



## Biological systems on a small scale

Pedersen, Lykke

*Publication date:*  
2012

*Document version*  
Early version, also known as pre-print

*Citation for published version (APA):*  
Pedersen, L. (2012). *Biological systems on a small scale*. Center for Ice & Climate: Faculty of Science, University of Copenhagen.

**Faculty of Science  
University of Copenhagen**

**PhD Thesis**



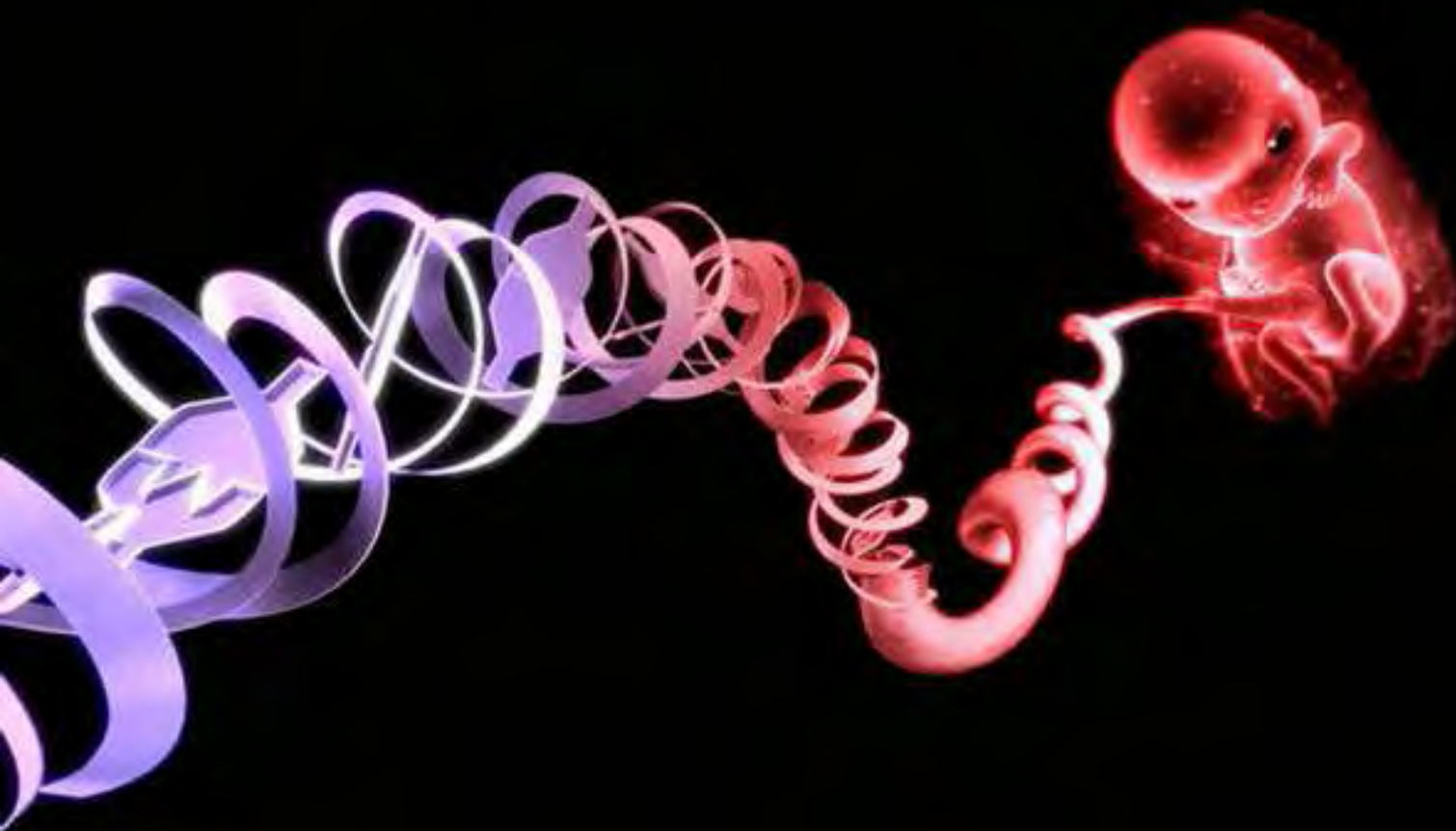
# **Biological systems on a small scale**

Embryo development and microRNA regulations

**Lykke Pedersen**

Center for Models of Life

Niels Bohr Institute, University of Copenhagen



**Supervisor:**  
Prof. Mogens Høgh Jensen  
University of Copenhagen

**Submitted:**  
March 28, 2012

The PhD School of Science  
Faculty of Science  
Niels Bohr Institute  
Center for Models of Life  
University of Copenhagen  
Denmark

Picture on frontage is adapted from <http://www.sciencephoto.com>

# Abstract

This thesis covers the studies of two distinct biological systems: embryo development and miRNA regulations.

A segmented body plan is the defining characteristic of vertebrates. The process of segmentation is carried out by a combination of changes in gene expressions and relative cell positions within the developing embryo. The genes influencing the segmentation in embryos are differentially expressed in cells depending on their anterior-posterior (head-to-tail) position in the embryo. In the anterior end somites are formed during a process called somitogenesis. The somites are small bundles of cells that appear in pairs along the future vertebral column, like beads on a string. They are the precursors of muscles, skeleton, and skin.

The Wnt signaling pathway is essential in controlling somitogenesis. Several Wnt target genes have oscillating expressions, which could constitute a segmentation clock. A clock that controls the very precise period of somite formation. Two theoretical models for the Wnt pathway are proposed within this thesis. These models explore features such as synchronization of neighboring cells in the anterior part of the embryo and termination of somitogenesis are explored. The mechanism behind the latter is unknown, but here a theoretical explanation of these is given.

MicroRNAs (miRNAs) are small RNA molecules that bind to protein coding messenger RNAs (mRNAs) and inhibit their translation. An inhibition that can either cause or prevent a disease to develop. Therefore miRNAs have the potential as important components in diagnosing and



treating several diseases such as cancer. The problem is to figure out which miRNAs bind to which mRNAs and simultaneously identify those that are deregulated upon development of a disease. The first is known as miRNA target prediction. In this thesis a solution to this problem is given by means of a new miRNA target prediction method that exploits the large amount of information stored in data for gene expressions. It is based on the use of independent component analysis in combination with sequence and pathway analyses. Ultimately, by this method it is possible to unravel the huge network that our genome comprises and predict miRNAs deregulated in type 1 diabetes and propose for miRNA mediated deregulations of pathways in ovarian cancer.

## Danish summary/Dansk resumé

Denne afhandling omhandler studier af to særskilte biologiske systemer: embryonal udvikling og mikroRNA regulering.

En segmenteret kropsopbygning er det mest karakteristiske for alle hvirveldyr. Selve processen bag segmenteringen sker som en følge af ændringer i genudtryk og relativ placering af celler i embryoet. Generne som påvirker segmenteringsprocessen er forskelligt udtrykt i embryoniske celler afhængigt af deres anterior-posterior (hoved-til-hale) placering. I den anteriore del af embryoet dannes somitterne under en proces kaldet somitogenese. Somitterne er små bundter af celler dannet i par langs den kommende rygsøjle, som perler på en snor. De er forstadierne til muskler, skelet og hud.

Wnt signalering er essentiel for kontrolleringen af somitogenese. Mange gener, som er styret af Wnt signalering, har oscillerende ekspressioner. Disse kunne udgøre et segmenteringsur. Et ur som styrer den meget præcise periode for dannelsen af somitter. To teoretiske modeller for Wnt signalering foreslås i denne afhandling. Ved hjælp af disse to modeller udforskes vigtigheden af synkronisering mellem nabo celler i den anterior del af embryoet. Desuden gives en teoretisk forklaring på, hvorledes segmenteringsprocessen stopper. Et fænomen der stadig ikke er fuldt forstået og beskrevet.

MikroRNAer (miRNAer) er små RNA molekyler, som binder til protein kodende messenger RNAer (mRNAer) og hæmmer translationen af disse. En hæmning, der enten kan forårsage eller forhindre sygdomsudvikling. Derfor har miRNAer et potentiale som vigtige komponenter for diagnostiseringen og behandlingen af bl.a. kræft. Problemet er at kæde de rigtige miR-

NAer sammen med de rigtige mRNAer og samtidig identificere dem, som ændrer ekspressions niveau når en sygdom udvikles. Det første er kendt som forudsigelse af miRNA interaktioner. I denne afhandling gives en mulig løsning af problemet i form af en ny metode til at forudsige miRNA interaktioner, som udnytter den store informations mængde gemt i data for genekspressioner. Metoden bygger på brugen af independent component analysis i sammenhæng med sekvens og metabolisk signalerings analyse. Ultimativt, vil denne metode kunne bruges til at optræfle det store netværk vores genom udgør og forudsige miRNAer, der er dereguleret i og type 1 diabetes og foreslå signalerings veje, som er dereguleret i æggestokkræft pga. ændringer i miRNA ekspressioner.

# Preface

This thesis is submitted in fulfilment with the requirements for obtaining a PhD degree at the Niels Bohr Institute, Faculty of Science, University of Copenhagen. The PhD study has been conducted under the supervision of Prof. Mogens Høgh Jensen at the Niels Bohr Institute. It was initiated in July 2008 and ended in March 2012 including 9 months of maternity leave.

The thesis includes four peer-reviewed papers Jensen et al. [1], Mengel et al. [2], Bang-Berthelsen et al. [3] and Pedersen et al. [4] and two submitted papers Pedersen and Hagedorn [5] and Pedersen et al. [6]. In the publication list full references for all papers are listed in the following pages.

The thesis is divided into two parts. The first part covers embryo development and the second miRNA target prediction. Each part is a synopsis of the most interesting and novel observations from the six papers included in the thesis. The reader is therefore encouraged to consult these papers included at the end of the thesis in Appendices B-G. For papers including supplementary files, these are included as well besides those that are excel-files. In Appendix A a short description of some general statistics is given, which is used in the second part of the thesis.

During my PhD studies I collaborated with the group of Jim McGuire at Hagedorn Research Institute on identifying diabetic nephropathy biomarkers. My contribution to this collaboration was the analysis of proteomic data by use of independent component analysis and the analysis is included in Overgaard et al. [7]; Hansen et al. [8]. Furthermore, I also took part in a collaboration with the group of Jens Høiriss Nielsen at the Department of Biomedical Sciences, University of Copenhagen. The aim of the study was

to identify miRNAs differentially regulated during the development of pancreas in mouse embryos and is published in Larsen et al. [9]. I contributed with an analysis of miRNA expression data. The papers of Overgaard et al. [7]; Hansen et al. [8]; Larsen et al. [9] are not included in the thesis and consequently they will not be discussed further.

All simulations and analyses performed in this thesis are carried by the author using the R environment for statistical computing and graphics (<http://www.r-project.org/>). Gene names are written in italic (*axin2*), mRNA names with low case letters (axin2), and protein names with a capital letter (Axin2).

## Acknowledgements

First of all I would like to thank Mogens Høgh Jensen, my supervisor, who have supported my ideas and guided me all through the PhD study. From the Niels Bohr Institute I would also like to thank Sandeep Krishna, now located at NCBS, India, who has been a great support during the study of embryo development.

Secondly, I would like to thank all the members of Flemming Pociots group at Glostrup Hospital for giving me the opportunity to work on their type 1 diabetes data in the development of the miRNA target prediction method. A special thanks goes to Claus H Bang-Berthelsen for sharing and passing on his deep knowledge of type 1 diabetes. In addition to this I would like to thank Peter H Hagedorn for his great cooperation and useful discussions during the development of the weighting scheme of independent components and the miRNA target prediction method.

Thirdly, I would like to thank Preethi Gunaratne, University of Houston, for providing the data of ovarian cancer. I have benefitted greatly from her broad knowledge of ovarian cancer and she is an inspiring person to collaborate with.

One year into my PhD studies I went for a six months stay at the Paulsson lab, Harvard Medical School. There I conducted research on stochastic models of biochemical reactions with Johan Paulsson as supervisor. Especially the stochasticity involved in two small molecules binding to a DNA string. Despite the many attempts I did not manage to achieve any results from this study that can be included in this thesis. But still I am grateful for the great hospitality and company of the Paulsson lab during the long working hours and their useful discussions.

Lastly, I would like to thank my family for their great support and infinite patience. You comprise the small things in life that matters for me.

# Publication list

## Publications included in this thesis

A name is given to each paper, which is used as its reference in the thesis.

### **AXIN2 paper [1]:**

P. B. Jensen, L. Pedersen, S. Krishna, and M. H. Jensen. *A Wnt oscillator model for somitogenesis*. Biophysical journal, 98(6):943–950, March 2010

### **OSC paper [2]:**

B. Mengel, A. Hunziker, L. Pedersen, A. Trusina, M. H. Jensen, and S. Krishna. *Modeling oscillatory control in NF-kappaB, p53 and Wnt signaling*. Curr Opin Genet Dev, 20(6):656–664, December 2010

### **T1D paper [3]:**

C. H. Bang-Berthelsen, L. Pedersen, T. Fløyel, P. H. Hagedorn, T. Gylvin, and F. Pociot. *Independent component and pathway-based analysis of miRNA-regulated gene expression in a model of type 1 diabetes*. BMC Genomics, 12:97, 2011

### **DKK1 paper [4]:**

L. Pedersen, M. H. Jensen, and S. Krishna. *Dickkopf1 - a new player in modelling the wnt pathway*. PLoS ONE, 6(10):e25550, 2011

**ICA paper [5]:**

L. Pedersen and P. H. Hagedorn. Ranking of Independent Components in Gene Expression Microarray Data Using Weighted Scoring. Under revision in PLoS ONE, 2011

**OVC paper [6]:**

L. Pedersen, W. Xiao, M. Jensen, and P. Gunaratne. Independent component analysis reveals possible microRNA regulated pathways in ovarian cancer. Submitted to PLoS Computational, 2012

**Publications not included in this thesis**

- [7] A. J. Overgaard, H. G. Hansen, M. Lajer, L. Pedersen, L. Tarnow, P. Rossing, J. N. McGuire, and F. Pociot. *Plasma proteome analysis of patients with type 1 diabetes with diabetic nephropathy*. Proteome Sci, 8:4, 2010
- [8] H. G. Hansen, J. Overgaard, M. Lajer, F. Hubalek, P. Højrup, L. Pedersen, L. Tarnow, P. Rossing, F. Pociot, and J. N. McGuire. *Finding diabetic nephropathy biomarkers in the plasma peptidome by high-throughput magnetic bead processing and MALDI-TOF-MS analysis*. Proteomics Clin Appl, 4(8-9):697–705, September 2010
- [9] L. Larsen, M. W. Rosenstjerne, L. W. Gaarn, A. Bagge, L. Pedersen, C. M. Dahmcke, J. H. Nielsen, and L. T. Dalgaard. *Expression and localization of microRNAs in perinatal rat pancreas: role of miR-21 in regulation of cholesterol metabolism*. PLoS ONE, 6(10):e25997, 2011

# Contents

<b>Abstract</b>	<b>i</b>
<b>Danish summary/Dansk resumé</b>	<b>iii</b>
<b>Preface</b>	<b>v</b>
<b>Publication list</b>	<b>vii</b>
<b>Contents</b>	<b>1</b>
<b>List of abbreviations</b>	<b>5</b>
 <b>I EMBRYO DEVELOPMENT</b>	 <b>7</b>
<b>1 Embryo segmentation</b>	<b>9</b>
1.1 Embryogenesis . . . . .	10
1.2 The Clock and Wavefront model . . . . .	11
1.2.1 The clock . . . . .	11
1.2.2 The wavefront . . . . .	13
1.2.3 Linking the clock and wavefront . . . . .	13
 <b>2 Oscillations in a dynamical system</b>	 <b>17</b>
2.1 Gene regulatory networks . . . . .	18
2.2 Feedback loops . . . . .	19
2.2.1 Stability analysis of feedback loops . . . . .	20



<b>3</b>	<b>Modeling embryo segmentation</b>	<b>23</b>
3.1	Existing models of segmentation . . . . .	24
3.2	A model for somitogenesis based on Axin2 . . . . .	24
3.2.1	The importance of Saturated degradation . . . . .	27
3.3	Dkk1 an inhibitor of Wnt signaling . . . . .	28
3.3.1	Incorporating a Wnt gradient in the PSM . . . . .	31
3.3.2	Synchronization of neighboring cells at the tail bud . .	32
3.3.3	Longer periods towards the end of somitogenesis . . .	33
3.3.4	Sustained or decaying oscillations - what is needed? .	34
<b>4</b>	<b>Concluding remarks Part I</b>	<b>37</b>
<b>II</b>	<b>MICRORNA TARGET PREDICTION</b>	<b>39</b>
<b>5</b>	<b>Complexity of miRNA target prediction</b>	<b>41</b>
5.1	microRNA biogenesis and regulation . . . . .	42
5.2	Predicting miRNA targets . . . . .	43
5.2.1	Accuracy of available target prediction methods . . .	44
5.2.2	Expression data used for miRNA target prediction . .	45
<b>6</b>	<b>miRNA target prediction using independent component analysis</b>	<b>47</b>
6.1	Principles of independent component analysis . . . . .	48
6.1.1	ICA applied to gene expression data . . . . .	49
6.1.2	Ranking of independent components . . . . .	50
6.2	miICA - miRNA target prediction based on ICA . . . . .	51
<b>7</b>	<b>miICA applied to a model of type 1 diabetes</b>	<b>55</b>
7.1	An introduction to type 1 diabetes . . . . .	56
7.2	Type 1 diabetes microarray data . . . . .	56
7.3	Independent components capture the expression of miRNAs .	56
7.3.1	miICA performs better than negative correlation . . .	58
7.3.2	Cooperativity between miRNAs . . . . .	58
7.4	Combining the effects of independent components . . . . .	59
7.5	Components map to diabetes related pathways . . . . .	60
7.6	Further insight into miRNA regulations . . . . .	61
<b>8</b>	<b>miICA reveals miRNA regulations in ovarian cancer</b>	<b>63</b>
8.1	Ovarian cancer microarray data . . . . .	64
8.2	The p53 pathway is deregulated in ovarian cancer . . . . .	64
8.3	miRNA regulations of the Wnt and mTOR pathways . . . . .	64
8.4	Fingerprints of miRNA regulations . . . . .	66
8.5	miR-374a and p53 form a new regulatory relationship . . . .	67

<b>9</b>	<b>Concluding remarks Part II</b>	<b>69</b>
	<b>Bibliography</b>	<b>71</b>
<b>A</b>	<b>Basic statistics</b>	<b>87</b>
A.1	ANOVA - analysis of variance . . . . .	87
A.2	Coefficient of determination . . . . .	87
A.3	Sensitivity and specificity . . . . .	88
A.4	Correction for multiple testing . . . . .	88
<b>B</b>	<b>AXIN2 paper</b>	<b>91</b>
<b>C</b>	<b>OSC paper</b>	<b>113</b>
<b>D</b>	<b>T1D paper</b>	<b>123</b>
<b>E</b>	<b>DKK1 paper</b>	<b>137</b>
<b>F</b>	<b>ICA paper</b>	<b>151</b>
<b>G</b>	<b>OVC paper</b>	<b>165</b>



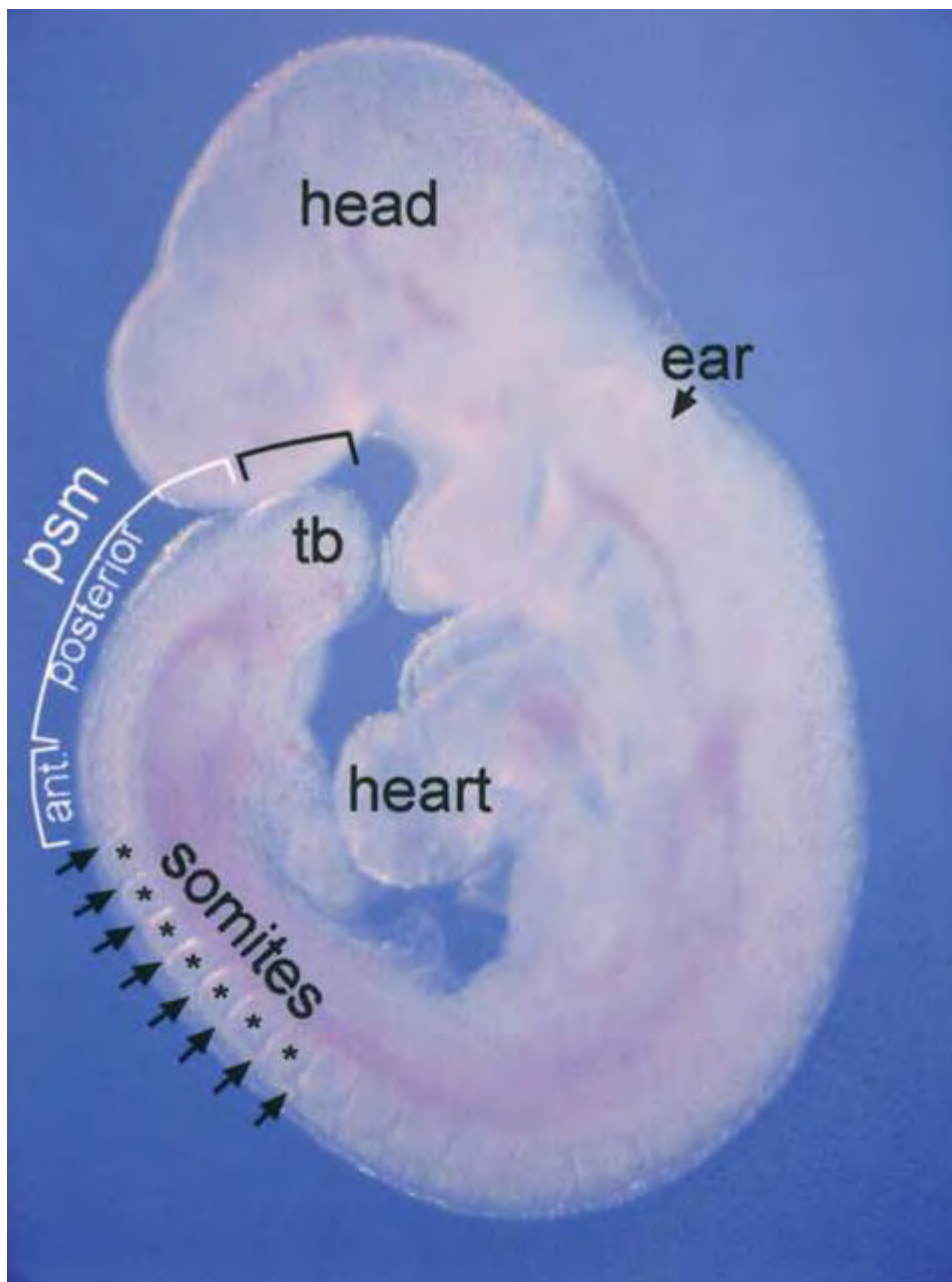
## List of abbreviations

3'UTR	3' untranslated region
5'UTR	5' untranslated region
ANOVA	Analysis of variance
APC	Adenomatous polyposis coli
AUC	Area under the curve
BSS	Blind source separation
CKI $\alpha$	Casein kinase
Dkk1	Dickkopf1
Dll	Delta-like receptor
FDR	false discovery rate
FGF	Fibroblast growth factor
FWER	Family wise error rate
GO	Gene Ontology
GRN	Gene regulatory network
GSK3 $\beta$	Glycogen synthase kinase
HTS	high-throughput sequencing

ICA	Independent component analysis
IC	Independent component
IL-1 $\beta$	Interleukin-1 $\beta$
KEGG	Kyoto Encyclopedia of Genes and Genomes
K-S test	Kolmogorov-Smirnov test
LRP5/6	Low-density lipoprotein receptor-related protein 5 and 6
miRNA	MicroRNA
mRNA	Messenger RNA
NGS	Next-generation-sequencing
NICD	Notch intracellular domain
ODE	Ordinary differential equation
OVC	Ovarian cancer
PCA	Principal component analysis
Pdx-1	Pancreatic and duodenal homeobox 1
PSM	Presomitic mesoderm
$R^2$	Coefficient of determination
RA	Retinoic acid
RISC	RNA- induced silencing complex
ROC	Receiver operating characteristic curve
SVM	Support vector machine
T1D	Type 1 diabetes mellitus
T2D	Type 2 diabetes mellitus
TCGA	The Cancer Genome Atlas
WRS test	Wilcoxon-rank-sum test

## Part I

# EMBRYO DEVELOPMENT





## CHAPTER

# 1

## Embryo segmentation

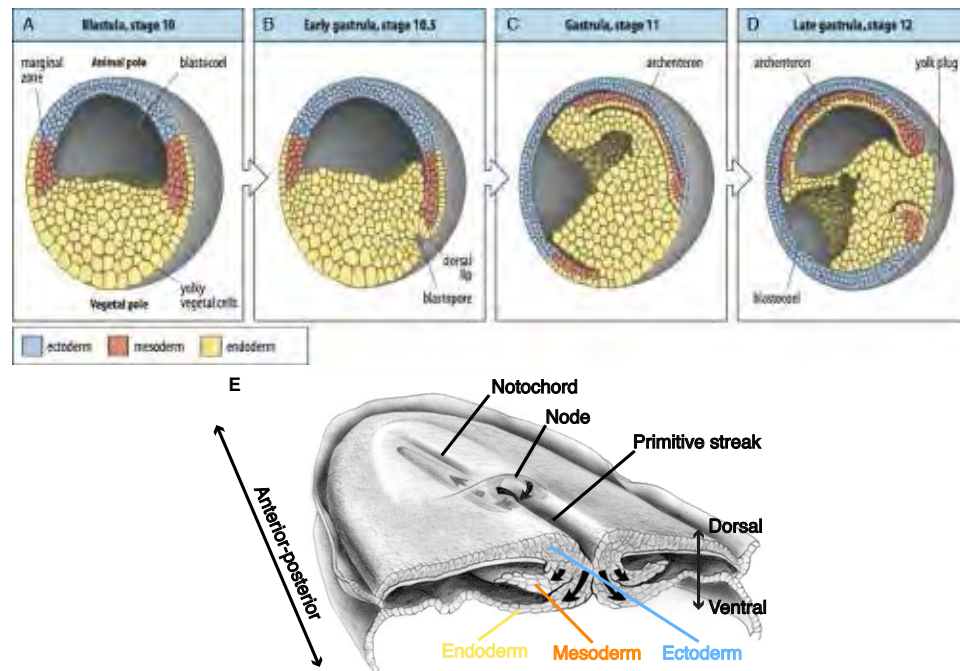
Everything has a life cycle, from the daily newspaper to the long life of tortoises, which live more than 150 years. For metazoan animals the life cycle starts with fertilization of an egg, and then follows cleavage, gastrulation, germ layer formation, organ formation, metamorphosis or birth, maturation and growth, and it ends with senescence. The part of the life cycle that comes in-between the formation of germ layers and organ formation is important for proper axis formation in vertebrate animals. Because, at this developmental stage the precursors of the vertebrates are formed. These are called somites and the formation of somites (somitogenesis) is the main focus for the chapters to follow.



## 1.1 Embryogenesis

At the earliest stage of embryo development (embryogenesis) the fertilized egg is cleaved into smaller cells (blastomeres)<sup>1</sup>. After cleavage several thousand blastomeres form a hollow sphere called the blastula (Fig. 1.1A). During gastrulation the one-layered blastula is rearranged to form a three-layered structure, the gastrula (Fig. 1.1D). The three layers are called ectoderm, endoderm and mesoderm and they are progenitors of specific future tissues, for example, part of the mesoderm becomes the skeleton. The rearrangement of the blastula happens by involution of the endoderm and mesoderm while the ectoderm spreads to surround the endoderm and mesoderm (Fig. 1.1A-D).

During gastrulation the primitive streak is formed as the indentation caused by the involution of the ectoderm and mesoderm (Fig. 1.1E). The primitive streak is the first embryonic axis marking the anterior-posterior direction. At the anterior end of the primitive streak there is a thickening of cells called Hensen's node or the primitive knot. The function of the node is to organize gastrulation and it moves posteriorly as the primitive



**Figure 1.1** Embryonic gastrulation.

(A-D) The involution of the endoderm and mesoderm during gastrulation. (E) A cross-section of the embryo displaying the three germ layers, the node, the primitive streak and the notochord. Figures adapted from Wolpert et al. [10].

<sup>1</sup>This section is based on Wolpert et al. [10] and Gilbert [11].

streak extends. The most dorsal cells in the region around the node forms a rigid rod-like notochord. The neural tube is formed from ectodermal cells overlying the notochord and the tube will develop into the spinal cord and brain. When regression of the primitive streak sets in, then the node moves towards the posterior end of the embryo and the primitive streak stops growing.

On each side of the notochord somites are formed in pairs as bundles of mesodermal cells (see Fig. 1.2). The mesodermal cells are part of the presomitic mesoderm (PSM), which lies on each side of the notochord and extends from the anterior most somite to the tail bud. The tail bud consist of stem cells that are remnants of the node, and it moves posteriorly as stem cells divide contributing to the elongation of the embryo.

The somites are segmented at regular time intervals that are species dependent, i.e., they are segmented every 90 min in frogs and every 120 min in mice. Also, the number of somites is species dependent. For example, humans have 33 somites and mice have 65 somites. The setting of somite number is unknown, as is the mechanism for how the somitogenesis terminates. On the other hand, the mechanism of how the somites are formed from cells within the PSM has been explained and is the subject for the next section.

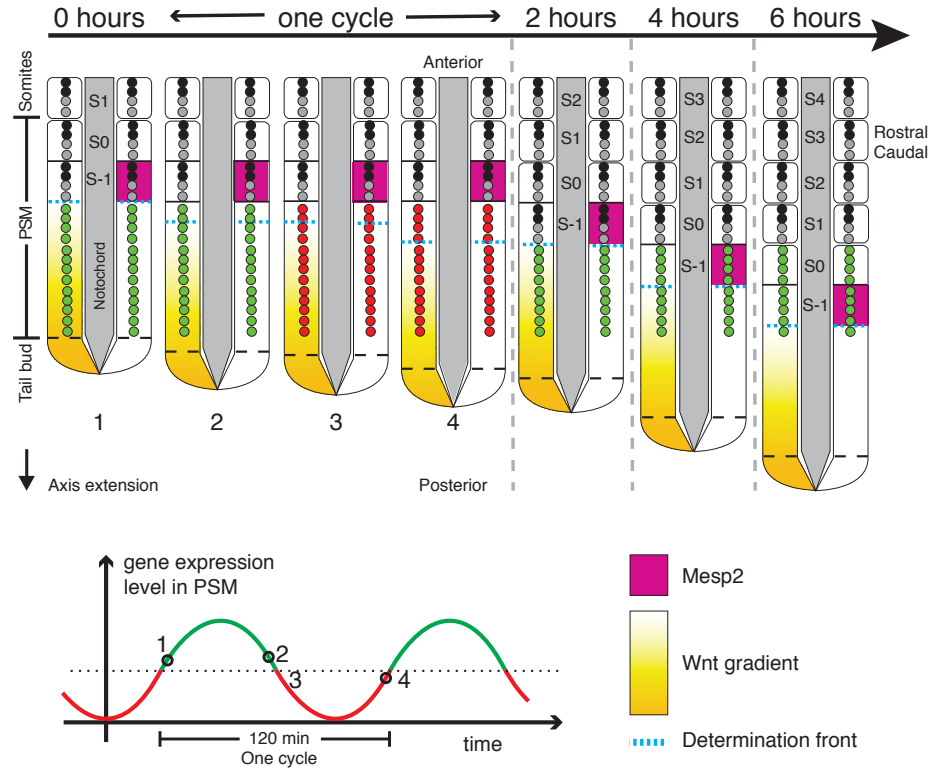
## 1.2 The Clock and Wavefront model

During the years different types of models have been proposed to explain the periodicity of somitogenesis. The most recognized is the Clock and Wavefront model proposed in 1976 by Cooke and Zeeman [12]. There the timing of the PSM cells is set by an oscillating clock and the position is set by a wavefront. With these informations the PSM cells know when and where to become part of a somite. Figure 1.2 (p. 12) is a schematic presentation of the Clock and Wavefront model.

### 1.2.1 The clock

The oscillating clock that provides the temporal information has first been discovered through the oscillation of the *Axin2* gene in mouse embryos [13]. Since then a more genome-wide approach has been taken and several oscillating genes have been found [14]. These genes belong to the three major pathways in somitogenesis: Wnt, Notch and FGF. Interestingly, the genes of the Wnt pathway oscillate out of phase with the genes of the FGF and Notch pathways, and FGF acts upstream of Wnt, which acts upstream of Notch [15].

The Wnt pathway is often referred to as the canonical Wnt/ $\beta$ -catenin pathway, since the main function of the Wnt signal is to destabilize  $\beta$ -catenin through a phosphorylation process induced by the formation of a destruction



**Figure 1.2** Schematic drawing of the Clock and Wavefront model.

The cells within the PSM express genes in an oscillatory manner with a period of 120 min. The Wnt gradient decreases from the posterior to the anterior end of the PSM (shades of yellow). Past the determination front (cyan line) somites can be formed. Depending on the level of expression the cells are either permissive to form somites (green) or not (red). In somite S-1 cells from the PSM form a prospective somite, where the border of the somite has not been defined yet. A border has been defined for somite S0, however the actual segmentation has not yet occurred as is the case for somite S1. The black and grey colors of cells within a somite refers to their rostral and caudal positioning, respectively. Mesp2 (magenta) is expressed in S-1 cells.

complex comprised of the kinases glycogen synthase kinase ( $GSK3\beta$ ) and casein kinase ( $CKI\alpha$ ), the scaffolding proteins Axin and adenomatous polyposis coli (APC) and  $\beta$ -catenin itself. Wnt signaling is initiated by the binding of a Wnt ligand to the receptor-coreceptor complex comprised of Frizzled and low-density lipoprotein receptor-related protein 5 and 6 (LRP5/6) on the cell membrane [16]. Besides being active during somitogenesis the Wnt pathway is also active in many diseases. For example, a constitutive activation of Wnt caused by mutations leads to the development of some cancers [17]. For a more detailed description of the Wnt signaling pathway see the

AXIN2 paper and the DKK1 paper.

Synchronization between neighboring cells in the PSM is a key factor for proper somite formation [18]. The Notch pathway provides the link between neighboring cells by controlling the level of intracellular Notch (NICD), which is part of the Notch signal transduction. Notch is a transmembrane receptor with ligands also being transmembrane proteins and among them are the Delta like receptors (Dll). Upon binding of Dll from one cell to the Notch receptor on another cell Notch signaling is transferred. Disruption of the Notch pathway in developing embryos results in a salt-and-pepper pattern, when staining Notch target genes [18; 19].

### 1.2.2 The wavefront

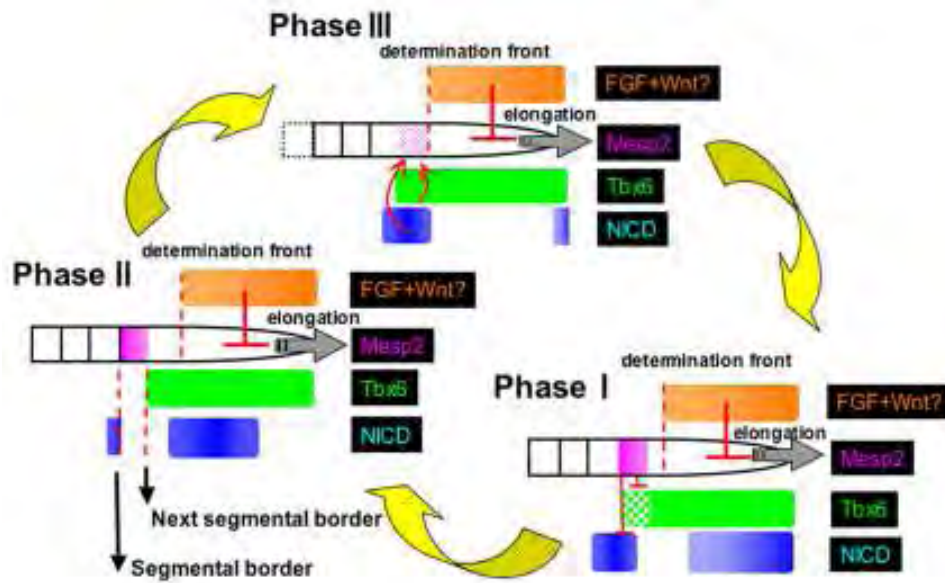
The wavefront is considered to be some morphogen gradient. There are three main gradients in the PSM, which can constitute the wavefront: Wnt, FGF and retinoic acid (RA) [20]. The Wnt and FGF gradients decrease anteriorly. At a certain level of FGF the cells become permissive to form a somite [21] (level 1 in Fig. 1.2). This level of FGF or stage of the FGF gradient is called the determination front (cyan line in Fig. 1.2). There are several members of the FGF family comprising the FGF gradient. *Fgf4* and *Fgf8* act in synergy to stop the oscillations, because separately they cannot constitute the determination front. The FGF members also act together with *Wnt3a*, which constitutes the Wnt gradient [13]. How the synergy between these two gradients is enabled is yet to be explored.

The third gradient of RA increases from the tail bud and throughout the PSM, thus opposing the gradient of Wnt and FGF. The FGF and RA gradients antagonize each other [22]: RA controls the expression of *fgf8*, and *Fgf8* in turn regulates the onset of RA synthesis. This interplay is thought to cause the right timing for the *Fgf8* level at the determination front.

### 1.2.3 Linking the clock and wavefront

The gene *Mesp2* is critical for stopping the oscillations and inducing the somite borders. It is believed to be the link between the temporal information from the clock and the spatial information from the wavefront [21]. The wavefront formed by a decreased level of FGF represses *Dll1* [23], which controls Notch signaling as described above. That is, as FGF levels decreases, Notch levels increases. Notch induces periodically expression of *Mesp2* [24] due to the periodic expression of Notch itself.

Just anteriorly to the determination front *Mesp2* is expressed in S-1 cells (colored magenta in Fig. 1.2), and the expression is initiated synchronously in the PSM due to the induction of Notch [21]. The posterior border of *Mesp2* expression is set by *Tbx6* [21]. Once activated by *Tbx6* *Mesp2* then represses translation of *tbx6* and an anterior border for the somite is set.



**Figure 1.3** Schematic illustration of Mesp2 regulations.

Somitogenesis is often divided into three phases depending on the phase of the oscillations. (Phase III) A certain FGF (orange) threshold sets the determination front and the posterior limit of the Mesp2 (pink) expression domain. When Notch activity (NICD, blue) reaches the anterior PSM, where Mesp2 transcription has been initiated in the cells with Tbx6 expression, then Mesp2 expression increases. (Phase I) Mesp2 protein accumulates and suppresses NICD and Tbx6 protein. (Phase II) When the next wave of Notch activity has just reached the anterior PSM region, the NICD, Mesp2 and Tbx6 expression domains are completely separated and the boundary between NICD and Mesp2 marks the segmental border. The boundary between Mesp2 and Tbx6 marks the next Mesp2 anterior limit. Adapted from Oginuma et al. [21].

Thus, Mesp2 is a post-translational regulator of Tbx6 because in Mesp2 null mutants, there is still a clear anterior border for the *tbx6* mRNA, but not for the Tbx6 protein [21]. FGF knock-out mouse showed no well defined anterior border of Tbx6 and no apparent oscillation of Notch. Thus, it is believed that at the determination front FGF reaches a certain level that triggers Tbx6 transcription to stop. If at the same time Notch signaling is high, then a somite border is formed [25]. Figure 1.3 schematically illustrates the regulation of Mesp2.

Another important role of Mesp2 is to set the rostral-caudal polarization of the somites. Because when the organs start to form the caudal part of a somite fuses with the rostral part of its posterior neighbor [26]. This resegmentation causes a proper formation of the skeleton and vertebrate. Thus, the somites need to have the correct rostral-caudal polarization. Once

somitogenesis terminates the organs start to form. The genes responsible for laying out the correct body plan, e.g., forming the right organs at the right place, are the Hox genes [27].



## CHAPTER

## 2

# Oscillations in a dynamical system

Whether it is the circadian clock, the inflammatory response of  $\text{Nf-}\kappa\text{B}$  or the segmentation process in embryos oscillations are observed. In the OSC paper it is deduced why oscillations are better in transmitting information than steady-state signals. In order to deduce this it is necessary to construct regulatory networks of the signaling pathways and investigate their behavior. In this chapter a broad overview of genetic regulation and feedback loops is given. The concepts of these are used in constructing the models of embryo segmentation described in Chapter 3.

This chapter includes results from:

B. Mengel, A. Hunziker, L. Pedersen, A. Trusina, M. H. Jensen, and S. Krishna. *Modeling oscillatory control in NF-kappaB, p53 and Wnt signaling*. Curr Opin Genet Dev, 20(6):656–664, December 2010



## 2.1 Gene regulatory networks

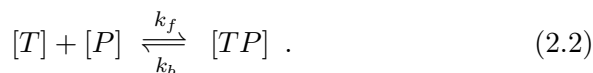
In a gene regulatory network (GRN) the chemical compounds, such as DNA, mRNA and proteins, constitute the nodes, and the links between these are their mutual interactions, e.g., complex formation, phosphorylation, and activation of transcription. The dynamics of a GRN can be simulated in numerous ways. The most common is to solve a system of ordinary differential equations (ODEs) to describe the dynamics of a GRN; if the system size is large enough so that fluctuations are equalled out. On the other hand if the system size is small and fluctuations have to be taken into account then Gillespie or Monte Carlo simulations can be performed to simulate a GRN.

In a deterministic formulation of a GRN the most basic dynamics needed to be included is the production/formation and degradation/breaking of a chemical compound. In terms of protein production the central dogma [28] states that DNA is transcribed into messenger RNA (mRNA), which in turn is translated into a protein. The transcription and translation are carried out at specific rates. In order to model a GRN it is often necessary to form a complex with a certain rate of formation and breaking. In the following a brief description is given for the mathematical formulation of protein production and degradation together with complex formation. This constitutes the basic tools used to write down the equations for the Wnt signaling pathway described in Chapter 3.

**Translation and degradation** The simplest form of protein dynamics include the translation of mRNA into protein and the degradation of protein. In the simplest example possible the translation is assumed to occur at a constant rate ( $c_{tsl}$ ) and degradation with a protein half-life of  $\tau_A$ . The ODE for the changes in protein concentration ( $A$ ) is then

$$\frac{dA}{dt} = c_{tsl}A_m - \frac{A}{\tau_A} . \quad (2.1)$$

**Transcription and complex formation** In the case of a transcription factor ( $T$ ) binding at the promoter ( $P$ ) of a gene the following chemical reaction is considered



The rates  $k_f$  and  $k_b$  are the rates at which the transcription factor binds to or dissociates with the promoter, respectively. At equilibrium these rates are equal and from the law of mass action follows

$$[T][P]k_f = [TP]k_b . \quad (2.3)$$

Defining  $K = \frac{k_b}{k_f}$  gives

$$K = \frac{[T][P]}{[TP]} \Rightarrow \frac{[TP]}{P_{\text{tot}}} = \frac{[T]}{K + [T]} , \quad (2.4)$$

wheres it is used that the total concentration of promoter is

$$P_{\text{tot}} = [P] + [TP] . \quad (2.5)$$

For low values of  $K$  the promoter sites are highly saturated even for low concentrations of  $[T]$ , i.e., almost all promoter sites have a transcription factor bound to it. In this case the ratio in Eq. 2.4 is approximately one meaning that the concentration of total promoter sites ( $P_{\text{tot}}$ ) equals the concentration of occupied promoter sites ( $[TP]$ ). The type of kinetics, where one species is a limiting factor, here the transcription factor, for a reaction is called Michaelis Menten kinetics.

In general when a complex forms as the formation of  $[TP]$  in Eq. 2.2 then the ratio between the breaking and formation rates ( $K = \frac{k_b}{k_f}$ ) is called the dissociation constant.

**Saturated degradation** The degradation stated in Eq. 2.1 is very simple and is proportional to the concentration of  $A$ . If there is a limit on the degradation, e.g., due to complex formation, then saturated degradation occurs. This is an example of Michaelis Menten kinetics and hence the mathematical formulation for saturated degradation of  $A$  is

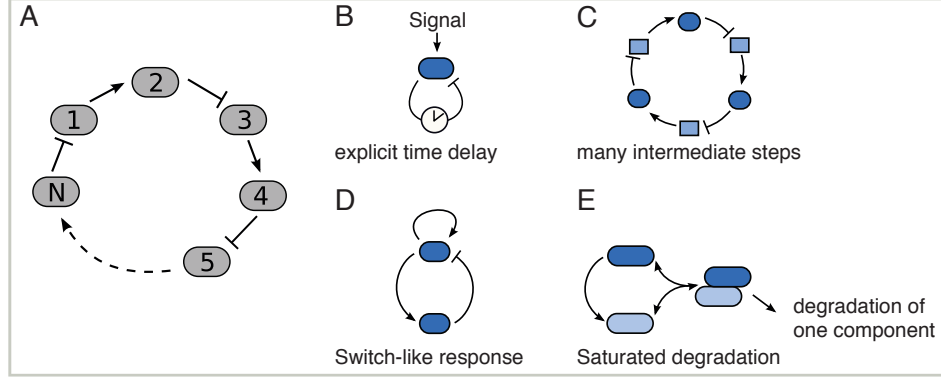
$$\frac{dA}{dt} = -v \frac{A}{A + K} . \quad (2.6)$$

Even for very high concentrations of  $A \gg K$  the maximum rate of degradation saturates to the constant  $v$ . Therefore it is called saturated degradation.

## 2.2 Feedback loops

A feedback loop is a closed and cyclic GRN with nodes either inhibiting or activating their next node. An example of a feedback loop is drawn in Fig. 2.1A. Depending on the regulatory interplay the loop can either be a negative feedback loop if the mutual regulation is inhibitory or positive if the mutual regulation is activating. Regarding the feedback loop in Fig. 2.1A it will be negative if  $N$  is odd and positive if  $N$  is even.

For oscillations to occur the minimum requirement is that the feedback loop must be negative and have a time delay [29] as in Fig. 2.1B. Other ways to obtain oscillations are exemplified in Fig. 2.1(C-E). An implicit time delay follows from a loop with many intermediate steps before a signal is passed



**Figure 2.1** Different kinds of feedback loops.

(A) Feedback loop with  $N$  nodes. The links between nodes alternates in being activating (arrows) or inhibiting (barred arrows). A feedback loop with  $N$  odd is a negative feedback loop. (B-E) Time delays in feedback loops can be incorporated by an explicit time delay (B), many intermediate steps (C), an additional positive feedback, which produces a switch-like response (D), and saturated degradation (E).

on (Fig. 2.1C). If a regulator needs to reach a certain threshold before acting in the loop it causes a time delay as in Fig. 2.1D. Saturated degradation, as shown in Fig. 2.1E, will be investigated further in the chapter to follow.

### 2.2.1 Stability analysis of feedback loops

Lewis [30] has proposed a very simple negative feedback loop with a time delay for the clock of zebrafish somitogenesis. In the model the regulations of an mRNA ( $m$ ) and a protein ( $p$ ) follow the system of ODEs below

$$\begin{aligned} \frac{dp(t)}{dt} &= am(t - \tau_p) - bp(t) = f(p, m) \\ \frac{dm(t)}{dt} &= \frac{k}{1 + \frac{p(t - \tau_m)^2}{p_0^2}} - cm(t) = g(p, m) , \end{aligned} \quad (2.7)$$

where  $\tau_p$  and  $\tau_m$  are the time delays caused by translation of mRNA ( $m$ ) into protein ( $p$ ) and transcription of the mRNA, respectively. The analytical solution to a linear two-dimensional system of ODEs is simple in the case of no time delays. Whereas if it is a system of delayed differential equations, a numerical solver is needed.

As an example of an analytical solution to a two-dimensional system of ODEs the non-delayed version of Eq. 2.7 is considered<sup>1</sup>. They are not linear and hence a linearization of the system is performed in order to investigate

<sup>1</sup>This section is based on Strogatz [31].

the behavior of  $p(t)$  and  $m(t)$  close to equilibrium where  $f(p, m) = g(p, m) = 0$ . The values of  $p$  and  $m$  satisfying this equality are called the fixed points and are denoted  $p^*$  and  $m^*$ . Very close to the fixed points of the system, a Taylor expansion around  $x = p - p^*$  and  $y = m - m^*$  can be done by

$$\begin{aligned} \frac{dx}{dt} &= f(p, m) = f(p^* + x, m^* + y) \\ &= \left( f + x \frac{\partial f}{\partial p} + y \frac{\partial f}{\partial m} \right) \bigg|_{(p^*, m^*)} + \mathcal{O}(x^2, y^2, xy) \\ &= \left( x \frac{\partial f}{\partial p} + y \frac{\partial f}{\partial m} \right) \bigg|_{(p^*, m^*)} + \mathcal{O}(x^2, y^2, xy) \end{aligned} \quad (2.8)$$

and similarly

$$\frac{dy}{dt} = \left( x \frac{\partial g}{\partial p} + y \frac{\partial g}{\partial m} \right) \bigg|_{(p^*, m^*)} + \mathcal{O}(x^2, y^2, xy) . \quad (2.9)$$

Omitting quadratic terms in Eqs. 2.8 and 2.9 the system of ODEs in Eq. 2.7 close to the fixed points can be written in matrix notation as

$$\begin{pmatrix} \frac{dp}{dt} \\ \frac{dm}{dt} \end{pmatrix} = \begin{pmatrix} \frac{dx}{dt} \\ \frac{dy}{dt} \end{pmatrix} = \begin{pmatrix} \frac{\partial f}{\partial p} & \frac{\partial f}{\partial m} \\ \frac{\partial g}{\partial p} & \frac{\partial g}{\partial m} \end{pmatrix} \bigg|_{(p^*, m^*)} \begin{pmatrix} x \\ y \end{pmatrix} = \mathbf{J}(p^*, m^*) \begin{pmatrix} x \\ y \end{pmatrix} . \quad (2.10)$$

Here  $\mathbf{J}$  is the Jacobian matrix. The system in Eq. 2.10 is straight forward to solve by finding the eigenvalues ( $\lambda$ ) and eigenvectors ( $\mathbf{v}$ ) of  $\mathbf{J}$ , since then

$$\mathbf{x}(t) = \begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = c_1 \exp(\lambda_1 t) \mathbf{v}_1 + c_2 \exp(\lambda_2 t) \mathbf{v}_2 \quad (2.11)$$

$$= \sum_{k=1}^2 c_k \exp(\alpha_k) (\cos(\beta_k t) + i \sin(\beta_k t)) \mathbf{v}_k . \quad (2.12)$$

The eigenvalues are written in complex form as  $\lambda = \alpha + i\beta$ .

Different classes of fixed points exist, e.g., they can be attractive or repellant. The class of a fixed point is dependent on the nature of the eigenvalues and -vectors. From Eq. 2.12 it is seen that the solution is oscillatory if  $\beta \neq 0$  and the sign of  $\alpha$  determines whether the solution is repelled away from the fixed point ( $\alpha > 0$ ) towards a limit cycle or attracted towards the fixed point ( $\alpha < 0$ ). Thus, to obtain sustained oscillations a complex conjugated pairs of eigenvalues with a positive real part is needed. This can, for example, happen through a Hopf bifurcation where a pair of complex conjugated eigenvalues crosses from the negative into the positive half-plane as a parameter is changed.

Returning to the example of Lewis [30] the Jacobian evaluated at the fixed points is

$$\mathbf{J}(p^*, m^*) = \begin{pmatrix} -b & a \\ \frac{-2p^*p_0^2k}{(p_0^2+p^{*2})^2} & -c \end{pmatrix} \quad (2.13)$$

The trace and determinant of this Jacobian are

$$\tau = -b - c \quad \text{and} \quad \Delta = bc + \frac{2p^*p_0^2ka}{(p_0^2+p^{*2})^2} , \quad (2.14)$$

respectively. From the trace and determinant the eigenvalues of  $\mathbf{J}$  are given by

$$\lambda_{\pm} = \frac{\tau \pm \sqrt{\tau^2 - 4\Delta}}{2} . \quad (2.15)$$

The physical quantities, i.e.,  $p$ ,  $b$ ,  $c$ , are non-negative, which gives  $\tau < 0$  and  $\Delta > 0$ . Therefore sustained oscillations cannot be obtained for this system, because the real part of  $\lambda_{\pm}$  will always be negative. This is an example of a famous theorem by Dulac, which states that sustained oscillations cannot be obtained in a two dimensional monotonic system. However, if time delays are introduced, it is possible. The classification of fixed points cannot be determined in the same manner when considering higher dimensions. However, the stability of the fixed points is still determined based on the eigenvalues of the Jacobian evaluated at the fixed points.

## CHAPTER

### 3

# Modeling embryo segmentation

After establishing the basic concepts of modeling a GRN and giving a description for classifying the dynamics of a GRN, in the previous chapter, I will now turn to actually model the GRN behind the Wnt signaling pathway. The first model, to be described in this chapter, is based on the Axin2 protein and in the section to follow a model based on the Wnt inhibitor Dickkopf1 (Dkk1) will be described. The models are described and discussed in details in the AXIN2 paper and the DKK1 paper, respectively, and therefore it is only the main and most interesting results of these two models that are presented in this chapter.

This chapter is based on:

P. B. Jensen, L. Pedersen, S. Krishna, and M. H. Jensen. *A Wnt oscillator model for somitogenesis*. Biophysical journal, 98(6):943–950, March 2010

L. Pedersen, M. H. Jensen, and S. Krishna. *Dickkopf1 - a new player in modelling the wnt pathway*. PLoS ONE, 6(10):e25550, 2011

### 3.1 Existing models of segmentation

In previous years a large range of models have been proposed to describe somitogenesis. Most of them are based on molecular networks with [30; 32–35] or without [36–38] an explicit time delay resulting in oscillations of the target genes. The goal of these models are to set up a clock, which in practice could set the time for the segmentation process. The space is then, in some models, set by including a morphogen gradient [39–41]. To mimic the growing embryo some models couple the oscillating molecular networks [18; 42]. In this way a 2D model of somitogenesis is constructed. Models not containing an explicit molecular network are usually based on coupling of oscillators arising using delayed coupling theory [43; 44], partial differential equations [45; 46] or coupling of phase oscillators [47; 48].

Both the Axin2 and Dkk1 models are based on a GRN without delays. They are structured in a simple manner in order to capture the basic dynamics of mRNAs and proteins during segmentation and propose for mechanisms of segmentation that are experimentally testable. The models are written based on known interactions of the Wnt/ $\beta$ -catenin signalling pathway. Furthermore it is assumed that the system size is large enough to exclude fluctuations. Therefore deterministic ODEs are written as to describe the Wnt signaling.

### 3.2 A model for somitogenesis based on Axin2

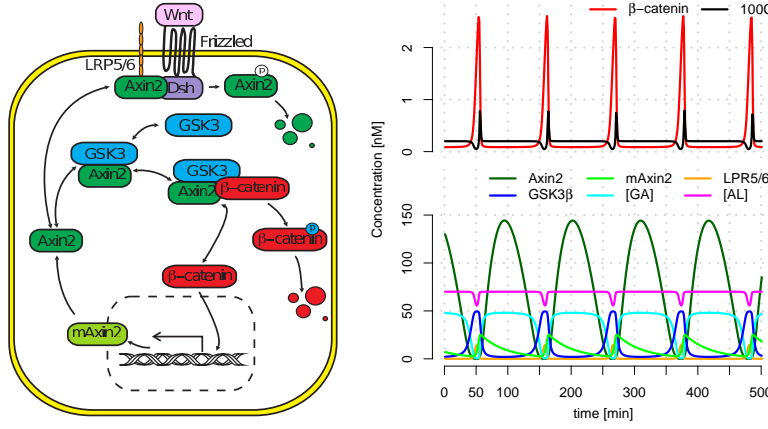
The first gene to be discovered with an oscillatory expression in the PSM of a mouse was Axin2 [13]. Axin2 takes part in the destruction complex phosphorylating  $\beta$ -catenin during Wnt signaling. Above it was stated that the GRN needs to be a negative feedback loop in order to oscillate. An oscillatory behavior is the goal for the model in order to mimic the oscillatory behavior of Axin2 in the PSM. The negative feedback loop of Wnt regulation comes in place because  $\beta$ -catenin is a transcription factor of Axin2 [49; 50]. When  $\beta$ -catenin gets phosphorylated then less Axin2 gets transcribed and less Axin2 will become available for the phosphorylation of  $\beta$ -catenin. Consequently the concentration of  $\beta$ -catenin increases, which then again leads to an increased production of Axin2 and the negative feedback loop is completed.

The model is shown in Fig. 3.1 and is described by the following equations

$$\frac{dC}{dt} = c_{fC}B[GA] - c_{bC}C - \alpha C \quad (3.1)$$

$$\frac{d[GA]}{dt} = c_{f[GA]}GA - c_{b[GA]}[GA] - c_{fC}B[GA] + c_{bC}C + \alpha C \quad (3.2)$$

$$\frac{dB}{dt} = S_B - c_{fC}B[GA] + c_{bC}C \quad (3.3)$$



**Figure 3.1** The Axin2 model in an oscillatory state.

Left: The model of Wnt signaling based on the regulation of Axin2. Right: The oscillations of the variables in the Axin2 model simulated with the default parameters from Table 3.A on p. 26. The sequential steps in the loop are as follows: a spike in  $\beta$ -catenin is followed by Axin2 building up. This leads to the formation of the destruction complex and consequently the degradation of  $\beta$ -catenin. Less Axin2 is then transcribed and more Axin2 is phosphorylated. As the concentration of Axin2 decreases  $\beta$ -catenin builds up.

$$\frac{dG}{dt} = -c_{f[GA]}GA + c_{b[GA]}[GA] \quad (3.4)$$

$$\begin{aligned} \frac{dA}{dt} = & -c_{f[GA]}GA + c_{b[GA]}[GA] + c_{tLA}A_m \\ & -c_{f[AL]}AL + c_{b[AL]}[AL] \end{aligned} \quad (3.5)$$

$$\frac{dA_m}{dt} = c_{tsA}B^2 - \frac{A_m}{\tau_{A_m}} \quad (3.6)$$

$$\frac{d[AL]}{dt} = c_{f[AL]}AL - c_{b[AL]}[AL] - \nu[AL] \quad (3.7)$$

$$\frac{dL}{dt} = -c_{f[AL]}AL + c_{b[AL]}[AL] + \nu[AL] , \quad (3.8)$$

where  $C$ ,  $[GA]$ ,  $B$ ,  $G$ ,  $A$ ,  $A_m$ ,  $[AL]$  are, respectively, concentrations of the destruction complex, GSK3 $\beta$ -Axin2 complex,  $\beta$ -catenin, GSK3 $\beta$ , Axin2, Axin2 mRNA and Axin2-LRP5/6 complex. There is a constant source of  $\beta$ -catenin ( $S_B$ ) and the phosphorylation of  $\beta$ -catenin happens with a constant rate ( $\alpha$ ). The proteins  $B$ ,  $G$ ,  $A$  and  $L$  are assumed to be stable [37] and therefore no half-lives are included for these.

Simulations of the Axin2 model in Eqs. 3.1-3.8 returns sustained oscillations with a period of 108 min (Fig. 3.1). The default parameters used to obtain these oscillations are listed in Table 3.A in the column denoted Axin2.



**Table 3.A** Default parameters for the Axin2 and Dkk1 models.

Some parameters are adapted from Lee et al. [37] others are chosen in order to obtain sustained oscillations of the models. Experimentally the dissociation constant for the LRP5/6-Dkk1 complex in the Dkk1 model has been measured to be  $K_{[LD]} = 0.4 - 0.5$  nM [51; 52].

Parameter	Process	Axin2	Dkk1	unit
$K_C$	Dissociation constant $C$	70	8	nM
$c_{bC}$	Breaking of $C$	7	7	$\text{min}^{-1}$
$\alpha$	Degradation of $\beta$ -catenin	200	2.2	$\text{min}^{-1}$
$K_{[GA]}$	Dissociation constant $[GA]$	6	1.5	nM
$c_{b[GA]}$	Breaking of $[GA]$	1.2	4	$\text{min}^{-1}$
$S_B$	Constant source of $\beta$ -catenin	0.4	1	nM/min
$\nu$	Degradation of Axin2/Axin	0.1	3.8	$\text{min}^{-1}$
$K_{[AL]}$	Dissociation of $[AL]$	0.008	-	nM
$c_{b[AL]}$	Breaking of $[AL]$	0.08	-	$\text{min}^{-1}$
$c_{tsA}$	Transcription of <i>axin2</i>	0.7	-	$(\text{nMmin})^{-1}$
$c_{tlA}$	Translation of Axin2	0.7	-	$\text{min}^{-1}$
$\tau_{Am}$	Half-life of Axin2 mRNA	40	-	min
$K_{[LGA]}$	Dissociation constant $[LGA]$	-	1	nM
$c_{b[LGA]}$	Breaking of $[LGA]$	-	10	$\text{min}^{-1}$
$K_{[LD]}$	Dissociation constant $[LD]$	-	0.5	nM
$c_{b[LD]}$	Breaking of $[LD]$	-	0.02	$\text{min}^{-1}$
$S_A$	Constant source of Axin	-	0.02	nM/min
$c_{tsl}$	Transcription of <i>dkk1</i>	-	0.02	$\text{min}^{-1}$
$c_{tsc}$	Translation of Dkk1	-	0.025	$(\text{Nm}^2 \text{min})^{-1}$
$\tau_{Dm}$	Half-life of dkk1 mRNA	-	8	min
$\tau_D$	Half-life of Dkk1	-	16	min
$GSK3\beta_{tot}$	Total G level	50	45	Nm
$L_{tot}$	Total L level	70	15	Nm

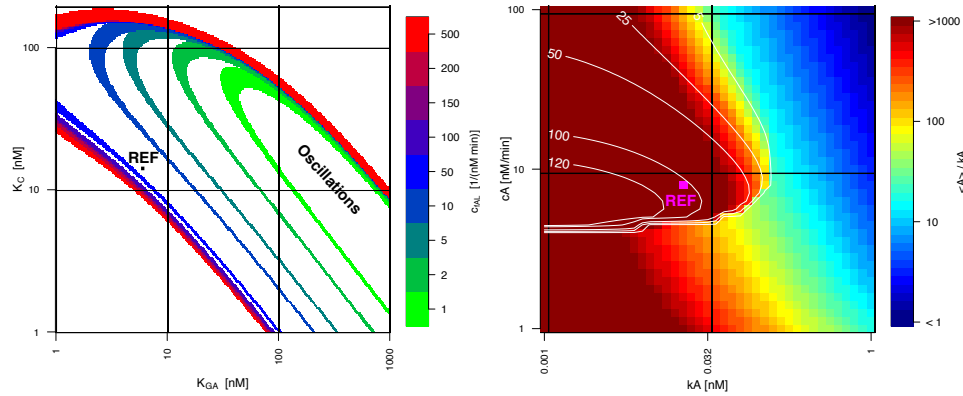
### 3.2.1 The importance of Saturated degradation

Besides the main negative feedback loop consisting of the six ODEs in Eqs. 3.1-3.6, there are two additional ODEs (Eqs. 3.7 and 3.8) describing the phosphorylation of Axin2 at the cell membrane. Upon Wnt signaling Axin2 is recruited to the LRP5/6 receptor, where it gets phosphorylated and degraded. This degradation of Axin2 is included in the model with a constant rate of degradation ( $\nu$ ).

By assuming that the timescales for breaking and formation of the  $[AL]$  complex are much faster than other processes in the model, i.e., there exist a quasi-equilibrium for the  $[AL]$  complex, then  $\frac{d[AL]}{dt} = 0$ . Looking at the default parameters in Table 3.A it is seen that the dissociation constant for the  $[AL]$  complex is orders of magnitudes smaller than the dissociation constants for the  $[GA]$  and destruction complexes. Once Axin2 is bound to the LRP5/6 complex then it rarely dissociates and consequently the  $[AL]$  complex will be in a state of quasi-equilibrium. Therefore it is reasonable to assume a state of quasi-equilibrium for the  $[AL]$  complex, where . The dynamics of Axin2 in Eq. 3.5 then becomes

$$\frac{dA}{dt} = -c_{f[GA]}GA + c_{b[GA]}[GA] + c_{tLA}A_m - c_A \frac{A}{k_A + A} . \quad (3.9)$$

The maximum rate at which Axin2 gets degraded is given by  $c_A = \nu L_{\text{tot}}$ ,



**Figure 3.2** Saturated degradation is necessary for sustained oscillations.

Left: The oscillating regime for varying values of  $c_{f[AL]}$ . The  $c_{b[AL]}$  value is also varied so as to keep the ratio  $c_{f[AL]}/c_{b[AL]}$  constant, i.e., to keep the binding strength of the  $[AL]$  constant. Oscillations occur within the tongue-shape bounded by the colored lines. Outside these boundaries all complex eigenvalues have negative real parts and therefore sustained oscillations cannot be obtained. Right: Colors show the ratio of average Axin2 concentration,  $\langle A \rangle$ , to  $k_A$  on the  $k_A - c_A$  parameter plane. White contour lines indicate the amplitude of oscillations (in nM) observed after 4 hours.

where  $L_{\text{tot}}$  is the total concentration of the LRP5/6 receptor.

The effect of saturated degradation in obtaining sustained oscillations of the Axin2 model is explored. The rates of  $[AL]$  formation and breaking are increased while keeping the dissociation constant unchanged in order to approach a quasi-equilibrium state of  $[AL]$ . As the rates increase the oscillatory regime of the system increases as seen from Fig. 3.2(left). The oscillatory regime lies within the growing area surrounded by the tongue-shape. At quasi-equilibrium the rates are very fast and following Fig. 3.2(left) the oscillatory regime is large for the reference value of  $c_{f[AL]} = 10 \text{ (nM min)}^{-1}$  resulting in a system stable in producing sustained oscillations.

In Eq. 3.9 the parameter  $k_A = \frac{c_{b[AL]} + \nu}{c_{f[AL]}}$  is an effective dissociation constant of  $[AL]$ . When the ratio  $A/k_A$  is large then saturation is approached and the degradation of Axin2 has reached its maximum as explained in Section 2.1. In Fig. 3.9(right) the ratio  $\langle A \rangle/k_A$  is plotted in the  $k_A - c_A$  plane. The white lines indicate the amplitude (in nM) of the oscillations at time  $t = 4$  hours. Only when  $\langle A \rangle/k_A \gg 1$ , where Axin2 is degraded by saturation, oscillations are observed. The reference state used in the AXIN2 paper lies within the oscillatory regime in Fig. 3.2 where it is marked REF.

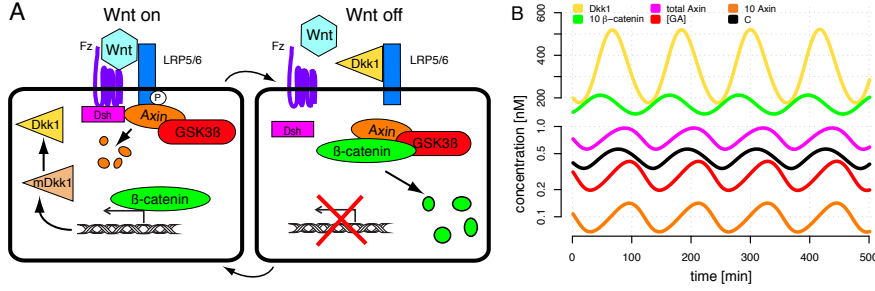
Since the oscillatory regime increases as the dynamics of  $[AL]$  approaches quasi-equilibrium and oscillations only occur when  $\langle A \rangle/k_A \gg 1$  it is stated that saturated degradation of Axin2 at the LRP5/6 receptor is necessary for oscillations, but it cannot be stated that it is sufficient to obtain sustained oscillations.

It is only possible to make this statement because the dynamics of Axin2 phosphorylation at the cell membrane is included in the Axin2 model. The model resembles a similar model proposed by Goldbeter and Pourquié [36]. There Michaelis Menten kinetics are used to describe most processes except for the simple complex formation of the  $[GA]$  complex. By including the degradation of Axin2 at the cell membrane we can give an argument as to why the degradation of Axin2 follows Michaelis Menten kinetics.

### 3.3 Dkk1 an inhibitor of Wnt signaling

The lack of phenotype for Axin2 knock-out mutants published by Yu et al. [53] led to my proposal of a Wnt model based on the Wnt inhibitor Dkk1. This protein is chosen based on its oscillatory behavior during embryo segmentation and that lowering the expression of Dkk1 causes a smaller and more irregular vertebrae in mouse embryos [54; 55], which is similar to a mouse phenotype where Wnt3a is overexpressed [56].

The regulation of Dkk1 is similar to the one described above regarding Axin2 regulations, in the sense that  $\beta$ -catenin is also a transcription factor for Dkk1 [57; 58]. Axin2 is substituted with Axin in the Dkk1 model because they are functionally equivalent [59]. A schematic illustration of the Dkk1



**Figure 3.3** The Dkk1 model in an oscillatory state.

(A) The Dkk1 model. Upon Wnt signalling Axin gets degraded and  $\beta$ -catenin promotes the transcription of Dkk1. When the Wnt signal is off then  $\beta$ -catenin gets degraded at the destruction complex, causing a decrease in Dkk1 transcription. (B) The oscillations of the variables using the default parameters in Table 3.A.

model is shown in Fig. 3.3A. When Wnt signaling is on Wnt is bound to the LRP5/6 receptor and causes phosphorylation of Axin at the cell membrane. Thus,  $\beta$ -catenin is stabilized and promotes Dkk1 transcription. Dkk1 inhibits the binding of Wnt to the LRP5/6 by binding to this receptor. When Wnt signaling is off then  $\beta$ -catenin gets phosphorylated and degraded at the destruction complex, which leads to less transcription of Dkk1. Hence, a negative feedback loop is established with Dkk1 at the core of the loop.

The structure of the Dkk1 model is similar to the Axin2 model and the equations for the Dkk1 model are

$$\frac{dB}{dt} = S_B + c_{bC}C - c_{fC}[GA]B \quad (3.10)$$

$$\frac{dC}{dt} = -\alpha C + c_{fC}[GA]B - c_{bC}C \quad (3.11)$$

$$\frac{d[GA]}{dt} = \alpha C + c_{bC}C - c_{fC}[GA]B + c_{b[LGA]}[LGA] \quad (3.12)$$

$$-c_{f[LGA]}L[GA] + c_{f[GA]}GA - c_{b[GA]}[GA] \quad (3.13)$$

$$\frac{dG}{dt} = \nu[LGA] + c_{b[GA]}[GA] - c_{f[GA]}GA \quad (3.14)$$

$$\frac{dL}{dt} = \nu[LGA] + c_{b[LGA]}[LGA] - c_{f[LGA]}L[GA] \quad (3.15)$$

$$+c_{b[LD]}[LD] - c_{f[LD]}LD \quad (3.16)$$

$$\frac{d[LGA]}{dt} = -\nu[LGA] + c_{f[LGA]}L[GA] - c_{b[LGA]}[LGA] \quad (3.17)$$

$$\frac{dD_m}{dt} = c_{tsc}B^3 - \frac{D_m}{\tau_{D_m}} \quad (3.18)$$

$$\frac{dD}{dt} = c_{tsl}D_m - \frac{D}{\tau_D} + c_{b[LD]}[LD] - c_{f[LD]}LD \quad (3.19)$$

$$\frac{d[LD]}{dt} = c_{f[LD]}LD - c_{b[LD]}[LD] \quad (3.20)$$

$$\frac{dA}{dt} = S_A + c_{b[GA]}[GA] - c_{f[GA]}GA \quad (3.21)$$

The new variables in the Dkk1 model are Dkk1 ( $D$ ), Dkk1 mRNA ( $D_m$ ), the LRP5/6-Dkk1 complex ( $[LD]$ ), the LRP5/6-GSK3 $\beta$ -Axin complex ( $[LGA]$ ), and Axin2 is substituted with Axin ( $A$ ). Using the default parameters listed in Table 3.A oscillations of the variables are obtained (Fig. 3.3B) together with a low concentration of total Axin, which has been measured to be low in *xenopus* embryos by Lee et al. [37].

