-01
42180	Cellular ketone metabolic process	*COQ7, DLAT, DLD, ACACB, PDHA1, PDHX*	1,46E-02	*DLAT, DLD, FABP3, GPI, ACACB, PDHA1*	8,05E-02
45454	Cell redox homeostasis	*TXNRD2, DLD, NNT, PTGES2*	1,46E-02	*TXNRD2, DLD, NNT, PTGES2*	4,91E-02
44282	Small molecule catabolic process	*CPT1B, ECI1, DLD, HK1, ACACB, ACADS, ACAT1*	1,88E-02	*CPT1B, ECI1, DLD, GPI, ACACB, ACADS, BCAT2, MCEE*	4,51E-02
98656	Anion transmembrane transport	*CLCN5, CPT1B, ACACB, SLC25A3, SLC1A3, VDAC1*	**2,31E-02**	*CPT1B, ACACB, SLC25A3, SLC1A3, VDAC1*	3,77E-01
6081	Cellular aldehyde metabolic process	*DLAT, DLD, PDHA1, PDHX*	2,59E-02	*DLAT, DLD, GPI, PDHA1*	8,73E-02
43648	Dicarboxylic acid metabolic process	*DLD, NMNAT3, MDH1, SLC1A3*	3,13E-02	*DLD, NMNAT3, MDH1, BCAT2, SLC1A3*	2,13E-02
16042	Lipid catabolic process	*CPT1B, ECI1, ACACB, ACADS, ACAT1, NCEH1*	3,65E-02	*CPT1B, ECI1, FABP3, ACACB, ACADS, NCEH1, MCEE*	6,59E-02
10257	NADH dehydrogenase complex assembly			*NDUFA3, NDUFB5, NDUFS1, OXA1L, NDUFA12*	**3,29E-03**
97031	Mitochondrial respiratory chain complex I biogenesis			*NDUFA3, NDUFB5, NDUFS1, OXA1L, NDUFA12*	**3,29E-03**

to share at least two thirds of their nodes with the corresponding clusters in Network 3. More precisely, 64.1% of the genes in cluster 1, 68.4% in cluster 2, 66% in cluster 3 and 82.4% in cluster 4, were observed in the corresponding clusters of Network 3. The other clusters in Network 0 (clusters 5, 6, 7, 8 and 9) were mainly spread each into two different clusters of Network 3. Additionally, the Normalized Mutual Information (NMI) value was calculated to quantify the similarity between clusterings for pairs of networks (Table 5). Interestingly, we observed that the clustering obtained in Network 0 was the most similar to the clustering obtained in Network 1 (NMI = 0.389). Similarly, the clustering in Network 1 was the most similar to the one obtained in Network 2 (NMI = 0.401), and the clustering in Network 2 was the most similar to the one obtained in Network 3 (NMI = 0.401). This finding suggests that clusterings become more consistent when introducing new biological information in each network inference iteration.

5.1.4 Functional enrichment analysis

To test the biological relevance of each cluster in Networks 0 and 3, a functional enrichment analysis was performed for each cluster from both networks. Significant GO terms for Biological Processes (GOBP) were observed in clusters 1 and 2 of Networks 0 and 3, and in clusters 3, 5 and 8 of Network 0 (Table 6 and Appendix 9 "Comparison of GOBP between Network 0 and Network 3"). Table 6 shows the four clusters presenting the non-redundant GOBP with the smallest False Discovery Rate (FDR). When comparing cluster 1 in Networks 0 and 3, eight common enriched GO terms were observed, mainly involved in extracellular matrix formation, embryonic development, metabolic processes and cellular response to stimulus. Besides, fourteen common enriched GOs were observed in cluster 2 of Networks 0 and 3. These GO terms were mainly involved in cellular respiration, energy metabolism, cellular metabolic processes and metabolism of fatty acids. Additionally, two GO terms were observed only in cluster 2 of Network 3, both involved in the mitochondrial respiratory processes. Interestingly, the smallest FDR were observed in Network 3: (i) for cluster 1 (containing all genes tested by 3D DNA FISH), referring to the "Extracellular structure" term (involving the Decorin gene *(DCN)*; FDR = 1.14e-08); (ii) for cluster 2, referring to the "Generation of precursor metabolites and energy" term (FDR = 1.32e-07) (Table 6).

These results suggest that our approach to network inference by incorporating *a priori* biological information enables us to obtain relevant GO terms while conserving the functional enriched terms found in the initial network (Network 0). Moreover, we unexpectedly observed that two (*IGF2* and *DCN*) of our seven target genes showed more significant GO terms in Network 3 than in the initial network. Specifically, *IGF2* was observed to be uniquely involved in the "Genetic imprinting" term in cluster 3 of Network 0 (FDR = 3.82e-02), while in cluster 1 of Network 3 it was found to be involved in two new significant GO terms, the one with the smaller FDR being "Skeletal system development" (FDR = 3.05e-03) (Table 6 and Appendix 9 "Comparison of GOBP between Network 0 and Network 3"). *DCN* was in turn observed to be involved in the "Sulphur compound metabolic process" term (FDR = 7.47e-03) in cluster 2 of Network 0, while in cluster 1 of Network 3 it appeared to be involved in the "Extracellular structure" term presenting the smallest FDR value (1.14e-08) of all clusters. Concerning *MEST*, *MYH3* and *DLK1*, also tested by 3D DNA FISH, even though the observed FDR

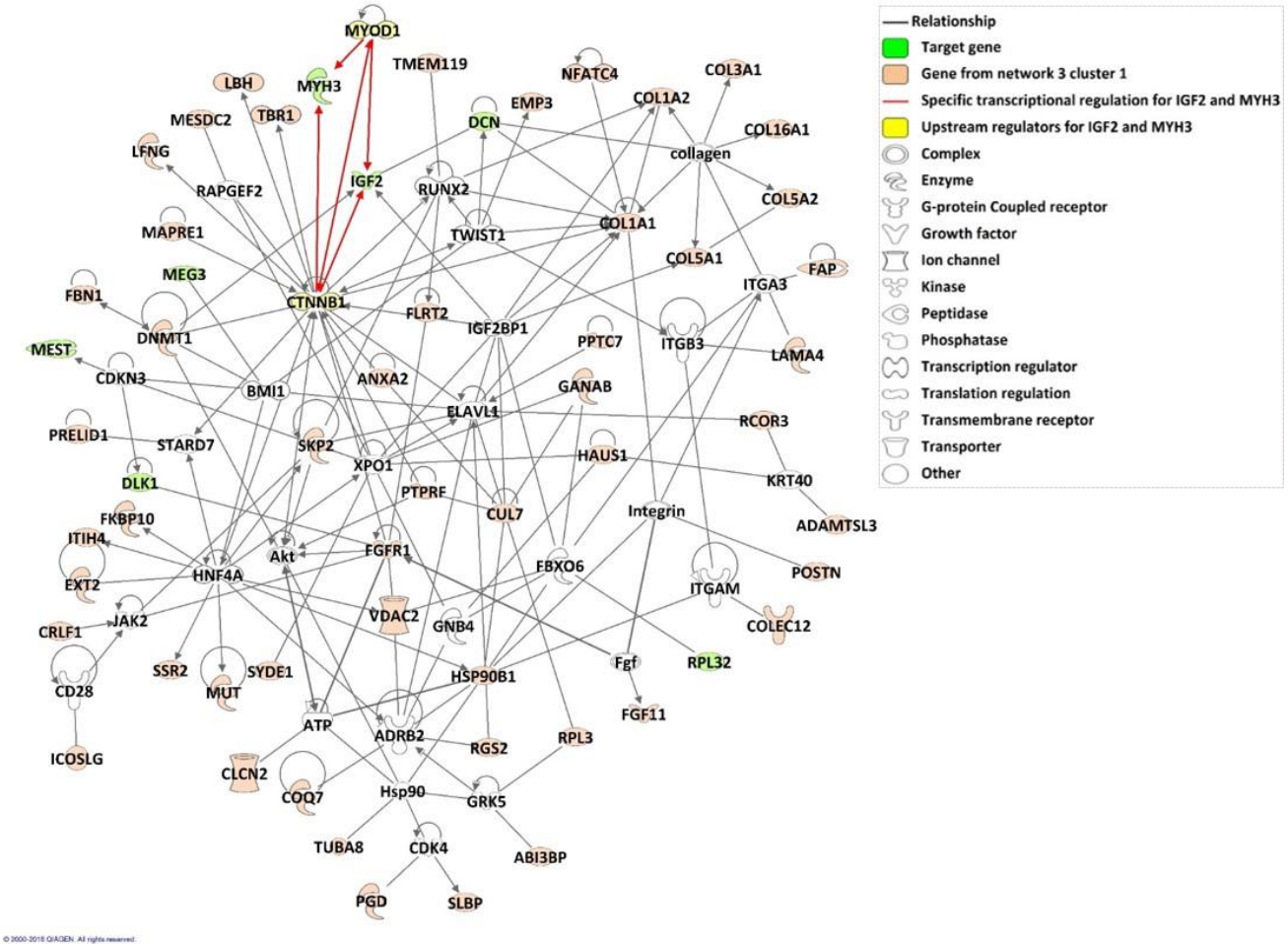


Figure 35. Reconstructed network of genes in cluster 1 of Network 3, based on Ingenuity Pathways Knowledge Base. *Nodes are displayed using various shapes that represent the functional class of the gene product. The reconstructed network was generated through the use of Ingenuity Pathway Analysis (IPA) (Ingenuity Systems; QIAGEN, Inc., Valencia, CA, USA).*

were higher than 5%, interesting GO terms were observed for these genes in cluster 1 of Network 3 (Appendix 9 "Comparison of GOBP between Network 0 and Network 3"). For instance, *MEST* was found to be involved in "Mesoderm development", *MYH3* in "Body morphogenesis", *DLK1* in "Notch signaling pathway" and *DCN* and *MYH3* were both found to be involved in "Muscle organ development".

Another functional analysis was performed with Ingenuity Pathway Analysis (IPA) specifically on cluster 1 of Network 3, which contains the target genes (*IGF2*, *DLK1*, *MEG3*, *RPL32*, *MEST*, *DCN* and *MYH3*). IPA proposed to connecting 49 (82%) out of 60 genes in a network including all target genes except *MEG3* and *MYH3*. *MYH3* was found in a small network with 8 out of 60 genes, and *MEG3* in another small network of only 1 out of 60 genes. Furthermore, *MYOD1* and *CTNNB1* were identified by upstream regulator analysis as potential transcriptional factors for a group of genes including *IGF2* and *MYH3*. As IPA offers the possibility of merging networks (if there are links between nodes in the Ingenuity Pathways Knowledge Base), a reconstructed network was obtained (Figure 35), and analyzed around the target genes. Fourteen genes, among them 7 genes from cluster 1 (including *DCN* and *IGF2*), were observed to be related to "Cell Morphology" (*p*-value = 1.75e-08). *DCN*, *DLK1* and *IGF2* were likewise involved in the "Quantity of cells" function with 31 genes, including 16 genes from cluster 1 (*p*-value = 2.48e-09). "Morphology of connective tissue cells" with 8 genes (*p*-value = 1.27e-04) included *DLK1* and *MEST*. "Formation of muscle", with 10 genes (*p*-value = 2.98e-05), involved *IGF2* and *MYH3* together with the two transcription factors *CTNNB1* and *MYOD1* (Appendix 3 "Biological network reconstructed following Ingenuity data analyses").

5.2 Discussion

We present here a new approach based on GGM that enables the user to introduce previously acquired biological knowledge to build gene co-expression networks. Since an observed correlation between two genes in the co-expressed gene network does not necessarily mean that these genes are related to a common biological process, we used information of gene nuclear co-localizations to reinforce observed links in the co-expressed gene network. Some studies have shown examples of co-expressed and co-localized genes being implicated in a particular process, e.g. the *Hbb* and *Hba* Klf1-regulated globin genes were found to be co-localized in specialized Klf1-enriched transcription factories of erythroid cells (Schoenfelder et al., 2010). Others have observed a role of co-expressed and co-localized genes in gene expression regulation, e.g. in the HUVECs endothelial cell line, *SAMD4A*, *TNFAIP2* and *SLC6A5* TNFα-induced genes were hierarchically transcribed when engaged in chromosomal interactions (Fanucchi et al., 2013).

In order to determine which pairs of genes would present a reinforced edge in the networks, we performed two negative controls (see "gene-gene associations" in the "Materials and Methods" section). As discussed in our previous study (Lahbib-Mansais et al., 2016), it can be difficult to define a suitable non-associating control. Sandhu *et al.* established a threshold of 2% (Sandhu et al., 2009), while others used the expected frequency of random co-localization based on the volume of the nucleus and

individual gene signals (<1%) (Osborne et al., 2004). This estimation of random co-localization does not take into account other constraints such as: (1) chromosomes occupy specific territories (Bolzer et al., 2005; Rieder et al., 2014); (2) transcriptionally silent domains reside at the nuclear periphery (Boyle et al., 2001); (3) chromatin regions are preferentially associated in topological domains (TADs) (Dixon et al., 2012). Fixing an arbitrary threshold of 10% was a more restrictive way of analysing co-expressed genes that might tend to interact preferentially. Consequently, the pair *IGF2-DCN* was given as not co-localized by enforcing the absence of an edge between both genes.

Testing the nuclear co-localization of *IGF2* and *RPL32* by 3D DNA FISH proved interesting, as this connection concerned an imprinted gene (*IGF2*, involved in muscle growth-related traits (Van Laere et al., 2003)) and a ribosomal protein coding gene *RPL32* (Young and Trowsdale, 1985). This experiment revealed that these genes are associated. Additionally, it was interesting to find co-localized pairs of genes such as *IGF2-MEST*, *(DLK1/MEG3)-MEST*, *(DLK1/MEG3)-DCN*, that were observed to be connected in co-expression networks in other studies (Al Adhami et al., 2015; Varrault et al., 2006), even though they were not directly connected by an edge in our network (Network 1) but via intermediary genes. Besides, surprising results showed the highest association we have ever observed between two genes (neither in the present study, nor in previous ones). This association concerns *MYH3* and *IGF2*. *MYH3* plays an important role in foetal muscle development (Schiaffino et al., 2015; Voillet et al., 2018), and encodes for the embryonic Myosin Heavy Chain (MYHC) 3 protein. To the best of our knowledge, no previous association between these two genes, whatever its origin (nuclear or functional), has ever been observed, even though the two genes are known to be involved in muscle development (Livingstone and Borai, 2014; Schiaffino et al., 2015). To determine the impact of the *a priori* co-localization information introduced to enforce the presence or the absence of an edge, we analysed the evolution from Network 0 to Network 3, first globally (with conserved edges and key genes) and then locally (with network clustering and functional enrichment). The global analyses revealed that 82% of edges in Network 0 were conserved in Network 3 and that the most important genes (with respect to network structure) in Network 0 were among those showing the highest values of betweenness and degree in Network 3. These findings suggest that the introduction of enforced edges is not linked to the appearance of major disturbances in the network structure. However, when focusing on the target genes analysed by 3D DNA FISH, we observed a general decrease in the degree value, meaning that *IGF2*, *DLK1*, *MEG3*, *RPL32*, *MEST*, *DCN* and *MYH3* were less connected with the rest of the other genes in Network 3. Despite this observed isolation concerning genes for which edges were enforced, this effect was not always accompanied by a loss of betweenness. In other words, reinforcing a limited number of edges did not change either the global network structure or the importance of target genes in the final network. In the local analysis, the NMI value revealed that the clusters resembled one another more with each new network inferred. In addition, four out of six clusters in the final network (Network 3) conserved more than 62% of genes in the corresponding clusters of Network 0. This concurred with the results of the functional enrichment analysis, which revealed that the GOs found were conserved between Networks 0 and 3. All these results support the evidence that our approach did not introduce any substantial disturbance. In fact, this iterative process brought substantial improvements; notably, it enabled us to obtain reliable networks in terms of relevant biological information, especially around our target genes. This was supported by the following findings: (1) the biological processes presenting the

smallest FDR were found in Network 3, even though one of them involved *DCN,* for which edge estimations were modified by the introduction of *a priori* information; (2) two new significant GO terms related to energy metabolism appeared in cluster 2 of Network 3; (3) two genes (*IGF2* and *DCN*) analysed by 3D DNA FISH were involved in biological processes with smaller FDR in Network 3 than in Network 0. Moreover, *IGF2* was found in an additional GO of Network 3, while only present in one GO of Network 0.

One of the most important goals of the present article was to elucidate the mechanisms that govern porcine skeletal muscle development in late gestation. Many studies have been performed in pig to address this question (Cagnazzo et al., 2006; Tang et al., 2015a; Voillet et al., 2014; Xu et al., 2012; Zhao et al., 2011, 2015). In our model, we proposed a final network (Network 3) in which enriched biological functions related to muscle development were observed. These observations were in agreement with the results obtained by Voillet *et al.*(Voillet et al., 2014). In addition, in the resulting IPA reconstructed network, we highlighted *MYOD1* and *CTNNB1* among the proposed transcription factors because they were especially interesting due to their connection to two important target genes, *IGF2* and *MYH3*. Although *MYOD1* and *CTNNB1* were not present in the 359 genes used for network inference, they were up-regulated at 90 days of gestation in all genotypes (Appendix "Gene expression profiles") (Voillet et al., 2014). *MYOD1* encodes for a myogenic factor that regulates skeletal muscle cell differentiation by activating transcription of muscle-specific target genes (for review (Berkes and Tapscott, 2005)). *CTNNB1* (β-catenin 1), encodes for a transcriptional co-activator that was found to be required for muscle differentiation in murine myoblasts by interacting directly with MyoD and promoting its binding to the E box elements enhancing its transcriptional activity(Kim et al., 2008). The co-expression and nuclear co-localization of *IGF2* and *MYH3* suggest they are each subjected to similar transcriptional regulation by these two transcription factors. The studies of (Shang et al., 2007) and (Ramazzotti et al., 2016) are in agreement with this hypothesis. Shang *et al.* revealed that in mesenchymal stromal cells from rats, an ectopic expression of *Ctnnb1* inhibits adipogenetic differentiation and induces the formation of long multinucleated cells expressing myogenic genes, such as *MyoD* and *Myhc,* by promoting the expression of skeletal muscle-specific transcription factors. Ramazzotti *et al.* observed that an overexpression and accumulation of β-catenin in the nuclei of differentiating murine myoblasts results in higher *MyoD* activation and *Myhc* induction. Additionally, *IGF2* was found to be up-regulated in pig during myogenesis and, more precisely, involved in primary and secondary muscle fibre differentiation (Zhao et al., 2011). Moreover, *Myod* and *Igf2* were observed to be involved in the switch between myogenic and adipose lineages in mouse (Borensztein et al., 2012). In addition, we found *IGF2* indirectly associated with *CTNNB1* (through the intermediary gene *IGF2 bp1*) in the reconstructed network. *IGF2 bp1* was not used for network inference but was found expressed at the 90th day of gestation (Appendix 10 "Gene expression profiles") (Voillet et al., 2014). Indeed, β-catenin was observed to induce *IGF2 bp1* in HEK293 cells (Noubissi et al., 2006), which in turn was observed to regulate *IGF2* mRNA subcellular location and translation in neurons (for review (Bell et al., 2013)). This suggests that in muscle cells, a similar mechanism could possibly be involved for the regulation of *IGF2* via the CTNNB1 transcription factor. Moreover, the long non-coding DNA of *MyoD* (*lncMyoD*), directly activated by MyoD, may negatively regulate *Igf2 bp1*-mediated translation of proliferation genes in murine myoblasts (Gong et al., 2015). This could explain how MyoD blocks

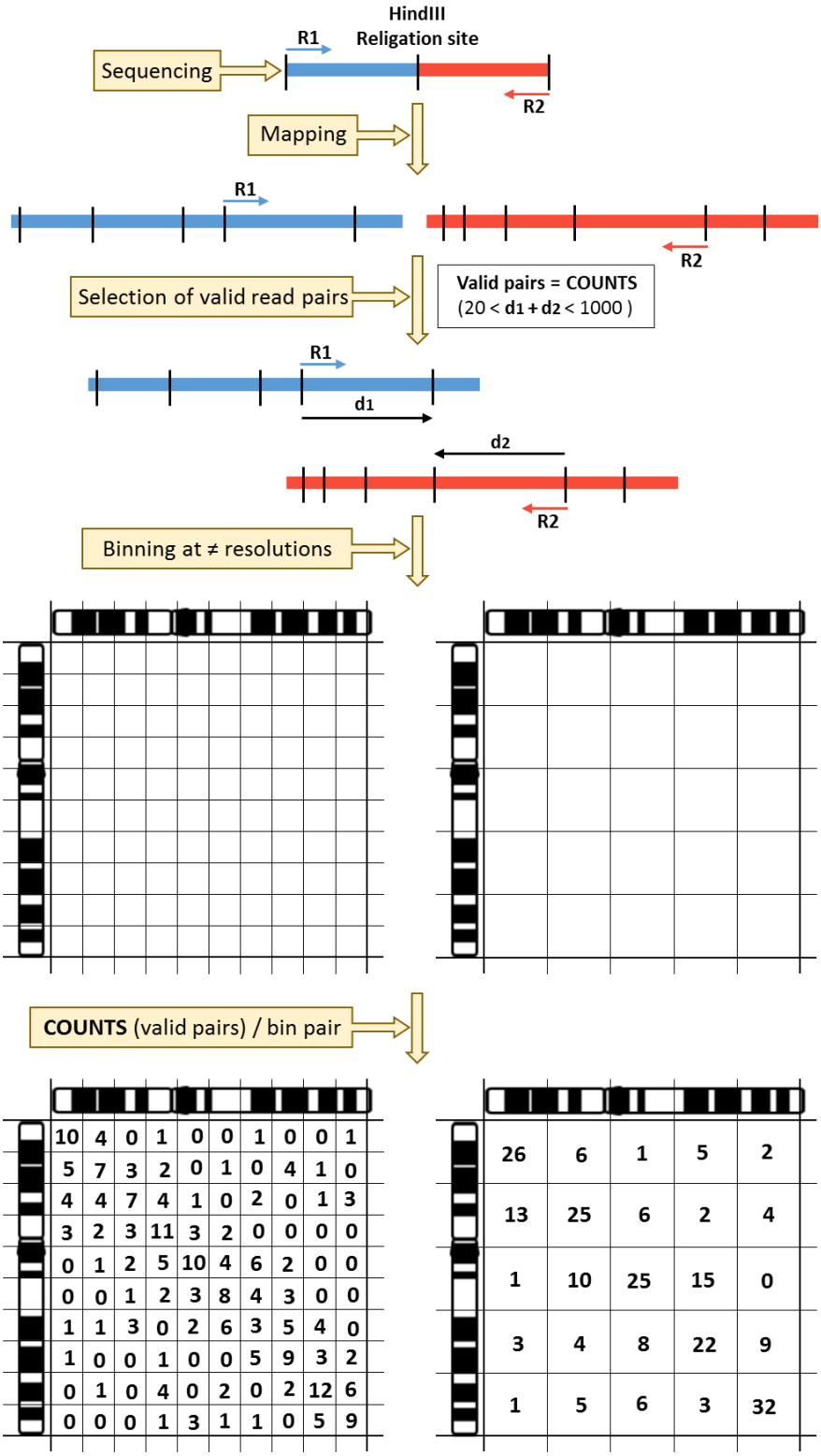


Figure 36. Summary of the main steps in data analysis. *The Hi-C hybrid fragment (result of a religation event) is paired-end (PE) sequenced (first line). Read1 (R1) and read2 (R2) are mapped to the reference genome (second line). Valid read pairs with an estimated insert size between 20 bp and 1 Kb (sum of distances from each read pair to their closest downstream HindIII genomic site) are kept, the others are discarded. The genome is segmented into genomic intervals of an equal number of bases (binning), the so-called "bins". Matrices are obtained at different resolutions (bin size) by adding the number of valid pairs (counts) per bin pair. The numbers in the matrices were made up for illustrative purpose and do not come from real data.*

proliferation to create a permissive state of differentiation. Moreover, *DLK1* and *MYOD1* were not connected in the reconstructed network. However, *DLK1* which encodes for a preadipocyte factor that inhibits adipocyte differentiation(Wang et al., 2010), might inhibit cell proliferation and enhance cell differentiation by regulating the expression of *MyoD* (Waddell et al., 2010). Combining all this information with the observed up-regulation at 90 days of gestation of the above-mentioned genes, our results highlight a network of interrelated genes associated with skeletal muscle regulation and that are mainly responsible for inhibition of proliferation and muscle differentiation.

A

	Total pairs	Mapped pairs	Valid pairs
Rep1-90	654,909,642	453,511,866	283,532,165
Rep2-90	685,369,147	461,032,971	270,028,254
Rep3-90	489,704,084	341,513,612	204,043,490
Mean-90	609,994,291	418,686,150	252,534,636
Rep1-110	494,332,893	362,904,795	233,274,518
Rep2-110	476,204,172	302,068,031	122,766,900
Rep3-110	646,908,804	446,570,196	272,611,287
Mean-110	539,148,623	370,514,341	209,550,902
Mean (all)	574,571,457	394,600,245	231,042,769
Total (all)	3,447,428,742	2,367,601,471	1,386,256,614

B

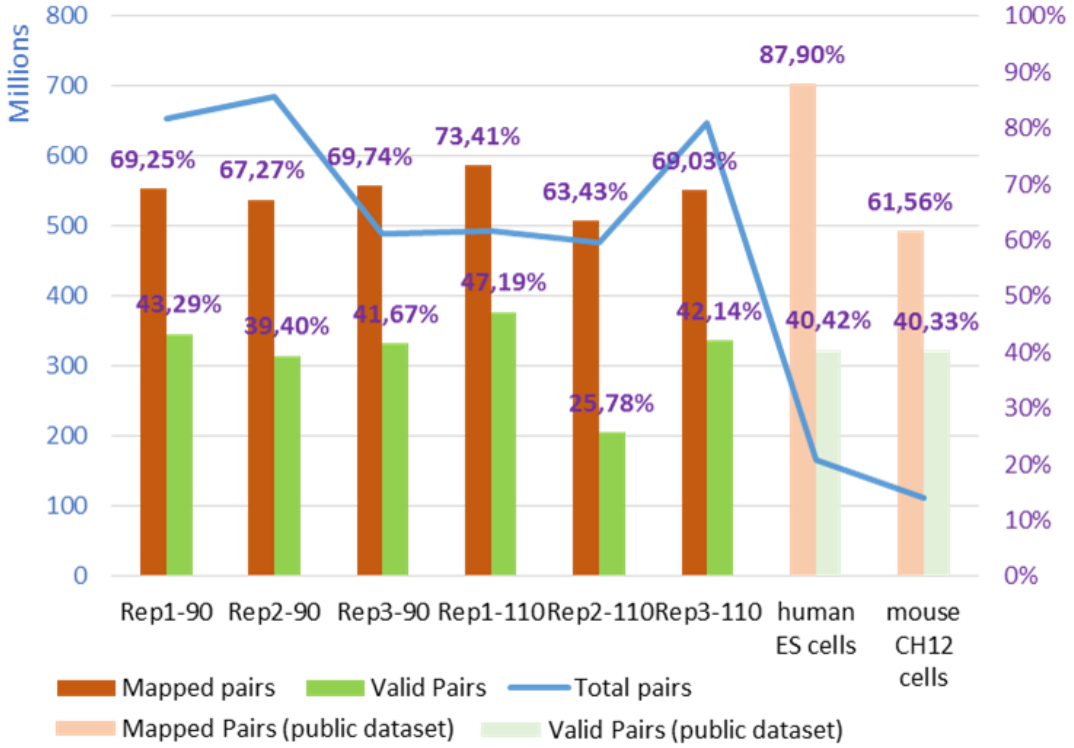


Figure 37. Hi-C read pairs statistics summary of the full dataset mapped to Sscrofa11. *(A) Total pairs: total number of sequenced read pairs. Valid pairs: uniquely mapped read pairs used to build the interaction matrices. The valid pairs have an estimated insert size between 20 bp and 1 Kb (sum of distances from each read pair to their closest downstream HindIII genomic site). (B) Percentage of mapped and valid pairs over the initial read pairs for each library. Two data sets (light colors) obtained from Hi-C assays performed in human ES cells* (Dixon et al., 2015) *and mouse CH12 cells* (Rao et al., 2014) *were analyzed with our bioinformatics pipeline, together with our six data sets (dark colors).*

6 Global genome organization assessed by Hi-C and gene expression

6.1 Results

6.1.1 Descriptive analysis of genome global organization in fetal muscle by Hi-C

In order to assess the 3D genome organization in fetal muscle, six Hi-C libraries (three per condition: 90 days and 110 days of gestation), called Rep1-90, Rep2-90, Rep3-90, Rep1-110, Rep2-110 and Rep3-110, were sequenced in two batches. First, an initial sequencing run was performed on 4 lanes of a HiSeq3000 to estimate the level of resolution that could be achieved in practice with these libraries. This first set was analyzed using the assembly version Sscrofa10 of the pig genome that was available at the time. Since a more recent version of the reference genome came out during the study, we reanalyzed this first dataset on the Sscrofa11 version. Later, in order to achieve a better resolution, we re-sequenced the same libraries over six new lanes. This full dataset was analyzed on the most recent assembly version Sscrofa11 and most of the results we present come from this entire set of data. However, in order to estimate the effect of the reference genome assembly on the analysis, we will sometimes compare results from the initial subset on both assembly versions. We will refer to these datasets as "subset_v10" and "subset_v11" in the text. In addition, in order to validate both our data and the analysis pipeline, we downloaded two public datasets from Hi-C assays performed in human and mouse cells (GEO Accessions SRR1030718 and SRR1658732 respectively) (Dixon et al., 2015; Rao et al., 2014). These two datasets, hereafter referred as "human ES cells" and "mouse CH12 cells" have been analyzed in parallel using their respective reference genomes (GRCh38 and GRCm38) as a control. The main steps of the analysis are summarized in Figure 36 (from raw data to matrix construction) as a remainder of the bioinformatics pipeline used to process Hi-C data (see Material and methods for more details).

6.1.1.1 Read statistics

6.1.1.1.1 Mapped pairs

Between ~ 476 M and 685 M read pairs were obtained per library after the two runs of sequencing (Figure 37A), which represents a total of ~ 3.45 billion of sequenced reads for the entire experiments. Around 63% - 73% of the read pairs could be mapped to the reference genome (Sscrofa11) (Figure 37B). All replicates showed a similar mapping ratio except for Rep2-110 which showed lower mapping rates. These mapping rates are lower than usually reported for human and mouse (Rao et al., 2014), as expected due to the lower quality of the porcine reference genome in terms of completion and assembly. Processing public datasets from human and mouse studies with our pipeline led to results in line with the literature (Figure 37B).

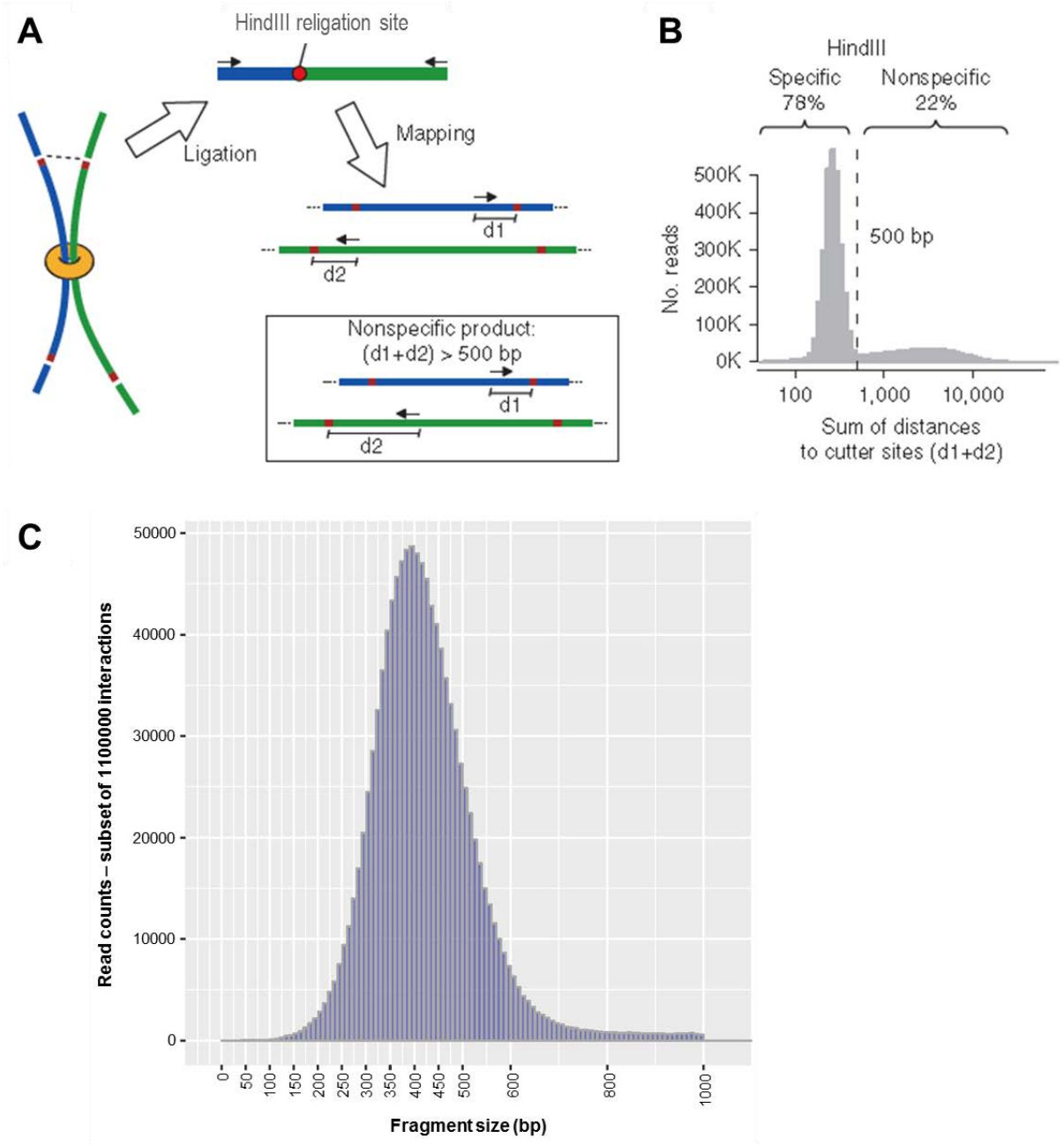


Figure 38. Selection of "valid read pairs" issue of a Hi-C religation event. *Upper panel (A and B): Images adapted from* (Yaffe and Tanay, 2011) *to illustrate how spurious ligation products are filtered out. (A) Hi-C ligation products are expected to map near restriction sites. The sum of distances from mapped Hi-C sequences to the nearest restriction sites (d1 + d2) is computed for each Hi-C paired read. (B) Then, the distribution of this sum of distances is reconstructed. (C) Example of the distribution of the sum of the distances from the reads to their closest downstream HindIII sites for Rep2-110. This distribution is consistent with the expected size of the fragments to be sequenced after the experimental step of size selection (see Materials and methods).*

6.1.1.1.2 Valid pairs

The next step of the Hi-C data analysis is to obtain, among the total read pairs, the so-called "valid pairs". The bioinformatics pipeline based on HiC-Pro (Servant et al., 2015) applies a filter to remove read pairs that do not have a mapping configuration consistent with a Hi-C religation event (see Materials and Methods). More precisely, since the Hi-C protocol aims to sequence chimeric fragments that contain a religation site, each read is expected to be mapped close to a HindIII restriction site on the genome. In addition, for a given read pair, the sum of the distances from the reads to their closest HindIII sites downstream (d1 + d2) should correspond to the size of the sequenced fragment (Figure 38A-B). The distribution of this sum is computed across all read pairs as a quality control, and pairs with extreme values (min and max threshold of 20 bp and 1 Kb respectively) are discarded (Figure 38C). This distribution corresponds to the insert size of the libraries (420- 520 bp), which was experimentally estimated after subtracting 120 bp (size of the adapter sequences) to the observed size (540 – 640 bp) of the libraries that was measured with the Fragment Analyzer (see Materials and Methods, Figure 27).

After applying this filter, between 122 M and 283 M valid pairs were obtained per library on the genome version Sscrofa11 (Figure 37A). This corresponds to ~ 26% – 47% of the total pairs (Figure 37B). Globally, libraries showed a good ratio (> 50%) of valid/mapped pairs, except Rep2-110 for which both mapping rates and proportion of valid pairs were lower compared to the other libraries. Despite this decrease on valid/mapped pair ratio, we still kept the Rep2-110 library because the fragment size distribution estimated from the valid pairs was consistent with a Hi-C religation event (Figure 38C), which supported the quality of the filtered data. The valid/mapped pair ratio obtained in pig was higher than the one observed in human, except for Rep2-110, meaning that our Hi-C libraries were generally more enriched in valid pairs than the data set obtained from human (Figure 37B). This could be the results of small variations in the experimental protocols for instance, or of intrinsic differences due to tissue and/or species specificities.

As mentioned above, a subset of the data was both analyzed on the 10 and on the 11 version of the pig genome. The results of these analyses are summarized in Figure 39. Running the pipeline by using the Sscrofa11 genome version allowed a ~ 8% - 10% increase on the mapping rates compared with Sscrofa10. These resulted in a ~2% - 4% increase on valid pairs, which is consistent with the considerable improvement on the sequence completion and assembly of the more recent genome version.

6.1.1.1.3 *Cis* and *trans* valid pairs

The valid pairs were then classified into *cis* and *trans* pairs depending on whether reads from the same pair mapped to the same or to different chromosomes respectively. Around 41% - 56% were classified into *trans* pairs and ~48% - 59% into *cis* pairs. Last, *cis* pairs were divided into short-range pairs (genomic distance within mapped reads ≤ 20 Kb; ~ 0.9% - 3.8%) and long-range pairs (genomic distance within mapped reads > 20 Kb; ~ 43% - 56%). These results were obtained using the Sscrofa11 genome version (Figure 40). Compared with results on the previous genome version, this represents ~4.5% less *trans* pairs (Figure 41). This difference could be explained by the assembly improvement. Indeed, read pairs with one read on a chromosome and the other one on an unplaced scaffold are

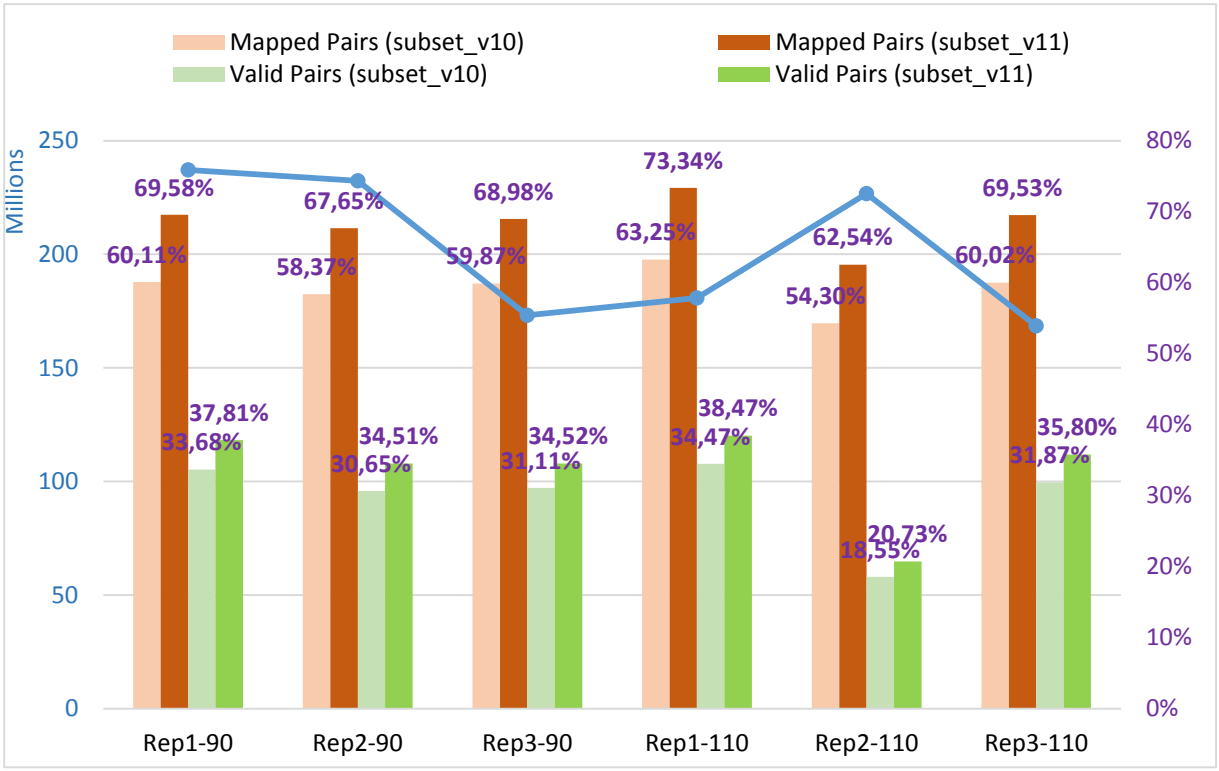


Figure 39. Results from a subset of the data on the previous genome version (Sscrofa10) and on the current genome version (Sscrofa11). *Comparison of the percentages of mapped and valid pairs obtained with the two assembly versions.*

A

	All cis valid pairs	Short-range cis valid pairs	Long-range cis valid pairs	Trans valid pairs
Rep1-90	154,656,274	7,258,570	147,397,704	128,875,891
Rep2-90	130,188,146	6,203,510	123,984,636	139,840,108
Rep3-90	120,983,843	7,721,962	113,261,881	83,059,647
Mean-90	135,276,088	7,061,347	128,214,740	117,258,549
Rep1-110	102,606,990	2,080,789	100,526,201	130,667,528
Rep2-110	659,34,998	3,241,306	62,693,692	56,831,902
Rep3-110	143,013,698	5,746,690	137,267,008	129,597,589
Mean-110	103,851,895	3,689,595	100,162,300	105,699,006
Mean (all)	119,563,992	5,375,471	114,188,520	111,478,778
Total (all)	717,383,949	32,252,827	685,131,122	668,872,665

B

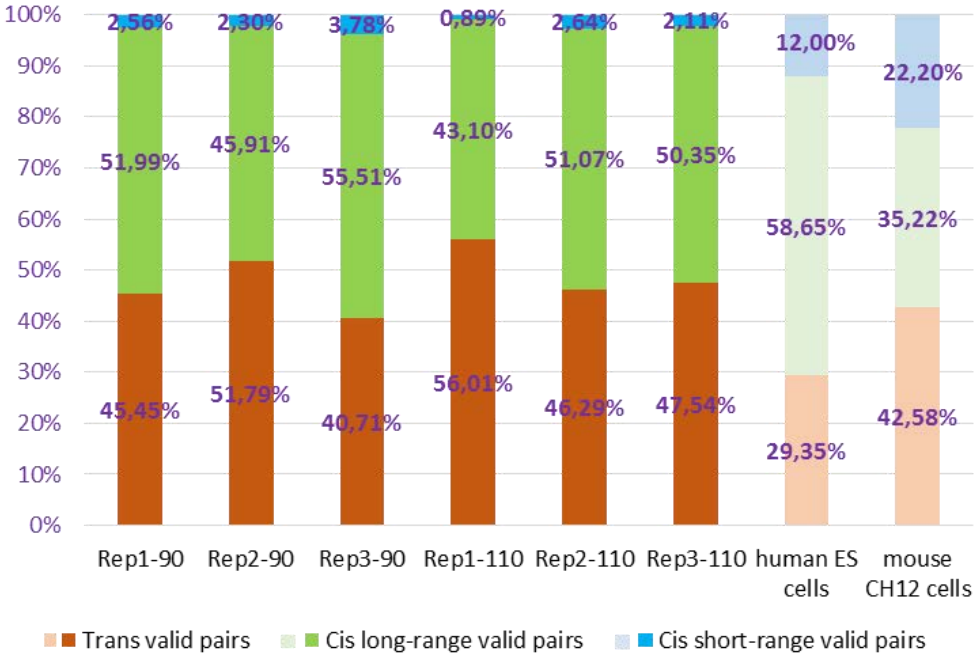


Figure 40. Valid read pairs per category after mapping the full dataset of reads on the Sscrofa11 genome version. *(A) Cis valid pairs: pairs with reads on the same chromosome (short-range: separated by a genomic distance ≤ 20 Kb; long-range: distance > 20 Kb). Trans valid pairs: pairs with reads on different chromosomes. (B) Percentage of valid pairs per category of our six datasets (dark colors). Results from two public datasets from human ES cells* (Dixon et al., 2015) *and mouse CH12 cells* (Rao et al., 2014)*, analyzed with our bioinformatics pipeline, are also shown (light colors).*

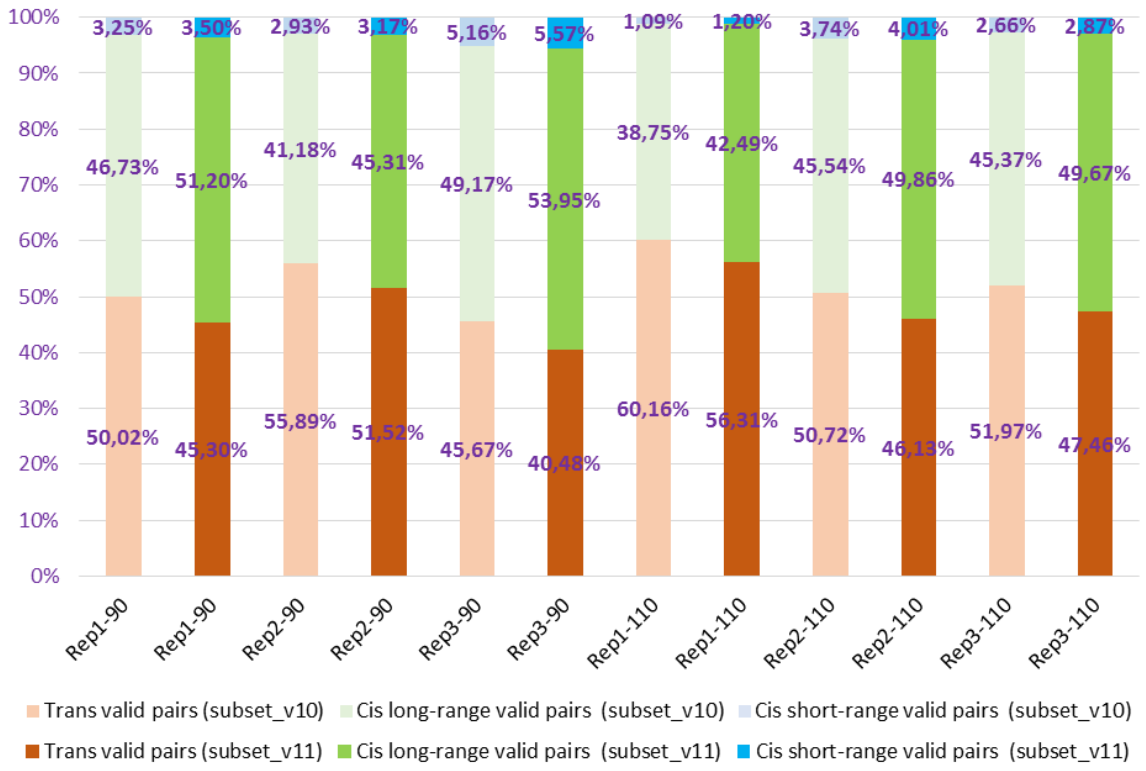


Figure 41. Results from a subset of the data on the previous genome version (Sscrofa10) and on the current genome version (Sscrofa11). *Comparison of the percentage of valid read pairs per category obtained with the two assembly versions.*

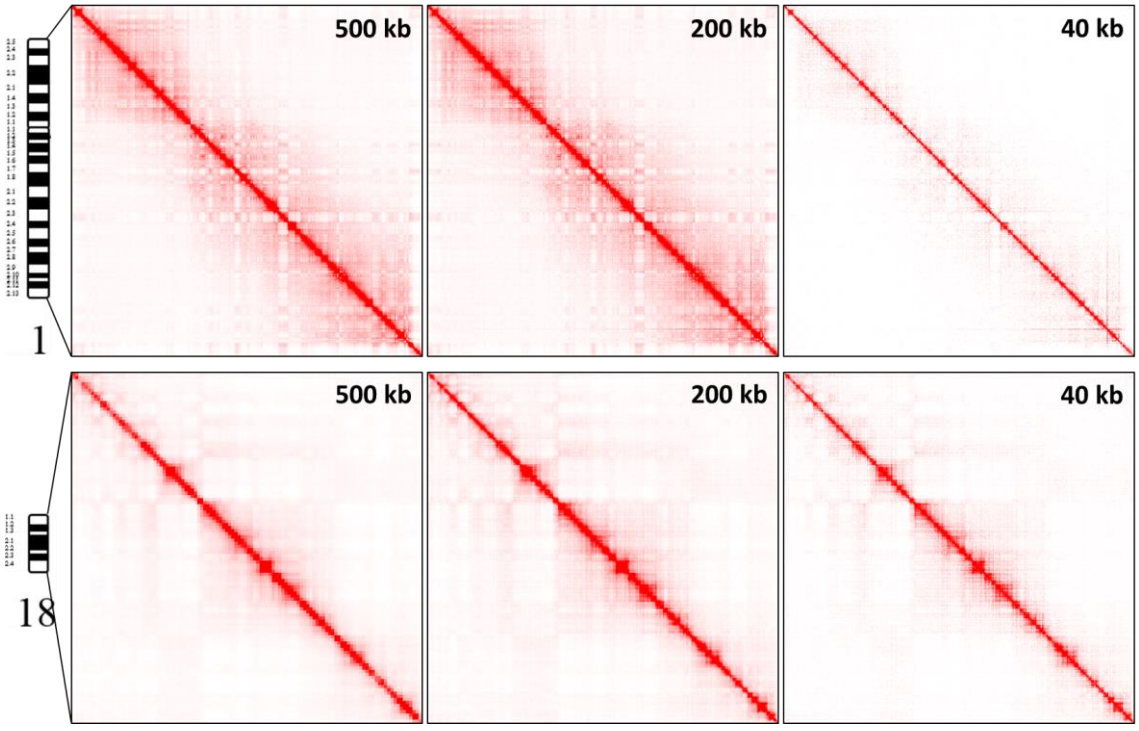


Figure 42. Hi-C contact matrices at different resolutions. *Illustration of Hi-C contact matrices of chromosomes 1 and 18 obtained at 500, 200 and 40 Kb resolution (Rep1-90). The darker the red color, the more contacts (valid read pairs) present in a pair of bins. Visualizations made with the HiTC R library.*

automatically classified as *trans* pairs. Consequently, if the number of unplaced scaffolds decreases between assembly versions, the number of *trans* pairs should decrease too. In fact, the Sscrofa10 version contains 4,562 unplaced scaffolds (7.54% of the genome; genome size ~ 2.8 Gb), vs. 583 unplaced scaffolds in the 11 version (2.65% of the genome; genome size ~ 2.5 Gb). Accordingly, the proportion of valid pairs that involved unplaced scaffolds (*i.e.* with at least one read mapped on a scaffold) dropped from 11.7% to 4.6% between versions, which can explain part of the drop in *trans* pairs.

Interestingly, we observed higher percentages of *trans* pairs in the porcine libraries than in the human one. Moreover, the percentage of *cis* short-range pairs were lower in pig than in human or mouse (Figure 40B). These unexpected results could be in part explained by differences on the assembly of the human and pig genomes, similarly to the differences observed between Sscrofa10 and Sscrofa11 as presented above. The human reference genome (GRCh38) contains 169 unplaced scaffolds (0.37% of the genome; genome size ~ 3.1Gb), compared with the 583 unplaced scaffolds in Sscrofa11 (2.65% of the genome). However, even if a small proportion of the increased values of *trans* pairs reported in fetal muscle pig can be explained by the quality of the reference genome, the percentages of *trans* pairs remain still high. This means that most of the differences in the *cis/trans* pair ratio underline specificities regarding the genome organization of the biological material (fetal pig muscle tissue vs. human embryonic stem cells).

6.1.1.2 Construction of genome-wide contact maps

The next step of the analysis is the generation of the contact matrices using valid read pairs. For that purpose, the genome was segmented into intervals of equal size (number of bases) called bins. To explore the data at different resolutions, we generated the matrices using several bin sizes (500 Kb, 200 Kb and 40 Kb). The larger the bin size, the lower the resolution, similarly to pixel size in pictures. Each cell of the matrix corresponds to a pair of bins, to which is associated the raw number of valid read pairs (referred as “counts” when talking about bin pairs) connecting the corresponding genomic intervals. The total number of bin pairs –hence the size of the matrix- therefore depends on the genome size and on the resolution (15,182,805, 83,650,645 and 1,973,647,378 bin pairs for the 500, 200 and 40 Kb resolutions respectively). An example of Hi-C contact matrices (also called interaction matrices or contact heatmaps) obtained at different resolutions for chromosomes 1 (the biggest one) and 18 (the smallest autosomal chromosome) is provided in Figure 42.

6.1.1.2.1 Main features of Hi-C matrices

Globally, Hi-C contact matrices share similar properties. The first one is that they are balanced (or symmetric) matrices, which means that the rows and columns of the matrix represent the same feature, in this case the same succession of genomic intervals. For instance, if we observe Figure 42, the 500 Kb resolution Hi-C matrix of chromosome 1 is displayed by dividing the length of the chromosome into 500 Kb genomic intervals to form the rows and columns respectively. The first row and column both represent the genomic interval Chr1:1-500,000 (assigned to bin1), the second ones represent the interval chr1:500,001-1,000,000 (assigned to bin2), and so on. Thus, the first cell of the matrix

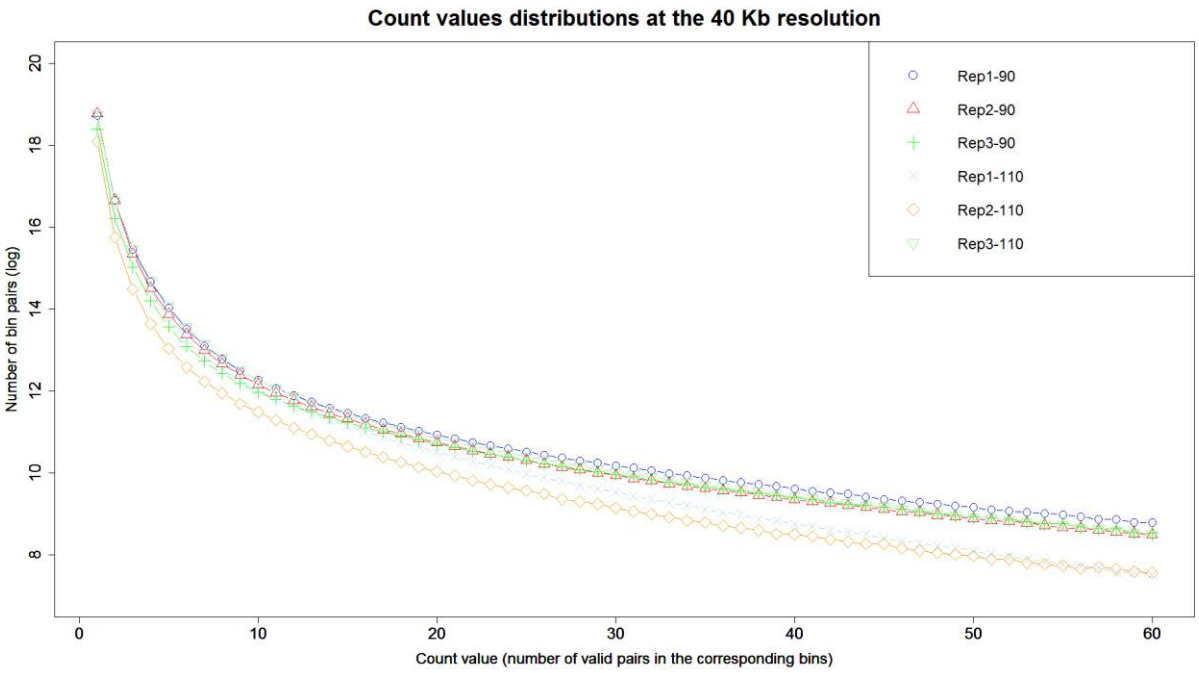


Figure 43. Distribution of count values in the 40 Kb matrices. *The number of occurrences of each count value (number of bin pairs with that value) is shown after log-transformation. Only positive values up to 60 are shown.*

Table 7. Percentage of cis and trans bin pairs in a virtual matrix with one count in each cell. *Values obtained in Hi-C matrices at different resolutions.*

	500 Kb	**200 Kb**	**40 Kb**
Total number of bin pairs	15,182,805	83,650,645	1,973,647,378
Cis bin pairs (%)	4.73	5.34	5.65
Trans bin pairs (%)	95.27	94.66	94.35

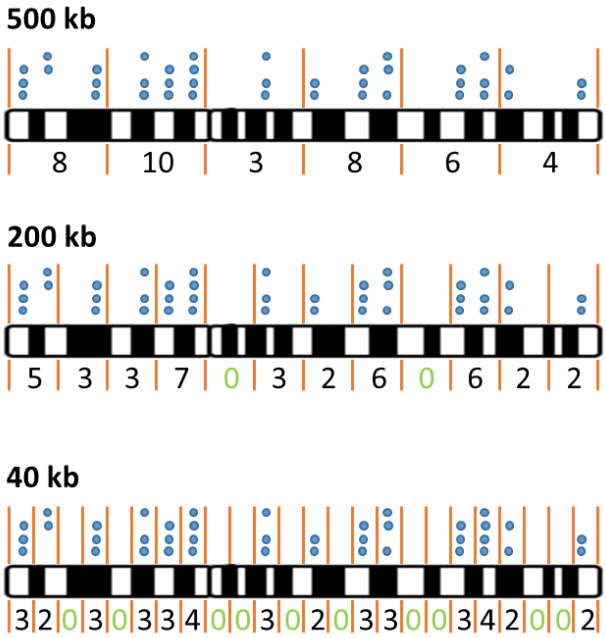


Figure 44. Schematic representation of the relationship between binning (resolution) and sparsity. *Read counts represented over one dimension genomic region to illustrate the positive correlation between resolution and sparsity on the Hi-C matrices. For an equal number of counts on a specific chromosome, when the last is divided into smaller genomic intervals (bins), the number of counts per bin decreases while the number of bins with no count increases. Consequently, the sparsity increases at higher resolutions.*

Table 8. Statistics of bin pairs counts of Hi-C matrices obtained at three different resolutions (R) for the six replicates. *The percentages of bin pairs with no count or at least one count were calculated over the total number of bin pairs (showed in Table 7). The percentages of cis and trans bin pairs with at least one count were calculated over the total bin pairs with at least one count.*

	R (Kb)	Rep1-90	Rep2-90	Rep3-90	Rep1-110	Rep2-110	Rep3-110	Mean
%bin pairs count=0	500	13.17	12.87	15.98	13.31	18.85	13.00	14.53
	200	29.66	28.83	42.15	27.48	52.32	29.21	34.94
	40	91.60	91.42	94.11	91.60	95.77	91.52	92.67
%bin pairs count>0	500	86.83	87.13	84.02	86.69	81.15	87.00	85.47
	200	70.34	71.17	57.85	72.52	47.68	70.79	65.06
	40	8.40	8.58	5.89	8.40	4.23	8.48	7.33
%cis bin pairs count>0	500	5.42	5.40	5.60	5.42	5.79	5.41	5.50
	200	7.44	7.33	8.88	7.23	10.47	7.41	8.13
	40	29.04	25.72	33.61	27.32	35.45	29.21	30.06
%trans bin pairs count>0	500	94.58	94.60	94.40	94.58	94.21	94.59	94.50
	200	92.56	92.67	91.12	92.77	89.53	92.59	91.87
	40	70.96	74.28	66.39	72.68	64.55	70.79	69.94

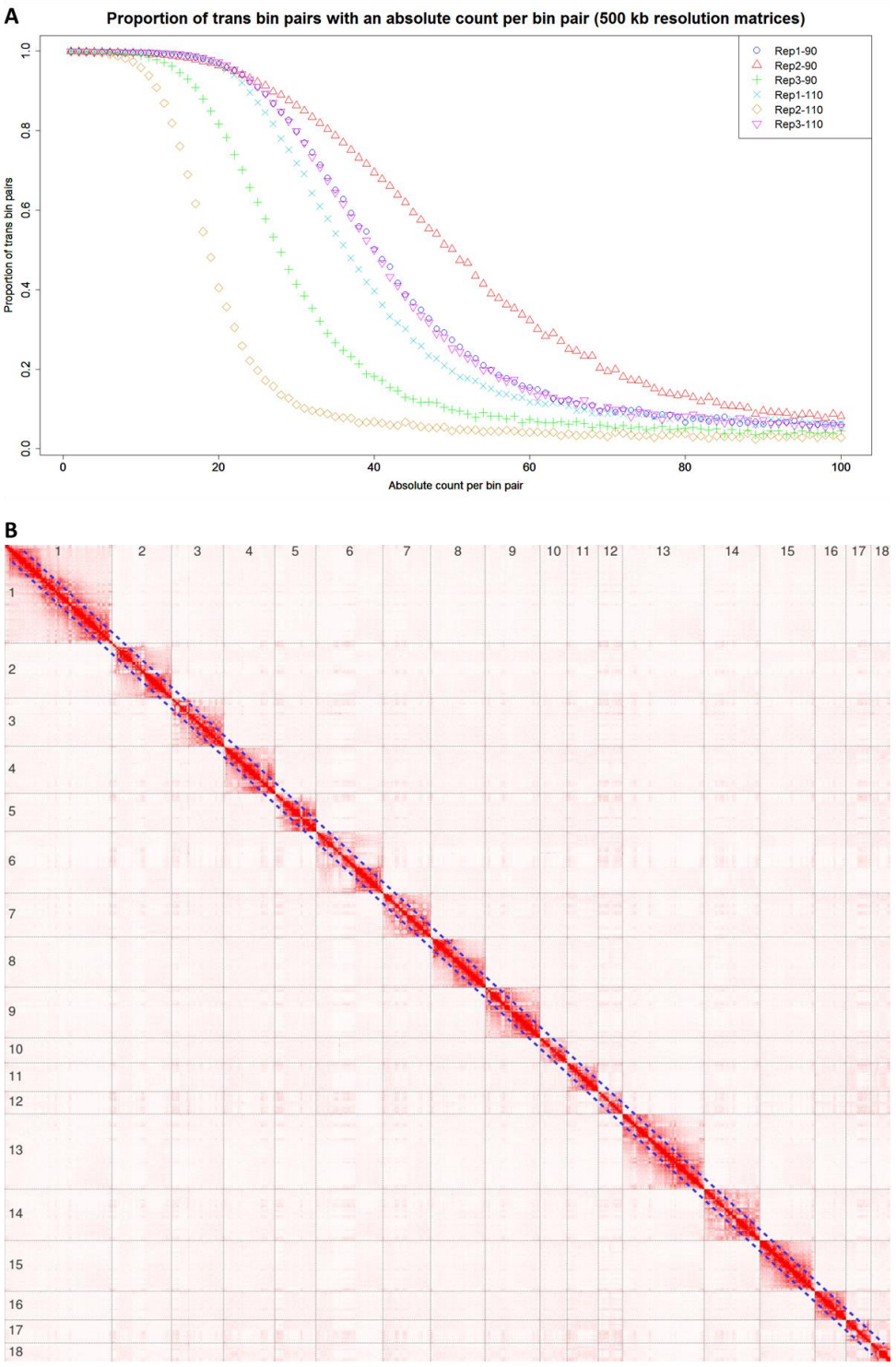


Figure 45. Distribution of counts in cis and trans bin pairs. *(A) Proportion of trans bin pairs among pairs of different count values. Proportions were computed within each set of bin pairs with 1 to 100 counts on the 500 Kb resolution matrix of each replicate. Pairs of bins with no count are not represented. Most of bin pairs harboring few counts correspond to genomic regions located on different chromosomes. Inversely, most of bin pairs harboring many counts correspond to genomic regions located on the same chromosomes. (B) Heatmap representation of a whole genome Hi-C matrix (sample Rep1-90 - only the 18 autosomes are represented). The matrix is proportional to the chromosome sizes. Interactions between proximal regions in the genomic space correspond to the area next to the diagonal (between dotted blue lines), where most of the high counts are located.*

(row1-column1) contains the number of counts (reported valid read pairs) between genomic regions mapped to bin1, the second one contains the reported counts between bin1 and bin2, and so on. These Hi-C matrices are symmetric because, for instance, the number of counts in the matrix cell "row1-column2" (bin1-bin2) is the same than in "row2-column1" (bin2-bin1).

