Marta Liber

ClockOME: Searching for oscillatory genes in early vertebrate development

Faculty of Medicine and Biomedical Sciences

2021

Marta Liber

ClockOME: Searching for oscillatory genes in early vertebrate development

Master in Biomedical Sciences

Work under the supervision of:

Doctor Guilhermina Isabel dos Santos Duarte

Professor Doctor Raquel Gláucia Varzielas Pego de Andrade

Faculty of Medicine and Biomedical Sciences

2021

Marta Liber

ClockOME: Searching for oscillatory genes in early vertebrate development

Declaração de autoria de trabalho

Declaro ser o(a) autor(a) deste trabalho, que é original e inédito. Autores e trabalhos consultados estão devidamente citados no texto e constam da listagem de referências incluída.

_____

(Marta Liber)

Acknowledgments

Writing a dissertation is also a form of art. It is a process that freed my imagination but at the same time encaged my work with scepticism, and I am grateful to everyone who contributed to my personal and professional growth throughout this breath-taking process. Now I am able to realize that I learned how to prize my curiosity, how to envision and deal better with everyday life choices, how to be confident and most important, how to be a real Scientist, driven by curiosity or if one prefers by kid´s questioning of 'why?'. Therefore, my most humble and sincere thanks to:

My superb supervisor Dr. Isabel Duarte that, throughout the writing of this dissertation transmitted support and assistance. She gave me the freedom to be me and throughout this project, she shaped a better me. Supervision of an immature student is not an easy fruit and Dr. Isabel never flinched. In this process, she was more than a teacher, she was a friend, a mentor and sometimes the only one who I knew will always show me the truth.

To the superhuman Ramiro Magno, who tirelessly defied my curiosity and with a calm and serenity showed me that science is mesmerizing. I am not able to transmit in any words my gratitude for the passion he allowed me to explore.

To the kindest teacher Dr. Raquel Andrade, that was always there to enlighten me and show that I am more than who I think I am. As she once said, I am here to drink from the well of wisdom of others and build my path learning from their mistakes.

To my beloved partner who never forgot to remind me that I have to write the thesis and never judge my choices and questions.

To all my friends that in their insane ways were always there and tirelessly believed in me.

As Richard Feynman once said "If you find science boring, you are learning it from the wrong teacher" and I must say that I was never bored.

Abstract

Embryo development is a dynamic process regulated in space and time. Cells must integrate biochemical and mechanical signals to generate fully functional organisms, where oscillatory gene expression plays a key role. The embryo molecular clock (EMC) is the best known genetic oscillator active in embryo segmentation, involving genes from the Notch, FGF, and WNT pathways. However, the list of cyclic genes is still incomplete mostly due to the challenges involved with studying periodic systems. Recently, such studies have become more feasible with the development of pseudo-time ordering algorithms that search for candidate oscillatory genes using large transcriptomics datasets sampled without explicit time measurements.

This study aims at finding candidate oscillatory genes - ClockOME - active in early chick embryo development.

Two *Gallus gallus* microarray transcriptomics datasets from Presomitic mesoderm (PSM), and one dataset from limb segmentation were gathered from GEO and ArrayExpress. To normalize these data from different experiments, an RData package - FrozenChicken - was developed to apply a frozen Robust MultiArray (fRMA) normalization to the data. Next the datasets were processed with Oscope (a pseudo-time ordering algorithm) to search for candidate periodic genes clustered by similar oscillatory behaviour. The clusters of predicted oscillators were then subject to functional enrichment and interaction network analyses to highlight the biological functions associated with these genes. Oscope predicted three clusters of oscillators: two in PSM (106 and 32 genes), and one in Limb (162 genes). Overall, the genes are associated with regulatory, morphological, and developmental processes. Mesp2, a gene involved with the EMC, was found in this dataset, validating the approach, however, the majority of genes are novel oscillatory candidates, associated with chromatin and transcriptional regulation, as well as protein and oxygen metabolism. The list of candidate oscillators represents a valuable resource for guided experimental validation to discover additional members of the chick EMC. Six genes have been proposed for high-priority experimental validation: SRC, PTCH1, NOTCH2, YAP1, KDR, CTR9.

Keywords: Oscillatory gene expression | Embryo development | *Gallus gallus* | Embryo molecular clock | Transcriptomics | Pseudo-time ordering algorithm

Resumo

O desenvolvimento embrionário é um processo dinâmico que envolve alterações moleculares no espaço e no tempo. As células embrionárias são constantemente expostas a estímulos bioquímicos e mecânicos, e respondem ao ambiente em que se encontram alterando o seu programa genético. Quando corretamente integradas, estas respostas celulares culminam com o desenvolvimento bem-sucedido de um organismo funcional. Assim, a embriogénese envolve processos moleculares estritamente regulados, sendo a expressão oscilatória de genes uma das formas possíveis para a regulação do comportamento das células ao longo do tempo. O relógio molecular embrionário é um conhecido oscilador genético, e está envolvido na segmentação do tecido paraxial embrionário. O conceito de relógio molecular foi inicialmente proposto em 1976 por Cooke e Zeeman, ao qual chamaram o modelo Clock and Wavefront (Relógio e Frente de Onda)[1]. Este modelo foi concebido para descrever teoricamente a formação rítmica de sómitos em ambos os lados da mesoderme paraxial (PSM) nos vertebrados, e baseia-se na existência de osciladores genéticos que regulam esse processo de segmentação da PSM ao longo do tempo. Para além do relógio, como diz o nome, o modelo inclui a existência de uma frente de onda, que determina espacialmente o comportamento das células presentes na mesoderme pré-somítica (PSM). Assim, os dois mecanismos guiam a diferenciação das células da PSM, que consequentemente sofrem transformações genéticas que precedem a formação dos sómitos. A base deste relógio molecular consiste na expressão periódica de genes que fazem parte das vias moleculares Notch, FGF e WNT. Contudo, a lista de genes envolvidos no relógio embrionário ainda não se encontra completa, facto este que se deve principalmente às dificuldades experimentais relacionadas com o estudo de sistemas periódicos quando não se conhece de antemão a periodicidade/ritmo da expressão dos genes envolvidos.

Com o advento de novas técnicas de transcriptómica que permitem o estudo dos valores de expressão de todos os genes simultaneamente, nomeadamente usando Microarrays, ou mais recentemente através de métodos de sequenciação, como RNA-sequencing ou Single-Cell RNA-sequencing, surge a oportunidade de procurar alargar a lista de genes com expressão oscilatória. Porém, estes métodos implicam a extração do RNA das células amostradas resultando na morte celular. Assim, este processamento inviabiliza o estudo das mesmas células ao longo do tempo, originando dados moleculares estáticos, isto é, os níveis de expressão obtidos representam uma única amostra temporal. Para o estudo de processos periódicos, seria então necessário fazer uma série temporal amostrando diferentes indivíduos ao longo do tempo

de desenvolvimento, aumentando grandemente o número de amostras biológicas necessárias para resolver o ciclo de oscilação para cada gene estudado.

Assim, sem informação temporal medida explicitamente, a expressão oscilatória de genes pode apenas ser estudada usando modelos matemáticos apropriados, nomeadamente através da aplicação de algoritmos de ordenação pseudo-temporal. Estes métodos ordenam as amostras ao longo do tempo de uma oscilação de forma a obter o padrão do comportamento cíclico para todos os genes cuja expressão oscila concomitantemente. Torna-se assim possível, bioinformaticamente, inferir o potencial oscilatório de genes medidos por estas técnicas de transcriptómica, sem informação temporal explícita.

Deste modo, o objetivo deste estudo é encontrar novos genes oscilatórios, a que coletivamente chamamos ClockOME, que estão ativos durante as primeiras etapas do desenvolvimento embrionário (somitogénese) da galinha, nos tecidos da mesoderme pré-somítica (PSM), e no membro superior (Limb); tecidos estes onde o relógio molecular foi descrito, atuando como regulador temporal das alterações genéticas subjacentes.

Para tal, recolheu-se 3 conjuntos de dados (datasets) de transcriptómica obtidos por microarray de dois repositórios de dados públicos: GEO (da instituição americana NCBI) e ArrayExpress (da instituição europeia EMBL-EBI). Dois datasets continham dados de mesoderme paraxial (PSM) – tecido onde ocorre a somitogénese; e um dataset de dados de obtidos do membro superior do embrião de galinha. Com o objetivo de normalizar os três datasets de forma a torná-los comparáveis (uma vez que são oriundos de processos experimentais diferentes), foi desenvolvido um pacote de R denominado “FrozenChicken: Promoting the meta-analysis of chicken microarray data” (publicado em 2021) (https://doi.org/10.1101/2021.02.25.432894). Este pacote contém dados sumarizados de 472 datasets de microarrays de embriões de galinha, tornando possível a normalização por fRMA (frozen Robust MultiArray) de microarrays de *Gallus gallus*. Após normalização e controlo de qualidade dos valores de expressão genética, os dados da PSM e do membro foram processados com o Oscope (algoritmo de ordenação pseudo-temporal), com o propósito de prever genes oscilatórios. Este algoritmo avalia todas as combinações de pares de genes, agrupando aqueles que apresentem padrões de expressão semelhantes, ou seja, cujos valores de expressão ao longo das amostras seguem trajetórias semelhantes, indiciando um período de oscilação potencialmente semelhante. Os clusters de genes previstos pelo Oscope foram posteriormente

submetidos a uma análise de enriquecimento funcional e a uma análise de interações funcionais, com o intuito de perceber o seu potencial papel biológico, e funções moleculares subjacentes.

O Oscope reportou três listas de genes potencialmente oscilatórios: dois grupos foram encontrados a partir dos dados da PSM (com 106 e 32 genes cada) e o terceiro grupo de 162 genes foi encontrado nos dados do membro superior. No total, a lista de genes que denominamos ClockOME é composta por 296 genes potencialmente oscilatórios, envolvidos em diversos mecanismos regulatórios importantes para o desenvolvimento embrionário e para a morfogénese. A maioria dos genes presentes nesta lista não estão descritos na literatura como sendo oscilatórios (novel candidates), representando, portanto, uma mais-valia para a comunidade científica que estuda o relógio molecular embrionário. Estes genes parecem estar associados a funções como remodelação da cromatina, regulação da transcrição, metabolismo proteico e metabolismo do oxigénio, sendo, portanto, bons candidatos para futura validação experimental. Notavelmente, o Oscope identificou com sucesso o Mesp2, um gene oscilatório bem descrito na literatura, mostrando assim a validade e o potencial desta abordagem teórica.

Em suma, este trabalho produziu uma lista de 296 genes potencialmente oscilatórios. Com base na sua novidade e na função molecular anotada, foi proposta uma lista de seis genes candidatos de particular relevância para validação experimental no futuro próximo, nomeadamente: SRC, PTCH1, NOTCH2, YAP1, KDR, CTR9. Assim, as listas resultantes do trabalho desta tese poderão agora guiar futuras experiências laboratoriais capazes de adicionar novos interactores moleculares ao atual modelo do relógio molecular embrionário.

Palavras-chave: Expressão oscilatória de genes | Desenvolvimento embrionário | *Gallus gallus* | Relógio molecular embrionário | Transcriptómica | Algoritmo de ordenação pseudo-temporal

Table of contents

List of Figures

List of Tables

List of Annexes

Abbreviations

A

AER Apical Ectodermal Ridge

AP Anterior-to-Posterior

aPSM anterior Presomitic Mesoderm

B

bHLH basic Helix-Loop-Helix

BMP Bone Morphogenic Protein

BP Biological Process

C

CC Cellular Component

cDNA complementary Deoxyribonucleic Acid

Ct Collection time

D

DF Determination Front

DV Dorsal-Ventral

E

EGF Epidermal Growth Factor

EMT Epithelial-to-Mesenchymal Transition

ENI Extended Nearest Insertion

ERK Extracellular signal-Regulated Kinase

F

FE Functional Enrichment

FGF Fibroblast Growing Factor

FI Functional Interaction

FISSEQ Fluorescent In Situ Sequencing

fRMA Frozen Robust Multi-Array Average

G

GEO Gene Expression Omnibus

GO Gene Ontology

GOI Gene of Interest

GRN Gene Regulatory Network

H

HDAC Histone Deacetylase

HES Hairy and Enhancer of Split

HH Hamburger and Hamilton

HOX Homeobox

HVG Highly Variable Gene

I

IDE Integrated Development Environment

IQR Interquartile range

K

K Cluster

L

LB Limb Bud

M

MCL	Markov Clustering Algorithm
MET	Mesenchymal-to-Epithelial Transition
MF	Molecular Function
mRNA	messenger Ribonucleic Acid
MSE	Mean Squared Error

N

NB	Negative Binomial
NCBI	National Center for Biotechnology Information
ncRNA	noncoding Ribonucleic Acid
NF-kB	Nuclear factor kB
NUSE	Normalized Unscaled Standard Error

P

P	Point
PCA	Principal Component Analysis
PD	Proximal-Distal
PI3K	Phosphoinositide 3-kinases
PORD	Progressive Oscillatory Reaction Diffusion
pPSM	Posterior Presomitic Mesoderm
PS	Primitive Streak
PSM	Presomitic Mesoderm
PTM	Post-Transcriptional Modification
PZ	Progress Zone

Q

QC	Quality Control

R

RA	Retinoic Acid
REviGO	Reduce + visualize Gene Ontology
RMA	Robust Multi-Array Average
RNA-seq	RNA Sequencing
RT-qPCR	Reverse transcription quantitative polymerase chain reaction

S

S	Sample
scRNA-seq	single-cell RNA-sequencing
SPR	Sliding Polynomial Regression
STRING	Search tool for retrieval of interacting Genes/Proteins

T

TGF-β	Transforming Growth Factor beta
TI	Trajectory Inference
TS	Two-Signal

U

UMI	Unique Molecular Identifier
UV	Ultraviolet

W

WNT	Wingless

Y

YAP Yes1 Associated Transcriptional Regulator

Symbols

ε error term

η minimal phase-shift

υ phase shift residual

ψ phase-shift

Chapter I
INTRODUCTION

1 | INTRODUCTION

Section 1 | Early Vertebrate Development

Building a successful organism requires the strict coordination of several cellular processes such as cell movement, proliferation, and differentiation, in order to achieve the correct number of cells (proliferation) that will shape the intricate structures that make up the tissues and organs (morphogenesis) of the final organism. To achieve a better understanding of the morphological changes taking place during vertebrate embryo development, three animal models have been classically studied: *Gallus gallus* (chicken), *Mus musculus* (mouse), and *Danio rerio* (zebrafish). This thesis will focus mainly on the chicken embryo, as this is a classical organism used for the study of early development, mostly due to its easy accessibility (external eggs), and the shared features with human early development[2]. However, the molecular basis of development seems to be similar across vertebrate species, including humans[3].

It is undisputed that embryo development is a complex dynamical system, following a temporal sequence of morphogenic events (**Figure 1.1**)[4]. These shaping events include cell rearrangement and proliferation, tissue and cell motility, or polarization and patterning of the tissues, and are mostly coordinated by gene regulatory networks (GRNs) that may be induced by various signaling molecules. These will cause cellular responses in a concentration dependent fashion, thus promoting the patterning of the embryo tissues (reviewed in: Tam & Loebel (2007)[5]; Uriu (2016)[6]; Dequéant (2006)[7]). This signaling is driven by internal cues like cell density and messenger RNA (mRNA) concentration, as well as by external conditions like oxygen levels, leading to a defined patterning of the tissue[6,8,9].

The temporal coordination of cellular events is equally important for proper embryo development, namely cell proliferation, differentiation, and movements, as well as signalling regulation[10]. Time is, therefore, an extra variable, which is often implicit, but requires a detailed examination to allow better comprehension of such dynamic events during embryogenesis.

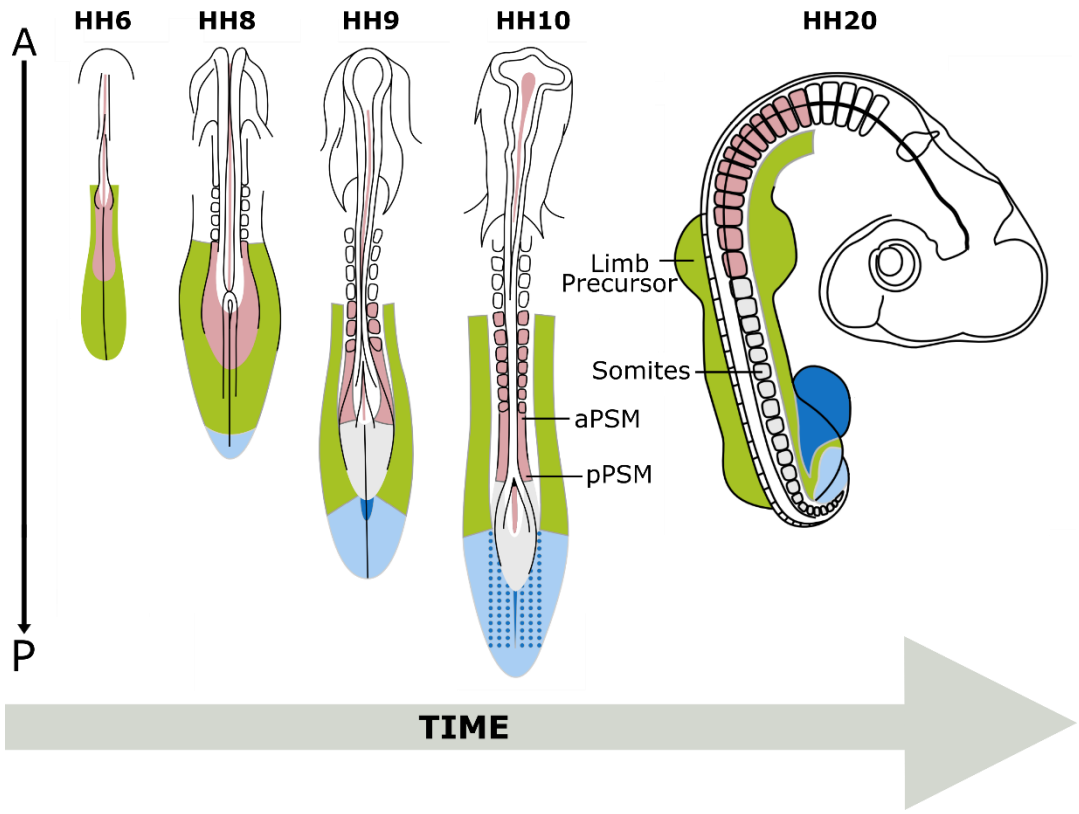

Figure 1.1 | Morphogenic events during early vertebrate development require tight spatial and temporal coordination to originate a functionally structured embryo. Schematic representation of the chicken embryo developmental timeline staged according to Hamburger and Hamilton (1951)[27], stages HH6, HH8, HH9, HH10, and HH20 (dorsal view). Shape changes are visible when comparing embryos from different stages, however, the correct organism architecture can only be achieved with proper spatiotemporal coordination between cellular and genetic processes like gene expression, cell differentiation or multiplication, cell architecture reorganization, cellular movements or tissue transformation. A - Anterior; P - Posterior. Adapted from "The Atlas of Chick Development", 3rd edition[30], Academic Press.

1.1 | Segmentation of the vertebrate body axis: a temporally regulated morphogenetic process

In order to understand the segmentation of the vertebrate embryo as a temporally regulated process, one must first describe the events taking place during these early stages of embryo development, namely Gastrulation, and Somitogenesis, as well as the underlying patterning of the embryo axis.

Gastrulation

In the avian and mammalian animal models, the formation of the embryonic body starts at the gastrulation phase, where a bilaminar disc of cells is transformed into a 3-layered embryo, through coordinated cell proliferation and movement through the primitive streak (PS)[11]. This process starts immediately after the formation of the PS.

Gastrulation: Primitive Streak

The vertebrate embryo development proceeds in a head-to-tail sequence, starting at the head (anterior part of the body), followed by the progressive formation of more posterior anatomical structures, up to the posterior limbs and/or tail[5,8]. The anterior-to-posterior (AP) body axis is established during the formation of the PS. The PS is an anatomical line of thickened epiblast that first appears at the caudal end of the embryo and extends cranially. In the avian embryo it arises from Koller's sickle (local thickening of epiblast cells) and the epiblast above it. At the anterior end of the PS, by a regional thickening of cells, forms the avian organizer, also called the Hensen's node. In the middle of the PS form an anatomical depression called primitive groove. The gastulating cells pass through the Hensen's node and the primitive groove to the internal layers of the embryo while acquiring their identity. The PS defines the major body axes: it divides the embryo along a midline, creating a base for the right-left body axis and also establishes the anterior-posterior axis since it extends in a posterior-anterior direction. Moreover, cells enter at the dorsal part of the PS and move towards its ventral side forming the dorsal-ventral separation[11].

During gastrulation, epiblast cells undergo epithelial-to-mesenchymal transition (EMT) characterized by change in the surface proteins of the cells, allowing the cell to migrate. Continuous proliferation of epiblast cells and their oriented movements towards and through the embryonic midline (i.e. the PS) originate new cellular material that will give rise to the mesoderm (mid cell layer) germ layer and endoderm (deeper cell layer) displacing the hypoblast, while ectoderm (upper layer) cells directly derive from the embryonic epiblast. Throughout the development, ectoderm tissue will originate epidermis and neural tissues, the endoderm develops into internal epithelial tissues from the respiratory and digestive system and the associated organs to the former, whereas the mesoderm gives rise to the notochord, blood vessels and somites - anatomical structures responsible for the origin of the skeleton, muscle and connective tissue of the axial part of the body, as well as the dermis of the back[11].

Gastrulation movements and cellular identity specification are guided by a key signaling family of FGF (Fibroblast Growing Factors) molecules. For example, Yang and colleagues (2002)[12] showed that FGF8 expressed in the posterior PS repeals migrating cells, while FGF4 in the anterior streak attracts cells. In parallel, FGF8 is involved in epiblast cell specification by regulating the expression of mesodermal genes like *snail*, *Brachyury*, and *Tbx6*. Spatial patterning of WNT (wingless) signaling is also important: Wnt5a in the posterior regions

promotes lateral cell migration, while Wnt3a inhibits the cellular movements in the anterior part of the embryo confining, them to become paraxial mesoderm[13,14]. However, the WNT signaling that guides the cellular movements through and across the PS are induced by FGF molecules secreted in the PS and hypoblast. Therefore, epiblast cells ingress through PS and differentiate by progressively changing their genetic profile in accordance to the FGF signalling patterns of the surrounding[8,11,15].

Somitogenesis

Early morphogenesis continues throughout the somitogenesis phase, forming the musculo-skeleton precursor elements of the embryo from paraxial presomitic mesoderm (PSM) (**Figure 1.1**). The mesoderm identity is acquired mostly due to BMP (Bone Morphogenic Proteins) gradient, with lower amounts found in the PSM[16,17]. The PSM identity is specified by the transcription of Brachyury (T), Tbx6, and Mesogenin, in addition to the *Foxc1* and *Foxc2* TF from the Forkhead (Fox) family[18–21]. At this stage, while gastrulation is still occurring at the posterior region of the embryo the somitogenesis is starting at the anterior part. New immature mesodermal progenitor cells are stably added to the posterior PSM (pPSM) through the PS[8,11]. Over time, pPSM cells become progressively positioned at the anterior part of the PSM (aPSM) and maturate along the way, due to external signaling[8,22]. Finally, the aPSM periodically segments into somites by a process called somitogenesis. Somites are transient structures, formed bilaterally, that later originate structures from the axial skeleton (vertebrae, muscles, dermis, and joints)[2]. Morphologically, each somite is a block of mesenchymal cells involved into an epithelial cell sheet. These epithelial cells derive also from PSM cells that had to undergo mesenchymal-to-epithelial transition - MET, induced by the Notch signaling molecules and regulated by *Mesp* TF expression[11,23].

Segmentation of the PSM region happens in a periodic way, synchronized between the left and right sides, where the period of formation of a somite pair is species-specific (90 minutes in the chicken embryo, 2 hours in mice and 4-6h in humans)[24–26]. Additionally, the period of the somitogenesis is fixed for a given species and the pace is maintained by the Notch signaling pathway that synchronizes the cellular behaviour[23,27]. The number of somite pairs can be used to differentiate embryonic stages like Hamburger and Hamilton (HH) stages in chicken (**Figure 1.1**)[28]. Therefore, the rhythmic pattern of the somite emergence requires a high level of temporal and spatial coordination between the cells, conceptualized in one of many theoretical models - the segmentation Clock[1], discussed in section 2. Molecular maturation of

the PSM cell is acquired by the mechanism involved in the anterior-to-posterior axis formation, which is discussed further[29]. As such, somitogenesis is an oscillatory process and will be discussed throughout this thesis.

Axis formation: Anterior-posterior patterning

Since the anterior region of the embryo (head) starts to develop first, the biochemical environment of the two opposing regions of the AP axis are different. As so, the posterior part of the embryo is maintained more immature and is characterized by the presence of FGF, WNT, Nodal and BMPs. These morphogenes establish a gradient, peaking at the posterior part of the embryo and gradually decreasing in the anterior direction, where their antagonists are expressed. Especially important for the somite PSM maturation are the gradients of Fgf8 and Wnt3a expressed in the embryo tailbud, that induce the expression of previously mentioned PSM markers (Tbx6, and Mesogenin). Later, also the Retinoic Acid (RA), synthesized by already formed somites, forms a gradient peaking anteriorly, and inhibits the FGF action.

Along the AP axis, the somites are morphologically equal but form different structures later in development. Therefore the PSM region is polarized before the somitogenesis. It is known that the FGF, and other posteriorly peaking morphogenes important during the somite formation, act through the Cdx family of TFs. These will activate specific HOX (Homeobox) genes depending on the biochemical signals of the surrounding. There are 4 HOX clusters (HOXA through HOXD), each containing up to 13 genes (Hox1 through Hox13) that are gradually expressed in order to specify the identity of the embryonic segments along the head-to-tail axis[31]. Overall, the expression of early Hox genes coordinates the identity of first somites, while later Hox genes give identity to the somites that are formed later. This colinear expression gives the segment identity along the AP axis in vertebrates.

1.2 | Segmentation of the vertebrate limb

The second embryogenesis process that is studied in this work is the anterior limb development. Limb formation begins when the somites are already formed and involves morphogen patterning of the tissue along 3 axes: proximal-distal (PD), antero-posterior (AP), and dorsal-ventral (DV). The limb bud (LB) skeletal structures (stylopod, zeugopod, and autopod) are formed sequentially from undifferentiated chondrogenic cells along the PD axis (reviewed in: Sheeba *et al.*, (2016)[32]) (**Figure 1.2**). The PD morphogenic signaling responsible for the outgrowth of the limb emerges from AER (apical ectodermal ridge), a thick region of

epithelial cells in the distal part of the LB (**Figure 1.2**). The segmentation of the limb is also controlled by opposing gradients of RA and FGF/WNT pathways, where RA induces differentiation in the proximal parts, and distal gradients of FGF and WNT synergistically promote proliferation of the LB and maintain the distal cells in an undifferentiated state (reviewed in: Sheeba *et al.*, (2016)[33]).

Therefore, the fact that both the trunk and the limb seem to be regulated by a common molecular clock with similar signaling molecules, allows the possibility of establishing parallelisms between somitogenesis and limb development, particularly regarding the temporal regulation of somitogenesis and limb patterning[32].

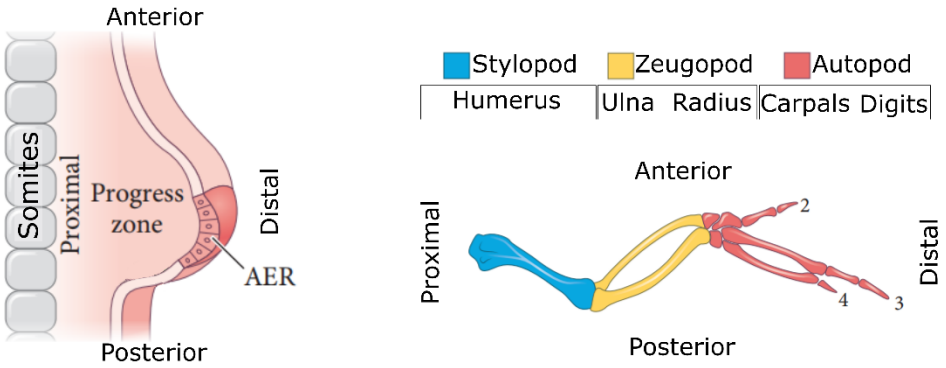

Figure 1.2 | Schematic representation of the chicken limb skeletal elements with the respective axes of development. AER = Apical ectodermal ridge; Adapted from Adapted from "Developmental Biology", 11th edition, Sinauer Associates, Inc[11]**.**

Section 2 | Oscillations in Biology

Biological architecture in both macro and microscopic systems can surprise us with distinct complexity levels, and its gradual dynamic changes range from rectilinear movements to apparently random behaviors. Particularly relevant for the study of biological systems, are oscillatory processes, i.e. sequences of repetitive molecular and/or cellular events periodically distributed in time. A system that generates these periodic variations is called an oscillator[34,35].

Biological systems present several examples of oscillatory mechanisms, such as the circadian rhythm, the cell cycle, cardiac beating, somitogenesis, as well as regulatory mechanisms like immune system activation in response to stressors or progenitor cell maintenance, that can be regulated in a cyclic manner[15,24,36–41]. Phenomena that proceed in a cycle (period) can be detected in different cell types allowing the organism to react to molecular clues in a timely fashion (reviewed in: Beta & Kruse (2017)[35]; Levine & Elowitz (2013)[37]).

Living cells are constantly presented with diverse environmental cues and to survive they have to be able to detect, decode, and generate an appropriate response to that environment. The signals in biology can be external to the organism, such as light, physical forces, temperature of the surrounding environment; or at the molecular level, the extracellular concentration of signaling molecules or even mechanical forces exerted on the cellular membrane[35,36]. Decoding those signals leads to activation of multiple pathways related to genetic regulation[37]. However, biological processes require a specific order of events and changes with an adequate pace/speed. With the absence of an external clock to measure time progression, the signaling can be internal to the system i.e. the oscillator is autonomous[4,23,35]. Therefore, at the molecular level, cells can host autonomous oscillators that, for example, in response to *stimuli* originate a rhythmic activation and deactivation of the molecules involved in the signal decoding[23].

An oscillatory response can be achieved by sequential increase and decrease in the concentration of particular mRNAs or proteins, changes in their localization and/or post-transcriptional modifications (PTMs), among others[37]. The kinetics of molecules that are activated or *de novo* synthesized in response to a signal is orchestrated by an ordered sequence of changes in the GRNs (reviewed in: Uriu (2016)[6]; Levine & Elowitz (2013)[37]).

Such oscillations in the levels of gene products can be achieved by a feedback mechanism in the mRNA transcription/translation cycle, collectively forming a genetic oscillator(reviewed in: Uriu (2016)[6]). As such, different phases of oscillation mean different mRNA levels that, in consequence, will trigger phase-specific genetic alterations inside each cell[6,34,35]. To illustrate this concept of cycles and genetic oscillators, one can imagine the time elapsed as a portion of an abstract cycle, corresponding to a particular mRNA level, therefore linking time with mRNA amount (**Figure 1.3**).

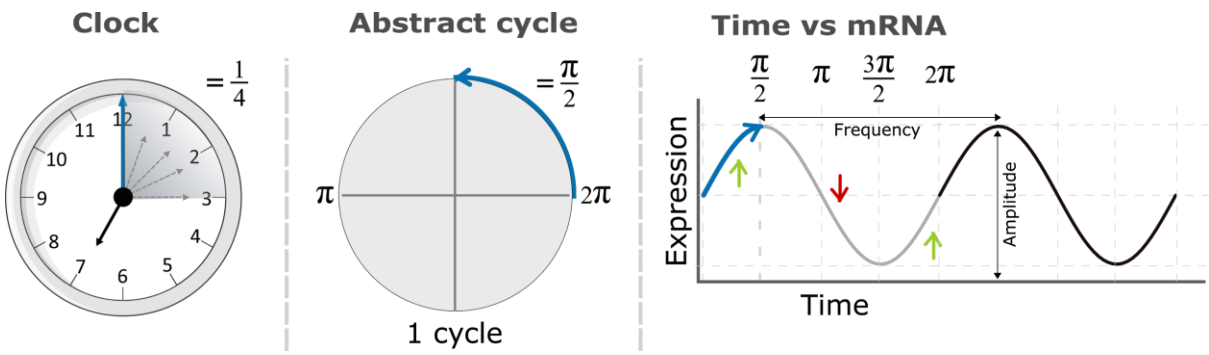

Figure 1.3 | Cycles and genetic oscillators. The minute's arm in a clock (blue arrow) illustrates an oscillatory process that repeats itself every 60 minutes (its period), corresponding to one full cycle. Continued on next page.

Figure 1.3 | Continued. Accordingly, 15 minutes in a clock means that a quarter of a full cycle was completed. Translating this information into a trigonometric circle (Abstract cycle), in that amount of time the cycle completed $\boldsymbol{\pi/2}$ of the full oscillation.When projecting this cycle into an *xy* axis system representing a toy genetic oscillator, where time is on the *x* axis and the levels of mRNA are on the *y* axis, we can follow one cycle of such a sinusoidal oscillation, from 0 to $\boldsymbol{2\pi}$. In this theoretical example, in the first quarter of the period of oscillation, the amount of mRNA product increases and reaches its maximum level.

Oscillatory responses to a biological signal can be found at different time scales in different systems (seconds, minutes, hours, days, months or even more). These oscillations depend on several variables including, the intensity and the duration (constitutive, pulsing) of the stimulus, the signal type (electrical, chemical, mechanical), among others. Some examples have been described in biology where the influence of these variables was detailed. For example, the constitutive presence of extracellular stimulus is required to induce oscillations in downstream pathways of ERK (extracellular signal-regulated kinase) in human epithelial cells and nuclear factor kB (NF-kB) in mouse fibroblasts[39,42]. ERK activation levels also depend on the extracellular concentration of epidermal growth factor (EGF), where increasing concentrations of EGF increased the frequency of activated ERK pulses. However oscillations were absent when cell density increases or the signal overflows (high concentration of EGF), in human mammary cell lines[39,40,42]. Similarly, in yeast, the transcription factor (TF) Crz1 oscillated in response to extracellular calcium, where the frequency of the pulsing is modulated by the concentration of calcium[43].

The same molecular pathway is able to drive multiple responses. This allows the cell to alternate between cellular states using the same set of molecules, preventing energy wasting by limiting cell response. As described in Cai & Elowitz (2008)[44], neuronal fate of mouse cells was induced by the continuous expression of Ascl1 gene, however, when its expression oscillates, the cells maintain its neuronal progenitor state possibly by interrupted expression of multiple fate-determining genes. Another interesting example is the mammalian tumor suppressor p53 expression: *p53* activation generates a sustained pulsatile feedback in response to Gamma irradiation that results in activation of a genetic programme leading to cell survival; whereas only one impulse is originated when cells are exposed to UV (ultraviolet) radiation, promoting apoptosis (programmed cell death)[38]. These evidences suggest that oscillatory mechanisms may work like bimodal shifts in several biological events, allowing the cell to choose between different genetic programmes in response to the surrounding environment and time progression, while also expanding its genetic capabilities, since one signaling input may generate various downstream effects. Additional examples and hypotheses of pulse-generating systems are described in Sonnen & Aulehla (2015)[15] and Levine & Elowitz (2013)[37].