Since the proposal of our Axin2 model new knowledge came to our mind about the dual role of GSK3 $\beta$  in Wnt signaling. Firstly, GSK3 $\beta$  takes part in the destruction complex leading to  $\beta$ -catenin degradation. Secondly, GSK3 $\beta$  also binds to Axin and when bound at LRP5/6 GSK3 $\beta$  aids in phosphorylating Axin at the cell membrane. Thus, GSK3 $\beta$  plays two conflicting roles in Wnt signaling [60] and it is unknown whether it is the same or two distinct  $[GA]$  complexes that take part in the two different phosphorylation processes. In the Dkk1 model they are equal.

The period of the oscillations is 116 min for the Dkk1 model at the reference state. It is mostly affected by changes in the phosphorylation and degradation of  $\beta$ -catenin ( $\alpha$ ) the source of Axin ( $S_A$ ) and the dynamics of the destruction complex ( $K_C$ ). For the Axin2 model we also observed that the period is strongly dependent on the dynamics of the destruction complex.

The Wnt signal is almost constantly turned off since the concentration of free and available LRP5/6 is very low, see Table 3.B. When Wnt is turned off then  $\beta$ -catenin is stabilized and therefore the concentration of free  $\beta$ -catenin will be high as seen from Table 3.B. The limiting factor for the formation of the destruction complex is Axin, which is only present at

**Table 3.B** Average concentrations for Dkk1 variables in the reference state.

Variable	average conc [nM]	Variable	average conc [nM]
$D_m$	1042.0	$C$	0.454
$D$	333.0	$[GA]$	0.293
$G$	44.2	$L$	0.0254
$B$	16.8	$A$	0.01011
$[LD]$	14.97	$[LGA]$	$5.26 \cdot 10^{-3}$

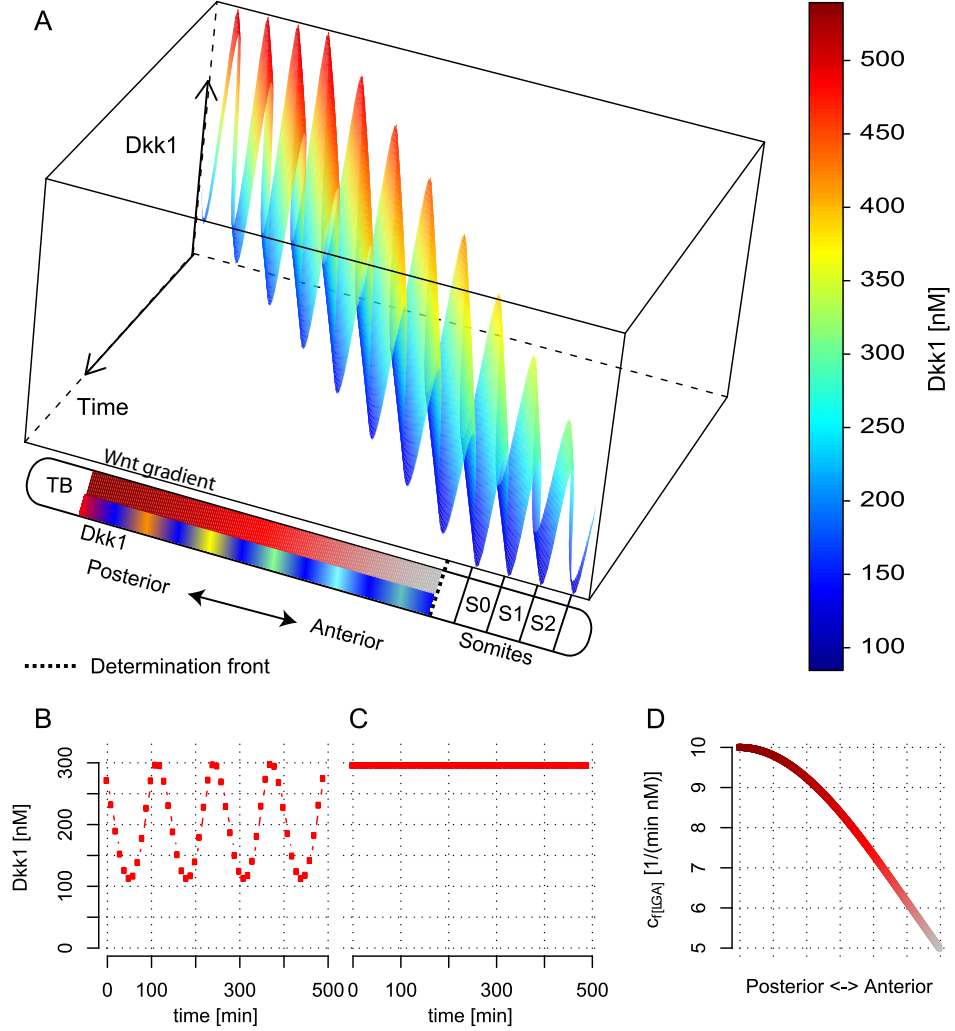
low concentrations. During Wnt signaling it is sequestered to the cell membrane where it is degraded. If it is not bound to LRP5/6 at the membrane Axin2 is bound to GSK3 $\beta$ , since little Axin2 is left unbound as seen from Table 3.B. As the concentration of Dkk1 builds up the level of  $\beta$ -catenin decreases and the level of Dkk1 will consequently decrease turning on a quick transition from phosphorylation of  $\beta$ -catenin in the destruction complex to Axin phosphorylation at the cell membrane. This transition occurs quickly because most of the Axin is constantly bound to GSK3 $\beta$ . However, there are no spiky oscillations as observed for  $\beta$ -catenin in the Axin2 model.

### 3.3.1 Incorporating a Wnt gradient in the PSM

The clock and wavefront model, described in Chapter 1, proposes a morphogen gradient to set the space once the time is set by a clock. In the Axin2 and Dkk1 models a morphogen gradient is introduced by letting the Wnt level decrease with time as to mimic the decrease of Wnt throughout the PSM. Wnt is only produced in the tail bud of the embryo and by diffusion and/or Wnt decay a gradient of Wnt is established in the PSM. In the AXIN2 paper the Wnt level is controlled by the parameter  $\nu$ , which is the rate of Axin2 phosphorylation. However, the Wnt level is crucial for the formation of the  $[LGA]$  complex, since if Wnt is bound to Frizzled then the  $[GA]$  complex cannot bind to LRP5/6 and form the  $[LGA]$  complex. The  $[LGA]$  complex is not included in the Axin2 model where Axin2 alone binds to LRP5/6.

The inclusion of a Wnt gradient is carried out through a time dependence for the rate of  $[LGA]$  complex formation, i.e.  $c_{f[LGA]}$  must be time dependent. The reason for choosing  $c_{f[LGA]}$  to be time dependent instead of  $\nu$ , as in the Axin2 model, is due to the fact that the Wnt signal is mediated by Wnt binding to LRP5/6. Simulating the Dkk1 model with a Gaussian profile of the Wnt level with a 2-fold decrease (Fig. 3.4D) causes the oscillations to decrease in amplitude, and the periods get slightly shorter, see Figure S2 of the DKK1 paper. Decreasing the Wnt level in the Axin2 model by decreasing  $\nu$  3-fold resulted in slightly longer periods.

Experimentally it has been shown that the wavelength of the oscillations decreases from the tail bud to the determination front [48]. Assuming that the wavelength is proportional to the period, then the period also decreases from the tail bud to the determination front. This supports the idea of including the Wnt gradient by a time dependency in  $c_{f[LGA]}$  instead of a time dependence in  $\nu$ . As a side note, the increase/decrease in the periods for the Axin2/Dkk1 model is not dependent on the Wnt gradient being linear or Gaussian. Experimentally, it has been measured that FGF (regulated by Wnt) exhibits a gradient in the PSM with a fold change of two to five [61].



**Figure 3.4** Synchronization of cells at the tail bud.

(A) Time series of the Dkk1 concentration with a decreasing value of  $c_{f[LGA]}$ . Space is introduced by letting a cell bud off from the tail bud every 10th min. The initial state of this cell is equal to that of its most anterior neighbour. The cells move relatively in the PSM. At the determination front (dotted line) the oscillations arrest. (B/C) The level of Dkk1 at the determination front with (B) and without (C) synchronization of neighboring cells at the tail bud. (D) The Gaussian profile used for  $c_{f[LGA]}$  in the anteroposterior direction. TB: tail bud. Si: Somite i, where S0 is the newly formed somite.

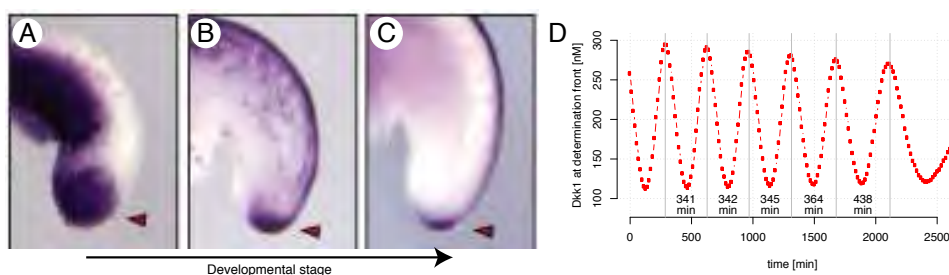
### 3.3.2 Synchronization of neighboring cells at the tail bud

The cells in the PSM are synchronized such that neighboring cells are oscillating in phase. This synchronization is important for proper somite for-

mation as explained in Chapter 1. In the DKK1 paper the cells within the PSM are coupled in a simple manner by letting a newly cell budded off from the tail bud having the initial state equal to the state of its most anterior neighboring cell. It can be compared to a wave during a football game on a stadium. Using this simple synchronization scheme we are able to explore the importance of synchronization by showing that the level of Dkk1 oscillates at the determination front (Fig. 3.4A,B), which is important for setting the boundaries of somites. If all cells are initialized with equal states then the level of Dkk1 at the determination front is constant (Fig. 3.4C). This constant value of Dkk1 depends on the initial concentration of Dkk1 at the tail bud. In Fig. 3.4C the level of Dkk1 at the determination front is rather high around 300 nM. If other initial values of Dkk1 are used then the constant Dkk1 level at the determination front can be changed.

### 3.3.3 Longer periods towards the end of somitogenesis

A decreasing level of Wnt in the tail bud has been measured experimentally [62] (Fig. 3.5(A-C)) along with an increase in the period [63] as the embryo approaches its full length. It is not known how the segmentation process actually terminates. The fact that the period increases along with a decrease in the amplitude of the oscillations, suggest that as the Wnt level decreases the oscillations arrest until the amplitude and/or period reaches a certain value, where the segmentation process then terminates. Using the synchronization scheme described above it is possible to simulate a decreasing level of Wnt in the tail bud. Performing this simulation returns oscillations of Dkk1 at the determination front with a decreasing amplitude and increasing period (from 341 min to 438 min, see Fig. 3.5D). On the other hand, if the Wnt level is increased in the tail bud then the periods shorten. These findings reject the hypothesis of Aulehla and Herrmann [64] stating that the



**Figure 3.5** The level of Wnt decreases in the tail bud.

(A-C) Chick embryos hybridised for Wnt3a expression show a decreasing Wnt3a expression in the tail bud as they develop. Adapted from Gibb et al. [62]. (D) As the Wnt level decreases in the tail bud synchronized PSM cells display a lengthening of the segmentation period and a decrease in Dkk1 amplitude.



increase in period is caused by an increase in the Wnt level.

In the synchronization scheme the differing length of the PSM observed experimentally [63] is not taken into account. Towards the end of somitogenesis the PSM shortens. Because the length of the PSM is reflected by the time the simulation is run the shortening of the PSM could simply be incorporated by letting the simulation run for a shorter time period. It is possible that incorporating this shortening of the PSM at the end of the segmentation process would show a more abrupt arrest of the oscillations. This is just speculations that needs supporting evidence in the form of numerical simulations.

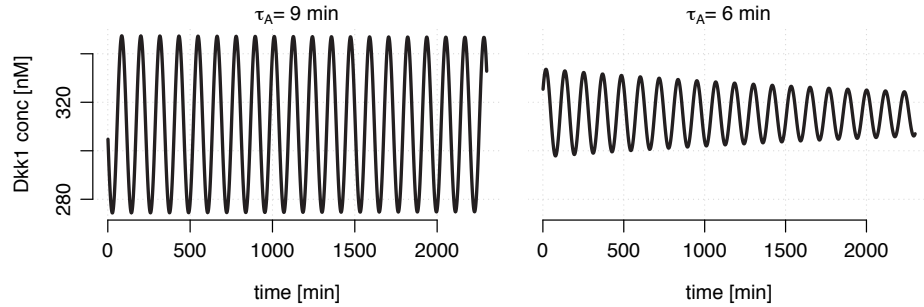
### 3.3.4 Sustained or decaying oscillations - what is needed?

The half-life of Axin is 6 min according to Lee et al. [37], where the measurements are performed on *xenopus*. In the Axin2 and Dkk1 models Axin is assumed to be stable within the timescales of segmentation, i.e. the half-life of Axin is effectively infinite. There are no measurements for the half-life of Axin in mouse or chicken embryos. Maybe it is longer. However, the measured 6 min half-life of Axin is within the timescales of segmentation, and the effect of incorporating a half-life of Axin is explored within this section.

If a half-life ( $\tau_A$ ) is included for the dynamics of Axin then

$$\frac{dA}{dt} = S_A + c_{b[GA]}[GA] - c_{f[GA]}GA - \frac{A}{\tau_A} . \quad (3.22)$$

If  $\tau_A = 6$  min is used together with the default parameters listed in Table 3.A, then the oscillations of Dkk1 are decaying. On the other hand if the half-life is slightly increased to  $\tau_A = 9$  min then the oscillations become sustained, see Fig. 3.6. Thus, it seems that the Dkk1 model is on the edge of a bifurcation. The type of bifurcation would be determined by investigating



**Figure 3.6** Introduction of an Axin2 half-life in the Dkk1 model.

The half-life of Axin takes the Dkk1 model from an oscillatory state (left) to a non-oscillatory state (right) as it is decreased from  $\tau_A = 9$  min to  $\tau_A = 6$  min.

the eigenvectors and eigenvalues around the fixed points as described in Chapter 2.

As discussed for the Axin2 model an increasing half-life of Axin2 mRNA increased the period of oscillations. This is also found for the Dkk1 model, where the period of oscillations increases slightly as  $\tau_A$  is increased above  $\tau_A = 9$  min. Interestingly, the period of Dkk1 oscillations lengthen as the half-life is decreased for  $\tau_A < 9$  min. Because the segmentation period increases as the embryo elongates one could speculate whether the bifurcation caused by decreasing  $\tau_A$  actually occurs in the embryo as to terminate the somitogenesis process. The decrease in Axin half-life could be caused by increasing the activity of the ubiquitin proteasome pathway [65].

Could it be that the half-life of Axin is changed during somitogenesis as to terminate the process through a bifurcation? Or is it a minimal change of the default parameters that takes the system from a non-oscillatory to an oscillatory state for  $\tau_A = 6$  min?



## CHAPTER

## 4

# Concluding remarks Part I

The general concern within the field of describing and modeling segmentation is the lack of an overall pacemaker. In both the Axin2 model and Dkk1 model the concentration of  $\beta$ -catenin needs to oscillate in order for Axin2 or Dkk1 to oscillate. From the observations of Aulehla et al. [66]  $\beta$ -catenin needs not to oscillate in the PSM, because if  $\beta$ -catenin is constitutively expressed then the target genes of Wnt and Notch still oscillate in the PSM. Even though  $\beta$ -catenin does not oscillate it has been shown that the Notch target gene Nrarp, which stabilizes LEF-1 [67], does oscillate [14]. LEF-1 does not oscillate in the PSM [62]. Notch and Wnt target genes oscillates out of phase. Thus, when Axin2 or Dkk1 is high, Nrarp will have a low expression, resulting in LEF-1 ubiquitination and consequently less Dkk1 and Axin2, because LEF-1 promotes the transcription of Wnt target genes. Thus, the  $\beta$ -catenin variable in both the Axin2 and Dkk1 models can be considered as a coarse-grained variable combining the effects of  $\beta$ -catenin, Nrarp and LEF-1.

Furthermore, it has been shown that the Notch pathway is not necessary for proper boundary formation of the somites and it is only required for synchronization of the cells in the posterior PSM [68]. Also, it is only Nrarp mRNA that oscillate and not Nrarp protein [69]. This might explain why LEF-1 does not oscillate in the PSM, but together with the story of Notch it also adds to the mystery story of finding an overall pacemaker of somitogenesis as mentioned in Aulehla et al. [66] and Dequéant and Pourquié [15].

Throughout this Part the aim has been to find sustained oscillations. However, could it be that the oscillations might not need to be sustained,

but only oscillatory for the time it takes the cell to move relatively from the tail bud to the determination front? This question puts another perspective on modeling somitogenesis, because a wider selection of models might be possible and the parameter space might become bigger. Actually, by only requiring damped oscillations it might be possible to circumvent the issue of  $\beta$ -catenin not oscillating in the PSM.

## Part II

# MICRORNA TARGET PREDICTION





## CHAPTER

## 5

# Complexity of miRNA target prediction

In a hand-written signature it is the small, personal curls and bows that uniquely define us. It is almost impossible to imitate a signature due to these. The same is true for our genetic signature where the small curls and bows could be represented by the small but regulatory important microRNAs (miRNAs). They target protein coding mRNAs by blocking their translation. microRNAs are commonly deregulated in many diseases and consequently their targets are deregulated, which contribute to the development of diseases. Therefore reliable and accurate microRNA target prediction methods are important for a better understanding of gene networks regulated by miRNAs. A better understanding of these might lead to new and improved treatments or maybe even a cure of diseases.

In this chapter a brief review of microRNAs and their regulatory mechanisms is given, along with an overview of the research field of computational miRNA target prediction. It is important to understand the mechanisms behind miRNA regulation in order to develop precise miRNA target prediction methods, and to understand the limitations of these.

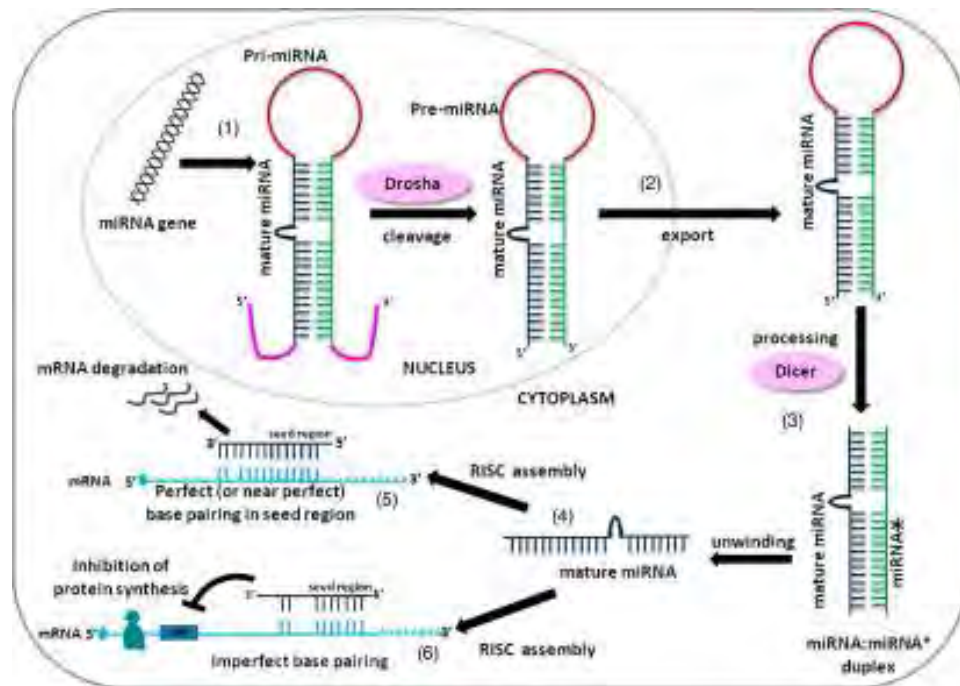


## 5.1 microRNA biogenesis and regulation

In 1993 Lee et al. [70] discovered a *C. elegans* gene (*lin-4*) which did not code for a protein and the effective gene consisted of only 22 nucleotides. Furthermore, the expressions of both *lin-14* mRNA and protein were downregulated upon *lin-4* expression, thus a translational repression was observed. The mechanism behind this inhibition was unknown, but when the 3'UTR of *lin-14* was deleted then a gain-of-function was observed [71].

The *lin-4* gene is the first miRNA to be discovered. miRNAs are a family of non-coding genes with a size of approximately 22 nucleotides. Despite their short length, and the fact that they comprise only 0.5-1% of the genes in a genome, they each target roughly 200 transcripts [72] and overall they regulate 60% of the human genome [73].

The genomic location of miRNAs can be intergenic, intronic or exonic



**Figure 5.1** From pri-miRNA to mature miRNA.

(1) The miRNA genes are transcribed by RNA polymerase II, and Drosha processing makes the pri-miRNAs into pre-miRNAs. These are exported out into the cytoplasm (2), where Dicer processing (3) incorporates the mature miRNA into the RISC complex (4). If there is nearly perfect complementarity between the 3'UTR and the miRNA then the mRNA is cleaved (5). If the mature miRNA sequence has imperfect complementarity with the 3'UTR of the target mRNA then translationally repression takes places (6). Figure adapted from Bernardo et al. [74].

[75]. Intergenic miRNAs are located alone or in clusters in genomic regions without a known transcription unit. Intronic and exonic miRNAs are located within introns or exons of known transcript units. The transcription of most miRNA genes are carried out by the RNA polymerase II as for protein-coding genes [76]. The transcribed miRNA forms an imperfect stem-loop called a pri-miRNA and cleavage of this gives a hairpin-loop called pre-miRNA (Fig. 5.1). The last step in obtaining a mature sequence of the miRNA involves a protein complex called Dicer that cleaves the pre-miRNA into a mature miRNA of about 22 nucleotides. The regulation of target gene expression includes an incorporation of the mature miRNA into the RNA-induced silencing complex (RISC). In the following it is the mature form of the miRNA that is referred to when miRNAs are mentioned.

The general knowledge is that the first 6-8 nucleotides, counted either from the first or second nucleotide in the 5' end of the miRNA, are of particular interest [77; 78]. Because these show near perfect complementarity between the miRNA and the 3'UTR of the target mRNA. They are called the seed region. When the miRNA binds to the mRNA the seed region acts like a rapid zip, that overcomes thermal diffusion, and is followed by a stabilizing thermodynamic step where more base pairs are formed between the miRNA and mRNA [77]. After forming imperfectly stem-loop structures with the target mRNA the miRNAs either degrades the mRNA by cleavage or inhibits the translation of the mRNA. The degree of complementarity in the miRNA:mRNA duplex determines the fate of the mRNA [79] (see Fig. 5.1). In plants the miRNAs usually binds with close to perfect complementarity. As a consequence their targets are cleaved [80].

## 5.2 Predicting miRNA targets

Starting in 2008 many papers have shown that miRNAs target coding regions just as well as they target 3'UTRs [81–83]. Furthermore, it has also been shown that miRNA seeds are not always present in the miRNA targets [81; 84–89]. Also, an upregulation of a targeted mRNA is seen when the miRNA binds to the 5'UTR [90]. These findings exemplifies the high complexity of miRNA regulations, which makes the computational task of predicting miRNA targets very difficult.

The list of available miRNA target prediction methods is long and counts more than 30 published methods. The basics of these are [73; 91; 92]:

1. **Base pairing:** There exist different types of base pairing between the miRNA and its target mRNA. Complementary binding at the seed region is sometimes followed by additional base pairing at the 3'end of the miRNA. Sometimes a mismatch or wobble is observed within the seed region, which is compensated by a long stretch of complementary base pairing in the 3'-end of the miRNA.

2. **Favourable seed sites:** It requires energy to unfold the mRNA and free energy is produced when the miRNA binds to the mRNA. Thus some seed sites are more energetically favourable than others [93]. The flanking region of the seed site is also important, since it has been observed that some motifs, e.g., AU-rich regions [94], in the flanking region improves the target prediction methods [95; 96].
3. **Expression analysis:** Usually there is a negative correlation between the expression of the miRNA and its target. Thus targets are sought among the mRNAs anti-correlating the most with the targeting miRNA.
4. **Conservation:** Both the seed region of the miRNA and its targets are well conserved among related species [97]. Therefore miRNA targets should be conserved across species and a phylogenetic analysis is often performed.
5. **Cooperativity:** miRNAs fine tune the expression of genes in a combinatorial manner, meaning that several miRNAs can target the same mRNA transcript or multiple mRNA transcripts [72; 98]. Furthermore, a cluster of co-expressed miRNAs can regulate functionally related genes [99].

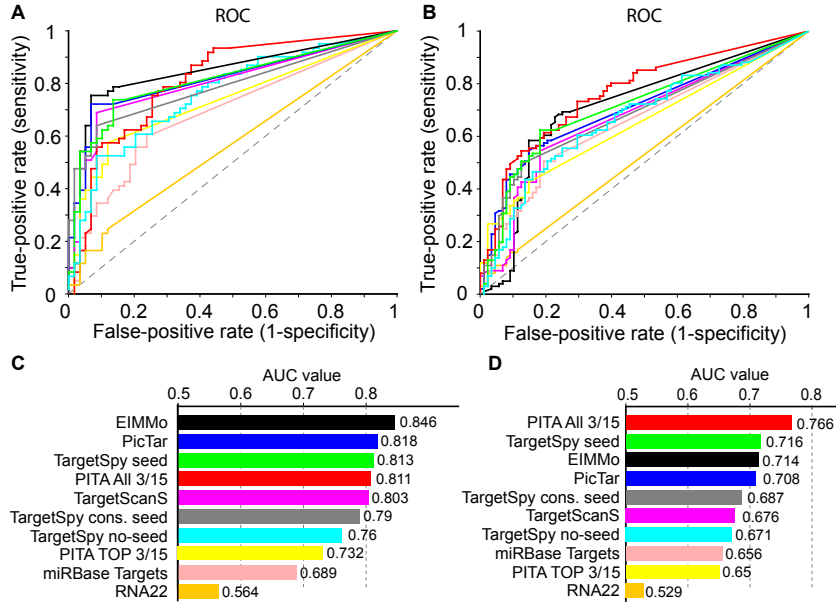
The features above are included to different extents in different methods and the mathematical scheme for predicting the targets also varies greatly between the methods. Some methods use simple enrichment tests [100], while others make use of different machine learning algorithms [101; 102], where one dataset is used to train the algorithm to classify miRNA targets, and then a second dataset is used to test the learning algorithm. Consequently, the overlaps between the target prediction methods are sparse. Kertesz et al. [93] report an overlap ranging from 19%-31%.

### 5.2.1 Accuracy of available target prediction methods

A miRNA target prediction is a binary test with a positive outcome if the mRNA is predicted as a target for the miRNA and a negative if the mRNA is not predicted as target. The aim of many miRNA target prediction methods is to get no false-positives or false-negatives, i.e. high sensitivities and specificities<sup>1</sup>. However, this is unattainable since there are a lot of unknown variables behind miRNA regulations that still needs to be mapped. If the method contains tunable parameters or is based on a machine learning algorithm, then it is common to optimize the method as to have the optimal trade-off between the sensitivity and specificity.

---

<sup>1</sup>Sensitivity and specificity are explained in App A.



**Figure 5.2** The performance for a selection of miRNA target prediction methods. The methods are tested by Sturm et al. [96] on a dataset consisting of validated miRNA:mRNA pairs from Stark et al. [103] (A,C) and Kertesz et al. [93] (B,D). References for methods: TargetSpy [96], EIMMo [104], miRanda [105], PicTar [72], PITA [93], RNA22 [106], TargetScanS [107], miRBase ([mirbase.org](http://mirbase.org)). The figure is adapted from Sturm et al. [96].

In Sturm et al. [96] a selection of methods with tunable parameters are tested on two validation datasets, see Fig. 5.2. In Fig. 5.2(A,B) the sensitivity is plotted against (1-specificity) for the different methods. This is a common plot as to validate a target prediction method and the area under the curves (AUC) measures the performance of the method (Fig. 5.2(C,D)). If all miRNAs predicted by the method is always “true” miRNA targets then AUC equals one, since there will be no false-positives. As can be seen from Fig. 5.2 it is very difficult to compare the performance of different methods, because it very much depends on the data that the methods are tested on. Methods that are trained on a specific dataset will be biased by the quality and contents of the dataset.

### 5.2.2 Expression data used for miRNA target prediction

Considering both the features incorporated into the methods of Fig. 5.2 and their AUC values indicate that incorporating multiple features increases the accuracy of the methods. This is also mentioned in the review on miRNA target prediction by Saito and Saetrom [91]. The method to be presented in this thesis uses expression data along with seed match and an optional

cooperativity feature. Thus it incorporates two-three different features. The method will be described in Chapter 6, but as an introduction a broad overview of existing methods including expression data will be given here.

Only a few of the published prediction methods incorporate gene expression data. This is mostly caused by the lack of gene expression data, where both miRNA and mRNA have been profiled within two different classes of samples. As the cost of performing a microarray experiment decreases more datasets of this kind will be available, e.g., through the large gene expression repository of the Cancer Genome Atlas (TCGA, [cancergenome.nih.gov](http://cancergenome.nih.gov)). Listed below are prediction methods incorporating gene expression data.

**TargetMiner** Bandyopadhyay and Mitra [102] propose a support vector machine (SVM) based classifier of miRNA targets. Putative miRNA targets are curated from the union of published miRNA targets. For these mRNAs they filter out those that positively correlate with miRNAs differentially expressed in a microarray experiment. Those mRNAs not filtered out need also to be differentially expressed under the same conditions as the miRNAs.

**HOCTAR** Gennarino et al. [108] consider intronic miRNAs, which are transcribed along with their targets. The expression of these intronic miRNAs must anti-correlate with the expression of their host genes.

**Stanhope** Stanhope et al. [109] apply a regression model to the expression data in order to filter out the contributions from systemic noise, other miRNAs binding to the target and the noise introduced by RISC binding to the mRNA.

**Genmir++** Potential miRNA targets are retrieved using an existing target prediction model. These potential targets are then used in a Bayesian learning algorithm to find miRNA targets, which anti-correlate in expression with the miRNA [101].

**Sylamer** van Dongen et al. [100] use expression data, where a miRNA is typically knocked down. Upon knockdown the potential miRNA targets will be upregulated. The idea behind Sylamer is that there is an enrichment of words complementary to the seed of the knocked down miRNA in 3'UTRs of the upregulated mRNAs.

Except for Stanhope all the methods search for anti-correlations between mRNA and miRNA expressions. Only Stanhope considers the fact that miRNA and targeted mRNA expressions might not anti-correlate.

## CHAPTER

# 6

### miRNA target prediction using independent component analysis

In this chapter a new and innovative method for predicting miRNA targets is proposed. It is called miICA. The innovative feature of miICA is that it is based on independent component analysis (ICA), which has been shown superior to clustering and principal component analysis (PCA) commonly used for analyzing microarray data [110–113]. In combination with ICA enrichments of putative miRNAs and biological pathways are sought. Ultimately miICA will be capable of unravel the huge genomic network that underlies the development of several diseases.

Furthermore, in a microarray experiment thousands of genes can be profiled and performing several microarray experiments for different experimental conditions provides a high-dimensional set of expression profiles. Turning these data into useful information is a research area on its own. In the ICA paper a method of implementing ICA and retrieving useful information from microarray data is described. A short overview of the ICA paper is given in this chapter.

This chapter contains results from:

L. Pedersen and P. H. Hagedorn. Ranking of Independent Components in Gene Expression Microarray Data Using Weighted Scoring. Under revision in PLoS ONE, 2011

## 6.1 Principles of independent component analysis

ICA was originally proposed for signal analysis [114]. It is a linear decomposition method used for blind source separation (BSS). In BSS it is the sources ( $\mathbf{C}$ ) behind the measured signals ( $\mathbf{E}$ ) that is wanted, but only the signals are known. For example, if the signal is a mixture of different sources, e.g., three microphones recording the sound from two speakers, then ICA can decompose the measured signal from the three microphones and retrieve the two original sources from the speakers. The retrieved sources of the speakers are called independent components (ICs). The matrix formulation of the problem is

$$\mathbf{E} = \mathbf{C}\mathbf{M} \quad \text{or} \quad E_{gs} = \sum_{t=1}^T C_{gt} M_{ts} , \quad (6.1)$$

where the mixing matrix ( $\mathbf{M}$ ) gives the linear combination of the sources resulting in the measured signals. It was found by Comon [114] that if the sources are assumed to be independent then the problem of finding both  $\mathbf{C}$  and  $\mathbf{M}$  becomes solvable. The sources are estimated by the ICs, which constitute the columns of the component matrix  $\mathbf{C}$ . The entries of  $\mathbf{C}$  ( $C_{gt}$ ) are denoted loads and the entries of  $\mathbf{M}$  ( $M_{ts}$ ) are denoted component weights. If  $\mathbf{E}$  is a  $G \times S$  matrix, i.e.  $g \in \{1, \dots, G\}$  and  $s \in \{1, \dots, S\}$ , then the maximum number of ICs that can be estimated is  $\min(G, S)$ .

From the matrix formulation of ICA it is seen that  $\mathbf{M}$  and  $\mathbf{C}$  are not uniquely defined or ordered. Hence three constraints are set on  $\mathbf{C}$  [114]:

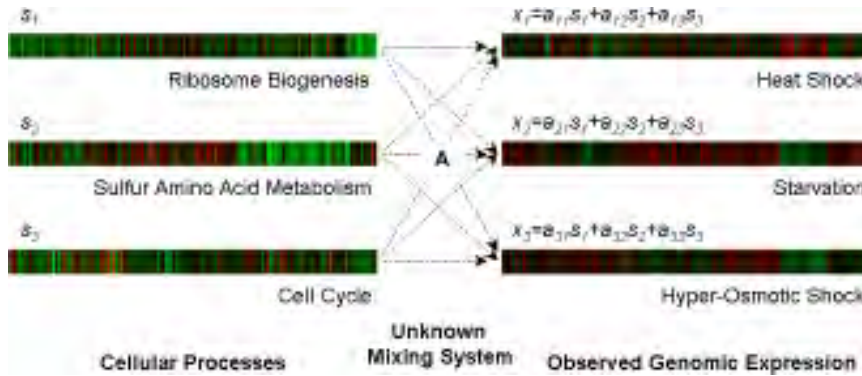
- Each IC is centered yielding a mean of zero and a standard deviation of one of the loads, i.e. for IC  $t$  then

$$\langle C_{.t} \rangle = 0 \quad \text{and} \quad \sigma(C_{.t}) = 1 . \quad (6.2)$$

- The eigenvalues for the covariance matrix of  $\mathbf{C}$  are ordered to give the order of the ICs.
- In each IC the largest modulus must be positive, i.e.

$$\text{sign}(C_{gt} \mid |C_{.t}| = \max(|C_{.t}|)) = 1 . \quad (6.3)$$

A last condition that needs to be fulfilled for ICA to be uniquely defined is that at most one of the ICs must have Gaussian distributed loads. This condition is almost always true for signals obtained by measuring a physical quantity, where the noises on the measurements are often Gaussian distributed but the sources themselves are non-Gaussian distributed.



**Figure 6.1** ICs capture biological pathways and cellular processes. Different biological processes are mixed to obtain the gene expressions for yeast (*Saccharomyces cerevisiae*) responding to diverse environmental transitions. Figure adapted from Lee and Batzoglou [111].

### 6.1.1 ICA applied to gene expression data

ICA is a method widely used to analyze microarray data. Assuming that the individual contributions on mRNA expressions are linearly separable and reasonable independent of each other, ICA can separate the various contributions to the mRNA expressions. The result of several microarray experiments can be contained in an expression matrix ( $\mathbf{E}$ ), where the columns correspond to different samples and the rows correspond to genes. The pre-processing steps performed in order to obtain  $\mathbf{E}$  will not be described in detail here, but see the ICA paper, the T1D paper, and the OVC paper for further details.

ICA has been proven to be superior to both clustering and PCA [110–113]. Both methods are widely used for the analysis of mRNA microarray data and recently, ICA was applied to miRNA profiling data as well [115]. When ICA is applied to expression data then the expression profile for each sample acts as a microphone. The speakers are the levels of genes within a transcriptional program (biological pathway, cellular component, etc). Actually, the ICs identified using ICA has been shown to capture the differential regulation of well-defined biological processes and metabolic pathways in breast cancer [113], acute myeloid leukemia [116], and Alzheimer’s disease [117]. When these are mixed together they will add up to give the expression profiles of the samples. In Fig. 6.1 this is exemplified for yeast (*Saccharomyces cerevisiae*) exposed to different environmental transitions such as heat shock and starvation.

Each gene is weighted in the ICs according to its gene weight ( $C_{gt}$ ), which in the previous was called a load as is common terminology. In the T1D paper the component and gene weights are called mixes and loads, respectively, as is tradition. In the OVC paper the terms gene and component



weights are introduced to use a more biological terminology.

There have been many proposals on how to select genes differentially expressed based on their gene weights [110; 111; 116]. In this thesis the method proposed by Liebermeister [110] is used. Liebermeister selects a gene  $g$  from IC  $t$  if

$$|C_{gt}| > 2\sigma(|C_{.t}|) , \quad (6.4)$$

where  $|C_{gt}|$  is the absolute gene weight of gene  $g$  in IC  $t$ ,  $\sigma$  is the standard deviation and  $|C_{.t}|$  is the absolute values of all the gene weights in IC  $t$ . A gene satisfying Eq. 6.4 is said to be active within IC  $t$ .

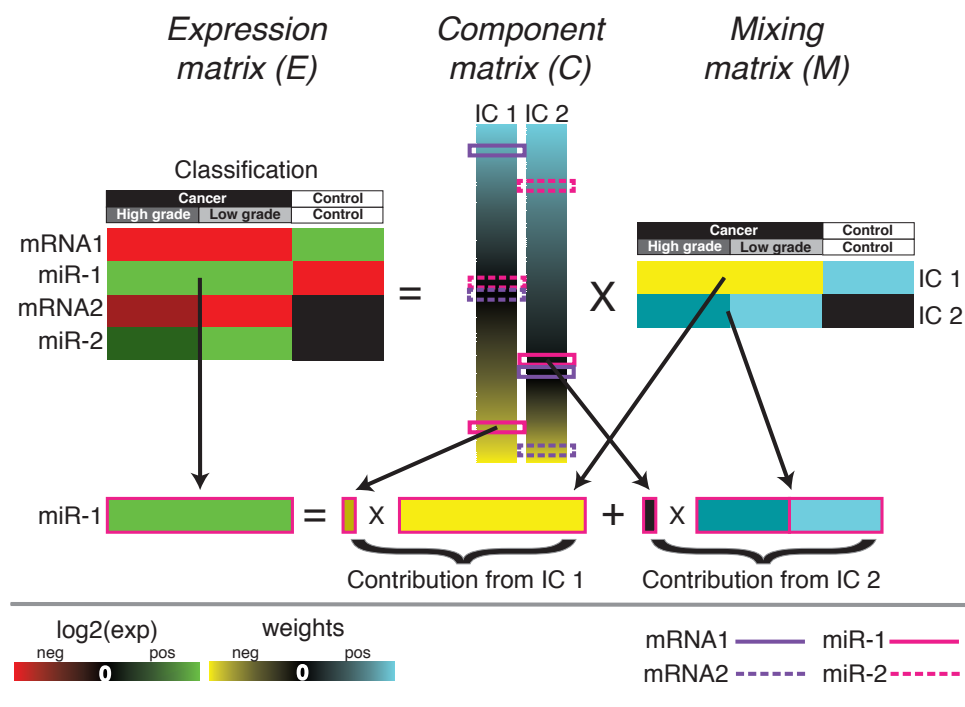
The ICs retrieved from a collection of gene expression profiles will be weighted differentially across the samples according to the values of  $M_{ts}$  in Eq. 6.1. The component weights measure the activity level of the transcriptional programmes across the samples, i.e.  $M_{ts}$  is the activity level of process  $t$  in condition  $s$ . For example, assume that an IC maps to a cancer related pathway then the weights of this IC will be very high in cancerous samples and negligible in non-cancerous samples. Thus the component weights can be used to classify samples according to a physiological, histological, etc. feature of the samples. Actually, tumor classification based on ICs seems to be a promising application of ICA [118; 119].

### 6.1.2 Ranking of independent components

One drawback of ICA is that no ranking of the ICs is given. In the ICA paper a method of ranking the ICs is described. It is based on a weighting scheme for the ICs, where the ranking is based on the ICs ability to separate samples according to their classification, the Gaussianity of the gene and component weights, the stability of the ICs and the biological concordance of the ICs. ICs map to biological processes or metabolic pathways, as described above, and the biological concordance measures the fraction of genes satisfying Eq. 6.4 that are also annotated by a biological pathway or process. Thus the biological concordance score measures the ICs capability to map to biological pathways or processes, and therefore the biological concordance feature of the ICA paper is important.

Other ranking methods have been proposed. Liebermeister [110] ranks the ICs according to the variance captured by the component weights and the negentropy of the gene weights. Similarly, Winther and Petersen [120] rank the ICs according to the energy  $\propto \sum M_{ts}^2$  that they each capture. Chiappetta et al. [121] rank the ICs according to the stability of these. The ranking method proposed in the ICA paper therefore includes ranking schemes already proposed in the literature, but it also includes some new and very important ranking features, such as the biological concordance.

In the ICA paper it is shown how a set of 492 genes correctly classifies cancer samples according to their grade with a accuracy of 96%. The genes are chosen based on their gene weights in the highest ranking ICs.



**Figure 6.2** ICA applied to mRNA and miRNA expressions.

The expression matrix is decomposed into a component matrix with ICs in the columns and a mixing matrix with weights of each IC in the rows. There are two clinical conditions cancer vs control and high vs low grade tumor. Two ICs are estimated with weights correlating with cancer (IC 1) and tumor grade (IC 2). mRNA1 and miR-1 are down/upregulated in cancer, thereby having negative/positive gene weights in IC 1, and mRNA2 and miR-2 are down/upregulated in low grade tumors, thereby having negative/positive gene weights in IC 2.

Red/green denotes negative/positive log2 expressions. Yellow/cyan denotes negative/positive gene and component weights.

The method presented in the ICA paper for ranking and identifying ICs and genes might therefore be interesting from a classification point-of-view. Furthermore, the described method of ranking ICs has also shown applicable in the analysis of proteome data in Overgaard et al. [7] and Hansen et al. [8]. The aim of these studies were to predict better biomarkers for diabetic nephropathy.

## 6.2 miICA - miRNA target prediction based on ICA

The most common way to incorporate expression analysis into miRNA target predictions is by finding anti-correlation between the expressions of a

miRNA and its target mRNA as in TargetMiner [102], HOCTAR [108], and Genmir++ [101] listed in Section 5.2.2. However, due to, e.g., multiple miRNA regulations, transcription factor binding and site accessibility the anti-correlation between the miRNA expression and its target mRNA expression doesn't need to be significant [109]. It has, for example, been observed experimentally that miRNAs can act both activating and inhibiting during the cell cycle of HEK293 cells [122]. As a way to separate these factors ICA is applied to the expression data.

Figure 6.2 is a schematic illustration for the application of ICA to an expression matrix (**E**) of both mRNAs and miRNAs. **E** is decomposed into a component matrix (**C**) with two ICs (IC 1 and IC 2) and a mixing matrix (**M**). IC 1 is weighted differently between cancerous and control samples and IC 2 between tumor grades. These regulation patterns are also observed for the expressions of mRNA1 and miR-1, which differ between cancer and normal samples (red versus green), and for the expressions of mRNA2 and miR-2, which differ between different tumor grades (shades of red/green). Consequently, mRNA1 and miR-1 lie at the ends of IC 1 and mRNA2 and miR-2 lie at the ends of IC 2.

Notice, when the gene weights of the miRNAs and mRNAs lie at the very ends of the ICs, as miR-1 and mRNA1 in IC 1, then the miRNAs lie oppositely to the targeted mRNAs due to the inhibitory effect from the miRNAs. Thus, in ICs weighted differentially according to a factor of interest, e.g. cancer, it is expected that miRNA targets will be overrepresented oppositely to the miRNA itself. Additionally, if there is an overrepresentation of genes annotated with a biological pathway in the same end as for the miRNA targets, then I propose for a regulatory relationship between the miRNA regulation and the biological pathway. This is the basis of miICA. The individual steps of miICA are described below.

### 1. Applying ICA to the expression matrices

There are several numerical implementations of ICA. Among them are the JADE algorithm [123], based on 4th order statistics, the RADICAL algorithm [124], based on mutual information, and the fastICA algorithm [125], based on maximization of negentropy. It has been found that different ICA algorithms including fastICA, RADICAL and JADE perform similar overall [111; 113]. Because of its nice implementation in R fastICA is often used when analyzing microarray data by ICA [110; 116–119; 121; 126].

fastICA returns slightly different estimates at each run, therefore consensus component and mixing matrices are calculated by running fastICA  $N = 20$  times. Each time estimating the maximum number of ICs to preserve the full information hidden in data [127; 128]. The fastICA algorithm and the calculation of consensus matrices are described in OVC paper and the T1D paper.

Sometimes the miRNA expression levels might be very different from the mRNA expression levels. This makes it impossible to apply ICA to the combined mRNA and miRNA expression matrix, as was done in Fig. 6.2. If the miRNA expression levels are very high then the miRNAs might only have very negative/positive gene weights in a single IC as was observed in the OVC paper. Oppositely, if the miRNA expressions are very low, then the gene weights of the miRNAs will never be significant as in the T1D paper. A way to circumvent this is to use the mixing matrix for the mRNAs ( $\mathbf{M}_m$ ), and then estimate a component matrix for the miRNAs ( $\mathbf{C}_{mi}$ ) by solving  $\mathbf{E}_{mi} = \mathbf{C}_{mi}\mathbf{M}_m$ , where  $\mathbf{E}_{mi}$  is the miRNA expression matrix.

## 2. Finding significant components

The samples are classified according to one or more factors. An ANOVA is applied to the component weights of each IC to test if they are significantly different between groups of samples. The resulting  $p$ -values are corrected for multiple testing. If the classification of the samples is unknown then the method of ranking ICs represented in the ICA paper can be used to find significant ICs. For the use of miICA it will especially be the bioconcordance score of the ICA paper that is of interest, because ICs mapping to biological pathways are of special interest.

## 3. Performing target enrichment analysis

A target matrix ( $\mathbf{T}$ ) is generated with entries denoting the numbers of seed matches between the mRNA sequence and the miRNA seed, e.g.,  $T_{mg} = 5$  corresponds to miRNA  $m$  having five seed matches in the mRNA sequence of gene  $g$ . A Wilcoxon rank sum test is then applied to each IC to test if the putative miRNA targets have significantly lower or higher gene weights than non-targets. If they have lower or higher gene weights in an IC then there is an enrichment of miRNA targets in the negative or positive end of that IC, respectively.

Upon enrichment analysis the  $p$ -values are FDR corrected for multiple testing. The default length of the seed is 6 nucleotides, but the use of 7mer or 8mer seeds is optional. The mature sequences of the miRNAs are downloaded from miRBase ([mirbase.org](http://mirbase.org)). The mRNA sequences are extracted by using the biomaRt package in R and they can either be 3'UTR, coding region or 5'UTR sequences. When multiple sequences are available for the same mRNA, then the longest is used.

The enrichment analysis is applied to the ICs, which have gene weights that are non-Gaussian. Hence the statistical test must be usable for non-Gaussian data. This is true for the Kolmogorov-Smirnov (K-S) and Wilcoxon-rank-sum (WRS) tests. Tests that require Gaussianity of data are, e.g.,  $t$ -test or  $z$ -test. The WRS-test is applied, because

data might be ordinal, i.e. not direct measurements, and the K-S test does not allow for ordinal data.

#### 4. Performing pathway enrichment analysis

A target matrix is created as above. The entries now denote if the gene is annotated by a certain pathway or cellular process, i.e.  $T_{pg} = 1$  corresponds to gene  $g$  being annotated by pathway/process  $p$  and if  $T_{pg} = 0$  then gene  $g$  is not annotated by pathway/process  $p$ . A major repository of biological pathways is the Kyoto Encyclopedia of Genes and Genomes (KEGG) [129] and in the Gene Ontology (GO) [130] a gene product is associated with its cellular component, biological process or molecular function. The Molecular Signatures Database, Broad Institute (MSigDB, [131]) is a metabolic pathway and molecular process repository, which is curated from existing repositories (such as GO and KEGG) and from results on pathways and processes published in the literature.

A large amount of data can be the output from step 1-4 of miICA if the dimensions of the miRNA and mRNA expression matrices are high, which is usually the case. Depending on the aim of the study different approaches can be undertaken in order to retrieve useful information from miICA. In the T1D paper the aim was to find differentially expressed miRNAs in a model of type 1 diabetes (T1D). In the OVC paper the aim of the study was to investigate which pathways the ICs map to and propose new links between miRNA regulations and deregulated pathways in ovarian cancer (OVC). In both papers miICA was validated by searching for regulatory mechanisms of T1D and OVC already published in the literature. Additionally, in the OVC paper the samples were profiled using two different microarray platforms enabling a direct validation of miICA. This is a unique opportunity, since it is seldom that the same set of samples is profiled twice.

## CHAPTER

## 7

# miICA applied to a model of type 1 diabetes

According to the World Health Organization (WHO) the number of diabetes patients is expected to be doubled in 2030. For the approximately 10% of diabetes patients with T1D the disease requires daily insulin therapy, which is not a cure of T1D and the patients are still at risk for developing late diabetic complications. Furthermore, diabetes is a costly disease for the global health care system. The cause of T1D is a destruction of the  $\beta$ -cells and cytokines are implicated in this destruction. The investigation of genes and miRNAs regulated by specific  $\beta$  cell transcription factors and/or by cytokines could give new insight to the mechanisms of T1D and hopefully guide the way for the prevention and targeted treatment of T1D.

The aim of the T1D paper was to investigate mRNAs and miRNAs deregulated during  $\beta$ -cell maturation and the effect of cytokines on this maturation. From this investigation novel miRNAs causing an increased sensitivity of cytokines in mature  $\beta$ -cells are identified.

This chapter is based on the findings in:

C. H. Bang-Berthelsen, L. Pedersen, T. Fløyel, P. H. Hagedorn, T. Gylvin, and F. Pociot. *Independent component and pathway-based analysis of miRNA-regulated gene expression in a model of type 1 diabetes*. BMC Genomics, 12:97, 2011

## 7.1 An introduction to type 1 diabetes

Diabetes Mellitus is a severe chronic disease characterized by defects in insulin production and/or response to insulin. Two main classes of diabetes exist: type 1 diabetes (T1D) and type 2 diabetes (T2D). T1D is caused by an immune-mediated destruction of the  $\beta$ -cells in genetically predisposed individuals and T2D is caused by a resistance to and/or abnormal secretion of insulin. The  $\beta$ -cells are found in the islets of Langerhans in the pancreas [132]. They are responsible for the production and release of insulin. The maturation of  $\beta$ -cells in the pancreas is dependent on different transcription factors of which one is the pancreatic duodenum homeobox (pdx-1) [133]. Cytokines, such as interleukin-1 $\beta$  (IL-1 $\beta$ ), are implicated in the pathogenesis of T1D as mediators for the destruction of  $\beta$ -cells [134].

## 7.2 Type 1 diabetes microarray data

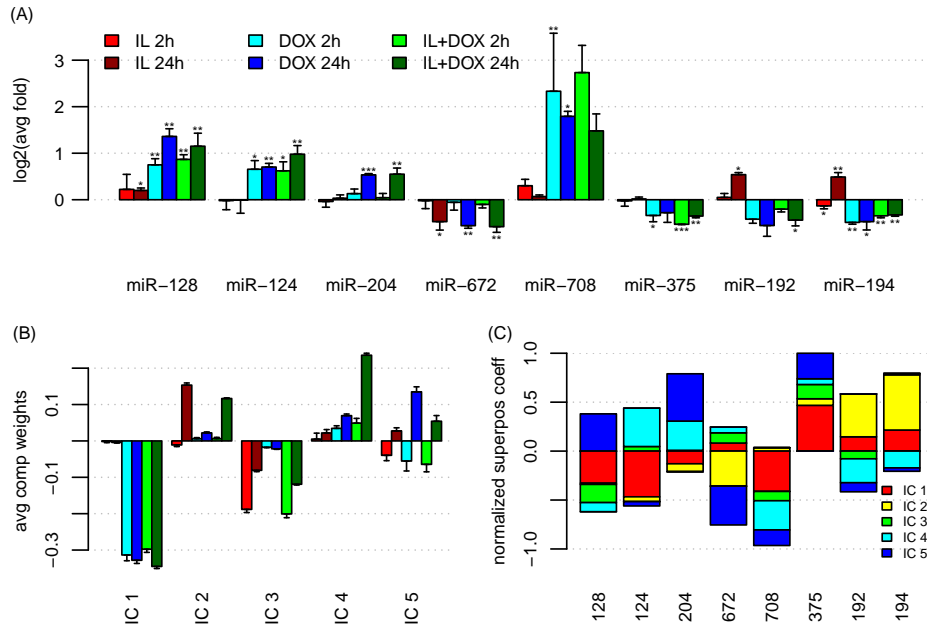
In this study the INSr $\alpha\beta$  rat cell line is used. In response to pdx-1 induction by doxycycline treatment the INSr $\alpha\beta$  cells progress from an immature ( $\alpha$ -cell phenotype) to a mature  $\beta$ -cell phenotype [133]. This maturation is accompanied by an increased sensitivity to the toxic effects of cytokines monitored by the addition of IL-1 $\beta$ . There are three experimental conditions in this experimental design, which are pdx-1 overexpression (denoted DOX) with or without additional IL-1 $\beta$  treatment (denoted IL and IL+DOX, respectively). Samples are taken 2h and 24h after induction of pdx-1 and/or IL-1 $\beta$  treatment, and there are three-four replicates within each of the six classes of samples. In total that gives 23 mRNA profiled and 18 miRNA profiled samples, where log2-ratios are calculated for each sample against a control reference. The miRNAs have very low expression levels as compared to the mRNAs. Therefore ICA is only applied to the mRNA expression matrix and 23 ICs are estimated.

## 7.3 Independent components capture the expression of miRNAs

Five ICs are differentially weighted between the six classes of samples ( $p < 10^{-4}$ , FWER corrected), see Fig. 7.1B. IC 1 is a Pdx-1 component weighted negatively upon pdx-1 induction. ICs 2 and 3 are cytokine components because they are non-present in DOX samples and weighted positively and negatively upon IL-1 $\beta$  treatment, respectively. The component weights of IC 4 correlate with the induction of pdx-1 and treatment with IL-1 $\beta$  after 24 h. IC 5 is a time component with component weights increasing from 2 h to 24 h in the three treatment conditions (DOX, IL, and IL+DOX).

There are eight miRNAs differentially expressed across the samples ( $p < 0.1$ , FWER corrected, see Fig. 7.1A), and these have expressions that are well explained by the five significant ICs. In fact a linear superposition of the five ICs captures  $\geq 97\%$  of their variance, see Fig. 7.1C and Table 7.A. Seven out of the eight differentially expressed miRNAs have putative targets enriched in one or more of the five significant ICs ( $q < 0.05$ , see Table 7.A).

The five significant ICs capture the expression of the eight miRNAs and seven of those have targets enriched among them. Therefore it is assumed that the transcriptional regulations within these five ICs are partly mediated by the eight miRNAs. Furthermore, targets of five (miR-128, 192, 194, 204, and 375) of the eight miRNAs were significantly regulated by cytokines in models of  $\beta$ -cell destruction in human islets [135], in rat  $\beta$ -cells [136] and in insulin producing INS-1E cells [137]. miR-672 does not have targets enriched in any of the five ICs. However, miR-672 is differentially expressed between  $\alpha$  and  $\beta$ -cells ( $p = 0.05$ , FWER corrected), see Supplementary Fig. 7 in the T1D paper. Together with miR-708 it is the first time that miR-672 has



**Figure 7.1** miRNA expressions and weights of ICs in a model of T1D.

There are three treatment conditions: IL (red), DOX (blue) and IL+DOX (green). (A) Log2-transformed fold changes for the eighth differentially expressed miRNAs between condition and control. \*:  $0.05 > q > 0.01$ , \*\*:  $0.01 > q > 0.001$ , \*\*\*:  $0.001 > q > 0$ . (B) The weights of the five ICs across the samples. (C) The coefficient for the linear superposition of the ICs giving the best fit of the miRNA expressions. The coefficients are scaled to have an absolute sum of one. The bars in (A) and (B) are standard deviations.



**Table 7.A** The eight miRNA regulations are composed of five ICs. The coefficient of determination ( $R^2$ ) is the variance explained by the linear superposition of the five ICs. See Appendix A for a definition of  $R^2$ . The row **ICA q** denotes the  $q$ -value for miRNA target enrichment in the end (**End**) of the IC (**IC**). The row **neg q** denotes the  $q$ -values for enrichment of negative correlation between the miRNA and its targets.

	miRNA							
	124	128	192	194	204	375	672	708
$R^2$	1.00	1.00	1.00	0.97	0.99	0.97	1.00	0.99
<b>ICA q</b>	4.4-3	0.026	7.E-45	0.012	0.012	0.015	0.27	7.5E-4
<b>IC</b>	IC 5	IC 5	IC 5	IC 2	IC 5	IC 1	IC 2	IC 5
<b>End</b>	neg	neg	neg	pos	neg	neg	pos	neg
<b>neg q</b>	2.7E-3	0.63	0.63	1.6E-3	0.63	2.7E-3	0.87	0.023

been examined in  $\beta$ -cells. Thus, all of the eight miRNAs are interesting candidates for further studies and especially miR-672 and miR-708 because of their novelty within  $\beta$ -cell research.

### 7.3.1 miICA performs better than negative correlation

Calculating the Pearson correlation coefficient between one of the eight miRNAs and the expressions of the mRNAs reveals that only four of the eight miRNAs have significant enrichment of negative correlation coefficients among their putative targets. Thus, it seems that miICA outperforms the use of negative correlation in miRNA target prediction, since the level of significance is also greater for miICA than for negative-correlation.

### 7.3.2 Cooperativity between miRNAs

In this study, small expression changes in the miRNAs are observed. However, even minute changes in miRNA expressions might have impact on mRNA expressions, and miRNAs acting in a cooperative manner can most likely induce biologically relevant expression changes in their targets. Cooperativity between miRNAs are investigated and it is found that miRNA pairs miR-375/672, miR-194/375, miR-192/375 and miR-124/194 have significant regulatory effects ( $q < 0.05$ ). Interestingly, miR-124 is upregulated and miR-194 is downregulated upon pdx-1 induction and only miR-194 is expressed in the IL cells (see Fig. 7.1A). Thus their regulation on common targets will be counteracting upon pdx-1 induction.

## 7.4 Combining the effects of independent components

As explained above the five ICs capture different aspects of  $\beta$ -cell maturation and their sensitivity to cytokines as was the aim of the study. Therefore it would be interesting to investigate if miRNA targets are significantly enriched in combinations of the five ICs. Instead of performing the enrichment analysis on the gene weights of putative targets the enrichment is now performed on the maximum rank of a gene within the ICs. First each gene is ranked within each IC. Secondly, in pairs of ICs the maximum rank for each gene is found. Thirdly, an enrichment analysis for these maximum ranks of putative miRNA targets is carried out.

As an example, enrichment of targets in the positive end of IC 3 and the negative end of IC 5 is explored. Genes that lies at the very positive end of IC 3 and at the very negative end of IC 5 will in general be downregulated in IL-1 $\beta$  treated cells and they become even more downregulated with time. To see this, consider the component weights of IC 3 in Fig. 7.1 and then add to this the component weights of IC 5 flipped around the x-axis. The following is an example of how the maximum rank of genes can be found. Assume that IC 3 has gene weights  $C_{.3}$  with ranks  $\rho(C_{.3})$

$$C_{.3} = \begin{bmatrix} -0.56 \\ -3.78 \\ -2.43 \\ 0.43 \end{bmatrix} \Rightarrow \rho(C_{.3}) = \begin{bmatrix} 3 \\ 1 \\ 2 \\ 4 \end{bmatrix}, \quad (7.1)$$

and that IC 5 has gene weights  $C_{.5}$  with ranks  $\rho(C_{.5})$

$$C_{.5} = \begin{bmatrix} 0.24 \\ 2.63 \\ 3.56 \\ -0.62 \end{bmatrix} \Rightarrow \rho(C_{.5}) = \begin{bmatrix} 2 \\ 3 \\ 4 \\ 1 \end{bmatrix}, \quad (7.2)$$

then the maximum rank for the gene weights in a combination of IC 3 and the inverse of IC 5 is

$$\max(\rho(C_{.3}), \rho(-C_{.5})) = \max\left(\begin{bmatrix} 3 \\ 1 \\ 2 \\ 4 \end{bmatrix}, \begin{bmatrix} 3 \\ 2 \\ 1 \\ 4 \end{bmatrix}\right) = \begin{bmatrix} 3 \\ 2 \\ 2 \\ 4 \end{bmatrix}. \quad (7.3)$$

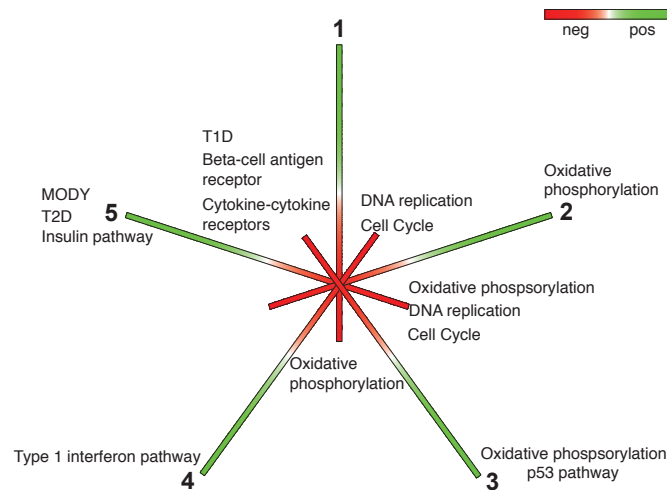
Targets of miR-128 are actually significantly enriched in the positive end of IC 3 combined with the negative end of IC 5 ( $q = 0.00090$ ). In the

linear superposition of ICs, IC 3 contributes negatively and IC 5 contributes positively to the expression of miR-128. This is in perfect agreement with its targets being enriched in the positive end of IC 3 and the negative end of IC 5. Interestingly, when performing enrichment of targets in pairs of the five ICs the positive ends of ICs 1, 2, and 3 together with the negative end of IC 5 show up as most significantly enriched for miRNA targets. In the negative end of IC 5 most of the miRNAs also have targets enriched as seen from Table 7.A.

## 7.5 Components map to diabetes related pathways

IC 3 mimics the sensitivity to IL-1 $\beta$  as being weighted negatively upon IL-1 $\beta$  treatment. Testing for enrichment of pathways there are several T1D related pathways, besides the T1D pathway, enriched in IC 3, e.g.  $\beta$ -cell antigen receptor, which promotes  $\beta$ -cell activation, in the negative end, and oxidative phosphorylation in the positive end. Genes involved in oxidative phosphorylation are downregulated in diabetes [138], which means that they should be downregulated in IL+DOX cells. IC 3 is weighted most negatively in IL-1 $\beta$  treated cells, and since targets for oxidative phosphorylation are enriched in the positive end of IC 3, these will be downregulated in IL-1 $\beta$  treated cells. This complies with oxidative phosphorylation being downregulated in T1D.

From the miRNA expressions in Fig. 7.1A and from a study on miRNA regulations in human pancreas [139] miR-375 is downregulated in  $\beta$ -cells. In agreement with this miR-375 targets are enriched in the negative end of IC 1, which corresponds to them being upregulated upon pdx-1 induction.



**Figure 7.2** Pathways enriched at the ends of the ICs.  
Red: negative end. Green: positive end.

Five of the eight miRNAs have targets enriched in the negative end of IC 5. IC 5 is enriched for genes annotated by oxidative phosphorylation in the negative end. When performing enrichments for miRNA targets in pairs of ICs, IC 5 shows up in 70 out of the 115 significant pairs with  $q < 0.05$ . In only one of these 70 pairs the enrichment of miRNA targets is in the positive end of IC 5. Thus miRNA targets are generally enriched in the negative end of IC 5, meaning that they are downregulated over time. Most pronounced in DOX cells, indicating that IL-1 $\beta$  might inhibit miRNA mediated regulations. This is of course pure speculations and needs further experimental validation. Though it does illustrate the hypothesis-generating feature of miICA.

## 7.6 Further insight into miRNA regulations

Next-generation-sequencing (NGS) is a type of high-throughput-sequencing (HTS) that will be used more frequently in the future to explore changes in gene regulation. In an NGS experiment RNA samples are sequenced and each sequence, called a RefSeq, is annotated and given a count measuring the presence of the RefSeq within the sample. However, it is not possible to annotate all RefSeqs due to SNPs or point mutations within the sequenced samples or maybe because the RefSeqs might not be of human origin. In a Chinese study exogenous plant miRNAs are found to be present within the tissue and sera of various animals [140]. The plant miRNAs are mainly acquired orally. For example, a rice miRNA was found to be enriched in chinese subjects.

It is believed that regulations from viruses contribute to the onset of T1D [141; 142]. In this context viral miRNAs are interesting, because it has been found that viral miRNAs are differentially regulated upon infection of the virus itself [143]. Therefore we want to investigate whether viral miRNA can mediate the onset of T1D. To achieve this we use a data set consisting of a large collection of blood samples taken from patients 3-6 months after the debut of T1D. In this data set two pools of samples are created, one with samples only from Denmark and one with samples from Europe. Together with a pool of Danish control blood samples NGS is performed. The task is then to test if some of the unannotated RefSeqs could be of viral origin. Especially, we want to see if we can match some of the unannotated RefSeqs with the sequences of Coxsackie virus A and B (Cox A and B) and viral miRNA. Cox A and B are strongly coupled to the development of T1D [144; 145] and are therefore of special interest.



## CHAPTER

## 8

# miICA reveals miRNA regulations in ovarian cancer

Early diagnosis and targeted treatment are important for all types of cancer. In the case of ovarian cancer (OVC) it is especially important because it is the fifth leading cause of death among females (in the US) [146] and it is usually diagnosed in a late stage [147]. From a societal point of view it is also important to have an early diagnosis tool and a targeted treatment in order to increase the chance of survival and the quality of life for the patients. From 2005-2009 there were 577 new occasions of OVC per year and 387 deaths per year in Denmark according to the Danish Cancer Society ([cancer.dk](http://cancer.dk)). The aim of applying miICA to a microarray dataset of OVC was initially to validate the method, because the samples have been profiled twice using two different microarray assays. However, as the analysis of the data proceeded new an interesting results were found.

This chapter is based on:

L. Pedersen, W. Xiao, M. Jensen, and P. Gunaratne. Independent component analysis reveals possible microRNA regulated pathways in ovarian cancer. Submitted to PLoS Computational, 2012

## 8.1 Ovarian cancer microarray data

The OVC data set contains a test and a validation set of mRNA expression profiles. The test set is used in the OVC paper to test miICA and the validation set is used to validate the results from the test set. The test set includes miRNA and mRNA expressions in 537 samples. Thus, 537 ICs are estimated in the first step of miICA. Out of the 537 samples there are 514 cancer, 15 recurrent and 8 normal samples. The 15 recurrent samples are taken from patients, where there is a paired sample at the initial onset of OVC. These matched pairs of samples make it possible to deeply investigate the mRNA and miRNA regulations at the recurrence of OVC as studied in the OVC paper.

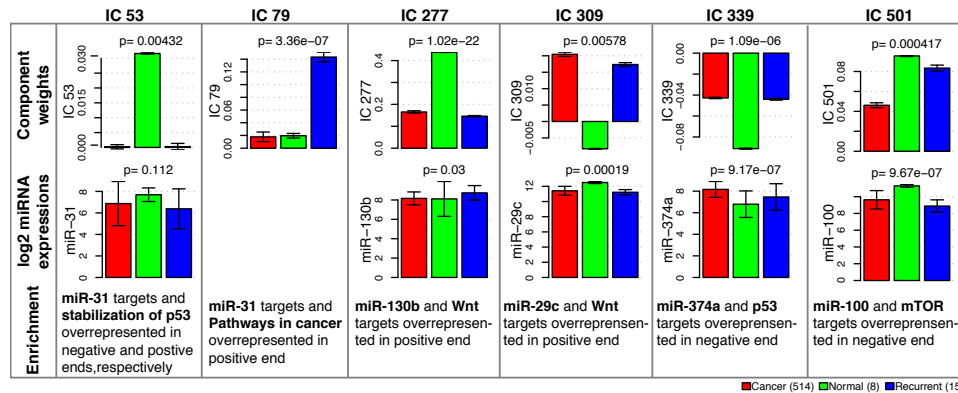
## 8.2 The p53 pathway is deregulated in ovarian cancer

p53 acts as a tumor suppressor and the p53 pathway is deregulated in 50% of all cancers [148]. The p53 pathway has been linked to high-grade serous OVC and about 96% of ovarian tumors from TCGA database exhibited mutations in p53 [149]. As a validation it is investigated whether miICA can rediscover the deregulation of the p53 pathway found experimentally in OVC [150]. There are 178 ICs that are weighted differentially between cancer, recurrent and normal samples. Of these, IC 339 shows the strongest enrichment of genes annotated by the p53 pathway. The enrichment is in the negative end of IC 339. This IC is not present in cancer and recurrent samples and is weighted negatively in normal samples (see Fig. 8.1). Since IC 339 is lowest weighted in normal samples and the p53 pathway is enriched in the negative end, genes annotated by the p53 pathway will in general be downregulated in cancerous samples. This complies with the fact that the p53 gene is deficient in about 96% of the TCGA ovarian tumors [149] with a lower p53 activity as a consequence. The finding of the p53 pathway as being deregulated in OVC based on the gene and component weights of IC 339 is a first validation of miICA.

## 8.3 miRNA regulations of the Wnt and mTOR pathways

The Wnt [151] and mTOR [152] pathways are deregulated in OVC. The miRNAs miR-29c [153] and miR-130b [151] targets the Wnt pathway and the tumor suppressor miR-100 [152] regulates the mTOR pathway.

The expression pattern of miR-130b is upregulated in recurrent samples compared to normal and cancer samples, see Fig. 8.1. Whereas, miR-29c is



**Figure 8.1** Components mapping to the p53, Wnt and mTOR pathways. Barplots show the component weights and log2-transformed miRNA expressions. Bars are mean values  $\pm$  variance. The enrichment for each IC is written below the bar plots. If the enrichment is in positive end then the enriched targets follow the regulations of the component weights. If the enrichment is in the negative end they anti-correlate with the regulations of the components weights.

downregulated in both recurrent and cancer samples. Thus the regulatory impact from these miRNAs on the Wnt pathway must be different and not surprisingly, they have targets enriched in two different ICs. In IC 277 targets of miR-130b are enriched in the positive end and in IC 309 targets of miR-29c are enriched in the positive end. Figure 8.1 shows barplots for the component weights of ICs 277 and 309. IC 277 is weighted highest and positively in normal samples. The component weights of 309 anti-correlate with this regulation in the sense that IC 309 is weighted negatively in normal samples and positively in recurrent and normal samples.