The second property of the Hi-C matrices is the high density of counts all along the diagonal, which represents the number of valid read pairs mapped to genomic regions located in the same genomic interval (bin) or consecutive genomic intervals (bins), as observed in Figure 42. This is an expected observation as proximal genomic regions in the linear sequence of the genome cannot be far from each other in the 3D nuclear space and thus, the probability of a Hi-C religation event between these regions is higher than between genomically distal regions. Hence, as we move away from the diagonal, the number of counts decreases along with the probability of contact between distal genomic regions. Far from the diagonal, inter-chromosomal bin pairs (*trans*) usually have low counts. These *trans* bin pairs represent the vast majority (~ 95%) of all bin pairs in a typical matrix (Table 7), meaning that most of the bin pairs have low or no counts. As the distribution of the count values shows in Figure 43, bin pairs with few counts (~ 1 – 5) are more abundant than those with many counts.

As mentioned above, the third property of the Hi-C matrices is the sparsity. In numerical analysis, a sparse matrix is defined as a matrix in which most of elements (in our case, most of counts) are zero, which is the opposite of a dense matrix. This can be observed in Figure 42, where outside the diagonal most of the matrix is white (absence of contacts). The sparsity increases as the resolution increases and pairs of bins become less dense in counts. In other words, for an equal number of counts in a contact matrix, when the genomic intervals (bins) are smaller (i.e. 40 Kb vs. 500 Kb resolution), the proportion of bins with no count increases (Figure 44). In our data for instance, the proportion of bin pairs with no count represents about 14.5% of the matrix at 500 Kb and 34.9 % at 200 Kb. At 40 Kb, it reaches 92.7 % of the matrix (Tables 7 and 8).

6.1.1.2.2 Proportion of *cis* and *trans* read pairs

As the *cis/trans* ratio is often mentioned as an informative statistic to describe Hi-C data (Dixon et al., 2015), we further analyzed the proportion of read pairs (counts) in *cis* and *trans* bin pairs. Within each set of bin pairs with a specific value (from 1 to 100 counts), we computed the proportion of *trans* bin pairs (Figure 45A). The first observation is that most of bin pairs (genomic regions) containing a low number of counts correspond to *trans* bin pairs. These results were expected due to the direct correlation between the number of counts and the genomic distance. By measuring digestion-religation events between proximal genomic regions in the 3D nuclear space, we are indirectly measuring the spatial distance between these genomic regions.

A second observation is that all curves in Figure 45A show a quite drastic transition from a high number of *trans* bin pairs with few counts, to a high number of *cis* bin pairs with many counts. This is reflected by the marked increase in count density around the diagonal in the heatmap representation of the Hi-C matrix genome-wide (the 25 longest scaffolds, including the 18 autosomes and the 2 sex chromosomes, Figure 45B). Moreover, we observed a shift between all six curves and a progressive

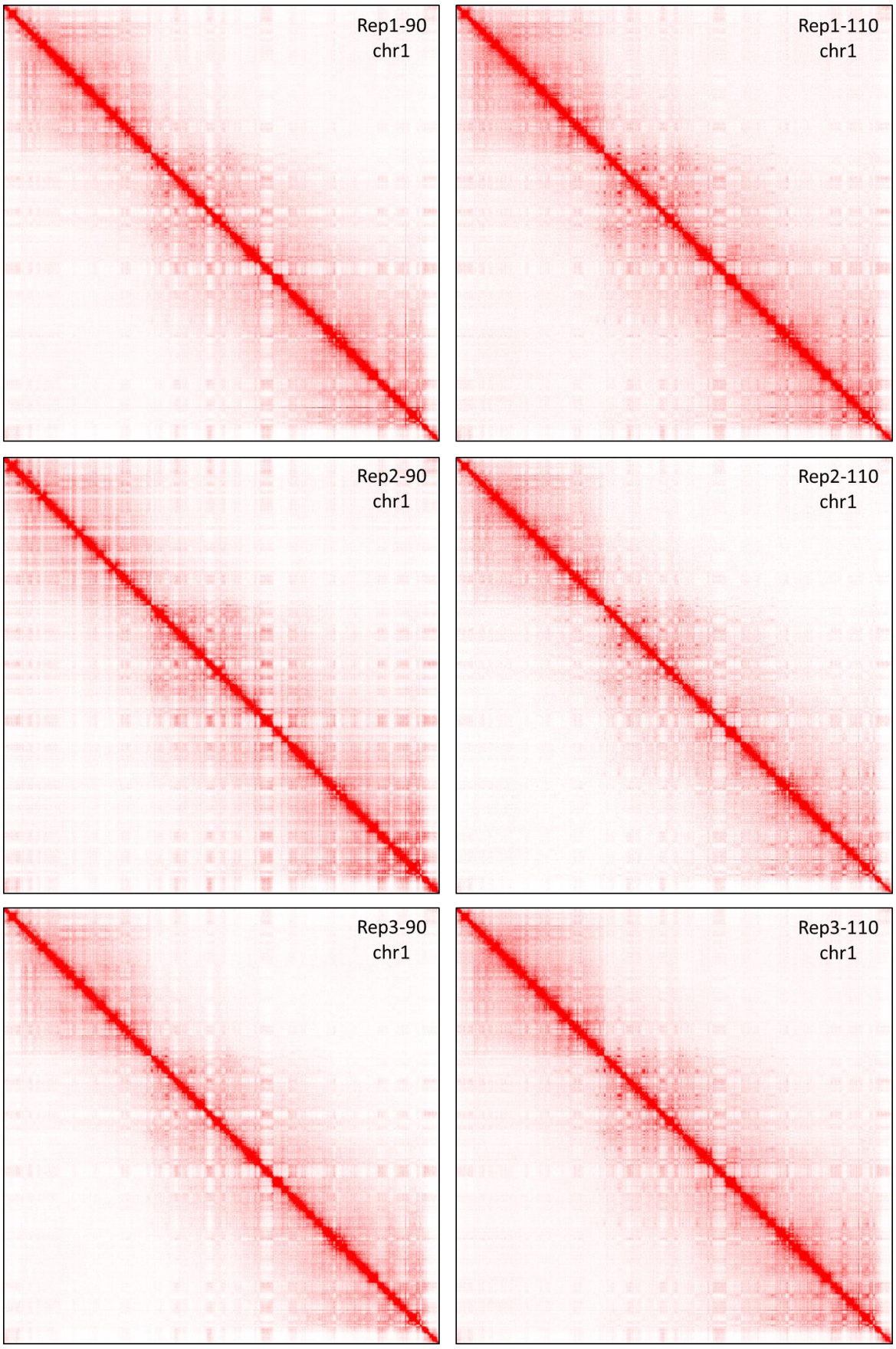


Figure 46. Individual Hi-C contact matrices for each replicate. *Illustration of Hi-C raw contact matrices of chromosome 1 obtained at 200 Kb resolution. Left column: the three replicates at 90 days. Right column: the three replicates at 110 days.*

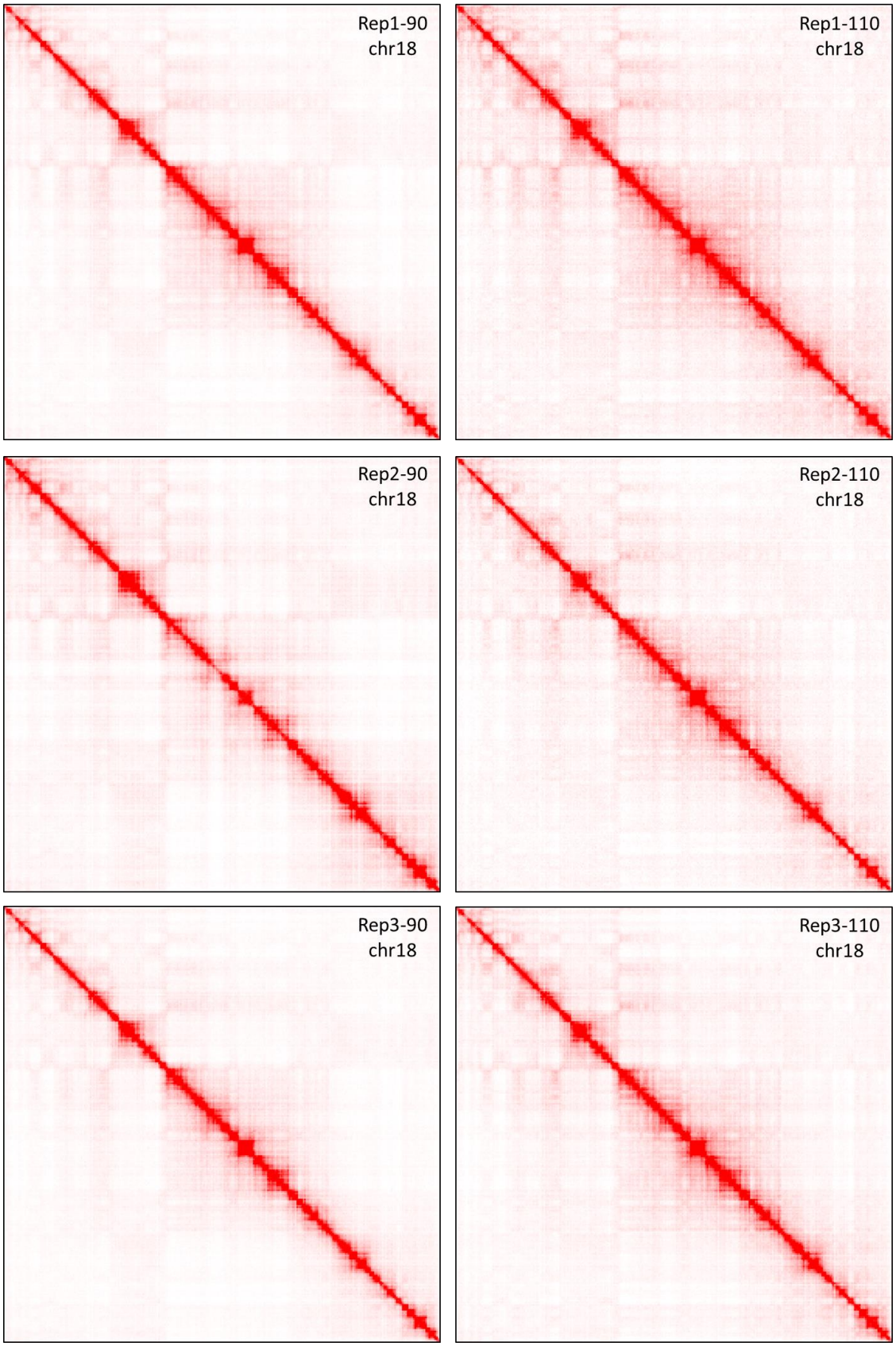


Figure 47. Individual Hi-C contact matrices for each replicate. *Illustration of Hi-C raw contact matrices of chromosome 18 obtained at 200Kb resolution. Left column: the three replicates at 90 days. Right column: the three replicates at 110 days.*

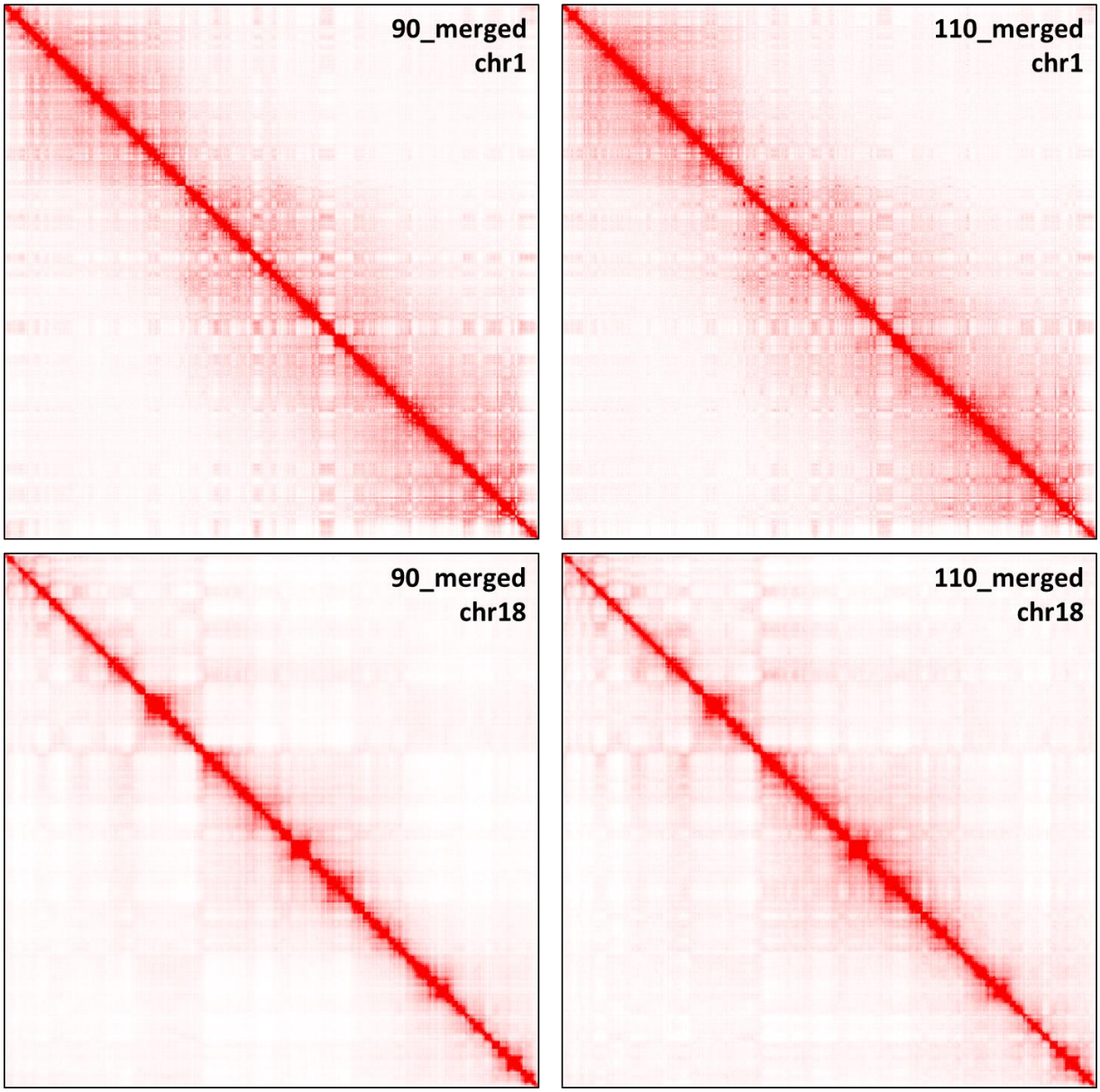


Figure 48. Merged Hi-C contact matrices. *The three Hi-C contact matrices of each condition (90 days and 110 days of gestation) were merged before normalization. Example of the merged matrices obtained at 200 Kb resolution for chromosomes 1 (top) and 18 (bottom).*

decrease in their slopes (Figure 45A). The first curve on the left corresponds to Rep2-110, the library with the lowest number of sequenced and mapped read pairs (Figure 37A), and the last curve on the right corresponds to Rep2-90, which shows the highest number of mapped read pairs. This means that the distribution of counts in *cis* and *trans* bin pairs is highly dependent on the genome coverage. In other words, as the coverage increases, we find more bin pairs with high counts in *cis* and more with small counts in *trans* and vice versa, and the observed transition on the cis/trans ratio between pair of bins with few and many counts becomes smoother due to the saturation of the matrix in counts. This emphasizes the need of an efficient normalization method to compare replicates and/or conditions.

6.1.1.2.3 Hi-C matrices comparison

Sixty Hi-C matrices were generated for each replicate: 3 resolutions x 20 chromosomes (18 autosomes + 2 sex chromosomes) (Appendix 11). To limit the effect of the sexual chromosomes on the results (due to experimental constraints, samples from both genders were collected) we focused our analysis on the 18 autosomes. A rough visual comparison of these Hi-C matrices shows that they look very different across chromosomes while very similar across the three replicates of the same condition (Figures 46 and 47) suggesting that results are reproducible. Considering the apparent similarity across replicates, merged matrices were generated for each condition (90_merged and 110_merged) by adding the raw counts of the individual matrices across animals for each pair of bins (Figure 48). Matrices from different conditions (90 vs. 110 days) also seem to be generally conserved overall, as previously reported in several studies ((Dixon et al., 2015; Rao et al., 2014; Sexton et al., 2012).

Globally, the general features we observed in our data (matrix sparsity, high density of counts over the diagonal) are consistent with the ones previously published in human and mouse. As far as we know this is the first characterization of the genome organization made by Hi-C in cells from fetal muscle tissues. Moreover, at that level of analysis, the apparent similarity between conditions globally shows a high level of conservation in the 3D genome structure as previously reported between different cell lines or tissues, even between different species (Rao et al., 2014).

6.1.1.3 Hi-C intra-matrices normalization

Hi-C matrices are subject to specific and non-specific biases that need to be corrected before being analyzed. The non-specific ones are the classical biases of sequencing (regions with high GC content) and mapping (repetitive DNA regions). In addition, the genomic density in restriction sites, which are the target of the restriction enzyme used in the DNA digestion step of the Hi-C experiment, is a specific source of biases when performing Hi-C assays. Regions that are enriched on the specific restriction site tend to be cut more frequently that those poor in restriction sites and consequently present a higher probability of Hi-C religation events. In order to remove such biases (GC content, mappability or restriction site density), all matrices were normalized using the non-parametric ICE method (Imakaev et al., 2012) implemented in HiC-Pro (Servant et al., 2015) (see Materials and Methods). Globally, the ICE (Iterative Correction and Eigenvector decomposition) normalization assumes that the bias for detecting contacts between two regions can be represented as the product of the individual biases of these regions. Briefly, this normalization is done in order to make all bins of a given matrix comparable by means of an iterative process. This method ensures that the total counts that involve a given bin (sum

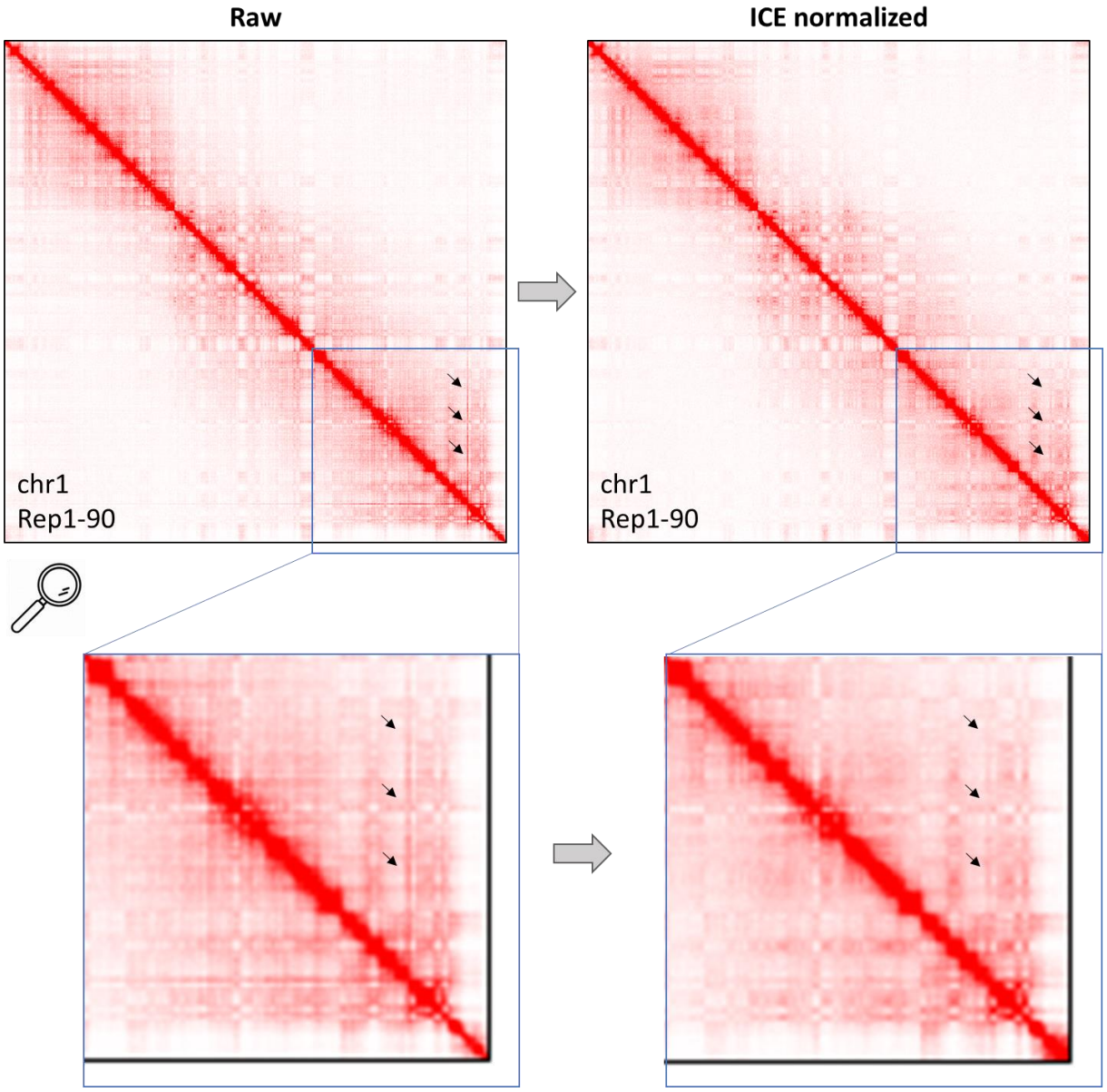


Figure 49. Normalization of Hi-C matrices. *Example of the Hi-C matrix of chromosome 1 before and after ICE normalization (Rep1-90) obtained at 200 Kb resolution. The "smoothing" effect of the normalization is visible at specific positions (example indicated by the black arrows).*

of all the values throughout a row or a column) is the same for all the bins. Figure 49 shows an example of Hi-C contact matrices before and after normalization.

The normalized matrices obtained at a 500 Kb resolution were used for the detection of large genomic compartments (known as A and B compartments), and those obtained at 40 Kb resolution were used to detect smaller genomic domains (TADs) (this will be further detailed in the next section). The 200 Kb contact matrices were obtained to work at an intermediate resolution. Accordingly to the "map resolution" definition described in (Rao et al., 2014), which refers to the smallest bin size such that 80% of the loci have at least 1,000 contacts, we achieved a good map resolution as 99.98%, 99.98% and 99.56% of bins in our 500, 200 and 40 Kb resolution matrices showed more than 1,000 contacts.

6.1.2 Identification of higher order chromosomal structures

6.1.2.1 A and B compartments

The so-called "A" and "B" compartments are large genomic regions often defined as "open active" and "close inactive" compartments respectively. "A" compartments are characterized as transcriptionally permissive, euchromatic, gene-rich and DNase I hypersensitive regions. Inversely, "B" compartments are considered as transcriptionally inert, heterochromatic, nuclear lamina-associated, gene-poor and DNase I insensitive (Bonora et al., 2014; Gibcus and Dekker, 2013). From our Hi-C data, we sought to investigate the compartmentalization of the genome in order to determine whether: (a) these functional compartments previously reported in other studies exist in porcine fetal muscle, and whether they are similar or they differ compared to model species, (b) they vary between the two conditions (90 days and 110 days of gestation). In brief, the method used to identify these compartments relies on a Pearson correlation matrix made from the bins to distinguish the two groups of genomic positions assigned to A and B compartments in each chromosome. The segmentation into A and B compartments can be observed in these correlation matrices in the form of a plaid pattern with red and blue stripes (see the corresponding section in Materials and Methods for more details). Correlation matrices for chromosomes 1 and 13 are shown in Figures 50 and 51. A complete set of all chromosome correlation matrices (obtained from the merged matrix of the 3 replicates at 90 days of gestation and the merged one at 110 days) is provided in the Appendix 12 and 13.

A first look at the resulting Pearson correlation matrices allowed us to note the presence of these A/B compartments in all chromosomes of our six samples, and to see that they look like the ones from the literature. Due to insufficient coverage and other filtering steps of the compartment calling method HiTC ((Servant et al., 2012), see Materials and methods), several genomic regions were not assigned to any compartment, resulting in white stripes in the visualizations. Nonetheless, we observed that these matrices appear to be quite similar within and between conditions (Figures 50 and 51) while, very different across chromosomes (Appendix 12 and 13). For instance, some chromosomes (i.e. chr1, chr3, chr15 and chr16) are highly segmented while others (i.e. chr5, chr6, chr8 and chr17) show quite large compartments.

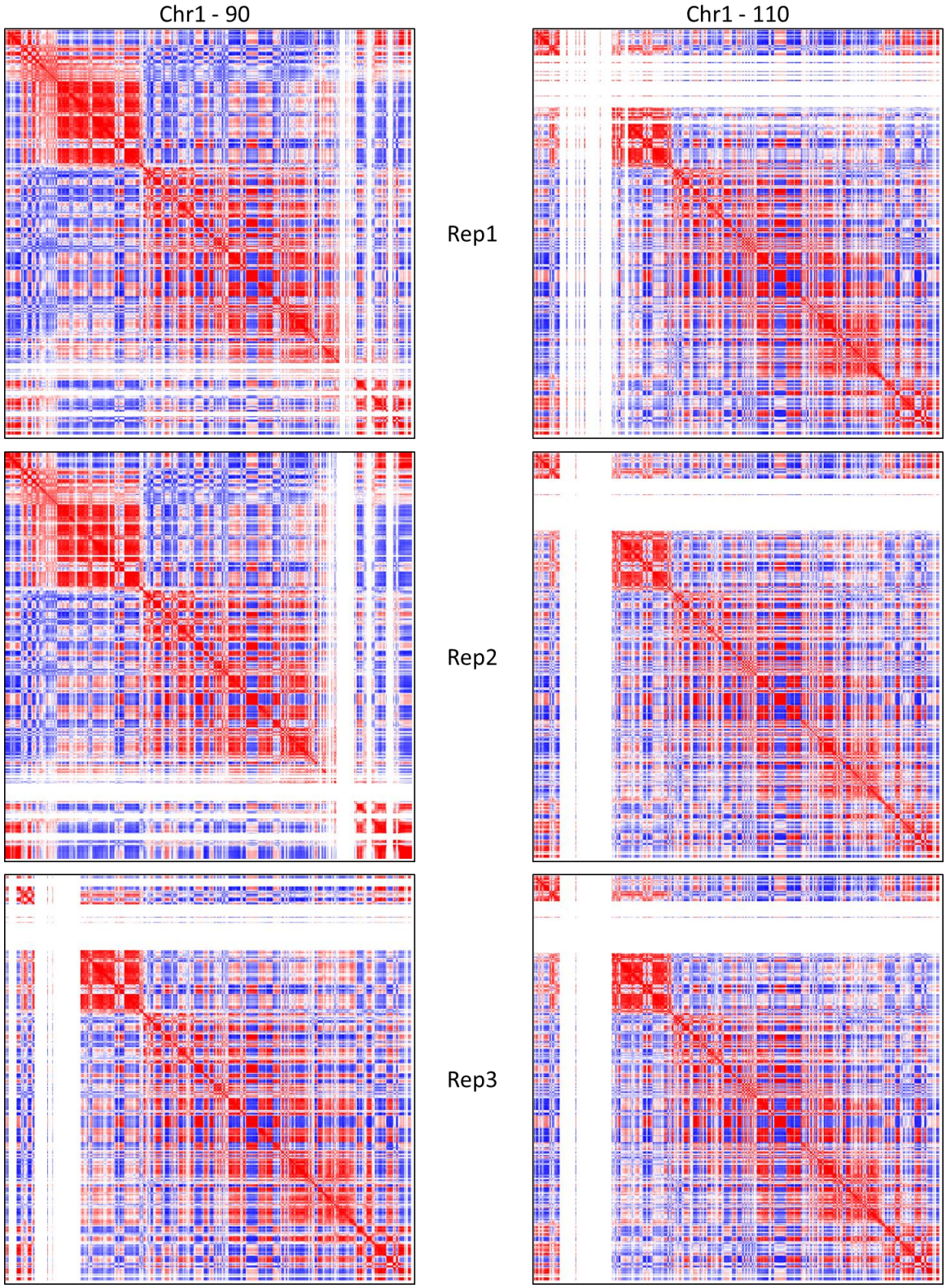


Figure 50. Hi-C A and B compartments for individual matrices (chromosome 1). *Illustration of Pearson correlation matrices obtained between bins on chromosome 1 at 90 days (left column) and 110 days (right column) of gestation, to predict A and B compartments. The color code in a pair of bins (cell of the matrix) represents the correlation coefficient between the normalized values of the corresponding bins (blue=low correlation, red=high correlation). Sharp transitions between blue/red stripes define boundaries between A and B compartments. Occasional white stripes characterize regions with no prediction from the method due to insufficient coverage and other filtering steps.*

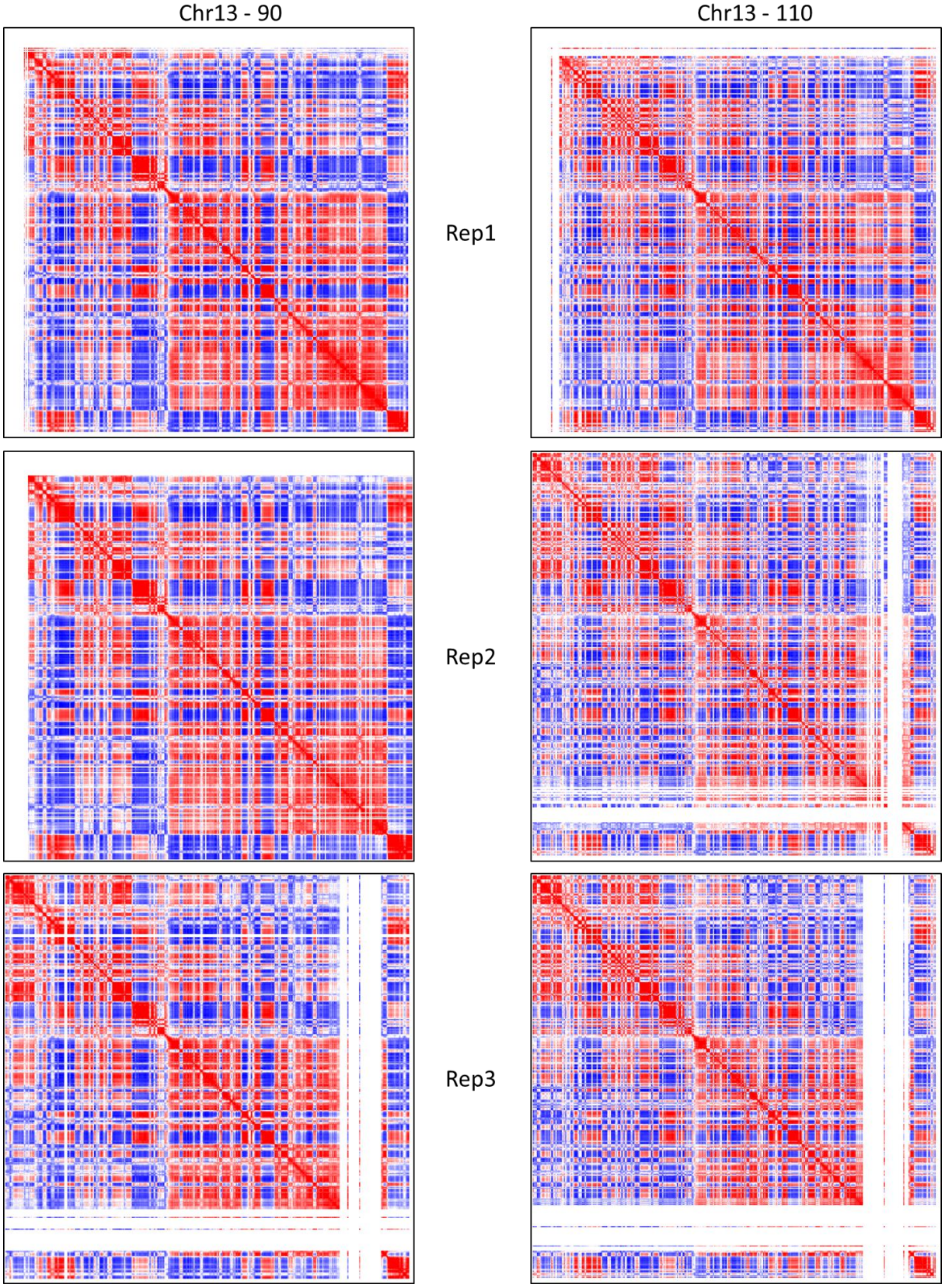


Figure 51. Hi-C A and B compartments for individual matrices (chromosome 13). *Same as above (Figure 49).*

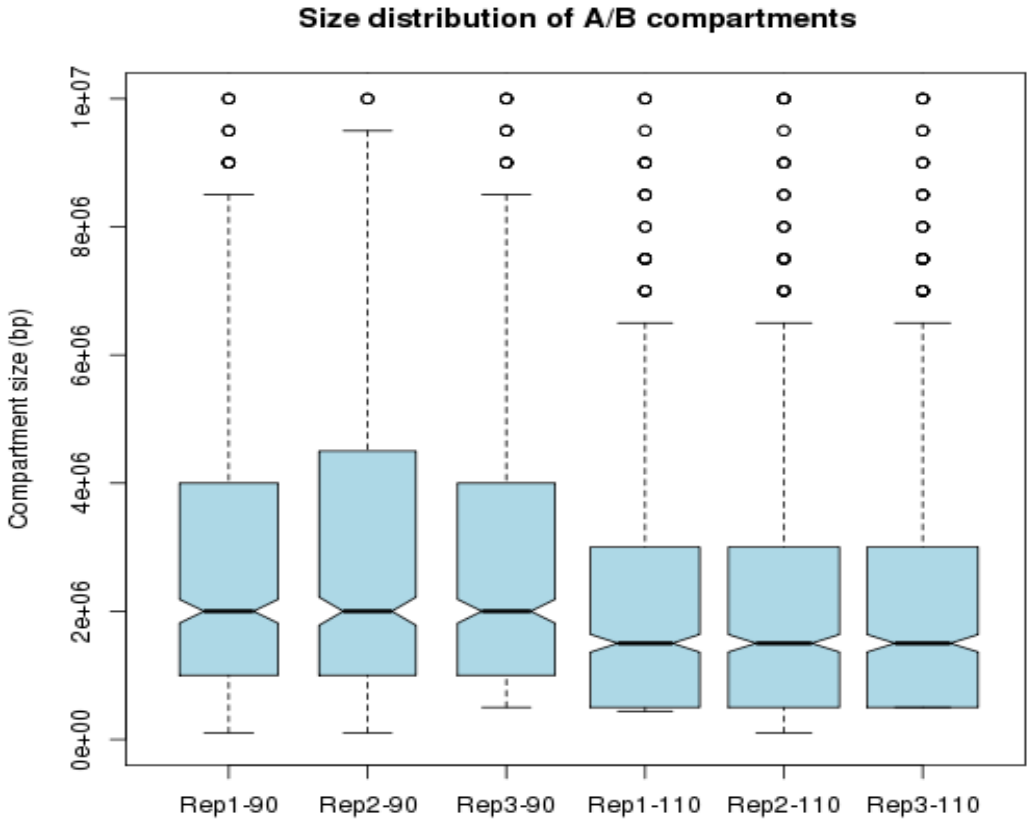


Figure 52. Size distribution of AB compartments for each replicate. *All pairwise comparisons of A/B compartments size distributions between replicates from the two conditions were significant and no significant differences on size distribution were found between pairs of the same condition.*

About 733 compartments per replicate were predicted on average, with a mean size between 1.5 Mb and 2 Mb, which are in the same order of magnitude that those reported from Hi-C experiments in human or mouse cells (Dixon et al., 2012; Lieberman-Aiden et al., 2009).

6.1.2.1.1 Differences in size and number of A/B compartments

Next, we investigated whether A and B compartments differ between conditions. To address this question, we first compared the number of A/B compartments across replicates. We noted that this number varies substantially (598, 586, 594 and 767, 781, 768 for the 3 replicates at 90 days and the 3 at 110 days respectively), and that the variability is higher between conditions than across replicates of each condition. In addition, analyzing the number of A and B compartments separately we found approximatively 51% of A and 49% of B compartments in each replicate. Consistently with the variability in number of A/B compartments observed between conditions, the compartments are remarkably larger at 90 days than at 110 days of gestation as shown in Figure 52. To confirm this observation, we tested the significance of all pairwise combinations across replicates and confirmed that the observed differences on size distribution inter-conditions were significant (Wilcoxon test, 3.166e-09 $\leq p$-value $\leq$ 4.83e-04) while no significant differences were found intra-conditions (0.1308 $\leq p$-value $\leq$ 0.8823). This outcome has two plausible explanations: it could be either, the result of a real biological difference of size between the two gestational ages, or it could be an artifact, resulting for instance from differences of sequencing depth, as previously observed with the number of counts in *cis* and *trans* bin pairs.

In order to explore the potential impact of sequencing depth on the size of compartments, we computed the correlation between the mean compartment size of each replicate and the number of valid read pairs. A relatively weak yet notable correlation was obtained (Pearson correlation coefficient r = 0.45), suggesting that the size of the detected compartments could be affected by the quantity of data. As we observed with the white stripes in the correlation matrices, regions with low coverage could be filtered out by the A/B compartment calling method. Such filtered bins, when present within compartments, would result in a fragmentation of the predicted compartments and consequently in a general shortening of their sizes. To answer if the observed difference in size distributions could be due to a difference in coverage via filtered bins, we computed the number of genomic regions (bin size: 500 Kb) that could not be assigned to an A or B compartment in each replicate. This number varies between 500 and 709, which represents between 11.3% and 16.9% of the total number of bins, meaning that considerable parts of the genome are not assigned to any compartment. This number of unassigned bins was negatively correlated with the mean size of the compartments across replicates (Pearson correlation r = -0.89), which is supportive of a potential impact of the available data on the size of the compartments via a fragmentation effect of the prediction method. We concluded that the significant difference in compartment sizes that we observed between conditions may be partially explained by an artefact of the prediction method via a difference in the quantity of available data. This emphasizes the importance of considering regions with enough information for any comparison between conditions.

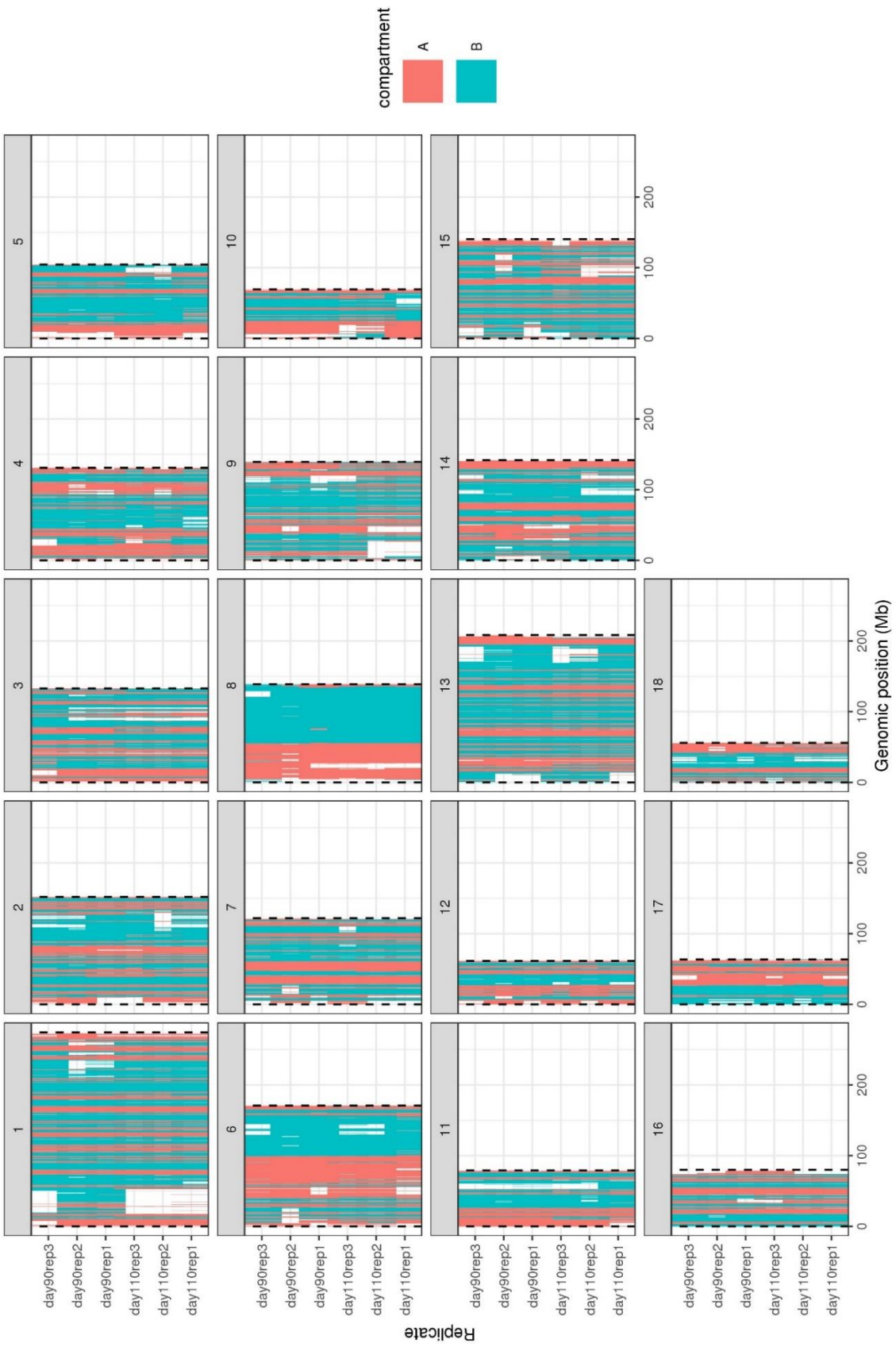


Figure 53. Distribution of Hi-C A and B compartments along each chromosome for each replicate. *Genome-wide overview of compartment labels per 500 Kb bin. A general consistency can be observed across replicates. Dotted lines delimit the beginning and the end of each chromosome. White regions are devoid of any called compartment.*

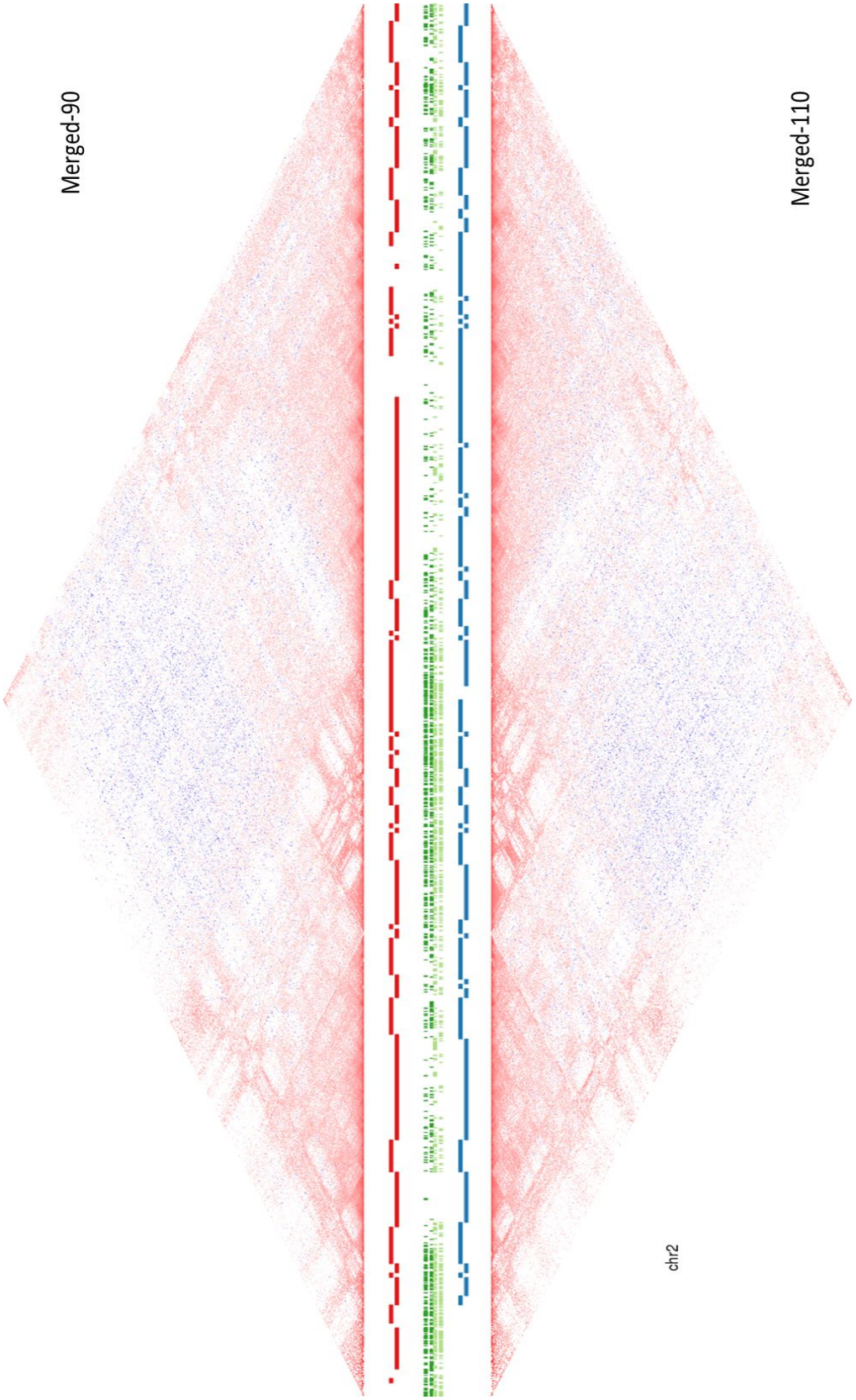


Figure 54. A/B compartments and gene annotation along the two Hi-C merged contact matrices (90 days vs. 110 days). *The figure shows as example the comparison of the two merged matrices obtained for chromosome 2. A/B compartments are represented as red intervals for the merged matrix obtained at 90 days of gestation, and as a blue intervals for the one obtained at 110 days. Gene annotation is represented in green, in the middle of the image.*

6.1.2.1.2 Differences in A/B compartments assignment

Since the compartment size could be affected by the sequencing depth, we wanted to make sure that it was not the case for the A/B assignation itself and sought to investigate: (a) the consistency of the A/B calling method across replicates from the same condition and (b) the differences in A/B calling between conditions. We therefore looked whether each genomic region (bin) was assigned to the same or to a different compartment type across replicates (Figure 53). Any bin containing missing data (lack of compartments assignment) in any of the six replicates was not considered for this analysis. Over 3,371 bins (bin size = 500 Kb), 2,809 (83.3%) were assigned to the same compartment type (either A or B) in all six replicates, which is consistent with a general conservation of the higher structural organization level of the genome even in different cells as previously observed in human (Barutcu et al., 2015). Interestingly, 94.1% and 90.7% of the bins were assigned to the same compartment type in all of the three replicates at 90 days and 110 days of gestation respectively. Considering pairwise comparisons, the average number of bins with the same label is 3,201 (95.0%) between replicates from the same group and 3,038 (90.1%) between replicates from different groups. Altogether, these results confirm the high consistency of the A and B compartments prediction method when the A/B information is available.

After verifying that the variability intra-condition was low, we focused our interest on analyzing the variability between conditions and, more precisely, on the proportion of bins switching from one compartment type to another. Again, considering only bins with assigned compartments, different approaches are possible. First, we performed the compartment calling after merging the three Hi-C contact matrices per condition, which allowed to increase the genomic coverage of the Hi-C matrix and therefore to reduce the number of unassigned bins (down to 7.6% and 8.2% for 90 days and 110 days respectively). The pairwise comparison of the A/B assignment between the merged matrices indicated that among the 4,026 bins (88.7%) with an assigned compartment in both conditions, 3592 of them (89%) have the same compartment (Figure 54). Among the remaining 444 variable bins (11.0%), which, at 500 Kb resolution, represents ~222 Mb of the genome, 181 (40.8%) indicate a switch from an A compartment at 90 days of gestation to a B compartment at 110 days (A ➔ B) and 263 (59.2%) showed a switch from B to A between the two gestational stages (B ➔ A). Alternatively, another approach to identify these compartment switches is to take advantage of our experimental design by considering the A/B compartment calling that was made on all the replicates separately. Among the 3,371 bins (74.3%) with an assigned compartment in all the 6 replicates, 2,809 (83.3%) have the same assignment. As expected, this proportion is slightly lower than the one we obtained from comparing only two merged matrices (88.7%) but still indicative of a strong consistency across replicates. To identify switching bins, we required a total consistency between the assignments within each condition, meaning a switch from a triple A to a triple B or the other way around. Eventually, 104 bins (3.1%) fulfilled this stringent condition (~52 Mb of the genome), among which 45 (43.3%) and 59 (56.7%) indicated an A ➔ B and a B ➔ A switch, respectively. Consistently, a large majority of them (92 out of 104) were also detected by the previously described approach using merged matrices. These bins, with a strong and consistent A➔B or B➔A switch, have been further investigated for integrative analyses with gene expression data in the section “Gene expression and nuclear organization”.

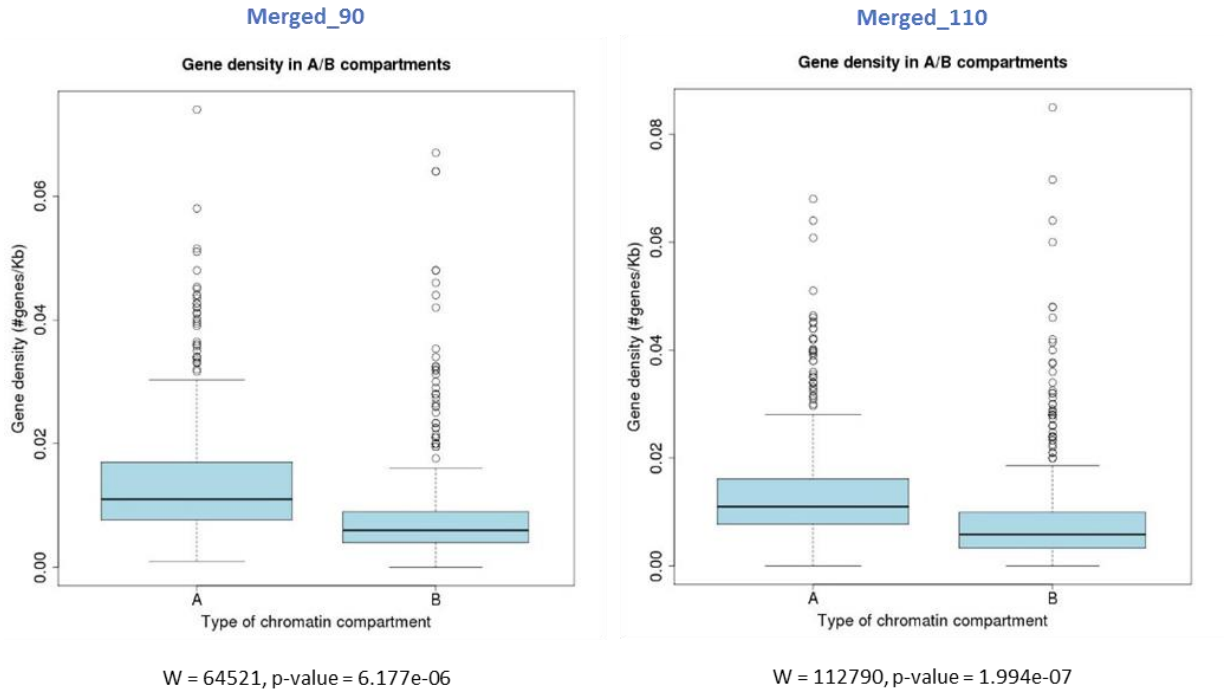


Figure 55. Gene density in A and B compartments. *The A and B compartments were estimated from the two merged matrices obtained at 90 days and 110 days of gestation. The differences observed between the two types of compartments were found significant (Wilcoxon test).*

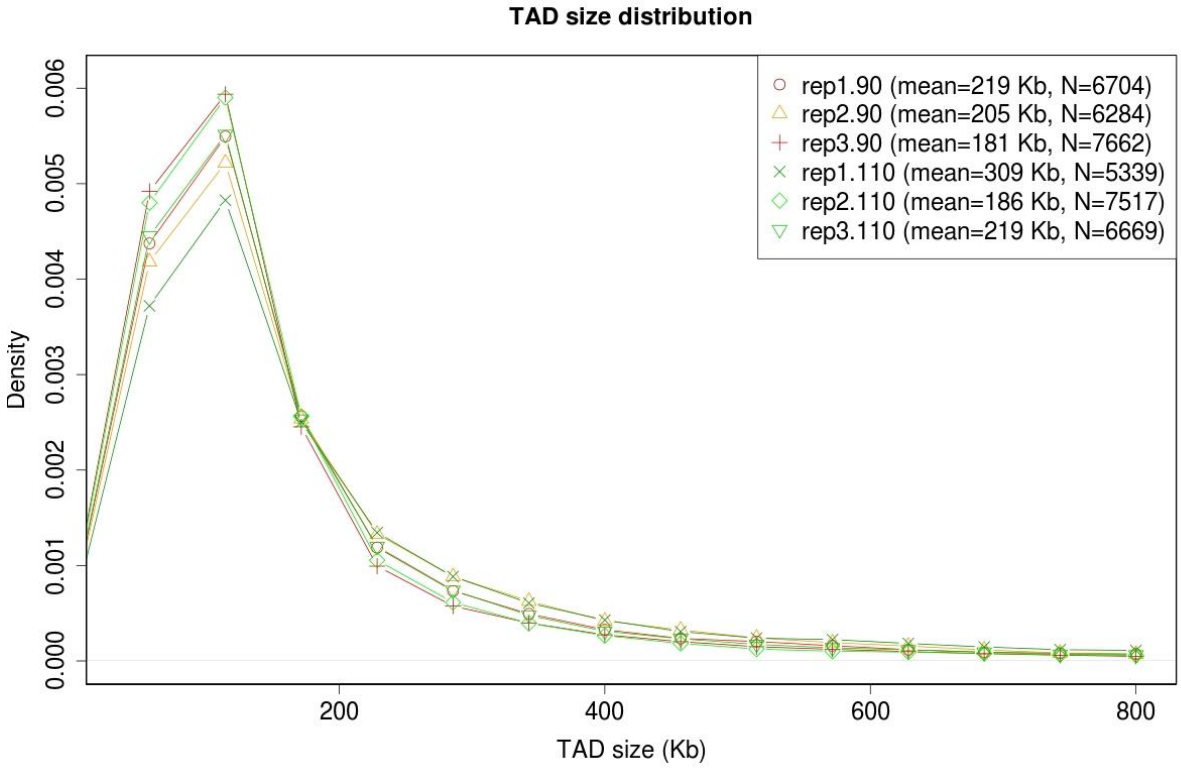


Figure 56. Size distribution of TADs for each replicate.