Oscillatory dynamics work not only as a timekeeping mechanism to ensure an ordered sequence and pace of time-delimited events, but can also provide spatial coordination to the surrounding cells (reviewed in: Webb & Oates (2016)[10]; Beta & Kruse (2017)[35]). To obtain a coherent behaviour among a population of genetically oscillating cells, they must become synchronized which can be achieved by communications between neighbouring cells (coupling). In oscillatory networks, coupling leads to adjustments and synchronization of frequencies between each individual cellular oscillator (reviewed in: Maroto *et al*., (2005)[23]; Beta & Kruse (2017)[35]). In consequence, coupling reduces the noise and variability between individual oscillators allowing the tissue to behave synchronously (reviewed in: Ebisuya & Briscoe (2018)[4]). When coupled in space, intracellular oscillators can originate wave-like patterns of gene expression (through periodic pulses of gene expression), that will induce changes in the genetic system of the cells in accordance with their position in the tissue[34,35].

2.1 | Oscillations in somitogenesis - Clock and wavefront model

The rhythmic nature of segmentation of the PSM in an anteroposterior direction in the vertebrate embryo - somitogenesis - is an important example of the morphological result of a genetic oscillatory process. The periodic bilateral segmentation of the PSM region into somites must follow a particular order of events with the correct tempo, where the period of segmentation is species-specific: for example, 90 minutes in the chicken embryo, 2 hours in mice, and approximately 5 to 6 hours in humans[24–26]. As a result, to form somites in the proper time and space, the underlying coordination of the PSM cells must be precise[5,8]. To describe such periodic segmentation in early vertebrate development, a theoretical model was put-forward by Cooke and Zeeman in 1976 - the "Clock and wavefront" model[1]. This model comprises two interconnected components: (i) the clock, composed by the genetic oscillators that give the temporal information, that means, whether a cell is in optimal conditions/genetically primed to start a morphogenetic transformation into somites; and (ii) the wavefront, that defines the positional information of where the cells are located relative to the full length of the embryo, indicating the location for the boundary of each somite pair. The interaction of both mechanisms drives PSM maturation and gives rise to a group of genetically competent cells, in the aPSM, able to undergo the transformation into somites, when the underlying network of genetic alterations is completed (reviewed in: Hubaud & Pourquié (2014)[45]).

The clock refers to autonomous oscillators found in pPSM cells. Experimentally it was found, in chicken embryos, that these genetic oscillators are originated by negative feedback loops in HES (hairy and enhancer of split) family of bHLH (basic helix-loop-helix) transcription factors, that repress their own transcription and act as Notch effectors[46]. Regular variations in the mRNA levels of avian *HAIRY1* gene in the PSM, correlates with the rhythmic generation of somites, suggesting that it may be establishing the pace of somitogenesis[24]. Several additional genes have since been reported to present cyclic expression in the PSM of multiple organisms (reviewed in: Bailey & Dale (2015)[47]). These are involved in the Notch, FGF, and WNT signaling pathways, however the specific players vary amongst different animal models[3,47,48]. More recently, oscillating genes associated with additional new pathways were reported in the human segmentation clock: BPMs; TGF-β (transforming growth factor beta); PI3K (Phosphoinositide 3-kinases); ephrin; HDAC (histone deacetylase); Hippo[26,48].

It is important to note that initially the individual oscillators are not synchronous, as experimentally it was shown that isolated PSM cells lack synchronization in the genetic circuitry, which becomes disorganized[23]. By cell-cell communication between neighboring cells, a synchronizing wave of gene expression sweeps the PSM, coupling the genetic oscillators. This is acquired by the interaction of membrane proteins, Delta and Notch, that induces the transcription of HES proteins in the target cell and shows to be essential and sufficient to drive the cell synchronization[23,49]. The oscillations in the Notch, FGF, and WNT are coupled across the PSM, where genes pertaining to the WNT pathway cycle in antiphase with cycling genes from Notch and FGF pathways, suggesting mechanisms of mutual inhibition between the genetic oscillators[7,50]. In consequence of synchronization, the clock signaling drives a set of coordinated genetic alterations in the PSM cells, that are specific to the clock phase, indicating when the cells are able to transform into somites (reviewed in: Hubaud & Pourquié (2014)[45]). As the embryo elongates, opposing gradients of morphogenetic pathways (RA peaking anteriorly and FGF/WNT posteriorly), create a threshold of signaling called the determination front (DF) in the aPSM, where the segmentation occurs (**Figure 1.4**). Already formed somites synthesize the RA that induces cell differentiation of the embryo anteriorly, and is not present in the posterior immature PSM tissue. The gradients of FGF and WNT pathway-related effectors are established across the pPSM preserving its highly motile immature status in a concentration dependent manner. A wave of gradual genetic alterations is formed that will create a spatial environment in the wavefront, where cells are now able to respond to clock signals (are competent) and can transform into somites[45,51]. The morphogenetic gradients are

extremely controlled, since interference in the concentration of any of the molecules lead to anomalies in the number, size or shape of the somites[52]. In summary, when a group of competent PSM cells, at the same phase of the oscillation cycle, reaches the DF, they undergo mesenchymal-to-epithelial transition (MET), and bud-off, forming a bilateral pair of somites (reviewed in: Hubaud & Pourquié (2014)[45]) (**Figure 1.4**). This is marked by the expression of the gene *Mesp2*, a TF required to initiate the formation of segmental borders, by repressing the Notch signaling[53,54].

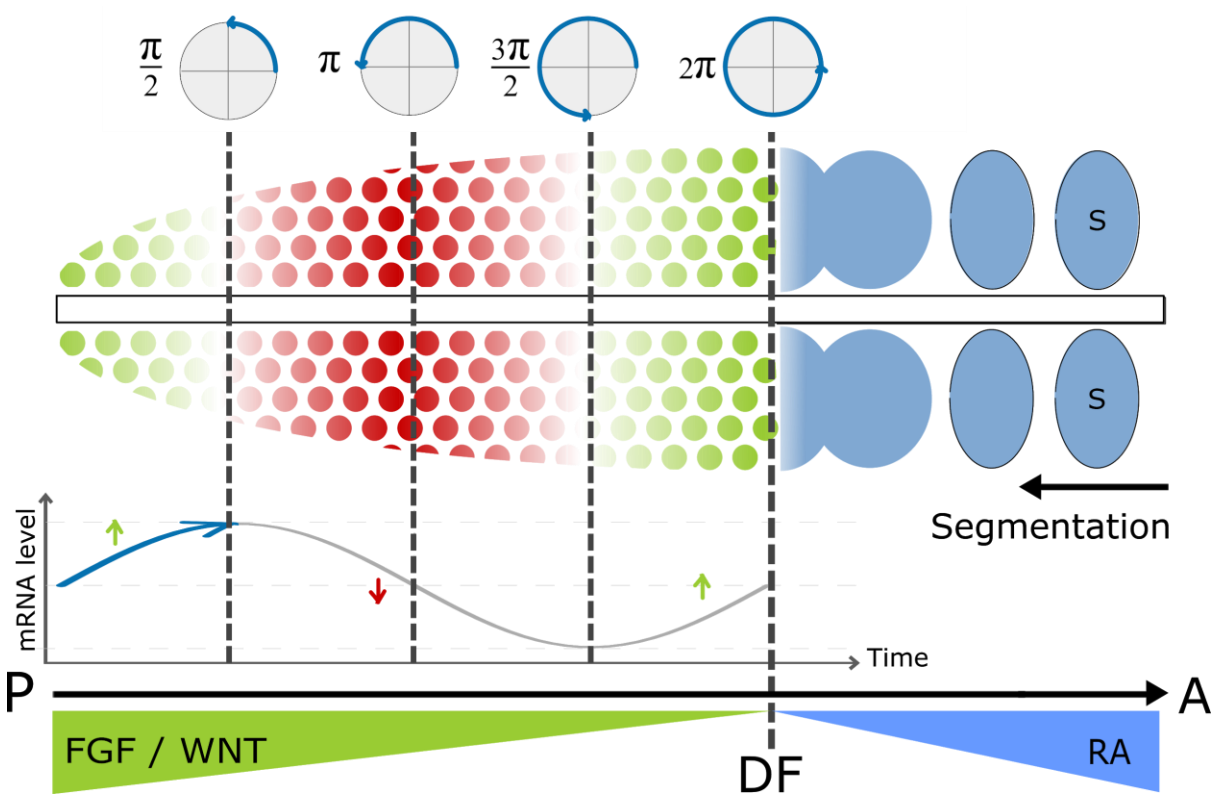

Figure 1.4 | Clock and wavefront pattern of gene expression in the PSM: temporal information translated into spatial cues. In the vertebrate embryo PSM, genetic oscillators in individual cells will locally synchronize through coupling via the Notch-Delta signaling pathway. As a result, the tissue differentiates in a specific place in response to a genetic programme given by the period of the oscillation and the extracellular morphogen signaling. The mRNA level wave displayed below, shows the oscillatory behaviour of the clock molecules. Morphogens like FGF, WNT and Retinoic Acid (RA) are expressed in opposite directions to create the determination front (DF), i.e. the place where the PSM segmentation (somite formation) will happen. A = anterior; P = posterior; S = somites; Green circles represent cells with increasing mRNA levels of the genetic oscillator; Red circles represent cells with decreasing mRNA levels of the same genetic oscillator shown in green.

Theoretically, the Clock and Wavefront model allows to infer the somite length (*S*) and total number (n). The *S* parameter is estimated by combining the clock period (*T*) and the wavefront velocity (*v*) in a way that $S = vT$. The total number of somites (n) can be stipulated by the period, taking into account the total duration of the segmentation process (d) in the system under study, i.e. $n = d/T$. Therefore, as reviewed in Pais-de-Azevedo *et al.*, (2018)[55], the

Cooke and Zeeman (1976)[1] model "places the period of the clock in a central role for determining the length and overall anatomy of the embryo somites".

2.2 | Oscillations in somitogenesis - Alternative models

Besides the Clock and Wavefront model, other alternatives have been proposed to describe the periodic formation of somites in the vertebrate embryo (reviewed in Pais-de-Azevedo *et al.,* (2018)[55]). Briefly, there are at least three major mechanisms that can also translate genetic oscillations into spatial tissue patterning: (i) A Turing-Hopf mechanism[56]; (ii) an oscillator phase-gradient model[57]; and (iii) the Progressive Oscillatory Reaction Diffusion (PORD) model which is also based on a Turing mechanism[58]. In general, contrary to the clock and wavefront model (where the wavefront is a causal agent for segmentation), in these alternative mechanisms, the wavefront is an emergent property of the system[55].

Experimentally it was possible to form ectopic somites without the contribution of oscillatory gene regulation, a.k.a. a genetic clock[59,60]. In consequence, the segmentation clock has been proposed to contribute to the timing of PSM segmentation and to prime the somite cell for further specialization, whereas the shape and size of the segments were controlled by the underlying cellular communication Notch pathway[59]. Furthermore, Hubaud et al. (2017) demonstrated that PSM cells, in culture, can transition from a non-oscillatory to an oscillatory state in response to cell density, and Notch and YAP (Yes1 Associated Transcriptional Regulator - downstream effector of Hippo pathway) signaling thresholds[49]. This argues that the segmentation clock is an excitable system with dynamic behaviour that cannot be satisfied by current models, instead of an autonomous molecular clock. Moreover, the authors propose that the traveling waves and the genetic oscillations are both emergent properties originated by the collective behaviour of a population of cells[49].

As a result of the alternative hypotheses available to explain oscillatory phenomena in early vertebrate development, further experimental research is imperative to elucidate which of the distinct models provides the most insight into the biological mechanisms underlying the periodic segmentation of the vertebrate embryo.

2.3 | Oscillations in limb patterning

Similarly to somitogenesis, the development of the limb requires tight control of the cellular and molecular events that shape the bone elements of the limb.

Different models intend to describe the segmentation of the limb across the PD axis: (i) the Progress Zone (PZ) model that, similarly to the clock and wavefront model, proposes that the progenitor cells have an internal timer[61]; (ii) the Two-Signal (TS) model that attributes the segmentation of limb to opposing gradients of RA and signaling from AER; and (iii) the Integrated Space-Time model that conciliates the PZ and TS models to pattern the limb tissue[62].

Briefly, the PZ model involves an internal cell-autonomous clock in the progenitor cells of the LB, distally located in the so-called progress zone. That area is under the influence of AER signaling and conditions the cells to respond in a time-dependent manner, i.e. promotes distal fate in proportion to the length of time the cells remain in the PZ by driving time specific genetic regulation. As a result, when the cell is no longer under the AER influence due to limb elongation (no signal incoming), their PD fate becomes fixed and the progenitor cells will follow the determined specification process (reviewed in: Sheeba *et al*., (2016)[33]). Such internal timer, in chicken embryos, was proposed to be the cyclically emerging *HAIRY2*, a bHLH TF, generated with a periodicity of 6 hours in stages HH20-28[63]. Likewise, each limb segment, with a developing time of 12h, requires 2 cycles of *HAIRY2* induced genetic regulations. In agreement with the segmentation clock, the synchronization of cell autonomous oscillators in the limb molecular clock is managed by Notch signalling, possibly by the pair of Serrate1 ligand and Notch2 receptor[32].

In the Two-Signal model the 3 bone structures of the limb are coded by the morphogene-established domains, therefore providing a molecule-based framework only. That is, with the embryo elongation, the most proximal region is under the influence of RA originating the stylopod, the distal region of LB is signaled by FGF driving the formation of the autopod, whereas the median part of the LB originates the zeugopod resultant from the combination of both morphogenic signals[64].

The more recent Integrated Space-Time model links the spatial morphogenic patterning (TS model) with the temporal precision of Hairy2 oscillations (PZ model)[62]. Sheeba *et al*., (2014) suggest that the antagonic RA and FGF gradients are spatially combined in the initial limb, promoting the constitutive expression of Hairy2, thus specifying the stylopode[62]. Over time, when the distal and proximal parts of the LB are distanced, due to tissue outgrowth, the previously mentioned morphogenetic gradients are separated in space. In consequence, a proper biochemical environment is created that allows the *HAIRY2* gene to systematically oscillate. From then on, LB progenitor cells are specified into distinct PD limb structures (zeugopod and

autopod) conditioned by the number of *HAIRY2* expression cycles experienced in the progress zone[62]. Further studies also support the importance of combining the morphogen levels and time measurements involved in the regulation of limb development[65].

Overall, additional studies are required to elucidate the details regarding the GRNs responsible for the modulation of the segmentation events taking place during limb formation. Moreover, the causality between the dynamics of limb structure formation and the periodicity of the limb molecular clock are also in need of further experimental clarification.

In general, there is an overwhelming amount of experimental evidence that shows that oscillations in gene expression coordinate and modulate the pace of progression through the developmental events, which are required for the proper organization and patterning of the embryo tissues, as well as its correct shaping (reviewed in: Ebisuya & Briscoe (2018)[4]). However, the molecular basis of genetic oscillations in development still presents a challenge as it is necessary to examine both the intracellular genetic networks and intercellular dynamics that involve, not only the movement of the cells, but also their proliferation and differentiation. For that, the knowledge of the molecular participants of this GRN, i.e., the set of genes that present oscillatory gene expression in each tissue, is essential.

Section 3 | Techniques used to study genetic oscillations

In order to properly describe a dynamical system, appropriate theoretical models must be applied to experimentally gathered data [6]. The study of patterns of oscillatory gene expression requires the usage of experimental methods that quantify gene expression levels. When a gene is transcribed to RNA, it originates a transcript that can be either coding (mRNA) or noncoding (ncRNA). Coding RNA will be translated to protein, and noncoding RNA will be further processed into ribosomal RNA, transfer RNA, short ncRNA, or long ncRNA (among others). Both proteins and ncRNAs can be regarded as the functional molecules that execute and regulate core cellular processes. Accordingly, measuring the expression values of a certain gene can be viewed as the most simple proxy measurement for the amount of its functionally active molecule present in the cell[66].

Cell biology relies largely on reproducible experimental observations. The ideal experimental technique to study gene expression oscillations would be a high throughput

analysis of gene expression values over time at the individual cell level. Although there are some recent technical advances moving toward this goal (e.g. FISSEQ: Fluorescent In Situ Sequencing [67], that allows live gene expression profiling in intact cells and tissues (albeit not in living); and more recently Live-seq that performs single-cell transcriptome profiling preserves cell viability[68]), the continuous expression profiling of all genes is not yet available. Currently available techniques (e.g. Microscopy, PCR, Microarrays, Sequencing) detailed in **Figure 1.5**, require experimental designs where scientists must choose between measuring continuously (across time) the expression of only a few genes, or increasing the sample size by measuring many genes simultaneously, while losing the temporal resolution. Consequently, the experimental design used to study oscillations must reflect the requirements of the hypothesis under study.

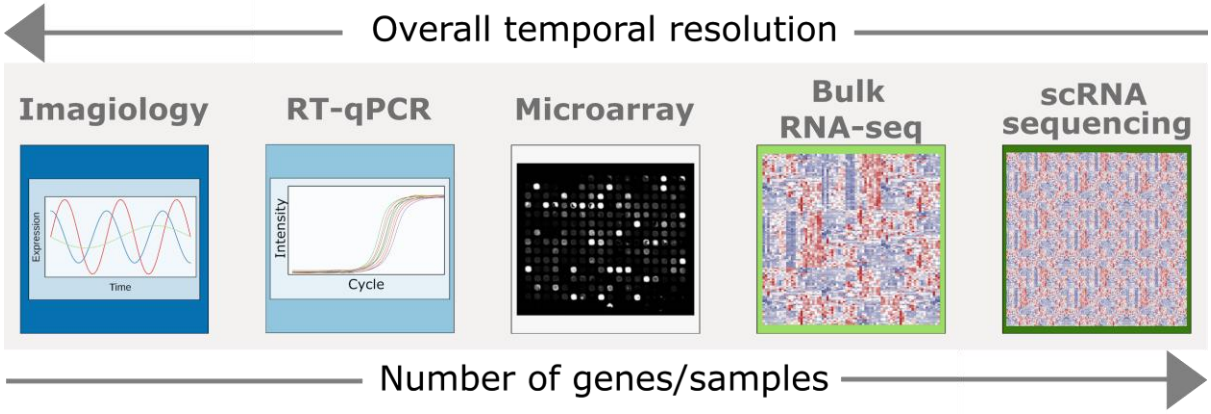

Figure 1.5 | Tradeoff between number of genes profilled and feasibility of temporal resolution. Low throughput techniques such as Imagiology allow experimental designs with high temporal resolution (the expression of a small set of genes can be followed continuously in time). Other techniques that require extraction of the molecules of interest (such as RT-qPCR, Microarray chips, and RNA-sequencing), allow the measurement of many genes simultaneously, increasing the throughput, but requiring experimental designs that use discrete time points per condition. This leads to a decrease in the temporal resolution of the experiment, since less time points are sampled, while increasing the overall cost of the experimental design.

3.1 | Imagiology approaches to gene expression dynamics

A particularly useful technique for the continuous monitoring of gene expression dynamics is direct measurement using imagiology techniques. This is possible by inserting a reporter gene (e.g. fluorescent) into the cell, which will be transcribed under the control of the endogenous promoter of the gene of interest (GOI). Thus, when the GOI is expressed, the reporter protein emits fluorescent light of a known wavelength, which can then be detected in real-time under a microscope. The fluorescence emitted will be proportional to the expression of the GOI, yielding continuous quantitative gene expression measurements (**Figure 1.5**

Imagiology). This procedure can be expanded to measure more genes simultaneously by using several fluorescent reporters, each emitting at a different wavelength. However, this approach is restricted to only a small set of genes due to technical limitations (such as spatial constraints of the cell)[69].

3.2 | RT-qPCR assessment of gene expression oscillations

Reverse transcription quantitative (real-time) polymerase chain reaction (RT-qPCR for short) is a laboratory technique that is able to specifically amplify complementary DNA (cDNA) sequences of interest, and quantify changes in their total amount (**Figure 1.5** RT-qPCR). This is a medium-throughput approach that can be applied to profile tens to a few hundreds of genes simultaneously, allowing for example differential gene expression analysis, as well as a variety of diagnostic uses (e.g. cancer, rare diseases).

For RT-qPCR, RNA is extracted from biological samples, reverse transcribed to yield cDNA, and then amplified using gene-specific primers that hybridize with the DNA sequence of interest to initiate the amplification process via a polymerase chain reaction. Fluorescent dyes (or sequence-specific probes, e.g. Taqman) are included in the reaction mix that will emit light proportionally to the amount of double stranded DNA produced in each amplification cycle. Since this is a hybridization-based technique, prior knowledge of each GOI sequence is necessary to design the primers. Therefore, the detection of novel transcripts with this technique is very limited. Gene expression data is obtained by quantifying the fluorescent signal followed by mathematical processing. The outcome of qPCR is the amount of mRNA of each GOI relative to the reference genes (whose expression is measured in all samples, and must be stable in all experimental conditions tested)[70]. Contrary to imaging, this technique does not allow the monitoring of a live sample over time. Therefore, it requires the sampling of timepoints, which lowers the time resolution of the experiments. Conversely, it has higher throughput, allowing the simultaneous quantification of up to a few hundred known genes.

3.3 | Microarray Chip gene expression quantification

Large-scale gene expression profiling was made possible by the development of DNA microarray assays. This is a hybridization-based technique used to simultaneously measure the expression values of thousands of genes in a biological sample, hence being a high-throughput technique. Microarray methods can be used for disease diagnosis, pharmaceutical studies, to discover differentially expressed genes between different conditions, among others[71].

Short DNA sequences (oligomers) complementary to known genes are attached to the glass surface of the microarray chip in a grid-like manner. In the most commonly used (commercially available) microarray platform (Affymetrix GeneChip™), each gene is represented by a set of 11 oligomers of 25 nucleotides each, called probes, that are scattered through the array to minimize spatial grid effects. Each array is composed of tens of thousands of probe sets. Extracted RNA molecules from the biological sample are reverse transcribed into cDNA, fluorescently labeled and introduced to the array plate allowing its hybridization with any complementary probe present. Since this technique is based on hybridization, it is imperative that the genome sequence of the target genes is known. Complementary base pairing between the probe and the sample nucleotide sequences emit fluorescent light. The fluorescent signal is then measured by a high-resolution scanner, and the amount of emitted signal *per* location is recorded (**Figure 1.5** Microarray)[72].

The signals from a microarray chip are not direct measurements of gene expression levels, since each gene is represented by a set of probes. Thus, proper bioinformatics analyses are necessary to produce the final output in a tabular format that contains the comparable relative values of gene expression per array. Briefly, three major steps are required: (i) feature extraction; (ii) quality control; and (iii) normalization. Feature extraction (usually performed by software provided by the microarray manufacturer) converts the fluorescence image scanned from the microarray into quantifiable values. It involves image processing with noise reduction and background removal, while also annotating the data with the gene IDs and sample names. For the Affymetrix GeneChip™ platform, it yields these data in .CEL format, which represents the raw data files available from such microarray chip studies. After feature extraction, microarray data analysis software packages can be used to make diagnostic plots (e.g. background signal, average intensity values, Principal Component Analysis (PCA), Normalized Unscaled Standard Error (NUSE) Plots, MA plots, RNA degradation plots, among others) to help identify possible physical artifacts, and problematic arrays, reporters or samples. Finally, microarray data normalization between different arrays from the same experiment is used to control for technical variation between assays, while preserving the biological variation[73,74].

One of the most widely used methods to normalise these data is the 'Robust Multi-Array Average' (RMA) method which entails the following steps: (i) probe specific correction of the perfect-match probes using a model based on observed intensity (which is the sum of signal and noise); (ii) Normalization of corrected Perfect Match probes using quantile normalization;

and (iii) summarization of the perfect matching probes using a median polish to calculate a single expression measurement per gene[75].

3.4 | Gene expression quantification using RNA sequencing (RNA-Seq)

Whole transcriptome sequencing-based techniques such as RNA Sequencing (RNA-seq) (**Figure 1.5** Bulk RNA-seq), and its derivative single-cell RNA-seq (scRNA-seq) (**Figure 1.5** scRNA sequencing), allow for the simultaneous expression profiling of all genes from the biological sample. These sequencing methods provide a comprehensive view of the transcriptome of the samples, and another advantage is the fact that they do not require prior knowledge about the primary sequence of the genes being measured, hence giving way to the discovery of novel transcripts or structural variations of the same protein coding gene, as well as non-coding genes like microRNAs or long non coding RNAs. The experimental protocol for both methods requires the mRNA isolation from the samples, followed by random primer amplification, originating libraries that hold the transcriptome of each sample. The libraries will then be sequenced and processed bioinformatically[66].

For bulk RNA-seq, during the library preparation, the total mRNA from pooled cells is used to construct the library that will be sequenced, meaning that the measured gene expression will be an average value from the whole cell population sampled. Accordingly, rhythmic signals potentially present in individual cells may cancel each other out in the resultant average measurements (**Figure 1.6**). This presents a major disadvantage, which is also applicable to RT-qPCR and Microarray studies, both of which require the extraction of RNA from large populations of cells (since the RNA extracted from a single cell would generate a signal well below the detection limit of these techniques)[66].

Conversely, scRNA-seq library preparation includes additional steps that isolate and individually label the RNA from each cell prior to sequencing. mRNA fragments are barcoded with Unique Molecular Identifiers (UMIs), i.e., short nucleotide sequences that identify the provenience of the transcript during the bioinformatics analysis, and assign it to its respective cell. Therefore, each library characterizes an individual cell's transcriptome. In this way, the scRNA-seq method could detect precise representations of gene expression levels in individual cells, i.e. not averaged for the whole population of cells sampled, thus increasing the probability of detecting oscillations in gene expression levels that would be masked if using any other bulk RNA expression analysis (**Figure 1.6**)[76,77].

Despite the clear advantages of these high-throughput approaches, the downside is that they are limited to the analysis of samples collected at discrete time points, hindering gene expression analysis over time.

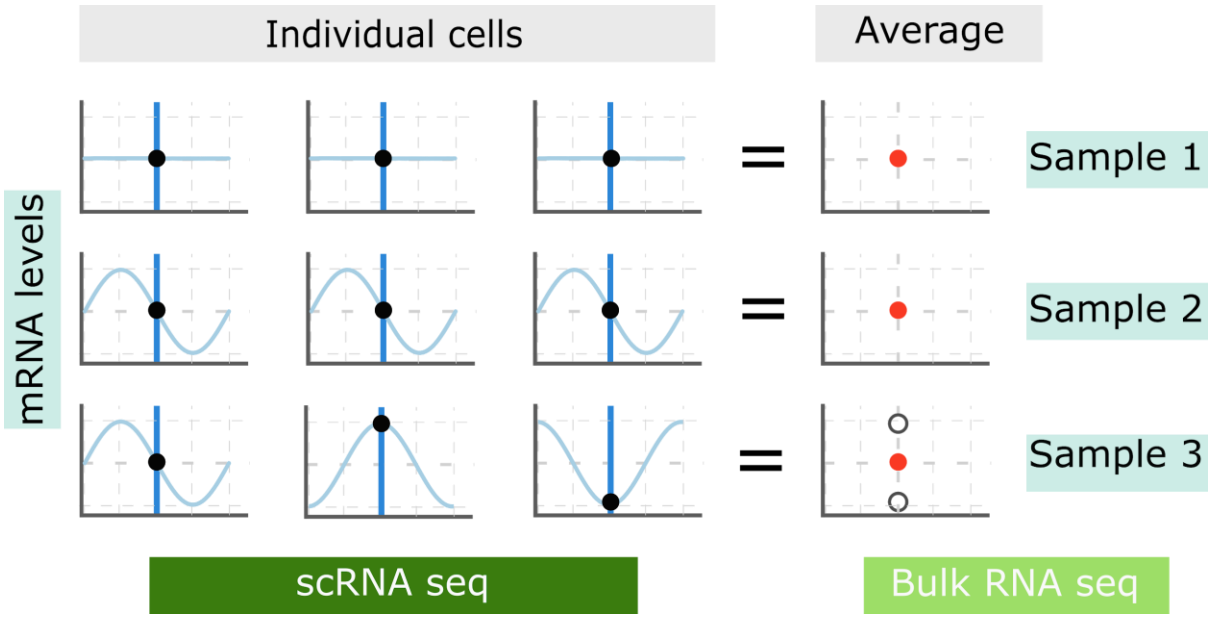

Figure 1.6 | Detection of gene expression oscillations in scRNA-seq versus bulk RNA-seq data. Levels of mRNA are constant in sample 1. In samples, 2 and 3, the gene expression oscillates at the single cell level. However, at the moment of sample collection (vertical blue line) the gene expression values from each cell in sample 3 are at different levels of the oscillation, whereas in sample 2 the cells are synchronized, and the mRNA levels are equal, in consequence of the synchronization. If bulk RNA-seq is performed, the expression value from all cells contained in each sample will be averaged to a single value (red dot), yielding the same output for all three situations, missing the information regarding the mRNA dynamics. Conversely, if measured at the single cell level, each cell accounts for its individual expression value (black dot) allowing the capturing of the variability in the expression patterns of the gene of interest.

Section 4 | How to infer time from static data

With the purpose of analyzing temporal changes in the GRNs underlying early embryo development, it is necessary to discuss the challenges of recovering dynamic behaviour from static data gathered by high-throughput techniques. The output of all sequencing methods is a "snapshot" representation of the cellular inner dynamics at the time of collection. The raw output comes in a gene-by-sample numerical matrix, with genes as rows and columns as samples, where in the case of scRNA-seq, columns correspond to individual cells. As a result, each sample represents the state of that gene at that specific moment of collection time (Ct).

In a tissue, gradual transcriptional changes occur in all cells allowing them to progress through different cellular stages and diverge into multiple cell fates, whose trajectories can be

indexed in a pseudotime ordering, i.e. an imaginary axis of progression through time. This ordering provides a surrogate measure of the cells' progression in time, which can be quantified. However these changes do not need to start at the same time in every cell. For example, even if a cluster of cells is running the same genetic cascade of events, each individual cell may execute it over a varying time scale. Thus, a population of cells is not synchronous by default, and in the context of biological oscillations, a cyclic gene can be detected at different points of the oscillation, even if the population of cells is collected at the same time[78] (**Figure 1.7**). Moreover, given time-static data, it is not possible to infer if the values of gene expression were measured at the ascending or descending phase of the oscillation. Consequently, the expression values, in the output matrix, comprehend a mixture of oscillatory states[78,79].

With the purpose of computationally reconstructing the pseudo-order of genetic events, a key assumption is that genes do not change drastically their expression patterns between the neighbouring cells and thus, samples holding similar transcriptional profiles are close in order[76]. Yet two genes oscillating with the same frequency that are phase shifted may have little similarity in the expression profile reconstructed computationally (**Figure 1.7**). As a result, during the study of dynamic gene expression, it is necessary to use a set of mathematical tools and assumptions to account for the asynchrony between the cells, and to recover oscillatory behaviour from a static set of measurements that do not contain any temporal resolution[78,79].

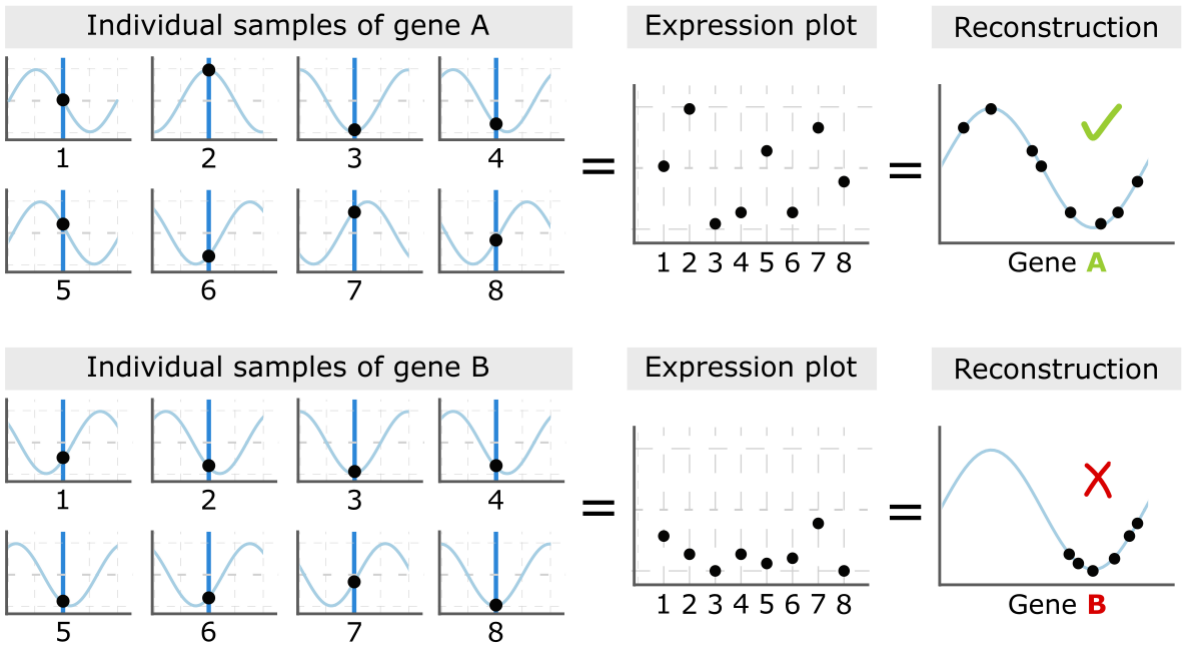

Figure 1.7 | Sample collection time limits the representation of oscillatory systems. Individual samples (black dots) collected at the same Ct (dark blue line), may be found at different oscillation times i.e., they started to pulse at different moments of time and thus, are phase-shifted. Continued on next page.

Figure 1.7 | Continued. Expression values from each sample, on the *y* axis, are plotted in a sample-expression scatter plot. The sampling order does not recapitulate the oscillation wave. As so, by mathematical modeling it is possible to reorder the expression values of a gene A in pseudotime to recover one oscillation cycle (blue sine wave). Gene B was only detected at the downphase of the oscillation and despite being a cyclic gene, it is not possible to reconstruct its oscillatory trajectory by only looking at values from individual samples, since those lead to a steadier pattern. Gene A and Gene B oscillate with the same frequency however when compared, their computationally reconstructed trajectories are not related, due to technical reasons (detection rate; bulk averaging; sample variability, etc). Ct = collection time.

Genetic information cannot be retrieved continuously since cells are lysed to extract the nucleic acids. One possible approach to deal with this technical challenge is to gather the gene expression values of equivalent samples in multiple consecutive timepoints, representing a time-series (**Figure 1.8**). With this approach, dynamic changes within the developing system can be interpreted as the progression from one state of the system to the following state. In this case genetic trajectories can be inferred with a greater precision, by comparing the data collected previously to the data in the following time of collection[78,80].