Despite their anti-correlated component weights ICs 277 and 309 are both enriched for the Wnt pathway in the positive end. This again might reflect two different controlling mechanisms, since miR-29c and miR-130b also have different regulatory patterns. Considering the component weights of ICs 277 and 309 in Fig. 8.1, it is seen that these contribute to an inhibition or activation respectively, of Wnt target genes in cancerous samples. In fact it is known that both activation or inhibition of Wnt target genes are observed in cancers [17; 154]. Dickkopf1 (Dkk1) is a known inhibitor of Wnt signaling as described in Part 1 and it is frequently overexpressed in OVC [154]. In ICs 277 and 309 Dkk1 has a negative and a positive gene weight, respectively, which in general results in an upregulation of Dkk1 in cancer. This complies with the experimental finding and strengthens the mapping of ICs 277 and 309 to the Wnt pathway.

Because miR-100 is a tumor suppressor it is downregulated in cancerous cells (Fig. 8.1). The targets of miR-100 are enriched in the negative end of



IC 501, which is lowest weighted in cancer samples. In the negative end of IC 501 targets of the mTOR pathway are also enriched and a link between miR-100 downregulation and upregulation of mTOR target genes in cancer can be established.

The validation set consists of 584 samples, because there is a greater number of cancer samples profiled using an Affymetrix microarray assay. This is the assay used for profiling the mRNAs in the validation set. miICA applied to the validation set documents one IC with overrepresentation of p53 genes corresponding to a downregulation of these, one IC documents a link between miR-130b and miR-29c and the Wnt pathway and yet another IC propose for a regulatory relationship between miR-100 and the mTOR pathway. These findings support and validate the results of miICA applied to the test set. This high degree of congruence observed between the results of the test and validation set gives a high level of confidence in the stability and predictive power of miICA.

## 8.4 Fingerprints of miRNA regulations

miICA can be used to find fingerprints of miRNAs that have been shown to be differentially regulated in the literature, but does not appear differentially regulated in the TCGA data. This can be caused by experimental and biological noise.

In cell lines where the p53 gene is dysfunctional or absent, miR-31 stops the cells from dividing wildly and induces apoptosis [150]. Transfecting miR-31 back into ovarian cancer cells actually turns off tumor growth. Actually, miR-31 is the most downregulated miRNA in OVC and consequently the targets of miR-31 should be upregulated in cancer samples. In the TCGA data miR-31 is not significantly, differentially expressed between cancer and normal samples ( $p = 0.11$ ). Figure 8.1 shows the expression of miR-31 in cancer and normal samples. There is a trend of miR-31 being downregulated in cancer samples, as expected, since miR-31 inhibits tumor growth.

There is one IC, which can document a regulatory relationship between miR-31 and the p53 pathway. IC 53 is positively weighted in normal samples and is non-present in cancerous samples ( $q = 0.0058$ , see Fig. 8.1). For the miR-31 targets to be upregulated in cancerous samples, they need to be enriched in the negative end of IC 53, which is actually the fact ( $q = 0.020$ ). Furthermore, in the positive end of IC 53 targets for the stabilization of p53 is enriched ( $q = 0.035$ ) meaning that they in general are downregulated in cancerous samples. This complies with the p53 being destabilized in OVC and the stabilization can be mediated by miR-31 through the transcriptional program mapped by IC 53.

## 8.5 miR-374a and p53 form a new regulatory relationship

In the sections above three examples were described as to validate the predictions of miICA. In this section new regulatory relationships are sought. The starting point is to use the above finding of IC 339 mapping to the p53 pathway.

Among the gene weights of IC 339 I search for miRNA targets, which are overrepresented in either the negative or positive ends. Table 8.A lists the nine miRNAs with  $q < 10^{-4}$  for the enrichment of targets in IC 339. Among these nine miRNAs only miR-374a,b and miR-154\* are differentially expressed between cancer, normal and recurrent samples. For their expressions see Fig. 8.2. Only miR-374a and miR-154\* are upregulated in both cancer and recurrent samples and miR-374a to a larger extent than miR-154\*. IC 339 has almost no weight in cancerous cells and above it was found that IC 339 documents that p53 target genes are downregulated in cancerous samples. Thus the very significant upregulation of miR-374a in cancerous cells together with its targets being enriched in the negative end of IC 339 might lead to a miR-374a mediated inhibition of p53 target genes in cancer. This inhibitory effect of miR-374a on the p53 pathway is a new regulatory relationship of OVC not yet prosed in the literature.

Recently, there has been a lot of research on p63, a p53 family member, because its role in cancer seems to be dual and very complex. It is known that depending on the isoform of p63 it can either activate or inhibit p53

**Table 8.A** miRNAs with the strongest enrichment of targets in IC 339.

1st column:  $q$ -values of the target enrichment analysis. 2nd column: the end of IC 339 in which the enrichment is found. 3rd column:  $q$ -values for the test of differentially expression between cancer, recurrent and normal samples.

	enrichment $q$ ( $\cdot 10^{-5}$ )	enrichment end of IC 339	diff ( $q$ )	exp
<b>miR-374a</b>	6.17	negative	$4.4 \cdot 10^{-6}$	
<b>miR-374b</b>	6.17	negative	$5.5 \cdot 10^{-3}$	
<b>miR-154*</b>	1.54	negative	0.033	
miR-487a	1.54	negative	0.24	
let-7a*	0.774	negative	0.053	
let-7b*	0.774	negative	0.22	
let-7f-1*	0.7.74	negative	0.23	
let-7f-2*	0.774	negative	0.77	
miR-744	0.290	positive	0.15	

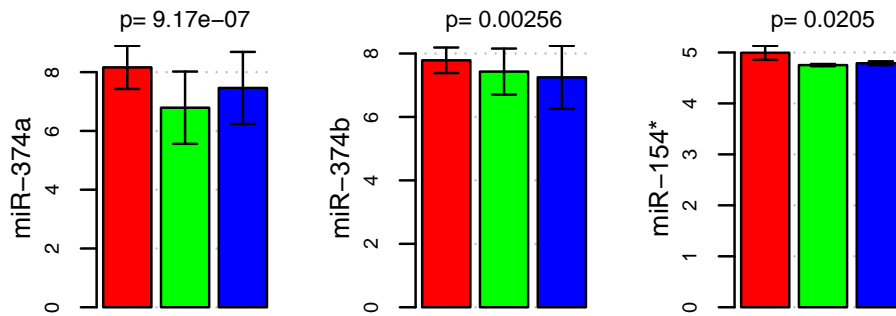


Figure 8.2 Log2-transformed miRNA expressions for miR-374a, - 374b, and -154\*.

target genes [155]. It is known that phosphorylated  $\Delta\text{Np63}\alpha$ , an isoform of p63, activates miR-374a and there is a seed match between miR-374a and p63. In IC 339 p63 lies among the 10 % genes with the most negative gene weights, i.e. it lies in the same end as genes annotated by the p53 pathway. In a recent study of Creighton et al. [156] it was found that miR-374a is strongly linked to the overall survival of patients in TCGA.

The fact that miR-374a correlates with better survival and miICA documents it to regulate the p53 pathway through p63 targeting is very promising for the possibility of treating OVC by use of miR-374a.

ICA has already been used as a tool for classification of tumors. Based on gene and pathway regulations found from miICA it would be nice to pinpoint miRNAs that can be used to diagnose ovarian cancer, because of its high mortality rate caused by a late discovery of the cancer. Actually, for the diagnosis of cutaneous T-cell lymphoma three miRNAs have been found to be the key factors [157]. Also, as mentioned above miR-31 and miR37a seem to be promising for the treatment of ovarian cancer, since miR-31 turns off tumor growth and miR-374a increases the sensitivity of chemotherapy [158] and correlates with a good prognosis for survival of OVC.

## CHAPTER

## 9

# Concluding remarks Part II

miICA is not a traditional miRNA target prediction method in the sense that it does not output a ranked list of miRNA targets as in PITA [93] and Sylamer [100]. The output of miICA can be a ranked list of miRNA targets, but the aim of miICA is also to retrieve mRNAs and miRNAs that are differentially expressed and investigate if these represent or regulate biological pathways important for the studied biological system. Thus, miICA is an all-in-one method for gene expression analysis.

The secondary structure of the mRNA [93] and the flanking region around the seed sites [95; 159] make some seed sites more favourable than others, as mentioned in Section 5.2. Therefore incorporating a measure for the structure and contents of the flanking region surrounding the seed site would improve the performance of miICA significantly. Furthermore, since high-throughput sequencing (HTS) becomes more common for gene profiling an adjustment of miICA would be desirable to enable the analysis of both microarray and HTS data by use of miICA. ICA has not yet been applied to HTS data. Though it seems plausible that ICA is applicable to HTS data, since the structure of microarray and HTS data are similar. They both measure gene expression levels.

Initial pre-processing of the expression data and creation of a classification scheme are the only steps that need to be performed before applying miICA. Even if miICA is also extended to include HTS data and an analysis of the seed site region. Therefore, miICA would be well suited as a web-based miRNA target prediction tool, where the user simply uploads the expression matrices and a classification scheme for the samples. This is the ultimate goal of the further development of miICA.



# Bibliography

- [1] P. B. Jensen, L. Pedersen, S. Krishna, and M. H. Jensen. *A Wnt oscillator model for somitogenesis*. Biophysical journal, 98(6):943–950, March 2010.
- [2] B. Mengel, A. Hunziker, L. Pedersen, A. Trusina, M. H. Jensen, and S. Krishna. *Modeling oscillatory control in NF-kappaB, p53 and Wnt signaling*. Curr Opin Genet Dev, 20(6):656–664, December 2010.
- [3] C. H. Bang-Berthelsen, L. Pedersen, T. Fløyel, P. H. Hagedorn, T. Gylvin, and F. Pociot. *Independent component and pathway-based analysis of miRNA-regulated gene expression in a model of type 1 diabetes*. BMC Genomics, 12:97, 2011.
- [4] L. Pedersen, M. H. Jensen, and S. Krishna. *Dickkopf1 - a new player in modelling the wnt pathway*. PLoS ONE, 6(10):e25550, 2011.
- [5] L. Pedersen and P. H. Hagedorn. *Ranking of Independent Components in Gene Expression Microarray Data Using Weighted Scoring*. Under revision in PLoS ONE, 2011.
- [6] L. Pedersen, W. Xiao, M. Jensen, and P. Gunaratne. *Independent component analysis reveals possible microRNA regulated pathways in ovarian cancer*. Submitted to PLoS Computational, 2012.
- [7] A. J. Overgaard, H. G. Hansen, M. Lajer, L. Pedersen, L. Tarnow, P. Rossing, J. N. McGuire, and F. Pociot. *Plasma proteome analysis of patients with type 1 diabetes with diabetic nephropathy*. Proteome Sci, 8:4, 2010.

- [8] H. G. Hansen, J. Overgaard, M. Lajer, F. Hubalek, P. Højrup, L. Pedersen, L. Tarnow, P. Rossing, F. Pociot, and J. N. McGuire. *Finding diabetic nephropathy biomarkers in the plasma peptidome by high-throughput magnetic bead processing and MALDI-TOF-MS analysis*. Proteomics Clin Appl, 4(8-9):697–705, September 2010.
- [9] L. Larsen, M. W. Rosenstjerne, L. W. Gaarn, A. Bagge, L. Pedersen, C. M. Dahmcke, J. H. Nielsen, and L. T. Dalgaard. *Expression and localization of microRNAs in perinatal rat pancreas: role of miR-21 in regulation of cholesterol metabolism*. PLoS ONE, 6(10):e25997, 2011.
- [10] L. Wolpert, T. Jessel, P. Lawrence, E. Meyerowitz, E. Robertson, and J. Smith. *Principles of development*. Oxford University Press, 3rd edition, 2006.
- [11] S. Gilbert. *Developmental biology*. Sinauer Associates, 7th edition, 2003.
- [12] J. Cooke and E. C. Zeeman. *A clock and wavefront model for control of the number of repeated structures during animal morphogenesis*. J Theor Biol, 58(2):455–476, May 1976.
- [13] A. Aulehla, C. Wehrle, B. Brand-Saberi, R. Kemler, A. Gossler, B. Kanzler, and B. G. Herrmann. *Wnt3a Plays a Major Role in the Segmentation Clock Controlling Somitogenesis*. Developmental Cell, 4(3):395–406, March 2003.
- [14] M.-L. Dequéant, E. Glynn, K. Gaudenz, J. Chen, A. Mushegian, and O. Pourquié. *A complex oscillating network of signaling genes underlies the mouse segmentation clock*. Science, 314(5805):1595–1598, December 2006.
- [15] M.-L. Dequéant and O. Pourquié. *Segmental patterning of the vertebrate embryonic axis*. Nat Rev Genet, 9(5):370–382, May 2008.
- [16] K. I. Pinson, J. Brennan, S. Monkley, B. J. Avery, and W. C. Skarnes. *An LDL-receptor-related protein mediates Wnt signalling in mice*. Nature, 407(6803):535–538, September 2000.
- [17] C. Y. Logan and R. Nusse. *The Wnt signaling pathway in development and disease*. Annu Rev Cell Dev Biol, 20:781–810, 2004.
- [18] K. Horikawa, K. Ishimatsu, E. Yoshimoto, S. Kondo, and H. Takeda. *Noise-resistant and synchronized oscillation of the segmentation clock*. Nature, 441(7094):719–723, June 2006.
- [19] Y. J. Jiang, B. L. Aerne, L. Smithers, C. Haddon, and D. Ish-Horowicz. *Notch signalling and the synchronization of the somite segmentation clock*. Nature, 408(6811):475–479, November 2000.

- [20] A. Aulehla and O. Pourquié. *Signaling Gradients during Paraxial Mesoderm Development*. Cold Spring Harbor Perspectives in Biology, 2(2):a000869–a000869, February 2010.
- [21] M. Oginuma, Y. Niwa, D. L. Chapman, and Y. Saga. *Mesp2 and Tbx6 cooperatively create periodic patterns coupled with the clock machinery during mouse somitogenesis*. Development, 135(15):2555–2562, August 2008.
- [22] R. Diez del Corral, I. Olivera-Martinez, A. Goriely, E. Gale, M. Maden, and K. Storey. *Opposing FGF and retinoid pathways control ventral neural pattern, neuronal differentiation, and segmentation during body axis extension*. Neuron, 40(1):65–79, September 2003.
- [23] M. B. Wahl, C. Deng, M. Lewandoski, and O. Pourquié. *FGF signaling acts upstream of the NOTCH and WNT signaling pathways to control segmentation clock oscillations in mouse somitogenesis*. Development, 134(22):4033–4041, November 2007.
- [24] Y. Saga, N. Hata, H. Koseki, and M. M. Taketo. *Mesp2: a novel mouse gene expressed in the presegmented mesoderm and essential for segmentation initiation*. Genes Dev, 11(14):1827–1839, July 1997.
- [25] Y. Niwa, H. Shimojo, A. Isomura, A. González, H. Miyachi, and R. Kageyama. *Different types of oscillations in Notch and Fgf signaling regulate the spatiotemporal periodicity of somitogenesis*. Genes Dev, 25(11):1115–1120, June 2011.
- [26] H. Aoyama and K. Asamoto. *The developmental fate of the rostral/caudal half of a somite for vertebra and rib formation: experimental confirmation of the resegmentation theory using chick-quail chimeras*. Mech Dev, 99(1-2):71–82, December 2000.
- [27] M. McGrew and O. Pourquié. *FGF signaling controls somite boundary position and regulates segmentation clock control of spatiotemporal Hox gene activation*. Cell, 106(2):219–232, 2001.
- [28] F. Crick. *Central dogma of molecular biology*. Nature, 227(5258):561–563, August 1970.
- [29] G. Tian, S. Krishna, S. Pigolotti, M. H. Jensen, and K. Sneppen. *Oscillations and temporal signalling in cells*. Phys Biol, 4(2):R1–17, June 2007.
- [30] J. Lewis. *Autoinhibition with transcriptional delay: a simple mechanism for the zebrafish somitogenesis oscillator*. Curr Biol, 13(16):1398–1408, August 2003.



- [31] S. Strogatz. *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. Studies in nonlinearity. Westview Press, 1994. ISBN 9780738204536.
- [32] S. Bernard, B. Cajavec, L. Pujo-Menjouet, M. C. Mackey, and H. Herzel. *Modelling transcriptional feedback loops: the role of Gro/TLE1 in Hes1 oscillations*. Phil Trans R Soc A, 364(1842):1155–1170, May 2006.
- [33] M. Campanelli and T. Gedeon. *Somitogenesis clock-wave initiation requires differential decay and multiple binding sites for clock protein*. PLoS Comput Biol, 6(4):e1000728, 2010.
- [34] M. H. Jensen, K. Sneppen, and G. Tiana. *Sustained oscillations and time delays in gene expression of protein Hes1*. FEBS Lett, 541(1-3):176–177, April 2003.
- [35] J. G. Rodríguez-González, M. Santillán, A. C. Fowler, and M. C. Mackey. *The segmentation clock in mice: interaction between the Wnt and Notch signalling pathways*. J Theor Biol, 248(1):37–47, September 2007.
- [36] A. Goldbeter and O. Pourquié. *Modeling the segmentation clock as a network of coupled oscillations in the Notch, Wnt and FGF signaling pathways*. J Theor Biol, 252(3):574–585, June 2008.
- [37] E. Lee, A. Salic, R. Krüger, R. Heinrich, and M. W. Kirschner. *The Roles of APC and Axin Derived from Experimental and Theoretical Analysis of the Wnt Pathway*. PLoS Biol, 1(1):e10, October 2003.
- [38] C. Wawra, M. Kühl, and H. A. Kestler. *Extended analyses of the Wnt/beta-catenin pathway: robustness and oscillatory behaviour*. FEBS Lett, 581(21):4043–4048, August 2007.
- [39] P. François, V. Hakim, and E. D. Siggia. *Deriving structure from evolution: metazoan segmentation*. Mol Syst Biol, 3:154, 2007.
- [40] R. E. Baker, S. Schnell, and P. K. Maini. *A clock and wavefront mechanism for somite formation*. Dev Biol, 293(1):116–126, May 2006.
- [41] M. Santillán and M. C. Mackey. *A proposed mechanism for the interaction of the segmentation clock and the determination front in somitogenesis*. PLoS ONE, 3(2):e1561, 2008.
- [42] M. Oginuma, Y. Takahashi, S. Kitajima, M. Kiso, J. Kanno, A. Kimura, and Y. Saga. *The oscillation of Notch activation, but not its boundary, is required for somite border formation and rostral-caudal patterning within a somite*. Development, 137(9):1515–1522, May 2010.

- [43] L. Herrgen, S. Ares, L. G. Morelli, C. Schröter, F. Jülicher, and A. C. Oates. *Intercellular coupling regulates the period of the segmentation clock*. Curr Biol, 20(14):1244–1253, July 2010.
- [44] L. Morelli, S. Ares, L. Herrgen, C. Schröter, F. Jülicher, and A. C. Oates. *Delayed coupling theory of vertebrate segmentation*. HFSP Journal, 3(1):55–66, 2009.
- [45] P. J. Murray, P. K. Maini, and R. E. Baker. *The clock and wavefront model revisited*. J Theor Biol, 283(1):227–238, August 2011.
- [46] A. J. Terry, M. Sturrock, J. K. Dale, and M. A. J. Chaplain. *A spatio-temporal model of Notch signalling in the zebrafish segmentation clock: conditions for synchronised oscillatory dynamics*. PLoS ONE, 6(2): e16980, 2011.
- [47] I. H. Riedel-Kruse, C. Müller, and A. C. Oates. *Synchrony dynamics during initiation, failure, and rescue of the segmentation clock*. Science, 317(5846):1911–1915, September 2007.
- [48] F. Giudicelli, E. M. Ozbudak, G. J. Wright, and J. Lewis. *Setting the tempo in development: an investigation of the zebrafish somite clock mechanism*. PLoS Biol, 5(6):e150, June 2007.
- [49] J. Behrens, J. P. v. Kries, M. Kühl, L. Bruhn, D. Wedlich, R. Grosschedl, and W. Birchmeier. *Functional interaction of beta-catenin with the transcription factor LEF-1*. Nature, 382(6592):638–642, August 1996.
- [50] V. Korinek, N. Barker, K. Willert, M. Molenaar, J. Roose, G. Wagenaar, M. Markman, W. Lamers, O. Destree, and H. Clevers. *Two members of the Tcf family implicated in Wnt/beta-catenin signaling during embryogenesis in the mouse*. Mol Cell Biol, 18(3):1248–1256, 1998.
- [51] M. V. Semenov, K. Tamai, B. K. Brott, M. Kühl, S. Sokol, and X. He. *Head inducer Dickkopf-1 is a ligand for Wnt coreceptor LRP6*. Curr Biol, 11(12):951–961, June 2001.
- [52] A. Bafico, G. Liu, A. Yaniv, A. Gazit, and S. Aaronson. *Novel mechanism of Wnt signalling inhibition mediated by Dickkopf-1 interaction with LRP6/Arrow*. Nature cell biology, 3:683–686, 2001.
- [53] H.-M. I. Yu, B. Jerchow, T.-J. Sheu, B. Liu, F. Costantini, J. E. Puzas, W. Birchmeier, and W. Hsu. *The role of Axin2 in calvarial morphogenesis and craniosynostosis*. Development, 132(8):1995–2005, April 2005.

- [54] B. MacDonald, M. Adamska, and M. Meisler. *Hypomorphic expression of Dkk1 in the doubleridge mouse: dose dependence and compensatory interactions with Lrp6*. Development, 131:2543–2552, 2004.
- [55] M. Mukhopadhyay, S. Shtrom, C. Rodriguez-Esteban, L. Chen, T. Tsukui, L. Gomer, D. D. Dorward, A. Glinka, A. Grinberg, S.-P. Huang, C. Niehrs, J. C. I. Belmonte, and H. Westphal. *Dickkopf1 is required for embryonic head induction and limb morphogenesis in the mouse*. Dev Cell, 1:423–434, 2001.
- [56] T. Greco, S. Takada, M. M. Newhouse, J. A. McMahon, A. P. McMahon, and S. A. Camper. *Analysis of the vestigial tail mutation demonstrates that Wnt-3a gene dosage regulates mouse axial development*. Genes Dev, 10:313–324, 1996.
- [57] A. Niida, T. Hiroko, M. Kasai, Y. Furukawa, Y. Nakamura, Y. Suzuki, S. Sugano, and T. Akiyama. *DKK1, a negative regulator of Wnt signaling, is a target of the beta-catenin/TCF pathway*. Oncogene, 23: 8520–8526, 2004.
- [58] J. M. González-Sancho, O. Aguilera, J. M. García, N. Pendás-Franco, C. Peña, S. Cal, A. García de Herreros, F. Bonilla, and A. Muñoz. *The Wnt antagonist DICKKOPF-1 gene is a downstream target of beta-catenin/TCF and is downregulated in human colon cancer*. Oncogene, 24(6):1098–1103, February 2005.
- [59] I. V. Chia and F. Costantini. *Mouse axin and axin2/conductin proteins are functionally equivalent in vivo*. Mol Cell Biol, 25(11):4371–4376, June 2005.
- [60] X. Zeng, K. Tamai, B. Doble, S. Li, H. Huang, R. Habas, H. Okamura, J. Woodgett, and X. He. *A dual-kinase mechanism for Wnt co-receptor phosphorylation and activation*. Nature, 438:873–877, 2005.
- [61] J. Dubrulle and O. Pourquié. *fgf8 mRNA decay establishes a gradient that couples axial elongation to patterning in the vertebrate embryo*. Nature, 427(6973):419–422, January 2004.
- [62] S. Gibb, A. Zagorska, K. Melton, G. Tenin, I. Vacca, P. Trainor, M. Maroto, and J. K. Dale. *Interfering with Wnt signalling alters the periodicity of the segmentation clock*. Dev Biol, 330(1):21–31, June 2009.
- [63] P. P. Tam. *The control of somitogenesis in mouse embryos*. J Embryol Exp Morphol, 65 Suppl:103–128, October 1981.

- [64] A. Aulehla and B. G. Herrmann. *Segmentation in vertebrates: clock and gradient finally joined*. Genes Dev, 18(17):2060–2067, September 2004.
- [65] S. Kim and E.-h. Jho. *The protein stability of Axin, a negative regulator of Wnt signaling, is regulated by Smad ubiquitination regulatory factor 2 (Smurf2)*. J Biol Chem, 285(47):36420–36426, November 2010.
- [66] A. Aulehla, W. Wiegraebe, V. Baubet, M. B. Wahl, C. Deng, M. Taketo, M. Lewandoski, and O. Pourquié. *A beta-catenin gradient links the clock and wavefront systems in mouse embryo segmentation*. Nature cell biology, 10(2):186–193, February 2008.
- [67] T. Ishitani, K. Matsumoto, A. B. Chitnis, and M. Itoh. *Nrarp functions to modulate neural-crest-cell differentiation by regulating LEF1 protein stability*. Nature cell biology, 7(11):1106–1112, November 2005.
- [68] E. M. Ozbudak and J. Lewis. *Notch signalling synchronizes the zebrafish segmentation clock but is not needed to create somite boundaries*. PLoS Genet, 4(2):e15, February 2008.
- [69] D. Wright, Z. Ferjentsik, S.-W. Chong, X. Qui, Y. Jiang, P. Malapert, O. Pourquié, N. V. Hateren, S. A. Wilson, C. Franco, H. Gerhardt, J. K. Dale, and M. Maroto. *Cyclic Nrarp mRNA expression is regulated by the somitic oscillator but Nrarp protein levels do not oscillate*. Developmental dynamics, 238:3043–3055, 2009.
- [70] R. C. Lee, R. L. Feinbaum, and V. Ambros. *The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14*. Cell, 75(5):843–854, December 1993.
- [71] S. R. Eddy. *Non-coding RNA genes and the modern RNA world*. Nat Rev Genet, 2(12):919–929, December 2001.
- [72] A. Krek, D. Grün, M. N. Poy, R. Wolf, L. Rosenberg, E. J. Epstein, P. MacMenamin, I. d. Piedade, K. C. Gunsalus, M. Stoffel, and N. Rajewsky. *Combinatorial microRNA target predictions*. Nat Genet, 37(5):495–500, May 2005.
- [73] D. Bartel. *MicroRNAs: target recognition and regulatory functions*. Cell, 136, 2009.
- [74] B. C. Bernardo, F. J. Charchar, R. C. Y. Lin, and J. R. McMullen. *A MicroRNA Guide for Clinicians and Basic Scientists: Background and Experimental Techniques*. Heart Lung Circ, 21(3):131–142, March 2012.

- 
- [75] A. F. Olena and J. G. Patton. *Genomic organization of microRNAs*. J Cell Physiol, 222(3):540–545, March 2010.
- [76] D. P. Bartel. *MicroRNAs: genomics, biogenesis, mechanism, and function*. Cell, 116(2):281–297, January 2004.
- [77] N. Rajewsky. *microRNA target predictions in animals*. Nat Genet, 38 Suppl:S8–13, June 2006.
- [78] B. P. Lewis, I.-h. Shih, M. W. Jones-Rhoades, D. P. Bartel, and C. B. Burge. *Prediction of mammalian microRNA targets*. Cell, 115(7):787–798, December 2003.
- [79] S. Yekta, I.-h. Shih, and D. P. Bartel. *MicroRNA-directed cleavage of HOXB8 mRNA*. Science, 304(5670):594–596, April 2004.
- [80] M. W. Rhoades, B. J. Reinhart, L. P. Lim, C. B. Burge, B. Bartel, and D. P. Bartel. *Prediction of plant microRNA targets*. Cell, 110(4):513–520, August 2002.
- [81] Y. Tay, J. Zhang, A. M. Thomson, B. Lim, and I. Rigoutsos. *MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation*. Nature, 455(7216):1124–1128, October 2008.
- [82] J. J. Forman, A. Legesse-Miller, and H. A. Collier. *A search for conserved sequences in coding regions reveals that the let-7 microRNA targets Dicer within its coding sequence*. PNAS, 105(39):14879–14884, September 2008.
- [83] A. Lal, H. H. Kim, K. Abdelmohsen, Y. Kuwano, R. Pullmann, S. Srikantan, R. Subrahmanyam, J. L. Martindale, X. Yang, F. Ahmed, F. Navarro, D. Dykxhoorn, J. Lieberman, and M. Gorospe. *p16(INK4a) translation suppressed by miR-24*. PLoS ONE, 3(3):e1864, 2008.
- [84] G. Easow, A. A. Teleanu, and S. M. Cohen. *Isolation of microRNA targets by miRNP immunopurification*. RNA, 13(8):1198–1204, August 2007.
- [85] D. Didiano and O. Hobert. *Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions*. Nat Struct Mol Biol, 13(9):849–851, September 2006.
- [86] D. Baek, J. Villén, C. Shin, F. D. Camargo, S. P. Gygi, and D. P. Bartel. *The impact of microRNAs on protein output*. Nature, 455(7209):64–71, September 2008.

- 
- [87] M. Selbach, B. Schwanhäusser, N. Thierfelder, Z. Fang, R. Khanin, and N. Rajewsky. *Widespread changes in protein synthesis induced by microRNAs*. Nature, 455(7209):58–63, September 2008.
- [88] J. G. Doench and P. A. Sharp. *Specificity of microRNA target selection in translational repression*. Genes Dev, 18(5):504–511, March 2004.
- [89] J. Vinther, M. M. Hedegaard, P. P. Gardner, J. S. Andersen, and P. Arctander. *Identification of miRNA targets with stable isotope labeling by amino acids in cell culture*. Nucleic Acids Res, 34(16):e107, 2006.
- [90] U. A. Ørom, F. C. Nielsen, and A. H. Lund. *MicroRNA-10a binds the 5'UTR of ribosomal protein mRNAs and enhances their translation*. Mol Cell, 30(4):460–471, May 2008.
- [91] T. Saito and P. Saetrom. *MicroRNAs—targeting and target prediction*. N Biotechnol, 27(3):243–249, July 2010.
- [92] H. Min and S. Yoon. *Got target? Computational methods for microRNA target prediction and their extension*. Exp. Mol. Med., 42(4): 233–244, April 2010.
- [93] M. Kertesz, N. Iovino, U. Unnerstall, U. Gaul, and E. Segal. *The role of site accessibility in microRNA target recognition*. Nat Genet, 39(10):1278–1284, September 2007.
- [94] D. Didiano and O. Hobert. *Molecular architecture of a miRNA-regulated 3' UTR*. RNA, 14(7):1297–1317, July 2008.
- [95] R. Heikham and R. Shankar. *Flanking region sequence information to refine microRNA target predictions*. J. Biosci., 35(1):105–118, March 2010.
- [96] M. Sturm, M. Hackenberg, D. Langenberger, and D. Frishman. *TargetSpy: a supervised machine learning approach for microRNA target prediction*. BMC Bioinformatics, 11:292, 2010.
- [97] R. C. Friedman, K. K.-H. Farh, C. B. Burge, and D. P. Bartel. *Most mammalian mRNAs are conserved targets of microRNAs*. Genome Res., 19(1):92–105, January 2009.
- [98] L. Hon. *The roles of binding site arrangement and combinatorial targeting in microRNA repression of gene expression*. Genome Biology, 8(R166), 2007.

- [99] M. Dews, A. Homayouni, D. Yu, D. Murphy, C. Seignani, E. Wentzel, E. E. Furth, W. M. Lee, G. H. Enders, J. T. Mendell, and A. Thomas-Tikhonenko. *Augmentation of tumor angiogenesis by a Myc-activated microRNA cluster*. Nat Genet, 38(9):1060–1065, September 2006.
- [100] S. van Dongen, C. Abreu-Goodger, and A. J. Enright. *Detecting microRNA binding and siRNA off-target effects from expression data*. Nat Meth, 5(12):1023–1025, November 2008.
- [101] J. C. Huang, Q. D. Morris, and B. J. Frey. *Bayesian inference of MicroRNA targets from sequence and expression data*. J Comput Biol, 14(5):550–563, June 2007.
- [102] S. Bandyopadhyay and R. Mitra. *TargetMiner: microRNA target prediction with systematic identification of tissue-specific negative examples*. Bioinformatics, 25(20):2625–2631, October 2009.
- [103] A. Stark, J. Brennecke, N. Bushati, R. B. Russell, and S. M. Cohen. *Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution*. Cell, 123(6):1133–1146, December 2005.
- [104] D. Gaidatzis, E. van Nimwegen, J. Hausser, and M. Zavolan. *Inference of miRNA targets using evolutionary conservation and pathway analysis*. BMC Bioinformatics, 8:69, 2007.
- [105] B. John, C. Sander, and D. S. Marks. *Prediction of human microRNA targets*. Methods Mol Biol, 342:101–113, 2006.
- [106] K. C. Miranda, T. Huynh, Y. Tay, Y.-S. Ang, W.-L. Tam, A. M. Thomson, B. Lim, and I. Rigoutsos. *A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes*. Cell, 126(6):1203–1217, September 2006.
- [107] B. P. Lewis, C. B. Burge, and D. P. Bartel. *Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets*. Cell, 120(1):15–20, January 2005.
- [108] V. A. Gennarino, M. Sardiello, R. Avellino, N. Meola, V. Maselli, S. Anand, L. Cutillo, A. Ballabio, and S. Banfi. *MicroRNA target prediction by expression analysis of host genes*. Genome Res., 19(3):481–490, March 2009.
- [109] S. A. Stanhope, S. Sengupta, J. den Boon, P. Ahlquist, and M. A. Newton. *Statistical use of argonaute expression and RISC assembly in microRNA target identification*. PLoS Comput Biol, 5(9):e1000516, September 2009.

- [110] W. Liebermeister. *Linear modes of gene expression determined by independent component analysis*. Bioinformatics, 18(1):51–60, 2002.
- [111] S.-I. Lee and S. Batzoglou. *Application of independent component analysis to microarrays*. Genome Biol, 4(11):R76, 2003.
- [112] A.-S. Carpentier, A. Riva, P. Tisseur, G. Didier, and A. Hénaut. *The operons, a criterion to compare the reliability of transcriptome analysis tools: ICA is more reliable than ANOVA, PLS and PCA*. Comput Biol Chem, 28(1):3–10, February 2004.
- [113] A. E. Teschendorff, M. Journée, P. A. Absil, R. Sepulchre, and C. Caldas. *Elucidating the altered transcriptional programs in breast cancer using independent component analysis*. PLoS Comput Biol, 3(8):e161, August 2007.
- [114] P. Comon. *Independent component analysis, A new concept?* Signal Processing, 36:287–314, 1994.
- [115] E. M. C. Ohlsson Teague, K. H. Van der Hoek, M. B. Van der Hoek, N. Perry, P. Wagaarachchi, S. A. Robertson, C. G. Print, and L. M. Hull. *MicroRNA-regulated pathways associated with endometriosis*. Mol Endocrinol, 23(2):265–275, February 2009.
- [116] A. Frigyesi, S. Veerla, D. Lindgren, and M. Höglund. *Independent component analysis reveals new and biologically significant structures in micro array data*. BMC Bioinformatics, 7:290, 2006.
- [117] W. Kong, X. Mou, and X. Hu. *Exploring matrix factorization techniques for significant genes identification of Alzheimer’s disease microarray gene expression data*. BMC Bioinformatics, 12 Suppl 5:S7, 2011.
- [118] X. W. Zhang, Y. L. Yap, D. Wei, F. Chen, and A. Danchin. *Molecular diagnosis of human cancer type by gene expression profiles and independent component analysis*. Eur. J. Hum. Genet., 13(12):1303–1311, December 2005.
- [119] D.-S. Huang and C.-H. Zheng. *Independent component analysis-based penalized discriminant method for tumor classification using gene expression data*. Bioinformatics, 22(15):1855–1862, August 2006.
- [120] O. Winther and K. B. Petersen. *Flexible and efficient implementations of Bayesian independent component analysis*. Neurocomputing, 71: 221–233, 2007.
- [121] P. Chiappetta, M. C. Roubaud, and B. Torrèsani. *Blind source separation and the analysis of microarray data*. J Comput Biol, 11(6): 1090–1109, 2004.



- 
- [122] S. Vasudevan, Y. Tong, and J. A. Steitz. *Switching from repression to activation: microRNAs can up-regulate translation*. Science, 318 (5858):1931–1934, December 2007.
- [123] J.-F. Cardoso. *High-Order Contrasts for Independent Component Analysis*. Neural Computation, 11:157–192, 1999.
- [124] E. G. Miller and J. W. Fisher. Independent Components Analysis by Direct Entropy Minimization. Technical Report UCB/CSD-3-1221, Computer Science Division, University of California, 2003.
- [125] A. Hyvärinen. *Fast and Robust Fixed-Point algorithms for Independent Component Analysis*. IEEE Trans. on Neural Networks, 10(3): 626–634, 1999.
- [126] C.-H. Zheng, Y. Chen, X.-X. Li, Y.-X. Li, and Y.-P. Zhu. *Tumor classification based on independent component analysis*. International Journal of Pattern Recognition, 20(2):297–310, 2006.
- [127] W.-C. Chang. *On Using Principal Components Before Separating a Mixture of Two Multivariate Normal Distributions*. Appl Statist, 32 (3):267–275, 1983.
- [128] K. Y. Yeung and W. L. Ruzzo. *Principal component analysis for clustering gene expression data*. Bioinformatics, 17(9):763–774, September 2001.
- [129] M. Kanehisa, S. Goto, and M. Furumichi. *KEGG for representation and analysis of molecular networks involving diseases and drugs*. Nucleic Acids Res, 38, 2010.
- [130] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. *Gene Ontology: tool for the unification of biology*. Nat Genet, 25:25–29, 2000.
- [131] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, and J. P. Mesirov. *Molecular signatures database (MSigDB) 3.0*. Bioinformatics, 27(12):1739–1740, June 2011.
- [132] E. Widmaier, H. Raff, K. Strang, and A. Vander. *Vander, Sherman, and Luciano’s human physiology*. McGraw-Hill Higher Education, 2004. ISBN 9780072834017.
- [133] H. Wang and C. B. Wollheim. *Pdx1 level defines pancreatic gene expression pattern and cell lineage differentiation*. J Biol Chem, 267 (27):25279–25286, 2001.

- [134] T. Mandrup-Poulsen, S. Helqvist, J. Mølvig, L. D. Wogensen, and J. Nerup. *Cytokines as immune effector molecules in autoimmune endocrine diseases with special reference to insulin-dependent diabetes mellitus*. *Autoimmunity*, 4(3):191–218, 1989.
- [135] R. Bergholdt, C. Brorsson, A. Palleja, L. A. Berchtold, T. Fløyel, C. H. Bang-Berthelsen, K. S. Frederiksen, L. J. Jensen, J. Størling, and F. Pociot. *Identification of Novel Type 1 Diabetes Candidate Genes by Integrating Genome-Wide Association Data, Protein-Protein Interactions, and Human Pancreatic Islet Gene Expression*. *Diabetes*, February 2012.
- [136] A. Cardozo, M. Kruhøffer, R. Leeman, and T. Ørntoft. *Identification of Novel Cytokine-Induced Genes in Pancreatic beta-Cells by High-Density Oligonucleotide Arrays*. *Diabetes*, 50, 2001.
- [137] B. Kutlu, A. Cardozo, and M. Darville. *Discovery of Gene Networks Regulating Cytokine-Induced Dysfunction and Apoptosis in Insulin-Producing INS-1 Cells*. *Diabetes*, 52, 2003.
- [138] V. Mootha, C. Lindgren, and K. Eriksson. *PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes*. *Nat Genet*, 34(3), 2003.
- [139] M. V. Joglekar, V. M. Joglekar, and A. A. Hardikar. *Expression of islet-specific microRNAs during human pancreatic development*. *Gene expression patterns*, 9(2):109–113, February 2009.
- [140] L. Zhang, D. Hou, X. Chen, D. Li, L. Zhu, Y. Zhang, J. Li, Z. Bian, X. Liang, X. Cai, Y. Yin, C. Wang, T. Zhang, D. Zhu, D. Zhang, J. Xu, Q. Chen, Y. Ba, J. Liu, Q. Wang, J. Chen, J. Wang, M. Wang, Q. Zhang, J. Zhang, K. Zen, and C.-Y. Zhang. *Exogenous plant MIR168a specifically targets mammalian LDLRAP1: evidence of cross-kingdom regulation by microRNA*. *Cell Res.*, 22(1):107–126, January 2012.
- [141] C. M. Filippi and M. G. von Herrath. *Viral trigger for type 1 diabetes: pros and cons*. *Diabetes*, 57(11):2863–2871, November 2008.
- [142] L. C. Stene, S. Oikarinen, H. Hyöty, K. J. Barriga, J. M. Norris, G. Klingensmith, J. C. Hutton, H. A. Erlich, G. S. Eisenbarth, and M. Rewers. *Enterovirus infection and progression from islet autoimmunity to type 1 diabetes: the Diabetes and Autoimmunity Study in the Young (DAISY)*. *Diabetes*, 59(12):3174–3180, December 2010.
- [143] X. Cai, A. Schäfer, S. Lu, J. P. Bilello, R. C. Desrosiers, R. Edwards, N. Raab-Traub, and B. R. Cullen. *Epstein-Barr virus microRNAs are*

- evolutionarily conserved and differentially expressed*. PLoS Pathog., 2(3):e23, March 2006.
- [144] M. S. Horwitz, L. M. Bradley, J. Harbertson, T. Krah, J. Lee, and N. Sarvetnick. *Diabetes induced by Coxsackie virus: initiation by bystander damage and not molecular mimicry*. Nature Medicine, 4(7): 781–785, July 1998.
- [145] F. Dotta, S. Censini, A. G. S. van Halteren, L. Marselli, M. Masini, S. Dionisi, F. Mosca, U. Boggi, A. O. Muda, S. D. Prato, J. F. Elliott, A. Covacci, R. Rappuoli, B. O. Roep, and P. Marchetti. *Coxsackie B4 virus infection of beta cells and natural killer cell insulitis in recent-onset type 1 diabetic patients*. Proc Natl Acad Sci USA, 104(12): 5115–5120, March 2007.
- [146] R. Siegel, D. Naishadham, and A. Jemal. *Cancer statistics, 2012*. CA Cancer J Clin, 62(1):10–29, January 2012.
- [147] D. C. Corney and A. Y. Nikitin. *MicroRNA and ovarian cancer*. Histol Histopathol, 23(9):1161–1169, September 2008.
- [148] F. Toledo and G. M. Wahl. *Regulating the p53 pathway: in vitro hypotheses, in vivo veritas*. Nat Rev Cancer, 6(12):909–923, December 2006.
- [149] Cancer Genome Atlas Research Network. *Integrated genomic analyses of ovarian carcinoma*. Nature, 474(7353):609–615, June 2011.
- [150] C. J. Creighton, M. D. Fountain, Z. Yu, A. K. Nagaraja, H. Zhu, M. Khan, E. Olokpa, A. Zariff, P. H. Gunaratne, M. M. Matzuk, and M. L. Anderson. *Molecular Profiling Uncovers a p53-Associated Role for MicroRNA-31 in Inhibiting the Proliferation of Serous Ovarian Carcinomas and Other Cancers*. Cancer Res, 70(5):1906–1915, 2010.
- [151] A. Sorrentino, C.-g. Liu, A. Addario, C. Peschle, G. Scambia, and C. Ferlini. *Role of microRNAs in drug-resistant ovarian cancer cells*. Gynecologic Oncology, 111(3):478–486, December 2008.
- [152] A. K. Nagaraja, C. J. Creighton, Z. Yu, H. Zhu, P. H. Gunaratne, J. G. Reid, E. Olokpa, H. Itamochi, N. T. Ueno, S. M. Hawkins, M. L. Anderson, and M. M. Matzuk. *A link between mir-100 and FRAP1/mTOR in clear cell ovarian cancer*. Mol Endocrinol, 24(2): 447–463, February 2010.
- [153] S. M. Hawkins, C. J. Creighton, D. Y. Han, A. Zariff, M. L. Anderson, P. H. Gunaratne, and M. M. Matzuk. *Functional MicroRNA Involved in Endometriosis*. Mol Endocrinol, 25(5):821–832, April 2011.

- [154] S. Wang and S. Zhang. *Dickkopf-1 is frequently overexpressed in ovarian serous carcinoma and involved in tumor invasion*. Clin. Exp. Metastasis, 28(6):581–591, August 2011.
- [155] E. R. Flores. *The roles of p63 in cancer*. Cell Cycle, 6(3):300–304, February 2007.
- [156] C. J. Creighton, A. Hernandez-Herrera, A. Jacobsen, D. A. Levine, P. Mankoo, N. Schultz, Y. Du, Y. Zhang, E. Larsson, R. Sheridan, W. Xiao, P. T. Spellman, G. Getz, D. A. Wheeler, C. M. Perou, R. A. Gibbs, C. Sander, D. N. Hayes, P. H. Gunaratne, and T. C. G. A. R. Network. Integrated analyses of microRNAs uncover tumor suppressor candidates in high-grade serous ovarian carcinoma. Under revision in PLoS ONE, 2012.
- [157] U. Ralfkiaer, P. H. Hagedorn, N. Bangsgaard, M. B. Løvendorf, C. B. Ahler, L. Svensson, K. L. Kopp, M. T. Vennegaard, B. Lauenborg, J. R. Zibert, T. Krejsgaard, C. M. Bonefeld, R. Søkilde, L. M. Gjerdrum, T. Labuda, A.-M. Mathiesen, K. Grønbaek, M. A. Wasik, M. Sokolowska-Wojdylo, C. Queille-Roussel, R. Gniadecki, E. Ralfkiaer, C. Geisler, T. Litman, A. Woetmann, C. Glue, M. A. Røpke, L. Skov, and N. Odum. *Diagnostic microRNA profiling in cutaneous T-cell lymphoma (CTCL)*. Blood, 118(22):5891–5900, November 2011.
- [158] Y. Huang, A. Chuang, H. Hao, C. Talbot, T. Sen, B. Trink, D. Sidransky, and E. Ratovitski. *Phospho-DNp63a is a key regulator of the cisplatin-induced microRNAome in cancer cells*. Cell Death Differ., 18(7):1220–1230, July 2011.
- [159] J. Wen, B. J. Parker, A. Jacobsen, and A. Krogh. *MicroRNA transfection and AGO-bound CLIP-seq data sets reveal distinct determinants of miRNA action*. RNA, 17(5):820–834, May 2011.
- [160] G. Glass and P. Peckham. *Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance*. Review of educational research, 42(3):237–288, 1972.
- [161] Y. Benjamini and Y. Hochberg. *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing*. Journal of the Royal Statistical Society. Series B (Methodological), 57(1):289–300, 1995.



## APPENDIX

### A

## Basic statistics

This appendix describes some basic statistical tools that are used within Part II.

### A.1 ANOVA - analysis of variance

The ANOVA is a hypothesis testing method based on general linear models. Assume that observations are done for samples which are classified according to some factor. In the simple case of classification according to two factors, then ANOVA simplifies to a  $t$ -test. When more factors are present then ANOVA is more desirable than multiple  $t$ -tests, due to the problem of multiple testing explained below. ANOVA is based on the comparison of the means within different classes of samples and especially the variance between these mean values. In general the measurements within a group of samples need to be Normal distributed, however away from normality an ANOVA can still be performed if the symmetry of the distribution is preserved [160].

### A.2 Coefficient of determination

The variance explained by fitting a regression model to data can be measured by the coefficient of determination. It is defined as the ratio between the sum of squares due to the regression model and the total sum of squares. Thus, the definition is

$$R^2 = \frac{\sum_j (\hat{y}_j - \bar{y})^2}{\sum_j (y_j - \bar{y})^2} , \quad (\text{A.1})$$

for a linear regression model  $y_j = \beta_0 + \sum_i \beta_i x_{ji} + \epsilon_i$ . Here  $\hat{y}_j$  is the predicted value,  $\bar{y}$  is the mean of the observations and  $\epsilon_i$  is an additive noise.

### A.3 Sensitivity and specificity

In a binary test it is common to operate with the terminology true-positive (TP), false-positive (FP), false-negative (FN) and true-negative (TN). Table A.A shows how this terminology can be transferred to miRNA target prediction. The aim is to obtain no false-positives and no false-negatives. To measure whether this is obtained the sensitivity and specificity are often calculated. The specificity/sensitivity of a prediction method is simply the probability that the test is positive/negative when the mRNA is actually validated as a target/non-target, i.e.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad \text{and} \quad \text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} . \quad (\text{A.2})$$

If the method does not include any variable parameters, then the sensitivity and specificity are just single numbers. If, on the other hand, the method contains tunable variables then it is common to plot the true-positive rate (sensitivity) and the false-positive rate (1-specificity) against each other in a receiver operating characteristic curve (ROC). Based on ROC the area under the curve can be calculated (AUC). A higher AUC value corresponds to a better performance of the method, because then it is possible to obtain a high number of true-positives without consequently increasing the number of false-positives.

**Table A.A** Contingency table for results of a miRNA target prediction method

		Test outcome	
		positive	negative
Experimental validation	target	TP	FN
	non-target	FP	TN

### A.4 Correction for multiple testing

The statistical tests performed in this thesis returns a  $p$ -value indicating the significance of the test. A  $p$ -value of 0.01 indicates that there is a 1% chance that the rejected null hypothesis is true. When analyzing 100 genes this means that 1 gene is expected to be a false-positive, i.e. a gene based on the  $p$ -value is differentially expressed but is not. With a larger sample size there is a problem of multiple testing. For example, when testing 10,000

genes, then 100 out of these will be false-positives. A correction of multiple testing is needed.

The False Discovery Rate (FDR) [161] is widely used in the analyses of gene expressions. The FDR gives the expected rate of false-positives, i.e. the fraction of falsely rejected null hypotheses. There is a tradition of labeling FDR corrected  $p$ -values with a  $q$  instead of a  $p$ , i.e.  $q < 0.05$  means that the FDR is less than 0.05.

The Bonferroni adjustment controls the Familywise Error Rate (FWER). The FWER is the probability of one false-positive among all hypotheses. Thus, the FWER is a more conservative error rate than the FDR. In this thesis the FDR correction is used when many tests are performed and the FWER is used for corrections of fewer tests.





APPENDIX

B

AXIN2 paper

## A Wnt Oscillator Model for Somitogenesis

Peter B. Jensen, Lykke Pedersen, Sandeep Krishna, and Mogens H. Jensen\*

Niels Bohr Institute, Copenhagen, Denmark

**ABSTRACT** We propose a model for the segmentation clock in vertebrate somitogenesis, based on the Wnt signaling pathway. The core of the model is a negative feedback loop centered around the Axin2 protein. Axin2 is activated by  $\beta$ -catenin, which in turn is degraded by a complex of GSK3 $\beta$  and Axin2. The model produces oscillatory states of the involved constituents with typical time periods of a few hours (ultradian oscillations). The oscillations are robust to changes in parameter values and are often spiky, where low concentration values of  $\beta$ -catenin are interrupted by sharp peaks. Necessary for the oscillations is the saturated degradation of Axin2. Somite formation in chick and mouse embryos is controlled by a spatial Wnt gradient which we introduce in the model through a time-dependent decrease in Wnt3a ligand level. We find that the oscillations disappear as the ligand concentration decreases, in agreement with observations on embryos.

### INTRODUCTION

Oscillations are ubiquitous in biological systems. Circadian (24 h) rhythms are well known, but recently ultradian (1–4 h) oscillations have been observed in the expression of genes involved in the immune system, programmed cell death, and embryo development (1,2). We focus on the latter, where the oscillations have an obvious physiological function: the segmentation clock provided by the oscillations is responsible for the periodic spacing of somites, structures that eventually become the vertebrae. We have made a model of the molecular network of genes in the Wnt signaling pathway, based on known experimental data. The model produces oscillations of the observed frequency and clarifies the essential ingredients required for such oscillations. In particular, our model shows how the oscillations can be stopped by a decrease in the Wnt ligand concentration. This is important because experiments have shown that a higher concentration of Wnt ligand demarcates the region where the somites form in the embryo (3).

The embryological process of somitogenesis in vertebrates is the rhythmic formation of vertebrae precursors known as somites in the anterior presomitic mesoderm tissue (PSM). As the embryo elongates in the posterior direction, adding new cells to the posterior PSM, the maturing anterior PSM sequentially buds off pairs of cell clusters—the somites. In mice, a new pair of somites forms approximately every 120 min (4). The strict periodicity of the process suggests the involvement of a cellular clock (the segmentation clock), the elucidation of which has been the focus of many studies. Underlying most models is the idea of locally coupled intracellular clocks controlled by a morphogen gradient in the PSM. This can supply cells with both the temporal information (cycle state) and spatial information (axial position) necessary to form distinct somites at the right

time and place. This general idea is known as the clock-and-gradient, or clock-and-wavefront model, originally proposed by Cooke and Zeeman (5).

Although species-dependent, the key to the clock operation seems to lie in the action and interaction of a number of cellular pathways, most notably Notch and Wnt. A number of oscillating Notch target genes has been identified and possible feedback loops have been proposed (reviewed in (6)). In particular, attention has been given to the autoinhibiting *hes/her* genes (7,8) and the glycosyltransferase *lfng* believed to alter Notch ligand susceptibility (9,10). In cells in the posterior two-thirds of the PSM, these genes demonstrate oscillation frequencies matching the somite segmentation frequency in different animals (120 min in mice, 90 min in chick embryos, i.e., ultradian periods).

Underlying the oscillations there must necessarily be a negative feedback loop (11). Most of the mathematical models of somite formation so far have focused on delay-driven negative feedback loops in Notch (12–15), or larger Wnt-Notch models operating with Dsh as a direct activator of *axin2* transcription and inhibitor of *hes* transcription (16). This potential role for Dsh is still undocumented experimentally in vertebrates, and circumvents essential parts of canonical Wnt signaling. In 2003, Aulehla et al. (2) discovered that *axin2*, a target gene of the canonical Wnt signaling pathway, displays a 120-min cyclic expression in mice completely out of phase with the Notch target genes. Axin2 is a known inhibitor of its own transcription, thus forming a negative feedback loop, and Aulehla et al. (2) suggests this loop in the Wnt pathway as the driving force behind the segmentation clock.

Here, we present a mathematical model based on this idea, where the Axin2 feedback loop is closed by its forming a complex with GSK3 $\beta$  and  $\beta$ -catenin (the destruction complex), which results in  $\beta$ -catenin degradation. We have determined most parameter values of the model from experimental data found in the literature, and have explored a range of values for those parameters we could not fix. In the

Submitted September 11, 2009, and accepted for publication November 18, 2009.

\*Correspondence: [mhjensen@nbi.dk](mailto:mhjensen@nbi.dk)

Editor: Herbert Levine.

© 2010 by the Biophysical Society  
0006-3495/10/03/0943/8 \$2.00

doi: 10.1016/j.bpj.2009.11.039

particular case where the complex of Axin2 and LRP5/6 coreceptor at the membrane is in quasiequilibrium, our model reduces to a form very similar to that described in Goldbeter and Pourquié (17). The association and dissociation rates of this complex have not been measured, so it is difficult to say whether the quasiequilibrium assumption is a good one in vivo. Our model allows us to investigate the robustness of the oscillations as one moves away from quasiequilibrium to slower association/dissociation of the complex. In addition, it clarifies how the nonlinear degradation terms used in the simpler model of Goldbeter and Pourquié (17) arise from the interaction of Axin2 with the LRP5/6 coreceptor. Lastly, as our model is less coarse-grained, the parameters represent specific physical processes and therefore several of them can be directly ascertained from experimental data. The overall goal of our model is to capture the oscillatory dynamics of key Wnt target genes in individual cells in the posterior two-thirds of the PSM.

## THE MODEL

### A Wnt oscillator

The key to canonical Wnt signaling (reviewed in Logan and Nusse (18)) is  $\beta$ -catenin. Accumulated  $\beta$ -catenin acts as a part of a transcription complex in the nucleus leading to transcription of Wnt target genes. The regulation of  $\beta$ -catenin levels in the cell is under the control of the destruction complex, which is composed of a large number of proteins. Two key players, whose role is well understood, are the Glycogen Synthase Kinase 3 $\beta$  (GSK3 $\beta$ ) and the scaffolding protein Axin. Phosphorylation of  $\beta$ -catenin by GSK3 $\beta$  leads to the degradation of  $\beta$ -catenin. Axin is known to boost this phosphorylation by several thousand-fold (19). Interestingly, whereas Wnt/ $\beta$ -catenin does not regulate transcription of Axin, it does regulate transcription of Axin2 (a homolog of Axin). Along with the oscillatory properties of Axin2 in the PSM tissue, this suggests that Axin2 is a likely component of the negative feedback loop driving oscillations.

Fig. 1 shows the selected key components of our mathematical model of the Wnt feedback network. Wnt signaling is initiated by the binding of a Wnt ligand to the Frizzled/LRP receptor-coreceptor complex on the cell membrane (20,21). How the signal is then mediated to the destruction complex and  $\beta$ -catenin regulation is still not clear, but it is known to involve the protein Disheveled (Dsh) (22). Dsh also binds to the Frizzled receptor, and Wnt activation recruits Axin to the LRP5/6 coreceptor (23). Additionally, Axin and Dsh can bind together and have been demonstrated to colocalize at the membrane (24). This sequestering and possible degradation of Axin at the membrane could lead to a decrease in the concentration of the destruction complex, leading to  $\beta$ -catenin accumulation. We have used this scenario in the model. Axin and Axin2 are functionally equivalent (25), thus Axin2 should be able to fill the role

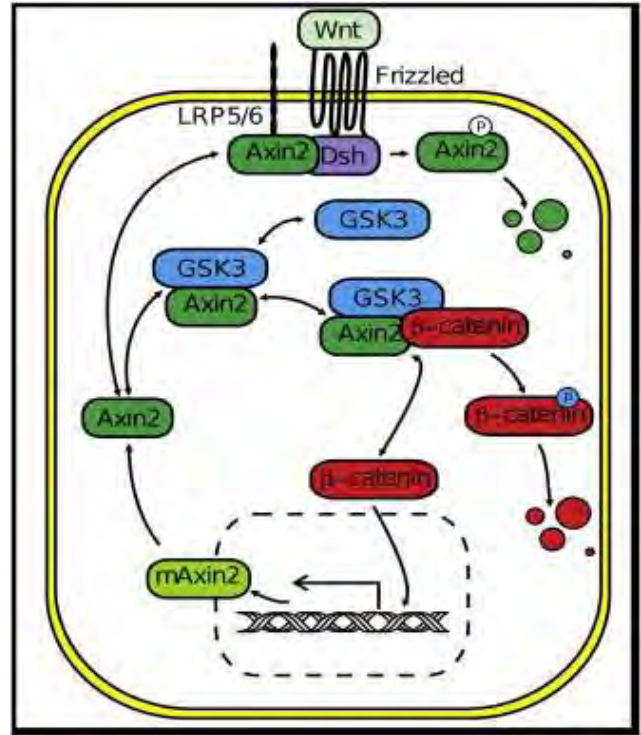


FIGURE 1 Illustration of the processes included in our model of the Wnt system. The core is a negative feedback loop:  $\beta$ -catenin (red) activates production of Axin2 (green) which, via a complex involving GSK3 $\beta$  (cyan), results in degradation of  $\beta$ -catenin. The pathway is controlled by Wnt ligands which, when bound to a receptor-coreceptor complex, catalyze the degradation of Axin2.

as scaffolding protein in the destruction complex and as binding partner to Dsh and LRP5/6.

In the model, we also include the full destruction complex (composed of Axin2,  $\beta$ -catenin and GSK3 $\beta$ ) and one of potentially three subcomplexes (Axin2 bound to GSK3 $\beta$ ). Other destruction complex components such as the priming kinase CK1 $\alpha$  and the other scaffolding protein APC do not play a role in the feedback loop and are therefore absorbed into the parameters. Dsh is similarly absorbed into parameters describing Axin2 binding and degradation at the receptor complex.  $\beta$ -catenin is present in a free form and as part of the destruction complex. Finally, there is GSK3 $\beta$ , which is present in free forms and bound forms in the destruction complex and subcomplex.

The processes shown in Fig. 1 are modeled using standard reaction kinetics, treating all complex formation reactions as reversible processes. This leads to a system of eight coupled ordinary differential equations,

$$\frac{dC}{dt} = c_{fC}B[GA] - c_{bC}C - \alpha C, \quad (1)$$

$$\frac{d[GA]}{dt} = c_{f[GA]}GA - c_{b[GA]}[GA] - c_{fC}B[GA] + c_{bC}C + \alpha C, \quad (2)$$

$$\frac{dB}{dt} = S - c_{fC}B[GA] + c_{bC}C, \quad (3)$$

$$\frac{dG}{dt} = -c_{f[GA]}GA + c_{b[GA]}[GA], \quad (4)$$

$$\begin{aligned} \frac{dA}{dt} = & -c_{f[GA]}GA + c_{b[GA]}[GA] + c_{tA}A_m - c_{f[AL]}AL \\ & + c_{b[AL]}[AL], \end{aligned} \quad (5)$$

$$\frac{dA_m}{dt} = c_{tsA}B^2 - \frac{A_m}{\tau_{Am}}, \quad (6)$$

$$\frac{d[AL]}{dt} = c_{f[AL]}AL - c_{b[AL]}[AL] - \nu[AL], \quad (7)$$

$$\frac{dL}{dt} = -c_{f[AL]}AL + c_{b[AL]}[AL] + \nu[AL], \quad (8)$$

where  $C$ ,  $[GA]$ ,  $B$ ,  $G$ ,  $A$ ,  $A_m$ , and  $[AL]$  are, respectively, concentrations of the destruction complex, GSK3 $\beta$ -Axin2 complex,  $\beta$ -catenin, GSK3 $\beta$ , Axin2, Axin2 mRNA, and the Axin2-LRP complex.

Subscripts of parameters ( $c$ ) are named according to the following system: A subscripted  $f$  denotes complex formation followed by the name of the complex being formed. Correspondingly, a subscripted  $b$  denotes dissociation of a complex.  $tA$  and  $tsA$  denote translation and transcription of Axin2.

Equations 1 and 2 model the breaking and formation of the destruction complex  $C$  and the subcomplex  $[GA]$ . The  $\alpha$ -term models destruction of  $C$  due to phosphorylation of  $\beta$ -catenin, whereas the  $c_{bC}$ -term is the spontaneous dissociation of the complex into  $[GA]$  and  $B$  (unphosphorylated  $\beta$ -catenin). Equation 3 deals with the concentration of free  $\beta$ -catenin.  $S$  represents constitutive production of  $\beta$ -catenin. Because free  $\beta$ -catenin has been shown to be extremely stable (26), no degradation term is included in Eq. 3. GSK3 $\beta$  is extremely stable and its total concentration is assumed to be fixed on the timescales involved (26); therefore, no source or sink terms are included in Eq. 4.

The concentrations of Axin2 and Axin2 mRNA are determined, respectively, by Eqs. 5 and 6. Translation is directly proportional to the amount of Axin2 mRNA transcript. In the activation of transcription by  $\beta$ -catenin, we use a Hill coefficient of 2, that is, we assume some cooperativity. In the absence of cooperativity, i.e., for a Hill coefficient of 1, there are no sustained oscillations, whereas higher values increase the region of parameter space where oscillations are found, but otherwise do not change our results. As there is some evidence for cooperativity in the action of  $\beta$ -catenin (27,28), using a value of 2 for the Hill coefficient seems reasonable.

$\tau_{Am}$  is the average lifetime of the Axin2 mRNA. Finally, Eqs. 7 and 8 deal with the binding of Axin2 to the LRP5/6

coreceptor and its subsequent degradation (the  $\nu$ -term). As with GSK3 $\beta$ , we assume a constant total concentration of LRP5/6. Note that activation of Wnt signaling, when wnt3a ligand binds to the receptor, is exclusively mediated down the pathway through the interaction of the LRP5/6 coreceptor with Axin2, and wnt3a ligand levels are therefore absorbed into the parameter  $\nu$ . Thus, keeping  $\nu$  constant corresponds to a constant level of Wnt signaling.

### Default parameter values

The literature contains limited quantitative measurements of concentrations and kinetics for the Wnt components in mouse PSM tissue. We therefore use results from general experiments on Wnt signaling. The primary source of parameter values is Lee et al. (26), where a number of component concentrations and dissociation constants are measured and calculated in *Xenopus* (frog) extracts, which has the same canonical Wnt signaling pathway as in mice. These include dissociation constants for the destruction complex and its primary subcomplex:

$$D_C = \frac{c_{bC}}{c_{fC}}$$

and

$$D_{[GA]} = \frac{c_{b[GA]}}{c_{f[GA]}}.$$

Lee et al. (26) also supplies the total concentration of GSK3 $\beta$  and the source of  $\beta$ -catenin synthesis and allows for a good estimation of the  $\alpha$ -value. The measurements in Lee et al. (26) are done for Axin, but as Axin2 is functionally equivalent to Axin (25), we assume it has the same binding affinities.

Table 1 shows the 14 parameters in the model, together with their estimated values and units. These default parameter values define a reference state.

The first six parameter values are from Lee et al. (26). From the values of  $D_{[GA]}$  and  $D_C$  given in Lee et al. (26),

**TABLE 1** Parameters in the eight-variable model of the Wnt system and their default values

Parameter	Process	Default value
$c_{fC}$	Binding of B to [GA] to form destruction complex C	0.1 nM <sup>-1</sup> min <sup>-1</sup>
$c_{bC}$	Dissociation of C into B and [GA]	7 min <sup>-1</sup>
$\alpha$	Dissociation of C due to destruction of $\beta$ -catenin	200 min <sup>-1</sup>
$c_{f[GA]}$	Binding of G to A to form [GA]	0.2 nM <sup>-1</sup> min <sup>-1</sup>
$c_{b[GA]}$	Dissociation of [GA] into G and A	1.2 min <sup>-1</sup>
$S$	Constant source of $\beta$ -catenin	0.4 nM min <sup>-1</sup>
$c_{f[AL]}$	Binding of A to L	10 nM <sup>-1</sup> min <sup>-1</sup>
$c_{b[AL]}$	Dissociation of [AL] into A and L	0.08 min <sup>-1</sup>
$c_{tsA}$	Transcription of <i>axin2</i> gene	0.7 nM <sup>-1</sup> min <sup>-1</sup>
$c_{tA}$	Translation of Axin2 mRNA	0.7 min <sup>-1</sup>
$\tau_{Am}$	Average lifetime of Axin2 mRNA	40 min
$\nu$	Degradation of Axin2 in [AL] complex	0.1 min <sup>-1</sup>
$GSK3\beta_{tot}$	Total G level	50 nM
$L_{tot}$	Total L level	70 nM

the values of  $c_{fC}$ ,  $c_{bC}$ ,  $c_{f[GA]}$ , and  $c_{b[GA]}$  are estimated. mRNA half-lives are known to range from a few minutes to several hours, thus the half-life of mRNA Axin2  $\tau_{Am}$  is set to 40 min. Looking at Eqs. 5 and 6 it is evident that  $c_{tIA}$  and  $c_{tSA}$  have a similar effect; the behavior of the system depends only on their product. Therefore, we simply assign them equal values. The constants concerned with the degradation of Axin2 are also unknown. The values in Table 1 were tuned to produce oscillations with an ultradian period of  $\sim 120$  min. The table also gives the total amounts of GSK3 $\beta$  and LRP5/6 in the reference state, which are needed as a boundary condition for the system, as neither source nor sink are included for these variables in our model.

## RESULTS

### Oscillatory behavior with default parameters

Fig. 2 shows a timeseries of the behavior of the eight variables of the Wnt system for default parameters (Table 1). The system is clearly oscillatory in all variables with a period close to 120 min. All variables are initially set to zero, except for GSK3 $\beta$  and LRP5/6 which are set to 50 nM and 70 nM, respectively.

The phases of the individual variables appear according to their relative positions in the negative feedback loop (as expected (11)). Starting with, for example, the onset of a  $\beta$ -catenin peak we observe Axin2 mRNA levels and then protein levels rise with a slight delay. The newly produced Axin2 then binds to the free GSK3 $\beta$  to form the [GA] complex. GSK3 $\beta$  and [GA] have mirror image profiles. As soon as [GA] levels go up, destruction complexes C will also form, causing a depletion of  $\beta$ -catenin. The kinetics of the destruction complex is extremely fast ( $\alpha = 200 \text{ min}^{-1}$ ). Hence any formed C breaks up very quickly with a resulting loss of  $\beta$ -catenin in the process.

### Robustness of the oscillations

Although the default parameters produce oscillations that are qualitatively similar to those observed in Wnt target genes and have the correct time period, it is worth examining what happens as the parameters are varied. We find in

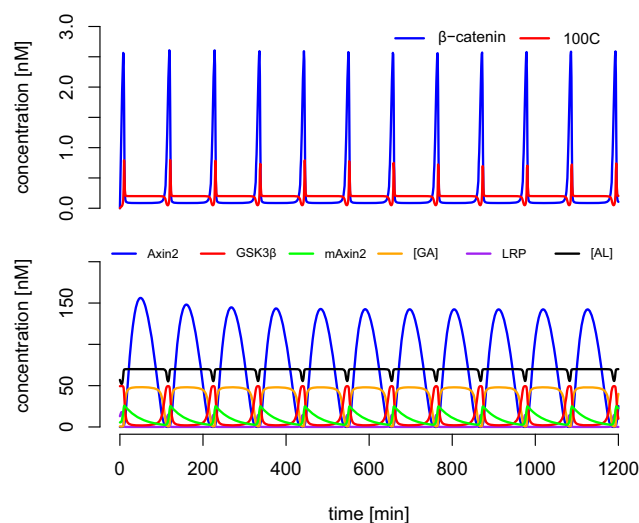


FIGURE 2 Oscillations of the Wnt model for default parameter values (see Table 1): the reference state. (Upper panel)  $\beta$ -catenin and C concentrations (the latter is multiplied by 100 for visibility). (Lower panel) Axin2, GSK3 $\beta$ , mAxin2, and [GA] concentrations, as a function of time.

general that the oscillations are quite robust to changes in parameter values. Their shape remains qualitatively the same, with short spikes in  $\beta$ -catenin-separating periods, where its level is close to zero. The period of the oscillations is quite sensitive to changes in some of the parameters, e.g.,  $D_C$ ,  $D_{[GA]}$ , and  $\tau_{Am}$ —although being almost independent of other parameters, e.g.,  $\alpha$  (see the Supporting Material).

Fig. 3 shows a plot of the amplitude and the oscillation period as a function of  $D_C$  and  $D_{[GA]}$ . These two parameters affect the formation and dissociation of the destruction complex and the [GA] complex. Robustness to other parameters is described in the Supporting Material. The reference state is indicated on the plot, as well as four other parameter sets. Timeseries corresponding to the latter are shown in Fig. 4.