6.1.2.1.3 Gene density in A/B compartments

In model organisms, A compartments have been reported to be transcriptionally active and gene-rich, while B compartments were transcriptionally inactive and gene-poor (Lieberman-Aiden et al., 2009). We estimated whether this general organization was preserved in our samples by computing the gene density in the A and B compartments estimated from the two merged matrices obtained at 90 days and 110 days of gestation. For that purpose, we used the reference gene annotation ENSEMBL (Figure 55) and considered for each bin the number of annotated genes in the corresponding 500 Kb. As reported in human and mouse, gene density was significantly higher in A vs. B compartments (Wilcoxon test: *p*-value < 6.2e-06). Similar results were obtained on the individual matrices (Appendix 14).

6.1.2.2 Topologically associated domains (TADs)

At a smaller scale than A/B compartments, we investigated the structural organization of the pig genome at the level of the Topologically Associated Domains (TADs). TADs are defined as chromatin domains enriched in self-interacting regions, with a frequency of intra-domain interactions higher than inter-domain interactions (Dixon et al., 2012; Matharu and Ahanger, 2015; Nora et al., 2012). They can generally be observed in contact heatmaps as darker triangles on each side of the diagonal (see Figure 46 for an example of visible TADs in our data). We used the armatus program (Filippova et al., 2014) in order to find TADs in the 40 Kb resolution matrices (see Materials and methods for more details). In a first step, TAD finding was performed on individual matrices of each replicate separately in order to assess the reproducibility of the results, then on the merged matrices to obtain a set of TADs for each condition. Globally, thousands of TADs could be identified in each replicate (from 4,941 to 7,176), with 78.9% of the genome being part of a TAD in at least one of the replicates. The average TAD size per replicate varies from 181 to 309 Kb (Figure 56), which is lower but in the same order of magnitude than the reported range of TADs found in human and mouse (median size: 880 Kb) (Dixon et al., 2012). Unlike for the A/B compartments, no strong correlation was found between the mean TAD size and the sequencing depth (valid pairs) across replicates (Pearson r=0.3).

6.1.2.2.1 CTCF and TADs

In several mammals, the CTCF DNA binding protein, which plays an important role in genome architecture, is enriched at TADs boundaries and involved in the mechanisms of loop formation (Björkegren and Baranello, 2018; Dixon et al., 2012; Rao et al., 2014). No information being available for pig, we wondered whether this protein could play a similar role in this species. Thus, we sought to identify the CTCF-binding sites in fetal porcine muscle and to map them in our detected TADs. In order to identify CTCF-binding genomic regions we tried two approaches, one experimental in *vitro* and one in *silico*. First, we performed ChIP-seq (Chromatin Immuno-Precipitation sequencing) assays on fetal muscle samples from the same animals than the ones used for Hi-C experiments. This method allows capturing chromatin regions associated to a protein of interest, in our case, the CTCF protein. In brief, all DNA-protein interactions are cross-linked and then, target sequences are enriched by using a specific antibody (see Materials and methods for more details). The six ChIP-seq libraries (Rep1-90, Rep2-90, Rep3-90, Rep1-110, Rep2-110 and Rep3-110) were pooled and sequenced on one lane of a HiSeq3000,

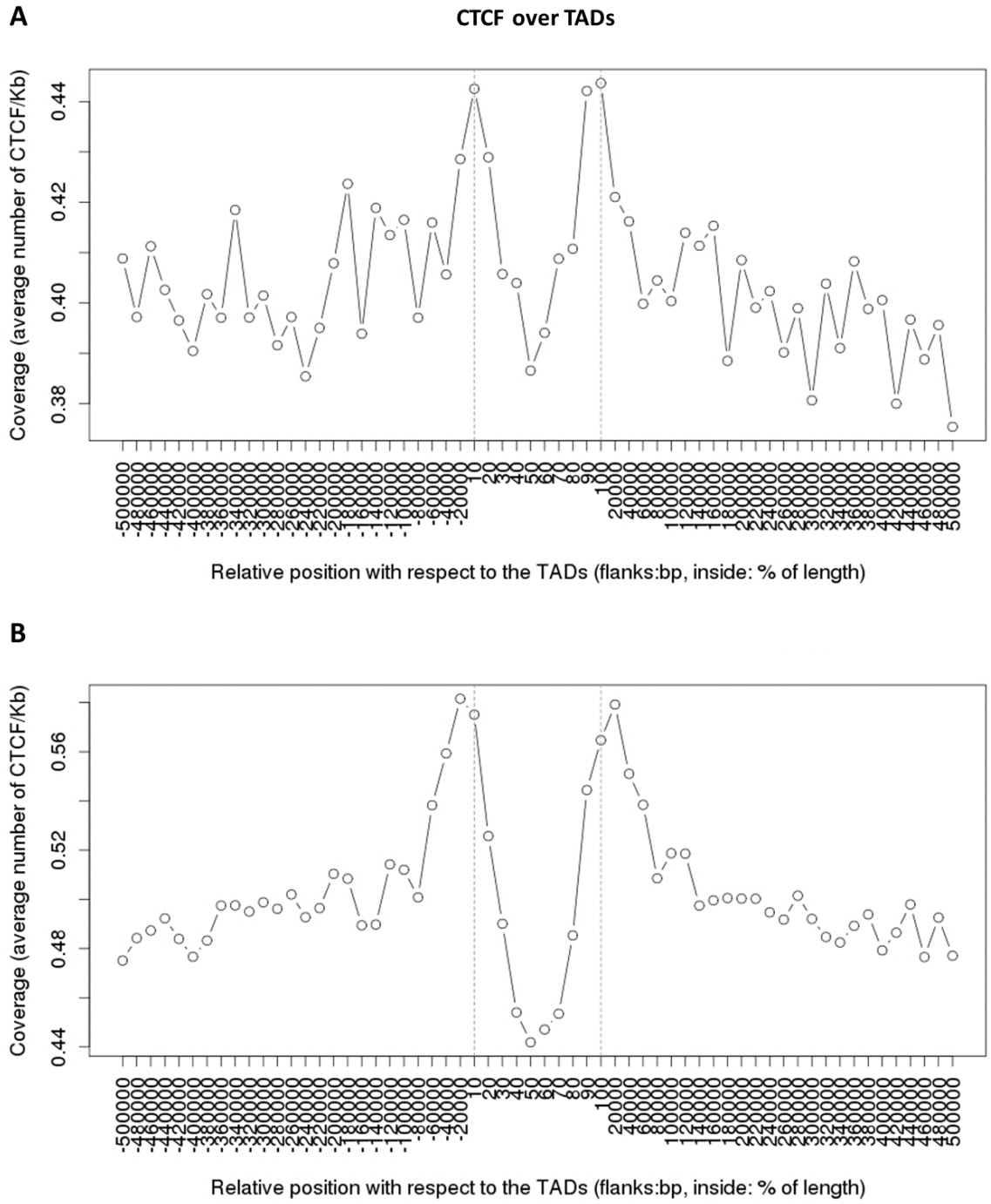


Figure 57. Genomic density profiles of predicted CTCF motifs around TADs. *Mean density of CTCF binding sites predicted on Sscrofa10 (A) and Sscrofa11 (B) relative to TADs positions detected on the corresponding genome versions. Dotted vertical lines represent the TAD boundaries and delimit the relative position inside the domains (from 0 to 100% of the TAD length). Outside of the domains, up to 500 Kb upstream and downstream flanking regions are represented. These density plots were obtained for Rep1-110, and similar profiles were observed for the other five samples.*

together with two input DNA libraries (input-1 and input-2) which are later required as a control for the data analysis. Input DNA represents a 10% fraction of the fragmented and cross-linked DNA set aside before any specific selection for CTCF-binding fragments, and which processing resumes at the reverse crosslink. Because the input DNA is essentially genomic DNA, it is used as a background sequencing control to compare with libraries that are enriched in CTCF-bound fragments. Between ~ 69 M and 96 M read pairs were sequenced per library. From them, around 76% - 81% and 73% - 82% could be mapped to the reference genome (Sscrofa11) for the three replicates at 90 and 110 days of gestation respectively. CTCF-enriched regions (called "CTCF peaks") were obtained by comparing the read mapping density along the genome between CTCF-immunoprecipitated libraries and input DNA using the MACS2 software (see "ChIP seq data analyses", Materials and methods). Between 909 and 5,491 CTCF peaks were predicted per replicate (mean size= ~ 340 bp). To control the quality of these data, we searched for the known sequence consensus of the CTCF recognition site in the peaks. To do this, we provided the FIMO motif detection software (Grant et al., 2011) with a model of the CTCF consensus binding site (PWM for Positional Weight Matrix, see Materials and methods), which is highly conserved in vertebrates (Kim et al., 2007). Unfortunately, only ~ 9% - 15% of the peaks contained the CTCF consensus motif sequence, contrary to an expected percentage of ~ 90% for a real enrichment (results obtained on porcine cell line samples from another project, data not shown). In light of these negative quality controls, we concluded that some issue could have happened during the ChIP-seq experiments (defective batch of CTCF antibody, some problem in the immunoprecipitation or size selection steps, not enough starting material, etc.) and we discarded these data.

In the absence of available CTCF ChIP-seq data obtained from the Hi-C muscle samples, we used an *in silico* approach to validate the biological relevance of the detected TADs. The CTCF consensus motif was used again, but instead of looking for binding sites in peaks only, we performed this time a genome-wide scan in order to identify all potential CTCF sites in the pig genome. Then, we computed the genomic density profiles of the predicted CTCF sites within and around TADs. This analysis was performed on the previous (Sscrofa10) and the current (Sscrofa11) reference genome versions. In both versions, the CTCF predicted sites tend to accumulate at TAD boundaries (see Figure 57 for Rep1-110; similar density profiles were obtained for the other five replicates, data not shown). Moreover, on the improved genome version (Sscrofa11), CTCF predicted sites were not only enriched at the TAD boundaries but also depleted inside TADs. Last, we performed the same analysis by considering the orientation of the predicted CTCF sites. Similar plots were obtained with asymmetrical peaks, showing a prevalence of "forward" CTCF sites at the beginning of the TADs and of "reverse" CTCF sites at the end of the TADs (Figure 58). This is supportive of the model provided in the literature for other mammals, where pairs of CTCF sites involved in DNA structures tend to display a convergent ("head-to-head") orientation (Rao et al., 2014). These results allowed us to validate the method used for TADs detection as well as the detected TADs. Moreover, the notable improvement of the CTCF enrichment in TAD borders we observed between the 10 and the 11 assembly versions emphasizes again the importance of a good genomic reference for this kind of study, in particular when considering large structural features like TADs.

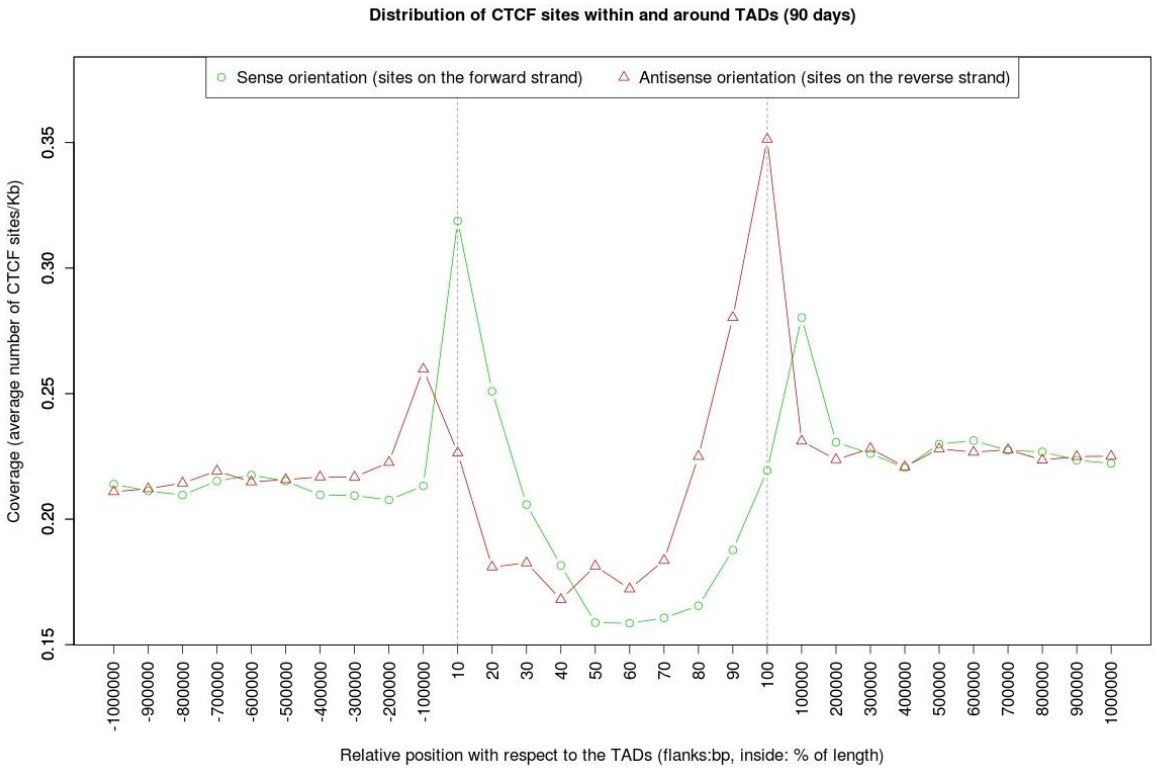


Figure 58. Genomic density profiles of forward and reverse predicted CTCF motifs around TADs. *Mean density of CTCF binding sites predicted on Sscrofa11 relative to TADs positions detected on the merged matrix at 90 days (the 110 merged matrix lead to similar results). TADs show an accumulation of forward CTCF sites at the beginning of the TAD and reverse CTCF sites at the end. The two shifted peaks correspond to boundaries of the upstream and downstream respective TADs.*

Table 9. *Number and proportion of tested bin pairs after the filtering step.*

Resolution (Kb)	**Number of bin pairs with at least one count**	
	Before filtering	After filtering
500	10,293,777	9,262,199 (89.98%)
200	63,872,799	3,844,272 (6.02%)
40	523,799,997	2,872,786 (0.55%)

6.1.3 Differential analysis of the genome organization

6.1.3.1 Global differences in the 3D genome organization of fetal muscle between 90 and 110 days of gestation

In order to investigate global changes occurring at the level of chromatin structure, we did a differential analysis to explore whether significant differences in the 3D genome organization exist between the two gestational ages. The differential analysis was performed on the raw matrices obtained for the 18 autosomes at 500, 200 and 40 Kb resolution.

As described in the Materials and Methods section, the first step was to discard pairs of bins with low read counts (5 per sample on average). Raw read counts from the remaining bin pairs (from 3 to 9 M, see Table 9) were then normalized in order to make matrices comparable across replicates. This inter-matrix normalization relies on the assumption that: (1) library sizes should be equal and, (2) MA plots should not show any strong trend between samples (Figure 59, Figure 60 and Appendix 15 and 16).

At each resolution, a Principal Component Analysis (PCA) was performed on the samples before and after normalization to investigate the organization of the data. Although the PCAs do not clearly show two distinct groups, which is expected since differences between conditions might not involve most of the genome, we observed that the resulting projection of the samples on the two first principal components (normalized data) allowed to separate the 90 and the 110 days samples along the first axis (Figure 61). Thus, we hypothesized that differences in conformation between gestational ages can explain part of the variability between samples, and that changes might occur at the level of the 3D genome organization between the 90th day and the 110th day of gestation.

6.1.3.2 Differential genome regions in late fetal muscle development

Using the normalized counts and the experimental design (2 groups, 3 replicates per group), a differential analysis has been conducted to identify pairs of genomic regions with a significant difference in the number of read pair connections between the two groups of samples. The analysis has been performed at the three resolutions (40, 200 and 500 Kb) as described in Materials and Methods.

A total of 10,183, 3,417 and 83 differential bin pairs were obtained at the 500, 200 and 40 Kb resolution respectively (Table 10). This represents a small proportion of the tested bin pairs (from 0.003 to 0.11%). Among them, between 82 and 95% bin pairs were located on the same chromosome at 500 and 200 Kb resolution, while only 58% at 40 Kb resolution. The observed differences between resolutions in both, the number and the *cis/trans* ratio of differential bin pairs, can be explained by the number of read counts per bin pairs. At smaller bin sizes, many bin pairs are filtered out because the number of counts per bin pair is low. Consequently, there are less remaining bin pairs to be compared between conditions and less differential bin pairs to be detected. Similarly, most of the bin pairs bearing low counts are those *in trans*, even after filtering. Consequently, small variations in the number of read counts per bin pairs between conditions may imply bigger contrasts in *trans* bin pairs than in *cis* bin pairs, thus increasing the number of differential bin pairs in *trans* at smaller bin sizes. This means that

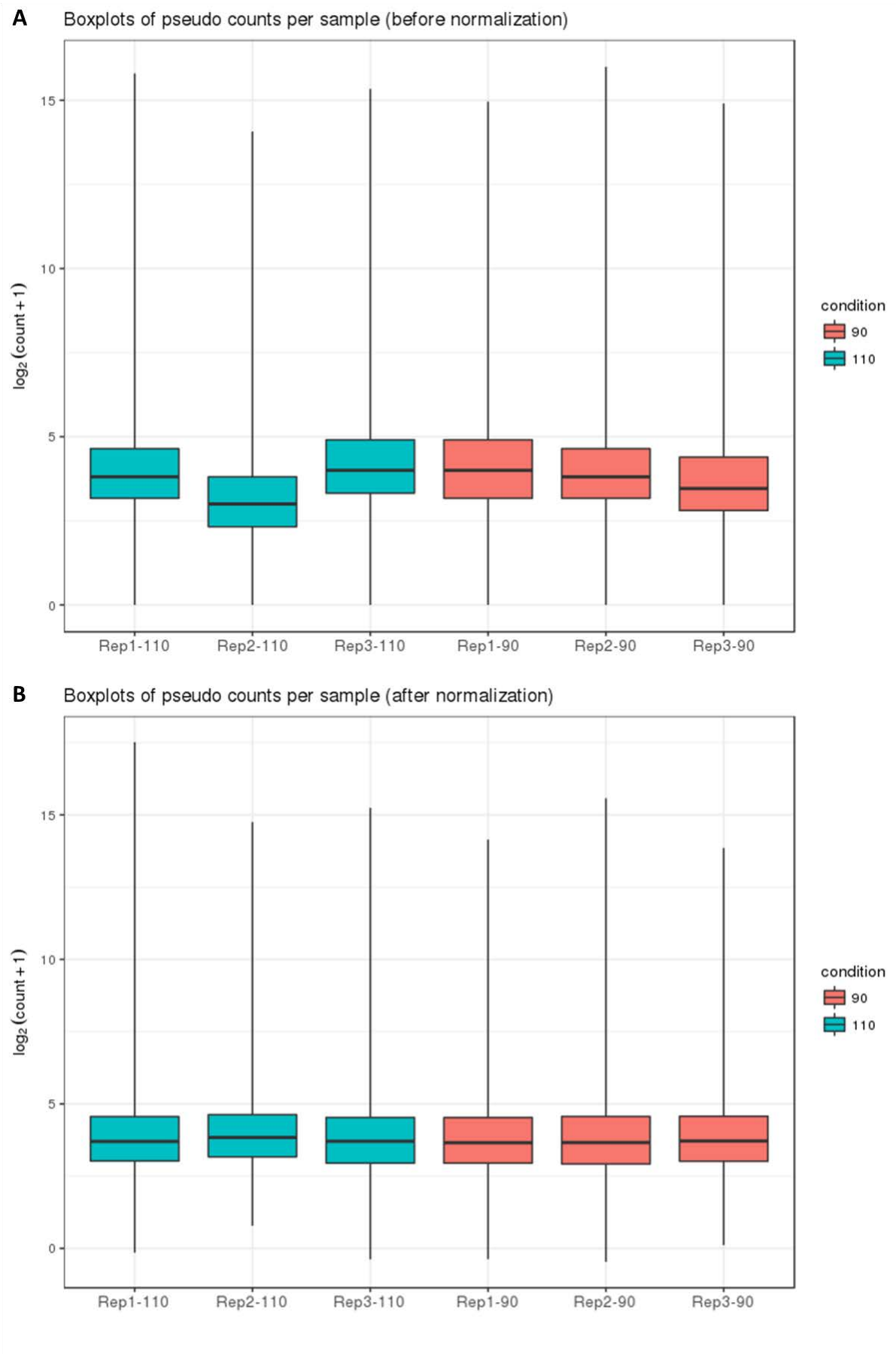


Figure 59. Distribution of raw (A) and normalized (B) counts per sample (200 Kb).

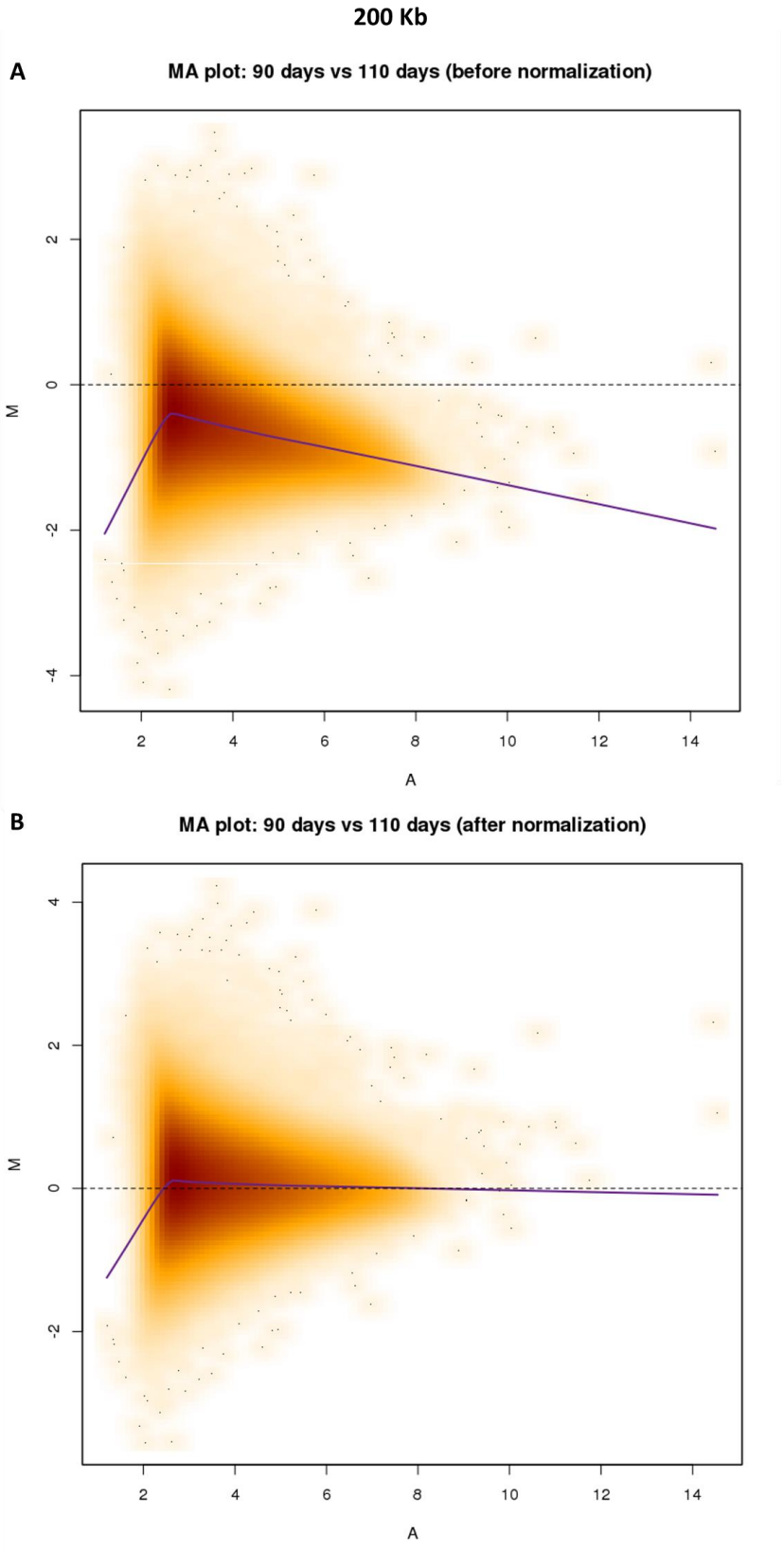


Figure 60. Global MA plot between samples at 90 and 110 days before and after normalization (200 Kb). *The MA plot represents for each bin pair (dots) the log-average count across all samples (A-value, x axis) and the average log-fold change between samples of different groups (M-value, y axis). The lowess fit (purple line) indicates the potential bias related to the average counting value (close to zero after normalization).*

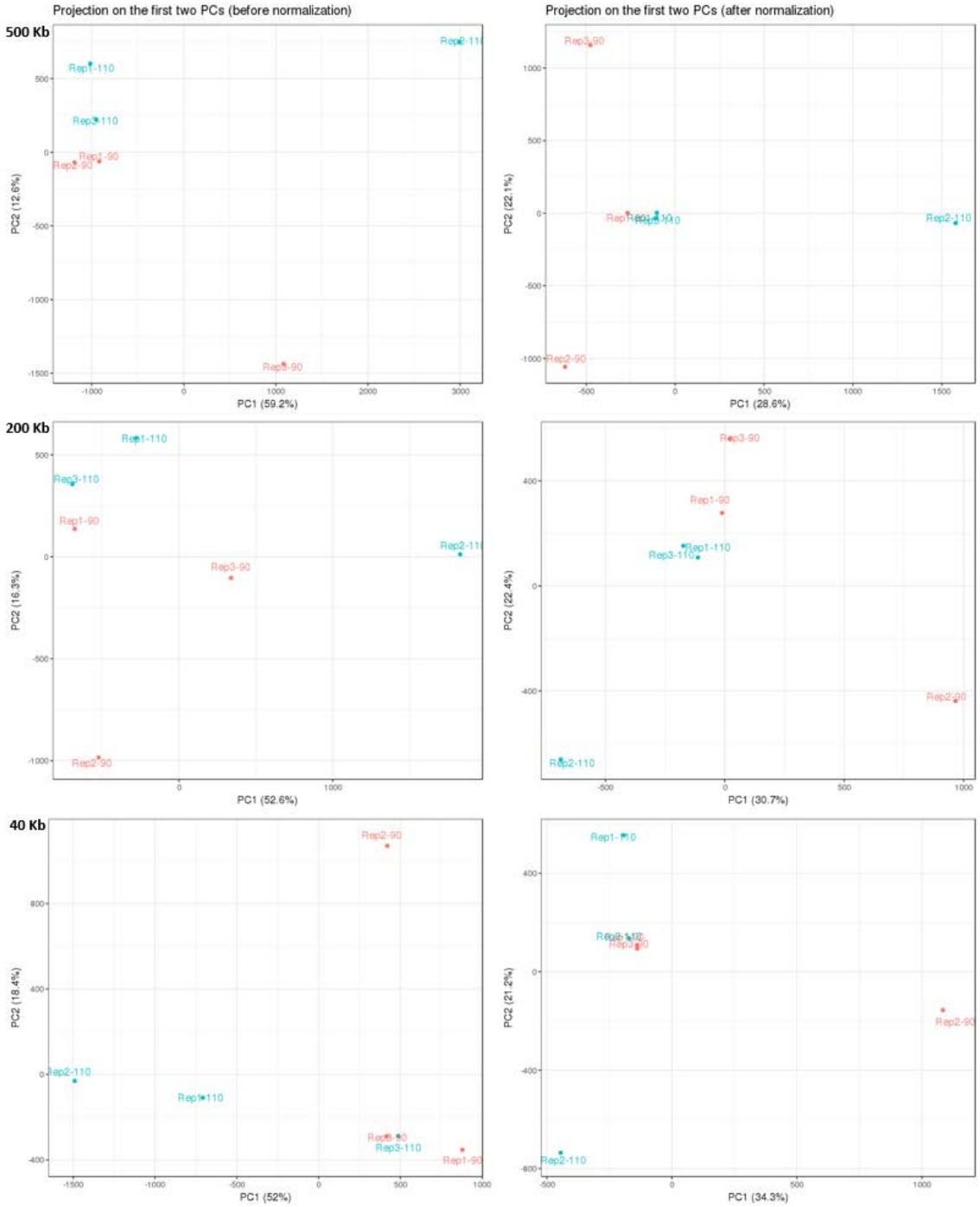


Figure 61. Principal component analysis of the samples using raw (left column) and normalized (right column) counts. *Data normalization resulted in separating the developmental stages along the first principal component (x axis). Replicates are shown in red (90 days of gestation) or blue (110 days).*

Table 10. Number and properties of the differential bin pairs

	500 Kb	**200 Kb**	**40 Kb**
Total bin pairs with any count	9,262,199	3,844,272	2,872,786
Differential bin pairs	10,183	3,417	83
% differential bin pairs	0.11	0.09	0.003
% differential bin pairs in *trans*	18.2	5.5	42.2
% differential bin pairs in *cis*	81.8	94.5	57.8
% differential bin pairs with logFC(+)	56.9	50.7	59.0
% differential bin pairs with logFC(-)	43.1	49.3	41.0

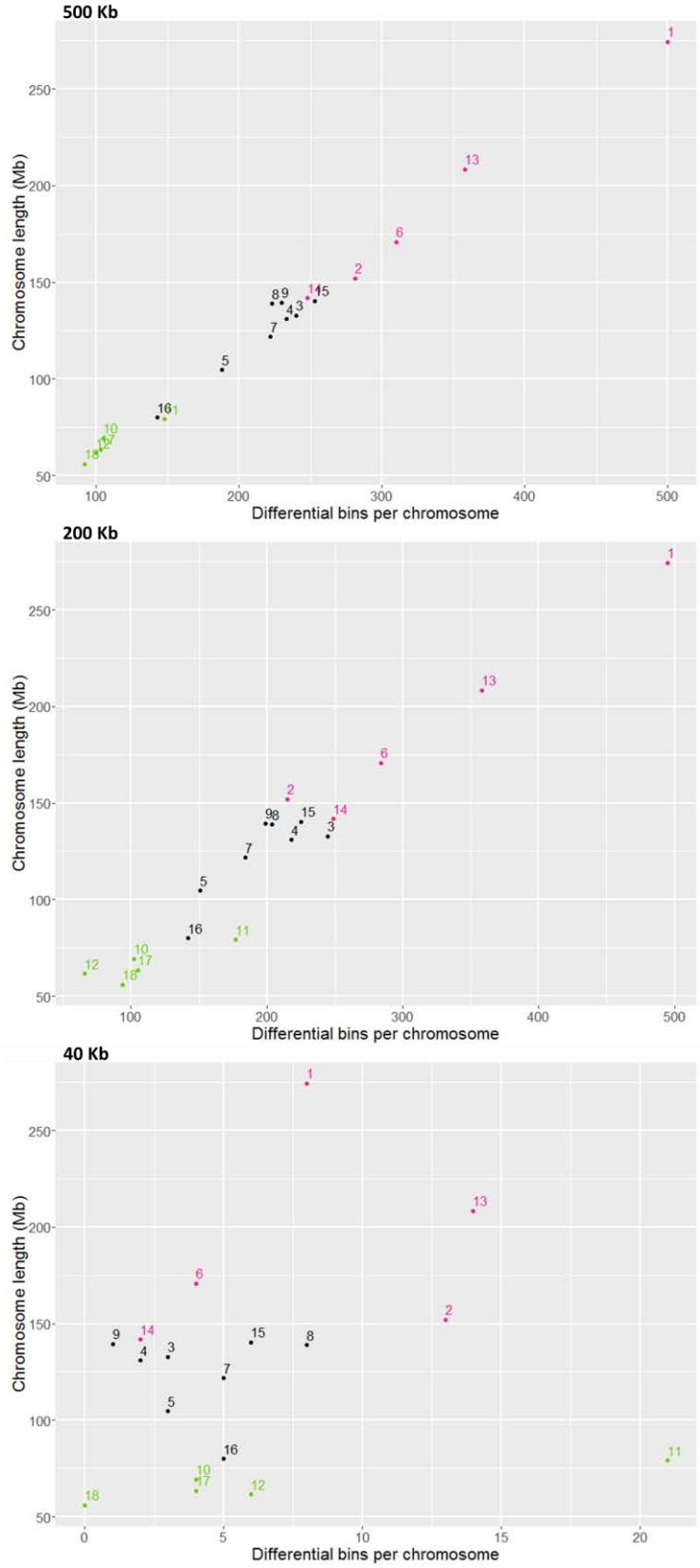


Figure 62. Distribution of differential bin pairs per chromosome at 500, 200 and 40 Kb resolution. *The number of differential bin pairs per chromosome are plotted against the chromosome length. The five smallest chromosomes are represented in green, and the five biggest ones in pink. The number of differential bins is globally correlated with the chromosome length. As expected because of lower counts, the higher the resolution the higher the variability.*

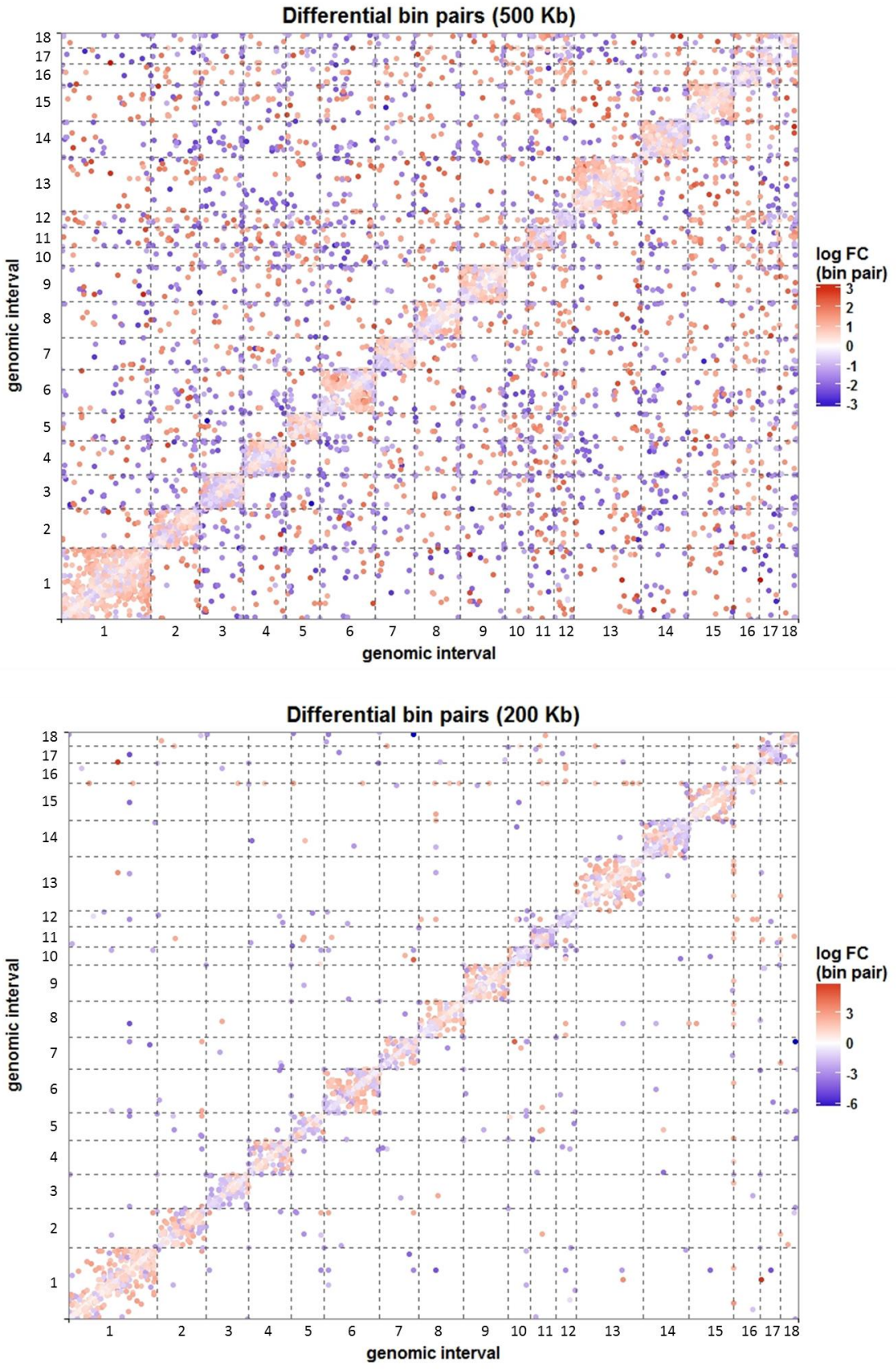


Figure 63. Differential bin pairs at 500 and 200 Kb resolution. *Each dot represents a bin pair which is associated to a log-fold change (logFC) value (blue-white-red gradient scale). Positive values of logFC correspond to genomic regions spatially closer at 110 days of gestation than at 90 days (red dots). Inversely, negative values correspond to genomic regions spatially closer at 90 days (blue dots). Only differential bin pairs of the 18 automosomes are represented.*

in order to detect local (fine) variations, the quantity of data needed to show significant changes between conditions is higher than for detecting global (large) changes. Approximately, 56% of the differential bin pairs on average showed a positive log-fold change (logFC), meaning that they bear significantly more counts at the 110 days condition than at the 90 days one. As the number of counts depends on the probability of a Hi-C religation event which in turn depends on the spatial distance between two genomic regions, these differential bin pairs with a positive logFC represent genomic regions that get closer from each other at 110 days of gestation than they were at 90 days. Inversely, ~44% of the differential bin pairs were found spatially closer at 90 days than at 110 days.

Then, we sought to investigate whether the distribution of these differential bin pairs across the chromosomes was homogeneous or whether some chromosomes were more represented than others. We observed that globally, the biggest chromosomes were those having the highest number of differential bin pairs (Figure 62). This confirms again that the results highly depend on the chosen resolution. Nevertheless, the global correlation between the chromosome sizes and the number of differential bin pairs suggests a widespread and homogenous distribution of the differential bin pairs along the genome. In order to normalize by the chromosome size, we computed for each chromosome the percentage of bins involved in at least one differential bin pair with respect to all bins in the chromosome. We noted that chromosome 11, with a relatively small length (79 Mb), presents the highest percentage of bins (93%, 45% and 1%) with respect to its size, involved in differential interactions in all three resolutions (500, 200 and 40 Kb respectively).

Then, among the bin pairs with a significant difference in read counts between conditions, we examined the distribution of *cis* and *trans* interactions, as well as the proportion of the ones with a positive logFC (significantly closer at 110 days) vs. negative logFC (significantly closer at 90 days). Figure 63 shows the positions of the differential bin pairs with a positive (red) and a negative (blue) logFC along the genome matrix (because of the low number of differential bins at 40 Kb only the other resolutions were shown). Globally, we observed a concentration of differential bin pairs along the diagonal (intra-chromosomes). The proportion of differential bin pairs with positive and negative logFC is highly heterogeneous across chromosomes (Appendix 17), suggesting various degrees of contribution to the topological difference between the developmental stages.

Another observation is that many differential bin pairs seem to be located close to the transitions between chromosomes (dotted vertical/horizontal lines, Figure 63), suggesting the presence of abundant *trans* interactions between terminal parts of the chromosomes. To investigate the positions of *cis* and *trans* interactions along the chromosomes with a better resolution, we used “circos plot” visualizations to represent significantly distal bin pairs. In these plots, we visualized separately the *cis* and *trans* differential bin pairs by representing the genome sequence as a circumference inside which, the relations between two genomic regions (a pair of differential bins) are represented as red (positive logFC) or blue (negative logFC) loops connecting the two regions (Figures 64 and 65). We first considered *cis* connections, and observed that some chromosomes show large genomic regions of differential bins with a specific logFC type (either positive or negative). For instance, chromosome 2 seems to be divided in two blocks: the first one (blue), smaller, with most of the bins having a negative log-fold change; the

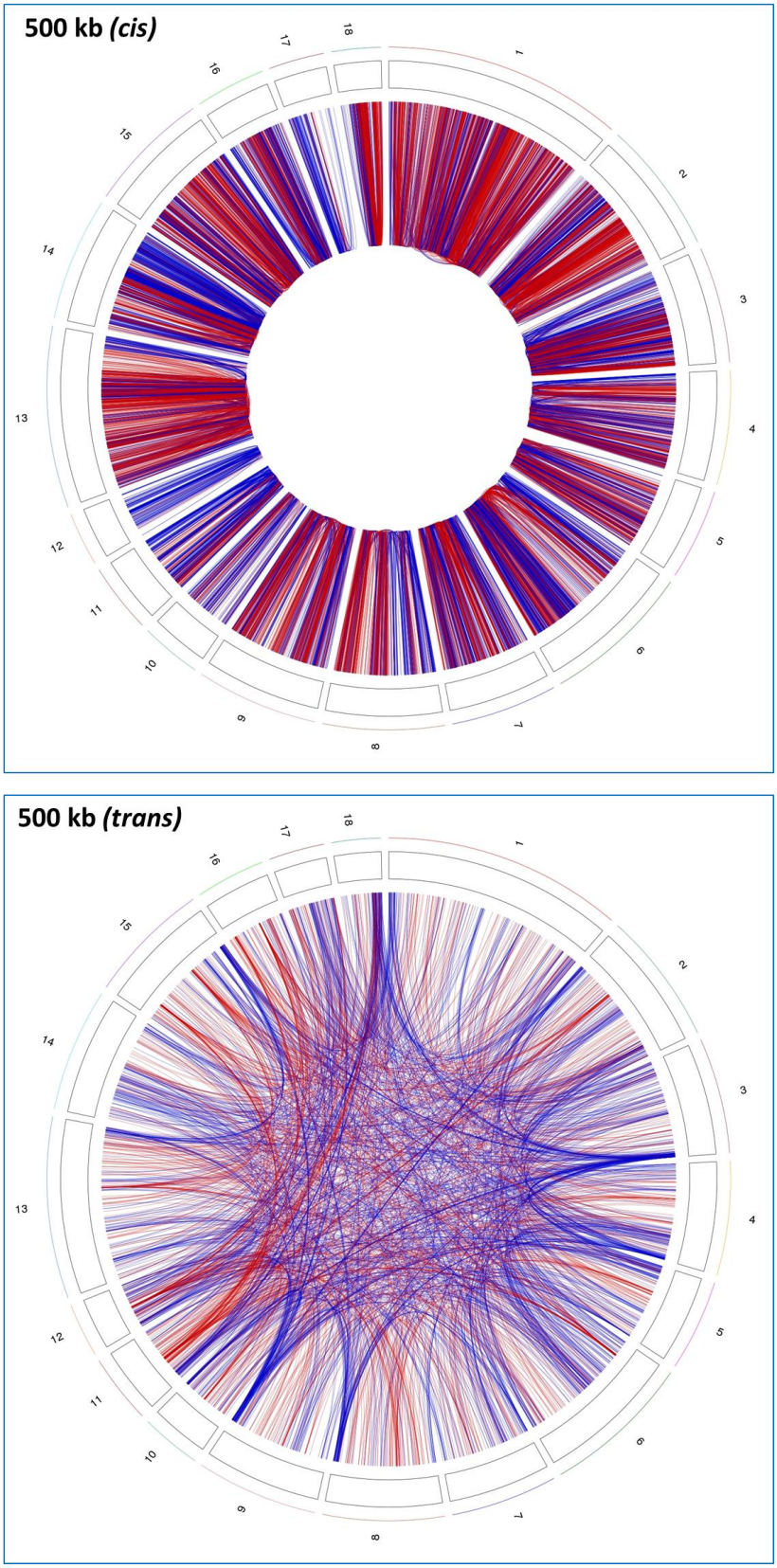


Figure 64. Distribution of differential bin pairs along the genome obtained at 500 Kb resolution. *Each bin pair is represented by a loop connexion between the two genomic regions (bins) involved in a differential bin pair. Chromosome positions are oriented clockwise. Differential bin pairs are represented in red (positive logFC) or blue (negative logFC). Upper panel: Differential bin pairs mapped to regions located on the same chromosome. Lower panel: Differential bin pairs mapped to regions located on different chromosomes.*

second one (red), mostly composed by bins with positive log-fold change (Figure 64, upper panel). Similarly, chromosome 3 seems to have a large blue region at the beginning, as well as chromosomes 6 and 14 show large blue regions in their second half. Other chromosomes show mixed profiles, alternating bins with positive and negative logFC (i.e. chromosomes 5 and 16). These results suggest that large portions of certain chromosomes contain genomic regions that behave in the same way by either becoming closer "condensation" or further "decondensation" from each other at 110 days of gestation with respect to their initial position at 90 days. These differentially distal regions define large chunks of adjacent regions behaving in the same way, similar as TADs or compartments but intrinsically dynamic because originating from a comparative analysis. Focusing on *trans* interactions allowed to confirm that many interchromosomal differential bin pairs seem to involve the extremities of the chromosomes, in particular with a negative logFC (Figure 64, lower panel). Moreover, these differential bin pairs seem to implicate the telomeric regions of both the "q" arm (*e.g.* chromosomes 3, 4, 8, 9, 10, 13 and 15) and the "p" arm (*e.g.* chromosomes 1, 2, 5 and 11), all with a prevalence of negative log-fold changes (blue connections). Density plots of *trans* vs. *cis* connections along each chromosome highlighted this trend for the *trans* connections to accumulate at the chromosome extremities (Appendix 18 and 19). In fact, by considering the first 5 and last bins of each chromosome as a "terminal region", about 4%, 10% and 38% of the *trans* interactions involved a terminal region at the 500, 200 and 40 Kb resolution respectively, while only 2%, 1% and 4% of the *cis* interactions did at the same resolution. This indicates a significant difference in the proximity of telomeric and subtelomeric regions between 90 and 110 days, being more proximal at 90 days than at 110 days, which might suggests a clustering of telomeric regions from different chromosomes in the 3D nuclear space at 90 days of gestation.

From this differential analysis, we concluded that it exists two global dynamic changes in muscle cells between the two gestational ages. The first one concerned intra-chromosomal interactions (global chunks of consecutive regions with a coordinated condensation/decondensation), and the second one concerned inter-chromosomal interactions with a strong component located at the chromosome extremities.

6.1.3.3 Functional analysis of differential bin pairs

We wanted to investigate in more details to which genes correspond the differential genomic regions and what their roles are. We wondered whether these differential regions are enriched in specific biological functions or not. In order to address this question, we performed a gene ontology (GO) analysis over all genes located in the differential bin pairs with a positive logFC, or a negative logFC obtained at 40, 200 and 500 Kb resolution. We searched for biological processes (BP), molecular functions (MF) and cellular components (CC) enriched among the human homologs of genes mapped to the differential bin pairs with respect to those mapped to all bin pairs. Obviously, this type of analysis highly relies on the quality of the genome annotation and on the proximity of the target genome with the human reference. Common functions enriched among genes found in differential bins with both, a positive and a negative logFC, were mainly biological processes referred to the synaptic transmission, signal transduction, metabolic processes and catalytic activity (Tables 11 and 12). The olfactory receptor activity was a specific biological process among differential bins with a positive log-fold change, while the response to stimulus was specific of differential bins with a negative log-fold change. No enriched functions were found at 200 and 500 Kb in differential bins with a negative log-fold change.

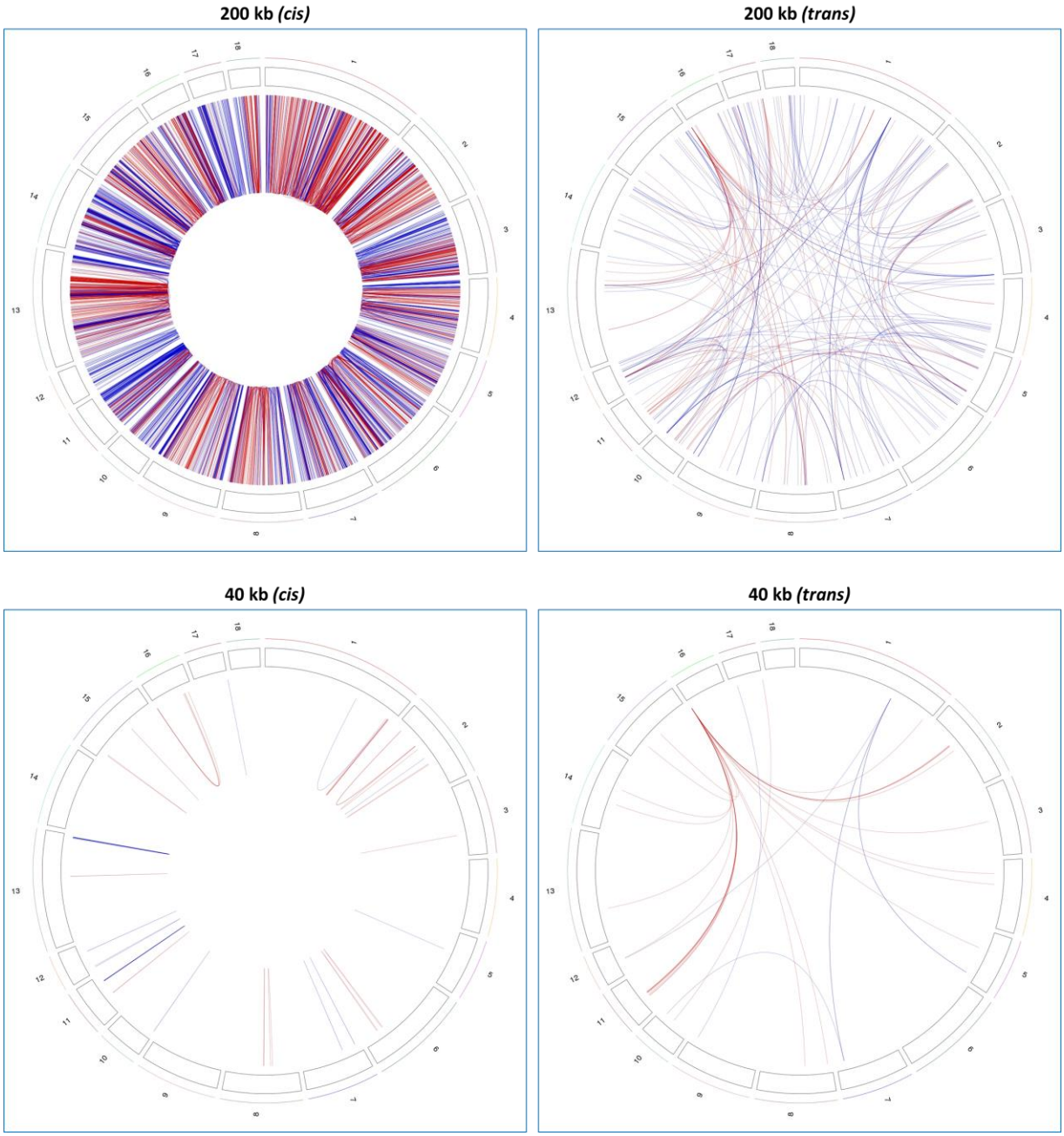


Figure 65. Distribution of differential bin pairs along the genome obtained at 200 and 40 Kb resolution. *Each bin pair is represented by a loop connexion between the two genomic regions (bins) involved in a differential bin pair. Chromosome positions are oriented clockwise. Differential bin pairs are represented in red (positive logFC) or blue (negative logFC). Left: Differential bin pairs mapped to regions located on the same chromosome. Right: Differential bin pairs mapped to regions located on different chromosomes.*

No apparent enrichment in muscle functions was detected from this analysis. This could be partly due to the specific nature of the topological differences between stages. Indeed, the main features revealed by the differential analysis (large regions of differential compaction in *cis* and telomere clustering in *trans*) tend to support a global reorganization of the genomic structure with a major structural component rather than a targeted gene expression regulatory program.

6.1.4 Gene expression and nuclear organization

In this last section we combined Hi-C and expression data to investigate whether changes in genome organization are linked to gene expression. To do that, we used the expression data obtained by microarray experiments in a previous study of the muscle transcriptome at 90 and 110 days of gestation (Voillet et al., 2014). In that transcriptome study, the expression data of 12,465 genes were measured by targeted microarray probes in samples from different breeds, including 8 Large White samples at 90 days of gestation and 9 Large White samples at 110 days. For each microarray probe, that study provided us two types of information: (1) an average expression value for both developmental stages (90 and 110 days); (2) statistical results from a differential analysis comparing both stages with a log-fold change (logFC) and an associated *p*-value. As the previous study was based on a former version of the reference genome and annotation -which highly impacts the microarray design- we decided to re-map the sequence of each probe on the more recent Sscrofa11 version and to keep only unambiguous matches with annotated exons from the Ensembl v90 annotation (see Methods). Importantly, while the previous study and the current project were conducted on different animals, we hypothesized that biological effects with strong and general impacts might be detected by broad integrative analyses.

6.1.4.1 Gene expression in A and B compartments

We first wanted to confirm the difference in gene expression that can be expected between A and B compartments. For each A or B compartment predicted from our Hi-C data at a given stage (90 or 110 days using the merged matrices), a mean expression value was computed by considering all the probes within the compartment across the 8 or 9 samples of the corresponding stage of gestation. The distributions of these average expression values in A and B compartments are shown in Figure 66. As shown in model organisms (Lieberman-Aiden et al., 2009; Rao et al., 2014), we observed a significantly higher gene expression in A vs. B compartments (Wilcoxon test, p-value < 2.2e-16 for both tests with 576 and 769 probes at 90 and 110 days respectively). Similar results were observed for the probes in A and B compartments obtained from the individual matrices (Appendix 20). This integration of transcriptome data from a previous project reveals a high consistency between 3D genome structure and function, considering that the transcriptome data were not obtained from the same animals than the ones used to perform the Hi-C experiments.

Table 11. Enriched GO terms found in genes mapped to differential bin pairs with a positive logFC.
Categories: BP (Biological Processes), MF (Molecular Functions) and CC (Cellular Components).

Resolu-tion	Category	GOBPID	Pvalue	Count	Size	Term
40 Kb	BP	0002091	0.002	1	5	negative regulation of receptor internalization
		0032482	0.002	1	5	Rab protein signal transduction
		0035418	0.004	1	11	protein localization to synapse
		0006677	0.004	1	13	glycosylceramide metabolic process
		0035640	0.004	1	13	exploration behavior
		0097503	0.006	1	20	sialylation
		0048488	0.007	1	22	synaptic vesicle endocytosis
	CC	0097060	0.002	2	200	synaptic membrane
		0030054	0.003	3	911	cell junction
		0060076	0.005	1	12	excitatory synapse
	MF	0008373	0.008	1	20	sialyltransferase activity
200 Kb	BP	0036150	3.00E-07	11	16	phosphatidylserine acyl-chain remodeling
		0007268	2.00E-06	120	641	synaptic transmission
		0036148	3.00E-06	10	16	phosphatidylglycerol acyl-chain remodeling
		0036152	6.00E-06	11	20	phosphatidylethanolamine acyl-chain remodeling
		0036149	8.00E-06	9	14	phosphatidylinositol acyl-chain remodeling
		0051966	1.00E-05	16	40	regulation of synaptic transmission, glutamatergic
		0050911	2.00E-05	25	85	detection of chemical stimulus involved in sensory perception of smell
		0051932	2.00E-05	13	30	synaptic transmission, GABAergic
		0052646	2.00E-05	13	30	alditol phosphate metabolic process
	CC	0005578	5.00E-06	65	303	proteinaceous extracellular matrix
		0044456	1.00E-05	72	353	synapse part
	MF	0004984	2.00E-05	25	85	olfactory receptor activity
500 Kb	BP	0050911	7.00E-08	70	85	detection of chemical stimulus involved in sensory perception of smell
	MF	0004984	6.00E-08	70	85	olfactory receptor activity
		0038023	5.00E-07	522	833	signaling receptor activity

Table 12. Enriched GO terms found in genes mapped to differential bin pairs with a negative logFC.

No enriched functions were found at 200 and 500 Kb in differential bins with a negative log fold change.