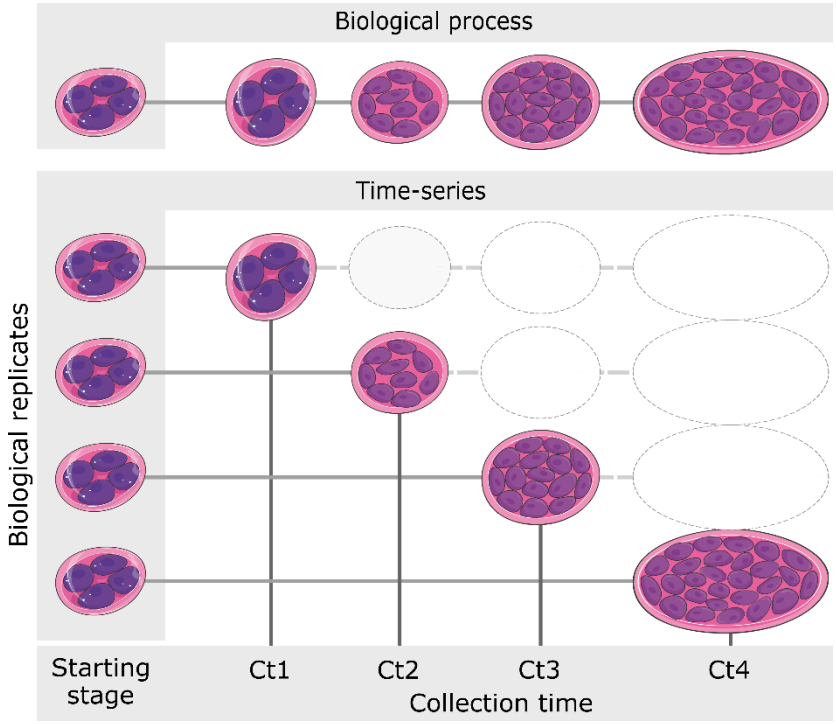

Figure 1.8 | Representation of a time-series design to study a biological process that is continuous. At the starting point, biological replicates are harvested equally and later used to extract genetic information in specific moments of time (collection time = Ct). The data obtained at different T represent different states of the same continuous system. The expression values are compared between Ct1, Ct2, Ct3 and Ct4 to reconstruct the behaviour of genes underlying a process in study. This is based on the assumption that the biological replicates are synchronous i.e. all cells start their mechanisms at the same moment.

Nevertheless, even though a well-characterized time-series can reveal different oscillatory genes on a genome-wide scale, time-series studies raise new concerns. Namely, a study with a time-series experimental design requires a great amount of initially synchronized samples so that the genes from all replicates are at the same phase of the system.

In both approaches, single matrix or time-series of expression measurements, due to the heterogeneity in gene specific periods and phases, the theoretical methods can fail to detect important variations within and between the genes. Hence, the sampling rate might not reflect correct oscillatory patterns or even recover artificial cyclic trajectories. Therefore, it is not trivial to recognize how truthfully a time-series illustrates the underlying continuous behaviour of the feature in question. Also, timed studies are not always possible given the time required for each sample collection, and also to the high cost of such experimental designs (reviewed in: Bar-Joseph *et al*., (2012)[80]).

For these reasons, the study of temporal biological events is a challenging field. Nowadays, computational analysis and mathematical modelling of transcriptomics data allows us to characterize and interpret large collections of experimental genome-wide information. Computational algorithms addressing RNA measurements rely on the asynchrony between the samples, while mathematical models are required to reconstruct the genetic trajectories. Finally, combining real time techniques with quantitative measurements will help to accurately describe the GRNs within the biological system[69,76,80].

Section 5 | Trajectory Inference Algorithm: Oscope

Is it possible to recover rhythmic information from static (snapshot-like) high-throughput genetic measurements, i.e. when the experimental design does not account for time between sampling points?

To approach this challenge, there are several trajectory inference algorithms (also called pseudotime inference algorithms) that aim at reconstructing the ordering of samples as they progress through a dynamic process (for a review see Saelens *et al*., (2019)[81]).

For this work we chose to apply the Oscope algorithm proposed by Ning Leng and colleagues (2015) which will be described in this section[76]. Oscope is a statistical approach to infer the oscillatory trajectory of genes from tabular, unsynchronized values[76]. It was designed to be applied to scRNA-seq data, however it can be used to infer the temporal ordering from a set of high-dimensional expression values obtained by any high-throughput technique. Briefly, to reconstruct a possible oscillatory trajectory for a gene, the Oscope reasoning is not uniquely based on the assumption that samples close in order of progression have similar expression

profiles (cell/sample synchrony). Rather than looking at the results as a progression towards an end point and reconstructing an individual line of events for each gene, by maximizing the similarities between the samples (case study in **Figure 1.7**), Oscope is guided by co-regulation information, i.e. by maximizing the similarities between the genes. This strategy results in a major advantage that consists in allowing phase-shift between different genes, enhancing the possibility of recovering an oscillatory behaviour[76,82].

Next, for groups of genes that have similar transcriptional profiles, the algorithm recovers a cluster-specific order of samples by minimizing the expression changes between them. This order will delineate the base-cycle for each cluster, that is, the profile of an oscillation that is repeated in time, for that particular cluster. Only then, the gene-specific trajectory is reconstructed based on the base-cycle of the cluster where the gene belongs. Noteworthy, Oscope does not focus on the magnitude of the oscillation or its scale of time. This algorithm focuses on recovering the cyclic profile of genes from unsynchronized genomic data and the variable order recovered by Oscope is not wall-clock time but progression through the cycle, hence the designation of pseudotime ordering strategy.

In the application process, in accordance to Leng *et al.* (2015)[76], the algorithm requires a previous processing step that consists of normalizing the quantified transcriptomic measurements, i.e. make the expression values comparable between samples, and rescaling the normalized expression values between -1 and +1 to facilitate the calculations. The user has the flexibility to choose which normalization method will be applied on the data however Oscope already provides a built in method. To reduce the computational demand and the time consumption, optionally, the user can select only highly variable genes (HVGs). Afterwards, it is possible to proceed to the main analysis, that is composed of three major steps: (i) Paired-sine model to find the putative co-cyclic genes; (ii) K-medoid clustering to cluster the groups of genes that oscillate with the same frequency; and (iii) ENI (Extended Nearest Insertion) algorithm to reconstruct the sample ordering and consequently the base-cycle for each cluster. A more detailed presentation of these steps is presented in the next subsections.

5.1 | Paired-sine model

A non-parametric test on co-regulation information between pairs of genes is used (hence the designation “paired”). Additionally, for simplicity reasons, the algorithm assumes that all oscillations are sinusoidal (hence the designation “sine”), which allows the modeling of

asynchronous data using trigonometric rules. If two genes, *g1* and *g2*, follow an oscillation with the same frequency but display a phase-shift (ψ), then not only in one sample (S_1) but in all samples (S_n) they will display approximately the sameψin the following formulation: $\psi_{gi,gj,s1}=\psi_{gi,gj,s2}=\ldots=\psi_{gi,gj,Sn}$. If this condition is not satisfied in all samples, then the dynamic changes between those two genes are not associated. Therefore, the rescaled expression values (X) (rescaled to [-1,1]), for each gene pair ($X_{g1,1}$ and $X_{g2,1}$), that oscillate in a codependent sinusoidal fashion in any sample (S), can be described by the following two-dimensional sinusoidal function:

$$\textbf{(eq. 1)} \qquad X^2_{g1,s} + X^2_{g2,s} - 2X_{g1,s}X_{g2,s}cos(\psi_{g1,g2}) - sin^2(\psi_{g1,g2}) = 0$$

In consequence, this demonstrates that the two dimensions, X_{g1} or X_{g2}, can be described as a function of each other. This means that the expression values are not time dependent in the paired-sine model. In other words, using the comparison between genetic profiles of two genes in all samples, it becomes possible to infer if these genes cycle with the same frequency across the dataset, independently of the sample order, time of oscillation or its starting point, or even the magnitude of expression for each gene.

Yet, eq.1 is true (i.e. corresponds to 0) only if the ψbetween *g1* and *g2* is exactly the same in all samples. Given no previous information about the $\psi_{g1,g2}$ Oscope estimates the optimal phase shift between the pair of genes under study, by testing all possibilities of ψ, ranging between 0 and 2π. The previous sinusoidal function (eq. 1) is fitted to each pair of genes, *g1* and *g2*, in each sample at a time, for a vector of ψ values. For each gene pair, the output of eq.1 is summed for all samples and is called the error term ($\varepsilon_{g1,g2}$). Mathematically, the $\varepsilon_{g1,g2}$ will be further transformed into genetic distance between *gi* and *gj* ($\varepsilon^2_{g1,g2}$), that can be calculated by the following eq.2:

$$\textbf{(eq. 2)} \qquad \varepsilon^2_{g1,g2} = \sum_s^0 0[X^2_{g1,s} + X^2_{g2,s} - 2X_{g1,s}X_{g2,s}cos(\psi_{g1,g2}) - sin^2(\psi_{g1,g2})]^2$$

The optimal $\psi_{g1,g2}$ is estimated by minimizing the ε^2 for each gene pair amongst all samples, i.e. the closer to 0 the more similar are the expression profiles between the genes.

As illustrated in **Figure 1.9**, two genes A (in blue) and B (in red), theoretically oscillate in time with sinusoidal trajectories. As gene A started to increase mRNA levels before the gene B, this indicates that they are phase-shifted, in this case $\psi = \pi/4$. Nevertheless, the expression values, that are plotted on the *y* axis, during the analysis are presented in a form of a gene-by-

sample matrix. So, Oscope fits the paired-sine model on the data to identify the optimal phase-shift between all pairs of genes across all samples, and to calculate the corresponding distances ($\varepsilon^2_{g1,g2}$) (**Figure 1.9** Top right panel).

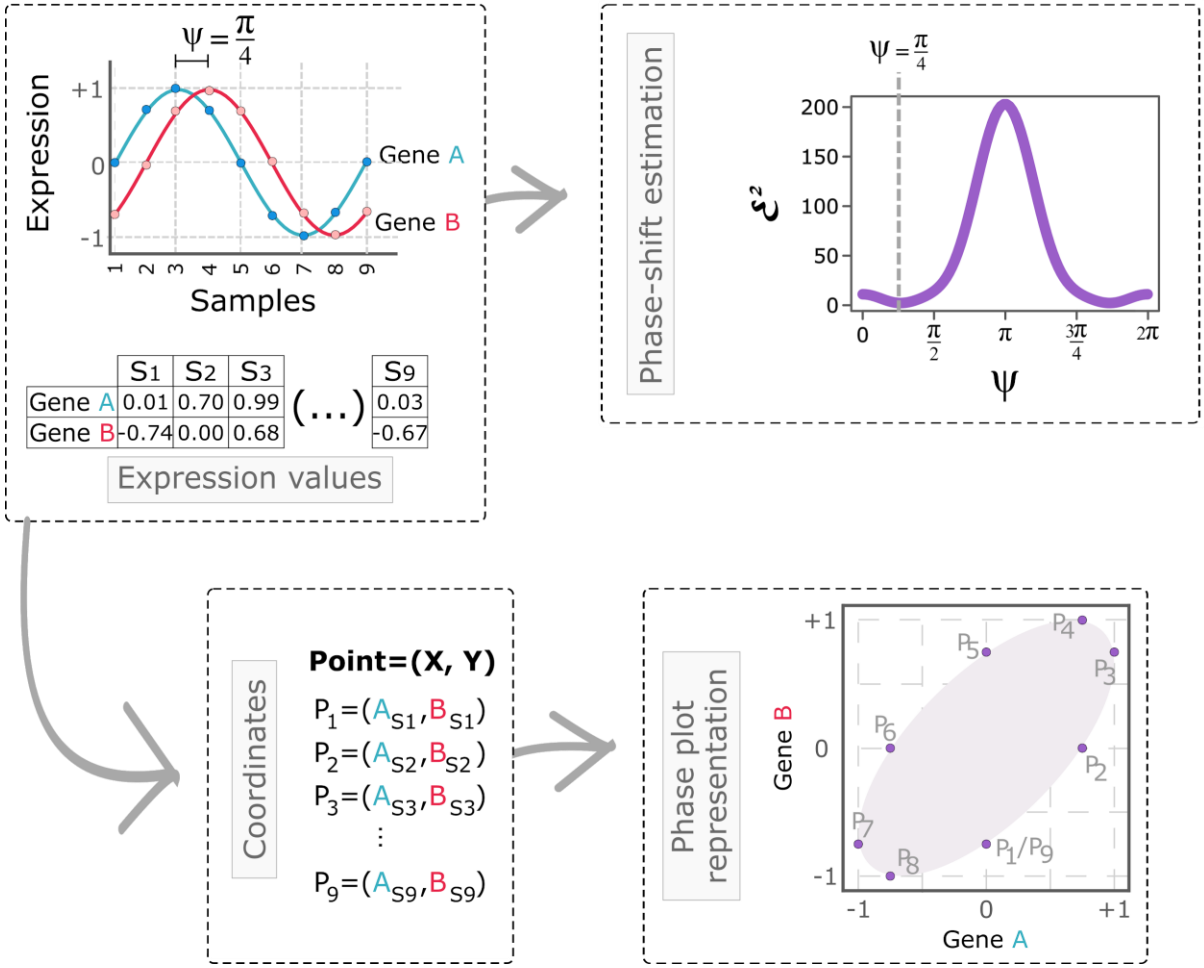

Figure 1.9 | Graphical representation of the Oscope modeling of 2 co-regulated oscillatory genes that are phase shifted by π/4. | Top left panel: scatter plot of genes A (blue) and B (red) expression values (*y* axis) across nine samples. Samples are artificially ordered to represent a sinusoidal pattern. The values are scaled between -1 and +1 to resemble the Oscope approach. Gene-by-sample matrix of expression values for gene A and B is presented. | **Top right panel:** Graphical representation of $\boldsymbol{\psi}$ (phase-shift) estimation. | **Bottom left panel:** table of *x* and *y* coordinates for each point gathered from the expression values for two genes that are predicted to oscillate with the same frequency. Coordinates are independent of order. | **Bottom right panel:** phase plot of gene A vs gene B scaled expression values (violet dots). If the genes A and B are phase-shifted by $\boldsymbol{\psi}=\boldsymbol{\pi/4}$, the pattern of points will resemble an ellipsoid with that direction. **P** = time-point; ***S*** = sample; $\boldsymbol{\psi}$= phase-shift.

Visually, one may interpret the Oscope reasoning by illustrating the expression values of two genes in a phase-plot. Rescaled expression values of both genes, X_{gA} and X_{gB}, are taken from each sample S and illustrated as one point (P) *per* sample (**Figure 1.9** Bottom). Therefore, sample order does not influence the results, since the genetic information for a pair of genes is gathered from one sample at the time. As a result, genes that oscillate with the same pace but with $\psi = \pi/4$, will form a closed trajectory, like the ellipsoid form in **Figure 1.9**.

The form represented in a phase-plot is dependent on the phase-shift between the pair of genes (**Figure 1.10**). Moreover, a closed ellipsoid trajectory can only be achieved if a pair of genes that is tested, cycles with the same frequency (**Figure 1.10 A**).

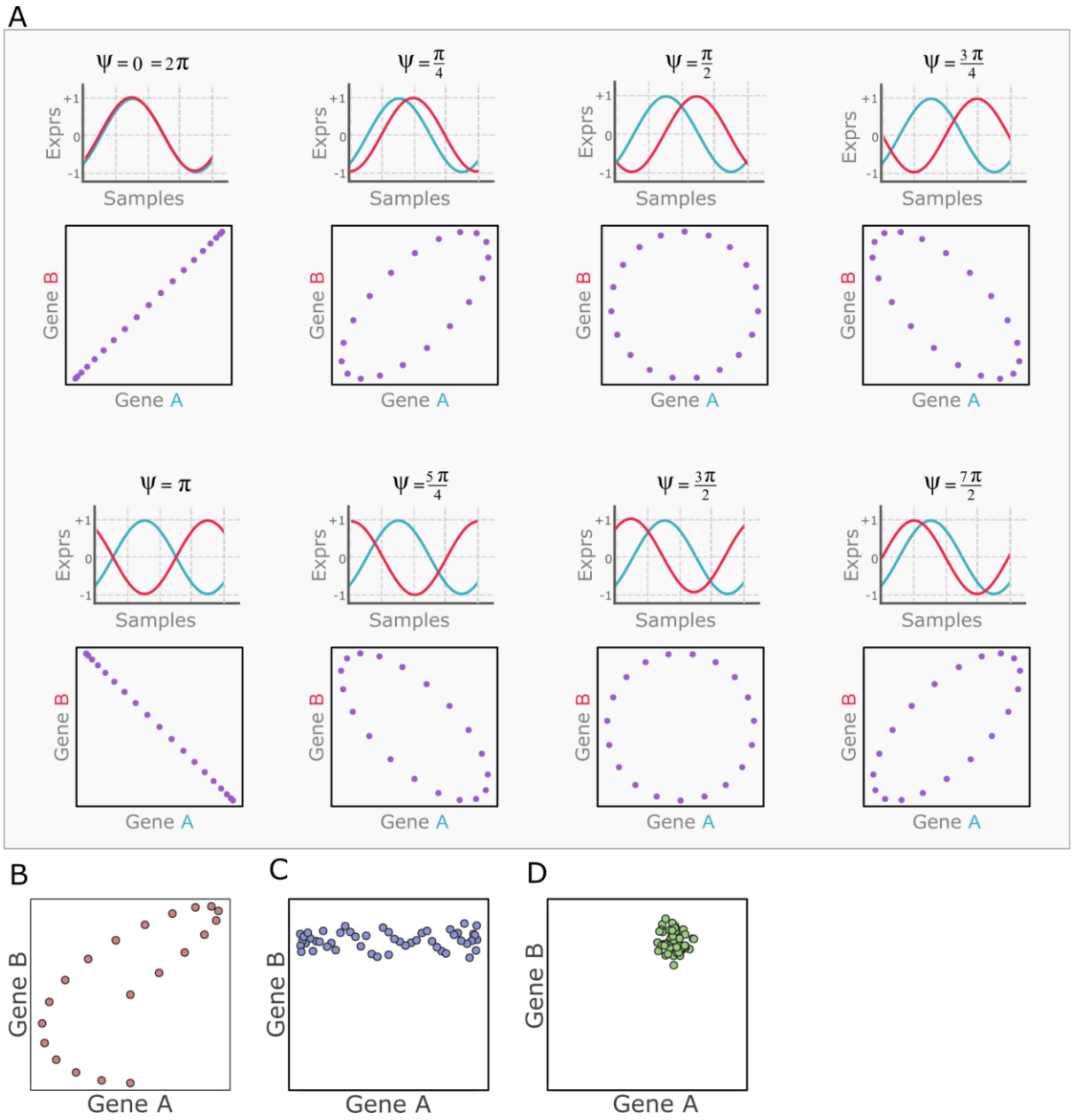

Figure 1.10 | Examples of phase-plot profiles for gene-pairs showing different expression profile scenarios. A | The scatter plot of two oscillatory genes A and B, with the same oscillation frequency but different $\boldsymbol{\psi}$ forms closed trajectories. It indicates repetition/periodicity, and each trajectory is dependent of the$\boldsymbol{\psi}$between the gene pair. On top, for different $\boldsymbol{\psi}$, are plotted gene expression profiles with correspondent phase plots below. **B |** Two oscillatory genes with different frequencies of oscillation, independently of the $\boldsymbol{\psi}$, will have an open trajectory. In this case, while gene A completes one oscillation ($2\boldsymbol{\pi}$), gene B only completed ¾ of the cycle ($\boldsymbol{3\pi/2}$). **C |** The scatter plot of an oscillatory gene A vs stably expressed gene B resembles a band. The gene A varies the expression values in accordance to up and down phase of oscillation on the *x* axis, while gene B is expressed around a stable value (on the y axis). **D |** The scatter plot of two genes with expression changes around stable values resembles a cluster of points. $\boldsymbol{\psi}$= phase-shift; Exprs = Expression.

As an example, by comparing pairs of genes, Oscope's approach allows to overcome misrepresented results, like in **Figure 1.7**, where two genes that cycle with the same periodicity but start to oscillate at different wall-clock times, are not related by the mathematical models only because their expression profiles were analyzed separately.

In the downstream analysis, the search is limited to pairs of genes that are highly related by mathematical modeling, i.e. the gene pairs with the smaller distance between them. Therefore, the sine score parameter is created by transforming the ε^2 into the sine score as follows: $-log_{10}(\varepsilon^2_{gi,gj})$. The sine scores are ranked and only the pairs with highest values (top T%) are selected for further analysis. By default, only the top 5% of gene pairs are assumed as good candidates for further clustering. However, the user has the freedom to guide the analysis under different parameters. The choice of an optimal T% value is not statistically supported and therefore can influence the outcome. Moreover, these values depend on the data that is under analysis, since if the samples are more homogeneous, for example form the same cell population, it is more likely to find genes with great pairwise similarity than in a heterogeneous population of cells, like in the examination of a whole embryo.

5.2 | K-medoid clustering

To cluster the candidates selected in the previous step, Oscope applies a K-medoid clustering algorithm, widely used for static sequencing data. This mathematical approach receives the dissimilarity metrics and the number of clusters (K) in the dataset as parameters. By testing different clustering options, the algorithm finds the optimal associations between the variables, by minimizing the dissimilarities between them. Oscope gives the $\varepsilon^2_{g1,g2}$ values as the genetic dissimilarities required by K-medoid. As a result, gene pairs with smaller ε^2 will be clustered together. Therefore, the clustering might find genes that are biologically co-regulated (and therefore functionally connected), or simply genes that share similar frequencies but are not functionally associated.

Since it is not always possible to know in advance the amount of clusters, Oscope offers the possibility to select a fixed number of K or to test different K values. The optimal K is determined by maximizing the silhouette distance - a measure of how similar a variable is to the attributed cluster. However, no further access to significance of each K is provided in the traditional Oscope pipeline.

It is worth remembering that the phase shift and the distances are mathematically estimated, therefore the clusters may present not only oscillatory dynamics. As an example, the expression values may induce the mathematical approach in the oscope algorithm to estimate a certain $\psi_{g1,g2}$ and a $\varepsilon^2_{g1,g2}$ close to 0 between genes that are linearly correlated i.e., genes that are not oscillatory but are codependent (**Figure 1.10** | $\psi = 0$ or $n\pi$). In consequence of high similarity in their dynamic trajectory, these genes will be clustered together and should not be analyzed in the downstream analysis. Therefore, the group of Ning Leng created two additional filters to examine the clustering performance[76]. Each cluster is tested for variability in sine scores and phase shifts in parallel.

Firstly, within-cluster sine score variations are tested to infer the significance of groups of genes found by the Oscope method. Arbitrary genes are selected from the matrix of rescaled expression values and their sine scores are calculated. The number of selected genes corresponds to the number of genes in the K of interest. The distributions of sine scores are compared between the original K and permuted group of genes. Exclusively the clusters where the original sine score median values are comprehended in the highest 10% of permuted sine scores will pass the filtering step. Clusters that meet these criteria prove that the clusters constructed by Oscope design are not randomly grouped genes.

Secondly, phase shift analysis was developed to avoid detecting clusters of genes with linear correlation. With this purpose, the authors created an additional parameter: phase shift residual (υ), that is calculated for each gene pair, *g1* and *g2*, in the K of interest. The following equation is applied in each sample and the minimum value is selected as$\upsilon$for a pair of genes:

(eq. 3) $\upsilon_{g1,g2} = \min((\pi - \eta_{g1,g2}), \eta_{g1,g2})$; where $\eta_{g1,g2}$ is $\psi_{g1,g2} \bmod \pi$.

Given that it is not possible to infer any direction from the static expression values, i.e., if the mRNA concentration was increasing or decreasing in the time of collection, $\upsilon_{g1,g2}$ is estimated to be as close to zero as possible. As such, the minimal possible $\psi_{g1,g2}$ is preferred in this filtering process, hence the need for $\eta_{g1,g2}$. This approach narrows the distribution of $\upsilon_{g1,g2}$ when compared to the original $\psi_{g1,g2}$ distribution. Only clusters where the 90th quantile of $\upsilon_{g1,g2}$ values is higher that the threshold of $\pi/4$are considered. Consequently, clusters having small within-group phase differences are rejected from the downstream search.

5.3 | Extended Nearest Insertion (ENI)

Once candidate genes are clustered, the Oscope algorithm proceeds to reconstruct the order of the samples for each cluster. Since the input data does not account for time differences between samples, it is not possible to distinguish between the earlier and the later cycles. As such only one oscillation that accommodates all the samples - the *base-cycle* is retrieved for each K. The base-cycle is reconstructed by applying the Extended Nearest Insertion (ENI) model developed for the Oscope method, however the algorithm allows it to be replaced by other probabilistic approaches for pseudotime reconstruction.

The ENI method was developed to start the search with only three randomly selected samples and constitutively a new sample is added to the analysis. Samples are set on a loop with no temporal direction to resemble a closed trajectory and capture the cyclic feature of the data. Therefore, for any number of samples under analysis, they are all evenly distributed on a cycle, and all possible combinations of order are studied every time that the number of samples increases. The differences between the original expression values in the sample and the expected values designed by the loop are examined. The best sample's position in the loop is selected by minimizing the differences between the samples' real and expected values taking into account all genes in the cluster. Thus, the spatial position of the samples in each cluster cycle can be different, which makes the base-cycle cluster-specific. Based on the base-cycle for each cluster, a gene-specific oscillatory profile can be visualized by plotting the expression values on a scatter plot (**Figure 1.11**).

To optimize the sample ordering along the oscillation, a 2-opt algorithm (which is a local search algorithm) is further applied. Arbitrary samples are selected and permuted, followed by a new estimation of the differences between quantified and expected expression values. If this difference is lower than in the previously determined order, a new sample order is adapted. The sample permutations continue for 1000 rounds or until the optimal order is achieved.

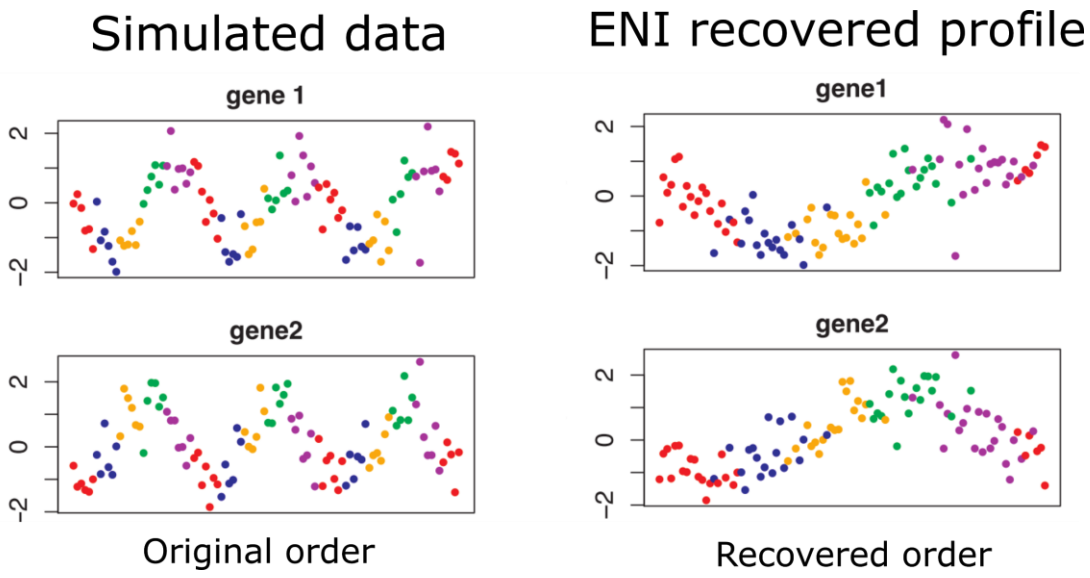

Figure 1.11 | Graphical representation of ENI order recovering for 2 different genes. Samples with simulated expression values found in different phases of oscillation are illustrated in different colors. ENI algorithm reconstructed a base-cycle for each gene accommodating all samples in one oscillation. (Adapted form Leng et al., 2015[76] - Supplementary material).

Overall, by comparing all the available pairs of genes, it is possible to find groups of genes with similar dynamic behaviour. After evaluating the data modeling approach underlying the Oscope algorithm, we conclude that, although developed for single-cell sequencing, this mathematical approach is also applicable to our bulk RNA-seq data, and therefore appropriate for the scientific challenge proposed for my thesis, i.e., to find oscillatory genes involved in the early stages of the chick embryo development, using static microarray transcriptomics data, for which no time series are available.

Section 6 | Main Goal and thesis outline

A developing embryo is composed by a great number of cells interacting continuously with each other in time and space, ultimately giving rise to a functional organism. Is it possible to study the temporal progression of the underlying gene networks regulating development at the scale of a whole organism? Emerging high-throughput sequencing technologies, allied with mathematical modeling, allows for data collection and analyses that help with understanding such complex questions.

The aim of this thesis was to apply a bioinformatics approach to finding the chick embryo ClockOME, i.e., a list of oscillatory genes active during the early periods of chicken

development, using publicly available transcriptomics datasets, and applying a pseudotime inference algorithm via a bioinformatics analysis pipeline.

Chapter I provides an overview of the biological subject and data collection techniques, as well as a brief introduction to the mathematical details of the algorithm applied to the data (Oscope).

Chapter II describes the workflow of the research, which started with the collection of microarray data and its preparation for the statistical analysis. This required the building of the FrozenChicken R data package, containing the gene expression vectors, that allowed the external normalization of the three chicken microarray datasets. To extract the oscillatory genes for which there is no temporal information, a statistical method called Oscope was applied. This method also inferred the pseudo-temporal trajectory of each candidate gene for the proposed ClockOME. Finally, it describes the genomic annotation of the results using the GO database, and their predicted interactome generated using the STRING database.

In **Chapter III**, the findings from the analysis are shown, starting by exploring the datasets collected to pursue the research. In total, 296 genes were predicted to be oscillatory, either in the limb or in the presomitic mesoderm of the chick embryo. These are relevant for a plethora of molecular functions, namely morphogenesis, development of different tissue types, and transcriptional regulation. The ClockOME gene list is also compared with previously published lists of oscillatory genes reported in the literature. Finally, six candidate oscillatory genes are proposed as high-confidence candidates for further experimental validation of their dynamic behaviour.

Chapter IV, summarizes the work and highlights the findings described in this thesis, framing its importance for the scientific community.

Finally, **Chapter V** provides some future work directions to advance the understanding of the role of oscillatory genes for the vertebrate embryo development.

Chapter II
METHODOLOGY

2 | METHODOLOGY

Data analysis programming environment

Aiming to search for oscillatory genes in early vertebrate development, it was necessary to develop an automated analysis pipeline that would be efficient and reproducible, i.e., also applicable to new datasets. Therefore, for this study we used the R programming language (version 3.6.3)[83], which is freely available, and particularly suited for statistical analysis and graphical data visualization. Additionally, this language is supported by an active development community, and expanded through numerous packages dedicated to OMICs and other biological data (present in the Bioconductor repository (https://bioconductor.org/)), hence greatly extending its functionality. To analyse, explore, and understand the data, the free open-source integrated development environment (IDE) RStudio® (version 1.1.463)[84] was used. All data analyses were undertaken using custom R scripts, implementing additional functions in order to develop a pipeline suited to answer our questions. As such, learning R programming and implementing the R analysis pipeline, presented in Annex 1, shows the programming work developed throughout the thesis.

All analyses were undertaken in a Linux environment, running the Ubuntu distribution (release 18.04.4 LTS)[85]. Final graphical illustrations developed for this dissertation were assembled and designed using the free open-source software Inkscape (version 0.92.4)[86].

2.1 | Data collection

In order to find genes that present oscillatory expression in early vertebrate development, we started by gathering publicly available transcriptomics datasets from *Gallus gallus*, obtained by microarray technology. This organism was chosen not only because of its long-standing tradition as a vertebrate embryo model organism, but also because this thesis was developed in close collaboration with the Temporal Control of Cell Differentiation Lab (Universidade do Algarve) that uses the chick as a model organism, hence allowing the future experimental validation of the candidate oscillatory genes. The technical decision to use microarray transcriptomics' datasets (instead of data from more recent techniques such as RNA-seq) was: (i) to avoid the anticipated time-consuming difficulties with mapping chicken genes to its newly annotated genome; and (ii) because the preliminary tests conducted using bigger

datasets (from Single-Cell RNA-seq) were not compatible with our currently available computational resources. The microarray datasets were searched in 2 public data repositories: ArrayExpress[87] and Gene Expression Omnibus (GEO[88]). The databases were queried (in February 2020) for expression profiling by array in *Gallus gallus* species, using the following key terms: "early vertebrate development", "gastrulation", "somitogenesis", and "segmentation". Next, the studies were selected for analysis based on raw-data availability (.CEL files). Also, we focused our analysis only on samples coming from the chicken limb or PSM that have not been experimentally manipulated (control conditions). Applying these criteria, 3 datasets were selected, with the following accession codes: GSE75798, and E-MTAB-406, both containing PSM data; and E-MTAB-4048 with limb data (**Table 1**). By combining these public datasets, it becomes possible to proceed with the meta-analysis in order to test new hypotheses without the need to re-run the transcriptomics laboratory experiments.

Table 1 | Collected datasets.

	E-MTAB-4048	***E-MTAB-406***	***GSE75798***
Repository	ArrayExpress	ArrayExpress	GEO
Platform	Affymetrix GeneChip Chicken Genome Array (A-AFFY-103)	Affymetrix GeneChip Chicken Genome Array (A-AFFY-103)	Affymetrix Chicken Genome Array (GPL3213)
Tissue	Limb	Right PSM	Bilateral PSM
Developmental Stage	HH* 20/24	HH* 12	HH* 12
Biological State	Somitogenesis	Somitogenesis	Somitogenesis
Samples (total)	34	75	19
Samples (used)	16	18	15
Citation	Anderson *et al.*, (2016)[89]	Krol *et al.*, (2011)[3]	Oginuma *et al.*, (2017)[9]
Year of dataset publication	2009	2011	2015

*HH: Hamburger and Hamilton embryo staging. (Hamburger V. and Hamilton HL. A series of normal stages in the development of the chick embryo. (1951) J Morphol. 88(1): 49-92. PMID 24539719)

All 3 experiments were conducted during the somitogenesis period of embryonic development, when somitogenesis clock gene expression oscillations are known to occur, and

measured using the same microarray chip – Affymetrix Chicken Genome Array (microarray platform description: GPL3213 in NCBI, and A-AFFY-103 in ArrayExpress).

The E-MTAB-4048 dataset holds samples from several chicken body parts used initially in Anderson *et al.*, (2016)[89]. Only 16 arrays were selected for downstream analysis. Samples were extracted from embryos at HH stages 20 and 24, and as shown in the **Figure 2.1 A**, the anterior limb bud was separated into the anterior and the posterior areas.

As for the dataset GSE75798, originally used in Oginuma et al. (2017)[9], the bilateral p-PSM was extracted, after the removal of the endoderm and the ectoderm tissues (**Figure 2.1 B**). Here, were selected 15 arrays corresponding to right and left PSM regions, from embryos in the HH12 stage.

Finally, the E-MTAB-406 dataset, described in Krol et al. (2011)[3], compiles arrays from different animal models. Accordingly, only the subset of 18 chicken arrays, from right PSM at HH12 stage, were selected for further analysis (since only the right pPSM samples were extracted for sequencing by the authors, after the removal of the endoderm tissue (**Figure 2.1 C**).