A mathematical analysis of the oscillatory properties of this system is presented in the Supporting Material. There we show that the model equations allow only one physically feasible steady state, for any choice of parameter values. Whether the system shows sustained oscillations at given

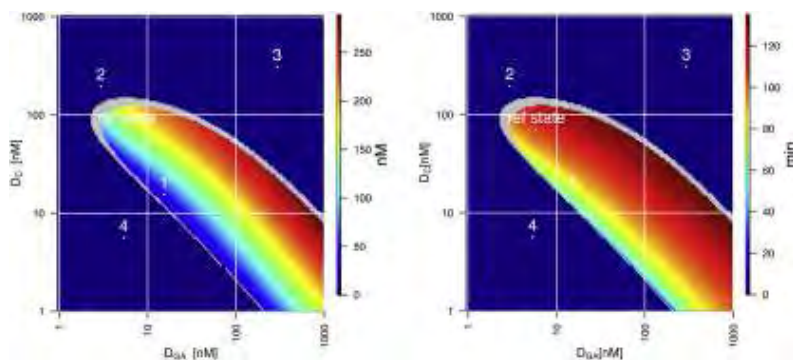


FIGURE 3 The  $D_C$  and  $D_{[GA]}$  parameter plane. (Left panel) The amplitude of the Axin2 oscillations. (Right panel) Oscillation period of Axin2. The shaded borderline indicates the boundary of sustained oscillations, derived from a stability analysis of the system. The reference state (see Fig. 2) is indicated by the dot labeled “ref state”. Timeseries corresponding to the four numbered dots are shown in Fig. 4.



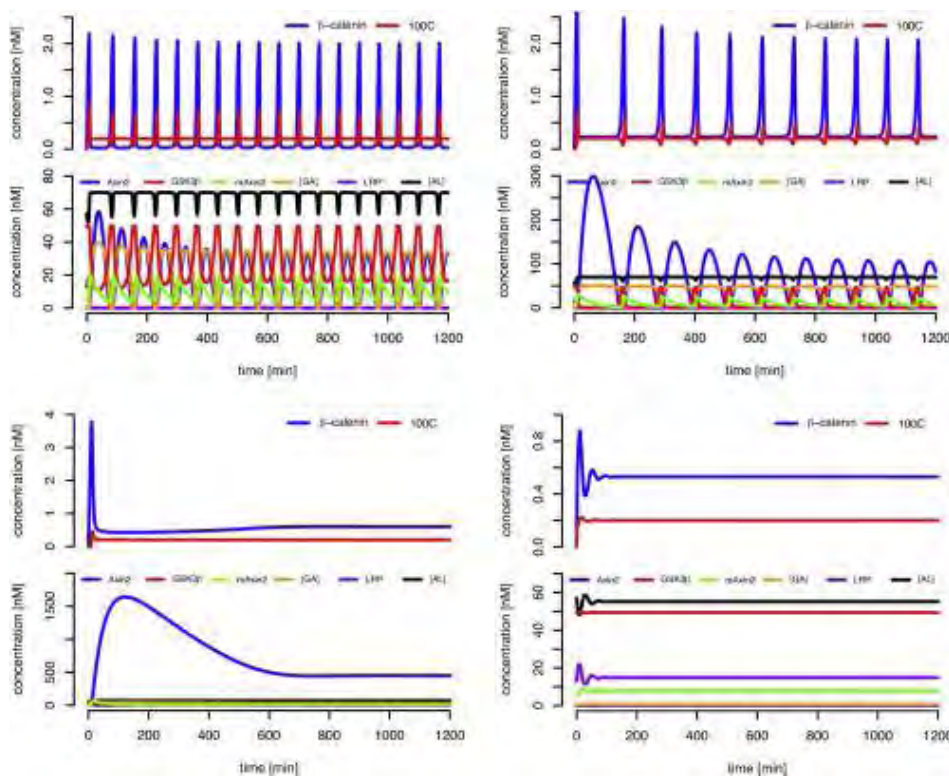


FIGURE 4 Four selected timeseries corresponding to the four numbered dots in Fig. 3. (Top left) State 1 shows sustained oscillations with a time period at approximately two-thirds of the reference state. (Top right) State 2 has very slowly damped oscillations (in practice, such oscillations would be indistinguishable from sustained oscillations in an embryo). (Bottom left and right) States 3 and 4 showing clearly damped oscillations.

parameter values can then be determined by examining the linear stability of the steady state via the eigenvalues of the Jacobian. This is explained in more detail in the [Supporting Material](#). The shaded contour in Fig. 3 shows the boundary of the oscillatory region as determined by this stability analysis.

With all other parameters fixed at their default values, we observe that there is a large domain across the diagonal of the  $D_C - D_{[GA]}$  plane where the system can oscillate with significant amplitude and a period of  $\sim 100$  min. The region of oscillation spans almost three orders of magnitude in both  $D_C$  and  $D_{[GA]}$ . We find that the reference state in fact lies close to the upper border of states with sustained oscillations.

### Importance of saturated degradation of Axin2

It turns out to be crucial for the oscillatory behavior that the rate of degradation of Axin2 is proportional not to  $A$ , the Axin2 concentration, but rather to  $[AL]$ , the concentration of the Axin2-LRP complex. This is because, in the model, the only route for degradation of Axin2 lies via the formation of this complex. The rates of association and dissociation of the  $[AL]$  complex have not been measured experimentally. We therefore examined the effect of varying these rates in our model.

From Fig. 5 it is seen that the oscillatory domain increases as the rate of association and dissociation of  $[AL]$  increases (while keeping the binding strength of the complex constant). In the limit of very fast association and dissociation,

the complex can be considered to be in quasiequilibrium, and we can reduce the model to a six-variable version, as shown in the [Supporting Material](#). This six-variable model is very similar to the model of Goldbeter and Pourqu   (17).

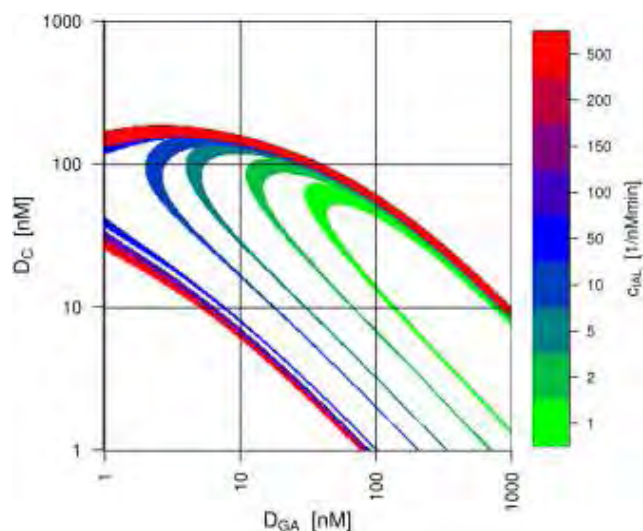


FIGURE 5 The oscillating regime for varying values of  $c_{[AL]}$ . The  $c_{b[AL]}$  value is also varied so as to keep the ratio  $c_{[AL]}/c_{b[AL]}$  constant, i.e., to keep the binding strength of the  $[AL]$  complex unvarying. Oscillations occur within the tongue-shape bounded by the colored lines, where the largest real part of the eigenvalues of the Jacobian is positive. Outside these boundaries all eigenvalues have negative real parts and therefore there can be no sustained oscillations.

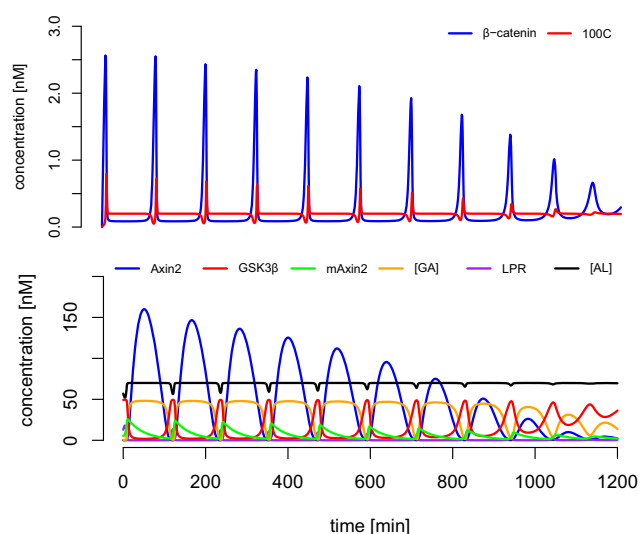


FIGURE 6 Oscillations of the Wnt variables as  $\nu$  is linearly decreased from 0.1 to 0.03  $\text{min}^{-1}$  in 1200 min. Note that the oscillations cease when the Wnt signal falls below a certain threshold level.

In the six-variable equations it is now clear that the degradation rate of Axin2 is not proportional to  $A$ . The degradation is instead given by  $-c_A A / (k_A + A)$ . We call this saturated degradation (29) because when  $A$  becomes much larger than the constant  $k_A$ , the degradation rate approaches a saturation value,  $c_A$ , beyond which it cannot increase further. This saturated degradation is a direct consequence of the complex formation between Axin2 and LRP5/6, especially evident in the case where the complex can be considered to be in quasiequilibrium. In the [Supporting Material](#), we show, by varying  $c_A$  and  $k_A$  values, that oscillations occur only when the average  $A$  concentration is much larger than  $k_A$ . Thus, saturated degradation is a necessary (but not sufficient) condition for oscillations.

### Onset of oscillations: integration with a Wnt gradient

In a growing embryo, new cells are added continuously to the posterior PSM. Older cells therefore have a relative movement toward the anterior PSM. It is known that there is a gradient of Wnt3a ligand across the embryo (2). If we assume that the gradient remains tied to the anterior end of the PSM, any given cell in the PSM effectively sees a slowly decreasing Wnt level with time. This provides a simple way of converting temporal periodicity to spatial periodicity required for precise placement of somites. [Fig. 6](#) shows how oscillations look inside a cell with this kind of decreasing Wnt signal (modeled by a decreasing  $\nu$ -value). When Wnt falls below a threshold level, of course, the oscillations cease. Before that, however, the oscillation amplitude steadily decreases, while the time period is quite robust.

The reason for this behavior can be explained mathematically using the stability analysis we did, described earlier,

which enables us to examine the nature of the transition between oscillations and steady states. For the parameter ranges we have explored, we find that when a biologically feasible steady state (wherein all concentrations are positive and finite) exists, then it is unique. For some parameter values there is no feasible steady state—in this case some concentrations eventually become infinite. For the default parameters, we find that of the six eigenvalues, one is zero, three are real and negative, and two are complex-conjugated with positive real parts. Thus, the unique fixed point is unstable. For all the parameter variations we have studied, where the system changes from a steady state to an oscillating state, we find the onset of oscillations coincides with a Hopf bifurcation (see the [Supporting Material](#)). That is, at the point of onset, the real part of a complex conjugate pair of eigenvalues becomes positive, whereas the imaginary part is finite and nonzero. Hopf bifurcations indicate that the oscillations have a finite time period as soon as they occur, which is why the time period is so robust in the time series of [Fig. 6](#), although the amplitude slowly increases as the system moves away from the Hopf bifurcation and deeper into the oscillating regime.

## DISCUSSION

The important features of our model are:

1. It is based on a negative feedback loop within the canonical Wnt signaling pathway, which involves Axin2 and the destruction complex.
2. It exhibits oscillations in the expression of Wnt target genes such as Axin2.
3. The oscillation periods match experimental values for parameters in realistic domains.
4. The system incorporates Wnt signaling in a way that can account for the absence of oscillations in regions where Wnt concentration is low.
5. The system has the potential for interaction with the Notch signaling pathway in a way that can explain why Notch target genes oscillate out of phase to Wnt targets (see below).

### What is necessary to produce the oscillations?

The oscillations arise from a combination of two features:

1. The core negative feedback loop visible in [Fig. 1](#): a high concentration of free  $\beta$ -catenin promotes high concentrations of Axin2 mRNA and protein, which promotes the forming of the [GA] subcomplex and the destruction complex. The latter degrades  $\beta$ -catenin, leading to a reduced production of Axin2 and so on.
2. The binding of Axin2 at the LRP5/6 coreceptor results in saturated degradation of Axin2. The existence of an upper limit to the degradation rate means Axin2 is degraded relatively slowly when its concentration is large. This



produces an effective time delay, which is required to make a negative feedback loop oscillate (29).

The structure of the model, based on negative feedback plus a time delay due to saturated degradation, is remarkably similar to that of the regulatory circuit producing oscillations in the immune response regulator NF- $\kappa$ B in mammalian cells (29). The tumor suppressor gene, p53, along with its repressor Mdm2, also exhibits oscillations and has the same features (1). This design could thus be quite a ubiquitous way of generating ultradian oscillations in widely varying biological contexts.

### Possible mechanism for interaction between Wnt and Notch pathways

In addition to its ability to phosphorylate and degrade  $\beta$ -catenin, GSK3 $\beta$  has also been shown to bind to the intracellular domain of the Notch receptor ( $N_{icd}$ ) (30,31), which is a transcription factor in Notch signaling. This involvement of GSK3 $\beta$  in both Wnt and Notch signaling provides a simple mechanism which could explain the observation (2) that Wnt and Notch targets oscillate out of phase with each other. The argument would proceed as follows: When Axin2 levels are high, GSK3 $\beta$  is sequestered in the destruction complex or in [GA], leaving little to bind to  $N_{icd}$ . Vice versa, when Axin2 levels are low, GSK3 $\beta$  is free to sequester  $N_{icd}$ . If this results in an activation of  $N_{icd}$ , as is the case for some Notch receptors (31), then there would be an increased transcription of Notch target genes. That is, Notch targets would show a high expression when Wnt targets, like Axin2, have a low expression, and vice versa, i.e., the oscillations of Notch targets would be out of phase with respect to Wnt targets. This scenario, in which the two pathways (in a sense) compete for GSK3 $\beta$ , of course requires more experimental evidence.

### Axin2 mutants

An important experimental observation that bears on this issue is that Axin2 deleted mice (Axin2 $^{-/-}$ ) have no dramatic segmentation phenotype (32). This suggests that Notch targets can still oscillate even in the absence of a Wnt oscillator such as the one proposed in this article. It has, however, been shown that Wnt signaling, at the level of  $\beta$ -catenin, is crucial for oscillations of both Wnt and Notch target genes (33). In loss of function mutants, neither Wnt nor Notch target genes oscillate; however, in stabilized (constant  $\beta$ -catenin level) mutants, only Wnt targets stop oscillating. This puts  $\beta$ -catenin upstream of all pathways and suggests for it a permissive rather than oscillatory role. Our proposed model could fit into this scenario, and can explain the observed oscillations in Wnt targets even if Notch targets have their own intrinsic oscillator. In Axin2 $^{-/-}$  mice, normal Axin, which is present in small concentrations, could resume its normal role from non-PSM tissue and supply the needed scaffolding for the destruction complex. This is sufficient

for canonical Wnt signaling and for a permissive role of  $\beta$ -catenin.

### SUPPORTING MATERIAL

Fourteen figures are available at [http://www.biophysj.org/biophysj/supplemental/S0006-3495\(09\)05997-9](http://www.biophysj.org/biophysj/supplemental/S0006-3495(09)05997-9).

The authors are grateful for discussions with Olivier Pourquié, Albert Goldbeter, Preethi Gunaratne, and Ertugrul Ozbudak.

This research is supported by the Danish Research Foundation and the Villum Kann Rasmussen Foundation.

### REFERENCES

1. Tiana, G., S. Krishna, ..., K. Sneppen. 2007. Oscillations and temporal signaling in cells. *Phys. Biol.* 4:R1–R17.
2. Aulehla, A., and B. G. Herrmann. 2004. Segmentation in vertebrates: clock and gradient finally joined. *Genes Dev.* 18:2060–2067.
3. Aulehla, A., C. Wehrle, ..., B. G. Herrmann. 2003. Wnt3a plays a major role in the segmentation clock controlling somitogenesis. *Dev. Cell.* 4:395–406.
4. Tam, P. P. 1981. The control of somitogenesis in mouse embryos. *J. Embryol. Exp. Morphol.* 65 (Suppl):103–128.
5. Cooke, J., and E. C. Zeeman. 1976. A clock and wavefront model for control of the number of repeated structures during animal morphogenesis. *J. Theor. Biol.* 58:455–476.
6. Cinquin, O. 2007. Understanding the somitogenesis clock: what's missing? *Mech. Dev.* 124:501–517.
7. Bessho, Y., R. Sakata, ..., R. Kageyama. 2001. Dynamic expression and essential functions of Hes7 in somite segmentation. *Genes Dev.* 15:2642–2647.
8. Hirata, H., S. Yoshiura, ..., R. Kageyama. 2002. Oscillatory expression of the bHLH factor Hes1 regulated by a negative feedback loop. *Science.* 298:840–843.
9. Brückner, K., L. Perez, ..., S. Cohen. 2000. Glycosyltransferase activity of Fringe modulates Notch- $\delta$  interactions. *Nature.* 406:411–415.
10. Chen, J., L. Kang, and N. Zhang. 2005. Negative feedback loop formed by Lunatic fringe and Hes7 controls their oscillatory expression during somitogenesis. *Genesis.* 43:196–204.
11. Pigolotti, S., S. Krishna, and M. H. Jensen. 2007. Oscillation patterns in negative feedback loops. *Proc. Natl. Acad. Sci. USA.* 104:6533–6537.
12. Lewis, J. 2003. Autoinhibition with transcriptional delay: a simple mechanism for the zebrafish somitogenesis oscillator. *Curr. Biol.* 13:1398–1408.
13. Hirata, H., Y. Bessho, ..., R. Kageyama. 2004. Instability of Hes7 protein is crucial for the somite segmentation clock. *Nat. Genet.* 36:750–754.
14. Jensen, M. H., K. Sneppen, and G. Tiana. 2003. Sustained oscillations and time delays in gene expression of protein Hes1. *FEBS Lett.* 541:176–177.
15. Bernard, S., B. Cajavec, ..., H. Herzel. 2006. Modeling transcriptional feedback loops: the role of Gro/TLE1 in Hes1 oscillations. *Philos. Transact. A Math Phys. Eng. Sci.* 364:1155–1170.
16. Rodríguez-González, J. G., M. Santillán, ..., M. C. Mackey. 2007. The segmentation clock in mice: interaction between the Wnt and Notch signaling pathways. *J. Theor. Biol.* 248:37–47.
17. Goldbeter, A., and O. Pourquié. 2008. Modeling the segmentation clock as a network of coupled oscillations in the Notch, Wnt and FGF signaling pathways. *J. Theor. Biol.* 252:574–585.
18. Logan, C. Y., and R. Nusse. 2004. The Wnt signaling pathway in development and disease. *Annu. Rev. Cell Dev. Biol.* 20:781–810.

19. Dajani, R., E. Fraser, ..., L. H. Pearl. 2003. Structural basis for recruitment of glycogen synthase kinase  $3\beta$  to the axin-APC scaffold complex. *EMBO J.* 22:494–501.
20. Bhanot, P., M. Brink, ..., R. Nusse. 1996. A new member of the frizzled family from *Drosophila* functions as a Wingless receptor. *Nature.* 382:225–230.
21. Pinson, K. I., J. Brennan, ..., W. C. Skarnes. 2000. An LDL-receptor-related protein mediates Wnt signaling in mice. *Nature.* 407:535–538.
22. Wallingford, J. B., and R. Habas. 2005. The developmental biology of Dishevelled: an enigmatic protein governing cell fate and cell polarity. *Development.* 132:4421–4436.
23. Mao, J., J. Wang, ..., D. Wu. 2001. Low-density lipoprotein receptor-related protein-5 binds to Axin and regulates the canonical Wnt signaling pathway. *Mol. Cell.* 7:801–809.
24. Fagotto, F., E. Jho, ..., F. Costantini. 1999. Domains of axin involved in protein-protein interactions, Wnt pathway inhibition, and intracellular localization. *J. Cell Biol.* 145:741–756.
25. Chia, I. V., and F. Costantini. 2005. Mouse axin and axin2/conductin proteins are functionally equivalent in vivo. *Mol. Cell Biol.* 25:4371–4376.
26. Lee, E., A. Salic, ..., M. W. Kirschner. 2003. The roles of APC and Axin derived from experimental and theoretical analysis of the Wnt pathway. *PLoS Biol.* 1:E10.
27. Behrens, J., J. P. von Kries, ..., W. Birchmeier. 1996. Functional interaction of  $\beta$ -catenin with the transcription factor LEF-1. *Nature.* 382:638–642.
28. Gottardi, C. J., and B. M. Gumbiner. 2004. Distinct molecular forms of  $\beta$ -catenin are targeted to adhesive or transcriptional complexes. *J. Cell Biol.* 167:339–349.
29. Krishna, S., M. H. Jensen, and K. Sneppen. 2006. Minimal model of spiky oscillations in NF- $\kappa$ B signaling. *Proc. Natl. Acad. Sci. USA.* 103:10840–10845.
30. Foltz, D. R., M. C. Santiago, ..., J. S. Nye. 2002. Glycogen synthase kinase- $3\beta$  modulates notch signaling and stability. *Curr. Biol.* 12:1006–1011.
31. Espinosa, L., J. Inglés-Esteve, ..., A. Bigas. 2003. Phosphorylation by glycogen synthase kinase- $3\beta$  down-regulates Notch activity, a link for Notch and Wnt pathways. *J. Biol. Chem.* 278:32227–32235.
32. Yu, H.-M. I., B. Jerchow, ..., W. Hsu. 2005. The role of Axin2 in calvarial morphogenesis and craniosynostosis. *Development.* 132:1995–2005.
33. Dunty, Jr., W. C., K. K. Biris, ..., T. P. Yamaguchi. 2008. Wnt3a/ $\beta$ -catenin signaling controls posterior body development by coordinating mesoderm formation and segmentation. *Development.* 135:85–94.

# Supplementary information

November 4, 2009

## 1 Robustness of oscillations and Stability analysis of the eight variable Wnt model

### 1.1 The fixed points of the Wnt model

The following are the equations of our Wnt model (see main text):

$$\frac{dC}{dt} = c_{fC}B[GA] - c_{bC}C - \alpha C \quad (1)$$

$$\frac{d[GA]}{dt} = c_{f[GA]}GA - c_{b[GA]}[GA] - c_{fC}B[GA] + c_{bC}C + \alpha C \quad (2)$$

$$\frac{dB}{dt} = S - c_{fC}B[GA] + c_{bC}C \quad (3)$$

$$\frac{dG}{dt} = -c_{f[GA]}GA + c_{b[GA]}[GA] \quad (4)$$

$$\frac{dA}{dt} = -c_{f[GA]}GA + c_{b[GA]}[GA] + c_{tLA}A_m - c_{f[AL]}AL + c_{b[AL]}[AL] \quad (5)$$

$$\frac{dA_m}{dt} = c_{tsA}B^h - \frac{A_m}{\tau_{Am}} \quad (6)$$

$$\frac{d[AL]}{dt} = c_{f[AL]}AL - c_{b[AL]}[AL] - \nu[AL] \quad (7)$$

$$\frac{dL}{dt} = -c_{f[AL]}AL + c_{b[AL]}[AL] + \nu[AL] \quad (8)$$

The fixed points of the eight variable Wnt model are found by equating Eqs. 1-8 equal to zero with the boundary conditions

$$G + [GA] + C - G_{\text{tot}} = 0 \quad , \quad (9)$$

and

$$L + [AL] - L_{\text{tot}} = 0 \quad , \quad (10)$$

where  $G_{\text{tot}}$  and  $L_{\text{tot}}$  are the total concentrations of GSK3 $\beta$  and LRP5/6, respectively. In the following an asterisk (\*) denotes a fixed point.

Firstly, we define

$$c_A = \nu L_{\text{tot}} \quad \text{and} \quad k_A = \frac{c_{b[AL]} + \nu}{c_{f[AL]}} \quad . \quad (11)$$

The fixed points of  $[AL]$  and  $L$  are:

$$\begin{aligned} [AL]^* &= L_{\text{tot}} \frac{A^*}{A^* + k_A} \\ L^* &= L_{\text{tot}} - [AL]^* \end{aligned}$$

The fixed points  $C^*$ ,  $A_m^*$  and  $[GA]^*$  are easily found from Eqs. 1, 6 and 9, respectively:

$$\begin{aligned} C^* &= \frac{S}{\alpha} \quad , \\ A_m^* &= c_{tsA} \tau_{Am} B^{*2} \\ [GA]^* &= G_{\text{tot}} - \frac{S}{\alpha} - G^* \quad . \end{aligned}$$

From Eq. 4 we have  $c_{f[GA]}G^*A^* = c_{b[GA]}[GA]^*$ . An expression for  $A^*$  is obtained by inserting this into Eq. 5 together with  $A_m^*$ :

$$c_{tLA}c_{tsA}\tau_{Am}B^{*2} = c_A \frac{A^*}{k_A + A^*} \Leftrightarrow A^* = \frac{\tau_{Am}c_{tLA}c_{tsA}k_AB^{*2}}{c_A - \tau_{Am}c_{tLA}c_{tsA}B^{*2}} = \frac{\tau k_AB^{*2}}{c_A - \tau B^{*2}} ,$$

where  $\tau$  is defined as  $\tau = \tau_{Am}c_{tLA}c_{tsA}$ . The fixed points of  $[GA]^*$  and  $C^*$  are inserted into Eq. 1, and  $G^*$  is found

$$\begin{aligned} 0 &= -c_{fC}B^* \left( G_{\text{tot}} - \frac{S}{\alpha} - G^* \right) + c_{bC} \frac{S}{\alpha} + S \Leftrightarrow \\ G^* &= G_{\text{tot}} - \frac{S}{\alpha} - \frac{c_{bC}S}{c_{fC}B^*\alpha} - \frac{S}{c_{fC}B^*} . \end{aligned}$$

The fixed point for  $\beta$ -catenin is the last one that needs to be found. To do this the fixed points of  $A^*$ ,  $[GA]^*$  and  $G^*$  are inserted into Eq. 4:

$$\begin{aligned} 0 &= -c_{f[GA]} \left( G_{\text{tot}} - \frac{S}{\alpha} - \frac{1}{B^*} \left( \frac{c_{bC}S}{c_{fC}\alpha} + \frac{S}{c_{fC}} \right) \right) \cdot \frac{\tau k_AB^{*2}}{c_A - \tau B^{*2}} \\ &\quad + c_{b[GA]} \frac{1}{B^*} \left( \frac{c_{bC}S}{c_{fC}\alpha} + \frac{S}{c_{fC}} \right) \Rightarrow \\ 0 &= -c_{f[GA]}\tau k_A \left( G_{\text{tot}} - \frac{S}{\alpha} \right) \frac{1}{c_AX^2 - \tau} + \frac{c_{b[GA]}S}{c_{fC}} \left( \frac{c_{bC}}{\alpha} + 1 \right) X \\ &\quad + \frac{\tau k_AS c_{f[GA]}}{c_{fC}} \left( \frac{c_{bC}}{\alpha} + 1 \right) \frac{X}{c_AX^2 - \tau} , \end{aligned} \quad (12)$$

where  $X = \frac{1}{B^*}$  to ease the notation. To get a simple cubic equation the constants  $\eta$ ,  $\theta$  and  $\mu$  are defined by

$$\eta = -c_{f[GA]}\tau k_A \left( G_{\text{tot}} - \frac{S}{\alpha} \right) \quad (13)$$

$$\theta = \frac{\tau k_AS c_{f[GA]}}{c_{fC}} \left( \frac{c_{bC}}{\alpha} + 1 \right) \quad (14)$$

$$\mu = \frac{c_{b[GA]}S}{c_{fC}} \left( \frac{c_{bC}}{\alpha} + 1 \right) . \quad (15)$$

The cubic equation for  $X$  is then

$$0 = \eta \frac{1}{c_AX^2 - \tau} + \theta \frac{X}{c_AX^2 - \tau} + \mu X \Leftrightarrow 0 = X^3 + \frac{\theta - \mu\tau}{\mu c_A} X + \frac{\eta}{\mu c_A} \quad (16)$$

with the discriminant

$$D = \left( \frac{\theta - \mu\tau}{3\mu c_A} \right)^3 + \left( -\frac{\eta}{2\mu c_A} \right)^2 \equiv Q^3 + R^2. \quad (17)$$

In the reference state (see Table 1 in the main text), the discriminant is negative. A negative  $D$  leads to three real solutions:

$$\begin{aligned} X_1 &= 2\sqrt{-Q} \cos(1/3\phi) = 1.81 \text{ nM}^{-1} , \\ X_2 &= 2\sqrt{-Q} \cos(1/3\phi + 2/3\pi) = -1.80 \text{ nM}^{-1} , \text{ and} \\ X_3 &= 2\sqrt{-Q} \cos(1/3\phi + 4/3\pi) = -0.00725 \text{ nM}^{-1} , \end{aligned}$$

with  $\cos(\phi) = \frac{R}{\sqrt{-Q^3}}$ . Since the concentration of  $\beta$ -catenin must be positive the (unique) fixed point in the

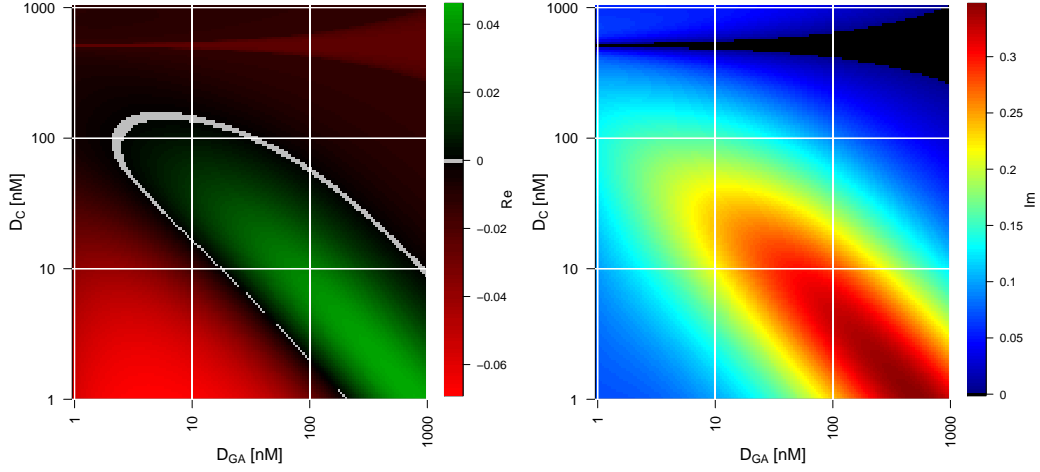


Figure 1: The real (left) and imaginary (right) parts of the complex conjugated eigenvalue with positive imaginary part. When the imaginary part is zero (black) the real part of the complex conjugated eigenvalues splits into two, i.e the real value shown decreases and then increases as  $D_C$  passes through the black region whereas the other complex conjugated eigenvalue behaves oppositely.

reference state is:

$$B^* = \frac{1}{X_1} = 0.592 \text{ nM} \quad (18)$$

$$G^* = G_{\text{tot}} - \frac{S}{\alpha} - \frac{c_{bC}S}{c_{fC}B^*\alpha} - \frac{S}{c_{fC}B^*} = 43.0 \text{ nM} \quad (19)$$

$$[GA]^* = G_{\text{tot}} - \frac{S}{\alpha} - G^* = 6.99 \text{ nM} \quad (20)$$

$$C^* = \frac{S}{\alpha} = 0.002 \text{ nM} \quad (21)$$

$$A^* = \frac{\tau k_A B^{*2}}{c_A - \tau B^{*2}} = 0.976 \text{ nM} \quad (22)$$

$$A_m^* = c_{tsA} B^{*2} \tau_{Am} = 9.82 \text{ nM} \quad (23)$$

$$[AL]^* = L_{\text{tot}} \frac{A^*}{A^* + k_A} = 68.7 \text{ nM} \quad (24)$$

$$L^* = L_{\text{tot}} - [AL]^* = 1.29 \text{ nM} \quad (25)$$

In general the solutions to a cubic equation can be written as

$$X_0 = U + T \quad \text{and} \quad X_{\pm} = -\frac{1}{2}(U + T) \pm \frac{1}{2}\sqrt{3}(U - T)i, \quad (26)$$

where  $U = \sqrt[3]{R + \sqrt{D}}$  and  $T = \sqrt[3]{R - \sqrt{D}}$ . The fixed point of the  $\beta$ -catenin concentration is the real valued maximum of  $1/X_0$  or  $1/X_{\pm}$ . When the fixed point  $B^*$  is found the rest follows from in Eqs. 19-25.

## 1.2 Eigenvalues of the Jacobian

The Jacobian of the Wnt system is given by

$$J = \begin{bmatrix} -c_{bC} - \alpha & c_{fC}B & c_{fC}[GA] & 0 & 0 & 0 & 0 & 0 & 0 \\ c_{bC} + \alpha & -c_{b[GA]} - c_{fC}B & -c_{fC}[GA] & c_{f[GA]}A & c_{f[GA]}G & 0 & 0 & 0 & 0 \\ c_{bC} & -c_{fC}B & -c_{fC}[GA] & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & c_{b[GA]} & 0 & -c_{f[GA]}A & -c_{f[GA]}G & 0 & 0 & 0 & 0 \\ 0 & c_{b[GA]} & 0 & -c_{f[GA]}A & -c_{f[GA]}G - c_{f[AL]}L & c_{tLA} & c_{b[AL]} & -c_{f[AL]}A & 0 \\ 0 & 0 & 2c_{tsA}B & 0 & 0 & -\frac{1}{\tau_{Am}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & c_{f[AL]}L & 0 & -c_{b[AL]} & c_{f[AL]}A & 0 \\ 0 & 0 & 0 & 0 & -c_{f[AL]}L & 0 & c_{b[AL]} & -c_{f[AL]}A & 0 \end{bmatrix}.$$

In the reference state there are eight eigenvalues of the Jacobian evaluated in the fixed points. Two are complex conjugated, two are zero and the other four are real.

The stability of the model is investigated by considering the complex conjugated eigenvalues in the plane spanned by the dissociation constants  $D_C = \frac{c_{bC}}{c_{fC}}$  and  $D_{[GA]} = \frac{c_{b[GA]}}{c_{f[GA]}}$ . We only vary the  $c_{fC}$  and  $c_{b[GA]}$ , whereas

the other constants are as in the reference state. The real and imaginary part of the complex conjugated eigenvalue with positive imaginary part is shown in Fig. 1. At the top of the heatmap in Fig. 1 (left) the complex conjugated eigenvalues splits into two real valued eigenvalues where the imaginary part is zero (colored in black). The range of  $D_C$  where all eigenvalues are real broaden as the value of  $D_{[GA]}$  increases.

There is a tongue shape in the  $D_{[GA]}$ - $D_C$  plane where the real part of the complex conjugated eigenvalue is positive. This tongue shape collides with the area in the  $D_{[GA]}$ - $D_C$  plane where the amplitude of the Axin2 oscillations is non-zero (see Fig. 3 in the main text). For a constant value of  $D_{[GA]}$  and increasing  $D_C$  the real part changes from negative to positive and then back again. This indicates that a Hopf bifurcation takes place. To test this we set  $D_{[GA]} = 10$  and find the eigenvalues of the Jacobian evaluated in the fixed points as  $D_C$  increases. In Fig. 2 the real part of the eigenvalues is plotted against the imaginary part. It is seen that there is one set of complex conjugated eigenvalues (a+b), two eigenvalues are always zero (g+h), and two are always negative (d+f). The eigenvalues in Fig. 2c,e are both negative and positive. However, they are negative when the complex conjugated eigenvalues have positive real parts. Thus a Hopf bifurcation may occur at  $D_C \approx 17$  and  $D_C \approx 136$  where the real part of the eigenvalues in Fig. 2ab is zero, see Fig. 3(left).

In the expression of  $A^*$ , Eq. 22, a singularity occurs when  $c_A = \tau B^{*2}$ . In Fig. 4 this singularity can be seen around  $D_C = 500$ , where  $A^*$  changes sign from positive to negative. The same change in sign occurs for  $G^*$ . The negative values of  $G^*$  and  $A^*$  are biologically improbable. The eigenvalues in Fig. 3(right) as a function of  $D_C$  reflect this change in sign as being discontinuous around  $D_C = 500$ . All the fixed points are shown as a function of  $D_C$  with  $D_{[GA]} = 10$  in Fig. 4.

When the complex conjugated eigenvalues have positive real parts and non-zero imaginary parts there should be sustained oscillations in the solutions of the Wnt model. To show this, timeseries for  $D_C = \{20, 30, 50, 100\}$  (sustained oscillations) and  $D_C = \{10, 450\}$  (no or decaying oscillations) are simulated. Within the range of  $D_C$  where the complex conjugated eigenvalues have positive real parts, the oscillations are in fact sustained, see Fig. 5, and for  $D_C = \{10, 450\}$  the system turns into a non-oscillating mode, see Fig 6. In the case of sustained oscillations the period of the oscillations increases as the value of  $D_C$  increases. This is caused by a simultaneous decrease in the absolute value of the imaginary parts of the complex conjugated eigenvalues, see Fig. 3(middle).

A behaviour similar to what is described in this section is observed when  $D_{[GA]} = \{100, 500\}$ . The two things that change are for which value of  $D_C$  the singularity of  $A^*$  occurs and for which range of  $D_C$  the complex conjugated eigenvalues have positive real part.

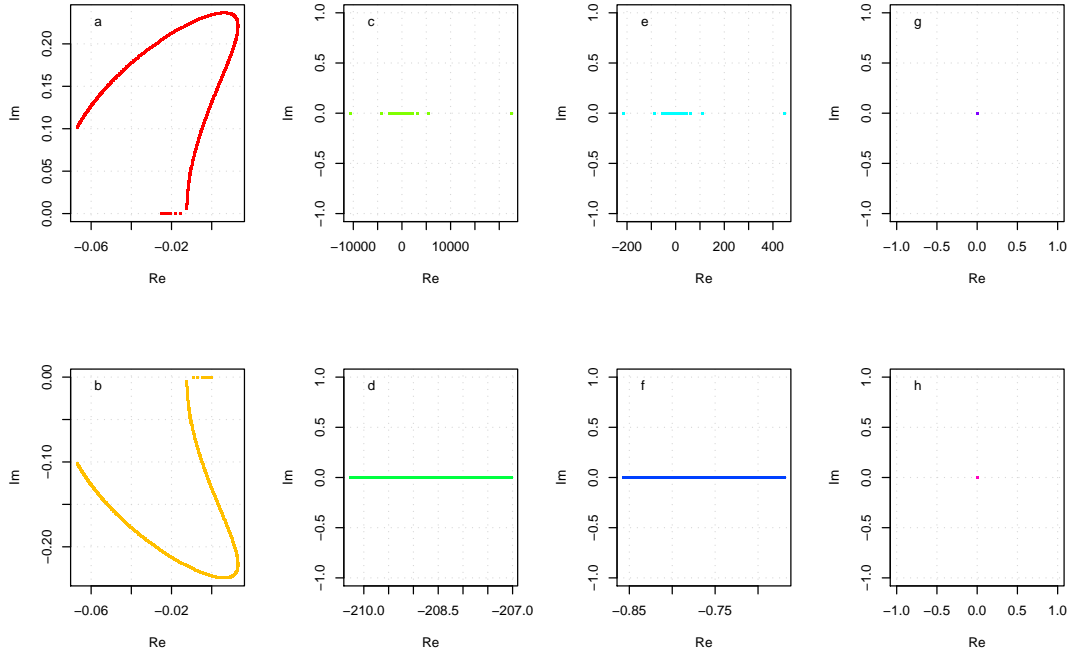


Figure 2: Real part against imaginary part of the eigenvalues of the Jacobian evaluated in the fixed points when  $D_{[GA]} = 10$ .

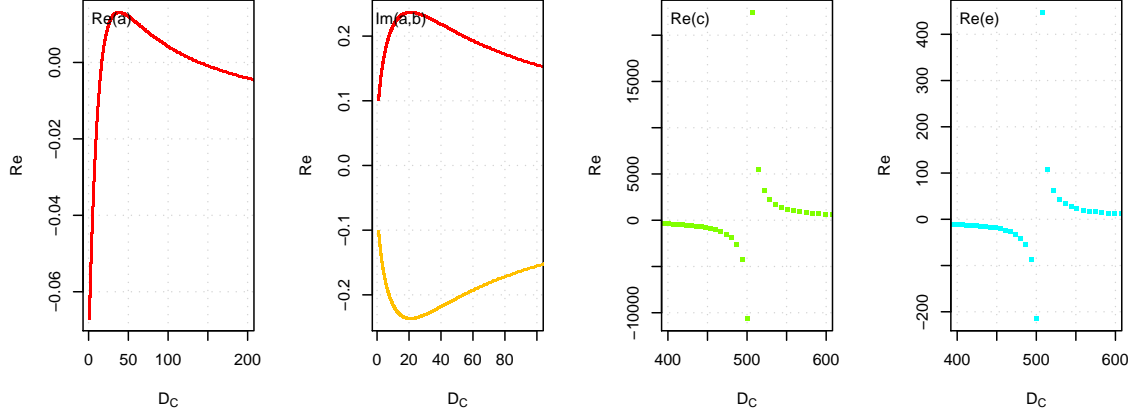


Figure 3: **Left** Real and imaginary parts of the eigenvalues in Fig. 2ab. **Right** The real parts of the eigenvalues in Fig. 2ce have a singularity at  $D_C \approx 500$ . All eigenvalues are real between  $D_C \approx 470 - 540$ .

### 1.3 Changing the half-life of Axin2 mRNA

If the Axin2 mRNA half-life is increased the tongue shape in the  $D_C - D_{[GA]}$  plane in Fig. 1 is narrowed and shifted downwards to the right, i.e. oscillations will occur in a more narrow range of  $D_C$  and  $D_{[GA]}$ , see Fig. 7.

At  $\tau_{Am}$  equal to 100 min there is one pair of complex conjugated eigenvalues of the Jacobian evaluated in the fixed points. The imaginary part of one of these is plotted as function of the real part in Fig. 8 (left). There is a resemblance to Fig. 2a except that the real part stays negative for all values of  $D_C$  and consequently the system will not display sustained oscillations. Timeseries for  $D_{[GA]} = 10$ ,  $D_C = 15$  and  $\tau_{Am} = 100$  min can be seen in Fig. 9, where the system is oscillating, but the oscillations decay. As for the complex conjugated eigenvalue the other real eigenvalues resembles the ones displayed in Fig. 2, i.e. one is always zero, two is constantly negative and one has a singularity.

Even though the system is not showing any sustained oscillations the characteristics such as the singularity

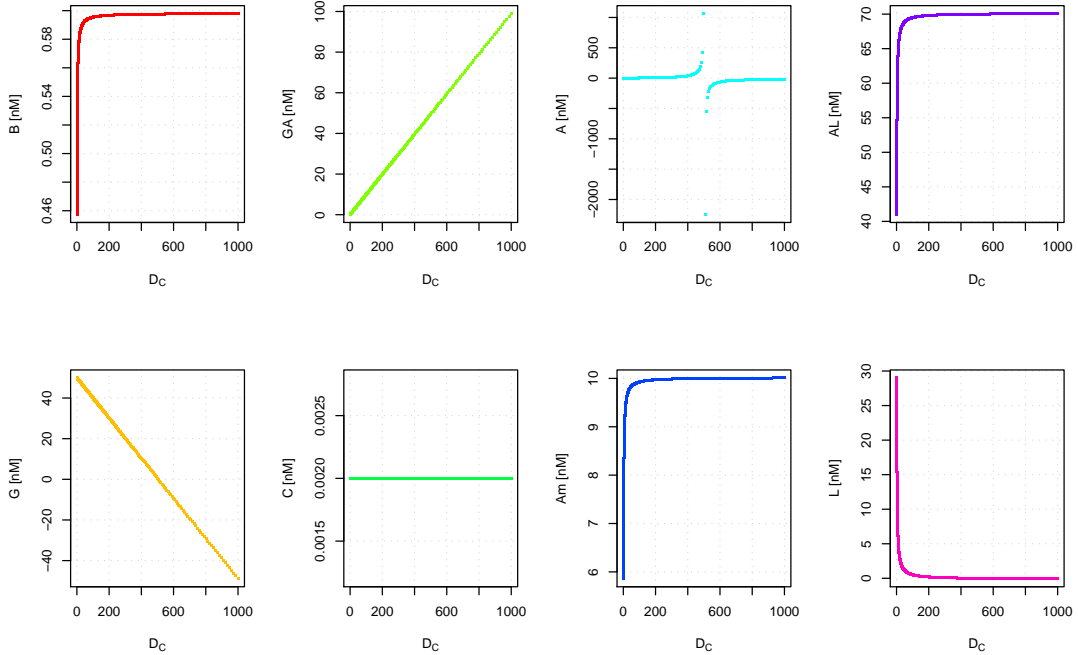


Figure 4: The fixed points of the Wnt model as a function of  $D_C$ . Other parameters are kept constant at the default values.

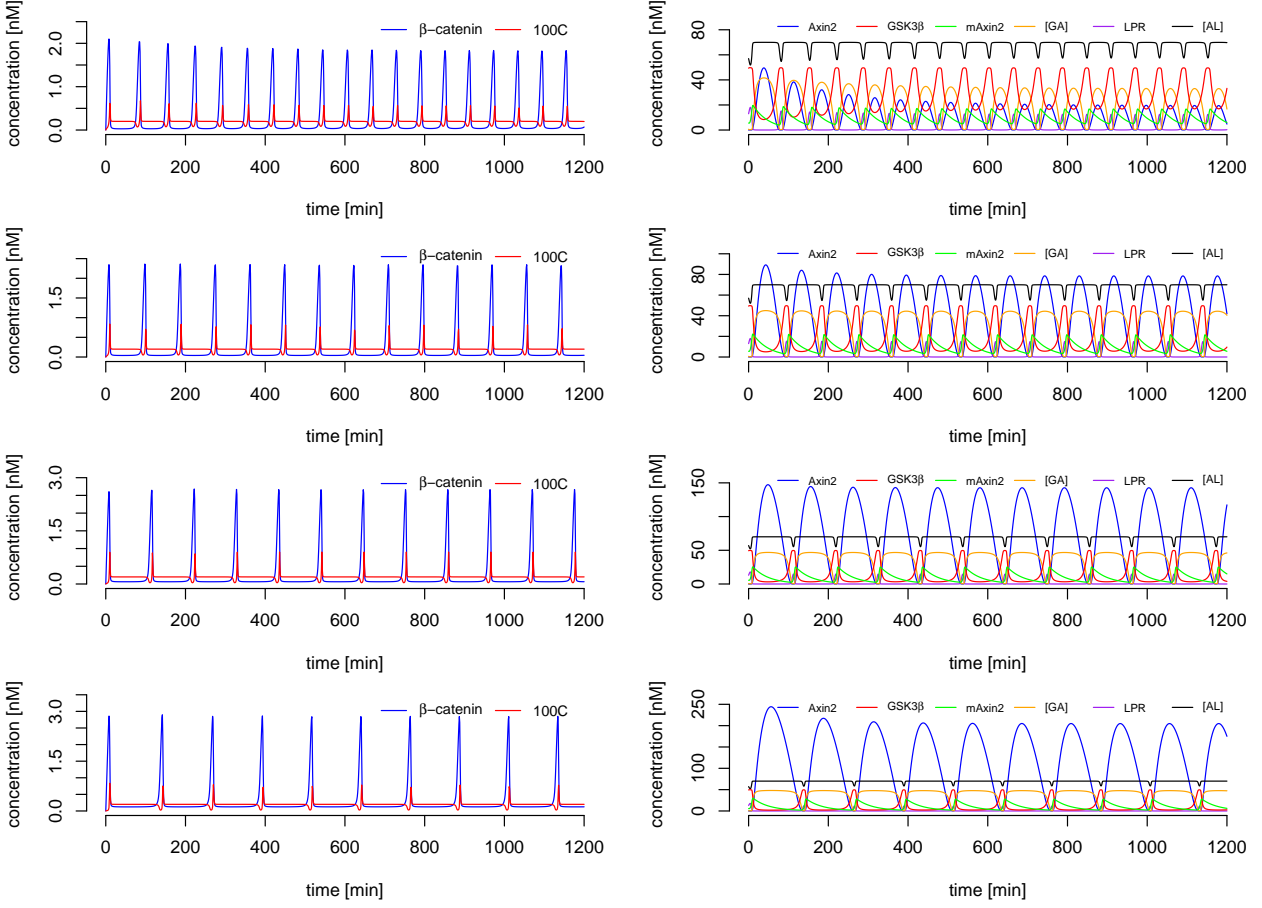


Figure 5: Timeseries for  $D_{[GA]} = 10$  and  $D_C = \{20, 30, 50, 100\}$  (top to bottom). In all panels there are sustained oscillations with periods increasing as  $D_C$  increases.

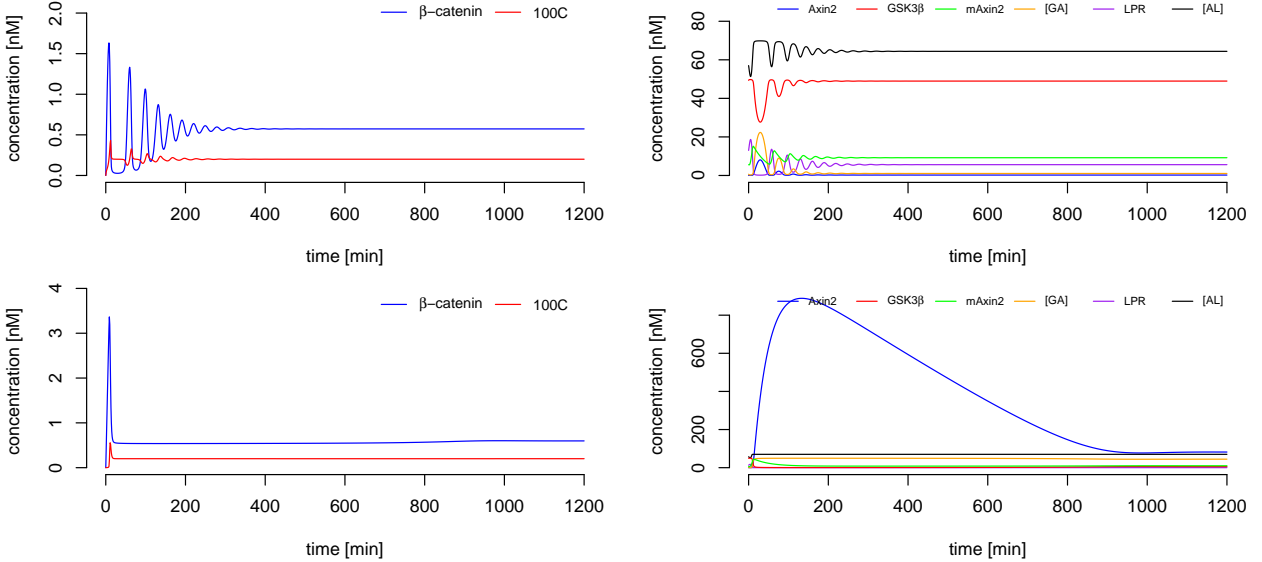


Figure 6: Timeseries for  $D_{[GA]} = 10$  and  $D_C = \{10, 450\}$  (A,B). In all panels the system evolves towards a non-oscillating mode.

of  $A^*$ , and the splitting of the complex conjugated eigenvalues into two real eigenvalues are still seen. The singularity occurs around  $D_C = 320$ , see Fig. 8 (middle, right). The splitting of the complex conjugated eigenvalues occurs approximately between  $D_C = 308$  and  $D_C = 330$ .

When  $D_{[GA]} = 100$  there are values of  $D_C$  for which the real part of the complex conjugated eigenvalues is



positive. Hence between  $D_C = 4$  and  $D_C = 22$  the system can display sustained oscillations which is shown in Fig. 10. Clearly the range of  $D_C$  where the system can oscillate is narrowed when the half-life of Axin is increased.

Again it is seen that the periods of the oscillations increase as the value of  $D_C$  increases (data not shown), and as described above, it is caused by a decrease in the absolute value of the imaginary part of the complex conjugated eigenvalues. Furthermore, the period of the oscillations is increased when  $\tau_{Am}$  is increased from 40 min to 100 min. This is expected since the balance between degradation and translation of Axin2 mRNA determines the concentration of Axin2. Thus when the half-life of Axin2 mRNA is increased then the time wherein Axin2 concentration is building up becomes longer resulting in a longer period of Axin2 oscillations, as seen in Fig. 11.

#### 1.4 The effect of $\beta$ -catenin degradation

The speed at which  $\beta$ -catenin is degraded from the destruction complex is determined by the value of the parameter  $\alpha$ . It simultaneously also measures the speed by which the destruction complex breaks apart. Its default value of  $200 \text{ min}^{-1}$  is very high, which is clearly evidenced from the time series in Fig. 2 (main article), where the concentration of the destruction complex C is nearly always zero. Around its default value, however,  $\alpha$  has little effect on the time period of the oscillations, as compared to parameters such as  $D_C$  or  $\tau_{Am}$  (see Figure 12).

Figure 13 shows plots of the amplitude of the oscillations of Axin2 in the  $D_C - D_{[GA]}$  parameterplane for three different values of  $\alpha$ . There is a clear trend that for values of  $\alpha$  lower than  $200 \text{ min}^{-1}$ , the domain of oscillations moves almost vertically down in the  $D_C - D_{[GA]}$  plane. This is not surprising, when we look at Eq.1 for the destruction complex. Only when  $\alpha$  becomes comparable in size to  $c_{bC}$  (which is equal to  $7 \text{ min}^{-1}$ ), do the dynamics begin to change, because normal breaking of the destruction complex (governed by  $c_{bC}$ ) becomes relatively strong.

Re-entering the oscillatory domain (moving from the dot marked with “ref state” to the one marked with “osc state”) requires a decrease of  $D_C$ , which is in reality accomplished by raising  $c_{fC}$ , the formation constant of C. The effect of a lower degradation rate and an increased formation of C is seen in the timeseries in Figure 13, where C shows an increase in concentration. A higher accumulation of destruction complex is thus needed in order to keep  $\beta$ -catenin levels low between its spikes.

As the oscillatory domain moves down in the  $D_C - D_{[GA]}$  plane, it maintains its profile. Because of the sinusoidal profile of the Axin2 oscillation, there is a good relationship between the amplitude and period of Axin2. Areas of the parameterplots with comparable amplitude generally have comparable periods. Periods of around 100 minutes, require Axin2 amplitudes approximately between 120 nM and 250 nM.

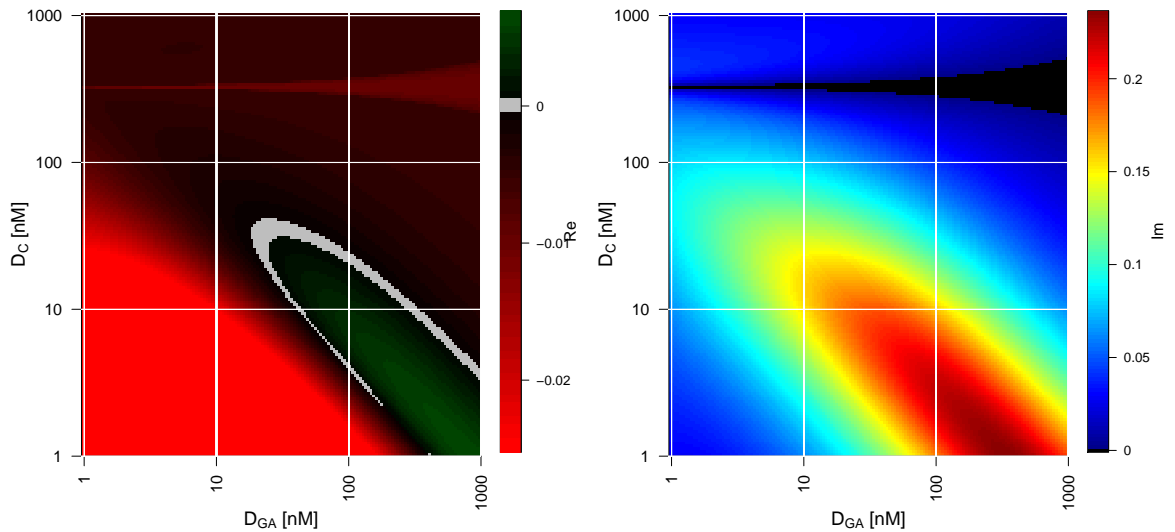


Figure 7: The real (left) and imaginary (right) parts of the complex conjugated eigenvalue of the Jacobian evaluated in the fixed points with  $\tau_{Am} = 100 \text{ min}$ . The black area indicates where the imaginary part is zero as in Fig. 1.

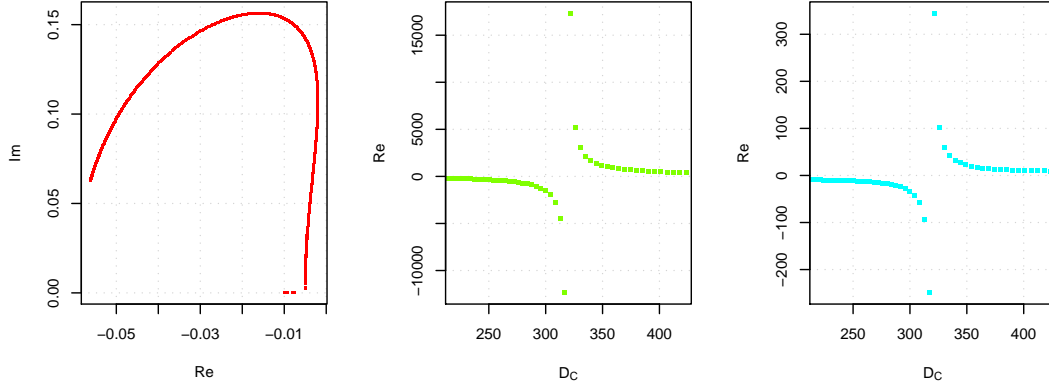


Figure 8: **Left** Real part against imaginary part of one of the complex conjugated eigenvalues of the Jacobian for  $D_{[GA]} = 10$  and  $\tau_{Am} = 100$  min. **Middle, Right** The eigenvalues with a singularity around  $D_C = 320$ .

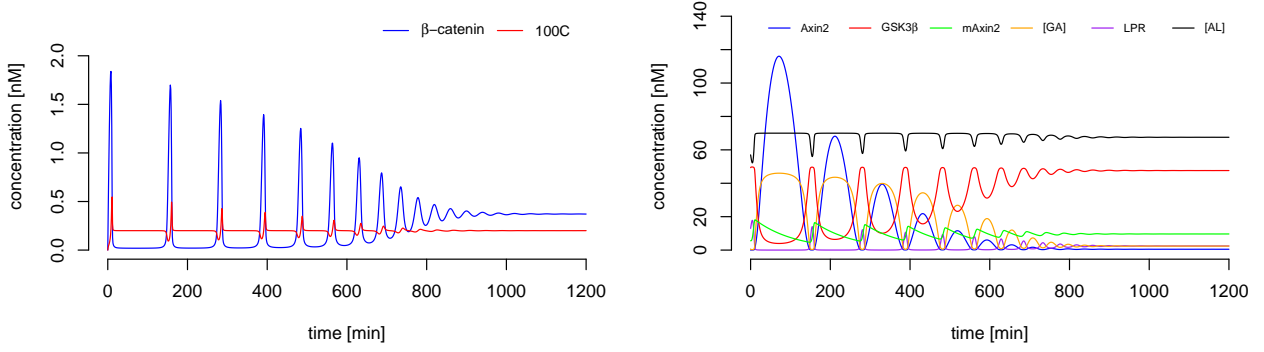


Figure 9: Timeseries of the Wnt model with  $D_{[GA]} = 10$ ,  $D_C = 15$  and  $\tau_{Am} = 100$  min. The oscillations decay because of a negative real part in the complex conjugated eigenvalues.

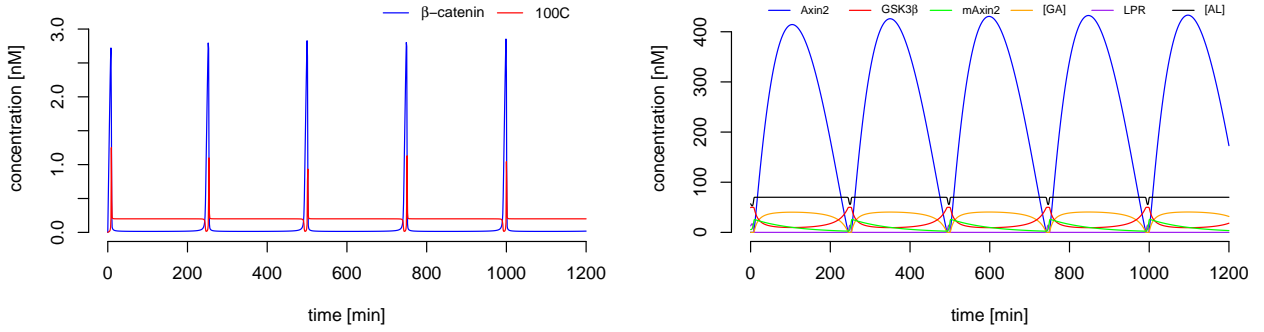


Figure 10: Timeseries of the Wnt model with  $D_{[GA]} = 100$ ,  $D_C = 10$  and  $\tau_{Am} = 100$  min.

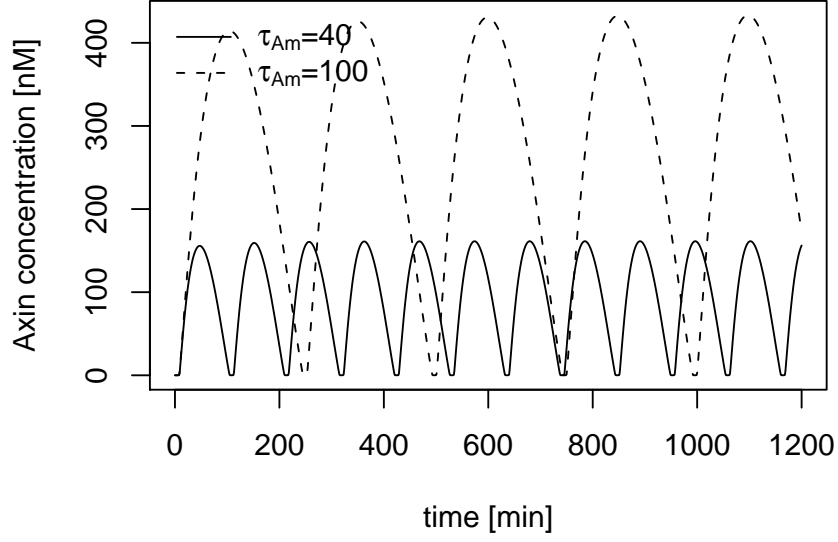


Figure 11: Timeseries of Axin2 (left) and Axin2 mRNA (right) when  $\tau_{Am}$  equals 40 (solid line) and 100 (dotted line). The values of  $D_C$  and  $D_{[GA]}$  are 10 and 100, respectively. The period of the oscillations is clearly longer when the Axin2 mRNA half-life is increased.

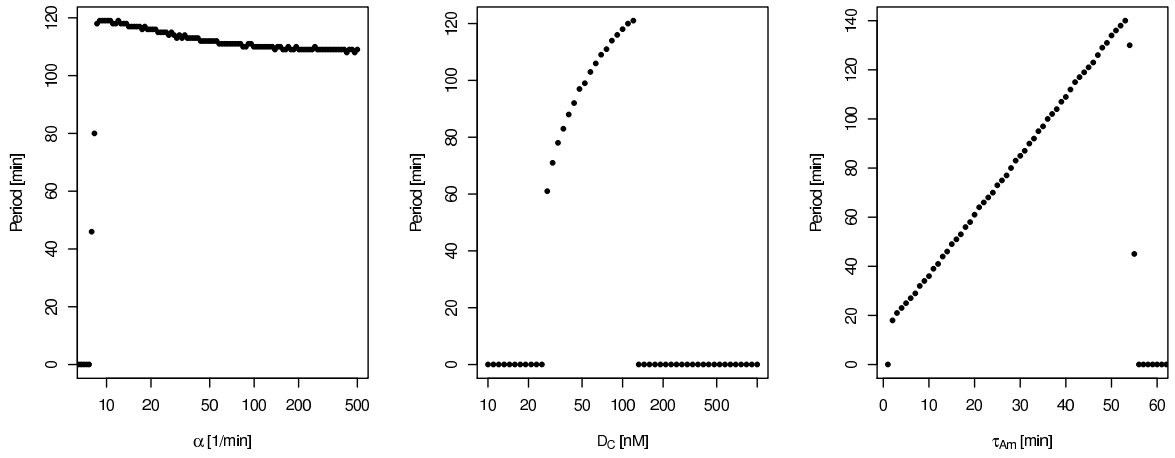


Figure 12: Effect of changes in  $\alpha$ ,  $D_C$  and  $\tau_{Am}$  on the period of oscillations.

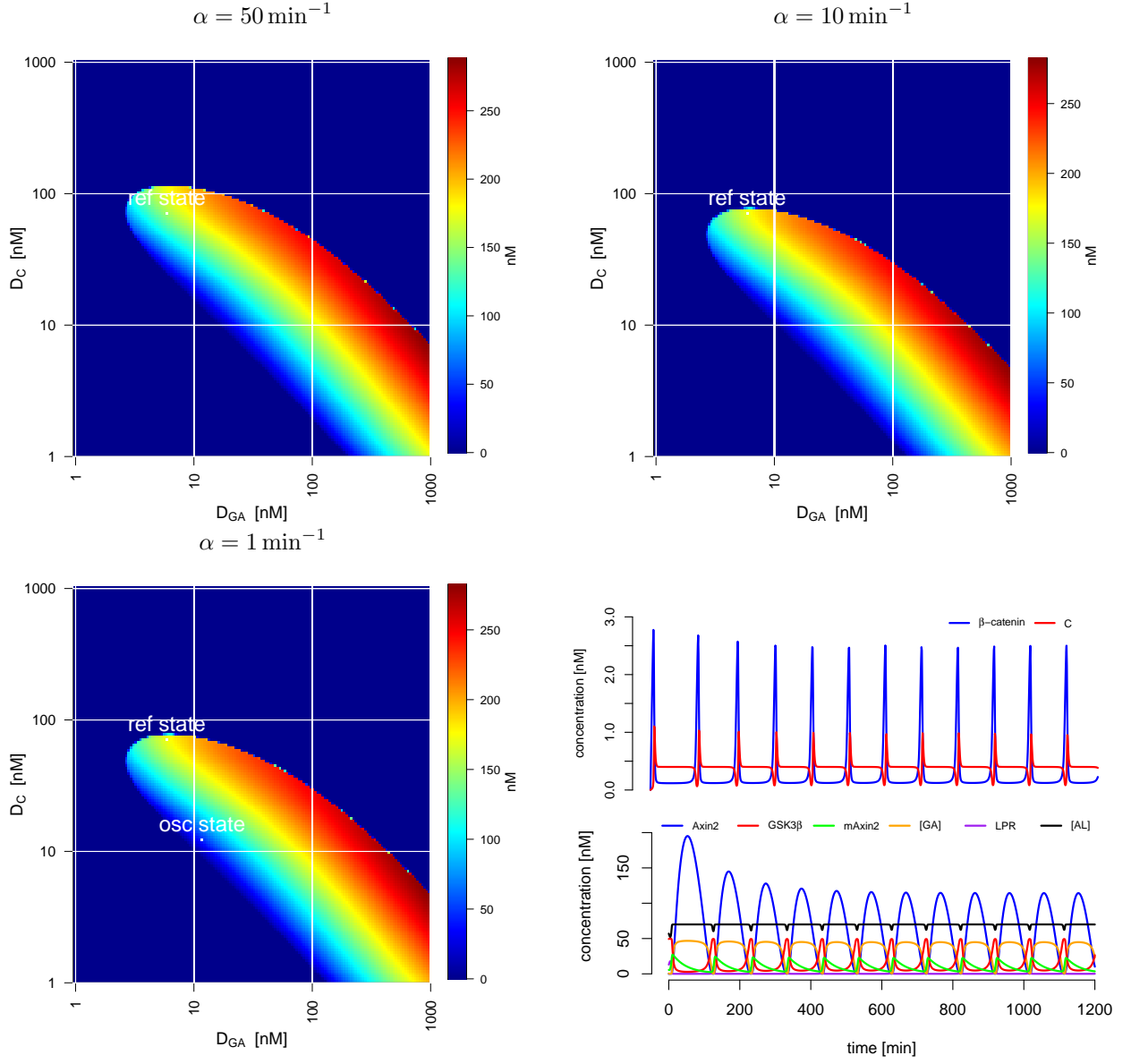


Figure 13: The effect of  $\alpha$  on the amplitude of Axin2 oscillations in the  $D_C$  and  $D_{[GA]}$  parameterplane. Dots marked “ref state” indicate the location of the reference state values in the  $D_C$ - $D_{[GA]}$  plane. The time series shown corresponds to parameter values given by the dot marked “osc state” for  $\alpha = 1 \text{ min}^{-1}$ .

## 2 Reduction from eight to six variables in the Wnt model

We begin with the equations of our eight variable model:

$$\frac{dC}{dt} = c_{fC}B[GA] - c_{bC}C - \alpha C \quad (27)$$

$$\frac{d[GA]}{dt} = c_{f[GA]}GA - c_{b[GA]}[GA] - c_{fC}B[GA] + c_{bC}C + \alpha C \quad (28)$$

$$\frac{dB}{dt} = S - c_{fC}B[GA] + c_{bC}C \quad (29)$$

$$\frac{dG}{dt} = -c_{f[GA]}GA + c_{b[GA]}[GA] \quad (30)$$

$$\frac{dA}{dt} = -c_{f[GA]}GA + c_{b[GA]}[GA] + c_{tLA}A_m - c_{f[AL]}AL + c_{b[AL]}[AL] \quad (31)$$

$$\frac{dA_m}{dt} = c_{tsA}B^h - \frac{A_m}{\tau_{Am}} \quad (32)$$

$$\frac{d[AL]}{dt} = c_{f[AL]}AL - c_{b[AL]}[AL] - \nu[AL] \quad (33)$$

$$\frac{dL}{dt} = -c_{f[AL]}AL + c_{b[AL]}[AL] + \nu[AL] \quad (34)$$

Assuming the complex  $[AL]$  is in quasi-equilibrium (i.e., assuming the association and dissociation timescales are much faster than other processes) gives:

$$\frac{d[AL]}{dt} = 0 \Leftrightarrow \nu[AL] = c_{f[AL]}AL - c_{b[AL]}[AL] , \quad (35)$$

which can be substituted for the last two terms in Eq. 31. Since  $L_{tot} = [AL] + L$  and  $[AL]$  are constant, then  $L$  is also constant giving  $\frac{dL}{dt} = 0$  in Eq. 34. To get rid of the substituted  $[AL]$  term in Eq. 31 we rewrite Eq. 35 and substitute  $L = L_{tot} - [AL]$ :

$$[AL] = \frac{c_{f[AL]}AL}{c_{b[AL]} + \nu} = \frac{c_{f[AL]}A}{c_{b[AL]} + \nu}(L_{tot} - [AL]) . \quad (36)$$

Finally, isolating  $[AL]$  in Eq. 36 and substituting  $k_A = \frac{c_{b[AL]} + \nu}{c_{f[AL]}}$  gives:

$$[AL] = \frac{A}{k_A}(L_{tot} - [AL]) \Rightarrow [AL] = \frac{L_{tot}A}{k_A + A} . \quad (37)$$

Substituting with  $[AL]$  gives the following equation for Axin2 time dependency:

$$\frac{dA}{dt} = -c_{f[GA]}GA + c_{b[GA]}[GA] + c_{tLA}A_m - \nu L_{tot} \frac{A}{k_A + A} . \quad (38)$$

None of the constants involved in the last term of Eq. 38 are known, and one could as well replace  $\nu L_{tot}$  by a single constant  $c_A$ . The final result is the reduction of the eight variable system to the following six variable system:

$$\frac{dC}{dt} = c_{fC}B[GA] - c_{bC}C - \alpha C \quad (39)$$

$$\frac{d[GA]}{dt} = c_{f[GA]}GA - c_{b[GA]}[GA] - c_{fC}B[GA] + c_{bC}C + \alpha C \quad (40)$$

$$\frac{dB}{dt} = S - c_{fC}B[GA] + c_{bC}C \quad (41)$$

$$\frac{dG}{dt} = -c_{f[GA]}GA + c_{b[GA]}[GA] \quad (42)$$

$$\frac{dA}{dt} = -c_{f[GA]}GA + c_{b[GA]}[GA] + c_{tLA}A_m - c_A \frac{A}{k_A + A} \quad (43)$$

$$\frac{dA_m}{dt} = c_{tsA}B^2 - \frac{A_m}{\tau_{Am}} \quad (44)$$

This system is very similar to the one proposed by Ref. [1]. In particular, the crucial ingredient in this six-variable model, as well as the model of Ref. [1] is the saturated degradation term in Eq. 35:  $-c_A \frac{A}{k_A + A}$ . We call

this saturated degradation, because when  $A$  is sufficiently large, the degradation rate saturates to a finite value,  $c_A$ . Our analysis shows that the source of this nonlinearity is in fact the quasi-equilibrated complex formation between  $A$  and  $L$ .