Resolu-tion	Category	GOBPID	Pvalue	Count	Size	Term
40 Kb	BP	0021897	7.00E-04	1	1	forebrain astrocyte development
		0051460	7.00E-04	1	1	negative regulation of corticotropin secretion
		1900011	7.00E-04	1	1	negative regulation of corticotropin hormone receptor activity
		0070593	1.00E-03	1	2	dendrite self-avoidance
		0071314	1.00E-03	1	2	cellular response to cocaine
		0060060	2.00E-03	1	3	post-embryonic retina morphogenesis in camera-type eye
		0097211	2.00E-03	1	3	cellular response to gonadotropin-releasing hormone
		0007162	2.00E-03	2	101	negative regulation of cell adhesion
		0048593	2.00E-03	2	102	camera-type eye morphogenesis
		0002125	3.00E-03	1	4	maternal aggressive behavior
		0035021	3.00E-03	1	4	negative regulation of Rac protein signal transduction
		0035865	3.00E-03	1	4	cellular response to potassium ion
		0007270	3.00E-03	2	115	neuron-neuron synaptic transmission
		0046929	4.00E-03	1	5	negative regulation of neurotransmitter secretion
		0042445	4.00E-03	2	137	hormone metabolic process
		0007406	4.00E-03	1	6	negative regulation of neuroblast proliferation
		0048842	5.00E-03	1	7	positive regulation of axon extension involved in axon guidance
	CC	0030424	6.00E-04	3	267	axon
		0031088	3.00E-03	1	4	platelet dense granule membrane
		0043196	4.00E-03	1	6	varicosity
		0005767	5.00E-03	1	7	secondary lysosome
	MF	0051424	7.00E-04	1	1	corticotropin-releasing hormone binding
		0016404	1.00E-03	1	2	15-hydroxyprostaglandin dehydrogenase (NAD+) activity
		0004719	2.00E-03	1	3	protein-L-isoaspartate (D-aspartate) O-methyltransferase activity
		0010340	3.00E-03	1	4	carboxyl-O-methyltransferase activity
		0008429	4.00E-03	1	5	phosphatidylethanolamine binding

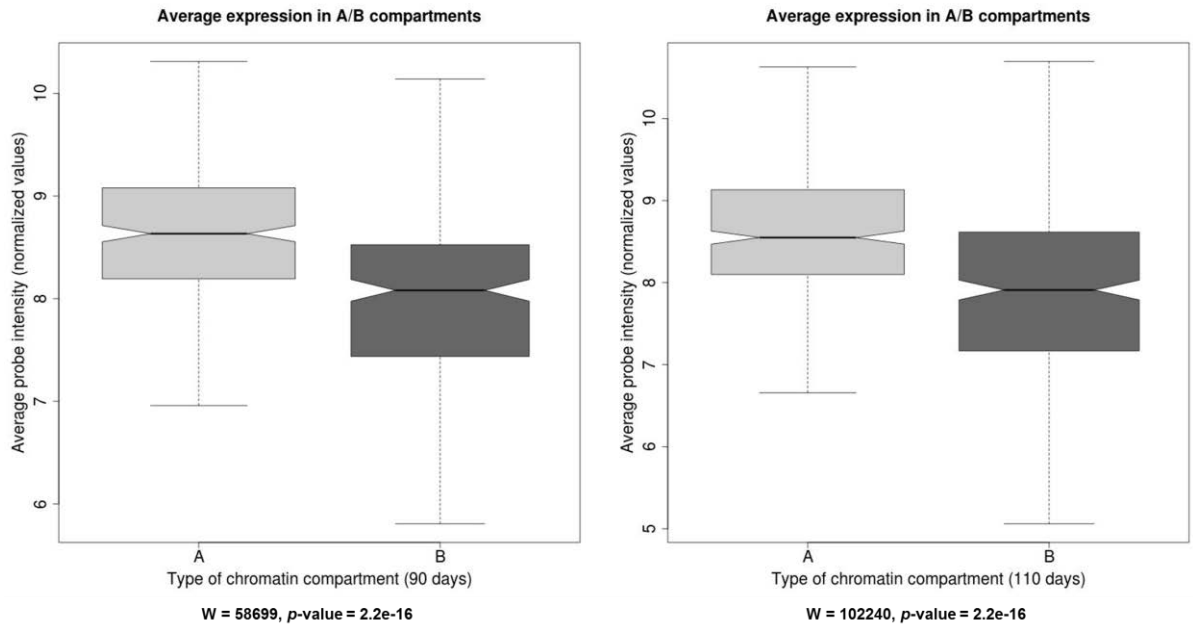


Figure 66. Average gene expression in AB compartments. *The A and B compartments were estimated from the merged matrices at 90 and 110 days of gestation. The gene expression data used to compute the average expression in both compartments was obtained from a muscle transcriptome study performed at 90 and 110 days of gestation respectively* (Voillet et al., 2014).

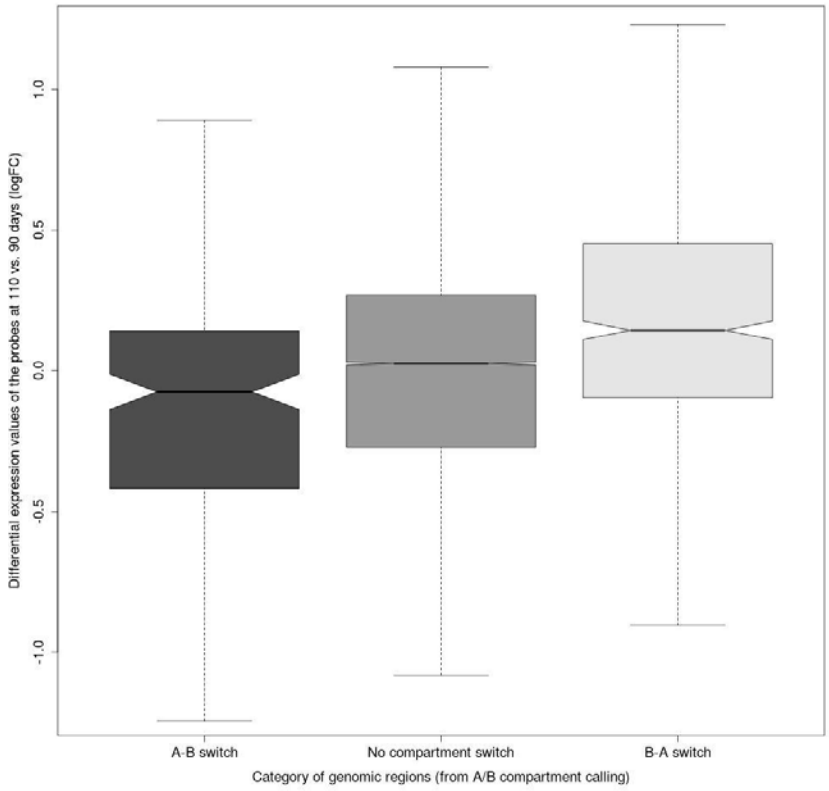


Figure 67. Distribution of differential expression values of probes mapped to genomic regions switching A/B compartment vs. probes mapped to regions with no compartment switch. *LogFC values of probes mapped to genomic regions: (left) switching from an A compartment at 90 days to a B compartment at 110 days (A-B switch); (middle) showing no compartment switch; (right) switching from a B compartment at 90 days to an A compartment at 110 days (B-A switch).*

6.1.4.2 Gene expression in A/B switching compartments

To better investigate potential correlations between high-order chromosomal compartments and gene expression, we compared results from the differential expression analysis with those from the compartment calling in the different stages. More precisely, we compared the distributions of the probe logFC values (110 days vs. 90) in distinct types of genomic regions according to the predicted A/B compartments. A positive logFC expression value indicates that the probe was significantly more expressed at 110 days than at 90 days. Inversely, a negative logFC expression value indicate that the probe was more expressed at 90 days than at 110 days. In particular, we considered the previously identified 2,809 regions (500 Kb bins) that showed a consistent A/B labelling in all 6 replicates on the one hand (see the "Differences in A and B compartments assignment" section) and the 104 regions with a consistent switch on the other hand. Within this second category, the distinction was made between A ➔ B and B ➔ A switches (switching sense: 90 ➔ 110 days). We could identify 26,083 probes in "conserved" regions, 200 in A ➔ B switches and 686 in B ➔ A switches. As shown in Figure 67, probes that mapped to A ➔ B switching regions showed lower log-fold changes than both probes in stable regions and probes in B ➔ A switching regions. Differences between these distributions were all statistically significant (Wilcoxon test, *p*-values equal to 1.2e-4 and 1.0e-15, respectively). In other words, genes in genomic regions that switch from an "active" state at 90 days of gestation to an "inactive" one at 110 days are likely to show a consistent decrease of expression, in line with previously reported results in human and mouse (Dixon et al., 2015).

Altogether, these results validate the biological relevance of the reported switching regions, as the observed differences in gene expression are consistent with the reported changes in the 3D genome structure. Although the differences in gene expression in these switching compartments were relatively subtle in terms of average logFC (-0.16 for A➔B vs. 0.23 for B➔A), they were significant. Moreover, as mentioned before, the expression values used in this analysis were not obtained from the same fetuses than those of the Hi-C experiments, supporting the hypothesis of a general and important regulatory mechanism.

6.1.4.3 Gene expression in differentially located genomic regions

To further study the relationship between expression and chromatin structure, we examined the expression profiles of genes located on the differential genomic regions responsible of the observed global differences in 3D genome structure between the two developmental ages (Figure 61). For that purpose, we investigated whether increases or decreases in gene expression could be associated with significant variations of the spatial proximity between genomic regions, similarly as we did with the switching A/B compartments. We therefore computed and compared again the distributions of differential expression values of probes mapped to different categories of genomic regions. Because some genomic region can be involved in both positive and negative log-fold changes (depending on the interacting partner), this time we considered separately regions that: (1) were not involved in any bin pair with a significantly different distance between conditions (2) were only reported in differential bin pairs with a significant *p*-value and a positive logFC, meaning a smaller 3D distance at 110 days vs. 90 days (3) inversely, regions that were only involved in significantly closer bin pairs at 90 days (negative

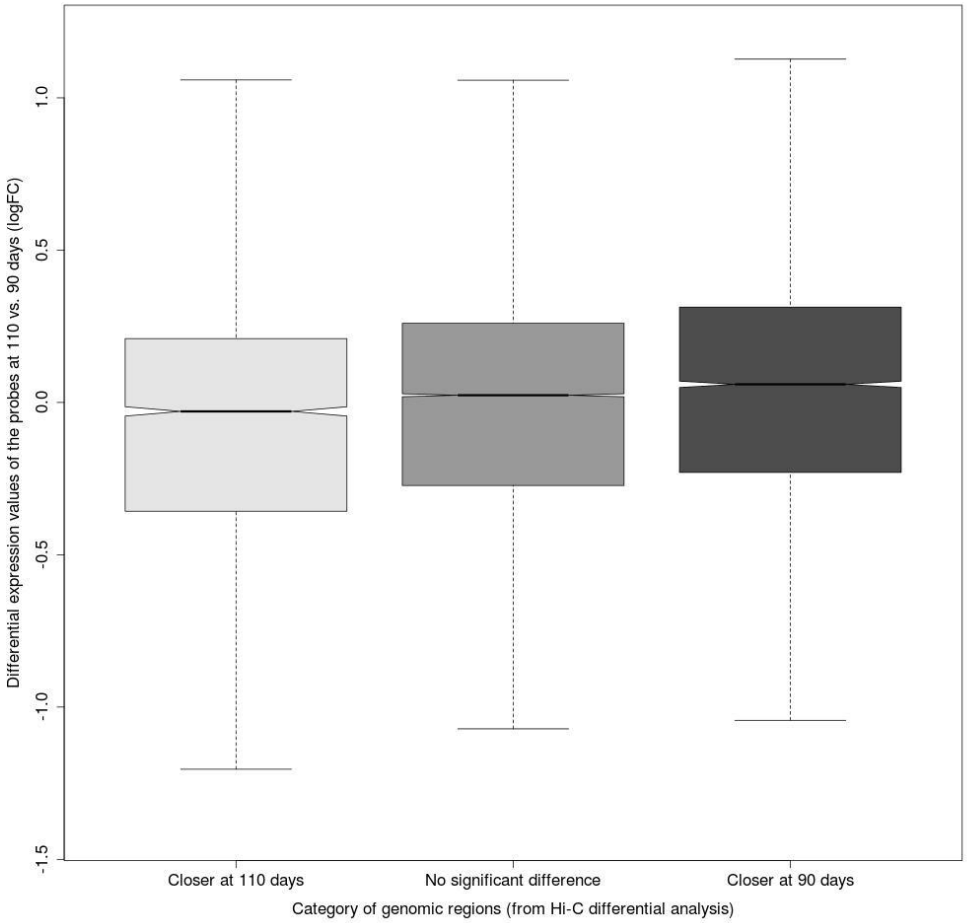


Figure 68. Distribution of differential expression values (logFC) of probes mapped to differentially located bin pairs (200 Kb resolution) with a positive or negative logFC vs. probes mapped to regions with no significant difference in spatial proximity.

logFC). The results of this analysis are shown in Figure 68. Interestingly, although the trends are relatively subtle again, the expression values of probes in genomic regions closer at either 90 days or 110 days of gestation are significantly lower both at 90 or 110 days (p-value < 2.2e-16, Wilcoxon test).

Altogether, these results suggest that the local variations in the 3D genome organization that we reported in fetal muscle may be involved in mechanisms of gene expression regulation occurring in late gestation, which supports our initial hypothesis.

6.2 Discussion

6.2.1 First insights in porcine muscle genome architecture at late gestation

The development of the High throughput chromosome conformation capture (Hi-C) method nearly 10 years ago (Lieberman-Aiden et al., 2009), has allowed to obtain a global view of the three dimensional (3D) architecture of genomes. Since the appearance of this approach, many studies about the 3D genome conformation have been carried on animal models such as human, mouse and fly (Dixon et al., 2012, 2015; Lieberman-Aiden et al., 2009; Rao et al., 2014; Sexton et al., 2012), as well as many other organisms such as fish, plants, yeast or bacteria (Dong et al., 2017; Kaaij et al., 2018; Kim et al., 2017; Marbouty and Koszul, 2015). However, to the best of our knowledge, no previous work has been published regarding the spatial organization of livestock genomes assessed by Hi-C. Although Hi-C assays were perfomed on goat blood samples, data were not used to study the chromatin structure but to improve instead *de novo* genome assembly of *Capra hircus* (Ghurye et al., 2017).

Indeed, the sequence and functional annotation of livestock genomes are not as well characterized as in model species. In order to fill this gap, the Functional Annotation of Animal Genomes (FAANG) initiative emerged recently to support and coordinate several projects on domesticated species (Andersson et al., 2015; Tuggle et al., 2016). As part of this initiative started the FR-AgENCODE project, a French pilot project centered on the functional analysis of pig, chicken, cattle and goat genomes (Foissac et al., 2018). This study has permitted to describe the global features of pig, chicken and goat genome architecture by obtaining the first Hi-C genome-wide contact maps in these species. Although this project has been proven highly valuable to describe the 3D genome organization of pig genome, the Hi-C assays were uniquely performed on a specific tissue (adult liver). Thus, this thesis has been conceived in parallel of the FR-AgENCODE project in order to investigate on another important tissue (muscle), in first place, the dynamic changes occurring at the level of the chromatin structure between two different conditions and next, to determine in which extent these variations in genome conformation are linked to differences in gene expression.

In the present project we focused on the study of porcine muscle tissue at 14 and 4 days before birth (the 90th and the 110th day of gestation respectively). During this period, important biological processes take place affecting the capacity of piglets survival at birth (Foxcroft et al., 2006; Rehfeldt and Kuhn, 2006; Rehfeldt et al., 2000). This makes our experimental design of special interest for the

agronomic research (in terms of meat production), as well as a potential biomedical model for human diseases and developmental issues considering the anatomical, physiological and genetic homologies between human and pig (Lunney, 2007). As far as we know, our study is the first performed on fetal muscle tissue regarding the 3D genome structure of muscle nuclei in this critical period of gestation. The closest study to our approach was published last year (Doynova et al., 2017) in which, Hi-C was used to capture the genome organization of mouse muscle progenitor cells before and after differentiation to myotubes. However, the experiments were performed on an *in vitro* model (cell lineage), thus not as close to real physiological conditions as we could investigate. Moreover, their *in vitro* model targeted a more premature stage of the muscle developmental process which includes the proliferation and differentiation of myoblast occurring during myogenesis (Buckingham, 2006; Chal and Pourquié, 2017; Yusuf and Brand-Saberi, 2012), while we focused on the maturity process of differentiated muscle fibers.

6.2.2 Adaptation of the *in situ* Hi-C protocol to porcine fetal muscle

To perform the Hi-C experiments we used the *in situ* Hi-C protocol (Rao et al., 2014) which reduces the risks to obtain spurious contacts due to random ligation events occuring in dilute solution (Nagano et al., 2015), when applying the initial *in solution* Hi-C protocol (Lieberman-Aiden et al., 2009). This is because in the *in situ* (or *in nuclei*) protocol, all main steps (crosslinking, digestion and religation of DNA) occur inside the nuclei. The *in situ* protocol was initially conceived to be applied in human embryonic and mouse lymphoblastoid cell lines. Indeed, most of the published work in this domain have been performed on cultured cell lines. One case of Hi-C experiments performed in fresh tissue was a study of chromosome conformation in developing human brain (Won et al., 2016). Our first challenge was to adapt the *in situ* Hi-C protocol to fetal muscle tissue preserved at -80°C in isopentane. The main difference when working with tissue compared with cell cultures is the need of a preliminary step of cell dissociation. Skeletal striated muscle is a quite particular tissue, as the specialized muscle fibers form a syncytium (fusion on thousands of individual muscle cells), which is surrounded by the sarcolemma (cell membrane) and the basement membrane (a support structure of connective tissue formed by collagen fibrils and laminin proteins) (Chal and Pourquié, 2017). Muscle nuclei have oval shapes because they are constrained by the basement membrane from the outside, and by myofibrils from the inside. As the preservation of nuclear shape is probably one of the most important factors when investigating 3D nuclear architecture, we paid special attention to this first step of nuclei isolation. Using an enzymatic treatment (collagenases) in order to dissociate both the sarcolemma and the basement membrane, resulted in nuclear lysis and loss of material. Indeed we found that when working with fetal frozen tissue, there is no need to use enzymatic digestion, a manual dissection of muscle fibers with scalpel blades followed by a filtration through a cell strainer and a quick fixation in order to preserve the nuclear shape, proved sufficient to isolate the nuclei. With these experiments, we demostrated that it is possible to perform Hi-C assays on frozen tissue, which may be advantageous in situations where, due to experimental contrains, the Hi-C assays cannot be performed on fresh tissue straight away after sampling.

6.2.3 High resolution porcine genome maps

As introduced in the "Hi-C resolution" section (see Chapter 3 of the bibliographic review), the sequencing depth is one of the most important factors to generate high resolution contact matrices. Thus, we sequenced our Hi-C libraries at high depth in order to obtain high resolutive 3D maps of the porcine genome. This allowed us to obtain a total of 3.45 billion read pairs for all our six libraries (575 M read pairs / library) from which, between 302 and 461 million per library were identified as valid read pairs. The Hi-C contact matrices described by (Rao et al., 2014) have been the most resolutive matrices obtained so far (between 395 M and 4.9 billion pairwise contacts in human maps from different cell lines and one mouse map). The number of sequenced read pairs obtained in some of these human libraries was in the same order of magnitude than the ones obtained in our experiment. However, the range of valid pairs was lower than the ones they obtained. This can be explained by: (1) differences in quality between the two reference genomes (affecting the mapping rates); (2) the selection criteria of valid pairs (pairwise contacts). Indeed, when analyzing one of their dataset with our pipeline, we obtained approximately the same proportion of valid/mapped pairs than in our libraries, which suggest that our selection criteria to discern read pairs issue from a Hi-C religation event may probably be more restrictive; (3) small variations or differences in performance of the experimental protocols, as observed in one of our Hi-C libraries (Rep2-110), which resulted less productive than the others. Despite this, even if lower, the quantity of valid data we obtained was in the same order of magnitude than the valid data reported in mouse (Rao et al., 2014). Indeed, comparing with the most relevant studies in this domain (Dixon et al., 2012, 2015; Lieberman-Aiden et al., 2009; Sexton et al., 2012), we obtained similar quantity of exploitable data, which underlines the high potential of our experimental design which, in addition, consisted in 3 replicates per condition, while in most of the studies, only one or two replicates were used.

Regarding Hi-C resolution we must discern between two confusing concepts: "matrix resolution" and "map resolution", as first described in (Rao et al., 2014). The first refers to the locus size (bin size) used to construct a contact matrix, and the second one was described as the smallest locus size such that 80% of the loci have at least 1,000 contacts. We obtained contact matrices at 500, 200 and 40 Kb matrix resolutions, on which at least 99% of loci have more than 1,000 contacts. This means that theoretically, we could have obtained Hi-C matrices at smaller bin sizes than 40 Kb while still keeping good mapping resolutions. However, we did not decrease the bin size because the criteria of map resolution to perform a differential analysis, as we did, must be higher in order to identify significant differences between the two conditions. Indeed, the number of differential bin pairs detected at 40 Kb was too low compared with 200 and 500 Kb. This suggests that at 40 Kb, the quantity of valid data is probably insufficient to target all relevant differences in chromatin structure existing between the two stages of development.

6.2.4 Main features of 3D genome folding in fetal muscle

Before building Hi-C contact matrices, we investigated the number of *cis* and *trans* read pairs among the total valid pairs. Interestingly, we obtained across all six replicates relatively higher

percentages of inter-chromosomal interactions (48% on average) than in the human and mouse dataset (Dixon et al., 2015; Rao et al., 2014) used for comparison (29% and 43% respectively). In addition the reported percentages of short-range (< 20 Kb) intra-chromosomal interactions (~ 2.4%) were much lower compared with the human and mouse datasets (12% and 22% respectively). We first wondered whether these discrepancies were a consequence of differences in the quality of the reference genomes. Thus, we compared the impact of a genome assembly improvement by running our pipeline on a small dataset with the previous (Sscrofa10) and current (Sscrofa11) genome versions. However, the observed decrease on *trans* bin pairs in Sscrofa11 with respect to Sscrofa10 was not important enough to elucidate all differences, and we suggested that cell-type specificities (i.e. nuclear shape or genome compaction) may explain the observed differences. Indeed, in the study performed in human embryonic stem (ES) cells and four ES derived lineages (from which we exported the human dataset (Dixon et al., 2015)) between 11% - 51% of *trans,* and 16% -53% of *cis* (< 500 pb) read pairs were reported, showing a high variability among the different cell types. The high percentages of *cis* short-range interactions observed in this study, might be due to the proliferating state of these progenitor stem cells, as during some points of the cell cycle (mitosis) the chromatin is found in its highest level of compaction forming chromosomes. In another Hi-C assay performed in human fetal brain tissue around 53% of *trans* valid pairs were reported (Won et al., 2016), similar to the 51% reported in neural progenitor cells (Dixon et al., 2015), but no information is available for the *cis*-short range in this tissue of differentiated cells. Another evidence of cell-type specificities on the number of *cis/trans* interactions, was the the results found in our aforementioned Fr-AgENCODE project (adult liver) (Foissac et al., 2018), where the percentages of *trans* (30% - 38%) and *cis* short-range valid pairs (6.4% – 6.8%) in hepatocytes were quite conserved across the three species (goat, chicken and pig). As mentioned before, muscle nuclei have a particular oval shape due to their peculiar location along the syncytium formed by the muscle fiber, which might explain the observed differences on the *cis/trans* ratio compared with other cell types. In fact, in a study performed in differentiated myotubes (Doynova et al., 2017), although the criteria for classifying *cis* short-range interactions was different (< 10 Kb instead of 20 Kb), the percentage was in the same order of magnitude (2.5% – 3%) than in our fetal muscle libraries. However, only 36-37% of *trans* read pairs were reported compared with our 41% - 52%. Nevertheless, we must consider that in this study, an *in vitro* instead of an *in tissue* model was used to investigate the genome organization of differentiated muscle cells, which might explain the differences. In addition, their Hi-C experiments were performed by using the *in solution* Hi-C protocol which has been proven to increase both experimental noise and bias and, more specifically, to reduce the reproducibility of long-range intra- and inter-chromosomal contacts (Nagano et al., 2015). Altogether, our results support the idea that the *cis/trans* contacts ratio may be more cell-type specific than species specific, and probably they can be explained by differences in nuclear shape and cell state (proliferating or quiescent cells).

In order to investigate the higher order structures of muscle genome, we obtained Hi-C contact matrices for individual chromosomes, as well as for the whole genome. The first observation was the high contrast between the density of intra- and inter-chromosomal contacts (Figure 45), displaying a clear delimitation of each chromosome in the whole genome matrix. These chromosomal structures correspond to the well-described chromosome territories (CTs) occupying discrete foci on interphase nuclei (Bolzer et al., 2005; Cremer and Cremer, 2001).

Then, we sought to investigate subchromosomal structures within the chromosome territories, the so-called A and B compartments. We identified about 682 compartments per replicate with a mean size between 1.5 and 2 Mb. These compartments are smaller than the ones previously observed in mouse (3 Mb median size) (Dixon et al., 2012). However, in our previous study on porcine liver (FR-AgENCODE project), we obtained A/B compartments which showed a mean size of ~ 3 Mb. These results suggest that there might be cell-type specific differences, apart from differences in the genomes and/or in the analysis method. Besides, we confirmed that, compared with the B compartments, A compartments show a higher density of genes, a higher gene expression, and a lower frequency of contacts, meaning that the chromatin is more decondensed (accessible), as previously described in (Lieberman-Aiden et al., 2009). We also observed that most of these compartments were highly conserved across all replicates, as previously reported in other studies (Barutcu et al., 2015; Doynova et al., 2017; Foissac et al., 2018), while their distribution is very heterogeneous across all chromosomes. For instance, some chromosomes (i.e. 1, 3, 15 and 16) seem highly segmented, while others (i.e. 5, 6, 8 and 17) show quite large compartments.

Beyond the CTs and the A/B compartments, we identified smaller chromatin structures defined as chromatin domains enriched in highly-self interacting regions, the so-called TADs (Dixon et al., 2012; Nora et al., 2012). TADs seem to play a role in coordinating the activity of groups of neighboring genes (Gibcus and Dekker, 2013). Indeed, TADs boundaries are enriched in insulator proteins (such as CTCF), histone marks associated to active promoters, and transcription start sites (TSS) (Dixon et al., 2012). Accordingly to this, we found a high density of genomic CTCF-binding sites around TAD borders, with a prevalence of “forward” CTCF sites at the beginning of the TADs and of “reverse” CTCF sites at the end of the TADs when considering the orientation of the CTCF-binding sites, as previously observed by (Rao et al., 2014). In fact, TAD boundaries have been suggested to be involved in the mechanism of loop formation, together with other proteins such as cohesin and RNAPII, which may need CTCF dimerization due to the convergent orientation of the two CTCF motifs present at the loop anchors (Björkegren and Baranello, 2018; Rao et al., 2014; Tang et al., 2015b). The mean size of our predicted TADs ranged between 181 and 309 Kb, which are considerably smaller than those initially described (~ 1 Mb) (Dixon et al., 2012; Nora et al., 2012), but similar in size than the “contact domains” and “physical domains” described in (Rao et al., 2014 and Sexton et al., 2012) respectively. Indeed, the TADs involved in loop formation have been proposed as “insulated neighborhoods” of approximately ~ 190 Kb, which can associate to form nested insulated neighborhoods through the formation of nested boundaries (Hnisz et al., 2016a). This suggests the existence of nested TADs organized in a hierarchical way, as previously described (Fraser et al., 2015), meaning that we possibly detected additional boundaries beyond those previously observed, as proposed by (Rao et al., 2014). This would explain the size difference. Another evidence that supports this hypothesis is that our TADs showed not only a high density of CTCF-binding sites at the boundaries, but also a depletion of these sites inside TADs, while, in Dixon et al. 2012, 85% of the CTCF-binding sites were found inside TADs. As just mentioned, this is probably because they did not find additional boundaries inside their TADs which might have contain CTCF-binding sites. In addition, these differences could also be explained because we used a different TAD detection method, the Armatus program (Filippova et al., 2014), instead of the directionally index

(DI) approach (Dixon et al., 2012). Indeed, Armatus has been proven to have higher sensitivity in recovering TAD boundaries than other methods (Forcato et al., 2017).

6.2.5 Major changes on chromatin conformation at late gestation

In this study, we have been able to detect dynamic changes in the chromatin structure of muscle nuclei occurring at late gestation (between the 90th and the 110th day). Some of these changes were global, (identification of many genomic regions showing a significant differential on the interaction frequencies); others were more specific, such as clustering of telomeric regions; and others were more subtle, such as the detection of few genomic regions switching between A and B compartments.

6.2.5.1 Switching compartments

Regarding the A and B compartments, although the vast majority of genomic regions have the same compartment assignment across replicates, 11% of them switched between the two conditions when comparing the two merged matrices. However, when being more restrictive by comparing uniquely genomic regions with a compartment assignment in all replicates, and with a total coherence between the three replicates of each condition, only 3.1% of the genomic regions switched compartment. These dynamic changes seem less important compared with some studies where extensive A/B compartment switches were observed. For instance, up to 25% of switches were reported between human embryonary stem (ES) cells and mesenchymal stem cells (MSCs) (Dixon et al., 2015), 12% between epithelial and a breast cancer cells (Barutcu et al., 2015), and 8% between progenitor and differentiated myotubes (Doynova et al., 2017). However, analyzing more in detail the different approaches used for compartment detection, we realized that these values are not strictly comparable. Indeed, in all these three studies, the genomic regions switching compartment were identified after merging all replicates for each condition, rather than requiring for a total consistency between replicates. Obviously, this approach leads to a different number of switching regions. Moreover, the A/B compartments were identified at different resolutions in each study. The choice of resolution might considerably affect the number and assignment of A/B compartments. Fine changes in compartment assignments that could not be detected at large bin sizes, might be easily detected when using smaller bin sizes, and consequently, the number of variable genome regions may increase. This would explain the high percentage (25%) of switching compartments found in (Dixon et al., 2015), as they used 40 Kb resolution matrices to determinate the A/B compartments, while we did the compartment calling on the 500 Kb resolution Hi-C matrices. Similarly, (Barutcu et al., 2015) obtained 12% of switches by using 250 Kb resolution matrices, and Doynova et al. observed 8% of switches in 400 and 500 Kb resolution matrices, the last being in the same order of magnitude than the number of changes we detected at similar resolutions.

Overall, when the different approaches are comparable, then it can be hypothesized that the magnitude of dynamic switches can be cell-type dependent, as reported in (Dixon et al., 2015), which observed huge differences in the number of switching compartments between different cell types. In this study, embryonic stem (ES) cells are derived in mesendoderm (ME) and mesenchymal stem cells (MSCs), being the first the initial progenitors and the last the most differentiated. ME cells and MSCs

showed 3.8% and 25% of switches with respect to ES cells respectively. It seems that the more divergent are the cell populations, the more important are the differences in chromatin structure. In this context, considering that we studied the dynamic changes between two populations of the same cell type (differentiated muscle fibers at two different points of the muscle maturation process), we observed a non negligible proportion (from 3% to 11%, depending on the approach) of genomic regions switching compartment types. These changes of chromatin state between the two conditions may potentially have a role in the regulation of gene expression, as variations in gene expression have been significantly associated to these switching compartments (further discussed bellow).

6.2.5.2 Dynamic interacting regions

Our differential analysis method allowed us to identify 10,183, 3,417 and 83 differential bin pairs at 500, 200 and 40 Kb resolution respectively between the two developmental stages. These differential bin pairs reflect dynamic interacting regions distributed all over the genome, and might be responsible of major changes in chromatin structure occurring between the 90th and the 110th day of gestation that explain the separation we observed between the two conditions (Figure 61). We were able to detect much more dynamic interacting regions, compared with the myogenesis *in vitro* model study performed in mice (Doynova et al., 2017), where only 55 differentially bin pairs were reported between myoblast and myotubes (400 Kb resolution). We could expect higher global changes during differentiation (myogenesis) than in our model of differentiated muscle fibers at two relatively close developmental stages. This is probably because we have a much better map resolution, which allowed us to identify more subtle changes. The functional analyses performed on these differential bin pairs show an enrichment in biological processes related to synaptic transmission, signal transduction, metabolic processes and catalytic activity. No apparent enrichment in muscle-associated functions was observed, unless the synaptic transmission refers in this case to the neuromuscular contraction. It must be considered that at 500 and 200 Kb resolution the differential genomic regions contain too many genes to be able to target specific genes. On the other side, althought at 40 Kb we could target more fine (gene scale level) differential genomic regions, the quantity of data was probably not enough to allow us identifying relevant differences in chomatin structure associated to expression regulatory programs. Further sequencing would be necessary in order to improve the results obtained in the differential analysis at 40 Kb resolution.

Among all differential bin pairs, we highlighted two interesting findings involving several related genomic regions. The first concerns large chromosomal adjacent regions and the second one involved telomeric regions of most of the chromosomes.

6.2.5.3 Differentially distal adjacent regions

Interestingly, we observed large genomic regions of adjacent differential bin pairs that exhibit the same dynamic behavior when comparing the two gestational ages. Specifically, we found two large clusters in chromosome 2 that seem to correspond each to a chromosome arm, with a high density of differential bin pairs with a negative log-fold change in the p arm, and a positive log-fold change in the q arm. This indicates that the p arm becomes less condensed at 110 days of gestation and the q arm more

condensed, which suggests that globally, genes located on the p arm may show a more inactive state than those located on the q arm at the end of gestation. Similar results were observed on the fly genome, where higher-order clusters corresponding to each chromosome arm were organized into active and inactive clusters (Sexton et al., 2012). However, unlike in our study, this was not associated to dynamic changes because this study was mostly focused on an exhaustive description of 3D folding features in the fly genome. Beyond these clusters found on chromosome 2, similar large structures were observed for instance in chromosomes 6 and 14 (both with a negative log-fold change) or 1 and 13 (positive log-fold change), however, they did not involve the whole chromosome arm. It remains to verifiy if genes located in these dynamic adjacent regions are related and/or whether they show a coordinated regulation of gene expression, in which case we could suggest that the chromatin remodeling of large adjacent regions explain in part a coordinated regulation of related genes.

6.2.5.4 Inter-chromosomal telomeres clustering

Another interesting finding was that many inter-chromosomal differential bin pairs involved the telomeric regions of many different chromosomes (at least nine among eighteen) located in either the p or the q arm. Some of them, such as telomeric regions in the q arms of chromosomes 3, 9 and 15, involved differential bin pairs with telomeric regions belonging to at least four different chromosomes. Moreover, most of them showed a negative log-fold change indicating than telomeres exhibit dynamic coordinated nuclear organization in muscle cells during late development. More specifically, this suggests that telomeres seem to be preferentially clustered at 90 days of gestation and might dissociate later at 110 days. Indeed, these telomeres changes could be possible since telomeres have been observed to display rapid movements in live human cells (Wang et al., 2008). Similarly, preferential contacts between telomeres have been reported in fly embryonic nuclei, but these contacts were not associated to dynamic changes (Sexton et al., 2012). In another study, telomeric and sub-telomeric regions were found to display more frequent interactions in epithelial cells than in breast cancer cells (Barutcu et al., 2015), however these interactions were only intra- but not inter-chromosomal, meaning that some chromosomes bend to bring in contact their two extremities. This phenomenon of telomeres clustering has been also observed in yeast meiotic and quiescent cells (Guidi et al., 2015; Lazar-Stefanita et al., 2017; Yamamoto, 2014). Also in yeast, the telomere clustering has been associated to the formation of foci in which silencing factors concentrate, and it has also been proved the dynamic nature of aggregation or dissociation of these clusters (Hozé et al., 2013). There are also evidences of telomere clustering in mammals both in somatic cells and gametes (Solov'eva et al., 2004). For instance in human cancer and mouse cell lines, dynamic associations and dissociations of a subfraction of telomeres have been also observed in quiescent mammalian cells (Molenaar et al., 2003). In human fibroblasts, telomeres are known to associate preferentially in interphase nuclei than in their cycling counterparts (Nagele et al., 2001), and long telomeres have been observed to be involved in forming chromosome loops that can affect the higher order chromatin structure and gene expression (Robin et al., 2014). Interestingly, this study was performed in human myoblasts where it was proposed that telomere length-dependent long-range chromosomal interactions may repress gene expression by silencing genes close to the telomere. Or it may inversely enhance gene expression by activating those genes when telomeres became shorter with cellular aging. Moreover, a strong clustering of telomeres has also been

reported in porcine neutrophils and lymphocytes (Yerle-Bouissou et al., 2009). Another study focused on the *cis* telomeric associations in neutrophils revealed that when telomeric associations occur, the assotiations of p and q arm from the same chromosome are more frequent (Mompart et al., 2013). Besides, on one side, the SMARCA4 subunit of the SWI/SNF complex, which has a potential role in tissue-specific gene regulation during embryonic development, has been suggested to play a role in three-dimentional organization of telomeric regions (Barutcu et al., 2016). On the other side, the ATPase subunit of this same SWI/SNF complex has also been found to be required for the formation of inter-chromosomal interactions contributing to changes in gene positioning during myogenesis and temporal regulation during myogenic transcription (Harada et al., 2015).

Our results of inter-chromosomal clustering of telomeric regions at 90 days of gestation, together with the aforementioned studies related to telomeres associations, suggest the possibility of a specific dynamic mechanism of gene expression regulation in fetal muscle cells through temporal formation-disruption of telomere clusters. Further studies by using 3D DNA FISH will be necessary to confirm this hypothesis.

Interestingly, similarly to telomeres yet less obvious, we observed that some differential bin pairs seem to involve the centromeric regions of few chromosomes (i.e. chromosomes 2, 5, 8, 10, 11 and 12) but in this case, they show a positive log-fold change. This suggest that centromeres might cluster preferentially at 110 days. This phenomenon of centromeres clustering has been previously observed in different studies (Botta et al., 2010; Sexton et al., 2012; Yerle-Bouissou et al., 2009). However, we were not able to prove it since their genomic location it is not available in the reference genome sequence.

6.2.6 Genome organization and gene expression

In order to investigate whether the observed structural changes in 3D genome folding (switching A/B compartments and differential bin pairs) were related to variations in gene expression, we integrated to our study muscle expression data obtained on fetuses of 90 days and 110 days gestational ages. Regarding A/B compartments, we observed that probes mapped to A ➔ B switching regions (switching sense: 90 days ➔ 110 days) showed significantly lower fold changes than those mapped to B ➔ A switching regions. This suggests that at 110 days of gestation, there is a downregulation of gene expression in these genomic regions, which seems to be associated to structural variations of the chromatin state (switch from an “active” state at 90 days of gestation to an “inactive” one at 110 days). Inversely, switches from B to A seem to be associated to upregulated genes in these genomic regions. This was in agreement with results reported previously in human and mouse (Barutcu et al., 2015; Dixon et al., 2015; Doynova et al., 2017; Won et al., 2016). Similarly, we found that the expression values of genomic regions (differential bin pairs) significantly closer either at 90 days or 110 days of gestation are significantly lower than in more distant regions. Although significant, the differences in gene expression, both in switching compartments and differential bin pairs, were subtle. This suggests that variations in chromatin structure, especially when considering large genomic regions, do not always imply a global regulation of gene expression but rather indicate fluctuations in the expression levels of

a subset of genes located in the interrogated region. In fact, the integration of gene expression and differentially located regions was done at 200 Kb resolution. When we used the differential bin pairs reported at 500 Kb, no significant differences in the logFC expression were found when comparing with stable regions (data not shown). This is probably because at 500 Kb the differentially located regions are too large to target genes potentially regulated (to many genes per genomic region) and/or because the genomic regions involved in differential genome conformation are not sufficiently specific of distancing/approaching phenomena. Despite these subtle but significant variations, these results strongly support our initial hypothesis that the differences in gene expression previously reported between the two developmental stages (Voillet et al., 2014), are at least in part associated to chromatin remodeling.

7 General conclusion

This project has permitted to explore the relations between 3D genome organization and gene expression. More specifically, it has allowed to shed light on the main changes occurring at the level of chomatin structure in porcine developmental muscle, which are associated at least in part with variations in gene expression.

Our first study, in which a single-cell approach (3D DNA FISH) was used to asses the nuclear proximity of a selected group of genes, allowed us to reveal interesting associations involving *IGF2*, *DLK1* and *MYH3* genes, all of them related to muscle development (Schiaffino et al., 2015; Van Laere et al., 2003; Waddell et al., 2010). Moreover, we developed an innovative approach of gene co-expression network inference in which, by means of integrating information of gene nuclear co-localizations, we were able to obtain consistent, robust and reliable gene co-expression networks. As these networks were build from genes differentially expressed in fetal muscle of two extreme breeds in terms of survival, the information generated by these networks, brought to light relevant functions involved in the development and maturity of the fetal muscle. In addition, we proposed the *MYOD1* and *CTNNB1* transcription factors as potential co-regulators of the aforementioned *IGF2* and *DLK1* genes that we found co-localized in muscle nuclei. Globaly, we proved that by combining biological information of spatial proximity between genes, with pairwise partial correlations between gene expression levels, we are able to highlight a network of muscle-specific interrelated genes.

In our second study, we investigated the 3D genome organization at a larger scale, by using a population-based method approach (Hi-C), which allowed to explore all genomic regions found in proximity in muscle cell nuclei. This study has permitted to provide the first 3D maps of the porcine muscle genome at 500, 200 and 40 Kb resolution, as well as to determine major chromatin structures such as the A/B compartments and TADs. More important, we have identified genomic regions showing significant differences in chromatin structure between the two gestational ages. Interestingly, a considerable proportion of these genomic regions involved the telomeric regions of several chromosomes, which seem to preferentially cluster at 90 days of gestation compared with 110 days. In addition, our data suggest that differences between conformations at the two developmental stages, can explain a part of the variability between conditions. Moreover, althought the A/B compartments were mostly conserved across replicates, we identified few genomic regions changing of compartment type between the two gestational ages. We proved an actual link between chromatin conformation and gene expression by first confirming that the gene expression was significantly higher in A vs. B compartments as expected. Second, we observed that swiching from an A compartment (at 90 days) to a B compartment (at 110 days) was accompaigned by a slight but significant decrease in gene expression at 110 days, which is consistent with the known genomic features of B compartments (related to close inactive regions associated to heterochromatic histone marks). Third, the genomic regions exhibing significant differences in chromatin conformation, showed as well subtle but significant differences in gene expression. More specifically, those regions significantly closer both at 90 days or at 110 days of

gestation, showed a significant decrease in gene expression than those regions significantly far from each other in the corresponding stage of gestation.

Altogether, these new insights would help us to understand possible mechanisms of gene expression regulation dependent on genome structure in fetal porcine muscle, which is a valuable information in the context of the agronomic research. Further functional studies will be still necessary to uncover which are those mechanisms (potentially involved in muscle development and the establishment of muscle maturity in pig). Meanwhile, this thesis has allowed to characterize the main structural changes occurring in the 3D genome organization at late gestation, where important variations in the expression of genes related to muscle maturation process have been described.

8 Perspectives

In order to further exploit our data, it would be interesting in the short term to identify A/B compartments at smaller matrix resolutions than we did, in order to explore the impact of resolution on the compartment assignments and, eventually, to identify new genomic regions switching compartment type. The obtainment of Hi-C matrices at smaller bin sizes (i.e. at 10 Kb resolution), if we still keep a good "map resolution", would allow us to search for loop structures as previously described in (Rao et al., 2014). As loop structures are known to be involved in mechanisms of gene expression regulation (by bringing genes and distal regulatory elements in proximity), these would be a first step to identify potential regulatory elements of target genes such as distal enhancers. In line with this, it would be highly valuable to integrate our Hi-C data with ChIP-seq (chromatin immunoprecipitation sequencing) data, which would allow us to capture DNA sequences bound by proteins (RNAPII, H3K36me3, H3K79me2, H3K27ac, H3K4me1, and H3K27me3) associated to transcriptionally active or inactive regions. As it would be worth repeating the ChIP-seq experiment targeting the CTCF protein, as well as the SMARCA4 subunit of the SWI/SNF complex which has been found to be involved in the telomere structure, but also found enriched in open chromatin regions and TAD boundaries (Barutcu et al., 2016).

Obviously, if we could further increase the sequencing depth of our Hi-C libraries, we will be able to achieve a better resolution in order to target structural variations of specific genes (even of the regulatory sequences of those genes), as well as to allow performing the differential analysis at lower resolutions. Moreover, if we could achieve such a level of resolution, we will be able to detect chromatin contacts between pairs of genes at the whole genome scale. This information, combined with the appropriate expression data, could be used in our model of network inference in order to extend the approach by using data of gene-gene interactions at the whole genome level, which will allow us to obtain highly relevant and informative gene co-expression networks.

In addition, it would be interesting to use expression data from RNA-seq assays, which would be more appropriate to be integrated with our Hi-C data. Indeed, the expression data used in our study were obtained from a porcine microarray, in which probes related to adipose tissue, immune system and skeletal muscle specific genes were overrepresented. On the microarray some genes were characterized by several probes while others were represented by a unique probe, and probes were designed when the available reference genome was still of relatively low quality. Using RNA-seq data would allow us to have a better representation of the whole genome transcripts, and a more accurate measure of gene expression levels, which would rend the expression and chromatin conformation data more comparable.

Because we studied a diploid genome, the results obtained for each chromosome are indeed a mixture of chromatin structures from the two homologs. It would be interesting to investigate whether we obtain the same chromosome folding patterns between the paternal and maternal homologs. In order to do this, we could detect allele-biased genomic regions in terms of chromatin structure by identifying SNPs overlapping to our reads. These new results could be integrated with the new RNA-seq data to

explore whether allelic imbalances in gene expression (i.e. genes subject to genomic imprinting) are associated to allelic differences on the interacting frequencies.

Finally, as we observed that the telomeric regions seem to preferentially cluster at 90 days of gestation, it would be interesting to perform 3D DNA FISH experiments in order to find out whether these clusters are just more prevalent at 90 days than at 110 days or whether they remarkably dissociate at 110 days.

9 References

Al Adhami, H., Evano, B., Le Digarcher, A., Gueydan, C., Dubois, E., Parrinello, H., Dantec, C., Bouschet, T., Varrault, A., and Journot, L. (2015). A systems-level approach to parental genomic imprinting: the imprinted gene network includes extracellular matrix genes and regulates cell cycle exit and differentiation. Genome Res. *25*, 353–367.

Andersson, L., Archibald, A.L., Bottema, C.D., Brauning, R., Burgess, S.C., Burt, D.W., Casas, E., Cheng, H.H., Clarke, L., Couldrey, C., et al. (2015). Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. Genome Biol. *16*, 57.

Aranda, S., Mas, G., and Di Croce, L. (2015). Regulation of gene transcription by Polycomb proteins. Sci. Adv. *1*, e1500737.

Archibald, A.L., Bolund, L., Churcher, C., Fredholm, M., Groenen, M.A.M., Harlizius, B., Lee, K.-T., Milan, D., Rogers, J., Rothschild, M.F., et al. (2010). Pig genome sequence--analysis and publication strategy. BMC Genomics *11*, 438.

Ayuso, M., Fernández, A., Núñez, Y., Benítez, R., Isabel, B., Barragán, C., Fernández, A.I., Rey, A.I., Medrano, J.F., Cánovas, Á., et al. (2015). Comparative Analysis of Muscle Transcriptome between Pig Genotypes Identifies Genes and Regulatory Mechanisms Associated to Growth, Fatness and Metabolism. PloS One *10*, e0145162.

Ballman, K.V., Grill, D.E., Oberg, A.L., and Therneau, T.M. (2004). Faster cyclic loess: normalizing RNA arrays via linear models. Bioinforma. Oxf. Engl. *20*, 2778–2786.

Bantignies, F., and Cavalli, G. (2011). Polycomb group proteins: repression in 3D. Trends Genet. TIG *27*, 454–464.

Barbosa-Morais, N.L., Irimia, M., Pan, Q., Xiong, H.Y., Gueroussov, S., Lee, L.J., Slobodeniuc, V., Kutter, C., Watt, S., Colak, R., et al. (2012). The evolutionary landscape of alternative splicing in vertebrate species. Science *338*, 1587–1593.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823–837.

Barutcu, A.R., Lajoie, B.R., McCord, R.P., Tye, C.E., Hong, D., Messier, T.L., Browne, G., van Wijnen, A.J., Lian, J.B., Stein, J.L., et al. (2015). Chromatin interaction analysis reveals changes in small chromosome and telomere clustering between epithelial and breast cancer cells. Genome Biol. *16*, 214.

Barutcu, A.R., Lajoie, B.R., Fritz, A.J., McCord, R.P., Nickerson, J.A., van Wijnen, A.J., Lian, J.B., Stein, J.L., Dekker, J., Stein, G.S., et al. (2016). SMARCA4 regulates gene expression and higher-order chromatin structure in proliferating mammary epithelial cells. Genome Res. *26*, 1188–1201.

Bell, J.L., Wächter, K., Mühleck, B., Pazaitis, N., Köhn, M., Lederer, M., and Hüttelmaier, S. (2013). Insulin-like growth factor 2 mRNA-binding proteins (IGF2BPs): post-transcriptional drivers of cancer progression? Cell. Mol. Life Sci. CMLS *70*, 2657–2675.

Belton, J.-M., McCord, R.P., Gibcus, J.H., Naumova, N., Zhan, Y., and Dekker, J. (2012). Hi-C: a comprehensive technique to capture the conformation of genomes. Methods San Diego Calif *58*, 268–276.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Ser. B Methodol. *57*, 289–300.

Berkes, C.A., and Tapscott, S.J. (2005). MyoD and the transcriptional control of myogenesis. Semin. Cell Dev. Biol. *16*, 585–595.

Berthelot, C., Muffato, M., Abecassis, J., and Roest Crollius, H. (2015). The 3D organization of chromatin explains evolutionary fragile genomic regions. Cell Rep. *10*, 1913–1924.

Bickmore, W.A. (2013). The spatial organization of the human genome. Annu. Rev. Genomics Hum. Genet. *14*, 67–84.

Björkegren, C., and Baranello, L. (2018). DNA Supercoiling, Topoisomerases, and Cohesin: Partners in Regulating Chromatin Architecture? Int. J. Mol. Sci. *19*.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. *2008*, P10008.

Bolzer, A., Kreth, G., Solovei, I., Koehler, D., Saracoglu, K., Fauth, C., Müller, S., Eils, R., Cremer, C., Speicher, M.R., et al. (2005). Three-Dimensional Maps of All Chromosomes in Human Male Fibroblast Nuclei and Prometaphase Rosettes. PLoS Biol. *3*.

Bonora, G., Plath, K., and Denholtz, M. (2014). A mechanistic link between gene regulation and genome architecture in mammalian development. Curr. Opin. Genet. Dev. *27*, 92–101.

Borensztein, M., Viengchareun, S., Montarras, D., Journot, L., Binart, N., Lombès, M., and Dandolo, L. (2012). Double Myod and Igf2 inactivation promotes brown adipose tissue development by increasing Prdm16 expression. FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol. *26*, 4584–4591.

Botta, M., Haider, S., Leung, I.X.Y., Lio, P., and Mozziconacci, J. (2010). Intra- and inter-chromosomal interactions correlate with CTCF binding genome wide. Mol. Syst. Biol. *6*, 426.

Boyle, S., Gilchrist, S., Bridger, J.M., Mahy, N.L., Ellis, J.A., and Bickmore, W.A. (2001). The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. Hum. Mol. Genet. *10*, 211–219.

Branco, M.R., and Pombo, A. (2006). Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. PLoS Biol. *4*, e138.

Buckingham, M. (2006). Myogenic progenitor cells and skeletal myogenesis in vertebrates. Curr. Opin. Genet. Dev. *16*, 525–532.

Burton, J.N., Adey, A., Patwardhan, R.P., Qiu, R., Kitzman, J.O., and Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat. Biotechnol. *31*, 1119–1125.

Butte, A.J., and Kohane, I.S. (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. Pac. Symp. Biocomput. Pac. Symp. Biocomput. 418–429.

Cagnazzo, M., te Pas, M.F.W., Priem, J., de Wit, A. a. C., Pool, M.H., Davoli, R., and Russo, V. (2006). Comparison of prenatal muscle tissue expression profiles of two pig breeds differing in muscle characteristics. J. Anim. Sci. *84*, 1–10.

Canario, L. (2006). Aspects génétiques de la mortalité des porcelets à la naissance et en allaitement précoce : relations avec les aptitudes maternelles des truies et la vitalité des porecelets (Paris, Institut national d'agronomie de Paris Grignon).

Cavalli, G., and Misteli, T. (2013). Functional implications of genome topology. Nat. Struct. Mol. Biol. *20*, 290–299.

Chal, J., and Pourquié, O. (2017). Making muscle: skeletal myogenesisin vivoandin vitro. Dev. Camb. Engl. *144*, 2104–2122.

Chaumeil, J., Micsinai, M., and Skok, J.A. (2013). Combined Immunofluorescence and DNA FISH on 3D-preserved Interphase Nuclei to Study Changes in 3D Nuclear Organization. J. Vis. Exp. JoVE.

Chen, K., Baxter, T., Muir, W.M., Groenen, M.A., and Schook, L.B. (2007). Genetic resources, genome mapping and evolutionary genomics of the pig (Sus scrofa). Int. J. Biol. Sci. *3*, 153–165.

Chuang, C.-H., Carpenter, A.E., Fuchsova, B., Johnson, T., de Lanerolle, P., and Belmont, A.S. (2006). Long-range directional movement of an interphase chromosome site. Curr. Biol. CB *16*, 825–831.

Clauset, A., Newman, M.E.J., and Moore, C. (2004). Finding community structure in very large networks. Phys. Rev. E Stat. Nonlin. Soft Matter Phys. *70*, 066111.

Clowney, E.J., LeGros, M.A., Mosley, C.P., Clowney, F.G., Markenskoff-Papadimitriou, E.C., Myllys, M., Barnea, G., Larabell, C.A., and Lomvardas, S. (2012). Nuclear aggregation of olfactory receptor genes governs their monogenic expression. Cell *151*, 724–737.

Cournac, A., Koszul, R., and Mozziconacci, J. (2016). The 3D folding of metazoan genomes correlates with the association of similar repetitive elements. Nucleic Acids Res. *44*, 245–255.

Cremer, T., and Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nat. Rev. Genet. *2*, 292–301.

Cremer, T., and Cremer, M. (2010). Chromosome territories. Cold Spring Harb. Perspect. Biol. *2*, a003889.

Cremer, C., Szczurek, A., Schock, F., Gourram, A., and Birk, U. (2017). Super-resolution microscopy approaches to nuclear nanostructure imaging. Methods San Diego Calif *123*, 11–32.

Crist, C.G., Montarras, D., and Buckingham, M. (2012). Muscle satellite cells are primed for myogenesis but maintain quiescence with sequestration of Myf5 mRNA targeted by microRNA-31 in mRNP granules. Cell Stem Cell *11*, 118–126.

Csárdi, G., and Nepusz, T. (2006). The igraph software package for complex network research. InterJournal Complex Syst. 1695.

Danon, L., Duch, J., Diaz-Guilera, A., and Arenas, A. (2005). Comparing community structure identification. J. Stat. Mech. Theory Exp. *2005*, P09008–P09008.

Davies, J.O.J., Telenius, J.M., McGowan, S.J., Roberts, N.A., Taylor, S., Higgs, D.R., and Hughes, J.R. (2016). Multiplexed analysis of chromosome conformation at vastly improved sensitivity. Nat. Methods *13*, 74–80.

Davies, J.O.J., Oudelaar, A.M., Higgs, D.R., and Hughes, J.R. (2017). How best to identify chromosomal interactions: a comparison of approaches. Nat. Methods *14*, 125–134.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. Science *295*, 1306–1311.

Deng, W., and Blobel, G.A. (2014). Manipulating nuclear architecture. Curr. Opin. Genet. Dev. *25*, 1–7.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions. Nature *485*, 376–380.

Dixon, J.R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J.E., Lee, A.Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., et al. (2015). Chromatin architecture reorganization during stem cell differentiation. Nature *518*, 331–336.

Dong, P., Tu, X., Chu, P.-Y., Lü, P., Zhu, N., Grierson, D., Du, B., Li, P., and Zhong, S. (2017). 3D Chromatin Architecture of Large Plant Genomes Determined by Local A/B Compartments. Mol. Plant *10*, 1497–1509.

Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C., et al. (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res. *16*, 1299–1309.

Dowen, J.M., Fan, Z.P., Hnisz, D., Ren, G., Abraham, B.J., Zhang, L.N., Weintraub, A.S., Schujiers, J., Lee, T.I., Zhao, K., et al. (2014). Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. Cell *159*, 374–387.

Doynova, M.D., Markworth, J.F., Cameron-Smith, D., Vickers, M.H., and O'Sullivan, J.M. (2017). Linkages between changes in the 3D organization of the genome and transcription during myotube differentiation in vitro. Skelet. Muscle *7*, 5.

Dundr, M., Ospina, J.K., Sung, M.-H., John, S., Upender, M., Ried, T., Hager, G.L., and Matera, A.G. (2007). Actin-dependent intranuclear repositioning of an active gene locus in vivo. J. Cell Biol. *179*, 1095–1103.

Edwards, D. (1995). Introduction to Graphical Modelling | David Edwards | Springer.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Falcon, S., and Gentleman, R. (2007). Using GOstats to test gene lists for GO term association. Bioinforma. Oxf. Engl. *23*, 257–258.

Fanucchi, S., Shibayama, Y., Burd, S., Weinberg, M.S., and Mhlanga, M.M. (2013). Chromosomal contact permits transcription between coregulated genes. Cell *155*, 606–620.

Feng, J., Liu, T., Qin, B., Zhang, Y., and Liu, X.S. (2012). Identifying ChIP-seq enrichment using MACS. Nat. Protoc. *7*, 1728–1740.

Filippova, D., Patro, R., Duggal, G., and Kingsford, C. (2014). Identification of alternative topological domains in chromatin. Algorithms Mol. Biol. AMB *9*, 14.

Foissac, S., Djebali, S., Munyard, K., Villa-Vialaneix, N., Rau, A., Muret, K., Esquerre, D., Zytnicki, M., Derrien, T., Bardou, P., et al. (2018). Livestock genome annotation: transcriptome and chromatin structure profiling in cattle, goat, chicken and pig. BioRxiv 316091.

Forcato, M., Nicoletti, C., Pal, K., Livi, C.M., Ferrari, F., and Bicciato, S. (2017). Comparison of computational methods for Hi-C data analysis. Nat. Methods *14*, 679–685.

Fortin, J.-P., and Hansen, K.D. (2015). Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. Genome Biol. *16*, 180.

Foxcroft, G.R., Dixon, W.T., Novak, S., Putman, C.T., Town, S.C., and Vinsky, M.D.A. (2006). The biological basis for prenatal programming of postnatal performance in pigs. J. Anim. Sci. *84 Suppl*, E105-112.

Fraser, J., Ferrai, C., Chiariello, A.M., Schueler, M., Rito, T., Laudanno, G., Barbieri, M., Moore, B.L., Kraemer, D.C.A., Aitken, S., et al. (2015). Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. Mol. Syst. Biol. *11*, 852.

Fudenberg, G., and Imakaev, M. (2017). FISH-ing for captured contacts: towards reconciling FISH and 3C. Nat. Methods *14*, 673–678.

Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y.B., Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H., et al. (2009). An oestrogen-receptor-alpha-bound human chromatin interactome. Nature *462*, 58–64.

Ghurye, J., Pop, M., Koren, S., Bickhart, D., and Chin, C.-S. (2017). Scaffolding of long read assemblies using long range contact information. BMC Genomics *18*, 527.

Gibcus, J.H., and Dekker, J. (2013). The hierarchy of the 3D genome. Mol. Cell *49*, 773–782.

Gong, C., Li, Z., Ramanujan, K., Clay, I., Zhang, Y., Lemire-Brachat, S., and Glass, D.J. (2015). A long non-coding RNA, LncMyoD, regulates skeletal muscle differentiation by blocking IMP2-mediated mRNA translation. Dev. Cell *34*, 181–191.

Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinforma. Oxf. Engl. *27*, 1017–1018.

Groenen, M.A.M., Archibald, A.L., Uenishi, H., Tuggle, C.K., Takeuchi, Y., Rothschild, M.F., Rogel-Gaillard, C., Park, C., Milan, D., Megens, H.-J., et al. (2012). Analyses of pig genomes provide insight into porcine demography and evolution. Nature *491*, 393–398.

Guidi, M., Ruault, M., Marbouty, M., Loïodice, I., Cournac, A., Billaudeau, C., Hocher, A., Mozziconacci, J., Koszul, R., and Taddei, A. (2015). Spatial reorganization of telomeres in long-lived quiescent cells. Genome Biol. *16*, 206.

Gustavsson, I. (1988). Standard karyotype of the domestic pig. Committee for the Standardized Karyotype of the Domestic Pig. Hereditas *109*, 151–157.

Hakim, O., John, S., Ling, J.Q., Biddie, S.C., Hoffman, A.R., and Hager, G.L. (2009). Glucocorticoid receptor activation of the Ciz1-Lcn2 locus by long range interactions. J. Biol. Chem. *284*, 6048–6052.