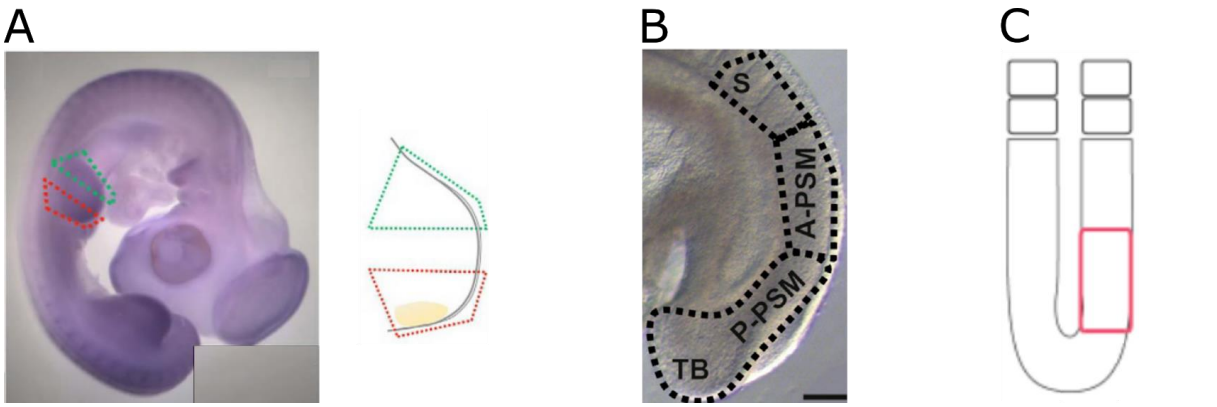

Figure 2.1 | Representation of the chick tissues used for RNA extraction prior to microarray hybridization. **A |** E-MTAB-4048 dataset - the limb bud was divided into 2 sections: the anterior limb (green box) and the posterior limb (red box). **B |** GSE75798 dataset - only pPSM was used for the microarray hybridization. The grafts included a tail bud (TB) and was extracted from both sides. **C |** E-MTAB-406 dataset - only the right p-PSM tissue was used for transcriptomics profiling. Adapted from A - Anderson *et al.*, (2016) [89]), B - Oginuma *et al.*, (2017)[9], and C - Krol *et al.*, (2011)[3].

After manually selecting the appropriate datasets, further data download (conducted in March 2020) and analysis was conducted programmatically. To extract datasets from the GEO database, we employed the *GEOquery*[90] package, and for the ArrayExpress database, the *ArrayExpress*[91] package was used.

To read in the raw-data from .CEL files into an *Affybatch* R object, we used the *affy* package (version 1.64.0)[92]. Probe intensity values coming from limb samples were joined into one Affybatch data object, where each array is individually stored as a sample. The same was performed for the arrays from PSM samples, which were selected from 2 different experiments. Expression values from both Affybatch objects were extracted with the *Biobase* package (version 2.46.0)[93] and retrieved as data matrices, where each array corresponds to one column. Finally, we obtained 2 matrices, one containing all 33 samples from the PSM tissue, and a second matrix with the 14 arrays from the limb tissue.

2.2 | Quality control

Before pursuing the analysis itself, one of the most important steps consists in the quality control (QC) of the samples, i.e. testing the quality of the data present in the arrays. This aims at improving the overall quality of the results, since including poor quality data in the analysis can negatively impact the results, and potentially lead to unreliable conclusions.

Accordingly, to test the quality of the previously selected samples, we used the *ArrayQualityMetrics* package (version 3.42.0)[94]. This package implements various methods that test the quality of the data present in the microarrays. Briefly, it performs quality tests firstly on the raw-data files, allowing the identification of potential technical errors. Next, the algorithm tests the quality of the normalized data, i.e., it assesses if the normalized expression values coming from different arrays are now directly comparable. Additionally, the package allows the visual inspection of various metrics to evaluate the individual quality of the arrays, and decide whether to eliminate some of the arrays from further analysis. The samples that fail to meet the criteria are called poor-quality arrays, behaving as outliers within the overall data.

In the selected dataset, both tissues presented outliers (**Annex 2**). Firstly, 2 arrays from the limb samples, (arrays HH24_PL_Post1692 and HH24_PL_Post1693) were identified as outliers by the NUSE (Normalized Unscaled Standard Error) metric, and consequently removed from further analyses (**Annex 2.1**). QC for the PSM samples also detected outliers in the samples coming from the GSE75798 data series. With NUSE metrics only one array was called outlier (array GSM1968022) and was removed from downstream analysis (**Annex 2.2**).

Additionally, individual array quality tested by MA plots (ratio intensity plot: M (log intensity ratio) as a function of A (average signal intensity)) also selected 7 additional outliers in the PSM arrays, including the one reported as outlier by NUSE. In these individual MA plots, these 7 arrays showed a trend in the lower range of A, indicating different background intensities. However, since the next step of the analysis will normalize the sample intensities, accounting for such technical biases, we decided to keep the 6 samples that were called as outliers exclusively in the MA plot but not with NUSE. Moreover, preserving these arrays was also important to keep the dataset at a meaningful size for our analysis.

After the QC step, the dataset used for further analyses was composed of 32 arrays from chick PSM tissue, and 14 arrays belonging to the limb.

2.3 | Data Pre-processing: Normalization of arrays from different experiments

Before any microarray dataset can be analysed, a pre-processing step must be performed. The most broadly used method for such data pre-processing is the Robust Multiarray Analysis (RMA) algorithm, that performs background correction, normalization, and summarization of the probe intensity levels, before converting them into relative gene expression levels (**Figure 2.3** - Panel 1)[75]. This method requires the simultaneous analysis of multiple arrays, usually all coming from the same experimental batch. However, microarray datasets independently pre-processed are not directly comparable. Accordingly, since our meta-analysis comprises arrays from 3 different experimental sets, they had to be jointly pre-processed. Given the imbalance in the number of arrays per experiment, the most accurate way to pre-process this dataset is to normalize it against a custom database of estimated probe-specific effects and variances, calculated from chicken microarrays present in public repositories. This method is termed Frozen Robust Multiarray Analysis (fRMA)[95].

The fRMA algorithm requires frozen RMA vectors, that are a large collection of microarray data from the same platform (i.e., from the same microarray chip model), that are precomputed together. This creates a universe of possible expression values for each microarray probe present in a particular microarray chip model. Due to the absence of ready-made frozen vectors for chicken microarrays, we developed an R package, named *FrozenChicken*, that holds the necessary probe intensity vectors needed to pre-process the 3 datasets selected for this project.

2.3.1 - FrozenChicken Development

The necessary data from microarray chips from *Gallus gallus* Affymetrix genome array was gathered programmatically from the GEO database using the E-utilities (Entrez Programming Utilities) software suite[96]. To filter this search, we only kept as datasets: (i) the GEO accession numbers of the experiments that had raw .CEL files available, and (ii) experiments conducted using the same commercially available chicken microarray chip, i.e., the platform id GPL3213 corresponding to the Affymetrix Chicken Genome Array. As shown in **Figure 2.2**, we collected 118 datasets (i.e., batches) with variable number of arrays *per* experiment. Next, we randomly selected 4 arrays from each dataset, hence creating a database with a batch size of 4. As a result, a total of 472 arrays were assembled into an R package titled *FrozenChicken* (containing the so-called frozen vectors). This package was built with the *frmaTools* package (version 1.34.0)[97], and it is freely available in GitHub (https://github.com/iduarte/frozenChicken).

The description of the process related to building the package, the R code, and usage examples may be explored in the publication "FrozenChicken: Promoting the meta-analysis of chicken microarray data", available in bioRxiv (https://doi.org/10.1101/2021.02.25.432894)[98], and Zenodo (https://doi.org/10.5281/zenodo.3765944)[99] (**Annex 3.1** and **3.2** respectively).

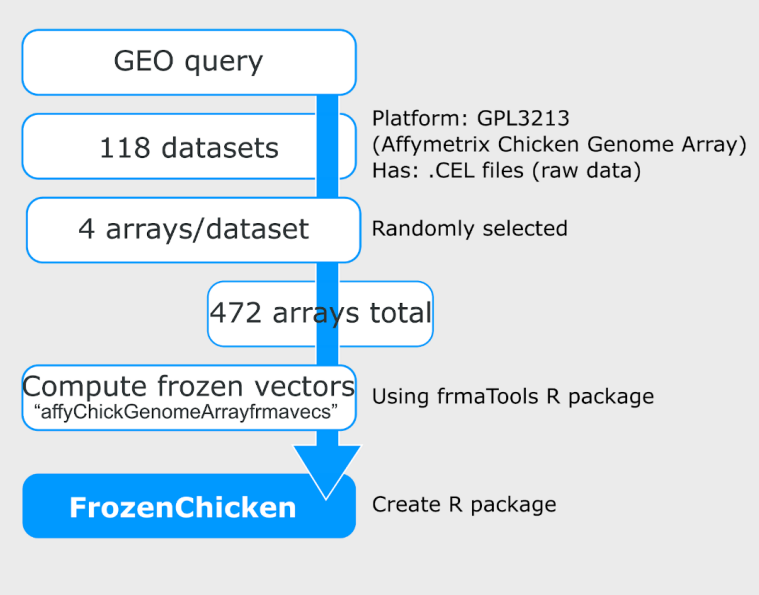

Figure 2.2 | Schematic representation of FrozenChicken package construction. 118 datasets, containing .CEL raw files, from GPL3213 microarray platform (Affymetrix Chicken Genome Array), were retrieved from the GEO database. Four arrays were selected randomly from each dataset. This resulted in 472 (Batch number x Batch size = 118 x 4) arrays that were computed into frozen vectors using the *frmaTools* package. Frozen vectors for chicken

Affymetrix genome array normalization are stored in the FrozenChicken R package (freely available in https://github.com/iduarte/frozenChicken).

2.3.2 - Data pre-processing and Normalization

Data pre-processing and normalization was performed via fRMA, using the frozen vectors specifically developed for this task, contained in the aforementioned *FrozenChicken* R Package[98]. This step normalized the probe intensities from the selected microarrays against an external database containing 472 Affymetrix chicken genome arrays. This approach accounts for various technical biases and batch effects, and retrieves gene expression values for each array in the form of a gene-by-array matrix. Importantly, this step includes the log2-transformation of the expression values.

2.4 | Annotation

When working with microarray data, one needs to retrieve additional information, also called metadata, about the genes that are coded by the microarray probes measured. This process of annotation consists of mapping each probe-set to a gene symbol, as well as gene description, and other relevant information, such as annotated gene function.

For the Affymetrix Chicken Genome Array, each chicken RNA transcript is mapped by a probe-set of 11 different oligonucleotides, each with 25 base-pairs in length. Therefore, to collect metadata for a gene, it is necessary to know which probes from the array represent the probe-set that together yield the expression value for each particular gene. The step of gathering the information about the mapping of the complete probe-set for each gene is called summarization. Then, each probe identifier (probe ID) can be mapped to a specific gene identifier (accession number) from a particular biological database. At this point, it is possible to map the accession numbers to the metadata stored in the genetic features database.

For this work, the gene identity and its functional information was annotated using the chicken database, version 3.2.3, hosted by the R Package *AnnotationDbi* (version 1.48.0) [100], containing accession numbers from the GeneBank database (https://www.ncbi.nlm.nih.gov/genbank/). Based on the probe ID (“PROBEID”) key type, the metadata retrieved was automatically queried for the following functional information: “SYMBOL” for the official gene symbol; “GENENAME” for the extended gene name; and

"ENSEMBL", for the ENSEMBL ID (unique gene identifier from the Ensembl database) for each gene. The probes with no official gene symbol (i.e., not yet fully annotated) were excluded. The final dataset comprised a total of 29 886 probes to be further studied.

2.5 | Exploring the data: Descriptive statistics

To visually explore the overall distribution of the gene expression values selected for further analysis, we conducted a brief descriptive statistical analysis on the normalized and annotated chicken data matrices containing log2 transformed gene expression measurements.

The distribution of gene expression values was visualized with a violin plot produced with *ggplot2* (version 3.3.1)[101] showing the median, as well as the first and the third quartiles for each sample. The statistical values are documented in the annexes for each dataset (**Annex 4**). To explore the gene variance between arrays, a Principal Component Analysis (PCA) was conducted in R. Through PCA we can make a low-dimensional representation of the samples using the first 2 principal components (PCs) of the data. This gives us reliable information about the variability between the samples. Additionally, this PCA projection makes it visually possible to infer relationships between the samples, namely if they group together. In order to observe the overall expression values and visually search for cues, a heatmap of the data was plotted with sample clustering to show similar gene expression profiles within the data. A heatmap is a graphical representation where the colour gradient depends on the intensity of gene expression. Columns represent samples and rows correspond to different probe sets.

The same descriptive statistics analysis was applied to the intermediate data matrices containing the HVGs (highly variable genes) used as input for the paired-sine model (**Annex 4.3** and **5.1** for PSM; **Annex 4.4** and **5.2** for Limb datasets). Also, this QC analysis was performed for the final data matrices, containing the clusters of genes selected as candidate oscillatory genes by the Oscope algorithm (see description below) (**Annex 4.5** and **5.3** for PSM K1; **Annex 4.5** and **5.4** for PSM K2; **Annex 4.7** and **5.5** for Limb K1).

2.6 | Identification of candidate oscillatory genes

In an attempt to discover inner cellular dynamic processes and search for genes that display an oscillatory behaviour in chick embryo development, we computationally modelled the expression values from the 3 transcriptomics datasets using a trajectory inference (TI) method named Oscope[76]. This algorithm, usable via the R package *Oscope* (version 1.16.0)[76] defines clusters of co-oscillatory genes, and then orders the arrays along an oscillatory trajectory for each of these clusters. This method uses as input a gene-by-sample matrix of expression values (**Figure 2.3**-Panel 2). The *Oscope* R package was originally designed for Single-Cell RNA-seq data and the algorithm was implemented with ineffective memory usage, hence requiring significant amounts of computational power, even for small microarray datasets. Accordingly, throughout this study we used the R package named *oscillation* (kindly developed by Dr. Ramiro Magno, and freely available in GitHub: https://github.com/ramiromagno/oscillation), which is a redesign and refactoring of the Oscope algorithm to be faster, more memory efficient, and more user-friendly.

2.6.1 – Trajectory Inference

Prior to the application of the Oscope method, additional pre-processing steps were applied to format the input matrix as required by the algorithm, namely the exponentiation of the expression values (to remove the log2 transformation applied by the fRMA normalization). Then, these data were used as input for the *oscillation* R package. Individual size factors *per* array were calculated by median normalization according to Anders & Huber (2010)[102]. The matrix of gene expression values was normalized with the *EBSeq* package (version 1.26.0)[103] using the pre-calculated vector of size factors. Next, the gene expression statistics were obtained with the 'gene_statistics' function from the *oscillation* package, which outputs the mean, median, variance, and q1, q2, and phi (parameter estimations related to the Negative Binomial (NB) distribution, a commonly assumed model for the distribution of gene expression counts) for each row of the matrix, i.e. for each probe-set representing a gene.

To be in accordance with the biological process being modelled, since oscillatory genes must vary their expression values, the normalized matrix of gene expression was initially filtered keeping only genes with highly variable expression values. This filtering step consisted of two phases: (i) first removing genes whose mean was contained in the first quartile (25%) of the means from the whole matrix (i.e. genes poorly expressed were removed); then (ii) a

linear regression was applied on the log-transformed values of the variances and the mean. Only genes falling above the fitted line, whose residuals are greater than zero i.e., with variability higher than expected, were considered as high variance genes, and therefore selected for further analysis. This resulted in a matrix of 10043 probe-sets from the limb dataset, and 9527 probe-sets from the PSM dataset. These were considered to be intermediate data matrices. QC statistics for these intermediate analyses are shown in the **Annex 5.1** (PSM) and **Annex 5.2** (Limb).

As required for the paired-sine model, described in Leng *et al*., (2015)[76] ,the expression values were rescaled between -1 and 1. As default, for outlier adjustment, extreme expression values for each gene were imputed to its upper (or lower) thresholds, where the upper threshold value corresponds to the 95th quartile, and the lower one to the 5th quartile. The paired-sine model estimates the optimal phase shift for gene pairs by minimizing the genetic distance (ε^2) in a way that a pair of genes that are genetically distant probably do not oscillate with the same frequency (**Figure 2.3** - Panel 3). This step is computed by the 'paired_sine_analysis' function from the *oscillation* package, and outputs a list of gene pairs with their corresponding optimal phase shift, the genetic distance, and the gene pair sine score.

Only the top 5% of gene pair candidates, with the higher sine scores, were selected for further analysis. The decision to keep the threshold value of 5%, despite being restrictive, aimed at keeping the strongest oscillatory signals, therefore increasing the possibility to find true oscillatory genes. Once the candidate genes were selected, Oscope runs a K-medoid clustering algorithm to cluster the genes into groups with similar frequencies, but allowing different phases (**Figure 2.3** - Panel 4). Previously calculated scores are used as dissimilarity matrices for the K-medoid algorithm. Here, a varying number of clusters can be tested and thus, the maximum number of clusters allowed was set to 15. Despite testing different numbers of possible groups, the optimal K value is obtained by maximizing the silhouette distance. To guarantee that only genes with oscillatory dynamics were extracted from the original input data the 'FlagCluster' function was used, filtering the expression values by sine and by shift.

The final step consists in performing a pseudotime ordering of the variables, in order to obtain the best oscillatory trajectory from the data for each cluster (**Figure 2.3** - Panel 5). To recover the best sample ordering that will fit one oscillation for each cluster, the ENI algorithm is applied. It achieves the best order by minimizing the MSE (Mean Squared Error) of a sliding

polynomial regression (SPR). Next, it runs the 2-opt local search algorithm that will stop iterating if the optimal order was found, or if there were no changes for 1000 iterations.

The visualization of individual gene profiles (as illustrated in **Figure 2.3** - Panel 6) was performed using the *ggplot2* R package[101]. The complete list of genes is listed in the **Annex 6** and all individual gene trajectories are displayed in **Annex 7**.

2.7 | Functional Enrichment

2.7.1 - Functional Annotation

Using the list of candidate oscillatory genes output from the aforementioned trajectory inference (TI) analysis, a downstream functional enrichment (FE) analysis was performed to attribute biological meaning to the results. We analysed the Gene Ontology (GO) biological category enrichment of the output genes using the *topGO R* package (version 2.38.1)[104]. All 3 clusters of putative oscillatory genes were examined separately, and mapped to ENTREZ gene identifiers using the chicken specific annotation package *org.Gg.eg.db*[105] (version 3.10.0). The Biological Process (BP), Molecular Function (MF), and Cellular Component (CC) annotations were explored, with node size of 5, meaning that GO categories with less than 5 annotated genes were discarded. For the enrichment analysis was employed the “elim" method since it is a conservative approach, and it yields a minimal number of false positive results, as shown by Alexa et al (2006) [104]. Regarding the statistical testing, we applied Fisher's exact test.

The FE analysis using the topGO provided a list of meta-data for each ontology (BP, MF, and CC) with: (i) the enriched GO category code; (ii) the extended category name (“Term”); (iii) the number of genes annotated with that GO term (“Annotated”) from the total universe of genes; (iv) the number of genes from our input list that are amongst those annotated genes (“Significant”); (v) the expected number of genes to be found by chance in this GO category (“Expected”); and (vi) the p-value for the enrichment. A new column named “Proportions", with the ratio between “Significant” number of genes and the total amount of genes from the imputed list that we annotated to the respective database, expressed in percentage, was added to the results. It was used to specify the size of the treemap rectangles representing each category. Additionally, the category names were shortened (for the 20 most significant GO terms only) and presented in a column named “Short Name”, in order to make

the treemap visually comprehensible. Complete tables containing all the aforementioned results are displayed in the **Annex 8**.

Treemap plots were created for summarization and visualization of the FE categories. This is a space-fitting visualization for hierarchical structures such as GO categories. The plots were obtained programmatically adapting the REviGO (Reduce + visualize Gene Ontology) (http://revigo.irb.hr/) R code, with the plot type set to be categorical.

2.7.2 – Protein interaction networks

To further explore the functional interactions between the proteins encoded by the candidate oscillatory genes, protein interaction networks were assembled (**Figure 2.3** - Panel 7). The interaction information data were obtained and computed in Cytoscape 3.8.0[106] using the list of gene symbols output from the Oscope algorithm (*oscillation* R package). The genes were mapped from *Gallus gallus* to their orthologues in *Homo sapiens*. For each cluster provided by Oscope, a network was gathered using data from STRING (Search tool for retrieval of interacting Genes/Proteins) (https://string-db.org/) in protein query mode. The functional interaction networks were computed with a confidence interval of 0.8, and with a maximum of 20 additional interactors. The EnrichmentMap[107] plugin from the Cytoscape App Store was used to construct and visualize the results. Original networks, presented in the **Annex 9**, were clustered with the Markov Clustering Algorithm (MCL) from the clusterMaker[108] Cytoscape plugin. MCL clustering was based on the edges attributes, and the resultant distance matrix was calculated with default parameters, i.e. with an inflation value of 2.0, and assuming undirected edges. The enrichment data was automatically retrieved, using default parameters, with the stringApp, a STRING[109] plugin for Cytoscape, based on the protein query.

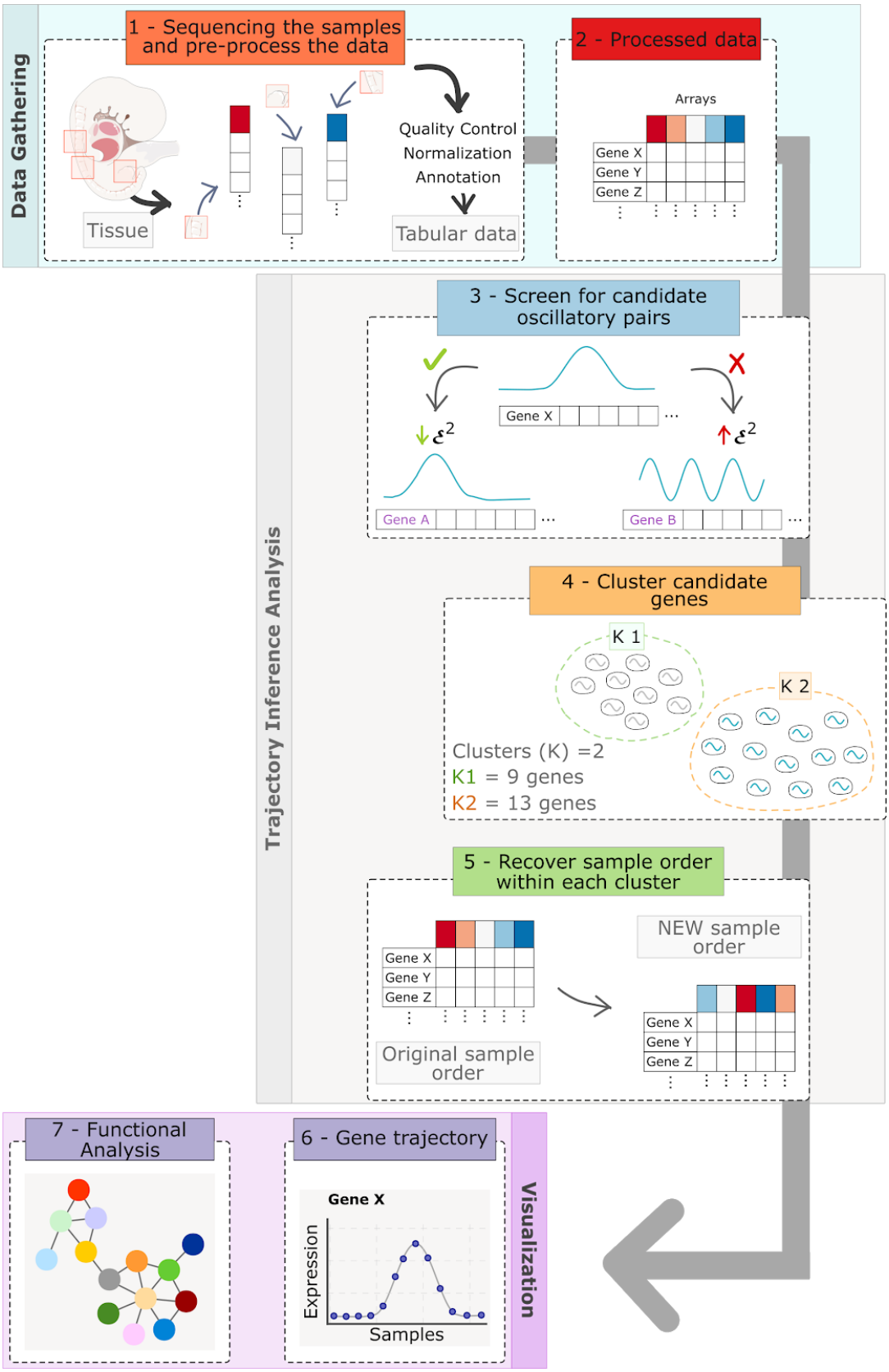

Figure 2.3 | ClockOME analysis workflow. Schematic representation of the major parts of the data analysis pipeline: **1.** Data gathering, **2.** Data pre-processing, **3-5** Trajectory inference; and **6-7** Data visualization and functional enrichment.

Chapter III
RESULTS
AND DISCUSSION

3 | RESULTS AND DISCUSSION

Aiming to shed new light and find clues about the biological oscillations taking place in early vertebrate embryogenesis, I went into a quest for the ClockOME. In this project I focused on searching for genes that are expressed in a rhythmic manner, using high-throughput transcriptomics data. The resultant list of candidate genes, that are expressed in an oscillatory way during the early development of chicken, will be called here the ClockOME.

Briefly, to find the ClockOME, I examined data coming from microarray experiments from samples collected during somitogenesis and limb development in the *Gallus gallus* animal model. The first step was to collect and prepare the data for the downstream analysis. To complete this step, I generated an R data package called "FrozenChicken" used for chicken microarray data normalization (**Section A**). Next, I applied a pseudo-time ordering algorithm called Oscope that outputs a list of oscillatory genes clustered according to their shared oscillatory behaviour (similar base cycle).

In **Section B** of this chapter, I start by a statistical description of the input data, and then I present the results found by Oscope. Simultaneously, in **Section C**, I present the biological insights gathered from the functional analysis conducted for the ClockOME genes. Throughout this section I discuss the biological outcomes, as well as some of the technical limitations of the approach. Finally, in **Section D**, I present a list of prioritized candidate oscillatory genes that I propose for subsequent experimental validation in the laboratory.

Section A | *"FrozenChicken"* – Data Normalization Vectors

The microarray data collected for this thesis belongs to three different experiments, as described in chapter 2. As a result, prior to any examination, it is necessary to process together the probe intensity values of the three experiments to make them statistically comparable. This can be achieved by applying the frozen Robust Multiarray Analysis (fRMA) normalization, an elegant procedure that uses data gathered from many different experiments using the same microarray platform to create a dataset of *frozen vectors* representing the variability allowed for each probe set in the array. These vectors are readily available for mouse[110,111], yeast[112] or human[113–116] data, however they were not available for our model organism, i.e. *Gallus gallus*.

Thus, the first step of this project was to generate an R data package named "FrozenChicken " that contains the frozen vectors required for the external normalization of chicken microarray data using the fRMA method. In other words, I created a dataset containing a reference window of probe intensity values coming from chicken microarray experiments (obtained with Affymetrix Chicken Genome Array- the commercially available chicken microarray chip). This window will be used as the normalization reference for the intensity values, and the resulting values correspond to the chicken expression estimates for each gene in each array. The arrays processed with these vectors will, therefore, be directly comparable.

The evident benefit of the *FrozenChicken* package is that it can be directly used in the processing of the raw microarray data, without the need for previous pre-processing. The package was firstly published in Zenodo (DOI:10.5281/zenodo.3765943, https://zenodo.org/record/3765944#.YIsr6LVKiUk) (**Annex 3.2**), and later in bioRxiv (https://doi.org/10.1101/2021.02.25.432894) (**Annex 3.1**). Also, the package was made freely available in GitHubin April 2020, for the benefit of the chicken research community. Consequently, in little over one year (on June 17th 2021), the package has had 95 total views and 1.834 downloads, showing that this package has been actively used by the target research community (**Annex 3.2**).

Section B | Data description

The first step in any data analysis pipeline is to explore the data using descriptive statistics to examine the distribution of the measured variables. Accordingly, in this section I present a detailed statistical description of the samples gathered during the data collection step (see **Chapter II**). Firstly, I describe the gene expression profile of each tissue: Presomitic Mesoderm (PSM) and Limb (**section B.1**), followed by the individual expression profiles for each dataset collected per tissue (**sections B.2** and **B.3**).

B.1 – Profiling gene expression at the level of the tissue

The goal of an initial statistical examination of the data is to determine if the collected data are appropriate for the subsequent analysis. Therefore, I proceeded to explore the statistical properties of the data collected for the PSM and Limb tissues. At this stage, the gene expression values had already been normalized by fRMA (see **Chapter II, section 2**), and log2 transformed. The PSM tissue comprises 32 samples (from two individual experiments[3,9]), and the limb comprises 14 biological samples[89]. **Figure 3.1** shows the gene expression distribution, and their corresponding descriptive statistics metrics, for each tissue (PSM and Limb).

	PSM	Limb
Minimum:	2.491	2.504
1st Quartile:	4.166	4.157
Median:	5.233	5.426
Mean:	5.905	6.072
3rd Quartile:	7.304	7.715
Maximum:	14.659	14.534

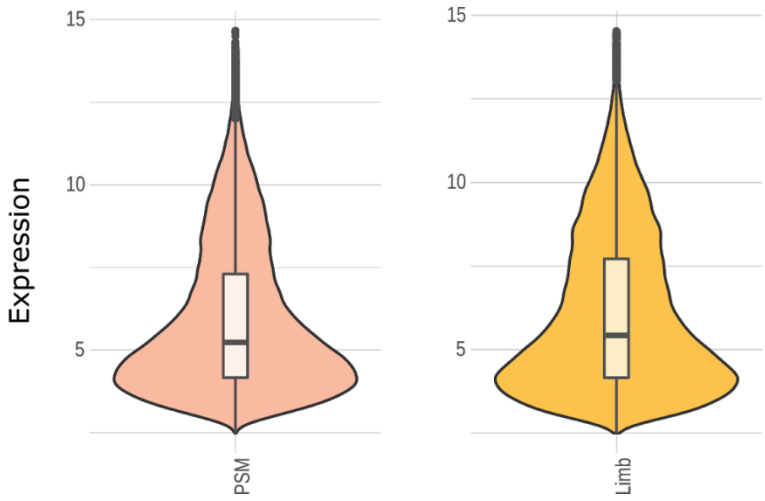

Figure 3.1 | Summary statistics for PSM and Limb samples. Box plot: median gene expression values and quartile ranges; Kernel density distribution: the width of the colored area corresponds to the proportion of the data points located there. Values are normalized (via fRMA) and log2 transformed. n = 32 for PSM; n= 14 for limb.

The distribution of the expression values from both tissues indicates that the data were correctly normalized since both datasets show the same expression range (between 2.5 and 15) (**Figure 3.1 B**). Additionally, this analysis shows that the log2 transformation has made the data more balanced (data not shown), although still showing, for both tissues, some positive

skewness (higher number of genes showing lower expression values, as it is usually observed for transcriptomics datasets). This can be seen in the shape of the violin plot (long tail towards higher expression values); and also expressed in the metrics reported since 75% of the data points (i.e., the third quartile) present expression values below 7.3 in PSM and 7.7 in Limb, which are values falling roughly at half of the maximum expression measured (circa 15). So, the preferred central tendency measure here will be the median since it represents the middle of the dataset.

Even though all samples came from the same animal model (chicken), the data represents 2 different developmental processes, occurring in different developmental stages and from 3 different experiments. Nevertheless, the data distribution is similar in both tissues. This suggests that the normalization with the newly developed FrozenChicken package was successful and thus the transcriptomics data are appropriate for the downstream analysis.

B.2 – PSM data description

To pursue the search for a list of oscillatory genes during somitogenesis I integrated the expression values of the chicken PSM samples from 2 datasets: GSE75798 and E-MTAB-406. A total of 33 microarray assays were gathered and submitted to Quality Control (QC). After removing 1 sample that failed to meet the QC conditions (see **chapter II; Annex 2**), a total of 32 samples was used for further analysis.

To explore the genetic expression from the 32 PSM samples selected, I performed a statistical cross-array comparison (**Figure 3.2**). A table with summary statistics of the log2 transformed PSM expression values is available in the **Annex 4.1**. No significant differences in the statistical distribution between the individual arrays were found (**Figure 3.2 A**). This shows that the fact that the data originates from two independent experiments is correctly normalized by fRMA, and the expected technical biases are effectively corrected, making the gene expressions directly comparable. As expected, the data distribution is positively skewed with the highest density below the median values, similarly to the overall PSM statistical profile (**Figure 3.2**). This reflects the fact that only a small number of genes are highly expressed, whereas the expression of most genes is low.

To visualize this multidimensional dataset with 29,886 probes, a PCA plot was generated (**Figure 3.2 B**). The PSM dataset is dispersed along the first component in two

different groups. Pink dots representing data coming from AE, described in Krol et al., (2011)[3]. These samples are in proximity to each other in this 2D embedding. Blue and green points, that represent the right and left parts of the PSM tissue, were described by Oginuma et al., (2017)[9] and are clustered together. As a result, assays coming from the same experiment were grouped and the main distinction between the two clouds of samples is their experimental origin. This leads to the conclusion that 25.66% of the variability in the data, presented along the Principal Component 1 (PC1), are created due to the batch effect.

PC2 also contributes with 7.43% of the variability between the samples. Overall, the first 2 components of the PCA represent only 33.42% of the variability between the 32 PSM samples. This percentage evidences that the expression profiles between the samples are very analogous, as expected due to the nature of the tissue. Thus, even slight differences in the expression patterns among the samples are considered. Additionally, no evidence to differentiate between the left or right PSM was found in the first two Principal Components (**Figure 3.2 B**). This was expected since the data arise from symmetrical halves of the same embryonic tissue.

In the heatmap, the darkest spots represent low expression values, and as the gene expression value rises, the coloring lightens (**Figure 3.2 C**). The coloring shows no major differences between the two experiments from which the PSM samples were collected. The results reflect the summary statistics for each array as shown in **Figure 3.2 A**, where the majority of genes are expressed lower than the median (purple and blue), and only a small portion is highly expressed (light green to yellow). Cross-array classification divided the samples (columns) into two major groups, the ones from AE and others from GEO. This mirrors the PCA clustering (**Figure 3.2 B**), where the samples were clustered depending on their experiment of origin.

Overall, mild variations were observed in the PSM tissue, mostly due to the batch effect, which was not completely corrected during the pre-processing normalization steps. Since the summary statistics did not show any outliers among the arrays, the downstream analysis was performed using all 32 PSM arrays, with no distinction between the data coming from Krol et al. (2011)[3] or Oginuma et al. (2017)[9].