Figure 14 shows evidence that saturated degradation is necessary for oscillations. The colormap shows the ratio of average  $A$  concentration to  $k_A$  on the  $k_A - c_A$  parameter plane. Superimposed on this are contour lines corresponding to the amplitude of Axin2 oscillations observed after  $t=4$  hours. It is clear that oscillations with significant amplitude only occur in regions where  $\langle A \rangle / k_A \gg 1$ , indicating that saturated degradation is necessary (though not sufficient) for oscillations. Note that because we plot the amplitude after  $t=4$  hours, we cannot distinguish between sustained oscillations or slowly damped oscillations that have not died out by 4 hours. However, this seems a reasonable procedure because for an embryo it is irrelevant whether the oscillations are truly sustained, or so slowly damped that they maintain a significant amplitude over many hours.

## References

- [1] Goldbeter A, Pourqu   O (2008) Modeling the segmentation clock as a network of coupled oscillations in the Notch, Wnt and FGF signaling pathways. *J Theor Biol* 252:574–585. doi:10.1016/j.jtbi.2008.01.006. URL <http://dx.doi.org/10.1016/j.jtbi.2008.01.006>.

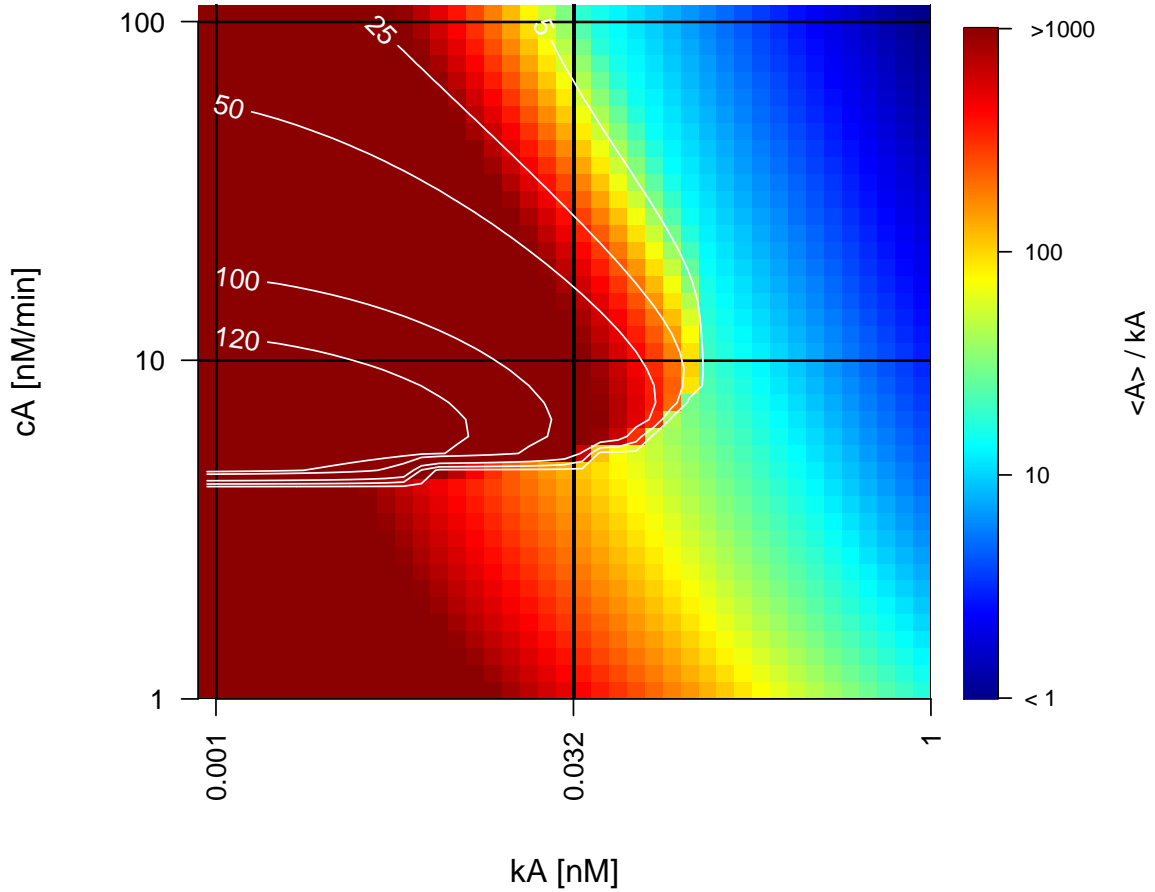


Figure 14: Colours show the ratio of average Axin2 concentration,  $\langle A \rangle$ , to  $k_A$ , which is an effective dissociation constant for the Axin2-LRP complex, on the  $k_A - c_A$  parameter plane. When the ratio is large the degradation of Axin2 is saturated. White contour lines indicate the amplitude of oscillations (in nM) observed after  $t=4$  hours. Note that oscillations are only observed in regions where  $\langle A \rangle / k_A \gg 1$ .



APPENDIX

C

OSC paper



# Modeling oscillatory control in NF- $\kappa$ B, p53 and Wnt signaling

Benedicte Mengel<sup>1</sup>, Alexander Hunziker<sup>1</sup>, Lykke Pedersen<sup>1</sup>, Ala Trusina<sup>1</sup>, Mogens H Jensen<sup>1</sup> and Sandeep Krishna<sup>1,2</sup>

Oscillations are commonly observed in cellular behavior and span a wide range of timescales, from seconds in calcium signaling to 24 hours in circadian rhythms. In between lie oscillations with time periods of 1–5 hours seen in NF- $\kappa$ B, p53 and Wnt signaling, which play key roles in the immune system, cell growth/death and embryo development, respectively. In the first part of this article, we provide a brief overview of simple deterministic models of oscillations. In particular, we explain the mechanism of saturated degradation that has been used to model oscillations in the NF- $\kappa$ B, p53 and Wnt systems. The second part deals with the potential physiological role of oscillations. We use the simple models described earlier to explore whether oscillatory signals can encode more information than steady-state signals. We then discuss a few simple genetic circuits that could decode information stored in the average, amplitude or frequency of oscillations. The presence of frequency-detector circuit downstream of NF- $\kappa$ B or p53 would be a strong clue that oscillations are important for the physiological response of these signaling systems.

## Addresses

<sup>1</sup> Center for Models of Life, Niels Bohr Institute, Copenhagen, Denmark

<sup>2</sup> National Centre for Biological Sciences – TIFR, Bangalore, India

Corresponding author: Krishna, Sandeep ([sandeep@nbi.dk](mailto:sandeep@nbi.dk))

**Current Opinion in Genetics & Development** 2010, **20**:656–664

This review comes from a themed issue on  
 Genetics of system biology  
 Edited by Jeff Hasty, Alex Hoffmann and Susan Golden

Available online 9th October 2010

0959-437X/\$ – see front matter  
 © 2010 Elsevier Ltd. All rights reserved.

DOI [10.1016/j.gde.2010.08.008](https://doi.org/10.1016/j.gde.2010.08.008)

## Introduction

Living cells continuously adjust gene expression patterns in response to changing environmental conditions. A simple way of encoding the presence of a stress or stimulus is to shift the concentration of a signaling molecule from one steady-state level to another. This scheme has potential disadvantages including the cost of continuous production of signaling molecules at a high level and unwanted cross-talk between pathways. It is not surprising, therefore, that cells often encode information about environmental changes in complex time-varying signals. This review deals with one subclass of such systems: those that exhibit oscillations.

The most obvious examples of periodic behavior are circadian rhythms [1,2] and cell cycles [3]. Much more rapid oscillations are seen in the levels of cellular calcium [4]. Many hormones also show intermittently periodic behavior and pulsatile secretion [5]. Figure 1 shows three more systems where oscillations have been observed: the NF- $\kappa$ B [6<sup>•</sup>,7<sup>•</sup>,8,9], p53 [10<sup>•</sup>,11<sup>•</sup>] and somitogenesis systems [12,13<sup>•</sup>,14], which are important for the immune response, cell growth/death and embryo development, respectively.

The purpose of this review is to provide a brief conceptual overview of deterministic mathematical models of such oscillations, and suggest how they can be used to explore the potential physiological role of oscillatory signals. Our emphasis will be on simplified models and the quintessential understanding they provide [15–17], rather than complex models which aim to reproduce experimental data in detail.

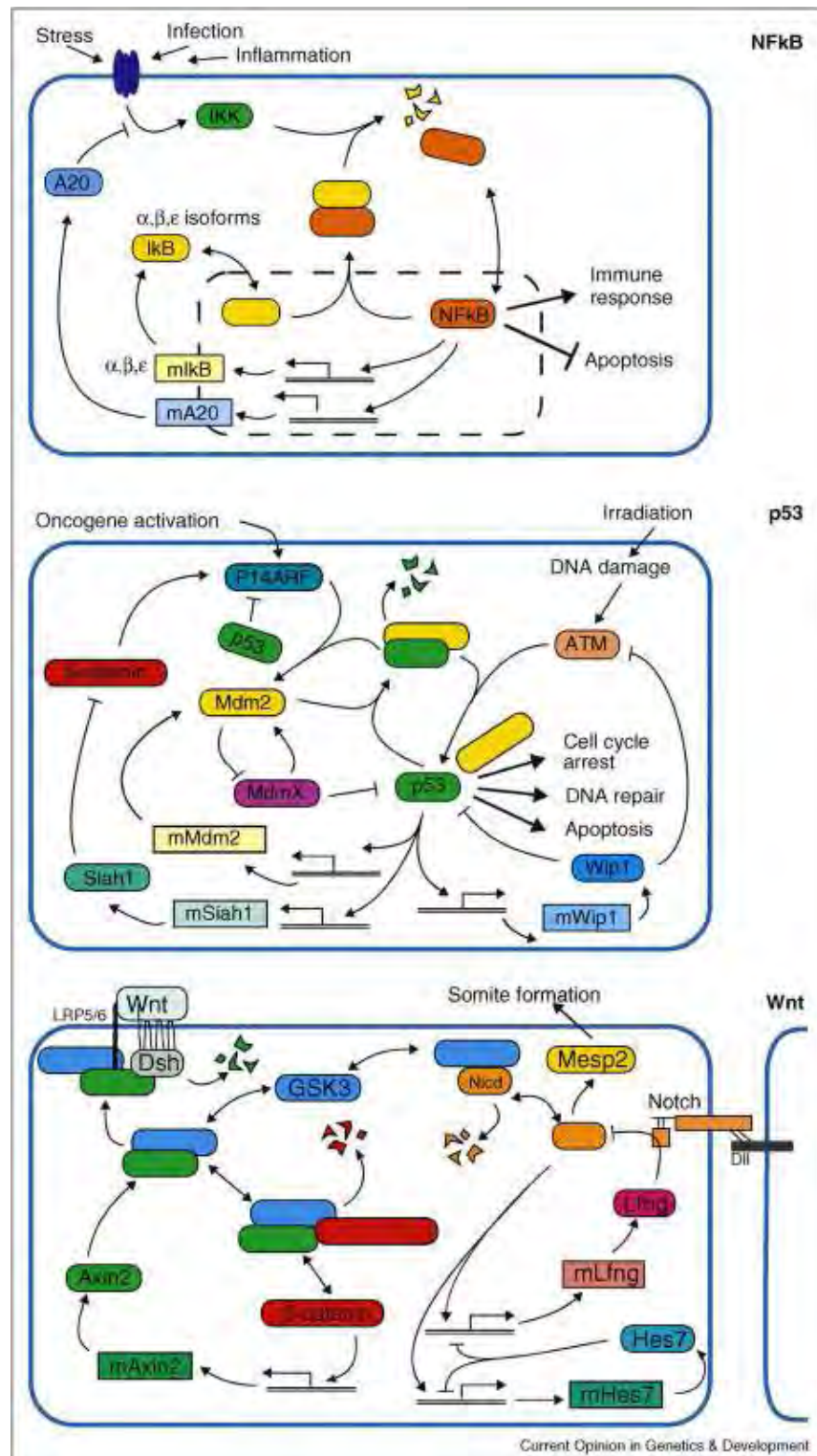
## Modeling oscillations

### Minimum ingredients for generating oscillations

The minimum requirement for oscillations is a negative feedback loop with a time delay [18]. A feedback loop is a closed cycle of nodes, representing genes, proteins, mRNA, etc. (henceforth: ‘regulators’), each affecting the concentration or activity of the next node through activatory or inhibitory links (Figure 2a). A negative feedback loop is the one with an *odd* number of inhibitory links. A small perturbation of one regulator will perturb the next one in the loop, which will increase or decrease the concentration of the next regulator, and so on, until the signal returns to the original regulator. The original perturbation will be cancelled if there are an odd number of inhibitory links. Thus, negative feedback tends to produce homeostasis. The faster the signal goes around the loop, the quicker the perturbations are nullified and therefore the tighter the homeostasis. However, if the signal goes around sufficiently slowly, that is, with a distinct time delay, then this homeostasis can be broken. Negative feedback will still try to counteract perturbations, but the delay can make the regulator concentrations repeatedly overshoot their homeostatic levels and so oscillate (Figure 2b).

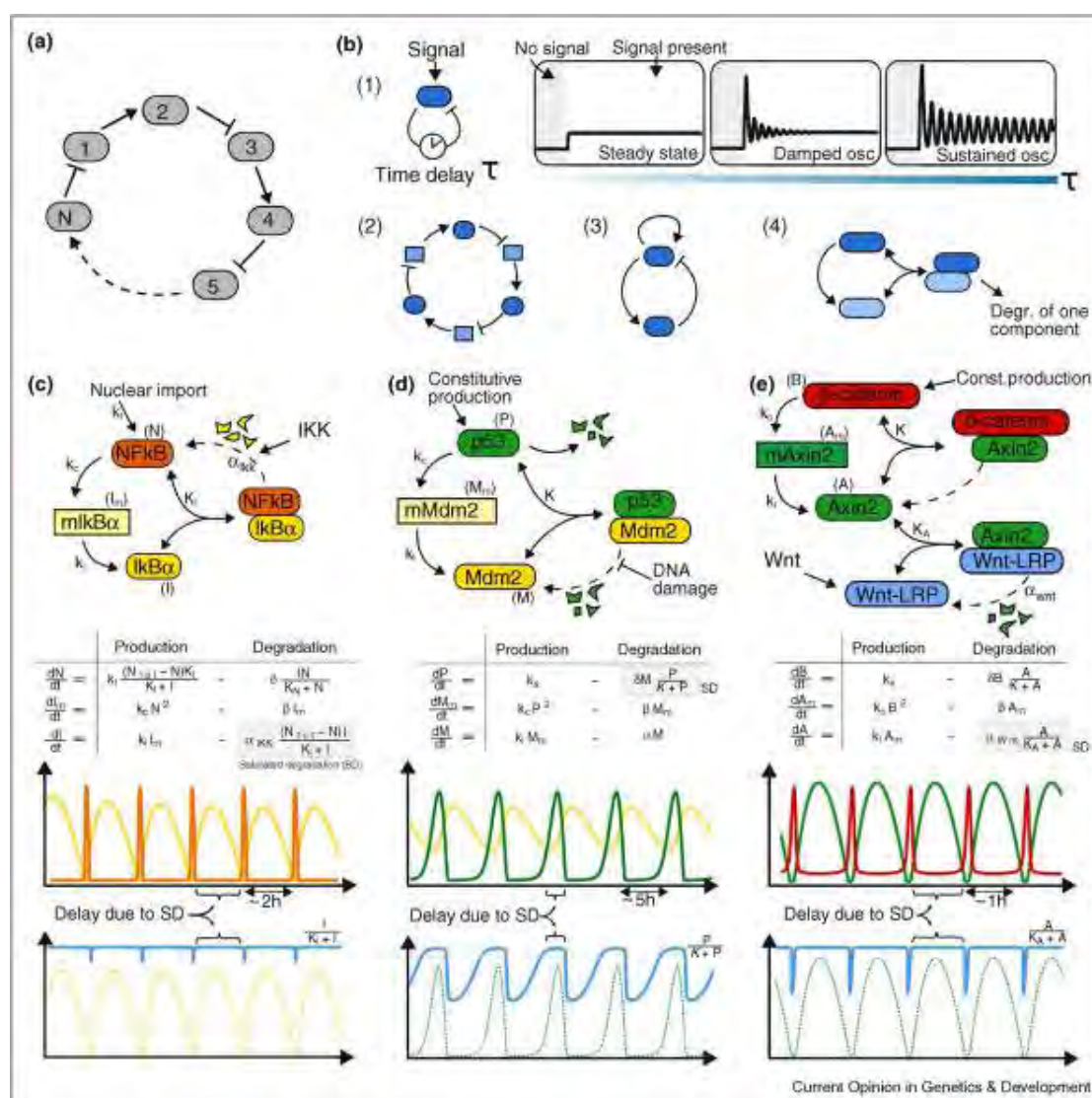
There are several ways to obtain an effective time delay [18], which will be elaborated upon in the next section:

Figure 1



The NF- $\kappa$ B, p53 and Wnt signaling systems. Schematic figures of some of the important proteins (rounded rectangles), mRNA (rectangles), genes (double lines), and interactions (activatory: ordinary arrows; inhibitory: barred arrows) in the three systems. Ordinary arrows are also used to represent conversion and transport between cellular compartments, and merging arrows indicate complex formation. More details of each network can be found in Refs. [59,60] (NF- $\kappa$ B), [61,62,53] (p53), [63,64] (Wnt). Experiments showing sustained and damped oscillations in these systems can be found in Refs. [6\*\*,7\*\*,8,9] (NF- $\kappa$ B), [10\*\*,11\*\*] (p53), [13\*\*,14] (Wnt).

Figure 2



Modeling negative feedback and time delay. **(a)** A general feedback loop containing activatory (ordinary arrows) and inhibitory links (barred arrows). Negative feedback loops are those that contain an odd number of inhibitory links. **(b1)** The simplest negative feedback loop: a single regulator inhibiting its own production/activity with an explicit time delay described by the equation  $dx/dt = \alpha/(K + x(t - \tau)) - \gamma x$ . As the time delay in this loop is increased, the behavior shifts from homeostasis to damped to sustained oscillations. **(b2 – b4)** Different ways of obtaining a time delay: more intermediate steps in the loop, an additional positive feedback which produces a switch-like response, and saturated degradation. **(c – e)** Core feedback loops generating oscillations in the NF- $\kappa$ B, p53 and Wnt systems. Rectangles denote mRNA and rounded rectangles represent proteins. Also shown are equations of three-variable models for each system. The NF- $\kappa$ B model is taken from Ref. [39]. The p53 and Wnt models have been simplified from the models of Refs. [42–44] by assuming all complexes are in quasi-equilibrium. Capital letters signify protein/mRNA concentrations, Greek letters are degradation rates, subscripted  $k$ 's denote maximal production rates, and subscripted  $K$ 's are dissociation constants. All these models use saturated degradation (terms in shaded boxes) to generate the sustained oscillations shown in the lower panels. The times when the relevant complexes are saturated are indicated by the flat peaks of the blue curves (bottom panels) that plot: **(c)**  $I/(K_I + I)$ , **(d)**  $P/(K + P)$ , **(e)**  $A/(K_A + A)$ . For a selection of more detailed deterministic models of these three systems see [65,66] (NF- $\kappa$ B), [67,68] (p53), [69,70] (Wnt).

- Processes that take a minimum amount of time.** For example, transcription and translation (Figure 2b.1).
- Many intermediate steps**, that is, a long feedback loop. Each step adds to the overall time delay (Figure 2b.2).
- Switch-like responses**, where a regulator must reach a threshold concentration before it acts on the next in the loop (Figure 2b.3).
- Saturated degradation**, where the degradation of a regulator is delayed by saturated complex formation (Figure 2b.4).



### Modeling negative feedback and time delay

The model of Hes oscillations in [19] is an example of case (1) above, where an explicit time delay in the production of Hes mRNA (which is inhibited by Hes protein) represents the amount of time taken for transcription, translation and then nuclear import of Hes protein (the identical model was published again in [20]). Models using explicit time delay have been widely used to model oscillations, for example, in respiration patterns and hematopoiesis [21], p53 [22<sup>\*</sup>], somitogenesis [23], insulin secretion [24], and the hypothalamic-pituitary-adrenal axis [25]. In our opinion, however, using explicit delays is somewhat ad hoc and we find it more satisfying to model the specific molecular processes that produce delay.

The conceptually simplest way to do this is case (2) above. For example, an oscillator formed by a loop of six nodes, involving three proteins, and their mRNA, each protein inhibiting transcription of the next (Figure 2b.2), has been modeled and experimentally realized in Ref. [26<sup>\*</sup>]. Another synthetic circuit that oscillates due to many intermediate steps was constructed in [27]. More complex is case (3) where one regulator affects only the next when it reaches a threshold concentration. The delay arises from the time taken for the threshold to be reached. Simple models of oscillations have implemented this using either a highly cooperative interaction with a large ( $\geq 8$ ) Hill coefficient [28] or positive feedback loops [29<sup>\*</sup>,30]. Both large Hill coefficients and positive feedback result in sigmoidal, switch-like responses and hence a threshold concentration below which the response is essentially zero. Combining positive and negative feedback has the advantage of making oscillations more robust and yet tunable [29<sup>\*</sup>,30], and has been used to model many phenomena, including cell cycles [31,32], circadian rhythms [2,16,33], division site localization in *Escherichia coli* [34], and p53 oscillations [10<sup>\*\*</sup>,35,36], as well as to design synthetic oscillators [37,38,27].

The next section elaborates on case (4), saturated degradation, which we find particularly interesting as it is seen in NF- $\kappa$ B, p53 and Wnt signaling. Note however that these time-delay mechanisms are not mutually exclusive. Systems typically use several of these mechanisms, each contributing to the overall delay.

### Saturated degradation models of NF- $\kappa$ B, p53 and Wnt signaling

The key negative feedback loop underlying sustained and damped oscillations in NF- $\kappa$ B, in both wild-type and genetically modified cells, involves the inhibitor protein I $\kappa$ B $\alpha$ . The feedback loop has two legs (Figure 2c): first, NF- $\kappa$ B activates I $\kappa$ B $\alpha$  production, second, I $\kappa$ B $\alpha$  inhibits NF- $\kappa$ B by binding to it and sequestering it in the cytoplasm. Leg (i) is active when there is little I $\kappa$ B $\alpha$ ,

so most NF- $\kappa$ B is free to enter the nucleus, causing I $\kappa$ B $\alpha$  levels to rise. Free NF- $\kappa$ B levels then fall rapidly as it gets bound to newly synthesized I $\kappa$ B $\alpha$ . In the model of Figure 2c [39], the amount of NF- $\kappa$ B–I $\kappa$ B $\alpha$  complex has a Michaelis–Menten form:  $N_c I / (K_I + I)$ , where  $I$  is the I $\kappa$ B $\alpha$  concentration and  $N_c$  is the total cytoplasmic NF- $\kappa$ B concentration. The binding is strong, that is,  $K_I$  is small, so I $\kappa$ B $\alpha$  levels quickly become large enough to saturate NF- $\kappa$ B, at which point the amount of complex becomes equal to  $N_c$  and independent of  $I$ . This is leg (ii) of the feedback. Now there is no further production of I $\kappa$ B $\alpha$ , so its levels will eventually fall. However, I $\kappa$ B $\alpha$  molecules that are bound to NF- $\kappa$ B are more susceptible to IKK-dependent degradation (due to stabilization of I $\kappa$ B $\alpha$  by NF- $\kappa$ B [40]) so the degradation rate (second term in the  $dI/dt$  equation in Figure 2c) depends not on the amount of I $\kappa$ B $\alpha$  present, but on the amount of the complex. Because the complex is saturated and equal to  $N_c$  most of the time (see blue curve in Figure 2c, bottom panel) we call this ‘saturated degradation’. If, instead, IKK-inducing stimuli led to equal degradation of both free and complexed I $\kappa$ B $\alpha$ , then the degradation rate would be proportional to  $I$ , which would make  $I$  fall exponentially fast. By contrast, with saturated degradation  $I$  falls slower than exponentially, resulting in the more rounded shape of  $I$  vs. time seen in Figure 2c (green curve). This provides a sufficient time delay to generate oscillations. The model of Figure 2c is of course a simplified one and it is important to check whether the assumptions made in simplifying the system are reasonable. For example, in wild-type cells, free I $\kappa$ B $\alpha$  is also degraded but the model ignores this. This could be included in the model as an additional degradation term, that is, proportional to  $I$  leading to an exponential, non-saturated, decrease of I $\kappa$ B $\alpha$  levels, which may neutralize the time delay provided by the saturated degradation pathway. To see which pathway of degradation is more important, one must compare the half-life of free I $\kappa$ B $\alpha$  with the rate of the NF- $\kappa$ B–I $\kappa$ B $\alpha$  complex formation. Using numbers for wild-type cells from Ref. [6<sup>\*\*</sup>,41], we find the average time for complex formation is of the order of tens of seconds, whereas the half-life of free I $\kappa$ B $\alpha$  is several minutes. Thus, we expect that saturated degradation is an important source of time delay in the NF- $\kappa$ B–I $\kappa$ B $\alpha$  feedback loop despite the presence of other non-saturated degradation pathways.

The p53–Mdm2 feedback loop (Figure 2d) is very similar: first, p53 activates Mdm2 production, second, Mdm2 inhibits p53 by binding to it. Mdm2 also causes poly-ubiquitination and thereby degradation of p53, again resulting in saturated degradation. Here, it is the transcription factor (TF) that has saturated degradation, rather than the inhibitor. This is the opposite to what happens with NF- $\kappa$ B, but the model of [42], a simplified version of which is shown in Figure 2d, shows that it does not matter for generating oscillations — the

time delay in this model occurs when p53 levels are high and fall relatively slowly, rather than when the inhibitor levels are high (see Figure 2d). Because p53 levels remain high for longer, Mdm2 levels also rise much higher than they would without saturated degradation. The extra time required for Mdm2 levels to fall in leg (ii) of the feedback provides an addition delay that helps oscillations. The half-life of p53 in the absence of Mdm2 is of the order of hours, whereas the average time for p53-Mdm2 complex formation is of the order of seconds (see [42] and references therein). So again, even though the p53-Mdm2 interaction is quite transient, we expect the saturated complex formation to be the dominant pathway for p53 degradation, rather than non-saturated pathways.

A third system where saturated degradation has been used to model oscillations is Wnt signaling (Figure 2e) [43,44]. Wnt and  $\beta$ -catenin are upstream controllers of all oscillating signals in the presomitic mesoderm [45] so it is useful to study the negative feedback loop involving Wnt,  $\beta$ -catenin and Axin2. The model of this loop in Figure 2e demonstrates an interesting variation on the ones in Figure 2c,d, showing that saturated degradation need not arise from the same complex that results in inhibition of the TF. In the model, Axin2 binds to the TF  $\beta$ -catenin and separately to the Wnt-activated LRP receptor complex. The former provides the negative feedback but the binding is weak and not saturated. The latter complex has a much smaller dissociation constant, that is, larger binding strength, and results in saturated degradation of Axin2. Thus, negative feedback and saturated degradation are separate ingredients that can be implemented independently in an oscillator.

Other simple models of oscillations using saturated degradation can be found in Refs. [46\*,47–49]. It is interesting that saturated degradation typically produces quite spiky oscillations, which have the advantage that the signaling molecule can achieve high levels without having to be produced at a high rate all the time. This brings us to the question of what these oscillations are useful for.

### Potential physiological role of oscillations

In some systems periodicity is an obvious requirement. A periodic spatial pattern is clearly necessary for proper somite spacing and temporal oscillations in Wnt and Notch targets are a way of generating the spatial periodicity [50\*,51]. Circadian clocks in cyanobacteria are useful for entraining metabolism, photosynthesis, cell division and global gene expression to the day–night cycle [52]. However, in NF- $\kappa$ B and p53 it is not obvious that the oscillations per se are important for the physiological response. For example, it has been suggested that p53 pulses might be a byproduct of pulsatility in ATM, an

upstream regulator of p53 required for proper DNA damage repair [53].

From the opposite angle, what benefit could oscillations provide in helping NF- $\kappa$ B and p53 produce gene expression patterns specific to distinct stimuli? One idea is that signals with complex temporal variation contain more information than steady-state signals and therefore can control downstream genes more subtly [39,54,53]. We elaborate on this below.

### Encoding information in oscillatory signals

Steady-state signals have a single adjustable characteristic, the level of the signal, while oscillations have many — average, amplitude, time period, spikiness, spike width, spike symmetry (see Figure 3a). Oscillations in NF- $\kappa$ B or p53 could thus encode more information than steady-states about which stimulus was triggering the system provided: first, different stimuli affect different parameters, and second, changing different parameters affects oscillation characteristics differently.

(i) In the p53 system, we know that different stimuli affect different sets of parameters. DNA damage affects Mdm2 activity and stability, hypoxia additionally alters the transcription rate of Mdm2, while other triggers like Nutlin change only the binding strength between p53 and Mdm2 (see references in [42]). In the NF- $\kappa$ B model, many triggers act through the IKK level which affects the degradation rate of I $\kappa$ B $\alpha$ , but different stimuli produce different profiles of IKK and thereby NF- $\kappa$ B [55,8].

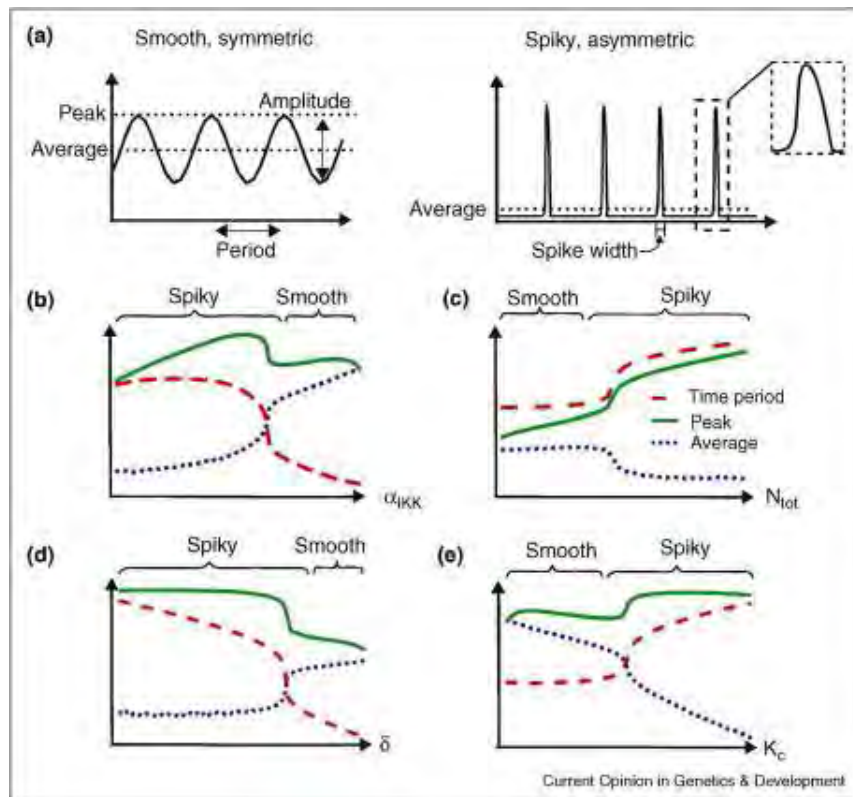
(ii) Figure 3b–e show that for the simple model of NF- $\kappa$ B described above, changing different parameters does indeed affect oscillation characteristics differently. The plots show that there are parameter regimes where one of the three characteristics, time period, average and peak, can vary while the other two remain constant. However, not all characteristics can be independently varied because there are correlations. For example, spikiness is correlated with larger time periods, lower averages and asymmetry of the spike shape. Experiments have also shown that characteristics of the temporal profile of nuclear NF- $\kappa$ B concentration, for example, the steepness of the initial increase and the later decline, are under the control of different regulators [56\*\*].

Similar behavior is seen in the p53 and Wnt models, so one can conclude that an oscillatory signal produced from a simple negative feedback loop with saturated degradation can indeed encode more information than steady-state signals.

### Decoding information from oscillatory signals

Next, it is necessary that different genes should respond to different characteristics of the oscillations. Consider

Figure 3



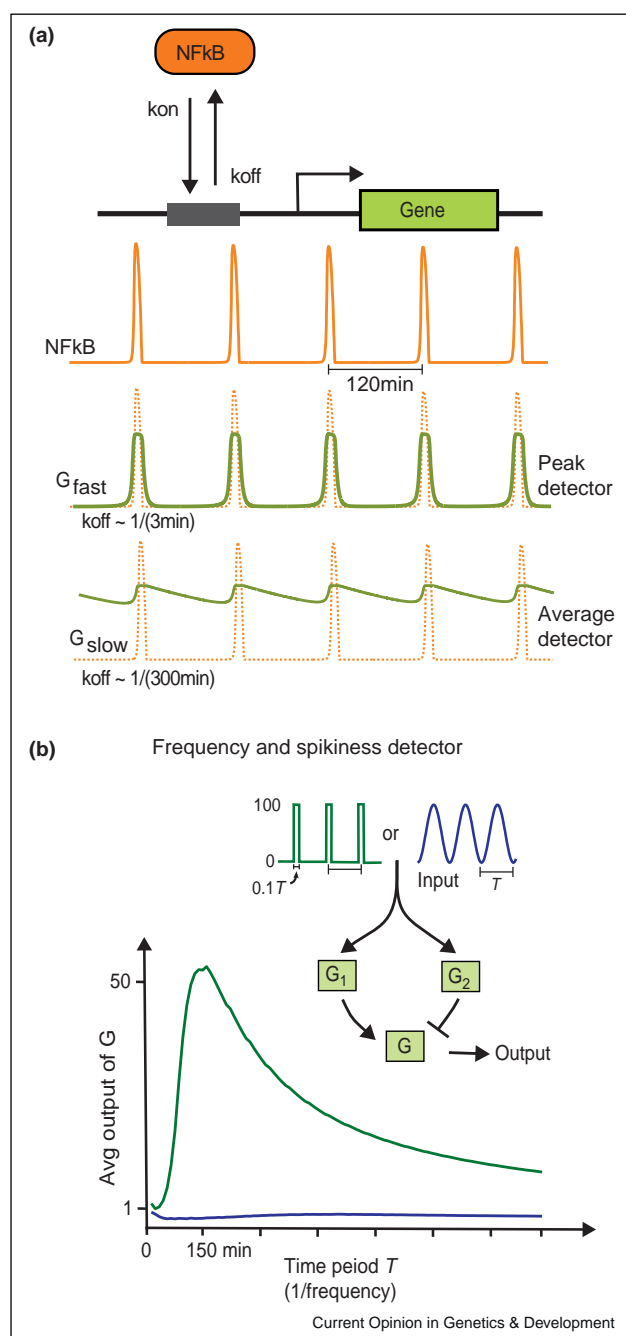
Encoding information in oscillations. (a) Characteristics of oscillatory signals that could potentially be used to encode information about external stimuli. (b–e) Time period (red dashed line), average (blue dotted line) and peak levels (green solid line) of NF- $\kappa$ B oscillations as selected parameters are varied in the model of NF- $\kappa$ B from Figure 2c. The braces mark the parameter ranges that produce smooth and spiky oscillations. Peak levels can be varied independently of time period and average by changing  $N_{tot}$  when oscillations are smooth. Similarly, average levels can be varied independently of the other two by changing  $k_c$  when oscillations are smooth. Finally, time period can be varied while average and peak remain constant by changing  $\delta$  when oscillations are spiky. For all the parameter ranges shown, we also found that increased spikiness is correlated with larger time periods, lower averages and asymmetry of the spike shape.

the simplest case where an oscillating TF binds to a single operator site (Figure 4a). Two parameters describe the binding: the association ( $k_{on}$ ) and dissociation ( $k_{off}$ ) rate constants. Figure 4a shows that the expression of a gene,  $G_{fast}$ , with sufficiently large  $k_{on}$  and  $k_{off}$ , will closely follow the oscillations. If the stimulus changes the peak level of the oscillations, the expression of this gene will follow that change;  $G_{fast}$  is a peak-detector. By contrast, the activity of  $G_{slow}$ , which has a much smaller  $k_{off}$ , will not follow the oscillations because its expression has not time to decline much before the next spike occurs. Thus, this gene's expression averages over many spikes.  $G_{slow}$  is therefore an average detector. Note that if we look at the concentration of the proteins encoded by  $G_{fast}$  and  $G_{slow}$ , then we also have to take into account their half-lives and those of the mRNAs. Thus, for example, if  $G_{fast}$  produces a very long lived mRNA or protein then the protein concentration would follow changes in the average of the input oscillations rather than the peak. By

contrast, even if  $G_{slow}$  produces a short-lived mRNA or protein it would remain an average detector.

The slightly more complex circuit of Figure 4b is a frequency detector. Two genes with twofold different  $k_{off}$  values respectively activate and inhibit a third, output gene. The average steady-state expression of this output gene has a maximum for a certain 'resonance' frequency (Figure 4b). Away from this resonance, especially for larger frequencies, the steady-state output falls dramatically. Here, the protein level will also show a similar resonance irrespective of its half-life because, assuming there is no complex translational or post-translational regulation, the average steady-state concentration of a protein is proportional to the average steady-state expression level of its gene. Interestingly, the spiky nature of the input oscillations is very important for this frequency resonance. Smooth oscillations with exactly the same time period and amplitude show a much weaker

Figure 4



Decoding information from oscillatory signals. (a) Regulation of gene expression by a transcription factor (TF) that associates with a single operator site with a rate constant  $k_{on}$  and dissociates with a rate constant  $k_{off}$ . We make the (conservative, for NF-κB and p53) assumption that the maximal concentration of the TF is 100-fold the operator concentration. Gene  $G_{fast}$  has  $k_{on} = 0.1 \text{ min}^{-1}$  per operator site,  $k_{off} = 0.3 \text{ min}^{-1}$  and follows the oscillations of the TF closely (green). Therefore the peak of  $G_{fast}$  expression follows changes in the peak level of oscillations.  $G_{slow}$  has  $k_{on} = 0.1 \text{ min}^{-1}$  per operator site,  $k_{off} = 0.003 \text{ min}^{-1}$  and its peak expression follows the average level of the oscillations. Note the vastly different responses in the two cases. (b)  $G_1$  ( $k_{on} = 0.1 \text{ min}^{-1}$  per operator site,  $k_{off} = 0.03 \text{ min}^{-1}$ ) and  $G_2$

steady-state response. This circuit is therefore a spikiness-detector as well.

## Outlook

Some of the interesting questions this discussion raises for future research on oscillatory control are:

(A) Relating to encoding information in oscillations:

- Which parameters, and which characteristics of oscillations, do different stimuli affect in the NF-κB and p53 systems?
- Can additional feedback loops enhance the encoding abilities of oscillations?

(B) Relating to decoding information from oscillations:

- Can other decoding circuits be constructed to count, say, the number of spikes in a signal, or distinguish between symmetry and asymmetry, or other characteristics of oscillations?
- Do any such circuits exist downstream of NF-κB or p53?

The ideal experiment to evaluate the necessity of oscillations would require being able to control the frequency, number and width of spikes produced when NF-κB or p53 is triggered, and to see how this affects the physiological response. Exactly such experiments have shown that varying the frequency of oscillations in calcium signaling [4] and hormone secretion [5] changes the physiological response. Similar experiments on the NF-κB system are just beginning to be actualized, and have reported some dependence of gene expression on frequency of oscillations [57]. Other experiments have shown the opposite, that expression of some NF-κB targets is unaffected in non-oscillating mutants [58]. These results are not necessarily contradictory, but until more comprehensive experiments become feasible for the NF-κB and p53 systems it might be useful to examine more carefully the genetic circuits downstream of these TFs. A frequency-detector circuit downstream of NF-κB or p53 would be a strong clue that oscillations are important for the physiological response.

( $k_{on} = 0.1 \text{ min}^{-1}$  per operator site,  $k_{off} = 0.06 \text{ min}^{-1}$ ) respectively activate and inhibit a third, output gene  $G$ :  $dG/dt = k(G_1/(1 + G_1/K_1))(1/(1 + (G_2/K_2)^2)) - \gamma G$ . For this circuit, we can find parameter values for the  $dG/dt$  equation such that the average expression of  $G$  has a maximum ('resonance') when the time period is around  $T = 150 \text{ min}$ , when the input is spiky square-pulse oscillations (green curve). The position of this maximum can be tuned by varying  $k_{off}$  values of  $G_1$  and  $G_2$ . With the same parameter values, when the input is smooth sine-wave oscillations of the same amplitude the response is much weaker (blue curve). See [47,71] for other frequency-detector circuits, involving protein phosphorylation.



## Acknowledgements

We thank Ian Dodd for many useful suggestions that greatly improved the manuscript, and the Danish National Research Foundation for funding.

## References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- of special interest
- of outstanding interest

1. Kitayama Y, Nishiwaki T, Terauchi K, Kondo T: **Dual kaic-based oscillations constitute the circadian system of cyanobacteria.** *Genes Dev* 2008, **22**:1513-1521.
2. Dong G, Golden SS: **How a cyanobacterium tells time.** *Curr Opin Microbiol* 2008, **11**:541-546.
3. Lu Y, Cross FR: **Periodic cyclin-cdk activity entrains an autonomous cdc14 release oscillator.** *Cell* 2010, **141**:268-279.
4. Zhu L, Luo Y, Chen T, Chen F, Wang T, Hu Q: **Ca<sup>2+</sup> oscillation frequency regulates agonist-stimulated gene expression in vascular endothelial cells.** *J Cell Sci* 2008, **121**:2511-2518.
5. Ciccone NA, Xu S, Lacza CT, Carroll RS, Kaiser UB: **Frequency-dependent regulation of follicle-stimulating hormone beta by pulsatile gonadotropin-releasing hormone is mediated by functional antagonism of bzip transcription factors.** *Mol Cell Biol* 2010, **30**:1028-1040.
6. Hoffmann A, Levchenko A, Scott ML, Baltimore D: **The I $\kappa$ B-NF- $\kappa$ B signaling module: temporal control and selective gene activation.** *Science* 2002, **298**:1241-1245.
- Experiments showing NF- $\kappa$ B nuclear oscillations in bulk populations of TNF stimulated fibroblasts. The paper also contains the first computational model of the NF- $\kappa$ B signaling system.
7. Nelson DE, Ihekweba AEC, Elliott M, Johnson JR, Gibney CA, Foreman BE, Nelson G, See V, Horton CA, Spiller DG *et al.*: **Oscillations in NF- $\kappa$ B signaling control the dynamics of gene expression.** *Science* 2004, **306**:704-708.
- Observation of nuclear NF- $\kappa$ B oscillations in individual transiently transfected HeLa cells using fluorescence microscopy. Oscillations were later also observed in stably transfected cells in Ref. [59].
8. Lee TK, Denny EM, Sanghvi JC, Gaston JE, Maynard ND, Hughey JJ, Covert MW: **A noisy paracrine signal determines the cellular NF- $\kappa$ B response to lipopolysaccharide.** *Sci Signal* 2009, **2**:ra65.
9. Bartfeld S, Hess S, Bauer B, Machuy N, Ogilvie LA, Schuchhardt J, Meyer TF: **High-throughput and single-cell imaging of NF- $\kappa$ B oscillations using monoclonal cell lines.** *BMC Cell Biol* 2010, **11**:21.
10. Geva-Zatorsky N, Rosenfeld N, Itzkovitz S, Milo R, Sigal A, Dekel E, Yarnithky T, Polak P, Liron Y, Kam Z *et al.*: **Oscillations and variability in the p53 system.** *Mol Syst Biol* 2006:2006.0033.
- Measurement of oscillations of fluorescently tagged p53 and Mdm2 proteins in gamma-irradiated human cells. The paper also examines several classes of models of the p53-Mdm2 negative feedback loop.
11. Hamstra DA, Bhojani MS, Griffin LB, Laxman B, Ross BD, Rehemtulla A: **Real-time evaluation of p53 oscillatory behavior in vivo using bioluminescent imaging.** *Cancer Res* 2006, **66**:7482-7489.
- The paper uses bioluminescent imaging to measure p53 oscillations triggered by ionizing radiation. Remarkably, they do this in vivo in intact mice.
12. Aulehla A, Herrmann BG: **Segmentation in vertebrates: clock and gradient finally joined.** *Genes Dev* 2004, **18**:2060-2067.
13. Dequeant ML, Glynn E, Gaudenz K, Wahl M, Chen J, Mushegian A, Pourquie O: **A complex oscillating network of signaling genes underlies the mouse segmentation clock.** *Science* 2006, **314**:1595-1598.
- Microarray studies of the mouse embryos during somite formation revealed a large number of oscillating genes. Genes in the Wnt pathway were found to be oscillating out of phase with genes of the Notch/FGF pathways.
14. Aulehla A, Wiegand W, Baubet V, Wahl MB, Deng C, Taketo M, Kewandoski M, Pourquie O: **A beta-catenin gradient links the clock and wavefront systems in mouse embryo segmentation.** *Nat Cell Biol* 2008, **10**:168-210.
15. Ingolia NT, Murray AW: **The ups and downs of modeling the cell cycle.** *Curr Biol* 2004, **14**:R771-R777.
16. Ferrell JE, Pomeroy JR, Kim SY, Trunnell NB, Xiong W, Huang CY, Machleder EM Jr: **Simple, realistic models of complex biological processes: positive feedback and bistability in a cell fate switch and a cell cycle oscillator.** *FEBS Lett* 2009, **583**:3999-4005.
17. Sneppen K, Krishna S, Semsey S: **Simplified models of biological networks.** *Annu Rev Biophys* 2010, **39**:43-59.
18. Tiana G, Krishna S, Pigolotti S, Jensen MH, Sneppen K: **Oscillations and temporal signalling in cells.** *Phys Biol* 2007, **4**:R1-R17.
19. Jensen MH, Sneppen K, Tiana G: **Sustained oscillations and time delays in gene expression of protein hes1.** *FEBS Lett* 2003, **541**:176-177.
20. Monk NAM: **Oscillatory expression of hes1, p53, and NF- $\kappa$ B driven by transcriptional time delays.** *Curr Biol* 2003, **13**:1409-1413.
21. Mackey MC, Glass L: **Oscillation and chaos in physiological control systems.** *Science* 1977, **197**:287-289.
22. Tiana G, Sneppen K, Jensen MH: **Time delay as a key to apoptosis induction in the p53 network.** *Europhys J B* 2002, **29**:135-140.
- The first use of delay differential equations to model oscillations in a genetic circuit consisting of a negative feedback loop with transcription/translation time delay. This model was for p53-Mdm2 oscillations, and a similar approach was used soon after to model oscillations in zebrafish somitogenesis (Ref. [23]) and auto-regulated Hes1 expression (Ref. [19]).
23. Lewis J: **Autoinhibition with transcriptional delay: a simple mechanism for the zebrafish somitogenesis oscillator.** *Curr Biol* 2003, **13**:1398-1408.
24. Li J, Kuang Y, Mason CC: **Modeling the glucose-insulin regulatory system and ultradian insulin secretory oscillations with two explicit time delays.** *J Theor Biol* 2006, **242**:722-735.
25. Walker JJ, Terry JR, Lightman SL: **Origin of ultradian pulsatility in the hypothalamic-pituitary-adrenal axis.** *Proc Biol Sci* 2010, **277**:1627-1633.
26. Elowitz MB, Leibler S: **A synthetic oscillatory network of transcriptional regulators.** *Nature* 2000, **403**:335-338.
- Experimental realization and computational modeling of the repressilator: the first synthetic genetic oscillator consisting of three proteins in a negative feedback loop, each repressing transcription of the next.
27. Stricker J, Cookson S, Bennett MR, Mather WH, Tsimring LS, Hasty J: **A fast, robust and tunable synthetic gene oscillator.** *Nature* 2008, **456**:516-519.
28. Goodwin BC: **Oscillatory behavior in enzymatic control processes.** *Adv Enzym Regul* 1965, **3**:425-438.
29. Tsai TY, Choi YS, Ma W, Pomeroy JR, Tang C, Ferrell JE Jr: **Robust, tunable biological oscillations from interlinked positive and negative feedback loops.** *Science* 2008, **321**:126-129.
- A theoretical study showing that relaxation oscillators consisting of a positive and a negative feedback loop have the advantages that their frequency can be varied over a wide range without affecting the amplitude, and that they are more robust than simple negative feedback oscillators. Ref. [30] shows that the spikiness and symmetry of these oscillations is also easily tunable.
30. Krishna S, Semsey S, Jensen MH: **Frustrated bistability as a means to engineer oscillations in biological systems.** *Phys Biol* 2009, **6**:1-8.
31. Pomeroy JR, Kim SY, Ferrell JE Jr: **Systems-level dissection of the cell-cycle oscillator: bypassing positive feedback produces damped oscillations.** *Cell* 2005, **122**:565-578.



32. Gerard C, Goldbeter A: **Temporal self-organization of the cyclin/cdk network driving the mammalian cell cycle.** *Proc Natl Acad Sci (USA)* 2009, **106**:21643-21648.
  33. Cookson NA, Tsimring LS, Hasty J: **The pedestrian watchmaker: genetic clocks from engineered oscillators.** *FEBS Lett* 2009, **583**:3931-3937.
  34. Howard M, Kruse K: **Cellular organization by self-organization: mechanisms and models for min protein dynamics.** *J Cell Biol* 2005, **168**:533-536.
  35. Ciliberto A, Novak B, Tyson JJ: **Steady states and oscillations in the p53/mdm2 network.** *Cell Cycle* 2005, **4**:488-493.
  36. Puzynski K, Hat B, Lipniacki T: **Oscillations and bistability in the stochastic model of p53 regulation.** *J Theor Biol* 2008, **254**:452-465.
  37. Atkinson MR, Savageau MA, Myers JT, Ninfa AJ: **Development of genetic circuitry exhibiting toggle switch or oscillatory behavior in *Escherichia coli*.** *Cell* 2003, **113**:597-607.
  38. Fung E, Wong WW, Suen JK, Bulter T, Gu Lee S, Liao JC: **A synthetic gene-metabolic oscillator.** *Nature* 2005, **435**:118-122.
  39. Krishna S, Jensen MH, Sneppen K: **Minimal model of spiky oscillations in NF- $\kappa$ B signaling.** *Proc Natl Acad Sci (USA)* 2006, **103**:10840-10845.
  40. Mathes E, O'Dea EL, Hoffmann A, Ghosh G: **NF- $\kappa$ B dictates the degradation pathway of I $\kappa$ B $\alpha$ .** *EMBO J* 2008, **27**:1357.
  41. O'Dea EL, Barken D, Peralta RQ, Tran KT, Werner SL, Kearns JD, Levchenko A, Hoffmann A: **A homeostatic model of I $\kappa$ B metabolism to control constitutive NF- $\kappa$ B activity.** *Mol Syst Biol* 2007, **3**:111.
  42. Hunziker A, Jensen MH, Krishna S: **Stress-specific response of the p53-mdm2 feedback loop.** *BMC Syst Biol*.
  43. Goldbeter A, Pourquie O: **Modeling the segmentation clock as a network of coupled oscillations in the Notch, Wnt and FGF signaling pathways.** *J Theor Biol* 2008, **252**:574-585.
  44. Jensen PB, Pedersen L, Krishna S, Jensen MH: **A wnt oscillator model for somitogenesis.** *Biophys J* 2010, **98**:943-950.
  45. Dunty WC, Biris KK, Chalamalasetty RB, Taketo MM, Lewandoski M, Yamaguchi TP: **Wnt3a/ $\beta$ -catenin signaling controls posterior body development by coordinating mesoderm formation and segmentation.** *Development* 2008, **135**:85-94.
  46. Bliss RD, Painter PR, Marr AG: **Role of feedback inhibition in stabilizing the classical operon.** *J Theor Biol* 1982, **97**:177-193.
- The first use of saturated degradation (a term introduced in Ref. [39]) to generate oscillations in a simple three-variable dynamical system. The model modifies the equally elegant Goodwin oscillator (Ref. [28]) to avoid the latter's use of a large Hill coefficient.
47. Goldbeter A, Bupont G, Berridge MJ: **Minimal model for signal-induced Ca<sup>2+</sup> oscillations and for their frequency encoding through protein phosphorylation.** *Proc Natl Acad Sci (USA)* 1990, **87**:1461-1465.
  48. Gerard C, Gonze D, Goldbeter A: **Dependence of the period on the rate of protein degradation in minimal models for circadian oscillations.** *Phil Trans R Soc A* 2009, **367**:4665-4683.
  49. Sneppen K, Lizana L, Jensen MH, Pigolotti S, Otzen D: **Modeling proteasome dynamics in parkinsons disease.** *Phys Biol* 2009, **6**:036005.
  50. Cooke J, Zeeman EC: **A clock and wavefront model for control of the number of repeated structures during animal morphogenesis.** *J Theor Biol* 1976, **58**:455-476.
- This paper introduced the clock-wavefront model to explain how temporal oscillations of gene expression could interact with morphogen gradients in the presomitic mesoderm to produce periodically spaced somites.
51. Pourquie O: **The segmentation clock: converting embryonic time into spatial pattern.** *Science* 2003, **301**:328-330.
  52. Yang Q, Pando BF, Dong G, Golden SS, van Oudenaarden A: **Circadian gating of the cell cycle revealed in single cyanobacterial cells.** *Science* 2010, **327**:1522-1526.
  53. Batchelor E, Loewer A, Lahav G: **The ups and downs of p53: understanding protein dynamics in single cells.** *Nat Rev Cancer* 2009, **9**:371-377.
  54. Kobayashi T, Kageyama R: **Dynamic advances in NF- $\kappa$ B signaling analysis.** *Cell Biol* 2009, **2**:1-2.
  55. Werner SL, Barken D, Hoffmann A: **Stimulus specificity of gene expression programs determined by temporal control of ikk activity.** *Science* 2005, **309**:1857-1861.
  56. Werner SL, Kearns JD, Zadorozhnaya V, Lynch C, O'Dea E, Boldin MP, Ma A, Baltimore D, Hoffmann A: **Encoding NF- $\kappa$ B temporal control in response to tnf: distinct roles for the negative regulators ikk $\alpha$  and a20.** *Genes Dev* 2008, **22**:2093-2101.
- The first demonstration that aspects of the temporal profile of nuclear NF- $\kappa$ B concentration are independently controlled by different regulators. This, in combination with the results of Ref. [57], shows that the complex temporal variation of nuclear NF- $\kappa$ B encodes information that can be used to generate gene expression patterns specific to different triggering stimuli.
57. Ashall L, Horton CA, Nelson DE, Paszek P, Harper CV, Sillitoe K, Ryan S, Spiller DG, Unitt JF, Broomhead DS et al.: **Pulsatile stimulation determines timing and specificity of NF- $\kappa$ B-dependent transcription.** *Science* 2009, **324**:242-246.
- A first attempt to manipulate the frequency of NF- $\kappa$ B oscillations and measure the effect on NF- $\kappa$ B targets. A comprehensive study, using such techniques, of how frequency and other characteristics affect gene expression patterns is exactly what is required to clarify whether oscillations per se are important for the physiological response.
58. Barken D, Wang CJ, Kearns J, Cheong R, Hoffmann A, Levchenko A: **Comment on 'oscillations in NF- $\kappa$ B signaling control the dynamics of gene expression'.** *Science* 2005, **308**:52a.
  59. Cheong R, Levchenko A: **Wires in the soup: quantitative models of cell signaling.** *Trend Cell Biol* 2008, **18**:112-118.
  60. Ankers JM, Spiller DG, White MRH CV, Harper: **Spatio-temporal protein dynamics in single living cells.** *Curr Opin Biotechnol* 2008, **19**:375-380.
  61. Harris SL, Levine AJ: **The p53 pathway: positive and negative feedback loops.** *Oncogene* 2005, **24**:2899.
  62. Tyson JJ: **Another turn for p53.** *Mol Sys Biol* 2006, **2**:0032.
  63. Aulehla A, Pourquie O: **Oscillating signaling pathways during embryonic development.** *Curr Opin Cell Biol* 2008, **20**:632-637.
  64. Ozbudak EM, Pourquie O: **The vertebrate segmentation clock: the tip of the iceberg.** *Curr Opin Genet Dev* 2008, **18**:317-323.
  65. Lipniacki T, Paszek P, Brasier AR, Luxon B, Kimmel M: **Mathematical model of NF- $\kappa$ B regulatory module.** *J Theor Biol* 2004, **228**:195-215.
  66. Cheong R, Hoffmann A, Levchenko A: **Understanding NF- $\kappa$ B signaling via mathematical modeling.** *Mol Sys Biol* 2008, **4**:1-11.
  67. Ma L, Wagner J, Rice JJ, Hu W, Levine AJ, Stolovitzky G: **A plausible model for the digital response of p53 to dna damage.** *Proc Natl Acad Sci (USA)* 2005, **102**:14266-14271.
  68. Batchelor E, Mock CS, Bhan I, Loewer A, Lahav G: **Recurrent initiation: a mechanism for triggering p53 pulses in response to dna damage.** *Mol Cell* 2008, **30**:277-289.
  69. Lee E, Salic A, Kruger R, Heinrich R, Kirschner MW: **The roles of apc and axin derived from experimental and theoretical analysis of the wnt pathway.** *PLoS Biol* 2003, **1**:10.
  70. Rodriguez-Gonzalez JG, Santillan M, Fowler AC, Mackey MC: **The segmentation clock in mice: interaction between the Wnt and Notch signalling pathways.** *J Theor Biol* 2007, **248**:37-47.
  71. Smolen P, Baxter DA, Byrne JH: **Frequency selectivity, multistability and oscillations emerge from models of genetic regulatory systems.** *Am J Physiol Cell Physiol* 1998, **274**:531-542.

## APPENDIX

### D

#### T1D paper

Supplementary files 2-6 and 8 are excel files, which can be downloaded from  
<http://www.biomedcentral.com/1471-2164/12/97>

RESEARCH ARTICLE

Open Access

# Independent component and pathway-based analysis of miRNA-regulated gene expression in a model of type 1 diabetes

Claus H Bang-Berthelsen<sup>1,2†</sup>, Lykke Pedersen<sup>3†</sup>, Tina Fløyel<sup>1,2</sup>, Peter H Hagedorn<sup>4,7</sup>, Titus Gylvin<sup>5</sup>, Flemming Pociot<sup>1,2,6\*</sup>

## Abstract

**Background:** Several approaches have been developed for miRNA target prediction, including methods that incorporate expression profiling. However the methods are still in need of improvements due to a high false discovery rate. So far, none of the methods have used independent component analysis (ICA). Here, we developed a novel target prediction method based on ICA that incorporates both seed matching and expression profiling of miRNA and mRNA expressions. The method was applied on a cellular model of type 1 diabetes.

**Results:** Microarray profiling identified eight miRNAs (miR-124/128/192/194/204/375/672/708) with differential expression. Applying ICA on the mRNA profiling data revealed five significant independent components (ICs) correlating to the experimental conditions. The five ICs also captured the miRNA expressions by explaining >97% of their variance. By using ICA, seven of the eight miRNAs showed significant enrichment of sequence predicted targets, compared to only four miRNAs when using simple negative correlation. The ICs were enriched for miRNA targets that function in diabetes-relevant pathways e.g. type 1 and type 2 diabetes and maturity onset diabetes of the young (MODY).

**Conclusions:** In this study, ICA was applied as an attempt to separate the various factors that influence the mRNA expression in order to identify miRNA targets. The results suggest that ICA is better at identifying miRNA targets than negative correlation. Additionally, combining ICA and pathway analysis constitutes a means for prioritizing between the predicted miRNA targets. Applying the method on a model of type 1 diabetes resulted in identification of eight miRNAs that appear to affect pathways of relevance to disease mechanisms in diabetes.

## Background

microRNAs (miRNAs) are a class of small non-coding RNAs that function as posttranscriptional regulators of gene expression by mediating translational inhibition or mRNA degradation. miRNA bind to “seed” sites, i.e. stretches of 6-8 nucleotides in the 3' untranslated region (UTR) of their target mRNAs. miRNAs regulate various cellular processes and appear to be involved in the development of many diseases.

Most approaches for miRNA target identification rely on either one or a combination of seed matching, site

accessibility and phylogenetic conservation [1]. In addition, some have incorporated target site location, multiple target sites and expression profiling [1]. A number of target prediction methods use expression profiling of both miRNA and mRNA expressions [2-5]. Most of the approaches are based on negative correlation, i.e. reciprocal expressions of miRNAs and their degraded target mRNAs.

High-dimensional biological data, such as microarray profiling data, are often interpreted as being composed of sets of transcriptional- or activity programs that explain some, or most, of the complexity in the data [6-8]. Various methods are being applied on profiling data e.g. principal component analysis (PCA) and clustering. In the last couple of years independent component analysis (ICA) has been extensively applied on

\* Correspondence: fpoc0001@glo.regionh.dk

† Contributed equally

<sup>1</sup>Glostrup Research Institute, Glostrup University Hospital, DK-2600 Glostrup, Denmark

Full list of author information is available at the end of the article

mRNA profiling data and recently, ICA was applied on miRNA profiling data as well [9].

ICA is a computational method for separating mixed independent signals and can be used to decompose the expression matrix into independent components [10]. This decomposition has been shown to be informative in several studies [11-15], and superior to clustering and PCA [16-19]. Apparently, the representation of gene expression as a mix of independent, possibly overlapping, transcriptional programs captures the differential regulation of well-defined biological processes and metabolic pathways [19,20].

Type 1 diabetes (T1D) is an immune-mediated disease characterized by insulin deficiency due to a specific destruction of the pancreatic  $\beta$ -cells. Pro-inflammatory cytokines are involved in the destruction through induction of apoptosis [21]. The endocrine cells in the pancreatic islets all arise from the same progenitor stem cell and the maturation of the different cell types is dependent on a number of factors, such as transcription factors and miRNAs, that are activated in a tightly regulated pattern [22,23]. One of the central transcription factors in pancreatic development is Pancreatic and duodenal homeobox 1 (Pdx-1) [24-26]. In the mature endocrine cells, Pdx-1 expression is restricted to the  $\beta$ -cell, where it is important for insulin expression [27].

In the present study, we have used a model of type 1 diabetes based on  $\beta$ -cell maturation and interleukin-1 $\beta$  (IL-1 $\beta$ ) sensitivity. In response to Pdx-1 induction the cells progress from an immature to a mature  $\beta$ -cell phenotype [28]. The  $\beta$ -cell maturation is accompanied by an increased sensitivity to the toxic effects of IL-1 $\beta$  that is reflected in both transcriptional and protein expression patterns [29-31]. Genes regulated by Pdx-1 are therefore believed to be involved in the acquired IL-1 $\beta$  sensitivity, and identification of these genes would provide knowledge about the mechanisms underlying this  $\beta$ -cell specific trait. Interestingly, a study investigating genomic targets of transcription factors in a  $\beta$ -cell line suggested that several miRNAs are under Pdx-1 regulation [32]. Furthermore, a number of miRNAs have been implicated in the regulation of pancreatic development and  $\beta$ -cell differentiation [22,33-36].

Here, we have developed a novel miRNA target prediction method that is based on ICA and combines seed matching and expression profiling. We comprehensively profiled both miRNA and mRNA expressions from the type 1 diabetes model. ICA was applied on the mRNA expression data for identification of miRNA targets. Our method was compared to negative correlation. We validated our observations by use of pathway analysis and human pancreatic islet preparations.

## Results

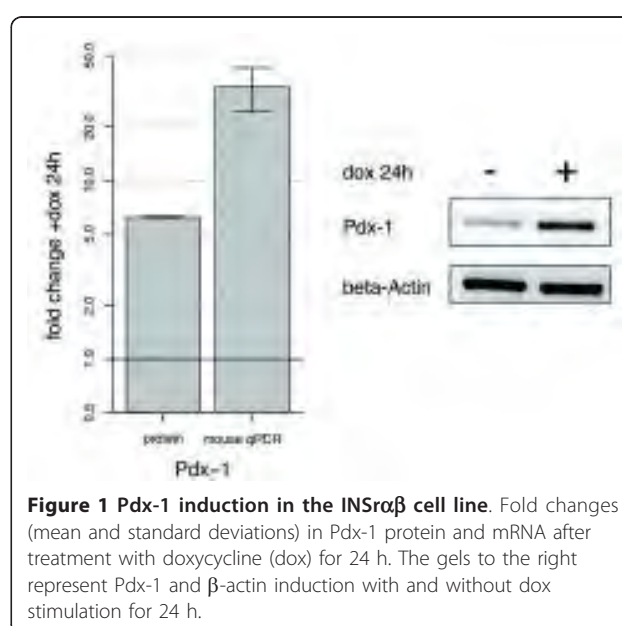
### $\beta$ -cell specific gene expression

To confirm that the INS $\alpha\beta$  cells progressed from a hybrid  $\alpha\beta$ -like phenotype to a more mature  $\beta$ -cell phenotype upon Pdx-1 induction as seen in previous studies [28,37], the expression levels of known insulin and Pdx-1 dependent genes were examined using the array data (Additional file 1). The observed expression profiles were in agreement with insulin and Pdx-1 regulations known from other studies.

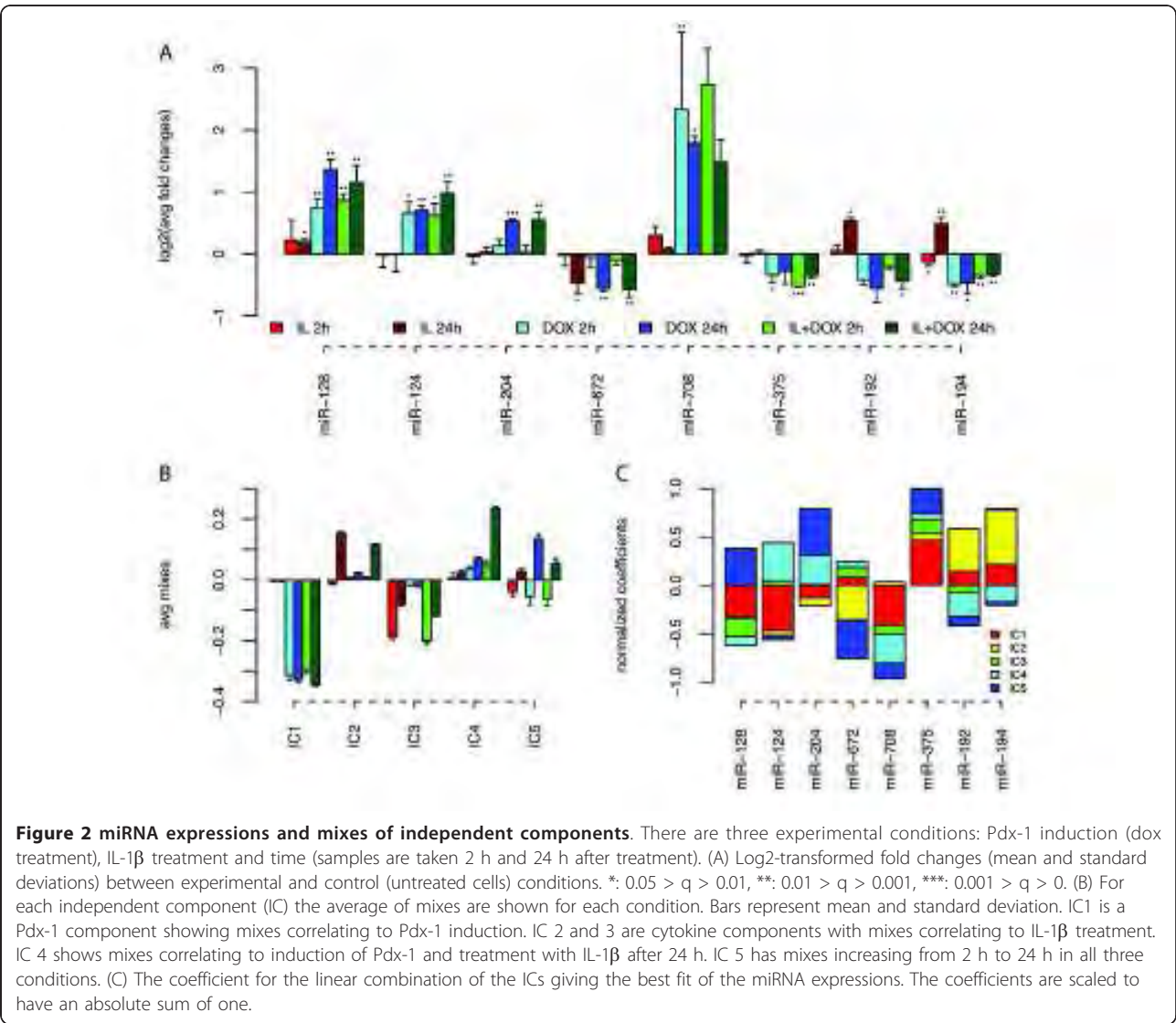
The dox-induced Pdx-1 expression was examined using qPCR and western blotting with mouse-specific primers and antibody, respectively. We found that dox treatment for 24 hours resulted in increased expression of both Pdx-1 mRNA (33 fold,  $p < 0.05$ ) and protein (6 fold,  $p < 0.05$ ) (Figure 1).

### miRNAs differentially expressed in a model of type 1 diabetes

miRNA expression profiling resulted in identification of eight miRNAs with differential expression in response to Pdx-1 induction and/or IL-1 $\beta$  exposure. All eight miRNAs (miR-124/128/192/194/204/375/672/708) showed significant ( $p < 0.05$ ) changes in expression between conditions and/or time points (Figure 2A). The eight miRNA expression profiles capture all three experimental conditions: dox-induced Pdx-1 expression, IL-1 $\beta$  exposure and time. For example, the miR-672 expression decreased significantly ( $p < 0.05$ ) with time independent of Pdx-1 induction and/or IL-1 $\beta$  treatment. The reversed pattern was seen for miR-204, though only for cells with induced Pdx-1 expression.



**Figure 1 Pdx-1 induction in the INS $\alpha\beta$  cell line.** Fold changes (mean and standard deviations) in Pdx-1 protein and mRNA after treatment with doxycycline (dox) for 24 h. The gels to the right represent Pdx-1 and  $\beta$ -actin induction with and without dox stimulation for 24 h.



A third example was the up-regulating effect of IL-1 $\beta$  treatment on the expressions of miR-128/192/194.