Han, J., Zhang, Z., and Wang, K. (2018). 3C and 3C-based techniques: the powerful tools for spatial genome organization deciphering. Mol. Cytogenet. *11*, 21.

Harada, A., Mallappa, C., Okada, S., Butler, J.T., Baker, S.P., Lawrence, J.B., Ohkawa, Y., and Imbalzano, A.N. (2015). Spatial re-organization of myogenic regulatory sequences temporally controls gene expression. Nucleic Acids Res. *43*, 2008–2021.

Harr, J.C., Gonzalez-Sandoval, A., and Gasser, S.M. (2016). Histones and histone modifications in perinuclear chromatin anchoring: from yeast to man. EMBO Rep. *17*, 139–155.

Herpin, P., Le Dividich, J., and Amaral, N. (1993). Effect of selection for lean tissue growth on body composition and physiological state of the pig at birth. J. Anim. Sci. *71*, 2645–2653.

Hnisz, D., Day, D.S., and Young, R.A. (2016a). Insulated Neighborhoods: Structural and Functional Units of Mammalian Gene Control. Cell *167*, 1188–1200.

Hnisz, D., Weintraub, A.S., Day, D.S., Valton, A.-L., Bak, R.O., Li, C.H., Goldmann, J., Lajoie, B.R., Fan, Z.P., Sigova, A.A., et al. (2016b). Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science *351*, 1454–1458.

Holwerda, S., and de Laat, W. (2012). Chromatin loops, gene positioning, and gene expression. Front. Genet. *3*, 217.

Hou, C., and Corces, V.G. (2012). Throwing transcription for a loop: expression of the genome in the 3D nucleus. Chromosoma *121*, 107–116.

Hozé, N., Ruault, M., Amoruso, C., Taddei, A., and Holcman, D. (2013). Spatial telomere organization and clustering in yeast Saccharomyces cerevisiae nucleus is generated by a random dynamics of aggregation-dissociation. Mol. Biol. Cell *24*, 1791–1800, S1-10.

Huang, F., Sirinakis, G., Allgeyer, E.S., Schroeder, L.K., Duim, W.C., Kromann, E.B., Phan, T., Rivera-Molina, F.E., Myers, J.R., Irnov, I., et al. (2016). Ultra-High Resolution 3D Imaging of Whole Cells. Cell *166*, 1028–1040.

Hughes, J.R., Roberts, N., McGowan, S., Hay, D., Giannoulatou, E., Lynch, M., De Gobbi, M., Taylor, S., Gibbons, R., and Higgs, D.R. (2014). Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. Nat. Genet. *46*, 205–212.

Humphray, S.J., Scott, C.E., Clark, R., Marron, B., Bender, C., Camm, N., Davis, J., Jenks, A., Noon, A., Patel, M., et al. (2007). A high utility integrated map of the pig genome. Genome Biol. *8*, R139.

Iannuccelli, E., Mompart, F., Gellin, J., Lahbib-Mansais, Y., Yerle, M., and Boudier, T. (2010). NEMO: a tool for analyzing gene and chromosome territory distributions from 3D-FISH experiments. Bioinformatics *26*, 696–697.

Imakaev, M., Fudenberg, G., McCord, R.P., Naumova, N., Goloborodko, A., Lajoie, B.R., Dekker, J., and Mirny, L.A. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat. Methods *9*, 999–1003.

Jin, W., Peng, J., and Jiang, S. (2016). The epigenetic regulation of embryonic myogenesis and adult muscle regeneration by histone methylation modification. Biochem. Biophys. Rep. *6*, 209–219.

Kaaij, L.J.T., van der Weide, R.H., Ketting, R.F., and de Wit, E. (2018). Systemic Loss and Gain of Chromatin Architecture throughout Zebrafish Development. Cell Rep. *24*, 1-10.e4.

Kim, C.-H., Neiswender, H., Baik, E.J., Xiong, W.C., and Mei, L. (2008). Beta-catenin interacts with MyoD and regulates its transcription activity. Mol. Cell. Biol. *28*, 2941–2951.

Kim, S., Liachko, I., Brickner, D.G., Cook, K., Noble, W.S., Brickner, J.H., Shendure, J., and Dunham, M.J. (2017). The dynamic three-dimensional organization of the diploid yeast genome. ELife *6*.

Kim, T.H., Abdullaev, Z.K., Smith, A.D., Ching, K.A., Loukinov, D.I., Green, R.D., Zhang, M.Q., Lobanenkov, V.V., and Ren, B. (2007). Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell *128*, 1231–1245.

Kind, J., Pagie, L., Ortabozkoyun, H., Boyle, S., de Vries, S.S., Janssen, H., Amendola, M., Nolen, L.D., Bickmore, W.A., and van Steensel, B. (2013). Single-cell dynamics of genome-nuclear lamina interactions. Cell *153*, 178–192.

Kocanova, S., Kerr, E.A., Rafique, S., Boyle, S., Katz, E., Caze-Subra, S., Bickmore, W.A., and Bystricky, K. (2010). Activation of estrogen-responsive genes does not require their nuclear co-localization. PLoS Genet. *6*, e1000922.

Kociucka, B., Cieslak, J., and Szczerbal, I. (2012). Three-dimensional arrangement of genes involved in lipid metabolism in nuclei of porcine adipocytes and fibroblasts in relation to their transcription level. Cytogenet. Genome Res. *136*, 295–302.

de Laat, W., and Duboule, D. (2013). Topology of mammalian developmental enhancers and their regulatory landscapes. Nature *502*, 499–506.

Lahbib-Mansais, Y., Barasc, H., Marti-Marimon, M., Mompart, F., Iannuccelli, E., Robelin, D., Riquet, J., and Yerle-Bouissou, M. (2016). Expressed alleles of imprinted IGF2, DLK1 and MEG3 colocalize in 3D-preserved nuclei of porcine fetal cells. BMC Cell Biol. *17*, 35.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Lazar-Stefanita, L., Scolari, V.F., Mercy, G., Muller, H., Guérin, T.M., Thierry, A., Mozziconacci, J., and Koszul, R. (2017). Cohesins and condensins orchestrate the 4D dynamics of yeast chromosomes during the cell cycle. EMBO J. *36*, 2684–2697.

Leenhouwers, J.I., Knol, E.F., Groot, D., N, P., Vos, H., and van der Lende, T. (2002a). Fetal development in the pig in relation to genetic merit for piglet survival. J. Anim. Sci. *80*, 1759–1770.

Leenhouwers, J.I., Knol, E.F., and van der Lende, T. (2002b). Differences in late prenatal development as an explanation for genetic differences in piglet survival. Livest. Prod. Sci. *78*, 57–62.

Lefaucheur, L., Milan, D., Ecolan, P., and Le Callennec, C. (2004). Myosin heavy chain composition of different skeletal muscles in Large White and Meishan pigs. J. Anim. Sci. *82*, 1931–1941.

van der Lende, T., Knol, E.F., and Leenhouwers, J.I. (2001). Prenatal development as a predisposing factor for perinatal losses in pigs. Reprod. Camb. Engl. Suppl. *58*, 247–261.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinforma. Oxf. Engl. *25*, 2078–2079.

Li, Q., Tjong, H., Li, X., Gong, K., Zhou, X.J., Chiolo, I., and Alber, F. (2017). The three-dimensional genome organization of Drosophila melanogaster through data integration. Genome Biol. *18*, 145.

Li, R., Liu, Y., Li, T., and Li, C. (2016). 3Disease Browser: A Web server for integrating 3D genome and disease-associated chromosome rearrangement data. Sci. Rep. *6*, 34651.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293.

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H., and Ecker, J.R. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell *133*, 523–536.

Livingstone, C., and Borai, A. (2014). Insulin-like growth factor-II: its role in metabolic and endocrine disease. Clin. Endocrinol. (Oxf.) *80*, 773–781.

Lun, A.T.L., and Smyth, G.K. (2015). diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. BMC Bioinformatics *16*, 258.

Lun, A.T.L., and Smyth, G.K. (2016). csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. Nucleic Acids Res. *44*, e45.

Lunney, J.K. (2007). Advances in swine biomedical model genomics. Int. J. Biol. Sci. *3*, 179–184.

Lupiáñez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., et al. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell *161*, 1012–1025.

Ma, H., Naseri, A., Reyes-Gutierrez, P., Wolfe, S.A., Zhang, S., and Pederson, T. (2015). Multicolor CRISPR labeling of chromosomal loci in human cells. Proc. Natl. Acad. Sci. U. S. A. *112*, 3002–3007.

Macdonald, W.A. (2012). Epigenetic mechanisms of genomic imprinting: common themes in the regulation of imprinted regions in mammals, plants, and insects. Genet. Res. Int. *2012*, 585024.

Marbouty, M., and Koszul, R. (2015). Metagenome Analysis Exploiting High-Throughput Chromosome Conformation Capture (3C) Data. Trends Genet. TIG *31*, 673–682.

Marti-Marimon, M., Vialaneix, N., Voillet, V., Yerle-Bouissou, M., Lahbib-Mansais, Y., and Liaubet, L. (2018). A new approach of gene co-expression network inference reveals significant biological processes involved in porcine muscle development in late gestation. Sci. Rep. *8*, 10150.

Matharu, N.K., and Ahanger, S.H. (2015). Chromatin Insulators and Topological Domains: Adding New Dimensions to 3D Genome Architecture. Genes *6*, 790–811.

Mathelier, A., Fornes, O., Arenillas, D.J., Chen, C.-Y., Denay, G., Lee, J., Shi, W., Shyr, C., Tan, G., Worsley-Hunt, R., et al. (2016). JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. *44*, D110-115.

Meinshausen, N., and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. Ann. Stat. *34*, 1436–1462.

Mercer, T.R., and Mattick, J.S. (2013). Understanding the regulatory and transcriptional complexity of the genome through structure. Genome Res. *23*, 1081–1088.

Meuleman, W., Peric-Hupkes, D., Kind, J., Beaudry, J.-B., Pagie, L., Kellis, M., Reinders, M., Wessels, L., and van Steensel, B. (2013). Constitutive nuclear lamina-genome interactions are highly conserved and associated with A/T-rich sequence. Genome Res. *23*, 270–280.

Molenaar, C., Wiesmeijer, K., Verwoerd, N.P., Khazen, S., Eils, R., Tanke, H.J., and Dirks, R.W. (2003). Visualizing telomere dynamics in living mammalian cells using PNA probes. EMBO J. *22*, 6631–6641.

Mompart, F., Robelin, D., Delcros, C., and Yerle-Bouissou, M. (2013). 3D organization of telomeres in porcine neutrophils and analysis of LPS-activation effect. BMC Cell Biol. *14*, 30.

Montastier, E., Villa-Vialaneix, N., Caspar-Bauguil, S., Hlavaty, P., Tvrzicka, E., Gonzalez, I., Saris, W.H.M., Langin, D., Kunesova, M., and Viguerie, N. (2015). System model network for adipose tissue signatures related to weight changes in response to calorie restriction and subsequent weight maintenance. PLoS Comput. Biol. *11*, e1004047.

Mourad, R., Hsu, P.-Y., Juan, L., Shen, C., Koneru, P., Lin, H., Liu, Y., Nephew, K., Huang, T.H., and Li, L. (2014). Estrogen induces global reorganization of chromatin structure in human breast cancer cells. PloS One *9*, e113354.

Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A., and Fraser, P. (2013). Single cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature *502*.

Nagano, T., Várnai, C., Schoenfelder, S., Javierre, B.-M., Wingett, S.W., and Fraser, P. (2015). Comparison of Hi-C results using in-solution versus in-nucleus ligation. Genome Biol. *16*, 175.

Nagele, R.G., Velasco, A.Q., Anderson, W.J., McMahon, D.J., Thomson, Z., Fazekas, J., Wind, K., and Lee, H. (2001). Telomere associations in interphase nuclei: possible role in maintenance of interphase chromosome topology. J. Cell Sci. *114*, 377–388.

Nezer, C., Moreau, L., Brouwers, B., Coppieters, W., Detilleux, J., Hanset, R., Karim, L., Kvasz, A., Leroy, P., and Georges, M. (1999). An imprinted QTL with major effect on muscle mass and fat deposition maps to the IGF2 locus in pigs. Nat. Genet. *21*, 155–156.

Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J., et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature *485*, 381–385.

Noubissi, F.K., Elcheva, I., Bhatia, N., Shakoori, A., Ougolkov, A., Liu, J., Minamoto, T., Ross, J., Fuchs, S.Y., and Spiegelman, V.S. (2006). CRD-BP mediates stabilization of betaTrCP1 and c-myc mRNA in response to beta-catenin signalling. Nature *441*, 898–901.

Nozawa, R.-S., and Gilbert, N. (2014). Interphase chromatin LINEd with RNA. Cell *156*, 864–865.

Ong, C.-T., and Corces, V.G. (2014). CTCF: an architectural protein bridging genome topology and function. Nat. Rev. Genet. *15*, 234–246.

Osborne, C.S., Chakalova, L., Brown, K.E., Carter, D., Horton, A., Debrand, E., Goyenechea, B., Mitchell, J.A., Lopes, S., Reik, W., et al. (2004). Active genes dynamically colocalize to shared sites of ongoing transcription. Nat. Genet. *36*, 1065–1071.

Óvilo, C., Benítez, R., Fernández, A., Núñez, Y., Ayuso, M., Fernández, A.I., Rodríguez, C., Isabel, B., Rey, A.I., López-Bote, C., et al. (2014). Longissimus dorsi transcriptome analysis of purebred and crossbred Iberian pigs differing in muscle characteristics. BMC Genomics *15*.

Perrin, J.-B., Ducrot, C., Vinard, J.-L, Hendrikx, P., and Calavas, D. (2011). Analyse de la mortalité bovine en France de 2003 à 2009. Inra Prod. Anim. 235–244.

Perruchot, M.-H., Ecolan, P., Sorensen, I.L., Oksbjerg, N., and Lefaucheur, L. (2012). In vitro characterization of proliferation and differentiation of pig satellite cells. Differentiation *84*, 322–329.

Phillips-Cremins, J.E., Sauria, M.E.G., Sanyal, A., Gerasimova, T.I., Lajoie, B.R., Bell, J.S.K., Ong, C.-T., Hookway, T.A., Guo, C., Sun, Y., et al. (2013). Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell *153*, 1281–1295.

Picard, B., Lefaucheur, L., Berri, C., and Duclos, M.J. (2002). Muscle fibre ontogenesis in farm animal species. Reprod. Nutr. Dev. *42*, 415–431.

Pombo, A., and Dillon, N. (2015). Three-dimensional genome architecture: players and mechanisms. Nat. Rev. Mol. Cell Biol. *16*, 245–257.

Qin, P., Parlak, M., Kuscu, C., Bandaria, J., Mir, M., Szlachta, K., Singh, R., Darzacq, X., Yildiz, A., and Adli, M. (2017). Live cell imaging of low- and non-repetitive chromosome loci using CRISPR-Cas9. Nat. Commun. *8*, 14725.

Ramazzotti, G., Billi, A.M., Manzoli, L., Mazzetti, C., Ruggeri, A., Erneux, C., Kim, S., Suh, P.-G., Cocco, L., and Faenza, I. (2016). IPMK and β-catenin mediate PLC-β1-dependent signaling in myogenic differentiation. Oncotarget *7*, 84118–84127.

Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., et al. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665–1680.

Rehfeldt, C., and Kuhn, G. (2006). Consequences of birth weight for postnatal growth performance and carcass quality in pigs as related to myogenesis. J. Anim. Sci. *84 Suppl*, E113-123.

Rehfeldt, C., Fiedler, I., Dietl, G., and Ender, K. (2000). Myogenesis and postnatal skeletal muscle cell growth as influenced by selection. Livest. Prod. Sci. *66*, 177–188.

Rieder, D., Trajanoski, Z., and McNally, J.G. (2012). Transcription factories. Front. Genet. *3*, 221.

Rieder, D., Ploner, C., Krogsdam, A.M., Stocker, G., Fischer, M., Scheideler, M., Dani, C., Amri, E.-Z., Müller, W.G., McNally, J.G., et al. (2014). Co-expressed genes prepositioned in spatial neighborhoods stochastically associate with SC35 speckles and RNA polymerase II factories. Cell. Mol. Life Sci. CMLS *71*, 1741–1759.

Robin, J.D., Ludlow, A.T., Batten, K., Magdinier, F., Stadler, G., Wagner, K.R., Shay, J.W., and Wright, W.E. (2014). Telomere position effect: regulation of gene expression with progressive telomere shortening over long distances. Genes Dev. *28*, 2464–2476.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinforma. Oxf. Engl. *26*, 139–140.

Rothbart, S.B., and Strahl, B.D. (2014). Interpreting the language of histone and DNA modifications. Biochim. Biophys. Acta *1839*, 627–643.

Sandhu, K.S., Shi, C., Sjölinder, M., Zhao, Z., Göndör, A., Liu, L., Tiwari, V.K., Guibert, S., Emilsson, L., Imreh, M.P., et al. (2009). Nonallelic transvection of multiple imprinted loci is organized by the H19 imprinting control region during germline development. Genes Dev. *23*, 2598–2603.

Schiaffino, S., Rossi, A.C., Smerdu, V., Leinwand, L.A., and Reggiani, C. (2015). Developmental myosins: expression patterns and functional significance. Skelet. Muscle *5*, 22.

Schneider, R., and Grosschedl, R. (2007). Dynamics and interplay of nuclear architecture, genome organization, and gene expression. Genes Dev. *21*, 3027–3043.

Schoenfelder, S., Sexton, T., Chakalova, L., Cope, N.F., Horton, A., Andrews, S., Kurukuti, S., Mitchell, J.A., Umlauf, D., Dimitrova, D.S., et al. (2010). Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. Nat. Genet. *42*, 53–61.

Schook, L.B., Beever, J.E., Rogers, J., Humphray, S., Archibald, A., Chardon, P., Milan, D., Rohrer, G., and Eversole, K. (2005). Swine Genome Sequencing Consortium (SGSC): a strategic roadmap for sequencing the pig genome. Comp. Funct. Genomics *6*, 251–255.

Schwartzman, O., Mukamel, Z., Oded-Elkayam, N., Olivares-Chauvet, P., Lubling, Y., Landan, G., Izraeli, S., and Tanay, A. (2016). UMI-4C for quantitative and targeted chromosomal contact profiling. Nat. Methods *13*, 685–691.

Servant, N., Lajoie, B.R., Nora, E.P., Giorgetti, L., Chen, C.-J., Heard, E., Dekker, J., and Barillot, E. (2012). HiTC: exploration of high-throughput "C" experiments. Bioinforma. Oxf. Engl. *28*, 2843–2844.

Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.-J., Vert, J.-P., Heard, E., Dekker, J., and Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. *16*, 259.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the Drosophila genome. Cell *148*, 458–472.

Shang, Y.C., Zhang, C., Wang, S.H., Xiong, F., Zhao, C.P., Peng, F.N., Feng, S.W., Yu, M.J., Li, M.S., and Zhang, Y.N. (2007). Activated beta-catenin induces myogenesis and inhibits adipogenesis in BM-derived mesenchymal stromal cells. Cytotherapy *9*, 667–681.

Sieben, C., Douglass, K.M., Guichard, P., and Manley, S. (2018). Super-resolution microscopy to decipher multi-molecular assemblies. Curr. Opin. Struct. Biol. *49*, 169–176.

Sodhi, S.S., Song, K.-D., Ghosh, M., Sharma, N., Lee, S.J., Kim, J.H., Kim, N., Mongre, R.K., Adhikari, P., Kim, J.Y., et al. (2014). Comparative transcriptomic analysis by RNA-seq to discern differential expression of genes in liver and muscle tissues of adult Berkshire and Jeju Native Pig. Gene *546*, 233–242.

Solinhac, R., Mompart, F., Martin, P., Robelin, D., Pinton, P., Iannuccelli, E., Lahbib-Mansais, Y., Oswald, I.P., and Yerle-Bouissou, M. (2011). Transcriptomic and nuclear architecture of immune cells after LPS activation. Chromosoma *120*, 501–520.

Solov'eva, L., Svetlova, M., Bodinski, D., and Zalensky, A.O. (2004). Nature of telomere dimers and chromosome looping in human spermatozoa. Chromosome Res. Int. J. Mol. Supramol. Evol. Asp. Chromosome Biol. *12*, 817–823.

Spichal, M., Brion, A., Herbert, S., Cournac, A., Marbouty, M., Zimmer, C., Koszul, R., and Fabre, E. (2016). Evidence for a dual role of actin in regulating chromosome organization and dynamics in yeast. J. Cell Sci. *129*, 681–692.

Stachecka, J., Walczak, A., Kociucka, B., Ruszczycki, B., Wilczyński, G., and Szczerbal, I. (2018). Nuclear organization during in vitro differentiation of porcine mesenchymal stem cells (MSCs) into adipocytes. Histochem. Cell Biol. *149*, 113–126.

Stevens, T.J., Lando, D., Basu, S., Atkinson, L.P., Cao, Y., Lee, S.F., Leeb, M., Wohlfahrt, K.J., Boucher, W., O'Shaughnessy-Kirwan, A., et al. (2017). 3D structures of individual mammalian genomes studied by single-cell Hi-C. Nature *544*, 59–64.

St-Pierre, J., Hivert, M.-F., Perron, P., Poirier, P., Guay, S.-P., Brisson, D., and Bouchard, L. (2012). IGF2 DNA methylation is a modulator of newborn's fetal growth and development. Epigenetics *7*, 1125–1132.

Sydor, A.M., Czymmek, K.J., Puchner, E.M., and Mennella, V. (2015). Super-Resolution Microscopy: From Single Molecules to Supramolecular Assemblies. Trends Cell Biol. *25*, 730–748.

Szczerbal, I., and Bridger, J.M. (2010). Association of adipogenic genes with SC-35 domains during porcine adipogenesis. Chromosome Res. Int. J. Mol. Supramol. Evol. Asp. Chromosome Biol. *18*, 887–895.

Szczerbal, I., Foster, H.A., and Bridger, J.M. (2009). The spatial repositioning of adipogenesis genes is correlated with their expression status in a porcine mesenchymal stem cell adipogenesis model system. Chromosoma *118*, 647–663.

Tan, J.Y., Smith, A.A.T., Ferreira da Silva, M., Matthey-Doret, C., Rueedi, R., Sönmez, R., Ding, D., Kutalik, Z., Bergmann, S., and Marques, A.C. (2017). cis-Acting Complex-Trait-Associated lincRNA Expression Correlates with Modulation of Chromosomal Architecture. Cell Rep. *18*, 2280–2288.

Tang, Z., Yang, Y., Wang, Z., Zhao, S., Mu, Y., and Li, K. (2015a). Integrated analysis of miRNA and mRNA paired expression profiling of prenatal skeletal muscle development in three genotype pigs. Sci. Rep. *5*, 15544.

Tang, Z., Luo, O.J., Li, X., Zheng, M., Zhu, J.J., Szalaj, P., Trzaskoma, P., Magalska, A., Wlodarczyk, J., Ruszczycki, B., et al. (2015b). CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. Cell *163*, 1611–1627.

Tolhuis, B., Palstra, R.J., Splinter, E., Grosveld, F., and de Laat, W. (2002). Looping and interaction between hypersensitive sites in the active beta-globin locus. Mol. Cell *10*, 1453–1465.

Tuggle, C.K., Giuffra, E., White, S.N., Clarke, L., Zhou, H., Ross, P.J., Acloque, H., Reecy, J.M., Archibald, A., Bellone, R.R., et al. (2016). GO-FAANG meeting: a Gathering On Functional Annotation of Animal Genomes. Anim. Genet. *47*, 528–533.

Uusküla-Reimand, L., Hou, H., Samavarchi-Tehrani, P., Rudan, M.V., Liang, M., Medina-Rivera, A., Mohammed, H., Schmidt, D., Schwalie, P., Young, E.J., et al. (2016). Topoisomerase II beta interacts with cohesin and CTCF at topological domain borders. Genome Biol. *17*, 182.

Van Laere, A.-S., Nguyen, M., Braunschweig, M., Nezer, C., Collette, C., Moreau, L., Archibald, A.L., Haley, C.S., Buys, N., Tally, M., et al. (2003). A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. Nature *425*, 832–836.

Varrault, A., Gueydan, C., Delalbre, A., Bellmann, A., Houssami, S., Aknin, C., Severac, D., Chotard, L., Kahli, M., Le Digarcher, A., et al. (2006). Zac1 regulates an imprinted gene network critically involved in the control of embryonic growth. Dev. Cell *11*, 711–722.

Vieux-Rochas, M., Fabre, P.J., Leleu, M., Duboule, D., and Noordermeer, D. (2015). Clustering of mammalian Hox genes with other H3K27me3 targets within an active nuclear domain. Proc. Natl. Acad. Sci. U. S. A. *112*, 4672–4677.

Villa-Vialaneix, N., Liaubet, L., Laurent, T., Cherel, P., Gamot, A., and SanCristobal, M. (2013). The Structure of a Gene Co-Expression Network Reveals Biological Functions Underlying eQTLs. PLoS ONE *8*.

Villa-Vialaneix, N., Vignes, M., Viguerie, N., and SanCristobal, M. (2014). Inferring Networks from Multiple Samples with Consensus LASSO. Qual. Technol. Quant. Manag. *11*, 39–60.

Voillet, V. (2016). Approche intégrative du développement musculaire afin de décrire le processus de maturation en lien avec la survie néonatale (Toulouse, INPT).

Voillet, V., SanCristobal, M., Lippi, Y., Martin, P.G.P., Iannuccelli, N., Lascor, C., Vignoles, F., Billon, Y., Canario, L., and Liaubet, L. (2014). Muscle transcriptomic investigation of late fetal development identifies candidate genes for piglet maturity. BMC Genomics *15*, 797.

Voillet, V., San Cristobal, M., Pere, M.-C., Billon, Y., Canario, L., Liaubet, L., and Lefaucheur, L. (2018). Integrated Analysis of Proteomic and Transcriptomic Data Highlights Late Fetal Muscle Maturation Process. Mol. Cell. Proteomics MCP.

Waddell, J.N., Zhang, P., Wen, Y., Gupta, S.K., Yevtodiyenko, A., Schmidt, J.V., Bidwell, C.A., Kumar, A., and Kuang, S. (2010). Dlk1 Is Necessary for Proper Skeletal Muscle Development and Regeneration. PLoS ONE *5*.

Wang, H., Maurano, M.T., Qu, H., Varley, K.E., Gertz, J., Pauli, F., Lee, K., Canfield, T., Weaver, M., Sandstrom, R., et al. (2012). Widespread plasticity in CTCF occupancy linked to DNA methylation. Genome Res. *22*, 1680–1688.

Wang, J., Duncan, D., Shi, Z., and Zhang, B. (2013). WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. Nucleic Acids Res. *41*, W77-83.

Wang, S., Su, J.-H., Beliveau, B.J., Bintu, B., Moffitt, J.R., Wu, C., and Zhuang, X. (2016). Spatial organization of chromatin domains and compartments in single chromosomes. Science *353*, 598–602.

Wang, X., Kam, Z., Carlton, P.M., Xu, L., Sedat, J.W., and Blackburn, E.H. (2008). Rapid telomere motions in live human cells analyzed by highly time-resolved microscopy. Epigenetics Chromatin *1*, 4.

Wang, Y., Hudak, C., and Sul, H.S. (2010). Role of preadipocyte factor 1 in adipocyte differentiation. Clin. Lipidol. *5*, 109–115.

Williamson, I., Berlivet, S., Eskeland, R., Boyle, S., Illingworth, R.S., Paquette, D., Dostie, J., and Bickmore, W.A. (2014). Spatial genome organization: contrasting views from chromosome conformation capture and fluorescence in situ hybridization. Genes Dev. *28*, 2778–2791.

de Wit, E., and de Laat, W. (2012). A decade of 3C technologies: insights into nuclear organization. Genes Dev. *26*, 11–24.

Won, H., de la Torre-Ubieta, L., Stein, J.L., Parikshak, N.N., Huang, J., Opland, C.K., Gandal, M., Sutton, G.J., Hormozdiari, F., Lu, D., et al. (2016). Chromosome conformation elucidates regulatory relationships in developing human brain. Nature *538*, 523–527.

Xie, X., and Percipalle, P. (2017). An actin-based nucleoskeleton involved in gene regulation and genome organization. Biochem. Biophys. Res. Commun.

Xu, Y., Qian, H., Feng, X., Xiong, Y., Lei, M., Ren, Z., Zuo, B., Xu, D., Ma, Y., and Yuan, H. (2012). Differential proteome and transcriptome analysis of porcine skeletal muscle during development. J. Proteomics *75*, 2093–2108.

Yaffe, E., and Tanay, A. (2011). Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat. Genet. *43*, 1059–1065.

Yamamoto, A. (2014). Gathering up meiotic telomeres: a novel function of the microtubule-organizing center. Cell. Mol. Life Sci. CMLS *71*, 2119–2134.

Yerle, M., Goureau, A., Gellin, J., Le Tissier, P., and Moran, C. (1994). Rapid mapping of cosmid clones on pig chromosomes by fluorescence in situ hybridization. Mamm. Genome Off. J. Int. Mamm. Genome Soc. *5*, 34–37.

Yerle-Bouissou, M., Mompart, F., Iannuccelli, E., Robelin, D., Jauneau, A., Lahbib-Mansais, Y., Delcros, C., Oswald, I.P., and Gellin, J. (2009). Nuclear architecture of resting and LPS-stimulated porcine neutrophils by 3D FISH. Chromosome Res. Int. J. Mol. Supramol. Evol. Asp. Chromosome Biol. *17*, 847–862.

Young, J.A., and Trowsdale, J. (1985). A processed pseudogene in an intron of the HLA-DP beta 1 chain gene is a member of the ribosomal protein L32 gene family. Nucleic Acids Res. *13*, 8883–8891.

Yusuf, F., and Brand-Saberi, B. (2012). Myogenesis and muscle regeneration. Histochem. Cell Biol. *138*, 187–199.

Zhan, Y., Mariani, L., Barozzi, I., Schulz, E.G., Blüthgen, N., Stadler, M., Tiana, G., and Giorgetti, L. (2017). Reciprocal insulation analysis of Hi-C data shows that TADs represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes. Genome Res. *27*, 479–490.

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. Stat. Appl. Genet. Mol. Biol. *4*, Article17.

Zhang, B., Kirov, S., and Snoddy, J. (2005). WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic Acids Res. *33*, W741-748.

Zhao, X., Mo, D., Li, A., Gong, W., Xiao, S., Zhang, Y., Qin, L., Niu, Y., Guo, Y., Liu, X., et al. (2011). Comparative Analyses by Sequencing of Transcriptomes during Skeletal Muscle Development between Pig Breeds Differing in Muscle Growth Rate and Fatness. PLoS ONE *6*.

Zhao, Y., Li, J., Liu, H., Xi, Y., Xue, M., Liu, W., Zhuang, Z., and Lei, M. (2015). Dynamic transcriptome profiles of skeletal muscle tissue across 11 developmental stages for both Tongcheng and Yorkshire pigs. BMC Genomics *16*, 377.

Zhao, Z., Tavoosidana, G., Sjölinder, M., Göndör, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Sandhu, K.S., Singh, U., et al. (2006). Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat. Genet. *38*, 1341–1347.

10 Appendix. Supplementary data.

Appendix 1. Description of the model used for network inference.

Description of the model used for network inference

Supplementary file for the article "A new approach of gene co-expression network inference reveals highly significant biological processes in pig muscle involved in the establishment of maturity"

This file describes the model used for network inference and puts it in perspective with other approach found in the litterature. It also explains the choices made for the different hyper-parameters of the method.

1 Network inference

Networks were inferred using Gaussian graphical models (GGM; Edwards, 1995) from $n = 61$ samples at gestational age 90. From expression data, GGM build a graph (or network) in which vertices are genes and edges represent a strong relationship between the gene expressions. GGM are based on the estimation of partial correlations (*i.e.*, correlations between two gene expressions knowing the expression of all the other genes). They were preferred over relevance networks (Butte and Kohane, 2000) because they better measure direct relations between gene expressions by accounting for the effect of all expression data and because they were found more efficient to group genes with a common function in a previous study (Villa-Vialaneix et al., 2013).

More precisely, if $X = (X_1, \ldots, X_p)$ denotes the random variables corresponding to the expression of p genes, GGM supposes that X follows a Gaussian distribution $\mathcal{N}(0, \Sigma)$ and aims at estimating

$$\mathrm{Cor}\left(X_j, X_{j'} | (X_k)_{k \neq j,j'}\right)$$

for every pair (j, j') in $\{1, \ldots, p\}$. A graph is obtained from these estimation by putting an edge between nodes corresponding to the variables X_j and $X_{j'}$ when this partial correlation is different from 0. It can be shown that estimating partial correlations is also equivalent to estimating $\beta_{jj'}$ in the following linear models:

$$X_j = \sum_{j'=1,\ldots,p, j' \neq j} \beta_{jj'} X_{j'}$$

and more precisely that

$$\beta_{jj'} \neq 0 \quad \Leftrightarrow \quad \mathrm{Cor}\left(X_j, X_{j'} | (X_k)_{k \neq j,j'}\right) \neq 0.$$

When the number of samples is smaller than the number of genes used for network inference (which is generally the case and which was the case for our problem), the estimation of the partial correlation or of the equivalent linear models are ill-posed problems. This issue is frequently addressed by adding a sparse (L_1) penalty to the maximum likelihood (ML) problems induced by the linear regression formulation: this is the Graphical Lasso (GLasso) (Friedman, Hastie, and Tibshirani, 2008). This method allows to simultaneously estimates the coefficients $\beta_{jj'}$ and to perform variable (here edge) selection among the possible candidates since Lasso penalty yields to provide sparse solution in which many coefficients $\beta_{jj'}$ are set to 0 by the maximization of the penalized likelihood.

Similarly to that approach, we used a model that included a sparse penalty (for edge selection) combined with two L_2 (smooth) penalties aiming at incorporating *a priori* information into the inference similarly to what is proposed in Villa-Vialaneix et al., 2014. More precisely, this led to the minimization over $\beta_{jj'}$ (for j

and j' varying from 1 to p) of

$$\underbrace{\frac{1}{2}\beta_j^T\widehat{\Sigma}_{\backslash j\backslash j}\beta_j + \beta_j^T\widehat{\Sigma}_{j\backslash j}}_{\text{pseudo maximum likelihood}} + \lambda \underbrace{\|\beta_j\|_1}_{L_1 \text{ (sparse) penalty}} + \mu \underbrace{\sum_{(k,j)\in E_1}(\beta_{jk}-1)^2}_{L_2 \text{ (smooth) penalty for co-localized edges}} + \mu \underbrace{\sum_{(k,j)\in E_2}(\beta_{jk}-0)^2}_{L_2 \text{ (smooth) penalty for non co-localized edges}} \quad (1)$$

in which $\widehat{\Sigma}$ is the empirical estimates of Σ, $\widehat{\Sigma}_{\backslash j\backslash j}$ is the same matrix deprived from row and column j, $\widehat{\Sigma}_{j\backslash j}$ is row j of the empirical covariance matrix deprived from entry j, E_1 is the list of known co-localized genes and E_2 is the list of genes known not to be co-localized. λ and μ are two positive hyper-parameters that respectively control the sparsity of the solution and its conformity to *a priori* co-localization of information.

The idea behind the model of Equation (1) is that edge estimation must be enforced for pairs of genes that are known to be co-localized whereas the absence of an edge must be enforced for pairs of genes that are known not to be co-localized.

2 Practical implementation of network inference

The same method, based on a bootstrapping scheme than the one described in (Villa-Vialaneix et al., 2014) was used to perform the inference while ensuring the robustness of the estimation: $B = 100$ bootstrap samples were drawn from the original dataset. Inference (*i.e.*, the minimization, for all $j = 1, \ldots, p$, of Equation (1)) was performed for every bootstrap sample and a fixed value of μ. The inference was performed for the complete set of values for λ along the regularization path (Friedman, Hastie, and Tibshirani, 2010). The value of λ that ensured at least T_1 edges in the network was kept (and T_1 was set to 20% of the number of pairs of nodes in the network). Only edges that appear in, at least, $T_2 = 15$ bootstrap samples were included in the final network.

Finally, μ was set to the minimum value such that all *a priori* information were recovered, which led to $\mu = 0.2$ in Network 2, $\mu = 0.3$ in Network 3, $\mu = 0.4$ in Network 3. All simulations were performed with the free statistical software R (R Core Team, 2017) (`https://cran.r-project.org`). The inference was performed using our own scripts (available at `https://github.com/tuxette/internet3D`) and the graphs were displayed and analyzed using the R package **igraph** (Csardi and Nepusz, 2006).

Bibliography

Butte, A. and I. Kohane (2000). "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements". In: *Proceedings of the Pacific Symposium on Biocomputing*, pp. 418–429. DOI: 10.1142/9789814447331_0040.

Csardi, G. and T. Nepusz (2006). "The igraph software package for complex network research". In: *InterJournal* Complex Systems. URL: `http://igraph.sf.net`.

Edwards, D. (1995). *Introduction to Graphical Modelling*. New York, USA: Springer.

Friedman, J., T. Hastie, and R. Tibshirani (2008). "Sparse inverse covariance estimation with the graphical lasso". In: *Biostatistics* 9.3, pp. 432–441. DOI: 10.1093/biostatistics/kxm045.

– (2010). "Regularization paths for generalized linear models via coordinate descent". In: *Journal of Statistical Software* 33.1, pp. 1–22.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `http://www.R-project.org`.

Villa-Vialaneix, N. et al. (2013). "The structure of a gene co-expression network reveals biological functions underlying eQTLs". In: *PLoS ONE* 8.4, e60045. DOI: 10.1371/journal.pone.0060045.

Villa-Vialaneix, N. et al. (2014). "Inferring networks from multiple samples with concensus LASSO". In: *Quality Technology and Quantitative Management* 11.1, pp. 39–60. URL: `http://www.cc.nctu.edu.tw/~qtqm/qtqmpapers/2014V11N1/2014V11N1_F3.pdf`.

Appendix 2. Gene description and cluster allocation.

Expression data of the 359 genes is available on the NCBI/GEO database with the following accession number GSE56301. The gene annotation was improved compared to the one given in the original publication (Voillet et al., 2014). Target genes tested by 3D DNA FISH are in red bold.

			Netwrork 0	Netwrork 1	Netwrork 2	Netwrork 3
Gene symbol	Gene description	ProbeName	cluster	cluster	cluster	cluster
ABCB7	ATP binding cassette subfamily B member 7 [Source:HGNC Symbol;Acc:HGNC:48]	gi|115551474|dbj|AK236152.1|	6	2	7	4
ABI3 bp	ABI family member 3 binding protein [Source:HGNC Symbol;Acc:HGNC:17265]	A_72_P039386	5	1	5	1
ABR	active BCR-related [Source:HGNC Symbol;Acc:HGNC:81]	A_72_P496416	3	3	1	3
ACACB	acetyl-CoA carboxylase beta [Source:HGNC Symbol;Acc:HGNC:85]	A_72_P582352	2	2	2	2
ACADS	acyl-CoA dehydrogenase, C-2 to C-3 short chain [Source:HGNC Symbol;Acc:HGNC:90]	A_72_P077821	2	2	7	2
ACAT1	acetyl-CoA acetyltransferase 1 [Source:HGNC Symbol;Acc:HGNC:93]	A_72_P414863	2	5	5	5
ADAMTSL3	ADAMTS like 3 [Source:HGNC Symbol;Acc:HGNC:14633]	A_72_P365958	1	5	1	1
ADH5	alcohol dehydrogenase 5 (class III), chi polypeptide [Source:HGNC Symbol;Acc:HGNC:253]	O1836	8	1	3	3
ADIPOR2	adiponectin receptor 2 [Source:HGNC Symbol;Acc:HGNC:24041]	O13159	9	2	1	2
AKAP11	A-kinase anchoring protein 11 [Source:HGNC Symbol;Acc:HGNC:369]	O1634	3	3	3	3
AKR7A2	aldo-keto reductase family 7 member A2 [Source:HGNC Symbol;Acc:HGNC:389]	O4483	4	4	4	4
ALKBH5	alkB homolog 5, RNA demethylase [Source:HGNC Symbol;Acc:HGNC:25996]	A_72_P016101	4	4	4	4
ANPEP	alanyl aminopeptidase, membrane [Source:HGNC Symbol;Acc:HGNC:500]	gi|47523627|ref|NM_214277.1|	9	6	2	5
ANXA2	annexin A2 [Source:HGNC Symbol;Acc:HGNC:537]	A_72_P554722	5	5	1	1
ANXA3	annexin A3 [Source:HGNC Symbol;Acc:HGNC:541]	A_72_P149216	3	3	3	3
ANXA5	annexin A5 [Source:HGNC Symbol;Acc:HGNC:543]	gi|115547936|dbj|AK234913.1|	5	5	2	5
APOO	apolipoprotein O [Source:HGNC Symbol;Acc:HGNC:28727]	A_72_P350828	2	2	7	2
ARF3	ADP ribosylation factor 3 [Source:HGNC Symbol;Acc:HGNC:654]	A_72_P154956	7	2	7	2
ARHGAP11A	Rho GTPase activating protein 11A [Source:HGNC Symbol;Acc:HGNC:15783]	A_72_P542588	7	5	7	5
ARL3	ADP ribosylation factor like GTPase 3 [Source:HGNC Symbol;Acc:HGNC:694]	A_72_P035416	5	5	5	5
ASB11	ankyrin repeat and SOCS box containing 11 [Source:HGNC Symbol;Acc:HGNC:17186]	A_72_P029861	6	6	4	6
ATAT1	alpha tubulin acetyltransferase 1 [Source:HGNC Symbol;Acc:HGNC:21186]	OTTSUST00000001027	2	2	7	2
ATF5	activating transcription factor 5 [Source:HGNC Symbol;Acc:HGNC:790]	O7841	4	4	4	4
ATP1B4	ATPase Na+/K+ transporting family member beta 4 [Source:HGNC Symbol;Acc:HGNC:808]	A_72_P223637	4	4	4	4
ATP2A1	ATPase sarcoplasmic/endoplasmic reticulum Ca2+ transporting 1 [Source:HGNC Symbol;Acc:HGNC:811]	A_72_P127906	2	2	7	2
ATP2A2	ATPase sarcoplasmic/endoplasmic reticulum Ca2+ transporting 2 [Source:HGNC Symbol;Acc:HGNC:812]	A_72_P601148	6	3	3	3
ATP2B4	ATPase plasma membrane Ca2+ transporting 4 [Source:HGNC Symbol;Acc:HGNC:817]	O12415	6	6	7	3
ATP5B	ATP synthase, H+ transporting, mitochondrial F1 complex, beta polypeptide [Source:HGNC Symbol;Acc:HGNC:830]	A_72_P563576	2	2	7	2
ATP5O	ATP synthase, H+ transporting, mitochondrial F1 complex, O subunit [Source:HGNC Symbol;Acc:HGNC:850]	A_72_P695881	2	6	6	6
ATP6V0A1	ATPase H+ transporting V0 subunit a1 [Source:HGNC Symbol;Acc:HGNC:865]	A_72_P081171	8	1	6	3
ATP6V0D1	ATPase H+ transporting V0 subunit d1 [Source:HGNC Symbol;Acc:HGNC:13724]	O10518	3	3	8	4
ATXN3	ataxin 3 [Source:HGNC Symbol;Acc:HGNC:7106]	O4703	4	3	8	4
B3GNTL1	UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase like 1 [Source:HGNC Symbol;Acc:HGNC:21727]	O12993	3	3	3	3
BAHD1	bromo adjacent homology domain containing 1 [Source:HGNC Symbol;Acc:HGNC:29153]	O10500	1	6	6	2
BCAT2	branched chain amino acid transaminase 2 [Source:HGNC Symbol;Acc:HGNC:977]	O9326	8	2	7	2
BLVRB	biliverdin reductase B [Source:HGNC Symbol;Acc:HGNC:1063]	A_72_P037556	8	3	3	3
BNIP1	BCL2 interacting protein 1 [Source:HGNC Symbol;Acc:HGNC:1082]	A_72_P081211	9	6	5	5
BSG	basigin (Ok blood group) [Source:HGNC Symbol;Acc:HGNC:1116]	A_72_P064421	8	2	3	2
BZW2	basic leucine zipper and W2 domains 2 [Source:HGNC Symbol;Acc:HGNC:18808]	O9786	4	4	4	4
C8G	complement C8 gamma chain [Source:HGNC Symbol;Acc:HGNC:1354]	A_72_P077791	8	3	3	4
CAMKK1	calcium/calmodulin dependent protein kinase kinase 1 [Source:HGNC Symbol;Acc:HGNC:1469]	O13469	5	5	5	5
CAPN10	calpain 10 [Source:HGNC Symbol;Acc:HGNC:1477]	A_72_P165216	6	6	2	2
CCNG1	cyclin G1 [Source:HGNC Symbol;Acc:HGNC:1592]	A_72_P313048	4	3	3	3

CD81	CD81 molecule [Source:HGNC Symbol;Acc:HGNC:1701]	A_72_P698441	1	6	6	6
CDK6	cyclin dependent kinase 6 [Source:HGNC Symbol;Acc:HGNC:1777]	O14748	2	2	7	2
CDK9	cyclin dependent kinase 9 [Source:HGNC Symbol;Acc:HGNC:1780]	A_72_P623883	4	4	4	4
CDKN1C	cyclin dependent kinase inhibitor 1C [Source:HGNC Symbol;Acc:HGNC:1786]	A_72_P240467	3	3	8	4
CDR2L	cerebellar degeneration related protein 2 like [Source:HGNC Symbol;Acc:HGNC:29999]	A_72_P263187	7	5	2	5
CELF1	CUGBP Elav-like family member 1 [Source:HGNC Symbol;Acc:HGNC:2549]	O5626	4	4	3	4
CEND1	cell cycle exit and neuronal differentiation 1 [Source:HGNC Symbol;Acc:HGNC:24153]	A_72_P572049	6	4	4	4
CEP128	centrosomal protein 128 [Source:HGNC Symbol;Acc:HGNC:20359]	A_72_P141351	3	3	3	3
CEP72	centrosomal protein 72 [Source:HGNC Symbol;Acc:HGNC:25547]	A_72_P410118	7	5	7	5
CHMP2A	charged multivesicular body protein 2A [Source:HGNC Symbol;Acc:HGNC:30216]	A_72_P272689	1	3	8	4
CHRD	chordin [Source:HGNC Symbol;Acc:HGNC:1949]	O14303	6	6	6	6
CISD1	CDGSH iron sulfur domain 1 [Source:HGNC Symbol;Acc:HGNC:30880]	A_72_P560709	8	2	7	2
CITED1	Cbp/p300 interacting transactivator with Glu/Asp rich carboxy-terminal domain 1 [Source:HGNC Symbol;Acc:HGNC:1986]	O10495	3	3	3	3
CKM	creatine kinase, M-type [Source:HGNC Symbol;Acc:HGNC:1994]	A_72_P650270	3	3	3	3
CLCN2	chloride voltage-gated channel 2 [Source:HGNC Symbol;Acc:HGNC:2020]	gi|6002628|gb|AF093592.1|AF093592	6	6	4	1
CLCN5	chloride voltage-gated channel 5 [Source:HGNC Symbol;Acc:HGNC:2023]	A_72_P080321	2	5	5	5
CLUH	clustered mitochondria homolog [Source:HGNC Symbol;Acc:HGNC:29094]	A_72_P275574	2	2	7	2
COL12A1	collagen type XII alpha 1 chain [Source:HGNC Symbol;Acc:HGNC:2188]	A_72_P496541	2	2	7	2
COL16A1	collagen type XVI alpha 1 chain [Source:HGNC Symbol;Acc:HGNC:2193]	A_72_P427844	1	5	1	1
COL1A1	collagen type I alpha 1 chain [Source:HGNC Symbol;Acc:HGNC:2197]	gi|115553423|dbj|AK236626.1|	1	5	1	1
COL1A2	collagen type I alpha 2 chain [Source:HGNC Symbol;Acc:HGNC:2198]	gi|115551911|dbj|AK236318.1|	1	5	1	1
COL3A1	collagen type III alpha 1 chain [Source:HGNC Symbol;Acc:HGNC:2201]	A_72_P107126	1	5	1	1
COL4A1	collagen type IV alpha 1 chain [Source:HGNC Symbol;Acc:HGNC:2202]	A_72_P316078	5	6	4	4
COL5A1	collagen type V alpha 1 chain [Source:HGNC Symbol;Acc:HGNC:2209]	A_72_P077826	1	3	1	1
COL5A2	collagen type V alpha 2 chain [Source:HGNC Symbol;Acc:HGNC:2210]	A_72_P657188	1	5	1	1
COLEC12	collectin subfamily member 12 [Source:HGNC Symbol;Acc:HGNC:16016]	A_72_P403848	1	5	1	1
COQ7	coenzyme Q7, hydroxylase [Source:HGNC Symbol;Acc:HGNC:2244]	O9545	2	2	7	1
COX6C	cytochrome c oxidase subunit 6C [Source:HGNC Symbol;Acc:HGNC:2285]	A_72_P764976	6	6	6	6
CPT1B	carnitine palmitoyltransferase 1B [Source:HGNC Symbol;Acc:HGNC:2329]	A_72_P273299	2	2	2	2
CRAT	carnitine O-acetyltransferase [Source:HGNC Symbol;Acc:HGNC:2342]	A_72_P670021	8	2	3	3
CREB3L4	cAMP responsive element binding protein 3 like 4 [Source:HGNC Symbol;Acc:HGNC:18854]	A_72_P402968	4	4	4	4
CRIM1	cysteine rich transmembrane BMP regulator 1 [Source:HGNC Symbol;Acc:HGNC:2359]	O1283	4	4	4	4
CRLF1	cytokine receptor like factor 1 [Source:HGNC Symbol;Acc:HGNC:2364]	gi|74360255|gb|AJ943513.1|AJ943513	5	5	1	1
CS	citrate synthase [Source:HGNC Symbol;Acc:HGNC:2422]	A_72_P537441	2	2	7	2
CUL7	cullin 7 [Source:HGNC Symbol;Acc:HGNC:21024]	A_72_P443764	1	5	1	1
CYB5R1	cytochrome b5 reductase 1 [Source:HGNC Symbol;Acc:HGNC:13397]	A_72_P004426	8	4	6	6
DBF4	DBF4 zinc finger [Source:HGNC Symbol;Acc:HGNC:17364]	A_72_P141791	2	5	4	4
DBNDD2	dysbindin domain containing 2 [Source:HGNC Symbol;Acc:HGNC:15881]	A_72_P245492	8	2	2	2
DCN	decorin [Source:HGNC Symbol;Acc:HGNC:2705]	A_72_P180841	2	5	8	1
DCUN1D2	defective in cullin neddylation 1 domain containing 2 [Source:HGNC Symbol;Acc:HGNC:20328]	A_72_P028376	5	5	5	2
DLAT	dihydrolipoamide S-acetyltransferase [Source:HGNC Symbol;Acc:HGNC:2896]	A_72_P704857	2	2	7	2
DLD	dihydrolipoamide dehydrogenase [Source:HGNC Symbol;Acc:HGNC:2898]	A_72_P592144	2	2	7	2
DLEC1	deleted in lung and esophageal cancer 1 [Source:HGNC Symbol;Acc:HGNC:2899]	A_72_P333958	4	3	4	3
DLK1	delta like non-canonical Notch ligand 1 [Source:HGNC Symbol;Acc:HGNC:2907]	A_72_P035731	3	3	8	1
DNAL4	dynein axonemal light chain 4 [Source:HGNC Symbol;Acc:HGNC:2955]	A_72_P072746	3	3	8	4
DNMT1	DNA methyltransferase 1 [Source:HGNC Symbol;Acc:HGNC:2976]	A_72_P688426	5	5	1	1
DPP4	dipeptidyl peptidase 4 [Source:HGNC Symbol;Acc:HGNC:3009]	gi|47523581|ref|NM_214257.1|	7	5	4	4
DPYSL3	dihydropyrimidinase like 3 [Source:HGNC Symbol;Acc:HGNC:3015]	O398	8	1	3	4
DVL2	dishevelled segment polarity protein 2 [Source:HGNC Symbol;Acc:HGNC:3086]	O10386	2	2	7	2
DYNC2H1	dynein cytoplasmic 2 heavy chain 1 [Source:HGNC Symbol;Acc:HGNC:2962]	A_72_P362953	8	1	8	5
DYNLL1	dynein light chain LC8-type 1 [Source:HGNC Symbol;Acc:HGNC:15476]	A_72_P088696	5	5	1	5
ECI1	enoyl-CoA delta isomerase 1 [Source:HGNC Symbol;Acc:HGNC:2703]	O14547	2	2	7	2
EEF1A1	eukaryotic translation elongation factor 1 alpha 1 [Source:HGNC Symbol;Acc:HGNC:3189]	A_72_P746511	4	4	4	4
EGFR	epidermal growth factor receptor [Source:HGNC Symbol;Acc:HGNC:3236]	gi|47522839|ref|NM_214007.1|	6	3	4	3

EHHADH	enoyl-CoA hydratase and 3-hydroxyacyl CoA dehydrogenase [Source:HGNC Symbol;Acc:HGNC:3247]	O9867	3	4	3	4
EIF4E3	eukaryotic translation initiation factor 4E family member 3 [Source:HGNC Symbol;Acc:HGNC:31837]	A_72_P054986	5	5	8	4
EMP3	epithelial membrane protein 3 [Source:HGNC Symbol;Acc:HGNC:3335]	O11983	1	1	1	1
ENC1	ectodermal-neural cortex 1 [Source:HGNC Symbol;Acc:HGNC:3345]	A_72_P311268	7	5	7	2
ENDOG	endonuclease G [Source:HGNC Symbol;Acc:HGNC:3346]	A_72_P196482	2	2	7	2
ENGASE	endo-beta-N-acetylglucosaminidase [Source:HGNC Symbol;Acc:HGNC:24622]	O10819	4	4	4	4
EPAS1	endothelial PAS domain protein 1 [Source:HGNC Symbol;Acc:HGNC:3374]	A_72_P441713	8	4	3	3
EPB42	erythrocyte membrane protein band 4.2 [Source:HGNC Symbol;Acc:HGNC:3381]	O5966	6	4	4	4
ESR1	estrogen receptor 1 [Source:HGNC Symbol;Acc:HGNC:3467]	A_72_P444427	3	3	3	3
EXT2	exostosin glycosyltransferase 2 [Source:HGNC Symbol;Acc:HGNC:3513]	A_72_P337783	5	5	1	1
EZH2	enhancer of zeste 2 polycomb repressive complex 2 subunit [Source:HGNC Symbol;Acc:HGNC:3527]	O10045	7	5	7	5
FABP3	fatty acid binding protein 3 [Source:HGNC Symbol;Acc:HGNC:3557]	A_72_P440921	6	2	7	2
FAP	fibroblast activation protein alpha [Source:HGNC Symbol;Acc:HGNC:3590]	A_72_P207492	7	5	7	1
FARSA	phenylalanyl-tRNA synthetase alpha subunit [Source:HGNC Symbol;Acc:HGNC:3592]	O12975	8	2	7	2
FBLN7	fibulin 7 [Source:HGNC Symbol;Acc:HGNC:26740]	A_72_P431264	7	5	7	5
FBN1	fibrillin 1 [Source:HGNC Symbol;Acc:HGNC:3603]	A_72_P088351	5	5	1	1
FDFT1	farnesyl-diphosphate farnesyltransferase 1 [Source:HGNC Symbol;Acc:HGNC:3629]	A_72_P098361	6	2	2	2
FGF11	fibroblast growth factor 11 [Source:HGNC Symbol;Acc:HGNC:3667]	O5710	1	2	1	1
FGFR1	fibroblast growth factor receptor 1 [Source:HGNC Symbol;Acc:HGNC:3688]	A_72_P046056	1	6	1	1
FGFR4	fibroblast growth factor receptor 4 [Source:HGNC Symbol;Acc:HGNC:3691]	gi|7055965|gb|AW485859.1|AW485859	9	5	1	5
FGGY	FGGY carbohydrate kinase domain containing [Source:HGNC Symbol;Acc:HGNC:25610]	A_72_P153686	9	2	7	1
FKBP10	FK506 binding protein 10 [Source:HGNC Symbol;Acc:HGNC:18169]	A_72_P154676	1	5	5	1
FLAD1	flavin adenine dinucleotide synthetase 1 [Source:HGNC Symbol;Acc:HGNC:24671]	O12800	8	2	2	2
FLRT2	fibronectin leucine rich transmembrane protein 2 [Source:HGNC Symbol;Acc:HGNC:3761]	O8020	5	5	1	1
FSTL1	follistatin like 1 [Source:HGNC Symbol;Acc:HGNC:3972]	O10038	2	2	2	5
FXYD6	FXYD domain containing ion transport regulator 6 [Source:HGNC Symbol;Acc:HGNC:4030]	O9076	6	6	1	4
FYN	FYN proto-oncogene, Src family tyrosine kinase [Source:HGNC Symbol;Acc:HGNC:4037]	A_72_P146946	1	5	5	5
GANAB	glucosidase II alpha subunit [Source:HGNC Symbol;Acc:HGNC:4138]	O12854	1	6	6	1
GFRA1	GDNF family receptor alpha 1 [Source:HGNC Symbol;Acc:HGNC:4243]	A_72_P002701	7	5	7	4
GHITM	growth hormone inducible transmembrane protein [Source:HGNC Symbol;Acc:HGNC:17281]	O11946	2	2	7	2
GLT8D1	glycosyltransferase 8 domain containing 1 [Source:HGNC Symbol;Acc:HGNC:24870]	A_72_P594034	1	6	6	6
GLUD1	glutamate dehydrogenase 1 [Source:HGNC Symbol;Acc:HGNC:4335]	A_72_P615755	4	4	4	4
GMPPB	GDP-mannose pyrophosphorylase B [Source:HGNC Symbol;Acc:HGNC:22932]	A_72_P230202	6	6	6	6
GNAI3	G protein subunit alpha i3 [Source:HGNC Symbol;Acc:HGNC:4387]	O12082	7	6	7	2
GPI	glucose-6-phosphate isomerase [Source:HGNC Symbol;Acc:HGNC:4458]	A_72_P146806	7	2	7	2
GPRASP2	G protein-coupled receptor associated sorting protein 2 [Source:HGNC Symbol;Acc:HGNC:25169]	A_72_P205372	7	6	7	2
GPSM1	G protein signaling modulator 1 [Source:HGNC Symbol;Acc:HGNC:17858]	A_72_P218247	3	3	3	3
GPX3	glutathione peroxidase 3 [Source:HGNC Symbol;Acc:HGNC:4555]	A_72_P671275	2	2	2	2
HAUS1	HAUS augmin like complex subunit 1 [Source:HGNC Symbol;Acc:HGNC:25174]	O10898	5	1	5	1
HES6	hes family bHLH transcription factor 6 [Source:HGNC Symbol;Acc:HGNC:18254]	A_72_P677517	5	5	5	5
HK1	hexokinase 1 [Source:HGNC Symbol;Acc:HGNC:4922]	O5171	2	5	5	5
HMCES	5-hydroxymethylcytosine binding, ES cell specific [Source:HGNC Symbol;Acc:HGNC:24446]	O4725	9	3	3	3
HMGB2	high mobility group box 2 [Source:HGNC Symbol;Acc:HGNC:5000]	A_72_P558174	5	2	7	2
HMGN1	high mobility group nucleosome binding domain 1 [Source:HGNC Symbol;Acc:HGNC:4984]	O9747	5	5	5	5
HSP90B1	heat shock protein 90 beta family member 1 [Source:HGNC Symbol;Acc:HGNC:12028]	A_72_P232707	1	5	1	1
HSPA13	heat shock protein family A (Hsp70) member 13 [Source:HGNC Symbol;Acc:HGNC:11375]	A_72_P435494	5	5	2	5
HSPA9	heat shock protein family A (Hsp70) member 9 [Source:HGNC Symbol;Acc:HGNC:5244]	A_72_P388638	2	2	7	2
IBA57	IBA57 homolog, iron-sulfur cluster assembly [Source:HGNC Symbol;Acc:HGNC:27302]	A_72_P067206	8	2	6	2
ICOSLG	inducible T-cell costimulator ligand [Source:HGNC Symbol;Acc:HGNC:17087]	A_72_P287144	1	5	1	1
IDH3G	isocitrate dehydrogenase 3 (NAD(+)) gamma [Source:HGNC Symbol;Acc:HGNC:5386]	O6582	8	4	4	4
IGF2	insulin like growth factor 2 [Source:HGNC Symbol;Acc:HGNC:5466]	A_72_P303139	3	3	8	1
IGSF1	immunoglobulin superfamily member 1 [Source:HGNC Symbol;Acc:HGNC:5948]	A_72_P466893	2	2	5	2
IGSF3	immunoglobulin superfamily member 3 [Source:HGNC Symbol;Acc:HGNC:5950]	O13413	7	5	2	5
IL12RB2	interleukin 12 receptor subunit beta 2 [Source:HGNC Symbol;Acc:HGNC:5972]	A_72_P077956	4	4	4	4