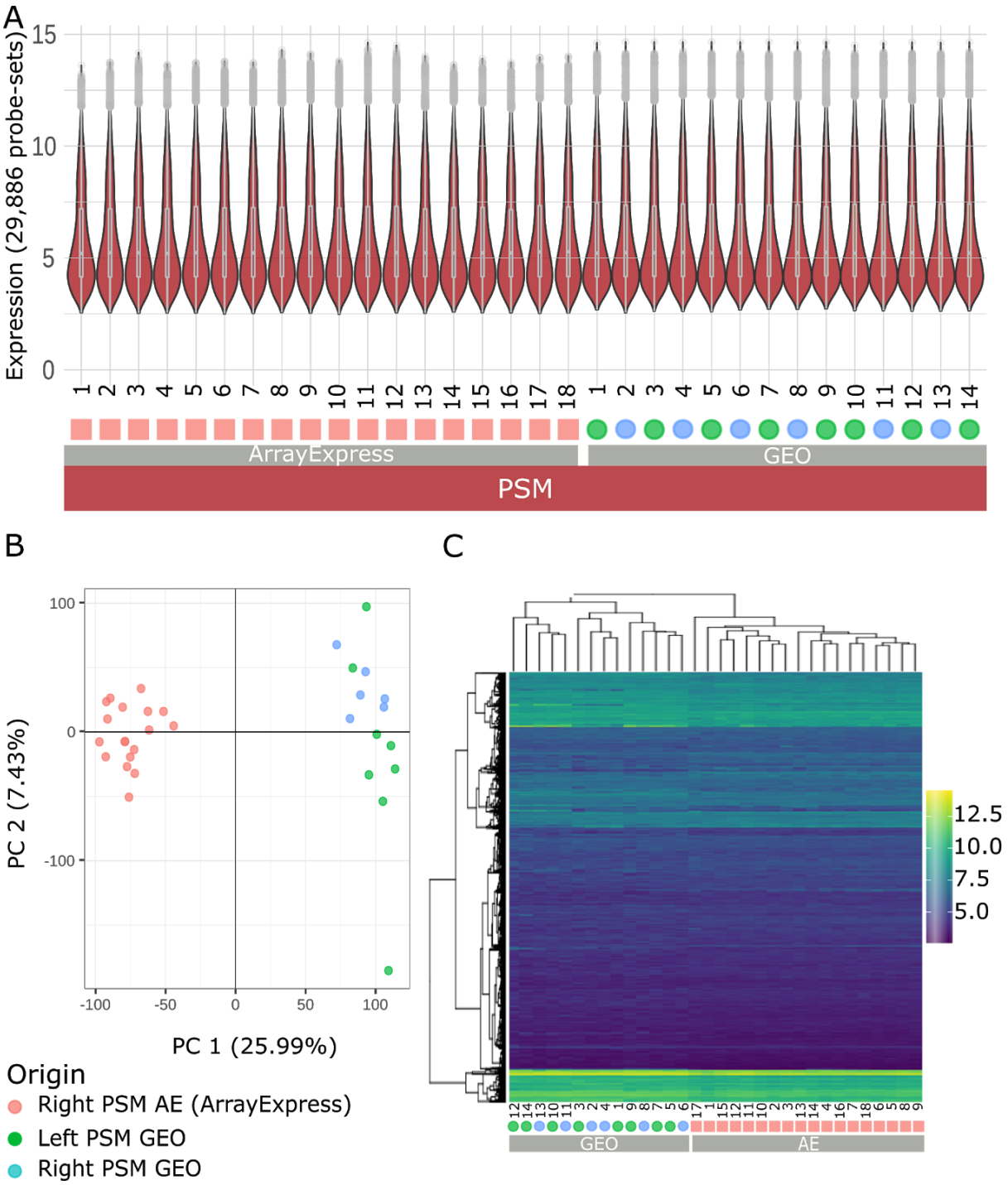

Figure 3.2 | Quality assessment of the initial set of log2-transformed expression values from 32 PSM samples. **A** | Comparison of the distribution of gene expression values. Each violin plot corresponds to one array, and the outline in black shows kernel probability density, i.e. the width of the colored area corresponds to the proportion of the data points located there. The inner boxplot indicates the median and IQR. Sample order has no meaning. **B** | PCA plot showing the cross-array variability across the 2 principal components. The *x-axis* corresponds to the first component and the *y* axis represents the second component. **C** | Heatmap visualization of the expression values. Lateral clustering is performed across genes. Sample clustering is displayed on top. Low gene expression is coded by darker colors, whereas high expressions are color-coded in lighter shades.

B.3 – Limb data description

The search for the oscillatory genes during limb development was performed with data gathered from E-MTAB-4048 deposited in the Array Express (AE) platform. A total of 16 assays were gathered and Quality Controlled. A total of 2 samples were removed, as described in chapter II, and after the pre-processing and annotation, the limb samples were represented by 14 arrays, with 29,886 probes. From the HH20 stage, 7 samples were from the limb bud: 3 from the posterior part (green) and 4 from the anterior part (pink). From the HH24 stage, 5 anterior (blue coded) samples and 2 samples from the posterior part of the limb bud (violet coded) were used (**Figure 3.3**).

Raw summary statistics can be found in the **Annex 4.2**, whereas a graphical representation of the median values, Inter Quartile Ranges (IQRs), and the density distribution for each array are shown in **Figure 3.3 A**. Following the overall limb profile, (**Figure 3.1**), the cross array comparison showed similar expression level distributions, with a right-skewed curve for all arrays. Once again, this shows that only a few genes are highly expressed in the dataset whereas most genes are expressed below the median.

To investigate the relationship between the samples, complementary graphics are presented (**Figure 3.3 B** and **C**). In the PCA, 29.71% of the variability between the limb samples was found in the first component (PC1) (**Figure 3.3 B**). The PC1 divides the samples into 2 groups according to the HH stage. On the left side, pink and green colors are closely grouped, representing the HH20 anterior and posterior parts of the limb bud, respectively. The right side shows the arrays coming from the HH24 stage (blue points representing the anterior part and violet points representing the posterior parts of the anterior limb bud).

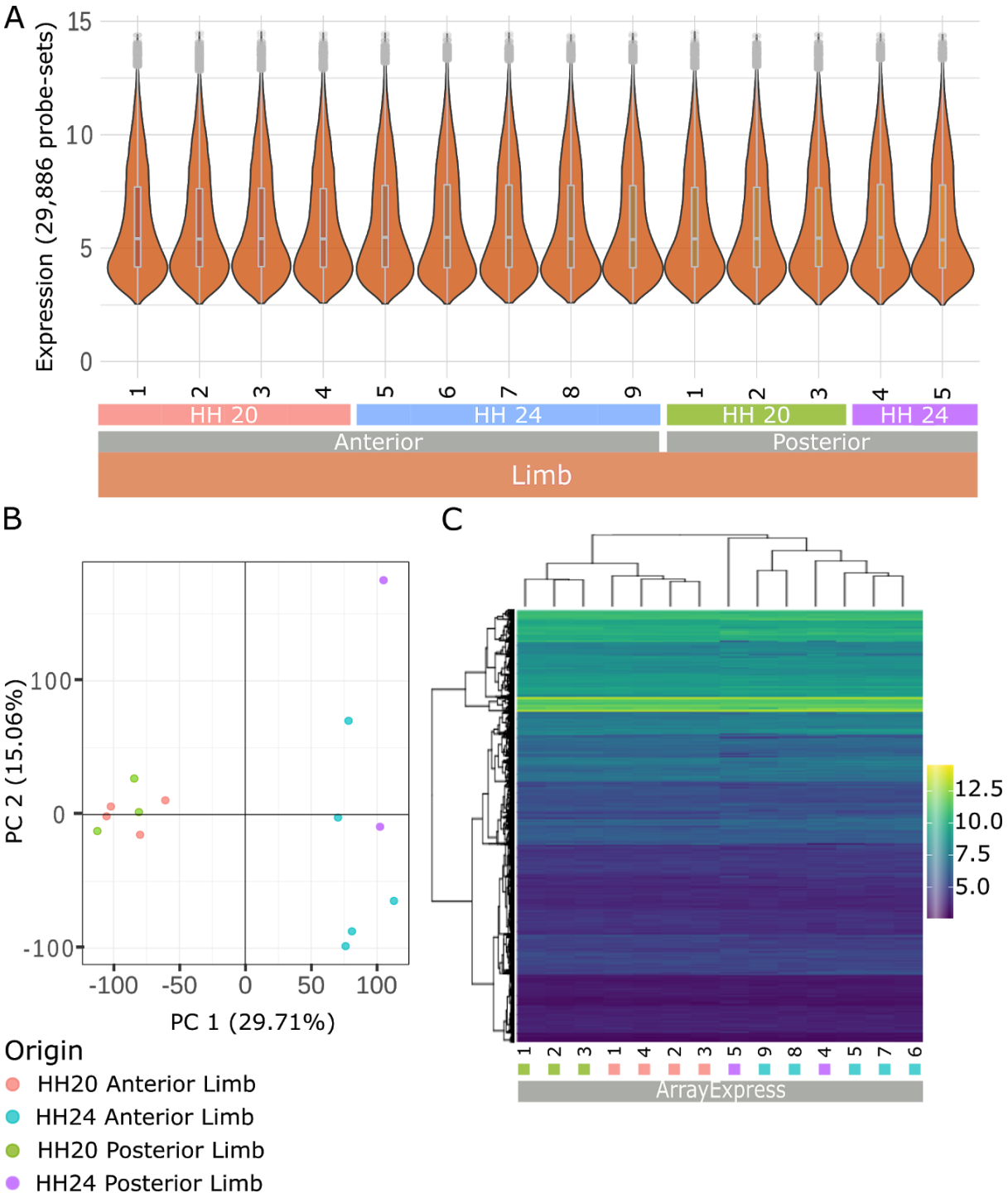

Figure 3.3 | Quality assessment of the initial set of log2-transformed expression values from Limb samples. **A** | Comparison of the distribution of gene expression values. Each violin plot corresponds to one array, and the outline in black shows kernel probability density, i.e. the width of the colored area corresponds to the proportion of the data points located there. The inner boxplot indicates the median and IQR. Sample order has no meaning. **B** | PCA plot showing the cross-array variability across the 2 principal components. The *x-axis* corresponds to the first component, and the *y* axis represents the second component. **C** | Heatmap visualization of the expression values. Lateral clustering is performed across genes. Sample clustering is displayed on top. Low gene expression is coded by darker colors, whereas high expressions are color-coded in lighter shades.

The second component (PC2) also explains an additional 15.06% of the variability between the data points, represented along the y axis. The left group containing only the HH20 samples is less dispersed when compared to the right cluster, exhibiting only samples from HH24. This may suggest that during the initial developmental stage of the limb the genetic expression across the anterior wing bud was more similar. On the other hand, the samples from the later developmental stage (HH24) have their genetic profiles more dispersed.

Heatmap visualization was also performed for this tissue (**Figure 3.3 C**). Throughout the color representation of the arrays, no major differences were found. However, similar to what was observed in the PCA, the clustering amongst the samples branched the arrays into two groups, dividing the data based on their developmental stage (HH20 versus HH24) (**Figure 3.3 B**).

Even though the arrays were retrieved from the same experiment (so no batch effects are expected), the samples present slightly different transcriptomic profiles. As the limb samples were not differentially treated, even smaller genetic changes are contributing to the observed variability between them. Overall, these data show that the major difference between the samples has a biological origin and resides in the fact that 2 different developmental stages are present: HH20 is evidently different from HH24 (two stages separated by more than 1 day of development).

In general, since the summary statistics for the limb samples did not show any outliers, all 14 samples were used to search for oscillatory genes.

Section C | Identification of oscillatory genes

This study intended to infer the oscillatory trajectory of gene expression in chicken embryos during early periods of development (somitogenesis and limb development). As previously mentioned, the transcriptomics information in the form of raw .CEL files was gathered from microarray datasets collected from publicly available platforms (ArrayExpress[87] and Gene Expression Omnibus (GEO[88]). A set of 14,555 fully annotated genes (represented by 29,886 probe sets) was used.

Next, to find cyclic genes - collectively called the ClockOME - I employed a pseudotime inference method called Oscope[76]. As required by the algorithm, input data was normalized, followed by filtering. Only highly variable genes (HVG) were selected to detect stronger oscillatory signals and to reduce the computation time of the study. In the end, 3 clusters of oscillatory genes were recovered: 2 in the PSM (PSM K1 and PSM K2) with 106 and 32 genes respectively, and 1 in the limb dataset (Limb K1) with 163 genes (**Figure 3.4**).

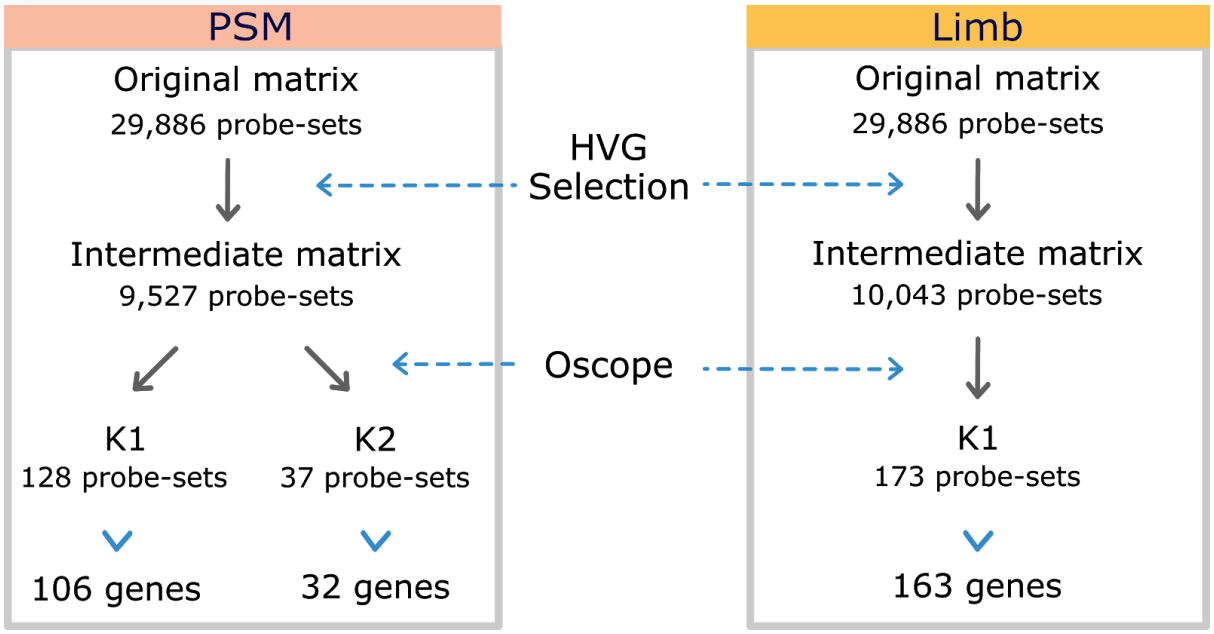

Figure 3.4 | Workflow applied to find oscillatory genes. Initial data in PSM and Limb tissues contained 29,886 probes. Only highly variable genes (HVG) were selected as input for the Oscope algorithm. The output from Oscope found 2 clusters of oscillatory genes in the PSM (PSM K1 and PSM K2) with 106 and 32 genes, respectively, and 1 cluster was found in limb (Limb K1) with 163 genes. HVG = Highly Variable Genes; K = cluster; PSM = presomitic mesoderm.

To evaluate the biological significance of these results, I explored the functional enrichment (FE) of the ClockOME genes using the Gene Ontology (GO) knowledgebase[117], a commonly used repertoire of functional terms (controlled vocabulary) regarding the functional

classification of genes and their products (proteins). GO annotations consist of three types of functional information, each stored in an individual database: (i) biological processes (GO BP) that consists of terms describing functions related to the biological mechanisms involving multiple genes, for example, signalling pathways, or DNA transcription; (ii) molecular functions (GO MF), that represent terms more specific for the molecular activities performed by the gene products, such as kinase activity, or histone deacetylase; and (iii) cellular component (GO CC), which has terms to describe the cellular anatomy and sub-cellular localization, rather than processes or pathways, such as mitochondrion, or plasma membrane[118].

For the functional analysis, firstly, I assigned the ClockOME genes to their respective GO categories within each database, separately for each cluster. As each gene has typically associated a unique identifier that maps between databases, to conduct the downstream FE analysis, the "ENSEMBL" accession number was used as identifier (ID) for each gene. It is important to note that each gene can belong to different categories, which results from the hierarchical structure of the GO annotation. Therefore, several biological mechanisms have a great number of genes associated, while a specific GO category involves only a small number of genes, related exactly to their biological function. To narrow down the magnitude of the resultant biological information, I only extracted GO annotations that have 5 or more hits. Likewise, only the top 20 categories (with highest numbers of hits from our clusters) were used to visually represent the FE results in the next sections. However, the full list of GO annotations can be found in the **Annex 8**.

In parallel, to explore the interaction between the proteins coded by the ClockOME genes, for each of the clusters found by Oscope, I created a network based on the STRING[119] database (Search Tool for the Retrieval of Interacting Genes/Proteins), that holds a wide range of functional interactions between proteins, originated from experimentally published data and computational predictions. The STRING database was queried for protein interactions using the gene symbol as query. These interactions can be direct (when the proteins interact directly with each other, for example as a pair of ligand and receptors), or indirect (when proteins are present in the same molecular pathway, or operate on the same substrate, however, do not interact physically). As such, I obtained several functional interaction (FI) networks (discussed below). These consist of simplified abstract representations (with no directionality), where nodes represent proteins, and edges (the links between nodes) represent a functional connection

between proteins. In order to expand the network to include other likely interactors, I allowed STRING to include up to 20 extra interacting proteins (onwards called interactors). This approach allows the finding of potential new interactors that could also be oscillatory or even regulators of the oscillatory processes described in this work. The proteins from our clusters that have no known interactions/interactors in STRING are not included in any networks, and are called orphans.

The resultant interaction networks are presented in the **Annex 9** and were further subdivided into smaller modules based on the strength of their connections. This aims at extracting biologically meaningful modules of functional interaction. Each of the modules, containing ClockOME gene-coded proteins and predicted interactors, was further studied individually. For each of the proteins contained in the modules, functional information based on GO BP categories was manually extracted from GeneCards[120], an integrative database with human genetic information. The most common GO BP category in each module was used for the overall functional characterization of the module, and used as module name (in the case of multiple equally enriched GO BP categories, the overall underlying mechanism was used to name the modules).

This section is structured in the following manner: I start by presenting the Oscope results in the form of a base cycle for each of the clusters. Then, I briefly discuss the list of candidate genes that, according to our results, might be involved in the timely regulation of somitogenesis and limb development in the chicken embryo. Next, I evaluate the biological significance of these results by showing and discussing the FE (functional enrichment) and FI (functional interactions) found. Finally, I compare my ClockOME genes to the list of cyclic genes previously reported in the literature, and conclude this section by presenting a ranked list of 6 candidate genes that I propose for experimental validation in the lab.

C.1 – PSM Cluster 1 Analysis

C.1.1 – The base cycle of PSM K1

The Cluster 1 of the PSM tissue (PSM K1) is composed of 106 genes (measured by 128 probe-sets) listed in the **Annex 6.1**. The descriptive statistics of the expression data for the genes in this cluster is presented in **Annex 5.3**, showing comparable summary metrics for all genes. Oscope calculated the base-cycle (i.e., the sample order) for this cluster (**Figure 3.5 A**), and all the genes were individually profiled by this pseudo time ordering algorithm (**Annex 7.1**). The base cycle was reconstructed on a gene expression rescaled space (values between 0 and 1), and for each gene individually, on the original expression values.

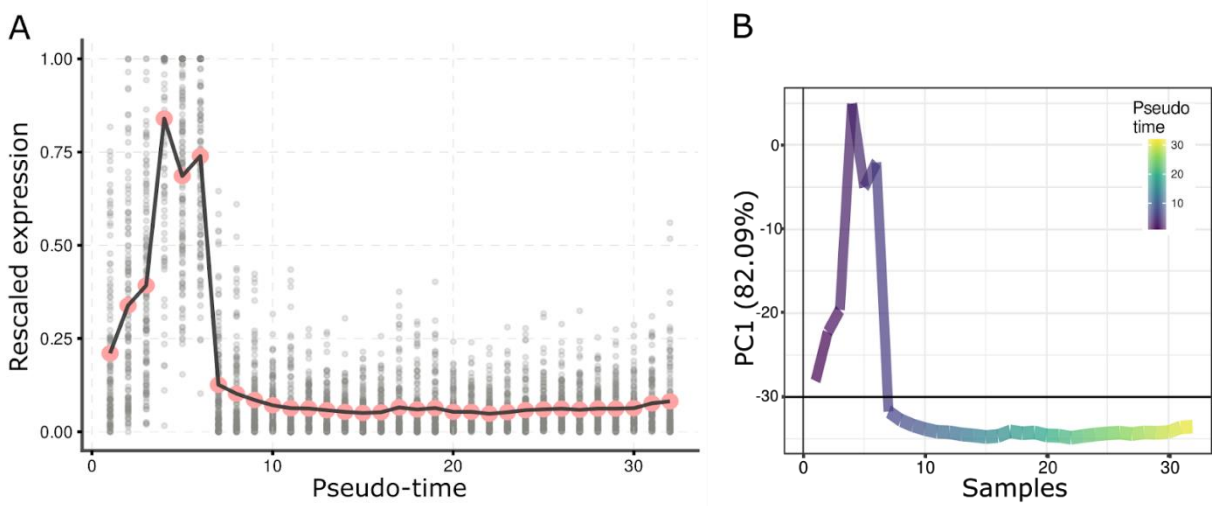

Figure 3.5 | Oscillatory trajectory recovered by Oscope for PSM cluster 1. A | Scatter plot of mean expression values for each sample (pink points), based on 128 probe-sets (background grey points). Sample order is given by the pseudo time ordering calculated by the ENI algorithm from Oscope (n = 32 samples). Expression values were rescaled between 0 and +1. **B** | PCA plot of PSM K1 genes showing the data variability contained in these genes. PCA was calculated with the rescaled expression values, showing that the PC1 explains 82.09% of the data variability. Samples are ordered by their base cycle ordering inferred by Oscope, where the 1st sample is colored dark violet, following a color gradient to the last sample in yellow.

The oscillation recovered for this cluster is characterized by a small rise in the mRNA concentration, drawing the up phase of the oscillation, with all the 106 genes from the cluster in the first 6 samples displaying this increased expression (**Figure 3.5 A**). Meanwhile, most of the samples (the other 26) contribute to the downstate of the oscillations (i.e., all 106 genes from this PSM cluster display a low level of expression in these samples). Such expression profile is shared by all the genes present in the PSM cluster 1, (grey dots shown in the plot) (**Figure 3.5 A**). By ordering the PSM samples in accordance with the inferred base cycle, a path was constructed based on the PCA analysis, using the rescaled information of the 128 probe-

sets from this cluster. Noteworthy, the reconstructed path took the first component variability and recapitulated the base cycle profile (**Figure 3.5 B**). With 82.09% of the genetic variability contained in the PC1, it is possible to conclude that, indeed, the samples attributed by the Oscope model to the up phase of the cycle are genetically distinct from the samples in the down phase of the trajectory (**Figure 3.5**). Individually, genes in PSM K1 behave with a similar profile (**Annex 7.1**).

Particularly relevant is the fact that this cluster includes *Mesp2*, a gene that has been experimentally found to oscillate during somite segmentation[54,121]. *Mesp2* is a bHLH TF (basic Helix-Loop-Helix Transcription Factor) acting downstream of the Notch pathway (See **Chapter 1, section B**)[54,121]. In this dataset, it is expressed according to the **Figure 3.6**. Moreover, in this cluster, genes known to be important for embryo development and morphogenesis were grouped together, with some of the features characteristic to somite segmentation: *MEOX1*[122]and *MEOX2*[123] (involved in somite morphogenesis)[120]; *PAX7* (myogenic marker expressed in somites[124,125]); *TCF15* (a bHLH TF necessary for the epithelialization of somites [126]); *TBX22* (member of T-box family of TFs indispensable for morphogenesis[127]), also *tbx6*[128] and *tbx16*[3] are known to oscillate in the PSM in the zebrafish animal model); *RIPPLY1* (required for somite segmentation[129]); *HEY1* (Notch induced bHLH TF of HES pathway[130]); and *TWIST1* (TF of the bHLH family, also part of the Notch mediated HES network[131]). Mutations in these genes may result in skeletal defects and various syndromes related to segmentation abnormalities[120,132].

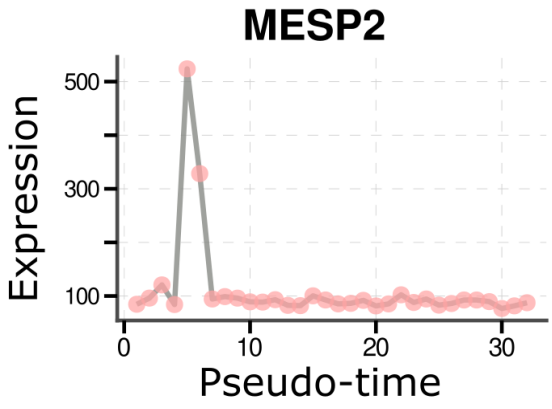

Figure 3.6 | Mesp2 gene expression in *Gallus gallus* PSM. The expression is shown along the pseudo-time order recovered by Oscope (samples ordered according to the predicted oscillation).

Another interesting set of genes included in the PSM K1 seem to be involved in neuronal development and signaling (for example, *PENK*[133], *NPY*[134], *GRIK3*[135], *GRM4*[136], *SOX8*[137],

LZTS3[138], and *ROBO2*[139]). Moreover, genes enrolled in different signal transduction pathways (*MAPK11*[140], *RHOU*[141], *BMP3*[142]) and cell-cell communication mechanisms (*JPH1*[143], *MMP9*[144], *PTCH1*[145]) were also found in this cluster. As a result, it was imperative to conduct a more global search for the functional properties, comprehending all the gene information available from prior research.

C.1.2 – Functional analysis | Morphogenic processes are enriched in PSM K1

After assigning GO annotations to the genes from PSM K1, not all the probes were annotated, meaning that for some candidate oscillatory genes there is no functional information. The full functional annotation information for each GO database is presented in **Table 2**. In total 143 GO BP, 24 GO MF, and 19 GO CC terms were found to be enriched in PSM cluster 1 (**Annex 8.1**).

Table 2 | Functional enrichment for GO terms for PSM K1 genes.

	GO Database	Database size (*)	# K1 genes present in database (**)	# Enriched GOterms
PSM K1 (n=106)	GO BP	3901	34	143
	GO MF	3606	32	24
	GO CC	4101	37	19

***Total number of chicken genes that are annotated in the database; ****Intersection between database size and the 106 genes present in K1; **GO** = Gene Ontology; **GO BP** = GO Biological Process; **GO MF** = GO Molecular Function; **GO CC** = GO Cellular Component.

To visualize the magnitude of ClockOME genes that contribute to each GO category, a treemap visualization is represented for top 20 GO terms individually in the **Figure 3.7** panel A, where GO BP are in blue, GO MF are in red, and GO CC are in green. A graphical counterpart is presented in panel B, where the information is based on the p-value of the enrichment analysis (**Figure 3.7 B**). Results of the FE analysis shows enrichment for biological processes related to morphogenesis and development. Also, signaling processes are enriched, namely the FGF pathway, which is known to be strongly related to the somitogenesis process[3,5,7,8,47,51,146]. Furthermore, genes in the PSM K1 are enriched for binding, a molecular function vital for molecular signaling. Whereas, in the GO CC classes, PSM K1 genes seem to be particularly localized to the plasma membrane, which is in accordance with the other results, since signaling proteins such as ligands and receptors are mostly found in the plasma membrane (**Figure 3.7**).

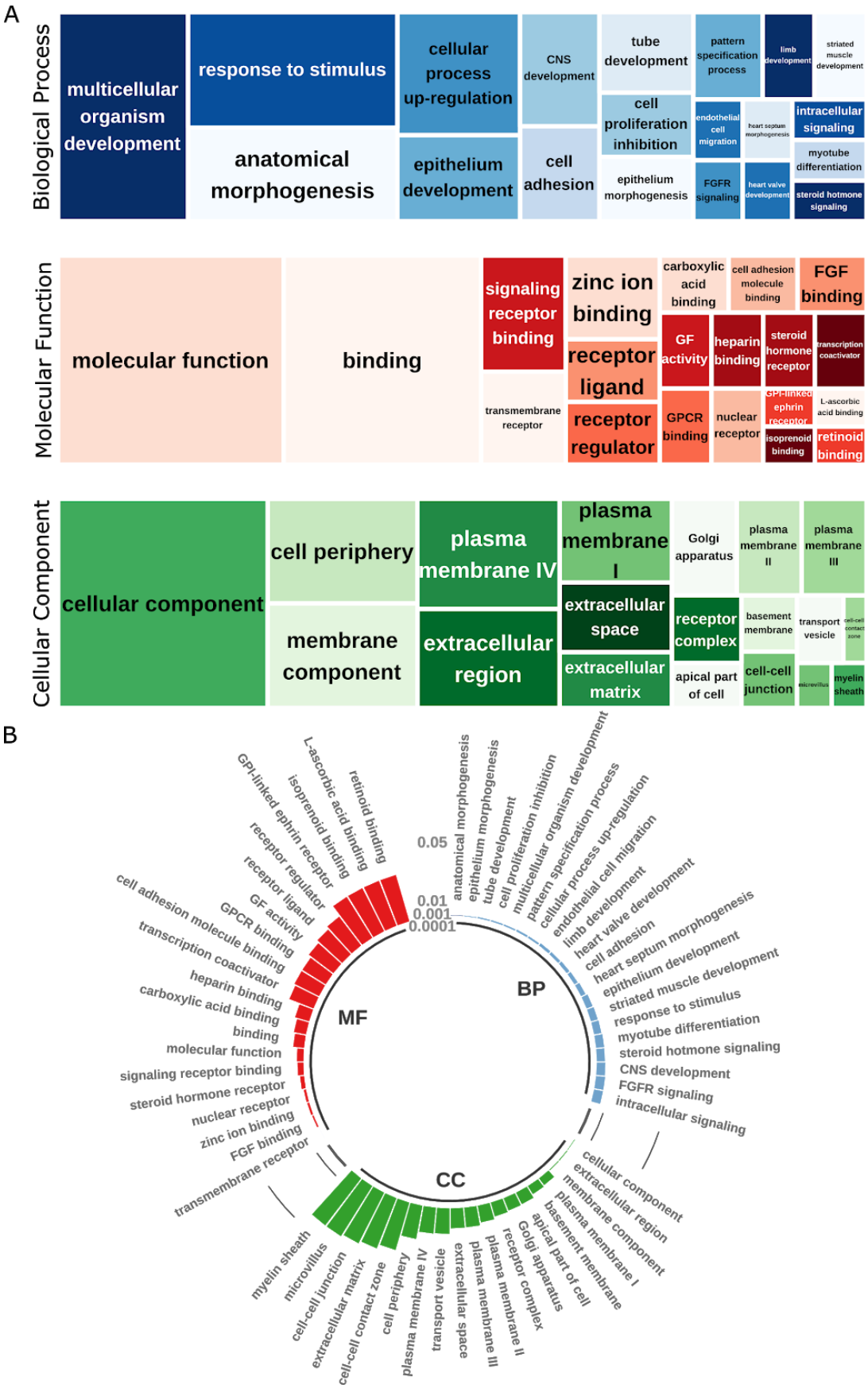

Figure 3.7 | Intra-cluster comparison of GO categories enriched for the PSM K1 genes. Continued on next page.

Figure 3.7 | Continued. **A |** Blue – Biological Processes (GO BP) with 143 categories in total; Red – Molecular Functions (GO MF) with 24 categories in total; Green – Cellular Components (GO CC) with 19 categories in total. Rectangle sizes are proportional to the number of ClockOME genes found for each GO category. The total number of genes from the ClockOME list annotated with terms from each GO database are: GO BP genes = 34; GO MF genes = 32; GO CC genes = 37. **B |** Circular plot shows the p-values for each GO term. All p values shown are below 0.05. The smaller the bar, the smaller the corresponding p-value. For simplicity only the top 20 enriched categories are displayed.

To gather more functional information from the PSM K1 gene networks, a STRING network was obtained (**Annex 9.1**). The PSM K1 network was subdivided into 4 smaller modules as shown in the **Figure 3.8**, which were named according to the most common biological process associated with the nodes (**Table 3**). The ClockOME proteins are pink-colored, while the predicted interactors are shown in blue.

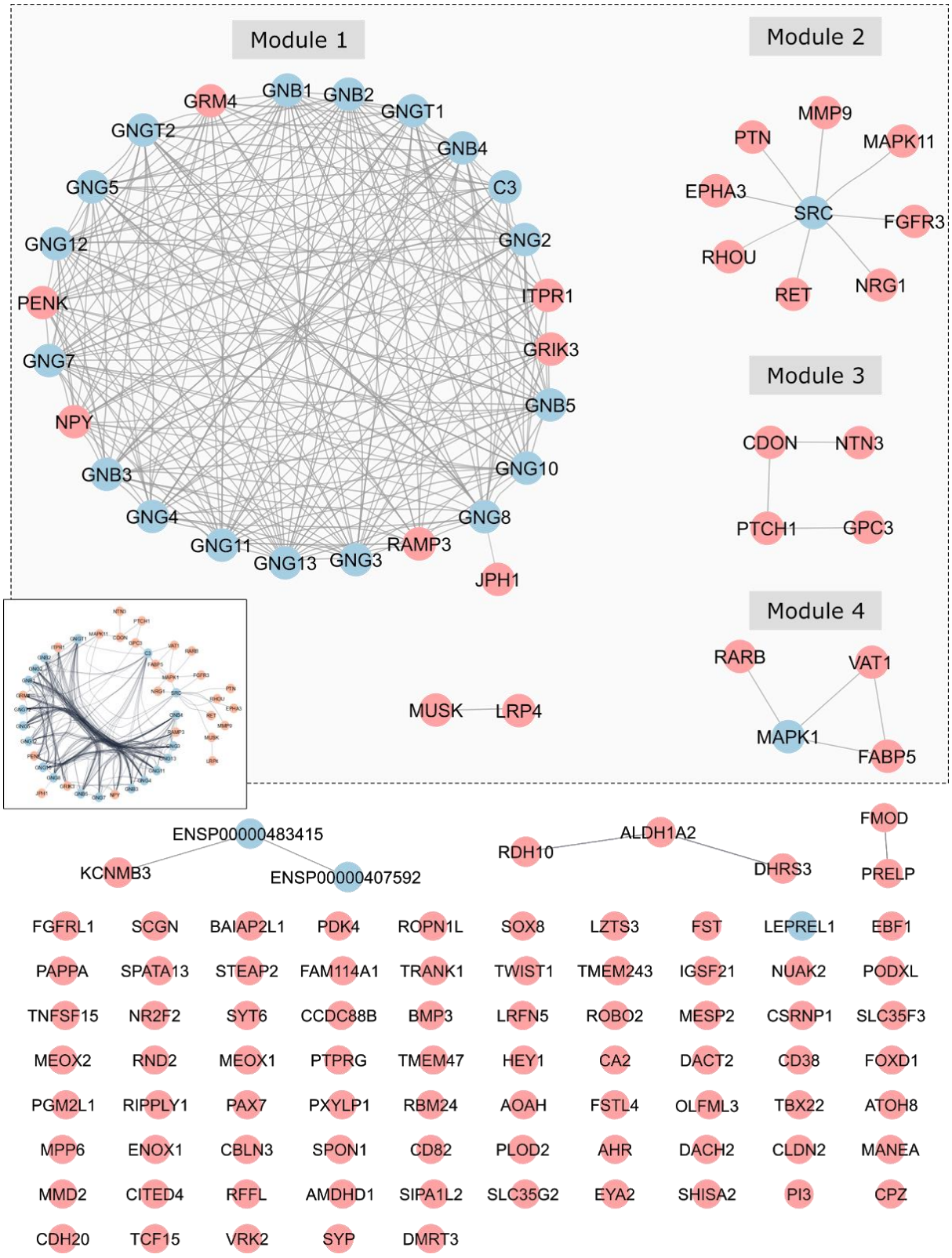

Figure 3.8 | Functional interaction network visualization for proteins coded by the genes from the PSM K1, with 20 additional predicted interactors. Four modules were found after MCL Clustering. Pink – ClockOME interactors; Blue – additional predicted interactors; Left-middle panel shows the original network with 44 nodes before the split into smaller modules using MCL clustering. Module 1 = 24n + 262e; Module 2 = 9n + 8e; Module 3 = 4n + 3e; Module 4 = 4n + 4e; n = nodes; e = edges.

Table 3 | Functional interaction modules in PSM K1.