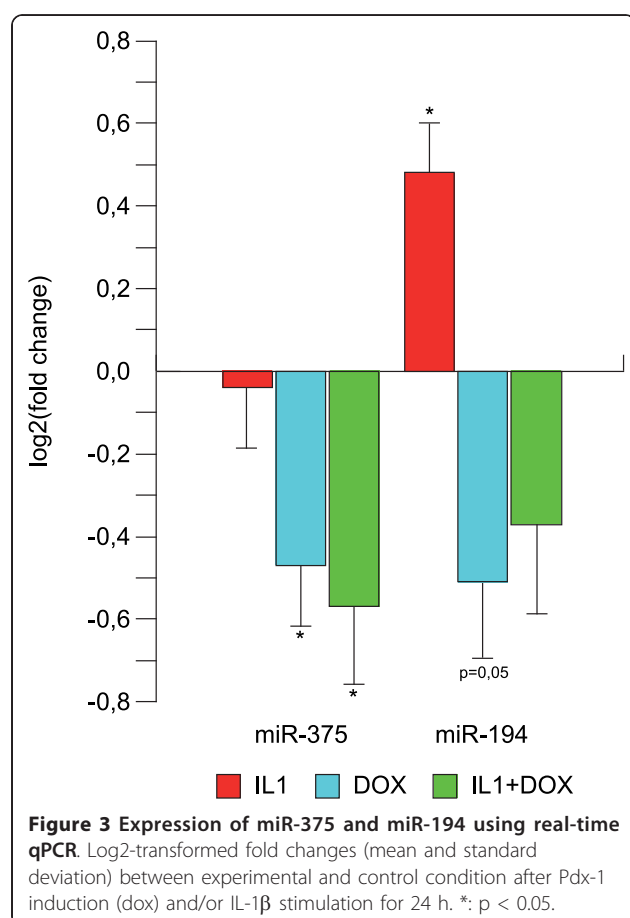
The expression of two of the eight miRNAs were validated in a new set of samples using real-time quantitative PCR. We found that the expression of miR-375 was significantly decreased in response to Pdx-1 induction (by ~50%,  $p < 0.05$ ), whereas the miR-194 expression was significantly increased in response to IL-1 $\beta$  treatment (by ~50%,  $p < 0.05$ ), see Figure 3. Furthermore, there was a tendency towards a decreased miR-194 expression in response to Pdx-1 induction (by ~50%,  $p = 0.05$ ). These expression changes confirmed the results from the array.

**miRNA target prediction using negative correlation or ICA**  
To assess the effect of the Pdx-1- and IL-1 $\beta$ -mediated miRNA activity, mRNAs were profiled in the same set of samples as used for the miRNA arrays.

Initially, targets were predicted for each of the eight miRNAs using 6mer seed matching. The total number of 6mer seed matches for each miRNA is listed in Table 1. For identification of miRNA targets using negative correlation, correlation coefficients were calculated between each pair of miRNA and mRNA profiles and the mRNAs were ordered according to these correlation coefficients. Four of the eight miRNAs showed a significant enrichment of sequence predicted targets among mRNAs with correlation coefficients close to -1 ( $q < 0.1$ ), Table 1. Additional file 2 lists all calculated correlation coefficients.

Applying ICA on the mRNA expression data resulted in identification of five highly significant ICs (Bonferroni corrected  $q < 0.0001$ ) with mixes shown in Figure 2B. See Additional file 2 for all loads, as well as the method section for further details on ICA.





For each of the eight miRNAs, the loads in each IC were tested for enrichment of predicted targets with a 6mer seed match. Seven of the eight miRNAs (including all four identified by negative correlation) showed a significant enrichment of sequence predicted targets in the IC loads ( $q < 0.07$ ), Table 1. Furthermore, we analyzed IC 1 for cooperativity between each pair of

**Table 1 Target prediction using negative correlation or ICA (loads)**

miRNA	# 6mer matches	Correlation (q-value)	Load (q-value)
miR-124	2215	0.0027	0.005
miR-128	2105	0.63	0.07
miR-192	1128	0.63	0.0007
miR-194	1110	0.0016	0.02
miR-204	1560	0.63	0.02
miR-375	1128	0.0027	0.04
miR-672	1060	0.87	0.5
miR-708	1675	0.023	0.0007

The table shows the number of 6mer seed matches in the array data for each of the eight miRNAs (2<sup>nd</sup> column), the q-values (corrected p-values) for negative correlation (3<sup>rd</sup> column) or load values (4<sup>th</sup> column). The q-values in the 4<sup>th</sup> column are the lowest q-value found among the five significant ICs.

miRNAs by testing the loads in IC 1 for enrichment of predicted targets with seed matches for both miRNAs (Additional file 3).

We found that the miRNA pairs miR-375/672, miR-194/375, miR-192/375 and miR-124/194 had a significant regulatory effect (Bonferroni corrected  $q < 0.05$ ) when testing for enrichment of their common targets in IC 1.

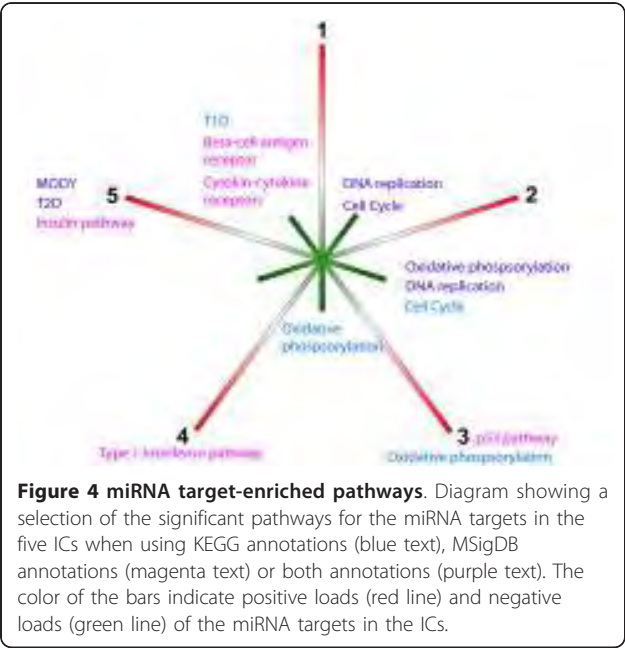
The miRNAs had small fold changes compared to the fold changes of the mRNAs. Therefore it was not possible to apply ICA on the miRNA data or on a combination of the mRNA and miRNA expression matrices. However, by forming a linear superposition of the ICs we identified the major contributions from the ICs on the miRNA expressions. Interestingly, all eight miRNAs could be represented (percent variance explained  $> 97\%$ ) by a superposition of the five ICs, as shown in Figure 2C (also see Additional file 4). The majority of predicted miRNA targets were present in two or three of the ICs. For example, miR-124 primarily has targets in component 1 and 4, which both are Pdx-1 affected components. Another example is miR-192 and miR-194 that both have targets in component 1, 2 and 4, which were affected by both Pdx-1 and IL-1 $\beta$ .

#### Independent components have clear biological profiles

To assess the biological relevance of the identified ICs, we tested the ICs for enrichment of genes known to be regulated by Pdx-1 or affected by IL-1 $\beta$  (Additional file 5). Mixes of IC 1 clearly correlated with the induction of Pdx-1, and genes regulated by Pdx-1 were over-represented in this IC ( $p < 0.047$ , data not shown). Interestingly, one group had low negative loads and another had high positive loads, indicating positive and negative regulation by Pdx-1, respectively. Similarly, ICs 3 and 4 correlated with the presence or absence of IL-1 $\beta$ . Specifically, IC 3 was enriched for known IL-1 $\beta$  affected genes among genes with low negative load ( $p < 0.0021$ , data not shown), whereas IC 4 was enriched for IL-1 $\beta$  affected genes with high positive load ( $p < 0.024$ , data not shown).

In a similar manner, we tested the ICs for enrichment of miRNA targets associated with metabolic or signalling pathways, as annotated by KEGG [38] or by the Molecular Signatures Database (MSigDB, <http://www.broadinstitute.org>). Specifically, we tested the genes with high positive or low negative loads in each IC for overrepresentation of annotated pathway genes.

When using KEGG terms only, 25 pathways were significantly enriched for miRNA targets in the five ICs ( $q < 0.05$ ). The most significant and diabetes-relevant pathways are shown in Figure 4 (all are listed in Additional file 6). Notable was the significance of genes with low loads in IC 3 belonging to the T1D pathway. IC

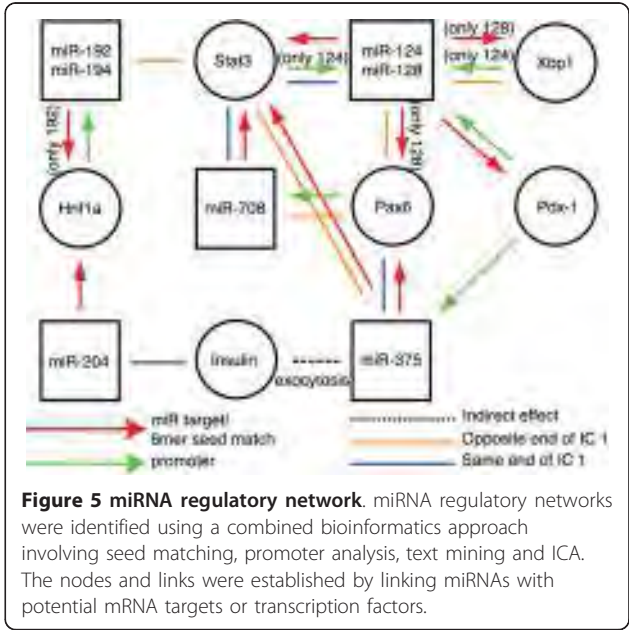


3 had a clear correlation with stimulation of IL-1 $\beta$ . Furthermore, the pathways maturity onset diabetes of the young (MODY), type 2 diabetes (T2D) mellitus and oxidative phosphorylation were significant. The first two have an obvious relation to diabetes, and the oxidative phosphorylation pathway has been shown to be related to both type 1 and type 2 diabetes [39]. When using MSigDB annotations, 150 pathways were significantly enriched for miRNA targets belonging to the five ICs ( $q < 0.05$ ). A selection of the pathways is shown in Figure 4 (all are listed in Additional file 6). Since KEGG is part of MSigDB it comes as no surprise that T2D, MODY and oxidative phosphorylation again showed up as significant. However, T1D did not show up as a significant pathway, probably due to correction of multiple testing, since MSigDB is a much larger repository. Also, dysregulation of genes involved in the p53 signaling pathway have been suggested to sensitize the cells to apoptotic stimuli [32,40]. In accordance with this, we find genes annotated with the p53-signalling pathway (using MSigDB) having significant low loads in IC 3.

**Identification of miRNA regulatory networks**

miRNAs and mRNAs can interact in regulatory networks. miRNAs can regulate mRNAs directly or indirectly through secondary factors. Furthermore, mRNA targets can act as transcription factors for miRNAs, thus forming a regulatory loop.

Using a combined bioinformatics approach involving seed matching, promoter analysis, text mining and ICA we identified miRNA regulatory networks (Figure 5). Using the miRNA database <http://www.mirbase.org>, we



found STAT3 binding sites in the promoter of the human miR-124 gene. This interaction was supported by ICA that showed that miR-124 and Stat3 both have negative contribution from IC 1 (Additional file 2 and Figure 2C) correlating with an up-regulation in response to Pdx-1 induction. Interestingly, the 3'UTR of Stat3 has seed matches for miR-124 (TargetScan and 6mer seed matching), indicating a potential feedback loop.

Furthermore, we found positive and negative contribution from IC 1 for Pax6 and miR-124/128, respectively (Additional file 2 and Figure 2C), meaning they have reciprocal expressions (down-regulation of Pax6 and up-regulation of miR-124/128) in response to Pdx-1 induction. Interestingly, the 3'UTR of Pax6 contains a 6mer seed match for miR-128 indicating potential repression of Pax6 by miR-128.

**Validation in human pancreatic islets and  $\beta$ -cell studies**

To verify the identified miRNA targets affected by Pdx-1 and/or cytokine stimulation, we used mRNA expression profiling data from studies on human pancreatic islets and rat  $\beta$ -cell lines and tested whether the data were enriched for ICA-identified targets of the eight miRNAs.

Using mRNA profiling data from untreated and cytokine treated human pancreatic islet preparations (unpublished data) we tested the cytokine-induced fold changes for enrichment of predicted miRNA targets. Interestingly, we found that predicted targets of miR-128/192/194/204/375 were significantly up-regulated in response to cytokine treatment as compared to non-targets ( $q = 0.007$ ,  $q = 0.004$ ,  $q = 0.033$ ,  $q = 0.001$  and

$q = 0.05$ , respectively). The results suggest that the five miRNAs may play a role in human pancreatic islets as well.

Additionally, we performed similar analysis of publicly available data from two previous studies on  $\beta$ -cells. Cardozo et al. [41] performed mRNA profiling on  $\beta$ -cells from 10 weeks old male Wistar rats, un-stimulated or stimulated with cytokines (IL-1 $\beta$  and/or IFN $\gamma$ ). Using their mRNA data, we found that targets of miR-192 were significantly down-regulated in response to IL-1 $\beta$  stimulation when comparing to non-targets ( $q = 0.02$ ). Kutlu et al. [42] performed mRNA profiling on insulin producing INS-1E cells, untreated or exposed to a combination of IL-1 $\beta$  and IFN $\gamma$ . In their data set, we found that the expressions of miR-128 targets were significantly increased in response to cytokine exposure when comparing to non-targets ( $q = 0.002$ ).

## Discussion

The motivation for applying a more advanced method like ICA for miRNA target prediction than negative correlation was that surprisingly few mRNAs have clear negative correlations with their targeting miRNAs. This is probably due to the mRNA profiles being influenced by a number of factors e.g. miRNA regulation, transcription factor binding and site accessibility. Here, ICA was used as an attempt to filter the factors influencing mRNA expression. Decomposition of microarray data using ICA has been shown to outperform other linear data representations, such as PCA [16-19]. Several target prediction methods have incorporated miRNA and mRNA profiling data. However, none of them uses ICA.

The miRNA array profiling identified eight miRNAs with differential expression in a type 1 diabetes model. Performing ICA on the mRNA expressions from the same samples resulted in five ICs that correlated with the experimental conditions studied. Comparing the two target prediction methods indicated that ICA was better at capturing miRNA activity than negative correlation. Seven miRNAs showed a significant enrichment of sequence-predicted targets when using ICA, as compared to only four by use of negative correlation. Interestingly, the ICs were enriched for miRNA targets with functional roles in diabetes-relevant pathways e.g. the pathways T1D, T2D, MODY, oxidative phosphorylation, insulin, cytokine-cytokine receptors and type 1 interferon. This supports that the eight miRNAs are implicated in disease mechanisms in diabetes. Additionally, targets of five of the eight miRNAs were significantly regulated by cytokines in models of  $\beta$ -cell destruction e.g. in human islets.

miRNAs fine tune the expression of genes in a combinatorial manner, meaning that several miRNAs can target the same mRNA transcript [43]. Furthermore, a cluster of co-expressed miRNAs can regulate

functionally related genes [44]. In this study, we observe small expression changes in the miRNAs. However, even minute changes in miRNA expressions might have impact on mRNA expression, and miRNAs acting in a cooperative manner can most likely induce biologically relevant expression changes in their targets. ICA can uncover these more complex interactions. Interestingly, it was recently suggested that cooperativity could be incorporated for prediction of target interactions between different miRNAs [1]. For the eight miRNAs we identified, there is a significant overlap in the mRNAs they target. We have incorporated cooperativity between miRNAs pairwise and identified four miRNA pairs (miR-375/672, miR-194/375, miR-192/375, miR-124/194) that had a significant co-regulatory effect on their common targets in IC 1.

Of the eight significant miRNAs, miR-124 and miR-375 have previously been identified in  $\beta$ -cells [33,35,45,46]. Further, the expression of miR-204 has been shown to be induced in insulinomas, where its expression correlated with insulin expression [47].

That miR-375 was significantly regulated strengthens our model since a previous study observed interaction between Pdx-1 (and NeuroD1) and the miR-375 locus [32]. However, no Pdx-1 consensus binding sites were identified, but binding elements for other transcription factors have been identified in the miR-375 locus [48]. The decreased miR-375 expression could, at least in part, be mediated through NeuroD1, since we observed a decreased NeuroD1 expression in response to Pdx-1 induction. Additionally, the decreased miR-375 expression is in agreement with miR-375 having a higher expression level in non- $\beta$ -cells than in  $\beta$ -cells [35]. This is also supported by our findings in  $\alpha$ -cells versus  $\beta$ -cells (Additional file 7). Interestingly, both miR-375 and Pax6, a key factor in  $\alpha$ -cell development, had negative loads in IC 1, i.e. both were down-regulated in response to Pdx-1. The decreased miR-375 expression is also in compliance with the function of miR-375 as a negative regulator of insulin exocytosis [45], since it correlates with the need for an increased insulin secretion in the mature  $\beta$ -cell phenotype. Similarly, miR-124 has been shown to modulate insulin secretion by targeting Foxa2 [33]. miR-194 is highly expressed in liver and in intestinal epithelial cells, where it is under regulation by Hnf1 $\alpha$  [49,50]. Interestingly, Hnf1 $\alpha$  is required for proper  $\beta$ -cell function and mutations in this gene cause MODY [23]. miR-192 is also expressed in the liver and is in cluster with miR-194, suggesting co-regulation [49,50]. Furthermore, miR-128 has been shown to induce apoptosis in kidney cells through interaction with Bax [51]. So far, miR-672 and miR-708 have not been examined in  $\beta$ -cells. The likely involvement of miR-128/192/194/204/708 in  $\beta$ -cell regulatory networks



(Figure 5) and the high expression in  $\beta$ -cells compared to  $\alpha$ -cells for miR-672/204 (Additional file 7) make them interesting candidates for further studies.

We have used ICA for bioinformatics investigation of the functional roles of the miRNAs and their targets. ICA in combination with pathway analysis indicates that the eight miRNAs, through their mRNA targets, are implicated in several diabetes relevant pathways.

The transcriptional changes mediated by miRNAs on their targets may not be entirely explained by direct repression but may also reflect indirect mechanisms such as activation by feedback and feed-forward transcriptional loops within regulatory networks [52,53]. miRNAs can be important players in these networks. By use of a combined bioinformatics approach we identified miRNA regulatory networks. The results suggest connections between seven of the eight miRNAs through interactions with key pancreatic transcription factors, cytokine signalling molecules and insulin (Figure 5).

## Conclusions

Using ICA, we have developed a novel miRNA target prediction method that incorporates seed matching and expression profiling. We believe that the method has advantages compared to simple negative correlation. Additionally, ICA in combination with bioinformatics approaches such as pathway analysis constitutes a means of prioritizing between the predicted miRNA targets for further investigations. To the best of our knowledge this is the first study that uses ICA for miRNA target prediction. Interestingly, applying the method on a model of type 1 diabetes resulted in identification of eight miRNAs that appear to directly or indirectly affect pathways of relevance to disease mechanisms in diabetes.

## Methods

### Cell cultures

INSr $\alpha\beta$ ,  $\beta$ TC-3 and  $\alpha$ TC-1 cells were cultured in complete RPMI-1640 medium (RPMI-1640 with Glutamax (Gibco BRL, Paisley, Scotland, UK), 100 U/ml penicillin and 100  $\mu$ g/ml streptomycin) supplemented with 10% FBS. 50  $\mu$ M  $\beta$ -mercaptoethanol was additionally added to the INSr $\alpha\beta$  cells. For the INSr $\alpha\beta$  cells, TeT system approved FBS (CLONTECH, Palo Alto, CA) was used, and the medium was supplemented with 150  $\mu$ g/ml geneticin and 100  $\mu$ g/ml hygromycin to maintain a pure culture of cells expressing the Pdx-1 plasmid construct. The INSr $\alpha\beta$  cells were treated with doxycycline (500 ng/ml; Sigma Aldrich, Saint Louis, MO, USA) for 24 h to allow for Pdx-1 expression after which the cells were cultured in the presence or absence of human recombinant IL-1 $\beta$  (40 pg/ml; BD Pharmingen, San Diego, CA) for additionally 2 h or 24 h.

### Human pancreatic islet preparations

Human islet preparations (n = 4) provided through the Juvenile Diabetes Research Foundation (JDRF) Islet Distribution Program (JDRF award 6-2005-1178) by Islet Cell Resource Centres in Milan (Italy) and Lille (France) were treated for 48 hours with IL-1 $\beta$  (1 ng/ul) or a combination of IL-1 $\beta$  (1 ng/ul), IFN $\gamma$  (20 ng/ul) and TNF $\alpha$  (8 ng/ul).

### RNA extraction and RT-qPCR

Total RNA was extracted using TRIzol reagent (Invitrogen, Carlsbad, CA) according to the manufacturer's instructions. cDNA was prepared using miScript reverse transcription kit (Qiagen, Hilden, Germany), TaqMan microRNA reverse transcription reagents (Applied Biosystems, Foster City, CA) or TaqMan reverse transcription reagents (Applied Biosystems), as described by the manufacturer.

In INSr $\alpha\beta$  cells (n = 5) the miRNA expression levels were analysed by use of miScript Primer Assays (Qiagen), whereas in  $\beta$ TC-3 (n = 5) and  $\alpha$ TC-1 (n = 5) cells miRNA expressions were analysed using TaqMan microRNA assays (Applied Biosystems). Gene expression levels were analysed using TaqMan gene expression assays (Applied Biosystems). All samples were analyzed in duplicates on a ABI 7900HT system (Applied Biosystems). Data were evaluated using the  $2^{-\Delta\Delta C_t}$  method [54], normalizing miRNA/gene expressions to an endogenous control, and subsequently comparing each miRNA/gene in treated vs. un-stimulated cells. miRNAs were normalized to let-7c, whereas genes were normalized to Ppia. The coefficient of variation (CV) for miScript assays ranged from 0.018 to 0.077.

### Western blotting

The INSr $\alpha\beta$  cells (n = 4) were washed with PBS and lysed. The protein concentrations of whole cell protein lysates were measured with Bio-Rad Protein Assay. Samples were boiled before loading on a NuPage gel (10% BisTris gels, Invitrogen) placed in an X-Cell Surelock system (Invitrogen). Proteins were separated and then transferred to a pre-soaked nitrocellulose membrane (Invitrogen), which was blocked in milk for one hour and then washed. The membrane was incubated in a 5 ml milk solution containing primary antibody, either monoclonal Pdx-1 mouse antibody [37] overnight or monoclonal mouse  $\beta$ -actin antibody (Abcam, Cambridge, MA; #ab6276) for one hour. Horseradish peroxidase conjugated anti-mouse antibodies (Cell signalling, #7076) were used as secondary antibodies. The membrane was incubated in LumiGlo (Cell Signaling Technology, Danvers, MA) solution and visualized using the LAS2000 system (Fujifilm Europa GmbH, Dusseldorf, Germany).

### Profiling and preprocessing of array data

The mRNA expressions from INSrαβ cells ( $n = 4$ ) were profiled on Affymetrix Rat Genome 230 2.0 arrays. For each probeset in each sample, log2-ratios were calculated between the expression of that probeset and the average of the control samples.

The miRNA expressions from INSrαβ cells ( $n = 3$ ) were profiled on Exiqon miRCURY LNA microRNA Array v.11.0 arrays and median normalized. The control samples ( $n = 3$ ) were pooled and used as a common reference on all arrays, and log2-ratios were calculated for each sample against this reference.

The mRNA expressions from human islet preparations ( $n = 4$ ) were analyzed on Affymetrix Human Genome U133 Plus 2.0 arrays. The β-cell data from Refs. [41] and [42] were downloaded from T1DBase <http://www.t1dbase.org>. These mRNA expressions had been profiled using the Affymetrix Rat Genome U34A array.

All CEL-files were preprocessed using the RMA package [55] in Bioconductor [56] with remapped Ensembl build 50 gene probesets [57].

For mRNA data from human islets and published β-cell studies, foldchanges  $\log_2(\text{cytokine}, 24 \text{ h})/(\text{untreated}, 24 \text{ h})$  were calculated for the average of the replicate samples in each condition.

### Sequence based miRNA target prediction

For 7764 out of the 9953 Ensembl gene probesets analyzed, 3'UTR sequences could be downloaded from Ensembl through Biomart [58]. Mature sequences of significantly changed miRNAs were downloaded from miRbase [59] and the 6mer seed matches extracted. Ensembl gene probesets with a given miRNA 6mer seed match in their 3'UTR were identified as targets of that miRNA.

### Independent component analysis

Mathematically,  $T$  transcriptional programs are active in the cells under study, each of which is a (column) vector  $\vec{C}_t$ ,  $t = 1, \dots, T$ , representing  $G_t$  gene inductions or repressions. We measure all  $G$  induced or repressed genes in  $S$  samples.

Each of these programs can be identified by a simple linear decomposition of the  $G \times S$  expression matrix  $E$ ,

$$E = CM \text{ or } E_{gs} = \sum_{t=1}^T C_{gt} M_{ts} \text{ or } \vec{E}_s = \sum_{t=1}^T C_t M_{ts}$$

where  $C$  is the  $G \times T$  matrix containing the transcriptional programs, in coefficient vector notation,  $C = (\vec{C}_1, \vec{C}_2, \dots, \vec{C}_T)$ , likewise  $E = (\vec{E}_1, \vec{E}_2, \dots, \vec{E}_S)$  and  $M$  is the  $T \times S$  matrix giving the linear mixing of each

program in each sample. Two-way clustering of correlated genes and samples into non-overlapping sets was represented by component vectors  $\vec{C}_t$  with discrepant sets of non-vanishing entries

$$C_{gt} \neq 0 \Leftrightarrow C_{gj} \approx 0 \quad \forall j \neq t.$$

ICA extracts  $n = \min\{G, S\}$  independent components. Since the number of samples was much lower than the number of genes, we assumed  $n = S$ . The number of interesting transcriptional programs  $T$  may be equal to, or lower than  $S$ . A common practice is to project  $E$  onto its first  $O$  principal component directions, and then extract  $O$  independent components in this subspace. This approach has several important drawbacks. First, the number of interesting transcriptional programs is not known beforehand, and one therefore needs to guess or resort to a recursive trial-and-error approach to choose an  $O$  that allows all  $T$  transcriptional programs to be extracted. Second, it is not obvious that a projection onto  $O < S$  principal component directions preserves all  $T$  transcriptional programs. It is possible to construct examples where two clusters in the data are only well separated in the subspace of the first and last principal components [60]. Clustering in a subspace spanned by principal components generally degrades cluster quality [61]. Finally, even when a projection onto  $T$  principal components allows all the  $T$  interesting transcriptional programs to be extracted by ICA, a way to explore and rank these independent components to clarify their significance is still needed. In the present study, the independent components were ranked according to their accordance with the experimental conditions.

The fastICA algorithm [62,63] in R <http://www.R-project.org> was used to estimate the component matrix and mixing matrix, by numeric maximization of the negentropy of the independent components. The negentropy was approximated by  $J(y) = (\langle G(y) \rangle - \langle G(v) \rangle)^2$ , where  $y$  is the distribution of loads,  $v$  is a random variable distributed as  $\mathcal{N}(0,1)$ , and the contrast function  $G(u)$  was here defined as  $G(u) = -\exp\left(\frac{-u^2}{2}\right)$ . By convention, the

loads in each component have mean 0 and variance 1.

Assuming that the individual contributions on mRNA expression are linearly separable and reasonable independent of each other, ICA can separate the various contributions making it possible to detect pathways regulated by the various miRNAs.

ICs with mixes, i.e. rows in  $M$  showing significant changes between the experimental conditions and/or time points, are identified by calculating F-statistics and requiring Bonferroni corrected p-values less than 0.0001.

## Statistics

Overall expression data were analysed using one-way analysis of variance (ANOVA; F-statistics). The eight significant miRNAs were selected based on an overall q-value < 0.1 (Additional file 8). Comparison between groups was done using t-test. Correction for multiple testing was performed either by applying a Bonferroni or a false discovery rate (fdr) correction to the p-values (q-value). A Bonferroni correction was chosen when few tests were performed, as was the case for the finding of significant ICs and miRNAs. When a larger number of tests were performed a fdr correction was applied.

The loads in each IC are considered non-Gaussian distributed. Loads/fold changes for genes that had seed match for the miRNAs or were assigned to a specific biological pathway were analysed using Wilcoxon rank sum test. For analysis of negative correlation, Pearson correlation coefficients were calculated between each pair of miRNA and mRNA profiles. The mRNAs were then ordered according to these correlation coefficients. P-values < 0.05 were considered significant.

## Additional material

### Additional file 1: Expressions of insulin and Pdx-1 dependent genes.

There are three experimental conditions: Pdx-1 induction (dox treatment), IL-1 $\beta$  treatment and time (samples are taken 2 h and 24 h after treatment). Log2-transformed fold changes (mean and standard deviation) between experimental and control conditions. \*: 0.05 > q > 0.01, \*\*: 0.01 > q > 0.001, \*\*\*: 0.001 > q > 0.

### Additional file 2: miRNA target prediction based on mRNA expression data.

For each Ensembl annotated gene on the Affymetrix array is given the gene symbol, gene description, loads in the five ICs, coefficients for correlation with the eight miRNA expression profiles and 6mer seed match with the eight miRNAs. We assumed that loads were significant if they had an absolute load greater than 2. For seed match, 1/0 denotes  $\geq 1$  seed matches and no seed match.

**Additional file 3: Cooperativity between miRNAs in IC 1.** The p-values (uncorrected) and q-values (corrected) for pairwise miRNA cooperativity in IC 1 are shown. 1/-1 denotes whether the individual miRNA has positive or negative loads in IC 1.

**Additional file 4: Coefficients for the linear superposition of the ICs giving the miRNA expression profiles.** R2 is the coefficient of determination.

**Additional file 5: Genes regulated by Pdx-1 and/or IL-1 $\beta$ .** 1/0 denotes regulation and no regulation based on text mining.

**Additional file 6: Pathways significantly enriched for genes with high positive or low negative loads in the ICs.** Sheet 1 shows KEGG pathways whereas sheet 2 shows MSigDB pathways. Only pathways with q-values < 0.05 (fdr corrected p-values) are highlighted.

**Additional file 7: Expression of the eight miRNAs in  $\alpha$ - versus  $\beta$ -cells.** Ratio of basal  $\Delta$ CT values for  $\alpha$ TC1 versus  $\beta$ TC3 cells. The ratio is found for un-stimulated cells. Bars are standard deviations and asterisks denote: \*\*: 0.01 > q > 0.001, \*\*\*: 0.001 > q > 0.

**Additional file 8: Overall q-values for the miRNA expressions.** FWER: family-wise error rate.

## List of abbreviations

**3'UTR:** untranslated region in the 3' end of the mRNA; **dox:** doxycycline; **FBS:** fetal bovine serum; **IC:** independent component; **ICA:** independent component analysis; **IFN $\gamma$ :** interferon gamma; **IL-1 $\beta$ :** interleukin 1 beta; **miR/miRNA:** microRNA; **MODY:** maturity onset diabetes of the young; **mRNA:** messenger RNA; **PBS:** phosphate buffered saline; **PCA:** principal component analysis; **RT-qPCR:** reverse transcription - quantitative polymerase chain reaction; **T1D:** type 1 diabetes; **T2D:** type 2 diabetes; **TNF $\alpha$ :** tumor necrosis factor alpha.

## Acknowledgements and Funding

This work was supported by grants from The Danish Diabetes association, EFSD (European Foundation for the Study of Diabetes) and the Danish Council for Strategic Research "Center for non-coding RNA in Technology and Health" (09-067036/DSF). The work of PHH was supported by a grant from the Lundbeck Foundation. We are grateful to Drs Haiyan Wang and Claes B. Wollheim University of Geneva for kindly providing the INS $\alpha$  cell line. We thank Bodil Bosmann Jørgensen, Hagedorn Research Institute, for technical assistance in the lab, and additionally we thank Jacob Hald also from Hagedorn Research Institute for providing the PDX-1 antibody. We thank Jens Højris Nielsen and Maiken Rosenstjerne at University of Copenhagen for providing access to array facilities and expert guidance in miRNA array methods. Drs. AE Karlén, M Kruhøffer and T Ørntoft are acknowledged for taking part in the design and carrying out the mRNA Affymetrix array experiments.

## Author details

<sup>1</sup>Glostrup Research Institute, Glostrup University Hospital, DK-2600 Glostrup, Denmark. <sup>2</sup>Hagedorn Research Institute, Niels Steensensvej 6, DK-2820 Gentofte, Denmark. <sup>3</sup>Center for Models of Life, University of Copenhagen, Blegdamsvej 17, DK-2100 Copenhagen, Denmark. <sup>4</sup>Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, DK-2800 Lyngby, Denmark. <sup>5</sup>Steno Diabetes Center, Niels Steensensvej 2, DK-2820 Gentofte, Denmark. <sup>6</sup>University of Lund, CRC, Skåne University Hospital, SE-20502 Malmö, Sweden. <sup>7</sup>Department of Molecular Biomedicine, LEO Pharma A/S, Industriparken 55, DK-2750 Ballerup, Denmark.

## Authors' contributions

FP and CHBB designed the study. The experimental work was performed by CHBB, TF and TG. The analyses and the manuscript draft were conducted by LP, CHBB and TF. ICA and microarray data analysis were conducted by LP and PHH. FP edited the paper and prepared it for final review. All authors read and approved the final manuscript and have no conflicting interest in the study.

Received: 9 August 2010 Accepted: 4 February 2011

Published: 4 February 2011

## References

1. Saito T, Sætrom P: **MicroRNAs - targeting and target prediction.** *New Biotechnology* 2010, **27**(3):243-249.
2. Sood P, Krek A, Zavolan M, Macino G, Rajewsky N: **Cell-type-specific signatures of microRNAs on target mRNA expression.** *PNAS* 2006, **103**(8):2746-2751.
3. Huang JC, Babak T, Corson TW, Chua G, Khan S, Gallie BL, Hughes TR, Blencowe BJ, Frey BJ, Morris QD: **Using expression profiling data to identify human microRNA targets.** *Nat Methods* 2007, **4**(12):1045-1049.
4. Dongen Sv, Abreu-Goodger C, Enright AJ: **Detecting microRNA binding and siRNA off-target effects from expression data.** *Nat Methods* 2008, **5**(12):1023-1025.
5. Cheng C, Li LM: **Inferring microRNA activities by combining gene expression with microRNA target prediction.** *PLoS ONE* 2008, **3**(4):e1989.
6. Perou CM, Sørli T, Eisen MB, Rijn Mvd, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, et al: **Molecular portraits of human breast tumours.** *Nature* 2000, **406**(6797):747-752.
7. Arbeitman MN, Furlong EEM, Imam F, Johnson E, Null BH, Baker BS, Krasnow MA, Scott MP, Davis RW, White KP: **Gene expression during the life cycle of Drosophila melanogaster.** *Science* 2002, **297**(5590):2270-2275.



8. Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Schölkopf B, Weigel D, Lohmann JU: **A gene expression map of *Arabidopsis thaliana* development.** *Nat Genet* 2005, **37**(5):501-506.
9. Ohlsson Teague EMC, Van der Hoek KH, Van der Hoek MB, Perry N, Wagaarachchi P, Robertson SA, Print CG, Hull LM: **MicroRNA-Regulated Pathways Associated with Endometriosis.** *Mol Endocrinol* 2009, **23**(2):265-275.
10. Comon P: **Independent component analysis, A new concept?** *Signal Processing* 1994, **36**:287-314.
11. Martoglio AM, Miskin JW, Smith SK, MacKay DJC: **A decomposition model to track gene expression signatures: preview on observer-independent classification of ovarian cancer.** *Bioinformatics* 2002, **18**(12):1617-1624.
12. Chiappetta P, Roubaud MC, Torrèani B: **Blind source separation and the analysis of microarray data.** *J Comput Biol* 2004, **11**(6):1090-1109.
13. Saidi SA, Holland CM, Kreil DP, MacKay DJC, Charnock-Jones DS, Print CG, Smith SK: **Independent component analysis of microarray data in the study of endometrial cancer.** *Oncogene* 2004, **23**(39):6677-6683.
14. Capobianco E: **Mining time-dependent gene features.** *J Bioinform Comput Biol* 2005, **3**(5):1191-1205.
15. Lutter D, Ugocsai P, Grandl M, Orso E, Theis F, Lang E, Schmitz G: **Analyzing M-CSF dependent monocyte/macrophage differentiation: expression modes and meta-modes derived from an independent component analysis.** *BMC Bioinformatics* 2008, **9**(1):100.
16. Liebermeister W: **Linear modes of gene expression determined by independent component analysis.** *Bioinformatics* 2002, **18**(1):51-60.
17. Lee SJ, Batzoglou S: **Application of independent component analysis to microarrays.** *Genome Biol* 2003, **4**(11):R76.
18. Carpentier AS, Riva A, Tisseur P, Didier G, Hénaut A: **The operons, a criterion to compare the reliability of transcriptome analysis tools: ICA is more reliable than ANOVA, PLS and PCA.** *Comput Biol Chem* 2004, **28**(1):3-10.
19. Teschendorff AE, Journée M, Absil PA, Sepulchre R, Caldas C: **Elucidating the altered transcriptional programs in breast cancer using independent component analysis.** *PLoS Comput Biol* 2007, **3**(8):e161.
20. Frigyesi A, Veerla S, Lindgren D, Höglund M: **Independent component analysis reveals new and biologically significant structures in micro array data.** *BMC Bioinformatics* 2006, **7**:290.
21. Eizirik DL, Flodstrom M, Karlén AE, Welsh N: **The harmony of the spheres: inducible nitric oxide synthase and related genes in pancreatic beta cells.** *Diabetologia* 1996, **39**(8):875-890.
22. Lynn FC, Skewes-Cox P, Kosaka Y, McManus MT, Harfe BD, German MS: **MicroRNA expression is required for pancreatic islet cell genesis in the mouse.** *Diabetes* 2007, **56**(12):2938-2945.
23. Bernardo AS, Hay CW, Docherty K: **Pancreatic transcription factors and their role in the birth, life and survival of the pancreatic beta cell.** *Mol Cell Endocrinol* 2008, **294**(1-2):1-9.
24. Jonsson J, Carlsson L, Edlund T, Edlund H: **Insulin-promoter-factor 1 is required for pancreas development in mice.** *Nature* 1994, **371**(6498):606-609.
25. Offield MF, Jetton TL, Labosky PA, Ray M, Stein RW, Magnuson MA, Hogan BL, Wright CV: **PDX-1 is required for pancreatic outgrowth and differentiation of the rostral duodenum.** *Development* 1996, **122**(3):983-995.
26. Stoffers DA, Zinkin NT, Stanojevic V, Clarke WL, Habener JF: **Pancreatic agenesis attributable to a single nucleotide deletion in the human IPF1 gene coding sequence.** *Nat Genet* 1997, **15**(1):106-110.
27. Ohlsson H, Karlsson K, Edlund T: **IPF1, a homeodomain-containing transactivator of the insulin gene.** *EMBO J* 1993, **12**(11):4251-4259.
28. Wang H, Maechler P, Ritz-Laser B, Hagenfeldt KA, Ishihara H, Philippe J, Wollheim CB: **Pdx1 level defines pancreatic gene expression pattern and cell lineage differentiation.** *J Biol Chem* 2001, **276**(27):25279-25286.
29. Nielsen K, Karlén AE, Deckert M, Madsen OD, Serup P, Mandrup-Poulsen T, Nerup J: **Beta-cell maturation leads to in vitro sensitivity to cytotoxins.** *Diabetes* 1999, **48**(12):2324-2332.
30. Nielsen K, Kruhoffer M, Orntoft T, Sparre T, Wang H, Wollheim C, Jørgensen MC, Nerup J, Karlén AE: **Gene expression profiles during beta cell maturation and after IL-1beta exposure reveal important roles of Pdx-1 and Nkx6.1 for IL-1beta sensitivity.** *Diabetologia* 2004, **47**(12):2185-2199.
31. Nielsen K, Sparre T, Larsen MR, Nielsen M, Fey SJ, Mose Larsen P, Roepstorff P, Nerup J, Karlén AE: **Protein expression changes in a cell system of beta-cell maturation reflect an acquired sensitivity to IL-1beta.** *Diabetologia* 2004, **47**(1):62-74.
32. Keller DM, McWeeney S, Arsenlis A, Drouin J, Wright CV, Wang H, Wollheim CB, White P, Kaestner KH, Goodman RH: **Characterization of pancreatic transcription factor Pdx-1 binding sites using promoter microarray and serial analysis of chromatin occupancy.** *J Biol Chem* 2007, **282**(44):32084-32092.
33. Baroukh N, Ravier MA, Loder MK, Hill EV, Bounacer A, Scharfmann R, Rutter GA, Van Obberghen E: **MicroRNA-124a regulates Foxa2 expression and intracellular signaling in pancreatic beta-cell lines.** *J Biol Chem* 2007, **282**(27):19575-19588.
34. Joglekar MV, Parekh VS, Hardikar AA: **New pancreas from old: microregulators of pancreas regeneration.** *Trends Endocrinol Metab* 2007, **18**(10):393-400.
35. Joglekar MV, Joglekar VM, Hardikar AA: **Expression of islet-specific microRNAs during human pancreatic development.** *Gene Expr Patterns* 2009, **9**(2):109-113.
36. Correa-Medina M, Bravo-Egana V, Rosero S, Ricordi C, Edlund H, Diez J, Pastor RL: **MicroRNA miR-7 is preferentially expressed in endocrine cells of the developing and adult human pancreas.** *Gene Expr Patterns* 2009, **9**(4):193-199.
37. Galbo T, Pedersen I, Floyel T, Bang-Berthelsen C, Serup P, Madsen O, Hald J: **Novel monoclonal antibodies against Pdx1 reveal feedback regulation of Pdx1 protein levels.** *European Journal of Histochemistry* 2010, **54**:e19.
38. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M: **KEGG for representation and analysis of molecular networks involving diseases and drugs.** *Nucleic Acids Res* 2010, **38** Database: D355-360.
39. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E, et al: **PGC-1[alpha]-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes.** *Nat Genet* 2003, **34**(3):267-273.
40. Gurzov EN, Germano CM, Cunha DA, Ortis F, Vanderwinden JM, Marchetti P, Zhang L, Eizirik DL: **p53 up-regulated modulator of apoptosis (PUMA) activation contributes to pancreatic beta-cell apoptosis induced by proinflammatory cytokines and endoplasmic reticulum stress.** *J Biol Chem* 2010, **285**(26):19910-19920.
41. Cardozo AK, Kruhoffer M, Leeman R, Orntoft T, Eizirik DL: **Identification of novel cytokine-induced genes in pancreatic beta-cells by high-density oligonucleotide arrays.** *Diabetes* 2001, **50**(5):909-920.
42. Kutlu B, Cardozo AK, Darville MI, Kruhoffer M, Magnusson N, Orntoft T, Eizirik DL: **Discovery of gene networks regulating cytokine-induced dysfunction and apoptosis in insulin-producing INS-1 cells.** *Diabetes* 2003, **52**(11):2701-2719.
43. Krek A, Grün D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, MacMenamin P, Piedade Id, Gunsalus KC, Stoffel M, et al: **Combinatorial microRNA target predictions.** *Nat Genet* 2005, **37**(5):495-500.
44. Dews M, Homayouni A, Yu D, Murphy D, Sevignani C, Wentzel E, Furth EE, Lee WM, Enders GH, Mendell JT, et al: **Augmentation of tumor angiogenesis by a Myc-activated microRNA cluster.** *Nat Genet* 2006, **38**(9):1060-1065.
45. Poy MN, Eliasson L, Krutzfeldt J, Kuwajima S, Ma X, Macdonald PE, Pfeffer S, Tuschl T, Rajewsky N, Rorsman P, et al: **A pancreatic islet-specific microRNA regulates insulin secretion.** *Nature* 2004, **432**(7014):226-230.
46. Poy MN, Hausser J, Trajkovski M, Braun M, Collins S, Rorsman P, Zavolan M, Stoffel M: **miR-375 maintains normal pancreatic alpha- and beta-cell mass.** *Proc Natl Acad Sci USA* 2009, **106**(14):5813-5818.
47. Roldo C, Missiaglia E, Hagan JP, Falconi M, Capelli P, Bersani S, Calin GA, Volinia S, Liu CG, Scarpa A, et al: **MicroRNA expression abnormalities in pancreatic endocrine and acinar tumors are associated with distinctive pathologic features and clinical behavior.** *J Clin Oncol* 2006, **24**(29):4677-4684.
48. Avnit-Sagi T, Kantorovich L, Kreda-Russo S, Hornstein E, Walker MD: **The promoter of the pri-miR-375 gene directs expression selectively to the endocrine pancreas.** *PLoS One* 2009, **4**(4):e5033.
49. Tang X, Gal J, Zhuang X, Wang W, Zhu H, Tang G: **A simple array platform for microRNA analysis and its application in mouse tissues.** *RNA* 2007, **13**(10):1803-1822.
50. Hino K, Tsuchiya K, Fukao T, Kiga K, Okamoto R, Kanai T, Watanabe M: **Inducible expression of microRNA-194 is regulated by HNF-1alpha during intestinal epithelial cell differentiation.** *RNA* 2008, **14**(7):1433-1442.

51. Adlakha Y, Saini N: **MicroRNA-128 downregulates Bax and induces apoptosis in human embryonic kidney cells.** *Cellular and Molecular Life Sciences* 2010, **1**-14.
52. Chen K, Rajewsky N: **The evolution of gene regulation by transcription factors and microRNAs.** *Nat Rev Genet* 2007, **8**(2):93-103.
53. Tsang J, Zhu J, van Oudenaarden A: **MicroRNA-Mediated Feedback and Feedforward Loops Are Recurrent Network Motifs in Mammals.** *Molecular Cell* 2007, **26**(5):753-767.
54. Livak KJ, Schmittgen TD: **Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method.** *Methods* 2001, **25**(4):402-408.
55. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP: **Summaries of Affymetrix GeneChip probe level data.** *Nucleic Acids Res* 2003, **31**(4):e15.
56. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, *et al*: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol* 2004, **5**(10):R80.
57. Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H, *et al*: **Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data.** *Nucleic Acids Res* 2005, **33**(20):e175.
58. Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, Kasprzyk A: **BioMart-biological queries made easy.** *BMC Genomics* 2009, **10**:22.
59. Griffiths-Jones S, Saini HK, Dongen Sv, Enright AJ: **miRBase: tools for microRNA genomics.** *Nucleic Acids Res* 2008, **36** Database: D154-D158.
60. Chang WC: **On Using Principal Components Before Separating a Mixture of Two Multivariate Normal Distributions.** *Appl Statist* 1983, **32**(3):267-275.
61. Yeung KY, Ruzzo WL: **Principal component analysis for clustering gene expression data.** *Bioinformatics* 2001, **17**(9):763-774.
62. Hyvärinen A, Oja E: **Independent component analysis: algorithms and applications.** *Neural Netw* 2000, **13**(4-5):411-430.
63. Hyvärinen A: **Fast and Robust Fixed-Point algorithms for Independent Component Analysis.** *IEEE Trans on Neural Networks* 1999, **10**(3):626-634.

doi:10.1186/1471-2164-12-97

**Cite this article as:** Bang-Berthelsen *et al*: Independent component and pathway-based analysis of miRNA-regulated gene expression in a model of type 1 diabetes. *BMC Genomics* 2011 **12**:97.

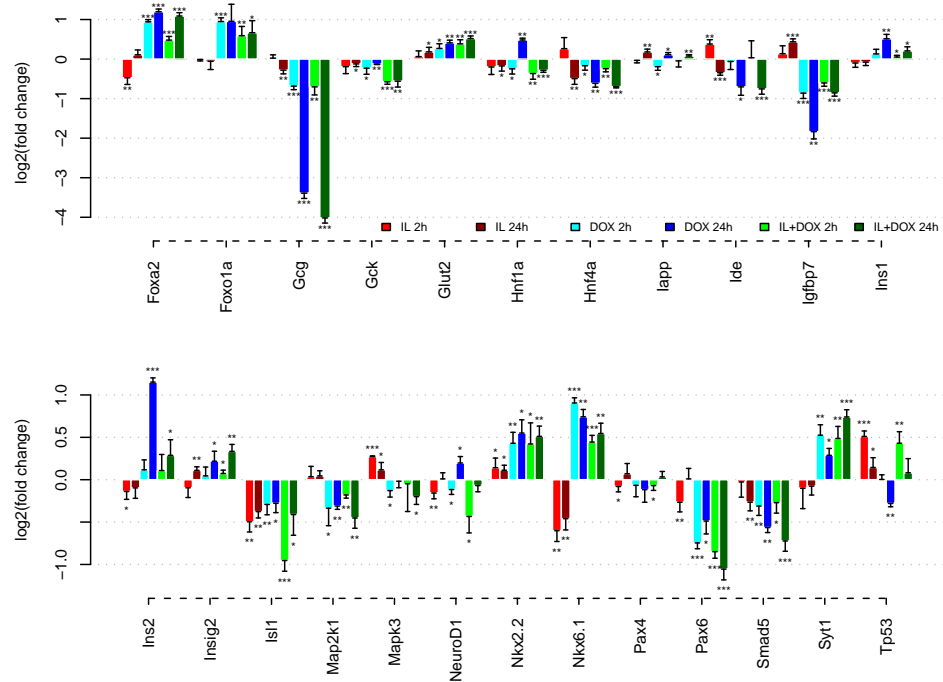
**Submit your next manuscript to BioMed Central and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

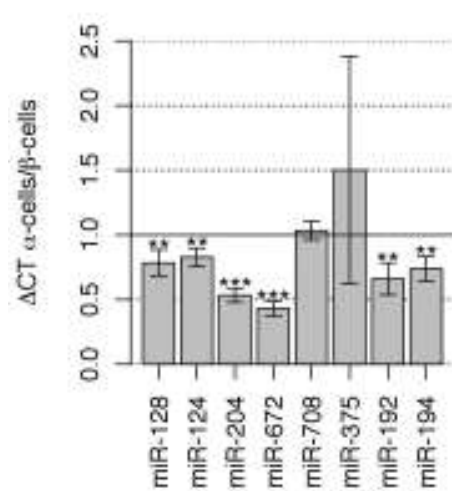
Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)



Supplementary Figure 1



Supplementary Figure 7





APPENDIX

E

DKK1 paper



# Dickkopf1 - A New Player in Modelling the Wnt Pathway

Lykke Pedersen<sup>1\*</sup>, Mogens Høgh Jensen<sup>1</sup>, Sandeep Krishna<sup>2</sup>

**1** Center for Models of Life, Niels Bohr Institute, Copenhagen, Denmark, **2** Theory and Modelling of Biological Systems, National Centre for Biological Sciences, Bangalore, Karnataka, India

## Abstract

The Wnt signaling pathway transducing the stabilization of  $\beta$ -catenin is essential for metazoan embryo development and is misregulated in many diseases such as cancers. In recent years models have been proposed for the Wnt signaling pathway during the segmentation process in developing embryos. Many of these include negative feedback loops where Axin2 plays a key role. However, Axin2 null mice show no segmentation phenotype. We therefore propose a new model where the negative feedback involves Dkk1 rather than Axin2. We show that this model can exhibit the same type of oscillations as the previous models with Axin2 and as observed in experiments. We show that a spatial Wnt gradient can consistently convert this temporal periodicity into the spatial periodicity of somites, provided the oscillations in new cells arising in the presomitic mesoderm are synchronized with the oscillations of older cells. We further investigate the hypothesis that a change in the Wnt level in the tail bud during the later stages of somitogenesis can lengthen the time period of the oscillations and hence the size and separation of the later somites.

**Citation:** Pedersen L, Jensen MH, Krishna S (2011) Dickkopf1 - A New Player in Modelling the Wnt Pathway. PLoS ONE 6(10): e25550. doi:10.1371/journal.pone.0025550

**Editor:** Johannes Jaeger, Centre for Genomic Regulation (CRG), Universitat Pompeu Fabra, Spain

**Received:** November 19, 2010; **Accepted:** September 7, 2011; **Published:** October 12, 2011

**Copyright:** © 2011 Pedersen et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** This research is supported by the Danish National Science Research Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

\* E-mail: lykkep@nbi.dk

## Introduction

A segmented body plan is a fundamental characteristic feature of vertebrates. The process of segmentation is carried out by a combination of changes in gene expression and relative anterior-posterior cell position in the presomitic mesoderm (PSM) [1]. In the anterior end of the embryo the somites are segmented at equally separated time points with species dependent periods. In mice the period is around 120 min and in frogs it is around 90 min.

In 1976 Cooke and Zeeman [2] proposed the clock and wavefront model to describe the segmentation process. The idea is that locally coupled oscillators are controlled by a morphogen gradient in the PSM. The oscillators are the clocks providing temporal information, e.g., cycle state, and the morphogen gradient is the wavefront providing spatial information about axial position. Until now three major pathways controlling the segmentation process have been found: the Notch, Wnt and FGF pathways. They all have target genes, which oscillates and, interestingly, Wnt target genes oscillate out of phase with Notch and FGF target genes [3]. These three pathways could be the clocks. There are decreasing gradients of wnt3a and fibroblast growth factor 8 (fgf8) starting from the tail bud through the PSM [4,5]. The two gradients act in synergy with each other during the somitogenesis [6,7]. The actual setting of the somites happens at the determination front, where the fgf8 level reaches a certain threshold. Cells past this determination front become permissive to form somites depending on their phase of oscillation [8].

In 2003 it was discovered by Aulehla et al. [4] that Axin2 oscillates during the segmentation process in developing mouse embryos. Since their discovery several models for the Wnt oscillator have been proposed [9–11] with Axin2 as a key variable.

However, while Axin2 is a negative regulator of the Wnt pathway, mice with a null mutation of Axin2 do not exhibit a segmentation phenotype – only malformations of skull structures [12]. Therefore, we propose a new model for the core negative feedback loop generating oscillations in the Wnt pathway, with Dickkopf1 (Dkk1), rather than Axin2, closing the feedback loop. Dkk1 has an oscillatory behavior during the segmentation process in mouse embryos [3] and lowered expression of Dkk1 results in smaller and more irregular vertebrae in mice [13,14]; similar to the phenotype produced by overexpression of Wnt3a.

## Analysis

### Modeling the Wnt/ $\beta$ -catenin pathway

During Wnt signaling  $\beta$ -catenin interacts with the TCF/LEF-1 DNA-binding proteins to promote transcription of Wnt target genes [15,16]. As for Axin2 the transcription factor for Dkk1 is  $\beta$ -catenin [17,18]. After transcription and translation Dkk1 goes through the cellular membrane where it can bind to the extracellular domains of the low-density lipoprotein receptor-related protein 5 and 6 (LRP5/6). When bound to LRP5/6, Dkk1 acts as an inhibitor of Wnt signaling by blocking the association between Wnt, Frizzled (Fz) and LRP5/6 [19]. Wnt acts as an inducer for the formation of this complex and Dkk1 is a competitor to this induction [20,21].

It has been proposed that the Wnt signal is transduced through the cell membrane by the binding of Dishevelled (Dsh) to the intracellular domain of the Fz receptor [22]. Axin and Dsh can bind together via their DIX domains [22] and they co-localize at the membrane [23] during Wnt signaling. Therefore Dsh bound to Fz may recruit Axin bound to the glycogen synthase kinase 3 (GSK3 $\beta$ ) to the LRP5/6 receptor [24], where a phosphorylation

of LRP5/6 is initiated. The LRP5/6 receptor has a binding site for Axin and upon Wnt signalling GSK3 $\beta$  (bound to Axin) phosphorylates LRP5/6, which requires Axin [24]. The phosphorylated LRP5/6 receptor may be able to recruit and more efficiently bind the Axin-GSK3 $\beta$  complex to the membrane and the phosphorylation process is thereby amplified [25].

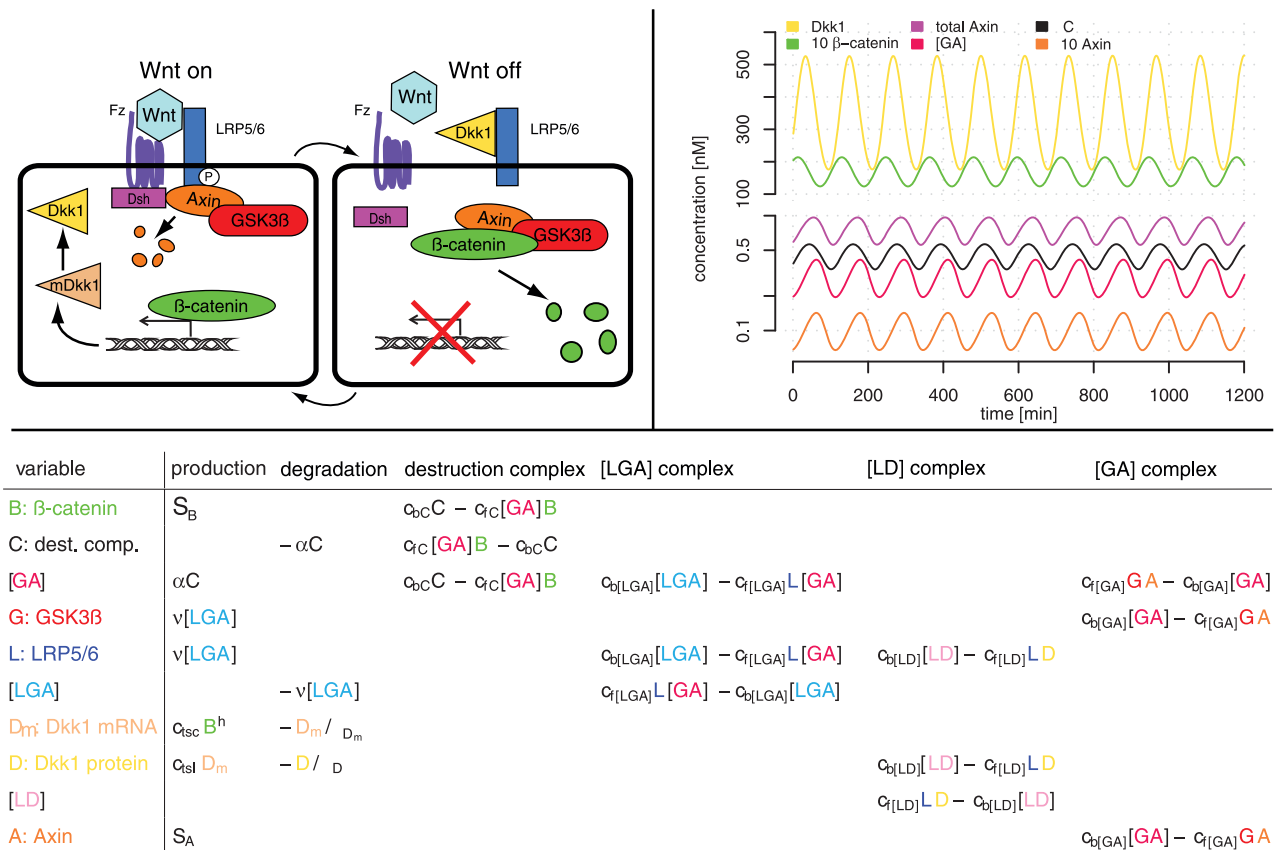
At the cell membrane Axin is phosphorylated by GSK3 $\beta$  and then degraded [26,27]. The degradation of Axin leads to a decrease in the formation of the destruction complex comprised of  $\beta$ -catenin, the two kinases GSK3 $\beta$  and casein kinase (CKI $\alpha$ ), and the scaffolding proteins Axin and adenomatous polyposis coli (APC). In the destruction complex  $\beta$ -catenin gets phosphorylated and subsequently degraded.

Interestingly enough GSK3 $\beta$  plays a dual role in controlling the Wnt signal. When the Wnt signal is off then GSK3 $\beta$  phosphorylates  $\beta$ -catenin in the destruction complex and when the signal is on then it phosphorylates Axin at the LRP5/6 receptor. Whether it is the same or distinct Axin-GSK3 $\beta$  complexes that carry out the phosphorylation of  $\beta$ -catenin and Axin is unknown [28].

Our model does not include the dynamics of the kinase CKI $\alpha$ , the scaffolding protein APC, the DNA binding proteins TCF/LEF-1 and the protein Dsh, since their dynamics are not a major part of the negative feedback loop. The dynamics of CKI $\alpha$  and

APC are included in the parameters governing the destruction complex, the TCF/LEF-1 dynamics are contained within the transcription of Dkk1, and the dynamics of Dsh are included in the formation of the complex consisting of Axin, GSK3 $\beta$  and LRP5/6 at the cell membrane. Figure 1 shows a simplified diagram of the proposed model and the associated equations. The variables  $C$ ,  $[GA]$ ,  $G$ ,  $B$ ,  $L$ ,  $D_m$ ,  $D$ ,  $[LD]$ ,  $A$  and  $[LGA]$  are the concentrations of the destruction complex, GSK3 $\beta$ -Axin complex, GSK3 $\beta$ ,  $\beta$ -catenin, LRP5/6, Dkk1 mRNA, Dkk1 protein, Dkk1-LRP5/6 complex, Axin and LRP5/6-Axin-GSK3 $\beta$  complex. The formation and breaking of a complex  $X$  are denoted by  $c_{fX}$  and  $c_{bX}$ , respectively. The transcription and translation rates of Dkk1 are given by the parameters  $c_{tsc}$  and  $c_{tsl}$ , respectively. The Hill coefficient on  $B$  regarding the transcription of  $D_m$  is associated to the amount of cooperativity between  $\beta$ -catenin and the TCF/LEF-1 complex. For example, no cooperativity would result in a Hill coefficient of one.

The concentration of GSK3 $\beta$  has been shown to be extremely stable [10] and consequently its total concentration,  $GSK3\beta_{tot}$ , is assumed to be constant during the time scales considered. The same assumption goes for the total concentration of LRP5/6, since the half-life of LRP6 is around 4.7 hours [29]. Therefore we have not included any source or sink for the concentrations of  $G$  and  $L$ .



**Figure 1. The diagram, simulation and equations of the Wnt model.** (Top, left): A diagram of the Wnt model with a feed-back loop over Dkk1. Included are only members of the Wnt pathway that are important for the understanding of the negative feedback loop. When the Wnt signal is on, then Axin gets degraded at the LRP5/6 complex and  $\beta$ -catenin can act as a transcription factor of the Wnt inhibitor Dkk1. Vice-versa, when the Wnt signal is off, due to inhibition by Dkk1, then  $\beta$ -catenin gets degraded. (Top, right): Simulated time series for a selection of variables from the model listed in the bottom panel. The total level of Axin (magenta) is low, which complies with the findings of [10]. (Bottom): Equations of the Wnt model split up in terms describing production, degradation and complex dynamics.

doi:10.1371/journal.pone.0025550.g001

Only constitutive sources,  $S_B$  and  $S_A$ , of  $\beta$ -catenin and Axin, respectively, are included in the model, because free (unphosphorylated)  $\beta$ -catenin and Axin is stable [10].

### Determination of the parameter values

The parameters used for our model are listed in Table 1. In Ref. [10] a model for the Wnt pathway in *Xenopus* is proposed and from this article estimates for the dissociation constants for the destruction complex ( $C$ ) and the  $[GA]$  complex, the degradation of  $\beta$ -catenin ( $\alpha$ ), the total concentration of GSK3 $\beta$ , and the sources of Axin and  $\beta$ -catenin are taken. Dkk1 binds to LRP5/6 with a high affinity; the dissociation constant has been measured to be around  $K_{[LD]} = 0.4 - 0.5$  nM [19,20].

The other parameters are estimated to produce oscillations with a period of around 120 min and a very low concentration of total Axin as found by Ref. [10]. The low concentration of total Axin is thought to act as a buffer to changes in the concentration of the other constituents, which may also take part in other signaling pathways [30].

In the activation of transcription of Dkk1 by  $\beta$ -catenin we assume cooperativity between  $\beta$ -catenin and TCF/LEF. The Hill coefficient is set to three. A Hill coefficient of two can produce oscillations, but this requires the affinity of  $\beta$ -catenin to the  $[GA]$  complex to be much higher than suggested by the dissociation constant value used in Ref. [10]. Therefore, we preferred to retain the parameter values of Ref. [10] but increase the Hill coefficient to three. In other models of the Wnt pathway, Hill coefficients of values two and five have been used [9,10,31].

## Results and Discussion

### Oscillations of the mRNA, protein and complex levels

The parameter values produce oscillations of the involved constituents with a period of around 120 min as seen in Fig. 1. In

addition, the concentration of Axin is very low compared to the concentrations of the other variables, as discussed above. The phase shifts between the different oscillatory components of the Wnt model can be explained by considering the sequential steps of the model. An increase in the  $[LGA]$  concentration leads to a decrease in the Axin concentration. This decrease causes a reduction in the concentration of the destruction complex and consequently an increase in the  $\beta$ -catenin concentration. This increase will after a while cause an increase in the Dkk1 concentration, which leads to an increase in the  $[LD]$  concentration. The high affinity of the  $[LD]$  complex leaves little free LRP5/6 behind to form a complex with  $[GA]$ . Thus the concentration of the  $[LGA]$  complex decreases, leading to an increase in Axin, and the cycle continues. The concentrations of GSK3 $\beta$  and  $[GA]$  are mirrors of each other, since a high concentration of  $[GA]$  will leave less free GSK3 $\beta$  behind.

Ref. [32] found no significant oscillations in the level of  $\beta$ -catenin. For our choice of parameters  $\beta$ -catenin shows an oscillatory behavior with an amplitude of approximately 5nM, which is not significantly low. A different set of parameters could possibly give a smaller amplitude of  $\beta$ -catenin but the general results, presented later, are not significantly altered by this. Even though  $\beta$ -catenin does not oscillate it has been shown that the Notch target gene *Nrarp*, which stabilizes LEF-1 [33], does oscillate [3]. LEF-1 does not oscillate in the PSM [34]. Notch and Wnt target genes oscillates out of phase. Thus, when Dkk1 is high, resulting in the inhibition of Wnt signaling, *Nrarp* will have a low expression, resulting in LEF-1 ubiquitination and consequently less Dkk1. The oscillatory behavior seen in Fig. 1 nearly resembles this. Thus, the  $\beta$ -catenin variable in our model can be considered as a coarse-grained variable combining the effects of  $\beta$ -catenin, *Nrarp* and LEF-1.

### Stability of the period and amplitudes of oscillations

In Fig. 2 the period and amplitude of the Dkk1 oscillations are plotted as a function of a selection of parameters. Changing  $K_C$  results in the most drastic changes in amplitude and period of Dkk1 (Fig. 2F). Even though the assembling of GSK3 $\beta$  and Axin is involved in three out of the four complexes in our model, changes in  $K_{[GA]}$  do not affect the period and only for  $K_{[GA]} > 1$  is the amplitude of Dkk1 oscillations really affected, see Fig. 2H.

The degradation of both Axin and  $\beta$ -catenin affects the amplitude. If more  $\beta$ -catenin is degraded then less Dkk1 is transcribed (Fig. 2B). Vice versa with the degradation of Axin. If more Axin gets degraded then less  $\beta$ -catenin gets phosphorylated at the destruction complex resulting in more transcription of Dkk1 (Fig. 2A). As these parameters affect the system in opposite directions, an experiment where the capability of GSK3 $\beta$  to phosphorylate Axin and  $\beta$ -catenin is tested, could shed light on the dual mechanism of GSK3 $\beta$  in the Wnt pathway.

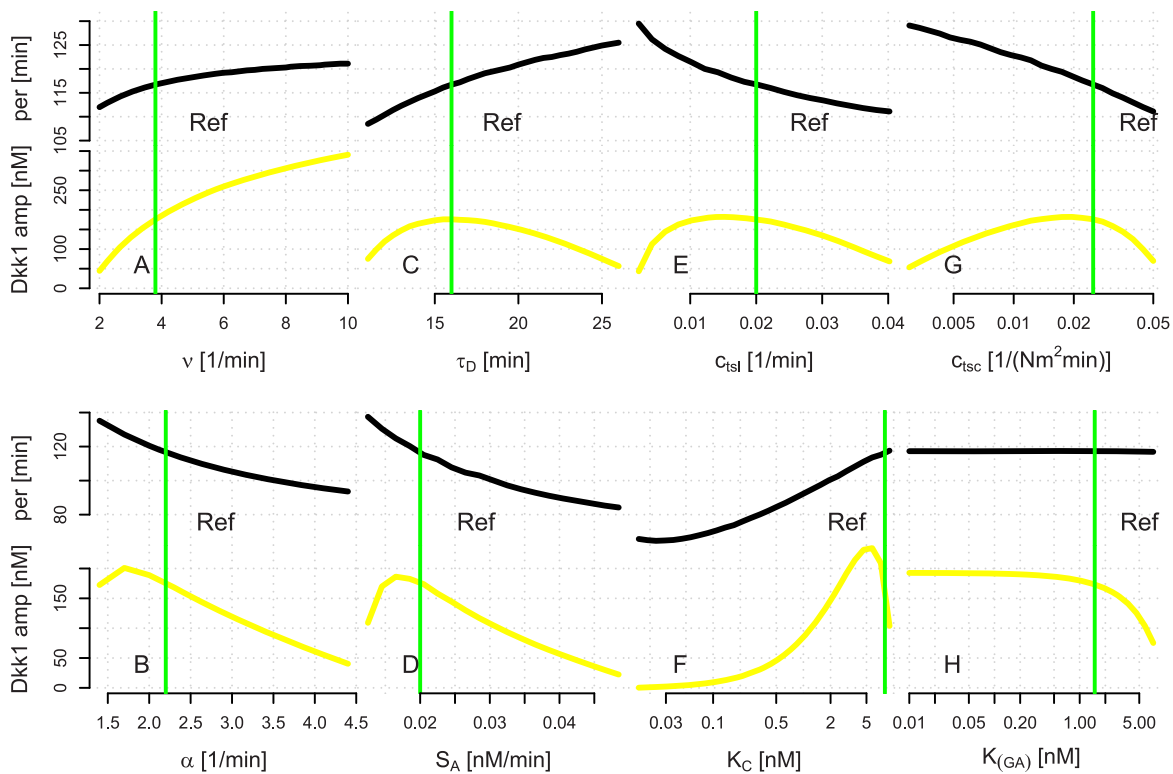
The half-life of Dkk1 affects the period – a shorter half-life leads to a shorter period (Fig. 2C). A similar effect of Axin2 half-life was seen in a previous model with an Axin2 negative feedback loop [11] similar to our Wnt model with a negative Dkk1 loop. Shorter periods are also found when the translation and transcription rates are increased (Fig. 2E,G). In comparing Figs. 2B,D it can be seen that altering  $\alpha$  is almost the same as altering  $S_A$ , which makes sense since the concentration of  $C$  is dependent on both the source of Axin and the rate of  $\beta$ -catenin phosphorylation.