IL18	interleukin 18 [Source:HGNC Symbol;Acc:HGNC:5986]	gi|47522819|ref|NM_213997.1|	2	2	5	5
INPP5F	inositol polyphosphate-5-phosphatase F [Source:HGNC Symbol;Acc:HGNC:17054]	A_72_P579372	4	3	3	4
INPP5J	inositol polyphosphate-5-phosphatase J [Source:HGNC Symbol;Acc:HGNC:8956]	A_72_P659813	5	3	3	4
IPO13	importin 13 [Source:HGNC Symbol;Acc:HGNC:16853]	O9886	8	4	3	3
IRAK1	interleukin 1 receptor associated kinase 1 [Source:HGNC Symbol;Acc:HGNC:6112]	A_72_P175791	4	4	4	4
ISYNA1	inositol-3-phosphate synthase 1 [Source:HGNC Symbol;Acc:HGNC:29821]	O6349	3	3	8	4
ITGA9	integrin subunit alpha 9 [Source:HGNC Symbol;Acc:HGNC:6145]	O14127	6	3	4	4
ITGB1	integrin subunit beta 1 [Source:HGNC Symbol;Acc:HGNC:6153]	O5530	7	6	7	2
ITIH4	inter-alpha-trypsin inhibitor heavy chain family member 4 [Source:HGNC Symbol;Acc:HGNC:6169]	O9843	3	1	5	1
JTB	jumping translocation breakpoint [Source:HGNC Symbol;Acc:HGNC:6201]	A_72_P387573	4	4	4	4
KCNC4	potassium voltage-gated channel subfamily C member 4 [Source:HGNC Symbol;Acc:HGNC:6236]	A_72_P055116	8	2	5	2
KCNG2	potassium voltage-gated channel modifier subfamily G member 2 [Source:HGNC Symbol;Acc:HGNC:6249]	O6864	6	4	4	2
KCNQ1	potassium voltage-gated channel subfamily Q member 1 [Source:HGNC Symbol;Acc:HGNC:6294]	A_72_P162961	3	3	3	3
KDELR2	KDEL endoplasmic reticulum protein retention receptor 2 [Source:HGNC Symbol;Acc:HGNC:6305]	O5029	7	5	7	5
KLF3	Kruppel like factor 3 [Source:HGNC Symbol;Acc:HGNC:16516]	A_72_P419349	1	3	3	3
LACTB2	lactamase beta 2 [Source:HGNC Symbol;Acc:HGNC:18512]	A_72_P680071	3	3	3	3
LAMA4	laminin subunit alpha 4 [Source:HGNC Symbol;Acc:HGNC:6484]	A_72_P379818	1	5	1	1
LAS1L	LAS1 like, ribosome biogenesis factor [Source:HGNC Symbol;Acc:HGNC:25726]	A_72_P531544	8	2	6	2
LBH	limb bud and heart development [Source:HGNC Symbol;Acc:HGNC:29532]	O3615	7	5	1	1
LDHD	lactate dehydrogenase D [Source:HGNC Symbol;Acc:HGNC:19708]	O7218	9	3	3	4
LFNG	LFNG O-fucosylpeptide 3-beta-N-acetylglucosaminyltransferase [Source:HGNC Symbol;Acc:HGNC:6560]	A_72_P232052	1	6	1	1
LHX2	LIM homeobox 2 [Source:HGNC Symbol;Acc:HGNC:6594]	A_72_P444597	3	3	3	3
LOXL2	lysyl oxidase like 2 [Source:HGNC Symbol;Acc:HGNC:6666]	A_72_P359358	7	5	2	5
LPAR4	lysophosphatidic acid receptor 4 [Source:HGNC Symbol;Acc:HGNC:4478]	A_72_P211927	7	5	7	4
LPIN2	lipin 2 [Source:HGNC Symbol;Acc:HGNC:14450]	A_72_P292904	7	5	7	5
LRRK1	leucine rich repeat kinase 1 [Source:HGNC Symbol;Acc:HGNC:18608]	A_72_P098446	7	6	7	2
MAFK	MAF bZIP transcription factor K [Source:HGNC Symbol;Acc:HGNC:6782]	A_72_P278329	8	1	3	4
MAGED1	MAGE family member D1 [Source:HGNC Symbol;Acc:HGNC:6813]	A_72_P149681	9	5	5	5
MAGED2	MAGE family member D2 [Source:HGNC Symbol;Acc:HGNC:16353]	A_72_P348613	2	5	5	5
MAGIX	MAGI family member, X-linked [Source:HGNC Symbol;Acc:HGNC:30006]	O7544	2	2	2	2
MAPRE1	microtubule associated protein RP/EB family member 1 [Source:HGNC Symbol;Acc:HGNC:6890]	A_72_P165071	5	5	5	1
MB	myoglobin [Source:HGNC Symbol;Acc:HGNC:6915]	A_72_P302979	3	3	3	3
MCEE	methylmalonyl-CoA epimerase [Source:HGNC Symbol;Acc:HGNC:16732]	A_72_P178091	8	3	3	2
MDH1	malate dehydrogenase 1 [Source:HGNC Symbol;Acc:HGNC:6970]	A_72_P303074	2	2	7	2
ME3	malic enzyme 3 [Source:HGNC Symbol;Acc:HGNC:6985]	A_72_P250467	8	1	3	3
MEG3	maternally expressed 3 (non-protein coding) [Source:HGNC Symbol;Acc:HGNC:14575]	A_72_P442171	4	4	8	1
MESDC2	Mesoderm Development LRP Chaperone	A_72_P501549	5	5	1	1
MESP1	mesoderm posterior bHLH transcription factor 1 [Source:HGNC Symbol;Acc:HGNC:29658]	A_72_P477678	3	3	3	3
MEST	mesoderm specific transcript [Source:HGNC Symbol;Acc:HGNC:7028]	A_72_P442223	2	5	8	1
MFAP5	microfibril associated protein 5 [Source:HGNC Symbol;Acc:HGNC:29673]	A_72_P293494	1	6	6	2
MGST3	microsomal glutathione S-transferase 3 [Source:HGNC Symbol;Acc:HGNC:7064]	O13326	6	2	7	4
MITF	melanogenesis associated transcription factor [Source:HGNC Symbol;Acc:HGNC:7105]	A_72_P444157	8	6	3	3
MPP6	membrane palmitoylated protein 6 [Source:HGNC Symbol;Acc:HGNC:18167]	A_72_P380523	3	3	3	3
MRPL37	mitochondrial ribosomal protein L37 [Source:HGNC Symbol;Acc:HGNC:14034]	O10627	4	3	3	4
MRPS2	mitochondrial ribosomal protein S2 [Source:HGNC Symbol;Acc:HGNC:14495]	O6203	9	2	7	2
MRPS28	mitochondrial ribosomal protein S28 [Source:HGNC Symbol;Acc:HGNC:14513]	O11321	4	2	7	2
MSANTD4	Myb/SANT DNA binding domain containing 4 with coiled-coils [Source:HGNC Symbol;Acc:HGNC:29383]	A_72_P089866	5	1	5	5
MSL2	MSL complex subunit 2 [Source:HGNC Symbol;Acc:HGNC:25544]	O12026	9	3	8	4
MUT	methylmalonyl-CoA mutase [Source:HGNC Symbol;Acc:HGNC:7526]	A_72_P441374	3	1	7	1
MXRA7	matrix remodeling associated 7 [Source:HGNC Symbol;Acc:HGNC:7541]	A_72_P298914	1	5	5	5
MYBL2	MYB proto-oncogene like 2 [Source:HGNC Symbol;Acc:HGNC:7548]	A_72_P350008	7	5	5	5
MYBPC2	myosin binding protein C, fast type [Source:HGNC Symbol;Acc:HGNC:7550]	O11393	4	3	3	4
MYH3	myosin heavy chain 3 [Source:HGNC Symbol;Acc:HGNC:7573]	A_72_P414973	3	3	3	1
MYPN	myopalladin [Source:HGNC Symbol;Acc:HGNC:23246]	A_72_P089766	6	6	6	4
NCAM1	neural cell adhesion molecule 1 [Source:HGNC Symbol;Acc:HGNC:7656]	A_72_P117001	9	5	3	3
NCEH1	neutral cholesterol ester hydrolase 1 [Source:HGNC Symbol;Acc:HGNC:29260]	A_72_P286854	2	3	3	2
NDP	NDP, norrin cystine knot growth factor [Source:HGNC Symbol;Acc:HGNC:7678]	A_72_P290889	3	3	3	3

NDUFA12	NADH:ubiquinone oxidoreductase subunit A12 [Source:HGNC Symbol;Acc:HGNC:23987]	O10344	8	2	7	2
NDUFA3	NADH:ubiquinone oxidoreductase subunit A3 [Source:HGNC Symbol;Acc:HGNC:7686]	O7196	8	2	7	2
NDUFB5	NADH:ubiquinone oxidoreductase subunit B5 [Source:HGNC Symbol;Acc:HGNC:7700]	A_72_P293684	8	2	7	2
NDUFS1	NADH:ubiquinone oxidoreductase core subunit S1 [Source:HGNC Symbol;Acc:HGNC:7707]	A_72_P089311	7	2	7	2
NES	nestin [Source:HGNC Symbol;Acc:HGNC:7756]	A_72_P002891	2	5	1	5
NFATC3	nuclear factor of activated T-cells 3 [Source:HGNC Symbol;Acc:HGNC:7777]	O1743	8	1	3	3
NFATC4	nuclear factor of activated T-cells 4 [Source:HGNC Symbol;Acc:HGNC:7778]	A_72_P119516	1	6	1	1
NFXL1	nuclear transcription factor, X-box binding like 1 [Source:HGNC Symbol;Acc:HGNC:18726]	A_72_P132466	3	3	3	3
NMNAT3	nicotinamide nucleotide adenylyltransferase 3 [Source:HGNC Symbol;Acc:HGNC:20989]	A_72_P147316	2	2	7	2
NNT	nicotinamide nucleotide transhydrogenase [Source:HGNC Symbol;Acc:HGNC:7863]	A_72_P397893	2	2	5	2
NT5DC1	5'-nucleotidase domain containing 1 [Source:HGNC Symbol;Acc:HGNC:21556]	A_72_P732133	4	4	3	4
NTMT1	N-terminal Xaa-Pro-Lys N-methyltransferase 1 [Source:HGNC Symbol;Acc:HGNC:23373]	A_72_P036826	2	2	7	2
NUP188	nucleoporin 188 [Source:HGNC Symbol;Acc:HGNC:17859]	A_72_P545332	6	6	4	6
OARD1	O-acyl-ADP-ribose deacylase 1 [Source:HGNC Symbol;Acc:HGNC:21257]	O9345	4	4	4	4
ODC1	ornithine decarboxylase 1 [Source:HGNC Symbol;Acc:HGNC:8109]	CUST_500_PI427286967	5	5	8	3
OGN	osteoglycin [Source:HGNC Symbol;Acc:HGNC:8126]	A_72_P255032	7	5	2	5
OXA1L	OXA1L, mitochondrial inner membrane protein [Source:HGNC Symbol;Acc:HGNC:8526]	O14558	2	2	7	2
P2RX5	purinergic receptor P2X 5 [Source:HGNC Symbol;Acc:HGNC:8536]	A_72_P236932	6	6	6	6
P4HA3	prolyl 4-hydroxylase subunit alpha 3 [Source:HGNC Symbol;Acc:HGNC:30135]	A_72_P367073	3	1	3	3
PAQR9	progestin and adipoQ receptor family member 9 [Source:HGNC Symbol;Acc:HGNC:30131]	A_72_P048236	4	4	3	4
PAX3	paired box 3 [Source:HGNC Symbol;Acc:HGNC:8617]	gi|46391809|gb|AY579430.1|	4	4	4	4
PAX7	paired box 7 [Source:HGNC Symbol;Acc:HGNC:8621]	A_72_P185391	3	3	3	3
PBX3	PBX homeobox 3 [Source:HGNC Symbol;Acc:HGNC:8634]	A_72_P543607	7	6	6	2
PCDH10	protocadherin 10 [Source:HGNC Symbol;Acc:HGNC:13404]	gi|90235591|gb|BX924774.2|BX924774	4	4	4	4
PDE4A	phosphodiesterase 4A [Source:HGNC Symbol;Acc:HGNC:8780]	A_72_P106041	3	3	3	3
PDHA1	pyruvate dehydrogenase alpha 1 [Source:HGNC Symbol;Acc:HGNC:8806]	A_72_P645767	2	2	7	2
PDHX	pyruvate dehydrogenase complex component X [Source:HGNC Symbol;Acc:HGNC:21350]	A_72_P190556	2	2	5	5
PDP1	pyruvate dehyrogenase phosphatase catalytic subunit 1 [Source:HGNC Symbol;Acc:HGNC:9279]	A_72_P199457	4	3	3	3
PEBP4	phosphatidylethanolamine binding protein 4 [Source:HGNC Symbol;Acc:HGNC:28319]	O620	6	3	3	6
PEG10	paternally expressed 10 [Source:HGNC Symbol;Acc:HGNC:14005]	A_72_P564724	3	3	8	4
PGD	phosphogluconate dehydrogenase [Source:HGNC Symbol;Acc:HGNC:8891]	O8797	7	6	7	1
PHKB	phosphorylase kinase regulatory subunit beta [Source:HGNC Symbol;Acc:HGNC:8927]	gi|74362209|gb|AJ945467.1|AJ945467	9	1	5	3
PHLDB1	pleckstrin homology like domain family B member 1 [Source:HGNC Symbol;Acc:HGNC:23697]	O9044	6	4	4	3
PINX1	PIN2/TERF1 interacting telomerase inhibitor 1 [Source:HGNC Symbol;Acc:HGNC:30046]	A_72_P101001	4	3	3	3
PKDREJ	polycystin family receptor for egg jelly [Source:HGNC Symbol;Acc:HGNC:9015]	O6639	4	4	4	4
PKIA	cAMP-dependent protein kinase inhibitor alpha [Source:HGNC Symbol;Acc:HGNC:9017]	A_72_P620119	5	1	3	5
PLXDC1	plexin domain containing 1 [Source:HGNC Symbol;Acc:HGNC:20945]	A_72_P357753	3	1	3	3
PLXNB2	plexin B2 [Source:HGNC Symbol;Acc:HGNC:9104]	A_72_P183761	1	6	6	6
PM20D1	peptidase M20 domain containing 1 [Source:HGNC Symbol;Acc:HGNC:26518]	A_72_P178791	3	3	3	3
PMEPA1	prostate transmembrane protein, androgen induced 1 [Source:HGNC Symbol;Acc:HGNC:14107]	A_72_P315448	7	6	1	5
PMPCB	peptidase, mitochondrial processing beta subunit [Source:HGNC Symbol;Acc:HGNC:9119]	A_72_P071856	1	6	6	6
POSTN	periostin [Source:HGNC Symbol;Acc:HGNC:16953]	A_72_P632486	1	5	1	1
PPA1	pyrophosphatase (inorganic) 1 [Source:HGNC Symbol;Acc:HGNC:9226]	A_72_P597416	4	4	4	4
PPP1CC	protein phosphatase 1 catalytic subunit gamma [Source:HGNC Symbol;Acc:HGNC:9283]	A_72_P671741	4	4	4	4
PPP1R14C	protein phosphatase 1 regulatory inhibitor subunit 14C [Source:HGNC Symbol;Acc:HGNC:14952]	A_72_P092786	4	4	4	4
PPP1R9A	protein phosphatase 1 regulatory subunit 9A [Source:HGNC Symbol;Acc:HGNC:14946]	A_72_P401473	3	4	4	4
PPTC7	PTC7 protein phosphatase homolog [Source:HGNC Symbol;Acc:HGNC:30695]	A_72_P099021	5	5	1	1
PRDX6	peroxiredoxin 6 [Source:HGNC Symbol;Acc:HGNC:16753]	A_72_P575489	4	4	4	4

PRELID1	PRELI domain containing 1 [Source:HGNC Symbol;Acc:HGNC:30255]	gi|40391838|gb|BP142367.1|BP142367	5	5	5	1
PRICKLE2	prickle planar cell polarity protein 2 [Source:HGNC Symbol;Acc:HGNC:20340]	A_72_P398358	7	5	7	5
PRKCQ	protein kinase C theta [Source:HGNC Symbol;Acc:HGNC:9410]	A_72_P442547	2	2	2	2
PTGES2	prostaglandin E synthase 2 [Source:HGNC Symbol;Acc:HGNC:17822]	O9510	2	2	2	2
PTGIR	prostaglandin I2 (prostacyclin) receptor (IP) [Source:HGNC Symbol;Acc:HGNC:9602]	A_72_P230127	8	1	3	5
PTMA	prothymosin, alpha [Source:HGNC Symbol;Acc:HGNC:9623]	A_72_P441011	6	5	7	4
PTPRF	protein tyrosine phosphatase, receptor type F [Source:HGNC Symbol;Acc:HGNC:9670]	A_72_P133066	1	6	1	1
PTPRJ	protein tyrosine phosphatase, receptor type J [Source:HGNC Symbol;Acc:HGNC:9673]	A_72_P404358	6	6	4	4
PYGM	glycogen phosphorylase, muscle associated [Source:HGNC Symbol;Acc:HGNC:9726]	A_72_P516452	8	2	7	2
RAB3A	RAB3A, member RAS oncogene family [Source:HGNC Symbol;Acc:HGNC:9777]	A_72_P549961	6	4	6	2
RAB3IP	RAB3A interacting protein [Source:HGNC Symbol;Acc:HGNC:16508]	A_72_P378108	3	3	3	3
RAP2B	RAP2B, member of RAS oncogene family [Source:HGNC Symbol;Acc:HGNC:9862]	gi|59806756|gb|DN113023.1|DN113023	7	5	7	4
RASA4	RAS p21 protein activator 4 [Source:HGNC Symbol;Acc:HGNC:23181]	A_72_P244002	3	4	3	3
RAVER1	ribonucleoprotein, PTB binding 1 [Source:HGNC Symbol;Acc:HGNC:30296]	A_72_P202207	7	6	7	2
RBM10	RNA binding motif protein 10 [Source:HGNC Symbol;Acc:HGNC:9896]	A_72_P418079	6	6	7	6
RBM15B	RNA binding motif protein 15B [Source:HGNC Symbol;Acc:HGNC:24303]	A_72_P046046	5	6	4	2
RCOR3	REST corepressor 3 [Source:HGNC Symbol;Acc:HGNC:25594]	A_72_P283769	2	2	5	1
RGS2	regulator of G protein signaling 2 [Source:HGNC Symbol;Acc:HGNC:9998]	A_72_P057121	3	1	5	1
RHOC	ras homolog family member C [Source:HGNC Symbol;Acc:HGNC:669]	A_72_P158146	5	5	5	5
RNF34	ring finger protein 34 [Source:HGNC Symbol;Acc:HGNC:17297]	A_72_P426884	1	3	3	3
RPL11	ribosomal protein L11 [Source:HGNC Symbol;Acc:HGNC:10301]	A_72_P041441	8	3	6	2
RPL3	ribosomal protein L3 [Source:HGNC Symbol;Acc:HGNC:10332]	O10058	5	5	1	1
RPL31	ribosomal protein L31 [Source:HGNC Symbol;Acc:HGNC:10334]	A_72_P294119	3	3	4	4
RPL32	ribosomal protein L32 [Source:HGNC Symbol;Acc:HGNC:10336]	A_72_P735568	3	3	8	1
RPS27A	ribosomal protein S27a [Source:HGNC Symbol;Acc:HGNC:10417]	A_72_P391738	5	3	8	4
RRAS2	related RAS viral (r-ras) oncogene homolog 2 [Source:HGNC Symbol;Acc:HGNC:17271]	A_72_P130806	5	3	8	3
RTN4IP1	reticulon 4 interacting protein 1 [Source:HGNC Symbol;Acc:HGNC:18647]	O13685	2	2	7	2
SATB1	SATB homeobox 1 [Source:HGNC Symbol;Acc:HGNC:10541]	O12598	9	5	2	5
SCN1B	sodium voltage-gated channel beta subunit 1 [Source:HGNC Symbol;Acc:HGNC:10586]	A_72_P058596	9	3	8	4
SCT	secretin [Source:HGNC Symbol;Acc:HGNC:10607]	A_72_P414018	5	5	5	5
SDC2	syndecan 2 [Source:HGNC Symbol;Acc:HGNC:10659]	A_72_P634951	5	5	5	5
SEC61B	Sec61 translocon beta subunit [Source:HGNC Symbol;Acc:HGNC:16993]	O9969	1	5	5	5
SELENBP1	selenium binding protein 1 [Source:HGNC Symbol;Acc:HGNC:10719]	A_72_P720208	3	3	3	3
SELO	Selenoprotein O	O10811	8	2	7	2
SGCE	sarcoglycan epsilon [Source:HGNC Symbol;Acc:HGNC:10808]	A_72_P136566	2	2	5	2
SKI	SKI proto-oncogene [Source:HGNC Symbol;Acc:HGNC:10896]	A_72_P079116	6	6	6	6
SKP2	S-phase kinase associated protein 2 [Source:HGNC Symbol;Acc:HGNC:10901]	O2895	7	5	7	1
SLBP	stem-loop binding protein [Source:HGNC Symbol;Acc:HGNC:10904]	A_72_P080921	5	5	1	1
SLC1A3	solute carrier family 1 member 3 [Source:HGNC Symbol;Acc:HGNC:10941]	O13275	2	2	5	2
SLC22A16	solute carrier family 22 member 16 [Source:HGNC Symbol;Acc:HGNC:20302]	O8716	4	4	3	4
SLC25A12	solute carrier family 25 member 12 [Source:HGNC Symbol;Acc:HGNC:10982]	A_72_P342738	5	5	5	5
SLC25A19	solute carrier family 25 member 19 [Source:HGNC Symbol;Acc:HGNC:14409]	A_72_P614041	4	3	3	4
SLC25A3	solute carrier family 25 member 3 [Source:HGNC Symbol;Acc:HGNC:10989]	A_72_P549036	2	2	7	2
SLC25A37	solute carrier family 25 member 37 [Source:HGNC Symbol;Acc:HGNC:29786]	A_72_P487571	4	3	3	4
SLC2A12	solute carrier family 2 member 12 [Source:HGNC Symbol;Acc:HGNC:18067]	O2906	3	3	3	3
SLC3A2	solute carrier family 3 member 2 [Source:HGNC Symbol;Acc:HGNC:11026]	gi|115554863|dbj|AK233675.1|	3	4	4	4
SLC41A1	solute carrier family 41 member 1 [Source:HGNC Symbol;Acc:HGNC:19429]	O11000	3	3	3	3
SLC46A3	solute carrier family 46 member 3 [Source:HGNC Symbol;Acc:HGNC:27501]	A_72_P131741	8	3	3	3
SLC9A2	solute carrier family 9 member A2 [Source:HGNC Symbol;Acc:HGNC:11072]	A_72_P088396	4	4	4	3
SMIM5	small integral membrane protein 5 [Source:HGNC Symbol;Acc:HGNC:40030]	A_72_P190151	4	3	3	4
SP7	Sp7 transcription factor [Source:HGNC Symbol;Acc:HGNC:17321]	A_72_P175886	4	4	4	4
SPAG11	Sperm Associated Antigen 11	A_72_P209602	5	1	3	5
SPG7	SPG7, paraplegin matrix AAA peptidase subunit [Source:HGNC Symbol;Acc:HGNC:11237]	gi|84128395|gb|CV874435.1|CV874435	8	2	2	2

SPI1	Spi-1 proto-oncogene [Source:HGNC Symbol;Acc:HGNC:11241]	gi|49274644|ref|NM_001001865.1|	4	4	3	4
SRI	sorcin [Source:HGNC Symbol;Acc:HGNC:11292]	A_72_P402828	1	1	6	3
SS18	SS18, nBAF chromatin remodeling complex subunit [Source:HGNC Symbol;Acc:HGNC:11340]	O5598	8	6	7	2
SSR2	signal sequence receptor subunit 2 [Source:HGNC Symbol;Acc:HGNC:11324]	O11071	1	5	5	1
ST5	suppression of tumorigenicity 5 [Source:HGNC Symbol;Acc:HGNC:11350]	O11683	8	3	3	3
STEAP3	STEAP3 metalloreductase [Source:HGNC Symbol;Acc:HGNC:24592]	A_72_P426009	4	3	3	4
STMN1	stathmin 1 [Source:HGNC Symbol;Acc:HGNC:6510]	A_72_P731693	5	5	5	5
STUB1	STIP1 homology and U-box containing protein 1 [Source:HGNC Symbol;Acc:HGNC:11427]	O14020	8	2	7	2
STXBP2	syntaxin binding protein 2 [Source:HGNC Symbol;Acc:HGNC:11445]	O6461	4	4	4	4
SVIL	supervillin [Source:HGNC Symbol;Acc:HGNC:11480]	O3767	8	1	3	3
SYDE1	synapse defective Rho GTPase homolog 1 [Source:HGNC Symbol;Acc:HGNC:25824]	A_72_P327048	1	5	1	1
SYT17	synaptotagmin 17 [Source:HGNC Symbol;Acc:HGNC:24119]	A_72_P138296	5	5	5	5
TBR1	T-box, brain 1 [Source:HGNC Symbol;Acc:HGNC:11590]	A_72_P471478	4	4	4	1
TENM1	teneurin transmembrane protein 1 [Source:HGNC Symbol;Acc:HGNC:8117]	A_72_P291424	2	2	7	2
TFRC	transferrin receptor [Source:HGNC Symbol;Acc:HGNC:11763]	A_72_P035411	4	4	4	4
TGFB3	transforming growth factor beta 3 [Source:HGNC Symbol;Acc:HGNC:11769]	gi|47523473|ref|NM_214198.1|	7	5	7	2
TIMP1	TIMP metallopeptidase inhibitor 1 [Source:HGNC Symbol;Acc:HGNC:11820]	O10493	5	3	8	4
TLE4	transducin like enhancer of split 4 [Source:HGNC Symbol;Acc:HGNC:11840]	A_72_P254987	9	5	5	5
TMEM119	transmembrane protein 119 [Source:HGNC Symbol;Acc:HGNC:27884]	A_72_P311568	1	6	2	1
TMEM150A	transmembrane protein 150A [Source:HGNC Symbol;Acc:HGNC:24677]	O13818	4	3	3	3
TMEM9	transmembrane protein 9 [Source:HGNC Symbol;Acc:HGNC:18823]	A_72_P266237	3	3	3	3
TNNT2	troponin T2, cardiac type [Source:HGNC Symbol;Acc:HGNC:11949]	O12845	2	5	5	5
TPBG	trophoblast glycoprotein [Source:HGNC Symbol;Acc:HGNC:12004]	A_72_P328828	5	5	4	1
TPM1	tropomyosin 1 [Source:HGNC Symbol;Acc:HGNC:12010]	A_72_P746394	4	4	4	4
TPPP3	tubulin polymerization promoting protein family member 3 [Source:HGNC Symbol;Acc:HGNC:24162]	O4242	5	5	5	5
TRAM2	translocation associated membrane protein 2 [Source:HGNC Symbol;Acc:HGNC:16855]	O9031	7	5	6	2
TRIM17	tripartite motif containing 17 [Source:HGNC Symbol;Acc:HGNC:13430]	O6090	4	4	4	4
TRIP6	thyroid hormone receptor interactor 6 [Source:HGNC Symbol;Acc:HGNC:12311]	gi|40800175|gb|CK452961.1|CK452961	3	3	3	3
TRNP1	TMF1-regulated nuclear protein 1 [Source:HGNC Symbol;Acc:HGNC:34348]	A_72_P210252	2	2	5	5
TSPAN7	tetraspanin 7 [Source:HGNC Symbol;Acc:HGNC:11854]	A_72_P499239	9	3	8	4
TUBA4A	tubulin alpha 4a [Source:HGNC Symbol;Acc:HGNC:12407]	A_72_P209947	4	4	3	4
TUBA8	tubulin alpha 8 [Source:HGNC Symbol;Acc:HGNC:12410]	O9364	2	1	7	1
TXNRD2	thioredoxin reductase 2 [Source:HGNC Symbol;Acc:HGNC:18155]	O1892	2	2	7	2
TYRO3	TYRO3 protein tyrosine kinase [Source:HGNC Symbol;Acc:HGNC:12446]	gi|115555033|dbj|AK240232.1|	7	5	7	5
UCP3	uncoupling protein 3 [Source:HGNC Symbol;Acc:HGNC:12519]	A_72_P443167	3	1	3	3
UNC5CL	unc-5 family C-terminal like [Source:HGNC Symbol;Acc:HGNC:21203]	A_72_P600728	2	2	7	2
URAD	ureidoimidazoline (2-oxo-4-hydroxy-4-carboxy-5-) decarboxylase [Source:HGNC Symbol;Acc:HGNC:17785]	O13849	4	4	4	4
USP25	ubiquitin specific peptidase 25 [Source:HGNC Symbol;Acc:HGNC:12624]	O3111	5	5	2	5
VASN	vasorin [Source:HGNC Symbol;Acc:HGNC:18517]	A_72_P379898	3	1	3	3
VCAN	versican [Source:HGNC Symbol;Acc:HGNC:2464]	A_72_P409933	2	5	5	5
VDAC1	voltage dependent anion channel 1 [Source:HGNC Symbol;Acc:HGNC:12669]	A_72_P114441	2	2	2	2
VDAC2	voltage dependent anion channel 2 [Source:HGNC Symbol;Acc:HGNC:12672]	A_72_P641463	1	2	1	1
VIM	vimentin [Source:HGNC Symbol;Acc:HGNC:12692]	A_72_P036391	2	2	7	2
XYLB	xylulokinase [Source:HGNC Symbol;Acc:HGNC:12839]	A_72_P173691	9	6	2	5
YWHAB	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein beta [Source:HGNC Symbol;Acc:HGNC:12849]	A_72_P330693	7	6	7	2
YWHAQ	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein theta [Source:HGNC Symbol;Acc:HGNC:12854]	gi|115546630|dbj|AK234468.1|	7	2	7	2
ZCCHC17	zinc finger CCHC-type containing 17 [Source:HGNC Symbol;Acc:HGNC:30246]	A_72_P046551	8	1	3	3
ZFP36L1	ZFP36 ring finger protein like 1 [Source:HGNC Symbol;Acc:HGNC:1107]	A_72_P601188	1	6	6	2
ZHX1	zinc fingers and homeoboxes 1 [Source:HGNC Symbol;Acc:HGNC:12871]	A_72_P500455	5	1	5	5
ZNF521	zinc finger protein 521 [Source:HGNC Symbol;Acc:HGNC:24605]	A_72_P292659	9	3	8	4

Appendix 3. Biological network reconstructed following Ingenuity data analyses.

(a) The genes list from cluster 1 of Network 3 was submitted to IPA (Ingenuity Pathway Analysis). Here, only consistent information from the proposed networks was used to reconstruct a biological network. The three first IPA proposed networks were merged. Only genes from cluster 1 of Network 3 (red bold) and nodes (genes/molecules, black bold) necessary to connect the cluster genes, were kept. Genes tested by 3D DNA FISH are in green bold. **(b)** Upstream Regulators analysis (extraction of relevant information) allowed to identify some transcription factors that could explain unexpected co-expression and nuclear co-localization (especially between IGF2 and MYH3 genes) identified in this study. **(c)** The list of genes from the final reconstructed network was submitted to IPA for biological interpretation (extraction of relevant information).

(a) IPA networks analysis

Analysis	Score	**Focus Molecules**	Top Diseases and Functions	Molecules in Network
Network 1	105	49	Cancer, Connective Tissue Disorders, Organismal Injury and Abnormalities	**ABI3 bp**, **ADAMTSL3**, ADGRG6, **ADRB2**, AMELX, **ANXA2**, AR-HSP90, **Akt**, Akt-Calmodulin-Hsp90-Nos3, CANX, **CD28**, **CDK4**, **CDKN3**, CFAP44, CHADL, **CLCN2**, COL12A1, COL13A1, **COL16A1**, COL19A1, **COL1A1**, **COL1A2**, COL20A1, COL21A1, COL22A1, COL23A1, COL24A1, COL25A1, COL27A1, COL28A1, **COL3A1**, COL4A4, COL4A6, **COL5A1**, **COL5A2**, COL6A5, COL6A6, COL9A2, COL9A3, **COLEC12**, **COQ7**, **CRLF1**, **CUL7**, Calmodulin-Hsp90-Nos3, Col10a1, Collagen type I, Collagen type III, Collagen type X, Collagen type XVIII, Collagen(s), **DCN**, **DLK1**, **DNMT1**, DUSP28, Dnajc7-Hsp90-Nr1i3, **ELAVL1**, **EMP3**, EPYC, **EXT2**, Erbb2 dimer, **FAP**, **FBN1**, **FBXO6**, FGF dimer, **FGF11**, FGF14, FGF22, **FGFR1**, **FKBP10**, FLRT1, **FLRT2**, Fgf, **GANAB**, GPR137, GPR146, **GRK5**, **HNF4A**, HPN, **HSP90B1**, Histone h3, Hsp84-2, Hsp90, Hsp90-Ppard, **ICOSLG**/LOC102723996, **IGF2**, ITGA10, ITGA11, **ITGA3**, **ITGAM**, **ITGB3**, **ITIH4**, **Integrin**, JAK2, KERA, KIAA0895, KLF12, **KRT40**, LAIR2, **LAMA4**, **MESDC2**, **MEST**, **MUT**, **NFATC4**, NUDT11, Nlrp4a, OPTC, **PGD**, PLXNC1, **POSTN**, **PPTC7**, **PTPRF**, **RCOR3**, RGD1560020_predicted, **RGS2**, **RPL3**, **RPL32**, **RUNX2**, **SKP2**, **SLBP**, SMIM12, SMIM7, **SSR2**, TMEM101, **TMEM119**, TMPRSS6, TRAM2, TSSK4, **TUBA8**, **TWIST1**, **VDAC2**, VN1R1, WWOX, XPNPEP2, Xap2-Hsp90-Ppara, **adenosine triphosphate**, bilirubin, collagen, factor XIII, riboflavin, ribose
Network 2	9	8	Cell Cycle, Cell Morphology, Cellular Assembly and Organization	ABCB1, ABLIM1, ACTR3, ADD1, ADD3, AICDA, AMER1, ANK3, ANP32A, ANP32B, AP3D1, APC/APC2, APC2, ARFGAP3, ARMC8, AXIN1, AXIN2, BCL3, BEGAIN, CA9, CBY1, CC2D1A, CCDC85C, CDCA8, CDH16, CEACAM1, CEP290, CEP350, CLASP1, CLASP2, CLINT1, CLTA, COPS8, CPSF4, CRYAB, CRYBG1, CSNK1D, **CTNNB1,** DES, DIAPH3, DMD, DNAJB4, DNAJB6, DNAJC11, DNAJC6, DNAJC8, DR1, DVL3, EIF5A, ERBIN, FANCG, FBRS, FERMT2, FOXC1, FOXC2, FOXO4, GFAP, GNB4, **HAUS1**, HAUS2, HAUS3, HAUS4, HAUS5, HAUS6, HAUS7, HAUS8, HMG20B, HNRNPM, HOOK2, HSPB11, HSPB8, HSPE1, **IGF2 bp1**, KIAA2013, KIF20A, L3 MbTL3, **LBH**, LEO1, **LFNG**, MAML1, MAPKAPK2, **MAPRE1**, MIS12, **MYH3**, MYH6, MYL4, MYLK, Macf1, NEURL2, NIPSNAP1, NIPSNAP2, NUMB, NUP62, PACSIN3, PDAP1, PDE4B, PDE4DIP, PIBF1, PKN1, PKP3, POC5, PPP1R13L, PPP1R2, **PRELID1**, PRKACB, PTH1R, RAB11FIP5, RANBP3, **RAPGEF2**, RPL21, Rnr, S100A4, SAA1, SCN5A, SEPT9, **SHOX2**, SIX1, SPAG5, STARD7, STIM1, STRN3, STXBP1, **SYDE1**, TADA3, TANC2, TBL1X, TBL1XR1, **TBR1**, TELO2, TFPT, TIAL1, TOB2, TRA, TRIM29, TRIM33, TTC26, TUBA4A, VPS52, XPO1, miR-92a-3p (and other miRNAs w/seed AUUGCAC)
Network 3	2	1	Cancer, Cell Cycle, Cellular Development	BMI1, **MEG3**
Network 4	2	1	Cellular Movement, Embryonic Development, Amino Acid Metabolism	GIPC1, TPBG
Network 5	2	1	Cellular Assembly and Organization, Cellular Function and Maintenance, Molecular Transport	EAF1, FGGY, NSFL1C

(b) IPA Upstream regulators analysis (extract)

Analysis	Upstream Regulator	p-value of overlap	Molecule Type	Number of genes	Gene names
Upstream Regulators	MYOD1	1.42E-02	transcription regulator	3	**IGF2**, **MYH3**, POSTN
Upstream Regulators	CTNNB1	1.50E-02	transcription regulator	6	COL1A1, **IGF2**, LBH, LFNG, **MYH3**, TBR1

(c) Biological functions analysis of the IPA reconstructed network.

Categories	Diseases or Functions Annotation	p-Value	Molecules	Molecules
Tissue Morphology	Quantity of cells	2,48E-09	31	ADRB2, CD28, CDK4, COL1A1, CRLF1, CTNNB1, CUL7, **DCN**, **DLK1**, DNMT1, ELAVL1, FGFR1, HNF4A, HSP90B1, ICOSLG/LOC102723996, **IGF2**, ITGA3, ITGAM, ITGB3, JAK2, LAMA4, LFNG, MYOD1, POSTN, PTPRF, RAPGEF2, RUNX2, SKP2, STARD7, TBR1, TWIST1
Cell Morphology	Sprouting	1,75E-08	14	ANXA2, CTNNB1, **DCN**, ELAVL1, FGFR1, **IGF2**, ITGA3, ITGB3, JAK2, LAMA4, NFATC4, PTPRF, RAPGEF2, RUNX2
Organ Development	Formation of muscle	2,98E-05	10	CTNNB1, ELAVL1, FGFR1, HSP90B1, **IGF2**, **MYH3**, MYOD1, NFATC4, RGS2, TWIST1
Tissue Morphology	Morphology of connective tissue cells	1,27E-04	8	ADRB2, CDK4, CLCN2, **DLK1**, ITGB3, **MEST**, POSTN, RUNX2

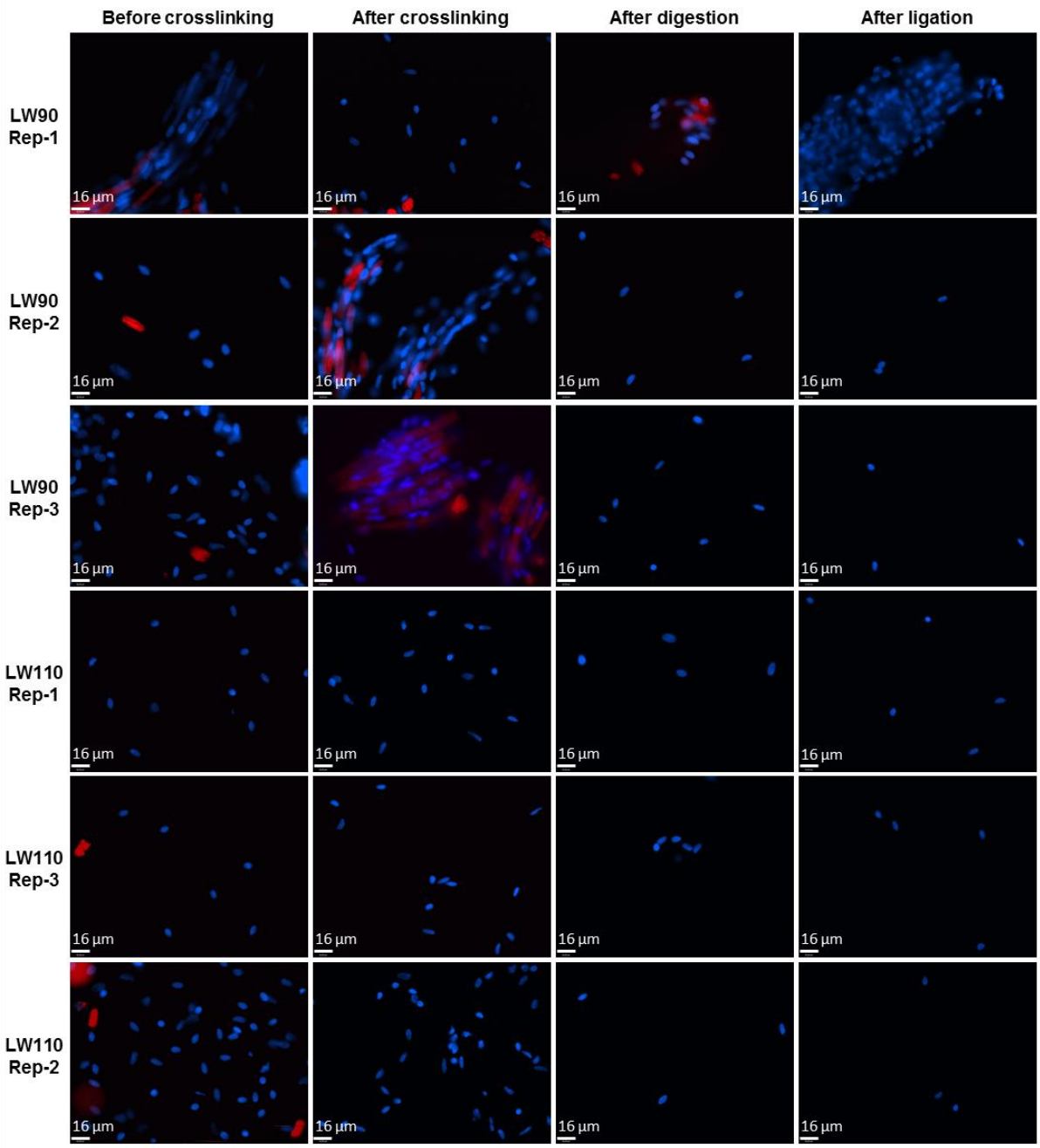
Appendix 4. Information about BACs used as probes for 3D DNA FISH experiments.

Information about BACs used as probes for 3D DNA FISH experiments.

Bacterial artificial clones (BACs) containing genes were isolated from porcine BAC libraries (available at the Biological Resources Center-GADIE, INRA, Jouy-en-Josas, France: http://crb-gadie.inra.fr/) using specific primers designed with Primer3 software (http://primer3.sourceforge.net/). *DCN was found non-imprinted in pig and imprinted in mouse (source: http://www.geneimprint.com/site/genes/Sus_scrofa_DCN.cdna) but its imprinting status is not clear in human. **Estimated size from human sequence. Fwd for (forward primers) and Rev for (reverse primers). ŧImprinted Gene Network (ING) of Varrault et al., 2006.

Gene Symbol (alias)	Full name	Imprinting Status	Expressed allele in Human	Present in IGN‡	Gene size (bp)	Porcine sequence	BAC number	Specific primers	Size PCR products	Localization in pig
IGF2	Insulin growth factor 2	imprinted	Paternal	yes	20491**	NM_213883	PigI-370D12	Fwd: TCCAGGTGTCATAGCGGAAG Rev: GCATCGTGGAAGAGTGCTG	458 bp	SSC2p17
DLK1	Delta-like 1 homolog (Drosophila)	imprinted	Paternal	yes	10870	EU597631	PigI-790B6	Fwd: GGCATCGTCTTCCTCAACAA Rev: CGCTGCTTAGATCTCCTC	194 bp	SSC7q26
MEG3 (GTL2)	Maternally expressed 3 (non-protein coding)	imprinted	Maternal	yes	81621**	EF468461	PigI-790B6	Fwd: ACCAGCCTACGAAGAAAGCC Rev: GGAGAATAAATGAGACGGTGAG	644 bp	SSC7q26
ZAR1	Zygote arrest 1	No imprinted	Both	no	4968	NM_001129956	PigI-427B5	Fwd: CCTGTTTCCTCCTCTGACGG Rev: CCACGTCTCGAATGCTGACT	169 bp	SSC8q11-12
MYH3 (MYHC)	Myosin Heavy Chain 3	No imprinted	Both	no	24813	XM_013981330.1	CH242-247C19	Fwd: TGGGTGCTCTTAATTTTCCCCT Rev: AAATGGAAGTGTTCGGCATAGC	151 pb	SSC12q
MEST (PEG1)	Mesoderm Specific Transcript	imprinted	Paternal	yes	60099	EF619473	SBAB-484H11	Fwd: GTCGATGACCAAGTTCCCGT Rev: TCTCAAAGATGGAGGCGTGC	424pb	SSC18
RPL32	Ribosomal protein L32	No imprinted	Both	yes	35317	EW027713	PigI-104F11	Fwd: CTGGCATTGGGATTGGTG Rev: TCATACTGTGCTGAGATTGCTC	104 bp	SSC13q24-33
DCN	Decorin	No imprinted*	Both*	yes	7103	NM_213920	PigI-849B12	Fwd: AACAACAACATCTCTGCAGTCG Rev: GGGAGCTACTTGTAGTTTCCGA	185 pb	SSC5qter
PRLR	Prolactin Receptor	No imprinted	Both	no	37230	NM_001001868.1	CH242-255C20	Fwd: ATAGCCCTTCAAAGCCACTG Rev: CTTGTCCAGGTTCGCTGC	1431 pb	SSC16

Appendix 5. Quality check of nuclear integrity in Hi C experimental steps.



Appendix 6. Evolution of the betweenness and degree values of a subset of genes from Network 0 to Network 3.

Genes are sorted by alphabetical order. Genes that were tested by 3D DNA FISH are in red bold.

	Network 0		Network 1		Network 2		Network 3		Comparison between Network 0 and Network 3 (% of variation)	
gene symbol	degree	betweeness	degree	betweeness	degree	betweeness	degree	betweeness	Degree	betweeness
ADIPOR2	15	646,65	14	487,32	15	628,78	14	660,97	-7	2
AKR7A2	19	492,63	17	436,71	15	474,10	14	291,90	-26	-41
CD81	17	551,17	18	616,7	15	478,76	17	600,58	0	9
CRAT	19	716,24	15	518,26	16	738,30	14	573,58	-26	-20
DCN	16	438,86	18	560,83	9	288,82	6	357,74	-63	-18
DLK1	10	103,52	6	81,7	5	74,22	5	24,13	-50	-77
DPP4	15	568,91	16	672,01	15	674,94	15	597,87	0	5
EGFR	16	624,92	12	375,87	12	385,35	11	354,78	-31	-43
GHITM	16	578,58	17	588,76	16	592,35	14	496,63	-13	-14
GLUD1	13	575,69	13	553,28	12	574,48	12	586,27	-8	2
IGF2	10	118,26	11	231,09	8	260,58	7	622,44	-30	426
LPAR4	14	464,31	17	644,76	18	812,81	16	798,82	14	72
MEG3	13	282,32	5	55,75	6	120,18	5	24,13	-62	-91
MESP1	12	228,49	14	320,34	14	483,27	14	775,31	17	239
MEST	13	148,2	12	121,44	10	345,69	7	385,27	-46	160
MRPS28	16	743	15	743,29	16	953,42	15	796,14	-6	7
MYH3	14	610,73	14	656,6	11	455,62	4	0,00	-71	-100
NMNAT3	17	562,63	18	664,84	16	473,55	17	573,15	0	2
RAVER1	16	613,84	16	665,73	16	696,35	16	745,66	0	21
RPL32	18	717,96	15	557,65	7	149,80	5	243,11	-72	-66
SELO	18	692,52	14	438,35	14	459,46	15	587,32	-17	-15
SYDE1	15	436,75	17	530,29	14	459,66	18	745,52	20	71
TFRC	15	595,1	15	534,83	13	437,38	17	846,81	13	42
TYRO3	20	785,95	18	659,9	16	603,94	17	700,03	-15	-11
YWHAB	20	670,22	17	470,35	17	538,41	17	547,17	-15	-18

Appendix 7. Clusters parameters.

Network	Cluster	Community sizes	Density	Transitivity
0	1	39	0,1525	0,2167
	2	57	0,1197	0,2062
	3	47	0,1582	0,2306
	4	51	0,1435	0,2502
	5	44	0,1533	0,2145
	6	28	0,1958	0,3115
	7	36	0,1714	0,237
	8	39	0,1498	0,2218
	9	18	0,2484	0,3169
1	1	27	0,1966	0,2404
	2	76	0,1035	0,1703
	3	74	0,1085	0,2253
	4	50	0,138	0,2371
	5	88	0,0925	0,2015
	6	44	0,129	0,2619
2	1	39	0,1404	0,2231
	2	24	0,2065	0,3279
	3	78	0,1012	0,189
	4	50	0,129	0,2285
	5	45	0,1384	0,1927
	6	24	0,1993	0,3456
	7	76	0,1	0,1848
	8	23	0,2174	0,4249
3	1	60	0,0977	0,2255
	2	86	0,0848	0,1587
	3	62	0,1163	0,205
	4	80	0,0972	0,2107
	5	56	0,1169	0,2191
	6	15	0,2762	0,4021

Appendix 8. Pairwise contingency tables between clusterings.

Percentage of genes for each cluster in Network 0 found in each cluster of Network 3. In bold, the most resembling values between clusters.

		Clusters in Network 3					
		1	2	3	4	5	6
Clusters in Network 0	1	**64,10**	7,69	7,69	2,56	7,69	10,26
	2	8,77	**68,42**	0,00	1,75	19,30	1,75
	3	14,89	0,00	**65,96**	19,15	0,00	0,00
	4	3,92	1,96	11,76	**82,35**	0,00	0,00
	5	**34,09**	6,82	4,55	11,36	**43,18**	0,00
	6	3,57	17,86	14,29	**32,14**	0,00	**32,14**
	7	11,11	**38,89**	0,00	11,11	**38,89**	0,00
	8	0,00	**48,72**	**33,33**	10,26	5,13	2,56
	9	5,56	11,11	16,67	**27,78**	**38,89**	0,00

Appendix 9. Comparison of GOBP between Network 0 and Network 3.

GO terms enriched in one of the clusters as well as all GO terms associated to one of the three target genes at least (even if not significantly enriched). In bold, the smallest FDR value for a given GOBP term when the difference between the FDR of the two clusters is higher than one order of magnitude. Genes tested by 3D DNA FISH are in red bold. * GO analysis by using a data base of redundant BP instead of non-redundant BP.