Cluster	Module	GOterm	Extended name of the GOterm	ClockOME genes only (*)	Full module (**)	Functional name
PSM K1	1	GO:0007186	G protein-coupled receptor signaling pathway	4/7	22/25	GPCR signaling
	2	GO:0006468	protein phosphorylation	4/8	5/9	Phosphorylation
		GO:0016310	phosphorylation	4/8	5/9	
	3	GO:0009887	animal organ morphogenesis	3/4	-	Morphogenesis
	4	GO:0043312	neutrophil degranulation	3/4	-	Neutrophil degranulation

*The number of ClockOME genes attributed to the GOterm out of the total number of ClockOME genes present in the respective module; **The number of genes attributed to the GOterm accounting for the full size of the module (ClockOME genes and the predicted interactors); GPCR = G protein-coupled receptors.

Module 1 is strongly related to the GPCR (G protein-coupled receptors) signaling pathway, which is known to be a molecular switch due to the ability of activation and inhibition of molecular pathways, involved in signal transduction from the exterior[147] (**Table 3**). The second module shows genes related to phosphorylation, a crucial process for the transduction of the signal in the cell, also present in GPCR signaling. Modules 3 and 4 are smaller and are composed of genes responsible for morphogenesis and neutrophil degranulation, respectively (**Table 3**).

Noteworthy, in module 4, is the fact that 3 out of 4 genes are related to the neutrophil degranulation category (**Table 3**). Could this mean that this process is somehow involved in somitogenesis? And if so, is it an oscillatory process active in the chicken PSM? Moreover, how far can one transpose the knowledge regarding the immune system between different animal models? Perhaps this finding does not necessarily mean that this aspect of the immune system is dynamic, but instead can be explained by the enriched protein phosphorylation activities involved in various biological processes like segmentation, cell migration, cell differentiation, or neutrophil degranulation (highlighted by the SRC protein family and GPCR signaling). Additionally, this could also reflect the fact that the same genes known in adults to be involved in immune functions (not yet present in the embryo), might present different roles during the developmental stages.

Overall, these results show that the PSM K1 holds multiple classic PSM marker genes. In general, morphogenesis is the dominant biological process while from the molecular

perspective, the features that belong to this cluster are various signal transduction pathways, regulated mostly by protein binding and phosphorylation.

C.2 – PSM Cluster 2 Analysis

C.2.1 – The base cycle of PSM K2

The second PSM cluster - PSM K2, groups 37 probe-sets that represent 32 genes (**Annex 6.2**). The descriptive statistics pertaining to these genes is presented in **Annex 5.4**, showing no significant differences between the samples. Similarly to the aforementioned results, Oscope calculated the base cycle for the PSM K2, which is displayed in **Figure 3.9**. PSM K2 genes were individually profiled along the pseudo-time and presented in the **Annex 7.2**.

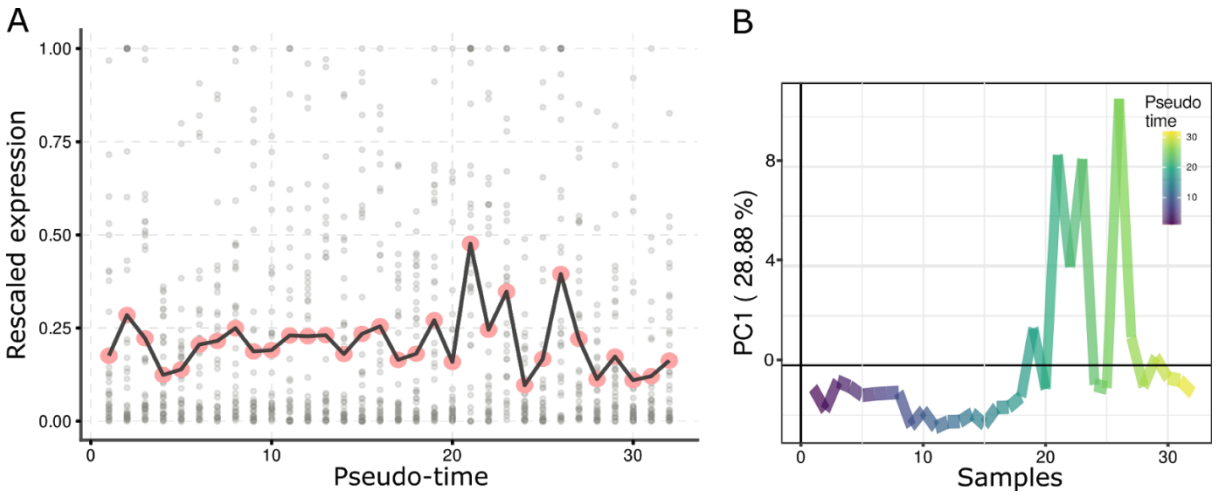

Figure 3.9 | Oscillatory trajectory recovered by Oscope for PSM cluster 2. A | Scatter plot of mean expression values for each sample (pink points), based on 37 probe-sets (background grey points). Sample order is given by the pseudo time ordering calculated by the ENI algorithm from Oscope (n = 32 samples). Expression values were rescaled between 0 and +1. **B |** PCA plot of PSM K2 genes showing the data variability contained in these genes. PCA was calculated with the rescaled expression values, showing that the PC1 explains 28.88 % of the data variability. Samples are ordered by their base cycle ordering inferred by Oscope, where the 1st sample is colored dark violet, following a color gradient to the last sample in yellow.

In comparison to cluster 1, the trajectory of the PSM cluster 2 seems to be noisier, showing more than one peak and less uniformity in the lower values. This outcome emerges because the expression values of the genes composing the cluster are not so uniform as in PSM K1, as can be seen by the cloud of grey points spread in the background of the expression profile plot (**Figure 3.9 A**). Additionally, the dimensionality reduction analysis (PCA) based on these 37 probe-sets, also confirms that the PSM K2 samples are not very dissimilar genetically, with only 28.88% of the data variability accommodated by the PC1 (**Figure 3.9 B**). Therefore, the

oscillation recovered by Oscope for PSM K2 is not so well sustained as for PSM K1. One possible explanation for such a misfit base cycle could be the fact that the genes present in PSM cluster 2 present a richer oscillatory pattern that cannot be recapitulated in detail with such a small number of samples (14 microarrays). Another likely possibility is that their oscillation departs too much from the simple sinusoidal function assumed by Oscope, and therefore are not well fitted by this model.

Despite showing a less convincing oscillatory pattern, and heterogeneous functional enrichment, some interesting genes showed up in this cluster, namely *PITX1*[148] and *PITX2*[149], that are involved in the establishment of left-right asymmetry, and diverse other morphogenetic processes. Also, a surprising finding was that the protein-coding genes for the hemoglobin-subunit were also present in this cluster (*HBA1*[150], *HBZ*[151]), although, to the best of our knowledge, never before described as oscillatory.

C.2.2 – Functional analysis | Cell-cycle, and metabolism are enriched processes in PSM K2

Given that the chicken genome still presents some incomplete genome annotations, I came across some genes in PSM K2 that were not yet functionally annotated. **Table 4** presents the FE condition, based on the genes present in PSM K2. A total of 79 GO BP, 22 GO MF, and 12 GO CC categories were found to be enriched in this cluster (**Annex 8.2**).

Table 4 | Functional enrichment conditions for PSM K2.

	GO Database	Database size (*)	K2 genes present in the database (**)	Enriched GOterms
PSM K2 (n=32)	GO BP	3901	14	79
	GO MF	3606	13	22
	GO CC	4101	15	12

*****Total number of chicken genes that are annotated in the database; ******Intersection between database size and the 106 genes present in K1; **GO** = Gene Ontology; **GO BP** = GO Biological Process; **GO MF** = GO Molecular Function; **GO CC** = GO Cellular Component.

The FE results can be observed in **Figure 3.10**. **Panel A** categorizes individual annotation terms based on the amount of ClockOME genes attributed to each term, while panel B shows the top 20 enriched GO terms based on the p-value. Enrichment of biological processes (in blue) involved in mitotic spindle and organelle dynamics suggests that the PSM K2 may

hold genes related to the cell cycle. Additionally, the enrichment in oxygen transport processes indicates that the genes in this cluster are also related to metabolic events taking place inside the cell. In red, the molecular functions enriched are globally related to binding to proteins and DNA, with additional emphasis on oxygen-related functions (binding and transportation). Finally, PSM K2 genes are enriched for diverse cellular structures, such as the Golgi apparatus and the cytoskeleton (in green), as well as cellular architecture and organelle trafficking (GO BP) (**Figure 3.10**).

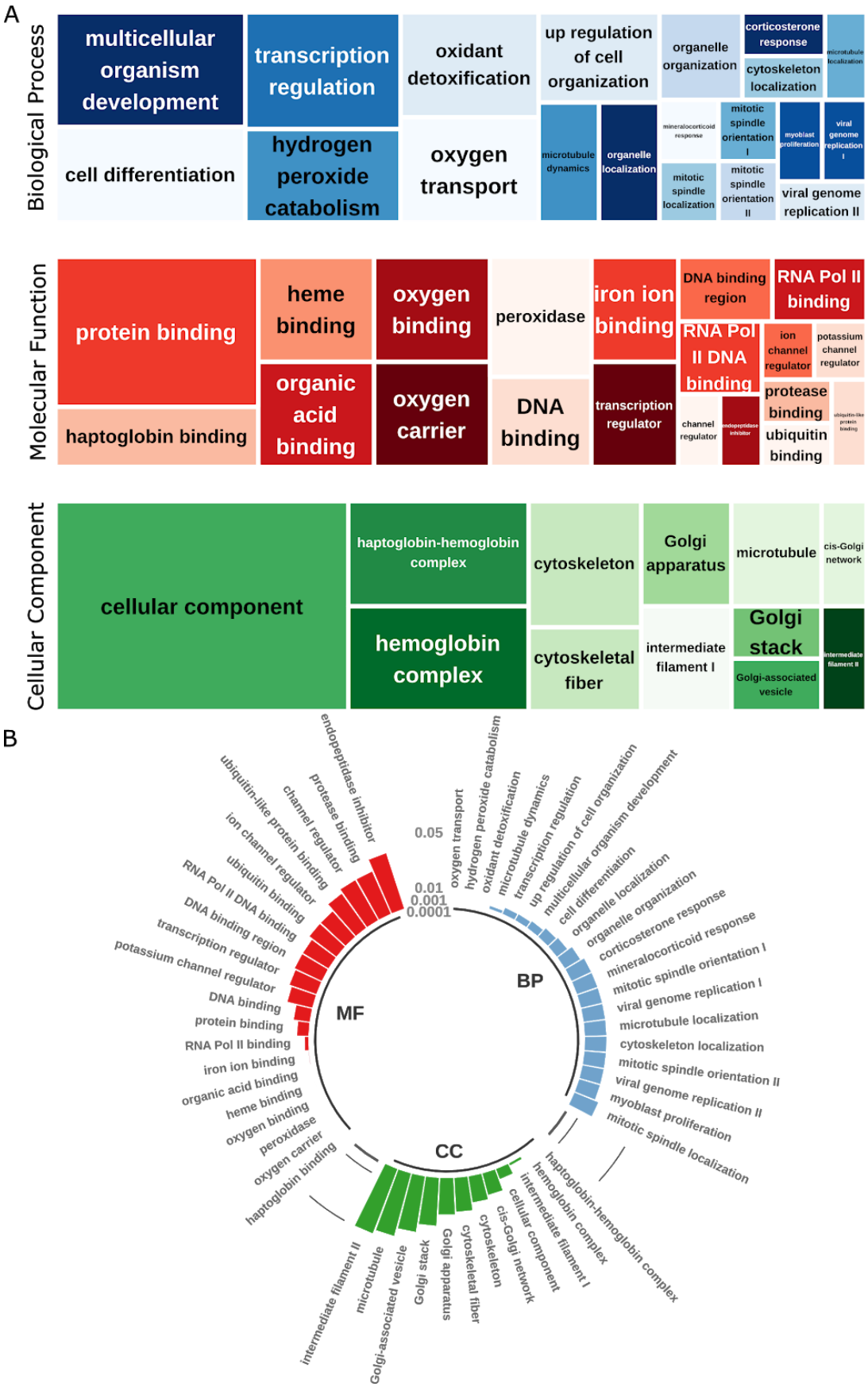

Figure 3.10 | Intra-cluster comparison of GO categories enriched for the PSM K2 genes. A | Blue – Biological Processes (GO BP) with 79 categories in total; Red – Molecular Functions (GO MF) with 22 categories in total; Green – Cellular Components (GO CC) with 12 categories in total. Continued on next page.

Figure 3.10 | Continued. Rectangle sizes are proportional to the number of ClockOME genes found for each GO category. The total number of genes from our list annotated with terms from each GO database are: GO BP genes = 14; GO MF genes = 13; GO CC genes = 15. **B |** The circular plot shows the p-values for each GO term. All values shown are below 0.05. For simplicity only the top 20 enriched categories are displayed.

The STRING functional interactions present in PSM K2 gene products are displayed in **Figure 3.11**. The initial network (**Annex 9.2**) was subdivided into 3 smaller modules with the most common biological processes being the ones shown in **Table 5**.

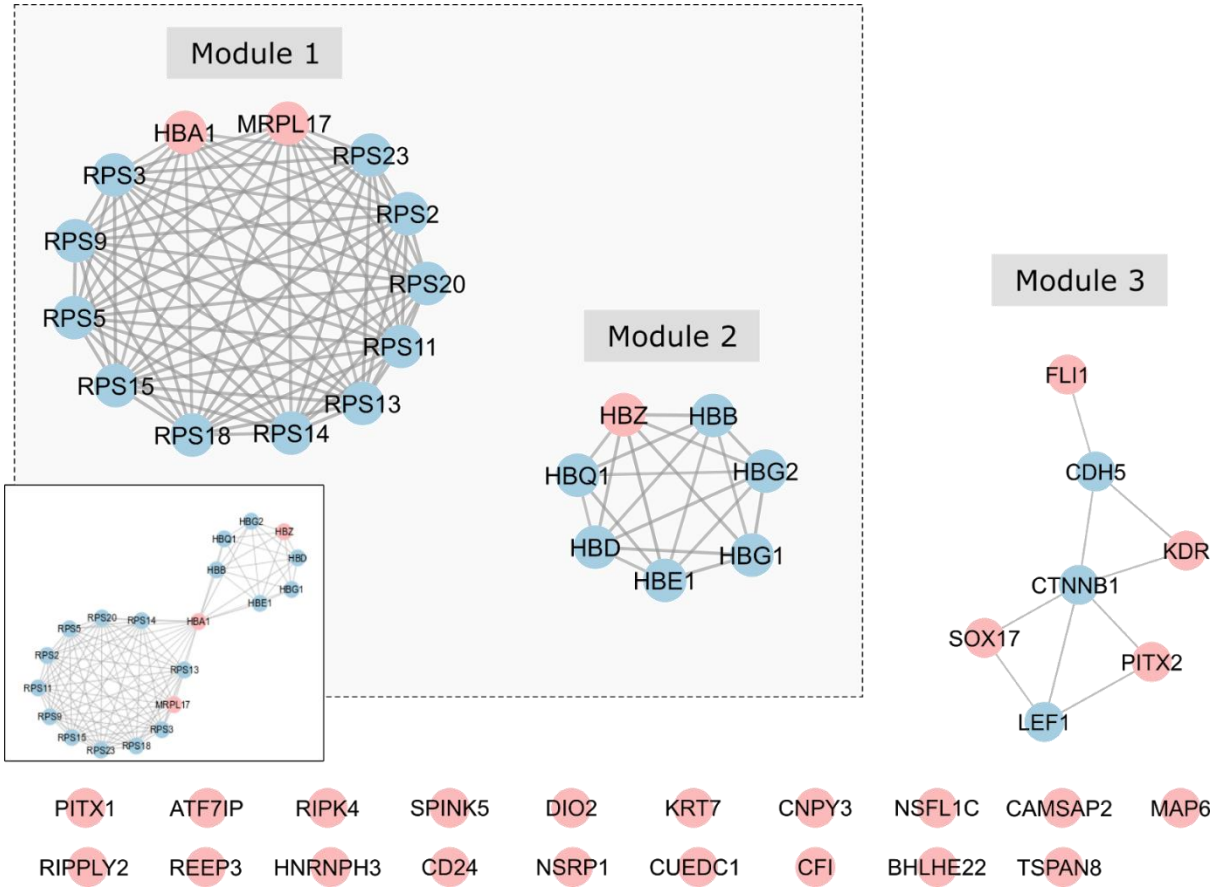

Figure 3.11 | Functional interaction network visualization for proteins coded by the genes from PSM K2, with 20 additional predicted interactors. Three modules were found after MCL Clustering. Pink – ClockOME interactors; Blue – additional predicted interactors; Left-middle panel shows the original network with 20 nodes before the split into smaller modules using MCL clustering. Module 1 = 13n + 77e; Module 2 = 7n + 20e; Module 3 – 7n + 9e; n = nodes; e = edges.

Module 1 is strongly related to mRNA translation processes, where all 11 predicted interactors are ribosomal protein subunits, highly interconnected (**Table 5**). Module 2 clusters genes are related to oxygen transport, while module 3 is composed of genes responsible for positive regulation of transcription, possibly related to cell differentiation (**Table 5**).

Table 5 | Functional interaction modules in PSM K2.

Cluster	Module	GOterm	Extended name of the GOterm	ClockOME genes (*)	Full module (**)	Functional name
PSM K2	1	GO:0006412	translation	1/2	12/13	mRNA translation
	2	GO:0015671	oxygen transport	1/1	7/7	Oxygen transport
		GO:0042744	hydrogen peroxide catabolic process	1/1	7/7	
	3	GO:0098869	cellular oxidant detoxification	1/1	7/7	Morphogenesis
	4	GO:0006355	regulation of transcription, DNA-templated	3/4	4/7	Transcription regulation

***The number of ClockOME genes attributed to the GOterm out of the total number of ClockOME genes present in the respective module; ****The number of genes attributed to the GOterm accounting for the full size of the module; **GPCR** = G protein-coupled receptors.

Taken altogether, the PSM K2 is composed of candidate cycling genes related to mRNA dynamics, whether involved in transcription or translation. This may be indicative of active cell differentiation and proliferation. Additionally, genes participating in processes related to cytoskeletal dynamics are represented, which may represent the extensive cellular rearrangements experienced by PSM cells for somite formation, such as epithelial-to-mesenchymal (EMT) transitions[2].

Metabolic processes related to oxygen transport are also strongly present in PSM K2, which could be explained by the high energy consumption of the developing embryo. Given the rapid cell proliferation in embryogenesis, metabolism is a critical biological process. Interestingly, genes related to oxygen transport are predicted to be oscillatory, which is curious, considering that the posterior part of the embryo is assumed to be maintained in an hypoxic environment regulated by the WNT and FGF pathways during the somitogenesis period[9]. Therefore, these results provide some hints that it could be important to further explore the oscillatory potential of the GRN underlying oxygen transport and cellular metabolism in the embryo.

C.3 – Limb Cluster 1 Analysis

C.3.1 – The base cycle of limb K1

Oscope retrieved a single cluster of candidate oscillatory genes in the limb tissue (Limb K1) with 173 probe-sets (corresponding to 163 genes) listed in **Annex 6.3**. Descriptive statistics

for Limb K1 is presented in **Annex 5.5**. Similarly to the analyses performed for the PSM clusters, the limb base cycle was reconstructed using the rescaled expression values and is displayed in **Figure 3.12 A**. Individual plots for the 163 genes are available for inspection in **Annex 7.3** (plotted using the original expression values, and samples sorted according to the pseudo time ordering inferred for this cluster).

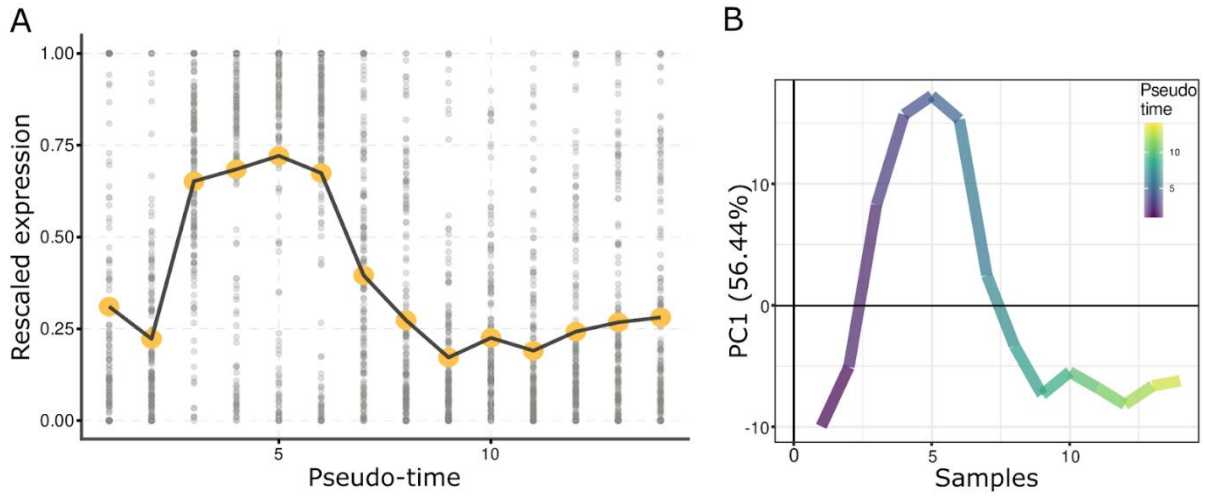

Figure 3.12 | Oscillatory trajectory recovered by Oscope for Limb cluster 1. A | Scatter plot of mean expression values for each sample (orange points), based on 173 probe-sets (background grey points). Sample order is given by the pseudo time ordering calculated by the ENI algorithm from Oscope(n = 14 samples). Expression values were rescaled between 0 and +1. **B** | PCA plot of Limb K1 genes showing the data variability contained in these genes . PCA was calculated with the rescaled expression values, showing that the PC1 explains 56.44% of the data variability. Samples are ordered by their base cycle ordering inferred by Oscope, where the 1st sample is colored dark violet, following a color gradient to the last sample in yellow.

Using the mean expression values for each sample, the limb base cycle recapitulates an oscillation (**Figure 3.12 A**), with a well-defined pattern. Globally, the individual gene expressions (grey points) follow the average gene expression (higher density of points around the mean represented by an orange dot). However, some outliers are visible (points scattered along the y axis), albeit with much lower density, conferring some confidence in the pattern of oscillation retrieved. The plot in **Figure 3.12 panel B** shows that 56.44% of data variability is explained by the Principal Component 1, and the path obtained from the sample ordering retrieved by oscope recapitulates the base cycle displayed in panel A, thus adding extra strength to the oscillation pattern described for the 163 genes present in Limb K1.

The cluster of candidate oscillatory genes found for the limb tissue presents great functional diversity. In this cluster there are genes involved in biological mechanisms like ubiquitination (*CTR9*[152], *USP42*[153], *CDC73*[154], *MIB1*[155], *RNF151*[156]); chromatin reorganization (*KAT6B*[157], *UHRF1*[158], *TOP3A*[159], *KDM1A*[160]); neuronal development (*LSAMP*[161], *GRIN3A*[162],

NLGN1[163]); cell-cycle progression (*NEK9*[164], *PCM1*[165], *CLSPN*[166], *G2E3*[167]) and other cellular mechanisms.

Noteworthy are 2 genes of the HOX family of TFs (genes that specify the identity of the embryonic segments along the head-to-tail axis[31] that were found to potentially oscillate in the limb, namely, *HOXD11*[168] and *HOXD13*[169], priming these genes for further experimental investigation in the limb.

C.3.2 – Functional analysis | Transcription regulation processes are enriched in Limb K1

The functional enrichment (FE) analysis of the Limb K1 is presented in **Table 6**. The ClockOME genes in this cluster are enriched for 89 GO BP, 22 GO MF and 33 GO CC terms (**Annex 8.3**).

Table 6 | Functional enrichment conditions for Limb K1.

	GO Database	Database size (*)	Limb genes present in the database (**)	Enriched GOterms
Limb K1 (n=163)	GO BP	3901	30	89
	GO MF	3606	28	22
	GO CC	4101	29	33

*****Total number of chicken genes that are annotated in the database; ******Intersection between database size and the 106 genes present in K1; **GO** = Gene Ontology; **GO BP** = GO Biological Process; **GO MF** = GO Molecular Function; **GO CC** = GO Cellular Component.

The visual representation of these results is displayed in **Figure 3.13 A**, showing the top 20 enriched GO terms. In panel B, the GO categories are ranked by p-value. The Limb K1 shows diverse biological regulatory pathways enriched. However, based on the p-value, the most enriched biological processes (in blue) are the ones involved in the regulation of DNA transcription (**Figure 3.13 B**). The molecular functions over-represented are globally related to binding, especially to DNA. Regarding the cellular component, the Limb K1 genes are very heterogeneous with respect to their cellular localization, being part of the nucleus, cytoplasm, and diverse intracellular organelles. Overall, the limb cluster collects various protein-coding genes engaged in transcriptional regulation, which is a broad classification, not amenable to draw one unique conclusion regarding the single defined function common to all these genes (**Figure 3.13**).

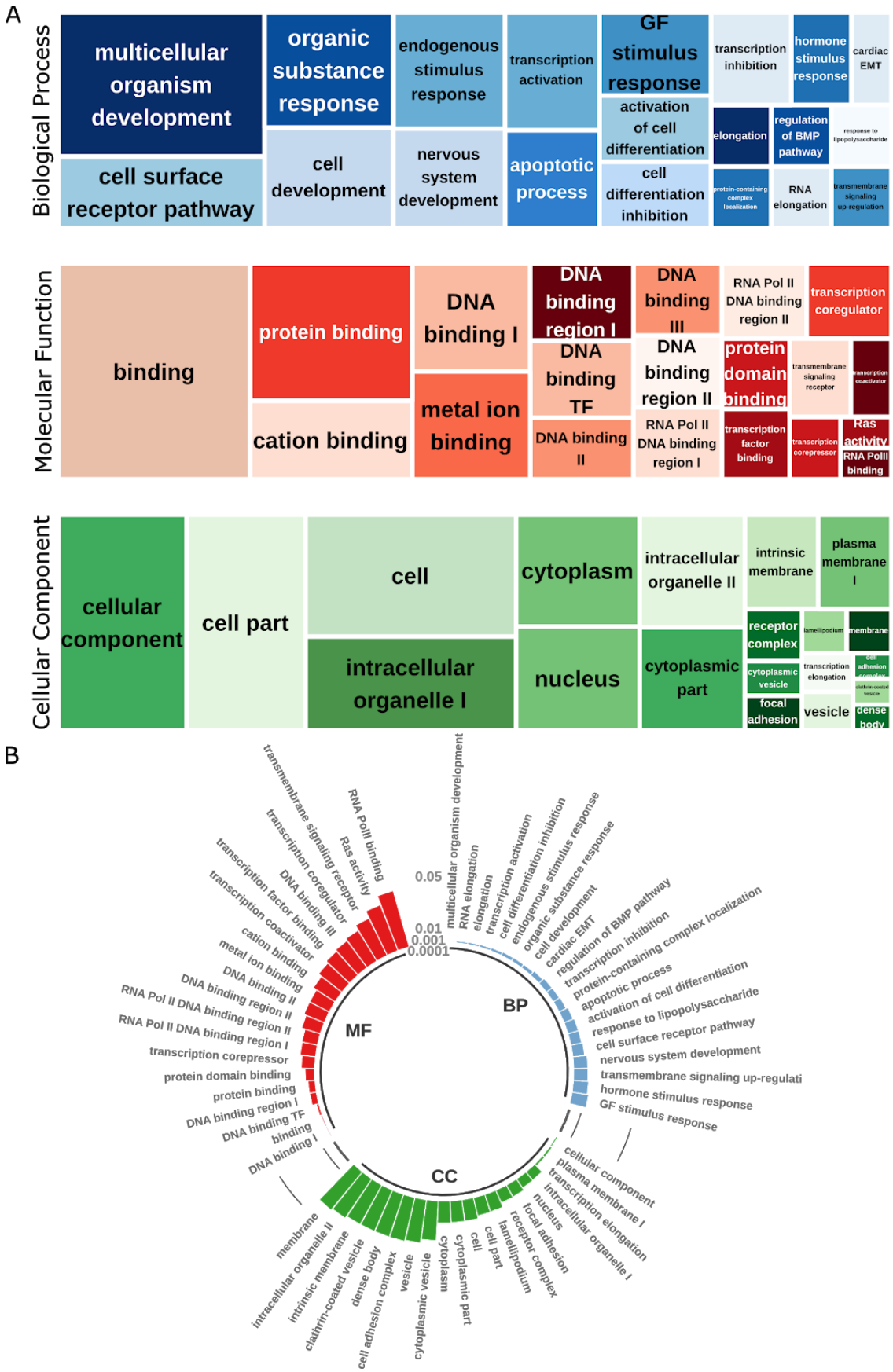

Figure 3.13 | Intra-cluster comparison of categories enriched for the Limb K1 genes. Continued on next page.

Figure 3.13 Continued.**A |** Blue – Biological Processes (GO BP) with 89 categories in total; Red – Molecular Functions (GO MF) with 22 categories in total; Green – Cellular Components (GO CC) with 33 categories in total. Rectangle sizes are proportional to the number of ClockOME genes found for each GO category. The total number of genes from our list annotated with terms from each GO database are: GO BP genes = 30; GO MF genes = 28; GO CC genes = 29. **B |** The circular plot shows the p-values for each GO term. All values shown are below 0.05. For simplicity only the top 20 enriched categories are displayed.

Following the functional interaction network construction (**Annex 9.3**), the limb cluster was divided into 6 modules (orange nodes correspond to ClockOME proteins, and blue nodes represent additional predicted interactors) (**Figure 3.14**). The functional classification for each module is presented in **Table 7**.

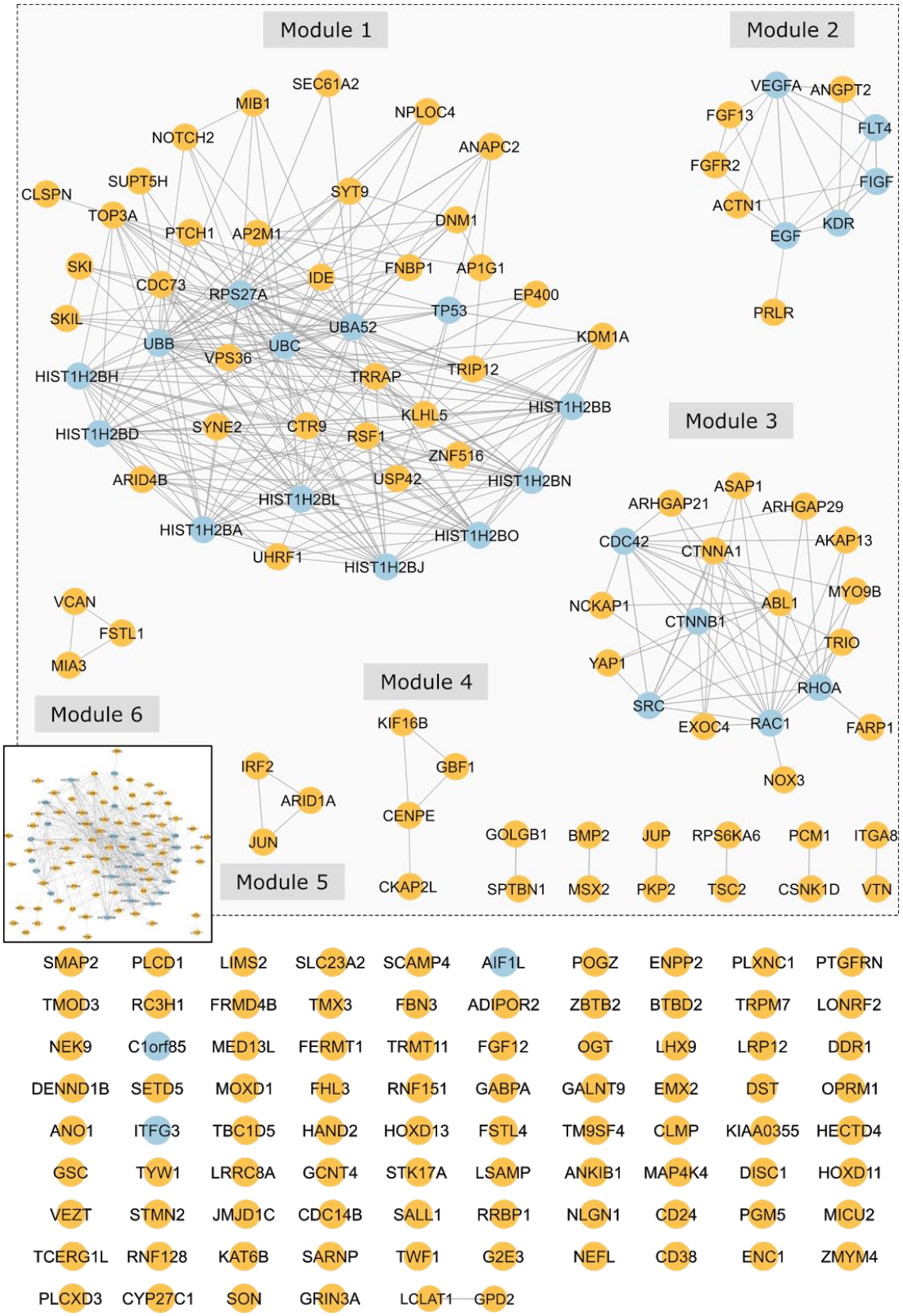

Figure 3.14 | Functional interaction network visualization for proteins coded by the genes from the PSM K2, with 20 additional predicted interactors. Three modules were found after MCL Clustering. Orange – ClockOME interactors; Blue – additional predicted interactors; Left-middle panel shows the original network with 20 nodes before the split into smaller modules using MCL clustering. Module 1 = 44n + 245e; Module 2 = 10n + 22e; Module 3 – 18n + 51e; Module 4 – 4n + 4e; Module 5 – 3n + 3e; Module 6 – 3n + 3e; n = nodes; e = edges.

Table 7 | Functional interaction modules in Limb K1.

Cluster	Module	GOterm	Extended name of the GOterm	ClockOME genes (*)	Full module (**)	Functional name
Limb	1	GO:0000122	negative regulation of transcription by RNA polymerase II	10/32	14/44	Transcription regulation
		GO:0016567	protein ubiquitination	7/32	17/44	
	2	GO:0001525	angiogenesis	3/6	7/10	Signal transduction
		GO:0007165	signal transduction	3/6	-	
		GO:0007275	multicellular organism development	3/6	6/10	
		GO:0008284	positive regulation of cell proliferation	3/6	7/10	
	3	GO:0051056	regulation of small GTPase mediated signal transduction	5/13	8/18	Signal transduction mediated by GTPase
	4	GO:0006890	post-Golgi vesicle-mediated transport	2/4	-	Organelle rearrangement
		GO:0006895	microtubule-based movement	2/4	-	
		GO:0007018	epidermal growth factor receptor signaling pathway	2/4	-	
	5	GO:0000122	negative regulation of transcription by RNA polymerase II	3/3	-	Negative regulation of transcription
	6	GO:0043687	post-translational protein modification	3/3	-	Metabolic processes
		GO:0044267	cellular protein metabolic process	3/3	-	

***The number of ClockOME genes attributed to the GOterm out of the total number of ClockOME genes present in the respective module; ****The number of genes attributed to the GOterm accounting for the full size of the module; **GTPase** = Guanosine triphosphate (GTP) hydrolase enzyme.