The amplitude of Dkk1 oscillations differs between the case with  $h = 3$  and  $h = 2$ , compare Figs. 1 and S1. Figure S1 shows the oscillations for  $h = 2$  and the parameters used for  $h = 2$  are listed in Table S1. A measure of the actual size of the Dkk1 amplitude along with a measure of the kinetics of the destruction complex

**Table 1.** Parameters in our model of the Wnt system and their default values.

Parameter	Process	Default Value
$K_C$	Dissociation constant $C$	8 nM
$c_{bC}$	Breaking of $C$	$7 \text{ min}^{-1}$
$\alpha$	Degradation of $\beta$ -catenin	$2.2 \text{ min}^{-1}$
$K_{[GA]}$	Dissociation constant $[GA]$	1.5 nM
$c_{b[GA]}$	Breaking of $[GA]$	$4 \text{ min}^{-1}$
$K_{[LGA]}$	Dissociation constant $[LGA]$	1 nM
$c_{b[LGA]}$	Breaking of $[LGA]$	$10 \text{ min}^{-1}$
$\nu$	Degradation of Axin	$3.8 \text{ min}^{-1}$
$K_{[LD]}$	Dissociation constant $[LD]$	0.5 nM
$c_{b[LD]}$	Breaking of $[LD]$	$0.02 \text{ min}^{-1}$
$S_B$	Constant source of $\beta$ -catenin	1 nM/min
$S_A$	Constant source of Axin	0.02 nM/min
$c_{tsl}$	Transcription of <i>dkk1</i>	$0.02 \text{ min}^{-1}$
$c_{isc}$	Translation of Dkk1 mRNA	$0.025 (\text{Mm}^2 \text{ min})^{-1}$
$\tau_{Dm}$	Average lifetime of <i>dkk1</i> mRNA	8 min
$\tau_D$	Average lifetime of Dkk1	16 min
$GSK3\beta_{tot}$	Total G level	45 nM
$L_{tot}$	Total L level	15 nM

doi:10.1371/journal.pone.0025550.t001



**Figure 2. Changes in the periods and amplitudes when varying parameter values.** The period (black) and amplitude (yellow) of Dkk1 oscillations as  $\nu$  (A),  $\alpha$  (B),  $\tau_D$  (C),  $S_A$  (D),  $c_{isl}$  (E),  $K_C$  (F),  $c_{isc}$  (G) and  $K_{[GA]}$  (H) are varied. The green lines refer to the values listed in Table 1. doi:10.1371/journal.pone.0025550.g002

could give a hint of the cooperativity between  $\beta$ -catenin and the TCF/LEF-1 complex. If the real Hill coefficient is two, then our model predicts that to produce oscillations they must also bind with a high affinity, and that the resultant oscillations of Dkk1 will have a relatively smaller amplitude than if the Hill coefficient is three.

### A spatial Wnt gradient induced by time variation

The oscillations of the variables in our model could function as the segmentation clock postulated in the clock and wavefront model [2]. We now investigate whether such oscillations can consistently be coupled to a wavefront, i.e. a spatial gradient of Wnt. As oscillating cells move through the PSM they effectively see a decreasing level of Wnt in time. Therefore we model the spatial Wnt gradient simply by a time dependent decrease in the parameter  $c_{f[GA]}$ . Other models have also been proposed with a gradient of a morphogen protein [35–37]. These models are complementary to ours in that they use abstract models of the clock, not any concrete mRNA, proteins and complexes interacting as in our model.

The reason for using  $c_{f[GA]}$  as the time dependent parameter to mimic a Wnt gradient is clear if we introduce the variable  $[LW]$  describing the binding of Wnt ( $W$ ) to the LRP5/6 receptor:

$$\frac{d[LW]}{dt} = c_{f[LW]}LW(t) - c_{b[LW]}[LW], \quad (1)$$

and if we assume steady state for the binding of Wnt to LRP5/6 then

$$[LW] = \frac{c_{f[LW]}LW(t)}{c_{b[LW]}} = \frac{LW(t)}{K_{[LW]}}. \quad (2)$$

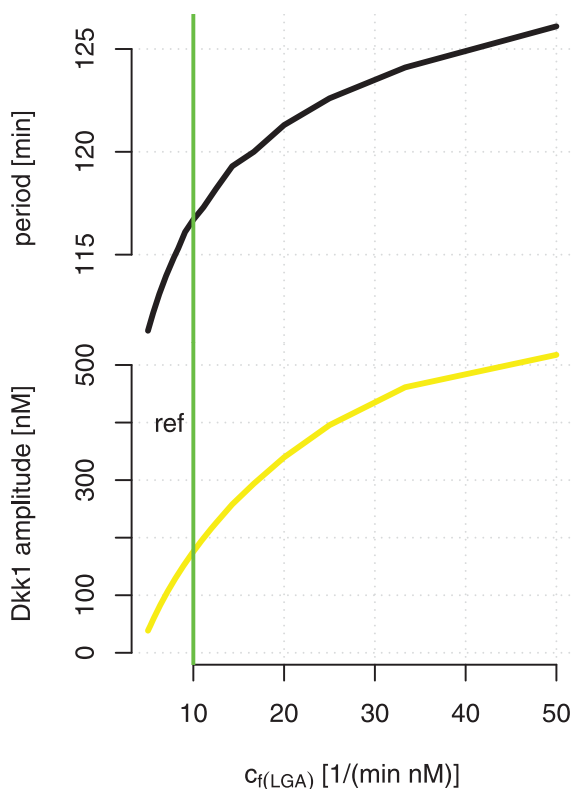
When substituting  $[LW]$  for  $L$  in the equations with terms governing the formation of the  $[LGA]$  complex, it can be seen that the rate constant for formation of  $[LGA]$  will be time dependent

$$c_{f[LGA]}(t) = \frac{c_{f[LGA]}W(t)}{K_{[LW]}}. \quad (3)$$

For simplicity we will assume that the Wnt gradient, and hence  $c_{f[LGA]}$ , has a Gaussian profile, which is what one would expect if the gradient was determined mainly by diffusive processes. For reference, Fig. 3 shows the amplitude and period of Dkk1 concentration for a range of  $c_{f[LGA]}$  values. The green line refers to the reference value of  $c_{f[LGA]}$ ; reducing it results in smaller amplitudes.

The length of the PSM is approximately constant during the formation of the first 15–20 somites in mice embryos [38]. The same is almost true for the size of the somites. At this stage the somites are  $\sim 100 \mu\text{m}$  and the PSM is about 1 mm, i.e. the PSM has a length corresponding to the length of 10 somites. Thus, a cell budded off in the tail bud at this stage will be segmented in around  $\sim 1100$ – $1300$  min.

It has been measured that FGF (regulated by Wnt) exhibits a gradient in the PSM with a fold change of two to five [5]. The fold change used for  $c_{f[LGA]}$  is two. Assuming that Wnt diffuses through



**Figure 3. The effect of the Wnt level on the period and amplitude.** The amplitude (solid) and the period (dashed) of Dkk1 oscillations for varying values of the parameter  $c_{f[LGA]}$ . The green vertical line denotes the value of  $c_{f[LGA]}$  in the reference state. doi:10.1371/journal.pone.0025550.g003

the PSM and setting the final value of  $c_{f[LGA]}$  equal to  $5 \frac{1}{\text{nMmin}}$  enable us to calculate a Gaussian profile of  $c_{f[LGA]}$  representing the Wnt gradient. In Fig. S2, Gaussian profiles of  $c_{f[LGA]}$  in the PSM are plotted with different initial values. A decreasing value of  $c_{f[LGA]}$  in the PSM will give rise to smaller amplitudes and slightly shorter periods. Experimentally it has been shown that the wavelength of the oscillations in the PSM decreases from the tail bud to the determination front [39]. If the oscillations are proportional to the wavelength, then the steepness of the  $c_{f[LGA]}$  profile sets the pace with which the wavelength decreases.

### Synchronization of neighbouring cells

Figure 4A shows the Dkk1 oscillations in an elongating embryo with a decreasing  $c_{f[LGA]}$  as described above. The elongation of the embryo is modeled in a very simplified manner. It is considered in only one dimension and a cell buds off from the tail bud at regular time intervals denoted by  $R$ . In Fig. 4 we use  $R = 10$  min. We further assume that the initial state of a newly budded cell in the PSM is the same as that of its anterior neighboring cell, i.e. if the state for a cell  $i$  at time  $t$  is denoted  $S_i(t)$ , then the initial state of a cell  $i+1$  is  $S_{i+1}(0) = S_i(R)$ . Thus, we effectively introduce a synchronization in the oscillations of adjacent cells. But this is put in by hand, rather than by an explicit coupling between cells in the model. The final level of Dkk1 at the determination front is oscillating (see Fig. 4B with a period of 120 min. It is the built-in synchronization that is the reason behind this. If, instead, all the cells were assigned the same initial state, i.e.  $S_{i+1}(0) = S_i(0)$ , then

the levels of Dkk1 at the determination front would also be equal, as in Fig. 4C. In a real embryo the synchronization, of course, occurs through coupling of the individual clocks in the PSM cells [40], which we have not modeled. However, the way we put in synchronization in our model is sufficient to demonstrate that its presence is necessary for a proper function of the segmentation process. The importance of synchronization has also been shown experimentally [41].

One could imagine that the synchronization of the clocks is not perfect. If the initial state is randomly chosen within the whole range of Dkk1 levels, then the oscillations of Dkk1 at the determination front are disrupted and no periodicity is visible (see Fig. S3B). If the initial state of a cell  $i+1$  is chosen randomly within the interval  $S_i(R - \frac{R}{2}) - S_i(R + \frac{R}{2})$ , then the period is almost unaltered (see Fig. S3A). Thus, the system seems to be robust to small changes in the synchronization.

In zebrafish the mechanism of synchronization is well understood by Delta-Notch interactions [42]. In the literature we have found models which couple the cells in zebrafish [40,42–45] by various mechanisms, such as coupling of phase oscillators and coupling of oscillating clock genes with a signalling protein. The model of Ref. [44] shows that even a weak coupling helps synchronization. In the embryo, mitosis and stochastic gene expression could result in nonlinear noise [40] that could disrupt the synchronization if they are not coupled strongly enough.

In the above simulations, a Gaussian profile of  $c_{f[LGA]}$  is used, since diffusion of Wnt is thought to be the main reason for the Wnt gradient in the PSM. The profile of  $c_{f[LGA]}$  would still be Gaussian if we also included a half-life of Wnt. If the Wnt gradient was only controlled by the half-life of Wnt then the levels of Dkk1 still oscillate at the determination front with a period of 120 min, when the model is simulated as above with  $R = 10$  min.

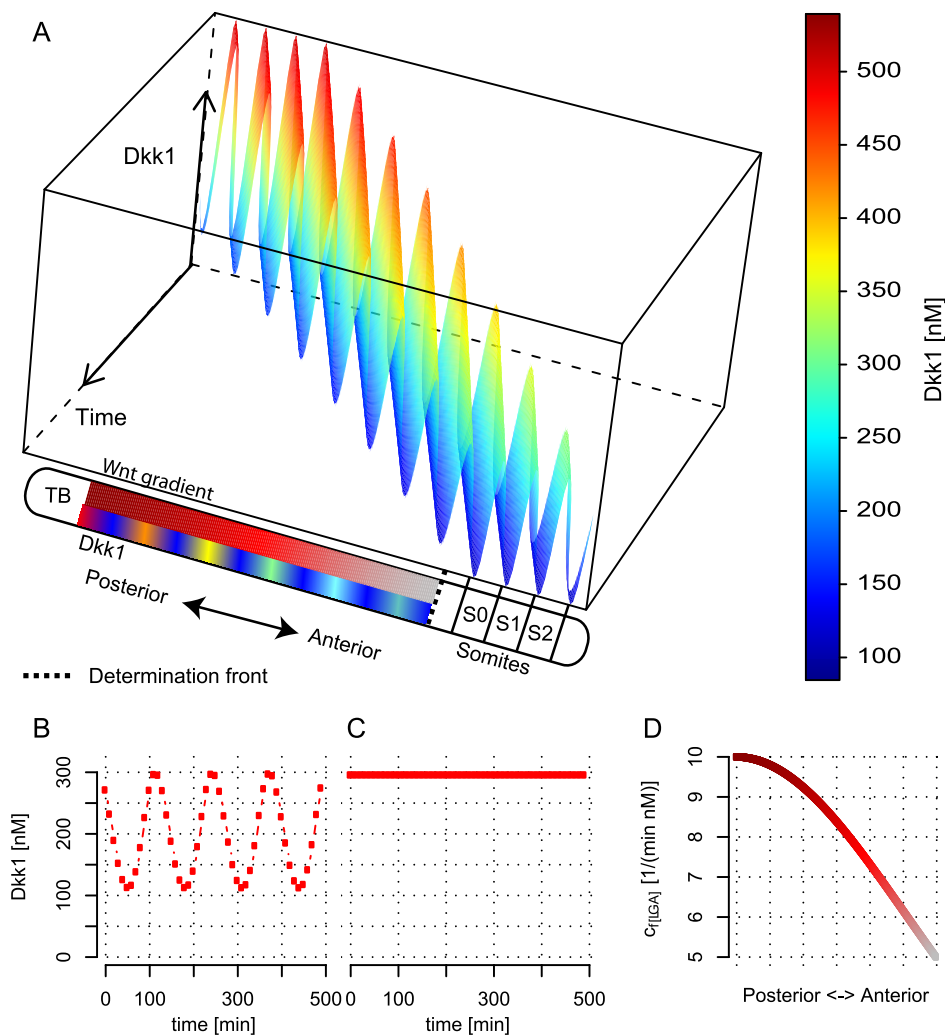
Because Ref. [32] found a decreasing gradient of  $\beta$ -catenin in the PSM, we also tried giving  $S_B$  a Gaussian profile in the PSM with a fold change of two from the tail bud to the anterior part of the PSM. This did not alter the 120 min oscillations of Dkk1 at the determination front, when the model was simulated as above with  $R = 10$  min.

### A decreasing Wnt level in the tail bud and an increase in period of segmentation

It is known that the period of somite formation increases [38] during late stages of somitogenesis in various organisms. In 2004 Aulehla and Hermann [46] hypothesized that an increase in the Wnt level of the tail bud could result in longer periods of the segmentation process observed in mice embryos. In our model, simulating an increasing level of Wnt ( $c_{f[LGA]}$  from  $10 \frac{1}{\text{nMmin}}$  to  $14 \frac{1}{\text{nMmin}}$ ) in the tail bud we do see that the period initially lengthens and the amplitude increases, see Fig. S4B. However, if the Wnt level is increased further the period decreases to as low as 30 min. An experiment where Wnt is upregulated in the tail bud would elucidate their hypothesis and our findings.

Recent experiments with chick embryos show the opposite – Wnt is downregulated in the tail bud at late stages of somitogenesis [34]. In our model, simulating a linear decrease of Wnt, through the parameter  $c_{f[LGA]}$ , from  $10 \frac{1}{\text{nMmin}}$  to  $5 \frac{1}{\text{nMmin}}$ , causes the period of Dkk1 oscillations to increase. The mechanism behind this increase in the period is different from that causing a small increase in the period as described above. There the Wnt level was decreased throughout the PSM, where here the initial level of Wnt in the tail bud is decreased, i.e. the gradient of Wnt in the PSM





**Figure 4. Synchronization of neighboring cells.** (A) Time series for the Dkk1 concentration. Space is introduced by letting a cell bud off from the tail bud every 10th min. Thus the cells move relatively in the PSM. At the determination front the oscillations arrest. (B/C) The level of Dkk1 at the determination front with  $R = 10$  min and synchronization between neighboring cells and with  $R = 10$  min and the cells have the same initial level of Dkk1 (C). (D) The Gaussian profile of  $c_{[LGA]}$  in the anteroposterior direction. TB: tail bud.  $S_i$ : Somite  $i$ , where  $S_0$  is the newly formed somite. doi:10.1371/journal.pone.0025550.g004

becomes less steep. The amplitude of the Dkk1 oscillations decreases at the same time, which is expected because the segmentation process does stop as the Wnt level diminishes.

## Outlook

We propose a Wnt model with Dkk1 as the core for the negative feedback loop which exhibits sustained oscillations of Dkk1. The clock and wavefront model was investigated using the Dkk1 oscillations as the clock and a Wnt gradient in the PSM as the wavefront. By simulating the elongating embryo we were able to test the importance of synchronization between neighboring cells. In addition, we could also show that small errors in the synchronization did not significantly disrupt smooth oscillations of the Dkk1 levels at the determination front. We could reproduce the experimental finding in chick embryos that downregulation of Wnt in the tail bud might lengthen the oscillation time periods during late stages of somitogenesis. The negative feedback loop involving Dkk1 introduced produces very similar behavior as the Axin2 negative feedback loops modeled

previously. Thus, it is conceivable that these two loops function together, providing some redundancy with respect to each other. This could explain why Axin2 null mutant mice do not show any segmentation phenotype and Dkk1 null mutant mice show only some irregularity in the vertebrae. Such redundancy has been seen in the case of *fgf4* and *fgf8*. Neither are individually essential, but removing both disrupts somitogenesis [7]. It would be interesting to see what phenotype a double knockout of both Axin2 and Dkk1 exhibits.

## Supporting Information

**Figure S1 Oscillations of the Wnt model with  $h = 2$ .** The Wnt model is simulated with a Hill coefficient of  $h = 2$ . The model still shows oscillations with a period of around 120 min, but the affinity of the  $\beta$ -catenin to bind the  $[GA]$  complex needs to be much higher than expected from experiments. The parameters used in this simulation can be found in Table S1.

(PDF)

**Figure S2 Gaussian profiles of the Wnt gradient.** (Top): Gaussian profiles of  $c_{f[LGA]}$  in the PSM are plotted with different initial values. A decreasing value of  $c_{f[LGA]}$  in the PSM will give rise to smaller amplitudes (middle) and slightly shorter/ almost constant periods (bottom). The abbreviation  $ref \pm i$  denotes the reference state value of  $c_{f[LGA]} \pm i$ .  
(PDF)

**Figure S3 Desynchronization of neighboring cells.** If the desynchronization between neighboring cells is small (A) then the oscillations of the Dkk1 level at the determination front is almost unaltered. However, if the desynchronization is strong then these oscillations are not appearing.  
(PDF)

**Figure S4 The Wnt level in the tail bud.** (A): When the Wnt level decreases in the tail bud, then the period of the Dkk1 level at

the determination front extends. (B): The period is also extended when the Wnt level increases in the tail bud, but the periods drop significantly below the period in the reference state.  
(PDF)

#### Table S1

(PDF)

#### Acknowledgments

The authors would like to thank Albert Goldbeter for useful discussions.

#### Author Contributions

Wrote the paper: LP SK MHJ.

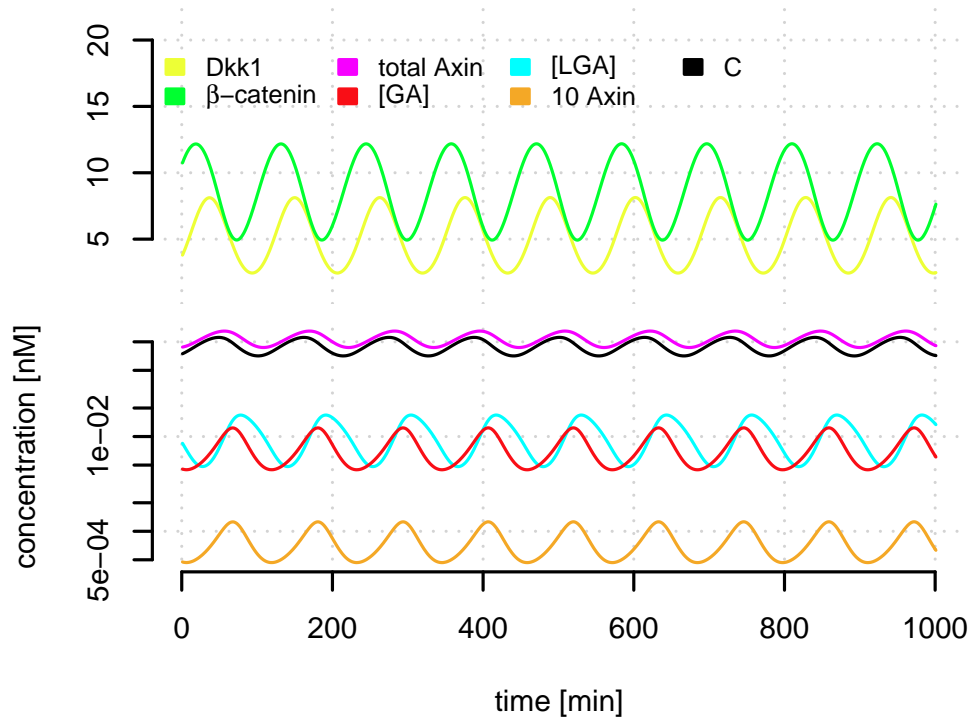
#### References

- Wolpert L, Jessel T, Lawrence P, Meyerowitz E, Robertson E, et al. (2006) Principles of development. Oxford University Press, 3rd edition.
- Cooke J, Zeeman EC (1976) A clock and wavefront model for control of the number of repeated structures during animal morphogenesis. *J Theor Biol* 58: 455–476.
- Dequ  ant ML, Glynn E, Gaudenz K, Wahl M, Chen J, et al. (2006) A complex oscillating network of signaling genes underlies the mouse segmentation clock. *Science* 314: 1595–1598.
- Aulehla A, Wehrle C, Brand-Saberi B, Kemler R, Gossler A, et al. (2003) Wnt3a plays a major role in the segmentation clock controlling somitogenesis. *Dev Cell* 4: 395–406.
- Dubrule J, Pourqu   O (2004) *fgf8* mRNA decay establishes a gradient that couples axial elongation to patterning in the vertebrate embryo. *Nature* 427: 419–422.
- ten Berge D, Brugmann S, Helms J, Nusse R (2008) Wnt and fgf signals interact to coordinate growth with cell fate specification during limb development. *Development* 135: 3247–3257.
- Naiche L, Holder N, Lewandoski M (2011) *Fgf4* and *fgf8* comprise the wavefront activity that controls somitogenesis. *PNAS* 108: 4010–4023.
- Dubrule J, McGrew M, Pourqu   O (2001) *Fgf* signaling controls somite boundary position and regulates segmentation clock control of spatiotemporal *hox* gene activation. *Cell* 106: 219–232.
- Goldbeter A, Pourqu   O (2008) Modeling the segmentation clock as a network of coupled oscillations in the notch, wnt and fgf signaling pathways. *J Theor Biol* 252: 574–585.
- Lee E, Salic A, Kr  ger R, Heinrich R, Kirschner MW (2003) The roles of *apc* and *axin* derived from experimental and theoretical analysis of the wnt pathway. *PLoS Biol* 1: E10.
- Jensen PB, Pedersen L, Krishna S, Jensen MH (2010) A wnt oscillator model for somitogenesis. *Biophysical journal* 98: 943–50.
- Yu HMI, Jerchow B, Sheu TJ, Liu B, Costantini F, et al. (2005) The role of *axin2* in calvarial morphogenesis and craniosynostosis. *Development* 132: 1995–2005.
- MacDonald B, Adamska M, Meisler M (2004) Hypomorphic expression of *dkk1* in the double-ridge mouse: dose dependence and compensatory interactions with *lrp6*. *Development* 131: 2543–2552.
- Mukhopadhyay M, Shtrom S, Rodriguez-Esteban C, Chen L, Tsukui T, et al. (2001) *Dickkopf1* is required for embryonic head induction and limb morphogenesis in the mouse. *Developmental cell* 1: 423–434.
- Behrens J, von Kries JP, K  hl M, Bruhn L, Wedlich D, et al. (1996) Functional interaction of beta-catenin with the transcription factor *lef-1*. *Nature* 382: 638–642.
- Korinek V, Barker N, Willert K, Molenaar M, Roose J, et al. (1998) Two members of the *tcf* family implicated in wnt/beta-catenin signaling during embryogenesis in the mouse. *Molecular and cellular biology* 18: 1248–1256.
- Niida A, Hiroko T, Kasai M, Furukawa Y, Nakamura Y (2004) *Dkk1*, a negative regulator of wnt signaling, is a target of the -catenin/tcf pathway. *Oncogene* 23: 8520–8526.
- Gonz  lez-Sancho JM, Aguilera O, Garc  a JM, Pend  s-Franco N, Pe  a C, et al. (2005) The wnt antagonist *dickkopf-1* gene is a downstream target of beta-catenin/tcf and is downregulated in human colon cancer. *Oncogene* 24: 1098–1103.
- Sem  nov MV, Tamai K, Brott BK, K  hl M, Sokol S, et al. (2001) Head inducer *dickkopf-1* is a ligand for wnt coreceptor *lrp6*. *Curr Biol* 11: 951–61.
- Bafico A, Liu G, Yaniv A, Gazit A, Aaronson S (2001) Novel mechanism of wnt signalling inhibition mediated by *dickkopf-1* interaction with *lrp6/arrow*. *Nature cell biology* 3: 683–686.
- Tamai K, Semenov M, Kato Y, Spokony R, Liu C, et al. (2000) Ldl-receptor-related proteins in wnt signal transduction. *Nature* 407: 530–5.
- Wallingford J, Habas R (2005) The developmental biology of dishevelled: an enigmatic protein governing cell fate and cell polarity. *Development* 132: 4421–4436.
- Fagotto F, Jho E, Zeng L, Kurth T, Joos T, et al. (1999) Domains of *axin* involved in protein-protein interactions, wnt pathway inhibition, and intracellular localization. *J Cell Biol* 145: 741–756.
- Zeng X, Huang H, Tamai K, Zhang X, Harada Y, et al. (2008) Initiation of wnt signaling: control of wnt coreceptor *lrp6* phosphorylation/activation via *frizzled*, *dishevelled* and *axin* functions. *Development* 135.
- Baig-Lewis S, Peterson-Nedry W, Wehrli M (2007) Wingless/wnt signal transduction requires distinct initiation and amplification steps that both depend on *arrow/lrp*. *Developmental Biology* 306: 94–111.
- Mao J, Wang J, Liu B, Pan W, Farr GH, et al. (2001) Low-density lipoprotein receptor-related protein-5 binds to *axin* and regulates the canonical wnt signaling pathway. *Mol Cell* 7: 801–809.
- Yamamoto H, Kishida S, Kishida M, Ikeda S, Takada S, et al. (1999) Phosphorylation of *axin*, a wnt signal negative regulator, by glycogen synthase kinase-3beta regulates its stability. *J Biol Chem* 274: 10681–10684.
- Zeng X, Tamai K, Doble B, Li S, Huang H, et al. (2005) A dual-kinase mechanism for wnt coreceptor phosphorylation and activation. *Nature* 438: 873–877.
- Sem  nov MV, Zhang X, He X (2008) *Dkk1* antagonizes wnt signaling without promotion of *lrp6* internalization and degradation. *J Biol Chem* 283: 21427–32.
- Logan CY, Nusse R (2004) The wnt signaling pathway in development and disease. *Annu Rev Cell Dev Biol* 20: 781–810.
- Wawra C, K  hl M, Kestler HA (2007) Extended analyses of the wnt/beta-catenin pathway: robustness and oscillatory behaviour. *FEBS Lett* 581: 4043–8.
- Aulehla A, Wieg  r W, Baubert V, Wahl MB, Deng C, et al. (2008) A beta-catenin gradient links the clock and wavefront systems in mouse embryo segmentation. *Nature cell biology* 10: 186–93.
- Ishitani T, Matsumoto K, Chitnis AB, Itoh M (2005) *Nrarp* functions to modulate neural-crest-cell differentiation by regulating *lef1* protein stability. *Nature cell biology* 7: 1106–12.
- Gibb S, Zagorska A, Melton K, Tenin G, Vacca I, et al. (2009) Interfering with wnt signalling alters the periodicity of the segmentation clock. *Dev Biol* 330: 21–31.
- Fran  ois P, Hakim V, Siggia ED (2007) Deriving structure from evolution: metazoan segmentation. *Mol Syst Biol* 3: 154.
- Baker RE, Schnell S, Maini PK (2006) A clock and wavefront mechanism for somite formation. *Dev Biol* 293: 116–26.
- Santill  n M, Mackey MC (2008) A proposed mechanism for the interaction of the segmentation clock and the determination front in somitogenesis. *PLoS ONE* 3: e1561.
- Tam PP (1981) The control of somitogenesis in mouse embryos. *J Embryol Exp Morphol* 65(Suppl): 103–128.
- Giudicelli F, Ozbudak EM, Wright GJ, Lewis J (2007) Setting the tempo in development: an investigation of the zebrafish somite clock mechanism. *PLoS Biol* 5: e150.
- Horikawa K, Ishimatsu K, Yoshimoto E, Kondo S, Takeda H (2006) Noise-resistant and synchronized oscillation of the segmentation clock. *Nature* 441: 719–23.
- Jiang YJ, Aerns BL, Smithers L, Haddon C, Ish-Horowicz D, et al. (2000) Notch signalling and the synchronization of the somite segmentation clock. *Nature* 408: 475–9.
- Ozbudak EM, Lewis J (2008) Notch signalling synchronizes the zebrafish segmentation clock but is not needed to create somite boundaries. *PLoS Genet* 4: e15.
- Campanelli M, Gedeon T (2010) Somitogenesis clock-wave initiation requires differential decay and multiple binding sites for clock protein. *PLoS Comput Biol* 6: e1000728.

44. Morelli L, Ares S, Herrgen L, Schröter C, Jülicher F, et al. (2009) Delayed coupling theory of vertebrate segmentation. *HFSP Journal* 3: 55–66.
45. Riedel-Kruse IH, Müller C, Oates AC (2007) Synchrony dynamics during initiation, failure, and rescue of the segmentation clock. *Science* 317: 1911–5.
46. Aulehla A, Herrmann BG (2004) Segmentation in vertebrates: clock and gradient finally joined. *Genes Dev* 18: 2060–2067.

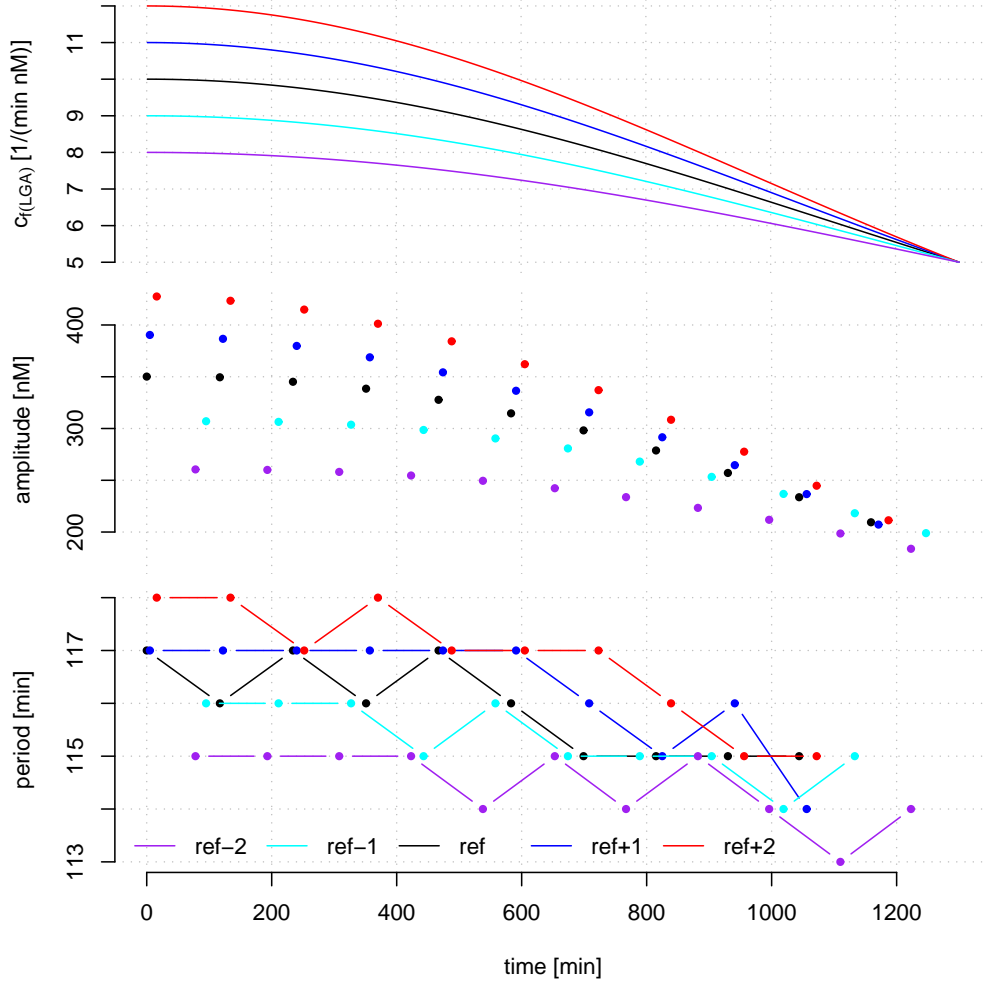


**Figure S1 - Oscillations of the Wnt model with  $h=2$**



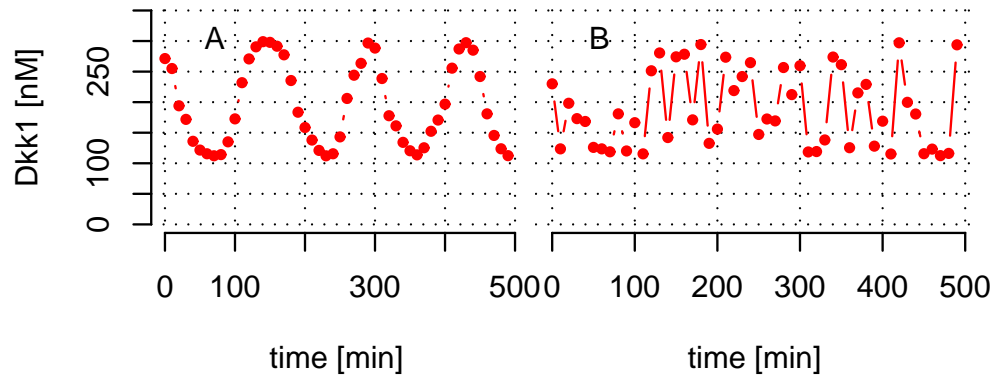
The Wnt model is simulated with a Hill coefficient of  $h = 2$ . The model still shows oscillations with a period of around 120 min, but the the affinity of the  $\beta$ -catenin to bind the  $[GA]$  complex needs to be much higher than expected from experiments. The parameters used in this simulation ca be found in Table S1.

**Figure S2 - Gaussian profiles of the Wnt gradient**



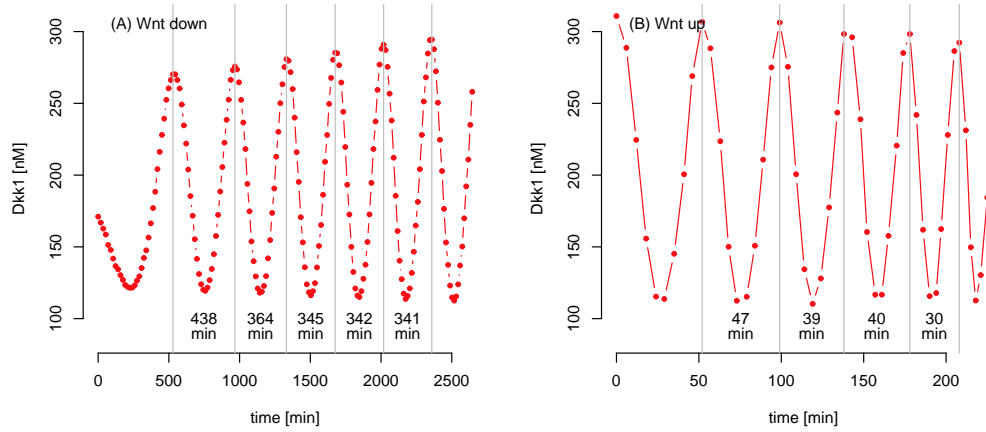
**Top:** Gaussian profiles of  $c_{f[LGA]}$  in the PSM are plotted with different initial values. A decreasing value of  $c_{f[LGA]}$  in the PSM will give rise to smaller amplitudes (**middle**) and slightly shorter/ almost constant periods (**bottom**). The abbreviation  $ref \pm i$  denotes the reference state value of  $c_{f[LGA]} \pm i$ .

**Figure S3 - Desynchronization of neighboring cells**



If the desynchronization between neighboring cells is small (**A**) then the oscillations of the Dkk1 level at the determination front is almost unaltered. However, if the desynchronization is strong then these oscillations are not appearing.

**Figure S4 - The Wnt level in the tail bud**



(A) When the Wnt level decreases in the tail bud, then the period of the Dkk1 level at the determination front extends. (B) The period is also extended when the Wnt level increases in the tail bud, but the periods drop significantly below the period in the reference state.

## Table S1

Parameters in our model of the Wnt system with  $h = 2$

Parameter	Process	Default Value
$K_C$	Dissociation constant $C$	0.4 nM
$c_{bC}$	Breaking of $C$	$16 \text{ min}^{-1}$
$\alpha$	Degradation of $\beta$ -catenin	$10 \text{ min}^{-1}$
$K_{[GA]}$	Dissociation constant $[GA]$	0.5 nM
$c_{b[GA]}$	Breaking of $[GA]$	$10 \text{ min}^{-1}$
$K_{[LGA]}$	Dissociation constant $[LGA]$	0.5 nM
$c_{b[LGA]}$	Breaking of $[LGA]$	$10 \text{ min}^{-1}$
$\nu$	Degradation of Axin	$0.2 \text{ min}^{-1}$
$K_{[LD]}$	Dissociation constant $[LD]$	0.5 nM
$c_{b[LD]}$	Breaking of $[LD]$	$0.02 \text{ min}^{-1}$
$S_B$	Constant source of $\beta$ -catenin	0.9 nM/min
$S_A$	Constant source of Axin	0.002 nM/min
$c_{tsl}$	Transcription of <i>dkk1</i>	$0.025 \text{ min}^{-1}$
$c_{tsc}$	Translation of Dkk1 mRNA	$0.025 (\text{Nm}^2 \text{ min})^{-1}$
$\tau_{Dm}$	Average lifetime of dkk1 mRNA	10 min
$\tau_D$	Average lifetime of Dkk1	10 min
$GSK3\beta_{tot}$	Total G level	50 nM
$L_{tot}$	Total L level	7 nM

## APPENDIX

### F

#### ICA paper

This paper is under revision in PLoS ONE.

# Ranking of Independent Components in Gene Expression Microarray Data Using Weighted Scoring

Lykke Pedersen<sup>1,\*</sup>, Peter H Hagedorn<sup>2,3</sup>

**1** Center for Models of Life, Niels Bohr Institute, University of Copenhagen, DK-2100 Copenhagen, Denmark

**2** Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, DK-2800 Lyngby, Denmark

**3** Present address: Santaris Pharma, DK-2970 Hørsholm, Denmark

\* E-mail: Corresponding lykkep@nbi.dk

## Abstract

Decomposition of gene expression microarray data using independent component analysis (ICA) has been used to identify and separate underlying transcriptional programs from observed gene expressions. In contrast to other linear data representations, such as principal component analysis (PCA) or linear discriminant analysis (LDA), however, ICA does not provide an implicit ranking of the identified components. We score each independent component according to its stability, contribution to overall gene expression, unsupervised information content, and, when applicable, ability to separate sample groups as well as representation in known biological processes and pathways. A weighted sum of these scores supplies a simple and robust ranking of the independent components. This is shown for synthetic microarray data where the transcriptional programs are known by design, as well as for three instructive biological examples. For each example, inspection of the individual score values provides additional insights into the underlying biology. Weighted scoring of a few key attributes of independent components provides a generally applicable and unbiased method for ranking, as well as exploring, such components in gene expression microarray data.

## Introduction

High-dimensional biological data, such as data from gene expression microarrays, can be interpreted as being composed of overlapping sets of transcriptional programs, or gene expression modules, that explain most of the differential expressions observed in the data. Based on the known regulatory control and functional role of genes in an identified transcriptional program, one can infer how overall biological processes and metabolic pathways are affected under different conditions [1–3]. Mathematically, suppose that  $T$  transcriptional programs are active in the samples under study, each of which is a (column) vector  $\vec{C}_t$ ,  $t = 1, \dots, T$ , representing  $G_t$  gene inductions or repressions. With microarrays, we measure the expression of all  $G$  (where  $G_t \subseteq G$ ) induced or repressed genes in  $S$  samples. Assuming that each of these programs can be identified by a simple linear decomposition of the  $G \times S$  expression matrix  $E$ , we can write

$$E = CM \text{ or } E_{gs} = \sum_{t=1}^T C_{gt} M_{ts} \text{ or } \tilde{E}_s = \sum_{t=1}^T C_t M_{ts}$$

where  $C$  is the  $G \times T$  matrix containing the transcriptional programs, in coefficient vector notation,  $C = (\vec{C}_1, \vec{C}_2, \dots, \vec{C}_T)$ , likewise  $E = (\vec{E}_1, \vec{E}_2, \dots, \vec{E}_S)$ , and  $M$  is the  $T \times S$  matrix giving the linear mixing of each program in each sample. The decomposition can be done in numerous ways. Two-way clustering of correlated genes and samples into non-overlapping sets can be represented by component vectors  $\vec{C}_t$  with discrepant sets of non-vanishing entries  $C_{gt} \neq 0 \Leftrightarrow C_{gj} \approx 0 \forall j \neq t$ . Principal component analysis (PCA) decomposes  $E$  such that the transcriptional programs, the principal components, are mutually orthogonal,  $\vec{C}_t \perp \vec{C}_j \forall j \neq t$ . Finally, linear discriminant analysis (LDA) decomposes  $E$  in a manner that

maximally separates the samples based on pre-defined classes. Both clustering, PCA, and LDA are used extensively when analyzing profiling data.

If statistical independence, instead of e.g. orthogonality, between transcriptional programs is assumed, one can decompose  $E$  using independent component analysis (ICA) [4]. This decomposition has been shown to be informative for microarray analysis by ourselves and others [5–11], and compares favorably to clustering and PCA [12–15]. Apparently, the representation of gene expression as a mix of independent, possibly overlapping, transcriptional programs, often matches the biological reality well [16]. Specifically, the independent components (ICs) identified using ICA has been shown to capture the differential regulation of well-defined biological processes and metabolic pathways, in breast cancer [15], acute myeloid leukemia [17], and Alzheimer’s disease [18]. Tumor classification based on ICs also seems to be a promising application of ICA [19, 20]. Recently, ICA has also been used to decompose the effects of multiple deregulated microRNAs on their partially overlapping sets of mRNA targets [10, 21].

ICA extracts  $n = \min\{G, S\}$  ICs. Since the number of samples is almost always much lower than the number of genes, we assume  $n = S$ . The number of interesting transcriptional programs  $T$  may be equal to, or lower than  $S$ . A common practice is to project  $E$  onto its first  $O$  principal component directions, and then extract  $O$  ICs in this subspace. This approach has several important drawbacks. First, the number of interesting transcriptional programs is not known beforehand, and one therefore needs to guess or resort to a recursive trial-and-error approach to choose an  $O$  that allows all  $T$  transcriptional programs to be extracted. And second, it is not obvious that a projection onto  $O < S$  principal component directions preserves all  $T$  transcriptional programs. In fact, as shown by Chang [22] one can easily construct examples where two clusters in the data are only well separated in the subspace of the first and last principal components. Complementing this, Yeung and Ruzzo [23] showed that clustering in a subspace spanned by principal components generally degrades cluster quality. Finally, even when a projection onto  $T$  principal components allows all the  $T$  interesting transcriptional programs to be extracted by ICA, one still needs a way to explore and rank these ICs to clarify their significance.

By extracting all  $S$  ICs and then identify and rank those that represent interesting transcriptional programs, these issues are avoided. One possible ranking that has been suggested is based on calculating the statistical overrepresentation of genes related to specific biological processes and metabolic pathways among the genes with the absolute largest loads  $|C_{gt}|$  in each component  $\vec{C}_t$  [13]. This, however, requires that all genes are annotated with respect to the transcriptional programs under study, which might often not be the case. Another possibility is to use the information contained in the loads  $C_{gt}$  of each component, along with the manner in which these ICs are represented in each sample given by the mixes  $M_{ts}$ . For specific applications of ICA, various methods for inherent identification by such load-based ranking have been used. Liebermeister [12] calculates the negentropy and captured variance for each component. The captured variance is similar to the energy, defined as  $E_t = \sum_s M_{ts}^2$ , and used to sort ICs by Winther and Petersen [24]. Chiappetta et al. [6] rank the ICs according to a stability score based on repeated runs of the algorithm. Finally, an ensemble feature selection approach that requires the data to be divided into training and test sets has been suggested [25].

Although each of these ranking schemes performs reasonably well for some studies, they are not, we find, generally applicable for all the various kinds of transcriptional regulations probed by gene expression microarrays. As introduced below, however, a simple weighted scoring scheme stably identifies the informative ICs across three highly different studies. It encompasses some of the already proposed methods for identification, as well as a few sensible new ones.



## Methods

### ICA algorithm and preprocessing

The fastICA algorithm [26] is used to estimate the component matrix  $\mathbf{C}$  and mixing matrix  $\mathbf{M}$ , by numeric maximization of the negentropy of the ICs. The negentropy is approximated by  $J(y) = (\langle G(y) \rangle - \langle G(v) \rangle)^2$ , where  $y$  is the distribution of loads,  $v$  is a random variable distributed as  $\mathcal{N}(0, 1)$ , and the contrast function  $G(u)$  is here defined as  $G(u) = -\exp(-u^2/2)$ . By convention, the loads in each component have mean 0 and variance 1.

The expression matrix  $\mathbf{E}$  is normalized by GCRMA [27] (for biological data), or quantile normalized (for synthetic data),  $\log_2$ -transformed, centered by subtracting the mean of each column and dividing by the standard deviation, and whitened, so that the columns are uncorrelated, by projecting it onto its  $S$  principal components (thus resulting in no loss of information).

### Synthetic data

The synthetic data consists of 20000 genes measured in 30 samples. The 30 samples are divided into six groups of five samples each (e.g. six individuals measured at five time points each). Six transcriptional programs determine the gene expressions. The effect of each transcriptional program is indicated in Figure 1A, which shows the true expressions. Besides individual groups of genes affected by only one program, two groups of genes are affected by two programs each. In Figure 1B is shown the standardized synthetic data with noise added as

$$y = \beta + x \exp(\eta) + \epsilon \quad (1)$$

where  $x$  is the true expression,  $y$  the expression with noise added,  $\epsilon$  the additive background noise, distributed as  $\mathcal{N}(0, \sigma_\epsilon^2)$ ,  $\eta$  the multiplicative noise, distributed as  $\mathcal{N}(0, \sigma_\eta^2)$ , and  $\beta$  the background [28]. The data presented in Calvano et al. [29] which have an experimental design reasonably similar to the setup of the synthetic data, are used to estimate the values of these parameters, giving  $\beta = -0.77$ ,  $\sigma_\epsilon^2 = 0.14$ , and  $\sigma_\eta^2 = 0.14$ .

### Feature selection

The standard deviation  $\sigma_t$  is calculated for the absolute values of the loads  $|C_{gt}|$  in each independent component. The genes with absolute loads greater than  $\alpha\sigma_t$  in component  $t$  are selected to represent that component [7, 12, 15]. For the biological examples we use the stringent choice of  $\alpha = 4$  [12].

### Scores and weights

Each independent component  $\vec{C}_t$  is given a score  $\lambda_{it}$  as defined below. To make different scores comparable, each score is individually scaled to sum to unity across ICs, i.e.  $\sum_t \lambda_{it} = 1$ .

**Informative loads**, score  $\lambda_{1t}$ . The non-normality of the distribution of loads  $C_{gt}$  in each component  $\vec{C}_t$ , calculated as the negentropy of the loads. The negentropy is approximated as suggested in the fastICA algorithm.

**Informative mixes**, score  $\lambda_{2t}$ . The non-normality of the distribution of mixes  $M_{ts}$  for each component, calculated as the negentropy of the mixes. As for the informative loads, the negentropy is approximated as suggested in the fastICA algorithm.

**Overall highest loads**, score  $\lambda_{3t}$ . The absolute value of either the 0.5% or the 99.5% quantile of the loads, depending on which is greatest.

**Component stability**, score  $\lambda_{4t}$ . Depending on the initial random guess for the un-mixing matrix (the inverse of  $\mathbf{M}$ ), the convergence of ICs that do not represent data structures, may be to local, instead of global, non-normality maxima. A component identifying structures in the data is expected to always

be extracted by ICA, that is, to always converge to the same non-normality maximum. To quantify this ICA is run 100 times. In each run, Spearman’s correlation coefficients are calculated between each of the original ICs  $\vec{C}_t$ , and the ICs in run  $r$ ,  $\vec{C}_{tr}$ . The highest absolute correlation coefficient  $\rho_{tr}$  calculated for each  $\vec{C}_t$  in run  $r$  is stored. The stored absolute correlation coefficients are averaged component-wise over all runs,  $\rho_t = 1/100 \sum_r \rho_{tr}$ .

**Sample separability**, score  $\lambda_{5t}$ . When samples can be separated into  $n$  classes (e.g. treatment groups), the proportion of the variability accounted for by this separation, denoted the coefficient of determination  $R^2$ , is calculated by fitting a one-way ANOVA to the mixes for each component.

**Biological concordance**, score  $\lambda_{6t}$ . When genes or features are annotated by GO-terms or mapped to KEGG-pathways, the genes selected as described under feature selection are divided into two groups, depending on each gene’s load being positive or negative. For each group, significantly overrepresented (family-wise error rate  $< 0.001$  using the Bonferroni correction) GO-terms and KEGG-pathways are found by hypergeometric testing. The fraction of selected genes annotated with significant terms or pathways is used as a measure of biological concordance. When significantly overrepresented GO-terms and KEGG-pathways are found for the negative loads only, by convention, all loads of that component as well as the corresponding row in the mixing matrix  $\mathbf{M}$ , are multiplied by  $-1$ .

**Weighting**. Using the standard deviation  $\sigma_i$  of each score across ICs to define weights, each weight is calculated as  $w_i = \sigma_i / \sum_i \sigma_i$ . For weights  $w_1, w_2, \dots, w_6$ , where  $\sum_i w_i = 1$ , the overall score of each independent component is calculated straightforwardly as  $\lambda_t = \sum_i w_i \lambda_{it}$  and ranked accordingly.

## Results

To show the performance of weighted scoring for exploring as well as ranking ICs, synthetic data where the transcriptional programs are known by design, as well as three instructive examples with biological data, are presented below.

### Synthetic data

For the synthetic data, the biological concordance score cannot be calculated. The equally weighted ( $w_i = 1/5$ ) sum of the remaining five scores defined in the methods section, correctly assign highest rank to the six ICs that represent the six transcriptional programs (Figure 2A). The informative loads score and the sample separability score contribute the most to the differentiation between the six top-ranking ICs and the rest. When inspecting the distribution of each score  $\lambda_{it}$  across ICs (the compound bars in Figure 2A), it is evident that the higher the spread of score values, the better that score differentiates between ICs that represent transcriptional programs and ICs that do not. By assigning weights according to the standard deviation of each score, the six top ranking ICs will consequently stand out even more clearly. This weighting by standard deviation is used for the three examples with biological data presented below.

Using  $\alpha = 4$  for feature selection identifies all 668 genes known to be affected and 212 non-affected in addition (giving a sensitivity of 100% but a positive predictive value (PPV) of only 75%). For  $\alpha = 5$ , which is the value used in Figure 2, 664 out of the 668 affected genes are identified as well as 9 non-affected (98% sensitivity and 97% PPV). Finally,  $\alpha = 6$  identifies 639 of the 668 affected genes and only one non-affected in addition (95% sensitivity and very close to 100% PPV).

To explore the performance of the fastICA algorithm, as well as the ranking of ICs, at different noise levels, the background,  $\beta$ , the variance of the additive noise,  $\sigma_\epsilon^2$ , and the variance of the multiplicative noise,  $\sigma_\eta^2$  (refer to Equation 1), are doubled (and also quadrupled in the case of  $\sigma_\eta^2$ ) one at a time. All six transcriptional programs are identified as separate ICs and ranked at the top when doubling  $\beta$  and  $\sigma_\epsilon^2$  (data not shown). When doubling  $\sigma_\eta^2$ , however, the fastICA algorithm does not identify the transcriptional program labelled with purple in Figure 1 (data not shown). The expression profile of this

transcriptional program is not varying significantly with time and is at a high level in only three samples out of 30. When increasing the value of  $\sigma_\eta^2$  the noise is increased especially on genes with high expression levels. This transcriptional program is therefore especially sensitive to this kind of noise degradation. The five other transcriptional programs are still ranked at the top. When quadrupling  $\sigma_\eta^2$ , none of the transcriptional programs are identified by the fastICA algorithm (data not shown).

### Example 1: Beta cell function

Type 1 diabetes results in a permanent destruction of insulin producing beta cells of the pancreas. To investigate beta cell function and development, a rat insulinoma cell line (INS-1 $\alpha\beta$ ) transfected with the mouse transcription factor pancreatic duodenum homeobox (*pdx1*) gene, was profiled [30]. Cells were selectively treated with doxycycline. The resulting overexpression of both mouse and rat *pdx1* genes activates a negative feedback loop affecting rat *pdx1* expression only. This depletes PDX1 protein below effective levels. Interleukin 1 $\beta$  (IL-1 $\beta$ ), an important mediator of the inflammatory response, was also selectively administered at various lengths of exposure. Since the samples are taken from treatment-synchronized, supposedly genetically identical, cell lines, the variability between biological replicates is low. The ranking of ICs with  $\alpha = 4$  for feature selection, and weighting based on standard deviations, is shown in Figure 3. The six top ranking ICs all show high sample separability scores and are readily interpreted in the context of presence/absence of PDX1 as well as time of IL-1 $\beta$  exposure. ICs 1 and 2 score especially high for informative loads and overall highest loads, signifying a clear presence of the significant genes in these ICs.

Component 1 correlate with time of IL-1 $\beta$  exposure, and genes belonging to the adipocytokine signaling pathway, and immune processes in general, are overrepresented. Component 2 correlate with PDX1, and the clear positive or negative presence of the component result in a high informative mixes score. Although genes annotated with system development are overrepresented in component 2, it is only a small fraction, 4%, of all the genes selected to represent that component. This might reflect that the functional role of most of these genes in relation to PDX1 and beta cell development has first been uncovered with this study and are therefore not part of the annotation yet. ICs 3 to 6 all display high biological concordance scores (between 25% and 77%), related to cell cycle processes (3), antigen processing and presentation (4), cell communication (5), and signal transduction (6), among others. Component 7 has a relatively high informative loads score, a high biological concordance score (cellular localization and intracellular transport, among others, being overrepresented), but a low sample separability score. This component probably represents a transcriptional program that is not clearly related to PDX1 and IL-1 $\beta$ , but to some other aspect of the study such as, e.g., phase in the cell cycle.

Figure 3C shows the loads of all 2302 genes present in one or more of the seven top-ranking ICs. Each component consists of a large group of genes unique to that component (especially component 7) as well as smaller groups of genes overlapping with one or more of the other ICs.

### Example 2: Inflammation

Inflammation is a hallmark of many diseases. To elucidate the transcriptional programs in blood leukocytes activated in response to systemic inflammation, four healthy human subjects were profiled at 0, 2, 4, 6, 9 and 24h after administration of bacterial endotoxin (eliciting an inflammatory response), and another four were profiled at the same times, having been administered sodium chloride, as control [29]. Leukocytes were recovered from blood samples by centrifugation. The within-group sample variability is higher than for the beta cell data, since individual people are more different than individual cell lines, and because of cell heterogeneity (lymphocytes, monocytes, and granulocytes present in varying amounts).

The ranking of ICs ( $\alpha = 4$ ) is shown in Figure 4. ICs 1, 2, and 4 show the same relative profile in Figure 4A, being positive in the controls and at 0h and 24h in the inflammation subjects, and negative between 2h and 9h. Figure 4B reveals that the mixes of each component are at different baselines.

Component 1 is always present, component 2 is absent between 2h and 9h in inflammation subjects, and positive otherwise, and component 4 is negative between 2h and 9h in inflammation subjects, and absent otherwise. Component 1 shows significant overrepresentation of genes related to translation of proteins in ribosomes, among others, component 2 to immune processes and translation also, and component 4 to cytolysis. Component 3 is positive at 2h in inflammation subjects, and otherwise negative (Figure 4B), defense responses and MHC class II receptor activity, among others, are overrepresented. Component 5 is positive at 2h, 4h, and 6h, with a peak at 4h, and otherwise absent. The biological concordance score is zero.

A lot of ICs besides the top five also show differences between inflammation and control patients. Moving down the list, the inter-group variability becomes progressively larger though, resulting in lower sample separability scores. Only 9 ICs have stability scores below 80% (the six top-ranking ICs have stability scores at or above 90%), reflecting that most if not all extracted ICs represent real structures in the data.

Figure 4C shows the loads of all 1405 genes present in one or more of the six top-ranking ICs. As for the beta cell data, each component consists of a large group of genes unique to that IC as well as smaller groups of genes overlapping with one or more of the other ICs.

### Example 3: Breast cancer

A clinically and prognostically important part of evaluating a breast cancer is to determine its histologic grade. A low grade cancer is well-differentiated, meaning that the tumor grows relatively slowly and have a well-defined shape. Oppositely, a high grade cancer is poorly differentiated, with breast cells that do not appear normal and tend to grow and spread more aggressively. Identification of the transcriptional programs that are active in breast cancers of various aggressiveness could be used to grade tumors more accurately. To this end, invasive breast carcinomas from 189 patients were profiled [31].

The ranking of ICs ( $\alpha = 4$ ) is shown in Figure 5, where 104 patients (the validation set in [31], but used as the training set here), and the top 16000 most varying transcript profiles, are included in the analysis. Component 1 is clearly negative for grades 1, and absent for grades 2 and 3. Immune system processes, antigen processing and presentation, as well as extracellular matrix structural constituent activity, are overrepresented. Component 2 is clearly most positive for grade 3, and closer to zero for grades 1 and 2, and structural constituents of ribosome are overrepresented. Figure 5C shows the loads of the 492 genes present in one or both of the two top-ranking ICs. Most genes are only present in one component.

The ability of these 492 genes to correctly classify the 104 samples, using  $k$ -nearest neighbors ( $knn$ , majority vote among 3 nearest neighbors), and leave-one-out cross-validation ( $loocv$ ), gives a classification accuracy of 96% for the grade 1 and 3 samples. When classifying the remaining 59 samples (the training set in [31] but used as the test set here) using  $knn$  and  $loocv$ , the classification accuracy is 93%. Out of the 492 genes, 377 are also measured in another breast cancer study that included 295 patients [32]. These 377 genes allow a  $knn$  classification accuracy of 70% when using  $loocv$ . These accuracies compare favorably to a similar recent study on the same data that used protein-protein interaction data to assist in the selection of features [33]. The method presented here for identifying ICs and genes might therefore be interesting from a classification point-of-view.

## Discussion

Any expression matrix decomposition requires subsequent evaluation of the components identified. The weighted scoring presented here provides a systematic framework for performing such an evaluation for ICs identified by ICA.

The example with synthetic data demonstrates that ICA is able to identify and separate the underlying transcriptional programs, even at high levels of noise, and that the weighted scoring assigns high ranks

to ICs representing these programs. The three biological examples show that different scores highlight different structures in the data. By weighting each score before summing the overall component score, particular properties of the data can consequently be emphasized. A heuristic weighting may for example focus on class comparison by having a high weight for sample separability. In a more unbiased manner, a weighting based on the standard deviations of each score across samples, provides a means for judging which scores differentiate the most between ICs.

The sample separability score assigns high scores to ICs that separate samples according to known classes. The score is seen to clearly differentiate ICs in all data, but less markedly so in the inflammation data, where most of the identified ICs correlate with inflammation to some degree.

The informative mixes score is intended as an unsupervised sample separability score. A component that identifies a transcriptional program will exhibit mixes that are defined by some regulatory control system. Such mixes will tend to be more ordered, having a lower entropy, than mixes for ICs not under such control. The degree of non-normality of the distribution of mixes can be used as a measure of this order [34]. The informative mixes score, as defined this way, works best for biological data. For the beta cell data, component 2, correlating with PDX1, and being essentially binary, have a high informative mixes score. In the same data, component 7, which has a much lower sample separability score than the top six ICs, has a reasonably high informative mixes score, hinting that this component represents a transcriptional program that is under control of something independent of PDX1 or IL-1 $\beta$ . For the inflammation data, the informative mixes also contributes to the differentiation between ICs, being generally highest for those ICs that are most ordered.

The informative loads score is based on the same idea as the informative mixes. The sharper the distinction between genes with absolute high loads and genes with absolute low loads, the higher the score. This tends to put emphasis on ICs representing smaller groups of genes and is seen to differentiate ICs clearly in both the synthetic as well as the beta cell data. Accompanying this, the higher the absolute loads of the group of genes primarily representing the IC, the more clearly the IC contributes to the expression of these genes. ICs 1 and 2 in the beta cell data, and IC 10 in the inflammation data, all display high absolute loads scores.

The stability score depends on the noise characteristics of the data as well as the number of active transcriptional programs. In the synthetic data, the score is clearly highest for ICs representing transcriptional programs. For the biological data, this score is not given much weight, reflecting that most ICs identify stable structures in the data, these being either transcriptional programs or biological noise.

The biological concordance score reflects the fraction of significant genes in an IC that are annotated with significantly overrepresented GO terms or KEGG pathways. For the beta cell data, only 15 out of the 54 identified ICs have a non-zero biological concordance score, 8 of them being among the top ten ranked ICs. Hence this score correlates with high values of the other scores. For the inflammation data, the same trend is seen. Here however, about 2/3 of all ICs have non-zero scores, reflecting that there are many more active transcriptional programs compared to the beta cell data. This is probably a general difference between samples taken from cell lines and tissues. In the breast cancer data, about half of the ICs have positive biological concordance scores, not obviously being correlated to the values of the other scores. This again probably reflects characteristics of the patients in the study other than tumor grade and estrogen receptor status..

## Conclusion

The sample separability and biological concordance scores provides readily interpretable information with respect to the experimental setup. The informative loads and overall highest loads provide additional insight into the manner in which the governing transcriptional programs influence genes, showing high scores for smaller, highly influenced, groups of genes. The informative mixes can in some cases help identify ICs representing transcriptional programs governed by factors other than those representing the experimental design. Finally, component stability provides a means for assessing the robustness of the

identified ICs. Together, the six scores presented allows comprehensive identification of relevant ICs in gene expression microarray data.

## Acknowledgments

We thank O. Winther, M. Aalund, K. Mortensen, J. V. Johansen, J. Bork-Jensen, C. Bang-Berthelsen, H. Flyvbjerg, and F. Pociot for useful discussions and comments.

## References

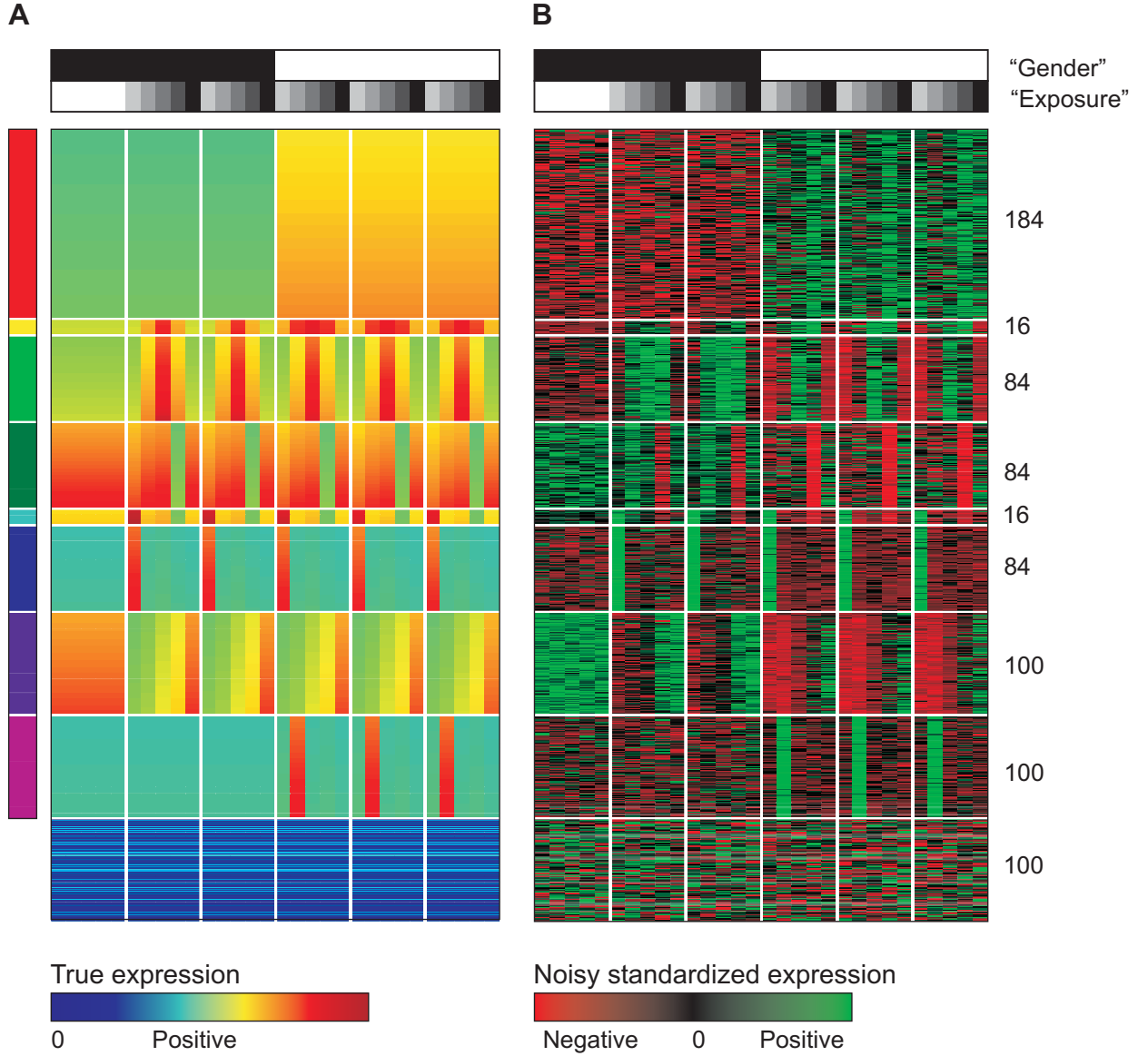
1. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, et al. (2000) Molecular portraits of human breast tumours. *Nature* 406: 747–752.
2. Arbeitman MN, Furlong EEM, Imam F, Johnson E, Null BH, et al. (2002) Gene expression during the life cycle of *drosophila melanogaster*. *Science* 297: 2270–2275.
3. Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, et al. (2005) A gene expression map of *arabidopsis thaliana* development. *Nat Genet* 37: 501–506.
4. Comon P (1994) Independent component analysis, a new concept? *Signal Processing* 36: 287–314.
5. Martoglio AM, Miskin JW, Smith SK, MacKay DJC (2002) A decomposition model to track gene expression signatures: preview on observer-independent classification of ovarian cancer. *Bioinformatics* 18: 1617–1624.
6. Chiappetta P, Roubaud MC, Torr sani B (2004) Blind source separation and the analysis of microarray data. *J Comput Biol* 11: 1090–1109.
7. Saidi SA, Holland CM, Kreil DP, MacKay DJC, Charnock-Jones DS, et al. (2004) Independent component analysis of microarray data in the study of endometrial cancer. *Oncogene* 23: 6677–6683.
8. Capobianco E (2005) Mining time-dependent gene features. *J Bioinform Comput Biol* 3: 1191–1205.
9. Lutter D, Ugocsai P, Grandl M, Orso E, Theis F, et al. (2008) Analyzing m-csf dependent monocyte/macrophage differentiation: expression modes and meta-modes derived from an independent component analysis. *BMC Bioinformatics* 9: 100.
10. Bang-Berthelsen CH, Pedersen L, Fl yel T, Hagedorn PH, Gylvin T, et al. (2011) Independent component and pathway-based analysis of miRNA-regulated gene expression in a model of type 1 diabetes. *BMC Genomics* 12: 97.
11. Wexler EM, Rosen E, Lu D, Osborn GE, Martin E, et al. (2011) Genome-wide analysis of a *wnt1*-regulated transcriptional network implicates neurodegenerative pathways. *Science signaling* 4: ra65.
12. Liebermeister W (2002) Linear modes of gene expression determined by independent component analysis. *Bioinformatics* 18: 51–60.
13. Lee SI, Batzoglou S (2003) Application of independent component analysis to microarrays. *Genome Biol* 4: R76.



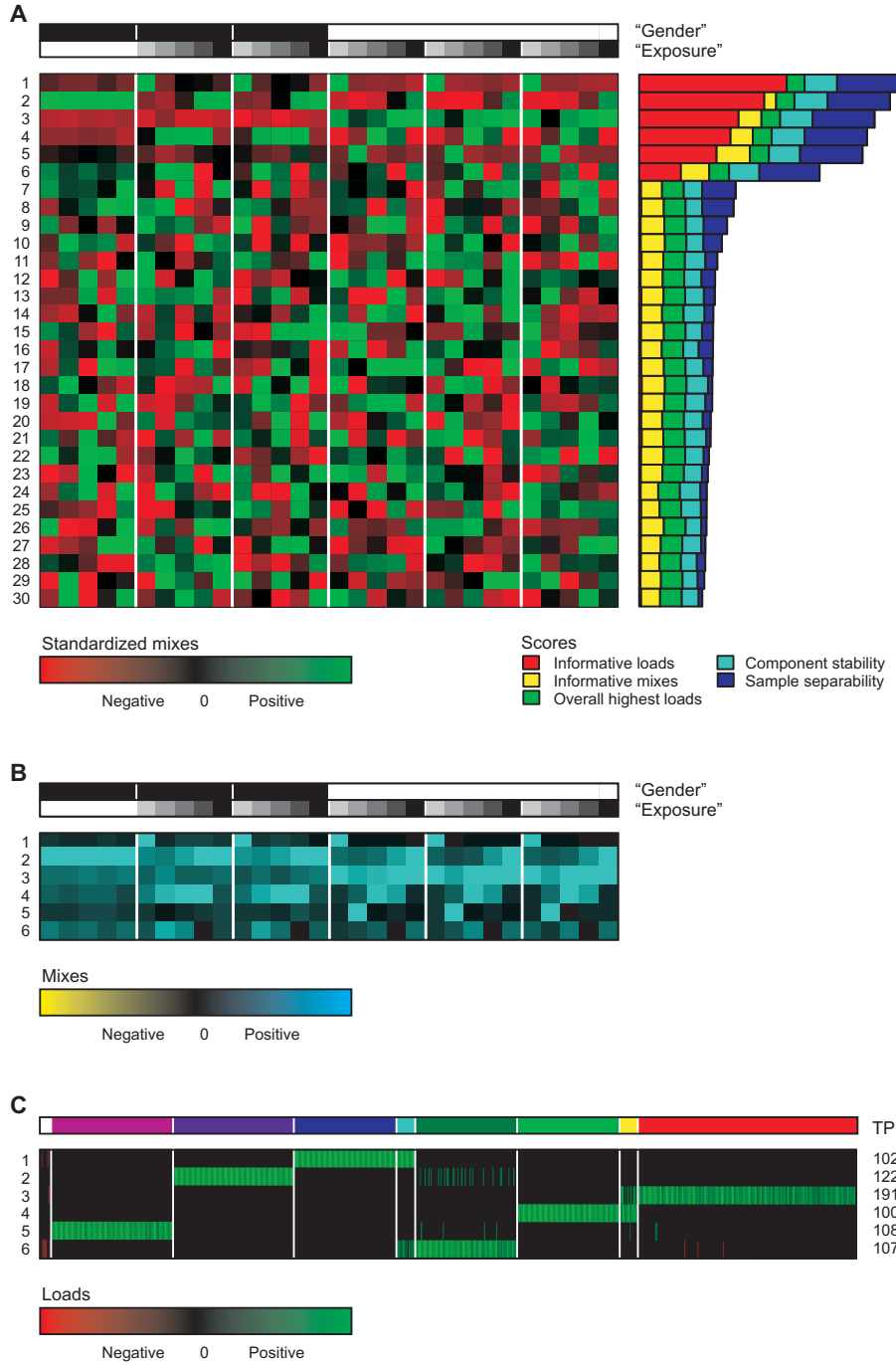
14. Carpentier AS, Riva A, Tisseur P, Didier G, Hénaut A (2004) The operons, a criterion to compare the reliability of transcriptome analysis tools: Ica is more reliable than anova, pls and pca. *Comput Biol Chem* 28: 3–10.
15. Teschendorff AE, Journée M, Absil PA, Sepulchre R, Caldas C (2007) Elucidating the altered transcriptional programs in breast cancer using independent component analysis. *PLoS Comput Biol* 3: e161.
16. Kong W, Vanderburg CR, Gunshin H, Rogers JT, Huang X (2008) A review of independent component analysis application to microarray gene expression data. *BioTechniques* 45: 501–520.
17. Frigyesi A, Veerla S, Lindgren D, Höglund M (2006) Independent component analysis reveals new and biologically significant structures in micro array data. *BMC Bioinformatics* 7: 290.
18. Kong W, Mou X, Hu X (2011) Exploring matrix factorization techniques for significant genes identification of Alzheimer’s disease microarray gene expression data. *BMC Bioinformatics* 12: S7.
19. Zhang XW, Yap YL, Wei D, Chen F, Danchin A (2005) Molecular diagnosis of human cancer type by gene expression profiles and independent component analysis. *Eur J Hum Genet* 13: 1303–1311.
20. Huang DS, Zheng CH (2006) Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. *Bioinformatics* 22: 1855–1862.
21. Ohlsson Teague EMC, Van der Hoek KH, Van der Hoek MB, Perry N, Wagaarachchi P, et al. (2009) MicroRNA-regulated pathways associated with endometriosis. *Molecular endocrinology* (Baltimore, Md) 23: 265–275.
22. Chang WC (1983) On using principal components before separating a mixture of two multivariate normal distributions. *Appl Statist* 32: 267–275.
23. Yeung KY, Ruzzo WL (2001) Principal component analysis for clustering gene expression data. *Bioinformatics* 17: 763–774.
24. Winther O, Petersen KB (2007) Flexible and efficient implementations of bayesian independent component analysis. *Neurocomputing* 71: 221–233.
25. Liu KH, Li B, Wu QQ, Zhang J, Du JX, et al. (2009) Microarray data classification based on ensemble independent component selection. *Computers in biology and medicine* 39: 953–960.
26. Hyvärinen A, Oja E (2000) Independent component analysis: algorithms and applications. *Neural Netw* 13: 411–430.
27. Wu Z, A IR, Gentleman R, Murillo FM, Spencer F (2004) A model based background adjustment for oligonucleotide expression arrays. *Johns Hopkins University, Dept of Biostatistics Working Papers Paper* 1.
28. Rocke DM, Durbin B (2001) A model for measurement error for gene expression arrays. *J Comput Biol* 8: 557–569.
29. Calvano SE, Xiao W, Richards DR, Felciano RM, Baker HV, et al. (2005) A network-based analysis of systemic inflammation in humans. *Nature* 437: 1032–1037.
30. Bergholdt R, Karlsen AE, Hagedorn PH, Aalund M, Nielsen JH, et al. (2007) Transcriptional profiling of type 1 diabetes genes on chromosome 21 in a rat beta-cell line and human pancreatic islets. *Genes Immun* 8: 232–238.

31. Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, et al. (2006) Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst* 98: 262–272.
32. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AAM, et al. (2002) A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* 347: 1999–2009.
33. Chuang HY, Lee E, Liu YT, Lee D, Ideker T (2007) Network-based classification of breast cancer metastasis. *Mol Syst Biol* 3: 140.
34. Hyvärinen A, Karhunen J, Oja E (2001) Independent component analysis. Adaptive and learning systems for signal processing, communications, and control. J. Wiley.

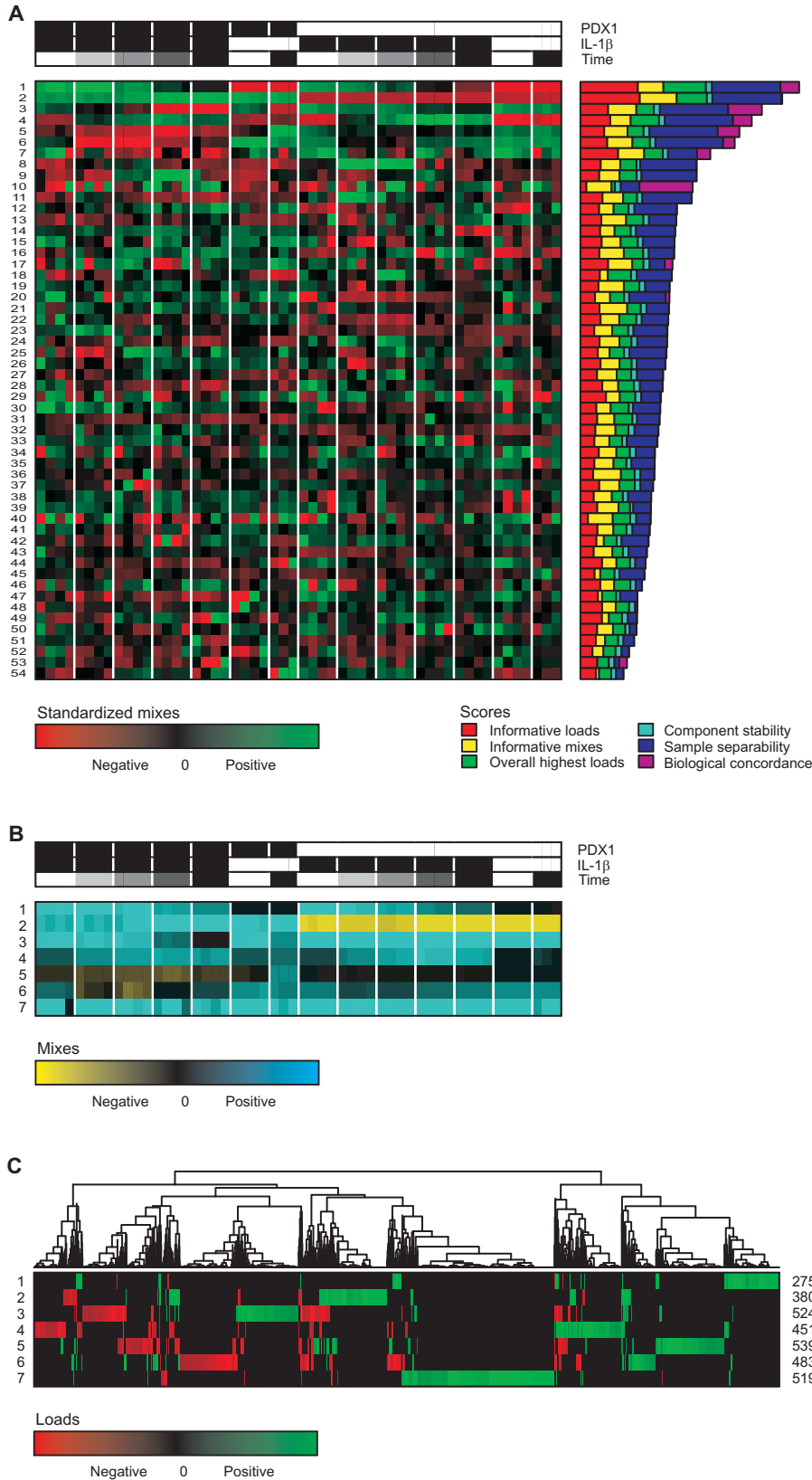




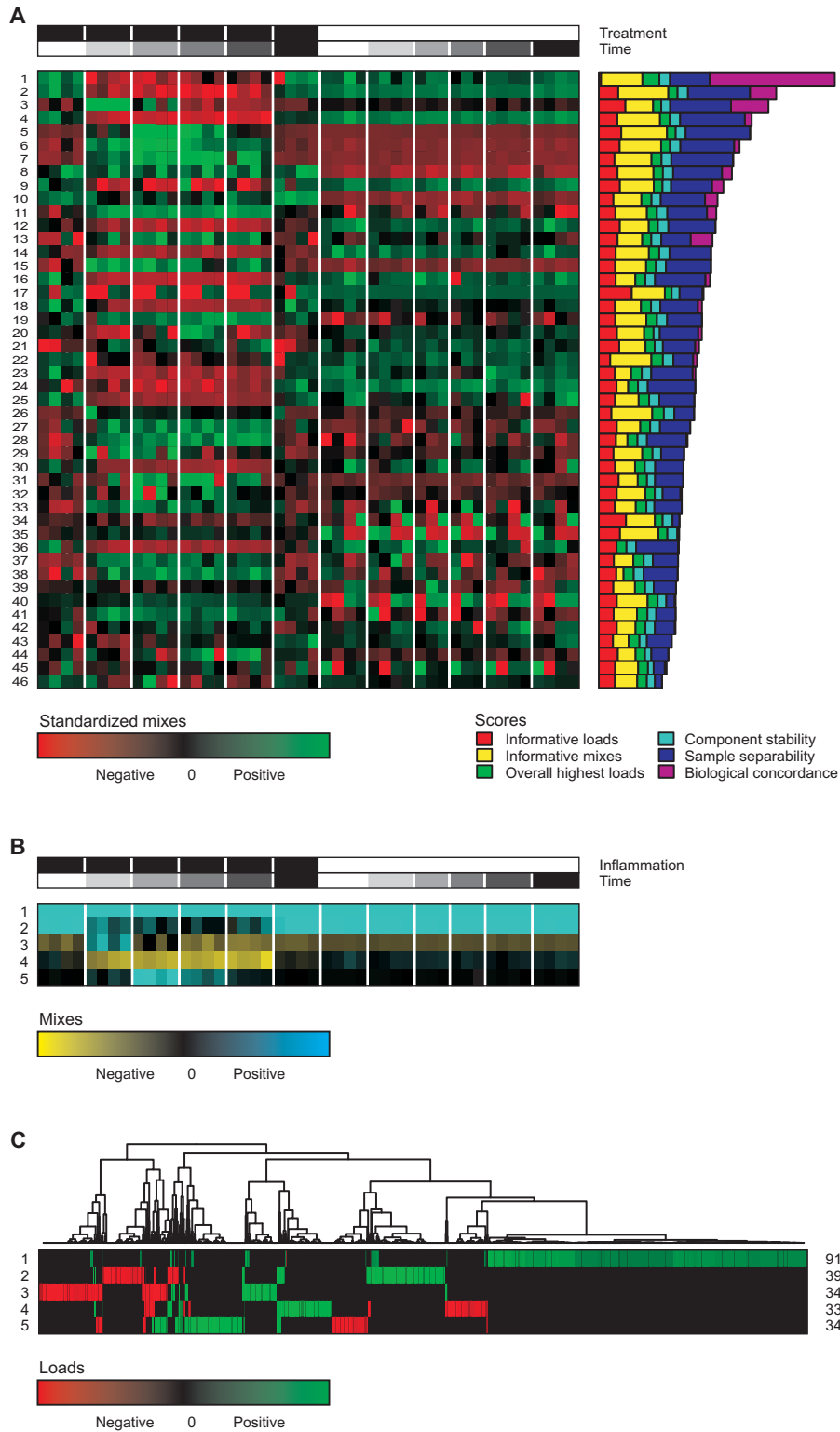
**Figure 1. Expression matrix for synthetic data.** Genes in rows and samples in columns. The 668 genes affected by one or more transcriptional programs are shown, as well as a selection of 100 unaffected genes. The horizontal colorpanels at the top separate samples according to gender (black or white) and effect of exposure (shades of grey or white). The six individual transcriptional programs are shown in the vertical colorpanel on the left as red, light green, green, blue, purple, and magenta. The combinations of transcriptional programs are shown as yellow and cyan. The number of genes in each program or combination of programs are shown to the right. Within each transcriptional program, the overall expression increases from top to bottom. **A.** True expressions. Red is high expression, blue is no or low expression. **B.** Noisy, standardized expressions. Black is the row-mean expression, red/green is the expression below/above row-mean in units of the standard deviation.



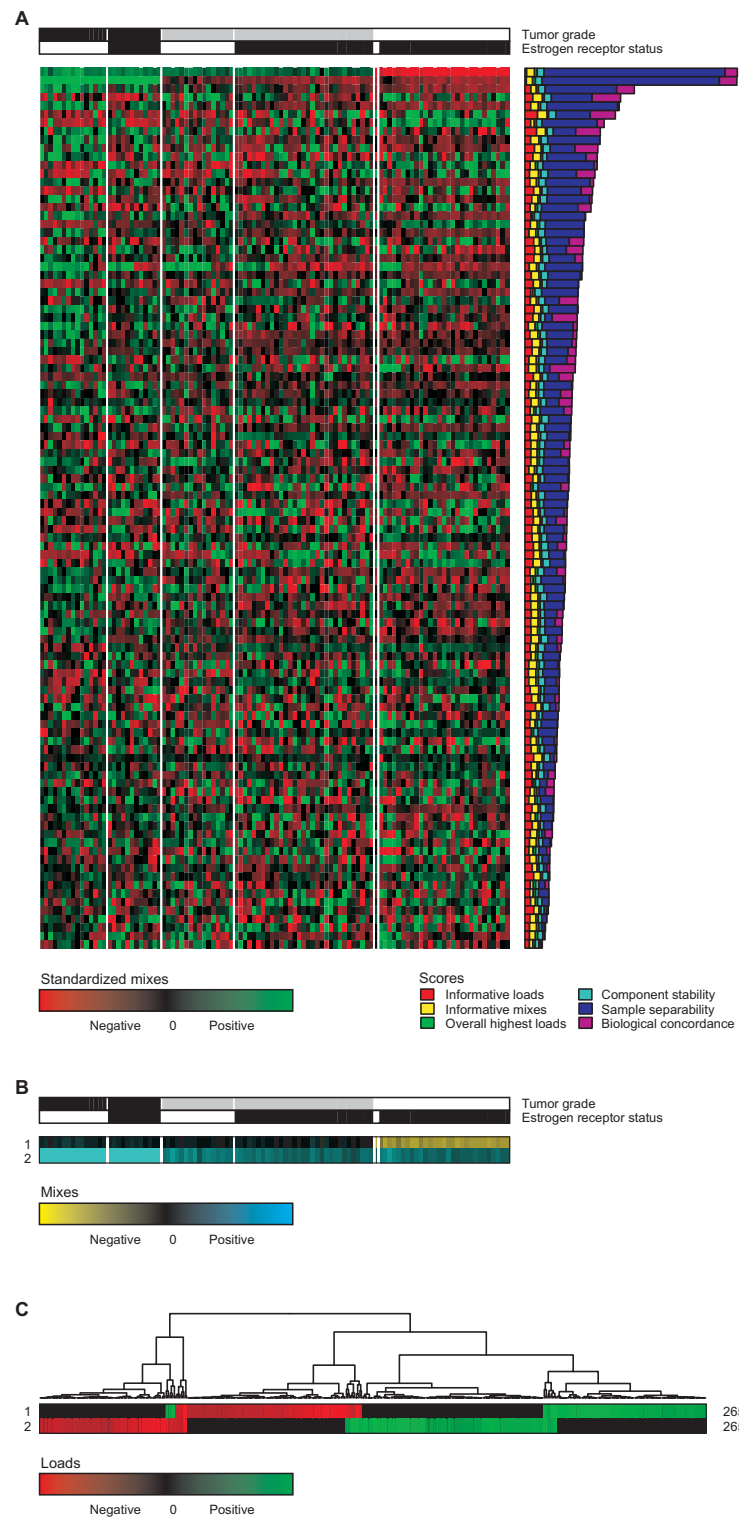
**Figure 2. Ranking of independent components in synthetic data.** **A.** Standardized mixing matrix highlighting changes across samples. Samples in columns, ICs in rows. The scores are weighted equally and are shown as compound bars to the right. On top of the mixing matrix are shown the sample types, as in Figure 1. **B.** Mixing matrix of the six top-ranking ICs. Actual mixes are shown to highlight the contribution in each sample. **C.** Component matrix. Genes in columns, ICs one to six in rows. Positive loads are green, negative red, and non-significant loads are black. The correspondence between ICs and transcriptional programs are shown in the horizontal colorpanel at the top, same colorcode as in Figure 1. The number of genes in each component are shown to the right.



**Figure 3. Ranking of independent components in beta cell data.** Same layout as in Figure 2. **A** and **B**. For PDX1 and IL-1 $\beta$ , black denotes presence and white absence. For time, white over grey to black denotes, 2, 4, 6, 12, and 24h, respectively. **C**. The 2302 genes with significant loads in one or more of the top six ICs are clustered hierarchically (Euclidean distance and average linkage).



**Figure 4. Ranking of independent components in inflammation data.** Same layout as in Figure 3. **A** and **B**. For treatment, black denotes bacterial endotoxin (inflammation) and white denotes sodium chloride (control). For time, white over grey to black denotes, 0, 2, 4, 6, 9, and 24h, respectively.



**Figure 5. Ranking of independent components in breast cancer data.** Same layout as in Figure 3. A and B. For tumor grade, black denotes grade 3, grey grade 2, and white grade 1. For estrogen receptor status, white denotes positive and black negative.

## APPENDIX

G

OVC paper

This paper has been submitted to PLoS Computational Biology.

# Independent component analysis reveals possible microRNA regulated pathways in ovarian cancer

Lykke Pedersen<sup>1,\*</sup>, Weimin Xiao<sup>2</sup>, Mogens H Jensen<sup>1</sup>, Preethi Gunaratne<sup>2</sup>,

**1 Center for Models of Life, Niels Bohr Institute, University of Copenhagen, Denmark**

**2 Department of Biology and Biochemistry, University of Houston, Texas, USA**

**\* E-mail: Corresponding author lykkep@nbi.dk**

## Abstract

Reliable microRNA (miRNA) target prediction methods, which can accurately reveal miRNA regulated pathways are important for a better understanding of gene networks regulated by miRNAs. Particularly, this is of interest for miRNA based diagnosis and the development of miRNA based treatment of various cancers.

Here we study miRNA and mRNA expression profiles for ovarian cancer samples from the cancer genome atlas (TCGA) project published by TCGA Research Network in Nature 2011. The expression profiles are analyzed by independent component analysis (ICA) followed by pathway enrichment and miRNA target analyses. The method described here links miRNA expressions to the deregulation of pathways in ovarian cancer and is applied to a test set of mRNA expressions and the findings are then validated by applying the method to a second set of mRNA expression profiles. With these methods, we retrieved miRNA and pathway relations already proposed in the literature as well as new possible relations.

We found that Wnt, mTOR and p53 pathways are highly enriched in both the test and validation data sets. These three pathways are already known to be deregulated in ovarian cancer. In addition we have linked these pathways to the differential expression of miR-29c, miR-100 and miR374a,b, respectively. While miR-100 has been previously linked to ovarian cancer this is the first time that miR-29a and miR-374a,b, which have been linked to other cancers, are linked to ovarian cancer. Furthermore, we identified components, which might be under miRNA regulations. This adds to the general belief that independent components map to biological pathways.

## Author Summary

Ovarian cancer is the fifth leading cause of death among females in the US and is usually diagnosed in an advanced stage [1,2]. Therefore an early diagnose is of interest to increase the chance of survival. One tool for diagnosis could be to evaluate the genetic signature that each of us inherits. There exist a class of genes called microRNAs, which were discovered in 1993 [3] and since then has been found to regulate over 60% of genes in the human genome [4]. They generally act to block the translation of mRNAs into proteins by binding to the 3-untranslated regions of protein coding gene transcripts and then they either cleave the transcript or obstruct translation [5]. miRNAs are commonly deregulated in many diseases, including cancer, and will consequently result in the mis-expression of their targets, which contribute to the development of diseases. The ability to correlate the regulation of microRNAs with the deregulation of genes causing ovarian cancer is the goal of this paper. This is achieved by analyzing genetic signatures, including miRNA and mRNA, of high-grade serous ovarian tumors by use of independent component analysis. By this method we can ultimately unravel the huge network that underlies the development of ovarian cancer.

## Introduction

Early diagnosis and targeted treatment are important for all types of cancer. Ovarian cancer (OVC) is especially important because it is the fifth leading cause of death among females in the US and it is usually diagnosed in a late stage [1, 2]. In the cancer genome atlas (TCGA) project, the expression profiles of both messenger RNA (mRNA) and microRNA (miRNA) from normal ovary and high-grade serous ovarian tumors are obtained, which can provide biosignatures for diagnosis and treatment targets by using bioinformatics analysis [6].

The miRNAs were discovered in 1993 by Lee et al. [3] and are small non-coding RNAs consisting of  $\sim 22$  nucleotides in their mature form. They inhibit the translation of its target mRNAs by binding to their 3'-untranslated regions (3UTRs) and control the expression of 60% percent of the human genes [4, 5]. In recent years, many publications demonstrate the important roles of miRNAs in the development of different diseases, e.g., cancers [7], diabetes [8–10] cardiovascular diseases [11, 12], infectious diseases [13]. In OVC, mir-9 has been shown to regulate the NF- $\kappa$ B pathway [14] and miR-31 regulates the p53 pathway [15]. Both pathways are deregulated in OVC.

The actual inhibition of a miRNA target commonly occurs through partial base pairing between the miRNA and the target mRNA. The sequences that undergo base pairing to form a miRNA:mRNA duplex are constrained to the 5'-seed region of the miRNA extending from nucleotides 2-8 and typically localized to the sequences in the 3'UTRs of the mRNA target. A minimum of six consecutive nucleotides must pair in order to form a productive miRNA:mRNA duplex, a seed match, that culminates in translational inhibition.

Many miRNA target prediction methods have been published during the last couple of years. The vast majority of them are based on finding seed matches between the miRNA and its targets. Others are (additionally) based on gene expression profiling. Table 1 lists features commonly included in miRNA target prediction methods. The list is curated from the reviews of Refs. [5, 16, 17]. Here we will use seed match finding together with expression profiling to investigate the role of miRNAs in controlling biological pathways related to OVC.

The most common use of expression profiles in miRNA target prediction methods is the search for negative correlations between the miRNA and its targets. Since due to the inhibitory effect of miRNAs. However, the anti-correlation between the miRNA and its target mRNA expression is not always significant due to many factors, e.g., miRNA regulation, transcription factor binding and site accessibility, et.al. [18]. Here we will apply independent component analysis to circumvent this and in order to find biological pathways potentially under the influence of miRNAs.

ICA is a linear decomposition method [19] widely used for the analysis of microarray data [20, 21], since it has proven been superior to both PCA and clustering analysis [20, 22, 23]. ICA decomposes the expression matrix into independent components (ICs), which often can be mapped to biological processes or pathways [21, 24, 25]. The expression matrix ( $\mathbf{E}$ ) has columns corresponding to expression profiles of the samples  $s$  and rows corresponding to the profiled genes  $g$ .

The general mathematical formulation of ICA can be expressed as

$$\mathbf{E} = \mathbf{C}\mathbf{M} \quad \text{or} \quad E_{gs} = \sum_{k=1}^N C_{gk} M_{ks} , \quad (1)$$

where  $g \in \{1, \dots, G\}$  and  $s \in \{1, \dots, S\}$  with  $G$  and  $S$  being the total number of genes and samples, respectively.  $\mathbf{C}$  is the  $G \times N$  component matrix comprised of the  $N$  ICs in the columns.  $\mathbf{M}$  is the  $N \times S$  mixing matrix. The ICs estimated from  $\mathbf{E}$  will be weighted differentially across the samples according to the component weights ( $M_{ks}$ ) in Eq. 1. These weights measure the level by which the ICs contributes to the total expression profile of the samples. How much an IC contributes to the overall expression of



a gene is measured by the quantity  $C_{gk}$  in Eq. 1. These entries of  $\mathbf{C}$  are denoted the gene weights. A given gene  $g$  will therefore have  $N$  gene weights and an IC will have  $S$  component weights.

A schematic illustration of ICA in relation to miRNA regulation is shown in Fig. 1. In Figure 1 the expressions of miR1 and mRNA1 are differentially regulated in cancer versus control samples, and miR2 and mRNA2 are deregulated according to tumor grade. The component weights of IC 1 are negative in cancer samples and positive in normal samples (yellow vs. blue) and the component weights of IC2 depend on tumor grade (shades of blue). Thus, these component weights are said to correlate with the classification of the samples. They also (anti)-correlate with the expression patterns of mRNA1/miR1 and mRNA2/miR2, respectively. Therefore, the gene weights of mRNA1/miR1 lies at the extremes of IC1 and the gene weights of mRNA1/miR2 lies at the extremes of IC2. IC1 and IC2 are informative, because they correlate with the clinical classification scheme.

To measure the effect of a miRNA within an IC, enrichments of its targets are performed. For example, if genes with a seed match for miR1 are significantly enriched in the positive end of IC1, miR1 is stated as active in IC1. Because the targets are enriched oppositely to miR1, which lies in the negative end of IC1. Additionally, if there is an enrichment of genes annotated with a biological pathway in the same end as the miRNA targets, the miRNA can be assumed to regulate that biological pathway. The regulations of course have to be put in the context of the weighting of the IC. Regarding miR2, the regulations of targets and biological pathways will be dependent on the grade of tumor and for miR1 it will be dependent on the disease state of the sample. This is the principles behind the miRNA target prediction method proposed here and we call it **miICA**.

The use of ICA in the analysis of microarray data is not novel. However, the novelty of **miICA** is the combination of ICA applied to the expression matrices of miRNAs and mRNAs together with enrichment analysis of targets and biological pathways. Using ICA enables the investigation of miRNA and pathway regulations across several conditions. This is a very important feature and will be used extensively in this article to explore the changes of expressions according to tumor development and treatment. It is also a feature distinguishing **miICA** from other miRNA target prediction tools using expression data [26, 27]. To our knowledge **miICA** is the first miRNA target prediction method to use independent component analysis to couple miRNA and pathway regulations with each other.

We have previously applied a version of **miICA** on a model for type 1 diabetes in order to explore the effects of miRNA regulations on type 1 diabetes. The results from **miICA** were validated by comparison with published results on type 1 diabetes [9]. There the results from **miICA** were validated by comparison with published results on type 1 diabetes. In this article we will also validate **miICA** by comparing the findings of **miICA** with published observations on deregulated miRNAs and pathways in OVC. Furthermore, since the TCGA data contains duplicate gene expression profiles, we can validate **miICA** by directly comparing the results of **miICA** when it is applied on the test set of Agilent mRNAs and the validation set of Affymetrix mRNA. The high degree of congruence we observe between these results give us a high level of confidence in our method.

## Results and Discussion

In TCGA ovarian cancer database, both mRNA and miRNA expressions of 537 samples (514 primary cancers, 15 recurrent and 8 normal samples) are profiled using an Agilent microarray assay. The number of mRNAs profiled is 17,282, which is greater than the number of samples. Therefore  $\min(537, 17282) = 537$  ICs are estimated. The primary task is to search for an IC with component weights that correlates with clinical conditions. Furthermore, the IC must contain gene weights that are significantly high or low for genes targeted by miRNAs and/or biological pathways. To find such an IC the following two analyses are performed:

1. An ANOVA is applied to each row of  $\mathbf{M}$  in order to test if the component weights for each IC correlate with the clinical conditions.

2. Two enrichment analyses are applied on the gene weights within each IC.

- One enrichment analysis searches for overrepresentation of putative miRNA targets with either very high or very low gene weights in each IC. Here putative miRNA targets are mRNAs with a six nucleotide seed match in their 3'UTR.
- Another enrichment analysis is performed on each IC in order to find overrepresentation of biological pathways among genes with either very high or very low gene weights.

In the Methods section a detailed description of the enrichment analyses is given.

In the following we present examples of such ICs, where we link the clinical conditions with the enrichments of biological pathways and miRNA targets.

The expressions of all miRNAs and mRNAs mentioned within this paper are plotted in Supplementary Figure 1 and 2, respectively. The ICs are simply named according to their order in **C**, e.g. IC 1 is the first column of **C**.

### Targets of the p53 pathway and miR-374a are enriched in IC 339

p53 acts as a tumor suppressor and the p53 pathway is deregulated in 50% of all cancers [28]. The p53 pathway has been linked to high-grade serous OVC and about 96% of ovarian tumors from TCGA database exhibited mutations in p53 [6]. Therefore it is very important that miICA can demonstrate a p53 deregulation in OVC.

The most significant enrichment of the p53 pathway in any IC correlating with the samples being either normal or cancerous is found in IC 339. Genes annotated by the p53 pathway are significantly overrepresented among genes with negative gene weights in IC 339 ( $q = 1 \cdot 10^{-3}$ ). Figure 2(top) shows the cumulative distribution function (CDF) for the gene weights of genes with and without an annotation for the p53 pathway. It is clear that the CDF for the gene weights of p53 annotated genes are shifted to the left, which indicates that these gene weights are significantly lower. IC 339 is clearly weighted negatively in normal samples and is non-present in cancerous samples ( $q = 1.1 \cdot 10^{-6}$ , see Fig. 4G). Hence, since the p53 annotated genes have negative gene weights in IC 339 they are downregulated in cancer versus normal samples. That is their expression pattern mirrors that of the component weights for IC 339. This follows the general knowledge of p53 inhibiting cancer development. The next step is to link the p53 regulation mapped by IC 339 to some miRNA regulations.

In IC 339 miR-374a,b have targets enriched with negative gene weights ( $q = 9.4 \cdot 10^{-6}$ ). This enables a coupling of miR-374a,b to the regulation of genes in IC 339. To ensure that the coupling is biological plausible the actual expressions of miR-374a,b need to be considered. The TCGA data shows an upregulation of miR-374a in cancerous samples ( $p = 9.17 \cdot 10^{-7}$ , 2.1-fold change, see Supplementary Figure 1). The component weights of IC 339 are also upregulated in cancerous samples. Thus the expression of miR-374a correlates with the component weights of IC 339. This is not observed for miR-37b, because miR-37b is upregulated in cancer samples and downregulated in recurrent samples ( $p = 0.00256$ ). A link between miR-374a and the p53 pathway can be created because two conditions are satisfied: (1) the expression of miR-374a follows the pattern of component weights for IC 339, which consequently leads to an overrepresentation of miR-374a targets among genes with negative gene weights in IC 339 and (2) genes annotated with the p53 pathway are also among the most negatively weighted genes in IC 339.

IGF1 has been demonstrated to be involved in the development of OVC [29]. Considering genes that are both targets of miR-374a and annotated by the p53 pathway IGF1, IGF1 has the second most negative gene weight in IC 339. Consequently it is expected that IGF1 is downregulated in cancer. The expression of IGF1 actually tends to be downregulated in recurrent samples, shown in Supplementary Figure 2. This observation of IC 339 significantly contributing to the expression profile of IGF1 underlines the plausible

miR-374a mediated regulation of the p53 pathway. IGF1 is demonstrated to be one potential target of miR-374a.

The expression of miR-374a was recently found to be strongly linked with overall survival in the TCGA ovarian tumor set [30] (under revision in PlosONE). Furthermore, miR-374a has been reported to be upregulated in colon cancer [31] and downregulated upon cisplatin treatment in squamous cell carcinoma (SCC) [32] and in non-small cell lung cancer [33]. These findings support the strength of miICA in being able to capture miRNA regulations that has just recently emerged in the literature.

## A regulatory relationship between miR-374a and p63

Recently, there has been a lot of research on p63, a p53 family member, because its role in cancer seems to be dual and very complex [34,35]. p63 activates miR-374a transcription [32], which above was linked to the p53 pathway. We are using the MSigDB annotation database for pathway enrichment analysis. In MSigDB p63 is not annotated with the p53 pathway. However, it is known that depending on the isoform of p63 it can either activate or inhibit p53 target genes [35]. In IC 339 p63 lies among the 10% most negatively weighted genes together with p53 annotated genes as described above. This indicates an activating role of p63 on the p53 target genes. On the Agilent microarray chip the probe for p63 cannot distinguish between the different isoforms of p63. By using miICA we are able to retrieve its activating role on p53 targets.

In 2011, Huang et al., [32] reported that miR-374a and  $\Delta Np63\alpha$ , an isoform of p63, are downregulated upon cisplatin exposure of head and neck squamous cell carcinoma (HNSCC) cells. Cisplatin is a chemotherapeutic agent. We want to explore if miICA is capable of demonstrating the regulatory relationship between p63 and miR-374 and their mutual downregulation upon chemotherapy in OVC. To demonstrate this we searched for ICs where p63 has a significant gene weight ( $C_{p63,k}$ ). A gene weight of gene  $g$  in the  $k$ th IC ( $C_{gk}$ ) is assumed to be significant if  $|C_{gk}| > 2\sigma(|C_{.k}|)$  [23], where  $\sigma$  is the standard deviation and  $|C_{.k}|$  are the absolute gene weights in the  $k$ th IC.

There are four ICs (ICs 22, 102, 231 and 425) with enrichments of miR-374a targets ( $p < 0.05$ ), a significant gene weight of p63 and component weights correlating with chemotherapy ( $p < 0.05$ ). There are two clinical conditions of the TCGA data related to chemotherapy. One is chemotherapy as the primary treatment and a second is additional chemotherapy after the return of OVC after a period of remission. Since p63 activates miR-374a transcription it is expected that the gene weights of miR-374a targets and the gene weight of p63 are of opposite sign in the four ICs. Without requiring it, in all these four ICs this is actually true.

The largest ratio of  $\frac{|C_{p63,k}|}{\sigma(|C_{.k}|)}$  is found in IC 231 and therefore IC 231 is assumed to contribute most significantly to the expression profile of p63 as compared to ICs 22, 102 and 425. Figure 4C shows the component weights of IC 231 when the samples are grouped according to additional chemotherapy. IC 231 is weighted highest in patients not receiving additional chemotherapy, where IC 231 has an positive average component weight. The gene weight of p63 in IC 231 is negative and consequently p63 is upregulated upon chemotherapy. The targets of miR-374a have significantly higher gene weights in IC 231 than non targets ( $q = 0.048$ ). The positive gene weights of miR-374a targets in IC 231 results in them being downregulated upon chemotherapy. The significance for the gene weight of p63 in the four ICs together with the enrichment of miR-374a targets indicate a deregulation of miR-374a and p63 upon chemotherapy. Interestingly, the results of miICA propose for an upregulation of miR-374a and p63 upon chemotherapy treatment, which contradicts the findings of Huang et al. [32].

## miR-933 is related to the recurrence of ovarian cancer

For 15 patients there exist paired cancer and recurrent samples. Considering only these 30 samples the component weights of IC 79 are significantly higher in recurrent samples as compared to cancer samples

(paired  $t$ -test,  $p = 0.043$ ), see Fig. 3A. In fact, IC 79 has the lowest  $q$ -value ( $q = 2.1 \cdot 10^{-5}$ ) of all ICs when testing the component weights between all cancer and recurrent samples.

In the negative end of IC 79 genes annotated with pathways in cancer are overrepresented and genes annotated by stabilization of p53 are overrepresented in the positive end. Pathways in cancer is a collection of pathways that are related to cancer, e.g., apoptosis ([www.genome.jp/kegg/](http://www.genome.jp/kegg/)). When testing for the difference in miRNA expressions between recurrent and cancerous samples miR-933 is most significantly differentially expressed ( $q = 1.6 \cdot 10^{-13}$ ). The expression of miR-933 follows the pattern for the component weights of IC 79, i.e. miR-933 is upregulated in recurrent samples (see Fig. 3B and Supplementary Figure 1). To follow the rationale from above the genes targeted by miR-933 must then be significantly overrepresented among genes with negative gene weights in IC 79. If so, then a link between pathways in cancer and the regulation of miR-933 can be proposed. In fact targets of miR-933 are significantly lower weighted in IC 79 than non-targets ( $q = 0.0098$ ) and therefore we propose a link between pathways in cancer and miR-933 through the regulations of genes with negative gene weights in IC 79.

Among genes with positive gene weights in IC 79, genes annotated by oxidative phosphorylation are significantly overrepresented ( $q < 10^{-16}$ ). This is an interesting observation, because their positive gene weights propose for oxidative phosphorylation annotated genes being upregulated in recurrent samples. Furthermore, it is the most significantly overrepresented pathway in the positive end of IC 79, as pathways in cancer is the most significantly overrepresented pathway in the negative end. In most cancer cells the energy production comes from utilization of glycolysis rather than oxidative phosphorylation. From IC 79 it seems that upon recurrence of OVC then oxidative phosphorylation becomes active.

IC 332 shows the most significant enrichment of oxidative phosphorylation annotated genes ( $q < 10^{-16}$ ). They have significantly lower gene weights than genes not annotated by oxidative phosphorylation. The component weights of IC 332 correlate with whether the tumor recurrence was first observed as metastasis or locoregional, see Fig. 4F. IC 332 is weighted highest in metastatic samples and since target genes of oxidative phosphorylation have negative gene weights they will downregulated in metastatic samples as compared to loco-regional samples. Thus, oxidative phosphorylation is most active in loco-regional tumors as compared to metastatic tumors.

### miR-9 relates to the recurrence of cancer and the p53 pathway

Through regulations of NF- $\kappa$ B1, miR-9 is a known inhibitor of OVC [14]. In the TCGA data miR-9 is only significantly, differentially expressed when comparing deceased and living patients, but not when comparing cancer and normal samples. By use of miCA, it might be possible to retrieve the inhibitory effect of miR-9 in relation to the development of OVC.

To find the fingerprint of miR-9 in the TCGA data we searched for ICs with a significant gene weight of miR-9. In 34 ICs the inequality  $\frac{|C_{\text{miR-9},k}|}{\sigma(|C_{\cdot,k}|)} > 2$  is true and in these it is tested if targets of miR-9 have gene weights that are significantly lower or higher than non-targets (requiring  $q < 0.05$ ). At the same time opposing signs of gene weights for miR-9 targets and miR-9 itself are needed to ensure an inhibitory effect of miR-9 on its targets. This leaves six ICs for further investigations. Among these only one IC is differentially weighted in cancerous and normal samples ( $q = 0.026$ ). It is IC 96 and its component weights are shown in Fig. 4B. The gene weight of miR-9 in IC 96 is negative, and because the component weights of IC 96 are highest in cancer samples, miR-9 will be consequently be downregulated in OVC as observed experimentally.

Since miR-9 is an inhibitor of OVC its targets should be upregulated in cancer versus normal samples. The targets of miR-9 are in fact overrepresented in the positive end of IC 96 and will consequently be expressed in a way similar to the component weights of IC 96. That is, miR-9 targets are upregulated in cancer. Stabilization of p53 is one pathway with target genes overrepresented in the positive end of IC 96. Therefore the pathways mapped by and genes in IC 96 might be regulated by miR-9 and a link between IC 96 and destabilization of p53 in cancer is plausible.

### IC 501 maps to the mTOR pathway

The mTOR pathway is druggable [36] and thus is a relevant pathway to investigate in relation to the treatment of OVC. The mTOR pathway is namely deregulated in OVC [37] and the expression of the tumor suppressor miR-100 is related to the deregulation of the mTOR pathway in OVC [37]. We want to investigate whether miICA is capable of documenting this regulatory relationship between the mTOR pathway and miR-100 for the TCGA data. There are only two ICs where that have both an enrichment of miR-100 targets and mTOR target genes. These are ICs 79 and 501. In both ICs the gene weights of miR-100 target and mTOR target genes are significantly lower than non-target genes ( $q < 0.001$  and  $p < 0.03$ , respectively). For IC 79 this comes as no surprise, since it has already been linked to the development of OVC by the results presented above.

Since miR-100 is a tumor suppressor gene it is downregulated in cancerous samples, see Supplementary Figure 1. The component weights of IC 501 are lower in cancer samples compared to normal samples ( $p = 4.2 \cdot 10^{-4}$ ), see Fig. 4H. Since the targets of miR-100 are overrepresented in the negative end of IC 501 they will in general be upregulated in cancer, which complies with miR-100 being downregulated in cancer. Thus, miICA is capable of documenting miR-100 as a tumor suppressor and also, since miR-100 and mTOR target genes both are enriched among genes with negative gene weights in IC 501 we propose for this IC mapping to a miR-100 mediated regulation of the mTOR pathway.

### A regulatory relationship between miR-31 and the p53 pathway

In cell lines where the p53 gene is dysfunctional or absent, miR-31 stops the cells from dividing wildly and induces apoptosis [15]. We want for miICA to retrieve this tumor suppressing function of miR-31 in the TCGA data. There are 48 ICs with a significant enrichment of miR-31 targets ( $q < 0.05$ ). For these 48 ICs we seek for enrichments of targets of either the p53 pathway or stabilization of p53. In total there are 7 ICs with an enrichment for p53 and miR-31 targets ( $q < 0.05$ ), but only 3 of these ICs (ICs 53, 195 and 277) are differentially weighted between cancerous and normal samples ( $q < 0.05$ ). Targets of miR-31 should be upregulated in cancer samples, because miR-31 is the most downregulated miRNA in OVC.

The target enrichment analysis in combination with the weighting of the ICs gives that targets of miR-31 in general are upregulated according to ICs 195 and 277, which does not comply with the experimental findings. However, IC 53 is positively weighted in normal samples and is non-present in cancerous samples ( $q = 0.0058$ , see Fig. 4A). For the miR-31 targets to be upregulated in cancerous samples, they need to be enriched in the negative end of IC 53, which is actually the fact ( $q = 0.020$ ). Furthermore, in the positive end of IC 53 targets for the stabilization of p53 is enriched ( $q = 0.035$ ) meaning that they in general are downregulated in cancerous samples. This complies with the p53 being destabilized in OVC. All this proposes for a connection of miR-31 to the stabilization of p53 through the gene regulatory pathway mapped by IC 53.

### Regulations of the Wnt pathway in ovarian cancer

There are two ICs (277 and 309) where gene weights for genes annotated by the Wnt pathway are enriched in the positive ends of the ICs ( $q = 0.015$  and  $q = 0.030$ , respectively). The component weights of IC 277 are highest and positive in normal samples ( $q < 10^{-16}$ ). For IC 309 the component weights are positive in cancerous samples and negative in normal samples ( $q = 0.022$ ), see Fig. 4(D,E). Thus the component weights of ICs 277 and 309 anti-correlate, even though the Wnt pathway is enriched in the positive end of both ICs. These two ICs may correspond to two different regulatory mechanisms of Wnt in OVC.

There are two known miRNAs targeting the Wnt pathway miR-29c [38] miR-130a,b [39]. miR-130a,b have targets significantly enriched with positive gene weights in IC 277 ( $q = 2.2 \cdot 10^{-5}$ ) and miR-29c have targets enriched with positive gene weights in IC 309 ( $q = 0.0011$ ). This complies with the Wnt



pathway also being enriched with positive gene weights in these two ICs. Recently, it has been shown that dickkopf-1 (Dkk1) an inhibitor of Wnt signalling [40, 41] is frequently overexpressed in OVC [42]. Dkk1 have negative and positive gene weights in ICs 277 and 309. Thus both these ICs contributes to the Dkk1 expression profiles with an upregulation of Dkk1 in cancer. The fact that targets of known Wnt related miRNAs are enriched in ICs 277 and 309 and the gene weights of Dkk1 in these two ICs complies with its inhibitory effect on the Wnt pathway underlines the mapping of these two ICs to the Wnt pathway.

## Congruence between the results of Agilent and Affymetrix mRNAs

When applying an enrichment analysis for overrepresentation of miRNA targets and pathway annotations on the component matrix obtained from the Affymetrix mRNA similar results are obtained as for the Agilent mRNA.

- In IC 512 the p53 pathway and targets of miR-374a,b are overrepresented with positive gene weights. The component weights of IC 512 are negative in cancerous samples and positive in normal samples ( $p < 0.005$ , see Fig. 5C). This is the opposite as for the component weights of IC 339 and the expression of miR-374a. In IC 339 the p53 pathway and targets of miR-374a were enriched with negative gene weights, where for IC 512 the enrichment is for positive gene weights. Thus, IC 512 maps to the p53 pathway in the same way as IC 339 mapped to the p53 pathway for the Agilent mRNAs.
- In ICs 355 and 516 there are enrichments of miR-100 targets ( $p = 0.048$  and  $p = 0.0074$ , respectively) and the mTOR pathway ( $p = 0.0075$  and  $p = 0.029$ , respectively) with negative gene weights. They both have negative component weights in all samples, but most negatively in cancerous samples ( $p = 2.0 \cdot 10^{-6}$  and  $p = 0.015$ , respectively, see Fig. 5A,D). Thus targets of miR-100 and the mTOR pathway are upregulated in cancerous samples. This complies with the fact that miR-100 is a tumor suppressor and is linked to the mTOR pathway as was observed for IC 501 regarding the Agilent mRNAs.
- The Wnt pathway was represented by two ICs for the Agilent mRNAs. For the Affymetrix mRNAs IC 385 have enrichments with positive gene weights of both miR-29c ( $p = 0.045$ ) and miR-130a,b ( $p = 0.019$ ) targets together with an enrichment of the Wnt pathway ( $p = 0.010$ ). IC 385 is almost non-present in normal samples and weighted positively in cancerous samples ( $p = 0.027$ , see Fig. 5B). These findings correspond to the finding of IC 309 mapping to the Wnt pathway for the Agilent mRNAs.

We are therefore able to retrieve the findings from our test set in our validation set and consequently we state that the results of miICA are reliable.

## Independent components mapping to miRNA regulations

As mentioned in the beginning it might not only be biological pathways but also genes highly regulated by miRNAs, which ICA identifies. IC 79 and IC 501 studied above could be such two ICs. They show enrichment of targets for 99% and 96% of the miRNAs and the pathways enriched include several cancer related pathways.

Thus from the results of miICA it is possible to retrieve genetic links between miRNA and mRNA regulations in relation to OVC, and we believe that some ICs map to miRNA regulated pathways.

## Methods

### Preprocessing of the mRNA and miRNA expression microarray data

The mRNA and miRNA expressions in OVC were downloaded from the TCGA database (<http://tcga-data.nci.nih.gov/tcga/>). The mRNA expressions profiled using the Agilent array 244K were downloaded as level 3 data, i.e. as a quantile normalized expression matrix, the mRNA expressions profiled using the Affymetrix U133A array were downloaded as level 1, i.e. as raw CEL-files, and the miRNA expressions profiled using the Agilent miRNA (v2) 8x15K array were downloaded as raw image files (level 1).

All the CEL-files were preprocessed using the RMA package [43] in Bioconductor [44] with remapped Ensembl build 50 gene probesets [45]. After quantile normalization of the mRNA expression matrix, missing entries were estimated using a nearest neighbor averaging function with the ten nearest neighbors. Only mRNAs with available 3'UTR sequences are included in the expression matrices for both Agilent and Affymetrix mRNAs. To select these the biomaRt package (vs2.8.1) in R (v.2.13) was used. The Agilent mRNAs were annotated with Agilent IDs and hgnc IDs were used to retrieve the 3UTR sequences. The PubMed gene repository ([www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)) was used to map between the different annotations. After exclusion of mRNAs without a 3'UTR sequence the Agilent mRNA expression matrix contains 15,466 genes and 537 samples and the Affymetrix mRNA expression matrix contains 11,869 genes and 584 samples.

The raw image files for the miRNA expressions were read using the AgiMicroRna package [46] in R. After loading the image files the quality of the array data were checked by looking at the MA and density plots for each array. Within each array a robust multichip analysis is performed to normalize the expression values. The log-transformed of the total gene signal is centered on the modes of the density distribution for each array. Then the standard deviation versus mean trend is removed before miRNAs having invariant expressions across the samples are found [47]. After this normalization, the miRNA expressions were filtered to only retain the miRNAs with a certain quality, e.g., the miRNA must be detected in more than 20% of the samples within a group of samples. In total, it leaves 510 miRNAs for further investigations. The mature sequences of these miRNAs were downloaded from miRBase (v18.0) ([www.mirbase.org/](http://www.mirbase.org/)).

When performing a microarray experiment for the expressions of miRNAs, it is the levels of both the active and precursor forms that are measured [48]. It is not possible to distinguish the mature sequence of a miRNA from its pre-miRNA sequence on the microarray. Therefore the miRNA expression levels from the microarray are only used quantitatively. Here we have set our results for the miRNA expressions against expression profiles for miRNAs already published. Thus, we are not only relying our findings on the miRNA microarray data.

### ICA applied on the expression matrices

The maximum number  $N = \min(S, G)$  of ICs is estimated to preserve the full information hidden in the data [49, 50]. fastICA [51] is one implementation of ICA in R and it was used in this analysis. It is an iterative algorithm based on the maximization of negentropy and it returns slightly different estimates of  $C$  and  $M$  for each application on the expression matrices. To meet this, fastICA is applied 20 times on the mRNA expression matrices and subsequently an average component matrix is calculated by taking the mean of the most correlating ICs from each of the 20 runs. Knowing the average component matrix, an average mixing matrix is easily calculated by the numerical solution to  $E = CM \Rightarrow M = C^{-1}E$ . For a detailed description of the fastICA algorithm see Ref. [51].

The miRNA expressions are higher compared to the Agilent mRNA levels. Thus a simple combination of the mRNA and miRNA expression matrices would cause the miRNAs to show up in only one IC when applying ICA to the combined matrices. ICA could in principle be applied on the two expression matrices

separately, but there are fewer miRNAs than samples leading to a problem of over-learning [52]. A way to apply ICA on the miRNA expression matrix  $E_{\text{miRNA}}$  is to use the mixing matrix for the mRNAs  $M_{\text{mRNA}}$  and then estimate a component matrix for the miRNAs  $C_{\text{miRNA}}$  by solving  $E_{\text{miRNA}} = C_{\text{miRNA}}M_{\text{mRNA}}$ . This is only possible for samples with both miRNA and Agilent mRNA expressions profiles. There are 584/537 samples with expression profiles for both miRNAs and Agilent/Affymetrix mRNAs.

## Enrichment analyses

From the 3'UTR sequences a target matrix (**T**) is generated with miRNAs in the columns and genes in the rows. The entries of the target matrix are the numbers of matches for the seed of a given miRNA in the 3'UTR of a given gene, i.e.  $T_{m,g} = 5$  corresponds to miRNA  $m$  having five seed matches in the 3'UTR of gene  $g$ . A Wilcoxon rank sum test is then applied to test if the gene weights of genes with targets are either significantly smaller or significantly larger than the gene weights of genes without targets for a specific miRNA. The enrichment test is applied on all ICs and the p-values were FDR (false discovery rate) corrected for multiple testing.

The same enrichment analysis was performed for the pathway annotation of the mRNA. Here we used the pathway annotation from the Molecular Signatures Database v3.0, Broad Institute ([www.broadinstitute.org/gsea/msigdb](http://www.broadinstitute.org/gsea/msigdb)). For a specific pathway, a Wilcoxon rank sum test was applied to test if genes annotated with that pathway have significantly low/high gene weights compared to genes not annotated by that pathway. The p-values were FDR corrected, as above, for multiple testing.

## Acknowledgments

This work is supported by the Danish National Research Foundation.

## References

1. Agarwal R, Kaye SB (2003) Ovarian cancer: strategies for overcoming resistance to chemotherapy. *Nat Rev Cancer* 3: 502–516.
2. Ozols RF (1992) Advances in the chemotherapy of gynecologic malignancies. *Hematol Oncol* 10: 43–51.
3. Lee RC, Feinbaum RL, Ambros V (1993) The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* 75: 843–854.
4. Friedman RC, Farh KKH, Burge CB, Bartel DP (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* 19: 92–105.
5. Bartel D (2009) MicroRNAs: target recognition and regulatory functions. *Cell* 136.
6. Cancer Genome Atlas Research Network (2011) Integrated genomic analyses of ovarian carcinoma. *Nature* 474: 609–615.
7. Cho WCS (2010) MicroRNAs in cancer - from research to therapy. *Biochim Biophys Acta* 1805: 209–217.
8. Lynn FC, Skewes-Cox P, Kosaka Y, McManus MT, Harfe BD, et al. (2007) MicroRNA expression is required for pancreatic islet cell genesis in the mouse. *Diabetes* 56: 2938–2945.

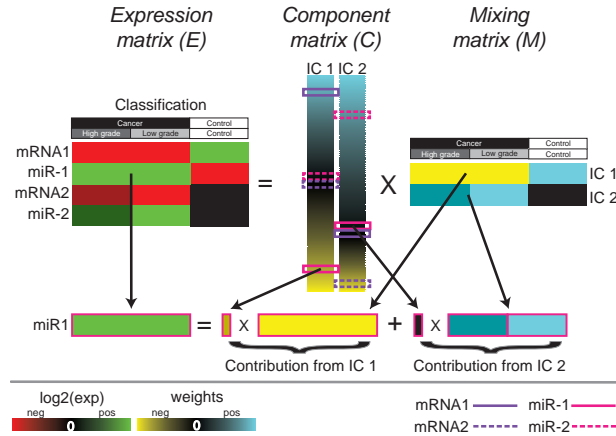


9. Bang-Berthelsen CH, Pedersen L, Fløyel T, Hagedorn PH, Gylvin T, et al. (2011) Independent component and pathway-based analysis of miRNA-regulated gene expression in a model of type 1 diabetes. *BMC Genomics* 12: 97.
10. Rottiers V, Najafi-Shoushtari SH, Kristo F, Gurumurthy S, Zhong L, et al. (2011) MicroRNAs in Metabolism and Metabolic Diseases. *Cold Spring Harb Symp Quant Biol* .
11. Martin K, O’Sullivan JF, Caplice NM (2011) New therapeutic potential of microRNA treatment to target vulnerable atherosclerotic lesions and plaque rupture. *Curr Opin Cardiol* 26: 569–575.
12. Fichtlscherer S, Zeiher AM, Dimmeler S (2011) Circulating microRNAs: biomarkers or mediators of cardiovascular diseases? *Arterioscler Thromb Vasc Biol* 31: 2383–2390.
13. Cai X, Schäfer A, Lu S, Bilello JP, Desrosiers RC, et al. (2006) Epstein-Barr virus microRNAs are evolutionarily conserved and differentially expressed. *PLoS Pathog* 2: e23.
14. Guo LM, Pu Y, Han Z, Liu T, Li YX, et al. (2009) MicroRNA-9 inhibits ovarian cancer cell growth through regulation of NF-kappaB1. *FEBS J* 276: 5537–5546.
15. Creighton CJ, Fountain MD, Yu Z, Nagaraja AK, Zhu H, et al. (2010) Molecular Profiling Uncovers a p53-Associated Role for MicroRNA-31 in Inhibiting the Proliferation of Serous Ovarian Carcinomas and Other Cancers. *Cancer Res* 70: 1906–1915.
16. Saito T, Saetrom P (2010) MicroRNAs–targeting and target prediction. *N Biotechnol* 27: 243–249.
17. Min H, Yoon S (2010) Got target? Computational methods for microRNA target prediction and their extension. *Exp Mol Med* 42: 233–244.
18. Stanhope SA, Sengupta S, den Boon J, Ahlquist P, Newton MA (2009) Statistical use of argonaute expression and RISC assembly in microRNA target identification. *PLoS Comput Biol* 5: e1000516.
19. Comon P (1994) Independent component analysis, A new concept? *Signal Processing* 36: 287–314.
20. Lee SI, Batzoglou S (2003) Application of independent component analysis to microarrays. *Genome Biol* 4: R76.
21. Teschendorff AE, Wang Y, Barbosa-Morais NL, Brenton JD, Caldas C (2005) A variational Bayesian mixture modelling framework for cluster analysis of gene-expression data. *Bioinformatics* 21: 3025–3033.
22. Carpentier AS, Riva A, Tisseur P, Didier G, Hénaut A (2004) The operons, a criterion to compare the reliability of transcriptome analysis tools: ICA is more reliable than ANOVA, PLS and PCA. *Comput Biol Chem* 28: 3–10.
23. Liebermeister W (2002) Linear modes of gene expression determined by independent component analysis. *Bioinformatics* 18: 51–60.
24. Frigyesi A, Veerla S, Lindgren D, Höglund M (2006) Independent component analysis reveals new and biologically significant structures in micro array data. *BMC Bioinformatics* 7: 290.
25. Kong W, Mou X, Hu X (2011) Exploring matrix factorization techniques for significant genes identification of Alzheimer’s disease microarray gene expression data. *BMC Bioinformatics* 12 Suppl 5: S7.
26. van Dongen S, Abreu-Goodger C, Enright AJ (2008) Detecting microRNA binding and siRNA off-target effects from expression data. *Nat Meth* 5: 1023–1025.

27. Huang JC, Morris QD, Frey BJ (2007) Bayesian inference of MicroRNA targets from sequence and expression data. *J Comput Biol* 14: 550–563.
28. Toledo F, Wahl GM (2006) Regulating the p53 pathway: in vitro hypotheses, in vivo veritas. *Nat Rev Cancer* 6: 909–923.
29. Terry KL, Tworoger SS, Gates MA, Cramer DW, Hankinson SE (2009) Common genetic variation in IGF1, IGFBP1 and IGFBP3 and ovarian cancer risk. *Carcinogenesis* 30: 2042–2046.
30. Creighton CJ, Hernandez-Herrera A, Jacobsen A, Levine DA, Mankoo P, et al. (2012) Integrated analyses of microRNAs uncover tumor suppressor candidates in high-grade serous ovarian carcinoma. Under revision in *PLoS ONE*.
31. Wang YX, Zhang XY, Zhang BF, Yang CQ, Chen XM, et al. (2010) Initial study of microRNA expression profiles of colonic cancer without lymph node metastasis. *J Dig Dis* 11: 50–54.
32. Huang Y, Chuang A, Hao H, Talbot C, Sen T, et al. (2011) Phospho-DNp63a is a key regulator of the cisplatin-induced microRNAome in cancer cells. *Cell Death Differ* 18: 1220–1230.
33. Vösa U, Voeder T, Kolde R, Fischer K, Völk K, et al. (2011) Identification of miR-374a as a prognostic marker for survival in patients with early-stage nonsmall cell lung cancer. *Genes Chromosomes Cancer* 50: 812–822.
34. Flores ER (2007) The roles of p63 in cancer. *Cell Cycle* 6: 300–304.
35. Yang A, Kaghad M, Wang Y, Gillett E, Fleming MD, et al. (1998) p63, a p53 homolog at 3q27-29, encodes multiple products with transactivating, death-inducing, and dominant-negative activities. *Mol Cell* 2: 305–316.
36. Hsieh AC, Costa M, Zollo O, Davis C, Feldman ME, et al. (2010) Genetic dissection of the oncogenic mTOR pathway reveals druggable addiction to translational control via 4EBP-eIF4E. *Cancer Cell* 17: 249–261.
37. Nagaraja AK, Creighton CJ, Yu Z, Zhu H, Gunaratne PH, et al. (2010) A link between mir-100 and FRAP1/mTOR in clear cell ovarian cancer. *Mol Endocrinol* 24: 447–463.
38. Hawkins SM, Creighton CJ, Han DY, Zariff A, Anderson ML, et al. (2011) Functional MicroRNA Involved in Endometriosis. *Mol Endocrinol* 25: 821–832.
39. Sorrentino A, Liu Cg, Addario A, Peschle C, Scambia G, et al. (2008) Role of microRNAs in drug-resistant ovarian cancer cells. *Gynecologic Oncology* 111: 478–486.
40. Seménov MV, Tamai K, Brott BK, Kühl M, Sokol S, et al. (2001) Head inducer Dickkopf-1 is a ligand for Wnt coreceptor LRP6. *Curr Biol* 11: 951–961.
41. Bafico A, Liu G, Yaniv A, Gazit A, Aaronson S (2001) Novel mechanism of Wnt signalling inhibition mediated by Dickkopf-1 interaction with LRP6/Arrow. *Nature cell biology* 3: 683–686.
42. Wang S, Zhang S (2011) Dickkopf-1 is frequently overexpressed in ovarian serous carcinoma and involved in tumor invasion. *Clin Exp Metastasis* 28: 581–591.
43. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, et al. (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* 31: e15.
44. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 5: R80.

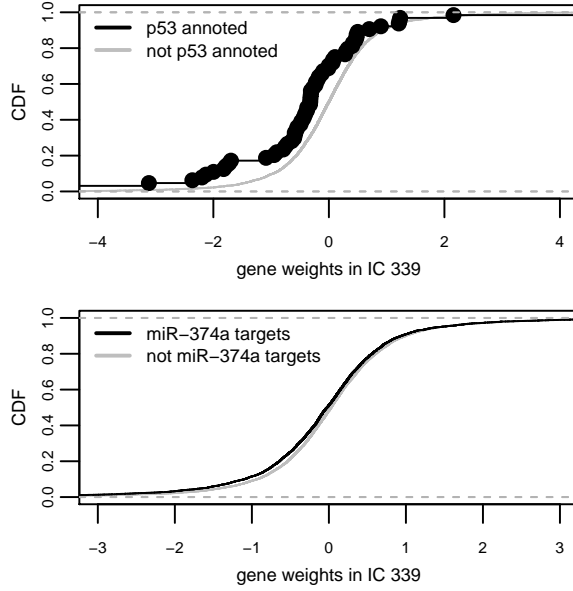
45. Dai M, Wang P, Boyd AD, Kostov G, Athey B, et al. (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res* 33: e175.
46. López-Romero P (2011) Pre-processing and differential expression analysis of Agilent microRNA arrays using the AgiMicroRna Bioconductor library. *BMC Genomics* 12: 64.
47. Pradervand S, Weber J, Thomas J, Bueno M, Wirapati P, et al. (2009) Impact of normalization on miRNA microarray expression profiling. *RNA* 15: 493–501.
48. Cheng C, Li LM (2008) Inferring microRNA activities by combining gene expression with microRNA target prediction. *PLoS ONE* 3: e1989.
49. Chang WC (1983) On Using Principal Components Before Separating a Mixture of Two Multivariate Normal Distributions. *Appl Statist* 32: 267–275.
50. Yeung KY, Ruzzo WL (2001) Principal component analysis for clustering gene expression data. *Bioinformatics* 17: 763–774.
51. Hyvärinen A, Karhunen J, Oja E (2001) Independent component analysis. Adaptive and learning systems for signal processing, communications, and control. J. Wiley. URL <http://books.google.dk/books?id=96D0ypDwAkkC>.
52. Särelä J, Vigário R (2001) Overlearning in Marginal Distribution-Based ICA: Analysis and Solutions. *J Comput Biol* 8: 557–569.
53. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E (2007) The role of site accessibility in microRNA target recognition. *Nat Genet* 39: 1278–1284.
54. Heikham R, Shankar R (2010) Flanking region sequence information to refine microRNA target predictions. *J Biosci* 35: 105–118.
55. Wen J, Parker BJ, Jacobsen A, Krogh A (2011) MicroRNA transfection and AGO-bound CLIP-seq data sets reveal distinct determinants of miRNA action. *RNA* 17: 820–834.

## Figure Legends

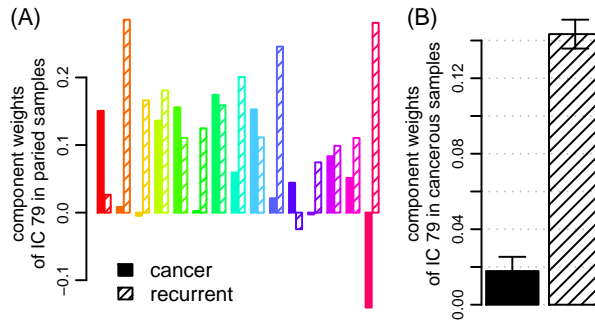


**Figure 1.** Schematic illustration for the application of ICA on an expression matrix.  $E$  is decomposed into two ICs and a mixing matrix ( $M$ ). mRNAs and miRNAs are weighted by gene weights in ICs 1 and 2 (purple and magenta squares). IC 1 and 2 are weighted in the samples according to their component weights in the rows of  $M$ . mRNA1 and miRNA1 are down/upregulated in cancer, therefore the gene weights of these are very high/low in IC 1, which is weighted negatively in cancer and positively in normal samples. mRNA2 and miRNA2 are down- and upregulated from high to low grade tumors, therefore the gene weights are very low/high in IC 2, which is weighted most positively in low grade tumors. The gene weights of miRNAs and mRNAs lies oppositely, since the expressions of these anti-correlate.

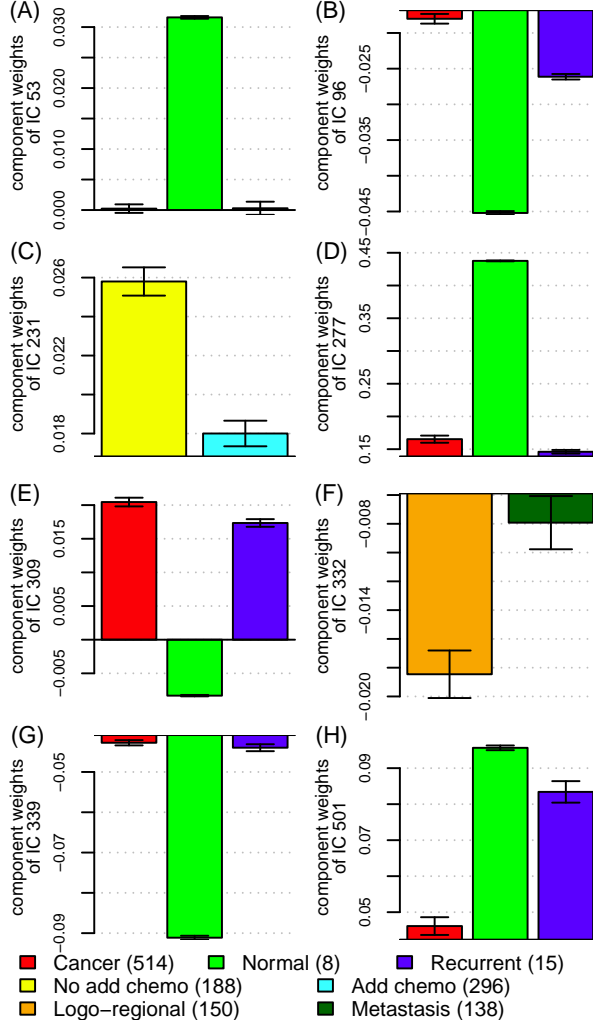
Logarithmic expression: red-green. Component and gene weights: yellow-cyan. Black/white: cancer/normal, grayscale: grade of tumor.



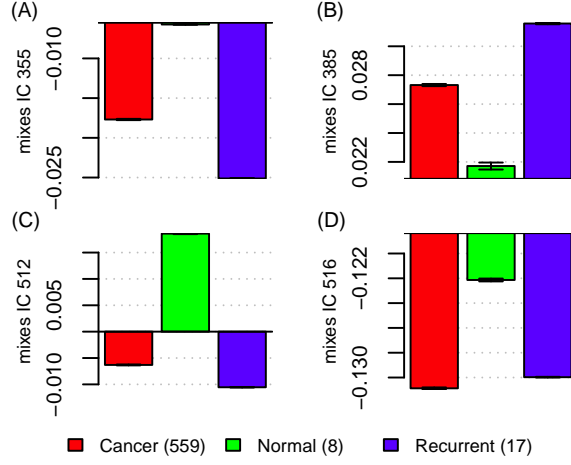
**Figure 2.** The cumulative distribution function (CDF) of gene weights in IC 339. Top: genes with (black) and without (grey) an annotation for the p53 pathway. Bottom: genes with (black) and without (grey) a seed match for miR-374a.



**Figure 3.** Component weights of IC 79 are highest in recurrent samples. (A) For 15 samples with both a cancer and recurrent profile the component weights of IC 79 are plotted. A paired t-test on the 15 paired samples gives a  $p$ -value = 0.043. (B) Barplot for the component weights of IC 79 in all cancer and recurrent samples. Bars are mean values and lines are  $\pm\sigma^2$ . Filled bars: cancer, striped bars: recurrence.



**Figure 4.** The mean of component weights in different ICs and for different classification schemes. The samples are grouped based on disease state (A,B,D,E,G-H), additional chemotherapy (C), and site of first tumor recurrence (F). In all panels the component weights significantly correlates with the classification of samples with  $p$ -values  $< 0.05$ . The numbers in parenthesis are the number of samples with a given classification. The bars are the mean values and the lines are  $\pm\sigma^2$ .



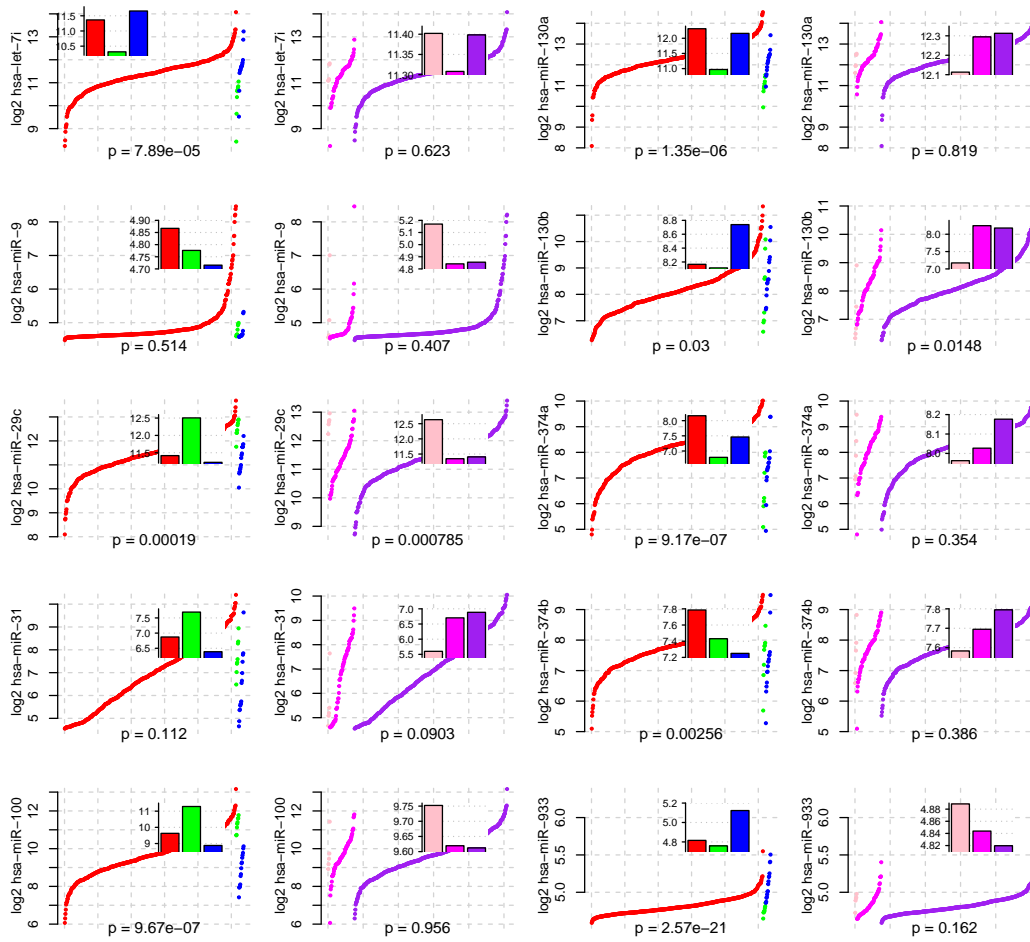
**Figure 5.** The mean weighting of ICs 385, 512 and 516 in cancer (red), normal (green) and recurrent (blue) samples. In all panels the component weights significantly correlates with the classification of samples with  $p$ -values  $< 0.05$ . In parenthesis are given the number of cancer, normal and recurrent samples for the Affymetrix mRNA profiles. The bars are the mean values and the lines are  $\pm\sigma^2$ .

## Tables

**Table 1. Features commonly included in miRNA target prediction methods**

Feature	Description
Seed match:	The formation of miRNA:mRNA duplexes by complementary base pairing.
Favourable seeds:	The secondary structure of the mRNA [53] and the flanking region around the seed sites [54, 55] make some seed sites more favourable than others.
Gene expressions:	A negative correlation between the expression of the miRNA and its target is expected due to the inhibitory effect of the miRNA.
Conservation:	The seed region of miRNAs and their targets are well conserved among related species [4].
Cooperativity:	Examine whether one or more seed sites resides within the miRNA targets.

## Supplementary Figure 1





## Supplementary Figure 2