		Network 0 - Cluster 1		**Network 3 - Cluster 1**	
Items GOID	**GOBP Terms**	**Genes**	**FDR**	**Genes**	**FDR**
0043062	Extracellular structure	*POSTN, COL1A1, COL1A2, COL3A1, COL5A1, COL16A1, LAMA4, MFAP5*	5,76E-05	*POSTN, COL1A1, COL1A2,COL3A1, COL5A1, COL5A2, COL16A1,* ***DCN****, FAP, FBN1, ABI3 bp, ANXA2, LAMA4*	**1,14E-08**
0071417	Cellular response to organonitrogen compound	*COL1A1, COL1A2,COL3A1, COL5A2, COL16A1, FYN, KLF3, ZFP36L1, HSP90B1*	**6,80E-04**	*COL1A1, COL1A2,COL3A1, COL5A2, COL16A1, DNMT1, FBN1,* ***IGF2****, HSP90B1*	1,16E-02
0045995	Regulation of embryonic development	*COL5A1, COL5A2, FGFR1, LAMA4, LFNG*	2,24E-03	*COL5A1, COL5A2, FGFR1, LAMA4, LFNG*	1,16E-02
0071559	Reponse to transformaing growth factor beta	*POSTN, COL1A1, COL1A2, COL3A1, FYN, ZFP36L1*	**2,35E-03**	*POSTN, COL1A1, COL1A2,COL3A1, FBN1*	1,24E-01
0044236	Multicellular organism metabolic process	*COL1A1, COL1A2,COL3A1, COL5A1, COL5A2*	2,35E-03	*COL1A1, COL1A2,COL3A1, COL5A1, COL5A2, FAP*	3,05E-03
0043588	Skin development	*COL1A1, COL1A2, COL3A1, COL5A1, COL5A2, ZFP36L1*	**3,18E-03**	*COL1A1, COL1A2, COL3A1, COL5A1, COL5A2*	1,44E-01
0001101	Reponse to acid chemical	*COL1A1, COL1A2,COL3A1, COL5A2, COL16A1, NFATC4*	1,17E-02	*COL1A1, COL1A2,COL3A1, COL5A2, COL16A1, DNMT1, NFATC4*	2,27E-02
0001501	Skeletal system development	*POSTN, COL1A1, COL1A2, COL3A1, COL5A2, FGFR1, TMEM119*	1,43E-02	*POSTN, COL1A1, COL1A2, COL3A1, COL5A2, FBN1, FGFR1, ANXA2, TMEM119,* ***IGF2***	3,05E-03
0001764	Neuron migration	*COL3A1, FGFR1, PLXNB2, FYN*	**2,82E-02**	*COL3A1, FGFR1, FLRT2*	4,37E-01
0071774	Reponse to fibroblast growth fator	*POSTN, COL1A1, FGFR1, ZFP36L1*	3,00E-02	*POSTN, COL1A1, FGFR1, FLRT2*	1,44E-01
0010975	Regulation of neuron projection development	*FGFR1, PLXNB2, FYN, NFATC4, PTPRF, CUL7*	**3,07E-02**	*TBR1, FGFR1, NFATC4, PTPRF, CUL7*	4,93E-01
0007498	Mesoderm development			*EXT2, FGFR1, MESDC2,* ***MEST***	1,24E-01
0010171	Body morphogenesis			*COL1A1,* ***MYH3***	4,37E-01
0060324	Face development			*COL1A1,* ***MYH3***	4,37E-01
0007517	Muscle organ development			*COL3A1, DCN, FGFR1,* ***MYH3***	8,35E-01

0007219	Notch signaling pathway			*POSTN, LFNG, **DLK1***	5,56E-01
0030199*	Collagen fibril organization	*COL1A1, COL1A2, COL3A1, COL5A1, COL5A2*	1,10E-04	*COL1A1, COL1A2, COL3A1, COL5A1, COL5A2, ANXA2*	1,02E-05
		Network 0 - Cluster 2		**Network 3 - Cluster 2**	
Items	**GOBP Terms**	**Genes**	**FDR**	**Genes**	**FDR**
0072350	Tricarboxylic acid metabolic process	*CS, DLAT, DLD, NNT, MDH1, PDHA1*	3,02E-06	*CS, DLAT, DLD, NNT, MDH1, PDHA1*	2,11E-05
0051186	Cofactor metabolic process	*COQ7, DLAT, DLD, NNT, HK1, ACACB, NMNAT3, ACAT1, MDH1, PDHA1, PDHX*	**2,97E-05**	*DLAT, DLD, IBA57, NNT, GPI, ACACB, NMNAT3, MDH1, PDHA1, FLAD1, MCEE*	1,34E-03
0072524	Pyridine-containig compound metabolic process	*DLD, NNT, HK1, NMNAT3, MDH1, PDHA1, PDHX*	**1,00E-04**	*DLD, NNT, GPI, NMNAT3, MDH1, PDHA1*	1,11E-02
0006631	Fatty acid metabolic process	*CPT1B, ECI1, DLAT, DLD, ACACB, ACADS, ACAT1, PDHA1, PTGES2, PDHX*	**1,00E-04**	*CPT1B, ECI1, DLAT, DLD, FABP3, ACACB, ACADS, PDHA1, ADIPOR2, PTGES2, MCEE*	1,17E-03
0006091	Generation of precursor metabolites and energy	*CS, DLAT, DLD, NNT, HK1, MDH1, OXA1L, ATP5B, PDHA1, SLC25A3*	1,09E-04	*CS, DLAT, DLD, NNT, GPI, MDH1, NDUFA3, NDUFB5, NDUFS1, OXA1L, ATP5B, PDHA1, SLC25A3, CISD1, NDUFA12, PYGM*	**1,32E-07**
0006090	Pyruvate metabolic process	*DLAT, DLD, HK1, PDHA1, PDHX*	5,42E-03	*DLAT, DLD, GPI, PDHA1, BSG*	2,32E-02
0006790	Sulfur compound metabolic process	*VCAN, **DCN**, DLAT, DLD, ACACB, ACAT1, PDHA1, PDHX*	**7,47E-03**	*DLAT, DLD, IBA57, ACACB, PDHA1, MCEE*	4,79E-01
0042180	Cellular ketone metabolic process	*COQ7, DLAT, DLD, ACACB, PDHA1, PDHX*	1,46E-02	*DLAT, DLD, FABP3, GPI, ACACB, PDHA1*	8,05E-02
0045454	Cell redox homeostasis	*TXNRD2, DLD, NNT, PTGES2*	1,46E-02	*TXNRD2, DLD, NNT, PTGES2*	4,91E-02
0044282	Small molecule catabolic process	*CPT1B, ECI1, DLD, HK1, ACACB, ACADS, ACAT1*	1,88E-02	*CPT1B, ECI1, DLD, GPI, ACACB, ACADS, BCAT2, MCEE*	4,51E-02
0098656	Anion transmembrane transport	*CLCN5, CPT1B, ACACB, SLC25A3, SLC1A3, VDAC1*	**2,31E-02**	*CPT1B, ACACB, SLC25A3, SLC1A3, VDAC1*	3,77E-01
0006081	Cellular aldehyde metabolic process	*DLAT, DLD, PDHA1, PDHX*	2,59E-02	*DLAT, DLD, GPI, PDHA1*	8,73E-02
0043648	Dicarboxylic acid metabolic process	*DLD, NMNAT3, MDH1, SLC1A3*	3,13E-02	*DLD, NMNAT3, MDH1, BCAT2, SLC1A3*	2,13E-02
0016042	Lipid catabolic process	*CPT1B, ECI1, ACACB, ACADS, ACAT1, NCEH1*	3,65E-02	*CPT1B, ECI1, FABP3, ACACB, ACADS, NCEH1, MCEE*	6,59E-02
0051235	Maintenance of location	*HK1, ACACB, **MEST**, ATP2A1*	5,62E-01		
0010876	Lipid localization	*CPT1B, ACACB, **MEST**, APOO*	7,54E-01		
0010257	NADH dehydrogenase complex assembly			*NDUFA3, NDUFB5, NDUFS1, OXA1L, NDUFA12*	3,29E-03
0097031	Mitochondrial respiratory chain complex I biogenesis			*NDUFA3, NDUFB5, NDUFS1, OXA1L, NDUFA12*	3,29E-03
0033108	Mitochondrial respiratory chain complex assembly			*NDUFA3, NDUFB5, NDUFS1, OXA1L, NDUFA12*	1,41E-02
0009141	Nucleoside triphosphate metabolic process	*DLD, HK1, ATP5B, ATP5O*	5,55E-01	*DLD, GNAI3, GPI, NDUFA3, NDUFB5, NDUFS1, ATP5B, NDUFA12*	**1,97E-02**
0097194	Execution phase of apoptosis	*ENDOG, PRKCQ*	6,59E-01	*CAPN10, ENDOG, HMGB2, PRKCQ*	**4,91E-02**
0055114*	Oxidation-reduction process	*COQ7, TXNRD2, CPT1B, CS, ECI1, DLAT, DLD, NNT, GPX3, HK1, ACACB, ACADS, MDH1, OXA1L, PDHA1, PTGES2, RTN4IP1*	2,23E-06	*TXNRD2, CPT1B, CS, ECI1, DLAT, DLD, FABP3, FDFT1, NNT, GPI, GPX3, ACACB, ACADS, MDH1, NDUFA3, NDUFB5, NDUFS1, OXA1L, PDHA1, CISD1, NDUFA12, PYGM, ADIPOR2, PTGES2, FLAD1, RTN4IP1*	**5,63E-09**
0006101*	Citrate metabolic process	*CS, DLAT, DLD, NNT, MDH1, PDHA1*	2,23E-06	*CS, DLAT, DLD, NNT, MDH1, PDHA1*	1,93E-05
0019752*	Carboxylic acid metabolic process	*CPT1B, CS, VCAN, ECI1, **DCN**, DLAT, DLD, NNT, HK1, ACACB, NMNAT3, ACADS,*	**2,23E-06**	*CPT1B, CS, ECI1, DLAT, DLD, FABP3, FARSA, NNT, GPI, ACACB, NMNAT3, ACADS,*	3,44E-05

		ACAT1, MDH1, PDHA1, SLC1A3, PTGES2, PDHX		*MDH1, PDHA1, BCAT2, SLC1A3, BSG, ADIPOR2, PTGES2, MCEE*	
0045333*	Cellular respiration	*CS, DLAT, DLD, NNT, MDH1, OXA1L, PDHA1*	5,33E-04	*CS, DLAT, DLD, NNT, MDH1, NDUFA3, NDUFB5, NDUFS1, OXA1L, PDHA1, CISD1, NDUFA12*	**2,65E-07**
0015980*	Energy derivation by oxidation of organic compounds			*CS, DLAT, DLD, NNT, MDH1, NDUFA3, NDUFB5, NDUFS1, OXA1L, PDHA1, CISD1, NDUFA12, PYGM*	1,88E-06

Network 0 - Cluster 3

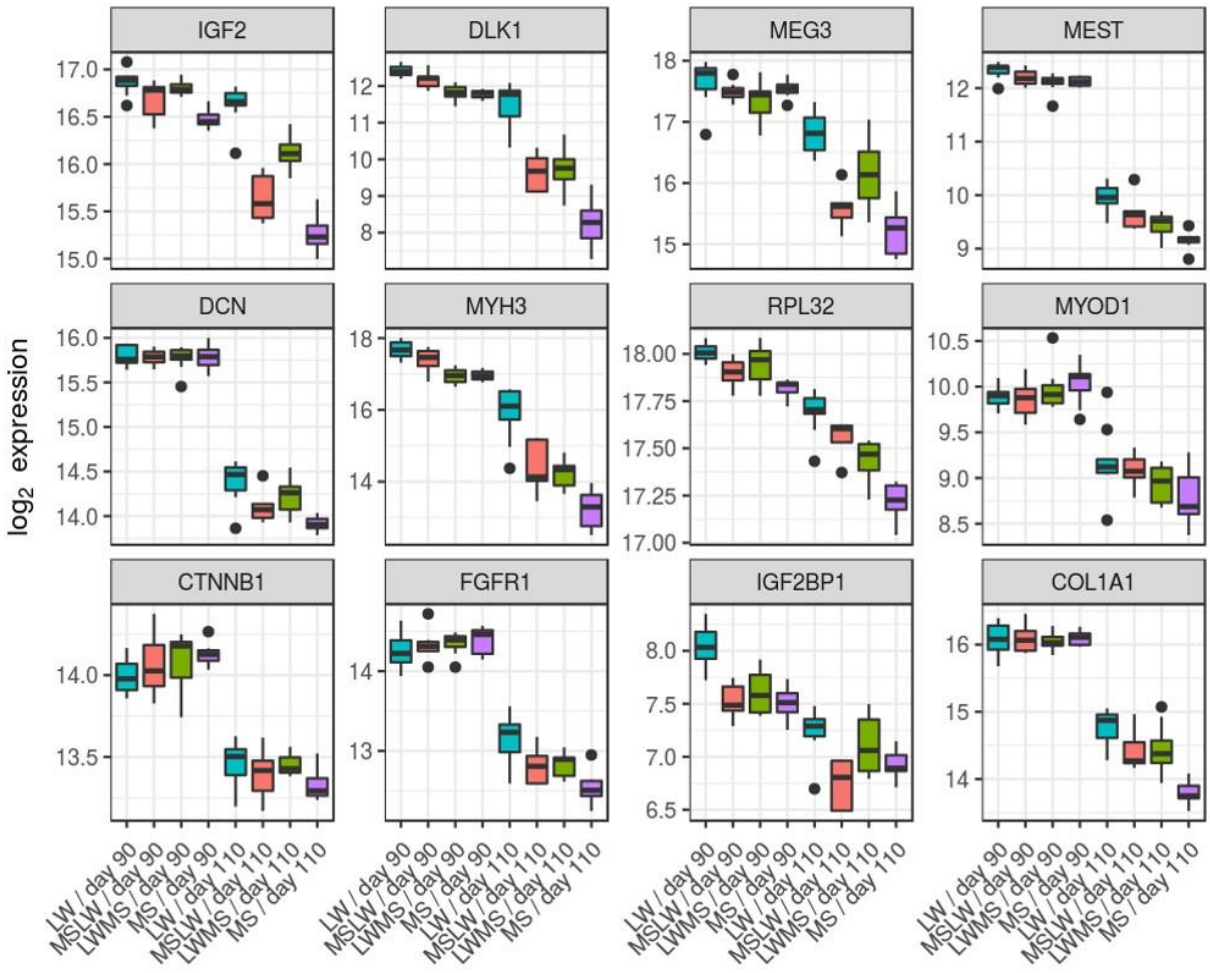
Items	GOBP Terms	Genes	FDR
0071514	Genetic imprinting	*CDKN1C, **IGF2**, KCNQ1*	3,82E-02
0050879	Multicellular organismal movement	*MB, **MYH3***	8,56E-01
0044270	Cellular nitrogen compound catabolic process	*PM20D1, PDE4A, RPL31, **RPL32***	1,00E+00
0006941*	Striated muscle contraction	*KCNQ1, MB, **MYH3**, RGS2*	5,47E-01

Network 0 - Cluster 5

Items	GOBP Terms	Genes	FDR
0031115*	Negative regulation of microtubule polymerization	*MAPRE1, INPP5J, STMN1*	1,06E-02
0046785*	Microtubule polymerization	*MAPRE1, INPP5J, STMN1, TPPP3*	3,41E-02

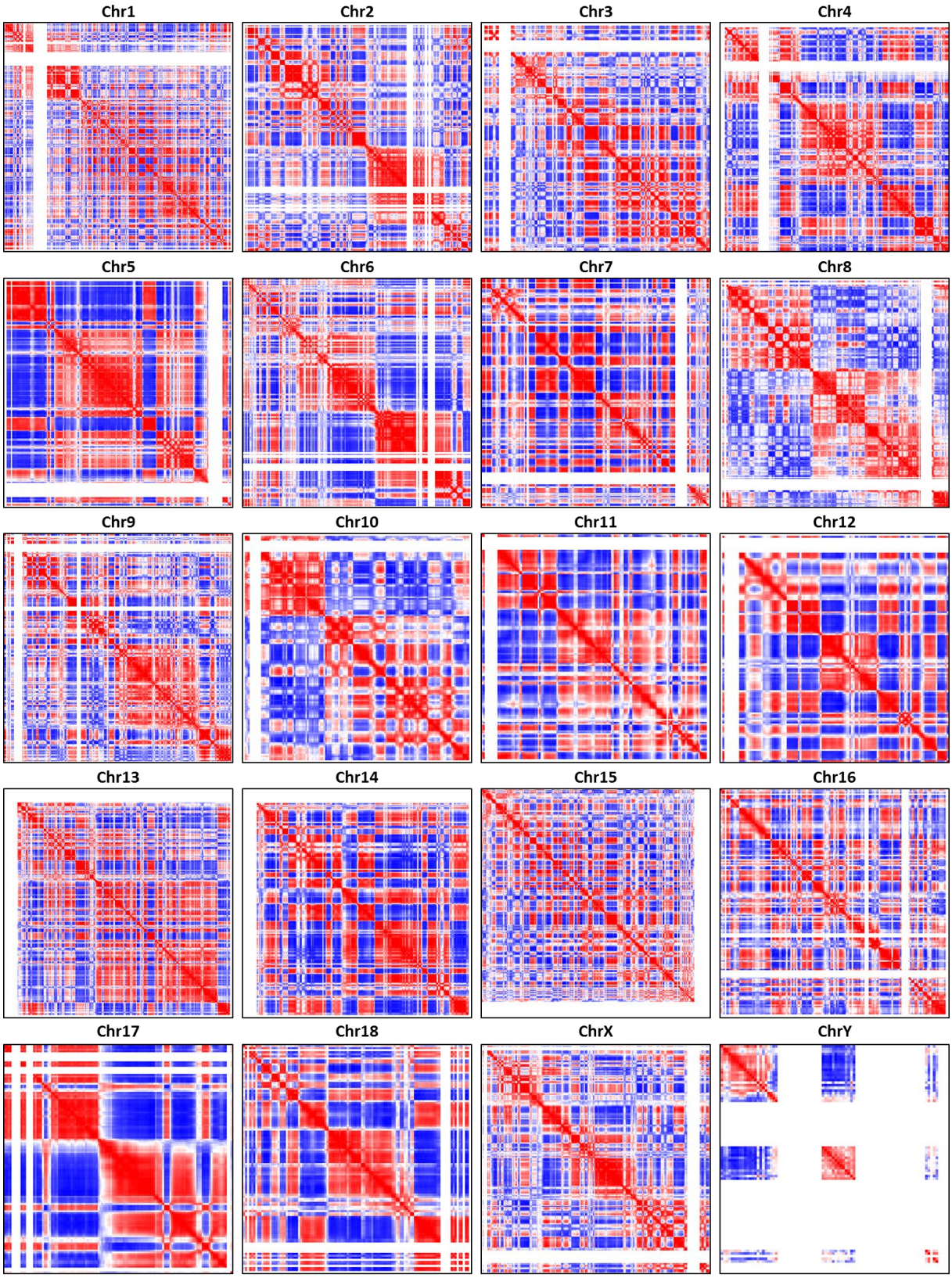
		Network 0 - Cluster 8		**Network 3 - Cluster 2**	
Items	**GOBP Terms**	**Genes**	**FDR**	**Genes**	**FDR**
0006091	Generation of precursor metabolites and energy	*ME3, IDH3G, NDUFA3, NDUFAB5, CISD1, NDUFA12, PYGM*	1,64E-02	*CS, DLAT, DLD, NNT, GPI, MDH1, NDUFA3, NDUFB5, NDUFS1, OXA1L, ATP5B, PDHA1, SLC25A3, CISD1, NDUFA12, PYGM*	**1,32E-07**
0055114*	Oxidation-reduction process	*ME3, ADH5, CRAT, IDH3G, NDUFA3, NDUFAB5, CYB5R1, CISD1, NDUFA12, PYGM, BLVRV, FLAD1*	7,25E-03	*TXNRD2, CPT1B, CS, ECI1, DLAT, DLD, FABP3, FDFT1, NNT, GPI, GPX3, ACACB, ACADS, MDH1, NDUFA3, NDUFB5, NDUFS1, OXA1L, PDHA1, CISD1, NDUFA12, PYGM, ADIPOR2, PTGES2, FLAD1, RTN4IP1*	**5,63E-09**
0015980*	Energy derivation by oxidation of organic compounds	*ME3, IDH3G, NDUFA3, NDUFAB5, CISD1, NDUFA12, PYGM*	8,17E-03	*CS, DLAT, DLD, NNT, MDH1, NDUFA3, NDUFB5, NDUFS1, OXA1L, PDHA1, CISD1, NDUFA12, PYGM*	**1,88E-06**
0045333*	Cellular respiration	*ME3, IDH3G, NDUFA3, NDUFAB5, CISD1, NDUFA12*	8,17E-03	*CS, DLAT, DLD, NNT, MDH1, NDUFA3, NDUFB5, NDUFS1, OXA1L, PDHA1, CISD1, NDUFA12*	**2,65E-07**

Appendix 10. Gene expression profiles from the normalized expression data from the transcriptome study of Voillet et al., 2014.

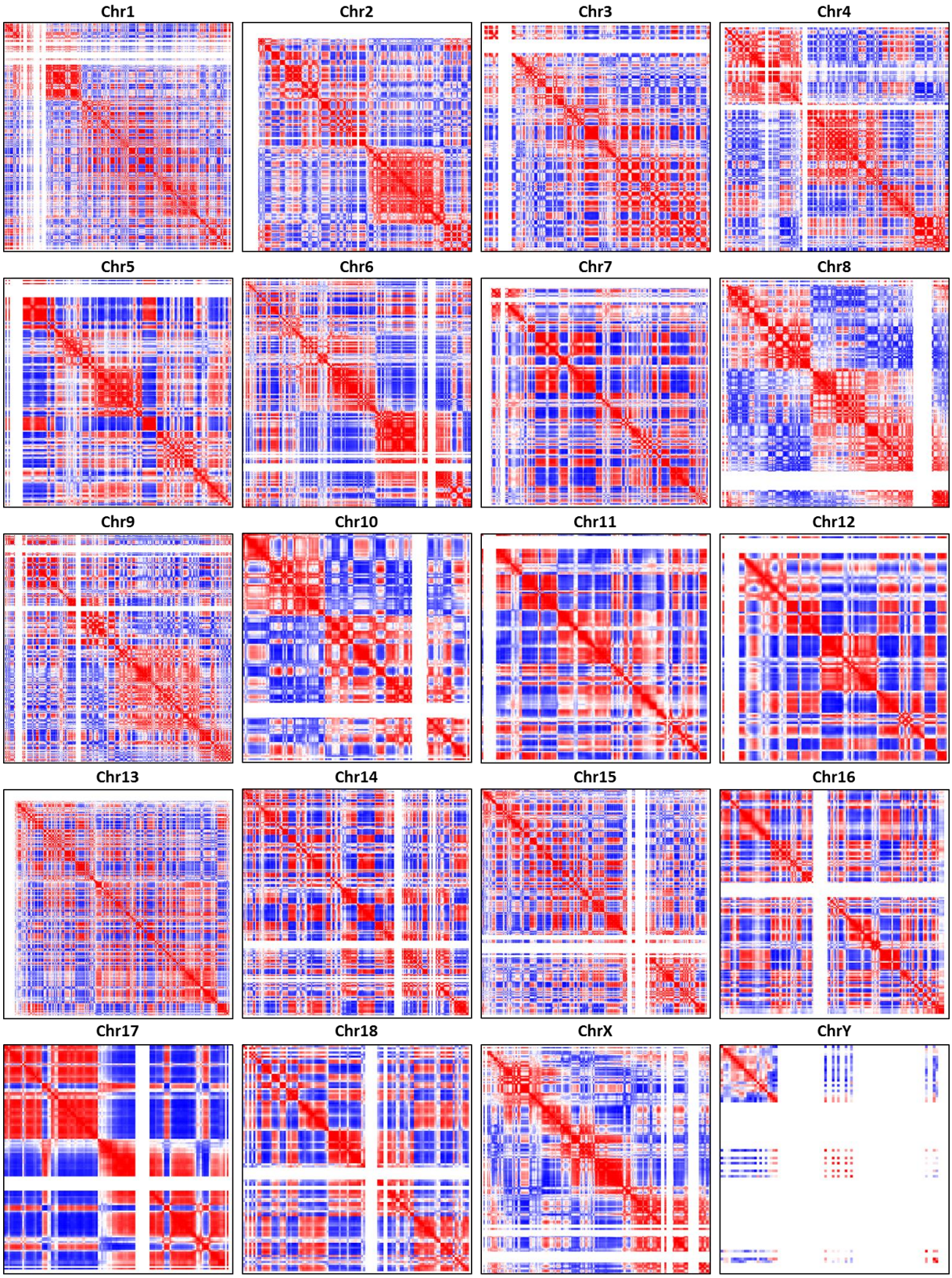


Appendix 11. Hi-C raw matrices of the 18 autosomic chromosomes and the 2 sex chromosomes obtained at 200 Kb resolution.

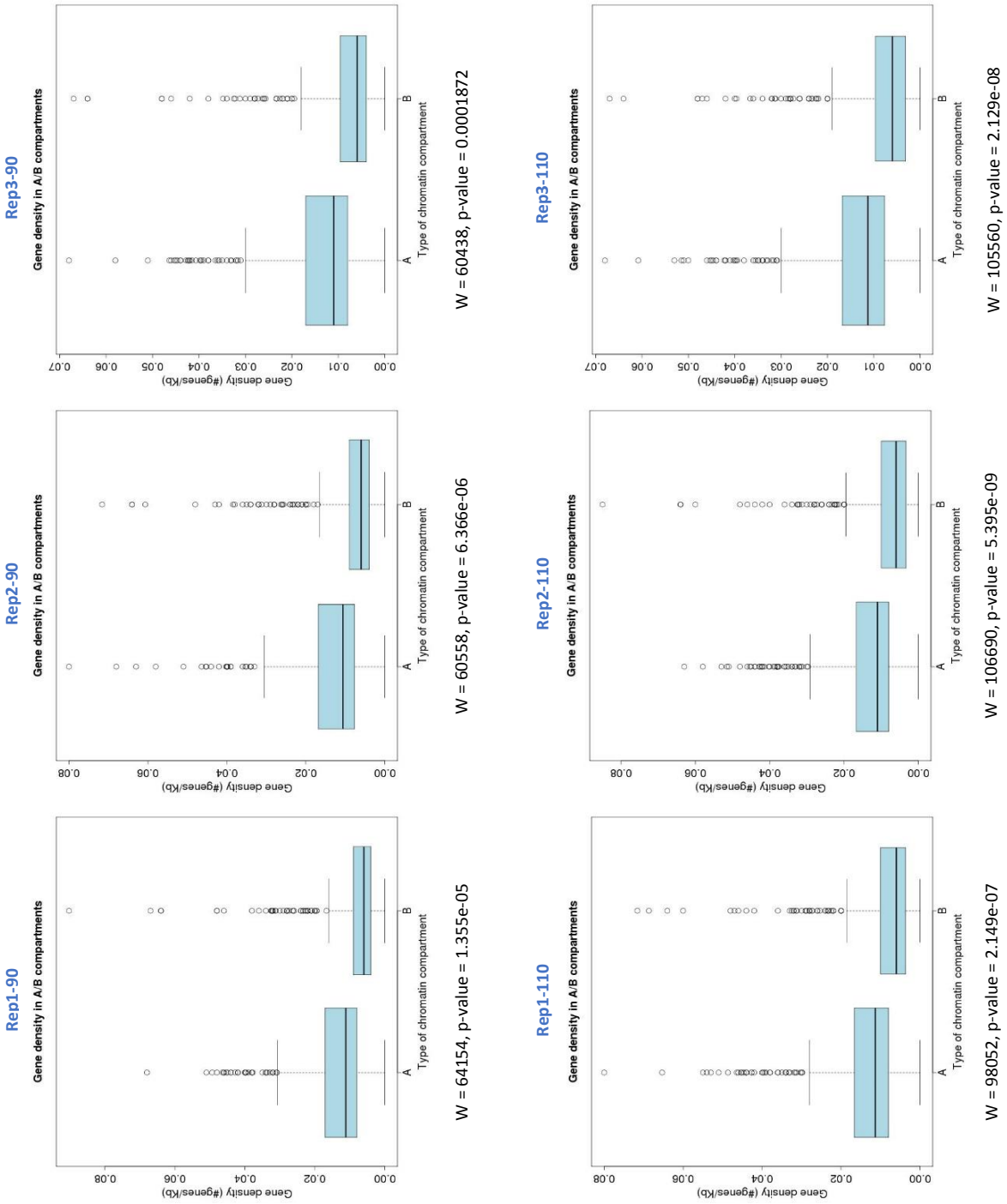
Apendix 12. Correlation matrices of the 18 autosomic chromosomes and the 2 sex chromosomes obtained from the merged-90 matrices.



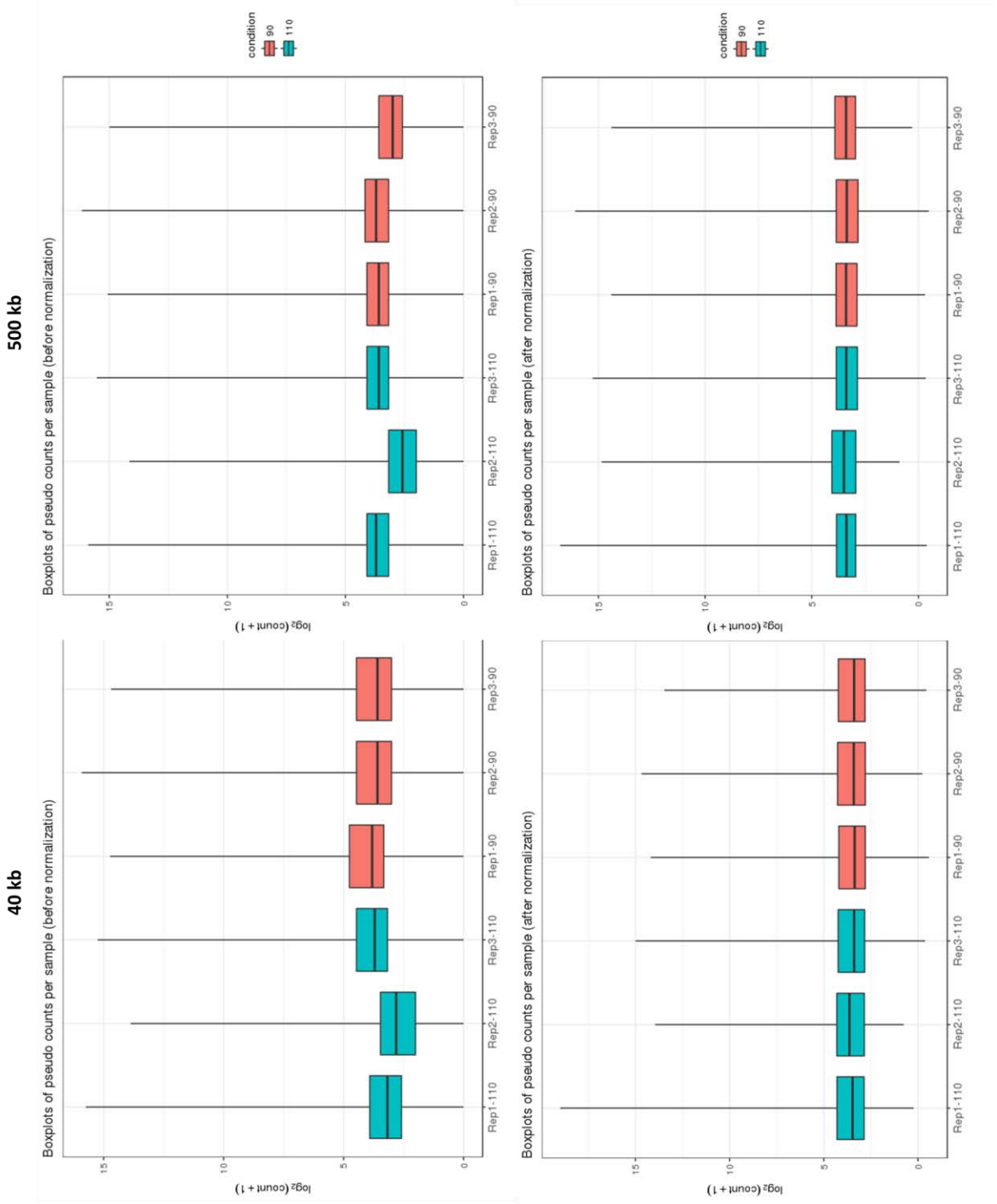
Appendix 13. Correlation matrices of the 18 autosomic chromosomes and the 2 sex chromosomes obtained from the merged-110 matrices.



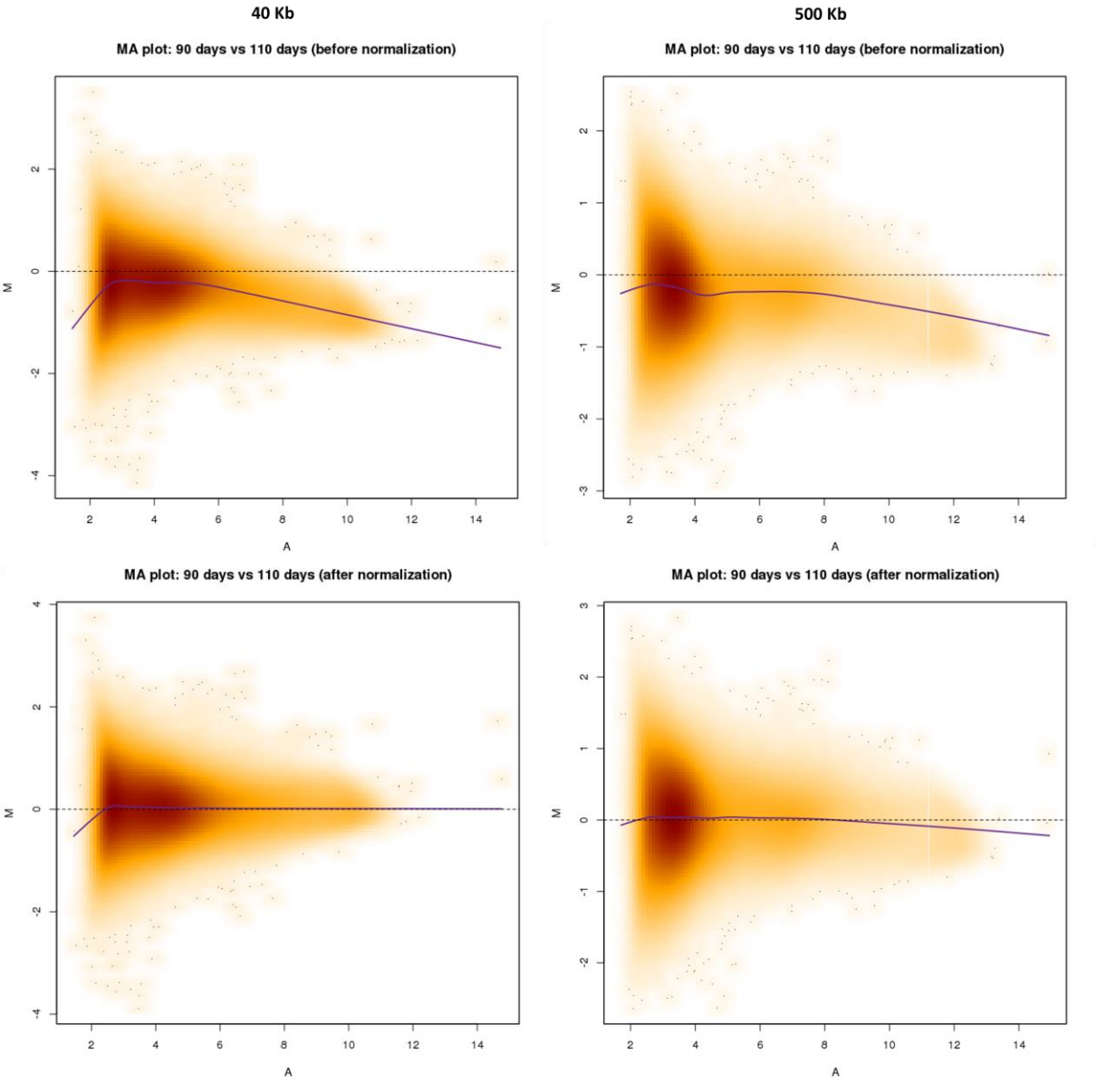
Appendix 14. Gene density in A and B compartments.



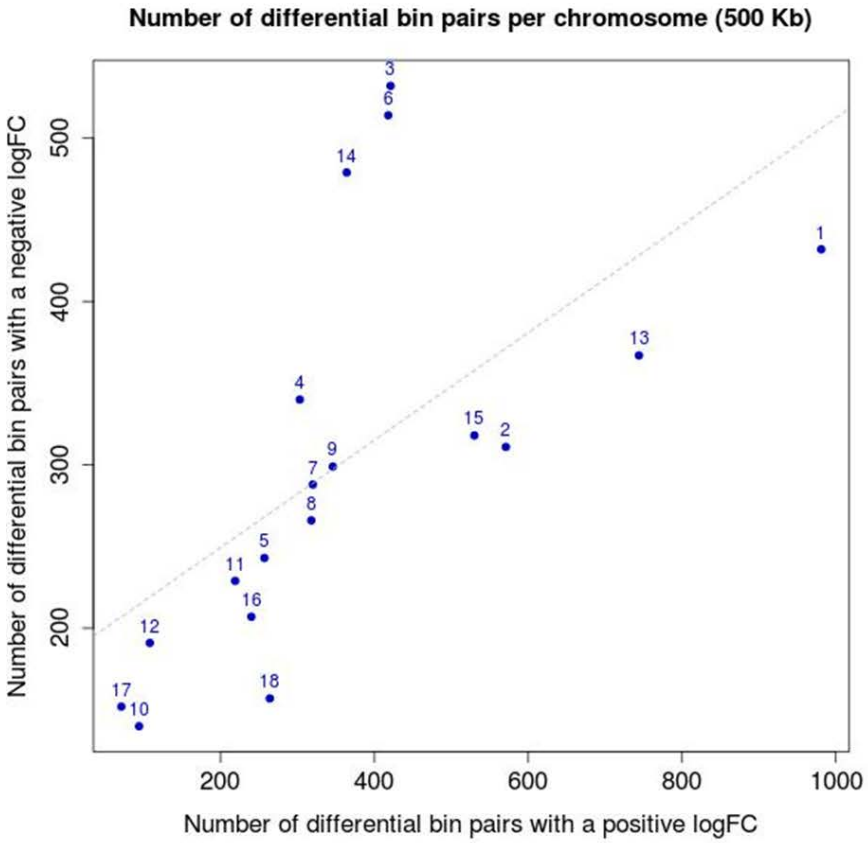
Appendix 15. Distribution of raw and normalized counts per sample.

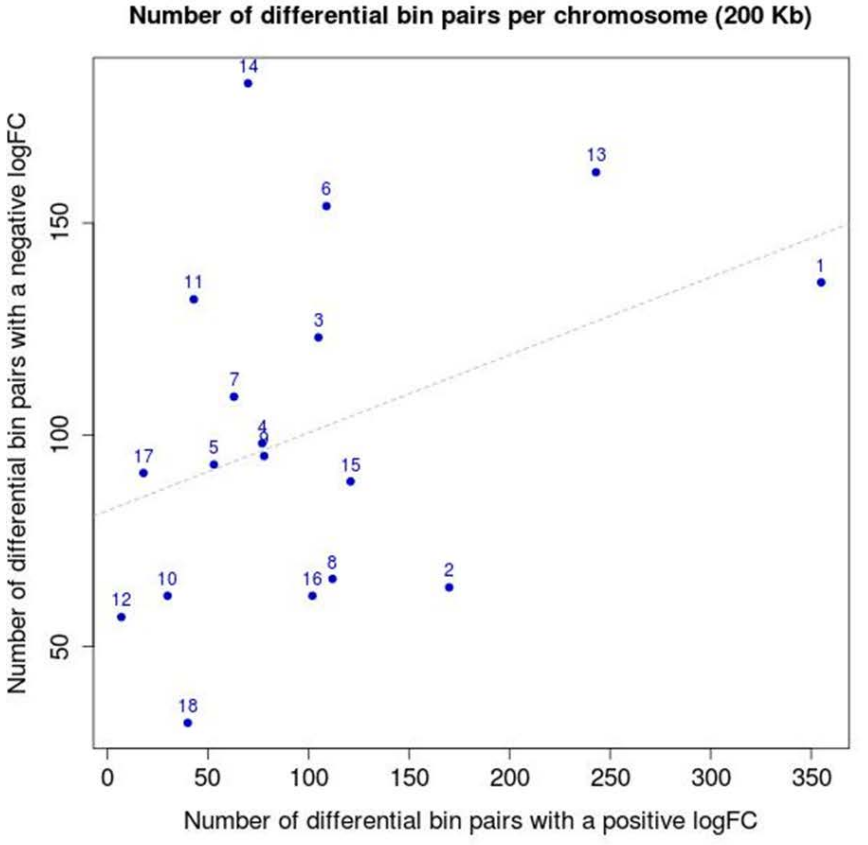


Appendix 16. Global MA plot between samples at 90 and 110 days before and after normalization.

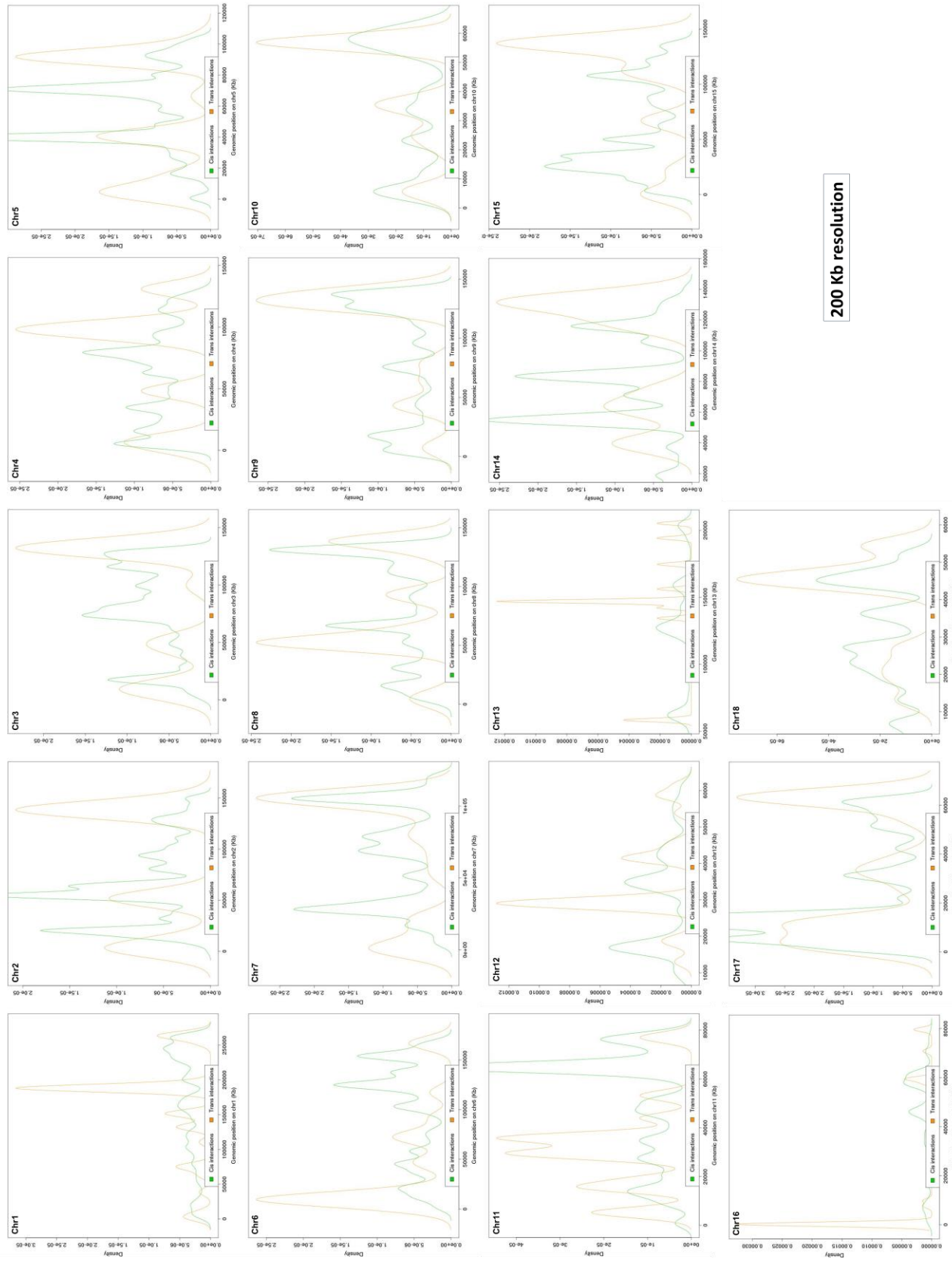


Appendix 17. Proportion of differential bin pairs with positive and negative logFC across chromosomes.

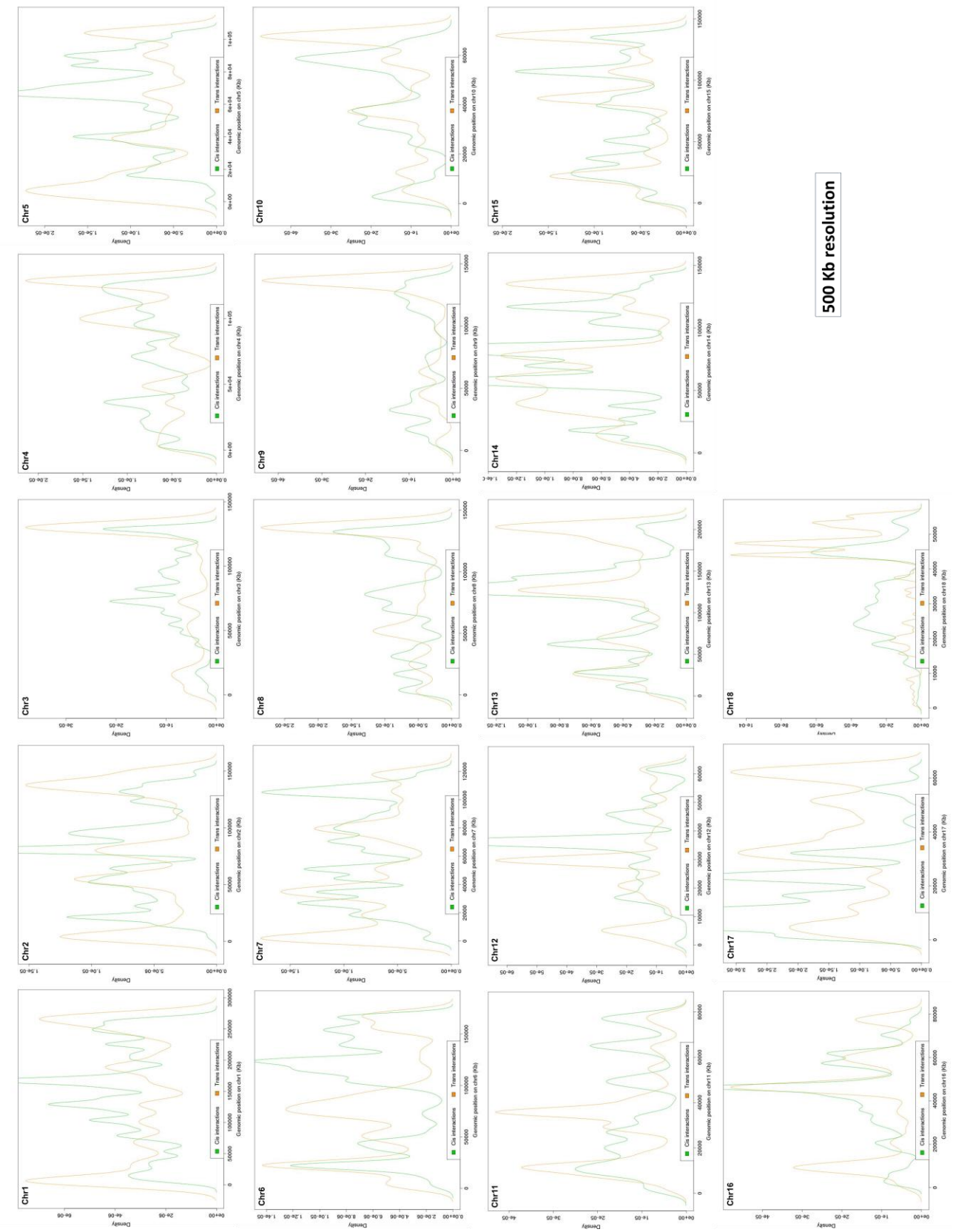




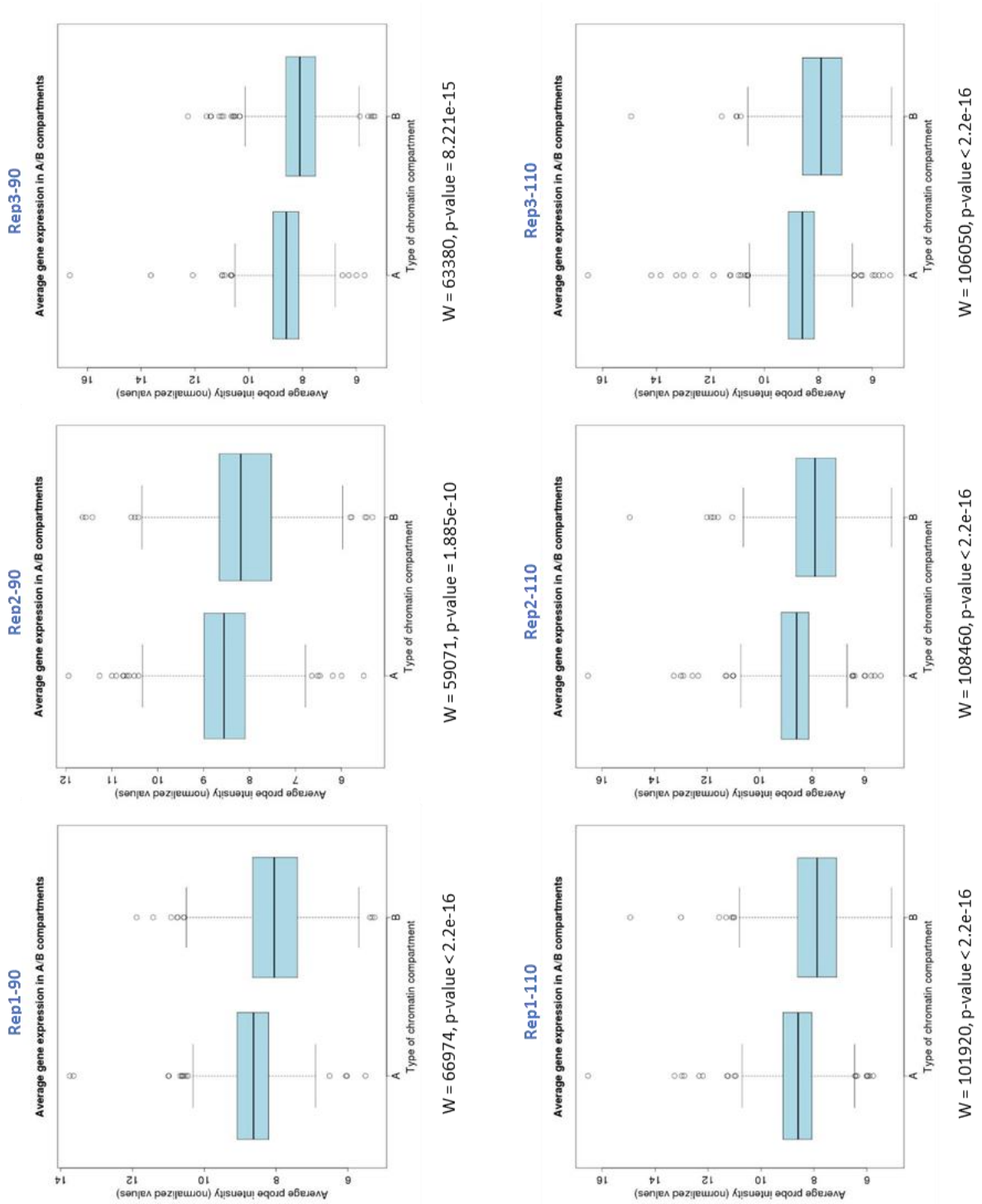
Appendix 18. Density plots of *trans* vs. *cis* connections along each chromosome at 200 Kb resolution.



Appendix 19. Density plots of *trans* vs. *cis* connections along each chromosome at 500 Kb resolution.



Appendix 20. Gene expression in A and B compartments.



Appendix 21. Published article in Scientific reports.

www.nature.com/scientificreports

OPEN

A new approach of gene co-expression network inference reveals significant biological processes involved in porcine muscle development in late gestation

Received: 11 March 2018
Accepted: 14 June 2018
Published online: 05 July 2018

M. Marti-Marimon[1], N. Vialaneix[2], V. Voillet[1], M. Yerle-Bouissou[1], Y. Lahbib-Mansais[1] & L. Liaubet[1]

The integration of genetic information in the cellular and nuclear environments is crucial for deciphering the way in which the genome functions under different physiological conditions. Experimental techniques of 3D nuclear mapping, a high-flow approach such as transcriptomic data analyses, and statistical methods for the development of co-expressed gene networks, can be combined to develop an integrated approach for depicting the regulation of gene expression. Our work focused more specifically on the mechanisms involved in the transcriptional regulation of genes expressed in muscle during late foetal development in pig. The data generated by a transcriptomic analysis carried out on muscle of foetuses from two extreme genetic lines for birth mortality are used to construct networks of differentially expressed and co-regulated genes. We developed an innovative co-expression networking approach coupling, by means of an iterative process, a new statistical method for graph inference with data of gene spatial co-localization (3D DNA FISH) to construct a robust network grouping co-expressed genes. This enabled us to highlight relevant biological processes related to foetal muscle maturity and to discover unexpected gene associations between *IGF2*, *MYH3* and *DLK1/MEG3* in the nuclear space, genes that are up-regulated at this stage of muscle development.

Cell type diversity in a given organism cannot be explained only by DNA sequences. *Cis*- and *trans*-acting regulatory sequences are not the only determinants of gene expression: other epigenetic mechanisms are also responsible for tissue-specific expression of genes. Indeed, more recently, numerous studies link the genome organization in the nucleus to an additional level of gene expression regulation[1–5]. It is known that in higher eukaryotes, genomes are organized into individual chromosomes that occupy discrete territories in the nucleus[6], which means that the distribution of the genome is not random. Moreover, interphase chromosome regions often loop out of their chromosome territories[7], and neighbouring chromosomes can intermingle, resulting in potential functional contacts between regions located on different chromosomes[2,4,8]. There is evidence that long-range interactions between genomic regions contribute to gene expression regulation[2] and might facilitate the consolidation of co-regulated genes in specialized foci of active RNA polymerase II as well as at nuclear speckles (pre-mRNA processing)[3–5]. These insights give us some clues about the contribution of the spatial genome organization in interphase nuclei to gene expression regulation (for review[9]).

Microscopy approaches such as 3D fluorescent *in situ* hybridization (FISH)[10,11], enable a global view of what is happening at the level of individual cells. Recently, we focused on this last item to study interchromosomal interactions between co-expressed genes belonging to the Imprinted Gene Network (IGN)[12]. We chose the genomic

[1]GenPhySE, Université de Toulouse, INRA, ENVT, Castanet Tolosan, France. [2]MIAT, Université de Toulouse, INRA, Castanet Tolosan, France. Y. Lahbib-Mansais and L. Liaubet jointly supervised this work. Correspondence and requests for materials should be addressed to L.L. (email: laurence.liaubet@inra.fr)

imprinting model because it can compare, in the same nucleus, the environment of an active allele with an allele maintained as repressed due to its imprinted status. We focused our analysis on *IGF2* because it is involved in pig muscle growth and fat deposition[13,14], being therefore a major gene of interest in the context of agronomic projects. In humans, IGF2 is well known to be a key element in foetal growth and development[15]. We highlighted associations between the expressed alleles of *IGF2* and *DLK1/MEG3* locus (*DLK1* being related to the control of muscle development and regeneration[16]), in foetal muscle and liver cells[17]. These results illustrate the implication in *trans*-interactions of genes associated with quantitative trait loci (QTLs) for growth traits, providing new evidence that genome organization could influence gene expression and phenotypic outcome in livestock species.

In this context, we focused on the study of the muscle maturity process (essential for the survival of piglets) to better understand how interesting phenotypes are elaborated, by combining transcriptome and co-localization data with network modelling. Indeed in pigs, and in general in mammals, one of the most critical period for survival is the perinatal period, and an important determinant of early mortality is maturity, defined as the stage of full development leading to survival at birth[18]. Piglet maturity involves biological processes occurring between the 90th day and the end of gestation, e.g. glycogen accumulation in muscle and liver, as well as maturation of tissues[19,20]. The maturity of skeletal muscles plays an important role in piglet survival at birth because of its involvement in motor functions and thermoregulation. On this subject, we previously performed a microarray analysis of foetal muscle to identify candidate genes for piglet maturity, which revealed genes that were differentially expressed between the 90th and the 110th day of gestation[21]. Using Pearson correlation a relevance gene co-expression network was built from these differentially expressed genes (DEGs) for four gestational ages. The network revealed and confirmed that: (i) genes involved in muscle development were up-regulated at the 90th day of gestation, (ii) at the 110th day, the enriched biological functions were involved in energy metabolism.

An increasing number of studies use gene co-expression networks to deal with large gene expression datasets in order to decipher biological processes[22–24]. Modelling co-expression with network models is useful for providing a global overview of the co-expression relationships between genes and enables a set of genes to be analysed globally with specific network tools. This approach has been found relevant for extracting biological information such as important genes with respect to their centrality in the network structure[25], densely connected groups of genes[26] or frequent motifs[27].

For the study described in this article, we developed a new method for the construction of a co-expression gene network with genes involved in the foetal muscle maturation process, using an original approach coupling a statistical model and observed data in an iterative process to further our understanding of the mechanisms involved in muscle development. More precisely, we combined gene expression data and gene spatial co-location, thus creating a new statistical method for graph inference. Our approach is based on Gaussian Graphical Models (GGMs[28]) that enable the computation of *partial correlations* and fit direct relations better than Pearson-based correlation networks. Such networks have been found to be more efficient for grouping genes with a common function[29]. This enabled us to obtain more reliable networks in which connections between genes were validated iteratively using biological evidence. In practical terms, we performed 3D DNA FISH experiments to test pairwise whether co-expressed genes (connected in the network) were co-localized in the 3D nuclear space.

The study enabled us to obtain a robust gene co-expression network that highlights significant Gene Ontology (GO) terms associated with biological processes related to foetal muscle maturity. In addition, unexpected associations were identified between *MYH3* and the imprinted loci *IGF2* and *DLK1*, which might help elucidate the mechanisms involved in the porcine muscle development process at the end of gestation.

Results

Data selection. The 44,368 probes from the expression dataset of the muscle transcriptome study from Voillet *et al.*[21] were found to correspond to 13,855 unique annotated genes, among which 1,131 unique genes were found to be differentially expressed between the two gestational ages and for the four genotypes characterizing the establishment of piglet maturity. Among them, 359 DEGs (Supplementary Table S1) were selected for being highly correlated with *IGF2*, *DLK1* and *MEG3* ($R^2 \geq 0.84$), also identified as DEG, and were used in all subsequent network inferences (see further details in "Materials and Methods", the section on "Microarray data description and pre-processing").

Network inference iteration and 3D FISH validations. The whole process involving the data selection, the network inference and the 3D FISH validations is summarized in Fig. 1. Network 0 was inferred with no *a priori* knowledge and contained 2,279 edges for 359 nodes (density: 3.55%). A sub-network extracted around the three target genes is shown in Fig. 2a.

Network 1 was built based on the triple co-localization of *IGF2, DLK1 and MEG3* found in our previous study[17]. This *a priori* information was used to reinforce the existence of an edge between the pairs *IGF2-DLK1*, *IGF2-MEG3* and *DLK1-MEG3* in Network 1 (sub-network in Fig. 2b), which contained 2,250 edges (density: 3.50%). In both graphs (Network 0 without *a priori* and Network 1 with *a priori*), we found a direct connection between the genes *IGF2* and *RPL32*. The *IGF2-RPL32* association was thus tested by 3D DNA FISH, because it involved one of our 3 initial target genes (*IGF2, DLK1* and *MEG3*), and because it was also found in the IGN of Varrault *et al.*[12]. The 3D DNA FISH assay revealed that *IGF2* and *RPL32* were associated in 20% of the analysed nuclei (Table 1, Fig. 3a).

Additionally, we used 3D DNA FISH to analyse *MEST* and *DCN* associations with each of the three target genes, because they were also connected in the IGN (Table 1 and Fig. 3b–e).

This new information about spatial co-localization in the nucleus was entered in our model as an *a priori* to build Network 2 (with 2,091 edges and 3.25% of density) (sub-network in Fig. 2c). Specifically, in addition to the three pairs *IGF2-DLK1*, *IGF2-MEG3* and *DLK1-MEG3* given as associated in Network 1, we gave the following pairs of genes as known to be co-localized: *IGF2-MEST* (34% of analysed nuclei presenting an

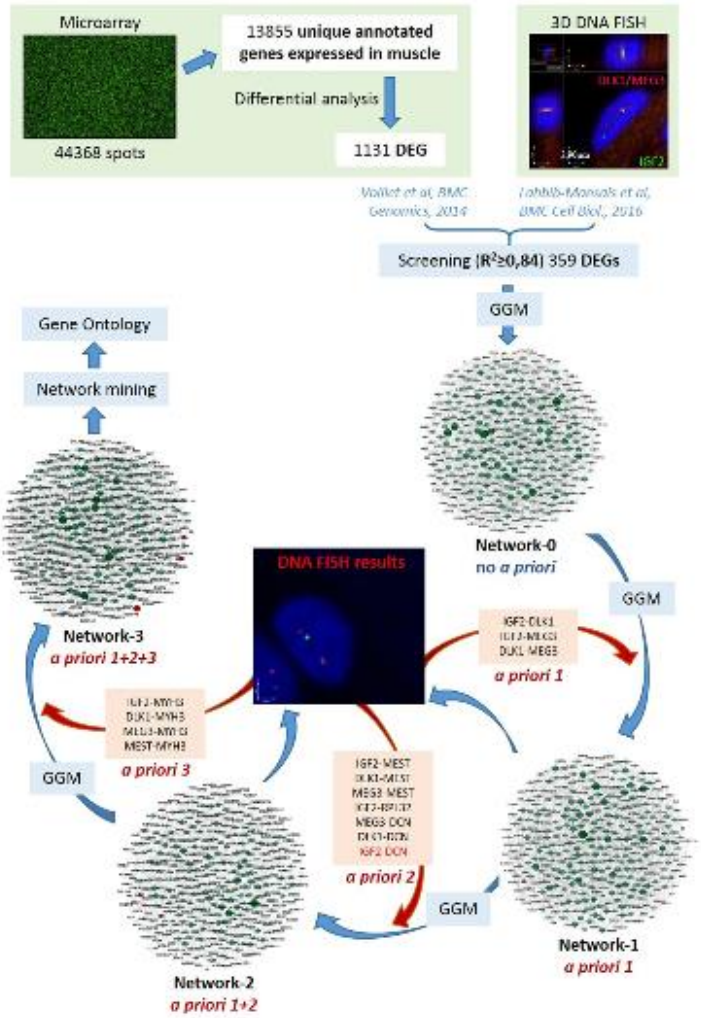


Figure 1. Experimental design. Published data are represented in green squares (microarray data and 3D DNA FISH data), statistical methods are represented in blue (GGM: Gaussian Graphical Models) and new information about spatial localization used for network inference is represented in red.

association), (*DLK1/MEG3*)-*MEST* (in 34% of analysed nuclei), (*DLK1/MEG3*)-*DCN* (in 15% of analysed nuclei) and *RPL32*-*IGF2* (in 20% of analysed nuclei). The pair *IGF2*-*DCN* was given as not co-localized (with 10% of nuclei presenting an association) (Table 1, Fig. 3b–e). *DLK1* and *MEG3* are two imprinted genes located in the same cluster, and are both present in the same Bacterial Artificial Chromosome (BAC) used for the 3D DNA FISH experiments, because of their proximity on the genomic sequence (Supplementary Table S2). Consequently, we considered *DLK1/MEG3* as a simple locus for all 3D DNA FISH analyses, even though they are considered to be single genes for network inference.

To obtain the last network (Network 3), we used 3D DNA FISH to test for associations involving *MYH3* because it was found to be connected to *DLK1* and *MEG3* in Network 0 and to *DLK1* in Network 1. We found *MYH3* associated with (i) *IGF2* in 52% of the analysed nuclei, (ii) *DLK1/MEG3* in 45% of the analysed nuclei, and (iii) *MEST* in 26% of the analysed nuclei (Table 1, Fig. 3f–h). Thus, in addition to the *a priori* information given in Networks 1 and 2, we gave the following new associations (*IGF2*-*MYH3*, *DLK1*-*MYH3*, *MEG3*-*MYH3* and *MEST*-*MYH3*) to infer Network 3 (2,091 edges, density = 3.25%) (Sub-network in Fig. 2d).