Based on these findings, module 1 is strongly related to transcription regulation and protein ubiquitination (**Table 7)**. In the second module, half of the ClockOME genes are involved in signal transduction for diverse biological processes, such as cell differentiation and angiogenesis. Module 3 holds genes that are involved in the regulation of signal transduction mediated by small GTPase molecules (**Table 7)**. Modules 4, 5, and 6 do not contain extra predicted interactors, and are thus composed uniquely by ClockOME genes (**Table 7)**. Module 4 seems to contain genes underlying organelle trafficking. Module 5 is

composed of 3 nodes connecting genes involved in the negative regulation of DNA transcription (**Table 7)**. Finally, Module 6, also composed of 3 nodes, is related to metabolic processes (**Table 7)**.

Therefore, the Limb K1 cluster focuses on transcription regulation and signal transduction processes, which is in line with the described GO molecular functions of binding to DNA, and localization to diverse cellular compartments.

C.4 – Comparing the ClockOME with Previously Published Data

The ClockOME identified in this work is composed of 296 candidate cyclic genes, distributed between 3 clusters: two for PSM and one for Limb. Interestingly, some genes were found both in the PSM clusters and in the Limb cluster (**Figure 3.15**), namely, *FSTL4*, *CD38*, and *PTCH1* are present in Limb and in PSM K1, and *LOC769756* and *CD24* are present in Limb and PSM K2. It should be noted that there are no common genes between the two PSM clusters, since they are mutually exclusive (resultant from the same dataset). Genes found in common between both tissues should be closely inspected given that PSM and Limb data are collected at different embryological stages (HH12 in the PSM and HH20/24 in the limb), providing extra confidence to their oscillatory potential.

To compare the ClockOME genes with previously published data for genes associated with the embryo molecular clock, I compiled a list of genes derived from 2 previously published studies, namely from Krol *et al*., (2011)[3] with *Gallus gallus* data, and from Matsuda *et al*., (2020)[26] with *Homo sapiens* data. In **Figure 3.15**, the Venn diagram shows the interactions between the datasets and lists the common gene IDs (as well as a short name description).

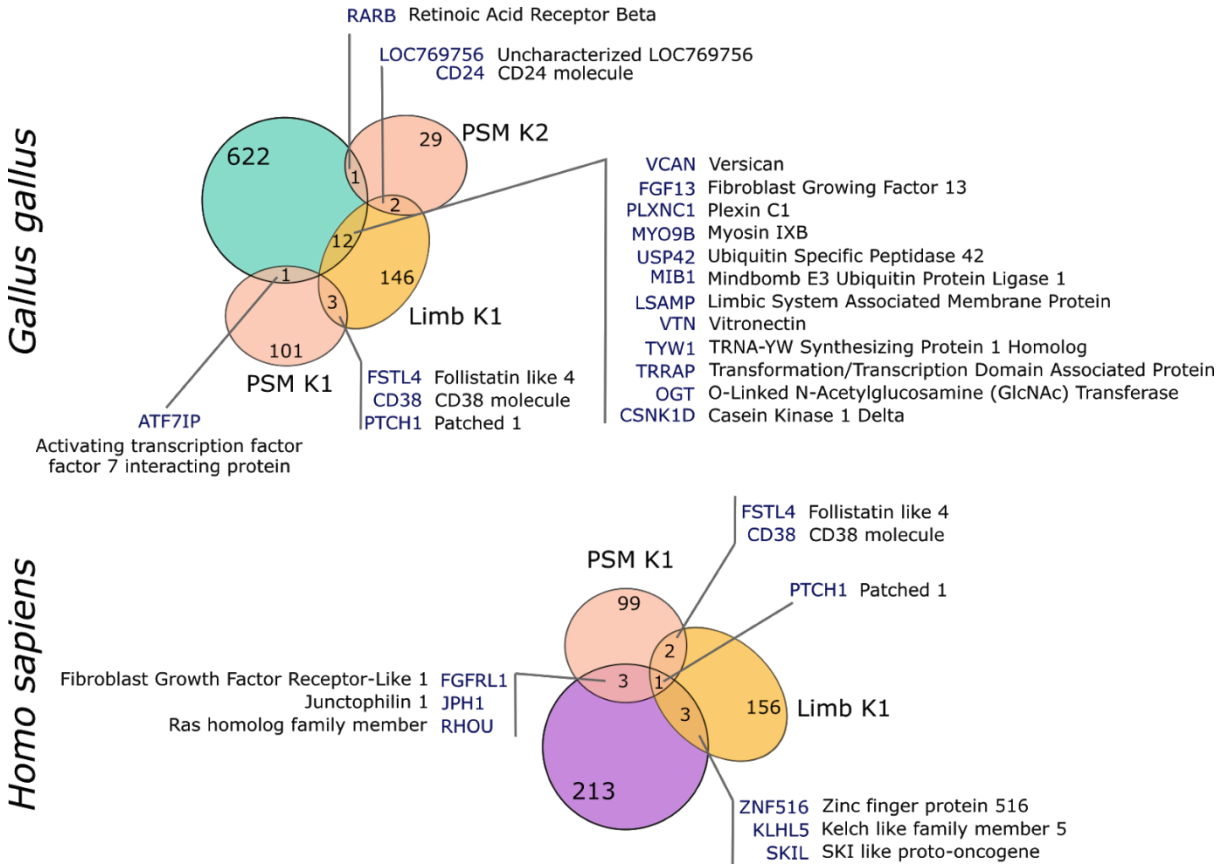

Figure 3.15 | Comparison between ClockOME genes and previously reported oscillatory genes. ***Gallus gallus*** **|** Overlap between Krol et al., (2011)[3] list of 636 chicken genes in cyan. The analysis was performed by microarray technology on the posterior PSM tissue. ***Homo sapiens*** **|** Overlap between Matsuda et al., (2020)[26] list of 220 human genes, in violet. The analysis was performed by RNA-seq on in vitro cultivated PSM tissue. Pink nodes represent PSM clusters; Orange nodes represent limb clusters.

The list of chicken clock genes, previously reported in 2011 by Krol et al., results from the evaluation of the microarray dataset E-MTAB-406, which is one of the two chicken PSM datasets used for this work[3]. Surprisingly, only 1 gene was found in common with PSM K1: *RARB* - a protein-coding gene that functions as a retinoic acid (RA) receptor and is found across different stages of chick development, including somitogenesis and limb formation[170,171]. Similarly, only one gene was found in common with PSM K2 – the *ATF7IP*. It codes for a multifunctional TF that, depending on the binding proteins, can act as a coactivator or corepressor of mRNA transcription, and it is associated with epigenetic changes in chromatin[172].

As for Limb K1, there were 12 shared genes (**Figure 3.15**). All these genes are associated with the molecular clock, either via direct association, like *MIB1*[155,173] that facilitates Notch signaling, or more indirectly like *VTN*[174] or *VCAN*[175] proteins, that are related to the extracellular matrix (ECM), and may play a role in intercellular communications.

Regarding the human genes reported by Matsuda et al. (2020)[26], the data were collected by high throughput RNA-sequencing analysis of Pluripotent Stem Cells (obtained by step-wise in vitro induction/differentiation of human presomitic mesoderm and its derivatives)[26]. The data were deposited in the GEO database, with accession number GSE116935[176]. This publication reports a list of 220 possibly oscillatory genes associated with the human segmentation clock. From the PSM cluster 1 in the ClockOME, 4 genes overlap with Matsuda's list, namely: *FGFRL1*, *JPH1*, *PTCH1*, and *RHOU* (**Figure 3.15**). Similarly, in the Limb cluster, 4 genes were found in common: *PTCH1*, *SKIL*, *KLHL5*, and *ZNF516* (**Figure 3.15**). No common genes were found between Matsuda's list and PSM K2. Noteworthy, the *PTCH1*[145] gene is present in all three lists (Matsuda's, PSM K1 and Limb K1) (**Figure 3.15**). This gene codes for a transmembrane receptor protein from the hedgehog signaling family, namely *Shh* (Sonic Hedgehog), which is an important morphogene involved in vertebrate embryo development for the establishment of the dorso-ventral axis during the body and limb formation[33,177].

A closer examination of the ClockOME genes shows that, from the list of known oscillatory genes involved in the molecular clock, only Mesp2 (which is known to robustly oscillate in mouse somitogenesis) was found. Other genes that have been established as oscillatory in early development, like *Hes1/5*, *Lfng*, *Dll*, *Dusp6/2*, *Axin2*, or *Brachyury*/*T*, were not found in the ClockOME, despite being present in the array. When searching for these genes in the datasets, I observed that they had been discarded in the early stages of the analysis due to the filtering parameters applied to retrieve only the stronger signals (i.e., these genes were removed because they were too stable for the variability filter).

The fact that these canonical oscillators were left out from the Oscope search may be explained by four factors: (i) the duration of the period; (ii) the oscillation pattern; and (iii) the oscillation level (cellular versus tissue oscillations) or (iv) the fact that the biological material was obtained from pools of tissues that can average out the oscillatory behaviour of genes that are not yet synchronized.

Regarding the period duration, the algorithm can only retrieve genes that are oscillating within the sampling window, which in this case is HH12 for PSM, and between HH20 and HH24 for the Limb. If these genes present periods higher than the HH sampling window, then they will be missed by the algorithm. Similarly, the shape of the oscillation can influence the success in finding oscillatory genes. For example, if the shape of the oscillation departs too much from the sinusoidal function (which is a very simple model), e.g., with several peaks and

troughs in each period, then the algorithm might be unable to correctly reconstruct one cycle, and therefore miss it altogether. Another example would be square waves (with long stable up and down values, with very fast transitions). These require a big number of samples in order to increase the resolution to find expression measurements with intermediate values, otherwise the gene just seems to either be switched on or off. Finally, and most importantly, these genes are known to oscillate at the cellular level, and our dataset measures bulk transcriptomics, meaning that it will only find genes that are oscillating at the tissue level (i.e., genes with synchronised expression between individual cells). I speculate that if the single cell RNA-seq expression dataset could have been used (as initially planned), these canonical oscillating genes would have been retrieved (contingent on their expression level being above the detection limit of single cell sequencing, circa 20% of the higher expressing genes).

Other alternative explanations could be related to the Oscope algorithm itself. As previously mentioned, the Oscope method was developed to account for large quantities of unsynchronized data coming from scRNA-seq, and its main feature is the paired-sine model, which is based on the co-expression of gene pairs, removing the necessity of timing in the sequencing studies. Nevertheless, its current implementation does not provide a statistical criterion for the process of selection of the oscillatory genes, rather an arbitrary T% of genes are considered downstream. The significance is only tested after the clustering. Considering this, the previously chosen T% threshold has a major impact on the results of clustering and consequently on the pseudo-time inference. Additionally, the algorithm is not able to distinguish between co-oscillatory genes with phase = 0 and linearly correlated genes. It uses a heuristic approach to filter by sine score and phase shift variability, which results in the removal of an entire cluster of genes when some of them could potentially be cyclic. Finally, the ENI method, used to estimate the pseudo temporal order of the samples, showed to be a computationally slow and poor probabilistic approach[82].

Some additional challenges were encountered for the analyses of these results. The first stems from the critical step which is the annotation of the genes. The chicken genome annotation process is still far from the detailed level achieved for other animal models, like mouse or zebrafish, thus these results cannot present a complete picture of the functional properties of all the candidate oscillatory genes found. In addition, the functional annotation process relies heavily on information from human genes, leaving room for inconsistencies and

sometimes even suppositions that the functional properties of these orthologues behave equally in chicken and in human cells.

Moreover, genes do not interact directly, rather employing their products - mostly the proteins (but also non-coding RNAs) - that regulate transcription, transport or even subtract abundance for molecular pathways (signalling, metabolism, etc). So, it is imperative to notice that the functional interaction networks analysed here show a static picture of the cell. They represent the functional interactions between gene products in abstract, and without any sense of time (i.e., the shape of the network might change dynamically in time or with cellular status, since not all interactions will occur simultaneously, despite being all presented in the same network). Additionally, such representations do not provide any directionality, i.e., does not depict how one gene transcript affects another, and does not indicate whether the gene transcription is reduced, or its product degradation increased. Thus, it is imperative to note that these interaction networks are just representative of a functional connection between proteins that are known (or predicted) so far, thus prone to future alterations, and do not carry direct regulatory information.

Overall, we find that the ClockOME genes that have been previously reported in the literature seem to be involved in cell proliferation and DNA dynamics (two fundamental processes for the proper development of the embryo). However, the ClockOME revealed several novel genes that have not yet been reported as cyclic, and can, therefore, expand the list of genes involved in the embryo molecular clock, hence increasing our knowledge about early vertebrate embryo development and other biological mechanisms not yet described as oscillatory. Accordingly, in the next section I present a list of 6 relevant candidates for future experimental validation.

Section D | Candidate oscillatory genes proposed for experimental validation

The ClockOME genes seem to be mostly engaged in cell differentiation, proliferation, and signaling, but also in oxygen metabolism, nervous system development, and epigenetic regulation. As such, for the validation-candidates, I highlight genes involved with diverse molecular processes (not only related to somitogenesis), to explore different molecular systems. Accordingly, after carefully examining the results, I propose a list of 6 protein-coding genes for experimental validation using appropriate laboratorial techniques (described in **Chapter 1, section C**). The expression patterns of these genes are presented in **Figure 3.16** using the original scale of expression.

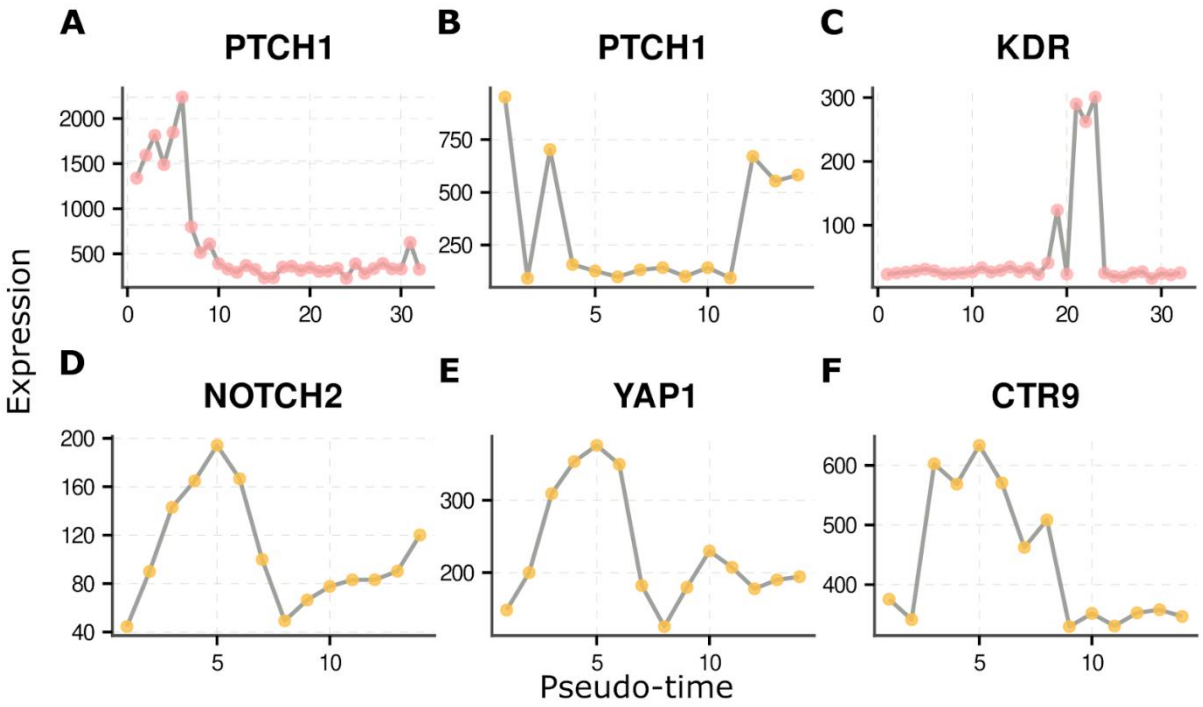

Figure 3.16 | Expression profiles of the ClockOME candidates for experimental validation. All plots are ordered by their respective base cycle (inferred by Oscope). **A |** PTCH1 gene found in PSM K1. **B |** PTCH1 gene in Limb K1. **C |** KDR (VEGFR2) found in PSM K2. **D |** NOTCH2 gene found in Limb K1. **E |** YAP1 gene found in Limb K1. **F |** CTR9 gene found in Limb K1.

D.1 – PTCH1

PTCH1 (Patched 1) is a protein coding gene that functions as a transmembrane receptor involved in the Sonic Hedgehog (*Shh*) signaling transduction pathway, which is a crucial pathway for embryo development since *Shh* signaling is required for the establishment of dorsoventral patterning and tissue homeostasis[145,178]. In the ClockOME, *PTCH1* is present in both tissues: in the PSM K1 (**Figure 3.8 Module 3** related to morphogenesis), and in Limb K1

(**Figure 3.14 Module 1** - Transcription regulation). However, the expression profiles found in PSM and in Limb (**Figure 3.16 A** and **B** respectively) show a different oscillatory pattern (however they should not be directly compared because their expression values are in different scales, and the oscillation was inferred from a very different number of samples, hence having different power to resolve the oscillation). Most importantly, the cycle was extracted from two different developmental stages and tissues, which indicates that different molecular networks might be active and therefore result in different dynamic behaviour for the same gene. One additional argument for its inclusion on this list is the fact that *PTCH1* is present in Matsuda's list of candidate human oscillatory genes, conferring extra confidence to its potentially cyclic behaviour[26]. Furthermore, FI networks of this gene in both clusters (PSM K1 and Limb K1) show that it is also functionally related to proteins involved in ubiquitination (known to be an important process for embryo development): *UBB*[179], *UBC*[180], and *UBA52*[181].

According to Geisha (an online repository of in situ hybridization images for genes expressed in chicken embryo in the first 6 days of development.), the *PTCH1* gene is expressed throughout early embryo development in different tissues, including PSM, somites, and limb bud, besides being expressed in different neuronal locations[171]. Experimentally, it has also been shown a relationship between the *Shh* and the Notch pathways in both directions, where Notch mechanistically upregulated the *Shh* downstream gene transcription (in competition with *PATCH1* regulation) and *Shh* induces *Hes1* expression (Notch effector gene)[178]. Moreover, the *PTCH1* gene was also found to be involved in the establishment of the correct PSM segmentation timing, and thus is involved in the somitogenesis clock[182]. In the limb, it was also studied as an important *Shh* signal transducer, guiding the correct limb tissue patterning[183]. So, *PTCH1* is a significant candidate for experimental validation, preferentially in a tissue-specific manner. Also, it would be interesting to examine its interactome and how it is maintained throughout the embryogenesis process.

D.2 – KDR

KDR (also known as *VEGFR2* (Vascular Endothelial Growth Factor Receptor 2)), is a transmembrane protein pivotal for angiogenesis[184]. Its activation is also involved in multiple cellular responses like permeability, migration, proliferation, and survival (reviewed in: Karaman *et al.,* Alitalo (2018)[185]). *KDR* signaling transduction regulates blood vessel formation and expansion, and thus it is also an important target in tumorigenesis. It was predicted in PSM K2 (**Figure 3.11 Module 3** - Morphogenesis), where it seems to be connected to WNT pathway-

related proteins**. Figure 3.16 panel C** shows its inferred oscillatory pattern, with a sharp peak in expression. This gene was also added to the Module 2 of the Limb K1 functional interaction network from STRING (**Figure 3.14 Module 2** - Signal transduction). *KDR's* expression in chick embryo development includes several different tissues, including the anterior PSM, but not the somites[171]. Moreover, *KDR* interacts with MAPK, AKT, SRC, and Notch pathways, making it a good candidate for experimental validation, particularly in the segmentation stages of embryo development[184,185].

D.3 – NOTCH2

NOTCH2 is a transmembrane receptor whose intracellular domain directly acts as a transcription regulator[186]. The Notch family of receptors is dominant across embryogenesis, guiding cellular behavior in a context-dependent way from survival and proliferation to death (reviewed in: Henrique & Schweisguth (2019)[173]). Similarly to *NOTCH1*, which has been demonstrated to coordinate oscillatory expression of HES genes in the PSM, *NOTCH2* was proposed to regulate the segmentation of the limb[32]. Accordingly, my results are in line with this hypothesis since this gene was identified in the list of Limb Oscillatory genes (**Figure 3.14 Module 1** - Transcriptional regulation). *NOTCH2* inferred oscillation pattern is shown in **Figure 3.16, panel D**.

NOTCH2 is related to ubiquitination proteins like *MIB1* (also predicted to be oscillatory in Module 1 of Limb K1). Experimentally, it was shown to function as tumor suppressor gene, and in its absence tumor cells seem to increase their mobility hence enabling metastization, through an increase in beta-catenin regulation (WNT pathway effector) and MAPK signaling, in different human cancer cells[187]. *NOTCH2* is widely expressed in chicken embryos, starting in the early stages of development in neuronal tissue, but also later in the somites, wings, dermis, hearth, and different regions of the future brain[171]. Moreover, Notch signaling is implicated in the models for self-sustained segmentation, and models that argue that segmentation is a programmed event, with no autonomous signaling. Accordingly, further studies of this gene and its pathway interactors are crucial for a better understanding of the segmentation process in vertebrate development.

D.4 – YAP1

The protein-coding gene *YAP1*[188] (Yes1 Associated Transcriptional Regulator) is a transcriptional regulator that activates and deactivates the transcriptional complex, depending on the context. It acts downstream of the Hippo pathway - a signaling pathway involved in the regulation of tissue and organ cell number by controlling the cell cycle progression and apoptosis, thus having a role in tumorigenesis. Additionally, the Hippo pathway has recently been implicated in human embryo segmentation[26]. In the literature, *YAP1* is described as being dependent on the mechanical cues that a cell is experiencing, which increases when a cell is experiencing dense connections with the extracellular environment[189]. Furthermore, it was also shown that *YAP1* counters the Notch signaling to maintain the stemness of epidermal cells, and it couples with Notch to allow the stable oscillatory behavior of PSM cells[49,189].

YAP1 was found in Limb K1 (**Figure 3.14 Module 3** - Signal transduction mediated by GTPases), where it is predicted to functionally interact with SRC (one of the following candidates). The base cycle for *YAP1* is shown in **Figure 3.16 panel E**.

D.5 – CTR9

CTR9 codes for a protein that accompanies the RNA polymerase II in the PAF1 complex (Paf1C), a protein complex responsible for the recruitment of histone modification factors, mRNA transcription, and elongation factors[152]. It is known to interact with multiple transcriptional co-regulators such as beta-catenin (WNT pathway) and Gli (*Shh* pathway), and to be essential for the transcription of HOX and WNT target genes[31,152]. Therefore, *CTR9* seems to be crucial for cell survival, maintenance of cell pluripotency (undifferentiated cellular state), and differentiation. In parallel, Paf1C is necessary for the epigenetic regulation on multiple histones and subsequent chromatin remodeling[152,190]. *CTR9* gene is present in Limb K1, interacting with Module 1 proteins related to transcriptional regulation (**Figure 3.14 Module 1** - Transcriptional regulation). Its predicted oscillation pattern is shown in figure **3.16 panel F**.

D.6 – SRC

SRC has been classified as non-receptor protein tyrosine kinase (PTK), with multiple activation domains. Molecularly, *SRC's* function is coupled to diverse receptors and is thought to regulate different intracellular pathways (GPCRs, MAPK, Akt, STAT), and so it is normally found near the cellular membrane. In consequence, it controls multiple cell fate decisions

ranging between immune system activation, cell adhesions or motility, cell-cycle progression or apoptosis, cell proliferation or differentiation[191,192]). *SRC* is not directly part of the ClockOME gene list, thus there is no inferred pseudo-temporal trajectory for this gene. Instead, it is a predicted protein that interacts with 8 genes from the PSM K1 list (**Figure 3.8 Module 2 -** Phosphorylation). The 8 proteins that functionally interact with SRC in module 2 from the PSM K1 are: *MMP9*, *MAPK11*, *FGFR3*, *NRG1*, *RET*, *RHOU*, *EPHA3*, *PTN* (**Figure 3.8**). Also, it is predicted to interact with 4 genes from the Limb K1 list (**Figure 3.14 Module 3** - Signal transduction mediated by GTPase).

Furthermore, the SRC protein is involved in *KDR* protein signaling, as well as with the *YAP1* and Notch transcription regulators (other candidates chosen for experimental validation). Yet, *SRC* has not been described as oscillatory, nor has it been directly associated with the embryo molecular clock. Importantly, it was shown to be implicated in different cancers as a proto-oncogene; important for the motor neuron guidance (in the chicken limb development[193]); and modulation of cytoskeletal responses via spatiotemporal activity[192]. Therefore, it would be important to verify if this gene's expression is itself cycling, or if it could be a regulator of other oscillatory genes relevant for early embryo development.

D.7 – Other relevant candidates

Besides the aforementioned candidates, several other genes are involved in central cellular processes, making them equally good candidates for experimental validation **(Table 8** and **Table 9**). Firstly, PSM K1 contains more genes related to somitogenesis, which makes it the best list to search for promising oscillatory candidates to expand the gene regulatory network related to cyclic somitogenesis events.

Two types of genes are proposed in **Table 8**: (i) genes already associated with the somitogenesis process (genes 1-12); and (ii) novel genes that are not associated with early development, but that showed up in this analysis as potentially oscillating, and therefore might be involved with the embryo clock in an unforeseen way (genes 13-18)**.** Here, I chose 6 genes that are described as being involved in neuronal functions, and therefore potentially interesting for the study of neural development in early embryogenesis.

Table 8 | Relevant candidates of PSM.

#	Gene Symbol	Extended gene name	Cluster
1	***TCF15***[194]	Transcription Factor 15	PSM K1
2	***RIPPLY1***[129]	Ripply Transcriptional Repressor 1	PSM K1
3	***FGFRL1***[195]	Fibroblast Growth Factor Receptor Like 1	PSM K1
4	***FGFR3***[196]	Fibroblast Growth Factor Receptor 3	PSM K1
5	***RHOU***[141]	Ras Homolog Family Member U	PSM K1
6	***MAPK11***[140]	Mitogen-Activated Protein Kinase 11	PSM K1
7	***RARB***[170]	Retinoic Acid Receptor Beta	PSM K1
8	***RDH10***[197]	Retinol Dehydrogenase 10	PSM K1
9	***HEY1***[130]	Hes Related Family BHLH Transcription Factor With YRPW Motif 1	PSM K1
10	***BMP3***[142]	Bone Morphogenetic Protein 3	PSM K1
11	***RIPPLY2***[198]	Ripply Transcriptional Repressor 2	PSM K2
12	***ATF7IP***[172]	Activating Transcription Factor 7 Interacting Protein	PSM K2
13	***FSTL4***[199]	Follistatin Like 4	PSM K1
14	***GRIK3***[135]	Glutamate Ionotropic Receptor Kainate Type Subunit 3	PSM K1
15	***NPY***[134]	Neuropeptide Y	PSM K1
16	***PENK***[133]	Proenkephalin	PSM K1
17	***SOX17***[200]	SRY-Box Transcription Factor 17	PSM K2
18	***RIPK4***[201]	Receptor Interacting Serine/Threonine Kinase 4	PSM K2

Similarly, for the limb tissue (**Table 9**), there are some interesting genes involved in different signaling pathways including Homeobox (HOX genes), BMP and FGF, which given their importance for the cell, might be good candidates for validation of their potential oscillatory behaviour. Further experimental characterization of these genes could potentially identify novel components of the segmentation clock network, as well as better illustrate the limb segmentation process, while also potentially increasing the knowledge of other molecular processes taking place in early vertebrate development.

Table 9 | Relevant candidates from Limb K1.

#	Gene Symbol	Extended gene name
1	***HOXD11***[168]	Homeobox D11
2	***HOXD13***[169]	Homeobox D13
3	***BMP2***[202]	Bone Morphogenetic Protein 2
4	***CTNNA1***[203]	Catenin Alpha 1
5	***FGFR2***[204]	Fibroblast Growth Factor Receptor 2
6	***PRLR***[205]	Prolactin Receptor
7	***TRRAP***[206]	Transformation/Transcription Domain Associated Protein
8	***MIB1***[155]	MIB E3 Ubiquitin Protein Ligase 1
9	***TOP3A***[159]	DNA Topoisomerase III Alpha

Advancement of the knowledge regarding the dynamical properties of gene transcription is important for basic science, but also to understand the mechanisms underlying developmental-related diseases. Additionally, it is imperative to broaden the research methodologies available, namely to provide a better *in vitro* environment to allow a faster

development of cell culture assays and organoids to enable the understanding of developmental processes, as well as the development of novel precision treatments in medicine. Moreover, it is likely that many diseases arise as a consequence of shifts in cyclic patterns of genes and proteins, making their oscillatory behaviour a good candidate as a disease marker to be used as a diagnostic tool. For this to happen, oscillations in biological systems must be further studied, both theoretically and experimentally.

Chapter IV

CONCLUSIONS

4 | CONCLUSIONS

This thesis reports the work developed with the goal of finding oscillatory genes in early vertebrate development. Briefly, as shown in **Figure 4.1**, I gathered and analysed three independent transcriptomics datasets derived from chicken embryo PSM and Limb, from which I obtained three clusters of candidate oscillatory genes: 138 genes in PSM, and 163 genes in the limb. These three gene sets were collectively termed ClockOME. Next, I performed a functional analysis for these genes, using GO categories and STRING interaction networks, showing that these genes are involved in broad biological processes, including molecular programs related to morphogenesis, development, and cytoskeletal rearrangements (**Figure 4.1**). Additionally, these functional analyses show that the ClockOME genes seem to be important for the transduction of external signals, and consequent transcriptional alterations involving gene expression regulation and protein synthesis. At the end, I present a ranked list of genes proposed for experimental validation, highlighting a few of the most promising candidates based on their novelty, functional description related to development, or other relevant biological features outside of segmentation, such as neurological development (**Figure 4.1**).

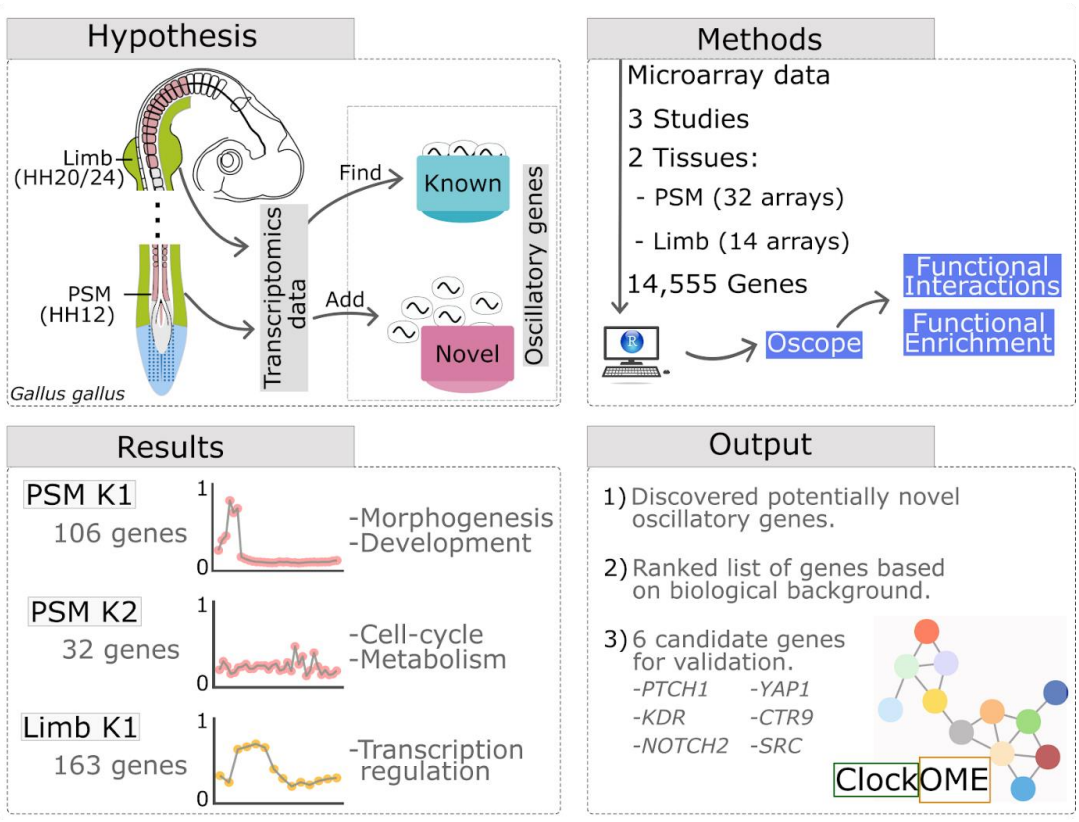

Figure 4.1 | Overview of the thesis outline. Continued on next page.

Figure 4.1 | Continued. This work aimed at discovering new candidate oscillatory genes in *Gallus gallus* early embryonic development. For this, I analyzed three microarray datasets sampled from PSM (Presomitic Mesoderm) and Limb tissues. In total, 14,555 genes per tissue were evaluated. I applied the Oscope algorithm to infer potential gene expression oscillatory behaviour. The analysis resulted into three clusters (K): two from PSM and one from Limb. In the end, this study provided a list of candidate oscillatory genes - the ClockOME - which was then short listed to a group of 6 candidate genes proposed for experimental validation. PSM = Presomitic mesoderm; K = cluster; HH = Hamburger and Hamilton embryo staging in *Gallus gallus.*

This study provides an overview of candidate genes periodically transcribed in early chick development, in the PSM tissue (during the somitogenesis), and in the Limb bud (HH20-24). These tissues and developmental stages were selected given their importance for the molecular embryo clock. The concept of a Molecular clock was introduced by the Clock and Wavefront model in 1976 by Cooke and Zeeman[1] to theoretically describe the rhythmic somite formation at both sides of the PSM tissue. This model postulates the existence of genetic oscillators that coordinate this timely segmentation. Additionally, it proposes a wavefront that provides the positional cues to the cells. Together, both mechanisms drive the maturation of the PSM cells, which in consequence are able to undergo molecular transformation into somitic structures. Further, such a molecular clock was also hypothesized to regulate other segmentation processes such as limb structures formation[61,63] which is not fully elucidated yet. Therefore, the dynamic biological mechanism of the molecular clock is still under investigation.

Transcriptomics data analysis is an important technology that revolutionized the way researchers measure gene expression. However, such measurements entail the extraction of RNA from the cells which are destroyed in the process, providing snapshots of the sampled system at particular times. Accordingly, the study of periodic or temporally sequential processes is not straightforward using this technology.

To tackle this challenge, and in the interest of studying the oscillatory genes involved in early chick embryogenesis, we applied a method belonging to the pseudotime ordering class of algorithms, that excludes the need for explicit temporal information in the samples. Given a sufficiently large number of samples, this method reorders the samples according to a sinusoidal model, therefore finding genes that fit this oscillatory function. Then, the algorithm finds other genes that share similar periodic functions, and groups them into the same cluster of candidate cyclic genes.

Using three chicken transcriptomics datasets, the statistical method applied, Oscope, resulted in the ClockOME gene list, comprising a total of 296 genes that seem to behave periodically in the microarray data from PSM and Limb. Following the biological annotations

for these genes, and their protein interactomes, three clusters of genes were found to be enriched for diverse biological functions. In the PSM tissue, 2 groups of genes were found, with K1 related to morphogenetic mechanisms, including segmentation clock-related features, while K2 was more based on the cellular rearrangements and oxygen metabolism functions. In the Limb, only one group of oscillatory genes was found with features contributing to transcription regulation and signal transduction events.

It should be pointed out here that the technical approach taken for this work has some critical points that might interfere with the robustness of the research. Firstly, and foremost, the use of microarray data which uses bulk RNA-seq, which is not ideal to search for oscillatory signals in non-synchronized tissues. Secondly, using the chicken animal model presents challenges related to its genome annotation, which is still not up to par with the one for other vertebrate model organisms.

Nevertheless, this work shows consistent results with previously reported studies. Namely, the ClockOME comprises *Mesp2*, a well-documented cyclic gene[23], as well as other genes related to signaling pathways (WNT, FGF, NOTCH) that dominate the embryo development both during somitogenesis and limb segmentation. These results illustrate the value of this approach, while providing new hints regarding the potentially oscillatory behaviour of well-established pathways active in early embryo development. Additionally, novel pathways and biological processes were brought forth to the cyclic embryogenesis arena, including GPCR signaling, ubiquitination, epigenetic regulation by histone modification, neuronal development, and sprouting, as well as oxygen and protein metabolism. Cellular mechanotransduction is also present, given the presence of features important for physical and architectural cell microenvironment in all three clusters.

Since the genetic expression is a process involving a hierarchy of events, it is not surprising that processes upstream (for example chromatin remodeling) and downstream (for example post-transcriptional modification) of transcription were mentioned in this work. For this it would be pivotal for such genes to be further studied, and their cycling behaviour experimentally validated, particularly the ones short listed: *SRC*, *PTCH1*, *NOTCH2*, *YAP1*, *KDR*, *CTR9*, as well as the other core embryogenesis genes present in the PSM cluster K1.

Overall, the advancement of the knowledge regarding the timed nature of gene transcription is important for the understanding of embryo development, but also in the fields

related to stem cell biology and regenerative medicine. I envision a future where the dynamic behavior of a protein, or a cluster of genes, could also be used as a marker to develop an optimized diagnostic and/or treatment for specific illnesses. Further experimental characterization of the ClockOME genes and their connectome could identify novel components of the segmentation clock network, illustrate the limb segmentation process, but also be translated into other biological systems and fasten the transition between traditional and precision medicine. In summary, this research sits at the beginning of a journey towards a comprehensive view of cyclic processes that are still not well understood, and that might hold the key to unknown biological potential in the modeling of cellular behavior.

Chapter V

FUTURE PERSPECTIVES

5 | FUTURE PERSPECTIVES

My work has opened the door for new avenues of investigation related to the oscillatory behaviour of genes involved in early chick embryo development. For future work, the first step would consist in experimentally validating the genes found in the ClockOME, prioritizing the list of 6 candidate genes presented in Chapter III section C.4. For this it would be interesting to perform imagiology assays. It could be interesting to have fluorescent reporter markers to measure in real time when the mRNA of each gene of interest is transcribed, and visually validate their oscillatory behaviour. For this, one possible approach would be performing explant assays, where one side of PSM is fixed to stop the developmental progression, while the other side of the tissue is further incubated for different amounts of time, and then immunohistochemically stained allowing the visualization of the gene transcription. This methodology also allows the discovery of the periodicity of the genes of interest (the incubation time taken to match the staining pattern present in the fixed embryo half).

Besides the experimental validation, I think that this work could be expanded by applying the same theoretical analysis I have described to different model organisms, and to different periods of development. For different organisms, such a study would further our knowledge regarding the oscillatory dynamics transversal to other vertebrate organisms. This work would be immediately feasible for example for the mouse embryo or the zebrafish that have mature genome annotations with richer functional information, and for which there are many publicly available transcriptomics datasets. Equally relevant, would be the assessment of a similar ClockOME gene list in human cell lines, since there is much to be learnt regarding the oscillatory behaviour of human genes. Regarding the application of this analysis to different developmental periods, it would be particularly interesting to examine oscillatory genetic behaviour during the gastrulation period, which is an important threshold period for the embryo with the establishment of cellular identity and significant genetic remodeling.

Finally, I think that in science it is always important to check the robustness of the results by applying different techniques to our data, and different data to the same algorithm. This is particularly relevant when using theoretical models, such as the case of Oscope that was selected for this work. Accordingly, I think that we should test our model assumptions with different pseudo-time ordering algorithms, such as OscoNet[82], or others deemed appropriate for the data type. Similarly, given that our dataset is small and composed by bulk RNA data (which is not ideal to discover oscillations in non-synchronized tissues), it would be very important to

apply Oscope to single cell RNA-sequencing data, and compare the results with the ones reported here.

5 | REFERENCES

1. Cooke, J. & Zeeman, E. C. A clock and wavefront model for control of the number of repeated structures during animal morphogenesis. *J. Theor. Biol.* **58**, 455–476 (1976).

2. Gilbert, S. F. *Developmental Biology*. (Sinauer Associates, Incorporated Publishers, 2014).

3. Krol, A. J. *et al.* Evolutionary plasticity of segmentation clock networks. *Development* **138**, 2783–2792 (2011).

4. Ebisuya, M. & Briscoe, J. What does time mean in development? *Development* **145**, (2018).

5. Tam, P. P. L. & Loebel, D. A. F. Gene function in mouse embryogenesis: get set for gastrulation. *Nat. Rev. Genet.* **8**, 368–381 (2007).

6. Uriu, K. Genetic oscillators in development. *Dev. Growth Differ.* **58**, 16–30 (2016).

7. Dequéant, M.-L. *et al.* A complex oscillating network of signaling genes underlies the mouse segmentation clock. *Science* **314**, 1595–1598 (2006).

8. Bénazéraf, B. & Pourquié, O. Formation and segmentation of the vertebrate body axis. *Annu. Rev. Cell Dev. Biol.* **29**, 1–26 (2013).

9. Oginuma, M. *et al.* A Gradient of Glycolytic Activity Coordinates FGF and Wnt Signaling during Elongation of the Body Axis in Amniote Embryos. *Dev. Cell* **40**, 342–353.e10 (2017).

10. Webb, A. B. & Oates, A. C. Timing by rhythms: Daily clocks and developmental rulers. *Dev. Growth Differ.* **58**, 43–58 (2016).

11. Gilbert, S. F. & Barresi, M. J. F. DEVELOPMENTAL BIOLOGY, 11TH EDITION 2016. *Am. J. Med. Genet. A* **173**, 1430–1430 (2017).

12. Yang, X., Dormann, D., Münsterberg, A. E. & Weijer, C. J. Cell movement patterns during gastrulation in the chick are controlled by positive and negative chemotaxis mediated by FGF4 and FGF8. *Dev. Cell* **3**, 425–437 (2002).

13. Aulehla, A. *et al.* Wnt3a plays a major role in the segmentation clock controlling somitogenesis. *Dev. Cell* **4**, 395–406 (2003).

14. Sweetman, D., Wagstaff, L., Cooper, O., Weijer, C. & Münsterberg, A. The migration of paraxial and lateral plate mesoderm cells emerging from the late primitive streak is controlled by different Wnt signals. *BMC Dev. Biol.* **8**, 63 (2008).

15. Sonnen, K. F. & Aulehla, A. Dynamic signal encoding—From cells to organisms. *Semin. Cell Dev. Biol.* **34**, 91–98 (2014).

16. Pourquié, O. *et al.* Lateral and axial signals involved in avian somite patterning: a role for BMP4. *Cell* **84**, 461–471 (1996).

17. Tonegawa, A., Funayama, N., Ueno, N. & Takahashi, Y. Mesodermal subdivision along the mediolateral axis in chicken controlled by different concentrations of BMP-4. *Development* **124**, 1975–1984 (1997).

18. van Eeden, F. J., Holley, S. A., Haffter, P. & Nüsslein-Volhard, C. Zebrafish segmentation and pair-rule patterning. *Dev. Genet.* **23**, 65–76 (1998).

19. Nikaido, M. *et al.* Tbx24, encoding a T-box protein, is mutated in the zebrafish somite-segmentation mutant fused somites. *Nat. Genet.* **31**, 195–199 (2002).

20. Windner, S. E., Bird, N. C., Patterson, S. E., Doris, R. A. & Devoto, S. H. Fss/Tbx6 is required for central dermomyotome cell fate in zebrafish. *Biol. Open* **1**, 806–814 (2012).

21. Wilm, B., James, R. G., Schultheiss, T. M. & Hogan, B. L. M. The forkhead genes, Foxc1 and Foxc2, regulate paraxial versus intermediate mesoderm cell fate. *Dev. Biol.* **271**, 176–189 (2004).

22. Schoenwolf, G. C. Tail (end) bud contributions to the posterior region of the chick embryo. *J. Exp. Zool.* **201**, 227–245 (1977).

23. Maroto, M., Dale, J. K., Dequéant, M.-L., Petit, A.-C. & Pourquié, O. Synchronised cycling gene oscillations in presomitic mesoderm cells require cell-cell contact. *Int. J. Dev. Biol.* **49**, 309–315 (2005).

24. Palmeirim, I., Henrique, D., Ish-Horowicz, D. & Pourquié, O. Avian hairy gene expression identifies a molecular clock linked to vertebrate segmentation and somitogenesis. *Cell* **91**, 639–648 (1997).

25. Bessho, Y. *et al.* Dynamic expression and essential functions of Hes7 in somite segmentation. *Genes Dev.* **15**, 2642–2647 (2001).

26. Matsuda, M. *et al.* Recapitulating the human segmentation clock with pluripotent stem cells. *Nature* **580**, 124–129 (2020).

27. Maroto, M., Bone, R. A. & Dale, J. K. Somitogenesis. *Development* **139**, 2453–2456 (2012).

28. Hamburger, V. & Hamilton, H. L. A series of normal stages in the development of the chick embryo. *J. Morphol.* **88**, 49–92 (1951).

29. Dequéant, M.-L. & Pourquié, O. Segmental patterning of the vertebrate embryonic axis. *Nat. Rev. Genet.* **9**, 370–382 (2008).

30. Bellairs, R. & Osmond, M. *Atlas of Chick Development*. (Elsevier Science, 2014).

31. Hox Genes in Development: The Hox Code. https://www.nature.com/scitable/topicpage/hox-genes-in-development-the-hox-code-41402/.

32. Sheeba, C. J., Andrade, R. P. & Palmeirim, I. Mechanisms of vertebrate embryo segmentation: Common themes in trunk and limb development. *Semin. Cell Dev. Biol.* **49**, 125–134 (2016).

33. Sheeba, C. J., Andrade, R. P. & Palmeirim, I. Getting a handle on embryo limb development: Molecular interactions driving limb outgrowth and patterning. *Semin. Cell Dev. Biol.* **49**, 92–101 (2016).

34. Oates, A. C., Morelli, L. G. & Ares, S. Patterning embryos with oscillations: structure, function and dynamics of the vertebrate segmentation clock. *Development* **139**, 625–639 (2012).

35. Beta, C. & Kruse, K. Intracellular Oscillations and Waves. *Annu. Rev. Condens. Matter Phys.* **8**, 239–264 (2017).

36. Voit, E. O. Signaling Systems. *A First Course in Systems Biology* 257–281 (2017) doi:10.4324/9780203702260-9.

37. Levine, J. H., Lin, Y. & Elowitz, M. B. Functional roles of pulsing in genetic circuits. *Science* **342**, 1193–1200 (2013).

38. Batchelor, E., Loewer, A. & Lahav, G. The ups and downs of p53: understanding protein dynamics in single cells. *Nat. Rev. Cancer* **9**, 371–377 (2009).

39. Shankaran, H. *et al.* Rapid and sustained nuclear-cytoplasmic ERK oscillations induced by epidermal growth factor. *Mol. Syst. Biol.* **5**, 332 (2009).

40. Albeck, J. G., Mills, G. B. & Brugge, J. S. Frequency-modulated pulses of ERK activity transmit quantitative proliferation signals. *Mol. Cell* **49**, 249–261 (2013).

41. Zhang, Y., Liu, H., Yan, F. & Zhou, J. Oscillatory dynamics of p38 activity with transcriptional and translational time delays. *Sci. Rep.* **7**, 11495 (2017).

42. Tay, S. *et al.* Single-cell NF-kappaB dynamics reveal digital activation and analogue information processing. *Nature* **466**, 267–271 (2010).

43. Cai, L., Dalal, C. K. & Elowitz, M. B. Frequency-modulated nuclear localization bursts coordinate gene regulation. *Nature* **455**, 485–490 (2008).

44. Imayoshi, I. *et al.* Oscillatory Control of Factors Determining Multipotency and Fate in Mouse Neural Progenitors. *Science* vol. 342 1203–1208 (2013).

45. Hubaud, A. & Pourquié, O. Signalling dynamics in vertebrate segmentation. *Nat. Rev. Mol. Cell Biol.* **15**, 709–721 (2014).

46. Hirata, H. *et al.* Oscillatory expression of the bHLH factor Hes1 regulated by a negative feedback loop. *Science* **298**, 840–843 (2002).

47. Bailey, C. & Dale, K. Somitogenesis in Vertebrate Development. *eLS* 1–15 (2015) doi:10.1002/9780470015902.a0003820.pub2.

48. Xi, H. *et al.* In Vivo Human Somitogenesis Guides Somite Development from hPSCs. *Cell Rep.* **18**, 1573–1585 (2017).

49. Hubaud, A., Regev, I., Mahadevan, L. & Pourquié, O. Excitable Dynamics and Yap-Dependent Mechanical Cues Drive the Segmentation Clock. *Cell* **171**, 668–682.e11 (2017).

50. Goldbeter, A. & Pourquié, O. Modeling the segmentation clock as a network of coupled oscillations in the Notch, Wnt and FGF signaling pathways. *J. Theor. Biol.* **252**, 574–585 (2008).

51. Bénazéraf, B. *et al.* A random cell motility gradient downstream of FGF controls elongation of an amniote embryo. *Nature* **466**, 248–252 (2010).

52. Dubrulle, J., McGrew, M. J. & Pourquié, O. FGF signaling controls somite boundary position and regulates segmentation clock control of spatiotemporal Hox gene activation. *Cell* **106**, 219–232 (2001).

53. Saga, Y., Hata, N., Koseki, H. & Taketo, M. M. Mesp2: a novel mouse gene expressed in the presegmented mesoderm and essential for segmentation initiation. *Genes Dev.* **11**, 1827–1839 (1997).

54. Morimoto, M., Takahashi, Y., Endo, M. & Saga, Y. The Mesp2 transcription factor establishes segmental borders by suppressing Notch activity. *Nature* **435**, 354–359 (2005).

55. Pais-de-Azevedo, T., Magno, R., Duarte, I. & Palmeirim, I. Recent advances in understanding vertebrate segmentation. *F1000Res.* **7**, 97 (2018).

56. Meinhardt, H. Models of Segmentation. in *Somites in Developing Embryos* (eds. Bellairs, R., Ede, D. A. & Lash, J. W.) 179–189 (Springer US, 1986).

57. Murray, P. J., Maini, P. K. & Baker, R. E. The clock and wavefront model revisited. *J. Theor. Biol.* **283**, 227–238 (2011).

58. Cotterell, J., Robert-Moreno, A. & Sharpe, J. A Local, Self-Organizing Reaction-Diffusion Model Can Explain Somite Patterning in Embryos. *Cell Syst* **1**, 257–269 (2015).

59. Dias, A. S., de Almeida, I., Belmonte, J. M., Glazier, J. A. & Stern, C. D. Somites without a clock. *Science* **343**, 791–795 (2014).

60. Row, R. H. *et al.* BMP and FGF signaling interact to pattern mesoderm by controlling basic helix-loop-helix transcription factor activity. *Elife* **7**, (2018).

61. Summerbell, D., Lewis, J. H. & Wolpert, L. Positional information in chick limb morphogenesis. *Nature* **244**, 492–496 (1973).

62. Sheeba, C. J., Andrade, R. P. & Palmeirim, I. Limb patterning: from signaling gradients to molecular oscillations. *J. Mol. Biol.* **426**, 780–784 (2014).

63. Pascoal, S. *et al.* A molecular clock operates during chick autopod proximal-distal outgrowth. *J. Mol. Biol.* **368**, 303–309 (2007).

64. Tabin, C. & Wolpert, L. Rethinking the proximodistal axis of the vertebrate limb in the molecular era. *Genes Dev.* **21**, 1433–1442 (2007).

65. Uzkudun, M., Marcon, L. & Sharpe, J. Data-driven modelling of a gene regulatory network for cell fate decisions in the growing limb bud. *Molecular Systems Biology* vol. 11 815 (2015).

66. Korpelainen, E., Tuimala, J., Somervuo, P., Huss, M. & Wong, G. *RNA-seq Data Analysis: A Practical Approach*. (CRC Press, 2014).

67. Lee, J. H. *et al.* Highly multiplexed subcellular RNA sequencing in situ. *Science* **343**, 1360–1363 (2014).

68. Chen, W. *et al.* Genome-wide molecular recording using Live-seq. *bioRxiv* 2021.03.24.436752 (2021) doi:10.1101/2021.03.24.436752.

69. Locke, J. C. W. & Elowitz, M. B. Using movies to analyse gene circuit dynamics in single cells. *Nat. Rev. Microbiol.* **7**, 383–392 (2009).

70. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25**, 402–408 (2001).

71. Microarray Technology. https://www.genome.gov/genetics-glossary/Microarray-Technology.

72. Lee, J. K. *Statistical Bioinformatics: For Biomedical and Life Science Researchers*. (Wiley, 2014).

73. Grant, G. R., Manduchi, E. & Stoeckert, C. J., Jr. Analysis and management of microarray gene expression data. *Curr. Protoc. Mol. Biol.* **Chapter 19**, Unit 19.6 (2007).

74. EMBL-EBI. Microarrays. https://www.ebi.ac.uk/training-beta/online/courses/functional-genomics-ii-common-technologies-and-data-analysis-methods/microarrays/.

75. Irizarry, R. A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264 (2003).

76. Leng, N. *et al.* Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments. *Nat. Methods* **12**, 947–950 (2015).

77. Single Cell Gene Expression - 10x Genomics. https://www.10xgenomics.com/products/single-cell-gene-expression.

78. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).

79. Gupta, A. & Bar-Joseph, Z. Extracting dynamics from static cancer expression data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **5**, 172–182 (2008).

80. Bar-Joseph, Z., Gitter, A. & Simon, I. Studying and modelling dynamic biological processes using time-series gene expression data. *Nat. Rev. Genet.* **13**, 552–564 (2012).

81. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).

82. Cutillo, L., Boukouvalas, A., Marinopoulou, E., Papalopulu, N. & Rattray, M. OscoNet: inferring oscillatory gene networks. *BMC Bioinformatics* **21**, 351 (2020).

83. Ripley, B. D. The R project in statistical computing. *MSOR connect.* **1**, 23–25 (2001).

84. RStudio. https://rstudio.com/.

85. Sobell, M. G. *A practical guide to Ubuntu Linux*. (Pearson Education, 2015).

86. Developers, I. W. Draw Freely. https://inkscape.org.

87. EMBL-EBI. ArrayExpress. https://www.ebi.ac.uk/arrayexpress/.

88. geo. Home - GEO - NCBI. https://www.ncbi.nlm.nih.gov/geo/.

89. Anderson, C. *et al.* A strategy to discover new organizers identifies a putative heart organizer. *Nat. Commun.* **7**, 12656 (2016).

90. Davis, S. & Meltzer, P. S. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* vol. 23 1846–1847 (2007).

91. Kauffmann, A. *et al.* Importing ArrayExpress datasets into R/Bioconductor. *Bioinformatics* **25**, 2092–2094 (2009).

92. Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20**, 307–315 (2004).

93. Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods* vol. 12 115–121 (2015).

94. Kauffmann, A., Gentleman, R. & Huber, W. arrayQualityMetrics—a bioconductor package for quality assessment of microarray data. *Bioinformatics* **25**, 415–416 (2008).

95. McCall, M. N., Bolstad, B. M. & Irizarry, R. A. Frozen robust multiarray analysis (fRMA). *Biostatistics* **11**, 242–253 (2010).

96. *Entrez Programming Utilities Help*. (National Center for Biotechnology Information (US), 2010).

97. McCall, M. N. & Irizarry, R. A. Thawing Frozen Robust Multi-array Analysis (fRMA). *BMC Bioinformatics* vol. 12 (2011).

98. Duarte, I., Liber, M., Magno, R. & Andrade, R. P. FrozenChicken: Promoting the meta-analysis of chicken microarray data. *bioRxiv* 2021.02.25.432894 (2021) doi:10.1101/2021.02.25.432894.

99. Duarte, I., Liber, M. & Andrade, R. P. *FrozenChicken: Promoting the meta-analysis of chicken microarray data*. (Zenodo, 2020). doi:10.5281/ZENODO.3765943.

100. Pagès, H., Carlson, M., Falcon, S. & Li, N. AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor. *R package version* **1**, (2019).

101. Wickham, H. *Ggplot2*. (Springer, 2011).

102. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).

103. Leng, N. *et al.* EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments.

Bioinformatics **29**, 1035–1043 (2013).

104. Alexa, A., Rahnenführer, J. & Lengauer, T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* **22**, 1600–1607 (2006).

105. Carlson, M. *org.Gg.eg.db*. (Bioconductor, 2017). doi:10.18129/B9.BIOC.ORG.GG.EG.DB.

106. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).

107. Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G. D. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS One* **5**, e13984 (2010).

108. Morris, J. H. *et al.* clusterMaker: a multi-algorithm clustering plugin for Cytoscape. *BMC Bioinformatics* **12**, 436 (2011).

109. Doncheva, N. T., Morris, J. H., Gorodkin, J. & Jensen, L. J. Cytoscape StringApp: Network Analysis and Visualization of Proteomics Data. *J. Proteome Res.* **18**, 623–632 (2019).

110. Matthew N. McCall, Rafael A. Irizarry. *mouse4302frmavecs*. (Bioconductor, 2017). doi:10.18129/B9.BIOC.MOUSE4302FRMAVECS.

111. Matthew N. McCall, Rafael A. Irizarry. *mouse430a2frmavecs*. (Bioconductor, 2017). doi:10.18129/B9.BIOC.MOUSE430A2FRMAVECS.

112. Matthew N. McCall, Rafael A. Irizarry. *ygs98frmavecs*. (Bioconductor, 2017). doi:10.18129/B9.BIOC.YGS98FRMAVECS.

113. Matthew N. McCall, Rafael A. Irizarry. *hgu133a2frmavecs*. (Bioconductor, 2017). doi:10.18129/B9.BIOC.HGU133A2FRMAVECS.

114. Matthew N. McCall, Sameer Chavan, Rafael A. Irizarry. *huex.1.0.st.v2frmavecs*. (Bioconductor, 2017). doi:10.18129/B9.BIOC.HUEX.1.0.ST.V2FRMAVECS.

115. Matthew N. McCall, Rafael A. Irizarry. *hugene.1.0.st.v1frmavecs*. (Bioconductor, 2017). doi:10.18129/B9.BIOC.HUGENE.1.0.ST.V1FRMAVECS.

116. Matthew N. McCall, Rafael A. Irizarry. *mogene.1.0.st.v1frmavecs*. (Bioconductor, 2017). doi:10.18129/B9.BIOC.MOGENE.1.0.ST.V1FRMAVECS.

117. Gene Ontology Resource. http://geneontology.org/.

118. Gene Ontology overview. http://geneontology.org/docs/ontology-documentation/.

119. STRING: functional protein association networks. https://string-db.org/.

120. GeneCards Human Gene Database. GeneCards - Human Genes. https://www.genecards.org/.

121. Cole, S. E., Levorse, J. M., Tilghman, S. M. & Vogt, T. F. Clock regulatory elements control cyclic expression of Lunatic fringe during somitogenesis. *Dev. Cell* **3**, 75–84 (2002).

122. GeneCards Human Gene Database. MEOX1 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=MEOX1&keywords=MEOX1.

123. GeneCards Human Gene Database. MEOX2 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=MEOX2.

124. Berti, F. *et al.* Time course and side-by-side analysis of mesodermal, pre-myogenic, myogenic and differentiated cell markers in the chicken model for skeletal muscle formation. *J. Anat.* **227**, 361–382

(2015).

125. GeneCards Human Gene Database. PAX7 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=PAX7&keywords=PAX7.

126. Burgess, R., Rawls, A., Brown, D., Bradley, A. & Olson, E. N. Requirement of the paraxis gene for somite formation and musculoskeletal patterning. *Nature* **384**, 570–573 (1996).

127. GeneCards Human Gene Database. TBX22 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=TBX22&keywords=TBX22.

128. Wanglar, C., Takahashi, J., Yabe, T. & Takada, S. Tbx protein level critical for clock-mediated somite positioning is regulated through interaction between Tbx and Ripply. *PLoS One* **9**, e107928 (2014).

129. GeneCards Human Gene Database. RIPPLY1 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=RIPPLY1&keywords=RIPPLY1.

130. GeneCards Human Gene Database. HEY1 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=HEY1&keywords=HEY1.

131. GeneCards Human Gene Database. TWIST1 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=TWIST1&keywords=TWIST1.

132. Nóbrega, A., Maia-Fernandes, A. C. & Andrade, R. P. Altered Cogs of the Clock: Insights into the Embryonic Etiology of Spondylocostal Dysostosis. *J Dev Biol* **9**, (2021).

133. GeneCards Human Gene Database. PENK Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=PENK&keywords=PENK.

134. GeneCards Human Gene Database. NPY Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=NPY&keywords=NPY.

135. GeneCards Human Gene Database. GRIK3 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=GRIK3&keywords=GRIK3.

136. GeneCards Human Gene Database. GRM4 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=GRM4&keywords=GRM4.

137. GeneCards Human Gene Database. SOX8 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=SOX8&keywords=SOX8.

138. GeneCards Human Gene Database. LZTS3 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=LZTS3&keywords=LZTS3.

139. GeneCards Human Gene Database. ROBO2 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=ROBO2&keywords=ROBO2.

140. GeneCards Human Gene Database. MAPK11 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=MAPK11&keywords=MAPK11.

141. GeneCards Human Gene Database. RHOU Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=RHOU&keywords=RHOU.

142. GeneCards Human Gene Database. BMP3 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=BMP3&keywords=BMP3.

143. GeneCards Human Gene Database. JPH1 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=JPH1&keywords=JPH1.

144. GeneCards Human Gene Database. MMP9 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=MMP9&keywords=MMP9.

145. GeneCards Human Gene Database. PTCH1 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=PTCH1&keywords=PTCH1.

146. Pourquié, O. The segmentation clock: converting embryonic time into spatial pattern. *Science* **301**, 328–330 (2003).

147. Trzaskowski, B. *et al.* Action of molecular switches in GPCRs--theoretical and experimental studies. *Curr. Med. Chem.* **19**, 1090–1109 (2012).

148. GeneCards Human Gene Database. PITX1 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=PITX1&keywords=PITX1.

149. GeneCards Human Gene Database. PTCH2 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=PTCH2.

150. GeneCards Human Gene Database. HBA1 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=HBA1&keywords=HBA1.

151. GeneCards Human Gene Database. HBZ Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=HBZ&keywords=hbz.

152. GeneCards Human Gene Database. CTR9 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=CTR9&keywords=CTR9.

153. GeneCards Human Gene Database. USP42 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=USP42&keywords=USP42.

154. GeneCards Human Gene Database. CDC73 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=CDC73&keywords=CDC73.

155. GeneCards Human Gene Database. MIB1 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=MIB1&keywords=MIB1.

156. GeneCards Human Gene Database. RNF151 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=RNF151&keywords=RNF151.

157. GeneCards Human Gene Database. KAT6B Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=KAT6B&keywords=KAT6B.

158. GeneCards Human Gene Database. UHRF1 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=UHRF1&keywords=UHRF1.

159. GeneCards Human Gene Database. TOP3A Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=TOP3A&keywords=TOP3A.

160. GeneCards Human Gene Database. KDM1A Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=KDM1A&keywords=KDM1A.

161. GeneCards Human Gene Database. LSAMP Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=LSAMP&keywords=LSAMP.

162. GeneCards Human Gene Database. GRIN3A Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=GRIN3A&keywords=GRIN3A.

163. GeneCards Human Gene Database. NLGN1 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=NLGN1&keywords=NLGN1.

164. GeneCards Human Gene Database. NEK9 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=NEK9&keywords=NEK9.

165. GeneCards Human Gene Database. PCM1 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=PCM1&keywords=PCM1.

166. GeneCards Human Gene Database. CLSPN Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=CLSPN&keywords=CLSPN.

167. GeneCards Human Gene Database. G2E3 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=G2E3&keywords=G2E3.

168. GeneCards Human Gene Database. HOXD11 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=HOXD11&keywords=HOXD11.

169. GeneCards Human Gene Database. HOXD13 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=HOXD13&keywords=HOXD11.

170. GeneCards Human Gene Database. RARB Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=RARB&keywords=RARB.

171. Geisha. in *The Visual Dictionary of Fashion Design* 121–121 (AVA Publishing SA Distributed by Thames & Hudson (ex-North America) Distributed in the USA & Canada by: English Language Support Office, 2007).

172. GeneCards Human Gene Database. ATF7IP Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=ATF7IP&keywords=ATF7IP.

173. Henrique, D. & Schweisguth, F. Mechanisms of Notch signaling: a simple logic deployed in time and space. *Development* **146**, (2019).

174. GeneCards Human Gene Database. VTN Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=VTN&keywords=vtn.

175. GeneCards Human Gene Database. VCAN Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=VCAN&keywords=vcan.

176. GEO Accession viewer. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE116935.

177. Dessaud, E., McMahon, A. P. & Briscoe, J. Pattern formation in the vertebrate neural tube: a sonic hedgehog morphogen-regulated transcriptional network. *Development* **135**, 2489–2503 (2008).

178. Stasiulewicz, M. *et al.* A conserved role for Notch signaling in priming the cellular response to Shh through ciliary localisation of the key Shh transducer Smo. *Development* **142**, 2291–2303 (2015).

179. GeneCards Human Gene Database. UBB Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=UBB&keywords=UBB.

180. GeneCards Human Gene Database. UBC Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=UBC&keywords=UBC.

181. GeneCards Human Gene Database. UBA52 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=UBA52&keywords=UBA52.

182. Resende, T. P. *et al.* Sonic hedgehog in temporal control of somite formation. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 12907–12912 (2010).

183. Sheeba, C. J., Palmeirim, I. & Andrade, R. P. Retinoic acid signaling regulates embryonic clock hairy2 gene expression in the developing chick limb. *Biochem. Biophys. Res. Commun.* **423**, 889–894 (2012).

184. GeneCards Human Gene Database. KDR Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=KDR&keywords=kdr.

185. Karaman, S., Leppänen, V.-M. & Alitalo, K. Vascular endothelial growth factor signaling in development and disease. *Development* **145**, (2018).

186. GeneCards Human Gene Database. NOTCH2 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=NOTCH2&keywords=notch2.

187. Baumgart, A. *et al.* Opposing role of Notch1 and Notch2 in a Kras(G12D)-driven murine non-small cell lung cancer model. *Oncogene* **34**, 578–588 (2015).

188. GeneCards Human Gene Database. YAP1 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=YAP1&keywords=YAP1.

189. Totaro, A. *et al.* YAP/TAZ link cell mechanics to Notch signalling to control epidermal stem cell fate. *Nat. Commun.* **8**, 15206 (2017).

190. Jaehning, J. A. The Paf1 complex: platform or player in RNA polymerase II transcription? *Biochim. Biophys. Acta* **1799**, 379–388 (2010).

191. Martin, G. S. The hunting of the Src. *Nat. Rev. Mol. Cell Biol.* **2**, 467–475 (2001).

192. Kerjouan, A. *et al.* Molecular flux control encodes distinct cytoskeletal responses by specifying SRC signaling pathway usage. *bioRxiv* 648030 (2019) doi:10.1101/648030.

193. Kao, T.-J., Palmesino, E. & Kania, A. SRC family kinases are required for limb trajectory selection by spinal motor axons. *J. Neurosci.* **29**, 5690–5700 (2009).

194. GeneCards Human Gene Database. TCF15 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=TCF15&keywords=TCF15.

195. GeneCards Human Gene Database. FGFRL1 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=FGFRL1&keywords=FGFRL1.

196. GeneCards Human Gene Database. FGFR3 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=FGFR3&keywords=FGFR3.

197. GeneCards Human Gene Database. RDH10 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=RDH10.

198. GeneCards Human Gene Database. RIPPLY2 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=RIPPLY2&keywords=RIPPLY2.

199. GeneCards Human Gene Database. FSTL4 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=FSTL4&keywords=FSTL4.

200. GeneCards Human Gene Database. SOX17 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=SOX17&keywords=SOX17.

201. GeneCards Human Gene Database. RIPK4 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=RIPK4&keywords=RIPK4.

202. GeneCards Human Gene Database. BMP2 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=BMP2&keywords=BMP2.

203. GeneCards Human Gene Database. CTNNA1 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=CTNNA1&keywords=CTNNA1.

204. GeneCards Human Gene Database. FGFR2 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=FGFR2&keywords=FGFR2.

205. GeneCards Human Gene Database. PRLR Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=PRLR&keywords=PRLR.

206. GeneCards Human Gene Database. TRRAP Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=TRRAP&keywords=TRRAP.

6 | MASTERS PORTFOLIO

Publications

- *__Liber, Marta__; *Duarte, Isabel; Andrade, Raquel P. (__2021__, Feb 26). *FrozenChicken: Promoting the meta-analysis of chicken microarray data.* __bioRxiv__. https://doi.org/10.1101/2021.02.25.432894 (* equal contribution). (Annex A3.1)

- * Duarte, Isabel, * __Liber, Marta__, & Andrade, Raquel P. (__2020__, April 25). *FrozenChicken: Promoting the meta-analysis of chicken microarray data* (Version 1.0). __Zenodo__. http://doi.org/10.5281/zenodo.3765944 (* equal contribution). (Annex A3.2)

Selected Talks

• "*FrozenChicken: Promoting the meta-analysis of chicken microarray data*" – Special DiA Meeting, Universidade do Algrve, Faro, Portugal. (Annex 10)	June, 2019
• "ClockOME: Searching for oscillatory genes in early vertebrate development" – **selected** for IMPSG 2021	July, 2021

Posters

• "FrozenChicken: Promoting the meta-analysis of chicken microarray data" – Special DiA Meeting, Universidade do Algarve, Faro, Portugal. (Annex 11)	June, 2019
• "ClockOME: Searching for oscillatory genes in early vertebrate development" – **selected** for IMPSG 2021 (Annex 12)	July, 2021

Seminars

• "How to find a clock without a watch" – LabClub, Universidade do Algarve, Faro, Portugal. (Annex 13)	March, 2020

Congress & Symposia attended

• Portuguese society for developmental biology - Special DiA Meeting (**poster selected for Short Talk**)	June, 2019

• Biomedical Engineering Summit	October, 2019
• International meeting of the Portuguese society of genetics (IMPSG)	January, 2020
• “From Cells to Embryo" International meeting	November, 2020
• 13th Berlin Summer Meeting: “Rising from the Ashes – Regeneration at the Single Cell Level”	December, 2020
• II International meeting of the Portuguese society of genetics (IMPSG-2021)	July, 2021

Courses & Workshops attended

• “Data and Text Processing in Health and Life Sciences - an example driven workshop using shell scripting”	July, 2019
• TRANSAUTOPHAGY Translational Workshop	January, 2020
• UAlg Open day	February, 2020
• Live Webinar: Interactive bioimage analysis with Python and Jupyter - NEUBIAS Academy	May, 2020
• Computer Science Fundamentals MicroBachelors	September – November, 2020