Network mining (network structure with key genes). For each network, two main numerical characteristics (degree and betweenness) were used to detect key genes with respect to the network structure. The degree of a node (in this case, of a gene) is the number of edges afferent to this gene. The betweenness of the node (gene) is the number of shortest paths between pairs of genes in the network that pass through that gene. High-degree

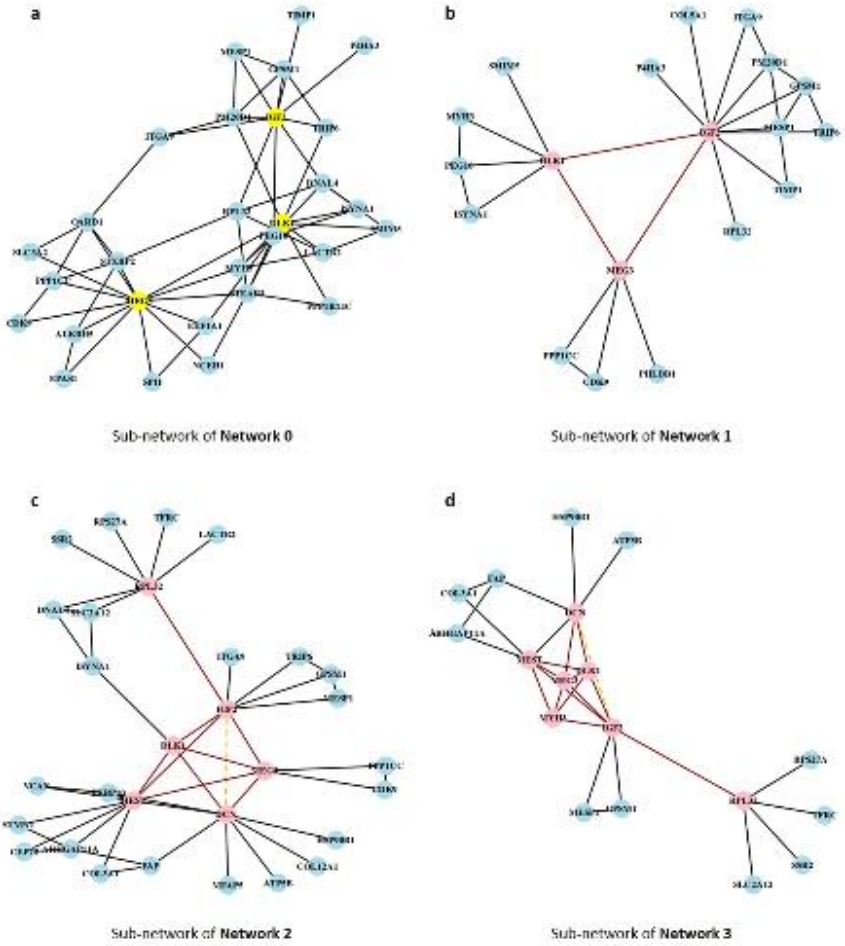


Figure 2. Analysis of gene associations. Pink nodes represent target genes, red edges represent the known associations observed by 3D DNA FISH and the dotted orange edge represents the observed as not associated after 3D FISH validations. Because networks are very dense and contain many genes, a sub-network restricted to the target genes and their direct neighbors is extracted from each network, and presented in this figure. (**a**) Network 0 is inferred without *a priori* information, and restricted to the nodes corresponding to *IGF2*, *DLK1* and *MEG3* (in yellow). To infer Networks 1, 2 and 3, new *a priori* information of spatial localization is introduced for the following pairs of genes: (**b**) *IGF2-DLK1*, *IGF2-MEG3* and *DLK1-MEG3* for Network 1; (**c**) *IGF2-MEST*, (*DLK1/MEG3*)-*MEST*, (*DLK1/MEG3*)-*DCN*, *RPL32-IGF2*, *IGF2-DCN* for Network 2; (**d**) *IGF2-MYH3*, *DLK1-MYH3*, *MEG3-MYH3* and *MEST-MYH3* for Network 3.

genes are connected to many other genes while high-betweenness genes are central and more likely to disconnect the network if removed. We analysed the evolution of the betweenness and degree from Network 0 to Network 3. Supplementary Table S3 shows a subset of 25 genes selected as key genes for the network structure because they showed a high betweenness or a high degree value or both a high betweenness and a high degree, or because they were among genes whose associations tested positive with 3D DNA FISH. Most of the genes presenting the highest betweenness values in Network 0, still kept or increased this numerical characteristic in Network 3 after network inference iterations. However, important changes were observed in some genes. For instance, *AKR7A2*, *DLK1*, *EGFR*, *MEG3*, *MYH3 and RPL32*, showed more than a 40% decrease in betweenness accompanied by a decrease in degree (>25%) when Network 3 was obtained. *DCN* showed a pronounced decrease in its degree while its betweenness was slightly modified. Interestingly, *MEST* and *IGF2* were found to have a mixed profile of betweenness and degree: in Network 3, we observed a 46% loss for *MEST* in gene connections, as compared to Network 0, while its betweenness increased by 160%. Similarly, a 30% loss of connections and a 426% gain in betweenness was observed for *IGF2*.

Network clustering. To analyse the evolution of the network structure from Network 0 to Network 3, clustering of the genes was performed on each network (for more details, see "Network mining and clustering" in

Gene associations	Number of nuclei analysed	Percentage of nuclei with signals			
		Distant (d > 1 µm)	Close (0, 5 < d ≤ 1 µm)	Co-localized (d < 0.5 µm)	Associated (d ≤ 1 µm)
MEST* - IGF2*	100	66	32	2	34
MEST* - (DLK1-MEG3)*	90	66	28	6	34
DCN - (DLK1-MEG3)*	73	85	15	0	15
RPL32 - IGF2*	80	80	16	4	20
DCN - IGF2*	98	90	7	3	10
IGF2* - MYH3	58	48	43	9	52
(DLK1-MEG3)* - MYH3	69	55	38	7	45
MEST* - MYH3	103	74	23	3	26
ZAR1 - IGF2*	61	92	8	0	8
ZAR1 - PRLR	63	92	8	0	8

Table 1. Association percentages of tested gene pairs. Associated signals (close + co-localized) are considered as those separated by a 3D distance (d) ≤ 1 µm, and are divided into two different classes: "close" signals (0.5 < d ≤ 1 µm), and "co localized" signals (d ≤ 0.5 µm). *Genes imprinted in pig.

"Materials and Methods" and Supplementary Tables S1 and S4). Four significant clusterings (p-value < 0.002) were obtained, one for each network. A total of nine clusters were obtained in Network 0, six in Network 1, eight in Network 2 and six in Network 3. Networks 0 and 3 were analysed in depth to search for any correspondence between clusters (Supplementary Table S5). Four clusters in Network 0 were found to share at least two thirds of their nodes with the corresponding clusters in Network 3. More precisely, 64.1% of the genes in cluster 1, 68.4% in cluster 2, 66% in cluster 3 and 82.4% in cluster 4, were observed in the corresponding clusters of Network 3. The other clusters in Network 0 (clusters 5, 6, 7, 8 and 9) were mainly spread each into two different clusters of Network 3. Additionally, the Normalized Mutual Information (NMI) value was calculated to quantify the similarity between clusterings for pairs of networks (Table 2). Interestingly, we observed that the clustering obtained in Network 0 was the most similar to the clustering obtained in Network 1 (NMI = 0.389). Similarly, the clustering in Network 1 was the most similar to the one obtained in Network 2 (NMI = 0.401), and the clustering in Network 2 was the most similar to the one obtained in Network 3 (NMI = 0.401). This finding suggests that clusterings become more consistent when introducing new biological information in each network inference iteration.

Functional enrichment analysis. To test the biological relevance of each cluster in Networks 0 and 3, a functional enrichment analysis was performed for each cluster from both networks. Significant GO terms for Biological Processes (GOBP) were observed in clusters 1 and 2 of Networks 0 and 3, and in clusters 3, 5 and 8 of Network 0 (Table 3 and Supplementary Table S6). Table 3 shows the four clusters presenting the non-redundant GOBP with the smallest False Discovery Rate (FDR). When comparing cluster 1 in Networks 0 and 3, eight common enriched GO terms were observed, mainly involved in extracellular matrix formation, embryonic development, metabolic processes and cellular response to stimulus. Besides, fourteen common enriched GOs were observed in cluster 2 of Networks 0 and 3. These GO terms were mainly involved in cellular respiration, energy metabolism, cellular metabolic processes and metabolism of fatty acids. Additionally, two GO terms were observed only in cluster 2 of Network 3, both involved in the mitochondrial respiratory processes. Interestingly, the smallest FDR were observed in Network 3: (i) for cluster 1 (containing all genes tested by 3D DNA FISH), referring to the "Extracellular structure" term (involving the Decorin gene *(DCN)*; FDR = 1.14e-08); (ii) for cluster 2, referring to the "Generation of precursor metabolites and energy" term (FDR = 1.32e-07) (Table 3).

These results suggest that our approach to network inference by incorporating *a priori* biological information enables us to obtain relevant GO terms while conserving the functional enriched terms found in the initial network (Network 0). Moreover, we unexpectedly observed that two (*IGF2* and *DCN*) of our seven target genes showed more significant GO terms in Network 3 than in the initial network. Specifically, *IGF2* was observed to be uniquely involved in the "Genetic imprinting" term in cluster 3 of Network 0 (FDR = 3.82e-02), while in cluster 1 of Network 3 it was found to be involved in two new significant GO terms, the one with the smaller FDR being "Skeletal system development" (FDR = 3.05e-03) (Table 3 and Supplementary Table S6). *DCN* was in turn observed to be involved in the "Sulphur compound metabolic process" term (FDR = 7.47e-03) in cluster 2 of Network 0, while in cluster 1 of Network 3 it appeared to be involved in the "Extracellular structure" term presenting the smallest FDR value (1.14e-08) of all clusters. Concerning *MEST*, *MYH3* and *DLK1*, also tested by 3D DNA FISH, even though the observed FDR were higher than 5%, interesting GO terms were observed for these genes in cluster 1 of Network 3 (Supplementary Table S6). For instance, *MEST* was found to be involved in "Mesoderm development", *MYH3* in "Body morphogenesis", *DLK1* in "Notch signalling pathway" and *DCN* and *MYH3* were both found to be involved in "Muscle organ development".

Another functional analysis was performed with Ingenuity Pathway Analysis (IPA) specifically on cluster 1 of Network 3, which contains the target genes (*IGF2*, *DLK1*, *MEG3*, *RPL32*, *MEST*, *DCN* and *MYH3*). IPA proposed to connecting 49 (82%) out of 60 genes in a network including all target genes except *MEG3* and *MYH3*. *MYH3* was found in a small network with 8 out of 60 genes, and *MEG3* in another small network of only 1 out of 60 genes. Furthermore, *MYOD1* and *CTNNB1* were identified by upstream regulator analysis as potential transcriptional factors for a group of genes including *IGF2* and *MYH3*. As IPA offers the possibility of merging networks (if there are links between nodes in the Ingenuity Pathways Knowledge Base), a reconstructed network

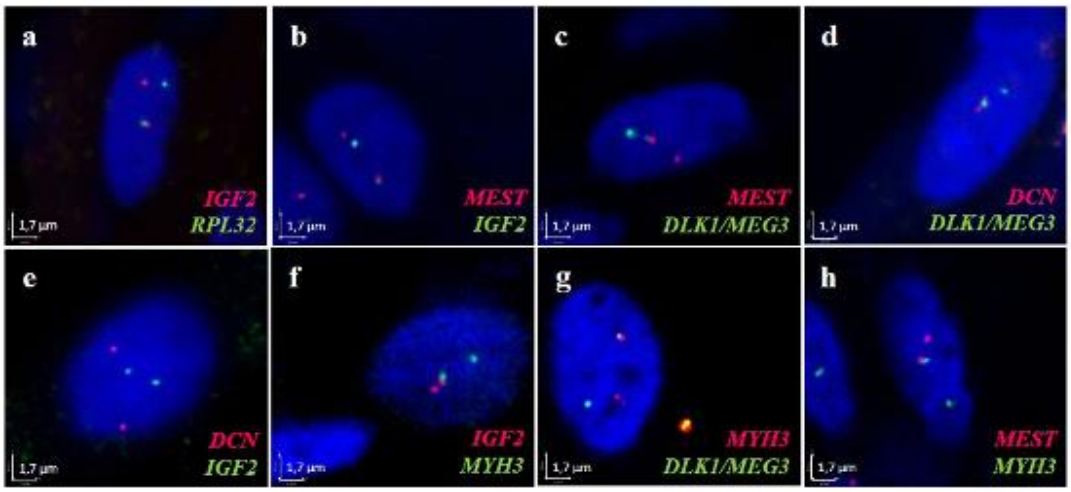


Figure 3. Analysis of gene associations by DNA FISH. Extended focus of 3D image sections from confocal microscopy and overlay of the 3 channels (blue, red and green) were obtained with Volocity v6.0 software (Perkin Elmer). The four signals in the nuclei correspond to the two alleles of each gene. Nuclei are counterstained with DAPI (blue). In all experiments, the percentage of association between genes was higher than 10% except for (e). Scale = 1.7 µm.

	Network 0	Network 1	Network 2	Network 3
Network 0	1	0.3893	0.3381	0.3244
Network 1	0.3893	1	0.4007	0.3923
Network 2	0.3381	0.4007	1	0.4152
Network 3	0.3244	0.3923	0.4152	1

Table 2. Normalized mutual information (NMI) between pairs of clusterings. NMI measure the similarity between two clusterings. The value is comprised between 0 and 1 and is equal to 1 when the two clusterings are identical.

was obtained (Fig. 4), and analysed around the target genes. Fourteen genes, among them 7 genes from cluster 1 (including *DCN* and *IGF2*), were observed to be related to "Cell Morphology" (p-value = 1.75e-08). *DCN*, *DLK1* and *IGF2* were likewise involved in the "Quantity of cells" function with 31 genes, including 16 genes from cluster 1 (p-value = 2.48e-09).

"Morphology of connective tissue cells" with 8 genes (p-value = 1.27e-04) included *DLK1* and *MEST*. "Formation of muscle", with 10 genes (p-value = 2.98e-05), involved *IGF2* and *MYH3* together with the two transcription factors *CTNNB1* and *MYOD1* (Supplementary Table S8).

Discussion

We present here a new approach based on GGM that enables the user to introduce previously acquired biological knowledge to build gene co-expression networks. Since an observed correlation between two genes in the co-expressed gene network does not necessarily mean that these genes are related to a common biological process, we used information of gene nuclear co-localizations to reinforce observed links in the co-expressed gene network. Some studies have shown examples of co-expressed and co-localized genes being implicated in a particular process, e.g. the *Hbb* and *Hba* Klf1-regulated globin genes were found to be co-localized in specialized Klf1-enriched transcription factories of erythroid cells[3]. Others have observed a role of co-expressed and co-localized genes in gene expression regulation, e.g. in the HUVECs endothelial cell line, *SAMD4A*, *TNFAIP2* and *SLC6A5* TNFα-induced genes were hierarchically transcribed when engaged in chromosomal interactions[30].

In order to determine which pairs of genes would present a reinforced edge in the networks, we performed two negative controls (see "gene-gene associations" in the "Materials and Methods" section). As discussed in our previous study[17], it can be difficult to define a suitable non-associating control. Sandhu *et al.* established a threshold of 2%[31], while others used the expected frequency of random co-localization based on the volume of the nucleus and individual gene signals (<1%)[5]. This estimation of random co-localization does not take into account other constraints such as: (1) chromosomes occupy specific territories[4,6]; (2) transcriptionally silent domains reside at the nuclear periphery[32]; (3) chromatin regions are preferentially associated in topological domains (TADs)[33]. Fixing an arbitrary threshold of 10% was a more restrictive way of analysing co-expressed genes that might tend to interact preferentially. Consequently, the pair *IGF2-DCN* was given as not co-localized by enforcing the absence of an edge between both genes.

GO ID	GOBP Terms	Network 0 - Cluster 1		Network 3 - Cluster 1	
		Genes	FDR	Genes	FDR
43062	Extracellular structure	*POSTN, COL1A1, COL1A2, COL3A1, COL5A1, COL16A1, LAMA4, MFAP5*	5,76E-05	*POSTN, COL1A1, COL1A2,COL3A1, COL5A1, COL5A2, COL16A1, DCN, FAP, FBN1, ABI3BP, ANXA2, LAMA4*	**1,14E-08**
71417	Cellular response to organonitrogen compound	*COL1A1, COL1A2,COL3A1, COL5A2, COL16A1, FYN, KLF3, ZFP36L1, HSP90B1*	**6,80E-04**	*COL1A1, COL1A2,COL3A1, COL5A2, COL16A1, DNMT1, FBN1, IGF2, HSP90B1*	1,16E-02
45995	Regulation of embryonic development	*COL5A1, COL5A2, FGFR1, LAMA4, LFNG*	2,24E-03	*COL5A1, COL5A2, FGFR1, LAMA4, LFNG*	1,16E-02
71559	Reponse to transforming growth factor beta	*POSTN, COL1A1, COL1A2, COL3A1, FYN, ZFP36L1*	**2,35E-03**	*POSTN, COL1A1, COL1A2,COL3A1, FBN1*	1,24E-01
44236	Multicellular organism metabolic process	*COL1A1, COL1A2,COL3A1, COL5A1, COL5A2*	2,35E-03	*COL1A1, COL1A2,COL3A1, COL5A1, COL5A2, FAP*	3,05E-03
43588	Skin development	*COL1A1, COL1A2, COL3A1, COL5A1, COL5A2, ZFP36L1*	**3,18E-03**	*COL1A1, COL1A2, COL3A1, COL5A1, COL5A2*	1,44E-01
1101	Reponse to acid chemical	*COL1A1, COL1A2,COL3A1, COL5A2, COL16A1, NFATC4*	1,17E-02	*COL1A1, COL1A2,COL3A1, COL5A2, COL16A1, DNMT1, NFATC4*	2,27E-02
1501	Skeletal system development	*POSTN, COL1A1, COL1A2, COL3A1, COL5A2, FGFR1, TMEM119*	1,43E-02	*POSTN, COL1A1, COL1A2, COL3A1, COL5A2, FBN1, FGFR1, ANXA2, TMEM119, IGF2*	3,05E-03
		Network 0 - Cluster 2		**Network 3 - Cluster 2**	
72350	Tricarboxylic acid metabolic process	*CS, DLAT, DLD, NNT, MDH1, PDHA1*	3,02E-06	*CS, DLAT, DLD, NNT, MDH1, PDHA1*	2,11E-05
51186	Cofactor metabolic process	*COQ7, DLAT, DLD, NNT, HK1, ACACB, NMNAT3, ACAT1, MDH1, PDHA1, PDHX*	**2,97E-05**	*DLAT, DLD, IBA57, NNT, GPI, ACACB, NMNAT3, MDH1, PDHA1, FLAD1, MCEE*	1,34E-03
72524	Pyridine-containig compound metabolic process	*DLD, NNT, HK1, NMNAT3, MDH1, PDHA1, PDHX*	**1,00E-04**	*DLD, NNT, GPI, NMNAT3, MDH1, PDHA1*	1,11E-02
6631	Fatty acid metabolic process	*CPT1B, ECI1, DLAT, DLD, ACACB, ACADS, ACAT1, PDHA1, PTGES2, PDHX*	**1,00E-04**	*CPT1B, ECI1, DLAT, DLD, FABP3, ACACB, ACADS, PDHA1, ADIPOR2, PTGES2, MCEE*	1,17E-03
6091	Generation of precursor metabolites and energy	*CS, DLAT, DLD, NNT, HK1, MDH1, OXA1L, ATP5B, PDHA1, SLC25A3*	1,09E-04	*CS, DLAT, DLD, NNT, GPI, MDH1, NDUFA3, NDUFB5, NDUFS1, OXA1L, ATP5B, PDHA1, SLC25A3, CISD1, NDUFA12, PYGM*	**1,32E-07**
6090	Pyruvate metabolic process	*DLAT, DLD, HK1, PDHA1, PDHX*	5,42E-03	*DLAT, DLD, GPI, PDHA1, BSG*	2,32E-02
6790	Sulfur compound metabolic process	*VCAN, DCN, DLAT, DLD, ACACB, ACAT1, PDHA1, PDHX*	**7,47E-03**	*DLAT, DLD, IBA57, ACACB, PDHA1, MCEE*	4,79E-01
42180	Cellular ketone metabolic process	*COQ7, DLAT, DLD, ACACB, PDHA1, PDHX*	1,46E-02	*DLAT, DLD, FABP3, GPI, ACACB, PDHA1*	8,05E-02
45454	Cell redox homeostasis	*TXNRD2, DLD, NNT, PTGES2*	1,46E-02	*TXNRD2, DLD, NNT, PTGES2*	4,91E-02
44282	Small molecule catabolic process	*CPT1B, ECI1, DLD, HK1, ACACB, ACADS, ACAT1*	1,88E-02	*CPT1B, ECI1, DLD, GPI, ACACB, ACADS, BCAT2, MCEE*	4,51E-02
98656	Anion transmembrane transport	*CLCN5, CPT1B, ACACB, SLC25A3, SLC1A3, VDAC1*	**2,31E-02**	*CPT1B, ACACB, SLC25A3, SLC1A3, VDAC1*	3,77E-01
6081	Cellular aldehyde metabolic process	*DLAT, DLD, PDHA1, PDHX*	2,59E-02	*DLAT, DLD, GPI, PDHA1*	8,73E-02
43648	Dicarboxylic acid metabolic process	*DLD, NMNAT3, MDH1, SLC1A3*	3,13E-02	*DLD, NMNAT3, MDH1, BCAT2, SLC1A3*	2,13E-02
16042	Lipid catabolic process	*CPT1B, ECI1, ACACB, ACADS, ACAT1, NCEH1*	3,65E-02	*CPT1B, ECI1, FABP3, ACACB, ACADS, NCEH1, MCEE*	6,59E-02
10257	NADH dehydrogenase complex assembly			*NDUFA3, NDUFB5, NDUFS1, OXA1L, NDUFA12*	**3,29E-03**
97031	Mitochondrial respiratory chain complex I biogenesis			*NDUFA3, NDUFB5, NDUFS1, OXA1L, NDUFA12*	**3,29E-03**

Table 3. Comparison of GOBP in clusters 1 and 2 between Network 0 and Network 3. GO terms enriched in one of the clusters as well as all GO terms associated to one of the three target genes at least (even if not significantly enriched). In bold, the smallest FDR value for a given GOBP term when the difference between the FDR of the two clusters is higher than one order of magnitude. Genes tested by 3D DNA FISH are in underline bold.

Testing the nuclear co-localization of *IGF2* and *RPL32* by 3D DNA FISH proved interesting, as this connection concerned an imprinted gene (*IGF2*, involved in muscle growth-related traits[14]) and a ribosomal protein coding gene *RPL32*[34]. This experiment revealed that these genes are associated. Additionally, it was interesting to find co-localized pairs of genes such as *IGF2-MEST*, *(DLK1/MEG3)-MEST*, *(DLK1/MEG3)-DCN*, that were observed

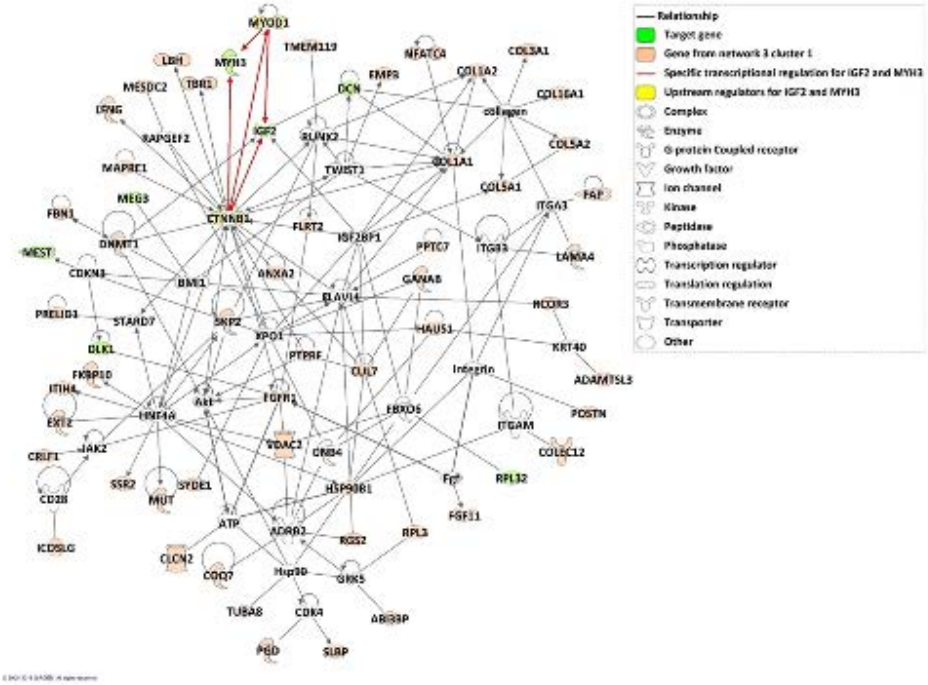


Figure 4. Reconstructed network of genes in cluster 1 of Network 3, based on Ingenuity Pathways Knowledge Base. Nodes are displayed using various shapes that represent the functional class of the gene product. The reconstructed network was generated through the use of Ingenuity Pathway Analysis (IPA) (Ingenuity Systems; QIAGEN, Inc., Valencia, CA, USA).

to be connected in co-expression networks in other studies[12,23], even though they were not directly connected by an edge in our network (Network 1) but via intermediary genes. Besides, surprising results showed the highest association we have ever observed between two genes (neither in the present study, nor in previous ones). This association concerns *MYH3* and *IGF2*. *MYH3* plays an important role in foetal muscle development[35,36], and encodes for the embryonic Myosin Heavy Chain (MYHC) 3 protein. To the best of our knowledge, no previous association between these two genes, whatever its origin (nuclear or functional), has ever been observed, even though the two genes are known to be involved in muscle development[35,37]. To determine the impact of the *a priori* co-localization information introduced to enforce the presence or the absence of an edge, we analysed the evolution from Network 0 to Network 3, first globally (with conserved edges and key genes) and then locally (with network clustering and functional enrichment). The global analyses revealed that 82% of edges in Network 0 were conserved in Network 3 and that the most important genes (with respect to network structure) in Network 0 were among those showing the highest values of betweenness and degree in Network 3. These findings suggest that the introduction of enforced edges is not linked to the appearance of major disturbances in the network structure. However, when focusing on the target genes analysed by 3D DNA FISH, we observed a general decrease in the degree value, meaning that *IGF2*, *DLK1*, *MEG3*, *RPL32*, *MEST*, *DCN* and *MYH3* were less connected with the rest of the other genes in Network 3. Despite this observed isolation concerning genes for which edges were enforced, this effect was not always accompanied by a loss of betweenness. In other words, reinforcing a limited number of edges did not change either the global network structure or the importance of target genes in the final network. In the local analysis, the NMI value revealed that the clusters resembled one another more with each new network inferred. In addition, four out of six clusters in the final network (Network 3) conserved more than 62% of genes in the corresponding clusters of Network 0. This concurred with the results of the functional enrichment analysis, which revealed that the GOs found were conserved between Networks 0 and 3. All these results support the evidence that our approach did not introduce any substantial disturbance. In fact, this iterative process brought substantial improvements; notably, it enabled us to obtain reliable networks in terms of relevant biological information, especially around our target genes. This was supported by the following findings: (1) the biological processes presenting the smallest FDR were found in Network 3, even though one of them involved *DCN*, for which edge estimations were modified by the introduction of *a priori* information; (2) two new significant GO terms related to energy metabolism appeared in cluster 2 of Network 3; (3) two genes (*IGF2* and *DCN*) analysed by 3D DNA FISH were involved in biological processes with smaller FDR in Network 3 than in Network 0. Moreover, *IGF2* was found in an additional GO of Network 3, while only present in one GO of Network 0.

One of the most important goals of the present article was to elucidate the mechanisms that govern porcine skeletal muscle development in late gestation. Many studies have been performed in pig to address this

question[21,24,38–41]. In our model, we proposed a final network (Network 3) in which enriched biological functions related to muscle development were observed. These observations were in agreement with the results obtained by Voillet *et al.*[21]. In addition, in the resulting IPA reconstructed network, we highlighted *MYOD1* and *CTNNB1* among the proposed transcription factors because they were especially interesting due to their connection to two important target genes, *IGF2* and *MYH3*. Although *MYOD1* and *CTNNB1* were not present in the 359 genes used for network inference, they were up-regulated at 90 days of gestation in all genotypes (Supplementary Fig. S7)[21]. *MYOD1* encodes for a myogenic factor that regulates skeletal muscle cell differentiation by activating transcription of muscle-specific target genes (for review[42]). *CTNNB1* (β-catenin 1), encodes for a transcriptional co-activator that was found to be required for muscle differentiation in murine myoblasts by interacting directly with MyoD and promoting its binding to the E box elements enhancing its transcriptional activity[43]. The co-expression and nuclear co-localization of *IGF2* and *MYH3* suggest they are each subjected to similar transcriptional regulation by these two transcription factors. The studies of Shang *et al.*[44] and Ramazzotti *et al.*[45] are in agreement with this hypothesis. Shang *et al.* revealed that in mesenchymal stromal cells from rats, an ectopic expression of *Ctnnb1* inhibits adipogenetic differentiation and induces the formation of long multinucleated cells expressing myogenic genes, such as *MyoD* and *Myhc*, by promoting the expression of skeletal muscle-specific transcription factors. Ramazzotti *et al.* observed that an overexpression and accumulation of β-catenin in the nuclei of differentiating murine myoblasts results in higher *MyoD* activation and *Myhc* induction. Additionally, *IGF2* was found to be up-regulated in pig during myogenesis and, more precisely, involved in primary and secondary muscle fibre differentiation[41]. Moreover, *Myod* and *Igf2* were observed to be involved in the switch between myogenic and adipose lineages in mouse[46]. In addition, we found *IGF2* indirectly associated with *CTNNB1* (through the intermediary gene *IGF2BP1*) in the reconstructed network. *IGF2BP1* was not used for network inference but was found expressed at the 90th day of gestation (Supplementary Fig. S7)[21]. Indeed, β-catenin was observed to induce *IGF2BP1* in HEK293 cells[47], which in turn was observed to regulate *IGF2* mRNA subcellular location and translation in neurons (for review[48]). This suggests that in muscle cells, a similar mechanism could possibly be involved for the regulation of *IGF2* via the CTNNB1 transcription factor. Moreover, the long non-coding DNA of *MyoD* (*lncMyoD*), directly activated by MyoD, may negatively regulate *Igf2bp1*-mediated translation of proliferation genes in murine myoblasts[49]. This could explain how MyoD blocks proliferation to create a permissive state of differentiation. Moreover, *DLK1* and *MYOD1* were not connected in the reconstructed network. However, *DLK1* which encodes for a preadipocyte factor that inhibits adipocyte differentiation[50], might inhibit cell proliferation and enhance cell differentiation by regulating the expression of *MyoD*[16]. Combining all this information with the observed up-regulation at 90 days of gestation of the above-mentioned genes, our results highlight a network of interrelated genes associated with skeletal muscle regulation and that are mainly responsible for inhibition of proliferation and muscle differentiation.

Conclusion

The innovative approach presented here has proven to be consistent, robust and reliable for the inference of gene co-expression networks in combination with gene nuclear co-localizations. The information generated by the final network brought to light relevant functions involved in the development and maturity of foetal muscle. In this context, the challenge for future studies will be to broaden this approach and render it more powerful by combining co-expression data with information about genome-wide interactions[51,52] to enforce edges in the network. This study also spotlights interesting gene associations in the three-dimensional nuclear space of muscle cells such as the associations found between *MYH3-IGF2* or *MYH3-(DLK1/MEG3)*. The three genes are up-regulated in LW at 90 days of gestation and are involved in muscle development. Determining through further functional studies whether and how these genes are co-regulated, will help us to understand the mechanisms involved in the establishment of pig muscle maturity.

Materials and Methods

Ethics Statement. All tissues sampled for the experiments were collected on pigs bred for another project (ANR-09-GENM-005-01, 2010–2015). The experiment authorization number for the experimental farm GenESI (Genetics, testing and innovative systems experimental unit) is A-17-661. The procedures performed in this study and the treatment of animals complied with European Union legislation (Directive 2010/63/EU) and French legislation in the Midi-Pyrénées Region of France (Decree 2001-464). The ethical committee of the Midi-Pyrénées Regional Council approved the experimental design (authorization MP/01/01/01/11). All the foetuses used in this study were males and were obtained by caesarean.

Microarray data description and pre-processing. Expression data were obtained from skeletal muscle for two foetal gestational ages (90 and 110 days of gestation) associated with four foetal genotypes (two extreme breeds for mortality at birth –Large White (LW) and Meishan (MS)– and two reciprocal crosses –MSxLW and LWxMS). The final dataset consisted of 44,368 probes for 61 samples under eight different conditions (four genotypes at two gestational ages). A precise description of the experimental design and data collection can be found in Voillet *et al.*[21]. Normalized expression data (log2-transformed) and sample information are available in NCBI (GEO accession number GSE56301).

Missing values were imputed with k-NN (R package "impute" function, with $k = 3$). Gene annotation was updated (nblast/NCBI July 2017, Sscrofa10.2) and the 40,847 annotated probes were found to correspond to 13,855 unique genes. For each gene, the probe with the highest average correlation with the other probes associated with the same gene was selected to serve as a representative in further statistical analyses.

Network inference. Networks were inferred using Gaussian Graphical Models (GGMs[28]) from $n = 61$ samples. From expression data, GGMs build a graph (or network) in which vertices are genes and edges represent

the conditional dependency structure between those genes. GGMs are based on the estimation of partial correlations (*i.e.*, correlations between two gene expressions when the expression of all the other genes is known). They were preferred over relevance networks[53] because they improve measurement of direct relations between gene expressions by accounting for the effect of all expression data, and because they were found to be more efficient for grouping together genes with a common function in a previous study[29].

Since the number of samples was smaller than the number of genes used for network inference, the models were fitted with a sparse penalty[54] to address the issues of high-dimensional data and edge selection. In addition, as many examples have shown that co-expressed genes occasionally tend to interact preferentially or consolidate in specialized foci of the nuclear environment[2–5], when *a priori* information about nuclear gene co-localization is available, the latter was included in the model using the approach described in Villa-Vialaneix *et al.*[55]. The details of the method and of the tuning of the different parameters are given in Supplementary Methods online.

Practical implementation of network inference. The starting point of the analysis was the inference of a network with no *a priori* information about co-localization. Since network inference based on partial correlation can only be performed with a limited number of genes (because of the number of samples) and since the number of unique genes ($p = 13{,}855$) was too great compared to the number of samples ($n = 61$), we applied two restrictions to the original list. First, we restricted the list to genes that were reported as differentially expressed (DEG)[21]. Secondly, among these DEGs, only those that had an absolute value for their correlation with either *IGF2*, *DLK1* or *MEG3* larger than 0.84 were kept. This final list contained 359 genes, provided in Supplementary Table S1.

Network inference iteration and 3D FISH validations. Based on network inference results or on genes found to be connected in the IGN of Varrault *et al.*[12], 3D DNA FISH experiments were performed to check whether pairs of genes of interest were co-localized in the 3D nuclear space. These experiments were conducted in an iterative manner with network inference. More precisely, network inference was performed with the following *a priori* conditions: (1) Network 0: was inferred with no *a priori* information, as a baseline for comparison; (2) Network 1: was inferred using *a priori* information from the triple association found in Lahbib-Mansais *et al.*[17] by giving the three pairs *IGF2-DLK1*, *IGF2-MEG3* and *DLK1-MEG3* as known co-localized genes. Network 1 was then used to propose candidate pairs of genes for testing by 3D DNA FISH for Network 2 (*IGF2-RPL32*) and Network 3 (*DLK1-MYH3*); (3) Network 2: in addition to the initial three pairs, Network 2 was inferred using *a priori* information provided by the results of the new 3D DNA FISH experiments by giving the pairs *IGF2-MEST*, *DLK1-MEST*, *MEG3-MEST*, *MEG3-DCN*, *DLK1-DCN*, and *RPL32-IGF2* as known to be co-localized and *IGF2-DCN* as known not to be co-localized; (4) Network 3: in addition to the 10 previous pairs, Network 3 was inferred using *a priori* information provided by the results of new 3D DNA FISH experiments by giving the additional pairs *IGF2-MYH3*, *DLK1-MYH3* *MEG3-MYH3* and *MEST-MYH3* as known co-localized genes.

All simulations were performed with the free statistical software R (https://cran.r-project.org). The inference was performed using our own scripts (available at https://github.com/tuxette/internet3D) and the graphs were displayed and analysed using the R package igraph (Csardi and Nepusz)[56].

Network mining and clustering. Nodes of importance to the network structure were obtained by computing the degree and the betweenness centrality measurement for every node. Node clustering was performed by applying the Louvain algorithm[57], which performs fast approximate optimization of the modularity[58]. All clusterings were found to be significant using the permutation test described in Montastier *et al.*[59] by generating 500 random networks with the same degree distribution (all clusterings were found to have a modularity larger than that obtained on the 500 random networks, p-value < 0.002). Clusters were compared using two methods: first, pairwise contingency tables between clusters were computed. Second, the normalized mutual information (NMI[60]) between pairs of clusterings was obtained. The NMI is a number between 0 and 1 measuring the similarity between two clusterings and is maximum (equal to 1) when the two clusterings are identical.

Functional analysis of the networks. Functional enrichment analysis based on GO was performed using the web tool Webgestalt (WEB-based GEne SeT AnaLysis Toolkit, http://www.webgestalt.org/option.php) updated on January 27, 2017[61,62]. The web tool uses the Fisher exact test and controls for the number of false positives among the declared significant GOs terms. The False Discovery Rate was used (Benjamini-Hochberg procedure[63], FDR < 5%). The analysis was performed using the Overrepresentation Enrichment Analysis (ORA) method, selecting non-redundant Biological Processes (BPs). The final network was analysed through the use of Ingenuity Pathway Analysis version 01–12 (updated on March 31st, 2018). Ingenuity Pathway Analysis (IPA, Ingenuity Systems; QIAGEN, Inc., Valencia, CA, USA, https://analysis.ingenuity.com/pa) contains a large bibliographic database (Ingenuity Pathways Knowledge Base) with various molecular relationships already identified between two genes (protein-protein interaction, ligand-receptor regulation, enzymatic modification, transcriptional expression regulation, etc.). The obtained network is a graphic representation of the molecular relationships between molecules. All edges are supported by at least one reference from the literature, or from canonical information stored in the Ingenuity Pathways Knowledge Base. The obtained networks were improved for representation using Path Designer. Nodes are displayed using various shapes that represent the functional class of the gene product. The Functional Analysis identified the biological functions, the canonical pathways and the upstream regulators that were the most relevant to the dataset. Molecules from the dataset that were associated with biological functions, canonical pathways or upstream regulators in the Ingenuity Knowledge Base were considered for the analysis. Fisher's exact test was used to calculate a right-tailed p-value determining the probability that each function and pathway assigned to that dataset is due to chance alone. The networks proposed by IPA were cleaned (some nodes/genes were discarded) in order to keep only the genes necessary to connect the

co-expressed genes. The three first networks were merged and regulation information was added to highlight transcription factors that could explain unexpected gene co-expression and nuclear co-localization (e.g. *MYH3* and *IGF2*; Supplementary Table S8).

Tissue preparation. Foetal muscle tissue was obtained from the *Longuissimus dorsi* muscle of 90-day gestation ♀MSxLW♂ pig and prepared as described in[17] with slight modifications. When needed, stored muscle fibre packets were permeabilised for 8 min in cytoskeleton extraction buffer (100 mM NaCl, 300 mM sucrose, 3 mM MgCl2, 10 mM PIPES pH 6.8) containing 0.5% Triton X-100 and then fixed in cold 4% paraformaldehyde for 5 min. After washing in cold PBS, muscle packets were manually dilacerated directly on Superfrost glass slides (CML, Nemours, France) to isolate individual fibres, and air-dried before adding DNA probes for *in situ* hybridization.

Probes construction. Bacterial artificial clones (BACs) containing genes were isolated from porcine BAC libraries (available at the Biological Resources Center-GADIE, INRA, Jouy-en-Josas, France http://abridge.inra.fr/) using specific primers designed with Primer3 software (http://primer3.sourceforge.net/) (Supplementary Table S2). For multiple-label experiments, approximately 120 ng of each BAC DNA was random-priming labelled directly by incorporation of dUTP Alexa Fluor (488 or 568) or indirectly with Biotin-6-dUTP detected by immuno-FISH (Bioprime DNA labelling kit, Invitrogen, Cergy Pontoise, France). Chromosomal localizations of all BAC probes were controlled by 2D DNA FISH on porcine metaphases prepared from lymphocytes according to standard protocols[64].

IGF2 had been localized previously on SSC2p17, *DLK1/MEG3* on SSC7q26 and *ZAR1* on SSC8q11-12[17]. In this study, additional genes were localized on pig metaphases: *MYH3* on SSC12q, *MEST* on SSC18, *RPL32* on SSC13q24-33, *DCN* on SSC5qter, and *PRLR* on SSC16 (Supplementary Table S2).

3D DNA-FISH on interphase nuclei. 3D DNA FISH experiments were conducted using specific probes to label each gene with a different colour as described in[17] with slight modifications. Probes were resuspended in hybridization buffer (50% formamide, 10% dextran sulphate, 2 mg/ml BSA, 2× SSC) at a final concentration of 110 ng/µl. Nuclear DNA and probes were simultaneously denatured at 74 °C for 7 min and then incubated overnight at 37 °C in a wet atmosphere (DAKO hybridizer). Washes were then performed with gentle agitation, first twice in 2× SSC at room temperature (RT) for 8 min, then twice for 3 min in 2× SSC, 50% formamide pH 7.0 at 40 °C, and finally twice for 15 min in 2× SSC, then in PBS at RT. When a biotin-labelled probe was used, biotins were detected by incubating the slides with streptavidin-Alexa 568 or 488 for 1 hour at RT.

Confocal microscopy and image analyses. Image stacks were captured at different depths with a Leica TCSSP2 confocal microscope (Leica Instruments, Heidelberg, Germany) equipped with an oil immersion objective (plan achromatic 63× N.A. = 1.4). The Z-stacks (around 60 confocal planes per capture) were acquired at 1024 × 1024 pixels per frame using an 8-bit pixel depth for each channel at a constant voxel size of 0.077 × 0.077 × 0.284 µm. Images were analysed with specific software for measuring the 3D distances (centre-to-centre) between signals (genes) (NEMO[65]) as described in[17]. Euclidean distances were computed with respect to the x, y and z resolutions. Given the resolution on the z axis, at least three pixels corresponding to 0.852 µm (0.284 × 3) were required for a high resolution between two separate signals; consequently, 1 µm was chosen as the upper cut-off for associated signals.

Gene-gene associations. In all 3D DNA FISH experiments, nuclei were only analysed when 4 signals (corresponding to the 2 alleles of each gene) were present. "Associated" signals were considered to be those separated by a distance (d) ≤ 1 µm, and were divided into two different classes: "close" signals (0.5 < d ≤ 1 µm), and "co-localized" signals (d ≤ 0.5 µm). The great majority of associations concerned uniquely one allele from each gene. To establish the threshold for distinguishing between associated and non-associated genes, two 3D DNA FISH experiments were performed as negative controls: first, between two genes (*ZAR1* and *PRLR*) located on different chromosomes and expressed at a very low level in muscle cells[21], second, between *IGF2* (highly expressed) and *ZAR1* (low expression)[17]. In both cases, the two genes were found to be associated in only 8% of the analysed nuclei. Considering this value as a sporadic association between loci not expected to be associated, a 10% value was arbitrarily chosen to distinguish between associated and non-associated genes.

Data availability. The data sets supporting the results of this article are available in the NCBI's Gene Expression Omnibus repository, and are available through GEO Series accession number GSE56301.

References

1. Pombo, A. & Dillon, N. Three-dimensional genome architecture: players and mechanisms. *Nat. Rev. Mol. Cell Biol.* **16**, 245–257 (2015).
2. Zhao, Z. *et al.* Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat. Genet.* **38**, 1341–1347 (2006).
3. Schoenfelder, S. *et al.* Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat. Genet.* **42**, 53–61 (2010).
4. Rieder, D. *et al.* Co-expressed genes prepositioned in spatial neighborhoods stochastically associate with SC35 speckles and RNA polymerase II factories. *Cell. Mol. Life Sci. CMLS* **71**, 1741–1759 (2014).
5. Osborne, C. S. *et al.* Active genes dynamically colocalize to shared sites of ongoing transcription. *Nat. Genet.* **36**, 1065–1071 (2004).
6. Bolzer, A. *et al.* Three-Dimensional Maps of All Chromosomes in Human Male Fibroblast Nuclei and Prometaphase Rosettes. *PLoS Biol.* **3** (2005).

7. Mahy, N. L., Perry, P. E. & Bickmore, W. A. Gene density and transcription influence the localization of chromatin outside of chromosome territories detectable by FISH. *J. Cell Biol.* **159**, 753–763 (2002).
8. Williams, A., Spilianakis, C. G. & Flavell, R. A. Interchromosomal association and gene regulation in trans. *Trends Genet. TIG* **26**, 188–197 (2010).
9. Bickmore, W. A. & van Steensel, B. Genome Architecture: Domain Organization of Interphase Chromosomes. *Cell* **152**, 1270–1284 (2013).
10. Chaumeil, J., Micsinai, M. & Skok, J. A. Combined Immunofluorescence and DNA FISH on 3D-preserved Interphase Nuclei to Study Changes in 3D Nuclear Organization. *J. Vis. Exp. JoVE*, https://doi.org/10.3791/50087 (2013).
11. Fudenberg, G. & Imakaev, M. FISH-ing for captured contacts: towards reconciling FISH and 3C. *Nat. Methods* **14**, 673–678 (2017).
12. Varrault, A. *et al.* Zac1 regulates an imprinted gene network critically involved in the control of embryonic growth. *Dev. Cell* **11**, 711–722 (2006).
13. Nezer, C. *et al.* An imprinted QTL with major effect on muscle mass and fat deposition maps to the IGF2 locus in pigs. *Nat. Genet.* **21**, 155–156 (1999).
14. Van Laere, A.-S. *et al.* A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. *Nature* **425**, 832–836 (2003).
15. St-Pierre, J. *et al.* IGF2 DNA methylation is a modulator of newborn's fetal growth and development. *Epigenetics* **7**, 1125–1132 (2012).
16. Waddell, J. N. *et al.* Dlk1 Is Necessary for Proper Skeletal Muscle Development and Regeneration. *PLoS ONE* **5** (2010).
17. Lahbib-Mansais, Y. *et al.* Expressed alleles of imprinted IGF2, DLK1 and MEG3 colocalize in 3D-preserved nuclei of porcine fetal cells. *BMC Cell Biol.* **17**, 35 (2016).
18. Institute of Medicine (US) Committee on Understanding Premature Birth and Assuring Healthy Outcomes. *Preterm Birth: Causes, Consequences, and Prevention*. (National Academies Press (US), 2007).
19. Leenhouwers, J. I. *et al.* Fetal development in the pig in relation to genetic merit for piglet survival. *J. Anim. Sci.* **80**, 1759–1770 (2002).
20. Foxcroft, G. R. *et al.* The biological basis for prenatal programming of postnatal performance in pigs. *J. Anim. Sci.* **84**(Suppl), E105–112 (2006).
21. Voillet, V. *et al.* Muscle transcriptomic investigation of late fetal development identifies candidate genes for piglet maturity. *BMC Genomics* **15**, 797 (2014).
22. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, Article17 (2005).
23. Al Adhami, H. *et al.* A systems-level approach to parental genomic imprinting: the imprinted gene network includes extracellular matrix genes and regulates cell cycle exit and differentiation. *Genome Res.* **25**, 353–367 (2015).
24. Tang, Z. *et al.* Integrated analysis of miRNA and mRNA paired expression profiling of prenatal skeletal muscle development in three genotype pigs. *Sci. Rep.* **5**, 15544 (2015).
25. Yu, H., Kim, P. M., Sprecher, E., Trifonov, V. & Gerstein, M. The Importance of Bottlenecks in Protein Networks: Correlation with Gene Essentiality and Expression Dynamics. *PLoS Comput. Biol.* **3** (2007).
26. Rives, A. W. & Galitski, T. Modular organization of cellular networks. *Proc. Natl. Acad. Sci. USA* **100**, 1128–1133 (2003).
27. Shen-Orr, S. S., Milo, R., Mangan, S. & Alon, U. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.* **31**, 64–68 (2002).
28. Edwards, D. *Introduction to Graphical Modelling*|*David Edwards*|*Springer* (1995).
29. Villa-Vialaneix, N. *et al.* The Structure of a Gene Co-Expression Network Reveals Biological Functions Underlying eQTLs. *PLoS ONE* **8** (2013).
30. Fanucchi, S., Shibayama, Y., Burd, S., Weinberg, M. S. & Mhlanga, M. M. Chromosomal contact permits transcription between coregulated genes. *Cell* **155**, 606–620 (2013).
31. Sandhu, K. S. *et al.* Nonallelic transvection of multiple imprinted loci is organized by the H19 imprinting control region during germline development. *Genes Dev.* **23**, 2598–2603 (2009).
32. Boyle, S. *et al.* The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Hum. Mol. Genet.* **10**, 211–219 (2001).
33. Dixon, J. R. *et al.* Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions. *Nature* **485**, 376–380 (2012).
34. Young, J. A. & Trowsdale, J. A processed pseudogene in an intron of the HLA-DP beta 1 chain gene is a member of the ribosomal protein L32 gene family. *Nucleic Acids Res.* **13**, 8883–8891 (1985).
35. Schiaffino, S., Rossi, A. C., Smerdu, V., Leinwand, L. A. & Reggiani, C. Developmental myosins: expression patterns and functional significance. *Skelet. Muscle* **5**, 22 (2015).
36. Voillet, V. *et al.* Integrated Analysis of Proteomic and Transcriptomic Data Highlights Late Fetal Muscle Maturation Process. *Mol. Cell. Proteomics MCP*, https://doi.org/10.1074/mcp.M116.066357 (2018).
37. Livingstone, C. & Borai, A. Insulin-like growth factor-II: its role in metabolic and endocrine disease. *Clin. Endocrinol. (Oxf.)* **80**, 773–781 (2014).
38. Cagnazzo, M. *et al.* Comparison of prenatal muscle tissue expression profiles of two pig breeds differing in muscle characteristics. *J. Anim. Sci.* **84**, 1–10 (2006).
39. Xu, Y. *et al.* Differential proteome and transcriptome analysis of porcine skeletal muscle during development. *J. Proteomics* **75**, 2093–2108 (2012).
40. Zhao, Y. *et al.* Dynamic transcriptome profiles of skeletal muscle tissue across 11 developmental stages for both Tongcheng and Yorkshire pigs. *BMC Genomics* **16**, 377 (2015).
41. Zhao, X. *et al.* Comparative Analyses by Sequencing of Transcriptomes during Skeletal Muscle Development between Pig Breeds Differing in Muscle Growth Rate and Fatness. *PLoS ONE* **6** (2011).
42. Berkes, C. A. & Tapscott, S. J. MyoD and the transcriptional control of myogenesis. *Semin. Cell Dev. Biol.* **16**, 585–595 (2005).
43. Kim, C.-H., Neiswender, H., Baik, E. J., Xiong, W. C. & Mei, L. Beta-catenin interacts with MyoD and regulates its transcription activity. *Mol. Cell. Biol.* **28**, 2941–2951 (2008).
44. Shang, Y. C. *et al.* Activated beta-catenin induces myogenesis and inhibits adipogenesis in BM-derived mesenchymal stromal cells. *Cytotherapy* **9**, 667–681 (2007).
45. Ramazzotti, G. *et al.* IPMK and β-catenin mediate PLC-β1-dependent signaling in myogenic differentiation. *Oncotarget* **7**, 84118–84127 (2016).
46. Borensztein, M. *et al.* Double Myod and Igf2 inactivation promotes brown adipose tissue development by increasing Prdm16 expression. *FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol.* **26**, 4584–4591 (2012).
47. Noubissi, F. K. *et al.* CRD-BP mediates stabilization of betaTrCP1 and c-myc mRNA in response to beta-catenin signalling. *Nature* **441**, 898–901 (2006).
48. Bell, J. L. *et al.* Insulin-like growth factor 2 mRNA-binding proteins (IGF2BPs): post-transcriptional drivers of cancer progression? *Cell. Mol. Life Sci. CMLS* **70**, 2657–2675 (2013).
49. Gong, C. *et al.* A long non-coding RNA, LncMyoD, regulates skeletal muscle differentiation by blocking IMP2-mediated mRNA translation. *Dev. Cell* **34**, 181–191 (2015).

50. Wang, Y., Hudak, C. & Sul, H. S. Role of preadipocyte factor 1 in adipocyte differentiation. *Clin. Lipidol.* **5**, 109–115 (2010).
51. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).
52. Dekker, J., Marti-Renom, M. A. & Mirny, L. A. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat. Rev. Genet.* **14**, 390–403 (2013).
53. Butte, A. J. & Kohane, I. S. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomput. Pac. Symp. Biocomput.* 418–429 (2000).
54. Meinshausen, N. & Bühlmann, P. High-dimensional graphs and variable selection with the Lasso. *Ann. Stat.* **34**, 1436–1462 (2006).
55. Villa-Vialaneix, N., Vignes, M., Viguerie, N. & SanCristobal, M. Inferring Networks from Multiple Samples with Consensus LASSO. *Qual. Technol. Quant. Manag.* **11**, 39–60 (2014).
56. Csárdi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal Complex Syst.* 1695 (2006).
57. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, P10008 (2008).
58. Clauset, A., Newman, M. E. J. & Moore, C. Finding community structure in very large networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **70**, 066111 (2004).
59. Montastier, E. *et al.* System model network for adipose tissue signatures related to weight changes in response to calorie restriction and subsequent weight maintenance. *PLoS Comput. Biol.* **11**, e1004047 (2015).
60. Danon, L., Duch, J., Diaz-Guilera, A. & Arenas, A. Comparing community structure identification. *J. Stat. Mech. Theory Exp.* **2005**, P09008–P09008 (2005).
61. Zhang, B., Kirov, S. & Snoddy, J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res.* **33**, W741–748 (2005).
62. Wang, J., Duncan, D., Shi, Z. & Zhang, B. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Res.* **41**, W77–83 (2013).
63. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to MultipleTesting. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).
64. Yerle, M., Goureau, A., Gellin, J., Le Tissier, P. & Moran, C. Rapid mapping of cosmid clones on pig chromosomes by fluorescence *in situ* hybridization. *Mamm. Genome Off. J. Int. Mamm. Genome Soc.* **5**, 34–37 (1994).
65. Iannuccelli, E. *et al.* NEMO: a tool for analyzing gene and chromosome territory distributions from 3D-FISH experiments. *Bioinformatics* **26**, 696–697 (2010).

Acknowledgements

This project received financial support from French National Agency of Research (PORCINET grant ANR-09-GENM005), from INRA Animal Genetics Division (Internet3D project) and from INRA GenPhySE (Génétique, Physiologie et Systèmes d'élevage) laboratory (Toulouse). The PhD fellowship of Maria Marti-Marimon is supported by the French Ministry of National Education. The PhD fellowship of Valentin Voillet was supported by the INRA Animal Genetics Division, the INRA Animal Physiology and Livestock Systems and the Occitanie region. The funders had no role in the study design, analyses, results interpretation and decision to publish. The authors would also like to thank Yvon Billon and collaborators (INRA experimental unit GenESI (UE1372)) for providing animals, and the T.R.I. Genotoul (Toulouse Réseau Imagerie, http://trigenotoul.com/en) imaging core facility in Toulouse (France) where 3D acquisitions were performed.

Author Contributions

Y.L.M., N.V. and L.L. conceived the integrative project. M.M.M., N.V., M.Y.B., Y.L.M. and L.L. conceived and designed the experiments. N.V. developed the statistical model. N.V. and V.V. performed network inference and statistical analyses. M.M.M. and Y.L.M. performed the FISH experiments. M.M.M., N.V., V.V., Y.L.M. and L.L. analyzed the data. M.M.M., N.V., M.Y.B., Y.L.M. and L.L. wrote the paper. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-28173-8.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Conformation 3D du génome et expression génique dans la cellule musculaire porcine en fin de gestation.

Maria Marti-Marimon – Vendredi 9 novembre 2018, Toulouse – Biologie moléculaire
UMR 1388 GenPhySE, INRA, 24 chemin de Borde Rouge – Auzeville, CS 52627, 31326 Castanet-Tolosan
Directeurs de thèse : Martine BOUISSOU-MATET et Sylvain FOISSAC.

Dans le secteur de l'élevage porcin, les truies ont été sélectionnées pendant des décennies pour leur prolificité afin de maximiser la production de viande. Cependant, cette sélection a été associée à une mortalité plus élevée des nouveau-nés. Dans ce contexte, le muscle fœtal squelettique est essentiel à la survie du porcelet, car il est nécessaire pour les fonctions motrices et la thermorégulation. Par ailleurs, la structure tridimensionnelle du génome s'est avérée jouer un rôle important dans la régulation de l'expression génique. Ainsi, dans ce projet, nous nous sommes intéressés à la conformation 3D du génome et l'expression des gènes dans les noyaux des cellules musculaires porcines à la fin de la gestation. Nous avons initialement développé une approche originale dans laquelle nous avons combiné des données transcriptomiques avec des informations de localisations nucléaires (évaluées par 3D DNA FISH) d'un sous-ensemble de gènes, afin de construire des réseaux de gènes co-exprimés. Cette étude a révélé des associations nucléaires intéressantes impliquant les gènes *IGF2*, *DLK1* et *MYH3*, et a mis en évidence un réseau de gènes interdépendants spécifiques du muscle impliqués dans le développement et la maturité du muscle fœtal. Nous avons ensuite évalué la conformation globale du génome dans les noyaux musculaires à 90 jours et à 110 jours de gestation en utilisant la méthode de capture de conformation de chromatine à haut débit (Hi-C) couplée au séquençage. Cette étude a permis d'identifier des milliers de régions génomiques présentant des différences significatives dans la conformation 3D entre les deux âges gestationnels. Fait intéressant, certaines de ces régions génomiques impliquent les régions télomériques de plusieurs chromosomes qui semblent former des clusters préférentiellement à 90 jours. Plus important, les changements observés dans la structure du génome sont associés de manière significative à des variations d'expression géniques entre le 90ème et le 110ème jour de gestation.

Mots-clés : Architecture nucléaire, muscle fœtal porcin, Hi-C, cartes de contact, 3D DNA FISH, réseau de co-expression génique.

3D genome conformation and gene expression in fetal pig muscle at late gestation.

In swine breeding industry, sows have been selected for decades on their prolificacy in order to maximize meat production. However, this selection is associated with a higher mortality of newborns. In this context, the skeletal fetal muscle is essential for the piglet's survival, as it is necessary for motor functions and thermoregulation. Besides, the three-dimensional structure of the genome has been proven to play an important role in gene expression regulation. Thus, in this project, we have focused our interest on the 3D genome conformation and gene expression in porcine muscle nuclei at late gestation. We have initially developed an original approach in which we combined transcriptome data with information of nuclear locations (assessed by 3D DNA FISH) of a subset of genes, in order to build gene co-expression networks. This study has revealed interesting nuclear associations involving *IGF2*, *DLK1* and *MYH3* genes, and highlighted a network of muscle-specific interrelated genes involved in the development and maturity of fetal muscle. Then, we assessed the global 3D genome conformation in muscle nuclei at 90 days and 110 days of gestation by using the High-throughput Chromosome Conformation Capture (Hi-C) method. This study has allowed identifying thousands of genomic regions showing significant differences in 3D conformation between the two gestational ages. Interestingly, some of these genomic regions involve the telomeric regions of several chromosomes that seem to be preferentially clustered at 90 days. More important, the observed changes in genome structure are significantly associated with variations in gene expression between the 90th and the 110th days of gestation.

Keywords: Nuclear architecture, porcine fetal muscle, Hi-C, contact maps, 3D DNA FISH, gene co-expression network.